jet INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2024, VOL. 70, NO. 2, PP. 341–348
Manuscript received February 14, 2024; revised June, 2024. doi: 10.24425/ijet.2024.149550

Procedurally generated AI compound media for expanding audial creations, broadening immersion and perception experience

Grzegorz Samson

***Abstract*—Recently, the world has been gaining vastly increasing access to more and more advanced artificial intelligence tools. This phenomenon does not bypass the world of sound and visual art, and both of these worlds can benefit in ways yet unexplored, drawing them closer to one another. Recent breakthroughs open possibilities to utilize AI driven tools for creating generative art and using it as a compound of other multimedia. The aim of this paper is to present an original concept of using AI to create a visual compound material to existing audio source. This is a way of broadening accessibility thus appealing to different human senses using source media, expanding its initial form. This research utilizes a novel method of enhancing fundamental material consisting of text audio or text source (script) and sound layer (audio play) by adding an extra layer of multimedia experience – a visual one, generated procedurally. A set of images generated by AI tools, creating a story-telling animation as a new way to immerse into the experience of sound perception and focus on the initial audial material. The main idea of the paper consists of creating a pipeline, form of a blueprint for the process of procedural image generation based on the source context (audial or textual) transformed into text prompts and providing tools to automate it by programming a set of code instructions. This process allows creation of coherent and cohesive (to a certain extent) visual cues accompanying audial experience levering it to multimodal piece of art. Using nowadays technologies, creators can enhance audial forms procedurally, providing them with visual context. The paper refers to current possibilities, use cases, limitations and biases giving presented tools and solutions.**

***Keywords*—procedural generation; generative media; multimodal art; audiovisual perception; text-to-image; transformers; large language models; latent diffusion models**

I. Introduction

THE advent of transformer networks has opened up new avenues for processing semantic constructs with a high degree of abstraction and transferring them across various domains and media [1]. Information encoded through text can now be transformed and mapped onto a different representation channel, such as the visual domain [2], [3]. This presents new frontiers for exploring the creation of transgressive multimodal messages, where an additional channel is built upon the source context, extending the content into a multimodal spectrum of representation [4].

A. Historical Context

In the past, individuals experimented with creating visual effects for existing audio material [5], [6]. This could be done either by preparing a visual message to play in parallel with the sound or by algorithmically generating a visual display that was more or less closely related to the audio message. The approach of preparing a visual message allowed for strong anchoring in the audio context, enabling semantic, high-contextual mapping of information [4]. However, this method was time-consuming and required human involvement throughout the process.

B. Algorithmic Generation

The algorithmic generation approach allowed for the creation of visualizations based on the physical features of the sound, represented through vision-generating algorithms [7]. This approach enabled the creation of visual context in real-time but resulted in the loss of the ability to represent semantic context. Without the use of deep learning models, we are unable to effectively process information with a high level of abstraction and context [8].

C. Current Possibilities

The recent availability of transformer models allows for new ways of processing and contextual representation of messages [1], [9]–[12]. Mapping audio material in the visual domain has gained new levels of exploration, especially for layers containing verbal messages.

II. Related Work

The subject of media creation using generative techniques has been recently extensively explored in academic research. There is a notable focus on generating coherent narratives in the textual domain using various deep learning architectures [13], as well as visualizing content in graphical form. Text-based story visualization has seen multiple attempts and studies, particularly employing Generative Adversarial Networks (GANs) [14]–[17]. Latent Diffusion Models (LDMs) offer even more promising results in the context of high-quality visualization [2], [13], [18], [19].

The issue of visually mapping sound semantics is a separate area. Semantic visual segmentation has been an object of

Grzegorz Samson is with the Feliks Nowowiejski Academy of Music in Bydgoszcz, Poland (e-mail: g.samson@amfn.pl).

research interest [20]–[22] but adaptation of audio to visual using deep learning transformers is yet underexplored area [23].

III. Theory

A. Main Assumptions

In this discussion of audial works based on semantic context such as audio plays, it is assumed that the source fundaments consist of textual material as a framework adapted into the audio domain. This textual message can be mapped from the text level to the visual level, serving as an additional and extending mode of communication and information representation from the perspective of audiovisual integration [4]. The main paradigms concerning the visualization of audio content in this research are:

- Mapping the semantic context (key visual elements),
- Mapping the aura, emotional context within audio.

B. Mapping Semantic Context

By processing information represented in textual form, Large Language Models (LLMs) [24] are capable of transforming it into a recognised semantic context, which can then be converted into a text prompt serving as input for text-to-image diffusion models [1]–[3]. This provides the opportunity to visualize concepts that may have a direct graphic representation [13].

C. Mapping Aura

Audiovisual context can trigger emotional response in art recipients [6], [25]. Visual context can effectively complement the auditory context, enhancing emotionally charging its message to significantly appeal to the human senses. Current image processing methods that use convolutional networks enable the identification and representation of elements with emotional values [10], [26]. LLMs are capable of processing information, generalizing, and transferring it to the abstract plane of emotions, and then proposing a method of visual mapping similarly to the mapping of other semantic context [27], [28].

D. Proposed Methods

Audial artistic forms incorporating semantic representation, such as audio plays, can benefit from various innovative possibilities for visual engagement, creating a visual cue for the recipient. For generation approaches, example methods may include:

- **Aura Panel Method**: Aura panels are generated to loosely convey the context-related mood, emotional context of a given scene.
- **Scenery Method**: Scenery panels are generated to convey the character of the place where the sound is located, immersing the listener in it. The place remains as an abstract visual cue, not exact adaptation.
- **Context Isolation Method**: Separate context panels are generated to represent the current context in isolation from the sequence of events and their interdependencies. This method focuses on presenting a key element that serves as a temporary visual cue, representing an idea.
- **Storyboard Method**: Storyboard panels are generated to represent distinct subsequent elements of the story. This allows for the representation of the entire palette of constituent sound elements, providing visual context.
- **Visual Novel Method**: Component elements are generated, which, when combined, create panels with a high level of customization. This is a set of assets that allows for the composition of visual context.

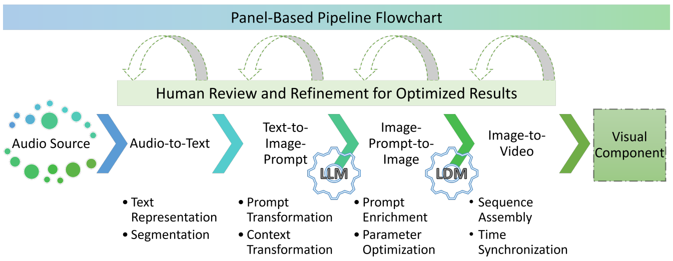


Fig. 1. Panel-Based Pipeline Conceptual Flowchart.

E. Abstract-Oriented Component Media Perception

As the level of complexity and context mapping increases, so do the expectations of the audience. The audience is capable of accepting the contextual conventionality of abstract visual projections to the abstract nature of music as there are many forms of visual representation of sound [5], [6]. The situation becomes problematic when a visualization of a context that has a tangible representation in reality appears [17].

F. Semantic-Oriented Component Media Perception

With advancements in technology, audiences are becoming accustomed to higher quality composite media [29]. The initial awe for the possibilities of creating procedural composite media may be replaced by expectations for their adequacy, coherence, and consistency. The admiration initially sparked by the discovery of such a level of accuracy in semantic representation evolves into a pursuit for further precision, detail, and coherence [13], [15], [16]. This evolution drives the continuous development and refinement of new technologies and underscores the objective of this work: to enhance and perfect the procedural generation process.

IV. Automated Pipeline Concept

A. Retrieving Semantic Context

Leveraging speech-to-text technologies, audio content can be transmuted into textual representations, thereby enabling semantic analysis and subsequent visual generation based on the interpreted data. This transformation essentially converts an audio-encoded signal into a text-based one, as transcription. While contemporary technologies can identify individual speakers, they fall short in capturing non-verbal cues that carry semantic meaning for humans, such as intonation, timbre, and phrasing, as well as ancillary auditory elements like background noise, sound events, and music. Deep learning techniques for speech processing are offering broader possibilities of speech recognition [30] and give promising results in speech-to-image transformation [31].

B. *Using Intermediary Semantic Source*

To maximize the contextual input for automating the generation of visual content, an intermediary textual medium is advisable. In the case of audio plays, the script serves this purpose, offering supplementary details about scenes, characters, and narrative elements that are conveyed in the audio format. This enriched context enhances the effectiveness of tools designed for visual media generation.

As for the conceptual process of creating visual components dependent on the audio medium, the following transformation approach is proposed and expanded as shown in Figure 1:

- Audial Medium
- Textual Medium (Intermediary Medium)
- Visual Medium

To process the textual medium into a visual medium, LDMs can be used, for which the textual medium can be prepared by a LLMs. This enables accurate representation of the semantic context originally present in the audio domain.

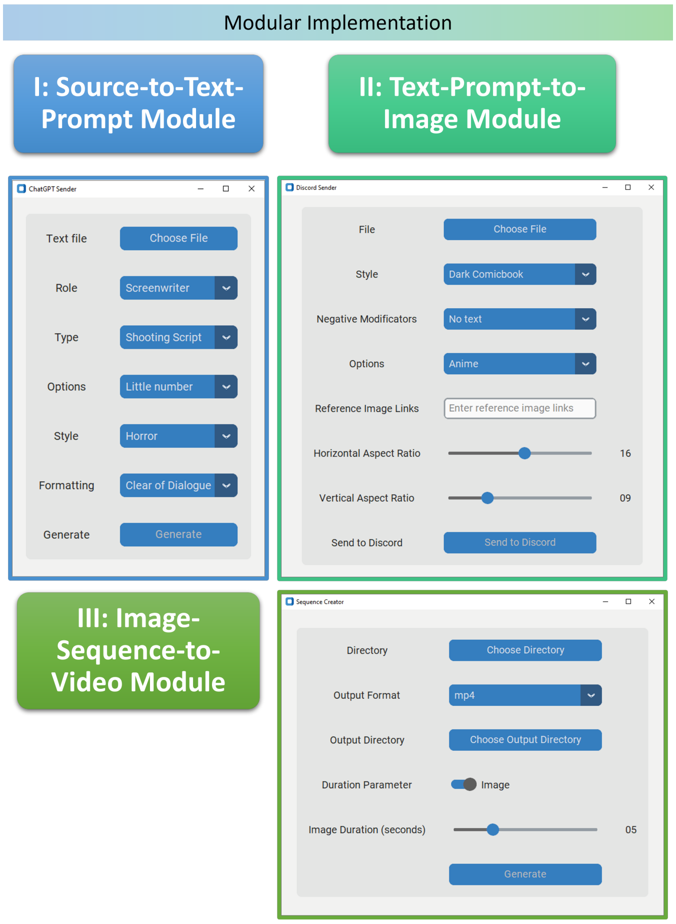
V. Pipeline Implementation

The implementation of the pipeline for procedurally generating images within the mentioned methods assumes modular automation. Consumer-grade technologies were chosen due to market standards and processing quality. The pipeline consists of three main modules:

- **Source-to-Text-Prompt Module**: This initial module processes the source material into text prompts using Language Learning Model. For this purpose, the ChatGPT API, specifically the **'gpt-3.5-turbo'** model (released on 01.03.2023) by OpenAI, is utilized.
- **Text-Prompt-to-Image Module**: The subsequent module converts the text prompts into graphics using Latent Diffusion Model. The Midjourney **'v5-2'** model (released on 22.06.2023) is employed for this task.
- **Image-Sequence-to-Video Module**: This concluding module provides rudimentary image-to-video conversion capabilities, focusing on the generation of frame sequences. It leverages Python libraries like OpenCV (Open Source Computer Vision Library) to accomplish this task.

A. *Automation and Codebase*

The automation of tool handling was implemented in Python, using the Model-View-ViewModel (MVVM) framework for the Graphical User Interface (GUI) as shown in Figure 2. The source code for the project is available at the following Open Science Framework repository link: https://osf.io/se8y6/?view_only=45a4e52c9dd44a0da12d52ea19302c8a


Modular Implementation
I: Source-to-Text-Prompt Module
II: Text-Prompt-to-Image Module
III: Image-Sequence-to-Video Module

ChatGPT Sender

Text file: Choose File
Role: Screenwriter
Type: Shooting Script
Options: Little number
Style: Horror
Formatting: Clear of Dialogue
Generate: Generate

Discord Sender

File: Choose File
Style: Dark Comicbook
Negative Modificators: No text
Options: Anime
Reference Image Links: Enter reference image links
Horizontal Aspect Ratio: 16
Vertical Aspect Ratio: 09
Send to Discord: Send to Discord

Sequence Creator

Directory: Choose Directory
Output Format: mp4
Output Directory: Choose Output Directory
Duration Parameter: Image
Image Duration (seconds): 05
Generate

Fig. 2. Modules Graphical User Interface with parameter configuration.

B. *Overview*

The presented solution is functional; however, difficulties in maintaining coherence at higher levels of semantic mapping have been described and are discussed separately in detail in the authors' Engineering Thesis [32]. The introduced pipeline procedurally generates panels that can be considered as materials or components. These components can be subjected to further fine-tuning.

VI. Pipeline Analisys

A. *Autonomous and Automated Generation*

While it is entirely possible to achieve complete automation in the realm of procedural generation, this might come at the expense of quality and fidelity to the original source material. Fully autonomous systems may offer a degree of engagement, but could fail to accurately represent the nuances of the source [16], [17], [33]. A more balanced approach might involve treating the generated elements as prefabricated building components. This allows for better alignment with the original sound and a more coherent integration of the generated visuals with the source material.

B. Human Involvement

At this stage of technological development of the procedural media generation process, perfection involves increasing the input of the human factor. The human factor is important in the process of supervision, material correction during text prompt and images generation, selection of appropriate panels, and their time synchronisation. The quality of the output material vary, depending on the human intervention in the process [33].

C. Types of Coherence

In evaluating the appropriateness of generated materials as image panels in relation to the source context, two types of coherence can be distinguished: vertical and horizontal. These types differ primarily in their orientation to time axes and complexity.

- **Vertical Coherence**: This form of coherence is relatively straightforward and focuses on the representation of specific content or information, isolated from its broader context [3]. The primary concern is temporal consistency at the immediate moment. For example, if a scene in the audio layer involves human interaction, a corresponding visual representation of a human would be displayed. In cases of low-level vertical coherence, distinctive features are not crucial. Different visual representations may be used for the same individual or concept, allowing for various depicting variants.
- **Horizontal Coherence**: In this type, the focus is on maintaining consistent distinctive features across different panels over time [15], [16]. This ensures the continuity of the representation of a given person or object by maintaining closely related characteristics that define its identity. The same principle applies to other components of the panels, such as the environment or other objects.

D. Levels of Coherence in Proposed Approaches

- **Aura Panel Approach**: This approach operates at a highly abstract level, where the LLM interprets the emotional tone of a specific audio context to produce visually related but loosely connected representations. The aim is to augment the auditory experience with non-intrusive visual elements that don't convey fixed semantic meaning.
- **Context Isolation Approach**: This approach assumes the breaking of cause-and-effect expectations and interdependencies in favor of visualizing individual elements that are the focus of attention. These elements directly correlate with the separated semantic context. This approach can be likened to an imaginary stream, where a visual representations of ideas and concepts are generated one after the other [34].
- **Scenery Approach**: This approach involves mapping through a scenery panel that can accompany audio information, serving as a backdrop for events. It is characterized by a relatively easy automation process and the possibility of greater horizontal coherence due to fewer key details that could distract the recipient.
- **Storyboard Approach**: This approach offers the creation of ready-made panels that only need to be time-synchronized with the audio material. It allows for a high degree of artistic freedom for the generator, enabling extraordinary aesthetic effects with minimal human intervention. It is important to note that depending on the implementation, high graphical quality may only be associated with vertical coherence, not the horizontal coherence of individual panels with each other, which may result in disengaging the audience during audiovisual projection.
- **Visual Novel Approach**: This approach involves using component elements to assemble panels, offering the promise of the highest level of coherence and customization at the cost of the highest level of human intervention by hand-picking individual elements.

E. Visual Elements as Component Collections

The generated visual elements can be conceptualized as collections of components that reflect the audio context. These panels serve not only as transformable entities but also as aggregations of individual components that can be isolated and manipulated. This is particularly evident in the proposed Visual Novel methodology, where libraries of substituent visual elements can be created to correspond with specific audio concepts. Two primary approaches to interpreting these panels are:

- **Raster-Based Interpretation**: In this approach, a panel is viewed as an assembly of pixel-based elements within the context of raster graphics. Specialized raster graphics editing software can dissect these panels into distinct components. Additionally, layers with focal points can be generated by re-running the panels through LDMs [35]–[37]. These layers can be overlaid to produce spatial effects like the parallax effect.
- **Vector-Based Interpretation**: Alternatively, a panel can be understood as a collection of elements in vector space. This requires a specific style of panel design with well-defined edges. Vector components offer more precise separation and modification capabilities compared to raster-based elements. Advances in AI-driven image processing technologies enable accurate conversions from raster to vector, expanding the possibilities for prototyping in visual media generation [38].

F. Strategies for Visual Component Generation

Listeners are generally limited to auditory perception for experiencing audio media. Procedural visual component generation techniques offer a way to augment these sensory channels [4], [6], [34]. The added visual elements serve as auxiliary components designed to enhance the overall perceptual experience.

The strategy for creating supplementary visual stimuli should be carefully calibrated. This involves considering both the methodologies employed and the intended level of interaction, as well as the appropriateness of the visual stimulus in relation to the auditory content. Two primary approaches to this are:

- **Simplicity and Impact**: Visual elements that are straightforward in terms of contextual complexity can be easier to produce technologically, yet still effectively enhance the aesthetic experience [5]. These elements can be further enriched by leveraging procedural generation techniques, making approaches like mood and scenery methods particularly promising.
- **Complexity and Coherence**: For more complex techniques that require high fidelity to the original audio, maintaining semantic coherence becomes crucial. Any inconsistencies can serve as distractions, detracting from the overall aesthetic enjoyment and focus on the auditory content.

VII. Fine-Tuning Suggestions

Fine-tuning visual components is a critical step in optimizing the viewer's experience. The process serves multiple purposes: it enhances immersion by creating more engaging visuals, it increases coherence between the audio and visual elements, and it elevates the overall aesthetic quality of the composite media [4], [6], [29], [34].

A. Role of Creators

The spectrum of creator involvement ranges from full creative control to varying degrees of automation. As technologies mature, creators find themselves in a dynamic interplay with automated systems, where both the generation and selection of visual elements can be either human-led, machine-assisted, or fully automated. This evolving landscape offers creators the flexibility to refine and adjust visual components, thereby enhancing the aesthetic experience. The nuanced relationship between human creators and automation technologies provides a rich tapestry for understanding how visual stimuli can be crafted to extend aesthetic sensations [33].

B. Current Technologies

Advancements in generative transformers are expanding the scope and flexibility for implementing various visual effects, playing a pivotal role in dynamizing the visual component [35], [39]–[42]. Unlike real photographs, which have a direct reflection in reality and can be naturally altered, these visual images are prefabricated products without a root source for fundamental modification. Within this technological landscape, it is feasible to incorporate a diverse range of elements into the transformation of visual media. Operations can be executed in the realm of virtual visual materials, eliminating the need for reference to any hypothetical, tangible visual source. This signifies a new paradigm in the creation of visual materials, one that leverages pre-existing, procedurally generated prefabricates for further refinement. The cornerstone of this methodology is the capability for autonomous generation of these components.

C. Panel Succession

The succession of panels in a visual representation, as shown in Figure 3, should be intricately tied to the accompanying sound. Some approaches include:

- **Fixed Time Changes**: In this approach, the panels change at fixed time intervals, providing a predictable yet potentially less dynamic visual experience.
- **Rhythmic Changes**: In this approach, the panels change in synchronization with the musical accents or beats of the accompanying sound.
- **Focus Changes**: In this approach, the panel changes to shift focus to the object of attention, following dynamic of dialogue or other interactive scenarios.

D. Panel Styling

Common styling can help maintain greater visual coherence [43]. Panels made in similar styling make it easier for the audience to compare them, creating the impression that they represent a common diegesis. Styling may be done outside the panel generation process, in a form such as passing images through an external filter (e.g., applying a blur, a filter suggesting dreaming). Styling using LDM can be achieved through prompting, e.g.:

- Art medium stylization.
- Composition.
- Graphic styling.
- Referring to a common visual reference.

E. Introducing Motion

The aim is to focus greater attention of the audience on the visual context, thereby increasing engagement and the impact of the stimulus [44]. One way to achieve this is by introducing motion, which can be categorized into three main types:

- **Overall Frame Movement**: This involves manipulations like zooming in and out within individual panels to create a sense of motion and depth.
- **Interframe Effects**: These include transitions and shifts between panels, as well as special effects that constitute a separate, continuous layer between individual panels. An example would be a fog effect that maintains and preserves motion continuity while obscuring the changing panels.
- **Intraframe Movement**: This refers to the movement or animation within a single frame or panel, such as the motion of characters or objects, which adds another layer of dynamism to the visual experience.

F. Intraframe Movement

This involves the motorization of elements within the panels creating points of interests:

- Adding artificial visual effects (commonly abbreviated as 'VFX'), e.g., related to weather (rain, fog), uniform movement of areas such as water, flame, leaves, wind.
- Adding depth of field and manually creating spatial movements in the frame by shifting layers relative to each other.

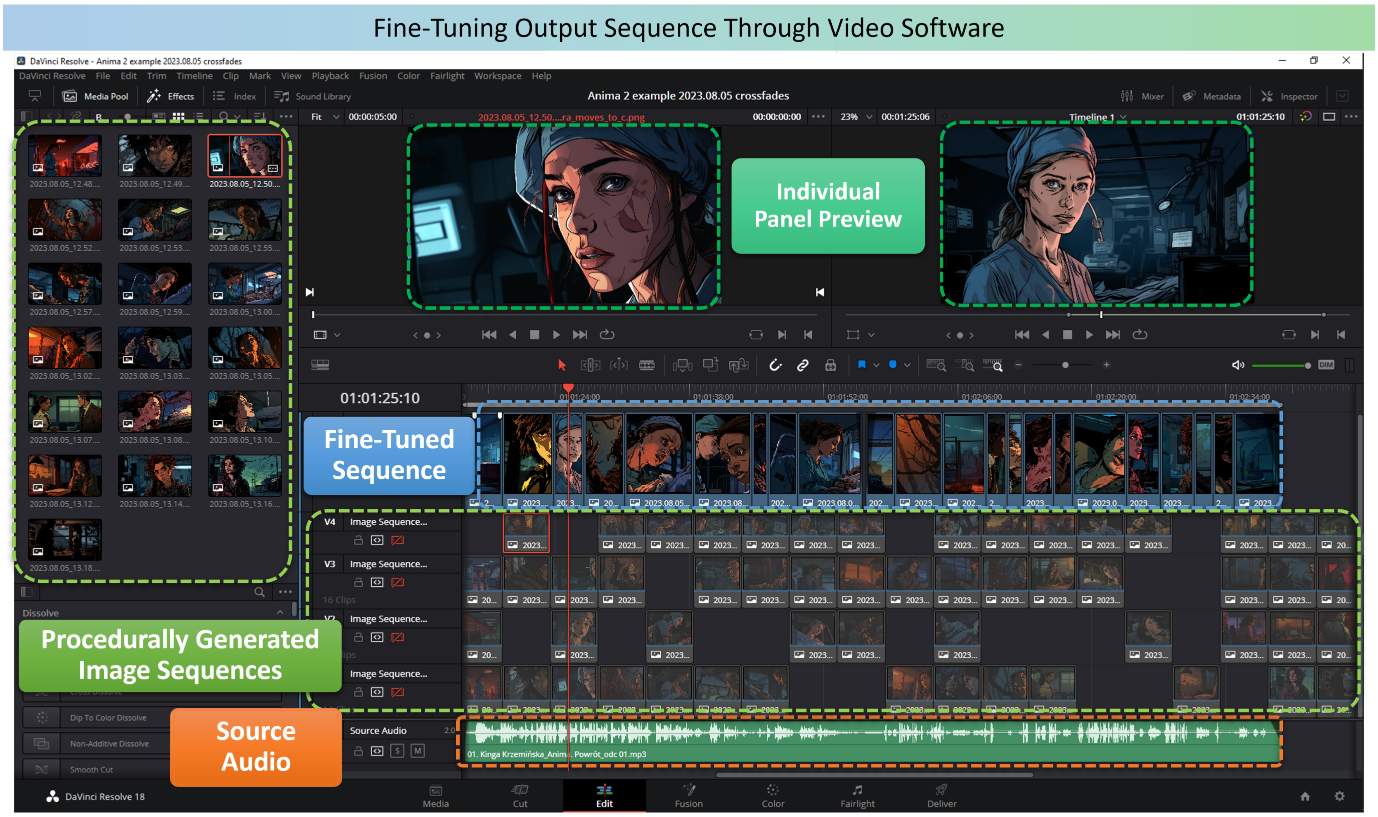


Fig. 3. Fine-tuning visual sequence through panel succession by puzzle-like composing 4 sets of sequences into one. An intermediary example demonstrating storyboard and context isolation method features available at: https://youtu.be/Ier6ROLrF4w. Additionally, showcase with transitions using storyboard method is available at: https://youtu.be/a1oYdO6oRIc.

- Animating characters:
  - Adding lip movement.
  - Adding eye movement.
  - Adding facial expressions, head movement.
  - Adding body movement.

G. *Generative Transformations*

Applying generative transformations to the image, such as:

- Using intermediary panels between the seed of the generator and the image, showing an animated process of generating a given panel or effects of transforming a ready panel (e.g., by adding seed and re-generating the final result).
- Generative transformation effects based on a seed created from the panel, a multi-version panel that changes (to varying degrees, in different ways).

VIII. Conclusion

This paper delved into the complex realm of composite visual media generation with a focus on augmenting audio contexts through the integration of visual elements. Utilizing advanced AI-driven techniques, the paper discussed the potential for enhancing traditional audio experiences with procedurally generated visuals and proposed methods for further refinement and exploration. The combination of auditory and visual elements through AI-driven procedural generation offers a promising avenue for enriching the sensory depth and emotional impact of artistic works. As these methods continue to evolve, the boundaries between technological feasibility and artistic significance will likely continue to blur, thereby expanding the possibilities in the field of composite media procedural generation.

A. *Possible Future Research Directions*

It is worth mentioning that the scope of visual layer generation is not confined solely to auditory stimuli with textual components. Semantic representation can extend to musical elements, encompassing harmonic and melodic progressions. LLMs possess the capability to process abstract musical concepts, such as harmonization or genre classifications [1]. These networks can analyze and interpret melodic-harmonic sequences, offering avenues for abstract-level visual interpretations akin to human musical perception. Emotional responses and reception patterns associated with specific musical structures can be analyzed, replicated, and processed [27], [28]. Such interpretation may not be straightforward and deterministic, reflecting the diverse ways in which music is perceived by individuals. This represents a further exploration into deriving visual contexts based on musical semantics, moving beyond wave analysis and sound spectrum to focus on the interpretation of musical language's semantic patterns.

B. Accelerated Advancements in Technology

The preponderance of recent citations in this article, from the current and previous year, underscores the rapid pace of innovation in this domain. This surge in technological expansion indicates a fertile ground for future research and development. As technology continues to evolve at an accelerated pace, there is a significant opportunity for groundbreaking findings that could revolutionize the creation of even more immersive compound media.

IX. Acknowledgements

This research draws its main inspiration and leitmotif from an author's engineering thesis conducted at the **Polish-Japanese Academy of Information Technology**, under the guidance of **Konrad Maciborski MEng**. As the supervisor, his insights and approach to scientific inquiry have been the cornerstone upon which this further exploration was built.

Furthermore, heartfelt thanks are extended to the **Feliks Nowowiejski Academy of Music in Bydgoszcz** for providing an enriching academic atmosphere that encourages scholarly growth and interdisciplinary collaboration. A special note of gratitude is directed towards Rector **Prof Elżbieta Wtorkowska** and Dean of the Department of Composition, Music Theory, and Sound Engineering, **Dr Szymon Godziemba-Trytek**, for their steadfast support and inspiration. Particular thanks are also due to **Dr Katarzyna Figat** for her indispensable assistance in kick-starting this research endeavor.

References

[1] S. Bubeck, V. Chandrasekaran, R. Eldan, J. A. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. Lee, Y.-F. Li, S. M. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang, "Sparks of artificial general intelligence: Early experiments with gpt-4," *arXiv.org*, 2023. [Online]. Available: https://doi.org/10.48550/arxiv.2303.12712

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *Computer Vision and Pattern Recognition*, 2021. [Online]. Available: https://doi.org/10.1109/cvpr52688.2022.01042

[3] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," *Neural Information Processing Systems*, 2022. [Online]. Available: https://doi.org/10.48550/arxiv.2205.11487

[4] C. Gao, J. J. Green, X. Yang, S. Oh, J. Kim, and S. V. Shinkareva, "Audiovisual integration in the human brain: a coordinate-based meta-analysis," *Cerebral Cortex*, vol. 33, no. 9, pp. 5574–5584, 11 2022. [Online]. Available: https://doi.org/10.1093/cercor/bhac443

[5] H. Lima, B. LimaHugo, C. G. R. dos Santos, S. G. R. Dos, S. MeiguinsBianchi, and B. S. Meiguins, "A survey of music visualization techniques," *ACM Computing Surveys*, 2021. [Online]. Available: https://doi.org/10.1145/3461835

[6] M. Tiihonen, E. Brattico, J. Maksimainen, J. Maksimainen, J. Wikgren, and S. Saarikallio, "Constituents of music and visual-art related pleasure – a critical integrative literature review," *Frontiers in Psychology*, 2017. [Online]. Available: https://doi.org/10.3389/fpsyg.2017.01218

[7] M. Mller, *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*, 1st ed. Springer Publishing Company, Incorporated, 2015. [Online]. Available: https://doi.org/10.1007/978-3-319-21945-5

[8] S. Latif, H. Cuayáhuitl, F. Pervez, F. Shamshad, H. S. Ali, and E. Cambria, "A survey on deep reinforcement learning for audio-based applications," *Artificial Intelligence Review*, vol. 56, no. 3, pp. 2193–2240, 2023. [Online]. Available: https://doi.org/10.1007/s10462-022-10224-2

[9] W. S. Peebles and S. Xie, "Scalable diffusion models with transformers," *arXiv.org*, 2022. [Online]. Available: https://doi.org/10.48550/arxiv.2212.09748

[10] S. Wu, T. Wu, F. Lin, S. Tian, and G. Guo, "Fully transformer networks for semantic image segmentation," *arXiv.org*, 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2106.04108

[11] L. Yang, Z. Zhang, and S. Hong, "Diffusion models: A comprehensive survey of methods and applications," *arXiv.org*, 2022. [Online]. Available: https://doi.org/10.48550/arxiv.2209.00796

[12] A. Ulhaq, N. Akhtar, and G. Pogrebna, "Efficient diffusion models for vision: A survey," *Cornell University - arXiv*, 2022. [Online]. Available: https://doi.org/10.48550/arxiv.2210.09292

[13] X. Pan, P. Qin, Y. Li, H. Xue, and W. Chen, "Synthesizing coherent story with auto-regressive latent diffusion models," *arXiv.org*, 2022. [Online]. Available: https://doi.org/10.48550/arxiv.2211.10950

[14] J. Zakraoui, M. Saleh, S. Al-Máadeed, and J. M. Alja'am, "A pipeline for story visualization from natural language," *Applied Sciences*, 2023. [Online]. Available: https://doi.org/10.3390/app13085107

[15] H. Chen, R. Han, T.-L. Wu, H. Nakayama, and N. Peng, "Character-centric story visualization via visual planning and token alignment," *Cornell University - arXiv*, 2022. [Online]. Available: https://doi.org/10.48550/arxiv.2210.08465

[16] Y. Z. Song, Y.-Z. Song, Y.-Z. Song, Z. R. Tam, Z. R. Tam, H.-J. Chen, H.-J. Chen, H.-H. Lu, H.-H. Shuai, and H.-H. Shuai, "Character-preserving coherent story visualization," *European Conference on Computer Vision*, 2020. [Online]. Available: https://doi.org/10.1007/978-3-030-58520-4_2

[17] S. Chen, B. Liu, B. Liu, B. Liu, B. Liu, B. Liu, J. Fu, R. Song, Q. Jin, P. Lin, P. Lin, X. Qi, C. Wang, and J. Zhou, "Neural storyboard artist: Visualizing stories with coherent image sequences," *arXiv: Artificial Intelligence*, 2019. [Online]. Available: https://doi.org/10.1145/3343031.3350571

[18] A. Maharana, D. Hannan, and M. Bansal, "Storydall-e: Adapting pretrained text-to-image transformers for story continuation," *European Conference on Computer Vision*, 2022. [Online]. Available: https://doi.org/10.48550/arxiv.2209.06192

[19] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Neural Information Processing Systems*, 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2105.05233

[20] J. Zhou, X. Shen, J. Wang, J. Zhang, W. Sun, J. Zhang, S. Birchfield, D. Guo, L. Kong, M. Wang, and Y. Zhong, "Audio-visual segmentation with semantics," *arXiv.org*, 2023. [Online]. Available: https://doi.org/10.48550/arxiv.2301.13190

[21] G. Irie, M. Ostrek, H. Wang, H. Kameoka, A. Kimura, T. Kawanishi, and K. Kashino, "Seeing through sounds: Predicting visual semantic segmentation results from multichannel audio signals," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2019. [Online]. Available: https://doi.org/10.1109/icassp.2019.8683142

[22] C. Liu, P. Li, X. Qi, H. Zhang, L. Li, D. Wang, and X. Yu, "Audio-visual segmentation by exploring cross-modal mutual semantics," *null*, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2307.16620

[23] G. Yariv, I. Gat, L. Wolf, Y. Adi, and I. Schwartz, "Audiotoken: Adaptation of text-conditioned diffusion models for audio-to-image generation," *arXiv.org*, 2023. [Online]. Available: https://doi.org/10.48550/arxiv.2305.13050

[24] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J. Nie, and J. rong Wen, "A survey of large language models," *arXiv.org*, 2023. [Online]. Available: https://doi.org/10.48550/arxiv.2303.18223

[25] T. Görne, "The emotional impact of sound: A short theory of film sound design," *null*, 2019. [Online]. Available: https://doi.org/10.29007/jk8h

[26] J. Z. Wang, S. Zhao, C. Wu, R. B. Adams, M. Newman, T. Shafir, and R. Tsachor, "Unlocking the emotional world of visual media: An overview of the science, research, and impact of understanding emotion," *Proceedings of the IEEE*, 2023. [Online]. Available: https://doi.org/10.1109/jproc.2023.3273517

[27] X. Wang, X. Li, Z. Yin, Y. Wu, L. J. D. O. P. L. O. Brain, Intelligence, T. University, D. Psychology, and R. University, "Emotional intelligence of large language models," *arXiv.org*, 2023. [Online]. Available: https://doi.org/10.48550/arxiv.2307.09042

[28] S. C. Patel and J. Fan, "Identification and description of emotions by current large language models," *bioRxiv*, 2023. [Online]. Available: https://doi.org/10.1101/2023.07.17.549421

[29] Z. Akhtar and T. H. Falk, "Audio-visual multimedia quality assessment: A comprehensive survey," *IEEE Access*, 2017. [Online]. Available: https://doi.org/10.1109/access.2017.2750918

[30] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, "A review of deep learning techniques for speech processing," *Information Fusion*, vol. 99, p. 101869, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253523001859

[31] J. Li, X. Zhang, X. Zhang, X. Zhang, X. Zhang, X. Zhang, C. Jia, J. Xu, X. Jizheng, L. Zhang, L. Zhang, L. Zhang, Z. Li, L. Zhang, Y. Wang, Y. Wang, W. Yue, Y. Wang, S. Ma, W. Gao, and W. Gao, "Direct speech-to-image translation," *arXiv: Multimedia*, 2020. [Online]. Available: https://doi.org/10.1109/jstsp.2020.2987417

[32] G. Samson, "Multimodal media generation: Exploring pipeline of procedural visual context-dependent media layer creation," Warsaw, p. 67, 2023, thesis (Engineering) - Polish-Japanese Academy of Information Technology, 2023. [Online]. Available: https://system-biblioteka.pja.edu.pl/Opac5/faces/Opis.jsp?ido=40788#

[33] J. Edwards, A. Perrone, and P. R. Doyle, "Transparency in language generation: Levels of automation," *CIU*, 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2006.06295

[34] R. Adaval, G. Saluja, and Y. Jiang, "Seeing and thinking in pictures: A review of visual information processing," *Consumer Psychology Review*, 2018. [Online]. Available: https://doi.org/10.1002/arcp.1049

[35] P. Gholami and R. Xiao, "Diffusion brush: A latent diffusion model-based editing tool for ai-generated images," *arXiv.org*, 2023. [Online]. Available: https://doi.org/10.48550/arxiv.2306.00219

[36] P. Li, Q. Huang, Y. Ding, and Z. Li, "Layerdiffusion: Layered controlled image editing with diffusion models," *arXiv.org*, 2023. [Online]. Available: https://doi.org/10.48550/arxiv.2305.18676

[37] X. Zhang, W. Zhao, X. Lu, and J. Chien, "Text2layer: Layered image generation using latent diffusion model," *arXiv.org*, 2023. [Online]. Available: https://doi.org/10.48550/arxiv.2307.09781

[38] X. Ma, Y. Zhou, X. Xu, B. Sun, V. Filev, N. Orlov, Y. Fu, and H. Shi, "Towards layer-wise image vectorization," *Computer Vision and Pattern Recognition*, 2022. [Online]. Available: https://doi.org/10.1109/cvpr52688.2022.01583

[39] M. Dorkenwald, T. Milbich, A. Blattmann, R. Rombach, K. Derpanis, and B. Ommer, "Stochastic image-to-video synthesis using cinns," *Computer Vision and Pattern Recognition*, 2021. [Online]. Available: https://doi.org/10.1109/cvpr46437.2021.00374

[40] Y. Hu, C. Luo, and Z. Chen, "Make it move: Controllable image-to-video generation with text descriptions," *Computer Vision and Pattern Recognition*, 2021. [Online]. Available: https://doi.org/10.1109/cvpr52688.2022.01768

[41] M. Stypulkowski, K. Vougioukas, S. He, M. Ziba, S. Petridis, and M. Pantic, "Diffused heads: Diffusion models beat gans on talking-face generation," *arXiv.org*, 2023. [Online]. Available: https://doi.org/10.48550/arxiv.2301.03396

[42] L. Shen, X. Li, H. Sun, J. Peng, K. Xian, Z. Cao, and G.-S. Lin, "Make-it-4d: Synthesizing a consistent long-term dynamic scene video from a single image," *arXiv.org*, 2023. [Online]. Available: https://doi.org/10.1145/3581783.3612033

[43] J. Wu, J. J. Y. Chung, and E. Adar, "Viz2viz: Prompt-driven stylized visualization generation using a diffusion model," *arXiv.org*, 2023. [Online]. Available: https://doi.org/10.48550/arxiv.2304.01919

[44] C. K. Praveen and K. Srinivasan, "Psychological impact and influence of animation on viewer's visual attention and cognition: A systematic literature review, open challenges, and future research directions." *Computational and Mathematical Methods in Medicine*, 2022. [Online]. Available: https://doi.org/10.1155/2022/8802542