INTERNATIONAL JOURNAL OF FILM AND MEDIA ARTS (2026) Vol. 11 Nº. 1 pp. 53-76

ijfma.ulusofona.pt
doi: 10.24140/ijfma.v11i1.10913

A MULTIMODAL FRAMEWORK FOR AUTOMATED BACKGROUND MUSIC GENERATION IN JAPANESE MANGA USING LARGE LANGUAGE MODELS

MEGHA SHARMA
The University of Tokyo, Japan

MUHAMMAD TAIMOOR HASEEB
Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates

GUS XIA
Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates

YOSHIMASA TSURUOKA
The University of Tokyo, Japan

Megha Sharma is a master of science at the University of Tokyo. She specialises in multimodal music retrieval & generation as well as ethical MIR and NLP. She is a research assistant at MBZUAI, United Arab Emirates. Her interest lies in methods of ethical work with AI, and in particular in generating music, predicting historical risks, and analysing social media outputs.
ORCID ID: 0000-0002-4313-7459

Corresponding author
Megha Sharma
ms22sharma@gmail.com

Acknowledgements
This work was made possible by the Aizawa Yamakata Matsui Lab at the University of Tokyo, who curated the Manga109 dataset.

Schedule for publication
Paper Submitted: 30th December 2025
Accepted for Publication: 3rd February 2026
Published online: 30 May 2026

Abstract

Recent waves of digitization have introduced a new form of enjoying comics: with music. We begin our analysis of this movement by discussing the multifarious modes of pairing music and comics as well as how music-comic pairs are received by readers and scholars. Our literature review reveals that this new movement has been growing sluggishly due to the time, effort, and cost required to produce music centered around the detailed story of a comic. In this work, we present a possible approach of using generative AI to automate the background music generation. The biggest challenge in this task is the lack of available data or a baseline. After extensive experimentation, we propose an audio generation pipeline that produces background music for an input manga (Japanese comic) book. By incorporating scene segmentation, longer context, and prompt engineering, we create a novel reading experience for manga readers by adding music as an additional stimulus. The pipeline begins with using the dialogues in a manga to detect scene boundaries and perform emotion classification using the characters' faces within a scene. Then, we use GPT-4o to translate this low-level scene information into a high-level music directive. Conditioned on the scene information and the music directive, another instance of GPT-4o generates page-level music captions to guide a state-of-the-art text-to-music model. This produces music that is aligned with the manga's evolving narrative. In a subjective evaluation, we find that participants prefer the proposed pipeline more than the baseline pipeline by a statistically significant margin across the metrics of relevancy, quality, and consistency. In particular, we find that our pipeline's largest contribution comes in providing consistency across pages as well as generating efficient text prompts for the music generation step. Alongside our development, ethical risks of generative AI in affecting transparency, bias, access and exploitation were examined and mitigated where possible. We provide output samples of the pipeline at: manga-to-music.github.io/M2M-Gen/

Keywords: Manga, Multimodal Analysis, Digital Comics, Generative AI, Content-Based Music Generation

Can comics hear their electric dreams? A popular digital comic titled *Purple Hyacinth*[1] begins with the sound of a single gunshot followed by soft piano. As the reader scrolls down the screen, an image of a fired gun enters the frame. *Purple Hyacinth* is one of many comics hosted on a digital platform, Webtoon, where artists are increasingly incorporating background music scores and sound effects as readers scroll through the comics. Current estimates indicate that Webtoon hosts over 160 comics with music[2]. Although sparse, music accompaniment to a comic predates popularisation of digital comics. Blin-Rolland's (2019) analysis of the French comic art form *bande dessinée* and Johnston's (2016) methodological account of composing music for comics demonstrates that music has long been integral to the comic creation process through live performances beginning in early 2000s. The combination of music with comics (what we refer to as music-comics), therefore, does not only exist but is also gaining rapid popularity through digital adaptations.

In our work, music-comics are operationalised as inherently asymmetrical, where the focus is on the unilateral impact of music on a comic. This relationship is grounded in existing literature that suggests that sound may influence the perception of visual stimuli, drawing attention to specific events through redundancy and emotional coherence between the auditory and visual stimuli (Salselas et al., 2021). A well-understood phenomena among filmmakers, music can, thus, take a functional role in influencing the reception of a film (Green, 2010). Theoretical analysis (Addis, 2017; Blin-Rolland, 2019; Summers, 2015) and reader perception studies (Zulkifli & Ahmad, 2024) of music-comic pairs align with this notion as well, suggesting that sound influences engagement with visual storytelling, as exemplified by how the audio of the beginning of *Purple Hyacinth* grasps the attention of the reader.

Our literature review reveals a significant gap in the exploration of background music for comics. Creating background music demands a novel composition that is relevant to the changing mood, theme, and pace of the comic, requiring composers to be highly creative and adaptable while collaborating with comic artists. For many comic artists, this engaging addition is exclusive to those who can find and employ musicians. An underexplored, albeit divisive, approach to resolving this resource disparity is to automate background music generation for comics using machine learning models. In this work, we study the feasibility of this novel task through implementation and evaluation using Japanese manga comics. In the Japanese media ecosystem, Manga is one part of the "manga–anime–live-action adaptation triangle" where intellectual property from one source is adapted to others (Pusztai, 2015). Manga has, thus, been a popular source material for animated films and TV (anime); 12 out of the 15 top anime ranked by a popular database of manga and anime are adapted from manga[3]. This positions manga as a likely candidate of cross-media adaptation, which includes integration of sound design among other features, and therefore, makes it a viable point of departure in studying automation of background music.

Our research question aims to answer if current established machine learning techniques can successfully generate background music for comics which is relevant, of good quality, and coherent. Lacking an end-to-end dataset and prior research on this specific task, we explore adjacent domains

1 This comic is accessible on Webtoon in English. Available at: https://www.webtoons.com/en/mystery/purple-hyacinth/list?title_no=1621&page=1

2 As per the count available from https://webtoon.fandom.com/wiki/Category:Webtoons_with_Music (Accessed March 2026). It should be noted however, that the website is yet to be updated, so the number may not be accurate and is only an underestimated approximation.

3 As per the ranking available at https://myanimelist.net/topanime.php (Accessed March 2026).

and establish a baseline to encourage future research and exploration. By leveraging recent advancements in vision, text, and music generation models, we formalize and explore the relationship between manga and music. The proposed system is designed to generate background music that aligns with the content and sentiment of a manga, evolving as the narrative unfolds[4]. We evaluate our work through two user studies, comparing our pipeline against a strong baseline and a random lower bound in different settings. The results of these studies reveal that the proposed system achieves higher reader satisfaction across the metrics of relevancy, quality and consistency as compared to the baselines.

In our modular approach, all but one model is publicly available, making our pipeline efficient for generating music for low-resource comics with minimal training. The technical decisions make the pipeline appropriate for character-driven stories with ample dialogue and expressive facial graphics.

Potentially, this research could be useful for comic artists interested in adding music to their comics and manga by automating the music generation process or for directors seeking to adapt comics to audiovisual media. While the system is fully autonomous, the pipeline is designed to allow for human monitoring and intervention at any step. Each step of our pipeline provides text outputs in English, allowing troubleshooting of possible errors through human-readable logs. Intervention is mindful of human agency in the creative process; however, the system is not immune to ethical dilemmas. In fact, concerns including exploitation and precarisation of musicians might worsen the resource disparity for amateur artists (Morreale, 2021). Therefore, adoption of this nascent technology must be navigated carefully, with active collaboration among artists, researchers, policymakers, and developers to ensure human dignity and creative integrity guide the innovation.

Literature Review

In this section, we go over the current literature to understand what constitutes a music-comic object and follow up on relevant applications of Artificial Intelligence (AI) models in generating music given a source media. Observations from this section inform our methodology and evaluation in the subsequent sections.

What are Music-Comics?

Academic scholarship on the construction, structure, and reception of comics has been gaining momentum in recent decades through the influential research of scholars including Natsume et al. (1997), McCloud (2006), Groensteen (2007, 2013), Hague (2014), Cohn (2010, 2013), and Berndt (2018). However, combinatory studies on music and comics are sparse in volume and varied in their conceptual approach. This fragmentation is a result of the ambiguity in defining music-comic. By drawing parallels in music and comics as languages, Addis (2017) formalizes 'Musicalization' in comics beyond the occurrence of musical elements within or with comics. However, the language of comics itself is arguably complicated to define, as both text and image play a role in creating a comic. An alternative position seeks to integrate both music and comics into a separate 'Musicomic' (Blin-Rolland, 2019). Derived from Mitchell's (2009) approach of unifying the image and text in graphical narratives as 'ImageText', the emergent 'Musicomic' is the product of resonance through simultaneous experience of music and comic (Blin-Rolland, 2019). Resonance, here, refers to a phenomenon of transient and momentous synchronisation of heterogenous elements to create a new perceptual experience. On achieving resonance, Blin-Rolland (2019) posits that the combined 'Musicomic' becomes a whole greater than the sum of its music and comic parts. One such possibility of resonance occurs when Ondaatje and Murch (2002) discuss how image and sound change each other as they are played

4 We provide sample outputs of the pipeline at: manga-to-music.github.io/M2M-Gen/

together (Chion, 1994). Here, a music-comic media creates a space for a reader to navigate different roles (such as viewer, reader, listener) and synchronise information from multiple stimuli (such as melody and tone in music with graphic style and panel in comics) to create a novel reading experience. Effectively, the resulting simultaneous experience of the music and comic for the reader can create a new perception distinct from the sum of its parts.

Instances of Music-Comics

Dyadic studies of music and comics vary both in scope and definition. Summer's (2015) analyses the presence of musical elements explicitly rooted in the comic print through dialogue (including singing), narration, onomatopoeia or even illustrations of the score in Alan Moore's comics. The diegesis of these stories often participate directly with these pre-existing or newly created music. Addis (2017) expands the relationship of music and comics beyond co-occurrence by defining comics through musical elements such as motifs, dissonance, and temporality. Music and comics are brought closer together, where one can be used to interpret the other. For example, a page is parallel to a musical phrase, and a panel is parallel to a musical bar. That said, music in such music-comic pairings functions more as a framework to interpret (or as between languages, translate) comics through music. Blin-Rolland (2019) develops music-comics further by proposing a unified approach for studying a music-comic pair as a singular 'musicomic' unit. The combination of music and comic here is synchronised by the reader, who experiences resonance as a novel perceptual synchronisation of music and comic elements with possible emotional catharsis. Multifarious factors such as degree of integration and synchronisation influence interaction for every reader (Blin-Rolland, 2019). For example, even a singular difference of reading speed can connect one musical phrase to a different panel for one reader compared to another, creating a new experience altogether. The experience is thus diverse and places more control with the reader, especially in music-comics; however, this also makes the evaluation of such experiences more elusive.

Background Music for Comics

Prior academic scholarship often engages with music found within the context of the comic, where musical elements are rooted implicitly or explicitly in the diegesis of the comic (Addis, 2017; Summers, 2015). However, pairing comics with an incidental musical score is not as straightforward. Some scholars argue that the concatenation of background music and comics compromises the temporal independence of comics and risks misrepresentation through ambiguity of background music (Goodbrey, 2015; Batinić, 2016; Glaude, 2023). By affecting the *tempo* of the reading pace, music could supersede a reader's agency, thereby violating fundamental aspects of a comic (Addis, 2017; Godek, 2006). Moreover, incoherent music accompaniment may be disruptive to a reader's engagement. A granular approach to synchronisation of constituent elements of comics (such as panel, dialogue etc.) and music (such as musical bar, melody etc.), therefore, creates multiple points of failure. Navigating this complex set of many-to-many interactions without circumscribing the temporal rhythm may be inherently limited. Therefore, we use 'Durational Extradiegetic sounds' as defined in Paolucci's (2019) classification of sounds with comics. Here, 'Durative' means that the sound lasts for several seconds and 'Extradiegetic' means that the music is outside the diegesis of the comic (what we refer to as non-diegetic). Notably, Paolucci (2019) argues that extradiegetic sounds are less prone to temporal dictation given their exogenous nature. Thus, we narrow our approach on building an environment of sound that affords open-ended cross-media engagement for the reader, with the purpose to improve visual saliency and aesthetic reception rather than dictate the pace and direction of reading. This approach builds on the insights in audio perception studies which suggest that sound can effectively improve visual saliency, consequently improving attention and engagement with visual events (Salselas, 2021). The legitimacy of non-diegetic background music for comics,

therefore, emerges in its ability to build a new experience, one incapable of existing with a comic in absence of the music. Addis (2017) reconciles earlier contention with this interpretation as she comments "experimentation with these sound technologies is however an interesting way to test the limits of comics as a medium" (p. 9).

Our exploration is further validated from perception studies that examine how engagement in comic reading is impacted with the addition of background music (Zulkifli & Ahmad, 2024). Using a popular Indonesian horror comic, *Kemala*[5], with an accompanying soundtrack, an independent-groups study compared the effect of reading the comic with and without music on the fear experienced by the participant. A total of one hundred participants were divided into two groups and asked to answer a questionnaire to rate their emotion of fear after reading the comic. The results indicated that participants who read the comic with the music playback reported significantly more fear than participants who read the comic without any music. Despite lacking precise control of the music playback over the comic, the music could still induce the relevant emotion of fear. Although this phenomenon is invariably observed and well known among artists, methodologies of constructing such music-comic pairs are sparse in the literature.

Practice-informed research contributes best to our understanding of how music is both created and placed with comics. Given that music-comics are relatively a recent development, practitioners often adapt approaches from other established fields. Game-based approaches often use navigation and diegesis in digital comics to trigger music (Goodbrey, 2015). By rooting music through controls outside (such as navigation buttons) or inside (such as buttons within the comic panel) the comic, music playback is in the control of the reader. In contrast, film-based approaches instead compose music by carefully constructing time as part of the comic reading process (Johnston, 2011). By structuring the comic-reading as a live performance, music playback is controlled by the musician as the comic is displayed behind the performers. In both scenarios, placement of the music and comic are most distinct. Nevertheless, music plays a similar function, communicating a narrative or a message (say, an emotion or style) in both scenarios. Film-music theory suggests that these functions can be categorised into three: 1) 'Formal Function': where music emphasizes or downplays transitions and sequences; 2) 'Narrative Function': where music supports the diegesis of the film, for example, setting a scene in India with sitar music; and, 3) 'Emotional Function': where music directs the emotion in a scene (Flach, 2012).

Without rooting the music into the illustration, narration, dialogue or other tangible sources in the comic, scoring a comic can lead to other challenges in deciding for instrumentation, style, genre, and mood. Untangling and addressing such questions are constructive to our understanding of what constitutes a comic and how music can reconcile sound with visual stimuli.

Background Music Generation with Machine Learning

Current literature lacks studies on the automated generation of background music for manga or other comics. Therefore, our insights are drawn from research on music generation and retrieval for games (Hutchings & McCormack, 2019), books (Shriram et al., 2022), and movies (Haseeb et al., 2024). Existing work indicates that emotional information can be crucial for generating a musical score for any medium. While adapting film soundtracks to books,

5 The comic is available on Webtoon in Bahasa Indonesia. Available at: https://www.webtoons.com/id/horror/kemala/list?title_no=2266

Shriram et al. (2022) apply a BERT[6] emotion classifier to match the sentiment of a given paragraph to the valence (polarity) of musical segments from the existing soundtrack. Hutchings and McCormack (2019) develop an adaptive system where concepts such as environment, object, and emotion act as weighted nodes on a map. Interaction with the game triggers activation of certain nodes, which then guide the style and melody of the music generated with the conditional music generation model. Results from these studies indicate that listening to music while reading a book or playing a game improved the immersive experience and found common ground in placing importance on the emotional function while retrieving or generating background music.

Apart from emotion, the narrative context can play a key role in generating background music for multi-modal media. The Hermann-1 movie-to-music pipeline inputs a movie scene and, through narrative and emotional context, develops a musical directive as a text prompt for a text-to-music model (Haseeb et al., 2024). The generated music complements the movie scene and achieves comparable results to the original soundtrack of the movie scene.

In another study exploring background music for manga, a fine-tuned AudioCLIP (Guzhov et al., 2022) is adapted to infer the relationship between manga and music by matching genre of the manga with a corresponding mood of the music (Sharma & Tsuruoka, 2023). However, the approach is limited in design. Constraining the emotional nuance of an entire page with a single annotation can affect the accuracy. Moreover, retrieving music from an existing dataset is also limited by the quality and scope of the available dataset. Although pre-existing music can cross-reference different sources of information to tie into the universe of the comic, retroactive addition of music to existing comics may be at risk of unintentional messaging.

Retrieval of existing music is more likely to be successful when selected during the construction of the comic. Successful implementations include, for example, references to pre-existing songs like "Roll Out the Barrel" in Alan Moore's *V for Vendetta* to set the *tempo* of the panels (Summers, 2015). We argue that music retrieval may be more meaningful during the comic creation process itself, and achieving reasonable consistency between independent pre-recorded music pieces and the comics may be more complicated than generation.

In this work, we draw inspiration from Hermann-1 (Haseeb et al., 2024), customizing its components to adapt to the structure of comics. We explore a perspective that a narrative context and deeper understanding of the sentiment in the comic are crucial to achieving relevance and consistency with the narrative. To the best of our knowledge, we are the first to address the challenge of generating background music for comics.

Methodology

Given the broad definition of music-comics found in the literature, we limit our scope to build a pipeline that generates empathetic non-diegetic instrumental music for a given comic. Here, empathetic means the mood and style of the music are in consonance with the mood and style of the comic (Chion, 1994). Given manga's specific affordance in cross-media adaptation, it is the subject of our implementation. For the purpose of our study, comics and manga refer to digital scans of print comics as was available to study in our dataset. The aim of our study is to test the feasibility of automated background music generation for comics.

Our proposed method improves upon preliminary experiments, contributing to understanding key features for the

6 A transformer-based deep learning architecture for language pre-training that can be adapted for other tasks including classification. For more information, please refer to Devlin et al. (2019).

final approach. For the sake of clarity and brevity of this article, we focus on the final method and its evaluation, rather than detailing the iterative approach. Results from early experiments suggested that the proposed system should include: 1) A pipeline-based approach: End-to-End machine learning models are trained on existing datasets with input-output pairs. In this case, the input would be an image (or other information) from a comic, and the output would be a piece of music. However, no such dataset is available. Hence, we employ a combination of models to build a pipeline that generates music from comics by translating them into another modality to act as an interpreter. 2) Polyphonic music: single-instrument and monophonic melodies seemingly limit the range of context given the diversity of genre and styles in manga. 3) Multi-modal conditioning: conditioning the input manga solely on a singular modality such as emotion also limits the applicability of the music generated.

Without a ground-truth dataset, defining the objective for a machine learning model can be challenging. Our proposed pipeline is guided by the processes of how we as humans compose for visual media. Findings from existing literature indicate that humans often segment long-form media into smaller, digestible scenes which usually have high consistency within itself but low consistency between each other. Analysis in mono-sensory stimulus reading comprehension (Brenner, 2010), as well as multi-sensory stimuli in film comprehension (Zacks et al., 2009) suggest that this perception model holds true for a diverse combination of sensory inputs.

Furthermore, evidence of consistency within scenes as opposed to between scenes have also been found in neuroscientific studies on comic analysis by observing the effect of disruptions within and between scenes (Cohn et al., 2012).

Given the wide applicability of the theory of event segmentation across perception of real-time events and consumption of long-form media, we follow this perception model to segment comics (as visual inputs) into scenes. Each scene can be segmented using what Cutting (2014) refers to as 'narrative shifts': shifts across dimensions of time, location or character. To avoid disruption of continuity in a scene, any additional modality such as music should also follow the consistency of the scenes. This is especially important given that music can often lower consistency in scene segmentation by viewers, referred to as 'diffusing attention' (Cutting, 2019).

Within each scene, emotional context in visual (facial) and textual (dialogue/narration) content play an important role in informing the mood of the background music. In absence of paired comic-music data, a comic-text-music pipeline was considered more comprehensible. Using an ekphrastic approach, free-form descriptions of the input images allow us greater flexibility in learning the context to inform the music generation step. A music track is generated for an image from a manga, with each image comprising two pages. During exploratory experiments, music generation for each panel was considered as well. However, the precision of using panels, which implies a sequential view-order of a page, could remove the effect, as intended by the author of the manga, of viewing the page as a whole. Addis (2017) argued against comics as simply 'sequential art'. Unlike strictly sequential events in film and music, temporality of comics is not linear, instead it is multi-dimensional through its format. A multi-panel spread on a page might split the reader's attention, emphasizing some panels before others. As Addis (2017) writes, "before we can think sequentially, we are encouraged to view this scene as a whole" (p. 17). In essence, a page containing a collection of panels may build a different experience than viewing a single panel at a time. Moreover, time in panels can be more complex than a single, uniform unit. Stylistic choices can render a single panel to illustrate either a passage of time or an instance of time. Time, i.e. the panel, is thus placed, shaped, and ordered on the page to convey the *tempo* and rhythm of the events. Hence, time is not only the sequence of panels, but the panels on the page themselves.

This phenomenon is well known among authors of comics and manga, who regularly experiment with panel layout, shape, size, and its boundaries to elicit meaning. Therefore, disjointing panels could potentially break a core aspect fundamental to the structure of a comic. Hence, rather than a bottom-up approach, we adopted a top-down approach. By breaking down the manga into self-contained scenes which are similar on a thematic scale, music can be generated for each page of the scene instead, preserving thematic adherence of music to the manga while paying attention to each page.

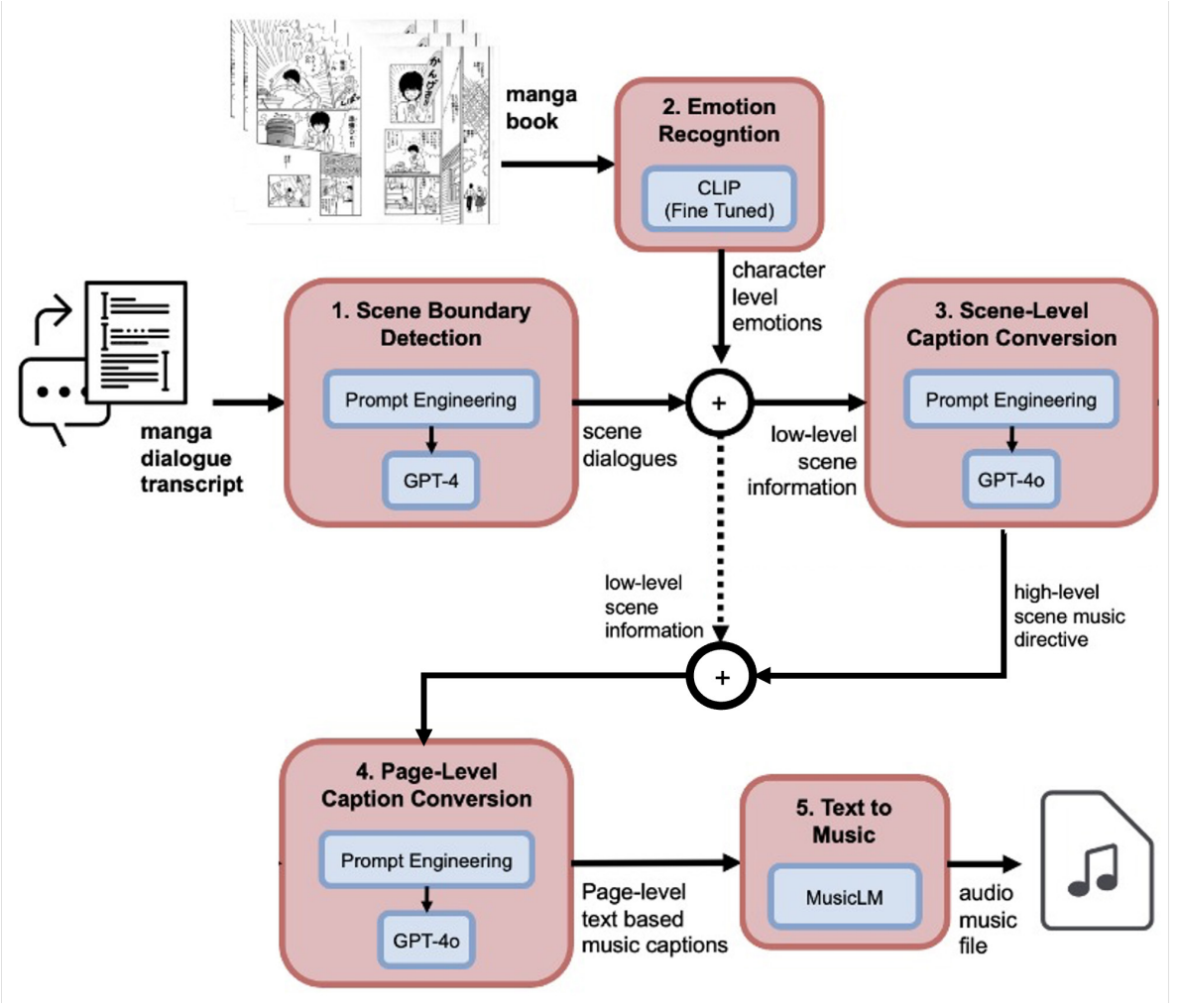

Figure 1 *Illustration of the Proposed Pipeline*
Note. The pipeline takes as input the images and dialogue transcript of a manga book and outputs an audio file containing tailored background music. Manga image taken from Manga109 (Aizawa et al., 2020; Matsui et al., 2017) © Yoshi Masako.

To train and evaluate the proposed method, we employ the Manga109 dataset (Aizawa et al., 2020; Matsui et al., 2017). Manga109 is one of the few large-scale, publicly available datasets on manga. Curated after receiving the permission of the authors of each manga to use for academic purposes, the dataset contains 109 volumes of Manga published between 1970 and 2012. The dataset also provides annotations of characters, faces, bodies, dialogues, and narrations. Although annotation tags are in English, the content of the manga as well as the extracted text are wholly in Japanese.

The larger problem of music composition for comics thus is divided into several intermediate components, where each step of the pipeline informs the subsequent, cascading information from images of comics-to-text to translate into instrumental music. The pipeline, which we call M2M-Gen (Manga-to-Music Generation), is illustrated in Figure 1. The components of the pipeline are described in detail below:

1. Scene Boundary Detection. To detect scene boundaries, we conducted prompt engineering to set the system role of the GPT-4 Series[7] to dissect an input manga book into scenes using the dialogue transcript. Our rationale to use Large Language Models (LLMs) like the GPT series is found in their ability to consistently predict scene boundaries in alignment with human listeners for scene segmentation (Kumar et al., 2023). We conduct a trial-and-error analysis of the system with several system prompts, and we utilise the following prompt in the first iteration of the model:

```
You are a Manga reader. A series of dialogs
will be provided from a Manga, please predict
the scene boundaries of the story using the
dialogs alone. The speaker of the dialog is
given.
An example dialog is given below:
[...]
The output for the above dialogs is divided
by the following lines:
1, 22, 68
Please Note:
1. Only output the line numbers and no other
text.
2. There should be a scene boundary every
time there is a change in location, time, or
characters.
3. There should be at least one scene bound-
ary for every 500 lines.
4. Always begin the output with line 1.
```

The model is conditioned with an example scene boundary of 90 lines of dialogue, from the Manga109 dataset (Aizawa et al., 2020; Matsui et al., 2017), annotated internally in the absence of a ground truth. Given the dialogues in Japanese of an input manga book, the model is prompted to predict the scene boundaries.

2. Emotion Recognition. As exemplified in our literature review, understanding the emotional tone of a scene can be crucial for creating appropriate background music. We use two sources of information to recognise emotions: faces and dialogues. In the current work, we utilise Ekman's six basic emotions (Ekman & Friesen, 1971). This is largely motivated by dataset availability as well as to provide a human-comprehensible data for inferring emotions as distinct categories (such as sad, happy, surprised).

For emotion recognition in faces, we use a dataset containing images of character faces from various manga annotated with their corresponding emotions known as KangaiSet (Théodose

7 The GPT-4 series are a series of closed-source and proprietary large language models developed by OpenAI. The models were accessed using an API key between March 2024 and August 2024. For more information, please refer to: https://platform.openai.com/docs/models/gpt-4 and https://platform.openai.com/docs/models/gpt-4o

& Burie, 2023). This dataset contains over 9,000 samples of faces, bodies and panels from the Manga109 dataset, with emotion annotations for each. These annotations include the six basic emotions alongside an additional class for neutral emotion. Although the dataset is considerably larger than the previous work, it is also severely imbalanced. The smallest emotion class (disgust) has only 30 samples, while the largest emotion class (neutral) has over 2,000. Hence, we evaluate the training of this component using F1 scores.

We experimented with various vision models such as the ResNet series (He et al., 2016), EfficientNet (Tan & Le, 2019), and Vision Transformer (Dosovitskiy, 2021). We found our best results with the CLIP model (Radford et al., 2021), a contrastive vision-language representation model that can be fine-tuned for classification. We use the Adam optimizer with betas set to 0.9 and 0.98 and a weight decay of 0.001. The learning rate is set to 1e−5, and the model is trained for 10 epochs with a batch size of 16. The fine-tuned model achieves a weighted F1 score of 70.2%.

3. Scene-Level Caption Conversion. This step leverages GPT-4o's ability to recognise musical directives given low-level scene information. Our system prompt consists of the following important information: 1) Instructions to create music directives for a manga scene; 2) Samples of music directives from demos of popular text-to-music models (Copet et al., 2023); and 3) The ideal length of the musical description. We adhere to a polite sentence structure, incorporating words like "please" as recent research suggests that LLMs perform better with moderately polite prompts (Yin et al., 2024). We optimized this approach through extensive prompt engineering. High-level scene music directive samples and input prompts are available on our demo page. The system prompt is as follows.

```
As a music composer for a manga you are given
dialogs and emotions of pages of the manga.
You are also given the emotions of the char-
acter faces.
Your job is to generate one description of
music that is suitable for a specific page
in the scene.
Please note that the given description will
be fed into a text-to-music mode so the ideal
description is 1-2 lines long.
Also please only generate one music caption
in the response.
Some example descriptions of music for the
text-to-music model are given below. Please
generate a description of music similar to
the below examples:
[...]
Please note that you only need to describe
the music.
```

The examples mentioned above in the system prompt were taken from the demo pages of text-to-music models MusicGen (Copet et al., 2023) and MusicLM (Agostinelli et al., 2023) to induce coherence with successful text prompts for the text-to-music step.

4. Page-Level Caption Conversion. To ensure that each page within a scene accurately reflects its unique emotional tone, we further break down a single scene-level caption (composite of multiple pages) into captions for each page within a scene. This step converts a scene music directive into a page music caption using an instance of the GPT-4o model with its role set as a music composer. Our prompt consists of the outputs of the three previous steps: 1) Scene dialogues; 2) Emotion recognition results; and 3) Music directive of the scene. This approach induces relevancy and consistency across pages within the same scene. In parallel, a unique caption for each page reflects the diversity in emotional tone or setting within the scene. Sample low-level music captions are available on our demo page. We use the following system prompt for the Page-Level Caption Conversion step.

```
You are given the dialogs and emotions from a scene spanning multiple pages in a manga. You will also be given a description describing the music for the scene.
Your task is to adapt the caption from the scene to the particular page in the prompt. This requires you to convert a general description to a more specific musical description of 1-2 lines for the page.
```

With the above system prompt, we give the following user prompt to the text-to-music model. This generates a unique but consistent caption for each page in the scene.

```
The dialogs in the image are as follows:
PAGE 5 藤田歩: [...
...]
PAGE 8 藤田歩: [...]
The faces have the following emotions: 69% neutral, 13% happiness.
The music description for the entire scene is: A light-hearted jazz trio piece with playful piano melodies, gentle rhythmic drums, and a warm comforting stand-up bass, capturing an atmosphere of casual conversations and a slice of everyday life.
Please Generate a music description for page 8 in the scene
```

5. Text-to-Music. In the final phase, we generate music for a page using page-level text-based music captions. We evaluated state-of-the-art text-to-music models, including MusicGen (Copet et al., 2023) and MusicLM (Agostinelli et al., 2023). For the final pipeline, we chose MusicLM[8] for our pipeline, as it received more favourable feedback in comparison to MusicGen. To ensure a seamless music transition between pages, we experiment with in-painting, conditioning on music generated for the previous page, and fading between musical pieces. Eventually, the final pipeline fades the music between scenes to allow for a seamless flow of music from one scene to another.

Evaluation

To evaluate the performance of the perception of the proposed system by humans, a comparison with other methods is necessary. However, in absence of existing methods from literature, we ablate the proposed pipeline, removing components until bare necessity to complete the task with minimal engineering. As illustrated in Figure 2, the baseline has three steps: 1) An instance of the GPT-4o model describes a manga page using the image and the dialogue; 2) The high-level conceptual description is converted into a low-level detailed music caption using another instance of GPT-4o, and 3) This musical description is fed into a text-to-music model. We fade the music between pages for a seamless transition. Progression of reading time across the comic is dynamic with every panel and page; however, research on viewing times for comic panels suggest a directly proportional relationship between the length of the music generated to the length of the text in the panels of the page (Ikuta et al., 2023). Hence, we train a linear regression model on the data provided by Ikuta et al. (2023) to predict the reading time given the text length. We then round it up to either 30 or 50 seconds per image, where the padded additional time is to account for variation in reading speeds.

Our primary results are derived from a subjective analysis, which is conducted using a Human-Subjective evaluation study. Since there is no dataset for this task, an objective assessment of the whole pipeline becomes infeasible. Two evaluation studies with human participants are conducted

8 MusicLM was accessed through an instance of the model found on the MusicFX platform between March 2024 and August 2024. Accessible at: https://labs.google/fx/tools/music-fx-dj

to test different hypotheses. The results from the first user study test the rejection of the null hypothesis, indicating whether our approach can outperform the competing baseline measured by human satisfaction. In this survey, the proposed pipeline takes context from the entire scene to generate music. The longer context used by the proposed method could contribute to improvement in the performance of relevancy as compared to the baseline and random methods. However, this raises questions like how the pipeline performs when the pages come from different scenes. To measure the difference in the perception of consecutive pages of manga between-scenes as compared to the results from within-scene, we conducted another survey where the images are captured from two different scenes (but consecutive pages) in the manga book.

User Study 1 Design

Existing literature does not provide a suitable dataset for objective evaluation of our proposed pipeline against human-composed background music. In this work, we hence rely on a subjective analysis. The study aims to measure reader satisfaction with the proposed system in comparison to baseline models. The null hypothesis states that there exists no statistical difference in preference across selected models for background music.

The evaluation corpus was populated using books from the Manga109 dataset (Aizawa et al., 2020; Matsui et al., 2017). The selection of books for the survey is tabulated in Table 1. Each sample was selected at random while reflecting the diversity of genre and decade of release. From each book, four consecutive pages (two images) were selected to construct a *scene* using the *scene boundary detection* module defined in the previous section.

The design of the subjective study is adopted from common practices in evaluation of audio generation models (Copet et al., 2023; Kreuk et al., 2023; Yang et al., 2023). In a within-subject setting, participants were asked to rate a *scene* with background music.

The background music was added to a scene using three sources: 1) The proposed model; 2) A baseline model, which is an ablated version of the proposed model; and 3) A random model, which randomly assigns music from a database of generated outputs from the proposed and baseline methods. The performance of the random model serves as a lower bound for performance of automated music generation models for comics.

After watching a video of the manga scene with background music, participants were requested to score the video on three metrics:

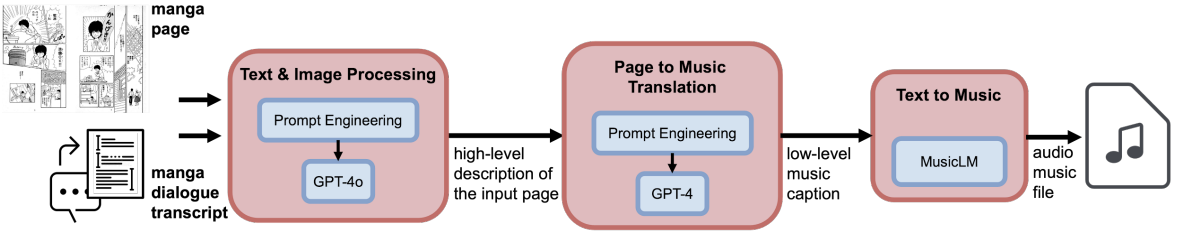

Figure 2 *Illustration of the Baseline Pipeline* Note. The baseline takes as input a manga page and dialogue transcripts and outputs an audio file containing tailored background music for the page. Manga image taken from Manga109 (Aizawa et al., 2020; Matsui et al., 2017) © Yoshi Masako.

- Relevancy (REL): How appropriate the music is for the scene,
- Quality (QUA): How "good" does the music sound, and
- Consistency (CON): How well does the music transitions between pages.

These metrics are derived from the functional role of music in background, as detailed in the 'Literature Review' section (Flach, 2012).

The relevancy metric here tests the narrative and emotional coherence of the music to the context of the corresponding manga, and the consistency metric tests the formal functionality of the music to smooth or contrast transitions.

The scoring was done on a 5-point scale from 1 to 5, where one means a very low score and five means a very high score for the metric. To mitigate bias and fatigue, the order of the scenes was shuffled, and participants were unaware of the source of the music. The user-study was designed as a within-subjects (repeated-measures) ANOVA with the sources acting as a three-level factor. One mean rating per source by each participant was computed by averaging ratings across multiple videos. The repeated-measures ANOVA tested whether average ratings differed across sources. If a statistically significant difference was found, a post-hoc analysis was conducted by comparing each pair using Wilcoxon's test. Since our ratings are ordinal, we used a non-parametric test.

The survey was published online for remote access to participants. Participation was based on voluntary consent, and participants were free to leave at any time during the survey. No demographic information was captured; however, all participants were required to be over 18 years of age and be able to read and understand Japanese.

Several approaches were adopted to recruit participants, including extending a call for participants on public platforms, university channels, and digital communities for Music Information Retrieval (MIR). Additionally, flyers for the survey were also circulated in-person. We only used the results from those who completed the entire survey.

User Study 1 Results

The survey to evaluate the three models was completed by 22 participants, all fluent in Japanese. All participants expressed their interest in manga and music to be at least above 2 (little interested) on a 5-point scale.

The results are illustrated in Figure 3. Both the baseline and proposed pipeline perform well above the random lower

Manga Name	Genre	Author	Release Decade
Nichijou Soup	Humor	Shindou Uni	1990s
ARMS	Science Fiction	Kato Masaki	1980s
Kuroido Ganka	Suspense	Taira Masami	1990s
Totteoki No ABC	Romantic Comedy	Aida Mayumi	1970s
Tasogare Tsushin	Horror	Tanaka Masato	1980s
Gakuen Noise	Battle	Inohara Daisuke	2000s

Table 1 *List of Books Selected for User Study 1*
Note. The information is taken from the metadata of Manga109 available at: http://www.manga109.org/en/explore.html.

bound. Results from the within-subject ANOVA test indicated that the source of the music for the manga had a significant effect on the ratings ($p<0.001$).

Post-hoc analysis revealed that the proposed pipeline is preferred over the random method ($p<0.05$) across all three metrics of relevancy, quality and consistency. Moreover, the proposed pipeline is preferred over the baseline ($p<0.05$) across relevancy and consistency.

The most notable difference is in relevance, where the proposed pipeline ($\mu=3.4$) and the baseline ($\mu=3.25$) outperform the random method ($\mu=2.34$) by nearly one point. In terms of quality, the differences are less pronounced but interesting, nonetheless. Although the difference in mean quality scores is smaller, this difference is unexpected given that the music generation model remained constant across all methods. This implies that not only did participants find the generated outputs of M2M-Gen more relevant, but on average participants also perceived the quality of the proposed pipeline to be better than the rest. Perhaps the greatest contribution of M2M-Gen lies in the consistency of the generated output, with a half-point difference in average scores compared to the baseline. Here, the baseline and random methods are comparable.

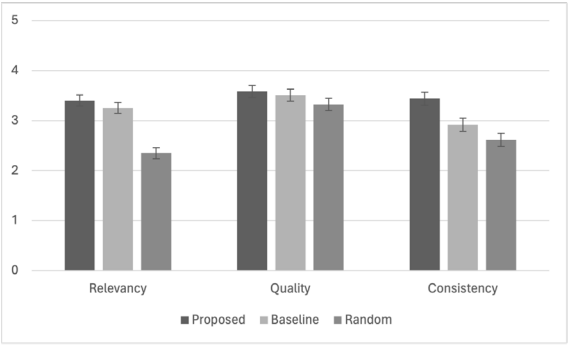

Figure 3 *Mean Scores in User Study 1*
Note. Scores are averaged across each metric (relevancy, quality, consistency) for proposed, baseline and random models for within-subjects study (error bars show standard errors).

User Study 2 Design

The initial survey focused on music generation within a single manga scene. Both images for each scene are provided with the same context to generate music. This could contribute to high scores for the consistency and the relevancy metric. However, this raises questions on how the pipeline performs when the pages come from different scenes. To evaluate the performance of the proposed method when transitioning between two scenes we conducted another user study. In this study, a transition between scenes is presented with music for each scene, where the music is independent of the narrative context of the neighbouring scene. The proposed pipeline takes context from two different scenes (but consecutive pages) of six different manga books.

User Study 2 Results

There were a total of 15 participants, and the survey was designed to be identical to the first. The results for between-scenes study are illustrated in Figure 4. Here again, the proposed method outperforms other methods by a statistically significant margin, contributing to our argument that the proposed method closely follows human satisfaction as compared to the baseline and the random methods. The within-subject ANOVA indicated that again, the source of music

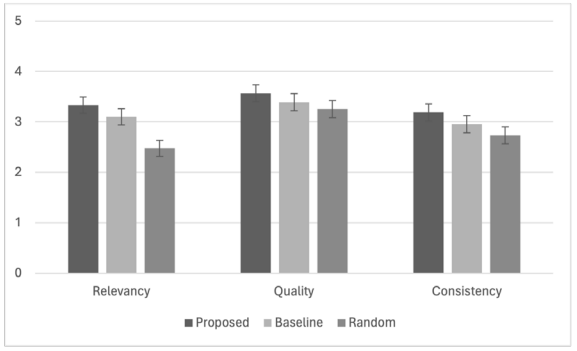

Figure 4 *Mean Scores in User Study 2*
Note. Scores are averaged across each metric (relevancy, quality, consistency) for proposed, baseline and random models for between-subjects study (error bars show standard errors).

had a significant effect on rating ($p<0.001$) even when given consecutive pages across different scenes. Further post-hoc analysis revealed that the proposed pipeline is preferred over the random method ($p<0.05$) across all three metrics of relevancy, quality and consistency, and is preferred over the baseline ($p<0.05$) across relevancy and consistency.

In fact, we find that across both the surveys studying the perception of within-scene and between-scenes, participants prefer the proposed model over the baseline. Our conclusions drawn from the previous study are further validated when tested in the new condition, as we find that the proposed model continues to be the preferred method for both quality and consistency. This survey emphasises performance evaluation for maintaining musical coherence during scene transitions and its adaptability to narrative shifts. Therefore, consistency of the generated music is of particular importance. Results suggest that the proposed pipeline was effectively capturing contextual nuances and ensuring more coherent and satisfying musical transitions. These findings not only validate our proposed method but also highlight its effectiveness in delivering consistent and contextually appropriate music, enhancing reading experience.

When comparing both studies, we first aimed to find statistical differences between the distributions across the independent variables of study (within-scene and between-scene), book (six books from each study), and method (proposed, baseline and pipeline).

To compare the effect of each variable on the ratings, we trained an ordinal logistic regression model, which outputs the significance of the effect of each independent variable on the dependent variable (here, ratings). Our results found that apart from the method, no independent variable had a statistically significant difference in the results. Hence, this deters us from conducting a statistical comparison between the studies.

Ethical Considerations

In this section, we reflect on the most relevant ethical considerations that guided this work. Our ethical principles are centred around human rights and data safety. Although this section appears after methodology and evaluation, in practice, implications of every decision were discussed in sync with the progress of the research.

The primary research question for this paper centres around the ability of current machine learning models to generate background music for comics. The usage of generative AI in this task can be excused as 'tools', where the responsibility of the ethical implications lies with who wields the tool. However, even conceptually, tools are never truly neutral (Xavier, 2025). Although absolute objectivity is appealing in research, truths in creative media are often plural, and neutrality is often impossible. The selection of an approach using machine learning models cascades the ethical issues stemming during the conception, training, and deployment of said model as we utilise it in further research. At the time of the study, this nuance was not overlooked during the development process; however, we could not find an open-source model or dataset to successfully output or train to generate reasonable music tracks from existing research. In a strictly academic setting, we instead built constraints to minimise potential damage as we inherit the responsibility of its deployment.

The dependence on large foundation models such as MusicLM (Agostinelli et al., 2023) and OpenAI's GPT-4 series (Hurst et al., 2024) employed in our study is not immune to issues around bias, transparency and accountability commonly found in such models (Jiao et al., 2025; Ma et al., 2024; Morreale, 2021). Although mitigations to circumvent issues of bias propagating in the outputs exist, such interventions are commonly required during training and before deployment, which was not possible given the nature of closed-source models (see, e.g., Fisher et al., 2025). Our focus instead, was

to foster solutions around transparency and accountability. As the large foundation models used in this system, MusicLM and GPT-4 series, are closed-source and proprietary models, our control over the specific instances and the output of the model was limited.

As a measure of transparency, the methodology of the experiments was documented in detail, including timeframes of access to proprietary models to provide traceability of the experiments.

Another important concern in applications of AI tools is the replacement and exploitation of human work (Huang et al., 2023; Morreale, 2021). Research also shows that standard practice for building and using training datasets in music generation indicate a severe lack of recognition, attribution and compensation for data creators, which include musicians, comic artists, annotators, etc. (Morreale et al., 2023). Where possible, training and evaluation data was obtained from open-sourced datasets with clear consent from the data creators. To avoid pitfalls from the black-box nature of end-to-end models, the design of the pipelines is intended such that it can be replaced by human intervention at any step. This design decision is based on a relevant concern that the pipeline should be in support of creators and researchers, not replace them in their expertise. Furthermore, the decision of translating an image of a manga into text, which then acts as a conditioning input for the music directive, is also motivated by the explainable nature of text in comparison to the more abstract image and music.

While the task of background music generation for comics aims to provide accessible tools for comic artists and readers to enhance the reading experience, potential risks of displacement of sound designers and musicians cannot be ignored. Trade-off between precarisation and productivity creates friction for AI adoption among musicians (Herington et al., 2026). While these conflicts are well-documented, effective intervention remains elusive. Instead of a straightforward approach, solutions require operationalising ecosystems where stakeholders in policy, industry, art, and AI development must act together to ensure AI-based tools do not infringe on human dignity and creative integrity (Ventayen, 2025). For AI developers, along with transparency in development and deployment, investment in educational intervention is critical to put ethical considerations into practice while artists should engage in exploration and evaluation of these systems (Ventayen, 2025). In our work, we try to provide complete technical details of the implementation, while prioritizing clarity for broad accessibility. Currently, our study is confined to the boundaries of academic research, however, implications of our results could potentially drive future research using more capable models. Hence, restraint and reflection are imperative to ensure human-centred values are not neglected in favour of short-term performance gains.

Other standard practices include employing strict privacy control during the user studies. At the foremost, participation was based on voluntary consent, and any participant could withdraw consent at any point after giving consent. Moreover, no personal or identifiable information was collected in the subjective surveys; participants were only asked to rate the music for the manga

Limitations and Future Work

This work serves to stimulate further research by highlighting new methods to construct music-comic pairings. At present, the system generates a music track offline and then the fixed-length music is arranged with the manga as a complete piece. However, in practice, readers of digital comics read at an individual pace, moving back and forth between pages. Research among theorists and practitioners foregrounds the significance of the autonomy of the reader to maintain control over the pace and navigation of the comic even in

music-comic pairings to preserve the fundamentals of what separates comics from other media, say films (Goodbrey, 2015; Groensteen, 2013). The contradiction of indeterminate temporality of a comic in comparison to definite temporal structure of music can threaten reader agency if off balanced. Goodbrey (2015) reconciles this conflict by suggesting looping short tracks; however, in practice, more complex narratives with multiple plots, subplots, characters, and themes can quickly lead to an infeasible production pipeline. Recent music generation models can generate music in a real-time and interactive setting (Caillon & Esling, 2021; Caillon et al., 2025); however, adapting live generation conditioned on reader behaviour while maintaining low latency and relevancy of musical tracks is still a difficult problem in machine learning. We aim that future work can examine the trade-off between reader control and placement of music across panels, pages, and chapters in comics.

Given the machine learning pipeline as well, the current evaluation method is limited in understanding performance of the proposed method against baselines. Without objective metrics in absence of Image-Music datasets, the subjective evaluation with user studies needs to be not only of much larger sample size but also tested under more liberal constraints. After all, readers tend to read comics and manga in leisure, in a location that is comfortable or convenient for them. Moreover, error propagation can be amplified in pipeline designs, such as incorrect emotion recognition may weaken the relevancy of the music generated.

The elusive nature of both music and comics also poses more questions for future work. As music and comics each admit a range of styles, the combinatorial space they produce expands considerably. In our work, music for a scene in a manga is generated with a parallel and empathetic approach, where the music is always faithful to the narrative (Chion, 1994). This consonance may deepen recognition and emotional attachment of the scene. In contrast, film composers in practice may alternatively play music as a counterpoint in asynchrony, where the music delivers a contradiction through an unexpected or experimental choice for the scene (Johnston, 2011). This dissonance may undermine engagement but can become memorable or offer an alternative perspective for the scene in the comic. For example, playing bassoon instrumental music over a battle scene influences reader perception by making the fight appear comedic in tone. LLMs, which generalise over a large corpus of text, risk redundancy and creative diversity through overly formulaic outputs (Wenger, 2026). In our work, this limitation could limit musical directives to only consist of a generalised set of common musical motifs. Although simple approaches to modify diversity such as temperature and presence penalty exist, further exploration may be necessary to understand creativity and productive dissonance in generative AI to expand dimensions of expression in music-comics.

Distinct from traditional print comics, digitisation has also evolved comics to embed features such as hypertext, panel/object movement, and incorporation of other media, commonly found in what Goodbrey (2017) refers to as 'digitally native comics'. These features also change perception and consumption of comics through digital interfaces (Berube et al., 2024). Therefore, experiments with varying user interaction interfaces could be used to discover digital features for constructing multi-modal experiences.

Finally, alternative approaches to background music generation may offer more insight in absence of standardisation for aligning music to comics. Research on film music highlights how thematic approaches can instead focus on recognising motifs in the media which can act as a grounding source of music when present (Stopar, 2024). By identifying key components such as characters, themes or objects, another approach could construct a pipeline to generate music for each component. The presence of the component, thus, may act as a trigger for the music.

Conclusion

This paper examines current theory around music-comics as an object, specifically examining practices in constructing a background musical score. Despite the growing popularity and effectiveness of music-comic pairings, we identify an evident gap in applications supported by machine learning techniques. Our research question aims to answer whether existing AI models can effectively create background, non-diegetic music for Japanese comics. Through extensive experimentation, we have developed a pipeline that extracts information from a manga book and translates it into low-level music conditions for each page, maintaining the coherency and relevancy of the generated music across pages and scenes. By converting different modalities to text before generating music, we can read intermediate directives that reveal how various elements of the image of the comic affect the music's mood and style. The pipeline also integrates state-of-the-art publicly available models, which reduces training requirements in the absence of a dataset. Incorporating scene segmentation, longer context, and prompt engineering, we are able to create a novel reading experience for manga readers by adding music as an additional stimulus. Our proposed model achieves higher human satisfaction across the ratings of relevancy, quality and consistency compared to the random lower bound. Thus, we conclude that it is indeed feasible to develop music generation pipelines using machine learning models. In parallel with the development process, ethical risks of AI are also examined and mitigated where possible. These potential pitfalls include risk of exploitation and precarisation of musicians, which informs our deployment practices of transparency, accountability and reflection.

Although we see this study as an important advancement in multi-modal background music generation for manga, offering a foundation for future research and artistic exploration, we recognise current limitations. These include lack of objective metrics and error propagation for the proposed system as well as reading pace control and a limited synchronous approach in the design of the system. We invite further research by suggesting alternate approaches and methods not explored in this paper. We conclude that although the process of creating background music for a manga may be largely stochastic; by building a system based upon fundamentals in the theory of multi-media music, our proposed pipeline can successfully generate background music for a comic that is relevant and cohesive.

References

Addis, V. (2017). The Musicalization of Graphic Narratives and P. Craig Russell's Graphic Novel Operas, The Magic Flute and Salomé. *Studies in Comics*, *8*(1), 7–28. https://doi.org/10.1386/stic.8.1.7_1

Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., & Frank, C. (2023). MusicLM: Generating Music from Text. *arXiv preprint arXiv:2301.11325.* https://doi.org/10.48550/arXiv.2301.11325

Aizawa, K., Fujimoto, A., Otsubo, A., Ogawa, T., Matsui, Y., Tsubota, K., & Ikuta, H. (2020). Building A Manga Dataset "Manga109" With Annotations for Multimedia Applications. *IEEE Multimedia*, 27(2), 8-18. https://doi.org/10.1109/MMUL.2020.2987895

Batinić, J. (2016). "Enhanced Webcomics": An Exploration of the Hybrid Form of Comics on The Digital Medium. *Image & Narrative*, *17*(5). https://imageandnarrative.be/index.php/imagenarrative/article/view/1384

Berndt, J. (2018). Anime in Academia: Representative Object, Media Form, and Japanese Studies. *Arts*, *7*(4), 56. https://doi.org/10.3390/arts7040056

Berube, L., Priego, E., Wisdom, S., Cooke, I., & Makri, S. (2024). "Moving With the Story": The Haptics of Reader Experience and Response to Digital Comics. *New Review of Hypermedia*

and Multimedia, *30*(1-2), 94–113. https://doi.org/10.1080/13614568.2024.2374291

Blin-Rolland, A. (2019). Becoming Musicomic: Music and Comics in Resonance. *Modern Languages Open*, *1*(2), 1–25. https://doi.org/10.3828/mlo.v0i0.235

Brenner, C. B. (2010). Event Segmentation and Memory Retrieval in Reading Comprehension (Publication No. 6) [Linguistics Honors Projects, Macalester College]. https://digitalcommons.macalester.edu/ling_honors/6/

Caillon, A., & Esling, P. (2021). RAVE: A Variational Autoencoder for Fast and High-Quality Neural Audio Synthesis. *arXiv preprint arXiv:2111.05011*. https://doi.org/10.48550/arXiv.2111.05011

Caillon, A., McWilliams, B., Tarakajian, C., Simon, I., Manco, I., Engel, J., & Roberts, A. (2025). Live Music Models. *The Thirty-ninth Annual Conference on Neural Information Processing Systems Creative AI Track: Humanity*. https://doi.org/10.48550/arXiv.2508.04651

Chion, M. (1994). *Audio-Vision: Sound on Screen* (C. Gorbman, Ed. & Trans.). Columbia University Press. ISBN 978-0231078993

Cohn, N. (2010). Japanese Visual Language: The Structure of Manga. In T. Johnson-Woods (Ed.), *Manga: An Anthology of Global and Cultural Perspectives* (pp. 187–203). Bloomsbury Publishing. https://digital.casalini.it/9781441107879

Cohn, N. (2013). *The Visual Language of Comics: Introduction to the Structure and Cognition of Sequential Images.* Bloomsbury Academic.

Cohn, N., Holcomb, P., Jackendoff, R., & Kuperberg, G. (2012). Segmenting Visual Narratives: Evidence For Constituent Structure in Comics. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *34*(34). https://escholarship.org/uc/item/1r60h6km

Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., & Défossez, A. (2023). Simple And Controllable Music Generation. *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*, *36*, Article 2066, 47704–47720. https://dl.acm.org/doi/abs/10.5555/3666122.3668188

Cutting, J. E. (2014). Event Segmentation and Seven Types of Narrative Discontinuity in Popular Movies. *Acta Psychologica*, *149*, 69–77. https://doi.org/10.1016/j.actpsy.2014.03.003

Cutting, J. E. (2019). Sequences In Popular Cinema Generate Inconsistent Event Segmentation. *Attention, Perception, & Psychophysics*, *81*(6), 2014–2025. https://doi.org/10.3758/s13414-019-01757-w

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., & Houlsby, N. (2021). An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *The Proceedings of the Ninth International Conference on Learning Representations.* https://doi.org/10.48550/arXiv.2010.11929

Ekman, P., & Friesen, W. V. (1971). Constants Across Cultures in the Face and Emotion. *Journal of Personality and Social Psychology*, *17*(2), 124. https://psycnet.apa.org/doi/10.1037/h0030377

Fisher, J., Appel, R. E., Park, C. Y., Potter, Y., Jiang, L., Sorensen, T., & Choi, Y. (2025). Political Neutrality in AI Is Impossible - But Here Is How to Approximate It. *arXiv preprint arXiv:2503.05728*. https://doi.org/10.48550/arXiv.2503.05728

Flach, P. S. (2012). *Film Scoring Today - Theory, Practice and Analysis* [Master's Thesis, The University of Bergen]. https://bora.uib.no/bora-xmlui/handle/1956/6016

Glaude, B. (2023). Listening To the Intrepid Alix: Audio Experience of a Classical Bande Dessinée. *Comicalités*. https://doi.org/10.4000/comicalites.8080

Godek, T. (2006, March 19). Music And Comics: Two Great Tastes That Taste Great Together?. *Comix Talk*. http://comixtalk.com/music_and_comics_two_great_tastes_taste_great_together/

Goodbrey, D. (2015). The Sound of Digital Comics. *Writing Visual Culture*, *7*(1), 1–17. ISSN 2049-718

Goodbrey, D. M. (2017). *The Impact of Digital Mediation and Hybridisation on The Form of Comics* [Doctoral Dissertation, The University of Hertfordshire]. http://merlin.herts.ac.uk/

Green, J. (2010). Understanding The Score: Film Music Communicating to And Influencing the Audience. *The Journal of Aesthetic Education, 44*(4), 81–94. https://doi.org/10.5406/jaesteduc.44.4.0081

Groensteen, T. (2007). *The System of Comics*. Univ. Press of Mississippi. https://www.jstor.org/stable/j.ctt2tvj7m

Groensteen, T. (2013). *Comics And Narration*. Univ. Press of Mississippi. http://www.jstor.org/stable/j.ctt24hvcv

Guzhov, A., Raue, F., Hees, J., & Dengel, A. (2022). AudioCLIP: Extending CLIP To Image, Text and Audio. *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 976–980. https://doi.org/10.1109/ICASSP43922.2022.9747631/

Hague, I. (2014). *Comics And the Senses: A Multisensory Approach to Comics and Graphic Novels*. Routledge. https://doi.org/10.4324/9781315883052

Haseeb, M. T., Hammoudeh, A., & Xia, G. (2024). GPT-4 Driven Cinematic Music Generation Through Text Processing. *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6995–6999. https://doi.org/10.1109/ICASSP48485.2024.10447950

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Herington, J., Borasi, R., Guerrero, B. J., Miller, D. E., Koerner, B., Han, Y. J., & Roberts, R. (2026). Musicians' Ethical Concerns About AI: An Interview Study. *AI & SOCIETY*, 41, 1075–1088. https://doi.org/10.1007/s00146-025-02601-6

Huang, R., Holzapfel, A., Sturm, B., & Kaila, A.-K. (2023). Beyond Diverse Datasets: Responsible Mir, Interdisciplinarity, And the Fractured Worlds of Music. *Transactions of the International Society for Music Information Retrieval*, *6*(1), 43–59. https://doi.org/10.5334/tismir.141

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A. J., Welihinda, A., Hayes, A., Radford, A., Mądry, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., Malkov, Y. (2024). GPT-4o System Card. *arXiv preprint arXiv:2410.21276*. https://doi.org/10.48550/arXiv.2410.21276

Hutchings, P. E., & McCormack, J. (2019). Adaptive Music Composition for Games. *IEEE Transactions on Games*, *12*(3), 270–280. https://doi.org/10.1109/TG.2019.2921979

Ikuta, H., Wöhler, L., & Aizawa, K. (2023). Statistical Characteristics of Comic Panel Viewing Times. *Scientific Reports*, 13(1), 20291. https://doi.org/10.1038/s41598-023-47120-w

Jiao, J., Afroogh, S., Xu, Y., & Phillips, C. (2025). Navigating LLM Ethics: Advancements, Challenges, And Future Directions. *AI and Ethics*, *5*(6), 5795–5819. https://doi.org/10.1007/s43681-025-00814-5

Johnston, P. (2011). The Polysynchronous Film Score: Songs for a Contemporary Score for FW Murnau's Faust (1926).

Screen Sound: The Australasian Journal of Soundtrack Studies, *3*, 89–105.

Johnston, P. (2016). Wordless! Music for Comics and Graphic Novels Turns Time into Space (and back again). *Southerly, 76*(1), 95–110. https://search.informit.org/doi/10.3316/informit.417488798236826

Kreuk, F., Synnaeve, G., Polyak, A., Singer, U., Défossez, A., Copet, J., Parikh, D., Taigman, Y., & Adi, Y. (2023). AudioGen: Textually Guided Audio Generation. *The Proceedings of the Eleventh International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.2209.15352

Kumar, M., Goldstein, A., Michelmann, S., Zacks, J. M., Hasson, U., & Norman, K. A. (2023). Bayesian Surprise Predicts Human Event Segmentation in Story Listening. *Cognitive Science*, *47*(10), e13343. https://doi.org/10.1111/cogs.13343

Ma, Y., Øland, A., Ragni, A., Sette, B. M. D., Saitis, C., Donahue, C., Lin, C., Plachouras, C., Benetos, E., Shatri, E., Morreale, F., Zhang, G., Fazekas, G., Xia, G., Zhang, H., Manco, I., Huang, J., Guinot, J., Lin, L., Wang, Z. (2024). Foundation Models for Music: A Survey. *arXiv preprint arXiv:2408.14340*. https://doi.org/10.48550/arXiv.2408.14340

Marion, P. (1997). Narratologie Médiatique Et Médiagénie Des Récits [Media Narratology and The Mediagenics of Narratives]. *Recherches en Communication*, *7*, 61–87. https://doi.org/10.14428/rec.v7i7.46413

Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T., & Aizawa, K. (2017). Sketch-Based Manga Retrieval Using Manga109 Dataset. *Multimedia Tools and Applications*, *76*(20), 21811–21838. https://doi.org/10.1007/s11042-016-4020-z

McCloud, S. (2006). *Making Comics: Storytelling Secrets of Comics, Manga and Graphic Novels*. Harper Collins Publishers.

Mitchell, W. J. T. (2009). Beyond Comparison. In Heer, J., & Worcester, K. (Eds.), *A Comics Studies Reader* (pp. 116–123). Univ. Press of Mississippi. https://lccn.loc.gov/2008016893

Morreale, F. (2021). Where Does the Buck Stop? Ethical And Political Issues with AI in Music Creation. *Transactions of the International Society for Music Information Retrieval*, *4*(1). https://transactions.ismir.net/articles/86

Morreale, F., Sharma, M., & Wei, I.-C. (2023). Data Collection in Music Generation Training Sets: A Critical Analysis. *The Proceedings of the 24th International Society for Music Information Retrieval Conference*, 37–46. https://archives.ismir.net/ismir2023/paper/000003.pdf

Natsume, F. (1997). *Manga Wa Naze Omoshiroi No Ka: Sono Hyōgen To Bunpō* [Why Manga is Interesting: Its Expression and Grammar]. Nippon Hōsō Shuppan Kyōkai.

Ondaatje, M., & Murch, W. (2002). *The Conversations: Walter Murch And the Art of Editing Film*. A&C Black.

Paolucci, P. (2019). Listening To Comics: When Digital Technology Makes the Ninth Art Audible. *Hybrid: Revue des arts et médiations humaines*, (6). https://doi.org/10.4000/hybrid.540

Pusztai, B. (2015). Adapting The Medium: Dynamics Of Intermedial Adaptation in Contemporary Japanese Popular Visual Culture. *Acta Universitatis Sapientiae, Film and Media Studies,* (10), 141–152. https://www.ceeol.com/search/article-detail?id=464729

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. & Sutskever, I. (2021). Learning Transferable Visual Models from Natural Language Supervision. *Proceedings of the 38th International Conference on Machine Learning*. 8748–8763. http://proceedings.mlr.press/v139/radford21a

Salselas, I., Penha, R., & Bernardes, G. (2021). Sound Design Inducing Attention in the Context of Audiovisual Immersive Environments. *Personal and Ubiquitous Computing, 25*(4), 737–748. https://doi.org/10.1007/s00779-020-01386-3

Sharma, M., & Tsuruoka, Y. (2023). Zero-Shot Music Retrieval for Japanese Manga. *Proceedings of the 16th International*

Symposium on Computer Music Multidisciplinary Research, 780–783. https://doi.org/10.5281/zenodo.10114097

Shriram, J., Tapaswi, M., & Alluri, V. (2022). Sonus Texere! Automated Dense Soundtrack Construction for Books using Movie Adaptations. *The Proceedings of the 23rd International Society for Music Information Retrieval Conference.* https://doi.org/10.48550/arXiv.2212.01033

Stopar, K. (2024). Traditional Language with a Flair for Innovation: Hans Florian Zimmer's Compositional Process. *Cinéma & Cie. Film and Media Studies Journal*, *24*(43), 71–91. https://doi.org/10.54103/2036-461X/23785

Summers, T. (2015). 'Sparks of Meaning': Comics, Music and Alan Moore. *Journal of the Royal Musical Association*, *140*(1), 121-162. https://doi:10.1080/02690403.2015.1008865

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks. *Proceedings of the 36th International Conference on Machine Learning, PMLR,* 97, 6105–6114. https://proceedings.mlr.press/v97/tan19a.html?ref=ji

Théodose, R., & Burie, J.-C. (2023). KangaiSet: A Dataset for Visual Emotion Recognition on Manga. In M. Coustaty & A. Fornés (Eds.), *Document Analysis and Recognition – ICDAR 2023 Workshops* (Vol. 14193, pp. 120–134). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-41498-5_9

Ventayen, R. J. M. (2025). AI Music and AI Artists: Understanding What Real Artists Think about the Rise of Digital Artists. Available at SSRN 5747885. https://ssrn.com/abstract=5747885

Wenger, E., & Kenett, Y. N. (2026). Large language models are homogeneously creative. *PNAS Nexus*, *5*(3). https://doi.org/10.1093/pnasnexus/pgag042

Xavier, B. (2025). Biases within AI: Challenging the illusion of neutrality. *AI & SOCIETY*, 40(3), 1545-1546. https://doi.org/10.1007/s00146-024-01985-1

Yang, D., Yu, J., Wang, H., Wang, W., Weng, C., Zou, Y., & Yu, D. (2023). Diffsound: Discrete Diffusion Model for Text-To-Sound Generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31, 1720–1733. https://doi.org/10.1109/TASLP.2023.3268730

Yin, Z., Wang, H., Horio, K., Kawahara, D., & Sekine, S. (2024). Should we respect LLMs? A cross-lingual study on the influence of prompt politeness on LLM performance. *Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)*, 9–35. https://doi.org/10.18653/v1/2024.sicon-1.2

Zacks, J. M., Speer, N. K., & Reynolds, J. R. (2009). Segmentation in Reading and Film Comprehension. *Journal of Experimental Psychology: General*, *138*(2), 307–327. https://doi.org/10.1037/a0015305

Zulkifli, A., & Ahmad, H. A. (2024). Fear Emotion of Reading the Horror Webtoon "Kemala" with Background Music. *Jurnal Minfo Polgan*, *13*(1), 1178–1184. https://doi.org/10.33395/jmp.v13i1.14003