Yewon Kim* KAIST Republic of Korea yewon.e.kim@kaist.ac.kr Sung-Ju Lee KAIST Republic of Korea profsj@kaist.ac.kr Chris Donahue Carnegie Mellon University United States chrisdonahue@cmu.edu

Abstract

Songwriting is often driven by multimodal inspirations, such as imagery, narratives, or existing music, yet songwriters remain unsupported by current music AI systems in incorporating these multimodal inputs into their creative processes. We introduce AMUSE, a songwriting assistant that transforms multimodal (image, text, or audio) inputs into chord progressions that can be seamlessly incorporated into songwriters' creative process. A key feature of AMUSE is its novel method for generating coherent chords that are relevant to music keywords in the absence of datasets with paired examples of multimodal inputs and chords. Specifically, we propose a method that leverages multimodal LLMs to convert multimodal inputs into noisy chord suggestions and uses a unimodal chord model to filter the suggestions. A user study with songwriters shows that AMUSE effectively supports transforming multimodal ideas into coherent musical suggestions, enhancing users' agency and creativity throughout the songwriting process.

CCS Concepts

• Human-centered computing \rightarrow Interactive systems and tools; • Applied computing \rightarrow Sound and music computing.

Keywords

Creativity Support Tool, Music, Songwriting, Human-AI Interaction, Machine Learning

ACM Reference Format:

Yewon Kim, Sung-Ju Lee, and Chris Donahue. 2025. AMUSE: Human-AI Collaborative Songwriting with Multimodal Inspirations. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 28 pages. https://doi.org/10. 1145/3706598.3713818

1 Introduction

Across creative pursuits, artists draw inspiration from a wide variety of concepts and experiences, relying on intuition to find connections between sometimes seemingly unrelated modalities. Within music, for example, Modest Mussorgsky composed *Pictures at an Exhibition* with direct inspiration from a series of paintings, and

*Work mainly done at Carnegie Mellon University as a visiting researcher.

\odot \odot

This work is licensed under a Creative Commons Attribution 4.0 International License. *CHI '25, Yokohama, Japan* © 2025 Copyright held by the owner/author(s). ACM ISBN 979-8-4007-1394-1/25/04

https://doi.org/10.1145/3706598.3713818

Björk's album *Biophilia* was inspired by ruminations on the relationships between music, nature, and technology. Increasingly, AI systems are capable of modeling such *multimodal* relationships, i.e., to understand not only *concrete* relationships between data modalities (e.g., *what objects are in this painting?*) but also more *abstract* ones (e.g., *what music might fit with this painting?*). Consequently, these systems are being applied across a growing range of creative pursuits, such as illustration [59], creative writing [18], and video editing [97].

Within the context of AI support for musical creativity, recent research has centered around two key pathways. First, AI systems have become capable of generating music audio from text descriptions [1, 21, 28, 36, 50, 82], leveraging the particularly intuitive multimodal relationship between music and text to offer users a compelling new form of musical control. However, because these systems output completed songs in audio rather than reusable musical elements (e.g., sequences of notes or chords), they are not particularly supportive to musicians who view iteration as central to both the creative process and to their sense of ownership over the outcome [67]. In contrast, symbolic music generation systems [30, 49, 76, 86, 92] are increasingly being incorporated into music editors to aid musicians in iterative composition workflows by generating reusable musical elements that agree with the surrounding musical context [32, 40]. While musicians find these types of systems helpful, these systems fall short in incorporating multimodal sources of inspiration, as they only take the surrounding musical context as input. Accordingly, musicians remain unsupported in incorporating multimodal sources of inspiration into their iterative creative processes.

In this work, we bridge this gap by building a generative musical support tool that can both incorporate multimodal inputs and facilitate creative iteration. Specifically, we present AMUSE (Figure 1), a songwriting assistant that transforms multimodal user inputs (images, text, audio) into reusable musical elements (chord progressions)¹. Motivated by a formative study with eight songwriters, AMUSE comprises two key functionalities: Chord Generator and Chord Transcriber. In the Chord Generator, we adopt a two-step chord generation approach where we (i) extract music keywords from user multimodal inputs and (ii) generate chord progressions that align with the given music keywords. The Chord Transcriber allows users to transcribe chords from a specified range of audio sources. Amuse is seamlessly integrated into Hookpad [42], a music editor that assists songwriters in the composition of the chords and melody that together characterize the basic essence of a piece of music. In combination with ARIA [32]-a contextual and unimodal AI songwriting assistant already integrated into Hookpad-Amuse

¹Sound examples and code are available at: https://yewon-kim.com/amuse

CHI '25, April 26-May 1, 2025, Yokohama, Japan

Yewon Kim, Sung-Ju Lee, and Chris Donahue



Figure 1: AMUSE transforms multimodal (image, text, or audio) inspirations into reusable musical elements (chord progressions) that songwriters can seamlessly incorporate into their creative process. AMUSE consists of two functionalities: Chord Generator (Left) and Chord Transcriber (Right). In the Chord Generator, user can generate music keywords from image/text inputs and generate musically coherent chord progressions based on the music keywords. These suggestions are generated by rejection-sampling the LLM-generated chord progressions using a unimodal chord model. The Chord Transcriber allows users to transcribe chords from a specified range of audio.

can support songwriters by incorporating multimodal sources of inspiration in addition to existing musical context.

A key challenge in developing AMUSE is the lack of obvious sources of training data consisting of paired multimodal inputs and reusable musical elements (i.e., chords). Text-to-music systems are trained on large corpora of text captions and music audio; however, to the best of our knowledge, no such corpora exist that include reusable musical elements as opposed to music audio. To build such a system without paired training data, we leverage the general capabilities of multimodal large language models (LLMs) by instructing GPT-40 [71] to generate chord progressions that might correspond to multimodal inputs. However, we find that chord progressions from GPT-40 are not musically coherent enough to directly serve users. Equipped with a unimodal dataset consisting only of humancomposed chord progressions, we train a language model prior over chords, which we use to filter out unrealistic chords generated by GPT-40 through a rejection sampling procedure. Through both quantitative and qualitative evaluation, we show that our method can generate diverse chord progressions that are relevant to user keywords and musically coherent.

To understand how AMUSE supports the songwriting process, we conducted a within-subjects study with 10 songwriters, where participants wrote 8-bar choruses based on songwriting prompts. Participants engaged in two conditions: one where they used both AMUSE and ARIA, and another where they used ARIA alone. Our study revealed that AMUSE effectively supports the transformation of multimodal inspirations into concrete musical elements, with participants feeling more guided and aligned with their creative goals than ARIA alone. Notably, although AMUSE was designed for early-stage use, participants employed it across various stages of the songwriting process, demonstrating diverse usage patterns. With AMUSE, participants experienced enhanced agency and creativity throughout the process. Overall, these findings suggest that AMUSE enriches the songwriting process by helping songwriters seamlessly integrate diverse multimodal inspirations into their creative workflows.

In summary, this paper presents the following contributions:

- AMUSE, an interactive songwriting assistant that supports users in transforming multimodal inspirations into reusable musical elements.
- (2) A novel method for generating diverse, relevant, and coherent chord progressions based on multimodal user inputs without paired training data.
- (3) Findings from a user study (N=10) on how and when multimodal inspirations are used throughout the songwriting process and their impact on the creative workflow.

CHI '25, April 26-May 1, 2025, Yokohama, Japan

2 Related Work

2.1 Human-AI Collaborative Music Creation

Research on human-AI collaborative music creation has largely centered on the symbolic music domain, where musical elements such as melodies and chords are represented as a sequence of symbols (e.g., MIDI). Most systems in this symbolic space employ *contextual* models [40, 45, 47, 49, 92], which generate continuation or infillation given user-provided musical elements. For example, Music Transformer [49] extends user-provided MIDI sequences by generating contextually coherent MIDI continuations. In addition, chord recommendation systems such as ChordRipple [46] and Chord-SequenceFactory [37] provide suggestions for chord progressions based on the current harmonic context, helping users build coherent and creative chord sequences over time.

In practice, human-AI collaborative music creation is facilitated by musical user interfaces, such as MIDI editors, where users can input and refine musical elements while collaborating with AI [7, 10, 41, 47, 61, 63, 89]. A notable example is Cococo [61], where users can manually input or edit musical notes, steer the contextual AI model to generate infills for incomplete sections and integrate AI-generated suggestions directly within the editor. These interfaces enable users to iteratively generate and refine AI suggestions, giving them agency in shaping the final output by alternating between manual input and AI suggestions [67]. Commercial tools exemplifying these human-AI collaborative creation interfaces include ARIA [32], which operates unimodally by generating musical elements-chords or melodies-based on user-provided contextual inputs written in the music editor. However, prior interfaces primarily focus on contextual musical inputs without considering other modalities (e.g., images, text) that can influence the creative process. Our study expands this scope by incorporating multimodal inputs into these *contextual* music co-creation workflows.

2.2 Text-to-Music Generative Models

While symbolic music AI focuses on modeling the representation of unimodal music data, one emerging area of research involves text-to-music generative models [1, 21, 28, 36, 50, 82]. These models learn multimodal representation between textual descriptions and music audio, allowing users to steer the model generations by text prompts (e.g., "smooth piano improvisation over a walking bass line"). Similarly, proprietary models like Suno [3] and Udio [93] allow users to input text prompts and full lyrics to generate complete songs. While these text-conditioned generative models offer users greater control and agency over the generation process, a key limitation is that the outputs are fully-rendered audio files, which are not easily integrated into the iterative workflows of music creation, such as those in digital audio workstations (DAWs). In contrast, our work focuses on human-AI collaborative music creation, where AI-generated outputs are meant to be part of an evolving, iterative creation process rather than final products.

2.3 Multimodality in Creativity Support Tools

Multimodal systems have been widely explored in HCI and creativity support tools. In the visual domain, different modalities are used to enhance the creation of both 2D [34, 59, 83] and 3D [13, 14, 23, 60]

visuals. For instance, Opal [59] supports the illustration creation process by incorporating inputs like news articles, keywords, tones, and styles, while in Attribit [14], users can input adjectives and emotional cues to generate 3D shapes. Multimodal systems have also been developed to assist with writing [12, 18, 52, 105]. For example, PandaLens [12] combines audio, gaze patterns, images, and verbal comments to generate narratives, while TaleBrush [18] allows for both text-based and sketch-based interactions in storytelling. Additionally, some writing systems provide multimodal feedback rather than multimodal input to support the creative process [9, 88]. For example, Fairytailor [9] retrieves sequences of images aligned with texts to stimulate ideation in storytelling. Despite these advancements, multimodal music creation remains a relatively unexplored and emerging area in creativity support tools. With AMUSE, we focus on integrating both contextual and multimodal inputs to explore how diverse communication channels can enhance human-AI collaborative music creation.

3 Formative Study

To inform the design of AMUSE, we conducted formative interviews with songwriters with two goals. Firstly, to learn about songwriters' initial ideation and song development process. Secondly, to explore potential AI supports that could enhance the songwriting process.

3.1 Participants and Procedure

We recruited eight participants who are active users of Hookpad [42] and ARIA [6, 32]. Hookpad is a widely-used online songwriting tool developed by Hooktheory [43] that enables users to create songs (chords and melodies) via keyboard or MIDI inputs. ARIA is an AI-powered songwriting assistant integrated into Hookpad, capable of generating melody or chord continuations and infills based on user-provided inputs. We selected these participants because Hookpad supports the earliest stages of songwriting, offering insight into the ideation process, while ARIA allows us to examine how users already incorporate contextual generation into their creative workflows.

All participants identified themselves as hobbyists, with songwriting years ranging from 1.5 years to 5 years (M=3.38, SD=1.19). These participants engaged in various genres of music creation: three participants in pop songs and others in rock, bluegrass, punk, classical, and jazz. Participants reported engaging in songwriting activities at least once a week and using ARIA at least once when writing songs using Hookpad.

During the interview, we asked participants about sources of inspiration and how they translate these inspirations into songs. We also asked their experiences and opinions on songwriting tools they have used (including ARIA) and desired AI functionalities to support songwriting. Participants were compensated with \$25 Amazon gift card for the 1-hour interview. The interviews were conducted online via Zoom in a semi-structured format and were recorded, manually transcribed, and coded through a thematic analysis.

3.2 Findings

3.2.1 **Inspirations from diverse modalities.** All of the songwriters mentioned drawing inspiration from various sources, which can be grouped into three categories: audio, narrative, and visual. Seven out of eight songwriters (P1-3, P5-8) said they often get ideas from listening to music. P3 noted, "I enjoy studying songs that are already proven to be great. I listen to them, take elements I like, and add my own twists." Three songwriters (P4, P7-8) mentioned drawing inspiration from narratives, such as stories and personal experiences. P4 explained, "I always start with a lyrical concept when writing a song. For example, I sometimes randomly turn to a page in a novel I'm reading and use that as the theme for the song." Similarly, P7 found inspiration in personal experiences: "I remember when my dog found a baby bird in a bush. It was a tense moment because we didn't know if my dog would attack the helpless bird. I connected that experience to a song I wrote in the harmonic minor scale." Lastly, two (P2, P8) mentioned being inspired by visuals, such as game scenes or movie aesthetics. P8 noted, "Watching a movie, the aesthetic can make me want to capture that in a song. For example, an action scene might inspire me to create something up-tempo or intense."

3.2.2 Transforming multimodal inspirations into musical elements. Once inspired, songwriters sought to translate their abstract sources of inspiration into concrete musical elements, such as chord progressions or melodies, to build upon. Seven out of eight participants (P2-8) mentioned improvising with instruments or their voices until they found something they liked. For instance, P4 explained, "I have a theme, that is, nostalgic memories triggered by scent. I would sing random melodies or try out many different chord combinations until I found something that matched my song concept." The process of improvisation was not always linear; sometimes, participants improvised first and connected their ideas to their initial inspiration. P7 noted, "I have a Google doc where I dump all my lyrical concepts, but I don't always start from there. Sometimes, I'll improvise on my guitar, come up with chord progressions I like, and begin writing. Later, I often find myself connecting the song to concepts from my list."

Another common approach was **analyzing music through listening sessions.** Four participants (P1, P3, P5-6) mentioned analyzing either existing songs or recordings of their own improvisations. P6 described, "I'll go to YouTube and listen to popular music. I'll note how sharps, flats, and chord progressions are used, to find what sounds pleasing to me. Then I'll transcribe them into Hookpad and expand on it to create something unique."

3.2.3 Seamless incorporation of modular AI suggestions. While all participants emphatically expressed the central importance of the manual aspects of the songwriting process, they also recognized AI tools as inspirational sources, as mentioned by seven participants (P2-8). AI suggestions were viewed as a "great starting point" (P4, P8) or useful for overcoming "writer's block" (P2, P3), as P2 remarked, "It's like having someone next to me, suggesting the next two bars." Participants appreciated AI tools like ARIA for offering modular suggestions that can be easily integrated into the songwriting process, particularly in preserving creativity, agency, and ownership in the process (P2-4). For example, participants described modular suggestions in MIDI format as helpful because "you have control over how and where to use, and you can edit the suggestions" (P4). Generating more than that, such as an entire audio track, was seen as intrusive, as it "takes the joy of songwriting from me" (P2). Similarly, P3 described these modular

suggestions as "*sparks that promote curiosity and creativity*," while generating too much "*dampens my creativity*." In addition, concerns about integrity were raised regarding audio track generations; two (P2-3) expressed fears that AI-generated audio might be "*potentially copied from copyrighted music*" (P3).

However, many participants found **ARIA's suggestions to be lacking context for their specific inspirations** (P1-5, P7-8). P4 explained, "I wish it would generate melodies that better fit my lyrical ideas. Most of the time, ARIA's outputs don't align with my song's theme." One participant (P5) used the text-to-music model, Udio [93], to address this but found it tedious, as it generated full audio track that could not be easily integrated into songwriting interfaces like Hookpad or digital audio workstation (DAW): "I found Udio really good at finding harmonies for specific styles or moods I'm looking for. I generated audio with a prompt 'country pop, uplifting, catchy anthem, introspective mood, love' and took the chord progression from there. But extracting the usable elements was a lot of work."

3.3 Design Goals

Guided by both our formative interviews and prior literature, we establish three Design Goals for our system:

- **DG1:** Incorporate multimodal inspirations. An AI songwriting system can support creative processes by enabling users to incorporate inspirations from diverse sources into their songwriting process. Based on the formative study, sources that users draw inspiration from are visual, narrative, and existing music.
- DG2: Transform abstract ideas into musical elements. The system should support transforming high-level, abstract ideas into concrete musical elements. As informed by prior studies, this transformation process should ensure user control over iterative generation of musical elements [84, 85], while aligning outputs with users' creative intentions [44].
- DG3: Provide modular musical building blocks. Consistent with prior studies [44, 81], participants in our formative study preferred modular, editable outputs that can be seamlessly integrated into their workflow. As such, the system should focus on generating modular musical elements—such as chord progressions—that can be refined and directly integrated into songwriting interfaces (e.g., DAWs).

4 AMUSE: Overview

Based on our design goals, we present AMUSE (Figure 2), a songwriting assistant that transforms high-level multimodal inspirations into chord progressions. AMUSE supports diverse sources of inspiration, including text, images, and audio (DG1), and translates these abstract inputs into coherent chord progressions (DG2) by either keyword-guided chord generation or audio transcription. By focusing on chord progressions as the system outputs (DG3), AMUSE provides users with fundamental components that can be seamlessly integrated into their compositions within the songwriting interface. In this section, we first provide an overview of AMUSE (§4.1), explain the interactions and technical details within AMUSE's interface (§4.2-§4.3), and implementation (§4.4).



A Hookpad Interface

Figure 2: Screenshot of the songwriting interface used in the user study. The main workspace is the Hookpad interface (A), where users can input melodies and chords using either the keyboard or MIDI devices. AMUSE (B), a Chrome extension, appears as a floating window within the Hookpad interface. Users can freely open, close, move, and resize this window. The current view of AMUSE displays the Chord Generator, which generates chord progressions from user-provided images or text. Participants also used ARIA (C), a tool for generating melodies and chords based on the content already written in the Hookpad interface.

4.1 System Overview

AMUSE is built as a Chrome extension that works on top of Hookpad [42], a widely-used online songwriting tool developed by Hooktheory [43] (see Figure 2). The decision to develop AMUSE as an companion tool to Hookpad rather than a standalone application was driven by the robust songwriting functionalities already present in Hookpad, such as MIDI input, playback, and instrumentation options, making it an efficient platform to build upon without the need to duplicate existing functionalities. Furthermore, Hookpad includes ARIA [32, 92], an AI songwriting assistant that generates melody/chord continuations or infills based on user-provided melodies and chords. ARIA is representative of an emerging family of contextual AI music assistants [47, 61, 62], i.e., ones that generate musical outputs based on existing musical material. In contrast, AMUSE is designed to facilitate multimodal transformations, complementary to the contextual inputs to ARIA. By integrating AMUSE with Hookpad and ARIA, we create a unique environment to study the interaction between these different forms of music AI assistance.

This setup not only allows us to explore the synergies that AMUSE brings to the previous songwriting interfaces but also provides valuable insights into their collaborative potential in enhancing the songwriting process.

AMUSE comprises two main features: Chord Generator (§4.2) and Chord Transcriber (§4.3). Chord Generator processes text and image inputs, extracting relevant musical keywords that users can select and refine. These keywords are then used to guide the generation of chord progressions that align with the mood and concepts embodied by the keywords. On the other hand, Chord Transcriber accepts audio inputs and transcribes them into chord progressions. When users select (click) chords suggested by AMUSE, they are instantly integrated into the Hookpad interface, where they can be played back and further refined. Users can change instrumentation, timbre, and play the sounds within the Hookpad interface, and continue songwriting based on AMUSE's suggestions.

CHI '25, April 26-May 1, 2025, Yokohama, Japan



Figure 3: Overview of Chord Generator in AMUSE. (A) Initial Interface: Users can upload an image or type text, which are used to generate music keywords. Users can also directly write music keywords in the keyword editor. (B) Keyword Extraction: Upon clicking the "Generate Keywords" button, AMUSE suggests music keywords based on the multimodal inputs. User-selected keywords are automatically pasted into the keyword editor ('acoustic, 'mellow,' and 'indie-pop' in the figure). (C) Keyword-based Chord Generation: Upon clicking the 'Generate Chords' button, AMUSE suggests four chord progressions based on the keywords. Users can choose between 3-bar or 4-bar progressions (default: 4 bars). The key is automatically detected from the song configuration in Hookpad (G Maj in the figure). Clicking a chord progression automatically pastes it into Hookpad (in the figure, 'Amaj7-Em7-A7-Dmaj7' and 'Bm9-E7-C#m7-F#m7' are selected), where users can play the audio and make further edits.

4.2 Chord Generator

4.2.1 Chord Generator Interface. The Chord Generator interface is illustrated in Figure 3. Upon opening the Chord Generator, users are prompted to provide an image, text input, or both, which are then used to extract relevant music keywords (Figure 3A). When users click the 'Generate Keywords' button, the system suggests keywords based on the multimodal inputs (Figure 3B). Users can select from the suggested keywords, with selected keywords automatically inserted into the keyword editor for further refinement. After finalizing the keywords, users can click the 'Generate Chords' button to receive chord progressions that align with the selected music keywords (Figure 3C). The interface reads the key configuration (e.g., G Maj) in the Hookpad editor interface and generates the chord progressions in the specified key. Users can also specify the number of bars (3 or 4) for the chord progressions. Upon clicking the button, AMUSE generates and displays four chord progressions, drawing on prior research that demonstrates multiple suggestions enhance creativity and are preferred in ideation tasks [11, 26, 55, 61]. We specifically provide four suggestions per query, following previous work on human-AI collaborative music creation system [61]. By

clicking on any of the suggestions, the chosen progression is automatically inserted into the Hookpad editor, where users can further modify or play the chords to integrate them into their compositions.

4.2.2 Keyword-based Chord Generation Pipeline. We adopted a keyword-based interaction framework between multimodal inputs and chord progressions to help users create meaningful semantic connections between two modalities that are otherwise difficult to link due to their abstract and morphologically different nature [88]. Keywords serve as an intermediary communication layer, inspired by practices in creativity support tools in the vision domain [17, 59, 60]. To extract music-related keywords that align with user inputs, we leverage the multimodal LLM, GPT-40 [71], instructing it with a list of music styles and genre-specific keywords crawled from a music keyword wiki [90]. The prompt used for keyword generation can be found in Appendix A.

For keyword-based chord generation, to ensure that the generated chords are diverse, musically coherent, and relevant to input keywords, we developed a rejection-sampling-based chord generation method that integrates an LLM with a language model prior trained over chord data. The detailed algorithm and its implementation are discussed in §5.



Figure 4: Overview of Chord Transcriber in AMUSE. (A) Initial Interface: Users can upload a local audio file or enter a YouTube URL. (B) Audio Inspiration Input: With an audio preview, users can select the desired segment for transcription by specifying start and end times (maximum 30 seconds). (C) Chord Transcription: AMUSE detects the key (shown as Gb Min) and chords of the selected audio segment. Users can play the audio segment and the chord is highlighted in sync with the playback (in this figure, C#m/E). Clicking on a chord pastes it into the Hookpad interface. If the "Convert to Hookpad Key" option is checked, the chords are transposed to match the key the user is working on (G Maj in the figure).

4.3 Chord Transcriber

The Chord Transcriber interface is illustrated in Figure 4. Users can either upload a local audio file or enter a YouTube URL for transcription (Figure 4A). Once the audio source is uploaded, users can preview the audio and set the transcription window, with a maximum length of 30 seconds (Figure 4B). Upon clicking the 'Transcribe' button, AMUSE processes the request with a 10-15 seconds delay and displays the detected key and chords for the selected audio segment (Figure 4C). The transcribed audio can be played back with the corresponding chords highlighted in sync with the playback. Users can click on any chord to automatically paste it into the Hookpad editor for further refinement. To accommodate key differences between the detected key and the key the user is working in, the 'Convert to Hookpad Key' option allows for automatic transposition of the chords to match the current working key in Hookpad. For chord transcription, we use an established chord transcription API provided by Music AI, a company that creates and provides AI platforms for music and audio production [2].

4.4 Implementation

We implemented the frontend of AMUSE (Chrome extension) in JavaScript and CSS. The backend was implemented as a Flask server, and we call the OpenAI API [72] for all LLM-based functionality (multimodal inputs to keywords, keywords to chords). Regarding the LLM configurations, we use gpt-4o-2024-05-13 and set the temperature to 1.0 for all components to promote creativity and diversity in generations.

5 Keyword-Based Chord Progression Generation

Directly generating chord progressions conditioned on music keywords presents a challenge due to the lack of paired training data linking music keywords to chord progressions. To address this issue, we propose a method that combines the general capabilities of LLMs and a unimodal chord generation model specifically trained on chord data. We first introduce the generation goals we aim to achieve (§5.1). We then detail the components of our pipeline designed to address these goals, including a prompting technique for diverse outputs (§5.2) and a rejection sampling methods for contextually coherent outputs (§5.3). Finally, we describe the implementation of our chord progression generation method deployed in AMUSE (§5.4).

5.1 Generation Goals

When generating chord progressions from user-provided keywords, we consider three desiderata: **diversity**, **relevance**, and **coherence**.

Diversity ensures that users receive a variety of options, as offering multiple suggestions has been shown to enhance creativity [55] and is often preferred in ideation tasks [11, 26, 61]. **Relevance** ensures that the generated chord progressions closely align with the user-provided keywords, reflecting the musical qualities or moods the user intends to capture. **Coherence** ensures that the chord progressions are musically coherent (i.e., follow real music data distribution). Ensuring both relevance and coherence is crucial as misaligned or low-quality suggestions can negatively impact user experience in creativity support tools [19, 38, 55, 64].

Formally, let $P(\mathbf{x}|\mathbf{c})$ be a target conditional distribution of chord progression x capable of generating diverse, relevant, and coherent suggestions given keywords control **c**. Modeling $P(\mathbf{x}|\mathbf{c})$ is limited by the scarcity of music data that contains paired keywords and chord progressions. As a result, we aim to find a good approximation of $P(\mathbf{x}|\mathbf{c})$ by using an *unpaired* chord progression dataset $\mathcal{D} = {\mathbf{x}_i}_{i=1}^N$ and an LLM (e.g., GPT-40 [71]) that has a general understanding of language and general music-related context but lacks the direct knowledge of harmonic structures seen in real-world music data (as it is not explicitly trained on music data). Specifically, we define $Q(\mathbf{x}|\mathbf{c})$ as an LLM that generates chord progressions \mathbf{x} given keywords c as prompts. We select an instruction-tuned LLM (GPT-40 [71]) for $Q(\mathbf{x}|\mathbf{c})$ due to its enhanced ability to follow user instructions that are critical for generating contextually relevant chord progressions from music keywords. These models are pretrained on large-scale text data from public webpages spanning diverse topics [70], which include music-related content (for instance, chord transcriptions [91] or music theory blogs [94]). As such, we assume the model has a general understanding of music and can infer relationships between musical elements (e.g., chords) and high-level descriptions (e.g., mood or style). While $Q(\mathbf{x}|\mathbf{c})$ is capable of keyword-controlled generation, we find this leads to a very sharp distribution (i.e., lacks diversity; we later show in our experiments that its outputs exhibit low diversity). We mitigate this lack of diversity by introducing a prompting technique for LLMs to generate diverse samples (§5.2). We also find the distribution $Q(\mathbf{x}|\mathbf{c})$ inherently has low coherence (i.e., deviates from real music distribution) as LLMs are not explicitly trained on music data. We thus present a rejection sampling strategy using a unimodal chord generation model $P(\mathbf{x})$ to select coherent outputs (§5.3). Based on these techniques, we describe the implementation of our chord progression generation process deployed in AMUSE (§5.4).

5.2 Prompting Technique For Diverse Chord Progressions

In creative tasks, generating diverse suggestions with LLMs typically involves multiple queries to the model [11, 26, 56]. However, instruction-tuned LLMs such as InstructGPT [73] and GPT-40 [71] are prone to a homogenization effect, where multiple queries often yield repetitive responses [5, 33, 74, 107]. This is particularly problematic in our context, where generating diverse outputs is crucial for fostering creativity.

To address this challenge of generating diverse chord progressions with instruction-tuned LLMs, we employ a prompting technique that instructs an LLM to generate a large collection of chord progressions. Our prompt takes user-written keywords, key (e.g., C maj) and bar counts (e.g., 4 for 4-bar chord progression) as inputs. The prompt then instructs the model $Q(\mathbf{x}|\mathbf{c})$ to generate *N* diverse chord progressions that reflect the mood and concept of the user keywords. To enhance diversity, the prompt contains instructions to vary chord components (e.g., root, quality, extensions, alterations), progression patterns (e.g., diatonic, chromatic), and cadences. The full prompt is detailed in Appendix A.2.

5.3 Chord Progression Generation By Rejection Sampling

Given *N* diverse chord progressions generated by $Q(\mathbf{x}|\mathbf{c})$ (§5.2), the most naive approach would be randomly sampling four progressions and showing them to the users. However, this is not the best approach as $Q(\mathbf{x}|\mathbf{c})$ generations could be less coherent, i.e., do not align with the statistical properties of chord progressions in real music data, as these language models are not explicitly trained on music data [70]. Consequently, relying solely on $Q(\mathbf{x}|\mathbf{c})$ for chord generation may produce noisy or less conventional progressions that do not meet the musical expectations of users. Therefore, our aim is to filter and select from the generations that are musically coherent, i.e., following real-world music data distribution. To achieve this, we employ rejection sampling [66], a Monte Carlo algorithm that allows us to sample from a target distribution that may be difficult to sample from directly (i.e., $P(\mathbf{x}|\mathbf{c})$), by using a proposal distribution that is easier to sample from (i.e., $Q(\mathbf{x}|\mathbf{c})$).

Method. We perform rejection sampling [66] to improve the coherence of $Q(\mathbf{x}|\mathbf{c})$:

$$u \sim \text{Unif}(0, 1)$$
, accept if $u < \frac{P(\mathbf{x}|\mathbf{c})}{M \cdot Q(\mathbf{x}|\mathbf{c})}$ else reject,

where M > 0 is a constant that ensures the inequality holds for all samples and scales the acceptance ratio. M is chosen such that $M \ge \max_{\mathbf{x}} \left(\frac{P(\mathbf{x}|\mathbf{c})}{Q(\mathbf{x}|\mathbf{c})} \right)$, ensuring that the acceptance rate remains valid across the entire sampling process.

Since we do not know ground-truth $P(\mathbf{x}|\mathbf{c})$, as well as the value of $Q(\mathbf{x}|\mathbf{c})$ (we can only sample from this distribution), we consider the following alternative using Bayes Rule:

$$\frac{P(\mathbf{x}|\mathbf{c})}{M \cdot Q(\mathbf{x}|\mathbf{c})} = \frac{P(\mathbf{c}|\mathbf{x})P(\mathbf{x})Q(\mathbf{c})}{M \cdot Q(\mathbf{c}|\mathbf{x})Q(\mathbf{x})P(\mathbf{c})} = \frac{P(\mathbf{c}|\mathbf{x})P(\mathbf{x})}{M \cdot Q(\mathbf{c}|\mathbf{x})Q(\mathbf{x})} \approx \frac{P(\mathbf{x})}{M \cdot Q(\mathbf{x})}$$

since $P(\mathbf{c}) = Q(\mathbf{c})$ as they are the same predefined keyword distributions and we assume $P(\mathbf{c}|\mathbf{x}) \approx Q(\mathbf{c}|\mathbf{x})$ as the generated progression \mathbf{x} from LLMs closely align with \mathbf{c} . Thus, we perform rejection sampling by calculating the ratio $\frac{P(\mathbf{x})}{M \cdot Q(\mathbf{x})}$, where both $P(\mathbf{x})$ and $Q(\mathbf{x})$ can be learned by training two deep neural networks with the dataset \mathcal{D} and a large collection of generated chord from LLMs, respectively.

Implementation. For $P(\mathbf{x})$ and $Q(\mathbf{x})$, we train two Long Short-Term Memory (LSTM) models. Specifically, for $P(\mathbf{x})$, we use the HookTheory dataset [31], which comprises 50 hours of melody and chord progression annotations derived from Hooktheory's TheoryTab database [27]. This dataset spans a wide range of genres, including pop, rock, EDM, jazz, and classical. For training $Q(\mathbf{x})$, we use chord progressions generated by GPT-40 [71]. We use the same prompt used for generating diverse chord suggestions (§5.2). To

Algorithm 1 Chord Progression Generation Pipeline in AMUSE

1: Input: Keywords c 2: Output: Accepted chord progressions Y 3: $\mathbf{X} \leftarrow \text{Candidate chord progressions } \{\mathbf{x}_i\}_{i=1}^N \text{ from } Q(\mathbf{x}|\mathbf{c})$ 4: $Y \leftarrow \emptyset$ 5: **for x** in **X** do Sample $u \sim \text{Unif}(0, 1)$ 6: if $u < \frac{P(\mathbf{x})}{MQ(\mathbf{x})}$ then ▹ Compute acceptance probability 7: $Y \leftarrow \widetilde{Y} \cup \{x\}$ \triangleright Accept the generated chord \mathbf{x}_i 8: end if 9: 10: end for 11: if |Y| < 4 then $\mathbf{Y} \leftarrow \mathbf{Y} \cup \{ \text{top-}k(\mathbf{x}_i : \frac{P(\mathbf{x}_i)}{MQ(\mathbf{x}_i)}) \}, \text{ where } k = 4 - |\mathbf{Y}|$ $\triangleright \text{ Add top-}k \text{ samples if } \mathbf{Y} \text{ is smaller than the target size}$ 12: 13: end if 14: Return Y

marginalize the influence of specific keywords during generation (i.e., generalize across diverse keywords), we randomly sample keywords from the music keyword wiki [90] for each generation. This process resulted in a total of 25,000 4-bar chord progressions with keywords abstracted into the underlying distribution. After fitting these models, the constant *M* in the rejection sampling algorithm was determined by calculating the ratio $\frac{P(\mathbf{x})}{Q(\mathbf{x})}$ over all chord progressions generated by GPT-40. We selected the 95th percentile value to avoid the influence of extreme outliers, resulting in M = 7.64. We provide training details and hyperparameter values in Appendix B.

5.4 Deploying Chord Progression Generation Pipeline in Amuse

The obtained constant M = 7.64 and the trained networks, $P(\mathbf{x})$ and $Q(\mathbf{x})$, are deployed on the backend server to sample LLMgenerated chord progressions. To approximately match four chord progressions, we set the number of chord generations to N = 30 in the deployment $(30/M \approx 4)$. During this process, each progression generated by $Q(\mathbf{x}|\mathbf{c})$ is assessed against the acceptance threshold determined by our rejection sampling method. To ensure we present at least four samples to users, if fewer than four samples are accepted, we select the top-k samples, k = 4 - *num accepted samples*. These samples are ranked by their acceptance probability, $\frac{\tilde{p}(\mathbf{x})}{M\tilde{q}(\mathbf{x})}$. The full sampling process is detailed in Algorithm 1.

6 Technical Evaluation

We conduct a technical evaluation of our chord progression generation approach. We assess whether our prompting technique results in **diverse** chord progressions (§6.1) as well as whether our rejection sampling can sample chord progressions that are **relevant** to keywords and musically **coherent** (§6.2).

6.1 Evaluation of Prompting Technique

To assess the effectiveness of our prompting technique (§5.2) in generating more **diverse** chord progressions, we conduct an automatic evaluation comparing our approach against a conventional method of generating multiple suggestions [11, 26, 56].

Study Setup. We compare the diversity of chord progressions produced by our prompting technique with those generated by the conventional method (*baseline*), which involves querying a model multiple times to generate individual suggestions. In the baseline method, a single 4-bar chord progression is generated per query and repeated 30 times, resulting in a set of 30 4-bar chord progressions. The prompt for the baseline is similar to ours but lacks instructions to generate diverse progressions in a batch (See the full prompt in Appendix A.3). In contrast, our technique generates all 30 4-bar chord progressions in a single batch using a single prompt.

To quantify the diversity of chord progressions, we employ Self-BLEU [108] as our metric. Self-BLEU score measures the diversity of the generated data, with a higher Self-BLEU score indicating less diversity among the chord progressions. Specifically, for each set of 30 4-bar chord progressions, we compute the BLEU score [75] for each individual progression by treating it as the hypothesis and the remaining 29 progressions as references. The Self-BLEU score is then obtained by averaging the BLEU scores across all 30 progressions in the set. This process is repeated across 100 distinct pairs of sets (i.e., 100 sets of 30 4-bar chord progressions generated by our technique and 100 corresponding sets generated by the baseline method). All chord progressions are generated in C, and the music keywords are randomly sampled from the music keyword wiki [90] for each generation. We report the average Self-BLEU score across these 100 pairs for each condition.

Results. Table 1 shows the result for the diversity. On average, our prompting technique (AMUSE) results in a lower Self-BLEU score compared to the conventional method (Baseline). This indicates that the chord progressions generated by our technique are more diverse than those generated by the baseline method.

6.2 Evaluation of Rejection Sampling Approach

We evaluate the **coherence** and **relevance** of sampled chord progressions generated through our rejection sampling approach (§5.3). Specifically, our goal is to assess whether the resulting progressions are (i) musically coherent and (ii) relevant to keywords. We first run an automatic evaluation to compare the distribution of the sampled progressions with that of a real music dataset [43], where we interpret closer match as greater musical coherence. We then perform a listening study with musicians to assess both the musical coherence and keyword relevance of the progressions.

6.2.1 Automatic Evaluation.

Study Setup. The aim of rejection sampling is to align LLMgenerated chord progressions with real music data distribution. We quantitatively assess this alignment by computing the Jensen-Shannon Divergence (JSD) [58] between the distributions of the generated chord progressions and real music data. JSD measures the similarity between two probability distributions, with values ranging from 0 (perfect similarity) to 1 (complete divergence). Intuitively, it captures how much the chord distributions from the model differ from those found in real music, offering insight into how well the generated chord progressions reflect real-world harmonic patterns. Specifically, we compare unigram and bigram distributions of chord tokens and use them to calculate JSDs. We use the HookTheory dataset [31] as the real music dataset. We pre-process the chords Method

Baseline

AMUSE

Table 1: AMUSE offers more diverse chord suggestions than the baseline by instructing an LLM to generate diverse chords in a batch, as indicated by the Self-BLEU [108] scores for chord progression generations. ↓ indicates lower values are better. Each value represents the Mean±Std score across 100 chord progression sets, where each set contains 30 chord progressions. Table 2: AMUSE generates keyword-conditioned chord progressions that are more coherent, i.e., closer to real music data distributions, as indicated by Jensen-Shannon Divergence (JSD) [58] between each chord generation method and music data [31]. \downarrow indicates that lower values are better. We compute JSD for unigram/bigram chord distributions. LSTM Prior refers to chords generated by an LSTM trained on music data ($P(\mathbf{x})$). GPT-40 refers to chords generated by GPT-40 ($Q(\mathbf{x}|\mathbf{c})$) with keywords marginalized. Amuse represents rejection-sampled GPT-40 using LSTM Prior.

Method	Keyword Conditional	Unigram \downarrow	Bigram↓
LSTM Prior (N=25,000)	No	0.15	0.30
GPT-40 (N=25,000)	Yes	0.42	0.57
Amuse (N=3,242)	Yes	0.27	0.46

and get N = 25,601 chord progressions, all transposed into C. For generating chord progressions with GPT-40, we randomly sampled music keywords from the music keyword wiki [90] to marginalize the effect of specific keyword choices. We calculate the JSD between the real music data and three different chord generation conditions:

Self-BLEU ↓

0.61±0.18

 0.30 ± 0.12

- LSTM Prior (N=25,000): Chord progressions generated by an LSTM model trained on the HookTheory data (P(x)).
- (2) GPT-40 (N=25,000): Chord progressions generated by GPT-40 (Q(**x**|**c**)) with keyword marginalized (i.e., prompted with random keywords).
- (3) AMUSE (N=3,242): Rejection-sampled GPT-40 using P(x) (LSTM Prior) and Q(x).

Results. Table 2 presents the alignment of each distribution with real music data, as measured by JSD. The results indicate that applying rejection sampling improves the alignment of the chord progressions generated by GPT-40 with the real music data distribution from the Hooktheory dataset. While the initial LLM-generated chord progressions (GPT-40) exhibited notable divergence from real-world chord patterns, the use of rejection sampling reduced this discrepancy.

6.2.2 Human Evaluation: Listening Study.

Study Setup. We conducted a listening study to further evaluate the qualitative performance of our rejection sampling approach against two baselines. We evaluated two key aspects of the generated chord progressions: (i) musical coherence and (ii) keyword relevance. Following the prior work [29], listeners were presented with pairs of 7-10s chord progression audio clips, each generated by different methods (ours or baselines).

The study comprised two types of tasks: First, participants listened to a pair of 7s 4-bar chord progressions played with piano sounds and were asked to indicate which of the two "sounds more pleasant and natural" (musical coherence). Second, participants were presented with a set of keywords and a pair of 7-10s, 4-bar chord progressions, with instrumentation and BPM adjusted to align with the keyword characteristics. They were then asked to indicate which of the two "better reflects the mood or concept of the given keywords" (keyword relevance).

We generated chord progressions for 10 keyword sets under three conditions: (i) our rejection sampling approach (AMUSE), (ii) keyword-conditioned LLM (GPT-40), and (iii) LSTM trained on music data without keyword conditioning (LSTM Prior). For AMUSE, we generated 30 4-bar chord progressions per keyword set using GPT-40, applied rejection sampling, and randomly selected one from the accepted samples. For GPT-40, we generated 30 4-bar chord progressions per keyword set and randomly selected one. For LSTM Prior, we generated a 4-bar chord progression using the LSTM model trained on HookTheory data [31] ($P(\mathbf{x})$). Each keyword set was processed five times, resulting in 10 keyword sets × 5 questions × $_{3}C_{2}$ pairs = 150 comparisons. Each task was evaluated by three unique listeners, yielding a total of 450 comparisons for each task (N=900 in total). Since the Shapiro-Wilk test indicated that the data was non-parametric, we analyzed these pairwise judgments using the Wilcoxon signed-rank test [20].

We recruited 45 unique listeners from Prolific [78], targeting individuals with a primary interest in music and 5+ years of experience in musical instruments. Participants were required to have audio devices and confirm they were in a quiet setting before starting the survey. Each participant answered 22 questions (10 on musical coherence, 10 on keyword relevance, and 2 attention checks) and was compensated £2, which corresponds to an approximate hourly rate of 12.13 USD.²

Results. Results for systems appear in Figure 5. For Musical Coherence (Figure 5a), our method (AMUSE) shows no statistical difference compared to LSTM Prior, which is the closest to real music distribution. Conversely, the analysis reveals that users did not prefer samples generated by GPT-40 in terms of musical coherence, with both LSTM Prior and AMUSE being significantly more preferred than GPT-40 (LSTM Prior > GPT-40: p=2.18e-5; AMUSE > GPT-40: p=0.009). For Keyword Relevance (Figure 5b), our system (AMUSE) was the most preferred by users (58% of cases against any other samples), significantly outperforming LSTM Prior (p=0.006).

This result, combined with our quantitative findings on diversity and coherence, suggests that AMUSE successfully achieves the desiderata we defined; our proposed method generates **diverse** and **coherent** chords that are **relevant** to the user input keywords.

²Among the 49 participants initially recruited, four failed the attention checks and were excluded from the analysis.



Figure 5: Results from our listening study where listeners indicated a preference between pairs of chord progression audio clips generated by two different methods among LSTM Prior, GPT-40, or Amuse. Each row indicates the % of times listeners preferred audio from that system compared to those from any other system (first column, N=300) and each system individually (other columns, N=150). Wilcoxon signed-rank test with these paired data reveals that for Musical Coherence, AMUSE shows no significant difference compared to LSTM Prior, which aligns closely with real music distributions (55.3% of LSTM Prior generations preferred over AMUSE). For Keyword Relevance, AMUSE is most preferred by users (58% of cases against any other samples), significantly outperforming LSTM Prior. (**p<.001, ***p<.001).

7 User Study

We conducted a user study to gain insights and feedback on the potential, limitations, and future opportunities of multimodal music generation tools to support songwriting creativity. This overarching goal broke down into three main research questions:

- **RQ1:** How does AMUSE support songwriters transforming multimodal inspirations into contextual musical elements?
- **RQ2:** How do songwriters incorporate AMUSE into their creative workflows?
- **RQ3:** How does AMUSE impact the songwriting experience and perceived quality of the compositions?

7.1 Participants

We recruited 10 songwriters through targeted email outreach to active Hooktheory users, ensuring that our study included genuine songwriters actively engaging with the platform. We collaborated directly with Hooktheory to reach out to the users. Eight identified themselves as hobbyists (i.e., write songs primarily for personal enjoyment) and two as professional (i.e., write songs as part of a career or generate income from songwriting). All were regular users of Hookpad and engaged in songwriting activity at least once a week. Six participants were regular users of ARIA and four had not used ARIA before. Participants composed in several different genres, including pop, rock, and blues. We detailed participants in Table 3. Participants were compensated with 50 USD Amazon gift card for the 2-hour study.

7.2 Study Procedure

To connect with real-world Hooktheory users across the globe, we conducted a remote study using Zoom. During the study, participants performed two songwriting tasks under different conditions: Table 3: Detailed background information of the participants in the user study (Section 7). For songwriting proficiency, we use self-reported levels. Hobbyist = I write songs primarily for personal enjoyment; Professional = I write songs as part of my career or to generate income from my songwriting.

Participant ID	Songwriting Proficiency	Songwriting Genre
P1	Hobbyist	Blues, Rock, Americana
P2	Hobbyist	Acoustic Pop, Symphonic
P3	Hobbyist	New Age, Classical
P4	Hobbyist	Indie, Alternative, Pop
P5	Hobbyist	Rock, Pop, Country
P6	Hobbyist	Рор
P7	Hobbyist	Pop, Jazz Pop
P8	Hobbyist	Rock, Pop
P9	Professional	Pop, R&B, Hip Hop
P10	Professional	Rock

Baseline and Assist. In the Baseline condition, participants were allowed to use only ARIA (reflects the *existing* AI-assisted songwriting practice), whereas in the Assist condition, participants had access to both AMUSE and ARIA (introduces *multimodal* inspiration support in addition to the conventional AI-assisted songwriting workflow). We opted not to compare to the manual composition without ARIA because our primary goal was to evaluate the added benefit of AMUSE in enhancing existing AI-assisted workflows rather than assessing AI versus non-AI approaches. The task for each session was to write an 8-bar chorus (chords and melodies) given one of two songwriting prompts: "Write an 8-bar chorus about the beginning of an unexpected friendship you had in your life" or "Write an 8-bar chorus about your favorite summer



Figure 6: Overview of the user study procedure. The study involved two songwriting sessions, each using a different set of tools. In each session, participants wrote an 8-bar chorus based on a given songwriting prompt. Following the sessions, participants took part in 25-minute semi-structured interviews about their experiences.

holiday memory." We adapted these prompts from the Songwriting subreddit [80]. Both task prompts and system conditions were randomized across the sessions to minimize the ordering bias.

The overall study procedure is illustrated in Figure 6. We started with an introduction and tool tutorial session, where users were given 20 minutes to familiarize themselves with AMUSE and ARIA. For each songwriting session, participants were given the songwriting prompt and had 25 minutes to write an 8-bar chorus using the assigned tool(s) with their screen shared. We asked participants to think aloud to learn their decision-making processes and reactions when interacting with the tools. The whole session, including the shared screen, was video recorded. Since AMUSE involves image inputs, we asked participants not to share any personal or sensitive pictures during the study. To ensure meaningful comparison between conditions, we required participants to use the assigned tool(s) at least once during the task; however, they had the flexibility to decide when and how to use the features, as well as how frequently to engage with them.

7.3 Measures

7.3.1 Questionnaires. We analyzed participants' responses to the two post-task surveys. These surveys asked participants to rate, on a seven-point Likert scale, the usefulness of the given systems (Baseline: 'ARIA alone'; Assist: 'both AMUSE and ARIA') in helping transform their inspirations into music and writing songs aligned with the given song prompts. Additionally, participants were asked to assess their self-perceived experience with the system using questions adapted from related works [62, 102]. We also collected participants' ratings on five aspects of the Creativity Support Index (CSI) [16] questionnaire, excluding the "Collaboration" question, as it is irrelevant to human-AI collaborations. We analyzed these Likert scale ratings using the Wilcoxon signed-rank test, as a Shapiro-Wilk test indicated the data was non-parametric [20]. We include the detailed survey questions in Appendix C.1.

7.3.2 *Interviews.* During the 25-minute semi-structured interviews, participants were asked about the difference between the two song-writing sessions and the perceived impact of the tools on their songwriting tasks. Detailed interview questions can be found in Appendix C.2.

We transcribed the semi-structured interviews and analyzed them using the constant comparative method [39]. Two authors independently open-coded two interview samples to identify key concepts and patterns. Axial coding was then performed to link these patterns [22], resulting in an initial codebook. The first author coded the remaining interview data while continuously refining the codebook. After the first coding round, another author coded two interview samples using the updated codebook for verification, and any issues were resolved through discussion. Throughout the process, the research team regularly discussed emerging themes while triangulating the interview data with quantitative analyses. We include the full codebook in Appendix C.3.

7.3.3 Interaction Logs. We collected the usage logs (i.e., participant actions with timestamps) to quantitatively analyze participant behaviors. We used this data to obtain statistics on tool usage, such as the time taken for each session, the number of keyword/chord generations in AMUSE, and the number of chord/melody generations in ARIA. For all these measures, we conducted a Shapiro-Wilk test to determine if the data was parametric and used a paired t-test for parametric data and a Wilcoxon signed-ranked test for nonparametric data. To qualitatively explain the user behaviors identified from the quantitative data, we analyzed screen recordings when necessary.

8 Findings

We present our findings on how participants transformed their multimodal inspirations into musical elements in §8.1 (RQ1), how AMUSE influenced participants' songwriting processes in §8.2 (RQ2), and how AMUSE impacted the songwriting experience and perceived quality of outcomes in §8.3 (RQ3). We summarize our main findings below:

- **RQ1:** Amuse effectively supports the transformation of multimodal inspirations into concrete musical elements, with participants feeling more guided and aligned with their creative goals compared to using ARIA alone.
- **RQ2:** Participants demonstrated diverse usage patterns of Amuse in the songwriting process, integrating it in various stages of their songwriting process.
- **RQ3:** With AMUSE, participants experienced enhanced agency and creativity throughout the songwriting process and perceived it as more efficient and easier. However, they found the compositions in both the Baseline and Assist conditions to be equally satisfactory.

8.1 Supporting Multimodal Inspirations

We first examine whether AMUSE successfully supports songwriters transforming multimodal inspirations into contextual musical elements (RQ1). Specifically, we present the effectiveness of AMUSE in the inspiration transformation process (§8.1.1) and the perceived

quality of suggestions (§8.1.2). We also report the participants' preferences for the Chord Generator over Chord Transcriber within AMUSE (§8.1.3) and the impact of music keywords in the system usage (§8.1.4). We describe relevant quantitative findings with qualitative insights.

8.1.1 Effective transformation of multimodal inspirations into musical elements. Overall, participants in Assist condition felt they could more easily transform their initial inspirations into concrete musical elements than Baseline (Inspiration Support in Figure 7; Assist=6.20±0.75, Baseline=4.60±1.36, z=0.00, p<0.01). Additionally, participants reported feeling more guided toward the task goal in Assist (Task Alignment in Figure 7; Assist=6.10±0.70, Baseline= 4.30 ± 1.62 , z=0.00, p=0.016). Six participants emphasized how Amuse helped better achieve their creative goals. Amuse's suggestions "fell in line with what I would have mentally constructed" (P5), allowing them to "focus more on the goal with a clear vision of how to write the rest of the song" (P2). In contrast, in Baseline condition, five participants mentioned feeling unsupported in achieving the task, highlighting a lack of direction. P8 stated, "Using ARIA alone, I had a very vague vision and didn't really know what I was going for." Similarly, P1 noted that without AMUSE, "the song ended up being the opposite of the style I originally intended to create with my lyrics".

8.1.2 Higher perceived quality of suggestions from Assist. While participants expressed satisfaction with the diversity of suggestions in both conditions (Diversity in Figure 7; Assist=6.10±0.94, Baseline=5.50±1.50; z=12.00, p=0.389), participants rated the quality of the outputs they received from Assist significantly higher than those they got from Baseline (Output Quality in Figure 7; Assist=6.40±0.66, Baseline=5.30±1.00, z=4.00, p=0.023). Interestingly, there was no significant difference in the average number of ARIA queries made between the conditions (Baseline= 3.60 ± 2.72 , Assist=2.30±1.16; z=10.0, p=0.130) or in the acceptance rates of those queries (Baseline=0.71±0.25, Assist=0.87±0.22, z=7.0, p=0.12). This similarity suggests that the improved perceptions of output quality may be more related to AMUSE's ability to generate suggestions that align with user intentions, leading to increased general satisfaction with the suggestions received throughout the songwriting process.

8.1.3 Preference for Chord Generator over Chord Transcriber. Participants demonstrated a clear preference for the Chord Generator over the Chord Transcriber within AMUSE. All participants utilized the Chord Generator, whereas only three used the Chord Transcriber in Assist (p < 0.001). Five participants praised the Chord Generator, describing the chords created with multimodal inspirations as "special" compared to those generated in Baseline. Three participants found the Chord Transcriber less creative than Chord Generator, with P5 noting, "With Chord Transcriber, you are picking pieces someone else created-it is not mine and feels close-ended. There's a risk of sounding like something that already exists. But the Chord Generator feels more open-ended and freeing." P6 suggested making Chord Transcriber's output unique by integrating it with the Chord Generator: "Transcribing the chords from the song I love would be good, but what I find challenging is changing these transcribed chords to make them your own. It would be



Figure 7: Distribution of participants' ratings on selfperceived songwriting experience. When equipped with AMUSE, participants felt they could more easily translate multimodal inspirations into musical elements (Inspiration Support), align their compositions with the given tasks (Task Alignment), and produce output of comparable quality (Output Quality). (*p<.05, **p<.01)

cool if you could input the Chord Transcriber's outputs into the Chord Generator to create new chords that match the vibe of the original song."

The low usage of Chord Transcriber might also be attributed to the task constraints, as two participants noted. P2 mentioned they could not think of a song to transcribe given the songwriting prompt: "I didn't use the transcribe feature which I think mainly because of the time constraint for me—I couldn't come up with a song that kind of fit the task that I could borrow from off the top of my head." P10 did not use the Chord Transcriber due to the absence of their instrument, stating, "I would have used it if I had my piano next to me. I would come up with whatever I felt and then upload the sound file to get the chords."

8.1.4 Enhanced transparency, agency, and explainability through music keywords. Five participants found the keywords helpful in the transformation process, particularly for the transparency and agency the keywords offered. Participants appreciated how the keyword control allowed them to see what the AI was interpreting from text or images, select keywords in consensus, and regenerate or edit them to better align with their creative goals. P5 noted, "Taking an image or story and turning it directly into chord progressions is an abstract jump, but with the keywords, the process becomes intuitive and predictable." P2 added, "Keywords put into words the idea that I was going for." Keywords also provided explainability for chord outputs. P6 explained, "When the chord progressions were weird, I realized I was using the keyword 'unexpected.' When I removed it, the generated chords were great." These findings reconfirm that the keywords could be an effective communication medium for connecting abstract relationships between different modalities as suggested in prior work [59, 60].



Figure 8: Interaction log timelines of all participants in the Assist condition. Interaction patterns were categorized into three distinct approaches: (i) kickstart with multimodal inspiration, (ii) ad hoc inspiration flow, and (iii) lyrics-centered brainstorming. 'Manual Editing' refers to user-initiated edits inside the Hookpad editor, including adjustments to melody, chords, instrumentation, and lyrics. 'Aria' refers to instances where participants queried ARIA to generate melodies or chords. 'AMUSE ChordGenerator' and 'AMUSE ChordTranscriber' refer to instances where participants were using each functionality. 'Accept Aria' and 'Accept AMUSE' refer to events where participants accepted each tool's suggestions, respectively. Empty spaces between the color bars indicate periods when the user was either listening to the composed song without taking any action, or explaining their thoughts out loud as part of the think-aloud protocol, again without taking any direct actions.

8.1.5 **Summary of findings**. The findings demonstrate that AMUSE effectively bridges the gap between multimodal inspirations and musical elements. Compared with using ARIA alone, participants felt more guided and found it easier to achieve their creative goals with AMUSE. The Chord Generator emerged as a preferred tool, offering unique chord progressions that participants found more inspiring than those from the Chord Transcriber. Positive user comments indicate that the music keyword pipeline further enhanced the experience by providing transparency, explainability, and control over the AI's output. Collectively, these findings confirm that AMUSE effectively supports songwriters in transforming their multimodal inspirations into contextual musical elements.

8.2 Impact on User's Songwriting Process

We describe how participants incorporated AMUSE into their songwriting process (RQ2). We first describe three AMUSE usage patterns observed in ASSist condition (§8.2.1), and explain how participants completed the task without AMUSE in Baseline condition, using only ARIA (§8.2.2). We report qualitative findings from analyzing screen recordings, think-aloud statements, and interviews.

8.2.1 **Diverse usage patterns of AMUSE in the songwriting process.** In the Assist condition, we identified three categories of approaches that participants used to find and integrate inspiration into their songwriting process: (i) kickstart with multimodal inspiration (N=6; P2-3, P5, P7-8, P10), (ii) ad hoc inspiration flow (N=3;

P4, P6, P9), and (iii) lyrics-centered brainstorming (N=1; P1). Participants engaged with AMUSE at diverse stages of their workflow, with its use generally concentrated in the early stages. In addition, when using AMUSE, participants reported a shift in their approach to ARIA, using it more for melody creation rather than for both chords and melody. The timeline of all participants' songwriting patterns is shown in Figure 8.

Kickstart with Multimodal Inspiration. Given the songwriting prompt, six participants (P2-3, P5, P7-8, P10) **quickly drew on personal inspirations and used AMUSE** to generate chord progressions that matched their ideas. After the initial generation with AMUSE, they did not return to AMUSE; instead, they developed the rest of the song by expanding or editing the chord progressions created with AMUSE and composing melodies, either manually or using ARIA. For example, P2 was inspired by a recent 'friendship at work created through a trauma bonding experience' and aimed to write a melancholy, storytelling song (see Figure 9(a) for the illustration). P2 accepted the four 4-bar chord progressions generated by AMUSE and combined them to create the entire chord progression of the song. They then developed the rest of the song by generating melodies with ARIA and editing those.

Ad hoc Inspiration Flow. In contrast to users who used AMUSE only at the first stage to kickstart their songwriting process, three participants (P4, P6, P9) constantly used AMUSE in their workflow.



Figure 9: Three different songwriting approaches in the Assist condition. (a) Kickstart with Inspiration: P2 generated chord progressions with AMUSE in the early stages of the process, creating the full chord progression before moving forward with other elements. (b) Ad hoc Inspiration Flow: P4 first used AMUSE with a text-based inspiration, which reminded them of Billie Eilish's song Birds of a Feather. P4 returned to AMUSE to build upon that inspiration. (c) Lyrics-Centered Brainstorming: P1 mapped out the lyrical content for each bar and used AMUSE to generate corresponding musical elements for each section.

P4 and P6 mentioned that **transforming their initial inspiration into chords with AMUSE sparked further ideas.** For instance, P4 described starting with inspiration from George Orwell's 1984 and then transitioning into Billie Eilish's *Birds of a Feather* (see Figure 9(b) for the illustration): "1984 is a book about conformity and structure, and I hoped to capture its serenity and finer ambiance. I thought keywords like 'dream pop' could aesthetically intertwine with my inspiration. As I listened to the chords, I wanted the next four bars to progress into a darker, non-stereotypical mood. Billie Eilish's Birds of a Feather came to mind because of its down-tempo, dreamy ambient soundscape, which would create a good resolution in bars five through eight."

P9, on the other hand, took a different approach and was not fully satisfied by AMUSE's generations. Rather than starting with AMUSE, P9 manually composed the chord progression for bars 1-4 and completed the melody with ARIA. The inspiration to reflect the



Figure 10: Interaction log timelines of all participants in the Baseline condition. Interaction patterns were categorized into two distinct approaches: manual experimentation until writer's block and generating until getting sparks. 'Manual Editing' refers to user-initiated edits inside the Hookpad editor, including adjustments to melody, chords, and lyrics. 'ARIA' refers to instances where participants queried ARIA to generate melodies or chords, and 'Accept ARIA' refers to events were participants accepted ARIA's suggestions. Empty spaces between the color bars indicate periods when participants were either listening to the composition without making changes or reflecting on their creative process as part of the think-aloud protocol, without engaging in direct editing.

mood of 'enjoying a 4th of July fireworks' came **midway through the songwriting process**, leading P9 to use AMUSE's Chord Generator for the last four bars. Later, P9 drew inspiration from Childish Gambino's song *Saturday* to infill the chord at bar 5, where he realized the chords for bars 1-4 and 5-8 did not fit well together. This led P9 to manually rewrite the progression for bars 5-8. Reflecting on the experience, P9 said, "*I could easily come up with the chord progression for the first few bars. I wanted to tell a story about the 4th of July fireworks using AMUSE and expected it to build on the chords I had already laid down, but I didn't like the result.*" Unlike ARIA, AMUSE does not have access to the full context of the user's composition, and we speculate that providing context would be necessary to better support users in the ad hoc inspiration flow.

Lyrics-Centered Brainstorming. One participant, P1, divided the song into three sections, assigning lyrical ideas to each part and using AMUSE to generate chords that aligned with those ideas (see Figure 9(c) for the illustration). P1 began by outlining the lyrical ideas, noting: 'Warm breeze salt in the air, cool water pouring over your hair, these are the worst lyrics one could think of for summer.' They then structured the song by planning bars 1-2 around a 'gentle summer beach,' bars 3-4 for 'cool water pouring over your hair,' and bars 4-8 to sing about the 'worst lyrics' theme. For the first two sections, P1 searched for relevant images online and used the Chord Generator to get chords matching the lyrical concepts. For the last four bars, he entered the lyric to the Chord Generator and selected two four-bar chord progressions that are singable. ARIA was used later in the process to assist with melody creation as P1 worked to bring the composition together. While most users relied on visual inputs to AMUSE, the instincts of this particular user centered strongly around lyrics, providing further evidence that systems should accommodate diverse multimodal inputs (DG1) to support a variety of users.

8.2.2 **Songwriting approaches without AMUSE**. In Baseline condition, where the participants wrote a song without AMUSE, participants either (i) manually experimented with chords and melodies until encountering writer's block (N=6; P1-3, P5-7), or (ii) relied on ARIA from scratch to find generations that align with inspiration (N=4; P4, P8, P9, P10). The timeline of all participants' songwriting process is shown in Figure 10.

Manual Experimentation Until Writer's Block. Six participants (P1, P2, P3, P5, P6, P7) began by creating chords or melodies manually. This process involved **experimenting with different combinations of musical elements** to meet their creative goals. Once they had set up the desired mood, they turned to ARIA for additional suggestions when they felt stuck. For example, P3 manually set up the full chord progressions and composed the melodies for the first five bars. Hoping "*ARIA will finalize the song*," P3 used it to generate the melodies for the last three bars. Different from the others (P1-3, P5-6) who started writing manually from the beginning, P7 first tried using ARIA for chords but found "*the chords do not sound summery*." Frustrated, P7 canceled generation and set up the full chord progressions manually and used ARIA later for melody generation.

Generating Until Getting Sparks. Four participants (P4, P8, P9, P10) relied on ARIA to spark their initial inspiration before writing any music manually, hoping to get a suggestion that aligned with their initial inspirations. However, this process often **resulted in suboptimal outputs that did not align with their intentions**. For example, P10 planned to write about 'the beginning of an unexpected friendship with my dog' but found that ARIA's chord suggestions were "*cool but bizarre for my dog*." After more attempts, P10 settled on a chord progression that "*didn't sound like a chorus, but was okay-ish*," and then manually adjusted it to better fit their goal.

8.2.3 **Summary of findings**. Based on the formative study findings, AMUSE was designed to support the early stages of songwriting by helping users brainstorm and transform initial ideas into musical elements. Interestingly, participants exhibited diverse usage patterns, leveraging AMUSE flexibly across different stages of their songwriting process rather than exclusively at the beginning—such as during ad hoc inspiration flows or after brainstorming lyrics with ARIA often used to generate melodies on chords created with AMUSE. In contrast, without AMUSE, users relied solely on ARIA to generate both melodies and chords to ease cognitive inertia in the process, often receiving suggestions that did not align with their creative intentions. These findings demonstrate how multimodal control prompts users to integrate inspirations dynamically, and thus, the importance of supporting multimodalities *throughout* the process to meet a wide range of creative workflows.

8.3 User Perception on Songwriting Process and Outcomes

We investigate AMUSE's impact on the songwriting experience and perceived quality of the compositions (RQ3). Concretely, we discuss AMUSE's impact on participants' sense of agency (§8.3.1), creativity (§8.3.2), and efficiency (§8.3.3) in the songwriting process. Lastly, we compare the overall satisfaction with final compositions across both conditions (§8.3.4). We describe relevant quantitative findings with qualitative insights.

8.3.1 Enhanced agency over the songwriting process. Overall, participants felt more in control over the process in Assist compared to the Baseline (Controllable in Figure 11; Assist=5.80±0.98, Baseline= 4.40 ± 1.62 , z=5.00, p=0.036). They also felt the process as more collaborative (Collaborative in Figure 11; Assist=5.80±0.98, Baseline= 3.90 ± 1.45 , z=2.00, p=0.014), indicating that the tools in the Assist condition fostered a stronger sense of partnership during the songwriting process compared to the Baseline condition. Five participants mentioned that they could communicate their creative intentions and have more agency over the workflow when equipped with AMUSE. For instance, P1 said, "I can be part of the creative process with AMUSE. You can influence what it generates because you get to tell it what you want it to consider." In contrast, two mentioned in Baseline they were simply "following what ARIA wants to do" (P2, P10). P10 explained, "With just ARIA, I wasn't able as much to create a mood or feel for my song. I didn't feel like I had much input-ARIA just sort of made up its own mind, and I had to just go along with it. It's like having a bad co-writer."

An interesting finding from the interviews, mentioned by three participants, was that having greater control during the early stages of songwriting with AMUSE led to a sense of disconnection later in the process. Participants expressed that they set higher expectations for ARIA, particularly regarding the relevance of its suggestions to their sources of inspiration after receiving suggestions from AMUSE. P3 mentioned, "Now that the chords are about a boy playing in the garden, I wished ARIA to generate something playful-maybe a melody with shorter notes and a quicker rhythm that would capture the image of a boy jumping. That was what I anticipated, but it didn't guite work." Similarly, P4 noted that using AMUSE and ARIA together felt less intuitive: "I felt like the chords were in place all the way through using AMUSE. But the only disconnection I felt was in how ARIA doesn't necessarily integrate with Amuse in the same way that Amuse integrates with ARIA. So it was almost like doing a 180-degree rotation when switching from AMUSE to ARIA because ARIA doesn't achieve quite the same results, given that it doesn't consider your inputs. [...] So I would say when AMUSE was there, I felt more enriched and musical, but using ARIA alone was more intuitive because I knew what to expect from it." These findings suggest that both systems-AMUSE and ARIAwould benefit from incorporating each other's control inputs, i.e., contextual and multimodal inputs, respectively.

8.3.2 Enhanced creativity with multimodal inspirations. When using both AMUSE and ARIA in the Assist condition, participants reported a greater sense of idea exploration (Exploration in Figure 11; Assist=6.10±0.83, Baseline=4.40±1.50, z=3.00, p=0.019) as well as expressiveness and creativity (Expressiveness in Figure 11; Assist=6.00±0.63, Baseline=4.20±1.66, z=4.00, p=0.028) compared to the Baseline. These results indicate that AMUSE enhanced participants' capacity to explore new musical ideas and freely express their creative intentions. This finding echoed in the interviews, where seven participants praised AMUSE for fostering creativity and offering greater freedom in musical expression. P4 and P7 described AMUSE as "liberating" their creative process; P4 said, "I feel like I'm actually guiding the tool toward what I want to create in a transformative, imaginative, and visually compelling world, almost as if I'm creating music as I type." In addition, P2 appreciated the integration of visuals in the songwriting process, saying, "I especially liked the picture input feature because, as the saying goes, a picture is worth a thousand words. Sometimes, it's easier to convey what you're aiming for with an image than with words."

8.3.3 **Perceived efficiency in the songwriting process.** Participants reported that the songwriting process felt more efficient in Assist during the interviews, despite no significant difference in actual completion times between conditions. Seven participants noted that the process **felt more efficient and easier** in the Assist condition. They described how AMUSE helped to "*streamline the process*" (P8, P9), contrasting it with their usual experience where "*it takes a while before getting a chord progression that I love*" (P6). P2 complimented, "*Honestly, I would use AMUSE almost every time*. *I really liked how it jump-started the process. I usually spend a lot of time and effort getting the right chords, which is tedious—being able to convey an idea quickly was really nice.*" However, the perceived efficiency **did not mean using AMUSE really sped up** CHI '25, April 26-May 1, 2025, Yokohama, Japan

Yewon Kim, Sung-Ju Lee, and Chris Donahue



Figure 11: Distribution of participants' ratings on self-perceived songwriting experience and Creativity Support Index (CSI). The left side focuses on AI-related aspects and composition output, while the right side represents CSI questionnaires. When equipped with AMUSE, participants felt that they had more control in songwriting processes, and felt more collaborating with the tool. In addition, participants felt more expressive during the Assist with more options to explore (*p<.05).

the process-there was no statistical difference in the task completion time between Assist and Baseline (Assist=17.39 \pm 5.45 min, Baseline= 17.61 \pm 5.27 min, z=26.00, p=0.922). This suggests that while AMUSE may enhance the perceived ease of the songwriting process, particularly by alleviating specific pain points like finding the right chords, it does not necessarily make the overall task faster.

8.3.4 Similarly satisfactory results in both conditions. Despite the reported differences in agency and creativity in the songwriting process, participants in both conditions often arrived at satisfying results. The post-survey results revealed that participants rated the quality of their compositions similarly in both conditions: they felt that the results were worth the effort (Results Worth Effort in Figure 11; Assist= 6.20 ± 0.98 , Baseline= 5.90 ± 1.37 , z=9.00, p=0.750), musically coherent (Musical Coherence in Figure 11; Assist= 6.00 ± 1.18 , Baseline= 6.00 ± 0.89 , z=10.50, p=1.00), and could be expanded to full songs (Potential for Development in Figure 11; Assist= 6.30 ± 0.90 , Baseline= 5.70 ± 1.35 , z=12.50, p=0.429). Moreover, participants found both conditions enjoyable (Enjoyment in Figure 11; Assist= 6.70 ± 0.46 , Baseline= 6.00 ± 1.10 , z=3.00, p=0.096).

One explanation for these similar levels of satisfaction, despite the differences in creative control, maybe the unexpected (and irrelevant) yet joyful, creative discoveries participants encountered in Baseline condition. We observed from the think-aloud data that, while transforming inspirations into musical elements with ARIA often led to off-topic suggestions, participants still **encountered moments of serendipity**. ARIA's generations were sometimes "*eye-opening*" (P1), even if they did not fit the original plan. In these cases, participants **deviated from their initial inspiration and still arrived at satisfying results**. For instance, P8 initially set out to express 'nostalgia for a childhood camp where my friends and I played card games,' but the final result was an epic heavy metal song. P8 reflected, "It didn't feel like myself—I kind of followed the genre ARIA gave me and adjusted my goal to suit it. I didn't feel like I had a clear idea going in. But, I'm still happy—the card games got dragons and crazy monsters in them, and these feel very important and epic from a kid's point of view." This observation suggests that participants were able to derive satisfaction from both adhering to and deviating from their original creative vision, indicating that serendipitous moments of discovery can be a key factor in creative satisfaction.

8.3.5 **Summary of findings**. Our study reveals that AMUSE positively impacts the songwriting experience by enhancing participants' agency, creativity, and perceived efficiency than using ARIA alone. However, some participants experienced a sense of disconnection when switching between the tools, which could be improved by incorporating each other's control inputs. Interestingly, despite the enhanced agency and creativity, the perceived quality of final compositions and enjoyment of the process were similar across both conditions. This similarity suggests that creative satisfaction may stem not only from control and alignment but also from unexpected discoveries throughout the process.

9 Discussion

We reflect on our findings from the design and evaluation of AMUSE and discuss the lessons learned and future opportunities for music creativity support systems.

9.1 Toward Contextual Multimodal Systems

Overall, we found that AMUSE effectively supported songwriters in exploring ideas while maintaining their sense of agency (§8.3). However, our study revealed moments of friction when participants transitioned from the multimodal tool (AMUSE) to the unimodal contextual tool (ARIA). Participants who engaged with AMUSE often expected ARIA to generate suggestions aligned with the sources of inspiration they inputted in AMUSE, only to find that ARIA could not reflect this prior context (§8.3.1). Moreover, while AMUSE was designed to support the early stages of songwriting, four out of 10 participants used it mid-process—after they had already developed a musical structure in the editor—even though AMUSE was not designed to account for existing user inputs. The global key conditioning of our chord generation system occasionally resulted in chord outputs that happened to be relevant to the surrounding context, but there were also instances where less contextually relevant outputs led to frustration (§8.2.1).

These frictions reflect a misalignment between user expectations and tool capabilities—participants formed assumptions about one tool's capability based on their experiences with the other, leading to frustration when their expectations were unmet [69], and consequently, a disruption in their creative flow [24, 25, 54]. For instance, interacting with AMUSE during the early stages of songwriting likely led participants to expect that these modalities would continue to play a role throughout the entire songwriting process. When transitioning to ARIA, however, their expectations clashed with ARIA's narrower focus on musical elements, which lacked the multimodality they had grown accustomed to. Similarly, some participants who used AMUSE mid-process expected it to consider both their multimodal inspirations and the evolving musical elements in the editor, which it did not do.

Addressing these discrepancies requires establishing consistency across tools used in the creative process [68]. An ideal system would be both multimodal and contextual, dynamically generating outputs that integrate both multimodal inspirations and the evolving musical context. This would ensure that users' creative expectations align with the system's capabilities. However, a key challenge in simultaneously achieving these dual goals lies in the absence of multimodal-contextual data (e.g., paired keyword-MIDI data) necessary for training models that can account for both users' multimodal and contextual inputs. Future work could involve building a dataset that captures the relationship between multimodal inspirations and musical elements. One potential approach is to collect interaction data [56], for example, by assigning specific songwriting tasks (e.g., 'compose a song reflecting the mood of an image') and logging the suggestions (e.g., melodies or chords) users accept from a contextual model like ARIA. Another approach might involve using multimodal models for the reverse task of generating relevant multimodal inspirations from musical elements to create paired data. These approaches would offer valuable multimodal-contextual data pairs to inform the development of more integrated systems.

9.2 The Role of Randomness in Human-AI Co-Creation

One of the interesting observations from our study was that participants felt equally satisfied with the final compositions and enjoyed the songwriting process similarly, regardless of whether they used only ARIA or both ARIA and AMUSE, despite AMUSE's outputs being considered as more controllable and better aligned with their creative intentions (§8.3.4). Notably, participants often described ARIA's suggestions as out-of-sync and random, yet they also reported moments of serendipity where these unexpected outputs prompted them to deviate from their original plans and still achieve satisfactory outcomes. This observation aligns with prior work showing that unrelated suggestions can be evocative and productive in creative contexts [51, 57, 100], with the surprising combinations of ideas provoking exploratory processes that lead to the production of creative ideas [53]. Participants may have viewed ARIA's random outputs not as bugs but as an inherent feature of the human-AI interaction, and this engagement with randomness likely stimulated cognitive reorganization and creative leaps that enriched their overall experience [88].

Another hypothesis is that participants' engagement styles would have played a crucial role in how they interacted with the systems and perceived their creative processes. As suggested by prior work in the creative writing domain [88], creative processes involve broadly different styles of engagement with suggestions, including reactive and proactive engagements. Reactive writers let AI suggestions actively shape their work, using the AI outputs to determine their creative direction. Proactive writers, by contrast, have a clear idea of what they want to create and tightly control their process. Similarly, participants with reactive engagement style in our study may have found ARIA's randomness more inspiring than AMUSE's controllable outputs, as it enabled exploration without requiring explicit decisions about the creative direction. However, it is important to note that our study was conducted in a controlled environment, where participants were given a songwriting prompt and instructed to use the assigned tool(s) at least once. This setup might have influenced how they engaged with the systems, potentially differing from how they would approach the tools in more natural settings. Future work could explore these dynamics in openended, real-world scenarios with a larger and more diverse group of participants. Such studies could uncover richer usage patterns and provide deeper insights into how randomness, controllability, and user engagement styles contribute to shaping creative satisfaction and the effectiveness of AI creativity support tools.

9.3 Reflection on the Chord Transcriber's Low Usage

While the majority of participants in the formative study highlighted existing music and transcribing elements like chords or melodies as a key ideation process (§3.2), the user study revealed a contrasting preference: participants generally favored the Chord Generator over the Chord Transcriber (§8.1.3). Although practical constraints—such as the absence of instruments and limited time to choose appropriate songs during the study—may have influenced this preference, the difference in how the two tools supported creative exploration may provide additional insights.

We hypothesize that the Chord Generator, by accepting nonmusical inputs such as images and texts, offered the creative freedom necessary to compose songs based on songwriting prompts (e.g., "Write an 8-bar chorus about your favorite summer holiday memory."), which was the focus of our user study task. Working to understand the abstract and open-ended relationship between non-musical inputs and musical outputs likely fostered imaginative and interpretative exploration. Theories of creative thought also support this hypothesis, which suggests that abstract thinking facilitates creative cognition by stimulating new associations and broadening the scope of possibilities [35, 99]. In contrast, the Chord Transcriber, which outputs literal transcriptions from musical inputs, provided a more concrete and constrained interaction. This specificity may limit participants' opportunities for abstract thinking and exploratory ideation. As a result, the Chord Generator's capacity to support interpretative and open-ended exploration likely made it better suited for our songwriting task, which emphasized creative ideation. Future work could explore contexts in which more literal, concrete tools like the Chord Transcriber are particularly effective.

Additionally, participants may have preferred the Chord Generator because it provided multiple suggestions, aligning with the design principles for creativity support tools that emphasize the importance of exploring various alternatives [81]. In comparison, the Chord Transcriber's singular, literal outputs may have constrained creative exploration. Although we designed the Chord Transcriber based on observations from the formative study—where users transcribed chords from audio and expanded upon them—participants in the user study appeared to expect the tool to support further exploration of the chords, as indicated by P6 in §8.1.3. Future work could consider expanding the Chord Transcriber's capabilities to encourage exploration within the creative space, such as offering multiple variations of transcribed chord progressions, like alterations or substitutions, alongside the literal transcriptions.

9.4 Real-time Multimodal Creativity Support

Like many existing AI music systems [61, 62, 89], AMUSE was designed for a static environment where users sit down at a computer to create music, and our experiments reflected this setting. However, music creation is inherently *dynamic*, often occurring in spontaneous moments, such as during improvisation on instruments [4, 96]. We observed similar behaviors in both our formative and user studies. The majority of formative study participants reported they rely on improvising on their instruments to manually conjure ideas (§3.2.2), and a user study participant noted that they would have improvised on the piano to develop ideas if they had access to it during the study session (§8.1.3).

To this end, a promising direction for the future development of human-AI music co-creation systems would involve building AI systems and interaction methods that facilitate real-time creative exploration. One intuitive approach is real-time accompaniment systems [8, 98, 104], which includes harmonizing with [104] or generating basslines [8] upon user melody inputs. Exploring how users can effectively interact with these systems to create a meaningful composition would be a valuable area of investigation.

Moreover, multimodal systems could further enhance this process by providing real-time feedback across different modalities, supporting the user's creative exploration of musical ideas. Similar to writing support tools that offer multimodal feedback to guide story development [9, 88], a real-time multimodal feedback system for music could offer lyrical ideas, visual cues, or rhythmic patterns that respond dynamically to the user's improvisations, enriching the creative process.

9.5 Expanding Multimodal Support Across Inputs and Outputs

In this study, we focused on generating chord progressions based on multimodal inspirations. While chord progressions are a critical building block of music, the music itself comprises multiple layers: lyrics, melody, chords, bassline, drums, structure, and instrumentation, to name a few. A promising research direction would involve expanding multimodal support to *output* other types of layers, each of which might constitute a reusable musical element. This could offer users greater flexibility in how they approach music composition and help them develop different layers of their songs simultaneously or in sequence.

Similarly, musicians might benefit from AI assistance supporting diverse forms of *input* beyond the text, image, and audio modalities considered in this work. For instance, users could initiate their creative process with rhythm (as sensory data) or motion information (as video data). Exploring how different modalities can inspire musical ideas could open up new possibilities for future research. The diversity of input/output directions pursued by participants in the AI Song Contest [48, 65] provides some evidence that greater flexibility on both ends would support the creative goals of more users.

9.6 Limitations

Several limitations of the study warrant consideration. First, AMUSE always displayed the Chord Generator first, potentially influencing participant preferences. Additionally, the controlled study environment, where participants were required to use each tool at least once, may not reflect how they would naturally interact with the tools in a more open-ended setting. The novelty effect of AMUSE also cannot be decoupled, as most participants were familiar with ARIA but not with AMUSE. Future studies could benefit from in-thewild, long-term investigations to capture more natural and organic use of AMUSE.

10 Conclusion

We present AMUSE, a songwriting assistant that transforms multimodal (image, text, audio) inspirations into reusable musical elements (chord progressions) that can be seamlessly incorporated into songwriters' creative process. Amuse offers two key functionalities: Chord Generator and Chord Transcriber. The Chord Generator allows users to extract music keywords from multimodal inspirations and generate chord progressions aligned with these keywords. The Chord Transcriber enables users to convert audio sources into chords. To ensure the generated chord progressions are diverse, musically coherent, and relevant to keywords, we developed a novel rejection-sampling-based approach that leverages the general capabilities of LLMs in combination with a unimodal chord generation model trained on real music data. Our technical evaluation and user study demonstrate that AMUSE effectively supports the transformation of multimodal inspirations into musical elements, enhancing the songwriting process by improving perceived agency, creativity, and efficiency. With advancements in music AI, we believe AMUSE opens new avenues for human-AI co-creation in music.

Acknowledgments

We thank our study participants for their valuable feedback on the design of AMUSE. We thank Sherry Tongshuang Wu, Nikolas Martelaro, Sihyun Yu, DaEun Choi, and Ryuhaerang Choi for their insights, discussions, and support. We also thank the Hooktheory team for facilitating connections with Hooktheory users and providing insightful feedback and resources for the system implementation. We disclose the use of generative AI tools in the process of writing this manuscript. These tools were used solely for editing the authors' own text, and the authors ensured the final content was free from plagiarism, misrepresentation, fabrication, and falsification. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00337007). This work was in part supported by Hooktheory.

References

- [1] Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. 2023. MusicLM: Generating Music From Text. arXiv:2301.11325 [cs.SD] https://arxiv.org/abs/2301.11325
- [2] Music AI. 2024. Music AI: AI Audio Models to Power Your Music Business. https://www.music.ai.
- [3] Suno AI. 2024. Suno AI. https://suno.com/.
- [4] Philip Alperson. 1984. On Musical Improvisation. The Journal of Aesthetics and Art Criticism 43, 1 (1984), 17–29. http://www.jstor.org/stable/430189
- [5] Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization Effects of Large Language Models on Human Creative Ideation. In Proceedings of the 16th Conference on Creativity & Cognition (Chicago, IL, USA) (C&C '24). Association for Computing Machinery, New York, NY, USA, 413–425. doi:10.1145/3635636.3656204
- [6] Aria. 2024. Introducing Aria Your personal AI co-creator for chords and melody. https://www.hooktheory.com/hookpad/aria.
- [7] NIC BECKER, RYAN LÓUIE, JOHN THICKSTUN, and PERCY LIANG. 2024. Designing Live Human-AI Collaboration for Musical Improvisation.
- [8] Christodoulos Benetatos, Joseph VanderStel, and Zhiyao Duan. 2020. BachDuet: A Deep Learning System for Human-Machine Counterpoint Improvisation. In Proceedings of the International Conference on New Interfaces for Musical Expression, Romain Michon and Franziska Schroeder (Eds.). Birmingham City University, Birmingham, UK, 635–640. doi:10.5281/zenodo.4813234
- [9] Eden Bensaid, Mauro Martino, Benjamin Hoover, and Hendrik Strobelt. 2021. FairyTailor: A Multimodal Generative Framework for Storytelling. arXiv:2108.04324 [cs.CL] https://arxiv.org/abs/2108.04324
- [10] Renaud Bougueng Tchemeube, Jeffrey John Ens, and Philippe Pasquier. 2022. Calliope: A Co-creative Interface for Multi-Track Music Generation. In Proceedings of the 14th Conference on Creativity and Cognition (Venice, Italy) (C&C '22). Association for Computing Machinery, New York, NY, USA, 608–611. doi:10.1145/3527927.3535200
- [11] Daniel Buschek, Martin Zürn, and Malin Eiband. 2021. The Impact of Multiple Parallel Phrase Suggestions on Email Input and Composition Behaviour of Native and Non-Native English Writers. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 732, 13 pages. doi:10.1145/3411764.3445372
- [12] Runze Cai, Nuwan Janaka, Yang Chen, Lucia Wang, Shengdong Zhao, and Can Liu. 2024. PANDALens: Towards AI-Assisted In-Context Writing on OHMD During Travels. In Proceedings of the CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 1053, 24 pages. doi:10.1145/3613904.3642320
- [13] Angel X. Chang, Mihail Eric, Manolis Savva, and Christopher D. Manning. 2017. SceneSeer: 3D Scene Design with Natural Language. arXiv:1703.00050 [cs.GR] https://arxiv.org/abs/1703.00050
- [14] Siddhartha Chaudhuri, Evangelos Kalogerakis, Stephen Giguere, and Thomas Funkhouser. 2013. Attribit: content creation with semantic attributes. In Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology (St. Andrews, Scotland, United Kingdom) (UIST '13). Association for Computing Machinery, New York, NY, USA, 193–202. doi:10.1145/2501988.2502008
- [15] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2022. HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection. arXiv:2202.00874 [cs.SD] https://arxiv.org/ abs/2202.00874

- [16] Erin Cherry and Celine Latulipe. 2014. Quantifying the Creativity Support of Digital Tools through the Creativity Support Index. ACM Trans. Comput.-Hum. Interact. 21, 4, Article 21 (jun 2014), 25 pages. doi:10.1145/2617588
- [17] DaEun Choi, Sumin Hong, Jeongeon Park, John Joon Young Chung, and Juho Kim. 2024. CreativeConnect: Supporting Reference Recombination for Graphic Design Ideation with Generative AI. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 1055, 25 pages. doi:10. 1145/3613904.3642794
- [18] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: Sketching Stories with Generative Pretrained Language Models. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 209, 19 pages. doi:10. 1145/3491102.3501819
- [19] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. 2018. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In Proceedings of the 23rd International Conference on Intelligent User Interfaces (Tokyo, Japan) (IUI '18). Association for Computing Machinery, New York, NY, USA, 329–340. doi:10.1145/3172944.3172983
- [20] WJ Conover. 1999. Practical nonparametric statistics. John Wiley & Sons, Inc, New York, NY, USA.
- [21] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2024. Simple and Controllable Music Generation. arXiv:2306.05284 [cs.SD] https://arxiv.org/abs/2306.05284
- [22] Juliet Corbin and Anselm Strauss. 2008. Basics of Qualitative Research (3rd ed.): Techniques and Procedures for Developing Grounded Theory. doi:10.4135/ 9781452230153
- [23] Bob Coyne and Richard Sproat. 2001. WordsEye: an automatic text-to-scene conversion system. In Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '01). Association for Computing Machinery, New York, NY, USA, 487–496. doi:10.1145/383259.383316
- [24] Mihaly Csikszentmihalyi. 1997. Flow and Creativity. NAMTA Journal 22, 2 (1997), 60–97. https://eric.ed.gov/?id=EJ547968
- [25] Mihaly Csikszentmihalyi. 1997. Flow and the psychology of discovery and invention. HarperPerennial, New York 39 (1997), 1–16.
- [26] Hai Dang, Sven Goller, Florian Lehmann, and Daniel Buschek. 2023. Choice Over Control: How Users Write with Large Language Models using Diegetic and Non-Diegetic Prompting. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 408, 17 pages. doi:10.1145/ 3544548.3580969
- [27] TheoryTab DB. 2024. TheoryTab DB: Tabs that show the theory behind songs. https://www.hooktheory.com/theorytab.
- [28] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A Generative Model for Music. arXiv:2005.00341 [eess.AS] https://arxiv.org/abs/2005.00341
- [29] Chris Donahue, Antoine Caillon, Adam Roberts, Ethan Manilow, Philippe Esling, Andrea Agostinelli, Mauro Verzetti, Ian Simon, Olivier Pietquin, Neil Zeghidour, and Jesse Engel. 2023. SingSong: Generating musical accompaniments from singing. arXiv:2301.12662 [cs.SD] https://arxiv.org/abs/2301.12662
- [30] Chris Donahue, Huanru Henry Mao, Yiting Ethan Li, Garrison W. Cottrell, and Julian McAuley. 2019. LakhNES: Improving multi-instrumental music generation with cross-domain pre-training. arXiv:1907.04868 [cs.SD] https: //arxiv.org/abs/1907.04868
- [31] Chris Donahue, John Thickstun, and Percy Liang. 2022. Melody transcription via generative pre-training. arXiv:2212.01884 [cs.SD] https://arxiv.org/abs/2212. 01884
- [32] Chris Donahue, Shih-Lun Wu, Yewon Kim, Dave Carlton, Ryan Miyakawa, and John Thickstun. 2025. Hookpad Aria: A Copilot for Songwriters. arXiv:2502.08122 [cs.SD] https://arxiv.org/abs/2502.08122
- [33] Anil R. Doshi and Oliver P. Hauser. 2024. Generative artificial intelligence enhances creativity but reduces the diversity of novel content. arXiv:2312.00506 [cs.HC] https://arxiv.org/abs/2312.00506
- [34] Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham W. Taylor. 2019. Tell, Draw, and Repeat: Generating and Modifying Images Based on Continual Linguistic Instruction. arXiv:1811.09845 [cs.CV] https://arxiv.org/abs/1811. 09845
- [35] Ronald A. Finke, Thomas B. Ward, and Steven M. Smith. 1992. Creative Cognition: Theory, Research, and Applications. The MIT Press, New York, NY, USA. doi:10. 7551/mitpress/7722.001.0001
- [36] Seth* Forsgren and Hayk* Martiros. 2022. Riffusion Stable diffusion for realtime music generation. https://riffusion.com/about
- [37] Satoru Fukayama, Kazuyoshi Yoshii, and Masataka Goto. 2013. Chord-Sequence-Factory: A Chord Arrangement System Modifying Factorized Chord Sequence Probabilities. https://api.semanticscholar.org/CorpusID:1099764

CHI '25, April 26-May 1, 2025, Yokohama, Japan

- [38] Katy Ilonka Gero and Lydia B. Chilton. 2019. Metaphoria: An Algorithmic Companion for Metaphor Creation. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (Glasgow, Scotland Uk) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/ 3290605.3300526
- [39] Barney Glaser and Anselm Strauss. 1999. Discovery of Grounded Theory: Strategies for Qualitative Research (1st ed.). Routledge, Abingdon, Oxon, UK. doi:10.4324/9780203793206
- [40] Gaëtan Hadjeres and Léopold Crestel. 2021. The Piano Inpainting Application. arXiv:2107.05944 [cs.SD] https://arxiv.org/abs/2107.05944
- [41] Gaëtan Hadjeres, François Pachet, and Frank Nielsen. 2017. DeepBach: a Steerable Model for Bach Chorales Generation. In Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70), Doina Precup and Yee Whye Teh (Eds.). PMLR, International Convention Centre, Sydney, Australia, 1362–1371. http://proceedings.mlr.press/v70/ hadjeres17a.html
- [42] Hookpad. 2024. Hookpad Songwriting Software: Create Amazing Music. https: //www.hooktheory.com/hookpad.
- [43] Hooktheory. 2024. Hooktheory: Create amazing music. https://www. hooktheory.com/.
- [44] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Pittsburgh, Pennsylvania, USA) (CHI '99). Association for Computing Machinery, New York, NY, USA, 159–166. doi:10.1145/302979.303030
- [45] Cheng-Zhi Anna Huang, Tim Cooijmans, Adam Roberts, Aaron Courville, and Douglas Eck. 2019. Counterpoint by Convolution. arXiv:1903.07227 [cs.LG] https://arxiv.org/abs/1903.07227
- [46] Cheng-Zhi Anna Huang, David Duvenaud, and Krzysztof Z. Gajos. 2016. ChordRipple: Recommending Chords to Help Novice Composers Go Beyond the Ordinary. In Proceedings of the 21st International Conference on Intelligent User Interfaces (Sonoma, California, USA) (IUI '16). Association for Computing Machinery, New York, NY, USA, 241–250. doi:10.1145/2856767.2856792
- [47] Cheng-Zhi Anna Huang, Curtis Hawthorne, Adam Roberts, Monica Dinculescu, James Wexler, Leon Hong, and Jacob Howcroft. 2019. The Bach Doodle: Approachable music composition with machine learning at scale. arXiv:1907.06637 [cs.SD] https://arxiv.org/abs/1907.06637
- [48] Cheng-Zhi Anna Huang, Hendrik Vincent Koops, Ed Newton-Rex, Monica Dinculescu, and Carrie J. Cai. 2020. AI Song Contest: Human-AI Co-Creation in Songwriting. arXiv:2010.05388 [cs.SD] https://arxiv.org/abs/2010.05388
- [49] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck. 2018. Music Transformer. arXiv:1809.04281 [cs.LG] https://arxiv.org/abs/1809.04281
- [50] Qingqing Huang, Daniel S. Park, Tao Wang, Timo I. Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, Jesse Engel, Quoc V. Le, William Chan, Zhifeng Chen, and Wei Han. 2023. Noise2Music: Text-conditioned Music Generation with Diffusion Models. arXiv:2302.03917 [cs.SD] https://arxiv.org/abs/2302.03917
- [51] Purnima Kamath, Fabio Morreale, Priambudi Lintang Bagaskara, Yize Wei, and Suranga Nanayakkara. 2024. Sound Designer-Generative AI Interactions: Towards Designing Creative Support Tools for Professional Sound Designers. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 730, 17 pages. doi:10.1145/3613904.3642040
- [52] Hyeongcheol Kim, Shengdong Zhao, Can Liu, and Kotaro Hara. 2020. LiveSnippets: Voice-based Live Authoring of Multimedia Articles about Experiences. In 22nd International Conference on Human-Computer Interaction with Mobile Devices and Services (Oldenburg, Germany) (MobileHCI '20). Association for Computing Machinery, New York, NY, USA, Article 31, 11 pages. doi:10.1145/3379503.3403556
- [53] Ziva Kunda, Dale T. Miller, and Theresa Claire. 1990. Combining Social Concepts: The Role of Causal Reasoning. *Cognitive Science* 14, 4 (1990), 551–577.
- [54] Tomas Lawton, Kazjon Grace, and Francisco J Ibarrola. 2023. When is a Tool a Tool? User Perceptions of System Agency in Human–AI Co-Creative Drawing. In Proceedings of the 2023 ACM Designing Interactive Systems Conference (Pittsburgh, PA, USA) (DIS '23). Association for Computing Machinery, New York, NY, USA, 1978–1996. doi:10.1145/3563657.3595977
- [55] Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A. Alghamdi, Tal August, Avinash Bhat, Madiha Zahrah Choksi, Senjuti Dutta, Jin L.C. Guo, Md Naimul Hoque, Yewon Kim, Simon Knight, Seyed Parsa Neshaei, Antonette Shibani, Disha Shrivastava, Lila Shroff, Agnia Sergeyuk, Jessi Stark, Sarah Sterman, Sitong Wang, Antoine Bosselut, Daniel Buschek, Joseph Chee Chang, Sherol Chen, Max Kreminski, Joonsuk Park, Roy Pea, Eugenia Ha Rim Rho, Zejiang Shen, and Pao Siangliulue. 2024. A Design Space for Intelligent and Interactive Writing Assistants. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA,

Article 1054, 35 pages. doi:10.1145/3613904.3642697

- [56] Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 388, 19 pages. doi:10.1145/3491102.3502030
- [57] Tuck Wah Leong, Frank Vetere, and Steve Howard. 2006. Randomness as a resource for design. In Proceedings of the 6th Conference on Designing Interactive Systems (University Park, PA, USA) (DIS '06). Association for Computing Machinery, New York, NY, USA, 132–139. doi:10.1145/1142405.1142428
- [58] J. Lin. 1991. Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory 37, 1 (1991), 145–151. doi:10.1109/18.61115
- [59] Vivian Liu, Han Qiao, and Lydia Chilton. 2022. Opal: Multimodal Image Generation for News Illustration. In Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (Bend, OR, USA) (UIST '22). Association for Computing Machinery, New York, NY, USA, Article 73, 17 pages. doi:10.1145/3526113.3545621
- [60] Vivian Liu, Jo Vermeulen, George Fitzmaurice, and Justin Matejka. 2023. 3DALL-E: Integrating Text-to-Image AI in 3D Design Workflows. arXiv:2210.11603 [cs.HC] https://arxiv.org/abs/2210.11603
- [61] Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J. Cai. 2020. Novice-AI Music Co-Creation via AI-Steering Tools for Deep Generative Models. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376739
- [62] Ryan Louie, Jesse Engel, and Cheng-Zhi Anna Huang. 2022. Expressive Communication: Evaluating Developments in Generative Models and Steering Interfaces for Music Creation. In Proceedings of the 27th International Conference on Intelligent User Interfaces (Helsinki, Finland) (IUI '22). Association for Computing Machinery, New York, NY, USA, 405–417. doi:10.1145/3490099.3511159
- [63] Martin E. Malandro. 2023. Composer's Assistant: An Interactive Transformer for Multi-Track MIDI Infilling. arXiv:2301.12525 [cs.SD] https://arxiv.org/abs/ 2301.12525
- [64] Enrique Manjavacas, Folgert Karsdorp, Ben Burtenshaw, and Mike Kestemont. 2017. Synthetic Literature: Writing Science Fiction in a Co-Creative Process. In Proceedings of the Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2017), Hugo Gonçalo Oliveira, Ben Burtenshaw, Mike Kestemont, and Tom De Smedt (Eds.). Association for Computational Linguistics, Santiago de Compostela, Spain, 29–37. doi:10.18653/v1/W17-3904
- [65] Lidia Morris, Rebecca Leger, Michele Newman, John Ashley Burgoyne, Ryan Groves, Natasha Mangal, and Jin Ha Lee. 2024. Human-AI Music Process: A Dataset of AI-Supported Songwriting Processes from the AI Song Contest.
- [66] Radford M. Neal. 2003. Slice sampling. The Annals of Statistics 31, 3 (2003), 705 767. doi:10.1214/aos/1056562461
- [67] Michele Newman, Lidia Morris, and Jin Ha Lee 0001. 2023. Human-AI Music Creation: Understanding the Perceptions and Experiences of Music Creators for Ethical and Productive Collaboration. 80-88 pages. doi:10.5281/zenodo.10265227
- [68] D.A. Norman. 2013. The Design of Everyday Things. MIT Press, Cambridge, MA, USA. https://books.google.co.kr/books?id=heCtnQEACAAJ
- [69] Richard L. Oliver. 1980. A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions. *Journal of Marketing Research* 17, 4 (1980), 460–469. http://www.jstor.org/stable/3150499
- [70] OpenAI. 2024. GPT-4o System Card. https://cdn.openai.com/gpt-4o-systemcard.pdf.
- [71] OpenAI. 2024. Hello GPT-40. https://openai.com/index/hello-gpt-40/.
- [72] OpenAI. 2024. OpenAI Platform. https://platform.openai.com/docs/overview.
- [73] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155 [cs.CL] https://arxiv.org/abs/2203.02155
- [74] Vishakh Padmakumar and He He. 2024. Does Writing with Language Models Reduce Content Diversity? https://openreview.net/forum?id=Feiz5HtCD0
- [75] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of* the 40th Annual Meeting on Association for Computational Linguistics (Philadelphia, Pennsylvania) (ACL '02). Association for Computational Linguistics, USA, 311–318. doi:10.3115/1073083.1073135
- [76] Christine McLeavy Payne. 2019. MuseNet. https://openai.com/blog/musenet/.
 [77] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv:2307.01952 [cs.CV] https://arxiv.org/abs/2307.01952
- [78] Prolific. 2024. Prolific | Quickly find research participants you can trust. https: //www.prolific.com/.
- [79] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark,

CHI '25, April 26-May 1, 2025, Yokohama, Japan

Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV] https://arxiv.org/abs/2103.00020

- [80] Reddit. 2024. Songwriting. https://www.reddit.com/r/Songwriting/.
- [81] Mitchel Resnick, Brad Myers, Kumiyo Nakakoji, Ben Shneiderman, Randy Pausch, Ted Selker, and Mike Eisenberg. 2018. Design Principles for Tools to Support Creative Thinking. doi:10.1184/R1/6621917.v1
- [82] Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Schölkopf. 2023. Moûsai: Text-to-Music Generation with Long-Context Latent Diffusion. arXiv:2301.11757 [cs.CL] https://arxiv.org/abs/2301.11757
- [83] Shikhar Sharma, Dendi Suhubdy, Vincent Michalski, Samira Ebrahimi Kahou, and Yoshua Bengio. 2018. ChatPainter: Improving Text to Image Generation using Dialogue. arXiv:1802.08216 [cs.CV] https://arxiv.org/abs/1802.08216
- [84] Ben Shneiderman. 2022. Human-Centered AI. Oxford University Press, Oxford, United Kingdom. https://global.oup.com/academic/product/human-centeredai-9780192845290
- [85] Ben Shneiderman, Catherine Plaisant, Maxine Cohen, Steven Jacobs, Niklas Elmqvist, and Nicholas Diakopoulos. 2016. Designing the User Interface: Strategies for Effective Human-Computer Interaction (6th ed.). Pearson, London, UK.
- [86] Ian Simon and Sageev Oore. 2017. Performance RNN: Generating Music with Expressive Timing and Dynamics. https://magenta.tensorflow.org/performancernn.
- [87] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. 2022. Make-A-Video: Text-to-Video Generation without Text-Video Data. arXiv:2209.14792 [cs.CV] https://arxiv.org/abs/2209.14792
- [88] Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L. Glassman. 2023. Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence. ACM Trans. Comput.-Hum. Interact. 30, 5, Article 68 (sep 2023), 57 pages. doi:10.1145/3511599
- [89] Minhyang (Mia) Suh, Emily Youngblom, Michael Terry, and Carrie J Cai. 2021. AI as Social Glue: Uncovering the Roles of Deep Generative AI during Social Music Composition. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 582, 11 pages. doi:10.1145/3411764. 3445219
- [90] suno.wiki. 2024. Style and Genre List. https://www.suno.wiki/faq/style-and-lyrics/styles-and-genres/.
- [91] Últimate Guitar Tabs. 2024. Misty Chords by Ella Fitzgerald. https://tabs. ultimate-guitar.com/tab/ella-fitzgerald/misty-chords-1206796.
- [92] John Thickstun, David Leo Wright Hall, Chris Donahue, and Percy Liang. 2024. Anticipatory Music Transformer. https://openreview.net/forum?id=EBNJ33Fcrl
- [93] Udio. 2024. Udio | AI Music Generator Official Website. https://udio.com/.
 [94] Unison. 2024. 7 Sad Chord Progressions That Instantly Get People Invested. https://unison.audio/sad-chord-progressions/.
- [95] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748 [cs.LG] https://arxiv.org/ abs/1807.03748
- [96] Ashley Walton, Auriel Washburn, Peter Langland-Hassan, Anthony Chemero, Heidi Kloos, and Michael Richardson. 2017. Creating Time: Social Collaboration in Music Improvisation. *Topics in Cognitive Science* 10 (11 2017). doi:10.1111/ tops.12306
- [97] Bryan Wang, Yuliang Li, Zhaoyang Lv, Haijun Xia, Yan Xu, and Raj Sodhi. 2024. LAVE: LLM-Powered Agent Assistance and Language Augmentation for Video Editing. In Proceedings of the 29th International Conference on Intelligent User Interfaces (Greenville, SC, USA) (IUI '24). Association for Computing Machinery, New York, NY, USA, 699–714. doi:10.1145/3640543.3645143
- [98] Zihao Wang, Kejun Zhang, Yuxing Wang, Chen Zhang, Qihao Liang, Pengfei Yu, Yongsheng Feng, Wenbo Liu, Yikai Wang, Yuntao Bao, and Yiheng Yang. 2022. SongDriver: Real-time Music Accompaniment Generation without Logical Latency nor Exposure Bias. In Proceedings of the 30th ACM International Conference on Multimedia (Lisboa, Portugal) (MM '22). Association for Computing Machinery, New York, NY, USA, 1057–1067. doi:10.1145/3503161.3548368
- [99] Thomas B Ward, Ronald A Finke, and Steven M Smith. 1995. Creativity and the mind: Discovering the genius within (1 ed.). Springer, New York, NY, USA. IX, 274 pages. doi:10.1007/978-1-4899-3330-0
- [100] Thomas B. Ward and E. Thomas Lawson. 2009. Creative Cognition in Science Fiction and Fantasy Writing. Cambridge University Press, Cambridge, 196–210.
- [101] Shih-Lun Wu, Chris Donahue, Shinji Watanabe, and Nicholas J. Bryan. 2024. Music ControlNet: Multiple Time-Varying Controls for Music Generation. IEEE/ACM Trans. Audio, Speech and Lang. Proc. 32 (May 2024), 2692–2703. doi:10.1109/TASLP.2024.3399026
- [102] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 385, 22 pages. doi:10.1145/3491102.3517582

- [103] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. 5 pages. doi:10. 1109/ICASSP49357.2023.10095969
- [104] Yusong Wu, Tim Cooijmans, Kyle Kastner, Adam Roberts, Ian Simon, Alexander Scarlatos, Chris Donahue, Cassie Tarakajian, Shayegan Omidshafiei, Aaron Courville, Pablo Samuel Castro, Natasha Jaques, and Cheng-Zhi Anna Huang, 2024. Adaptive Accompaniment with ReaLchords. In Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235), Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (Eds.). PMLR, Valencia, Spain, 53328–53345. https://proceedings.mlr.press/v235/wu24c.html
- [105] Zihan Yan, Chunxu Yang, Qihao Liang, and Xiang 'Anthony' Chen. 2023. XCreation: A Graph-based Crossmodal Generative Creativity Support Tool. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (San Francisco, CA, USA) (UIST '23). Association for Computing Machinery, New York, NY, USA, Article 48, 15 pages. doi:10.1145/3586183.3606826
- [106] Sihyun Yu, Weili Nie, De-An Huang, Boyi Li, Jinwoo Shin, and Anima Anandkumar. 2024. Efficient Video Diffusion Models via Content-Frame Motion-Latent Decomposition. arXiv:2403.14148 [cs.CV] https://arxiv.org/abs/2403.14148
- [107] Yiming Zhang, Avi Schwarzschild, Nicholas Carlini, Zico Kolter, and Daphne Ippolito. 2024. Forcing Diffuse Distributions out of Language Models. arXiv:2404.10859 [cs.CL]
- [108] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A Benchmarking Platform for Text Generation Models. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (Ann Arbor, MI, USA) (SIGIR '18). Association for Computing Machinery, New York, NY, USA, 1097–1100. doi:10.1145/3209978.3210080

CHI '25, April 26-May 1, 2025, Yokohama, Japan

A Prompts

We list the GPT-40 prompts used in Amuse and its evaluation.

A.1 Keywords Extraction

System Prompt

You are a creative assistant that generates song keywords. You will receive one or more of the following: an image URL, a text note, and user-written keywords. Your task is to create keywords that capture the essence of the input, using style-words, genres, and song-types from the provided keyword list. You may also create new relevant keywords if necessary. Respond only with the keywords, avoiding any additional commentary or formatting.

Instructions:

 Analyze the inputs: For an image, identify visual elements that suggest musical themes and emotions. For text, identify the main themes, emotions, or ideas. For user-written keywords, expand upon them with related music styles, genres, and types.
 Generate keywords: Use the provided keyword list, but feel free to create new, relevant keywords if they better capture the input.

3. Format the output: Provide the keywords in a comma-separated list without any additional text or formatting.

Keyword List:

Style: dance, festive, groovy, mid-tempo, syncopated, tipsy, atmospheric, cold, dark, doom, dramatic, sinister, adjunct, art, capriccio, mellifluous, nü, progressive, unusual, anthemic, ${\tt emotional,\ happy,\ jubilant,\ melancholy,\ sad,\ aggressive,\ banger,}$ power, stadium, stomp, broadway, cabaret, lounge, operatic, storytelling, torch-lounge, theatrical, troubadour, vegas, ethereal, majestic, mysterious, ambient, cinematic, heat, minimal, slow, sparse, german schlager, glam, glitter, bedroom, chillwave, intimate, sadcore, carnival, distorted, glitchy, haunted, hollow, musicbox, random, arabian, bangra, calypso, chalga, egyptian, hindustani, hōgaku, jewish music, klezmer, matsuri, middle east, polka, russian navy song, suomipop, tribal Genre: appalachian, bluegrass, country, folk, freak folk, western, afro-cuban, dance pop, disco, dubstep, disco funk, edm, electro, high-nrg, house, trance, downtempo, synthwave, ${\tt trap, \ cyberpunk, \ drum'n'bass, \ electronic, \ hypnogogical, \ phonk,}$ synthpop, techno, bebop, gospel, frutiger aero, jazz, latin jazz, RnB, soul, bossa nova, forró, mambo, salsa, tango, afrobeat, dancehall, dub, reggae, reggaeton, black metal, deathcore, death metal, heavy metal, heavy metal trap, metalcore, nu metal, power metal, pop, pop rock, kpop, jpop, classic rock, blues rock, emo, glam rock, hardcore punk, indie, industrial rock, punk, rock, skate rock, skatecore, funk, hiphop, rap

Type: elevator, jingle, muzak, adan, adjan, call to prayer, gregorian chant, i want song, hero theme, strut, march, military, villain theme, lullaby, nursery rhyme, sing-along, toddler, adagio, andante, allegro, acoustic guitar, bass, doublebass, electricbass, electric guitar, fingerstyle guitar, percussion, chaotic, noise, stuttering, glissando trombone, legato cello, orchestral, spiccato violins, staccato viola, symphonic, 1960s, barbershop, big band, classic, doo wop, girl group, salooncore, swing, americana, christmas carol, a cappella, arabian ornamental, dispassionate, melismatic, monotone, narration, resonant, spoken word, sprechgesang, sultry, scream, torchy, vocaloid

Examples

 $\ensuremath{\textbf{User:}}$ Image: [A base64-encoded image of a city skyline at sunset]

System: ambient, lo-fi house moods, mellow, smooth jazz, chillwave, urban pop, dreamy, city pop

User: Image: [A base64-encoded image of a long-distance couple reuniting at the airport] | Text Note: Met after three years at the airport. Felt like a movie scene. I scanned the crowd, and there you were, smiling. I knew I was home.

System: cinematic, indie, nostalgic, intimate, troubadour, americana, orchestral, piano

User: Text Note: "Hope" is the thing with feathers - That perches in the soul - And sings the tune without the words - And never

stops at all. | User Keywords: hopeful
System: emotional, ethereal, acoustic guitar, folk, atmospheric,
storytelling ballad, soft-spoken harmonies

A.2 Chord Progression Generation: AMUSE

System Prompt

You are a musical assistant generating chord progressions based on user-provided keywords, key, mode, and bar. The keywords describe the genre, style, and song type. The key specifies a root note that is in [C, G, D, A, E, B, F#, Db, Ab, Eb, Bb, F]. The mode specifies a scale that is in [Maj, Min, Dor, Phr, Lyd, Mix, Loc, Hmin, Phdm]. The bar specifies the number of chords to generate for each progression. Your task is to create 30 diverse chord progression sconforming to the keywords, key, and mode. Each progression should consist of the same number of chords as the bar input, with each chord separated by a space ' and each progression on a new line.

Instructions:

1. Analyze Chord Functions: Determine the functions of chords in the given key and mode. Tonic (I, vi) provides resolution and stability. Subdominant (IV, ii) creates movement away from the tonic. Dominant (V, vii°) creates tension that needs to resolve to the tonic.

2. Analyze the Keywords: Determine the chord components and progression patterns based on the keywords. For example, for jazz-related keywords, consider using seventh chords, altered chords, and common jazz progressions like ii-V-I. For keywords like 'sadness' or 'emotional,' use minor chords, diminished chords, and progressions that create tension.

3. Generate 30 Chord Progressions: Create 30 distinct chord progressions that fit the specified key and mode and match the keywords. Each progression should be unique and align with the bar parameter (i.e., if bars = 4, each progression should have 4 chords). Ensure diversity by varying the chord components (root, quality, extensions, alterations, etc.), progression patterns (diatonic/chromatic), and cadences.

Each chord text can have the following components, in order:

1. Root Note: A-G, with optional accidentals (#, b, x).

2. Chord Quality: maj, min, aug, dim.

3. Extensions: Specific chord extensions such as 6/9, 7, 9, 11, 13.

4. Suspended Chords: Suspended chords such as sus2, sus4, sus#2, sus#4.

5. Added Notes: Added notes such as add2, add4, add6, add9, add11, add13.

6. Altered Notes: Alterations such as b5, #5, b9, #9, #11, b13.

7. Slash Chords: Alternate bass notes such as /E, /G#, /Bb, /Dx.

Ensure the chord progressions are musically coherent, stylistically appropriate, and diverse. Include extensions, suspensions, adds, altered notes, slash chords as needed to achieve maximum diversity. Use both diatonic and chromatic chords to enhance the progressions. Respond only with the chord progressions, avoiding any additional commentary or formatting.

Examples:

User keywords: dreamy, jazz, soft | Key: B | Mode: Maj | Bars: 4 Example Progressions: C#m7 F#7 Bmaj9 d#dim/C\nEmaj7 A#m7b5 D#m7 G#7 User keywords: singer-songwriter, acoustic, emotional | Key: F# | Mode: Maj | Bars: 3

Example progressions: F# B/F# C#/G#\nC# D#msus2 D#m/ User keywords: orchestral, adventurous, epic | Key: D | Mode: Min | Bars: 4

Example progressions: dm gm/Bb gm dm\nBb F C C#dim

Generate 30 progressions for each user input, following the above guidelines and ensuring diversity and musical

coherence. Keep the chord format the same as the examples provided (e.g., G, Amaj7, Cm are valid formats, but Gmaj, Cmin are invalid formats).

A.3 Chord Progression Generation: Baseline

System Prompt

You are a musical assistant generating chord progressions based on user-provided keywords, key, mode, and bar. The keywords describe the genre, style, and song type. The key specifies a root note that is in [C, G, D, A, E, B, F#, Db, Ab, Eb, Bb, F]. The mode specifies a scale that is in [Maj, Min, Dor, Phr, Lyd, Mix, Loc, Hmin, Phdm]. The bar specifies the number of chords to generate for each progression. Your task is to create a chord progression conforming to the keywords, key, and mode. The progression should consist of the same number of chords as the bar input, with each chord separated by a space ' ' and the progression on a new line.

Instructions:

1. Analyze Chord Functions: Determine the functions of chords in the given key and mode. Tonic (I, vi) provides resolution and stability. Subdominant (IV, ii) creates movement away from the tonic. Dominant (V, vii°) creates tension that needs to resolve to the tonic.

2. Analyze the Keywords: Determine the chord components and progression patterns based on the keywords. For example, for jazz-related keywords, consider using seventh chords, altered chords, and common jazz progressions like ii-V-I. For keywords like 'sadness' or 'emotional,' use minor chords, diminished chords, and progressions that create tension.

3. Generate a Chord Progression: Create a chord progression that fits the specified key and mode and matches the keywords. The progression should align with the bar parameter (i.e., if bars = 4, the progression should have 4 chords).

Each chord text can have the following components, in order:

1. Root Note: A-G, with optional accidentals (#, b, x).

2. Chord Quality: maj, min, aug, dim.

3. Extensions: Specific chord extensions such as 6/9, 7, 9, 11, 13.

4. Suspended Chords: Suspended chords such as sus2, sus4, sus#2, sus#4.

5. Added Notes: Added notes such as add2, add4, add6, add9, add11, add13.

Altered Notes: Alterations such as b5, #5, b9, #9, #11, b13.
 Slash Chords: Alternate bass notes such as /E, /G#, /Bb, /Dx.

Ensure the chord progression is musically coherent and stylistically appropriate. Include extensions, suspensions, adds, altered notes, and slash chords as needed to achieve a rich and satisfying progression. Use both diatonic and chromatic chords to enhance the progression. Respond only with the chord progression, avoiding any additional commentary or formatting.

Examples:

User keywords: dreamy, jazz, soft | Key: B | Mode: Maj | Bars: 4 Example Progression: C#m7 F#7 Bmaj9 d#dim/C User keywords: singer-songwriter, acoustic, emotional | Key: F# | Mode: Maj | Bars: 3 Example progression: F# B/F# C#/G# User keywords: orchestral, adventurous, epic | Key: D | Mode:

Min | Bars: 4 Example progression: dm gm/Bb gm dm

Generate a progression for each user input, following the above guidelines and ensuring musical coherence. Keep the chord format the same as the examples provided (e.g., G, Amaj7, Cm are valid formats, but Gmaj, Cmin are invalid formats).

Table 4: CLAP scores calculated using the synthetic audio used to evaluate keyword relevance in the listening study (Section 6.2). AMUSE achieves higher scores (↑), indicating better relevance compared to the baselines.

Method	$CLAP \uparrow$
LSTM Prior	0.2626
GPT-40	0.2661
Amuse	0.2685

B Chord Progression Generation Details

B.1 Rejection Sampling Implementation Details

 $P(\mathbf{x})$ consists of two stacked LSTM layers with 512 embedding and hidden dimensions, a learning rate of 1e-5, and a dropout rate of 0.2. Similarly, $Q(\mathbf{x})$ uses two stacked LSTM layers but with 256 embedding and hidden dimensions, the same learning rate of 1e-5, and a dropout rate of 0.2. The distribution is smoothed with a temperature value of 1.7 during rejection sampling.

B.2 Keywords Used for Listening Study

We list 10 sets of keywords used for the "Keyword Relevance" section of the listening study.

- urban, hip hop, trap
- powerful, rock, heavy metal
- emotional ballad, sad
- bossa nova, latin jazz, samba
- acoustic, folk, country
- soul, r&b, groovy
- smooth, jazz, swing
- lo-fi, dreamy, ambient
- orchestral, epic, cinematic

In addition to this, we included the attention check question with keywords "beautiful, calm, piano" where users select between consonance and dissonance chord progressions.

B.3 Additional Evaluation on Keyword Relevance

We report the automatic evaluation results for keyword relevance using the CLAP score [15, 103] to complement the qualitative findings from our listening study (Section 6.2.2). The CLAP score measures how well audio aligns with a text prompt by computing the pairwise cosine similarity between text and audio embeddings extracted by CLAP. CLAP is a dual-encoder model, with one encoder processing text inputs and the other processing audio inputs. These embedding spaces are jointly trained using a contrastive learning objective [95]. For evaluation, we feed the synthetic audio (used in the listening study for keyword relevance) into the CLAP audio encoder and set the CLAP text encoder input to "*An audio track of [keywords] music*," building on prior work [101, 103].

Table 4 presents the results: LSTM prior, GPT40, and AMUSE achieved CLAP scores of 26.3%, 26.6%, and 26.9%, respectively. Note that scores in the range of 20%-30% are typical for cosine similarity metrics in multimodal contrastive models (e.g., CLIP [79]) [77, 87, 101, 106]. In contrast, in our listening study, listeners rated those

systems as having better pairwise keyword relevance 41.4%, 50.7%, and 58.0% of times. For each pair in the listening study, we measured the correlation between differences in CLAP scores and listener judgments. This resulted in a percent agreement of 48.67% and a Pearson correlation of 0.05 (p=0.25). We suspect that the CLAP audio encoder cannot meaningfully embed our synthetic audio, resulting in a weak correlation overall. Although CLAP scores and listener judgments follow a similar trend (favoring AMUSE), the listening study provides a more reliable evaluation in this setting.

C User Study

C.1 Survey Questions

For the post-task surveys in the user study, participants were asked to rate their agreement with the following statements on a sevenpoint Likert scale (1=Strongly Disagree, 7=Strongly Agree). [system(s)] was either "Aria alone" (Baseline) or "Amuse and Aria together" (Assist).

- **Inspiration Support**: "Using [system(s)], I could easily translate abstract ideas (e.g., imagery, emotions) into concrete musical components."
- **Task Alignment**: "Using [system(s)], I felt the tool(s) guided me toward a composition that felt connected to the task prompt."
- Think Through: "Using [system(s)], I could think through what kinds of outputs I would want to complete the task goal, and how to complete the task."
- **Output Quality**: "Using [system(s)], the outputs I received from the tool(s) were of sufficient quality to be useful in my composition."
- **Diversity**: "Using [system(s)], I was able to receive a wide range of suggestions that I wouldn't have composed on my own."
- **Collaborative**: "Using [system(s)], I felt I was collaborating with the tool(s) to come up with the composition."
- **Controllable**: "Using [system(s)], I felt I had control over the songwriting process. I could steer the tool(s) towards the task goal."
- **Ownership**: "I feel that the composition I created using [system(s)] is mine."
- Musical Coherence: "I feel that the composition I created using [system(s)] is musically coherent."
- **Potential for Development**: "I feel that the composition I created using [system(s)] could serve as a strong foundation for a full song that I could expand on."
- Enjoyment: "Using [system(s)], I was very engaged in the songwriting activity I enjoyed and would do it again."
- **Exploration**: "Using [system(s)], it was easy for me to explore many different options, ideas, designs, or outcomes."
- **Expressiveness**: "Using [system(s)], I was able to be very expressive and creative while doing the activity."
- Immersion: "Using [system(s)], I was able to concentrate on the activity, and I forgot about the tool that I was using."
- **Results Worth Effort**: "Using [system(s)], what I was able to produce was worth the effort required to produce it."

The questions for think-through, collaborative, and controllable are adapted from the previous work that measured user perceptions on AI systems [102]. The questions for ownership and musical coherence are adapted from the related work on human-AI music co-creation [62]. The questions for enjoyment, exploration, expressiveness, immersion, and results worth effort are Creativity Support Index [16] measures.

C.2 Interview Questions

We list the questions used for the semi-structured interview after the two songwriting sessions with baseline and AMUSE.

- Can you walk me through your songwriting process when using both Aria and AMUSE together?
- Comparing the baseline and AMUSE, what were the main differences you noticed in the songwriting process?
- How did using AMUSE with Aria influence your final composition compared to using Aria alone? Did it lead to any unexpected or particularly satisfying results?
- Were there any creative challenges you faced during the process that the tools helped (or could have helped) you overcome? How did they assist (or not) in solving specific problems?
- Which features in AMUSE did you find most helpful or not? Can you describe specific scenarios where they were particularly helpful or unhelpful?
- How do you envision using tools like Aria and AMUSE in your future songwriting projects? What role do you see them playing in your creative process?
- What specific improvements or features would you suggest for either Aria, AMUSE, or both to better support your songwriting process? Are there any gaps or frustrations you encountered that you want addressed?

C.3 Thematic Analysis Codebook

We include the codebook used for the thematic analysis of user experience on AMUSE and ARIA. Table 5 presents the codes related to user experiences in the ASSISt condition where users used both AMUSE and ARIA. Table 6 presents the codes related to user experiences in the Baseline condition where users used ARIA alone. Table 5: Codebook summarizing the dimensions, descriptions, and example quotes from participants regarding the benefits, challenges, and feature usability of AMUSE integrated with ARIA in the Assist condition.

Dimensions and Codes	Code Description	Example Quote	
What are the benefits of using AMUSE with ARIA?			
Controllable outcomes	Users can steer the tool outputs to align with their creative intentions.	<i>"AMUSE was definitely more helpful when starting from scratch. The ideas it generated fell in line with I would have mentally constructed."</i>	
Diverse outcomes	The tools generate novel and di- verse outputs, providing inspira- tions or surprises to users.	"AMUSE surprised me with its use of chromaticism, push- ing me to think beyond diatonic patterns. It's been fun and engaging to work with ideas I wouldn't have come up with on my own."	
Foster imagination	The tools facilitate imaginative thinking and creative expression.	"AMUSE did allow for a more generative ideation. I felt more creatively connected to what I was actually doing musically from a very imaginative perspective."	
Streamline workflow	Users find creation process efficient and easy with the tools.	"I really liked how it jump-started the process. I usually spend a lot of time and effort getting the right chords, which is tedious—being able to convey an idea quickly was really nice."	
Agency over workflow	User has a sense of autonomy in their interactions with the tools.	"It felt like unlocking a communication channel where I can tell what I wanted, and compared to AI writing the whole thing for me, with AMUSE, I had control in the writing process."	
What are the challenges of usi	ng Amuse with Aria?		
Disconnection between tools	Switching between AMUSE and ARIA disrupts the creative work- flow.	"The only disconnection I felt was in how ARIA doesn't necessarily integrate with AMUSE in the same way that AMUSE integrates with ARIA. So it was almost like doing a 180-degree rotation when switching from AMUSE to ARIA because ARIA doesn't achieve quite the same results, given that it doesn't consider your inputs."	
Which features in AMUSE did users particularly helpful or unhelpful?			
Music keywords are helpful	Users find music keywords in Chord Generator helpful.	"Turning an image directly into chords feels like a big abstract leap. Generating keywords from the image helps me connect with it—asking, 'What keywords come to mind for this image?' makes the process feel more familiar and predictable."	
Image input is helpful	Users find image input in Chord Generator helpful.	"I especially liked the picture input feature because, as the saying goes, a picture is worth a thousand words. Sometimes, it's easier to convey what you're aiming for with an image than with words."	
Chord Transcriber is unhelpful	Users find Chord Transcriber unnec- essary or unhelpful.	"I didn't use the transcribe feature which I think mainly is because of the time constraint for me—I couldn't come up with a song that kind of fit the task that I could borrow from off the top of my head."	

Table 6: Codebook summarizing the dimensions, descriptions, and example quotes from participants regarding the benefits and challenges of using ARIA alone in the Baseline condition.

Dimensions and Codes	Code Description	Example Quote
What are benefits of using A	RIA alone?	
Intuitive workflow	Users find the standalone ARIA ex- perience straightforward.	"I would say when AMUSE was there, I felt more enriched and musical, but using ARIA alone was more intuitive because I knew what to expect from it."
Unexpected discoveries	Users receive unexpected, serendip- itous results that inspire new direc- tions from ARIA.	"I feel like ARIA generates something randomly so it could or could not work. But in this case, the fact that it worked and it actually sounded good was pretty nice experience."
What are the challenges of u	sing Aria alone?	
Uncontrollable outcomes	Users find ARIA's output random or irrelevant, failing to align with the user's creative intentions.	<i>"ARIA is a bit random – its suggestions don't sync up exactly with what I am writing."</i>
Lack of agency over workflow	Users feel they lack autonomy in their interactions with ARIA.	"With just ARIA, I wasn't able as much to create a mood or feel for my song. I didn't feel like I had much input—ARIA just sort of made up its own mind, and I had to just go along with it. It's like having a bad co-writer."