"It's more of a vibe I'm going for": Designing Text-to-Music Generation Interfaces for Video Creators

Noor Hammad
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
nhammad@cs.cmu.edu

C. Ailie Fraser
Adobe Research
Seattle, Washington, USA
fraser@adobe.com

Erik Harpstead
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
eharpste@cs.cmu.edu

Jessica Hammer
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
hammerj@andrew.cmu.edu

Mira Dontcheva
Adobe Research
Seattle, Washington, USA
mirad@adobe.com

Abstract

Background music plays a crucial role in social media videos, yet finding the right music remains a challenge for video creators. These creators, often not music experts, struggle to describe their musical goals and compare options. AI text-to-music generation presents an opportunity to address these challenges by allowing users to generate music through text prompts; however, these models often require musical expertise and are difficult to control. In this paper, we explore how to incorporate music generation into video editing workflows. A formative study with video creators revealed challenges in articulating and iterating on musical preferences, as creators described music as "vibes" rather than with explicit musical vocabulary. Guided by these insights, we developed a creative assistant for music generation using editable vibe-based recommendations and structured refinement of music output. A user study showed that the assistant supports exploration, while direct prompting is more effective for precise goals. Our findings offer design recommendations for AI music tools for video creators.

CCS Concepts

• **Human-centered computing** → **User centered design**; **Interactive systems and tools**; • **Applied computing** → **Sound and music computing**.

Keywords

Music Generation, Video Editing, Generative AI

ACM Reference Format:
Noor Hammad, C. Ailie Fraser, Erik Harpstead, Jessica Hammer, and Mira Dontcheva. 2025. "It's more of a vibe I'm going for": Designing Text-to-Music Generation Interfaces for Video Creators. In *Designing Interactive Systems Conference (DIS '25), July 05–09, 2025, Funchal, Portugal.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3715336.3735814

1 Introduction

Music is an essential component of storytelling in videos. From epic feature-length films to social media shorts, the strategic use of music can profoundly influence a video's emotional impact and memorability [38, 61]. On social media platforms, the vast majority of videos include music (in 2023, 85% of videos on TikTok and 84% of videos on YouTube contained music [21]), though the style and prominence of music in social videos varies widely. Most "narrative-style" videos such as vlogs, explainer/how-to videos, interviews, and podcasts feature dialogue, voiceover, and/or other audio. For these videos, background music can help underscore a narrative, add energy or emotion, emphasize key moments, and attract viewers [61].

While some narrative-style videos are made by professional teams, most are made by hobbyist or professional-aspiring individuals who handle every aspect of video production themselves [10, 29, 48]. These creators typically do not have expertise in composing or producing music, so they rely on existing songs as background music for their videos. However, finding the right song for a given video is a challenging and tedious process [28]. Creators must be careful to avoid using copyrighted music without permission, so they often rely on popular royalty-free stock music libraries such as Epidemic Sound [50] and Artlist.io [8]. As prior work has found [28, 44, 55], searching these libraries is hard because users have to describe their desired music in words and/or filter by a limited selection of facets (e.g., genre, mood). Creators must also ensure that the song they choose fits the duration and narrative arc of their video, which requires careful previewing of potential options. This can quickly become time-consuming and frustrating.

Recent breakthroughs in AI music generation present new opportunities for creators to make their own songs using only a text prompt [3, 15, 20, 27, 42, 49]. While this technology is promising, prior work has not yet explored how it might be leveraged in the context of video editing and how to best help users overcome the challenges in prompting generative models [22, 66]. This paper provides an initial investigation into this burgeoning design space, exploring the potential and challenges of incorporating generative music into video editing workflows. First, we conducted a formative study with eleven social narrative video creators to understand the role music plays in their videos, their current workflows and challenges, and their experience with existing music generation technologies. Our findings revealed that while many creators lack formal musical training, they often have strong opinions about the kind of music they want. However, challenges arise when they try to articulate these preferences as queries on music search platforms,

DIS '25, Funchal, Portugal

ACM ISBN 979-8-4007-1485-6/25/07
https://doi.org/10.1145/3715336.3735814

as creators primarily describe music in terms of its emotional impact or "vibe", rather than technical aspects like instruments, genre, or tempo. We also found that participants valued creative control as well as the ability to explore options and iterate, but current music search interfaces make exploration feel aimless and unproductive.

Based on our formative study as well as prior work, we distill a set of design goals for interfaces that bring text-to-music generation into the hands of video creators. Through an iterative design process, we designed and implemented a creative assistant that embodies these goals with a focus on balancing automated guidance with user control. This assistant automatically recommends an editable "vibe" based on the user's video, translates this vibe text into generated music, and provides a structured interface for iterating on the generated music. Using this prototype as a design probe, we conducted a user study with seven video creators who generated music with and without the creative assistant to understand how this interaction paradigm compares to directly prompting a music generation model. Our findings suggest that vibe-based assistance is useful when creators are open to new ideas, while direct prompting is valuable in situations where creators already have specific song attributes in mind.

In this work, we make the following contributions:

- Findings from a formative study with social video creators that underscore the importance of emotional impact or "vibe" in describing music for videos
- The design and implementation of a vibe-based creative assistant that integrates text-to-music generation into video editing workflows
- Design recommendations for generative music interfaces for video creators, including offering initial guidance based on users' videos, surfacing intermediate structures, and allowing navigation between layers of control

2 Related Work

Despite the importance of music in videos, there is limited existing HCI research on the needs and workflows of video creators when it comes to music [28]. In this section, we first discuss the general challenges of human-AI co-creation, then discuss common challenges and potential solutions from prior work regarding music search and recommendation, and AI music generation.

2.1 Human-AI Co-Creation

Iteration is a key part of the creative process [22, 68, 69], but when it comes to prompting large models, iteration can often feel unproductive due to their opaque nature [22, 34, 64, 66]. Recent work has explored how AI tools can better support productive iteration by designing human-AI co-creation tools across creative domains including writing [14, 18], image creation [34], video editing [58, 60], and music creation [28, 35, 69].

One common challenge with interacting with large AI models is developing an effective prompt. Prompting a model in natural language might sound intuitive, but in reality, writing effective prompts can be challenging for non-experts due to the "black box" nature of AI models [9, 22, 64, 66] Prior work has shown that effective prompts often conform to specific structures [9, 22, 64]. Human-AI co-creation tools therefore often provide the user with prompt templates [14] or transform user input to improve the quality of a prompt [58, 60, 69].

Prior work has also shown that breaking complex tasks down into smaller, more concrete subtasks can improve model transparency, feelings of control, and output quality [64]. Human-AI co-creation systems break the creative process into concrete stages to help users more easily guide AI output and iterate within individual stages [18, 58, 60, 69]. Our design process was informed by this prior work, and we demonstrate through a design probe how similar strategies might be applied to the creative task of generating music for an edited video in order to give users creative control and enable productive iteration.

2.2 Music Search and Recommendation

Frid et al. [28] identified common workflows and challenges in adding music to videos through two formative questionnaires with social video creators. They found that most creators source music from free or paid online music libraries, and the main factor they consider when searching for music is its emotion or mood. Most popular stock music libraries rely on keyword-based search, but constructing a search query with keywords can be difficult, especially for non-music experts who don't have a specific song in mind [44, 55]. Filtering songs by attributes limits the user's expressivity and control to a small set of platform-defined attributes (usually genre, instruments, mood, and theme [8, 50]). Research on multimodal text-audio embeddings shows promise for more effective natural language music search [30, 37], but would still require the user to articulate a concrete description of their goal. Some music platforms also support searching for "similar" music to an input song [2], but this is only useful if the user has an example song in mind.

Prior work has proposed methods to automatically recommend music for a given video, typically by training a model on an existing dataset of music-video pairs [25, 33, 39, 45, 46, 55, 65]. This can be helpful for situations where the user does not have a specific idea in mind. However, most of these models either consider only the visual attributes of the video, not existing audio such as dialogue [39, 46, 55], or they explicitly focus on the domain of music videos, where music is at the forefront and there is typically no other audio [25, 33, 45]. Approaches also exist for recommending music based only on the textual content of an audio story [62], but for narrative-style videos, both the visuals and the dialogue/voiceover are important [61]. Inspired by this prior work, our work demonstrates how both a video's transcript and visual data can be leveraged to recommend styles of music to generate.

Most of the above methods are fully automatic pipelines that do not allow the user to specify their intent or iterate on the recommended music [33, 45, 46, 55, 65]. There are a few exceptions; McKee et al. [39] allow the user to include a natural language text description of their desired music, and MuseChat by Dong et al. [25] allows the user to iterate once on the recommended song with a natural-language response. However, these systems were not evaluated with users, so it is unclear how they would fit into a video creator's workflow.

2.3 AI Music Generation

Research and development of text-to-music generation models has exploded in popularity recently, both in academic research [3, 15, 20, 26, 27, 40–42, 49] and in commercial platforms (e.g., Stable Audio [5], MusicFX [31], Suno [54], Udio [56], and Riffusion [4]). Most commercial interfaces provide an open-ended text box as the main user input, often with suggested example prompts and keywords the user can try. These models present a new opportunity for non-music experts to quickly create and iterate on their own personalized songs. While AI-generated music has been an active area of research in the Computer Music research community for decades [19], the more recent text-to-music models are an exciting development for two main reasons: 1) they generate high-fidelity raw audio files, as opposed to symbolic (MIDI) music which consists of a limited number of prescribed instrument tracks [28]; and 2) they take natural language as input, as opposed to lower-level music parameters such as time signature and pitch [28]. However, these benefits also have trade-offs. First, while high-fidelity raw audio provides richer and more realistic soundscapes than MIDI music, it is hard to control the output of these models at a fine-grained level with a single text prompt [67, 69]. Second, although these models can take any text as input, like other generative models [9, 22, 64] they tend to perform better when the input has more structure and mentions specific musical properties such as instruments and tempo [6, 67] (see Stable Audio's user guide [6] and the examples typically shown in commercial products [5, 31, 54, 56]).

Prior work has proposed interactions to help users express their intent and control music generation models such as steering along specific structured dimensions [36], conversational dialogue for iteration [69, 70], and drawing notes as input [35]. Zhang et al. [69] demonstrated how an LLM can help interpret the user's intention in order to select the best fitting AI model for the type of edit they want to make to a song (e.g., add or remove tracks, re-generate sections, transpose key). Recent work has also begun to explore how generative models themselves might directly support more user control by providing additional inputs such as images [17], audio [24, 63], and other time-varying attributes like dynamics or rhythm [40, 41, 63]. However, none of this work has considered how music generation might apply to video creators, who have more constraints around the music they want based on their video, such as duration and timing [28].

2.4 Music Generation for Videos

Some recent work has developed AI models that can directly generate music given a video as input, by learning a correspondence between music and video representations [23, 53]. Similar to the aforementioned research on automatic music recommendation, these methods show promise for helping users get started without needing to explicitly describe their intent, but they do not address how users might control and iterate on this initial output.

To our knowledge, the only work that has specifically explored interactions for video editors to generate music is Frid et al.'s Music Creation by Example [28]. This paper proposed an interaction paradigm for non-music-expert video creators to generate symbolic (MIDI) music based on a reference song, and iterate on the output by combining different instrument tracks from different generated songs. Our work is largely inspired by this paper and its focus on bringing powerful AI systems into the hands of non-expert users with a careful balance of automation and control. We build on this work by exploring how interfaces for high-fidelity text-to-music generation can strike a similar balance, to bridge the gap between how video creators conceptualize music and the language generative models expect. Because symbolic music generation systems require lower-level input parameters, a main contribution of Music Creation by Example was to extract low-level features from a reference song and use those as input parameters to generate new music. In contrast, since modern music generation models use text prompts as input, our approach focuses on helping the user express their intent in natural language and translating that intent into a structured prompt that the model can understand. We believe these two approaches can be complementary, as both text and reference songs are likely to be important input modalities for video creators.

3 Formative Study

To design music generation interfaces for video creators, we first needed to understand their current experiences with AI music generation, the role music plays in their videos, their workflows, and the challenges they face. To this end, we conducted a formative study with eleven experienced social video creators. Our study focused on the following questions:

- What workflows and strategies do video creators employ when adding music to their videos?
- What challenges do video creators face when adding music to their videos?
- How do video creators conceptualize music?
- What are video creators' perceptions of AI music generation?

3.1 Participants

We targeted social media video creators who produce narrative-style videos that include background music, such as vloggers, how-to YouTubers, and travel content creators. Our inclusion criteria required participants to be at least 18 years old, manage their own production and editing processes, and have publicly available content on a platform such as Youtube, Instagram, or TikTok. To reach this audience, we posted a recruitment message on X as well as corporate and university Slack channels, engaged with specialized content creator communities such as Discord servers for YouTubers, and sent cold emails to YouTubers whose work aligned with our study criteria. Prospective participants completed a screener survey to ensure eligibility, which included questions about their video editing experience, links to their video channels, frequency of music use in videos, and their preferred platforms for finding music.

In total, eleven participants took part in the study. Table 1 summarizes the diverse range of content types and platforms used by our participants.

3.2 Procedure

Each study session was conducted remotely via Microsoft Teams, consisting of a 25-minute semi-structured interview, a 25-minute task walkthrough, and concluding with a ten-minute semi-structured interview. Prior to the study, participants were asked to prepare

Participant	Content Type	Content Platform(s)	Music Platform(s)
P1	Get Ready With Me	TikTok, Instagram	CapCut stock music
P2	Travel, vlog	YouTube	YouTube, SoundCloud
P3	Sport highlights	YouTube, Instagram	YouTube
P4	Short film, vlog	YouTube, Instagram	Spotify, Artlist.io
P5	Vlog, TV show reviews	YouTube	Vllo stock music, SoundCloud
P6	DIY crafts	TikTok, Instagram	Instagram
P7	Vlog, travel, short film	YouTube	Epidemic Sound
P8	Vlog	YouTube	Epidemic Sound
P9	Vlog, travel	YouTube	YouTube
P10	Vlog, video diary	YouTube	Epidemic Sound, YouTube
P11	Travel, documentary	YouTube, Instagram	YouTube

Table 1: Information about our formative study participants, including types of videos they make, the platforms they post their videos on, and the platforms they use for music.

an in-progress project to share during the task walkthrough. In total, sessions lasted approximately 60 minutes and were recorded and transcribed for analysis. Participants received a $30 gift card as compensation.

The first semi-structured interview focused on participants' existing workflows and challenges. Questions covered the role of music in their videos, when in the editing process they typically add music, their current strategies for finding and adding music to their videos, what challenges they face during this process, and what an ideal workflow might look like.

Next, participants completed a task walkthrough and think-aloud, adding music to the in-progress video project they had prepared. We interspersed the walkthrough with additional questions about the challenges in music selection, strategies for resolving issues, and their approach to finalizing a song choice. The task walkthrough concluded with a "music supervisor" question: we asked participants to describe their desired music as if they were delegating the music search and selection process to an expert.

We closed the session with additional semi-structured questions focusing on participants' experiences with AI music generation systems. These questions included whether they had used music generation technologies before, their impressions of such tools, and their requested features for future music-generation systems.

3.3 Results

Two members of the research team transcribed and cleaned session transcripts. We then employed an inductive coding approach to identify key themes [11].

Overall, music played a critical role in participants' videos in several ways: establishing the creator's brand identity (P6, P11); "captur[ing] the essence" of the content (P1), particularly by augmenting storytelling (P3); and promoting viewer engagement (P3, P5, P9). Our interviews revealed the challenges that participants face when attempting to achieve these musical goals. We elaborate on these challenges as well as the remainder of the study's findings below.

3.3.1 Rough-cut first workflow is commonly used for narrative-style videos. In line with prior work [28], most participants used a *rough-cut first* approach, where they edited raw footage into a draft video before integrating music (P1, P2, P3, P5, P6, P7, P8, P9, P10, P11). While some participants were thinking about the music they wanted as early as filming (P1, P4), most began thinking about music only after completing their rough cut, allowing the video's emerging structure to guide their music selection process.

3.3.2 Participants conceptualize music in terms of "vibe". As prior work suggests [28], participants focused on the mood they wanted the music to evoke, rather than referring to musical attributes like instruments or genre. Notably, all participants used the word "vibe" when describing their goals for the emotional resonance of a piece of music. While the term is somewhat ambiguous, it generally refers to the "mood or character of a place, situation, or piece of music and the way they make you feel" [13] and is a common colloquialism in modern English. Participants needed the vibes of the music to align with the "overall feeling" (P6) of the video. For example, P4 described their desired music as "beach sunset summer vibes", which illustrates the flexibility of the term "vibe" when layering multiple moods or atmospheres into a single description. When asked about future interfaces, participants imagined getting music recommendations based on a video's vibe, which they felt would streamline their process (P1, P2, P5, P7, P8).

Some participants also referenced other media to help articulate their vision, whether by pointing to existing artists (e.g. Taylor Swift), soundtracks (e.g., Interstellar), or broader categories like movie styles and genres (e.g., Disney). For instance, P5 described her desired music as "Disney meets Romcom," remixing familiar genres to communicate the tone she wanted. This form of musical reference differs from what previous research has identified [28], as participants are not simply pointing to specific songs but invoking broader cultural touchstones, such as styles, moods, or iconic associations (e.g., "Interstellar" or "Disney") that immediately convey a shared understanding of the desired tone.

3.3.3 Music must align with the video. When asked to describe their desired music to a music supervisor, nine out of eleven participants began by providing an overview of the video's narrative to offer context, including both verbal descriptions and scrubbing through their footage while screensharing. Some participants went a step further, offering a detailed breakdown of how they wanted the music to evolve throughout the video, describing specific points where the music should crescendo or transition to a different mood (P1, P8). In other words, narrative video creators needed music to align with the video's *structure* and *story*, in addition to producing the right vibe.

3.3.4 Music search is messy and inefficient. In line with previous findings [28], participants attempted to develop search strategies that addressed known challenges including finding royalty-free music, finding music of the right length and mood, and finding music that would need minimal editing. Strategies included browsing well-known music platforms such as Spotify or Epidemic Sound (P4, P7, P8), using stock libraries integrated into their video editing apps such as CapCut or Vllo (P1, P5), and exploring music from other creators' videos (P2, P4, P6, P10, P11).

Overall, participants were dissatisfied with the search strategies available to them, feeling that the process was time-consuming and aimless. They could not predict whether their search was "going in the right way" (P10), and often felt that finding the right music was a matter of luck rather than skill. Although they wanted to search for music based on vibes, they hesitated to use keywords for searching (P1, P2, P4, P5, P6, P8, P9, P10, P11). It was hard for them to translate "something so abstract like a vibe into something that assists you to find it" (P1) and some believed they were "bad at figuring out [...] keywords" (P2), in part because existing search tools often failed to interpret their keywords accurately. For example, the keywords "Spanish Romantic" returned results fitting only one of the terms, but not both (P2). Even when they found a song that matched their desired vibe, it often turned out to be too short or too long for their video, making it unusable (P1, P2, P8, P10). These challenges led many participants to reuse music to save time (P1, P2, P5, P9. P10), at the cost of settling for less-than-perfect music (P10).

3.3.5 Music results can fail when validated with the video. After searching, participants validated their song choice by listening to the music while watching their video (P2, P3, P4, P5, P7, P8, P9). Rather than listening to songs all the way through, participants typically sampled short segments—ranging from two to thirty seconds—to determine whether a track matched the desired vibe (P1, P7, P8, P9). Some participants scrubbed through songs, jumping to the middle or end for brief evaluations (P4, P5). If the "vibes aren't vibing" (P2), the music would be rejected.

Because participants often evaluated multiple songs, they did not want to go to the effort of importing all their music options into their video timeline, so instead played them from their original source. However, the context switching required to validate songs against the video can also be chaotic. P2, for example, described how relying on YouTube often left her with numerous tabs open that "usually have no relation to one another" as she tried to validate potential songs against her rough cut (P2). Participants desired a more streamlined approach to comparing multiple songs against their videos, which was seen as an important part of the creative process (P1, P2, P6, P7, P8, P10, P11). As P7 put it, "I'm not going to be sure [of my song choice] because I haven't kissed enough frogs to get my prince. You know, there's actually a lot of value in kissing frogs."

3.3.6 Participants see potential in music generation, but want to maintain creative control. Most participants were unfamiliar with AI music generation technology (P1, P2, P5, P6, P7, P8, P10, P11), and none had used it in their videos. Nonetheless, participants believed that music generation could address key challenges in their search and evaluation processes, such as helping them create music that matches their desired vibe (P1, P8), enabling quick comparisons between multiple options (P5, P7), and creating personalized instrumentals that alleviate the need to reuse the same songs in every video (P2, P10). Participants also expressed interest in using AI for manipulating musical structure, such as producing variable-length songs tailored to their needs (P1, P8, P9, P10) or aligning the instruments or sections with their video's narrative (P1, P4, P5, P6).

Despite the interest in music generation to help with their process, particularly the potential of natural-language interfaces (P2, P5, P7, P10), participants also wanted to maintain creative control over features such as song length, instruments, and the overall vibe (P1, P3, P4). P4 argued that relinquishing creative control is "what a lot of creatives are scared of," because the process of engaging with the musical material is a valuable creative act. Instead, P4 envisioned music generation tools "that can enhance your ability to be creative, instead of just doing it all for you."

4 Design and Prototyping

Based on our formative study findings as well as prior work, we identified several key design goals to guide our exploration of integrating music generation into video editing workflows. With these goals in mind, we designed and implemented a creative assistant as a design probe. This section details the goals and process that went into this design, followed by a description of the prototype we built.

4.1 Design Goals

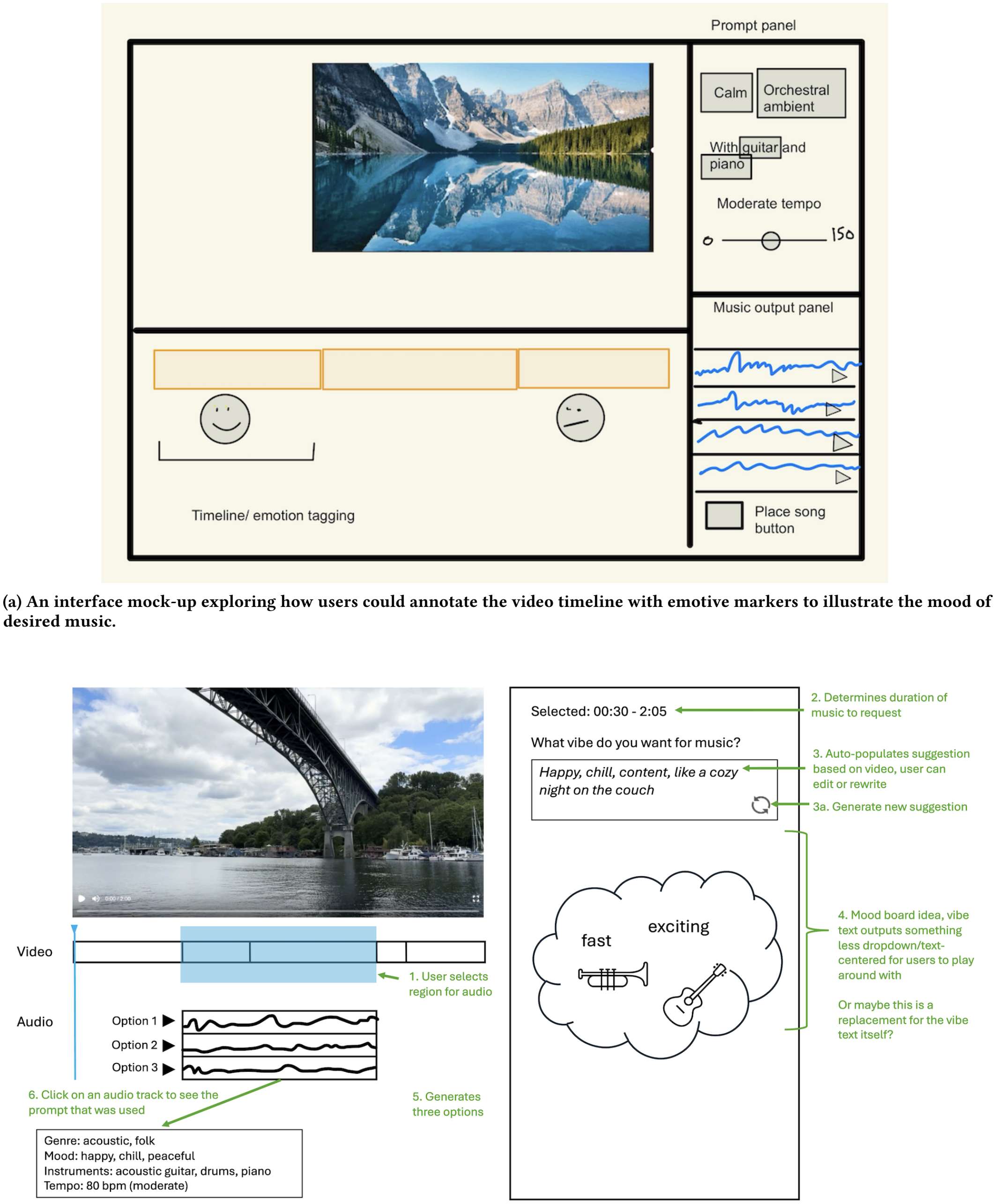
(DG1) **Support a rough cut-first approach:** For narrative-style videos, the most prevalent workflow among participants was to add background music to the video after it has been edited.

(DG2) **Center on vibe and emotion:** Participants consistently described music in terms of its "vibe," corroborating related work [28, 44]. Thus, interfaces should abstract lower-level musical attributes (e.g., genre, tempo, instruments) in favor of higher-level descriptors like vibe and emotion.

(DG3) **Consider video context:** Music must account for the narrative and content of the video [28, 61]. Participants in our formative study framed their music choices by describing their video's content, emphasizing the need for music that aligns seamlessly with the story being told in their videos.

(DG4) **Enable music validation with video:** Validating candidate songs with the video before adding them to the timeline was a critical step for participants. This process includes ensuring the music matches the video's vibe and aligns with its duration [28].

(DG5) **Support structured exploration of multiple music options:** Participants emphasized the importance of exploring multiple song options, but their current strategies for doing so were overwhelming and lacked direction. Interfaces that add structure to the exploration process can help it feel more productive.



(a) An interface mock-up exploring how users could annotate the video timeline with emotive markers to illustrate the mood of desired music.

Video
Audio
Option 1
Option 2
Option 3
Selected: 00:30 - 2:05
What vibe do you want for music?
Happy, chill, content, like a cozy night on the couch
fast
exciting

1. User selects region for audio

2. Determines duration of music to request

3. Auto-populates suggestion based on video, user can edit or rewrite

3a. Generate new suggestion

4. Mood board idea, vibe text outputs something less dropdown/text-centered for users to play around with

Or maybe this is a replacement for the vibe text itself?

5. Generates three options

6. Click on an audio track to see the prompt that was used

Genre: acoustic, folk
Mood: happy, chill, peaceful
Instruments: acoustic guitar, drums, piano
Tempo: 80 bpm (moderate)

(b) An interface mock-up exploring how a mood board could help users illustrate desired characteristics of music.

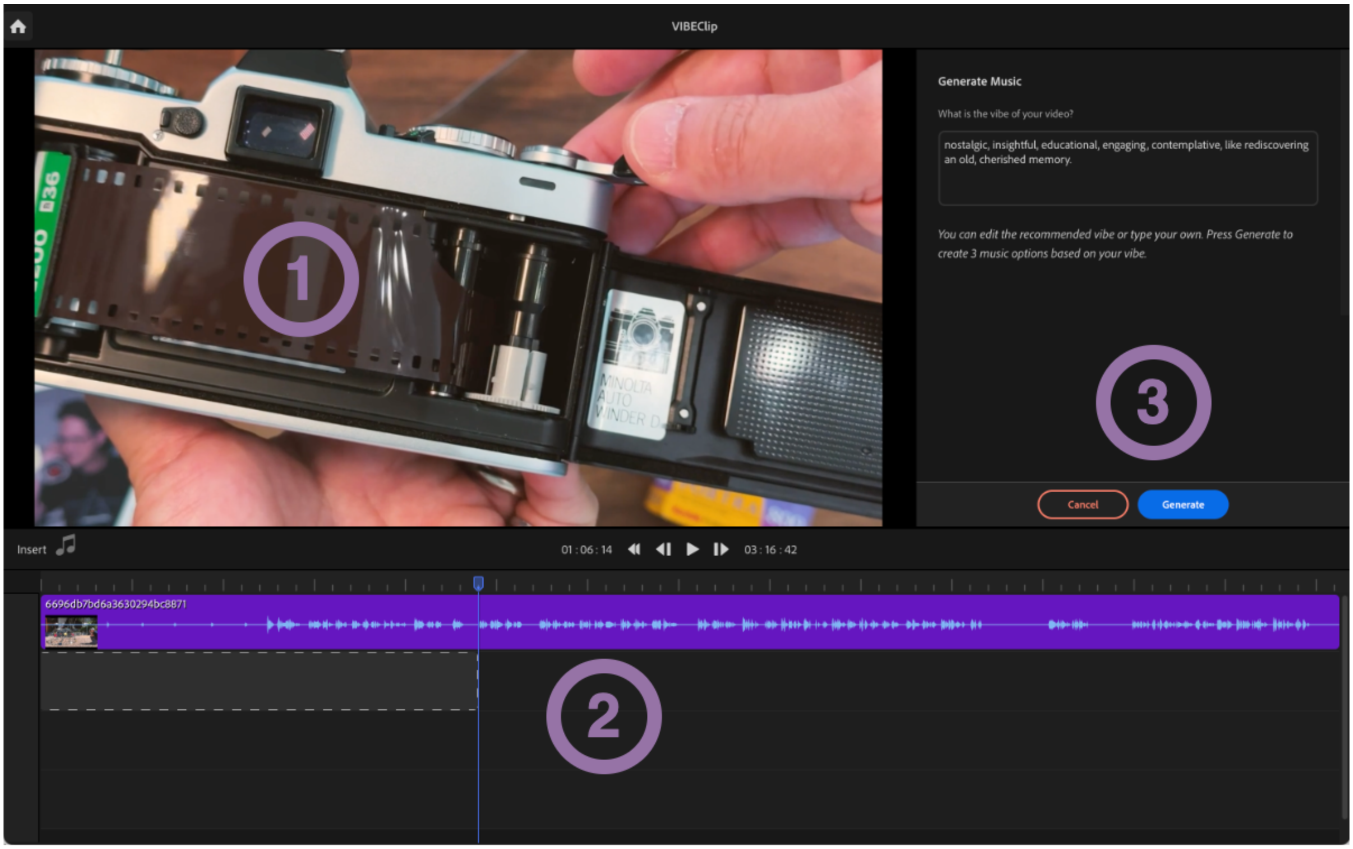
Figure 1: Early prototype sketches created during our iterative design process

4.2 Design Process

With these design goals in mind, we conducted an iterative design process to brainstorm how an interface for text-to-music generation could support video creators. Central to our exploration was determining the inputs the system should process. These fell into two categories: *implicit system inputs for automated assistance* and *explicit user inputs to articulate intent.*

For implicit system inputs, we explored leveraging video data to inform music guidance (DG3) such as:

- Transcript and Audio: Emotion, topic, and narrative cues
- Visual Information: Scene descriptions, prominent objects, people, colors, and actions
- Editing Features: Cut points, cut pacing, text effects, and sound effects



Generate Music

What is the vibe of your video?

nostalgic, insightful, educational, engaging, contemplative, like rediscovering an old, cherished memory.

You can edit the recommended vibe or type your own. Press Generate to create 3 music options based on your vibe.

Figure 2: The creative assistant prototype featuring a (1) video player, (2) a multi-track timeline, and (3) a music generation panel.

For explicit user inputs, we explored modalities that allow users to express their intent or refine system-generated suggestions. These included:

- Natural language text: E.g., “uplifting, cinematic”
- Visual sketches: E.g., Graphing emotional curves over time
- Audio-based inputs: Humming or uploading reference songs
- Structured input to specify musical attributes: E.g., drop downs to select genre, instruments, tempo, etc.

One early sketch introduced emotion tagging directly on the timeline, where emoji-like characters represented the changing vibe of the video (Figure 1a). Though this visual approach helped convey temporal changes, it lacked the expressivity needed for the types of nuanced, verbal vibe descriptions our formative study participants used. We excluded the modalities of humming and sketching because none of our participants mentioned or used them, and we did not want to overwhelm users with novelty given that AI music generation was already a novel technology for most participants. In a later mock-up, we tried supplementing natural language text input with additional “mood board” features, allowing users to drag and drop musical elements (e.g., instruments or themes) (Figure 1b). In the final design, we replaced the mood board with structured text boxes (Figure 3) to give users the flexibility to type their own text while also producing structured prompts that the music generation model can understand [67]. In other words, although we explored many options in our design process, we chose natural language text as the primary input modality due to its approachability and prevalence in current generative AI interfaces, and because it aligned effectively with participants’ existing practices for expressing musical intent.

Eventually, we converged on an interaction paradigm where users can upload a rough-cut video (DG1) and select a section of the video timeline to define the music’s duration (DG4). The system automatically recommends a vibe based on the video content in that section (DG3). Users can then edit the textual representation of the vibe before submitting it as input for music generation (DG2). The system generates multiple different songs based on this vibe, and users can then refine finer details of each song in a more structured interface (DG5), which is visible alongside the video (DG4). Like most commercial music generation interfaces today [5, 31, 54, 56], our approach provides the user with suggestions to get them started, but unlike commercial interfaces, these suggestions are tailored to their specific video.

This approach provides a higher-level abstraction of music at the start of the process to support the user’s mental model, while providing more structure and control over musical attributes during the refinement stage. By structuring the music generation process around multiple intermediate stages, our design supports a human-in-the-loop paradigm, striking a balance between automated guidance and user control, and aligning with principles from human-AI collaboration research [64]. Importantly, the user is always able to modify or rewrite any AI-generated text to ensure it captures their intent.

4.3 Creative Assistant Prototype

We implemented a prototype that embodies our design in a creative assistant powered by a generative AI pipeline (Figure 2). The prototype comprises a video player, a music generation panel, and a multi-track timeline. These interface elements and the core interactions are drawn from existing video editor paradigms (e.g., Premiere Pro, Final Cut Pro, CapCut) but are limited in functionality to only the music-related tasks described below. The timeline displays the user's video on the first track, and all generated music is rendered on audio tracks below (DG4).

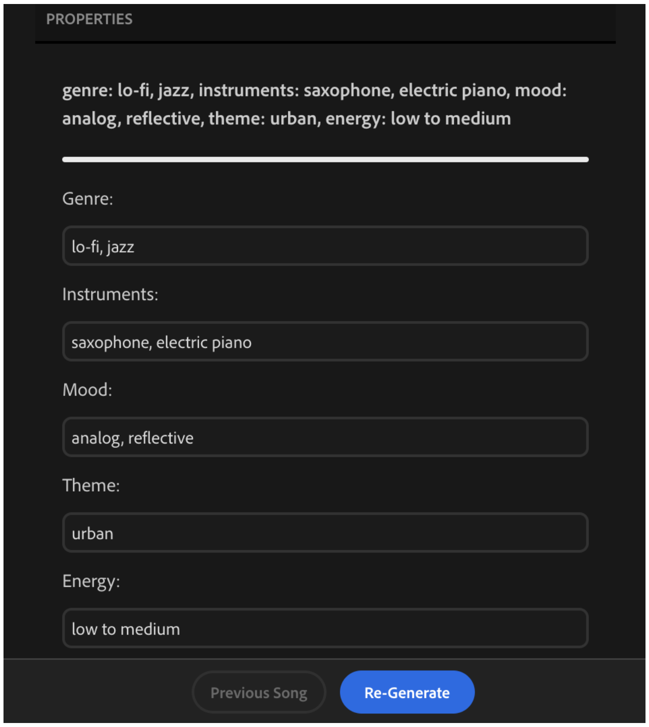
4.3.1 Video understanding. As an offline pre-processing step, we compute two metadata features from the user's video: a transcript of all spoken dialogue in the video, and visual scene captions comprising a textual description of the visual contents of each salient scene in the video. This metadata is used in the following two steps (DG1, DG3).

4.3.2 Vibe recommendation. The user starts by selecting a region on the timeline where they would like to add music. This determines the duration of the music that will be generated and its position on the timeline (DG4). Once a region is selected, a recommended "vibe description" will automatically appear in the music generation panel (DG2). This recommendation is generated by prompting a large language model (LLM) to describe the overall vibe/emotion of the selected section of the video (see A.1 for our prompt template). The prompt first includes the transcript and visual captions from the selected portion of the video to target the user's selection, followed by the full transcript and visual captions of the entire video. We include the full transcript and captions to ensure that the LLM has access to the broader narrative of the video with the goal of increasing the relevance of the vibe descriptions (DG3). Since "vibe" is an inherently ambiguous and subjective concept, we also allow the user to edit the recommended vibe description to ensure it captures their intent before moving to the next step.

4.3.3 Translating vibe into structured music prompts. When the user is satisfied with their vibe description, they can press the "Generate" button, which translates the vibe description into three distinct music prompts and submits each prompt to a text-to-music generation model. To achieve this translation, the system prompts an LLM to generate three music-generation prompts that incorporate information about the selected clip and the overall vibe (DG2, DG4). We generate three prompts at once to provide the user with multiple options for exploration (DG5) that capture different potential interpretations of the vibe, since there is no one correct way to map an abstract vibe to a concrete music description.

The structure we ask the LLM to use for each prompt is as follows:

- "genre": "1-2 genre descriptors",
- "instruments": "1-2 instrument descriptors",
- "mood": "1-2 words describing the mood of the video",
- "theme": "1-2 words describing the video theme",
- "energy": "energy level of the video"


PROPERTIES

genre: lo-fi, jazz, instruments: saxophone, electric piano, mood: analog, reflective, theme: urban, energy: low to medium

Genre: lo-fi, jazz

Instruments: saxophone, electric piano

Mood: analog, reflective

Theme: urban

Energy: low to medium

Figure 3: Interface shown in the music generation panel when the user selects a song for refinement. They can optionally edit any of the prompt's properties before clicking "Re-Generate," after which the "Previous Song" button will be enabled. If the user clicks "Previous Song," the UI will update to show the previous prompt, and the "Re-Generate" button will change to "Next Song," allowing them to navigate between versions.

This prompt structure is based on publicly available prompt guidance for the popular music generator Stable Audio [6], common music attributes used by popular stock music libraries [8, 50, 52], as well as trial-and-error testing during development. See A.2 for our prompt template.

4.3.4 Generating and previewing music. The system submits three requests to the text-to-music service, each with a prompt as per the previous section, along with a randomly-generated seed. The seed is a random number that serves as a starting point for AI generation. We use a different seed for each prompt, following the typical default behavior for generative AI services [1]. Once they are ready, the three generated music tracks are placed on the timeline in the region the user had selected, in three vertically-aligned tracks below the video track. Users can mute and unmute each music item, drag and drop it in the timeline, and delete it using the Delete key. Finally, they can use the playhead to seek and the play/pause button to play the video, as they would in a regular video editor. The video plays in sync with any unmuted music items (DG4).

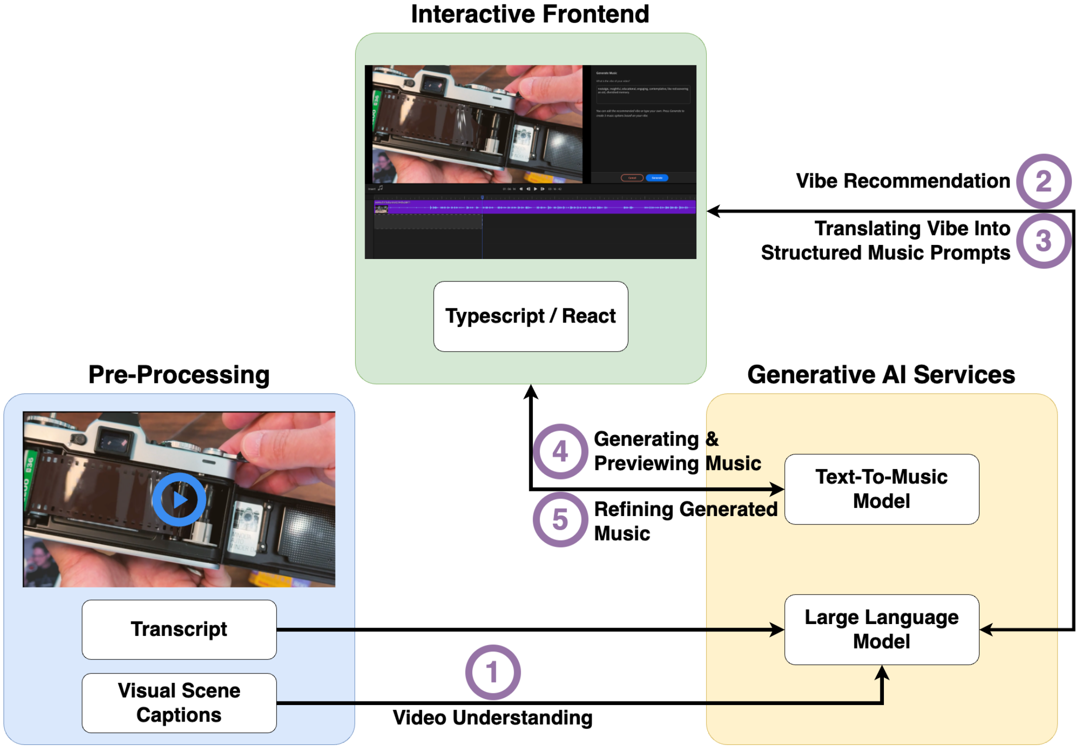


Figure 4: Overview of our prototype implementation, starting from the (1) Video Understanding step that generates transcripts and captions for (2) the vibe recommendation. Once the user submits the vibe (3) the system translates the vibe into structured music prompts for (4) the text-to-music model to generate three songs. To iterate on the songs, the user can (5) refine the structured prompt.

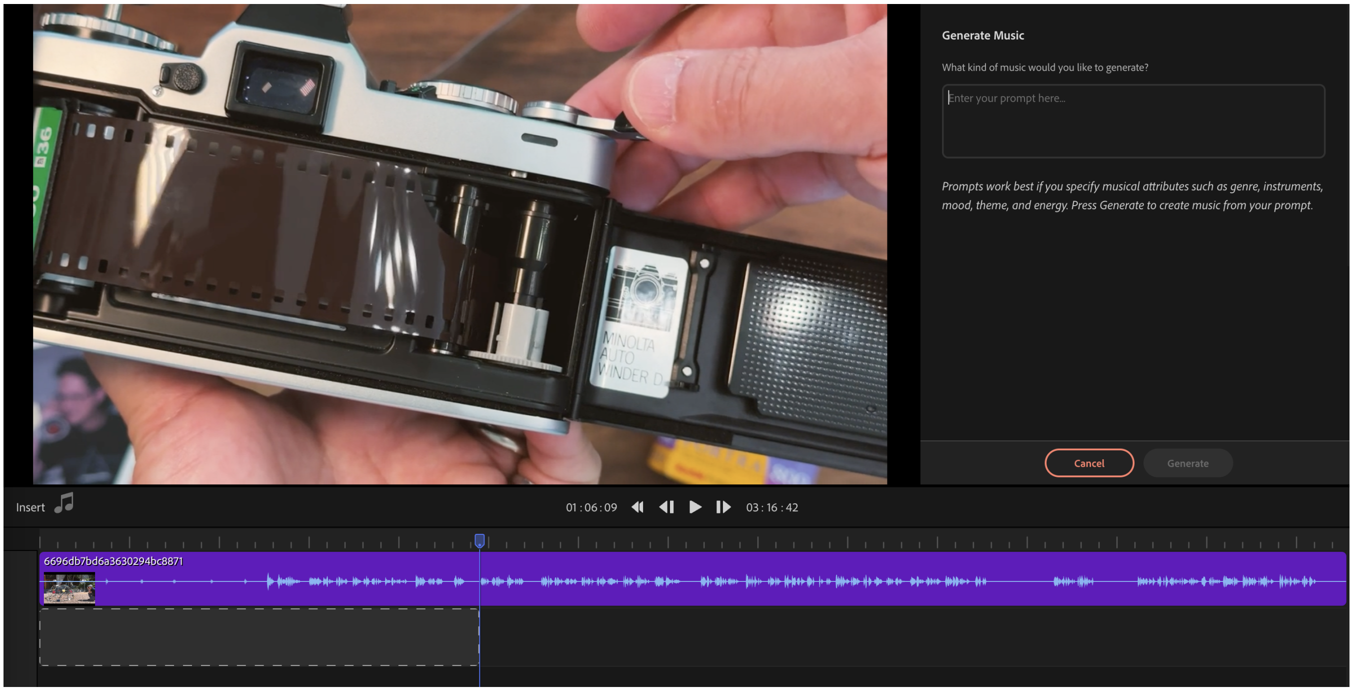
4.3.5 Refining generated music. The user can select a music item on the timeline to see the prompt that generated it in the music generation panel. Since the prompt conforms to a specific structure (see section 4.3.3), the interface breaks the prompt into independently editable text boxes for genre, instruments, mood, theme, and energy (DG5) (Figure 3). If the user clicks "Re-Generate" without modifying the prompt at all, the system uses a new random seed in its request to the text-to-music service; otherwise the same song would be generated again. If they did modify the prompt, it uses the same seed as before. This can help maintain consistency across iterations when making minor tweaks to the prompt (e.g., changing one instrument), as it means the generative model will start from the same initial number, making it more likely to produce a similar song. The generated result replaces the selected music item on the timeline, and the user can toggle between iterations using the "Previous song" and "Next song" buttons to hear how their prompt changes affected the resulting music (DG4, DG5).

4.3.6 Prototype Implementation. Figure 4 illustrates the core components of our prototype's implementation. First, the user uploads a rough cut video. Our offline pre-processing step computes a transcript of the video using Speechmatics [51] and visual scene captions using CLIP [43] and LLaVA [47]. We used GPT-4-Turbo as the LLM that generates the vibe recommendation (section 4.3.2) and the structured music prompts (section 4.3.3). To generate music, the prototype uses a custom text-to-music generation model, trained on licensed instrumental-only music. The model takes a text prompt and desired duration as input, and generates up to five minutes of instrumental music. Finally, all user interactions are done through our web interface, developed using Typescript, React, and MobX.

We note that the implementation details of these features are not part of this paper's contribution. In practice, each element could be replaced by other technical solutions, e.g., a different text-to-music model, as long as the components of the system can communicate as per Figure 4.

5 User Study

Our design and prototyping phase yielded a high-fidelity design probe which allowed us to explore different aspects of the emerging design space, given limited current understanding of how users interact with music generation tools in the context of video editing. We invited seven video creators to use our probe along with a similar prototype with a standard text-prompt interface. We introduced the second prototype as a baseline to compare how our probe, whose design enables assisted human-in-the-loop interactions, compares to the type of interface music generation tools use today. We examined how participants experienced the features of both prototypes and how they compare to their current workflows for sourcing music.


Generate Music

What kind of music would you like to generate?

Enter your prompt here...

Prompts work best if you specify musical attributes such as genre, instruments, mood, theme, and energy. Press Generate to create music from your prompt.

Figure 5: The baseline prototype used in the user study. The music generation panel features a single text box where the user enters their prompt.

Below we discuss the study details, including the baseline prototype, and the study results.

5.1 Baseline Prototype

Since none of our formative study participants had any previous exposure to existing AI music generation systems, we wanted to explore how directly prompting a text-to-music model compares to the assisted workflow found in the probe. We thus created a prototype that does not include any video analysis or automated assistance (Figure 5). As in the probe, the user selects a region of the timeline where they want music. Once a region is selected, a text box appears in the music generation panel, where the user can enter a natural-language prompt describing the music they want. When the user clicks "Generate", their prompt is sent to the music generation service. Unlike the probe, only one song is generated at a time.

When the user selects a music item on the timeline, they can see and modify the prompt that generated it in the music generation panel. As in the probe, clicking "Re-Generate" will submit the new prompt to the music generation service. We chose to keep the "Previous Song" and "Next Song" buttons from the probe since existing music generation interfaces also allow users to explore their previous prompts and results [5, 31, 54, 56].

5.2 Participants

As with the formative study, we targeted social video creators who produce narrative-style videos with background music, with one added constraint: participants were recruited internally from a large tech company due to the prototypes' use of an internally deployed text-to-music model. Interested individuals filled out a screener survey, where we asked about the types of videos they create, how they source their music, and their general video production practices. Selected participants were emailed consent forms and scheduling information. In total, we recruited seven participants, three of whom had also participated in the formative study. Importantly, the formative study did not include exposure to generative music or mention the prototypes, ensuring that repeat participants approached this user study without prior bias or familiarity with the systems being tested.

Table 2 summarizes the content types and platforms used by our participants in the user study.

5.3 Procedure

Study sessions were conducted remotely via Microsoft Teams, with participants sharing their screens and audio to allow for real-time observation. In total, sessions lasted approximately 80 minutes and were recorded and transcribed for analysis. Participants received a $50 gift card as compensation.

Since our prototype assumes a rough-cut-first workflow (DG1), participants were asked to submit two in-progress videos in rough-cut form, simulating the point in their workflow just before they begin searching for and adding music. They submitted videos ahead of time so we could pre-process them and add them to our internal database.

Each session began with two 20-minute think-aloud tasks, one with each prototype (presented in random order). In each task, participants had 15 minutes using the prototype to add music to

Participant	Content Type	Content Platform(s)	Music Platform(s)
P1	DIY Crafts	TikTok, Instagram	Instagram
P2	How-To	YouTube	Epidemic Sound
P3	Vlog	YouTube	Epidemic Sound
P4	Vlog	YouTube	Adobe Stock
P5	Get Ready With Me	TikTok, Instagram	CapCut stock music
P6	Vlog	Instagram	Adobe Stock
P7	How-To	YouTube/Internal Company Videos	YouTube

Table 2: Information about our user study participants, including types of videos they make, the platforms they post their videos on, and the platforms they use for music.

one of their submitted videos, followed by 5 minutes to complete a post-task survey. The post-task survey comprised twenty questions on a seven-point Likert Scale. We designed the survey by adapting relevant elements from Cherry & Latulipe’s Creativity Support Index (CSI) [16], Amershi et al.’s Guidelines for Human-AI Interaction [7], and similar post-task surveys from related work [28, 58]. See the Supplementary Materials for the full list of survey questions.

After completing both think-aloud tasks, the session concluded with a 40-minute semi-structured interview, which complemented the survey by providing deeper insights into participants’ experiences. We asked about their overall impressions of the systems, their preferences, and how the tools compared to their current workflows. Observational notes were also revisited during the interview.

5.4 Data Analysis

While the small sample size makes statistical testing inappropriate, subtle variations emerged across specific survey questions, which warranted further investigation to understand participants’ nuanced perceptions. To investigate these trends, we used exploratory data visualizations including divergent stacked bar charts and slope graphs. To derive qualitative insights from the interview data, we conducted a thematic analysis of the transcripts using an inductive coding approach.

5.5 Results

While both prototypes were well-received, participants had conflicting opinions on which system offered better creative control. The main takeaways from our user study are as follows:

- The design probe was helpful for getting started and exploring possibilities, mainly due to its vibe recommendations and multiple output suggestions.
- The baseline prototype was most helpful when participants had a specific musical idea in mind, and was preferred by participants with musical expertise.
- Participants disagreed on which prototype gave them more agency and control; some found the design probe’s structured prompt easier to control, while others found it overly constraining and preferred the baseline’s direct prompting approach.

5.5.1 Vibe recommendations and structured prompts enable productive exploration. All participants found the vibe recommendations in the design probe to be useful (Figure 6), describing the system as a great starting point for the music process that helped them generate ideas and explore possibilities (P2, P4, P7). Several participants likened the probe to a conversational assistant, offering a helping hand when needed (P2, P4, P5, P7). One participant highlighted the vibe recommendation as an excellent way to establish “a solid enough playground to mess around with things” (P5). This participant also found the suggested prompts helpful: “It was good to have a bit of training wheels, I think, because I don’t know the words to necessarily use when it comes to describing instruments, for example” (P5). Overall, exploration emerged as one of the probe’s key strengths (Figures 7 and 8): “having the three [songs] laid out is beneficial if you’re trying to explore totally different themes” (P1). Participants pointed out that the three music suggestions allowed them to consider genres they might not have explored otherwise (P2, P3, P5) while making it easy to track different iterations (P6, P7).

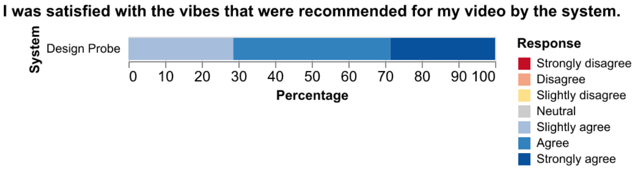


Figure 6: Participants’ responses when asked about the recommended vibe for their videos

5.5.2 System preference depends on user expertise and situation. Four out of seven participants preferred the design probe for their workflows (P2, P3, P4, P7), citing its flexibility and ability to suggest ideas they might not have considered. Two participants enjoyed using both systems but noted that the baseline was especially useful in scenarios where they had a specific musical idea in mind and needed a more direct approach to realize it (P1, P5). P6, who has more musical expertise than the other participants, was critical of both systems due to the inability to define musical attributes directly. Despite this, they acknowledged being impressed by the design probe’s capacity to accurately recommend music that matched their intended vibe.

Participants emphasized that their choice of system would likely depend on situational factors (P1, P3, P5, P6, P7). For example, “if [their goal] was exploration, the [probe] is more ideal” (P1). In contrast, the baseline would be preferable in situations where they already have “a very clear idea of the type of music” (P3) in mind

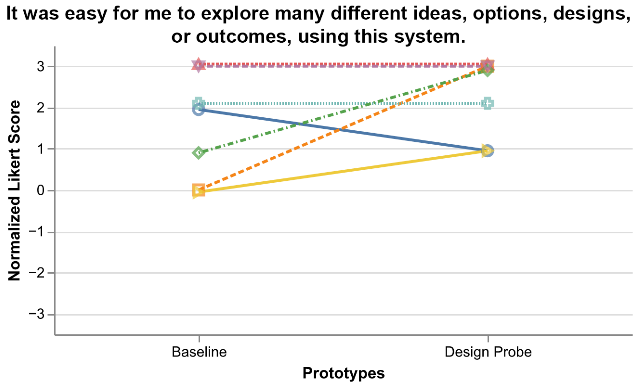


Figure 7: Participants' responses when asked about ease of exploring ideas using each prototype

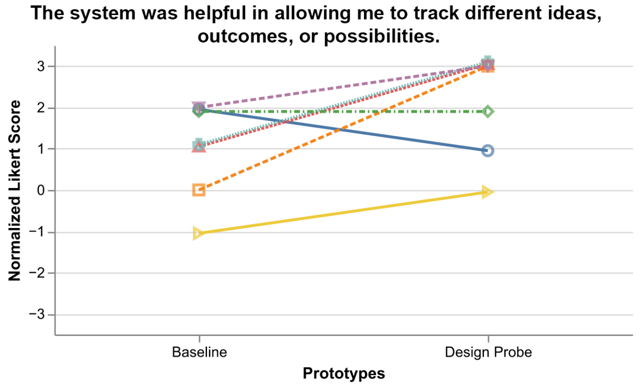


Figure 8: Participants' responses when asked about their ability to track ideas throughout their use of each system

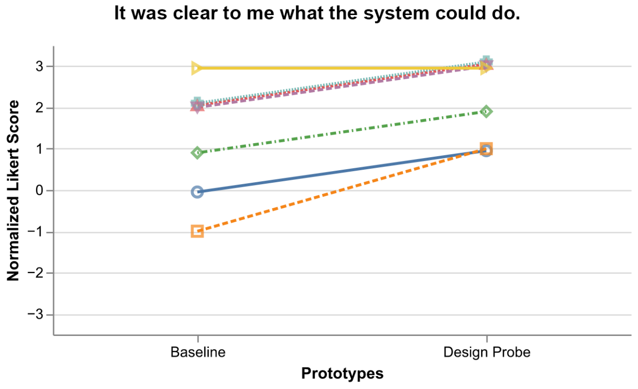


Figure 9: Participants' responses when asked about system clarity

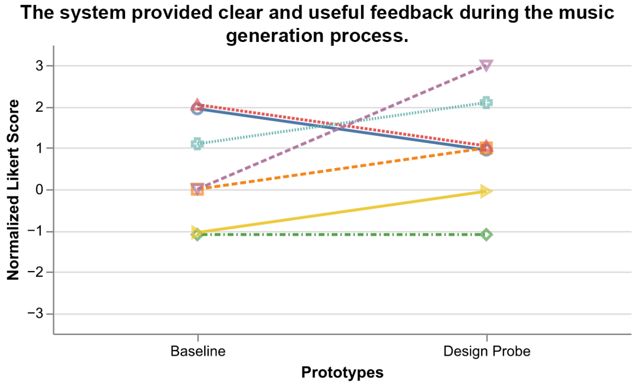


Figure 10: Participants' responses when asked about feedback provided by both systems

or even "already [have] a track to edit" (P1). However, the ability to describe a specific musical direction using the baseline "would be limited by how musical the user is" (P7). Accordingly, one of the most commonly requested improvements was for the system to "progressively [disclose] more and more controls" (P4), and to allow users to move between different levels of abstraction depending on their use case. P7 echoed this, stating that "it would be neat if there could be an open mode [i.e., the baseline] and guided mode [i.e., the design probe]", so they could choose when they want to use the vibe recommendations versus directly prompt the model.

5.5.3 Participants disagree on which system provides more control. Participants had varied experiences when it came to control, transparency, and creativity. While the survey results suggested that the design probe's feature set was clearer (Figure 9), there was some disagreement regarding the feedback provided by both systems (Figure 10).

Some participants found the design probe to be more transparent than the baseline prototype, as it provided insight into how the system was generating music recommendations (P2, P3, P4, P7) and gave them more output parameters to tinker with (P2, P4, P7). "I felt like I was co-creating because it got my creativity flowing more and I could see what [the system] was thinking [...] It's giving me a bigger look inside the black box of AI" (P7). These participants felt that the baseline system required a significant amount of trial and error to achieve desired results (P2, P3, P7).

Conversely, the other participants felt *less* agency with the design probe, saying that its structured interface limited their ability to refine or control music output (P1, P6). These participants found the baseline prototype easier to guide toward their desired result due to its focus on a single-song output and resemblance to familiar LLM interfaces and music search interfaces (P1, P5).

Participants also shared ideas for future features, some that would increase the amount of automated assistance, and others that would provide more fine-grained control. Interestingly, these ideas did not align with which prototype they preferred. For example, both P1 and P3 wanted the system to automatically synchronize the video with their music (increasing assistance), while both P5 and P6 wanted more precise control over the musical structure, such as specifying where a beat should drop (increasing control).

6 Discussion and Takeaways

Through our formative and user studies, we gained a deeper understanding of how video creators conceptualize music and explored how a creative assistant can facilitate music generation for videos. Below, we present three key takeaways that can serve as actionable recommendations for designers of future music generation systems.

6.1 Offer Initial Guidance Based on User Content

A significant challenge for video creators is describing the music they envision in words. Our findings suggest that offering *personalized recommendations* based on the user's content at the start of the music generation process can be a powerful way to address this problem. Personalized recommendations are more effective than predefined examples, as they directly align with the user's unique context, enabling them to begin their creative process with guidance rather than guesswork.

In addition to being personalized, initial guidance should align with the user's mental model of a task, so that they can easily modify it to suit their intent. Our approach of describing the vibe of a video was one way to achieve this, but future systems could explore other user-friendly representations such as the timeline emotion tagging and mood boards we explored in our design process. Alternatively, systems could directly generate music given a video by leveraging video-to-music generation models [23, 53], but would need to provide mechanisms for users to iterate on their outputs and guide them towards their creative goals.

6.2 Surface Intermediate Structure for Productive Iteration

Prompting generative AI models is challenging [66], especially for users with limited expertise in the model's domain. In the context of text-to-music generation, these challenges are exacerbated by the inherently abstract nature of describing music and the subjectivity of music interpretation. Our findings suggest that surfacing *intermediate structures* can help users iteratively refine their output.

Under the hood, a text-to-music model implicitly interprets and transforms the text prompt in order to arrive at a coherent musical output. In our baseline prototype, the model's interpretation of prompts remained opaque, and participants struggled to infer how their input influenced the output when refining their prompts. While this mismatch between user conceptualization and model interpretation was partially due to users' limited musical literacy, it also stems from the complexity of mapping subjective descriptions to structured musical outputs. Our design probe introduced a structured intermediate representation that allowed users to iterate on their input more effectively. By making the system's interpretation more explicit, this approach provided users with concrete axes of control, supporting more productive exploration and refinement.

While we explored one potential representation for intermediate structure, many other representations are possible. For example, Music Creation by Example [28] employs a grid-like interface that allows users to mix and match five attributes across five generated outputs. Future work should consider other interpretable intermediate structures that provide clear opportunities for refinement and empower users to engage more effectively with generative music tools.

6.3 Allow Users to Navigate Between Levels of Control

Our study revealed conflicting opinions among participants regarding which prototype provided better control. Even within the same session, participants' preferences shifted depending on their task or creative intent. This finding highlights the importance of offering users multiple *levels of control* rather than a one-size-fits-all interface solution.

While some participants found the design probe easier to control because of its structured refinement interface, others felt the predefined categories were overly constraining. Similarly, some participants found the baseline's open-ended text box familiar and straightforward, while others found it daunting and unproductive without clear feedback. Future systems must account for these varying needs, offering intuitive ways to move between structured and unstructured modes of interaction. Systems could allow users to toggle between high-level guidance (e.g., vibe-based recommendations) and detailed customization (e.g., control over tempo, instrumentation, or structure) to seamlessly navigate between levels of abstraction as their needs evolve during the creative process. Such systems could also provide educational scaffolding to help address the gap in musical literacy observed among non-expert users, such as tooltips or interactive explanations of musical attributes. This dual role—both as a creative tool and a learning platform—could empower users to make more informed decisions about music while building their confidence in interacting with generative models.

7 Limitations and Future Work

Our study has several limitations that point to opportunities for future research. First, the small number of participants and the exploratory nature of our design probe limit the generalizability of our findings. Rather than conducting a full evaluation, we aimed to observe how video creators interacted with our creative assistant to uncover initial insights into this design space. While our exploratory visualizations of the survey data suggest some potential trends, larger studies with a more diverse participant pool are needed to validate the effectiveness of our proposed workflow. Second, our study only tested one music generation model, and some participants had concerns about the quality of the generated music. While other models may vary in their ability to interpret natural language prompts and quality of their outputs, music generation technology is rapidly improving in general. Future work could compare different models to better understand their relative strengths and weaknesses. Finally, we did not directly compare our creative assistant to participants' existing music search workflows, nor explore how the two might work together. Future work should consider when one approach might be preferred over the other, or even how the two could be blended to combine their strengths—for example, music "outpainting" technology [41] could be used to generatively extend a short stock song to match the duration of a user's video.

More broadly, our design probe represents one point in a broad and largely unexplored design space. We therefore highlight some of the choices we made to emphasize opportunities for future work. First, our prototype requires the user to initiate music generation by selecting a region on the timeline; future versions could enable a mixed-initiative experience by proactively suggesting regions to add music to based on the video's structure. Second, our prototype only provides song-level control over music generation, though some participants wanted to specify how the dynamics or mood of a song should change over time. Future work could explore how we might leverage model improvements that enable finer-grained temporal control (e.g., dynamics or musical structure [40, 41, 63]) by automatically suggesting temporal parameters based on the user's video or providing intuitive mechanisms for users to specify such parameters (e.g., sketching to specify musical structure). Third, our current method of reusing the same seed for a modified prompt does not guarantee that the re-generated result will sound similar, especially if more than a few words in the prompt are changed, because the seed only controls the start of the generative process. Future systems could use the previously-generated song as additional input to guide re-generation, and leverage model optimization techniques to provide specific editing controls such as "inpainting" (replacing a selected region of a song while leaving the rest unchanged) [40, 41] and adding or removing instruments [68]. Finally, future work could also explore how AI music tools might apply to other types of videos, beyond our focus on narrative-style content.

7.1 Ethics and Social Impact

While generative AI tools have the potential to democratize creativity, many creative professionals are also concerned about the potential of such tools to disrupt their livelihoods [57]. Our design process and choice to focus on background music for social videos was informed by the desire to enable more creators to do what they otherwise could not: create and customize unique music for their own videos. Our hope is that by reshaping the creative process in this way, our approach might diversify the landscape of music in narrative videos, as creators may be less likely to reuse the same stock songs in multiple videos.

However, the diversity of a model's generated music depends on its training data, which often under-represent non-Western music [59]. In addition, many AI music generation models have come under recent scrutiny for training on copyrighted music without permission [12, 32], thus producing music that directly competes with the data it was trained on. While training models only on licensed music can help ensure that creators have agreed to the use of their music as training data, questions remain around how this broader shift toward generated content might impact music creators, and how to ensure that generated music represents a broad range of styles and cultures. By addressing these questions, we can build AI tools that empower creators while fostering equitable and sustainable artistic communities.

8 Conclusion

This paper explored the challenges and opportunities of incorporating AI-generated music into video editing workflows for narrative social video creators. Insights from a formative study informed the design of a vibe-based creative assistant that balances automated guidance with user control. Using this creative assistant as a design probe, we conducted a user study that compared it against open-ended text prompting. Our results showed that the creative assistant helped spark ideas and enabled exploration. However, it also highlighted the need for creators with specific musical goals to have more direct control. These findings emphasize the importance of flexibility in the design of text-to-music interfaces for video creators, whose needs of exploration, control, and creativity can change depending on expertise and video content. As text-to-music models continue to evolve, our work underscores the need for a design space for music generation interfaces for video workflows that meet the needs of video creators.

Acknowledgments

We would like to thank Nick Bryan, Ding Li, Adolfo Hernandez Santisteban, Kim Pimmel, and our participants for their contributions to this work.

References

[1] Adobe. 2024. Seeds - Adobe Firefly API. https://developer.adobe.com/firefly-services/docs/firefly-api/guides/concepts/seeds/ Accessed: 2025-1-17.

[2] Adobe. 2025. Advanced Search Features. https://stock.adobe.com/search-features Accessed: 2025-1-10.

[3] 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

[4] Riffusion AI. [n. d.]. *Riffusion*. https://www.riffusion.com/

[5] Stability AI. 2023. *Stable Audio*. Stability AI. https://stability.ai/stable-audio

[6] Stability AI. 2025. *Stable Audio - User Guide*. https://stableaudio.com/user-guide/prompt-structure Accessed: 2025-01-19.

[7] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. 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, 1–13. https://doi.org/10.1145/3290605.3300233

[8] Artlist.io. 2016. Artlist.io. https://artlist.io Accessed: 2025-1-19.

[9] Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang, and Alexander Rush. 2022. PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Valerio Basile, Zornitsa Kozareva, and Sanja Stajner (Eds.). Association for Computational Linguistics, Dublin, Ireland, 93–104. https://doi.org/10.18653/v1/2022.acl-demo.9

[10] Syed Balkhi. 2023. 2025's Creator Economy Statistics That Will Blow You Away. https://www.wpbeginner.com/research/creator-economy-statistics-that-will-blow-you-away/ Accessed: 2025-1-19.

[11] V Braun and V Clarke. 2006. Using thematic analysis in psychology. *Qual. Res. Psychol.* (2006).

[12] Blake Brittain. 2024. Music labels sue AI companies Suno, Udio for US copyright infringement. *Reuters* (June 2024). https://www.reuters.com/technology/artificial-intelligence/music-labels-sue-ai-companies-suno-udio-us-copyright-infringement-2024-06-24/

[13] Cambridge Dictionary. 2025. Vibe Definition. https://dictionary.cambridge.org/dictionary/english/vibe

[14] Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahman, and Smaranda Muresan. 2024. Creativity Support in the Age of Large Language Models: An Empirical Study Involving Emerging Writers. arXiv:2309.12570 [cs.HC] https://arxiv.org/abs/2309.12570

[15] Ke Chen, Yusong Wu, Haohe Liu, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2024. MusicLDM: Enhancing Novelty in text-to-music Generation Using Beat-Synchronous mixup Strategies. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1206–1210. https://doi.org/10.1109/ICASSP48485.2024.10447265

[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 (June 2014), 25 pages. https://doi.org/10.1145/2617588

[17] Sanjoy Chowdhury, Sayan Nag, K J Joseph, Balaji Vasan Srinivasan, and Dinesh Manocha. 2024. MeLFusion: Synthesizing Music from Image and Language Cues using Diffusion Models. arXiv:2406.04673 [cs.CV] https://arxiv.org/abs/2406.04673

[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. https://doi.org/10.1145/3491102.3501819

[19] Miguel Civit, Javier Civit-Masot, Francisco Cuadrado, and Maria J. Escalona. 2022. A systematic review of artificial intelligence-based music generation: Scope, applications, and future trends. *Expert Systems with Applications* 209 (2022), 118190. https://doi.org/10.1016/j.eswa.2022.118190

[20] 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

[21] Jenny Crawford. 2024. 2023 TikTok data report: 85% of videos on TikTok contain music. https://pex.com/blog/2023-tiktok-data-report-85-of-videos-on-tiktok-contain-music/ Accessed: 2025-1-19.

[22] Hai Dang, Lukas Mecke, Florian Lehmann, Sven Goller, and Daniel Buschek. 2022. How to Prompt? Opportunities and Challenges of Zero- and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models. arXiv:2209.01390 [cs.HC] https://arxiv.org/abs/2209.01390

[23] Shangzhe Di, Zeren Jiang, Si Liu, Zhaokai Wang, Leyan Zhu, Zexin He, Hongming Liu, and Shuicheng Yan. 2021. Video Background Music Generation with Controllable Music Transformer. In *Proceedings of the 29th ACM International Conference on Multimedia* (Virtual Event, China) *(MM '21)*. Association for Computing Machinery, New York, NY, USA, 2037–2045. https://doi.org/10.1145/3474085.3475195

[24] 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

[25] Zhikang Dong, Bin Chen, Xiulong Liu, Pawel Polak, and Peng Zhang. 2024. MuseChat: A Conversational Music Recommendation System for Videos. arXiv:2310.06282 [cs.LG] https://arxiv.org/abs/2310.06282

[26] Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. 2024. Long-form music generation with latent diffusion. arXiv:2404.10301 [cs.SD] https://arxiv.org/abs/2404.10301

[27] Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. 2024. Stable Audio Open. arXiv:2407.14358 [cs.SD] https://arxiv.org/abs/2407.14358

[28] Emma Frid, Celso Gomes, and Zeyu Jin. 2020. Music Creation by Example. 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. https://doi.org/10.1145/3313831.3376514

[29] Josh Howarth. 2022. 30+ Incredible Creator Economy Statistics (2024). https://explodingtopics.com/blog/creator-economy-stats Accessed: 2025-1-19.

[30] Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel P. W. Ellis. 2022. MuLan: A Joint Embedding of Music Audio and Natural Language. arXiv:2208.12415 [eess.AS] https://arxiv.org/abs/2208.12415

[31] Google Labs. 2023. *MusicFX*. https://labs.google/fx/tools/music-fx

[32] Elias Leight. 2024. Google Trained Its AI on Copyrighted Music, Sources Say — Now It's Trying to Make Deals. http://www.billboard.com/pro/google-youtube-trained-ai-copyrighted-music-before-deals Accessed: 2025-1-19.

[33] Jen-Chun Lin, Wen-Li Wei, James Yang, Hsin-Min Wang, and Hong-Yuan Mark Liao. 2017. Automatic Music Video Generation Based on Simultaneous Soundtrack Recommendation and Video Editing. In *Proceedings of the 25th ACM International Conference on Multimedia* (Mountain View, California, USA) *(MM '17)*. Association for Computing Machinery, New York, NY, USA, 519–527. https://doi.org/10.1145/3123266.3123399

[34] Vivian Liu and Lydia B Chilton. 2022. Design Guidelines for Prompt Engineering Text-to-Image Generative 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 384, 23 pages. https://doi.org/10.1145/3491102.3501825

[35] 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. https://doi.org/10.1145/3313831.3376739

[36] 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. https://doi.org/10.1145/3490099.3511159

[37] Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas. 2022. Contrastive Audio-Language Learning for Music. arXiv:2208.12208 [cs.SD] https://arxiv.org/abs/2208.12208

[38] Masahiro Matsuo, Fumi Masuda, Yukiyoshi Sumi, Masahiro Takahashi, Atsushi Yoshimura, Naoto Yamada, and Hiroshi Kadotani. 2019. Background music dependent reduction of aversive perception and its relation to P3 amplitude reduction and increased heart rate. *Front. Hum. Neurosci.* 13 (June 2019), 184. https://pubmed.ncbi.nlm.nih.gov/31316359/

[39] Daniel McKee, Justin Salamon, Josef Sivic, and Bryan Russell. 2023. Language-Guided Music Recommendation for Video via Prompt Analogies. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14784–14793. https://doi.org/10.1109/CVPR52729.2023.01420

[40] Zachary Novack, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas Bryan. 2024. DITTO-2: Distilled Diffusion Inference-Time T-Optimization for Music Generation. arXiv:2405.20289 [cs.SD] https://arxiv.org/abs/2405.20289

[41] Zachary Novack, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J. Bryan. 2024. DITTO: Diffusion Inference-Time T-Optimization for Music Generation. arXiv:2401.12179 [cs.SD] https://arxiv.org/abs/2401.12179

[42] Zachary Novack, Ge Zhu, Jonah Casebeer, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J. Bryan. 2024. Presto! Distilling Steps and Layers for Accelerating Music Generation. arXiv:2410.05167 [cs.SD] https://arxiv.org/abs/2410.05167

[43] OpenAI. 2021. CLIP: Connecting text and images. https://openai.com/index/clip/ Accessed: 2025-1-19.

[44] Jeongeun Park, Hyorim Shin, Changhoon Oh, and Ha Young Kim. 2024. "Is Text-Based Music Search Enough to Satisfy Your Needs?" A New Way to Discover Music with Images. 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 504, 21 pages. https://doi.org/10.1145/3613904.3642126

[45] Laure Pretet, Gael Richard, and Geoffroy Peeters. 2021. Cross-Modal Music-Video Recommendation: A Study of Design Choices. arXiv:2104.14799 [cs.MM] https://arxiv.org/abs/2104.14799

[46] Laure Prétet, Gaël Richard, Clément Souchier, and Geoffroy Peeters. 2023. Video-to-Music Recommendation Using Temporal Alignment of Segments. *IEEE Transactions on Multimedia* 25 (2023), 2898–2911. https://doi.org/10.1109/TMM.2022.3152598

[47] Microsoft Research. 2023. LLaVA: Large Language and Vision Assistant. https://www.microsoft.com/en-us/research/project/llava-large-language-and-vision-assistant/ Accessed: 2025-1-19.

[48] Stephen Robles. 2024. How Much Do YouTubers Make? (Earnings and Examples). https://riverside.fm/blog/how-much-do-youtubers-make Accessed: 2025-1-19.

[49] 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

[50] Epidemic Sound. 2009. Epidemic Sound. https://www.epidemicsound.com Accessed: 2025-1-19.

[51] Speechmatics. 2023. Speechmatics AI Speech Technology. https://www.speechmatics.com/ Accessed: 2025-1-19.

[52] Adobe Stock. 2020. Adobe Royalty Free Stock Music and Audio. https://stock.adobe.com/audio Accessed: 2025-1-19.

[53] Kun Su, Judith Yue Li, Qingqing Huang, Dima Kuzmin, Joonseok Lee, Chris Donahue, Fei Sha, Aren Jansen, Yu Wang, Mauro Verzetti, and Timo I. Denk. 2024. V2Meow: Meowing to the Visual Beat via Video-to-Music Generation. arXiv:2305.06594 [cs.SD] https://arxiv.org/abs/2305.06594

[54] Suno. 2023. *Suno*. https://suno.com

[55] Dídac Surís, Carl Vondrick, Bryan Russell, and Justin Salamon. 2022. It's Time for Artistic Correspondence in Music and Video. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10554–10564. https://doi.org/10.1109/CVPR52688.2022.01031

[56] Udio. 2024. *Udio*. https://www.udio.com/

[57] Veera Vimpari, Annakaisa Kultima, Perttu Hämäläinen, and Christian Guckelsberger. 2023. "An Adapt-or-Die Type of Situation": Perception, Adoption, and Use of Text-to-Image-Generation AI by Game Industry Professionals. *Proc. ACM Hum.-Comput. Interact.* 7, CHI PLAY, Article 379 (Oct. 2023), 34 pages. https://doi.org/10.1145/3611025

[58] 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. https://doi.org/10.1145/3640543.3645143

[59] Qinyuan Wang, Bruce Gu, He Zhang, and Yunfeng Li. 2024. Fairness of Large Music Models: From a Culturally Diverse Perspective. In *2024 IEEE 9th International Conference on Data Science in Cyberspace (DSC)*. 706–712. https://doi.org/10.1109/DSC63484.2024.00105

[60] Sitong Wang, Zheng Ning, Anh Truong, Mira Dontcheva, Dingzeyu Li, and Lydia B Chilton. 2024. PodReels: Human-AI Co-Creation of Video Podcast Teasers. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 958–974. https://doi.org/10.1145/3643834.3661591

[61] David Winter. 2024. Amplifying Stories: The Vital Role of Audio in Video Narratives. https://lwks.com/blog/amplifying-stories-the-vital-role-of-audio-in-video-narratives Accessed: 2025-1-19.

[62] Minz Won, Justin Salamon, Nicholas J. Bryan, Gautham J. Mysore, and Xavier Serra. 2021. Emotion Embedding Spaces for Matching Music to Stories. arXiv:2111.13468 [cs.IR] https://arxiv.org/abs/2111.13468

[63] 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. https://doi.org/10.1109/TASLP.2024.3399026

[64] 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. https://doi.org/10.1145/3491102.3517582

[65] Jing Yi, Yaochen Zhu, Jiayi Xie, and Zhenzhong Chen. 2023. Cross-Modal Variational Auto-Encoder for Content-Based Micro-Video Background Music Recommendation. *IEEE Transactions on Multimedia* 25 (2023), 515–528. https://doi.org/10.1109/TMM.2021.3128254

[66] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. 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 437, 21 pages. https://doi.org/10.1145/3544548.3581388

[67] Yongyi Zang and Yixiao Zhang. 2024. The Interpretation Gap in Text-to-Music Generation Models. arXiv:2407.10328 [cs.SD] https://arxiv.org/abs/2407.10328

[68] Yixiao Zhang, Yukara Ikemiya, Woosung Choi, Naoki Murata, Marco A. Martínez-Ramírez, Liwei Lin, Gus Xia, Wei-Hsiang Liao, Yuki Mitsufuji, and Simon Dixon. 2024. Instruct-MusicGen: Unlocking Text-to-Music Editing for Music Language Models via Instruction Tuning. arXiv:2405.18386 [cs.SD] https://arxiv.org/abs/2405.18386

[69] Yixiao Zhang, Akira Maezawa, Gus Xia, Kazuhiko Yamamoto, and Simon Dixon. 2024. Loop Copilot: Conducting AI Ensembles for Music Generation and Iterative Editing. arXiv:2310.12404 [cs.SD] https://arxiv.org/abs/2310.12404

[70] Yixiao Zhang, Gus Xia, Mark Levy, and Simon Dixon. 2021. COSMIC: A Conversational Interface for Human-AI Music Co-Creation. In *NIME 2021*. https://nime.pubpub.org/pub/in6wsc9t

A LLM Prompts

This section contains the prompts used by our creative assistant prototype for generating vibe recommendations and structured music prompts. We used the `1106-preview` version of GPT-4-Turbo.

A.1 Vibe Recommendation

```
Describe the overall vibe/emotion of a specific section of
my video. Give me five adjectives that describe it, and one
short phrase describing how a viewer should feel while
viewing the video. Here is an example for a travel vlog:
exciting, energetic, adventurous, happy, curious, like a
really fun concert. DO NOT INCLUDE ANY PREAMBLE. Just give
me the adjectives and the phrase in a single comma
separated sentence.

Here is the transcript for the specified section of the
video:

"${filteredTranscriptSentences}"

For context, here is the full transcript from the whole
video:

"${transcriptSentences}"

Here is an array describing the visual features of the
video in JSON format. Each item in the array represents one
salient visual scene in the video, and describes the visual
attributes of that scene.
This array is of the visual scenes from the specified
section in the video:

${filteredVisualScenes}

For context, here is the full captions for the whole video:

${visualScenes}
```

Where:

- `transcriptSentences` = a string containing the entire transcript of the video with punctuation
- `filteredTranscriptSentences` = a string containing only sentences in the transcript that occur within the region of the video selected by the user
- `visualScenes` = an array of strings, where each string represents one visual scene and contains a few sentences describing the visual attributes of that scene
- `filteredVisualScenes` = an array of strings, containing only the scenes that occur within the region of the video selected by the user

A.2 Translating Vibe into Structured Music Prompts

```
Come up with three prompts for a music generation model for
music that will go with a specified section of the video
where the generated music will play. Here is a summary of
the mood/vibe of the video:

"${vibeRecommendation}"

The model accepts prompts that use the following
properties: genre (1-2 genres), mood, energy, instruments
(1-2 instruments), and theme. Each prompt should produce a
different kind of music, but they should all still relate
to the mood/vibe of the video.

For "theme", you can use words like ${themes}.

Use language that is easy to understand and do not
hallucinate. Don't include words such as 'Sure! Here's a
short description for the music' or 'Generate a song
that...'.

Write the prompts so that they match the following template
in JSON format:

{
  "prompts": [
    {
      "genre": "1-2 genre descriptors",
      "instruments": "1-2 instrument descriptors",
      "mood": "1-2 words describing the mood of the video",
      "theme": "1-2 words describing the video theme",
      "energy": "energy level of the video"
    },
    {
      ...
    },
    {
      ...
    }
  ]
}

The output must be able to be parsed as JSON. Do not start
the response with "{". Start your response with {

To help with creating the prompt, use the following
transcript and visual captions from the video. Here is the
transcript for the specified video section:

"${filteredTranscriptSentences}"

For context, here is the full transcript from the whole
video:

"${transcriptSentences}"

Here is an array describing the visual features of the
video in JSON format. Each item in the array represents one
salient visual scene in the video, and describes the visual
attributes of that scene.

This array is of the visual scenes from the specified
section in the video:

${filteredVisualScenes}

For context, here is the full captions for the whole video:

${visualScenes}
```

Where:

- `vibeRecommendation` is the output from A.1, potentially modified/rewritten by the user
- `themes` is the following hardcoded list of themes, which was used in the training of our text-to-music model but is based more generally on common themes that appear in stock music for video (and could be omitted or replaced by any list of example themes): `Business/Corporate, Vlog, Commercial, Trailer, Weddings, Documentary, Food, Education, Lifestyle, Road Trip, Travel, Gaming, Sport & Fitness, Slow Motion, Urban, Party, Fashion, Science, Medical, Industry, Drone Shots, Landscape, Nature, Podcast, Tropical, Love Song, Action Adventure`
- `transcriptSentences`, `filteredTranscriptSentences`, `visualScenes`, `filteredVisualScenes` are the same as in A.1 above