VidTune: Creating Video Soundtracks with Generative Music and Video-Based Thumbnails

Mina Huh
University of California, Berkeley
Berkeley, California, USA
minahuh@berkeley.edu

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

Dingzeyu Li
Adobe Research
Seattle, Washington, USA
ding@dingzeyu.li

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

Bryan Wang
Adobe Research
Seattle, Washington, USA
bryanw@adobe.com

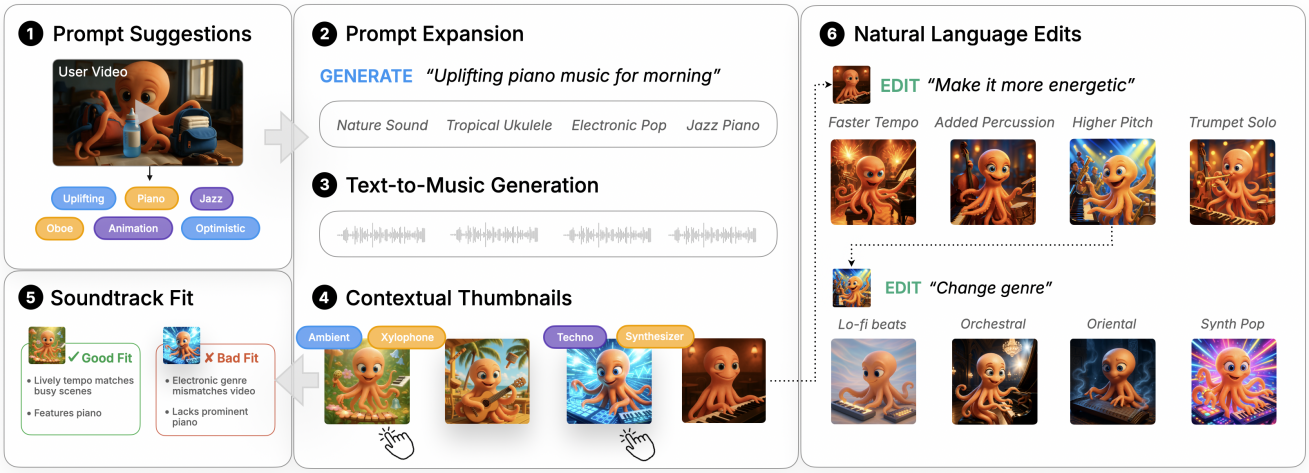

Figure 1: VidTune is an interactive system that helps video creators generate soundtracks. VidTune provides prompt suggestions based on the input video (1) and expands the user's prompt (2) to generate a diverse set of candidate music tracks (3). Each track is presented with *contextual thumbnails* (4) for efficient music preview in-context, and users can hover over each track to see reusable keywords and a *soundtrack fit check* (5). Creators can iteratively refine tracks with language-guided edits (6), which VidTune turns into updated candidate tracks.

Abstract

Music shapes the tone of videos, yet creators find it hard to find soundtracks that match their video's mood and narrative. Recent text-to-music models let creators generate music from text prompts, but our formative study (N=8) shows creators struggle to construct diverse prompts, quickly review and compare tracks, and understand their impact on the video. We present VidTune, a system that supports soundtrack creation by generating diverse music options from a creator's prompt and producing contextual thumbnails for rapid review. VidTune extracts representative video subjects to ground thumbnails in context, maps each track's valence and energy onto visual cues like color and brightness, and depicts prominent genres and instruments. Creators can refine tracks with natural language edits, which VidTune expands into new generations. In a controlled user study (N=12) and an exploratory case study (N=6), participants found VidTune helpful for efficiently reviewing and comparing music options and described the process as playful and enriching.

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**; • **Computing methodologies** → **Artificial intelligence**.

Keywords

Text-to-Music Generation, Soundtrack Creation, Video Tools

ACM Reference Format:
Mina Huh, C. Ailie Fraser, Dingzeyu Li, Mira Dontcheva, and Bryan Wang. 2026. VidTune: Creating Video Soundtracks with Generative Music and Video-Based Thumbnails. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/3772318.3791572

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791572

1 Introduction

Generative music is increasingly used in video production [41, 77], letting creators generate custom soundtracks from natural language prompts. However, current text-to-music workflows often struggle to meet the demands of creating video soundtracks. Creators may lack a clear sense of what music would fit their video, and even when they do, they often struggle to articulate these intentions as effective prompts [41, 102]. Our formative study with 8 video creators surfaced these same breakdowns in practice, with participants relying on only a few familiar descriptors. Moreover, as these models rapidly produce many alternatives, reviewing, organizing, and aligning tracks with the video's structure became a major bottleneck [35]. The temporal nature of audio forced them to listen through each track to compare and recall favorites, making it hard to skim and triage options. These issues were even more pronounced for deaf and hard-of-hearing (DHH) creators, who lacked accessible cues to explore and validate the music.

One way to mitigate these challenges is to represent audio in visual forms that reveal similarities and differences at a glance, in ways that are difficult to achieve through listening alone. Motivated by this, we present VidTune, a text-to-music generation tool that uses *contextual thumbnails* to visually summarize generative music outputs in the context of the user's video. In VidTune, each thumbnail takes the form of both a static image and a short animated video generated from it. These thumbnails map music attributes such as mood, energy, and instrumentation onto visual style and composition with subjects anchored in the user's video, giving each track a concise and meaningful visual summary. This explicit audio-visual mapping helps address gaps in current tools, which rely on vague titles, generic cover art, and waveforms that may not be intuitive for many creators, including non-music experts and those with limited hearing.

Grounded in prior work and our formative insights, we designed VidTune around four capabilities: helping creators **explore** by expanding user prompts into musically diverse alternatives, **review** with contextual thumbnails that visually distinguish different tracks, **refine** by steering generations with natural language edits, and **manage** large sets of outputs via an overview of the music space that reveals explored directions and potential next steps.

Our technical evaluation shows that VidTune's prompt expansion algorithm can increase musical diversity and that its thumbnails more reliably reflect the generated music than a baseline. We also evaluated VidTune in a within-subjects study with 12 video creators, who compared VidTune against a baseline interface similar to existing AI music generation tools. Participants described VidTune as more expressive and enjoyable to use, and felt the results were more worth the effort. Finally, an exploratory case study (N=6) using creators' own videos further shows how the thumbnails were perceived in situ and highlights VidTune's potential to make music generation more accessible to a broader set of users.

Taken together, our work makes the following contributions:

- VidTune, a system that enables creators to explore, interpret, and refine generative music for video soundtracks.
- An AI pipeline that generates *contextual thumbnails* by capturing key elements of the input video and the generated music tracks to support *visual sensemaking* of music.
- Empirical findings from a controlled study (N=12) demonstrating VidTune's advantages over a baseline, and an exploratory study (N=6) highlighting emergent uses and broader applications for contextual thumbnails.

2 Related Work

VidTune is informed by prior work on video-based music generation, music visualization, and sensemaking of generative AI output.

2.1 Music Generation for Videos

Recent advances in AI assist creators in composing melodies [30], harmonizing chords [53, 59], and even writing lyrics [30, 86]. Modern text-to-music models [23, 86, 92] also make end-to-end music generation more accessible to novices [63] and creators with hearing impairments [20] by allowing them to describe desired moods or styles in natural language. Other work conditions audio directly on video content: synthesizing *diegetic audio* aligned with on-screen actions [32, 37, 84], and composing *background music* to match narrative emotion [29, 81, 82, 101], or visual rhythm [56, 58, 91]. More recently, commercial text-to-video models (*e.g.,* Sora [73], Veo [27]) can produce both visuals and synchronized audio from a single prompt, effectively generating end-to-end videos with soundtracks. However, these systems keep music generation opaque and non-interactive for creators.

Beyond optimization-centric approaches, recent HCI systems help video creators easily express their music goals by suggesting text prompts from video analysis [41], allowing users to provide example songs as inspiration [34], or iteratively updating music based on a conversational dialogue with the user [107]. These systems highlight the importance of supporting high-level user input, previewing music in the context of the user's video, and balancing automated support with user control. However, they mainly focus on initial prompting and refinement of a few outputs, with less attention to how creators can efficiently understand, compare, and evaluate the large number of options produced in music creation workflows [35, 46]. We address this gap by designing workflows for in-context review and comparison of generative soundtracks.

2.2 Music Visualization

Music is inherently temporal and non-visual, which makes it harder to skim or grasp at a glance compared to images or video. Visualizations can help by externalizing key attributes to augment the listening experience [62], evoke synesthetic effects [54], and improve accessibility [104]. Symbolic and signal-level encodings such as scores and waveforms are commonly used for composition, analysis, and performance [60, 68], and video editing and stock music platforms often rely on waveform displays to help users navigate and edit music. While useful for precise editing, they do not convey perceptual qualities of music such as mood or vibe. In contrast, semantic visualizations emphasize mood, style, or structure to aid listening and exploration [52, 60] by mapping rhythm, timbre, and affect onto colors, motion, scenes, and art styles [14, 26, 49, 62, 90]. Semantic visualizations can also aid music *creation* by generating real-time images to inspire MIDI-based composition [103], or in commercial tools like Suno [86] and Udio [92] by adding thumbnail

images to generated songs. However, these approaches are primarily designed to enrich the experience of creating a single track, rather than to help creators select from multiple generated soundtracks. We investigate what video creators seek when choosing a generative soundtrack and propose semantic visualizations that embed musical attributes into imagery anchored in the user's own video. Building on accessible music visualization work that maps musical mood and structure onto colors and shapes [28, 69, 97], or expressive faces [97], we also explore how VidTune's thumbnails can function as a non-auditory accessibility channel in selecting generative soundtracks.

Even with semantic visual aids, exploring and understanding a large set of music options can be overwhelming, especially as generative models make it possible to generate infinitely many candidates. Most music search and recommendation interfaces show results as ranked lists [15, 83], and music generation platforms similarly display outputs as lists of tracks ordered by recency [86, 92], but a linear layout does not convey relationships or organization across the space of choices [36, 78]. Prior work has used 2D maps to visualize music spaces for exploring large collections of songs [51, 76, 87, 94] and visualizing users' taste [36, 78]. Building on these approaches, we introduce a music map adapted for generative soundtrack selection, where the tracks generated for a given video are represented as thumbnails in a 2D space that enables exploration, comparison, and iteration.

2.3 Sensemaking of Generative AI Outputs in Creative Work

Generative AI enables creators to explore design spaces more efficiently by streamlining creative processes [47, 96] and generating diverse variations, including ones beyond their own capabilities [21]. In creative workflows, prior work shows that seeing alternatives side by side helps people notice differences [47], reflect on trade-offs, and achieve higher-quality outcomes [31]. Prior HCI systems operationalize this idea by expanding users' input prompts to generate more diverse results [10, 13, 42].

With many alternatives, interfaces must help users make sense of them. Recent sensemaking tools organize generations into hierarchies [3, 85] and 2D similarity maps [10], and guide navigation with constraints and filters that help users traverse large collections [50, 66, 67]. Some focus on comparison, supporting users in understanding differences across outputs [9, 47, 48]: for instance, GenAssist describes similarities and differences across generated images [48], and VideoDiff visualizes alternative video edits side by side [47].

These systems focused on visual and textual domains, where alternatives can be quickly skimmed and compared with thumbnails, storyboards, or text summaries [6, 7, 39, 98]. Recent work on generative music has begun to explore constructing diverse prompts and iteratively refining tracks [41, 102], but these interfaces typically present results as waveforms [41] or spectrograms [102] that are difficult for non-experts to interpret and still require sequential listening. VidTune addresses this gap by presenting soundtrack options as contextual thumbnails that encode key musical attributes and providing organization mechanisms to support rapid comparison and navigation without listening to every track.

3 Formative Study

To understand the strategies and challenges of using AI to generate music for videos, we conducted a formative study with 8 video creators. The formative study consisted of semi-structured interviews and a music generation task using participants' own videos. Unlike prior work that surveyed creators [34] or observed existing workflows without text-to-music tools [41], we observed creators scoring their own videos with a state-of-the-art text-to-music model to surface strategies and challenges that arise in actual use.

3.1 Method

Participants. We recruited 8 video creators who add music to their videos on a regular basis (P1-P8, §A Table 6). Participants were recruited via mailing lists and compensated 40 USD for a 1.5-hour remote study conducted via Zoom. Participants had an average of 15 years of video creation experience (SD=7.43) and created a wide variety of videos, including short films, commercials, vlogs, and social-media reels. We intentionally sampled a range of experience levels to capture how both less-experienced and skilled creators approach AI music tools. 3 participants regularly used music generation tools including Suno [86] and Udio [92] (P1, P2, P6), 1 participant occasionally used them (P4), and 4 participants had never used them (P3, P5, P7-P8). We included 3 creators (P5, P7-P8) who were deaf or hard-of-hearing (DHH) as they encounter distinct perceptual challenges in generating music and can inform the design of an inclusive tool.

Procedure. We first conducted a semi-structured interview asking participants how they currently find or generate music for their videos. Then, participants completed a 30-minute task where they used Suno [86] to generate music for their own videos which ranged from 2 to 17 minutes in length. While there are many music generation tools available in the market today, we selected Suno as it is one of the most widely used text-to-music platforms with over 25 million users [89]. For participants who had not used Suno, we provided a short tutorial on how to write prompts, configure generation settings, and export music. Participants could generate unlimited tracks and add multiple pieces to different parts of the video. With DHH creators, we adopted their preferred communication channels (Zoom chat, live captions) and encouraged them to ask questions about each generated music to make decisions, which researchers answered in real time. We analyzed the questions to understand what information DHH creators seek in music generation, thereby guiding what attributes future systems should surface [48].

We transcribed the interviews and participants' comments during the task and grouped their strategies and challenges into 4 stages of the music-generation workflow: 1) writing prompts, 2) reviewing music, 3) iterating on outputs, and 4) organizing them.

3.2 Findings

Current practices are tedious. All participants described music as essential to shaping the story and audience experience. P8 noted *"Most of my audience is hearing and I want them to also enjoy my videos."* Similarly P3 stated *"[Soundtracks] are very important, but with long videos, it's harder to come up with multiple music tracks."*

Table 1: Types of information that our formative study participants considered when evaluating generative music for their video soundtracks. Items were coded from participants' explanations of how they evaluated tracks and, for DHH creators, from the questions they asked about each track.

Category	Criteria	Example Question
Contextual Qualities	Prompt alignment	*Does this music align with my text prompt/desired style?*
	Video alignment	*How well does this music fit the visuals and mood of my video scene?*
Musical Qualities	Genre & Style	*What is the genre or musical style of this track?*
	Mood & Emotion	*What mood or emotion does this music convey?*
	Energy & Tempo	*What is the music's overall energy level and tempo like?*
	Instruments	*What instruments are prominently featured in this music?*
	Structure & Progression	*How does the music's structure and progression unfold over time?*
	Errors & Quality	*Are there generation artifacts or quality issues?*

To source soundtracks for their videos, 5 participants searched stock libraries (P2-4, P6, P8), 2 composed music themselves (P6-P7), and 4 used generative music tools (P1-P2, P4, P6). Among DHH creators, P8 often selected from recommended music on short-form platforms. To review whether the music is right for the video context, she checked lyrics, examined where the track had been used, and read audience comments about how the music felt. The other two DHH creators often delegated soundtrack selection to friends and family (P5, P7).

Despite these strategies, participants found it tedious to search for a track that suited their goals, which made generative music particularly appealing. As P6 explained, *"When I have much time, I'd make my own music using FL Studio, but it's hard now because I create and upload more often. Stock tracks all sound too similar; I don't want to use a song people recognize from other channels."* P1, who frequently uses AI music, highlighted its increasing quality: *"Now people don't know it's AI made. It doesn't sound like AI."*

Coming up with diverse prompts is hard. Creators with musical backgrounds (P1-P2, P4) used more concrete, musically specific prompts (*e.g.,* genre, instrumentation, tempo, progression cues), while those with less musical background (P3, P5-P8) wrote more abstract or usage-oriented prompts tied to the video content and desired audience reaction. 3 participants wanted to tailor style to their audience but were unsure what fits. P6 noted, *"This is a huge challenge for me... I am 46 years old; what I generate might feel outdated to students. What do they listen these days?"* Because starting from a blank prompt was challenging, participants drew on references and recommendations – browsing YouTube examples (P3) and asking ChatGPT for ideas (P2, P6, P8). 2 participants also tried including artist names (P2—*"Miley Cyrus"*; P1—*"Hans Zimmer"*) as a concise way to convey style, but found these blocked by Suno due to copyright restrictions. P4 expressed a desire to branch further: *"I really want to see more different music styles but don't know how to get there."* P1 noted that long prompts or explicit timing cues were often not reliably followed [57] and wanted ways to filter for generations that adhere to requests in the prompt.

These observations echo Hammad et al. [41] on the need for video-grounded prompt suggestions. Further, our findings highlight the importance of helping users expand and explore broader options to discover what fits.

Reviewing generated music is time-consuming. Table 1 summarizes what creators consider when reviewing and selecting soundtracks. Beyond low-level musical qualities, participants evaluated whether a track adhered to the prompt and how well it suited the footage. To evaluate music in context with their video, participants listened to music with a side-by-side still frame of the video (P3-P4, P6-P7) or imported tracks into a video editor to preview (P1-P2). The Suno version used in our study did not allow users to specify duration and produced 2–4-minute tracks. Novices tended to listen longer (*"I need 40–90 seconds to know if I like it,"* – P3), whereas others skipped around to quickly understand the entire song.

Participants emphasized listening beyond a quick skim. P2 explained, *"As songs are usually much longer than my scene, I need to listen through the whole piece to find the section that fits best."* P6 also noted *"I need to verify the music holistically. But humans' instinct is to quickly judge, so that's hard. Similar to how AI-generated images have deformations when you look closely, generated music can be the same but harder to notice as it's audio. [...] Once, I heard female vocals gradually turn into to male later in the track and that was weird."*

With the ability to quickly generate many options, participants described *reviewing* as the bottleneck. The default titles and thumbnails provided by the platform were not informative enough to convey the music or distinguish tracks, so participants resorted to sequential listening. P3 remarked, *"Thumbnails aren't catchy or descriptive at all, and all titles are "Untitled" unless I manually type in. Just showing the prompt isn't enough."* Consequently, participants forgot how previous tracks sounded or which they preferred, often replaying tracks multiple times (P6 repeating *"Was it this one?"*) to find them again. Both P4 and P6 wanted visual snippets to preview music, and P6 added *"I don't want it [future system] to just choose the best music for my video [...] Overly suggesting can harm our creativity. I just want help in reviewing."*

DHH creators encountered additional hurdles in reviewing generated music. Without familiar social or context cues (*e.g.,* artist name, comments, prior uses), it was harder to evaluate generated tracks: *"Does it actually sound like reggae music [as in the prompt]?*

[...] there's no artist or comments, so I cannot really trust this music" (P8). Both P5 and P7 wanted more visual support like colors or shapes to quickly convey emotion and intensity, as dense text descriptions can be slower to read. P5 stated *"We deaf people love visual stuff. [...] I like music visualizers that convey emotion and intensity. But I can also see abstract visuals being more confusing when I'm selecting music not just enjoying it."*

Iterating is necessary but non-trivial. Participants were rarely satisfied with the first pass and went through multiple iterations – either refining the prompt or using Suno's audio edit feature – to align generation with their music goal. P2 wanted more diverse options for edits, explaining *"I wish it would generate more than two, like four or more."* As text-to-music models often fill in unspecified details, participants added constraints to their prompts to steer results (*e.g., "no strings," "remove brass"*). Many participants found it challenging to rewrite full prompts, wanting instead to provide feedback as *critique*: *"This is too sad and slow. How should I change to make it less sad?"* (P2). To narrow down candidates, participants marked tracks they liked and unliked. P6 noted that pointing to examples is easier than articulating how to change: *"Picking what I don't like is easy, but describing why I don't like is hard"*.

Even after selecting a favorite, participants kept generating alternatives. P3 explained, *"I still make backups because once I change the cut [in the video editor] or change the overall vibe, I might need a different option."* Participants tried to steer generation to match the video scene by prompting exact timing or tempo, but found the results did not adhere to the prompt. After edits, they wanted clear before/after comparisons: *"Explicitly state what changed; right now I have to manually check"* (P3).

Organization gets harder as generations pile up. Participants generated an average of 15 (SD = 5.44) tracks each during the study. As sets grew, managing versions, favorites, and use-cases became non-trivial. To avoid missing good music candidates, P6 listened to all generated tracks from the beginning. P5 found her favorites were spread far apart in the list so she kept scrolling back and forth to compare and decide.

As the default instrumental tracks appeared as *"Untitled"*, 3 participants adopted ad-hoc titling for management, but found it inefficient (P2-P3, P7). P7 titled by intended placement in the video (*e.g.,* "podcast intro", "outro") and P3 used brief style-based labels (*e.g.,* "piano blues", "sad harp") but dropped the practice after two trials, finding it challenging to come up with a title every time. They also wanted help selecting based on intent and fit, as P3 described: *"I want it [future system] to watch my video and shortlist only the ones that are a good match. So I can prioritize what to listen to."*

3.3 Design Goals

Based on our observations, we distill 5 core design goals for VidTune to harness the creative flexibility of generative music for videos. These goals address key user tasks in music generation: exploration (D1, D2), evaluation (D3), refinement (D4), and organization (D5).

- **D1.** Provide contextualized music prompt suggestions
- **D2.** Generate diverse but relevant music options
- **D3.** Facilitate efficient music evaluation and comparison
- **D4.** Support iterative refinement through user feedback
- **D5.** Enable flexible management of generated music alternatives

4 VidTune Interface

Guided by the design goals, we developed VidTune (Figure 2), an interactive system that supports video soundtrack generation with text-to-music models. Users can generate diverse music options tailored to their videos (D1, D2), quickly review and compare tracks with contextual thumbnails (D3), and iteratively refine them with *edit*, *vary*, and *blend* interactions (D4). As users generate more tracks, VidTune simplifies management through filtering, searching, and grouping by similarity (D5). Building on Hammad et al. [41], which shows that creators want to validate music in the context of their video, VidTune integrates a video player.

Writing prompts with contextual suggestions. Once the user uploads a video, they enter basic project details: title, video type, target audience, and soundtrack goal. On the left are the video player and filmstrip (Figure 2A-B), where users can select scene segments for which to generate music. Based on this selection, VidTune displays *prompt suggestions* (Figure 2D) relevant to the current scene and the user-provided project details. Suggested keywords span instruments (*e.g.,* guitar, bass), genres (*e.g.,* jazz, indie pop), and vibes (*e.g.,* upbeat, cheerful). Users can click to add suggestions, refresh to see alternatives, or type custom text. VidTune generates 4 candidate music tracks from a single prompt (Figure 2E) by expanding the prompt into four different alternatives and generating one track along with a descriptive title for each.

Reviewing generations with contextual thumbnails. Each generated track is paired with a *contextual thumbnail* that anchors on core subjects from the user's footage (*e.g.,* the octopus character) and conveys genre, mood, tempo, and instrumentation through artistic style, color, and motion effects. Contextual thumbnails let users rapidly skim candidate tracks, spot differences between tracks, and help remember tracks [19]. As the user plays each track, its thumbnail animates into an 8-second loop (Figure 3) making tempo and energy more salient and memorable. Rather than animating all thumbnails at once, we only animate the currently playing track to avoid visual clutter. The static thumbnails still convey tempo and energy through motion cues such as blur and speed lines.

Each thumbnail includes a *fit check*, a subtle green/red indicator signaling alignment between the generated music and the user's prompt and video. On hover, VidTune explains the rationale (Figure 2G). Hovering over the thumbnail also reveals *reusable keywords* (Figure 2F): prominent attributes specific to the current track that are not already specified in the original prompt or title. Users can click a keyword to add it to the prompt for future generations. If the user likes a track, they can save it to the scene with the ⊕ icon. The track appears in *Saved Tracks* (Figure 2C) and its waveform is shown on the filmstrip (Figure 2B). VidTune automatically adds short fade-in/out at boundaries for smoother transitions, and the video player previews the video and selected track together. Users can attach multiple alternatives to the same scene and switch between them using the carousel arrows.

Iterating with natural language. When users hover over a thumbnail, *Edit* and *Vary* buttons appear. To edit a track, users can type in a natural language instruction (*e.g., "Make it more calm"*). VidTune responds with 4 new tracks that implement the user's request in different ways (*e.g.,* reduced percussion, slowed tempo, softer

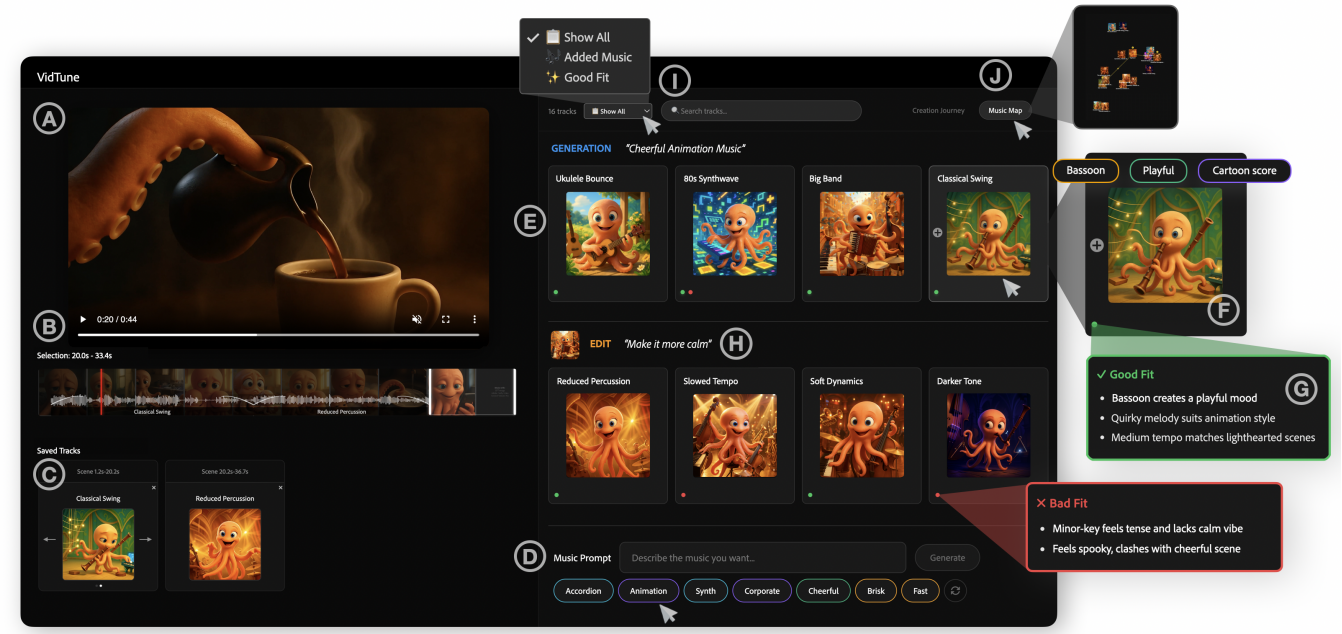

Figure 2: VidTune Interface: The video player and timeline let users choose a scene to add music and preview candidates in sync with the video (A-C). VidTune surfaces prompt suggestions based on the selected scene and user goal (D), then expands the prompt to generate 4 candidates with *contextual thumbnails* (E). On hover, users see reusable prompt keywords (F) and a *fit check* (G). Users can iterate with natural-language-edits (H), organize generations via filter/search (I), or view a *music map* for similarity-based exploration (J).

dynamics, or darker tone). Each new track has a title summarizing the change (Figure 2H). Clicking *Vary* generates 4 new tracks conditioned on the selected track's audio. These variations typically preserve high-level properties such as genre, tempo, and instrumentation while changing local phrasing and texture, giving users quick alternatives to a track they already like.

Managing generation alternatives. As generations accumulate, users can scroll through the grid and narrow down options with *filters* ("Show All", "Added Music", "Good Fit") or *search* by keywords such as instruments or vibe (Figure 2I). For a broader view, users can switch to the *Music Map* (Figure 4), which replaces the grid view with a 2D layout that arranges tracks by audio similarity. While prior maps for music visualization render nodes as dots [78] or artist photos [36], our contextual thumbnail nodes encode rich audio cues (*e.g.*, instruments, mood), making neighborhoods more interpretable and easier to skim at a glance. VidTune's map overlays a yellow dashed path indicating the sequence of tracks already placed in the video. Users can select nearby tracks to explore close variants or choose multiple tracks and click *Blend* to generate a new option that combines their qualities.

5 VidTune Algorithms

5.1 Suggesting and Diversifying Prompts

VidTune automatically suggests prompts for text-to-music models grounded in the user's video context, and expands them into diverse, non-conflicting variations.

Suggesting prompts. VidTune first segments the video into short scenes with a Large Multimodal Model (LMM), capturing story changes such as emotions, events, or settings. For each scene, the system generates 5 keywords per category (instruments, genres, vibes, and energy levels, inspired by the categories in [41]), and when users select a region in the filmstrip, a random subset (1–2 per category) is surfaced to support quick prompting. Clicking "refresh" shows a different random subset from the pre-analyzed set. As users begin placing music tracks on the timeline, VidTune uses an LMM to generate a detailed caption of each selected track. These captions are incorporated as additional context for future keyword generation, allowing subsequent suggestions to reflect both the video scene and the user's emerging preferences. We include the full prompts used for generating the suggestions in the Appendix.

Expanding prompts. To help users explore diverse possibilities, we designed an algorithm that expands each prompt into N variations and generates a corresponding set of candidate tracks (Alg. 1). First, we use an LMM to generate N diverse, non-conflicting suffix modifiers (2–3 words each) spanning genre, instrumentation, mood, and energy. Each modifier is appended to the user's original prompt to form a complete prompt for the text-to-music model. For example, given the prompt *"piano solo"*, the system may suggest modifiers such as *"meditative ambient"* or *"jazz swing"*, but not *"guitar"* or *"strings"*, which would contradict the solo-instrument constraint.

Once candidate tracks are generated, VidTune evaluates their suitability to guide selection. Each track is embedded in CLAP [100],

Figure 3: VidTune's animated thumbnail sampled at 0.5 FPS. As the thumbnail animates, the character switches to instruments that enter later in the audio (A, B), movements convey rough tempo (C), and visual effects convey energy (D).

a multimodal model that maps audio and text into a shared representation space for semantic similarity comparisons. We use CLAP embeddings to score each track by (i) *scene fit*, measuring similarity to a vibe description of the scene (derived from the LMM and embedded as CLAP-text), and (ii) *taste fit*, measuring similarity to the mean embedding of tracks the user has previously added to the timeline. We empirically set $N = 6$ to balance variety and generation speed, and present the top 4 ranked tracks to the user.

Algorithm 1 Prompt Expansion

Inputs:
  Q_t: user query
  $\mathbf{c}$: scene vibe embedding (CLAP-text)
  $\mathbf{u}$: taste prior (mean CLAP-audio of added tracks)
Generate modifiers: $\mathcal{S} \leftarrow \text{Gemini}(Q_t)$ N non-conflicting suffixes
Form prompts: $\mathcal{P} \leftarrow \{\, Q_t \oplus s \mid s \in \mathcal{S} \,\}$
for $p \in \mathcal{P}$ **do**
  $a(p) \leftarrow \text{T2M}(p)$ music generation
  $\mathbf{z}(p) \leftarrow \text{CLAP\_Audio}(a(p))$
end for
Scene fit: $R(p, \text{scene}) = \cos(\mathbf{z}(p), \mathbf{c})$
Taste fit: $R(p, \text{taste}) = \cos(\mathbf{z}(p), \mathbf{u})$
Score: $R(p) = \alpha R(p, \text{scene}) + (1 - \alpha) R(p, \text{taste})$
Filter: $\mathcal{P}' \leftarrow \text{TopK}(\mathcal{P}, 4 \text{ by } R(p))$
Output: 4 tracks $p \in \mathcal{P}'$, ordered by $R(p)$

5.2 Generating Contextual Thumbnails

VidTune's contextual thumbnails allow creators to instantly see and compare how different musical options would feel in the context of their actual footage. To generate the thumbnails, VidTune 1) extracts the core semantic visual anchor from the video (*e.g.,* key subject or scene) and 2) analyzes each track for musical attributes including genre, valence, energy, instruments, and tempo. Then, we generate each thumbnail by blending the musical attributes directly into the style of the visual anchor (Figure 5).

Identifying the visual anchor. Our pipeline first uses an LMM to identify at most three core visual anchors in the video. When the anchor is an animated or non-human character (*e.g.,* a cartoon figure, animals), VidTune recreates it as closely as possible to maintain consistency with the video. When the anchor is a real human, instead of reproducing their likeness, VidTune generates a stylized avatar that reflects key appearance details (*e.g.,* hair color, clothing) while avoiding direct modification of real people [18, 60] and the associated uncanny valley concerns [70]. To maintain consistency across multiple thumbnails, VidTune uses an LMM to generate a post-description of each detected anchor, which is reused to render visually coherent images. When no clear protagonist is present, the LMM identifies the central object or theme of the scene. For example, in a travel vlog, it detects "Paris" as the theme and selects a representative frame (*e.g.,* Eiffel Tower) as the anchor (Figure 5). In videos centered on a core object – such as an advertisement or a product demo – the object itself is chosen as the anchor.

Analyzing music and generating thumbnails. In order to produce visual information from music tracks in a consistent way, we developed a mapping framework (Table 2) grounded in color–emotion correspondence [75, 93], links between tempo/energy and perceived motion speed [26], and affective associations of genre as settings [55]. Beyond aggregating prior works, this framework operationalizes them into a single, composable schema that yields consistent, multi-attribute thumbnails for music. Whenever a new track is generated, we prompt an LMM to describe its musical character using this schema, conditioning the prompt with the mapping rules and a few examples [11]. From this description, we then generate a *stylistic modifier* – a short phrase that translates the music's qualities into visual effects. Finally, the stylistic modifier is appended to the visual anchor's description to construct a complete prompt for the text-to-image model. For example, in Figure 5, the anchor *'a shot of the Eiffel Tower'* is combined with a modifier derived from upbeat electronic music to become: *'A cinematic shot of the Eiffel Tower at night. The artistic style is futuristic electronic, with the surrounding scene transformed into a vibrant neon cityscape* (Genre & Style). *The image is bright and highly saturated with a warm color palette of*

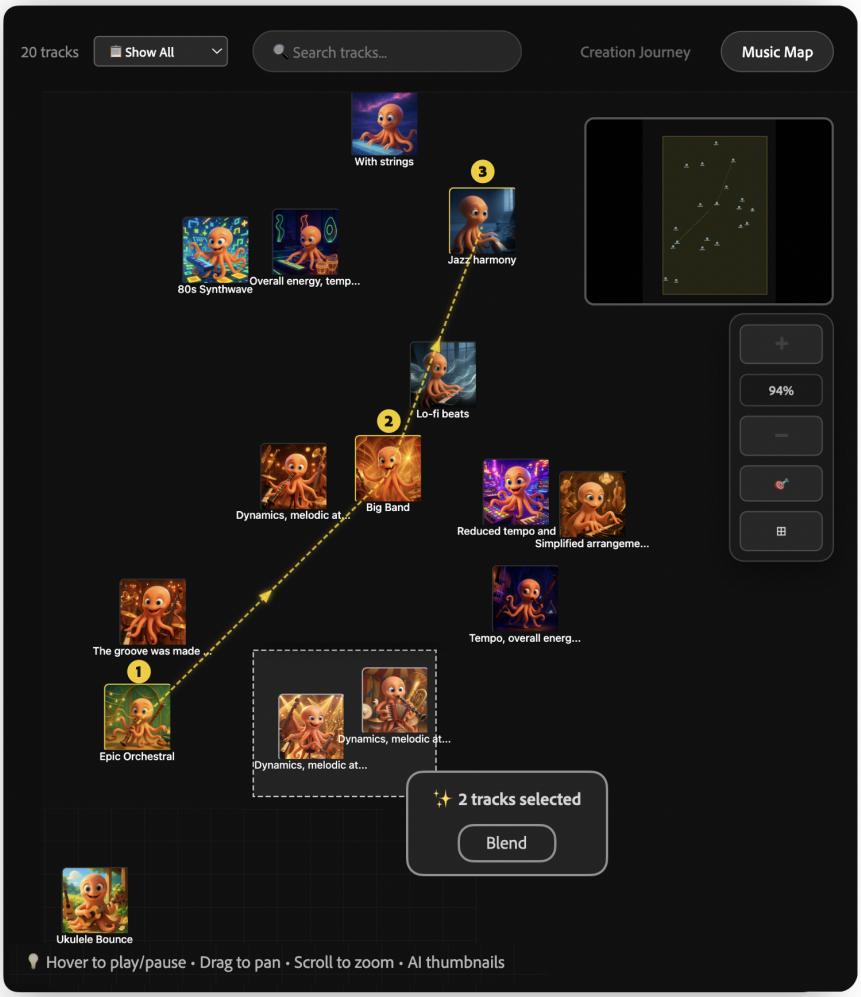

Figure 4: VidTune's Music Map arranges generated tracks in a 2D space by audio-embedding similarity (CLAP [100]), revealing families of related music at a glance. A dashed path shows the sequence of tracks added to the current video. Users can multi-select tracks and use *Blend* to generate similar variations.

gold and magenta (Energy & Valence). *There is a dynamic sense of implied motion, with subtle light streaks and motion blur to reflect the fast tempo* (Tempo)'. Finally, we generate an animated thumbnail with a video model that takes the static thumbnail as the first frame and uses the same music-informed style prompt, so that tempo descriptors (*e.g.,* rapid light streaks, slow drifting movement) are reflected in the animation speed and motion intensity.

5.3 Representing the Music Space

To provide a global view of all generated tracks, VidTune projects them into a two-dimensional music map. Each track is embedded in CLAP [100], and we compute audio–audio similarities between tracks and apply t-SNE [65] to position them so that nearby points correspond to CLAP-similar tracks. In practice, we observed that local neighborhoods tend to group tracks with similar genre, instrumentation, and mood, so we treat the map as an approximate similarity layout for exploration rather than as interpretable axes for specific musical features. As users generate or save new tracks, the map updates dynamically, allowing them to understand explored regions and identify gaps for further exploration.

5.4 Refining Music Generations

VidTune supports 3 operations for iterative refinement: **edit**, **vary**, and **blend**. **Edit** requests (*e.g., "make it more energetic"*) are interpreted by the LMM into 4 alternative strategies (*e.g., "increase tempo," "add percussion," "brighter chord progression"*). For each strategy, the LMM rewrites the track's structured music description (from §5.2) to preserve core attributes while reflecting the requested modification, and uses the rewritten descriptions to regenerate the 4 edited candidates. **Vary** regenerates new options by providing the audio of a selected track as input to condition the model, producing close variations to the original. **Blend** combines multiple tracks

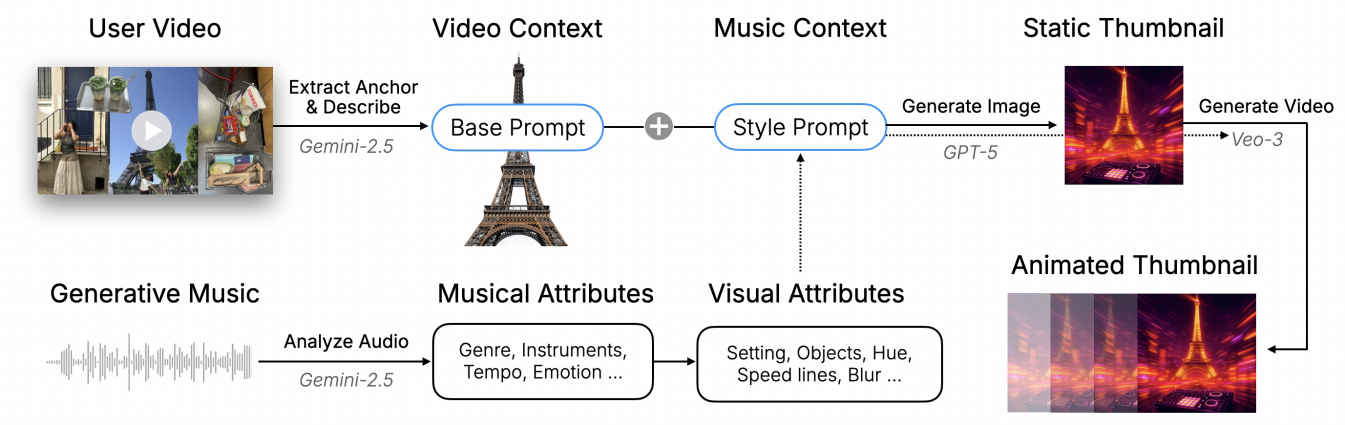

Figure 5: From a user video, VidTune extracts an anchor subject and description to form a *base prompt* grounded in the footage. It analyzes the generated music to infer musical attributes, maps them to visual cues, and fuses these into a *style prompt*. The fused prompt is used to generate static and animated thumbnails that reflect both the video context and the music.

Table 2: Mapping of musical attributes to AI-generated thumbnails in VidTune.

Musical Attribute	Visual Mapping Rule
Genre & Style	Background scene and artistic style *e.g.,* electronic → neon cityscape
Instruments	Protagonist performing instrument (size proportional to prominence)
Tempo	Implied motion *e.g.,* fast → speed lines & blur
Emotion & Mood	Facial expressions and body language matching emotion
Valence	Visual filter: hue/tint adjustment *e.g.,* positive → warm, negative → cool
Energy	Visual filter: brightness and saturation *e.g.,* high → bright, low → dim

by prompting the LMM with their descriptions to produce a unified prompt capturing their shared traits. This blended prompt is then used to generate 4 new tracks that inherit qualities of both inputs. Together, these interactions help users progressively refine the soundtrack while maintaining coherence across scenes.

5.5 System Implementation

We implemented VidTune as a full-stack application. The interface is built in React and TypeScript and uses native Web APIs [71] for video playback and audio fade-in/out effects. The backend is a Flask REST server [74] that processes audio with FFmpeg [33] and runs CLAP [100] in PyTorch [79] for embedding generation and similarity scoring. The algorithms use the LMM `gemini-2.5-flash` [88] for video analysis (prompt suggestions, thumbnail anchor extraction) and music analysis, `gpt-5-image` [72] for generating thumbnail images, and Veo-3 [27] (`veo-3.0-fast-generate-preview`) for generating animated thumbnails. We used a custom text-to-music model trained on licensed instrumental music, which takes either a text prompt or an audio file as input along with desired duration, and generates music with the given duration (VidTune uses the duration of the selected scene). Model implementations are not our core contribution, and any component could be swapped for an alternative (*e.g.,* a different text-to-music model).

6 Technical Evaluation

We conducted technical evaluations of VidTune's core algorithmic components, focusing on (1) whether prompt expansion increases music diversity and (2) how well contextual thumbnails convey the corresponding music.

6.1 Diversity of Music from Expanded Prompts

Method. We built a test corpus of 20 diverse music prompts that generated **160** music tracks. To derive prompts that cover diverse soundtrack contexts, we sampled 5 videos (V1-V5 in Table 7). For each video, we selected 2 scenes and authored both a short prompt (*e.g., 'lively busy music'*) and a long prompt (*e.g., 'Upbeat and quirky indie pop, 130 bpm. A driving acoustic guitar and bouncy bassline create a positive, energetic rhythm. Features a bright pizzicato string melody suggesting multitasking.'*) for each scene. For each prompt, we generated 8 tracks using the same text-to-music model as VidTune: 4 from the original prompt (*Baseline*) and 4 from LLM-expanded variations (*VidTune*). We embedded audio with CLAP (512-d) [100] and computed 2 measures per prompt:

- *Mean pairwise cosine distance* within the set (higher = broader spread).
- *Cluster separation* (higher = clearer split into clusters).

For cluster separation, we use k=2 as a minimal check of whether outputs separate into more than one group, serving as a lightweight, interpretable complement to distance.

Results. Overall, we did not find a statistically significant difference in diversity between VidTune and the baseline (Table 3). For *short* prompts, VidTune's diversity was noticeably higher, with

significance in cluster separation (cosine: t=1.96, p=0.079; cluster sep.: t=2.42, p<0.05). Results for *long* prompts showed no reliable differences between conditions. The lack of diversity gains for *long* prompts likely reflects that these prompts already encode rich, specific constraints (instrumentation, tempo, style, mood), and even with well-crafted variations, differences can be buried by the dense specification. Yet, as seen in our formative study, users tended to use short abstract prompts to explore breadth and long detailed prompts for specific targets. It is therefore reasonable that VidTune's expansions most effectively increase variety when user intent is open-ended.

6.2 Music–Thumbnail Correspondence

To assess how well VidTune's thumbnails both represent their underlying music and support comparison, we conducted a controlled evaluation guided by the following research questions:

- **RQ1**: How well do VidTune's thumbnails **represent** their corresponding music?
- **RQ2**: How well do VidTune's thumbnails help users **associate** the right music in the presence of similar alternatives?

Baseline. Generating visual thumbnails to represent music is not new, and commercial tools already use them. We include this practice as a baseline for comparison, but these tools do not disclose how thumbnails are produced. Our analysis of tracks and their thumbnails generated in the formative study shows that they often loosely reflect explicit prompt terms (*e.g.,* "rainbow" leading to rainbow imagery) or the general vibe of the music (*e.g.,* relaxed tracks paired with sky or beach scenes). We approximate this by providing both the music prompt and the generated track to an LMM [88] along with instructions to generate an image prompt, which is then used to create a thumbnail image with the same text-to-image model [72] as VidTune.

Method. We collected **1,600** judgments from 20 annotators (80 trials each: 40 ratings, 40 selections). Stimuli comprised 40 audio clips (20 prompts × 2 variations) and 80 thumbnails (VidTune + baseline per clip). We reused the same 5 videos and 20 prompts (short and long) from §6.1 and the videos covered diverse types of anchors for thumbnail generation (*e.g.,* animated characters, landmarks, human avatars). For each prompt, we randomly sampled 2 of the 4 VidTune-generated clips as the audio stimuli. For each clip, we produced a VidTune thumbnail and a baseline thumbnail.[1] Including both short and long prompts enabled the selection task to probe associations under easier (more distinct) and harder (more similar) audio conditions.

Annotators self-identified as experts (N=1), intermediate (N=11), and novices (N=8) in music. We used a within-subjects design with counterbalancing, and the task order was randomized. Thumbnails were shown without video context to test whether our thumbnail alone establishes strong music–image correspondence. Annotators completed:

- **Rating (RQ1):** Given one audio clip and one thumbnail (VidTune or baseline), rate on a 1–7 scale how well the image represents the audio (7 = most representative).
- **Selection (RQ2):** Given one audio clip and two thumbnails, choose the image that matches the audio (2AFC). Both thumbnails were from the same condition (either VidTune or baseline) with one from the target audio and one foil from a different audio variation of the same prompt, mirroring typical use.

Table 3: Diversity metrics by prompt length. Bold indicates the larger mean (μ) within each Baseline–VidTune pair and * indicates p<.05 with paired t-test.

	Total		Short prompts		Long prompts	
	μ	σ	μ	σ	μ	σ
Pairwise cosine distance (0-2)						
Baseline	0.19	0.07	0.22	0.09	**0.17**	0.05
VidTune	**0.22**	0.10	**0.29**	0.12	0.16	0.04
Cluster separation						
Baseline	18.87	4.89	20.27	4.89	**17.47**	3.59
VidTune	**20.15**	5.32	**23.77***	6.21	16.53	2.33

Table 4: Results for Rating and Selection by prompt length (within-subject, n=20). Bold indicates the larger mean (μ) within each Baseline–VidTune pair. Significance is denoted as $^*p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$ (paired t-test).

	Total		Short prompts		Long prompts	
	μ	σ	μ	σ	μ	σ
Rating (1–7)						
Baseline	4.63	0.62	4.72	1.80	4.55	1.92
VidTune	**4.99***	0.74	**5.14****	1.52	**4.83**	1.56
Selection (%)						
Baseline	69.8	11.6	76.1	42.7	63.3	48.3
VidTune	**88.2*****	7.3	**93.5*****	24.7	**83.0*****	37.7

Results. Table 4 shows that VidTune thumbnails outperformed the baseline on both the *Rating* and *Selection* tasks. We analyzed within-subject differences using paired t-tests and confirmed all effects with Wilcoxon signed-rank tests. In the rating task, VidTune thumbnails scored significantly higher than the baseline ($t(19)$ = 2.28, p = 0.035, d = 0.51), with 13 of 20 annotators preferring them. Differences in preference may stem from annotators judging thumbnails without the original video context or preferring abstract illustration over concrete cues (*e.g.,* explicit instruments). In the selection task, VidTune thumbnails yielded significantly higher 2AFC accuracy than the baseline, with a large effect size ($t(19)$=7.14; p<0.001; d=1.60). Every annotator performed as well or better with VidTune thumbnails, demonstrating that VidTune's thumbnails reliably encode music and surface distinctions, even when the two audio variants were generated from the same prompt.

[1] Study materials and annotation interface details provided as supplementary materials.

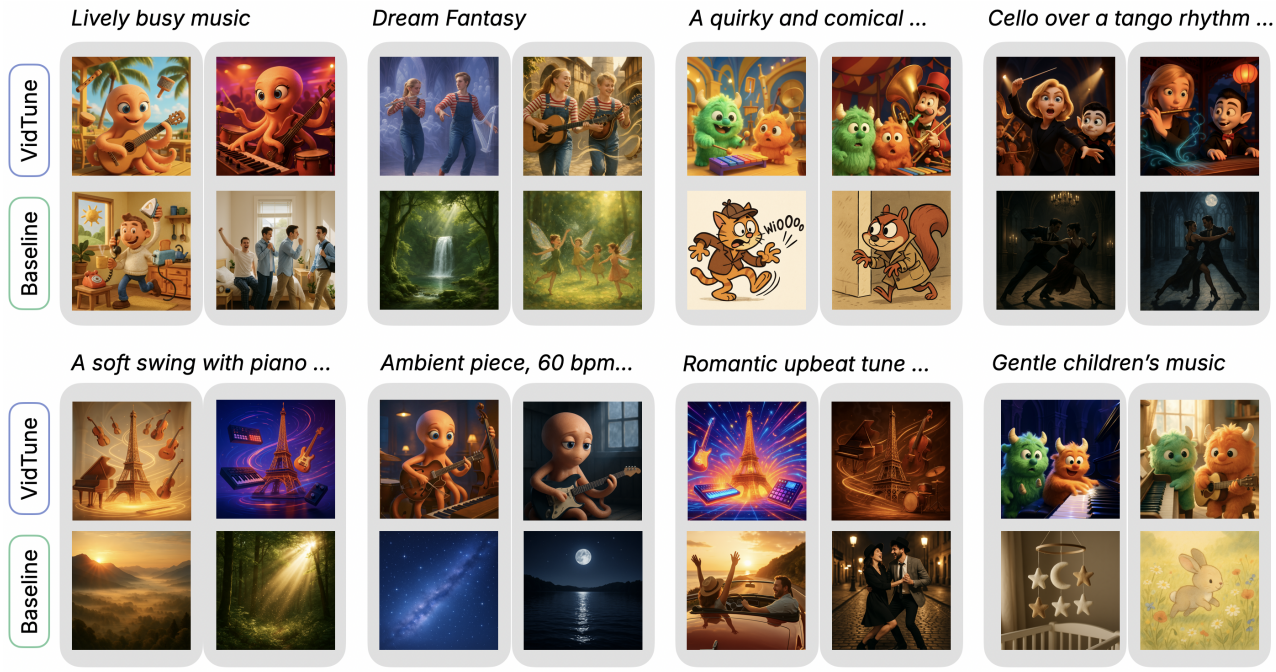

Figure 6: Example thumbnails from the pipeline evaluation study. For each prompt (column), the top row shows VidTune thumbnails and the bottom row shows the baseline. VidTune tends to convey music-relevant cues (*e.g.*, genre, instruments), while the baseline leans toward literal prompt imagery or abstract scenes.

Figure 6 illustrates qualitative examples from VidTune and the baseline as presented to annotators. For VidTune's thumbnails, differences between candidates were often more distinguishable, conveyed through color palettes, background settings, facial expressions, and depicted instruments. In the baseline thumbnails, prompts with salient visual components (*e.g.*, tango rhythm, comical) were often literalized, overshadowing musical attributes. Of the 40 baseline thumbnails, only 1 featured instruments present in the audio, 28 featured a character or central object, and 12 were abstract scenes. Distinctions were slightly clearer for *short* prompts (which produced more divergent audio). For instance, the prompt *"Romantic upbeat tune"* yielded one electronic-leaning and one jazzier variation, and the two baseline thumbnails slightly captured that split in mood.

7 Controlled User Study

To further understand how VidTune supports soundtrack generation, we conducted a within-subjects study with 12 video creators comparing VidTune to a baseline.

7.1 Method

Participants. We recruited 12 participants with diverse video creation experience using mailing lists (P9-P20, §A Table 6). 5 described themselves as proficient (P11, P14-P16, P18), having an average of 14 years of video editing experience (SD=6.46), and 7 were amateurs (P9-P10, P12-P13, P17, P19-P20) with 4.86 years (SD=2.27) of experience. The study lasted 1.5 hours, conducted either remotely via Microsoft Teams/Zoom or in-person based on participant preference, and we compensated $60.

Baseline. To examine the benefits of VidTune relative to existing designs in text-to-music generation, we implemented a baseline (Figure 7) inspired by widely used commercial systems [86, 92]. Because these tools do not document whether they diversify prompts internally, we approximate their typical observable user experience: a single free-form text prompt that yields multiple tracks and simple cover images. For controlled comparison, both the baseline and VidTune use the same underlying music model, generate four tracks

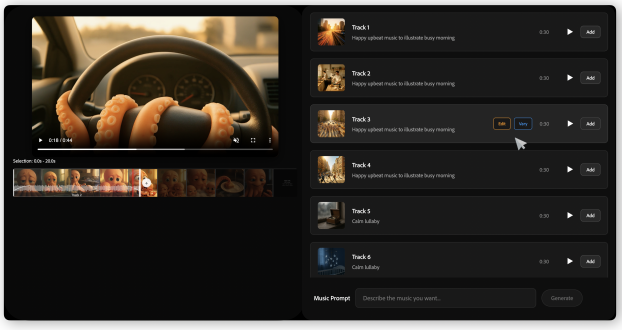

Figure 7: Baseline interface, designed to resemble the UI and features of existing music generation tools.

per prompt, and share the same video player pane and timeline so users can preview music in sync with their video and see tracks with fade-in/out transitions.

For thumbnails, we use the baseline generation pipeline from our pipeline evaluation (§6.2), which conditions an LMM on the user prompt and generated music and then calls the same text-to-image model [72] as VidTune. This produces generic *vibe-oriented* thumbnails without video context or explicit mappings from musical attributes, which are specific to VidTune's contextual thumbnails. The baseline also supports iterative refinement with *edit* and *vary*: *vary* uses the same audio-seed variation mode as VidTune, while *edit* creates a single new prompt with an LLM from the original prompt and edit request and then generates four tracks, but without VidTune's structured diversification.

Materials. We selected 3 videos (V1, V2, V6 in §A Table 7) from an independent filmmaker's short animated films, all generated using tools including Adobe Firefly [1], GPT [72], and Veo 3 [27]. Each video was between 45–60 seconds long without narration. V1 was used in the tutorial session for both VidTune and baseline conditions. The main study used V2 and V6, which were similar in length, visual complexity, and scene changes.

Procedure. We began with an interview on participants' video creation practices and typical approaches to adding music. Participants then received a 10-minute tutorial on both VidTune and the baseline interface (V1). Each participant used both interfaces, with up to 25 minutes per interface, to add music to one of two study videos (V2, V6). The assignment of videos and the order of interfaces were counterbalanced and randomly assigned. After each session, participants completed a post-stimulus survey measuring cognitive load (NASA-TLX [43]), creativity support (CSI [17]), and custom items related to visual thumbnails. We omitted inapplicable items (*e.g.,* physical demand) and overlapping ones across scales (*e.g.,* NASA-TLX *Effort* vs. CSI *results worth the effort*).

8 Results

Figure 8 summarizes ratings for cognitive load and creativity support, and Figure 9 reports thumbnail-specific ratings. In this section, we synthesize findings on VidTune's exploration support (§8.1), review support (§8.2), and additional observations on how participants work with VidTune (§8.3).

8.1 Exploration Support

On average, participants created 28 tracks in VidTune (SD = 10.65; 1.83 edits, 1.80 variations, 0.17 blends) compared to 16 tracks in the baseline (SD = 1.82; 1.67 edits, 1.00 variations).

Guided exploration reduces effort. When generating music with VidTune, all participants utilized prompt suggestions and often added their own keywords or explanations to refine the prompt. As music generation required waiting time in both conditions, participants often started new generations in parallel or revisited earlier outputs. Participants using VidTune reported significantly higher result-worth-effort (μ=6.00, σ=0.60 vs. μ=5.08, σ=1.68; Z=-2.08; p<0.05). They explained that VidTune's expanded prompts produced more diverse outputs, helping them explore possibilities and find tracks that matched their goals, whereas the baseline often returned 4 very similar tracks, leading them to generate more options before reaching a desirable one. 3 participants also noted that VidTune made iteration easier by allowing them to easily reuse prompts by clicking on keywords (P8, P12-P13).

Requests for expanding expressive range. We did not observe statistically significant differences in perceived exploration support or satisfaction between VidTune and the baseline. 4 participants wanted more fine-grained control, such as reusing only part of a generated track (P14), extracting and editing stems directly (*e.g.,* only increasing piano volume) (P13, P15, P18), or using more advanced fade types (P18). 3 participants also input sound effects based on the video scene (*e.g.,* keyboard typing, traffic noises), but noticed the model omitted them (P11-P13). 2 participants noted that outputs sometimes missed the requested style and wanted brief system explanations on what failed and how to steer it, so they could redirect generation (P17, P19).

8.2 Contextual Thumbnails for Reviewing Music

Figure 9 shows that participants found VidTune's thumbnails significantly more useful for understanding music (μ=5.5, σ=1.51 vs. μ=3.33, σ=1.78; Z=-2.38; p<0.05), comparing different tracks (μ=5, σ=1.41 vs. μ=3, σ=1.71; Z=-2.47; p<0.05), and remembering them (μ=5.42, σ=1.51 vs. μ=3.17, σ=1.8; Z=-2.33; p<0.05).

Contextual thumbnails accurately represent music. Participants described a close match between the music and VidTune's thumbnails, as P16 explained: *"When a thumbnail had a dark background and fireworks, it matched the music really well, so I'd say it was pretty accurate."* P17 noted that thumbnails can sometimes be more helpful than titles in understanding music: *"Terms like Celtic jazz or Synthwave are hard to grasp when I just read but much easier when I see them visually."* P9 found that the baseline's thumbnails *"conveyed only the 'vibe' of a song,"* while VidTune's thumbnails *"also communicate an 'acoustic representation' "*, showing that VidTune further helps anticipate genre, instrumentation, and intensity.

Contextual thumbnails aid review burden. VidTune required significantly less temporal demand (Figure 8, μ=5.33, σ=1.44 vs. μ=3.67, σ=1.67; Z=2.13; p<0.05). While both systems had similar end-to-end generation times – and VidTune had additional delay for prompt expansion and music analysis – participants still experienced VidTune as less rushed. With VidTune, 3 participants often skipped tracks and selectively listened to tracks with thumbnails that visually resembled ones they had already liked (P9, P16, P21); *"because I can tell which would fit my video better by looking at them"* (P9). In contrast, all 12 participants in the baseline condition listened to all generated tracks, often replaying them because it was hard to visually preview and filter options. P12 explained, *"When it [Baseline] gives many songs that look the same, I feel like I must review all of them."* showing that VidTune reduced this burden by presenting more distinct thumbnails. Similarly, P10 noted about the baseline thumbnails *"Sometimes it's not what I imagined when I clicked to listen."*

Contextual thumbnails serve as reliable memory cues. Participants noted that VidTune's thumbnails support recall [19]. P11

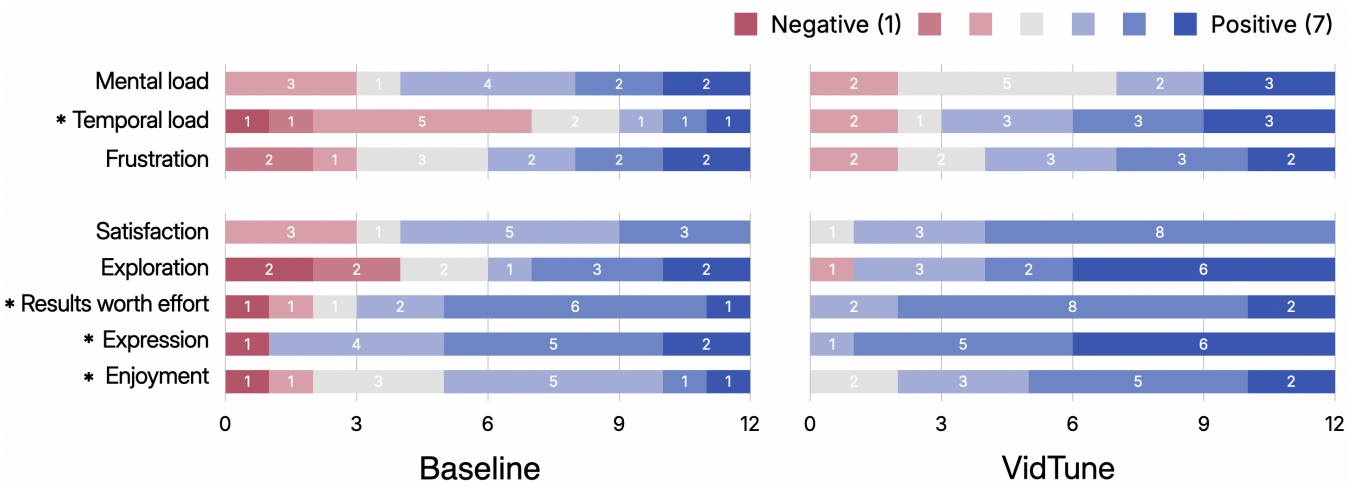

Figure 8: Distribution of rating scores for the Baseline and VidTune (1 = negative, 7 = positive) in the two tasks on cognitive load and creativity support. Higher values indicate more positive feedback. * indicates statistical significance from Wilcoxon tests ($p < 0.05$).

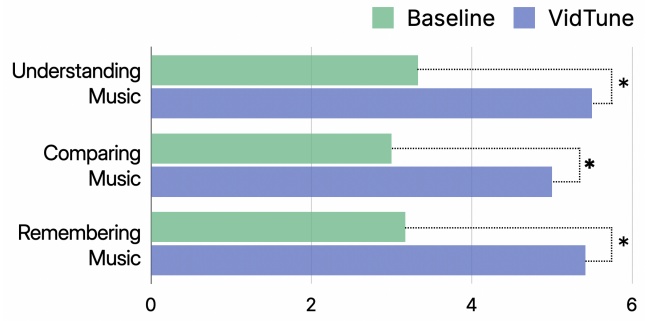

Figure 9: Rating scores on thumbnail-specific questions for the Baseline and VidTune (1 = negative, 7 = positive) across the two tasks. Higher values indicate more positive feedback. Asterisks denote statistical significance from Wilcoxon tests (* $p < 0.05$).

stated *"It's 100% helping when finding what music I liked. Names [titles] are not that useful, thumbnails are more useful."* By contrast, baseline thumbnails offered little support for comparison or recall: *"Unless I really pay attention, it's hard to remember what music I listened to. It's too simple"* (P11). Additionally, inconsistent visual styles in the baseline further challenged comparing tracks using thumbnails: *"It's a bit too random. Some are illustrations, some are real photos. I cannot compare when one image is showing a group of people partying and the other shows an animated cat"* (P9). With abstract imagery in baseline thumbnails, participants often forgot which thumbnail matched which track and had to re-listen to identify it.

Contextual thumbnails bring a more engaging, personalized experience. Participants enjoyed using VidTune significantly more (μ=6.42, σ=0.67 vs. μ=5.42, σ=1.56; Z=-2.13; p<0.05) and felt more expressive (μ=5.58, σ=1.00 vs. μ=4.50, σ=1.51; Z=-2.13; p<0.05). Although our design goals did not explicitly aim to increase enjoyment, participants described VidTune's thumbnails as making the music creation process more enjoyable and personally meaningful. P17 remarked, *"More pleasing, more fun to interact with. Usually with video editing, I always just look at randomly colored blocks, waveforms, and it's more like a task. But this makes the process itself more interesting."* P11 similarly emphasized the sense of personal connection: *"It helps because it uses my characters, feels more connected. Feels more like this is generating music for MY own video."*

With animated thumbnails, 9 participants found they offered an additional sense of playfulness; as P14 mentioned, *"Not sure how professional editors would think, but as a novice creator, this [animation] is just more pleasing."* Similarly, P16 said they were *"more memorable as they're more dynamic."* 2 participants also noticed additional instruments in the background when the thumbnail animated, and felt that the motion aligned closely with the energy of the track (P10, P16). However, 3 participants felt the additional motion did not add much to the static image, describing it as an extra but not essential (P9, P11) or even distracting (P20).

Limitations. Despite these benefits, participants also pointed out several limitations of VidTune's thumbnails. 2 participants wanted the option to hide thumbnails and see more detailed textual descriptions of each track, as P20 explained, *"I'm not a visual person, I always prefer to read things."* P18 observed that the thumbnails sometimes contradicted the titles, which caused confusion. For example, a track titled *Ukulele Folk* was paired with a thumbnail showing a xylophone. The actual audio had sound qualities that could be interpreted as either instrument, as music generation models often blend overlapping acoustic features [2]. Similarly, P12 mentioned that recurring characters in the thumbnails caused them to look similar. Comparison became harder after multiple rounds of prompt iterations, when the generated samples started to converge and looked nearly identical.

Broader use cases. Participants suggested additional creative use cases for VidTune's thumbnails. P15 envisioned using them as a *visual scaffold* for iteration *"Sometimes I only remember instruments by the visual and not the name, so I want to draw on [VidTune's*

thumbnails] or paste [images of] instruments on it." 4 participants wanted to reuse them as creative assets – for example, P18 wanted to export thumbnails for teasers and ending credits of his videos, or use them as emojis with music for fun communication. P9 and P14 mentioned their potential as music videos or album covers for low-budget artists. Others highlighted educational and collaborative potential: P9 noted that combining audio with visual thumbnails could help children learn music *"like concept cards"*, and P16 explained that thumbnails could make it easier to communicate musical choices with deaf collaborators.

8.3 Additional Observations

Fit checks are adopted selectively. We observed varied use of VidTune's fit check feature in participants' workflows. 3 participants actively relied on fit checks, using them to filter out weaker candidates (P10) or to reassure themselves after adding a track to the video (P9, P12). In contrast, some chose not to use them at all, as P13 explains *"I didn't read the fit check because it might bias me. Even if the music isn't bad, if it tells me it's a bad fit I might not consider it."* Similarly, P16 noted, *"Fit checks don't really influence my decision. At the end, it's more of a creative choice, so there isn't a right or wrong."*

Music map facilitates cross-music consistency. In both conditions, all but one participant (P20) added multiple tracks to the final video and emphasized the importance of maintaining consistency to avoid abrupt transitions between scenes. 3 participants used the music map to spot gaps between tracks added to the video (P8, P15, P16). They verified that their tracks were not too distant from one another and compared the thumbnail styles of nearby alternatives in the map. When a candidate felt off, they switched to a closer neighbor or a previously-saved track. P16 used the map's blend feature to generate middle-scene music from the first and last tracks.

Workflows are iterative. Participants shaped their soundtracks through small, repeated adjustments. First, they iterated within a scene by requesting changes to mood, speed, and instrumentation (adding or removing) in both conditions. In the baseline, P18 tried repeating the original prompt as an edit when the first generation missed key elements. In VidTune, P12 tried prompting *"add [instrument] from option 2"* but found the system did not support referencing attributes from another candidate. After choosing one track, participants iterated to generate a next track that kept style consistent while still allowing emotion to shift.

We observed two iteration strategies: (1) reusing prompt keywords with small, context-specific tweaks (*e.g.,* emotion, tempo, instrumentation), and (2) starting from a favored track, making quick edits or variations, and reusing those across later scenes. Two participants using *Vary* (P14, P15) sometimes struggled to choose among similar options, and wanted clearer text descriptions of differences. Based on these observations, we speculate that the lack of significant differences in mental load (VidTune: μ=4.75, σ=1.48; baseline: μ=4.92, σ=1.44; Z=-0.31; p=0.75) and frustration (VidTune: μ=5.08, σ=1.38; baseline: μ=4.58, σ=1.73; Z=0.68; p=0.50) reflects the need for iterative generation and review, as well as occasional misalignment between steering and intent.

9 Exploratory Case Study

The controlled user study (§7) examined how VidTune compares to a baseline on pre-selected videos. To understand how VidTune supports soundtrack creation across a more diverse set of personal projects, we further conducted an exploratory case study with creators using their own footage (Table 5).

9.1 Method

We recruited 6 creators (2 professionals, 4 amateurs) through mailing lists and social media (P8, P11, P16, P20–P22). One participant (P8) had also joined the formative study, and three others (P11, P16, P20) took part in the controlled study. Consistent with the formative study, we included a deaf creator (P8) to understand whether VidTune's visual approach could improve accessibility for soundtrack generation. All communication with P8 was text-based. Before the study, we collected an edited video (without music) from each participant. Participants' videos ranged from 1-4 minutes and spanned diverse genres (Table 5). 3 out of 6 videos had narration (V8-V10). Figure 10 illustrates resulting thumbnails grounded in participants' videos, including speaker avatars, characters, animals, landmarks, and products. During a 1hr study, we provided a tutorial of VidTune, invited participants to add music to their own video with VidTune, and asked semi-structured interview questions about their experience. We compensated $40.

9.2 Findings

Perceptions on reusing video content. P20, who saw her product calendar appear alongside a mascot in the thumbnail, described the experience as more personal. During the study, she mainly focused on thumbnails instead of switching back and forth to the video player, explaining *"I already know my video since I made it, and because the thumbnails include context, I can just review them instead of rewatching."* P8 was excited to see her dog featured in the thumbnail: *"It's so cute, I want to share all of these [thumbnails] with my family and on my channel."* P16, who generated music for her daily vlog featuring herself, expressed mixed feelings. She noted, *"It's cool that it [VidTune] got my purple hair and black knit correct. But some of the thumbnails don't feel like me."* While she still preferred stylized avatars over photorealistic generations, she explained that even minor mismatches could be distracting because she could easily spot differences in her own appearance.

Accessibility to diverse audiences. P21 used VidTune to add music to his 5-year-old daughter's sketch-style storytelling videos. He found it easier to generate child-appropriate music than searching

PID	VID	Video Type	Duration	Thumbnail Anchor
P8	V6	Vlog (animal)	1:01	Animal
P11	V7	Vlog (scenic views)	1:05	Landscapes
P16	V8	Vlog (talking head)	3:41	Speaker avatars
P20	V9	Product demo	1:36	Mascot & Product
P21	V10	Child Storytelling	3:09	Sketched character
P22	V11	Generated Ads	1:00	Characters & Product

Table 5: Videos used in the exploratory case study.

Figure 10: VidTune thumbnails grounded in participants' own videos from the exploratory study. Diverse video anchors are used as thumbnail subjects.

stock libraries, with VidTune suggesting relevant keywords After seeing the visual thumbnails, P16 was enthusiastic about how the sketch was animated into a character that performs instruments. He highlighted how VidTune could lower the barriers to music generation for children and novices: *"Kids will love it, and parents will love it. There are more and more young creators now, and tools like this help them easily add music and learn music."*

P8, a deaf creator, explained how thumbnails improve accessibility. *"Very helpful. This [VidTune] lets us see the music. I used to only use tracks with lyrics as I can read them and understand, but now I can visualize."* As she got closer to finalizing creation, she wanted more detailed information to confirm her choices—especially since the music would be shared publicly—and suggested that interactive question-answering could support this. She also appreciated the fit check, which increased confidence in her selections, and noted that she wants to share VidTune with many other deaf creators who often struggle with adding music.

Compatibility with current workflows. P11, who often reuses the same track across multiple videos, asked whether VidTune could adapt thumbnails to new contexts – *"If I use this music for another video about cars, will the thumbnails change to cars instead of sailboats? That would be cool, but I can also see it losing some of the 'memory' I had."* She also wanted to use trending songs, either to show them in VidTune or to generate similar tracks, explaining that while novel music is valuable, familiar references can sometimes help audiences connect more quickly. P22, a professional filmmaker who frequently uses stock music, noted that VidTune-style thumbnails could help people skim large catalogs on stock sites. However, he emphasized that as a professional, he is accustomed to reviewing music with waveforms and therefore often prefers this standardized representation. Both P11 and P20 requested lightweight organization aligned with their current practice, such as renaming tracks and adding notes in their own words.

10 Discussion

10.1 Scope and Design Boundaries

Users. VidTune supports a diverse range of users from novices to those with limited hearing, who can benefit from prompt suggestions, contextual thumbnails, and fit-evaluation tools. In our studies, more experienced creators also found the thumbnails useful for rapid skimming and communication, but noted that they would still rely on existing tools for fine-grained control, so we view VidTune as a complementary exploration aid rather than a full expert workstation. Although we did not study children directly, some participants suggested that the thumbnails could help users with shorter attention spans such as children.

Musical Content. VidTune targets non-lyrical (instrumental) music. Extending to vocal tracks would introduce additional review needs (*e.g.,* lyric semantics, vocal style, timing, and prosody) and require multimodal support beyond images, such as highlighting key lyrics synchronized to beats (*e.g.,* karaoke-style phrasing).

Control mechanisms. VidTune currently accepts text prompts and does not support reference audio. It also omits structural scaffolding before generation (*e.g.,* outlining beats/chords [45]), fine-grained mixing/mastering [24], and automatic video editing to fit music [106]. Our formative study and prior work [41] show that editors often try to match music beats to specific cuts, but precise beat-sync requires either retiming the video (adjusting or inserting cuts) or fine-grained temporal control over generated music. Since current text-to-music models mainly allow control over global attributes such as mood and overall tempo rather than beat-level changes, We focus on helping creators select and place soundtracks at the scene level, leaving fine-grained alignment to downstream editing tools. Future work could explore how to incorporate model advances in parameterized controls [99] and beat/phrase/stem extraction while maintaining simple, intuitive interactions.

10.2 Challenges in Generative Abundance

By producing many candidates in a short time with minimal user input, generative models shift creative effort from making to choosing. This shift introduces new challenges. First, homogenization can emerge [4, 5, 42], especially for novices who are unaware of the broader space or how to reach it. We therefore set diversity as a design goal (D2) and built VidTune to expand users' prompts and provide diverse music generations. Second, with abundant choices, review and comparison become the bottleneck, both in terms of time and cognitive load. Building on sensemaking work across text [38, 80], image [3, 48], and video [47], VidTune makes generative music glanceable via contextual thumbnails.

Beyond reviewing burden, a selection-centric workflow could dilute agency and ownership [47, 105] as users may feel less attached to an artifact they spent less time on. Future tools can mitigate this by adding intentional (good) friction [16, 25] that elicits meaningful user input before assistance and by personalizing suggestions to the creator's evolving tastes and project context. As GenAI reshapes creative practices, interfaces should support creators to explore broadly, compare efficiently, and feel ownership over the result.

10.3 Making Music More Visual

VidTune is grounded in the idea of making music more visual – translating soundtracks into forms that can be skimmed, compared, and remembered through sight, drawing on the speed and parallelism of visual perception. Our study showed that thumbnails capturing key musical qualities helped users understand the music [61], filter irrelevant tracks, aid recall [19], and—when contextualized with their video—foster a more personal and enjoyable experience. That said, visualizing music carries inherent limitations. Proxies may not fully capture the richness of sound, mappings between musical and visual features could vary across cultures and individuals [95], and users may risk privileging what *"looks right"* over what truly sounds right. It is therefore crucial to frame such visuals as augmentations to listening, not replacements for it.

While we used thumbnails primarily for sensemaking in soundtrack creation, participants were excited about use cases beyond, including richer music listening with thumbnails or repurposing them for music videos or album covers. Another potential is *music-driven storytelling*, where thumbnails expand into storyboards to spark narrative ideas from music. Finally, future work could also explore music editing through visual manipulations of thumbnails, *e.g.,* enlarging a cello to amplify its volume or substituting it with a double bass to alter orchestration.

10.4 Building with Imperfect Gen AI Models

VidTune relies on LMMs for scene analysis and prompt suggestions, a text-to-music model for generating soundtracks, and image and video generation models for generating thumbnails. These models have limitations and are prone to errors. For instance, suggesting music prompts based on the scene can reflect bias in how LMMs interpret visual context [12]. Music generators may miss user intent for culturally specific styles or underrepresented genres – (*e.g.,* collapsing diverse African traditions into generic "tribal drums"), echoing longstanding concerns about dataset and modeling bias in language models [8] and image models [64]. Finally, biases reported in image generation can also surface when associating musical qualities with visuals in thumbnail generation. For instance, when multiple characters are involved, mappings can drift toward stereotypical assignment (*e.g.,* male characters for "energetic", female for "calm") [42, 64]. To work around and address these issues, future tools should offer transparent, user-facing explanations of how recommendations are generated. They can also intentionally diversify options when uncertainty is high, and provide a feedback loop for users to flag when outputs do not align with requests.

10.5 Role of Playfulness in Creative Tasks

While we did not set out to design VidTune as a playful system, participants reported unexpected enjoyment: reusing their own characters and scenes in thumbnails made reviewing feel personal, and the visuals reduced the tedium of sequential listening. Prior work on gamification shows that making progress visual can sustain attention and momentum [40]; VidTune's thumbnails provide that visibility. As generative AI systems increasingly automate production and shift effort toward intent articulation and review, we argue that creativity tools should preserve the *fun* parts with the creators. In practice, this means offloading tedious steps while making input and review more engaging. VidTune addresses the review side, while complementary works have explored playful input (*e.g.,* toy-play interactions for storytelling with LLMs [22]). We believe future creativity tools should cultivate enjoyable micro-moments to make the process as rewarding as the outcome.

11 Conclusion

We introduced *VidTune*, a system that helps video creators generate soundtracks by expanding prompts, visualizing music with video-grounded thumbnails, and supporting iterative refinement. By making music more *visible in context*, VidTune shifts the effort from sequential listening to rapid, informed comparison. Through a technical evaluation, we showed that VidTune's contextual thumbnails more faithfully reflected musical attributes than baseline thumbnails. In a user study (N=12) comparing VidTune to a baseline text-to-music interface, participants reported that VidTune made it easier to understand, compare, and remember music candidates, and described the workflow as more expressive and enjoyable. Finally, a case study (N=6) with creators' own videos showed that participants valued seeing their footage reflected in the thumbnails and felt this made the generated music feel more tailored to their content. Broadly, this work demonstrates how generative AI shifts soundtrack selection for video from a retrieval task to a creative endeavor, with VidTune illustrating how music creation can be more visual, playful, and accessible for diverse creators.

Acknowledgments

We thank Nick Bryan for his valuable feedback, and Gabi Duncombe for generously allowing us to use her videos in this work. Mina Huh is supported by a Google Ph.D. fellowship.

References

[1] Adobe. 2025. Adobe Firefly. https://www.adobe.com/sensei/generative-ai/firefly.html Generative AI for images and design.

[2] 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

[3] Shm Garanganao Almeda, JD Zamfirescu-Pereira, Kyu Won Kim, Pradeep Mani Rathnam, and Bjoern Hartmann. 2024. Prompting for Discovery: Flexible Sense-Making for AI Art-Making with Dreamsheets. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.

[4] 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*. 413–425.

[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] Aadit Barua, Karim Benharrak, Meng Chen, Mina Huh, and Amy Pavel. 2025. Lotus: Creating short videos from long videos with abstractive and extractive summarization. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*. 967–981.

[7] Benjamin B Bederson. 2001. PhotoMesa: a zoomable image browser using quantum treemaps and bubblemaps. In *Proceedings of the 14th annual ACM symposium on User interface software and technology*. 71–80.

[8] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021), 610–623.

[9] Karim Benharrak and Amy Pavel. 2025. HistoryPalette: Supporting Exploration and Reuse of Past Alternatives in Image Generation and Editing. *arXiv preprint arXiv:2501.04163* (2025).

[10] Stephen Brade, Bryan Wang, Mauricio Sousa, Sageev Oore, and Tovi Grossman. 2023. Promptify: Text-to-image generation through interactive prompt exploration with large language models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–14.

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[12] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability, and Transparency (Proceedings of Machine Learning Research, Vol. 81)*. PMLR, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

[13] Alice Cai, Steven R Rick, Jennifer L Heyman, Yanxia Zhang, Alexandre Filipowicz, Matthew Hong, Matt Klenk, and Thomas Malone. 2023. DesignAID: Using generative AI and semantic diversity for design inspiration. In *Proceedings of The ACM Collective Intelligence Conference*. 1–11.

[14] Chuer Chen, Nan Cao, Jiani Hou, Yi Guo, Yulei Zhang, and Yang Shi. 2023. MusicJam: Visualizing Music Insights via Generated Narrative Illustrations. *arXiv preprint arXiv:2308.11329* (2023).

[15] Jiawei Chen, Luo He, Hongyan Liu, Yinghui Yang, and Xuan Bi. 2024. Background music recommendation on short video sharing platforms. *Information Systems Research* 35, 4 (2024), 1890–1908.

[16] Zeya Chen and Ruth Schmidt. 2024. Exploring a behavioral model of "positive friction" in human-AI interaction. In *International Conference on Human-Computer Interaction*. Springer, 3–22.

[17] Erin Cherry and Celine Latulipe. 2014. Quantifying the creativity support of digital tools through the creativity support index. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21, 4 (2014), 1–25.

[18] Bobby Chesney and Danielle Citron. 2019. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.* 107 (2019), 1753.

[19] Terry L Childers and Michael J Houston. 1984. Conditions for a picture-superiority effect on consumer memory. *Journal of consumer research* (1984), 643–654.

[20] Youjin Choi, JaeYoung Moon, JinYoung Yoo, and Jin-Hyuk Hong. 2025. Exploring the Potential of Music Generative AI for Music-Making by Deaf and Hard of Hearing People. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–20.

[21] John Joon Young Chung, Shiqing He, and Eytan Adar. 2021. The intersection of users, roles, interactions, and technologies in creativity support tools. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference*. 1817–1833.

[22] John Joon Young Chung, Melissa Roemmele, and Max Kreminski. 2024. Toyteller: Toy-playing with character symbols for ai-powered visual storytelling. In *Adjunct Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–5.

[23] Geoffrey Cideron, Sertan Girgin, Mauro Verzetti, Damien Vincent, Matej Kastelic, Zalán Borsos, Brian McWilliams, Victor Ungureanu, Olivier Bachem, Olivier Pietquin, et al. 2024. Musicrl: Aligning music generation to human preferences. *arXiv preprint arXiv:2402.04229* (2024).

[24] Michael Clemens and Ana Marasović. 2025. MixAssist: An Audio-Language Dataset for Co-Creative AI Assistance in Music Mixing. *arXiv preprint arXiv:2507.06329* (2025).

[25] Anna L Cox, Sandy JJ Gould, Marta E Cecchinato, Ioanna Iacovides, and Ian Renfree. 2016. Design frictions for mindful interactions: The case for microboundaries. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*. 1389–1397.

[26] Abe Davis and Maneesh Agrawala. 2018. Visual rhythm and beat. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–11.

[27] Google DeepMind. 2025. Veo 3. https://deepmind.google/technologies/veo/ Text-to-video generative model.

[28] Jordan Aiko Deja, Alexczar Dela Torre, Hans Joshua Lee, Jose Florencio Ciriaco IV, and Carlo Miguel Eroles. 2020. Vitune: A visualizer tool to allow the deaf and hard of hearing to see music with their eyes. In *Extended Abstracts of the 2020 CHI conference on human factors in computing systems*. 1–8.

[29] 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*. 2037–2045.

[30] Shuangrui Ding, Zihan Liu, Xiaoyi Dong, Pan Zhang, Rui Qian, Conghui He, Dahua Lin, and Jiaqi Wang. 2024. Songcomposer: A large language model for lyric and melody composition in song generation. *arXiv preprint arXiv:2402.17645* (2024).

[31] Steven P Dow, Alana Glassco, Jonathan Kass, Melissa Schwarz, Daniel L Schwartz, and Scott R Klemmer. 2010. Parallel prototyping leads to better design results, more divergence, and increased self-efficacy. *ACM Transactions on Computer-Human Interaction (TOCHI)* 17, 4 (2010), 1–24.

[32] Yuexi Du, Ziyang Chen, Justin Salamon, Bryan Russell, and Andrew Owens. 2023. Conditional generation of audio from video via foley analogies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2426–2436.

[33] FFmpeg Developers. 2025. FFmpeg. https://ffmpeg.org/ Accessed Sep. 9, 2025.

[34] 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*. 1–13.

[35] Yue Fu, Michele Newman, Lewis Going, Qiuzi Feng, and Jin Ha Lee. 2025. Exploring the Collaborative Co-Creation Process with AI: A Case Study in Novice Music Production. *arXiv preprint arXiv:2501.15276* (2025).

[36] Pavel Gajdusek and Ladislav Peska. 2021. SpotifyGraph: visualisation of user's preferences in music. In *International Conference on Multimedia Modeling*. Springer, 379–384.

[37] Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba. 2020. Foley music: Learning to generate music from videos. In *European Conference on Computer Vision*. Springer, 758–775.

[38] Katy Ilonka Gero, Chelse Swoopes, Ziwei Gu, Jonathan K Kummerfeld, and Elena L Glassman. 2024. Supporting sensemaking of large language model outputs at scale. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–21.

[39] Dan B Goldman, Brian Curless, David Salesin, and Steven M Seitz. 2006. Schematic storyboarding for video visualization and editing. *Acm transactions on graphics (tog)* 25, 3 (2006), 862–871.

[40] Juho Hamari, Jonna Koivisto, and Harri Sarsa. 2014. Does gamification work?–a literature review of empirical studies on gamification. In *2014 47th Hawaii international conference on system sciences*. Ieee, 3025–3034.

[41] 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 *Proceedings of the 2025 ACM Designing Interactive Systems Conference*. 2738–2754.

[42] Evans Xu Han, Alice Qian Zhang, Hong Shen, Haiyi Zhu, Paul Pu Liang, and Jane Hsieh. 2025. POET: Supporting Prompting Creativity and Personalization with Automated Expansion of Text-to-Image Generation. *arXiv preprint arXiv:2504.13392* (2025).

[43] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[44] Allison Ho. 2025. Paris, France 2025. https://www.youtube.com/watch?v=Rijmgs7fIKs. YouTube video; duration 15:03.

[45] 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*. 241–250.

[46] 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 preprint arXiv:2010.05388* (2020).

[47] Mina Huh, Ding Li, Kim Pimmel, Hijung Valentina Shin, Amy Pavel, and Mira Dontcheva. 2025. VideoDiff: Human-AI Video Co-Creation with Alternatives. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*.

1–19.
[48] Mina Huh, Yi-Hao Peng, and Amy Pavel. 2023. GenAssist: Making image generation accessible. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–17.
[49] Fathinah Izzati, Xinyue Li, Yuxuan Wu, and Gus Xia. 2025. MusiScene: Leveraging MU-LLaMA for Scene Imagination and Enhanced Video Background Music Generation. *arXiv preprint arXiv:2507.05894* (2025).
[50] Youngseung Jeon, Seungwan Jin, Patrick C Shih, and Kyungsik Han. 2021. FashionQ: an ai-driven creativity support tool for facilitating ideation in fashion design. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.
[51] Carles Fernandes Julià and Sergi Jordà. 2009. SongExplorer: A Tabletop Application for Exploring Large Collections of Songs.. In *ISMIR*. 675–680.
[52] Richard Khulusi, Jakob Kusnick, Christofer Meinecke, Christina Gillmann, Josef Focht, and Stefan Jänicke. 2020. A survey on visualizations for musical data. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 82–110.
[53] Yewon Kim, Sung-Ju Lee, and Chris Donahue. 2025. Amuse: Human-AI Collaborative Songwriting with Multimodal Inspirations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–28.
[54] ChungHa Lee and Jin-Hyuk Hong. 2025. musicolors: Bridging Sound and Visuals For Synesthetic Creative Musical Experience. *arXiv preprint arXiv:2503.14220* (2025).
[55] Arto Lehtiniemi and Jukka Holm. 2012. Using animated mood pictures in music recommendation. In *2012 16th International Conference on Information Visualisation*. IEEE, 143–150. doi:10.1109/IV.2012.34
[56] Ruiqi Li, Siqi Zheng, Xize Cheng, Ziang Zhang, Shengpeng Ji, and Zhou Zhao. 2024. Muvi: Video-to-music generation with semantic alignment and rhythmic synchronization. *arXiv preprint arXiv:2410.12957* (2024).
[57] Sifei Li, Weiming Dong, Yuxin Zhang, Fan Tang, Chongyang Ma, Oliver Deussen, Tong-Yee Lee, and Changsheng Xu. 2024. Dance-to-Music Generation with Encoder-based Textual Inversion. In *SIGGRAPH Asia 2024 Conference Papers*. 1–11.
[58] Sifei Li, Binxin Yang, Chunji Yin, Chong Sun, Yuxin Zhang, Weiming Dong, and Chen Li. 2024. VidMusician: Video-to-Music Generation with Semantic-Rhythmic Alignment via Hierarchical Visual Features. *arXiv preprint arXiv:2412.06296* (2024).
[59] Hyungui Lim, Seungyeon Rhyu, and Kyogu Lee. 2017. Chord generation from symbolic melody using BLSTM networks. *arXiv preprint arXiv:1712.01011* (2017).
[60] Hugo B Lima, Carlos GR Dos Santos, and Bianchi S Meiguins. 2021. A survey of music visualization techniques. *ACM Computing Surveys (CSUR)* 54, 7 (2021), 1–29.
[61] Hugo B. Lima, Carlos G. R. Dos Santos, and Bianchi S. Meiguins. 2021. A Survey of Music Visualization Techniques. *ACM Comput. Surv.* 54, 7, Article 143 (July 2021), 29 pages. doi:10.1145/3461835
[62] Vivian Liu, Tao Long, Nathan Raw, and Lydia Chilton. 2023. Generative disco: Text-to-video generation for music visualization. *arXiv preprint arXiv:2304.08551* (2023).
[63] 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*. 1–13.
[64] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2023. Stable Bias: Analyzing Societal Representations in Diffusion Models. arXiv:2303.11408 [cs.CY] https://arxiv.org/abs/2303.11408
[65] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.
[66] Justin Matejka, Michael Glueck, Erin Bradner, Ali Hashemi, Tovi Grossman, and George Fitzmaurice. 2018. Dream lens: Exploration and visualization of large-scale generative design datasets. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.
[67] Justin Matejka, Tovi Grossman, and George Fitzmaurice. 2014. Video lens: rapid playback and exploration of large video collections and associated metadata. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. 541–550.
[68] Sköld Mattias. 2023. *Sound Notation: The visual representation of sound for composition and analysis*. Ph. D. Dissertation. KTH Royal Institute of Technology.
[69] Thomas Barlow McHugh, Abir Saha, David Bar-El, Marcelo Worsley, and Anne Marie Piper. 2021. Towards inclusive streaming: Building multimodal music experiences for the deaf and hard of hearing. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.
[70] Masahiro Mori, Karl F MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & automation magazine* 19, 2 (2012), 98–100.
[71] Mozilla Developer Network. 2025. Web APIs. https://developer.mozilla.org/en-US/docs/Web/API Accessed Sep. 9, 2025.
[72] OpenAI. 2024. GPT Image Generation. https://platform.openai.com/docs/guides/images Accessed Sep. 9, 2025.
[73] OpenAI. 2025. Sora. https://openai.com/index/sora/. Text-to-video generative model. Accessed Sep. 9, 2025.
[74] Pallets Projects. 2025. Flask Documentation (3.x). https://flask.palletsprojects.com/ Accessed Sep. 9, 2025.
[75] Stephen E Palmer, Karen B Schloss, Zoe Xu, and Lilia R Prado-León. 2013. Music–color associations are mediated by emotion. *Proceedings of the National Academy of Sciences* 110, 22 (2013), 8836–8841.
[76] Elias Pampalk, Andreas Rauber, and Dieter Merkl. 2002. Content-based organization and visualization of music archives. In *Proceedings of the tenth ACM international conference on Multimedia*. 570–579.
[77] Sarah Perez. 2025. YouTube rolls out a free AI music-making tool for creators. https://techcrunch.com/2025/04/10/youtube-rolls-out-a-free-ai-music-making-tool-for-creators/ Accessed Sep. 11, 2025.
[78] Savvas Petridis, Nediyana Daskalova, Sarah Mennicken, Samuel F Way, Paul Lamere, and Jennifer Thom. 2022. Tastepaths: Enabling deeper exploration and understanding of personal preferences in recommender systems. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*. 120–133.
[79] PyTorch Foundation. 2025. PyTorch. https://pytorch.org/ Accessed Sep. 9, 2025.
[80] Mohi Reza, Nathan M Laundry, Ilya Musabirov, Peter Dushniku, Zhi Yuan "Michael" Yu, Kashish Mittal, Tovi Grossman, Michael Liut, Anastasia Kuzminykh, and Joseph Jay Williams. 2024. Abscribe: Rapid exploration & organization of multiple writing variations in human-ai co-writing tasks using large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.
[81] Steve Rubin and Maneesh Agrawala. 2014. Generating emotionally relevant musical scores for audio stories. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. 439–448.
[82] Steve Rubin, Floraine Berthouzoz, Gautham Mysore, Wilmot Li, and Maneesh Agrawala. 2012. UnderScore: musical underlays for audio stories. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. 359–366.
[83] Epidemic Sound. [n. d.]. Bring your story to life | Music & SFX for videos | Epidemic Sound. https://www.epidemicsound.com/
[84] Kun Su, Xiulong Liu, and Eli Shlizerman. 2020. Audeo: Audio generation for a silent performance video. *Advances in Neural Information Processing Systems* 33 (2020), 3325–3337.
[85] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2024. Luminate: Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–26.
[86] Suno. [n. d.]. Suno | AI Music. https://suno.com/
[87] Takumi Takahashi, Satoru Fukayama, and Masataka Goto. 2018. Instrudive: A Music Visualization System Based on Automatically Recognized Instrumentation.. In *ISMIR*. 561–568.
[88] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).
[89] Daniel Tencer. 2024. Suno, after being sued by the majors for copyright infringement, launches V4, which it says takes AI music creation 'to the next level'. *Music Business Worldwide* (Nov. 2024). https://www.musicbusinessworldwide.com/suno-after-being-sued-by-the-majors-and-hiring--timbaland-as-strategic-advisor-preps-launch-of-v4-claimed-to-be-a-new-era-of-ai-music-generation12/
[90] Purva Tendulkar, Abhishek Das, Aniruddha Kembhavi, and Devi Parikh. 2020. Feel the music: Automatically generating a dance for an input song. *arXiv preprint arXiv:2006.11905* (2020).
[91] Zeyue Tian, Zhaoyang Liu, Ruibin Yuan, Jiahao Pan, Qifeng Liu, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. 2025. Vidmuse: A simple video-to-music generation framework with long-short-term modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 18782–18793.
[92] Udio. [n. d.]. Udio AI Music Generator. https://www.udio.com/
[93] Patricia Valdez and Albert Mehrabian. 1994. Effects of color on emotions. *Journal of experimental psychology: General* 123, 4 (1994), 394.
[94] Rob van Gulik, Fabio Vignoli, and Huub van de Wetering. 2004. Mapping music in the palm of your hand, explore and discover your collection. In *Proceedings of the 5th International Conference on Music Information Retrieval*. Queen Mary, University of London London.
[95] Robert Walker. 1987. The effects of culture, environment, age, and musical training on choices of visual metaphors for sound. *Perception & psychophysics* 42, 5 (1987), 491–502.
[96] 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*. 699–714.
[97] Yubo Wang, Fengzhou Pan, Danni Liu, and Jiaxiong Hu. 2023. Music-to-facial expressions: emotion-based music visualization for the hearing impaired. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 16096–16102.

[98] Allison Woodruff, Andrew Faulring, Ruth Rosenholtz, Julie Morrsion, and Peter Pirolli. 2001. Using thumbnails to search the web. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 198–205.

[99] 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

[100] 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. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[101] Zhifeng Xie, Qile He, Youjia Zhu, Qiwei He, and Mengtian Li. 2025. FilmComposer: LLM-Driven Music Production for Silent Film Clips. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 13519–13528.

[102] Hiromu Yakura and Masataka Goto. 2023. IteraTTA: An interface for exploring both text prompts and audio priors in generating music with text-to-audio models. *arXiv preprint arXiv:2307.13005* (2023).

[103] Meng Yang, Maria Teresa Llano, and Jon McCormack. 2024. Exploring Real-Time Music-to-Image Systems for Creative Inspiration in Music Creation. *arXiv preprint arXiv:2407.05584* (2024).

[104] HyeonBeom Yi, Dasom Choi, Suhyeon Yoo, Youngmi Song, Jun Woo Lee, Chi Yoon Jeong, and Sungyong Shin. 2025. Toward More Inclusive Music Experience: Understanding Deaf and Hard-of-hearing Individuals' Everyday Music Activities. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*. 375–388.

[105] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*. 841–852.

[106] Xinyu Zhang, Dong Gong, Zicheng Duan, Anton van den Hengel, and Lingqiao Liu. 2025. Let Your Video Listen to Your Music! *arXiv preprint arXiv:2506.18881* (2025).

[107] Yixiao Zhang, Akira Maezawa, Gus Xia, Kazuhiko Yamamoto, and Simon Dixon. 2023. Loop copilot: Conducting ai ensembles for music generation and iterative editing. *arXiv preprint arXiv:2310.12404* (2023).

A PARTICIPANTS

Table 6: Demographics of study participants (Formative study (N=8): P1-P8; Evaluation study (N=12): P9-P20; Exploratory study (N=6): P8, P11, P16, P20-P22)

PID	Gender	Age	Job	Videos Created	AI Music Experience	Hearing
1	Male	43	Magician	Demonstrations, AI films	Regular	Hearing
2	Female	42	AI filmmaker	AI films, Documentaries	Regular	Hearing
3	Male	34	Illustrator	Motion graphics, commercials	Never	Hearing
4	Female	26	Motion designer	Reels, Sports highlights	Occasional	Hearing
5	Female	45	ASL teacher	Teaching materials	Never	Deaf (Profound)
6	Male	46	Design instructor	Student testimonials	Regular	Hearing
7	Female	36	Podcast producer	Podcasts	Never	Deaf (Moderate)
8	Female	35	ASL teacher	Teaching materials, Travel vlogs	Never	Deaf (Profound)
9	Female	22	Technical manager	Vlogs	Never	Hearing
10	Female	24	Product manager	Vlogs, Short-forms	Never	Hearing
11	Female	45	Content creator	Sports recaps, How-To videos	Never	Hearing
12	Female	26	Software engineer	Skits, Weddings	Never	Hearing
13	Male	27	Account manager	Music videos, AI films	Regular	Hearing
14	Female	27	Filmmaker	Music videos, Animated films	Regular	Hearing
15	Male	23	Technology analyst	Vlogs, Short-forms	Occasional	Hearing
16	Female	30	Content creator	Vlogs, UX Tutorials	Regular	Hearing
17	Female	30	Campaign manager	Short-forms	Never	Hearing
18	Male	45	Designer	Explainers, Travel Vlogs	Occasional	Hearing
19	Female	33	Software engineer	Music performance videos, Vlogs	Never	Hearing
20	Female	23	Product manager	Educational/lecture videos	Occasional	Hearing
21	Male	38	Freelance video editor	Animations, Family videos	Occasional	Hearing
22	Male	56	Filmmaker	Animations, AI films, Ads	Regular	Hearing

B VIDEO MATERIALS

Table 7: Videos used in the technical evaluation (Section 6) and controlled user evaluation (Section 7). All generated films were provided by an independent filmmaker and are used with consent. Participants' own videos used in the exploratory case study are in Table 5.

Video ID	Video Type	Duration	Thumbnail Anchor	Studies used
V1	AI-generated film	0:32	Animated Characters	§6, §7
V2	AI-generated film	0:44	Animated Characters	§6, §7
V3	AI-generated film	0:42	Animated Humans	§6
V4	AI-generated film	2:00	Animated Humans	§6
V5	Vlog (CC-licensed) [44]	15:03	Landmark	§6
V6	AI-generated film	0:58	Animated Characters	§6, §7

A Pipeline Prompts

1. Video Analysis & Prompt Suggestions

Task: Analyze uploaded videos to extract scenes with timestamps and music guidance

Prompt Template:
Analyze this video and segment it into meaningful scenes. For each scene, provide detailed music composition guidance.

Please return a JSON response with this EXACT structure:

```
{
  "videoAnalysis": {
    "totalDuration": "MM:SS",
    "scenes": [
      {
        "sceneId": 1,
        "startTime": "MM:SS",
        "endTime": "MM:SS",
        "duration": "MM:SS",
        "sceneDescription": "Detailed description of what's happening visually",
        "vibeDescription": "Emotional atmosphere, mood, and feeling of the scene",
        "musicKeywords": {
          "genres": ["array of 10+ music and video genres that would fit this scene"],
          "instruments": ["array of 10+ instruments that would enhance this scene"],
          "moods": ["array of 10+ emotional/mood descriptors for music"],
          "energy": ["array of 10+ energy level and tempo descriptors"]
        }
      }
    ],
    "overallKeywords": {
      "dominantGenres": ["top 4 genres for the entire video"],
      "primaryMoods": ["top 4 moods for the entire video"],
      "recommendedInstruments": ["top 4 instruments for the entire video"],
      "energyProfile": "overall energy description",
      "tempoRange": "recommended tempo range"
    }
  }
}
```

Requirements:
1. Segment video into 2–6 logical scenes based on visual/narrative changes
2. Each scene should be 15–60 seconds long
3. Provide 10+ options for each music keyword category per scene
4. Focus on music production keywords that would help generate appropriate background music
5. Consider cinematic, documentary, and video content music styles
6. For genres: include both music genres (electronic, orchestral, acoustic) and video/cinematic genres (thriller, romance, action)
7. For energy: include both energy levels (high, medium, low) and tempo descriptors (fast, slow, building, driving)
8. Ensure keywords are diverse and cover different musical approaches for each scene

Video duration: {videoDuration} seconds

Return ONLY valid JSON, no additional text or explanation.

Template Variables: {videoDuration}

Table 8: Video analysis prompt.

2. Music Analysis: Fit Check & Tags (Part 1)

Task: Analyze generated music tracks using audio input for fit evaluation and tagging

Prompt Template:
You are a music expert with the ability to listen to and analyze audio. Listen carefully to this music track and provide comprehensive analysis based on what you actually hear in the audio.

Base your analysis only on what you hear in the attached audio file. Ignore any unrelated text context.
You must identify and include all instruments and sounds you hear in the audio.

Reference info (for final fit evaluation only):
– Original User Prompt: "{originalPrompt}"
– User Video Metadata: title "{title}", type "{videoType}", audience "{audience}", goal "{soundtrackGoal}"
– Video Scene Data (current reference): description "{sceneDescription}", vibe "{sceneVibe}", duration "{sceneDuration}" (seconds); start "{sceneStart}"–end "{sceneEnd}" (seconds)

Analysis task:
Listen to the attached audio file and analyze what you hear:

1. Fit evaluation: Evaluate how well it fits the original user prompt and the current scene/video context.
   – Good fit: provide 2–3 specific reasons (4–7 words each) why the music works well
   – Bad fit: provide 1–2 specific reasons (4–7 words each) why the music does not work well

Examples of good responses:
   – "Acoustic guitar fits family content mood"
   – "Medium tempo aligns with scene energy"
   – "Genre matches dominant video themes"

Examples of bad responses:
   – "Tempo too fast for slow scenes"
   – "Genre mismatch with video content"
   – "Energy level too high for calm scenes"

Avoid vague statements like "doesn't match", "unspecified genre", "not suitable".

2. Prominent tags: Generate 3 tags based on what you hear in the audio.
   – 1 genre tag (Purple: #8B5CF6)
   – 1 mood/emotion tag (Green: #10B981)
   – 1 instrument/technical tag (Blue: #06B6D4)

Instrument tag requirement: choose the most audible/prominent instrument (drums, piano, guitar, bass, strings, brass, synthesizer, etc.).

Table 9: Music analysis prompt (part 1).

3. Music Analysis: Image Description & Details (Part 2)

3. Image description: Create a complete visual description based on what you hear in the music.
– Include all instruments you detect
– Show the protagonist character interacting with all instruments
– Match visual energy to musical energy
– Refer to the character as "the protagonist character"
– Use 4–5 sentences covering all instruments and elements

Image prompt generation instructions:
For each track, create a detailed image prompt using this structure:

Base character: "${protagonist}"

Then add these elements based on the music analysis (examples are guidance, not fixed choices):

1. Visual scene & genre (examples):
– Jazz: "A warm, intimate jazz club with soft amber lighting and swirling smoke patterns"
– Electronic: "A futuristic urban landscape under a neon glow, with energy conduits crisscrossing through digital architecture"
– Orchestral: "A grand natural amphitheater surrounded by majestic mountains and flowing waterfalls"
– Rock: "A dynamic concert stage with dramatic lighting and electric energy crackling through the air"
– Ambient: "A serene, ethereal space with floating geometric forms and gentle light emanations"
– Folk: "A cozy, rustic environment with warm wooden textures and natural, earthy elements"

2. Protagonist performance (examples):
– ${videoType === 'octopus' ? 'Her tentacles gracefully embody the musical instruments' : 'She gracefully embodies the musical instruments'} – [describe specific instruments from the caption]
– "The visual prominence and size of each instrument effect correlates with its prominence in the music"

3. Implied motion & tempo (examples):
– Fast/energetic: "rapid, energetic movements with sharp light trails, motion blur, and dynamic speed lines"
– Slow/calm: "flowing, graceful movements with smooth light trails and gentle undulating patterns"
– Medium: "rhythmic, measured movements with steady light pulses and balanced visual flow"

4. Emotion (examples):
– Happy/joyful: "joyful expression with bright, sparkling eyes and an uplifted, confident posture"
– Calm/peaceful: "serene expression with peaceful, steady eyes and relaxed, flowing posture"
– Energetic: "enthusiastic expression with intense, focused eyes and dynamic, powerful gestures"
– Mysterious: "enigmatic expression with knowing, slightly narrowed eyes and graceful, controlled movements"

5. Color palette: Follow overall trends from color theory; use harmony or contrast to reinforce mood.
– Warm, high-energy palettes often leverage complementary contrasts (reds/oranges against blues) to emphasize vibrancy.
– Calm or peaceful palettes favor analogous, low-saturation hues for balance and unity.
– Dark or mysterious palettes rely on deep values and limited contrast to build intrigue.
Examples:
– Positive/energetic: "vibrant, warm colors dominated by golden yellows, energetic oranges, and uplifting blues"
– Calm/peaceful: "soft, pastel tones with gentle blues, warm creams, and subtle lavenders"
– Mysterious/dark: "dark, rich colors with deep purples, midnight blues, and ethereal silver accents"

End with: "Style: 3D animation, Pixar-quality rendering, cinematic lighting, highly detailed, vibrant and expressive."

Return only a JSON object with this format:
{ "fitAnalysis": {...}, "prominentTags": [...], "imageDescription": "...", "detailedMusicDescription": "..." }

Detailed music description must cover tempo, time signature, key, instrumentation, form, genre specificity, articulation, production, emotional character, and distinctive elements. The description should be comprehensive enough to accurately recreate or edit the musical content.

Template Variables: {originalPrompt}, {trackTitle}, {fullQuery}, {videoMetadata}, {sceneData}

Table 10: Music analysis prompt (part 2).

4. Edit Request Expansion

Task: Expand a user's edit request into 4 creative variations that minimally modify the original while reflecting the user's intent.

Prompt Template:
You are a music production expert. A user wants to edit an existing music track with this request: "{editRequest}"

Original music description: "{originalMusicDescription}"
Original prompt: "{originalPrompt}"

Critical requirements:
1. Intent-focused editing: ensure the user's intent is clearly reflected throughout the description. If they say "make it calmer", the overall feel should be noticeably calmer.
2. Strategic changes: modify all relevant elements that support the user's intent (tempo, energy, instruments, dynamics, mood).
3. Preserve core identity: keep the genre and main instrumentation unless specifically requested to change; adapt energy/mood/tempo as needed.

Your task: create 4 conservative variations that minimally modify the original description.

Examples of intent-focused editing (illustrative, not fixed choices):
Original: "Upbeat electronic dance track with pounding drums and bright synths"
User request: "make it calmer"
– Variation 1 Title: "Calmed Down" (description: "Relaxed electronic ambient track with soft drums and mellow synths")
– Variation 2 Title: "Reduced Energy" (description: "Gentle electronic chillout track with subtle drums and warm synths")
– Variation 3 Title: "Slower Pace" (description: "Laid-back electronic downtempo track with gentle drums and soothing synths")
– Variation 4 Title: "Softened" (description: "Peaceful electronic meditation track with quiet drums and ethereal synths")

Title requirements (must follow exactly):
– Maximum 2–4 words
– Simple verbs: Added, Removed, Slower, Faster, Softer, Louder, Calmer, Energetic
– No abstract terms: avoid "overall", "rhythmic dive", "dynamic", "enhanced", "textural"

Title examples (copy these patterns):
– User says "add piano" → "Added Piano"
– User says "make slower" → "Slower Tempo"
– User says "more energy" → "More Energy"
– User says "remove drums" → "No Drums"
– User says "calmer" → "Calmed Down"
– User says "louder" → "Louder Volume"

Return only a JSON response with this exact format:

```
{
  "variations": [
    { "description": "Original description with targeted changes throughout to fulfill user intent", "title": "MAXIMUM 2–4 WORDS describing the change", "emphasis": "What specifically was changed" },
    { "description": "Original description with targeted changes throughout to fulfill user intent", "title": "MAXIMUM 2–4 WORDS describing the change", "emphasis": "What specifically was changed" },
    { "description": "Original description with minimal targeted changes", "title": "Just the edit change (2–3 words max)", "emphasis": "What specifically was changed" },
    { "description": "Original description with minimal targeted changes", "title": "Just the edit change (2–3 words max)", "emphasis": "What specifically was changed" }
  ]
}
```

Template Variables: {editRequest}, {originalMusicDescription}, {originalPrompt}

Table 11: Edit request expansion prompt.

5. Blend-Based Expansion

Task: Given two or more music descriptions, infer their common attributes, produce one detailed common description, and suggest 4 short variations that remain similar to the given examples.

Prompt Template:
You are a music production expert. Identify common attributes across multiple reference music descriptions and propose closely related variations.

Input music descriptions (2 or more):
{
"desc_1": "{musicDescriptions[0]}",
"desc_2": "{musicDescriptions[1]}",
"desc_N": "{musicDescriptions[N]}"
}

Analysis goal: derive a unified, detailed common description that captures overlapping style, genre/sub-genre, tempo/energy, instrumentation roles, rhythmic feel, harmonic palette, production aesthetics (space/reverb/stereo image/dynamics), and emotional character that are shared by the inputs. Preserve elements that consistently recur; avoid outliers unique to only one example.

Variation goal: propose four short postfix-style variations (2–4 words each) that keep the music similar to the references while exploring small, controlled differences (tempo nudges, instrument emphasis shifts, articulation tweaks, mix emphasis). Titles must follow the same rules as in edit expansion: short, literal, non-abstract.

Examples (illustrative, not fixed choices):
References suggest: "Organic lo-fi chill with soft drums, warm electric piano, mellow bass, gentle sidechain, relaxed 80–90 BPM, intimate stereo, cozy mood."
– Variation: "Added Piano" (emphasis: "bring out warm e-piano voicings")
– Variation: "Softer Drums" (emphasis: "brush/loose kit, reduced transients")
– Variation: "Deeper Bass" (emphasis: "rounder low end, sustained notes")
– Variation: "Airy Reverb" (emphasis: "slightly longer tails, wider space")

Return only a JSON object with this exact format:
{
"commonDescription": "Detailed unified description capturing shared attributes (genre/sub-genre, tempo/energy, rhythmic feel, instrumentation roles, harmonic tendencies, production aesthetics, emotional character).",
"variations": [
{ "title": "2–4 words", "emphasis": "what to tweak while staying similar" },
{ "title": "2–4 words", "emphasis": "what to tweak while staying similar" },
{ "title": "2–4 words", "emphasis": "what to tweak while staying similar" },
{ "title": "2–4 words", "emphasis": "what to tweak while staying similar" }
]
}

Template Variables: {musicDescriptions}

Table 12: Blend-based expansion prompt.