MoSound: An Interactive Tool for Generative Sound Design in Motion Graphics

Jialin Huang
George Mason University
Fairfax, Virginia, USA
jhuang26@gmu.edu

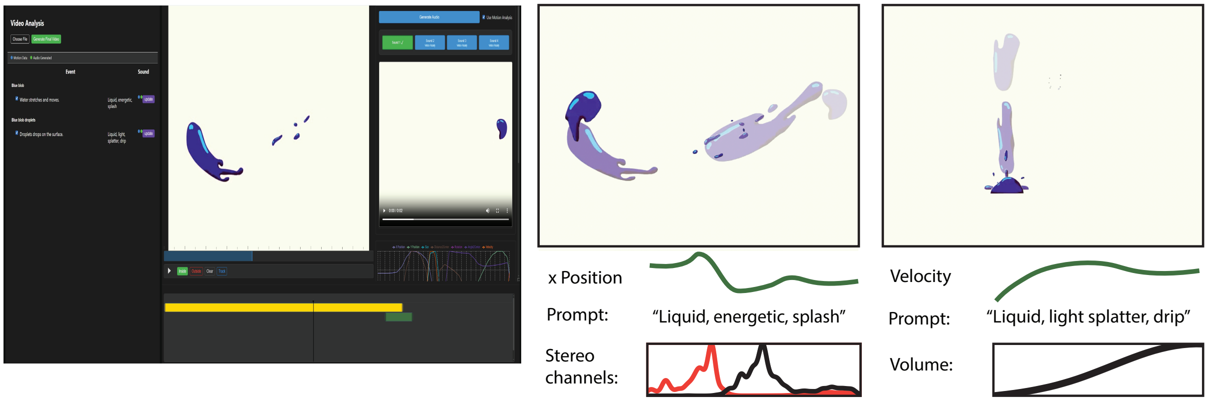
Prem Seetharaman
Adobe Research
San Francisco, California, USA
seethara@adobe.com

Timothy Richard Langlois
Adobe Research
Seattle, Washington, USA
tlangloi@adobe.com

Li-Yi Wei
Adobe Research
San Jose, California, USA
lwei@adobe.com

Rubaiat Habib Kazi
Adobe Research
Seattle, Washington, USA
rhabib@adobe.com

Yotam Gingold
Computer Science
George Mason University
Fairfax, Virginia, USA
ygingold@gmu.edu



Event	Sound
Blue blob	
☑ Water stretches and moves.	Liquid, energetic, splash
Blue blob droplets	
☑ Droplets drops on the surface.	Liquid, light, splatter, drip

(A) User Interface with multiple audio events

(B) Mapping **x** to stereo panning

(C) Mapping **v** to volume

Figure 1: *Generating effects sounds from motion graphics videos via MoSound.* Given a motion graphics video, our interactive system extracts the key events and generates the corresponding sound effects. (A) The user interface of our system, which includes views of the automatically extracted visual events, their timing, and descriptions of the graphics and motions. From there, the user can choose how to map the visual events to the sound effects, preview the generated sound effects, and export the final audio. (More details are in Figure 3.) (B,C) Several keyframes of an example video, along with the corresponding *guide sounds* from the extracted motions, suggested prompts, and final *effect sounds* generated by MoSound. Please see the accompanying video for animation and sound effects. Water video © Alejandro Imondi.

Abstract

Motion graphics, which bring logos, text, and other illustrations to life, are greatly enhanced with sound effects. Sound design for motion graphics presents unique challenges due to their short, abstract nature. Sound designers must identify opportunities for adding sound, decide on the sound's character to match the visual graphics, synchronize sounds with events, and align sonic properties with motions. We introduce MoSound, an interactive system that helps with all steps of this creation process. We designed the interface of MoSound based on formative studies with practitioners and implemented the system as a combination of visual event detection, spatial attribute mapping, and generative sound stylization. We demonstrate MoSound on a variety of examples, showing that it is capable of creating high quality soundtracks while being accessible to novices.

CCS Concepts

• **Computing methodologies → Motion processing**; • **Human-centered computing → Sound-based input / output**.

Keywords

sound synthesis, motion graphics, generative AI, vision-language models

CHI '26, Barcelona, Spain
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791162

ACM Reference Format:
Jialin Huang, Prem Seetharaman, Timothy Richard Langlois, Li-Yi Wei, Rubaiat Habib Kazi, and Yotam Gingold. 2026. MoSound: An Interactive Tool for Generative Sound Design in Motion Graphics. 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, 13 pages. https://doi.org/10.1145/3772318.3791162

1 Introduction

Motion graphics are short animations in which logos, text, and other simple illustrations are brought to life. They are often enhanced with sound effects that add vitality [14], emotional impact [31], and memorability [47]. The sound design for a motion graphic is an integral part of its appeal and recognizability. For example, Netflix named its annual public-facing event "Tudum" after its iconic sound logo. Other examples of sound logos include the HBO Static Angel, the Intel jingle, the MGM lion, the THX Deep Note, and the 60 Minutes ticking clock. Designing sounds for motion graphics presents unique challenges due to the often short, abstract visual content and the need to match the visual and auditory events in timing and character. Pre-existing rhythmic audio like music will not, in general, match the events. Good sound design provides *synchresis*, "the spontaneous and irresistible mental fusion, completely free of any logic, that occurs between a sound and a visual when these occur exactly at the same time" [14].

Both the research literature (e.g. [37–39, 66]) and commercial products (e.g., [1, 2, 7]) have explored tools for creating and editing motion graphics. These tools help with visual motion generation, allowing users to add motions to static illustrations via a variety of input modalities, including text prompts, timeline sequences, and node-based interfaces. However, they only consider visual rather than sound design.

There are multiple tasks that a sound designer must accomplish: identifying opportunities for adding sounds, deciding on the character of the sounds based on the visual content, creating the sounds, synchronizing the sounds' timing with the graphical events, and aligning the sounds' precise sonic properties with the motions.

Current automatic or low-guidance generative techniques (e.g., [11–13, 24, 35, 42, 68]) are one-shot generation pipelines. They provide controllability at the clip level (text prompts, reference audio, etc.) but don't support direct frame-level editing or adjustment. Only MMAudio [12] reliably generated reasonable results but only when provided with precise and well-crafted prompts, which MoSound is designed to automate. See a detailed comparison in Section 5.2.

MoSound aims to help with all steps of this process while providing flexible user control for novices and experts. MoSound presents users with a list of automatically identified events, populated by a vision language model (VLM)'s analysis of the video frames (Figure 3a). Events are accompanied by a description of a sound effect (Figure 3a) and its placement on the timeline (Figure 3f). Users can modify these suggestions, adjust timing, and identify objects for motion tracking (Figure 3b). Users can map motion characteristics (Figure 3d) such as position and velocity to the properties volume and stereo panning (Figure 3e). A sound synthesis model [22] generates sound effects from the textual descriptions and optional mapped properties (Figure 3c).

We conducted a user study to evaluate our pilot prototype (Section 6). Overall, participants are satisfied with the generated sound effects and found the overall interface fun and easy to use. Experts appreciated the ability to quickly create sound effects, though noted that more precise controls are needed for MoSound to replace professional tools. Results from the user study along with those created by the authors are provided in the supplemental materials.

The contributions of our work are as follows:

- MoSound, a human-in-the-loop interactive workflow that facilitates sound designs for motion graphics videos by combining visual event detection, motion-to-sound mapping, sound effect suggestion, and user control.
- A motion-to-sound mapping between visual events and generative sound effects, which enables users to create a wide range of sound designs for different types of motion graphics videos.
- Insights from a user study in which experts and novices use MoSound, demonstrating that the mixed-initiative nature of the design lowers barriers for novices and offers experts valuable prototyping tools.

2 Related Work

2.1 Motion graphics design and authoring

Despite their often short temporal durations, motion graphics tend to be very challenging to create, given the multiple stages involved and the steep learning curves for various tools faced by novice designers [30]. Thus, simplifying and unifying the authoring process is a key goal of many recent works, including end-to-end tools [28, 29, 39], automation [54], design spaces [25, 49], and specific aspects such as motion [37, 56, 67] and color [48]. However, these prior works predominantly focused on motion and visuals instead of sound, which our work addresses.

2.2 Sound design

Sound can be an integral part of logos for branding and marketing [52] and a key modality to convey information, aesthetic, and emotional qualities in interactions [26, 45]. However, sound design can be a quite involved process, requiring both technical expertise and artistic creativity. INVISO [9] is an interface for designing virtual sonic environments from existing sounds that can be interactively explored by spatial positioning. In contrast, we focus on synthesizing new sounds to enhance linear motion graphics videos. Kamath et al. [32] studied the adoption of generative AI for professional sound designers and made several observations on their practices and recommendations for interface design. They focused on the synthesis of individual sounds controlled by two kinds of sliders (domain- and technology-specific), including recommendations for interacting with the model itself (e.g., perceptually linear controls, using spectrograms to visualize sounds). In contrast, we focus on harnessing sound synthesis for the multiple events in a motion graphic video with an interface suitable for novices and experts.

2.3 Sound generation

Sounds can be generated in a variety of ways, including physical recording, simulation, and generative AI.

Physical sounds. Foley [5] is a sound effect creation technique that uses physical objects to create sounds that are recorded and synchronized with visual content. It is popular in film production but can require high expertise and significant physical labor. Sha et al. [46] explored sounds as a form of human-matter interaction, where sound is used not only to convey information, but also to create and modify the experience of a physical space or object. We focus on a purely digital scenario, leveraging recent advances in generative AI to create sounds for an input video.

Simulated sounds. To avoid manipulating physical objects, a variety of works have explored sound generation by simulating the underlying physical phenomena [6, 10, 34, 70]. These approaches can produce highly realistic sounds perfectly synchronized with visual objects. However, these approaches are not directly applicable to abstract visuals such as animated logos, which are not typically grounded in physical reality. Furthermore, for motion graphics, cinematic effects are often more important than physical realism; even for logos that do resemble physical objects, the sound is often exaggerated for effect. Our system addresses these limitations by generating sound effects which may not be physically realistic, but are based on the motion graphics video, rather than relying on physical simulation or field recordings.

Generative sounds. Generative AI provides an alternative approach to sound synthesis beyond field recording and physical simulation. Some works rely solely on text prompts to generate sound effects, similar to one-shot text-to-image generation [11–13, 36, 65, 68]. With these approaches, it is difficult to control the timing and style of the generated sound effects. This is true for OpenAI's Sora 2 and Google's Veo 3, which generate both video and audio from a text prompt and optional start and end frames. Other works provide more flexible sonic modalities to control the sound generation process [15, 17, 19, 22, 41, 55, 63], as well as synthesis driven by sketching [57], gestures [20], and scene objects [27, 35, 51]. AutoFoley [23], an early such work, is a deep learning-based approach for generating synchronized sound tracks for silent captured videos. The method is trained on a dataset of captured natural scenes with a discrete catalog of sound effects (e.g., footstep, typing, car, horse, etc.) and may not be applicable to abstract motion graphics. MoSound leverages this line of work for motion graphics needs. Namely, MoSound generates text and sound prompts for Sketch2Sound [22], a state-of-the-art approach which stylizes the visual-to-sound curves into high-quality sound effects. Our approach to mapping properties of motion into sound resembles data sonification techniques [27, 50, 61, 62]. However, our mapped sounds are not the end result; rather, they are used as guidance for sound effect synthesis. We compare to recent one-shot generative approaches [11–13, 35, 68] in Section 5.2.

2.4 Music generation

Beyond individual sound clips, music can be generated to provide a more holistic experience. Amuse [33] supports collaborative songwriting by generating chord progressions and melodies from multi-modal inspirations such as images, text, and audio. Other work has embedded generative models directly into professional workflows, for example through MMM-C [53], a one-knob plugin for Cubase that enables AI-assisted composition and has been evaluated with expert composers. Beyond these settings, systems such as AffectMachine-Classical [4] generate real-time classical music conditioned on emotional states, while ConL2M [69] focuses on controllable lyrics-to-melody generation. In contrast to these systems, which primarily aim to assist in creating musical structures, our work targets the under-explored domain of event-based sound design, emphasizing temporal precision and synchronization of generated sound effects with motion in motion graphics.

3 Formative Steps

We conducted a design study to better understand the challenges of existing tools and workflows. While prior works investigated automatic or minimally guided generative audio tools, we aimed to understand the creative possibilities and necessary controls for such methods within a audio design workflow for our domain. Based on our observation and insights, we formulate the design goals for our system.

3.1 Method

We used a mixed-method approach for our formative steps, consisting of expert interviews and surveying existing workflows. We interviewed three expert sound designers (P1–P3) with at least 10 years of experience in sound design for animation and motion graphics. The participants are proficient with motion graphics tools (e.g., After Effects) and sound design tools (e.g., Tsugi, Adobe Audition, Reverb). The interviewers were semi-structured, conducted remotely for approximately 60-80 minutes.

We recruited participants through our professional network using purposive sampling. Participation was voluntary and uncompensated. We anonymized all responses in reporting.

3.2 Observations and Design Goals

Audio-Motion Synchronization. A key challenge for creating sound effects for motion is the accurate audio-video synchronization (such as, directionality, velocity, appearance, disappearance, swinging) that is critical to believability and immersion (P3). Yet it constitutes the most tedious aspect of the design. Using traditional tools (i.e., After Effects), designers have to painstakingly synchronize these two channels using keyframes in the timeline editor (P1–P3). Other tools, such as Tsugi, allow designers to control the properties of sound by drawing, where the position, speed, and other properties of the mouse movements are evaluated to generate a sound effect that matches the drawn motion (P3). However, precise synchronization with the video motion is still challenging and tedious.

Creative controls and exploratory workflows. Due to the abstract nature of motion graphics, artistic interpretation and creative controls are essential for sound effects design (P1–P3). Given a motion graphics video, sound designers prioritize the most significant objects and events that actively contribute to the narrative, audience attention, emotional impact, and brand identity (P3). However, this process requires considerable experience and expertise, and familiarity with the nuanced intricacies of audio design and mixing. For amateurs, it is challenging to identify the right events and express sonic properties with natural language for searching or generating sound effects.

Iterative and multi-layered workflow. Sound design relies on iteration and layering, where designers repeatedly test variations and combine multiple tracks to build complexity (P1, P3). Iteration allows users to explore alternative directions and gradually refine results, while layering enables richer compositions by blending overlapping sounds. Together, these practices foster creativity, encourage experimentation, and mirror the way complex effects are constructed in professional sound design.

Generative Sound Synthesis. Traditionally, sound designers rely on existing audio libraries and Foley for source material. A basic organic sound is shaped with multiple layers of effects to produce dynamic modulations and variations (P3). However, novices lack such libraries and Foley abilities. Moreover, they do not know how to describe or search for their desirable sounds.

Based on the above observations, we define the following design goals for our interactive system.

G1: Synchronization of audio and video: Align the precise sonic properties of sounds with motions and synchronize their timing with graphical events.

G2: Events Detection: Empower designers with intuitive tools to discover and select key graphical events in the video, offering both quick exploration and precise customization for audio design.

G3: Multilayered Timeline: Enables users to stack, balance, and refine multiple sound clips in parallel. It supports the iterative and exploratory nature of sound design and the richness that comes from blending overlapping sounds.

G4: Generative Sound Synthesis: Provide users with on-demand, unique sound effect synthesis using Generative AI. Generative sound synthesis presents an opportunity for rapid exploration and unique sound effect creation. Generating sound descriptions allows novices to participate effectively.

4 MoSound System

Guided by our formative study and design requirements, we created MoSound, an interactive system for generating sound designs for motion graphics videos. At a high level, users import a motion graphics video, review and refine automatically detected events, explore motion-to-sound mappings, test generated sound candidates, and compose them with the input video.

MoSound is implemented with a React-based web interface and a Python backend. Users begin by uploading a motion graphics video, which is analyzed to detect key visual events. For each detected event, motion tracking algorithms extract salient motion features from the corresponding video clip. These extracted motion features are then mapped to sound properties—such as volume and panning—to synthesize a *guide sound*. This *guide sound* serves as a conditioning signal for a text-to-audio generation model, which produces sound effects that are temporally and semantically aligned with the motion graphics. Once the sound effects are generated, they are composited with the original video, resulting in a new video file enhanced with motion-consistent audio. This final output is made available for user download.

4.1 Interface

Before walking through the interface, we provide an overview of the end-to-end workflow in Figure 2. The diagram outlines how MoSound analyzes video frames, extracts events and motion features, enables user-defined mappings, and synthesizes audio that is assembled into a layered timeline.

The interface can be seen in Figure 3 and in the accompanying video. When using MoSound, the user begins by loading a motion graphics video such as a logo animation. MoSound uses a vision-language model (VLM) to identify objects and events involving those objects (Section 4.2).

The objects and their events populate the **Event Panel** along the left side of the interface (A). Each event has a start time, an end time, a visual description (e.g., "Sphere squishes and flattens."), and a proposed sound effect description (e.g., "Liquid, smooth, squish"). The descriptions are editable in-place. (Sound effect descriptions created by experts and expected by the generative model we use are shorter than the text prompts commonly used by image synthesis techniques (§4).) Users can manually add additional events by clicking "Add New Event."

The user selects events to add to the timeline with checkboxes (A). Selected events appears in the **Timeline View** (F, along the bottom of the interface) with the suggested start time and duration. The user can adjust the start and duration in the Timeline View. (The Event Panel contains an "Update" button to re-generate the event and sound description from the adjusted batch of frames.)

To generate sound for an event, the user double-clicks on it in the Timeline View. The relevant frames appear in the **Motion Analysis Panel** (B, along the top of the interface). The Motion Analysis Panel provides access to sound generation and optional motion tracking. The Generate Audio button generates four candidate *effect sounds* (C) based on the event's duration and the sound description text (Section 4.4). The candidate effect sounds are presented side-by-side. For in situ perceptual evaluation, we play each effect sound multiplexed with the video clips for a synchronized audio-video preview. This allows users to compare how well each effect sound aligns with the visual motion. The user can choose from these four options or click the Generate Audio button again for a new set of choices, with or without changes to the descriptions, motion mappings, and event start/end.

The Motion Analysis Panel also provides access to motion tracking for shaping the generated audio to precisely match the visual motion (Section 4.3). To perform motion tracking, the user marks one (or more) points as inside the shape (e.g., inside a water droplet) in one (or more) frames (B). The user can optionally also mark points as outside the shape. The Track button initiates a tracking procedure which segments every frame of the clip and analyzes the motion of the segmented object. This populates the Motion Analysis plot (D) with curves for various properties, such as position, rotation, scale, and velocity. The user can then select a motion property (E) to map to a sound property (stereo panning or volume) and control its smoothing and range. The user can add additional mappings as desired and preview the *guide sound*, which is the audio generated from the motion curves. The guide sound is provided to the sound synthesizer (Section 4.4) along with the event and sound description to produce the corresponding *effect sound*.

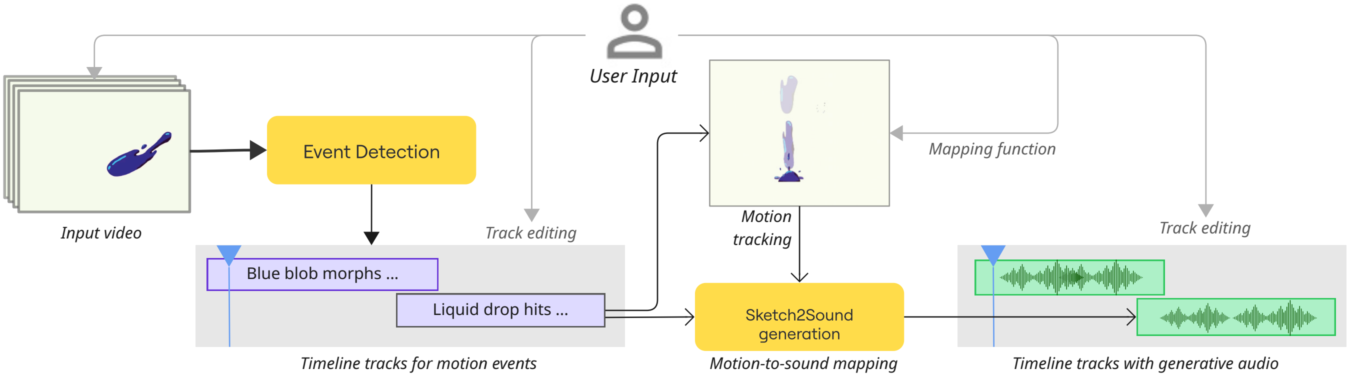
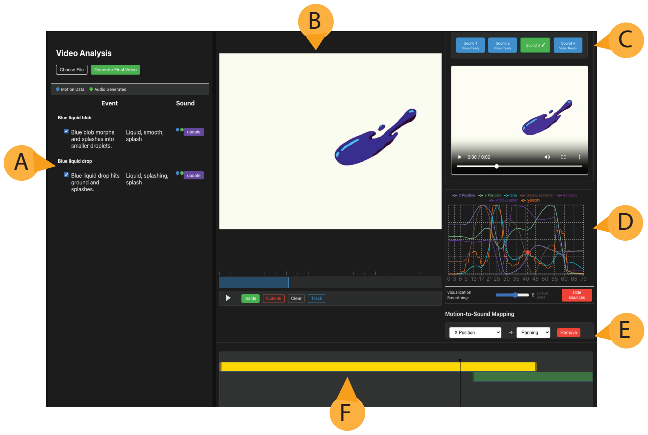


Figure 2: Overview of the MoSound pipeline. The system processes the input video through event detection and motion analysis, enables user-driven motion-to-sound mapping through control curves, and synthesizes sound effects via Sketch2Sound before assembling them into a layered timeline. Water video © Alejandro Imondi.



Event	Sound
Blue liquid blob	
☑ Blue blob morphs and splashes into smaller droplets.	Liquid, smooth, splash
Blue liquid drop	
☑ Blue liquid drop hits ground and splashes.	Liquid, splashing, splash

Figure 3: The MoSound user interface. Automatically detected events are shown along with suggested sound descriptions in the Event Panel (A). The Timeline View (F) shows the chosen events. The Motion Analysis Panel allows users to track objects (B), generate sounds (C), and choose motion mappings (D,E). Water video © Alejandro Imondi.

Previewing the guide sound is useful for immediate feedback, since re-generating the effect sounds takes nearly 10 seconds. The guide sound is provided as a prompt along with the descriptions for the sound generation model.

Finally, MoSound combines the individual effect sounds to form the soundtrack for the entire video. At any point, the user can play or export the entire video with its soundtrack.

4.2 Event Detection

The video is first downsampled to either 10 Hz or one-third of the original frame rate, whichever is lower, and every 20 consecutive frames are grouped into a single batch for analysis by a visual-language model (GPT-4o). This batching strategy allows the model to focus on isolated motion segments and extract sound-related events more accurately.

For every batch analysis, the system prompt positions the model as an expert in motion graphics and sound design. The user prompt provides the frame window and instructs the model to identify key transitions only—ignoring frames without motion—and to describe each event with precise frame indices, object name, motion description, and a structured sound effect (material/texture, style/adjective, event). Each new prompt includes all previous model responses to maintain temporal context across clips. The complete prompts are provided in the supplemental materials

After processing all clips, a summarization prompt instructs the model to merge events across clips, enforce consistency in object naming and formatting, and produce a complete chronological list of sound-mapped events using the same structured format. These results populate the interface's **Event Panel**. Formally, each detected event consists of:

- **Start and end frames** (defining the temporal window),
- **Object name** (e.g., *Blue blob*"),
- **Motion description** (e.g., *Squishes and flattens*"),
- **Sound description**, which combines:
  - **Material/Texture:** Auditory or physical quality of the object (e.g., *liquid*", *metallic*"),
  - **Style/Adjectives:** Tone or energy qualifiers (e.g., *smooth*", *sharp*"),
  - **Event:** A succinct sound action or onomatopoeia (e.g., *whoosh*", *pop*").

4.3 Motion Tracking

To generate a sound that aligns precisely with visual motion, MoSound constructs a motion-consistent *guide sound*—a synthesized audio signal shaped directly by object dynamics in the video. The process involves three key stages: (1) user-guided segmentation and tracking to isolate the object of interest, (2) motion feature extraction from the tracked object masks, and (3) user-defined mapping of these motion features to sound properties. This guide sound acts as a dynamic control signal for sound generation, ensuring both semantic and temporal consistency between audio and motion.

First, users activate motion tracking for a selected event on the timeline. A short video clip is extracted based on the event's start and end frames. Then users could annotate one or more frames by clicking on points inside (and optionally outside) the object of interest. These annotations are passed to the backend segmentation module, Segment Anything Model 2 (SAM 2) [43]. As immediate feedback, the clicked frames are segmented to allow users to verify the annotation quality. Once satisfied, users trigger full-clip segmentation, enabling SAM 2 to propagate the mask across all frames for reliable object tracking.

Then we extract motion features from the sequence of segmentation masks produced by SAM 2. These features are computed on a per-frame basis:

- **Centroid position and velocity magnitude**
- **Rotation magnitude**, estimated from changes in the mask's principal components
- **Object size**, derived from the square root of the average eigenvalue of the mask's pixel covariance matrix
- **Distance and angle to the frame center**

Each feature is smoothed using a configurable moving average filter and normalized to a $[0, 1]$ range. The resulting motion curves are displayed in the **Motion Analysis Panel**, where users can preview and tune the extracted dynamics for subsequent sound mapping.

These curves are combined with the user's chosen sound property mappings (e.g., volume or stereo panning) to create a guide sound for sound effect synthesis. The guide sound captures the timing and dynamics of the motion and is passed as a conditioning input to the generative model.

4.4 Sound Effect Synthesis

In our early experiments, we tested simpler generative approaches that produced continuous tonal or music-like audio. While these methods could generate background textures, they often failed to produce convincing event-based effects such as footsteps. In particular, discontinuous or transient sounds were difficult to align with the precise actions in a scene, leading to mismatches between the visuals and the audio.

To overcome these limitations, we adopted Sketch2Sound [22], a state-of-the-art text-to-audio model that integrates text prompts with interpretable, time-varying control signals such as loudness, brightness, and pitch. In Sketch2Sound, these temporal dynamics take the form of a time-aligned control waveform that specifies how properties such as loudness and spatial balance evolve throughout the clip, enabling the model to condition audio generation on this waveform so that sound events naturally follow the timing and structure of visual motion. This makes it especially effective for MoSound 's needs, where users often want short, temporally precise effects (e.g., drops, impacts) rather than long continuous sound textures in motion graphics. Moreover, Sketch2Sound is lightweight to integrate and supports flexible degrees of temporal precision. We apply Sketch2Sound directly without architectural modifications and provide it with the following control parameters (prompts):

- The structured sound and event description (Sec. 4.2).
- The duration of the clip.
- (Optional) Guide sound.

Sketch2Sound outputs four candidate effect sounds. The sounds are baked together with the video clips to allow direct comparison in the interface.

Guide Sound. The amplitudes of motion curves (Sec. 4.3) are transformed to become the time-varying control signals for Sketch2Sound. Sketch2Sound accepts a stereo waveform that conditions the audio output based on each channel's loudness. MoSound combines multiple features (e.g., velocity and x-position) to be combined into that curve, enabling richer mappings while keeping the interface lightweight.

The smoothed and normalized motion curves are resampled to match the audio rate and used to shape a synthesized waveform. For example, mapping a rising velocity to volume results in a sound that becomes perceptibly louder as the object speeds up; similarly, mapping the x-position to stereo panning creates a spatial effect where the sound moves between the left and right channels. To avoid artifacts, the first and last 0.2 seconds of the waveform are tapered with linear fade-in and fade-out ramps.

5 Technical Evaluation

We have used MoSound to create sound effects for a variety of motion graphics. Examples are visualized in Figures 1 and 4, which took us between 10-40 minutes each to create. The full videos are provided in the supplemental materials.

5.1 Running Time and Clip Length Scalability

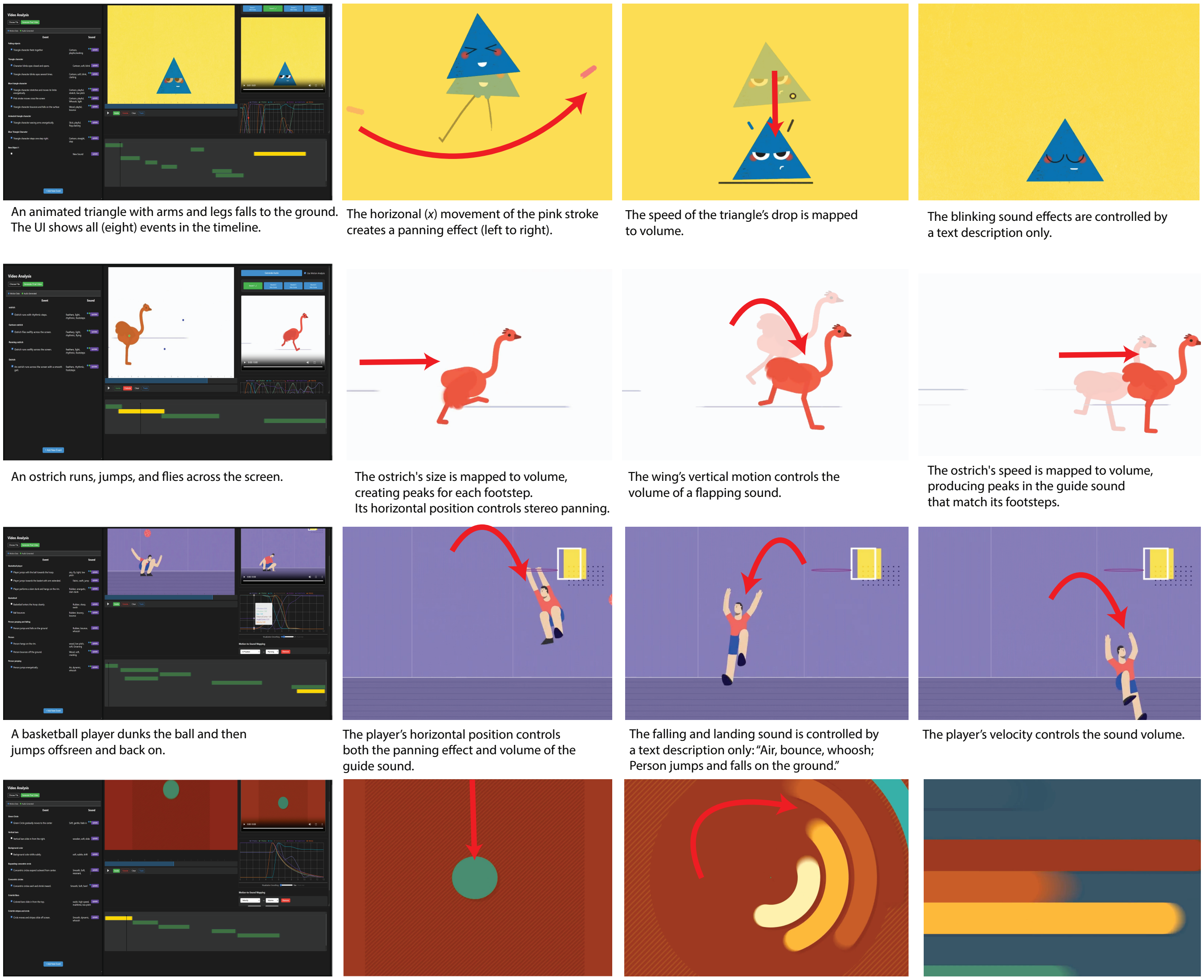
Event Detection takes approximately 5 seconds to process each second of input video. Motion tracking via SAM 2 takes approximately 9 seconds of processing per second of video in the event. Sound synthesis typically takes 10 seconds to generate four candidate effect sounds for a one-second video. All other operations in MoSound take negligible time. This includes generating guide sounds, playing or exporting the effect sounds along with the video, and handling UI events.

The sound-synthesis model does not impose a strict architectural limit on duration; however, empirical evidence from public examples and our own observations indicates that high-quality, temporally coherent generation is most reliable for clips up to roughly 30 seconds. Beyond this range, synthesized audio may gradually exhibit temporal drift or reduced consistency. As a result, MoSound is most effective for short motion-graphics sequences, which aligns with common practice in professional workflows where animations are brief and event-dense.

5.2 Comparison and Ablation with Generative Sound Techniques

For comparison, we ran five recent generative sound techniques on our motion graphics examples, namely FoleyCrafter [68], MMAudio [12], LOVA [13], Soundify [35], and MultiFoley [11]. With a generic prompt ("Add sound effects to this motion graphic video"), these approaches produce silence, incoherent babbling, or noise. The generated audio can be found in the supplemental materials. These degenerate outputs do not merit a quantitative evaluation.

As an ablation, we combined MoSound's analysis with other approaches' sound synthesis. We generated specific prompts by



An animated triangle with arms and legs falls to the ground. The UI shows all (eight) events in the timeline.

The horizonal (*x*) movement of the pink stroke creates a panning effect (left to right).

The speed of the triangle's drop is mapped to volume.

The blinking sound effects are controlled by a text description only.

An ostrich runs, jumps, and flies across the screen.

The ostrich's size is mapped to volume, creating peaks for each footstep. Its horizontal position controls stereo panning.

The wing's vertical motion controls the volume of a flapping sound.

The ostrich's speed is mapped to volume, producing peaks in the guide sound that match its footsteps.

A basketball player dunks the ball and then jumps offsreen and back on.

The player's horizontal position controls both the panning effect and volume of the guide sound.

The falling and landing sound is controlled by a text description only: "Air, bounce, whoosh; Person jumps and falls on the ground."

The player's velocity controls the sound volume.

Abstract shapes expand, spin, and move.

The velocity of the green circle controls the volume of the guide sound, which leads to a drum sound effect.

The sound effect is controlled only by a text description: "Smooth, Soft, Swirl; Concentric circles swirl and shrink inward."

The sound effect is controlled only by a text description: "Smooth, dynamic, whoosh; Circle moves and stripes slide off screen."

Figure 4: Sounds effects created with MoSound for motion graphics (top to bottom: Triangle, Ostrich, Basketball, Color). Automatic generative sound techniques produce silence, incoherent babbling, or noise. Triangle/Ostrich/Basketball videos © Alejandro Imondi. Color video © Enchanted Studios (Adobe Stock asset #217867437).

using MoSound to generate a list of events and suggested sounds and summarized them with ChatGPT. For some videos, MMAudio (and sometimes MultiFoley) can produce results comparable to MoSound, but only when provided with precise and well-crafted prompts—something that MoSound is designed to generate automatically. (The other approaches generated incoherent output.) All competing methods struggled with abstract motion graphics that lack recognizable objects, such as scenes based purely on color or shape transformations (e.g., Figure 4, 4th row). MoSound can still generate creative and meaningful sound effects in these cases. Other systems also lack control over motion-driven cues like steps or drops, which MoSound handles by directly mapping visual motion to sound properties.

This ablation motivates our choice of Sketch2Sound, whose high-quality output and controllability meet our requirements.

We considered the applicability of recent video (with audio) synthesis approaches, OpenAI's Sora 2 [42] and Google's Veo 3 [24]. However, their APIs do not allow providing an input video and generating only a soundtrack. As input, they take a text description and optional first or last frames. To test applicability, we ran Sora 2 Pro with the first frame from a motion graphic video (the triangle from Figure 4), a text description of the entire video (generated by MoSound), and event descriptions, suggested sound effects, and precise start times and durations. Sora 2 generated a new video with different content than the original. As the events are not the same, the generated soundtrack cannot be used for the original video. This provides further motivation for MoSound.

5.3 Guide Sound and Final Sound Consistency

We quantitatively evaluated how well the final sound effects follow the temporal and spatial structure specified by the guide sounds. In particular, we focus on three aspects: (1) the timing of salient events, (2) the overall loudness dynamics, and (3) the left–right panning behavior. We measured the following on the 15 generated sounds in the examples shown in Figures 1 and 4 that used a guide sound. (One guide sound was mono, so the panning evaluation was conducted on 14 generated sounds.)

Across all clips, the median onset deviation between the guide and final sounds is 0.032 seconds (measured via `librosa` [40]). This indicates that the start times of salient events in the final audio remain tightly aligned with those in the guide and are within typical perceptual thresholds for asynchrony perception [18]. The mean (± SD) envelope correlation between the guide and final sounds is 0.536 ± 0.332 (comparing root-mean-square values). This shows a consistent similarity in overall loudness dynamics while still allowing the generative model to introduce local variation driven by text prompts and stochastic sampling. Panning is preserved even more strongly: the mean correlation between inter-channel level difference (ILD) curves is 0.731 ± 0.191, demonstrating that the left–right motion of the final sounds closely follows the panning encoded in the guide. Since the guide sound directly encodes the user-defined stereo panning, the high ILD correlation indicates that the spatialization behavior of the final audio preserves the intended left–right motion.

Taken together, these results show that our guide-based design effectively preserves the motion-derived temporal and spatial structure in the final audio, while still supporting flexible stylistic control through text conditioning.

5.4 Event Detection Accuracy

We quantitatively evaluated event detection accuracy by comparing VLM-generated events to user-edited events. We assess accuracy along two complementary dimensions: semantic similarity and temporal agreement. Together these measures indicate whether the VLM provides conceptually relevant and temporally useful guidance. Because users frequently adjust, merge, or refine the VLM's suggestions, our measurements cannot assume a correspondence between the VLM suggestions and the user's events.

To obtain robust estimates, we used five example videos with user-edited events (Figures 1 and 4). For each example, we repeated our evaluation five times with different VLM suggestions. The reported measures are the averages across these runs.

Semantic similarity. For each user event in an example video, we compute its semantic similarity to all VLM-suggested events (for that run) and take the maximum value (indicating the best match, since we lack a correspondence). We define the semantic similarity between two events (one VLM and one user) as the average similarity of the three attribute descriptions: the main object, event, and suggested sound. Attribute description similarity is defined as the cosine similarity between the text embedded in a latent space (SentenceTransformer model all-MiniLM-L6-v2 [44]). This measurement varies between -1 (worst match) and 1 (perfect match) and reflects whether the VLM produced a conceptually relevant description of the event, independent of when it occurs. Across all 25 generated events, the mean semantic similarity was 0.505 ± 0.141. This indicates that VLM suggestions frequently capture substantial aspects of the event semantics that users ultimately retain, such as the type of motion, the object involved, or the qualitative sound characteristics. The moderate spread of values reflects the diversity of scene structures, spanning from concrete physical interactions to more abstract or stylized animations.

Temporal alignment. To measure broad temporal alignment, we compute the Jaccard similarity (intersection over union or IoU) between the union of all VLM event intervals and the union of all final event intervals. The mean temporal IoU is 0.337 ± 0.128. Although modest, these values are expected given that the events are short (typically 0.2–0.6 seconds), making IoU highly sensitive to even small onset differences. This metric should therefore be interpreted as a coarse indicator of timeline overlap rather than precise temporal accuracy.

Center alignment error. To provide a more interpretable measure of temporal deviation, we compute the absolute distance between each final event's center time and the closest VLM event center. The mean center alignment error is 0.660 ± 0.247 seconds, with most runs falling between 0.5 and 1.0 seconds. This shows that VLM suggestions usually fall in the general neighborhood of the final event locations but are not tightly aligned. This looseness is consistent with the fact that VLMs are not trained for fine grained temporal grounding and that our examples span a wide range of motion dynamics.

Taken together, these results show that VLM suggestions provide semantically meaningful but temporally coarse guidance. Users typically refine timing and specificity during editing, while the VLM helps surface relevant conceptual events and sound characteristics as starting points for the design process.

6 User Study

6.1 Protocol

To evaluate MoSound, we conducted a user study in which three experts (E1–E3) and four novices (N1–N4) used MoSound and provided feedback. The expert users had 22, 12, and 30 years of professional sound design experience, respectively. We began each session by teaching the participants how to use MoSound. Participants were given a step-by-step live demonstration using the example from Figure 1, in which we thoroughly explained each button and the logic of MoSound. We then asked them to create sound effects themselves on the same video. (E2 and E3 felt comfortable skipping this step.) Next, participants chose a video from a set of motion graphic videos we collected with permission from rightsholders and used MoSound to create sound effects. The authors played the role of an interactive help system; participants were told to ask questions as needed. After using MoSound, participants were asked to fill out a questionnaire to provide feedback about the system. The questionnaire included numerical and open-ended questions. We analyzed these qualitative responses using thematic analysis to identify recurring user perceptions about MoSound's workflow and controls.

The survey questionnaire, responses, and user-created videos can be found in the supplemental materials. Our protocol was approved by our institutional ethics board.

6.2 Quantitative Evaluation

All users were able to successfully use MoSound to add sound effects. On average, users took 9 minutes creating sounds effects for the tutorial video and 21 minutes for the free choice video. For the free choice video, users created 3–7 events. Responses to our quantitative Likert questions (Q1–Q7) are visualized in Figure 5. The p-values shown are computed via a t-test against the theoretical mean response of neutral and adjusted with a Holm-Bonferroni family-wise error correction. Users were satisfied and found appealing the quality of generated sounds (Q1, Q2), the events and their descriptions (Q3), the effectiveness of the motion tracking (Q5), and its role in producing motion-synchronized audio (Q6). The weakest aspect of MoSound is the accuracy of automatic event placement on the timeline (Q4).

6.3 Comparison to Professional Tools

Although our study lacked a direct comparison to expert tools, expert participants E1 and E2 estimated that recreating the audio effects manually in their preferred professional tools (e.g., ProTools) would have taken 30 minutes, involving sound sourcing, editing, mixing, and synchronization. In contrast, E1 created the artifact using our system in 5 minutes, which included creative exploration and refinement. E1 commented, *"I was able to get to my [minimum viable product] MVP design faster than with my tool."* E1: *"MoSound makes it easier for people to do the tracking and syncing. In other tools [Tsugi], you have to watch the video and you (manually) draw it. But its hard to be accurate."* E2 worked more slowly and methodically as in his typical workflow, marking every sound opportunity with an event. E2 wished for finer controls throughout the interface, like the ability to zoom and make frame-by-frame adjustments in the timeline, or the ability to edit motion curves directly. E3 tried the same video twice, aiming to explore the full range of possibilities offered by MoSound. E3 highlighted the system's strengths in generating appealing foreground sounds, but wished for a way to create ambient and atmospheric layers. E3 also wished for fine-grained controls (e.g., layer mixing, panning, timeline precision, attack decay sustain release controls). E3 emphasized that he viewed MoSound as an augmentation rather than a replacement of traditional practices, and praised its scalability and potential for professional integration. E3 had particular expertise in vocal Foley and suggested allowing users to record effects.

6.4 Event Detection and Sound Synthesis

Both novices and experts added additional events and modified the event and sound descriptions. Novices did this when the automatic generation omitted objects they wished to consider. E1 created duplicate, overlapping events in order to add an ambient sound. Professional sound designers often create appealing sounds by layering samples from an existing audio library. E2 commented that the generated effect sounds were surprisingly appealing and usable as is. (E2 still recommended changing MoSound to allow choosing more than one candidate effect sound.) E2 appreciated that MoSound suggested events but wished for more accurate timeline placement. E2 created many additional events to cover the scene with individual short-duration clips to ensure that all relevant events were adequately captured and represented. *"Half of the job of a sound designer is go through through every single frame and watch for the things happening."* To enhance audio design, E1 proposed allowing the system greater creative freedom, suggesting it could generate more detailed and nuanced audio elements (e.g., "magical shimmering") that are not explicitly depicted in the video. Beyond simply adding multiple clips, E3 emphasized that professional sound design depends on careful layering and mixing. E3 questioned: *"Can I explicitly control the loudness mix? Can I synthesize a combined sound? A lot of times we listen back 100 times to get the balance."* This highlights that while MoSound currently supports event addition and duplication, experts expect more fine-grained control over dynamic mixing parameters. N1 wish for multi-object tracking and object identity detection across clips for consistent sound synthesis.

6.5 Event and Sound Descriptions

The suggested event and sound descriptions were deemed appropriate by all users (Q3). This functionality sped up the creative process and facilitated exploration, especially for novice users who often find describing sound to be difficult. E1 commented, *"You don't know how to describe what you're seeing . . . a lot of things are really abstract. So being able to have like a video understanding is really helpful."* and *"sometimes it's really hard to find, to describe the sounds in the library."*. E3 noted that *"video analysis and custom user events is a really powerful idea"* and emphasized the need for automatic ways to evaluate and refine generated outputs, or to direct the UI to generate an automatic description for a manually selected time range.

6.6 Motion Tracking

Study participants had an overall favorable opinion of MoSound's object tracking (Q5). All participants had a favorable opinion of motion synchronization (Q6). Novice users were able to synchronize their sounds to motion, a fairly advanced and tedious endeavor. E2 wished for the ability to edit the motion curves directly. E3 strongly valued the automatic motion-tracking-to-audio automation feature, calling it a *"million dollar idea"* and suggesting it could stand alone as a plugin for professional software.

6.7 Creative Control

Participants commented on their sense of creative control and agency, even with the available automation and defaults. N1 stated that MoSound *"make[s] me feels like a powerful designer"*. N3 remarked that *"there [are] a lot of options to make adjustment[s] for the sound effect"*. E1: *"I felt that I have full creative control over it. (The system) just made it easier for me, but I still was able to make decisions on what type of sound and with how I wanted to move, and focus on the details. You unleash more creative power to me because it (the default) already sounds good and then it was all about enhancing and*

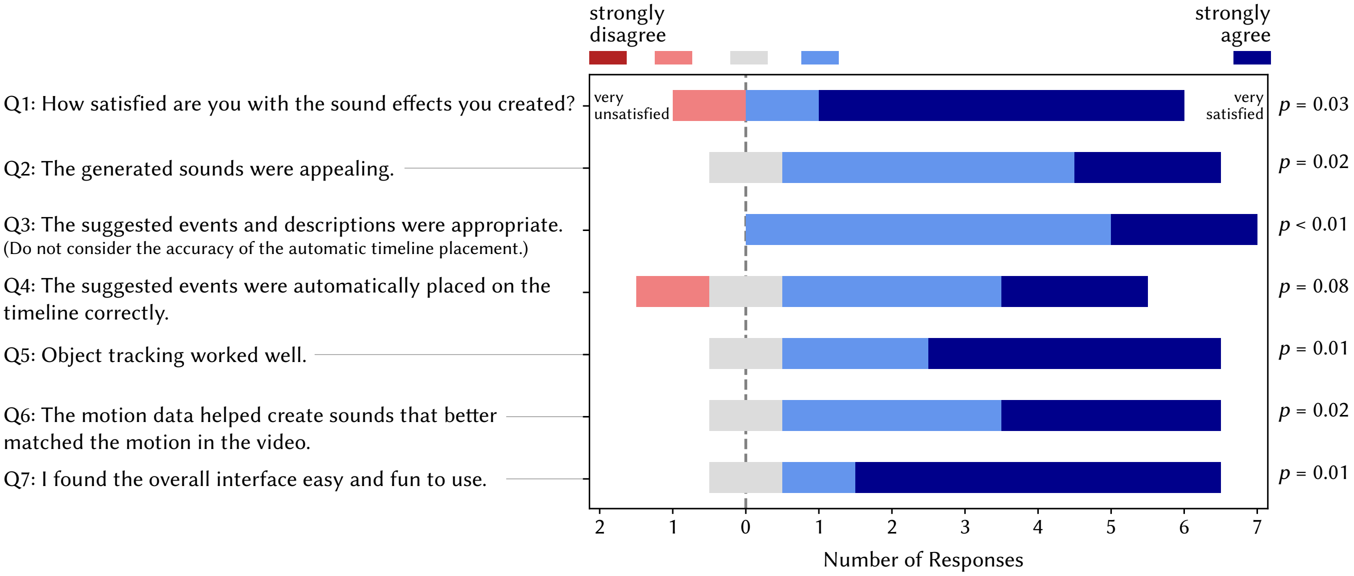


Figure 5: Quantitative Likert questions from our user study (Section 6). p values were adjusted with a Holm-Bonferroni family-wise error correction.

adding details." E2 suggested that MoSound is more appropriate for users with medium expertise, *"but not for a really professional level like Hollywood movies"*. While N2 appreciated how MoSound transforms a complex video into multiple manageable segments for sound design, E3 expressed a preference for maintaining long, continuous control across the entire sequence, reflecting different creative approaches to structuring audio.

7 Discussion and Implications for Design

MoSound shows how automatic event detection, motion curve extraction, and generative audio synthesis can be combined into a single workflow that lets users control sound timing and dynamics directly from the video. This integration, rather than any individual component, gives users a level of synchronized and editable motion driven control that existing text-only or one-shot workflows do not support. Novices treated MoSound as a replacement for traditional workflows, relying on its end-to-end automation to produce soundtracks without prior expertise. Professionals, by contrast, approached it as an augmentation, valuing the motion-tracking and curve-extraction features but also identifying gaps around layering, precision, and multimodal control that point toward future opportunities.

7.1 Generative Audio tools as Replacement or Augmentation

Our findings indicate that MoSound plays different roles depending on user expertise. For novices, the system acts as a practical replacement for traditional workflows, not because it removes complicated control but because it offers a guided, semi-automatic process that remains editable at every step. Instead of one-shot generation, MoSound surfaces event candidates, proposes sound descriptions, and produces motion-aligned audio that users can refine, adjust, or regenerate. This scaffolded pipeline enabled beginners to create synchronized soundtracks without requiring prior technical or artistic knowledge, while still keeping them in control of timing, content, and stylistic choices.

For professionals, however, MoSound was less suited as a replacement, since expert workflows rely on fine-grained layering, mixing, and integration with existing tools. This is in line with recent findings for generative audio tools [32]. Participants valued the system's automatic motion tracking and curve extraction as a useful augmentation, but also noted the lack of detailed controls such as track-based loudness balancing, spatial panning, and envelope shaping (e.g., ADSR). While effective for short, event-based sounds, MoSound was less suited for continuous textures or complex layering. These gaps reflect a broader limitation of current AI models, which are optimized for high-level semantic generation rather than the low-level parametric control demanded in professional practice.

Together, these divergent patterns illustrate a broader implication for generative sound and motion tools: the same system can simultaneously act as a point of entry for beginners and as an automation layer for experts. Designing for this dual role requires balancing end-to-end scaffolding with opportunities for expert control. This insight could be transferable to the development of future AI-assisted creative tools across domains where user experience varies widely.

7.2 Expanding Motion Features Beyond Current Parameters

MoSound currently maps a limited set of motion descriptors (position, velocity, size, and rotation) into sound parameters. While these proved sufficient for generating basic time-varying audio control signals , our findings point to opportunities for expanding the parameter space in future systems. In particular, incorporating physics-inspired descriptors such as collisions, tension, or spring dynamics could enable more expressive and intuitive mappings. For example, the sharp increase in pitch during a car braking event is difficult to express through position or velocity curves alone, but can be naturally described by a rising tension parameter. The importance of timbre variations is widely recognized [60], and can be captured by simulation methods [64, 70]. However, such methods are specific to certain scenarios, and do not apply to general motion graphics. Also, these parameters may be especially valuable for animation or film sound design, where the physical dynamics of motion influence the viewer's sense of perceptual plausibility. Sound synthesis would need to be adapted to consider additional properties of the guide sound. Automatically detecting and quantifying such properties over the length of the video would require new visual understanding models.

Sound designers in animation, film, and games often describe motion using concepts such as impact strength or tension buildup, even though current tools do not explicitly represent these cues. If future audio systems could capture such higher-level motion semantics, either by detecting them automatically or by letting designers annotate them, then generated sound could better reflect how professionals think about timing and expressive emphasis. Building this intermediate layer between geometric motion and sound synthesis would allow tools to interpret motion in terms of meaningful patterns, supporting workflows that are more consistent and transferable across creative domains without increasing interface complexity.

7.3 Multimodal Workflows

MoSound demonstrates the potential of multimodal authoring by combining motion features with text prompts to generate synchronized audio. This approach already provided users with a new expressive channel, where visual motion could guide time-varying audio control signals of the generated audio while text captured semantic intent. In current practice in animation, film, and games, sound designers rarely describe what they want purely in words or numeric parameters. Instead, they rely on vocal sketches [22] or vocal Foley (E3), directly drawing on the screen or uploading reference sounds [3, 16, 57], or importing audio tracks. These examples are not just convenient workarounds; they are a primary way professionals encode timbre, pacing, and style. This suggests that in audio design, creative intent is naturally expressed as a mixture of voice, movement, and concrete sound examples rather than as text alone.

For generative audio systems, this reframes what a "prompt" should be. Systems that only accept textual descriptions or abstract controls risk ignoring the forms of expression that practitioners already use to think, communicate, and iterate about sound. Treating vocal sketches, reference sounds, and other example-based inputs as first-class prompts would bring generative methods closer to established workflows in animation and film sound design. It would also open up a design space where models are trained and evaluated not only on output quality, but on how well they respond to the multimodal traces that designers naturally produce when working with sound.

7.4 Events and Sound Description Accuracy

Our current pipeline relies on GPT-4o for generating event and sound descriptions. While this approach provided a baseline for semantic alignment, the outputs were often imprecise. Descriptions lacked nuanced detail, temporal boundaries were approximate and mismatched, and events were treated as isolated occurrences rather than part of a continuous scene. Moreover, the generated descriptions tended to emphasize salient objects rather than identifying the core object driving the action.

This observation points toward a broader design implication for AI-assisted creative tools: when models generate semantic structures such as events, annotations, or timelines, these outputs must remain legible, editable, and available at multiple levels of granularity. Providing clearer temporal grounding and more precise object references would improve accuracy, while structured but flexible representations would allow users to refine model suggestions without discarding them.

Such principles generalize beyond sound design. Systems for animation authoring, video editing, and multimodal content creation increasingly depend on model-generated intermediate representations. Ensuring that these representations are transparent, correctable, and incrementally refinable is essential for integrating generative models into creative workflows where precision and interpretability matter.

7.5 Intellectual Property Concerns

Intellectual property concerns pose not only legal questions but also structural barriers to adopting MoSound in professional settings. As U.S. Copyright Office guidance notes, AI-generated audio may not qualify for copyright protection [58], and training on unlicensed datasets can expose users to liability if generated outputs resemble copyrighted material [59]. When the provenance of model-generated audio is opaque, users cannot determine whether the output is copyrightable, whether it unintentionally resembles items in the training set, or whether it can be safely deployed in commercial pipelines. This uncertainty limits the integration of generative audio tools regardless of their technical capability.

To address these barriers, future systems need mechanisms that reduce legal ambiguity at its source. One direction is to train or fine-tune models exclusively on open-licensed or user-owned sound libraries, making the provenance of generated material auditable. Approaches such as Datamind Audio's Concatenator [16] demonstrate the feasibility of workflows that restrict models to user-provided datasets, ensuring that generated sounds inherit clear licensing status. Allowing users to select a "transparent-data" model would give creators greater confidence that outputs can be cleared for commercial use.

A complementary direction is to incorporate model-side safeguards that prevent close reproduction of training examples. Techniques such as dataset de-duplication [8] or embedding-level repulsion [21] that steers generation away from copyrighted clusters can further reduce the risk of memorized patterns. These approaches proactively reduce memorization risk within the model itself, helping ensure that generated audio does not inadvertently mirror copyrighted material.

Although our focus is on sound effects, these strategies generalize to AI systems that generate images, animation, or video. As generative models increasingly support creators of all kinds building copyright-aware training regimes and copyright-safe generation modes will be essential for enabling broader and safer adoption in everyday creative workflows.

8 Conclusion

We have presented MoSound, an interactive tool for enriching motion graphics with sound effects. MoSound assists users with all stages of sound design. It makes use of a VLM to automatically generate events in the form of event and sound descriptions placed on the timeline. MoSound allows users to track objects and provide precise sound guidance based on motion features. Finally, MoSound leverages a state-of-the-art sound synthesis technique to create effect sounds for the motion graphics sound track. Beyond this pipeline, our work also clarifies how motion cues in motion graphics can meaningfully guide generative sound, outlining a motion-to-sound design space specific to this domain.

Our study shows that MoSound lowers barriers for novices by automating end-to-end workflows, while offering experts a novel prototyping tool and augmentations in the form of motion-driven automation data. Our observations clarify how different users engage with generative workflows and inform how future systems might better align AI capabilities with real production practices.

Building on MoSound's strength in generating short, event-based effects, future work should explore continuous and consistent textures and sound identity, fine-grained layering, more precise video analysis, richer motion parameters, and multimodal input. An important goal for future work is to better bridge AI-assisted automation with expert practice, combining the flexibility of AI analysis and generation with the precision, personalization, and workflow integration required by professionals when creating production-ready soundtracks.

Acknowledgments

We are grateful to our reviewers for their time and helpful suggestions. We are also grateful to the designers whose input helped us shape and understand our work. Authors Huang and Gingold were supported in part by the United States National Science Foundation (IIS-2402893) and a gift from Adobe Systems Inc.

References

[1] Adobe. 2025. Adobe After Effects. https://www.adobe.com/products/aftereffects.html

[2] Adobe. 2025. Adobe Express. https://www.adobe.com/express/

[3] Adobe. 2025. Firefly Generate Sound Effects. https://firefly.adobe.com/generate/sound-effects

[4] Kat R Agres, Adyasha Dash, and Phoebe Chua. 2023. AffectMachine-Classical: a novel system for generating affective classical music. *Frontiers in Psychology* 14 (2023), 1158172.

[5] Vanessa Theme Ament. 2014. *The Foley grail: The art of performing sound for film, games, and animation.* Routledge.

[6] Steven S. An, Doug L. James, and Steve Marschner. 2012. Motion-driven concatenative synthesis of cloth sounds. *ACM Trans. Graph.* 31, 4, Article 102 (July 2012), 10 pages. doi:10.1145/2185520.2185598

[7] Canva. 2025. Canva. https://www.canva.com/

[8] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)*. 5253–5270.

[9] Anil Çamcı, Kristine Lee, Cody J. Roberts, and Angus G. Forbes. 2017. INVISO: A Cross-platform User Interface for Creating Virtual Sonic Environments. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (Québec City, QC, Canada) *(UIST '17)*. Association for Computing Machinery, New York, NY, USA, 507–518. doi:10.1145/3126594.3126644

[10] Jeffrey N. Chadwick, Changxi Zheng, and Doug L. James. 2012. Precomputed acceleration noise for improved rigid-body sound. *ACM Trans. Graph.* 31, 4, Article 103 (July 2012), 9 pages. doi:10.1145/2185520.2185599

[11] Ziyang Chen, Prem Seetharaman, Bryan Russell, Oriol Nieto, David Bourgin, Andrew Owens, and Justin Salamon. 2025. Video-Guided Foley Sound Generation with Multimodal Controls. *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR).*

[12] Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. 2025. MMAudio: Taming Multimodal Joint Training for High-Quality Video-to-Audio Synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference.* 28901–28911.

[13] Xin Cheng, Xihua Wang, Yihan Wu, Yuyue Wang, and Ruihua Song. 2024. LoVA: Long-form Video-to-Audio Generation. *arXiv preprint arXiv:2409.15157* (2024).

[14] Michel Chion. 1994. *Audio-Vision: Sound on Screen.* Columbia University Press, New York, NY.

[15] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and controllable music generation. *Advances in Neural Information Processing Systems* 36 (2023), 47704–47720.

[16] Datamind Audio. 2025. Concatenator. https://datamindaudio.ai

[17] Ninon Devis, Nils Demerlé, Sarah Nabi, David Genova, and Philippe Esling. 2023. Continuous descriptor-based control for deep audio synthesis. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 1–5.

[18] Norman F Dixon and Lydia Spitz. 1980. The Detection of Auditory Visual Desynchrony. *Perception* 9, 6 (Dec. 1980), 719–721. doi:10.1068/p090719

[19] H Flores Garcia, P Seetharaman, R Kumar, and B Pardo. 2023. VampNet: Music Generation via Masked Acoustic Token Modeling. 24th International Society for Music Information Retrieval Conference.

[20] Jules Françoise, Olivier Chapuis, Sylvain Hanneton, and Frédéric Bevilacqua. 2016. SoundGuides: Adapting Continuous Auditory Feedback to Users. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 2829–2836. doi:10.1145/2851581.2892420

[21] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. 2023. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision.* 2426–2436.

[22] Hugo Flores García, Oriol Nieto, Justin Salamon, Bryan Pardo, and Prem Seetharaman. 2025. Sketch2Sound: Controllable Audio Generation via Time-Varying Signals and Sonic Imitations. arXiv:2412.08550 [cs.SD] https://arxiv.org/abs/2412.08550

[23] Sanchita Ghose and John Jeffrey Prevost. 2021. AutoFoley: Artificial Synthesis of Synchronized Sound Tracks for Silent Videos With Deep Learning. *IEEE Transactions on Multimedia* 23 (2021), 1895–1907. doi:10.1109/TMM.2020.3005033

[24] Google. 2025. Veo 3. https://aistudio.google.com/models/veo-3

[25] Chris Harrison, Gary Hsieh, Karl D.D. Willis, Jodi Forlizzi, and Scott E. Hudson. 2011. Kineticons: using iconographic motion in graphical user interface design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 1999–2008. doi:10.1145/1978942.1979232

[26] Leona M Holloway, Cagatay Goncu, Alon Ilsar, Matthew Butler, and Kim Marriott. 2022. Infosonics: Accessible Infographics for People who are Blind using Sonification and Voice. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 480, 13 pages. doi:10.1145/3491102.3517465

[27] Jialin Huang, Alexa Siu, Rana Hanocka, and Yotam Gingold. 2023. ShapeSonic: Sonifying Fingertip Interactions for Non-Visual Virtual Shape Perception. In *SIGGRAPH Asia 2023 Conference Papers* (Sydney, NSW, Australia) *(SA '23)*. Association for Computing Machinery, New York, NY, USA, Article 80, 9 pages. doi:10.1145/3610548.3618246

[28] Amir Jahanlou and Parmit K Chilana. 2022. Katika: An End-to-End System for Authoring Amateur Explainer Motion Graphics Videos. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 502, 14 pages. doi:10.1145/3491102.3517741

[29] Amir Jahanlou and Parmit K Chilana. 2024. How Example-Based Authoring of Motion Graphics Impacts Creative Expression: Differences in Perceptions of Professional and Casual Motion Designers. In *Proceedings of the 16th Conference on Creativity & Cognition* (Chicago, IL, USA) *(C&C '24)*. Association for Computing Machinery, New York, NY, USA, 347–357. doi:10.1145/3635636.3656197

[30] Amir Jahanlou, William Odom, and Parmit Chilana. 2021. Challenges in Getting Started in Motion Graphic Design: Perspectives from Casual and Professional Motion Designers. In *Proceedings of Graphics Interface 2021* (Virtual Event) *(GI 2021)*. Canadian Information Processing Society, 35 – 45. doi:10.20380/GI2021.06

[31] Patrik N Juslin and Daniel Västfjäll. 2008. Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and brain sciences* 31, 5 (2008), 559–575.

[32] Purnima Kamath, Fabio Morreale, Priambudi Lintang Bagaskara, Yize Wei, and Suranga Nanayakkara. 2024. Sound Designer-Generative AI Interactions: Towards Designing Creative Support Tools for Professional Sound Designers. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 730, 17 pages. doi:10.1145/3613904.3642040

[33] 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.

[34] Timothy R. Langlois, Changxi Zheng, and Doug L. James. 2016. Toward Animating Water with Complex Acoustic Bubbles. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2016)* 35, 4 (July 2016). doi:10.1145/2897824.2925904

[35] David Chuan-En Lin, Anastasis Germanidis, Cristóbal Valenzuela, Yining Shi, and Nikolas Martelaro. 2023. Soundify: Matching Sound Effects to Video. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (, San Francisco, CA, USA,) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 18, 13 pages. doi:10.1145/3586183.3606823

[36] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. In *International Conference on Machine Learning*. PMLR, 21450–21474.

[37] Vivian Liu, Rubaiat Habib Kazi, Li-Yi Wei, Matthew Fisher, Timothy Langlois, Seth Walker, and Lydia Chilton. 2025. LogoMotion: Visually-Grounded Code Synthesis for Creating and Editing Animation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 157, 16 pages. doi:10.1145/3706598.3714155

[38] Jiaju Ma and Maneesh Agrawala. 2025. MoVer: Motion Verification for Motion Graphics Animations. *ACM Transactions on Graphics* 44, 4 (Aug. 2025). doi:10.1145/3731209.

[39] Jiaju Ma, Li-Yi Wei, and Rubaiat Habib Kazi. 2022. A Layered Authoring Tool for Stylized 3D animations. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 383, 14 pages. doi:10.1145/3491102.3501894

[40] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science conference*, Vol. 8.

[41] Zachary Novack, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J Bryan. 2024. DITTO: Diffusion inference-time T-optimization for music generation. In *Proceedings of the 41st International Conference on Machine Learning*. 38426–38447.

[42] OpenAI. 2025. Sora 2. https://openai.com/index/sora-2/

[43] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. 2024. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714* (2024). https://arxiv.org/abs/2408.00714

[44] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[45] Davide Rocchesso, Stefania Serafin, Frauke Behrendt, Nicola Bernardini, Roberto Bresin, Gerhard Eckel, Karmen Franinovic, Thomas Hermann, Sandra Pauletto, Patrick Susini, and Yon Visell. 2008. Sonic interaction design: sound, information and experience. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems* (Florence, Italy) *(CHI EA '08)*. Association for Computing Machinery, New York, NY, USA, 3969–3972. doi:10.1145/1358628.1358969

[46] Xin Wei Sha, Adrian Freed, and Navid Navab. 2013. Sound design as human matter interaction. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (Paris, France) *(CHI EA '13)*. Association for Computing Machinery, New York, NY, USA, 2009–2018. doi:10.1145/2468356.2468718

[47] Ladan Shams and Aaron R Seitz. 2008. Benefits of multisensory learning. *Trends in cognitive sciences* 12, 11 (2008), 411–417.

[48] Xinyu Shi, Yinghou Wang, Yun Wang, and Jian Zhao. 2024. Piet: Facilitating Color Authoring for Motion Graphics Video. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 148, 17 pages. doi:10.1145/3613904.3642711

[49] Yang Shi, Xingyu Lan, Jingwen Li, Zhaorui Li, and Nan Cao. 2021. Communicating with Motion: A Design Space for Animated Visual Narratives in Data Videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 605, 13 pages. doi:10.1145/3411764.3445337

[50] Alexa Siu, Gene S-H Kim, Sile O'Modhrain, and Sean Follmer. 2022. Supporting Accessible Data Visualization Through Audio Data Narratives. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 476, 19 pages. doi:10.1145/3491102.3517678

[51] Xia Su, Jon E. Froehlich, Eunyee Koh, and Chang Xiao. 2024. SonifyAR: Context-Aware Sound Generation in Augmented Reality. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 128, 13 pages. doi:10.1145/3654777.3676406

[52] Dallas Taylor, Kenny Malone, Casey Emmerling, Jess Jiang, and James Sneed. 2025. TikTok's Trojan Horse Strategy. https://www.npr.org/2025/10/22/nx-s1-5582749/tiktok-massive-music-sonic-brand-sound-indentity-logo

[53] Renaud Bougueng Tchemeube, Jeff Ens, Cale Plut, Philippe Pasquier, Maryam Safi, Yvan Grabit, and Jean-Baptiste Rolland. 2025. Evaluating human-AI interaction via usability, user experience and acceptance measures for MMM-c: A creative AI system for music composition. *arXiv preprint arXiv:2504.14071* (2025).

[54] John R Thompson, Zhicheng Liu, and John Stasko. 2021. Data Animator: Authoring Expressive Animated Data Graphics. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 15, 18 pages. doi:10.1145/3411764.3445747

[55] Zeyue Tian, Yizhu Jin, Zhaoyang Liu, Ruibin Yuan, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. 2025. AudioX: Diffusion Transformer for Anything-to-Audio Generation. arXiv:2503.10522 [cs.MM] https://arxiv.org/abs/2503.10522

[56] Tiffany Tseng, Ruijia Cheng, and Jeffrey Nichols. 2024. Keyframer: Empowering animation design using large language models. *arXiv preprint arXiv:2402.06071* (2024).

[57] Tsugi GK. 2020. DSP Motion. https://tsugi-studio.com/web/en/products-dspmotion.html

[58] U.S. Copyright Office. 2025. *Copyright and Artificial Intelligence, Part 2: Copyrightability*. Technical Report. U.S. Copyright Office.

[59] U.S. Copyright Office. 2025. *Copyright and Artificial Intelligence, Part 3: Generative AI Training*. Technical Report. U.S. Copyright Office.

[60] Kees van den Doel and Dinesh K. Pai. 1998. The Sounds of Physical Shapes. *Presence* 7, 4 (1998), 382–395.

[61] Bruce N. Walker and Gregory Kramer. 2005. Mappings and metaphors in auditory displays: An experimental assessment. *ACM Trans. Appl. Percept.* 2, 4 (Oct. 2005), 407–412. doi:10.1145/1101530.1101534

[62] R. Wang, C. Jung, and Y. Kim. 2022. Seeing Through Sounds: Mapping Auditory Dimensions to Data and Charts for People with Visual Impairments. *Computer Graphics Forum* 41, 3 (2022), 71–83. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14523 doi:10.1111/cgf.14523

[63] Zeyu Xie, Xuenan Xu, Zhizheng Wu, and Mengyue Wu. 2024. Picoaudio: Enabling precise timestamp and frequency controllability of audio events in text-to-audio generation. *arXiv preprint arXiv:2407.02869* (2024).

[64] Kangrui Xue, Ryan M Aronson, Jui-Hsien Wang, Timothy R Langlois, and Doug L James. 2023. Improved Water Sound Synthesis using Coupled Bubbles. *ACM Trans. Graph.* 42, 4 (aug 2023).

[65] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. 2023. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), 1720–1733.

[66] Hui Ye, Chufeng Xiao, Jiaye Leng, Pengfei Xu, and Hongbo Fu. 2025. MoGraphGPT: Creating Interactive Scenes Using Modular LLM and Graphical Control. arXiv:2502.04983 [cs.HC] https://arxiv.org/abs/2502.04983

[67] Sharon Zhang, Jiaju Ma, Jiajun Wu, Daniel Ritchie, and Maneesh Agrawala. 2023. Editing Motion Graphics Video via Motion Vectorization and Transformation. *ACM Trans. Graph.* 42, 6, Article 229 (Dec. 2023), 13 pages. doi:10.1145/3618316

[68] Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, and Kai Chen. 2024. FoleyCrafter: Bring Silent Videos to Life with Lifelike and Synchronized Sounds. (2024).

[69] Zhe Zhang, Yi Yu, and Atsuhiro Takasu. 2023. Controllable lyrics-to-melody generation. *Neural Computing and Applications* 35, 27 (2023), 19805–19819.

[70] Changxi Zheng and Doug L. James. 2011. Toward high-quality modal contact sound. *ACM Trans. Graph.* 30, 4, Article 38 (July 2011), 12 pages. doi:10.1145/2010324.1964933