LUMIA: A Handheld Vision-to-Music System for Real-Time, Embodied Composition

Connie Cheng[*1], **Chung-Ta Huang**[*1], **and Vealy Lai**[2]

[1]Harvard University, Cambridge, MA 02138,
connie_cheng@gsd.harvard.edu, chungta_huang@gsd.harvard.edu
[2]Massachusetts Institute of Technology, Cambridge, MA 02139, laiv@mit.edu

Abstract

Lumia is a handheld, camera-inspired device for real-time music generation from visual input. Scenes captured by the user are analyzed by GPT-4 Vision to extract descriptors such as mood, objects, genre, and tempo, which are combined with chosen instrumentation into prompts for Stable Audio 2.0 [Evans et al., 2024]. Generated 15-second stereo loops are layered live via a Tone.js engine with beat-aware, equal-power crossfades, enabling embodied, improvisational composition that fuses perception, language, and sound.

1 Description of the Work

Lumia operates as a multimodal co-creative system that merges human framing with AI inference. When a user captures an image, GPT-4 Vision OpenAI et al. [2024] extracts contextual and affective descriptors—objects, mood, genre, and tempo—returned in structured form. These are merged with user-selected instrumentation into a single, optimized text prompt. Stable Audio 2.0 uses this prompt to generate a 15-second stereo loop, matched in tempo and mood to the captured scene. The system layers loops in real time with beat-aware, equal-power crossfading for seamless transitions, enabling embodied improvisation outside the traditional DAW.

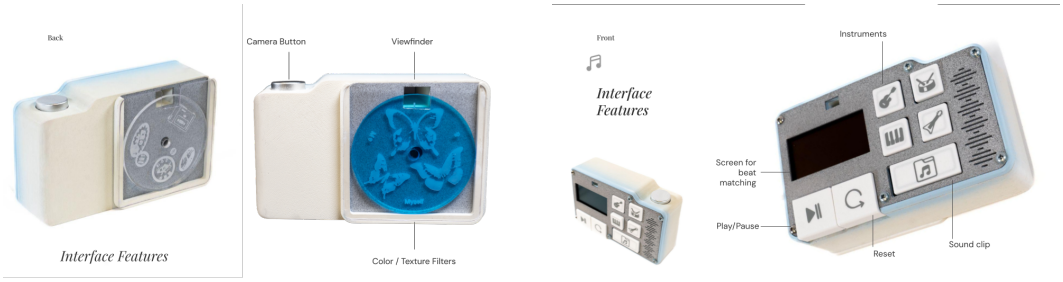


Figure 1: Front view (left) and back view (right) of the Lumia device.

Underlying this interaction is the paradigm of large language models (LLMs), which transform natural language inputs into structured, contextually rich outputs by leveraging high-dimensional representations learned from vast multimodal datasets. In Lumia, this capability is instantiated in a tangible, interactive form, aligning with the concept of *Large Language Objects* Coelho and Labrune [2024], where the reasoning and generative abilities of LLMs are embedded within physical interfaces. This embodiment allows users to engage with AI systems not only as abstract computational tools but as situated creative partners, making the process more transparent, intentional, and human-centered.

2 Role of AI/ML

Lumia operates as a multimodal co-creative system that merges human framing with AI inference. When a user captures an image, GPT-4 Vision extracts contextual and affective descriptors such as

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Creative AI Track.

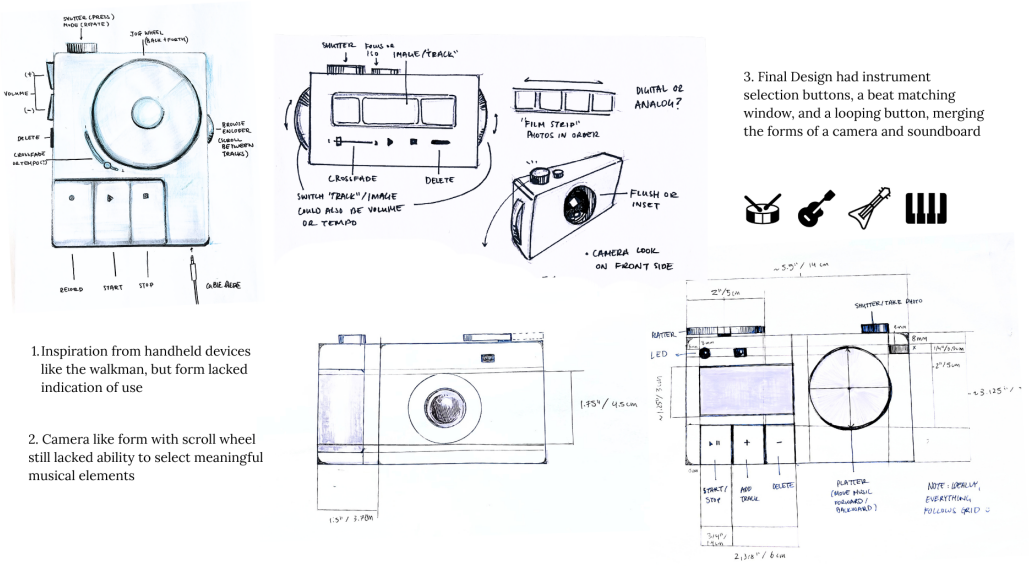


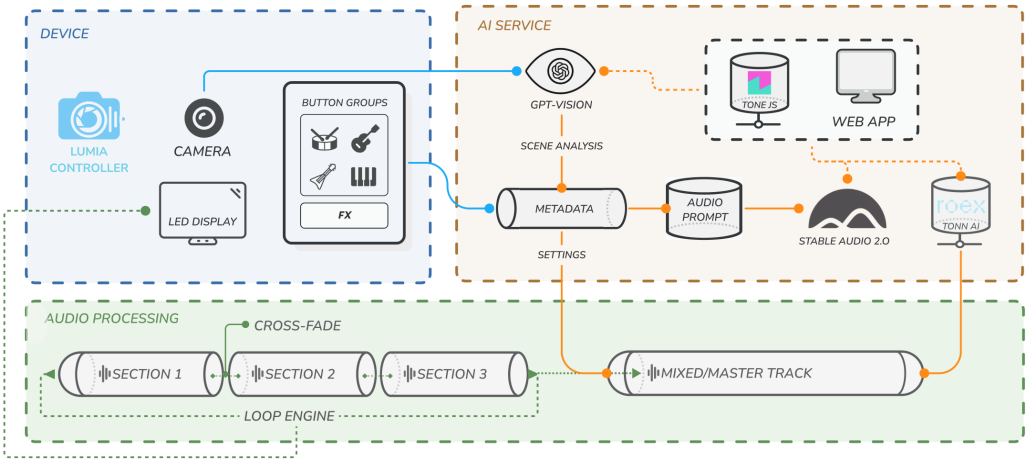
Figure 2: Form evolution

objects, mood, genre, and tempo, returning them in structured form. These descriptors are merged with user-selected instrumentation into a single, optimized text prompt that guides the generative process. Stable Audio 2.0 consumes this prompt to produce a 15-second stereo loop, matched in tempo and mood to the captured scene. The loop is not static; it becomes one element within a layered, evolving composition. The system continuously arranges these loops in real time, employing beat-aware scheduling and equal-power crossfading to achieve phase-coherent, seamless transitions. This design enables embodied improvisation in any setting, removing the constraints of a traditional DAW and replacing them with a portable, tactile, and immediate form of interaction. From a technical perspective, the pipeline begins with image capture and captioning, where the frame is transmitted to GPT-4V (average latency ~1.2 s, ~120 input tokens, ~200 output tokens) for analysis. The prompt construction stage applies section-specific modifiers (for example, “higher energy, catchy hook” for a chorus) and variation cues (such as “motif development”) to create musical continuity. Stable Audio 2.0 then generates a fixed-length stereo WAV at the desired BPM and style (average latency ~3.8 s, $0.14 per clip). These clips are streamed to the Tone.js-based playback engine, which manages real-time layering and ensures smooth overlaps using crossfade functions

$$g_{out}(n) = \cos\left(\frac{\pi n}{2N}\right), \quad g_{in}(n) = \sin\left(\frac{\pi n}{2N}\right)$$

so that $g_{out}^2(n) + g_{in}^2(n) = 1$ at all points. Optional stems can be sent to the Tonn API for rapid mix preview and mastering, integrated at bar boundaries.

AI/ML systems in Lumia serve distinct roles: GPT-4 Vision enables semantic and affective perception from imagery, Stable Audio 2.0 executes high-fidelity generative synthesis, and the playback engine’s adaptive scheduling logic supports dynamic, context-aware arrangement.



3 Addressing the Theme of *Humanity*

In line with the theme of *Humanity*, this work asks: what does it mean to be human when creativity is increasingly co-authored with machines? LUMIA's design philosophy is centered on fostering intentional, situated human–AI co-creation. User control, through framing, instrumentation choices, and scene selection, actively shapes and constrains the AI's interpretive and generative processes, ensuring that each musical outcome reflects both human intent and machine contribution. The device's portable, camera-like form factor enables music-making in diverse, lived environments, grounding creative output in real-world contexts and personal experience. This form is an intentional nod to photography and fieldwork: just as the advent of photography democratized image-making and sparked radical artistic innovation, AI now acts as the "next camera," transforming creative paradigms and challenging notions of artistic value in an era where beautiful works can be generated instantly. We believe that value lies in the deliberate choices and embodied experiences that guide creation, and Lumia was crafted to embody that ethos. By uniting transparency in prompt construction, user agency in creative direction, and a familiar, approachable interaction model, Lumia promotes cultural diversity, counters stylistic homogenization, and avoids anthropomorphizing AI, framing it as a responsive instrument rather than an autonomous author. Looking forward, we aim to further tighten the coupling between gesture and digital outcome, ensuring that physical actions directly modulate the qualities of generated audio, deepening the embodied nature of human–machine collaboration.

4 Author Biographies

Connie Cheng* (Harvard Graduate School of Design) is an industrial designer and HCI researcher interested in human-AI interaction with AI tooling, particularly for creative and educational settings.

Chung-Ta Huang* (Harvard Graduate School of Design) is a creative technologist specializing in the design of immersive entertainment and music systems, with a research interest in multimodal interaction.

Vealy Lai (Massachussets Institute of Technology) is an undergraduate in the Mechanical engineering department and does accessibility related robotics research at the D'arbeloff lab.

References

M. Coelho and J.-B. Labrune. Large language objects: The design of physical ai and generative experiences. *Interactions*, 31(4):43–48, June 2024. ISSN 1072-5520. doi: 10.1145/3672534. URL `https://doi.org/10.1145/3672534`.

OpenAI, J. Achiam, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024. URL `https://arxiv.org/abs/2303.08774`.