

Mirelo's Picture Gap

Sound for video is not the same as scoring for picture.

Charlie Fisher, Film & TV Composer. Bleeding Fingers Music / Sony Music Publishing. Human X, Human X: Part Two (Netflix, Amazon Prime). Music+Sound Awards Rising Star 2025.

Executive Summary

Most professional creatives working at delivery level have tried AI audio tools. None of them use them for delivery yet. That is not a coincidence.

The gap between AI audio generation and professional adoption is a specification problem. The capability exists. The market is measurable. What is missing is a precise, professionally grounded definition of what the tool needs to do, built from inside the workflow rather than projected onto it from outside.

For a company building the generative audio layer for visual content, the professional film and television market is the proof of concept. If the model meets the demands of the most exacting use case in audio-for-picture, it can meet any use case below it. The ceiling, once broken, raises the floor for every user beneath it.

Mirelo has built the correct foundation. The professional layer above it has not yet been tested at delivery standard. This paper is about that distance.

Mirelo's Architectural Bet

No other tool in this space has made the right architectural decision: sound should respond to what is on screen, not be described in isolation from it. The SFX model proves the thesis works in practice. It is a functioning proof of concept for what the music layer needs to become.

One distinction is worth noting early. SFX generation stays objective. A footstep, a door, an impact has a correct answer relative to the image. Musical generation is different. The machine/user creative balance is harder to calibrate, and getting it wrong in either direction loses the professional. Too much machine and the output has no authorial identity. Too little and the tool adds no value. That calibration is the central design challenge the music layer faces, and it requires someone who understands both sides of it from practice, not theory.

Professional Adoption as the Forcing Function

Consumer numbers tell you the product works. Professional billing tells the industry what to build next.

When a professional creative works to a locked picture edit, to a delivery standard set by a dubbing stage, and delivers with the output, they establish something no engagement metric can. A tool tested at that standard and found trustworthy becomes the benchmark the industry measures itself against. Formula One engineering does not exist to make racing cars. It exists because designing for the absolute limit of performance produces knowledge that filters through every tier beneath it. Professional audio-for-picture is the same forcing function.

The professional audio-for-video market spans advertising production, broadcast drama, theatrical film, gaming, and documentary. It does not behave like consumer subscriptions. A locked cue generates performance royalties across territories for years after the session closes. A tool embedded at that level does not get swapped out when a competitor ships a new feature. The switching cost is the entire career.

The SFX Model as Proof of Concept

Mirelo's SFX model already does what the music model does not. It reads a video and generates audio synchronised to visual events, responding to timing, energy, and spatial context on screen.

The music module, currently powered by the 11 Labs model while Mirelo's own model is offline, works differently: submit a text prompt, choose from two or three generations. Picture lock, timecodes, frame hits, layered generations, and a way to develop a musical variation are all absent. The video conditioning that defines Mirelo's SFX model is not present in the music layer.

This is not a criticism of where the product is. The current music module does something useful. It establishes tonal and textural direction for a scene, which is a genuine starting point for a creative workflow. Part of Mirelo's user base does not speak in musical language, and a tool that can infer a viable sonic world from visual input alone has real value for that user. The diagnostic point is narrower: the SFX model treats the image as its primary input; the music module treats it as backdrop. **In professional picture work, image leads.** Closing that gap, so that the music layer reads the picture the way the SFX layer already does, is the specific unlock this paper is about.

The Mirelo model, when it returns, may close some of this distance. What it will need to close it entirely is domain knowledge from inside the professional creative workflow. Someone who can define not just what the output should sound like, but how it should respond to the picture in each case, across the full range of users who cannot specify that themselves.

Accuracy, Not Art

The debate about AI and musical creativity asks the wrong question. Before a professional can engage the creativity question, the tool needs to handle the objective layer.

Music has an internal logic that is not a matter of taste. Pitch relationships are verifiable. Harmonic voicings either support the existing chord or they do not. A rhythmic element either locks to the grid or it drifts. These are parameters with correct and incorrect values, the same way a variable in a function has a correct and incorrect state.

The visual layer contains the same kind of information. Edit pace, lighting quality, shot scale, spatial depth. A professionally trained creative reads these as starting points for sonic inference. A slow-cut sequence with intimate framing suggests one sonic world; a rapid montage with wide exterior shots suggests another. These are heuristics, not rules. Sometimes the most powerful choice runs against the visual grain. They are still data. **Mirelo's video conditioning is already performing this inference for SFX. The music layer needs to perform it too, with the understanding that the creative still owns the final call.**

Once the machine handles the objective layer and the creative retains all subjective decision-making, the authorship question stops blocking professional revenue. The US Copyright Office has confirmed that AI assistance does not bar copyrightability when the human contributes sufficient creative control. If the machine is responsible only for verifiable accuracy, the human contribution is the work, and the revenue reaches the right person.

Sound DNA: Getting the Genre Right

Every professional audio genre has an acoustic and timbral identity, a set of production decisions so consistently applied that experienced listeners recognise them immediately. Not right or wrong in an absolute sense, but intentional. The difference between a sync-ready orchestral cue and a convincing orchestral approximation is not compositional. It is the sum of decisions about ensemble texture, dynamic range, spatial placement, and the relationship between elements. Decisions that define the genre and signal to a music supervisor that the person who made them knows the territory.

Consumer AI outputs miss this not because the compositional material is wrong, but because the genre DNA is absent or inconsistent. The gap is not about room size. It is about whether the output reads as belonging to its genre in the way that a professional ear requires.

There is a path through this that Mirelo is positioned to take. The SFX model already infers acoustic space from visual data: the character of a large interior, the proximity of a source, the ambient density of an environment. The same spatial inference applied to the music layer is the beginning of an answer. A scene's visual environment contains acoustic information. A model that reads the picture can use it. The professional market is waiting for a tool that does.

Three Possibilities

These three are a starting point, not a ceiling. Many more possibilities exist once the visual intelligence Mirelo has already built is connected to the music layer above it.

Possibility 01: Iterative editing comes to the music layer.

The SFX model can now regenerate a section without touching the rest. A glitchy second of audio gets fixed. The surrounding material stays exactly as it was. That capability, applied to the music layer, changes everything about how a professional cue gets built and revised. A director asks for the emotional peak four seconds earlier to hit a cut. You move it. The harmonic context holds. The surrounding architecture stays intact. That is not a feature. It is the definition of a professional instrument. Mirelo has already proved the architectural conviction. The music layer is the next build.

Possibility 02: The creative becomes an architect, not a selector.

Professional music for picture is built, not chosen. A three-minute cue has an architecture: an opening, a build with internal logic, a resolution that arrives because the structure demanded it. Constructed across melodic, harmonic, textural, and rhythmic layers, each decision informing the next. Layering, variation expansion, and iterative cue development would close the distance between a mood selector and a compositional environment. That distance is the distance between a consumer tool and a professional instrument.

Possibility 03: Synchronisation makes delivery possible.

The defining requirement of professional picture work is audio locked to specific visual moments. Frame hits, cut points, beats that land on specific images. **In professional scoring, music exists for the image.** Timecodes are the language of that relationship, and they need to be flexible. Picture often changes after the first pass, and the music needs to adapt without being rebuilt from scratch.

The SFX model synchronises to picture. The music model does not. Closing that gap is the single change with the highest leverage on professional adoption, and Mirelo's own architecture has already proven it is possible.

One Decision. Three Unlocks.

Unlock 01: The professional market opens. Advertising production, broadcast drama, theatrical film, documentary, gaming. Every sector where audio is commissioned to serve a specific visual purpose and measured against a delivery standard that cannot be approximated. The specification layer is the moment those markets become reachable, not all at once, but compounding from the first locked delivery.

Unlock 02: Authorship becomes defensible. The rights infrastructure that makes professional audio financially sustainable depends on a clear distinction between what the machine did and what the human did. The specification layer makes that distinction objective. Machine owns the accurate execution of verifiable parameters. Creative owns every subjective decision. The human contribution is the work, and the revenue reaches the right person across territories, for years after delivery.

Unlock 03: The standard gets set. Mirelo is defining a category that does not yet have a benchmark. **The first company to meet the professional delivery standard does not enter a market. It creates one.** Every tool that follows will be measured against what Mirelo built. The media creatives, music supervisors, and directors who adopt it first will carry it into every production that comes after. That is not a feature advantage. That is a structural position, and Mirelo is closer to it than anyone else in this space.

The Specification Layer That Must Be Built

The gap sits between disciplines. It cannot be closed by an audio research engineer learning music theory, or a product designer learning sync licensing. It requires someone who has never had the luxury of treating these as separate domains.

There is a distinction worth drawing here. Musical training (conservatoire composition, classical performance, academic grounding) is genuine expertise. It is also a different specialisation to professional media delivery. The media delivery workflow is sync briefs, broadcast technical standards, DAW integration, picture lock protocols, dubbing stage preparation, library submission requirements, and the specific revision language of directors and music supervisors. That knowledge is not taught. It is accumulated under live briefs, at delivery deadlines, inside the workflow the model needs to reach.

That is what determines whether the specification layer gets built correctly, or approximated by people making informed guesses about a workflow they have studied but not worked inside. Before these architectural decisions are made, they are open. After they ship, they are expensive to reverse.

Mirelo has built further toward the correct answer than any other company in this space. The video conditioning is real, it works, and it proves the thesis. The music layer above it needs to reflect that ambition.

A Personal Note

I have tested what is live. I do not know what the Mirelo model does that ElevenLabs does not. It was not available to test. What I know is what the professional music layer needs to look like above either of them, because I am the creative it needs to reach. I am scoring for picture under active briefs right now. I can hear the distance between what exists and what delivery requires. The tool described in this paper is one I want for myself as much as I want it to exist for the industry.

Charlie Fisher: Film and TV composer. Credits include Human X and Human X: Part Two (Netflix, Amazon Prime, 1,000+ cinemas, 20+ countries). Active composer with Bleeding Fingers Music (the Hans Zimmer / Sony Music Publishing joint venture) and Extreme Music / KPM (Sony's global production music library). Music+Sound Awards Rising Star 2025. PRS member, earning sync and performance royalties across broadcast territories. Built production software in HTML/JS and React Native. Evaluates AI music generation tools against professional delivery standards. 5M+ organic social media impressions.

charliefishermusic.com · @charliefishermusic