Whose Intent? Distributed Authorship, Cross-Modal Constraints, and the Limits of Generative Music Systems in Screen Scoring

Chenxi Bao
Dynamix, China
cloudingcxb17@gmail.com

Jingrui Han
Auckland University of Technology
Qingdao Film Academy
jingrui.han@aut.ac.nz

Zhaokai Wang
Shanghai Jiao Tong University, China
wangzhaokai@sjtu.edu.cn

Clinton Watkins
Auckland University of Technology, New Zealand
clinton.watkins@aut.ac.nz

Abstract

While general text-to-music generation has achieved substantial success, video-to-music generation continues to struggle with widespread adoption in professional industrial contexts, such as film scoring. The dominant technical society usually attributes this bottleneck to the lack of high-quality paired datasets and evaluation challenges. In this paper, we argue that the root cause is fundamentally paradigmatic. Most current AI music systems are constrained by "single-agent" and "endogenous time" assumptions, severely misaligning with the inherently multi-agential, highly contextualized reality of professional screen scoring. Through a systematic theoretical critique and a practice-based expert case study, we deconstruct the professional film scoring pipeline into its constituent multi-agential negotiations. We articulate three structural frictions hindering AI deployment: the intentional gap between directors' creative language and algorithmic parameters, the strict subordination of musical architecture to exogenous visual timelines, and the cross-modal conflicts inherent in the final dub stage. Ultimately, we propose actionable Human-Computer Interaction (HCI) design implications, advocating for a paradigm shift from building "autonomous music generators" to developing "collaborative boundary objects" capable of supporting the relational labor and distributed authorship required in industrial workflows.

1 INTRODUCTION

In recent years, driven by the advancement of AI technologies and the expansion of related research communities, researchers have increasingly turned their attention to the field of music AI. Within this domain, general music generation, such as text-to-music generation, has achieved substantial success in both academic and commercial spheres(Copet et al., 2023; Evans et al., 2024). However, in the realm of video-to-music generation, although Ji et al. (2025); Wang et al. (2025) has comprehensively reviewed the huge amount of recent relevant works, this technology still struggles to find widespread adoption in professional industrial contexts, such as film scoring.The evaluation of generative music systems remains an open methodological challenge, compounded in professional screen scoring by the absence of evaluation frameworks that account for cross-modal and multi-stakeholder constraints.

From a technical perspective, this bottleneck is often considered as being attributed to the scarcity of high-quality paired datasets and the inherent difficulties in evaluating generated music. However, from a composer's point of view, this friction stems from a fundamental misalignment between the interaction paradigms of current AI technologies and the established collaborative workflows of the film scoring industry.

Specifically, contemporary AI music co-creation systems typically operate under the assumption of a "single user" who manipulates internal musical parameters (e.g., harmony, tempo, and style). However, the real-world screen scoring process is inherently multi-agential. The final musical outcome is not dictated by a single individual, but is rather a negotiated consensus reached among the director, the editor, and the composer through iterative spotting sessions.

Furthermore, distinct from traditional music generation models, the temporal structure of music in screen scoring is exogenous rather than endogenous; it must be strictly subordinated to the visual edit (e.g., hard cuts, hit points). However, current video-to-music generation works still predominantly focus on obvious external parameters and do not pay attention to the intrinsic cross-modal parameters embedded within the film scoring workflow. Consequently, this renders AI systems incapable of comprehending such cross-modal temporal imperativeness.

While it might be tempting to attribute the current bottlenecks in AI screen scoring to a mere lack of high-quality paired data, we argue that the root cause is fundamentally paradigmatic. In this paper, we systematically analyze existing technical mechanisms alongside real-world scoring practices. By examining this landscape through the dual lenses of both the computer scientist and the composer, we articulate the structural gaps between current generative systems and industrial deployment. Ultimately, we aim to shift the research focus from merely expanding datasets to rethinking interaction workflows, providing practical design implications to guide future research.

Specifically, our core contributions are three-fold:

Theoretical Critique: We systematically deconstruct the "single-agent" and "endogenous time" assumptions in current generative models, demonstrating why scaling datasets alone cannot resolve the misalignment with industrial workflows.

Practice-Based Analysis: Drawing on real-world film scoring case studies, we articulate the hidden relational labor and cross-modal parameters (e.g., hard cuts, hit points) that define the professional pipeline.

Design Implications: We propose actionable HCI design guidelines, advocating for a paradigm shift from building "autonomous music generators" to developing "collaborative boundary objects" that integrate into multi-agential workflows.

2 TECHNICAL AFFORDANCES

2.1 The exploration of AI film scoring has unfolded over several years, with the V2M-survey(Wang et al., 2025) cataloging research prior to March 2025. Observing this timeline reveals that the evolution of AI film scoring is essentially a gradual shift away from strictly end-to-end processes. This progression inevitably points to one ultimate outcome: the AI film scoring workflow will eventually merge with, and fully reflect, the traditional real-world film scoring workflow.

Early studies, such as CMT (Di et al., 2021) and V-MusProdZhuo et al. (2023), focused heavily on translating a video's narrative and rhythmic features directly into symbolic music (MIDI) generation. During this period, researchers primarily viewed the relationship between video and music as a direct mathematical link. However, since the artistic and emotional connection between images and sound is far more complex than a simple formula, the results naturally fell short of expectations.

As time went on,the focus shifted toward direct audio synthesis. Projects from this era, such as V2Meow(Su et al., 2024) and VidMuse(Tian et al., 2025), aimed to directly align visual and audio elements within a latent space. Alternatively, V2M-Zero(Lin et al., 2026) generated music by comparing the internal similarity curves of both the music and the video. While this approach allowed the music to reflect both the overall tone and specific moments of a video to some extent, it remained an end-to-end learning method at its core, heavily relying on the existing capabilities of models like MusicGen(Copet et al., 2023) or Stable Audio(Evans et al., 2024).

The arrival of Large Language Models opened a new chapter. MTM(Wang et al., 2024), Mozart's Touch(Li et al., 2025) and SONIQUE(Zhang and Fuentes, 2025) began to use LLMs to first translate the video into text descriptions, which were then used to guide the generation of music. This approach successfully achieved a degree of decoupling and allowed users to edit the text descriptions, introducing a welcome element of human-computer interaction. However, text naturally contains less information than video or audio, for instance, it is notoriously difficult to capture the exact content or feeling of a piece of music or a visual scene in mere words. Consequently, this method acts as a "compress then sample" process, leading to significant information loss. The music produced this way often shows shifts through time, but the overall music quality remains unpolished and not professional.

Within the past year, researchers have increasingly recognized the need to move toward professional film scoring workflows. FilmComposer(Xie et al., 2025a), for instance, introduced a dedicated model for "spotting" (marking exactly where music should change to match the video). It also pioneered a multi-agent approach to handle video analysis, melody writing, orchestration, and mixing. While it still utilized MusicGenCopet et al. (2023), it introduced a transcription model to convert the generated audio melody into MIDI format for further orchestration and mixing. Moreover, the project included a user-friendly interface, allowing creators to easily edit each element. As a result, the music produced by this system already captures a sense of cinematic atmosphere.

Yet, a critical question remains: how can we facilitate collaboration among the various creative roles within a professional film scoring workflow, such as composers, directors, and music editors? Creating a shared workflow or model that all these professionals can use together is an unresolved and fascinating challenge. Recently, AutoMV(Tang et al., 2025) introduced a "Screenwriter Agent" and a "Director Agent" to draft scene scripts, storyboard lists, and character profiles. This is similar to the concept mentioned above, but its primary goal was to generate music videos based on songs, rather than composing music for visual media. Therefore, if artificial intelligence is to be fully integrated into the professional film scoring workflow, a systematic and thorough analysis of this workflow is the essential next step.

2.2 Dataset Analysis

Previous research has employed a variety of approaches to dataset construction, with the vast majority of datasets being compiled through automatic algorithmic extraction. Datasets such as BGM909(Di et al., 2021), SymMVZhuo et al. (2023), for instance, have gathered video and audio clips specifically from music videos (MVs). While this approach easily yields a massive scale of data, it overlooks a crucial reality: music videos are visuals crafted around pre-existing music, designed to amplify the auditory experience and often heavily infused with the director's subjective stylistic choices. Film scoring, conversely, involves composing music for the video, using sound to support and elevate the visual narrative. The underlying production logic of the two mediums is entirely reversed. Consequently, datasets built on MVs fundamentally struggle to capture the true, intrinsic relationship between music and videos as it exists in film.

Other studies have opted for broader, more diverse video datasets, such as DISCO-MV and DVMSetLi et al. (2024); Lin et al. (2025), which encompass a wide array of formats including commercials, promotional clips, and animations. However, this sheer variety creates a chaotic learning environment for the model, which make it incredibly difficult for the AI to learn the precise, nuanced styles required for professional film scoring.

Encouragingly, recent years have seen a gradual shift toward dataset construction rooted in the professional film industry. VidMuse(Tian et al., 2025), for example, gathers a collection of trailers, documentaries, and cinematic clips. Taking it a step further, MusicPro-7k(Xie et al., 2025a) directly commissioned professional musicians to compose original scores for silent excerpts from the LSMDC classic film database, complete with manually annotated sync points.

Yet, a significant gap remains: to date, no film scoring dataset contains detailed and comprehensive metadata reflecting the involvement of multiple creative roles (such as directors and music editors). While some datasets might include global information like a director's name, the absence of rich, role-specific data makes it difficult to build unified, multi-agent models capable of assisting in a collaborative creative process. FilmComposer(Xie et al., 2025a) attempted to address this lack of detailed role data by using prompt-engineered AI agents to simulate roles like the director. This creative solution marks an exciting shift in the field: instead of struggling to find datasets with highly detailed labels, researchers are now using AI agents to create expert instructions in real-time. However, a fundamental bottleneck remains. While AI agents can simulate individual roles to patch dataset deficiencies, these technical workarounds do not equate to understanding the actual, messy sociological dynamics of a scoring room. If we are to evolve these tools from autonomous generators into shared boundary objects that genuinely facilitate seamless interaction, we must step outside the algorithmic domain and systematically examine the real professional workflow.

3 EXPERT CASE STUDY

As the preceding technical analysis reveals, the trajectory of AI film scoring has outgrown simple end-to-end generation, yet it remains conceptually constrained. Contemporary AI music co-creation systems largely operate under what we term a single-agent assumption: the idea that an individual user possesses a legible creative intention that can be progressively articulated through interactions with musical parameters (Louie et al., 2020, 2022; Young et al., 2022).

However, when these generative systems are transposed into the professional screen scoring workflow, the single-agent assumption collapses under the structural realities of the film industry. We argue that this "paradigmatic friction" is not merely a technical shortfall in music quality, but a fundamental misalignment in how creative agency is conceptualized. In professional scoring, the "user" is not a bounded individual, but a *distributed network of agents* whose intents are often contradictory, non-musical, and constantly shifting in response to the visual edit.

We identify three core dimensions where contemporary AI music systems encounter these frictions:

- **The Intentional Gap (Distributed vs. Single Agent):** While current systems expect a user to provide clear, parameter-based prompts, directorial intent is often articulated through metaphors ("make it feel like a cold morning") or narrative subtext. The composer's labor, therefore, is not just musical production, *relational translation*—a process of bridging the gap between a director's subjective vision and the system's acoustic output.
- **Cross-modal Relationality (Relational vs. Autonomous):** In standalone generation, music is evaluated by its internal coherence. In screen scoring, musical "success" is defined by its relationship to non-musical elements—dialogue frequencies, Foley textures, and character glances. Current AI models, lacking "visual-dialogue awareness," fail to treat music as a subordinate narrative layer.
- **The Tyranny of the Frame (Exogenous vs. Endogenous Time):** Generative models typically operate on internal rhythmic logic (beats, measures). Screen scoring, however, is governed by *exogenous time*—the absolute, frame-accurate timeline of the locked cut. A single 12-frame shift in an edit can render a generative structure obsolete, a constraint that current "black-box" temporal models cannot accommodate.

To illustrate these frictions in detail, we present a case study in screen scoring, breaking down the professional workflow into its interactions between multiple agents, deconstructing the professional scoring workflow into its constituent multi-agential negotiations.

Terminological Note

Before proceeding, we clarify several domain-specific terms drawn from professional screen scoring practice that recur throughout this analysis (Buhler et al., 2010; Cooke, 2008).

A *spotting session* is a structured viewing meeting in which the director, editor, sound supervisor, and composer watch a near-final cut of the film together to determine where music should begin and end, and what narrative or emotional function it should serve at each point.

A *locked cut* refers to the version of a film's edit that has been formally approved by the director and will undergo no further structural changes. It marks the point at which the composer may begin writing to picture in earnest, since only a stable timeline permits frame-accurate synchronization.

A *hard cut* refers to an instantaneous transition between two shots with no transitional effect—the most common editorial cut in film. In scoring terms, a hard cut frequently functions as an absolute temporal boundary: the music must either conclude precisely at the cut or begin anew from it, with no ambiguity.

A *hit point* (or *sync point*) denotes a specific visual event—such as a cut, a physical gesture, or a dramatic action—to which the music must be synchronized with frame-level precision. Hit points constitute the primary temporal anchors around which a film score is structured.

SMPTE timecode is the industry-standard frame-level time-addressing format, expressed as hours:minutes:seconds:frames (e.g., `01:14:08:12`). All in-points, out-points, and hit points in a spotting session are recorded in SMPTE, making it the shared coordinate system across picture, dialogue, and music.

A *mock-up* is a preliminary realization of a cue produced by the composer using MIDI sequencing and orchestral sample libraries, presented to the director before live recording. It serves as the primary *boundary object* for iterative feedback: concrete enough to provoke a response, malleable enough to be revised.

A *temp track* (or *temporary track*) is existing music—typically drawn from other film scores or recordings—which an editor places against the cut during post-production before an original score is composed. Directors frequently develop strong aesthetic attachments to temp tracks, creating an implicit reference frame that the composer must navigate and, where necessary, displace.

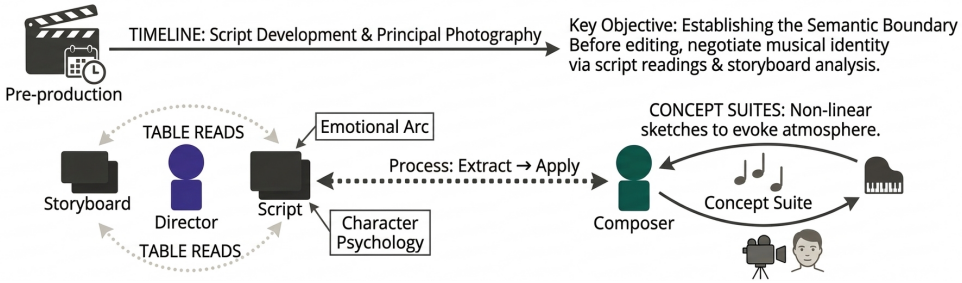

Figure 1: Phase 0: Narrative Grounding & Aesthetic Priming (Pre-production).

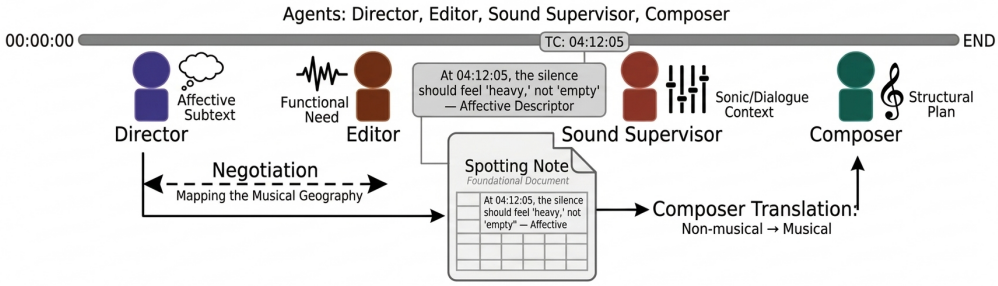

Figure 2: Phase 1: The Spotting Session (Negotiating Intentionality).

A *dub stage* is the facility at which the final mix takes place, where a *re-recording mixer* integrates the music, dialogue, and sound effects into a unified sonic track. The re-recording mixer holds ultimate authority over the relative levels of each element, and may significantly reduce the presence of the score where other sonic layers take precedence.

Finally, *stems* refer to the individual component layers of a mixed score—such as strings, brass, or percussion delivered in isolation—which are submitted to the dub stage so that the re-recording mixer can balance each layer independently against dialogue, foley, and ambient sound. Any system that outputs only a single, non-decomposable audio file is structurally incompatible with this stage of the workflow.

3.1 Practice-Based Analysis: The Screen Scoring Workflow

3.1.1 Methodological Note

The analysis presented in this section draws on the professional screen scoring experience of one of the authors, supplemented by informal observational insights gathered through participation in industry production contexts. It is positioned as an *expert reflective case study*, a methodology employed in early-stage design research and computational creativity work when formal datasets or controlled user studies are not yet available or methodologically appropriate. The cases presented are reconstructed accounts selected for their theoretical relevance to the structural mismatches identified in this paper; they are not transcribed records and were not collected under institutional ethical review. They are subject to self-report bias and carry no claim of statistical generalizability. Their function is illustrative rather than evidential: each case is intended to instantiate a structural pattern, not to establish a generalizable empirical finding.

Critically, the cases are not presented as evidence that current AI tools universally fail in film scoring contexts. For standalone music generation, early ideation, and single-user compositional sketching, existing systems offer genuine creative utility (Larsen and Zhu, 2024). The critique developed below is scoped specifically to *professional multi-agent screen scoring workflows*, characterized by locked-cut production timelines, distributed authorial authority, and the requirement for frame-accurate, stem-decomposable output. Where a case touches on a task that current AI tools handle adequately, this is noted explicitly.

We frame this workflow not as a linear act of composition, but as a distributed negotiation governed by exogenous temporal constraints and cross-modal feedback loops. Each phase below identifies a specific friction point and draws an implication for AI system design.

3.1.2 Phase 0: Narrative Grounding (Pre-production)

Timeline: Script Development & Principal Photography | **Key Objective:** Establishing the Semantic Boundary.

Before editing begins, the musical identity of a film is negotiated through script readings and storyboard analysis. The composer participates in table reads to develop an understanding of the film's emotional arc and character psychology, producing "Concept Suites"—non-linear musical sketches intended to evoke the film's atmosphere rather than accompany specific scenes. These sketches may serve as on-set stimulus during principal photography, functioning as a relational anchor within an evolving multi-agent interpretive process (Figure 1).

Implication: At this stage, no locked visual material exists against which a generative system could align its output. Current AI music systems are input-conditioned on audio, video, or text prompts; they have no architectural mechanism for representing the open-ended semantic constraints being negotiated between director, composer, and produc-

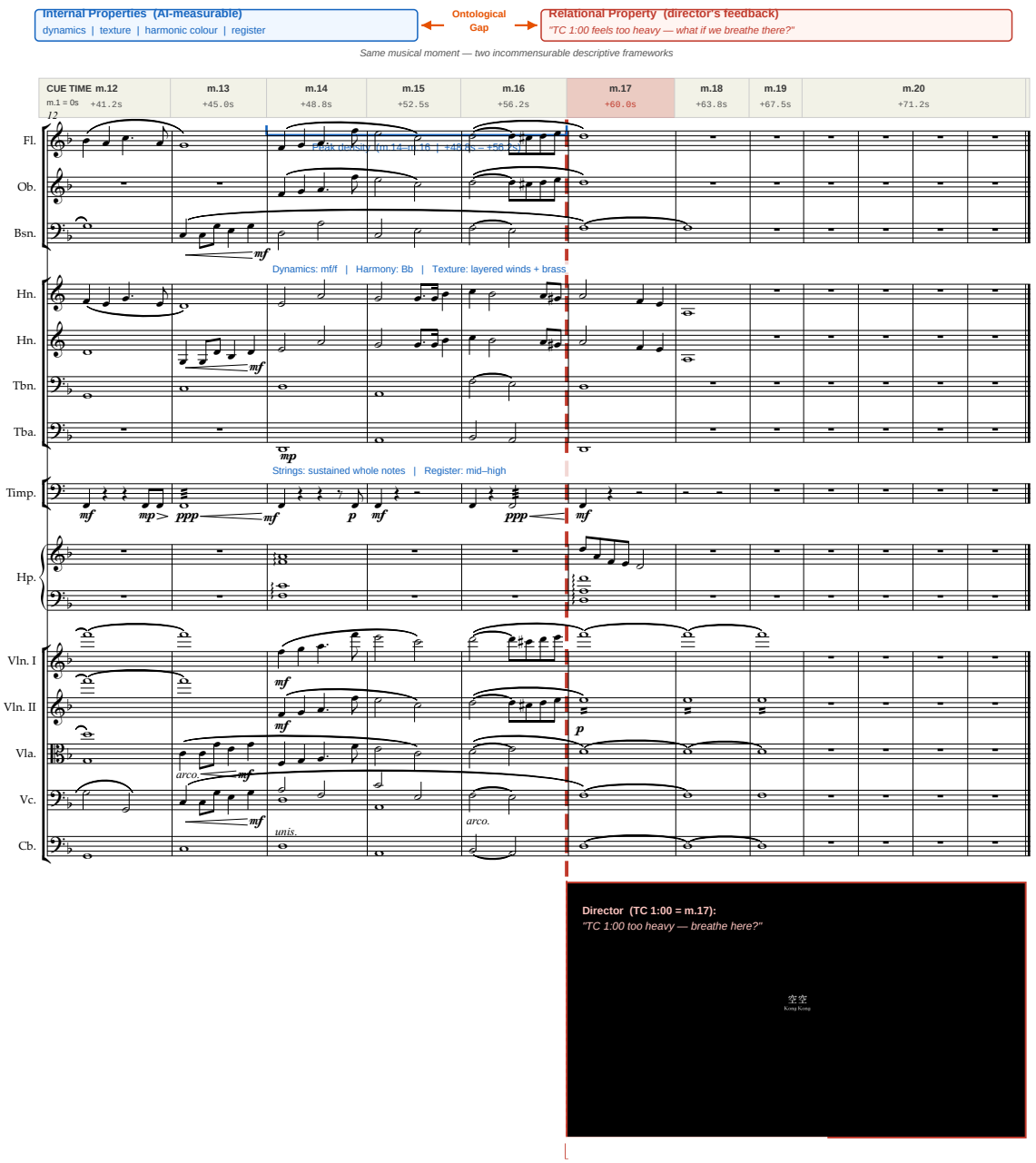

Figure 3: Scoring discussions can be understood as a process of distributed meaning-making (Davis et al., 2016), in which collaborators construct a shared interpretive framework through metaphorical, narrative, and cross-modal communication.

tion team prior to any fixed timeline. The absence of a *pre-cut constraint representation layer*—a structure capable of encoding evolving narrative and emotional parameters before a visual edit exists—means that AI tools are not well-aligned with this phase of the workflow, regardless of their generative quality.

3.1.3 Phase 1: The Spotting Session (Negotiating Intentionality)

Timeline: Week 1–2 | **Agents:** Director, Editor, Sound Supervisor, Composer.

Once the locked cut is established, the team conducts a spotting session to determine where music should appear, what narrative function it should serve, and what emotional register it should occupy (Figure 2). The resulting spotting notes record precise in/out timecodes in SMPTE format.

The following exchange, reconstructed from professional experience, illustrates how directorial instruction operates in practice. The session involved a film centred on familial relationships:

Director: "The moment at 1:00 feels too heavy—what if we let it breathe there?"

Composer: "Do you mean pull the strings back, or let the silence do more work?"

Director: "More the second—the image already carries the weight. The music shouldn't confirm it (Figure 3)."

This exchange is analytically significant for two reasons. First, the director's instruction simultaneously encodes a temporal coordinate, a dramatic valence, and a relational stance toward the image, none of which is reducible to a musical parameter. Second, when the composer attempts

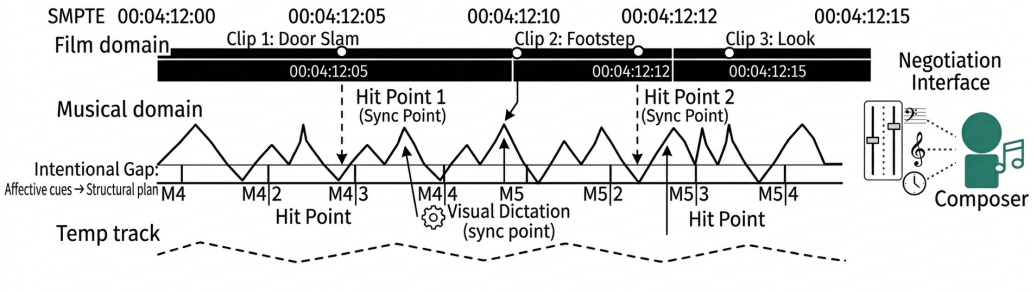

Figure 4: Phase 2: Mock-up & Temporal Subordination (The Exogenous Logic).

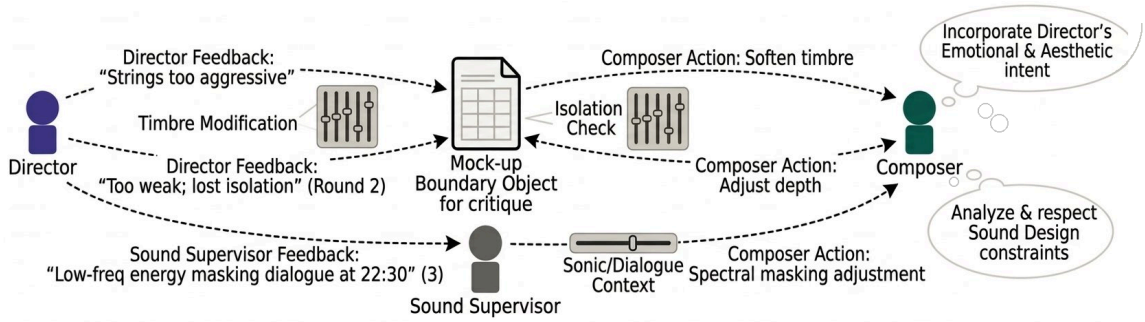

Figure 5: Phase 3: Iterative Refinement & Cross-modal Conflict (The Feedback Loop).

to translate it into actionable terms, the director reframes the response relationally: the music's role is defined not by its own content, but by what the image is already doing.

It is worth noting that AI tools can contribute usefully to the earlier, more generative phase of this process: given a textual description of a scene's emotional register, a system such as MusicGen can produce reference material that gives the director and composer a shared sonic object to react against. This is a legitimate and productive use. The limitation emerges at the point where the director's instruction becomes relational and cross-modal, as in the exchange above. At that point, the system has no representational structure for encoding "the music should not confirm what the image is already doing"—a directive that defines a relationship between sonic and visual layers rather than a property of the music in isolation.

Implication: Current AI music interfaces face structural limitations in multi-agent spotting contexts because they are designed to receive musical instructions and return musical outputs. They do not provide mechanisms for encoding the *interval between music and image* that professional spotting sessions are centrally concerned with establishing.

3.1.4 Phase 2: Mock-up & Temporal Subordination

Timeline: Week 3–6 | **Constraint:** Frame-Accurate Synchronization.

In screen scoring, temporal structure is exogenously imposed by the visual track. The locked cut establishes a rigid framework including hard cuts, scene boundaries, and hit points to which the score must conform with frame-level precision (Figure 4). The following case illustrates the nature of this constraint. In a session scoring a boxing film, a hard cut was placed at `2:14:08`—the moment of impact—after which the frame cut to black. The director's instruction was unambiguous: *"The music has to end with the cut."* This single directive subordinates all preceding musical architecture—phrase length, harmonic resolution, orchestral release—to one visual event at a fixed timecode.

In this case, the composer used a generative tool to sketch the climactic passage leading to the cut. The system produced material with strong internal momentum—rhythmically coherent, dynamically appropriate, and affectively well-matched to the scene. Evaluated as a piece of music, the output was successful. The failure emerged at the architectural level: the phrase's natural resolution fell approximately four seconds past `2:14:08`. The system had no mechanism for treating the cut as a hard temporal constraint; it generated music according to its internal rhythmic logic and produced a structurally complete phrase that was nonetheless incompatible with the visual edit. The composer was required to manually re-edit the audio at the cost of its internal coherence before it could be used (Figure 6).

This case is not presented as evidence that the tool was poorly designed for its intended purpose. For standalone generation, its output was of high quality. The mismatch is structural: the tool was designed for a context in which temporal structure is endogenously determined, and was applied to one in which it is exogenously fixed.

Implication: The fragility of the temporal contract in screen scoring is considerable: a 12-frame editorial shift can render an entire musical structure functionally obsolete. AI systems face structural limitations in professional locked-cut workflows because they treat temporal boundaries as post-hoc alignment targets rather than as generative priors (Xie et al., 2025b).

3.1.5 Phase 3: Iterative Refinement & Cross-modal Conflict

Timeline: Week 7–9 | **Agents:** Director, Composer, Sound Supervisor.

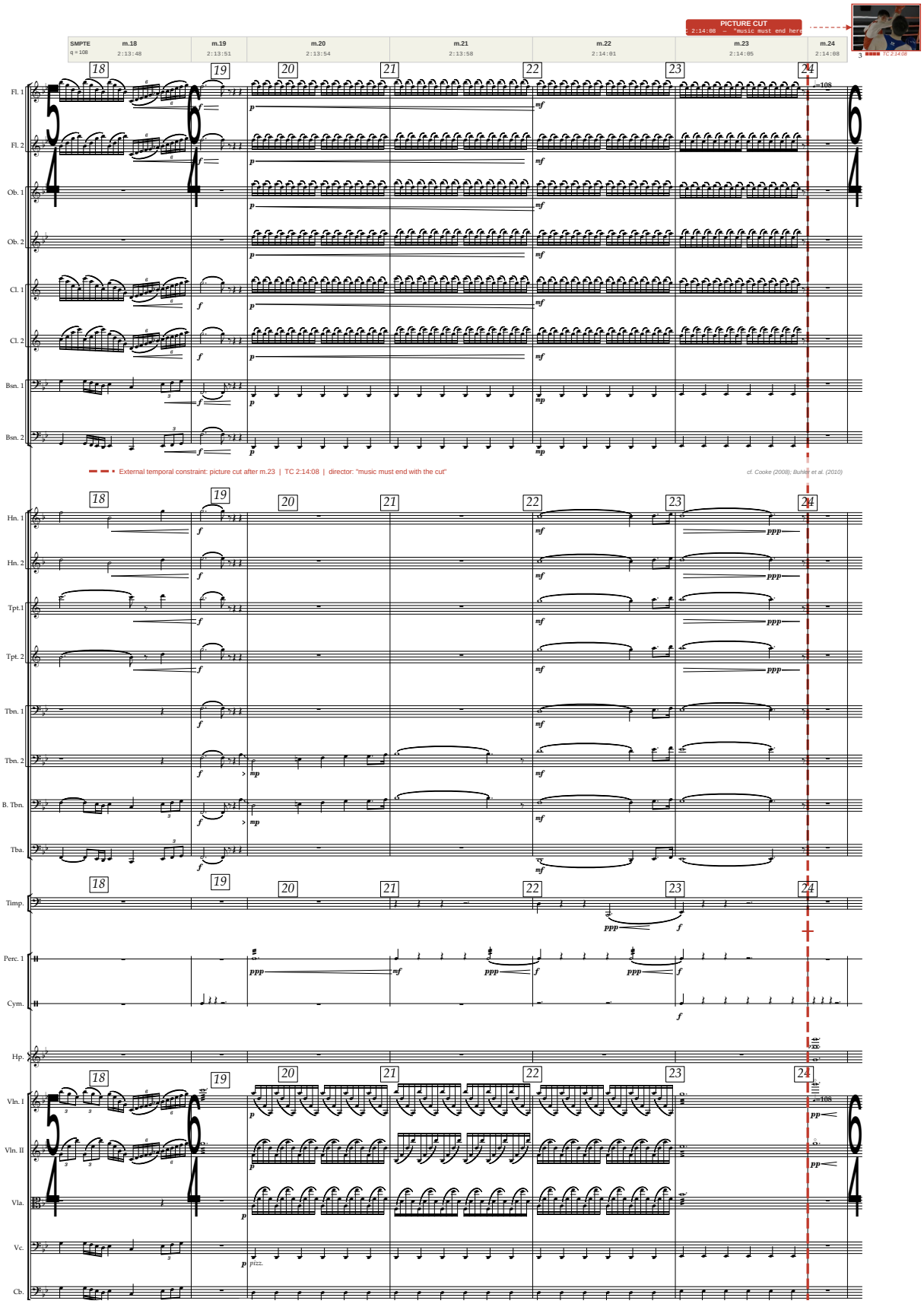

Figure 6: Temporal constraints of this kind are not reducible to harmony, texture, or phrase structure. They are defined by a visual event. Current generative systems are not well-aligned with such cross-modal constraints.

The mock-up functions as a boundary object for iterative critique: sufficiently concrete to provoke response, sufficiently malleable to be revised. The following re-

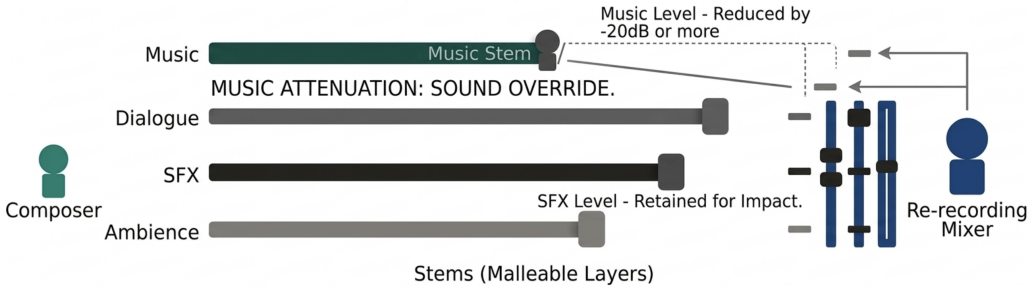

Figure 7: Phase 4: Final Mix & Dissolution of Authorship (The Sonic Integration).

constructed exchange, from a session scoring a maritime tragedy, illustrates the multi-modal nature of the constraints the composer must hold simultaneously:

Round 1: Director: "The strings are too aggressive. This person is not fighting—they are floating."

Round 2: Composer reduces texture and softens attack. Director: "Now it is too weak. It feels peaceful, but it should feel like being forgotten."

Round 3: Sound Supervisor: "The sustained cello is masking the interior monologue at 22:30. If we cannot hear the character's thoughts, the scene loses its anchor."

Each objection originates in a different modal register: the first concerns the relationship between musical gesture and bodily state; the second, the distinction between two forms of stillness; the third, the hierarchy between musical and verbal narrative. None is resolvable through musical reasoning alone, and each revision that satisfies one constraint risks violating another (Figure 5).

Implication: This multi-round, cross-modal negotiation cannot be adequately supported by systems that optimise toward a single audio-quality objective. The mock-up's function as a shared negotiation object depends on its decomposability across modal registers, which is a property that non-stem-decomposable AI outputs structurally do not preserve.

3.1.6 Phase 4: The Dub Stage & Final Mix

Timeline: Week 10–12 | **Agents:** Re-recording Mixer, Sound Designer, Director.

At the dub stage, the re-recording mixer integrates music, dialogue, and sound effects into a unified sonic track and holds ultimate authority over the relative levels of each element. A composer's thematic climax may be attenuated significantly to accommodate dialogue or ambient sound. The final work is not a music file; it is a total sonic narrative in which the score is one layer among several, and in which the composer's authorial agency is, by design, partial (Figure 7).

Implication: Any system that outputs only a single, non-decomposable audio file is not compatible with this stage of the professional workflow, regardless of its generative quality. Stem-level decomposability is a structural requirement for integration with industry mixing practice, not an optional feature.

4 DISCUSSION

4.1 The Single-Agent Assumption and Distributed Creative Intent

This paper argues that the central limitation of current AI-assisted film scoring systems is not a lack of generative capability, but a structural mismatch in how creative intent is represented within multi-agent screen scoring workflows. We identify two core limitations in existing systems. First, they are grounded in a single-agent assumption, in which creative intent is presumed to be unified, stable, and expressible through parameter-based interaction. This assumption does not hold in professional screen scoring, where intent is distributed across multiple stakeholders and continuously negotiated through iterative production processes. Even in single-user co-creative contexts, the dynamics of control and delegation are complex (Oh et al., 2018); in distributed multi-stakeholder workflows, these dynamics multiply across roles with structurally incompatible forms of authority.Second, current systems lack a representational layer for cross-modal, relational directives. In professional practice, musical decisions are frequently driven by metaphorical and narrative instructions that define relationships between sound and image, rather than properties of music in isolation.

Existing systems such as FilmComposer (Xie et al., 2025b) represent a step toward more structured film scoring pipelines through modular design and dataset construction. However, they remain grounded in a single-point optimization paradigm, in which multi-agent components are aligned toward a unified system objective rather than treated as distinct epistemic positions within a creative negotiation process.

The design problem facing AI-assisted film scoring is therefore not primarily one of generative capability, but of coordination: how to support the iterative negotiation of creative intent across a distributed network of agents who operate in different modal registers, hold different forms of authority, and communicate through different representational conventions. In reframing the problem in these terms, this paper does not claim to advance a new theory of creative collaboration. Instead, it situates existing theoretical frameworks—boundary object theory(Star and Griesemer, 1989) and distributed cognition (Hutchins, 1995) within CSCW literature—against the architectural assumptions of contemporary generative AI music systems.

4.2 Boundary Objects and Systemic Misalignment

Boundary object theory was originally developed to explain how artifacts function across communities of practice with divergent interpretive frameworks, enabling coordination without requiring consensus. In professional screen

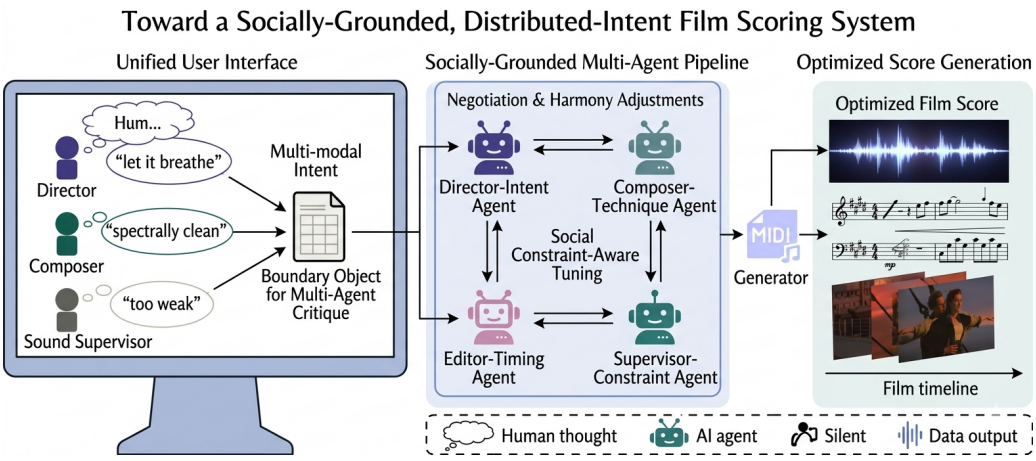

Figure 8: Distributed Creative Intent: A Socially-Grounded Model for Collaborative AI Film Scoring

scoring, the mock-up already functions as such an object: it is sufficiently concrete to support directorial evaluation, sufficiently malleable for compositional revision, and sufficiently structured to be assessed against dialogue, sound design, and editorial constraints.

However, existing CSCW frameworks did not anticipate a generative setting in which the artifact is produced by a system premised on a single-user interaction model. In such systems, only one interpretive perspective is operationalized during generation, while other stakeholders are excluded from the generative process and relegated to post-hoc adjustment. When a boundary object is produced under these conditions, it ceases to function as a shared object of negotiation and instead becomes a fixed deliverable, structurally resistant to iterative reinterpretation.

5 SYSTEM DESIGN

5.1 Implications for System Design

Future research should therefore move beyond user-centric interaction models and instead focus on representational infrastructures for distributed creative intent in audiovisual production contexts. At minimum, this implies three requirements: (1) an intent translation layer that encodes directorial instruction in relational, cross-modal terms rather than purely musical descriptors; (2) interaction interfaces that support concurrent input from multiple stakeholder roles; and (3) output representations that remain decomposable at the stem level, preserving downstream agency in the dubbing and mixing stages (Figure 8).

These are not proposals for a fully specified system. Rather, they are design requirements that emerge from applying established coordination theories to a generative AI context in which those theories have not yet been operationalized.Existing guidelines for human-AI interaction (Amershi et al., 2019) are premised on a single-user interaction model. The screen scoring workflow requires an extension of this framework to account for concurrent, multi-stakeholder input with divergent forms of authority.Recent work such as CoLyricist (Yoshida et al., 2026) demonstrates that workflow-aligned AI support can improve creative outcomes in single-user lyric writing contexts. The professional screen scoring pipeline presents a more complex case, in which workflow alignment must simultaneously accommodate multiple stakeholders with divergent modal registers and authority structures.

5.2 Temporal Sovereignty and the Limits of Endogenous Time

The second structural mismatch concerns the temporal logic governing music generation. Film musicology has long established that screen scoring is constitutively defined by exogenous temporality: Cooke (2008) and Buhler et al. (2010) characterise it as the subordination of musical time to the absolute temporal authority of the visual edit. We do not extend this concept; rather, we relocate it from the domain of film music analysis into that of computational constraint architecture, identifying the specific failure modes that result when generative systems premised on endogenous temporal logic are applied to an inherently exogenous temporal domain.

Generative music models treat temporal structure—phrase length, cadential placement, formal proportion—as emergent properties of learned musical logic, with no external authority constraining their placement. This assumption becomes a structural failure mode in screen scoring, where hit points, hard cuts, and scene boundaries encoded in SMPTE timecode are not approximation targets but fixed coordinates requiring frame-level precision. The failure is independent of output quality: a passage can be internally coherent, affectively appropriate, and stylistically well-formed, and remain unusable if its resolution falls past the hard cut at which the music must end. The composer must then re-edit the generated audio—destroying its coherence—or discard it. The system has not reduced compositional labour; it has displaced it.

This exposes not a gap in musicological theory but a translation failure between established theory and generative system design. Existing video-to-music alignment approaches treat temporal correspondence as a post-hoc optimization target, generating music according to internal logic before fitting it to the visual timeline—a sequencing architecturally reversed with respect to screen scoring practice, in which the locked cut is the primary compositional constraint. Relocating exogenous temporality as an operational system constraint yields a precise design requirement: hit points and scene boundaries must function as *generative priors*, active during generation itself, rather than as post-generative correction targets. This is not a new theoretical claim about film music. It is a new claim about what film music theory demands of generative AI architecture.

6 CONCLUSION

In this paper, we've argued that the primary barrier to AI in professional film scoring is not mere audio quality, but a fundamental mismatch between system design and industrial reality. While current models assume a "single user" with a clear musical intent, professional scoring is a collaborative "team sport" defined by messy, iterative negotiations between directors, editors, and composers.

Our analysis shows that scoring is less about "generating notes" and more about translating abstract narrative goals into sound. This process requires handling rigid constraints like frame-accurate sync and dialogue masking—that "black-box" generators simply cannot accommodate through more data alone.

Moving forward, we suggest that AI research should move from building "autonomous composers" to creating "shared workspaces." This means prioritizing stem-level control and interfaces that facilitate communication across different creative roles. Ultimately, the goal is not to build a single end-to-end model, but to build a bridge that supports the high-stakes, multi-agential reality of the film industry.

7 ACKNOWLEDGMENT

Language revision and figure illustration in this paper were assisted by AI technologies. The intellectual content, research design and main arguments remain original and human-authored.

REFERENCES

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., and Horvitz, E. (2019). Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Glasgow Scotland Uk. ACM.

Buhler, J., Neumeyer, D., and Deemer, R. (2010). *Hearing the Movies: Music and Sound in Film History*. Oxford University Press, New York.

Cooke, M. (2008). *A History of Film Music*. Cambridge University Press, 1 edition.

Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., and Défossez, A. (2023). Simple and controllable music generation. *Advances in neural information processing systems*, 36:47704–47720.

Davis, N., Hsiao, C.-P., Yashraj Singh, K., Li, L., and Magerko, B. (2016). Empirically Studying Participatory Sense-Making in Abstract Drawing with a Co-Creative Cognitive Agent. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 196–207, Sonoma California USA. ACM.

Di, S., Jiang, Z., Liu, S., Wang, Z., Zhu, L., He, Z., Liu, H., and Yan, S. (2021). Video background music generation with controllable music transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2037–2045.

Evans, Z., Carr, C., Taylor, J., Hawley, S. H., and Pons, J. (2024). Fast timing-conditioned latent audio diffusion. In *Forty-first International Conference on Machine Learning*.

Hutchins, E. (1995). *Cognition in the Wild*. MIT Press.

Ji, S., Wu, S., Wang, Z., Li, S., and Zhang, K. (2025). A comprehensive survey on generative ai for video-to-music generation. *arXiv preprint arXiv:2502.12489*.

Larsen, A. H. and Zhu, J. (2024). Ideary: Facilitating Electronic Music Creation with Generative AI. In *Designing Interactive Systems Conference*, pages 275–278, IT University of Copenhagen Denmark. ACM.

Li, J., Xu, T., Chen, X., Yao, X., Han, J., and Liu, S. (2025). Mozart's touch: a lightweight multimodal music generation framework based on pre-trained large models. In *International Conference on AI-Generated Content (AIGC 2024)*, volume 13649, pages 198–207. SPIE.

Li, S., Yang, B., Yin, C., Sun, C., Zhang, Y., Dong, W., and Li, C. (2024). Vidmusician: Video-to-music generation with semantic-rhythmic alignment via hierarchical visual features. *arXiv preprint arXiv:2412.06296*.

Lin, Y.-B., Casebeer, J., Mai, L., Mahapatra, A., Bertasius, G., and Bryan, N. J. (2026). V2m-zero: Zero-pair time-aligned video-to-music generation. *arXiv preprint arXiv:2603.11042*.

Lin, Y.-B., Tian, Y., Yang, L., Bertasius, G., and Wang, H. (2025). Vmas: Video-to-music generation via semantic alignment in web music videos. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1155–1165. IEEE.

Louie, R., Coenen, A., Huang, C. Z., Terry, M., and Cai, C. J. (2020). Novice-AI Music Co-Creation via AI-Steering Tools for Deep Generative Models. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Honolulu HI USA. ACM.

Louie, R., Engel, J., and Huang, C.-Z. A. (2022). Expressive Communication: Evaluating Developments in Generative Models and Steering Interfaces for Music Creation. In *27th International Conference on Intelligent User Interfaces*, pages 405–417, Helsinki Finland. ACM.

Oh, C., Song, J., Choi, J., Kim, S., Lee, S., and Suh, B. (2018). I Lead, You Help but Only with Enough Details: Understanding User Experience of Co-Creation with Artificial Intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Montreal QC Canada. ACM.

Star, S. L. and Griesemer, J. R. (1989). Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, 19(3):387–420.

Su, K., Li, J. Y., Huang, Q., Kuzmin, D., Lee, J., Donahue, C., Sha, F., Jansen, A., Wang, Y., Verzetti, M., et al. (2024). V2meow: Meowing to the visual beat via video-to-music generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4952–4960.

Tang, X., Lei, X., Zhu, C., Chen, S., Yuan, R., Li, Y., Oh, C., Zhang, G., Huang, W., Benetos, E., et al. (2025). Automv: An automatic multi-agent system for music video generation. *arXiv preprint arXiv:2512.12196*.

Tian, Z., Liu, Z., Yuan, R., Pan, J., Liu, Q., Tan, X., Chen, Q., Xue, W., and Guo, Y. (2025). Vidmuse: A simple video-to-music generation framework with long-short-term modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18782–18793.

Wang, B., Zhuo, L., Wang, Z., Bao, C., Chengjing, W., Nie, X., Dai, J., Han, J., Liao, Y., and Liu, S. (2024). Multimodal music generation with explicit bridges and retrieval augmentation. *arXiv preprint arXiv:2412.09428*.

Wang, Z., Bao, C., Zhuo, L., Han, J., Yue, Y., Tang, Y., Huang, V. S.-J., and Liao, Y. (2025). Vision-to-music generation: A survey. *arXiv preprint arXiv:2503.21254*.

Xie, Z., He, Q., Zhu, Y., He, Q., and Li, M. (2025a). Filmcomposer: Llm-driven music production for silent film clips. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13519–13528.

Xie, Z., He, Q., Zhu, Y., He, Q., and Li, M. (2025b). FilmComposer: LLM-Driven Music Production for Silent Film Clips. arXiv:2503.08147 [cs].

Yoshida, M., Li, B., Zhao, S., Zhou, Q., Hu, S., Chen, X. A., and Peng, N. (2026). CoLyricist: Enhancing Lyric Writing with AI through Workflow-Aligned Support. In *Proceedings of the 31st International Conference on Intelligent User Interfaces*, pages 1387–1410, Paphos Cyprus. ACM.

Young, H., Dumoulin, V., Castro, P. S., Engel, J., and Huang, C.-Z. A. (2022). Compositional Steering of Music Transformers. In *Workshops at the 27th International Conference on Intelligent User Interfaces (IUI Workshops)*.

Zhang, L. and Fuentes, M. (2025). Sonique: Video background music generation using unpaired audio-visual data. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zhuo, L., Wang, Z., Wang, B., Liao, Y., Bao, C., Peng, S., Han, S., Zhang, A., Fang, F., and Liu, S. (2023). Video background music generation: Dataset, method and evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15637–15647.