Material Biases as Compositional Resources in Text-to-Audio Models

David Piazza
Centre for Interdisciplinary Research in Music Media and Technology
Montréal, Canada
david.piazza@umontreal.ca

Abstract

Text-to-audio models shape composition through data, annotations, text encoders, and diffusion controls. We treat these mediations as material, operationalizing and combining shallow prompting and partial denoising in Stable Audio Open. A reproducible CLAP audit of 360 matched clips found that abstract prompts were closer than descriptive prompts to same-seed unconditioned controls in 83–93% of comparisons; reducing denoise from 1.0 to 0.4 lowered exact-prompt similarity in both conditioned families. These results support weaker grounding and reduced prompt adherence under this configuration, not aesthetic or compositional value. We then trace the material through montage and the realized 17-minute, fully improvised performance *Conflagrations*, whose shared corpus was accessed automatically through descriptor-driven Max/MSP mosaicking and manually through Mosaïque in Ableton Live. The contribution is an integrated material-system analysis, matched audit, and situated account of compositional reuse.

1 INTRODUCTION

Text-to-audio (TTA) systems present prompt-to-sound generation as a relatively direct operation, even though their outputs are structured by training corpora, metadata practices, model architectures, and sampling controls. The sounds they produce reflect specific material conditions and statistical regularities learned during training (Coelho, 2025b; McCormack et al., 2023). For composition, this shifts attention away from output fidelity alone and toward how those conditions can be interrogated, situated, and re-used within practice (Ben-Tal et al., 2021; Gioti et al., 2022; Kaila and Sturm, 2024).

We propose that the biases of a TTA system constitute its *material signature*. This signature is the practice-facing structure of recurrent associations among prompt text, metadata, latent audio features, and generated sound behaviours. It is material insofar as it organizes what kinds of sounds become reachable, adjacent, or difficult to reach under particular prompting and sampling regimes. Our hypothesis is that a practice organized around a model's constitutive limits yields different music than one organized around its apparent capabilities. This orientation draws on Simondon's account of technical individuation, in which a technical object is understood through its genesis and its *associated milieu* (Simondon, 2017), with Carey providing a bridge to modular-synthesis practice (Carey, 2024).

Our investigation focuses on Stable Audio Open (Evans et al., 2024), an openly available TTA model whose training data, architecture, and weights are documented and locally deployable. Through composition of an electroacoustic performance work with the model, we analyze how its training corpus, frozen T5 text encoder, and latent diffusion process jointly produce a characteristic material signature.

We operationalize and combine two established exploratory controls: *Shallow Prompting*, which uses abstract, polysemic, and willfully divergent text to probe how strongly language anchors generation away from default tendencies, and *Partial Denoising*, which starts from an intermediate point in the sampler schedule. Implemented in ComfyUI, these controls informed the production of material subsequently reorganized in *Conflagrations*. The realized concert system coupled live Buchla 200e synthesis in quadraphonic space with automatic descriptor-driven mosaicking in Max/MSP and manual corpus navigation in Ableton Live.

We make three contributions:

1. A material-system account of Stable Audio Open shaped by training data, annotation practices, text encoding, and diffusion controls.
2. A reproducible matched protocol for testing prompt specificity and denoising depth.
3. A situated account of reusing the resulting material in the performed realization of *Conflagrations*.

The companion page includes audio examples, the ComfyUI workflow, generation logs, public audit artifacts, and documentation of the realized concert system: https://davidpiazza.github.io/AIMC2026-1/.

2 RELATED WORK AND CONTEXT

2.1 Text-to-Audio Generation

Contemporary TTA systems demonstrate strong text-to-audio coherence across musical and sonic domains. While other models like MusicLM (Agostinelli et al., 2023) and MusicGen (Copet et al., 2023) represent major advances in text-conditioned music generation, Stable Audio Open (Evans et al., 2024) is particularly relevant to artistic research because it is openly available, locally deployable, and trained on documented Creative Commons–licensed audio.

These systems inherit design assumptions from diffusion modelling: Denoising Diffusion Probabilistic Models

(DDPM) generate data by progressively denoising a noisy signal (Ho et al., 2020), and latent diffusion architectures shift generation to compressed latent spaces for efficiency (Rombach et al., 2022). For composition, this implies that interventions in conditioning, denoising schedules, and latent initialization are the operations a composer may use to influence sound morphologies and the perceptual ambiguity of the model's outputs.

2.2 Technical Individuation

Simondon's account of technical individuation is our primary theoretical source (Simondon, 2017); Carey's application to modular synthesis provides a musical-practice bridge (Carey, 2024, 2023). A technical object is understood through its genesis rather than as a fixed instrument. Concretization names the movement from abstract functions toward an internally resonant technical ensemble, while an associated milieu is the relational field through which that ensemble functions.

Mapping these concepts onto generative audio requires caution. We treat the training data, annotations, and training procedures as genetic conditions of the model, and T5, the Diffusion Transformer (DiT), and the Variational Autoencoder (VAE) as a coupled technical ensemble. This account does not imply that a trained network is equivalent to Simondon's canonical machines; it rather seeks to identify relations that are otherwise hidden by the prompt-to-output interface.

The performance associated a different milieu: performer, Buchla 200e, quadraphonic space, Max/MSP Data Knot objects, Mosaïque, and monome arc. Its different corpus-access paths organized distinct but concurrent performance possibilities. The generated files were not active model inference but materials carried from the model's genetic conditions into this later ensemble.

2.3 Practice-Led AI Music

Process-centered music-AI research evaluates systems through engagements with composers, performers, and audiences rather than through fidelity metrics alone (Ben-Tal et al., 2021). Gioti et al. borrowing from Deleuze and Guattari, model ML-assisted creation as a heterogeneous assemblage of nonlinear human-machine decisions (Gioti et al., 2022), while Dahlstedt argues that creativity gains meaning through entangled interaction in a creative process extended in time rather than interaction with a system that merely mimics the artifacts of that process (Dahlstedt, 2026). Vigliensoni and Fiebrink emphasize that large-model defaults can constrain artist intention if interaction design remains too abstract, as is often the case with TTA systems (Vigliensoni and Fiebrink, 2024). Coelho's empirical study of these systems shows that descriptor constellations can micro-locate stable stylistic regions and inducible artist-proximate behaviour, which holds implications for attribution and consent (Coelho, 2025b). Taken together, these perspectives motivate the approach we adopt here: we treat the model's biases and affordances as material to be investigated, mapped, and re-appropriated through compositional practice.

This also places the paper near creative forms of model auditing and probing. In this context, auditing means a situated attempt to reverse-engineer how a system behaves under musically relevant interactions; the CLAP audit below therefore supports a practice-led method rather than replacing listening with a metric. It asks whether the prompt and denoising regimes used compositionally also leave measurable traces in a separate audio–text representation.

3 STABLE AUDIO OPEN AS MATERIAL SYSTEM

3.1 Architecture

Stable Audio Open is built on three main components (Evans et al., 2024). A VAE operates directly on audio waveforms, compressing stereo audio at a 21.5 Hz latent rate: 47.6 seconds of audio becomes a sequence of approximately 1024 latent frames with 64 channels. A DiT is trained to generate latent sequences from noise, conditioned on text embeddings. A frozen T5-base text encoder converts text prompts into contextualized token embeddings, which condition the DiT through cross-attention.

The architectural separation between these components is consequential for creative appropriation (Sturm and Ben-Tal, 2021). The T5 encoder was pre-trained on a large web text corpus and is used frozen in Stable Audio Open; it was not fine-tuned on audio data (Raffel et al., 2020; Evans et al., 2024). T5 produces contextualized token representations, but because the encoder itself never sees audio during pre-training or during Stable Audio Open's training, the structure of the conditioning signal is shaped by textual rather than sonic regularities: a prompt's representation reflects how its words co-occur and relate in web text, not how their referents sound. The DiT must then bridge this linguistically organized conditioning signal to the sonically organized latent space learned by the VAE, which itself only learns about audio. The bridging is necessarily partial and shaped by the training data: the model learns strong text-to-audio correlations only for the distribution of text-audio pairs it has seen during training.

3.2 Training Data and Distributional Biases

The model is trained on 472,618 recordings from Freesound.org (6,330 hours) and 13,874 tracks from the Free Music Archive (970 hours), all under Creative Commons licensing (Evans et al., 2024). The text conditioning is derived from randomly assembled subsets of metadata tags, titles, and descriptions associated with each recording.

These sources carry specific distributional biases. In the Freesound subset, categories such as "multisample," "single-note," and "synthesizer" dominate. In the FMA subset, "instrumental," "electronic," and "experimental" genres predominate, while categories such as "classical," and "indie-rock" occupy the long tail. These distributions shape what the model has learned to associate with particular textual descriptors: a prompt referencing "ambient" sound will activate correspondences learned primarily from the kinds of "ambient" recordings that Freesound contributors uploaded and tagged. The model thus produces outputs shaped by the recordings and annotations through which that category was learned. Because text annotations are drawn from user-generated metadata—tags assigned by uploaders, titles chosen for searchability, descriptions written for community sharing and legibility—they reflect the folksonomy of those platforms rather than any systematic, perceptually-grounded taxonomy. The resulting model's interface between text and sound is therefore doubly mediated: first by the audio that contributors chose to record, upload, and license under Creative Commons, and sec-

ondly by the English-language descriptive conventions of those communities.

3.3 Known Behavioural Properties

Evans et al. report several behavioural properties that follow from these material conditions (Evans et al., 2024): the model does not produce intelligible speech, is biased toward isolated sound events, struggles with producing sustained musical structure as well as with prompts containing connectors, and performs better with English prompts. Each of these maps onto the conditions already described: the absence of speech-coherent training pairs, the Freesound corpus's bias toward short, tagged samples rather than sustained compositions, and the English-language folksonomy of the annotations. From our compositional standpoint, these are the boundaries of a specific sedimentation of data, annotation, and architecture, which we aim to make audible.

4 HEURISTICS FOR EXPOSING MATERIAL BIASES

Most user-facing TTA workflows foreground precision: the goal is to describe the desired output as accurately as possible. We contend that artistic exploration of complex systems can benefit from partial control instead, preserving enough stability to curate material and enough instability to reveal unexpected structures. In diffusion-based TTA systems, sampler family, scheduling algorithm, denoising depth, guidance scaling, prompt granularity, and seed interact in ways that resist clean isolation as independent variables. Hence, we consider a repeatable procedure for finding and returning to productive regimes can be more useful than a complete axis-by-axis parameter map.

We therefore frame the two operations below as *navigational heuristics* rather than novel controls or controlled experiments. Prompt artists already optimize specialized vocabulary, circulate prompt templates, and cultivate model glitches as styles (Chang et al., 2023; McCormack et al., 2023); semantic destabilization is also established in AI-art and music practice (Coelho, 2025a; Cros Vila and Sturm, 2025). Partial denoising follows SDEdit-style interventions in diffusion trajectories (Meng et al., 2022; Rombach et al., 2022). Our distinction is to operationalize and combine these practices as model-specific audit instruments, test them with a matched protocol, and follow the exposed material through compositional reuse.

4.1 Generation Workflow

Our generation process is a repeatable workflow for documenting and returning to productive regimes:

1. Initialize at nominal Stable Audio Open inference settings (DPM++ 3M SDE sampler, exponential scheduler, denoise $= 1.0$, CFG $= 7.0$, 50 steps, seed randomized at each pass).
2. Set prompt and lock seed.
3. Generate and listen.
4. Change one control parameter, regenerate, and compare.
5. When an interesting sound is found, unlock seed and generate to test whether the regime generalizes across seed variations.
6. If it generalizes, launch a batch rendering job at the selected regime.

The workflow probes control interactions rather than isolating independent variables or optimizing toward a universal target. Seed-locked listening enables relative comparison within a fixed latent neighborhood; seed-unlocked listening checks whether observed properties are regime-level regularities or seed-specific.

4.2 Shallow Prompting

Shallow Prompting uses low-specificity, abstract, or polysemic prompts rather than likely training descriptors such as instrument labels or genre tags. The goal is not to guarantee interpolation between recognizable categories, but to test how strongly different kinds of language can anchor the model away from default tendencies. This aligns with semantic-destabilization strategies in AI-music pedagogy, where prompt ambiguity is treated as a deliberate compositional control (Coelho, 2025a).

The compositional interest of Shallow Prompting lies in a specificity spectrum produced by prompt content and CFG scale. At one end, specific source/event descriptors with moderate CFG (around 7.0) reveal microlocations: stable, inducible regions of the kind Coelho documents in Udio's TTA model (Coelho, 2025b). At the other end, empty-prompt generation with CFG $= 0.0$ serves as an operational unconditioned control for comparison. Between these poles, generic stylistic descriptors and abstract prompts can expose weaker grounding: they may produce hybrid or ambiguous timbral identities in exploratory listening, while the matched audit below shows abstract prompts remaining close to the unconditioned control while still producing nonzero text-conditioned displacement.

4.3 Partial Denoising

Partial Denoising constrains the standard denoising trajectory (Ho et al., 2020) by starting from an intermediate point in the n-step diffusion process. The step count determines the granularity of the noise schedule, while the denoise amount sets the point from which denoising begins. Reducing the denoise amount means the model operates only over the fine-grained tail of the schedule, where large-scale structural decisions have presumably not been made. The resulting sounds have diffuse structural detail and incomplete object identity, suggesting recognizable sources without fully committing to them.

In this workflow, "partial denoising" refers specifically to the `denoise` parameter exposed by ComfyUI's `KSampler` node for Stable Audio Open. The workflow sends an `EmptyLatentAudio` source to the sampler; with fixed seed and latent initialization, reducing `denoise` truncates the effective denoising trajectory so that generation begins from a later point in the sampler schedule.

Partial Denoising thus makes the tension between seeded noise initialization and text conditioning both explicit and unresolved. At full denoising with sufficient steps, the model can usually reconcile the seeded initialization with the conditioning signal, producing outputs that generally match the prompt's semantic content. With partial denoising, the output retains broad traces of the initialization's statistical tendencies and the conditioning's semantic refinement without collapsing into either.

Table 1: Precedents and the narrower distinction of the present workflow.

Precedent	Established practice	Distinction here
Prompt optimization (McCormack et al., 2023; Chang et al., 2023)	Specialized vocabulary steers outputs toward intended content or style.	Prompt specificity is varied as an independent audit factor rather than optimized for fidelity.
Prompt templates (Chang et al., 2023)	Reusable textual structures delimit a family of outputs.	Fixed prompt slots and matched seeds permit comparisons across prompt families and denoise values.
Glitch aesthetics (Chang et al., 2023)	Model failures become recognizable artistic styles.	Instability is documented as model- and setting-specific, not claimed as a new style or universally desirable failure.
Semantic destabilization (Coelho, 2025a; Cros Vila and Sturm, 2025)	Ambiguous language disrupts direct communication with a model.	Abstract prompts are compared with descriptive and empty-prompt controls to test grounding strength.
Latent editing (Meng et al., 2022; Rombach et al., 2022)	Editing or partial-noise trajectories retain information while redirecting an image.	ComfyUI's denoise endpoint is tested in TTA with empty latent audio and fixed seeds; no novelty is claimed for truncation itself.
Model probing (Gioti et al., 2022; Coelho, 2025b)	Artistic interaction maps inducible regions and system tendencies.	A CLAP audit is paired with a situated account of how the probed audio was reused in performance.

Our experiments indicate regime-dependent behaviour rather than separable parameter causality. In our listening, denoise values around 60% can remain musically legible under some schedulers while producing nondescript noise under others, with further modulation by sampler family and prompt composition. We report these as qualitative observations rather than controlled comparisons, and therefore avoid universal claims about individual controls, reporting instead the protocol for locating morphologically interesting regimes.

In short, Shallow Prompting relaxes semantic determination to probe prompt grounding strength; Partial Denoising begins from an intermediate point in the noise schedule to expose the tension between unconditioned and conditioned behaviour in a specific latent zone. Both are available in standard ComfyUI (ComfyUI Contributors, 2026) model sampling nodes that expose diffusion controls: seed, classifier-free guidance (CFG) scale, sampler family, scheduler, denoise amount, step count, and initialization source.

4.4 Combining the Two Heuristics

When Shallow Prompting and Partial Denoising are combined, valid T5 embeddings produced by prompts that do not correspond clearly to common source, event, or genre descriptors expose how weakly grounded conditioning behaves. In exploratory listening, this produced unstable materials that did not fully stabilize as source, texture, or gesture. The matched audit below tests whether those compositional procedures also leave measurable traces in CLAP audio space: abstract prompts are compared with descriptive prompts and empty-prompt controls, while denoise values test how truncating the denoising trajectory changes their relation to the unconditioned control.

4.5 Matched Prompt–Denoise Audit

We ran a matched audit across prompt families and denoising depth. It generated 360 fifteen-second clips: three prompt families, with five prompt slots per family, six matched seeds per slot, and four denoise values (1.0, 0.7, 0.55, 0.4). The descriptive family used "Nature ambience, birds," "City soundscape, traffic, noise," "Ambient synth loop," "Cat purring, feline, kitty," and "Jazz piano chords, loop." The abstract family used "Solar ekphrasis," "Cosmic infoldings," "Colliding blackholes," "Fragile entropic structure," and "Entangled bilateral transfer functions." The unconditioned family paired each prompt slot with an empty prompt and CFG $= 0$. The unconditioned family used thirty unique seed instances per denoise level, matched by prompt slot and seed index to the conditioned families. All conditioned runs used CFG $= 8$, 60 steps, DPM++ 3M SDE GPU, and the Karras scheduler.

Each clip was embedded with the LAION-CLAP HTSAT-unfused checkpoint (Wu et al., 2023). CLAP is itself a learned audio–text model with its own dataset and annotation biases, so we use it as a comparative machine-listening instrument rather than as ground truth. For each denoise value we computed prompt-family silhouette in CLAP audio space, a 1000-sample shuffled-label null as a scale reference, centroid distance from the unconditioned family, within-family and within-prompt dispersion, audio–text similarity to each prompt text, and best-match prompt-family confusion. The public clip-level similarity data, paired endpoint table, metric CSVs, configuration, and manifest underlie all numerical results below.

Within this configuration, prompt-family separation was small and became negative at denoise $= 0.4$, although it remained above the shuffled-label reference. Descriptive centroid distance from the unconditioned family moved from 0.548 to 0.215 between denoise $= 1.0$ and 0.4, while abstract distance moved from 0.063 to 0.045. More importantly, same-seed comparisons showed abstract clips closer to their matched unconditioned controls than descriptive clips in $28/30$, $26/30$, $25/30$, and $25/30$ comparisons across denoise 1.0, 0.7, 0.55, and 0.4 (83–93%). Under these prompts and settings, abstract language therefore behaved as a weakly grounded conditioning regime rather than a distinct intermediate family.

Reducing denoise from 1.0 to 0.4 also lowered mean exact-prompt CLAP similarity from 0.428 to 0.336 for

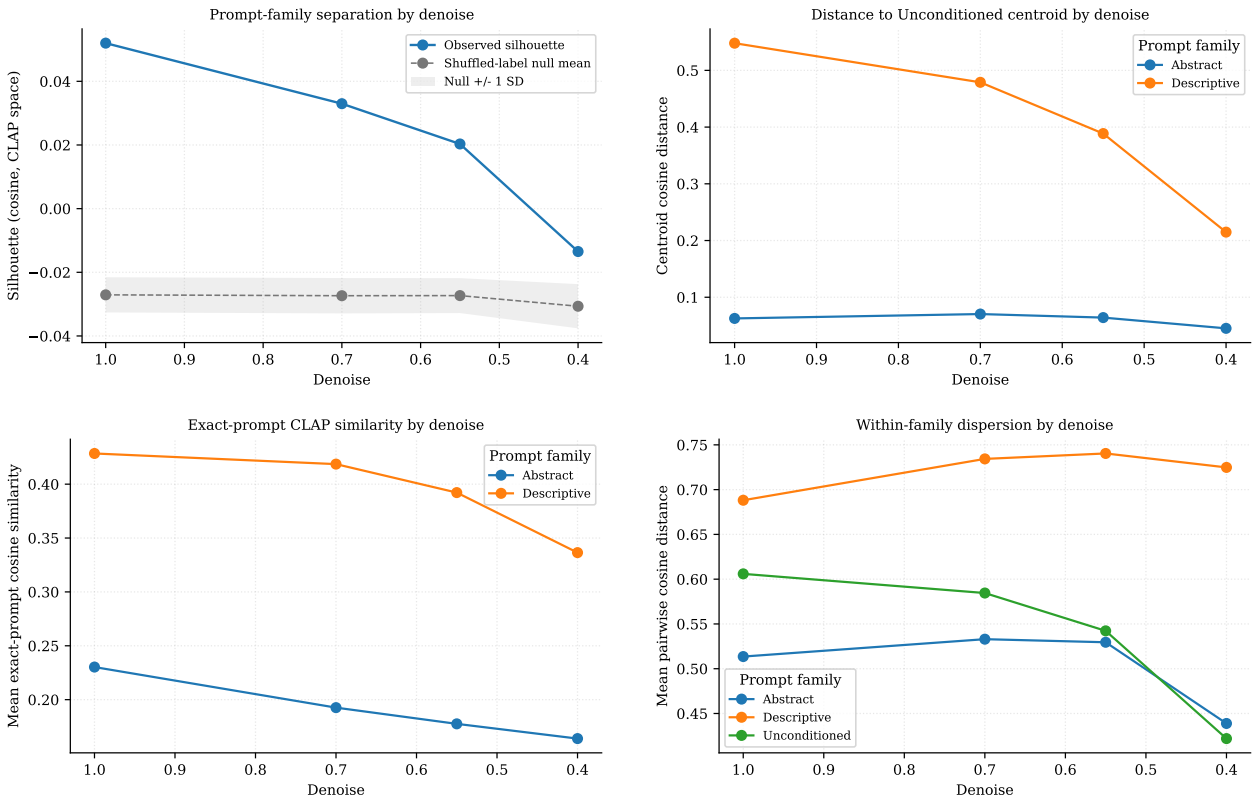

Figure 1: Matched prompt–denoise audit metrics in CLAP space. Top left: prompt-family separation against a shuffled-label reference. Top right: centroid distance from each conditioned family to the unconditioned controls. Bottom left: exact-prompt CLAP similarity, where each clip is compared only to the prompt that generated it. Bottom right: within-family dispersion.

descriptive prompts and from 0.230 to 0.164 for abstract prompts. In prompt-stratified paired endpoint comparisons, the changes were -0.092 (95% percentile interval $[-0.148, -0.037]$) and -0.066 ($[-0.106, -0.026]$), respectively; four of five prompts declined in each family. This supports reduced prompt adherence under partial denoising in this configuration.

5 PERFORMANCE: *CONFLAGRATIONS*

Conflagrations was performed on 9 May 2026 at Galerie Eisode during Festival formes·ondes. The public concert performance lasted 17 minutes. The system used two concurrent access paths into a shared collection of generated corpora, and its form was fully improvised.

5.1 Material Continuity

Material production began with iterative cycles of probing, expansion, and curation using the protocol described above. Seed-locked sessions mapped how the controls behaved under neighbouring settings; seed-unlocked expansion generated candidate sets for comparison. Stable Audio Open outputs then entered an acousmatic montage pass alongside Buchla recordings and other sources. Editing, effects, layering, and mixing transformed the generated clips before live reuse. The generated layers and fragments emerging from that pass became the corpus collection accessed by both concert paths. This maintained material continuity without presenting model outputs as isolated prompt demonstrations.

5.2 Dual-Path System

The Buchla 200e produced live synthesized audio moving through a quadraphonic field. In the automatic path, a Max/MSP patch containing Data Knot objects loaded the shared corpus collection and performed real-time descriptor analysis for audio-proximity mosaicking. Because the analysis operated across the four-channel array, movement of the Buchla sound changed which corpora were activated or queried according to its position in the field. Spatial synthesis thus provided both audible material and an indirect control signal for corpus selection.

Concurrently, a single Mosaïque instance in Ableton Live held all generated corpora at once (Thibault et al., 2025). Mosaïque provided direct navigation of corpus-based concatenative material (Schwarz, 2007; Tremblay et al., 2021). The encoders of a monome arc were mapped to Mosaïque corpus-query points, giving the performer a manual entry path independent of the Max/MSP analysis. The same transformed source collection could therefore be reached indirectly through descriptor proximity and spatial position, or directly through arc gestures.

5.3 Improvised Form and Situated Reflection

The form had no fixed scene progression or predetermined large-scale trajectory. All corpora were concurrently available through Mosaïque, while the Max/MSP path changed corpus access in response to the moving Buchla signal. Large-scale contrast emerged through improvised alternation, combination, and tension between these automatic and manual entry points.

Table 2: Claims supported by the matched audit and their evidentiary boundaries.

Claim	Measure	Result	Boundary
Abstract prompts are weakly grounded.	Same-seed conditioned-to-control distance, 30 pairs per denoise value.	Abstract clips were closer in 83–93% of comparisons.	Applies to these prompts, seeds, Stable Audio Open settings, and CLAP audio space.
Lower denoise reduces adherence.	Exact-prompt CLAP similarity paired by prompt and seed at endpoints.	Mean changes: descriptive -0.092; abstract -0.066; four of five prompts declined in each family.	Establishes reduced CLAP prompt adherence.

From the performer-author's situated perspective, having multiple control entry points helped in improvising a contrasting form. This is a practitioner observation about the realized interaction design, not an audience finding or a metric-backed conclusion.

6 DISCUSSION

6.1 Biases as Constitutive Properties

In the framework developed in sec:related, the Freesound and FMA corpora, their English-language annotations, and the procedures that coupled them to audio are genetic conditions of the model. Their traces are carried into inference through the T5–DiT–VAE ensemble. Treating these biases as constitutive properties does not make them neutral; it directs attention to the specific relations that make a sound reachable and therefore available for critique and reuse.

Shallow Prompting exposed differences in grounding strength, while Partial Denoising reduced measured adherence to the exact prompt. Partially denoised outputs are not literally Simondonian metastable states: a denoise endpoint is a sampler setting, not an apparent-equilibrium state in Simondon's sense (Simondon, 2017; Carey, 2024). "Unresolved potential" is therefore a bounded compositional analogy for material that was initially interesting and remained open to later selection, montage, and performance.

The associated milieu made that later potential concrete through performer–Buchla–quad space–Data Knot–Mosaïque–arc relations. Automatic descriptor proximity and manual corpus queries did not merely reproduce model inference; they organized two concurrent ways of encountering its transformed traces.

6.2 On Homogenization

One objection is that a practice organized around a model's biases may reinforce stylistic homogenization. Our situated proposition is that the coupling among model, prompt, curation policy, and interaction regime matters: selecting weakly grounded material and subjecting it to montage and dual-path improvisation may redirect attention away from genre-prototypical outputs. The present audit does not test homogenization, however, and the performance observation cannot demonstrate resistance to it. Establishing that broader claim would require comparative repertoire, listener, or longitudinal practice evidence beyond this case.

6.3 Model Transparency and Artistic Agency

This practice depends on model transparency. Stable Audio Open's documented training data, open architecture, and local deployability make it possible to relate observed tendencies to inspectable material conditions (Evans et al., 2024).

This has implications for artistic agency in AI-assisted composition. Creativity in human-machine systems can emerge through entangled interaction extended in time rather than from model scale or output fidelity alone (Dahlstedt, 2026; Vigliensoni and Fiebrink, 2024). Our framework adds material-level understanding to interface control: agency includes interrogating and re-appropriating the conditions that produce outputs, not merely selecting them.

6.4 Limitations: Authorship, Provenance, and the Limits of "Openness"

The audit reported in sec:audit is intentionally narrow. It uses one model, three prompt families, ten non-empty prompt texts, one sampler/scheduler configuration, one conditioned CFG value, and a fixed matched-seed design. Other guidance values, schedulers, models, prompt sets, or longer materials may behave differently. The audit supports weaker grounding for the abstract prompts and reduced exact-prompt adherence at lower denoise under this configuration. It does not establish greater diversity, aesthetic quality, compositional usefulness, artistic productivity, audience response, resistance to homogenization, or cross-model generality. No listener study was conducted.

Using an open, documented model with local render logs supports clearer provenance. Human decisions structured heuristic design, parameter selection, curation, montage, corpus construction, and performance interpretation. Evidence of inducible artist-proximate regions also reinforces the need to report likeness risk and attribution posture (Coelho, 2025b). Disclosure cannot be reduced to an "AI used / not used" label (Barnett, 2023); it should include prompt text, CFG, sampler, scheduler, denoise amount, seed policy, and subsequent transformation. For *Conflagrations*, the model-generated material was rendered before the concert, transformed in an acousmatic montage pass, and reused through the two corpus-access paths; no model inference occurred onstage.

The status of Stable Audio Open's training data also deserves direct acknowledgement. Stable Audio Open is documented as trained on Creative Commons–licensed material (Evans et al., 2024), but formal licensing should not be treated as equivalent to informed consent for large-scale generative model training. Freesound contributors plausibly did not anticipate training models whose outputs may compete with, displace, or stylistically flatten the practices they participated in. A workflow built on an "open" model therefore inherits a consent gap that openness alone does not close.

7 CONCLUSION

This paper has treated the biases of Stable Audio Open as compositional resources while narrowing what can be claimed from them. It operationalized and combined established exploratory controls, documented their effects with a matched CLAP audit, and traced the resulting material through montage into the performance of *Conflagrations*. The audit supports weaker grounding and reduced prompt adherence in our configuration; the artistic value of the dual-path, fully improvised concert remains situated practitioner evidence. Composition here is not prompt output as endpoint, but the auditing and re-articulation of a model's material conditions within a new associated milieu. That practice depends on transparent, locally deployable models and remains entangled with unresolved questions of consent and provenance.

ACKNOWLEDGEMENTS

Stable Audio Open generated source audio as described in the method. OpenAI Codex assisted with copy-editing and restructuring, LaTeX formatting, deterministic audit post-processing code, and companion-page markup. The author supplied the research content and retains responsibility for all analyses, interpretations, and final text. We would like to thank CIRMMT and Laboratoire formes·ondes for their support.

REFERENCES

Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., Sharifi, M., Zeghidour, N., and Frank, C. (2023). MusicLM: Generating music from text. arXiv preprint arXiv:2301.11325.

Barnett, J. (2023). The Ethical Implications of Generative Audio Models: A Systematic Literature Review. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 146–161, Montreal QC, Canada. ACM.

Ben-Tal, O., Harris, M. T., and Sturm, B. L. T. (2021). How Music AI Is Useful: Engagements with Composers, Performers and Audiences. *Leonardo*, 54(5):510–516.

Carey, B. (2023). Ergodynamics of the open machine: Material engagement and modular synthesis performance practice. *Organised Sound*, 28(2):237–247.

Carey, B. (2024). Metastable inventions: Simondonian concretisation and technical invention in modular synthesis practice. *Organised Sound*, 29(3):315–326.

Chang, M., Druga, S., Fiannaca, A. J., Vergani, P., Kulkarni, C., Cai, C. J., and Terry, M. (2023). The Prompt Artists. In *Proceedings of the 15th Conference on Creativity and Cognition*, pages 75–87, New York, NY, USA. Association for Computing Machinery.

Coelho, G. (2025a). AI in Music and Sound: Pedagogical Reflections, Post-Structuralist Approaches, and Creative Outcomes in Seminar Practice. In *Proceedings of the 2025 AI Music Creativity Conference (AIMC 2025)*, Brussels, Belgium. Zenodo. Conference paper. https://doi.org/10.5281/zenodo.16946639.

Coelho, G. (2025b). The Artist Is Present: Traces of Artists Residing and Spawning in Text-to-Audio AI. In *Proceedings of the 2025 AI Music Creativity Conference (AIMC 2025)*, Brussels, Belgium. Zenodo. Conference paper. https://doi.org/10.5281/zenodo.16946623.

ComfyUI Contributors (2026). ComfyUI. https://github.com/comfyanonymous/ComfyUI. GitHub repository. Accessed: 2026-02-09.

Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., and Defossez, A. (2023). Simple and controllable music generation. In *Advances in Neural Information Processing Systems*, volume 36, pages 47704–47720. Curran Associates, Inc.

Cros Vila, L. and Sturm, B. (2025). (Mis)Communicating with Our AI Systems. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–9, New York, NY, USA. Association for Computing Machinery.

Dahlstedt, P. (2026). Entangled interaction with minimal algorithms: Is interactive process, and not model size, the key to more creative machines? In *Generative Artificial Intelligence and Creativity*, pages 53–67. Academic Press.

Evans, Z., Parker, J. D., Carr, C. J., Zukowski, Z., Taylor, J., and Pons, J. (2024). Stable Audio Open. arXiv preprint arXiv:2407.14358.

Gioti, A.-M., Einbond, A., and Born, G. (2022). Composing the Assemblage: Probing Aesthetic and Technical Dimensions of Artistic Creation with Machine Learning. *Computer Music Journal*, 46(4):62–80.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc.

Kaila, A.-K. and Sturm, B. L. T. (2024). Agonistic Dialogue on the Value and Impact of AI Music Applications. In *Proceedings of the 2024 AI Music Creativity Conference (AIMC 2024)*, Oxford, UK. Conference paper. https://aimc2024.pubpub.org/pub/p3rww87r.

McCormack, J., Cruz Gambardella, C., Rajcic, N., Krol, S. J., Llano, M. T., and Yang, M. (2023). Is Writing Prompts Really Making Art? In Johnson, C., Rodríguez-Fernández, N., and Rebelo, S. M., editors, *Artificial Intelligence in Music, Sound, Art and Design*, pages 196–211, Cham. Springer Nature Switzerland.

Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. (2022). SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *International Conference on Learning Representations (ICLR)*. OpenReview. https://openreview.net/forum?id=aBsCjcPu_tE.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Schwarz, D. (2007). Corpus-based concatenative synthesis. *IEEE Signal Processing Magazine*, 24(2):92–104.

Simondon, G. (2017). *On the Mode of Existence of Technical Objects*. Univocal. University of Minnesota Press, Minneapolis, MN. Translated by Cécile Malaspina and John Rogove.

Sturm, B. L. T. and Ben-Tal, O. (2021). Folk the Algorithms: (Mis)Applying Artificial Intelligence to Folk Music. In Miranda, E. R., editor, *Handbook of Artificial Intelligence for Music: Foundations, Advanced Approaches, and Developments for Creativity*, pages 423–454. Springer International Publishing, Cham.

Thibault, D., Piazza, D., Allard, M., and Arseneault, M. (2025). Navigating Abstract Timbre Spaces: Instrumental Affordances of Concatenative Synthesis in Mosaïque. In *Proceedings of the 2025 AI Music Creativity Conference (AIMC 2025)*, Brussels, Belgium. Zenodo. Conference paper. https://doi.org/10.5281/zenodo.16946851.

Tremblay, P. A., Green, O., and Roma, G. (2021). Enabling programmatic data mining as musicking: The Fluid Corpus Manipulation toolkit. *Computer Music Journal*, 45(2):9–23.

Vigliensoni, G. and Fiebrink, R. (2024). Data- and interaction-driven approaches for sustained musical practices with machine learning. *Journal of New Music Research*, 53(1-2):19–32.

Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., and Dubnov, S. (2023). Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.