# World-Schema Alignment: Conditioning Data as a Compositional Strategy in Generative Audio
# Paper type: Case study

**Frederick Rodrigues**
School of Architecture Design and Planning,
The University of Sydney
Australia
fred.rodrigues@sydney.edu.au

## Abstract

Generative audio systems are increasingly deployed in sonic art, yet conditioning-based control architectures and their supporting data designs remain underexamined in relation to computational creativity (CC) and worldbuilding approaches. Unconditioned generative models can support world-like readings through curation or selection of datasets and audio output, and sonification-style mappings can overlay datasets onto synthesis engines. However, these approaches provide no explicit, model-internal representation of world variables or structured relations, leaving the world implied rather than enacted. Text-prompt steering similarly relies on interpretative language rather than rule-consistent parameter spaces aligned with an authored world model. This paper advances an approach, situated in CC discussions of representation, control, and constraint-consistent generation, in which conditioning data encodes a fictional world's rule system as an operational ontology. The conditioning space structures an invented world through variables, constraints, and relational dynamics. The trained model then renders these relations as audible, under specified conditions. Reflecting on a recent creative work that partially realised this approach, this paper offers a practice-led account of conditioning design as composition and proposes transferable design principles for aligning worldbuilding concepts with model and dataset structure, derived from the work's development and reflective analysis.

## Introduction

Generative audio has moved from a specialist research activity to a widely available cultural technique (Pons et al. 2025), with state of the art models now routinely producing convincing sound effects, soundscapes, and music from prompts and example audio (Lam et al. 2023; Evans et al. 2024; Copet et al. 2023). Yet in much contemporary sonic art practice, the dominant modes of 'control' over such models remain either through traversal of latent space (Tahiroğlu, Kastemaa, and Koli 2021; Pons et al. 2025) offering navigable variation without explicit ontological commitment, control external to the model (selection, curation or montage of media) (Scurto and Chemla–Romeu-Santos 2024), or linguistic mediation through text prompting in contemporary text-to-audio systems (Evans et al. 2024), including recent artistic works that explore prompt-based control (Azimi 2025). The gap in potential practice highlighted by this research, is that many works that use generative models engage in worldbuilding, or invite world-like readings of the sonic output (Scurto and Postel 2023; Scurto and Chemla–Romeu-Santos 2024). However, in most cases the 'world' remains implied, reconstructed by the listener from outputs, rather than enacted by a stable set of variables and relations that constrain what the system can and cannot do.

By "worldbuilding," this paper refers to the artistic construction of an encompassing situation or domain within which sounds are framed as belonging together and as emerging from conditions (Martin and Sneegas 2020). A world, in this sense, need not be realistic, complete, or consistent with the physical universe; it may be partial, abstract, or speculative. What distinguishes worldbuilding from the curation or orchestration of discrete sounds is that it proposes a structured domain of relations between entities, conditions, and constraints within a defined system (Ekman and Taylor 2016). In computational terms, this resembles the specification of a possibility space whose dimensions and dependencies govern how a generative system can behave. It establishes a map of possible conditions, behaviours, or perspectives through which an audience can situate their listening. Worldbuilding therefore concerns the construction of an experiential domain defined by relational principles, however minimal or fictional, rather than relying on aesthetic coherence or interpretative association alone to produce a sense of situatedness.

This paper addresses a structural gap in much generative audio practice: while world-like readings are often produced through outputs and framing, these worlds' organising relations are seldom specified as model-internal variables and constraints (Scurto and Postel 2023; Scurto and Chemla–Romeu-Santos 2024; Tahiroğlu, Kastemaa, and Koli 2021). It does so by treating conditioning (Mirza and Osindero 2014) in generative audio not as an auxiliary steering mechanism, but as a site of representational commitment. When a work constructs a world, understood as a domain organised by relational principles, conditioning mechanisms offer a means of articulating those relations explicitly as structured variables, constraints, and dependencies, rather than relying solely on interpretative coherence. In

this framing, conditioning becomes a compositional tool, a mechanism through which a world's organising logic can be embedded within the generative process itself. In computational creativity terms, this shifts attention from the production of novelty and toward the shaping of structured possibility, foregrounding representation, control, and constraint-consistent generation (Wiggins 2006).

The limitations of the prevailing paradigms are particularly visible in works that rely on generative models without explicit conditioning schemas. Many contemporary artistic systems employ models such as RAVE, a variational autoencoder widely adopted in creative audio practice (Caillon and Esling 2021). These systems can produce coherent and stylistically consistent audio, yet they cannot internally represent variables corresponding to world-level dimensions. Variation occurs within a learned latent space rather than along explicitly defined axes of relational conditions. Text-to-audio systems can name world variables through prompting (Evans et al. 2024), but linguistic description does not guarantee their operational encoding as stable, rule-consistent parameters, and numerical specificity often degrades under tokenisation (Akhtar et al. 2023). These approaches are not flawed in themselves; they become limiting only when the artistic aim is not only stylistic exploration but the enactment of a structured world.

A related distinction emerges when considering data sonification practices, where artists construct detailed correspondences between datasets (representative of a world) and synthesis systems (Hermann, Hunt, and Neuhoff 2011). Such work often involves careful translation between two domains, building perceptual and conceptual links that allow one system to be experienced through another. The relational structure remains distributed across the mapped domains: the data retain their own logic, and the synthesis engine its own behavioural rules (Flowers et al. 2001). By contrast, world-schema conditioning seeks to integrate relational structure into the generative mechanism itself, so that the defined variables and dependencies shape the behaviour of the model from within rather than being translated across systems.

In the approach proposed in this paper, conditioning design functions as a form of worldbuilding through operational ontology. The conditioning schema acts as a formal specification of a world's structure, defining its variables and the dependencies that organise and respond to those variations. The claim is not that a generative model becomes a scientifically accurate simulator, but that it can act as a renderer of an authored rule system, provided the conditioning space is coherently defined and the dataset is structured so that the world relations are learnable by the model (Mirza and Osindero 2014). This shifts aspects of compositional control away from the direct manipulation of sonic outputs and toward the design of variables, constraints, and training structure. In doing so, the artist's work moves from shaping individual sounds to shaping the relational logic of the world itself.

The contribution of this paper is therefore a conceptual and methodological articulation of conditioning design as world construction. It proposes that generative systems can be evaluated not only in relation to external ground truth, but also in relation to the coherence and enactment of an authored world-schema. The compositional intervention thus operates at the level of world configuration rather than acoustic adjustment, and worldbuilding becomes an operational design procedure: the design of a parameterised domain whose internal relations are enacted by the generative process itself. Through a reflective, practice-led analysis of Synthetic Ornithology and the design of EAGLE's conditioning space, the paper derives a set of transferable principles for aligning worldbuilding concepts with model and dataset structure. The broader aim is to reposition conditioning and schema design as primary sites of creative authorship within computational creativity, particularly in practices where the central artefact is not a single output, but a structured generative capacity.

## World-Schema Alignment in Generative Systems

The preceding discussion introduced world-schema alignment as a way of understanding conditioning design as a form of world construction within generative audio systems. For this concept to be analytically useful, however, it must be specified more precisely in computational and structural terms. What does it mean for a generative system to enact a world rather than merely evoke one? What elements of a system must align in order for world variables and relations to have operational force within the generative process?

This section formalises the concept of world-schema alignment introduced above. It first defines what constitutes a world-schema, then examines how such a schema must be realised materially through dataset design, and finally clarifies the role and limits of the generative model as renderer rather than simulator. The aim is to specify the technical and structural conditions under which conditioning design can plausibly enact an authored world.

### From Implied Worlds to Enacted Worlds

Generative systems frequently produce outputs that invite world-like interpretation without explicitly encoding such a schema (Scurto and Postel 2023). Coherence may arise through stylistic consistency, curatorial framing, or audience inference. For artists interested in worldbuilding, however, conceptual consistency may be desirable: in such a scenario, the relations that organise the world should also structure the generative process that produces it (Ekman and Taylor 2016). When these relations are embedded within the system itself, the world model can exist as an operational structure independent of individual outputs, capable of continuously generating new representations of the same world without further artistic intervention. World-schema alignment therefore requires that the coordinates organising that coherence be internally represented within the generative mechanism.

When worldbuilding is implemented within a generative system, it entails the formal definition of a possibility space: a set of variables, their domains, and the relational constraints that organise the system's world. A world is enacted

when explicit variables corresponding to world dimensions are provided (by the artist, or audience or another controlling mechanism) and when variation along those dimensions produces systematic, interpretable regularities within defined bounds. The critical feature is structured dependency rather than strict determinism: repeated conditioning states should yield outputs recognisably situated within the same region of the defined possibility space, even if stochastic variation is present (Mirza and Osindero 2014).

Under this framing, conditioning becomes the structural locus of ontological commitment. The schema defines the axes of variation before generation occurs, and the model operationalises those relations in sound. Crucially, this alignment does not occur at the level of the model alone. It emerges through the coordination of three layers: the authored world schema, the structure of the dataset through which that schema becomes learnable, and the generative model that renders those relations.

In practice, these layers are rarely developed in isolation. For artists working with generative systems, the design of a world schema and the construction of a dataset would be iterative and mutually dependent processes (Scurto and Chemla–Romeu-Santos 2024). An initial schema may propose a set of world variables or dimensions, yet attempts to assemble or generate training data may reveal that variables are insufficient, too difficult to represent with clear acoustic consequences, or that additional dimensions are required. The schema may therefore be revised in step with the development of the dataset. While the three layers described below are analytically distinct, this artistic practice would likely move back and forth between them before a stable configuration is reached.

## Schema, Dataset, and Learned Rendering

World-schema alignment can therefore be understood as a three-layer alignment model linking ontology, dataset structure, and generative mechanism. Each layer plays a distinct role in determining whether world variables operate as meaningful dimensions within the generative system.

**Layer 1: Authored World Schema** This layer consists of a conceptual and creative specification of a world through explicitly defined variables, their domains, and the structural relations that organise them. In this context, a "world" may consist of any set of interacting conditions an artist chooses to define, whether realistic, speculative, partial, or internally inconsistent. These variables ultimately shape how the generative system behaves and how audiences interpret the resulting sound.

A world schema includes:

- Named variables representing world dimensions (e.g., temporal cycles, environmental factors, spatial coordinates).
- Defined ranges or bounded domains for each variable.
- Optional conceptual relations between variables, where the schema specifies how one world condition may influence another.

In artistic practice these variables are rarely defined in isolation. They are typically embedded within a conceptual framework that gives the schema meaning, situating the defined variables within a broader artistic or speculative context through which audiences interpret the generated sound (Ekman and Taylor 2016). Designing such a schema can therefore be understood as an act of composition. Like many computerised compositional tools, it is structural, abstract, and technically oriented, yet it remains deeply creative, forming part of the conceptualisation of the sonic artwork itself.

Relations between variables do not necessarily need to be encoded explicitly within the schema itself. In many cases, interactions emerge through the acoustic consequences present in the dataset and subsequently learned by the generative model. However, artists may choose to define conceptual dependencies between world variables when constructing a fictional or speculative domain. When such relations are intended to exist within the world, they must ultimately be represented through structured variation in the dataset so that their consequences become learnable by the model (Mirza and Osindero 2014).

Within the framework proposed here, these variables describe the state of the world rather than the properties of the resulting sound. They therefore function as world-level descriptors whose sonic consequences emerge indirectly through the relations encoded in the dataset and learned by the model. The schema therefore reflects a commitment to a particular ontology: it determines which dimensions of the world are considered meaningful and how variation across those dimensions may organise the behaviour of the generative system.

**Layer 2: Dataset Realisation** A world-schema remains abstract unless its variables correspond to structured variation in training data. The dataset must exhibit systematic acoustic differences aligned with changes in the conditioning variable (Engel et al. 2020). If a variable does not produce learnable differences across the dataset, it cannot function as an operational dimension of the generative space (Engel et al. 2020). This layer is an emerging compositional act in art practice involving generative models, even when structured condition values are not in use (Scurto and Chemla–Romeu-Santos 2024). In the case of world-schema aligned approaches, curating, selecting, recording, synthesising, or structuring material must be done in a way that the defined world dimensions have audible consequences (Engel et al. 2020). Simply labelling audio with fictional properties does not guarantee enactment.

Two broad technical pathways exist for dataset realisation:

1. Pre-recorded structured audio, where world variables are inferred, creatively repurposed or recontextualised from contextual metadata already aligned with recorded acoustic variation. In such cases the variables do not necessarily retain their original semantic meaning; instead, they function as coordinates within the constructed world-schema, allowing existing acoustic variation to be situated within a newly defined relational domain.

2. Procedurally constructed or synthetic datasets, where the artist generates audio under controlled parameter regimes to ensure systematic alignment between variables and sound.

The latter approach may produce tighter control over relational structure, but raises a crucial issue: why train a generative model rather than using the synthesis, or audio creation apparatus directly. The value of a learned renderer lies in its capacity to interpolate across sparsely sampled regions of the conditioning space, produce stochastic variation within defined constraints, and extend relational structure beyond discrete exemplars (Mirza and Osindero 2014; Rombach et al. 2022; Evans et al. 2024). The resulting artefact becomes not a fixed mapping between parameters and sound, but a generative field governed by the authored schema. In this sense, dataset construction becomes a second compositional layer, translating the conceptual structure of the world schema into structured acoustic variation.

Once the schema and dataset have been stabilised, model training becomes a comparatively distinct stage. The training process remains technically demanding, but its structure is largely determined by the relational design embedded in the preceding layers. In this sense, the generative model operates as a renderer of an already specified relational domain rather than as the primary site of world construction.

**Layer 3: Learned Renderer** The trained generative model maps conditioning vectors to sound (Evans et al. 2024; Mirza and Osindero 2014). It does not define ontology; it learns statistical regularities from the structured dataset. Its role is to enact, in audible form, the relations implied by the authored schema.

Conditioning may be implemented through concatenation of conditioning vectors, feature-wise modulation, cross-attention mechanisms, or other architectural strategies (Vaswani et al. 2017; Rombach et al. 2022; Perez et al. 2018). The specific implementation matters less than the structural requirement: the conditioning variables must participate directly in shaping generative behaviour rather than functioning as post hoc annotations. In this case, ontology does not reside solely in model weights, nor solely in metadata. Instead it emerges from the alignment between schema definition, dataset structure, and model capacity; the generative process both affirms and enacts the world-schema.

### Technical Constraints and Limits

World-schema conditioning does not permit arbitrary assignment of variables to audio material. A schema acquires operational force only when conditioning variables correspond to audible regularities present in the dataset and when those regularities are learnable by the model.

Several technical constraints follow from this requirement:

- Variables must correspond to measurable or constructed acoustic consequences (Engel et al. 2020).
- The dataset must exhibit sufficient variation across the conditioning space (Mirza and Osindero 2014).
- Coverage across variable ranges must be balanced (Mirza and Osindero 2014).
- Architectural capacity must be sufficient to model dependencies between variables without collapsing them into trivial correlations (Rombach et al. 2022).
- Overfitting must be mitigated to prevent memorisation of narrow correlations (Goodfellow, Bengio, and Courville 2016).

Without such alignment, the system risks functioning as an indexed archive rather than a structured generative world. World-schema conditioning therefore requires coherence between ontology, dataset design, and model implementation. When these layers align, world variables operate as meaningful dimensions of variation within the system, allowing a generative model to enact an authored relational domain rather than merely evoke one.

This coordination defines what this paper terms world-schema alignment: a design framework for conditioning-based generative systems in which ontology, dataset construction, and generative architecture are deliberately co-designed. Within this framework, conditioning is not simply a control interface applied to a trained model, but a structural mechanism through which an authored world becomes generative. The model then functions not as the origin of that world, but as its renderer, producing audible outcomes situated within the relational domain defined by the schema.

Methodologically, the paper develops world-schema alignment as a conceptual framework and then examines its operation through reflective, practice-led analysis of Synthetic Ornithology. The case is not presented as validation of ecological claims, but as an account of how schema design, dataset structure, and conditioning-integrated generation can be composed to enact an authored world, and where that alignment becomes fragile.

## Operationalising World-Schema Alignment

The framework described above can be examined in practice through the development of *Synthetic Ornithology*, an interactive installation built around a conditioning-based generative model, EAGLE. This section analyses Synthetic Ornithology(Rodrigues 2025) as an operational instantiation of world-schema alignment. The focus is on the system's three aligned layers: how the conditioning schema was specified, how dataset structure supported relational learnability, and how the EAGLE model rendered those relations in sound. The discussion also identifies points where alignment weakened in practice, clarifying both the practical requirements and the limits of conditioning design as worldbuilding.

The approach is motivated by, and grounded in, the practice-led development of Synthetic Ornithology, an interactive installation that uses a bespoke generative model Environmental Audio Generation for Localised Ecologies (EAGLE) to generate birdsong-focused soundscapes from speculative climate scenarios (Rodrigues 2025). The system is framed as a plausible 'forecasting' interface, where audiences choose a location, date, and climate conditions to hear a simulated sonic outcome. The work intentionally

occupies the conceptual space between scientific simulation and speculative fiction: both present rule-governed systems and invite exploration of possible futures (Frigg 2010). The installation is therefore positioned as neither empirical simulator nor pure fiction, but as a parafictional construct whose credibility depends on the apparent structural coherence of the world it enacts (Lambert-Beatty 2009).

## Authored World Schema

Synthetic Ornithology operationalises world-schema alignment through a deliberately reduced environmental ontology. The authored conditioning schema is structured around non-sonic world variables that function as coordinates of the proposed and simplified world. The schema includes location and time, operationalised as latitude, longitude, minute-of-day, and day-of-year, alongside climate related variables; temperature, humidity, pressure, and wind speed. These parameters are not ancillary metadata but the primary axes along which the generative space is organised. The dataset is curated so that variation across these coordinates corresponds to systematic variation in acoustic outcomes, allowing the model to internalise their relational influence.

These variables function as coordinates of an abstracted ecological domain. They do not describe acoustic properties directly; rather, they specify environmental conditions under which acoustic regularities may emerge. In this sense, the conditioning variables are explicit descriptors of the constructed world, while their influence on sound remains indirect. This dual relationship between world description and sonic consequence is central to the idea of world-schema alignment: conditioning parameters describe the state of the world, while the model learns how those states give rise to audible structure. The schema therefore encodes a claim: that systematic variation in birdsong-focused soundscapes can be organised along these axes.

The enacted world is radically simplified. A small set of environmental variables stands in for the immense complexity of ecological systems, population dynamics, interspecies interaction, and distributed simultaneity. The system cannot plausibly function as a predictive ecological model; the interface and setting borrow simulation rhetoric as an aesthetic frame, not as an epistemic claim. Instead, it constructs a reduced but internally coherent relational domain whose plausibility derives from structured internal logic rather than empirical completeness.

This reduction is deliberate. The schema does not attempt to capture ecological reality in full, but to define a set of environmental dimensions sufficient to describe the constructed world in which the generative system operates. It is precisely this reduction that enables the work to adopt the form of simulation while functioning as an act of world-building. The model renders the audible consequences of variation across these dimensions without claiming to reproduce ecological dynamics.

## Dataset Realisation and Relational Learnability

The schema acquires operational force only through dataset structure. Synthetic Ornithology uses a curated corpus of birdsong-focused soundscape recordings sourced from archival repositories and annotated with temporal, geospatial, and climate metadata.

The dataset was sourced from xeno-canto[1] and the Macaulay Library[2].. For this research, only entries from Australia and with complete timestamps and locations were used. To minimise variation in format and audio levels, all recordings were converted to 16-bit 44.1 kHz WAV format, and a DC offset removal filter, a high-pass filter at 60 Hz with a 24 dB/octave roll off, and normalisation to -0.1 dBFS were applied to all entries. Climate data for each entry was collected via the OpenWeatherMaps API, using the GPS locations and timestamps supplied with the archival material. The processed dataset is presented as a navigable archive at audioweather.com[3].

The dataset is structured so that variation across the conditioning space corresponds to systematic acoustic differences. Three constraints govern this realisation:

- ***Relational density:*** Each conditioning dimension must exhibit sufficient variation across the dataset for the model to detect regularities.
- ***Cross-variable distribution:*** Conditioning variables must not be perfectly correlated in the dataset. If temperature only appears within narrow latitudinal bands, the model cannot disentangle spatial from climatic influence.
- ***Audible consequence:*** A variable is only operational if variation across it corresponds to measurable or perceptible acoustic change, such as shifts in species presence, density, or spectral energy distribution.

These conditions ensure that the variables defined in the authored schema correspond to learnable acoustic structure within the dataset, allowing the conditioning space to function as an operational domain rather than a purely descriptive annotation layer.

## Learned Rendering in EAGLE

EAGLE implements conditioning through numerically encoded world variables integrated directly into the generative architecture. Conditioning vectors representing the environmental schema are embedded and supplied to the model during training and inference, allowing variation along these dimensions to participate in shaping generative output.

Unlike text-prompted systems that rely on linguistic tokenisation, EAGLE operates entirely on continuous numeric conditioning. This allows precise control over scalar variables such as temperature or time-of-day and avoids the discretisation inherent in token-based prompts. Conditioning is therefore less dependent on linguistic interpretation and more directly embedded as a structural component of the generative process.

The EAGLE model architecture is based on Stable Audio Tools (Evans et al. 2024) and has two key components: an audio encoder-decoder implemented as a Generative Adversarial Network with Residual Vector Quantisation

[1] https://xeno-canto.org
[2] https://www.macaulaylibrary.org
[3] https://audioweather.com/

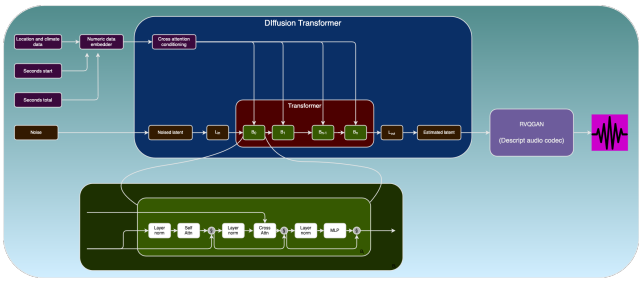


Figure 1: A flow diagram of the EAGLE model architecture for generating audio.

(RVQGAN), based on the Descript Audio Codec (Kumar et al. 2023); and a diffusion transformer with numeric conditioning supplied through a data embedder. EAGLE uses a cross-attention mechanism (ensuring the consistent application of conditioning data across diffusion steps and sequences) to apply conditioning data with a Classifier-Free Guidance (CFG) scale. The CFG scale modulates the influence of conditioning data during generation, allowing outputs to be more or less strongly constrained by the supplied environmental variables.

Figure 1 illustrates the EAGLE architecture during generation. On the left, the conditioning parameters are used to guide the denoising of Gaussian noise. The diffusion transformer iteratively applies learned transformations aligned with the conditioning data, producing a latent representation that is passed to the RVQGAN and decoded into audio.

The model learns statistical correlations between conditioning vectors and acoustic structure. It does not encode causal ecological laws. Its role is to render structured regularities present in the dataset such that:

- Repeated conditioning states tended to produce outputs recognisably situated within the same region of the conditioning domain.
- Variation within that region remains stochastic rather than fixed.
- Gradual changes in conditioning values produce continuous shifts in acoustic detail.

Within this framework the model does not define the ontology. The ontology is authored at the schema level and realised through dataset structure. The model acts as a learned renderer of that structured domain, producing audible outcomes whose variation is shaped by the relational logic encoded across the conditioning variables.

When schema design, dataset structure, and conditioning-integrated generation remain aligned, the system does not merely evoke a world through stylistic suggestion. It can enact a relational domain whose variation is organised by the authored schema.

## Misalignment, Fragility, and Technical Limits

In Synthetic Ornithology, the system enacted the authored schema with varying degrees of fidelity, shaped by dataset structure, metadata quality, and the way conditioning was integrated into the architecture. From a technical standpoint, misalignment described the points at which the nominated world variables stopped functioning as reliable dimensions of variation. From an artistic standpoint, however, misalignment did not necessarily register as failure. Because the work operated in speculative worldbuilding rather than empirical simulation, there was no external ground truth against which the world's sonic behaviour could be verified.

In practice, some outputs appeared to reflect weak or ambiguous correlations between proposed climate conditions and audible features, and some speculative parameter combinations likely pushed inference beyond regions well represented in the training corpus. Several interacting factors contributed to this fragility, including uneven distribution of the dataset across geography, season, time of day, and weather; artefacts within the recordings such as variations in microphone type, recording technique, and anthropogenic noise; climate metadata that had been reconstructed rather than directly measured; and the schema's deliberate reduction of ecological complexity. Artistically, however, these issues were acceptable: the system continued to support the parafictional framing and maintained a stable sense of worldhood even when the precise mapping between conditions and sound was uncertain.

Despite these sources of technical fragility, the work demonstrated that world-schema alignment can function effectively as a compositional and conceptual tool. The generated audio remained legible as a consequence of changing world states, allowing audiences to experience the system as enacting a coherent domain rather than producing arbitrary variation. In this sense, the success of Synthetic Ornithology was that the schema, dataset, and generative process together sustained a stable sense of worldhood. This case suggests that world-schema conditioning remains a viable and productive method for generative worldbuilding in sonic arts.

## Evaluation and samples

The EAGLE model's outputs were evaluated for fidelity and perceptual realism using a Mean Opinion Score (MOS) survey. The survey grouped participants according to their familiarity with birdsong and audio production, allowing responses from professional or specialist listeners to be compared with those from non-professional listeners. All participants listened to 15 randomly selected audio samples from a set of 9 real and 19 generated soundscapes. For each audio sample, participants responded to these five questions using a Likert scale ranging from "Strongly Disagree" to "Strongly Agree."

1. "The sounds in this recording appear natural and lifelike."
2. "This audio recording creates a sense of being in a real environment."
3. "This audio is real and not generated by an artificial intelligence model."
4. "The audio in this recording is pleasant to listen to."
5. "The audio in this recording is of high-quality and free from artefacts."

Likert responses were converted to numeric values from 1 to 5. For the naturalness, environmental presence, pleasantness, and quality questions, higher scores indicate more favourable listener responses. For the realness question, higher scores indicate that the sample was more likely to be perceived as non-synthetic. Responses from all listening events were grouped according to whether they corresponded to a real or generated file, and average scores were then calculated for each category. With 37 respondents each listening to 15 samples, the survey collected data from 555 listening events.

The similarities in responses to real and generated samples suggest that the model's outputs were perceived as broadly comparable to real recordings within the terms of the survey. In the context of this paper, these results support the claim that EAGLE can produce outputs that sustain the perceptual plausibility of the world it renders.

Table 1 presents the survey results. Professional respondents showed a similar response pattern to non-professional respondents, with generated audio scoring close to real samples across most criteria. This evaluation is included not as a comprehensive technical validation of the model, but as evidence that its outputs can maintain a stable sense of worldhood within the creative work.

| | Section | Natural. | Env. | Real. | Pleas. | Quality | Average |
|---|---|---|---|---|---|---|---|
| **All** | **Real** | 4.14 | 4.14 | 3.58 | 3.37 | 3.58 | 3.76 |
| **All** | **Generated** | 3.94 | 3.95 | 3.47 | 3.58 | 3.30 | 3.65 |
| **Pro** | **Real** | 4.24 | 4.46 | 3.70 | 3.91 | 4.12 | 4.08 |
| **Pro** | **Generated** | 4.02 | 4.10 | 3.62 | 3.80 | 4.02 | 3.91 |

Table 1: Mean listener ratings from the MOS survey comparing real and generated soundscape samples.

Within the creative work, the model operates continuously and the generated output varies constantly. Examples of misalignment therefore appear unpredictably. In *Synthetic Ornithology*, however, some degrees of misalignment contribute positively to the work by producing ambiguity, surprise, or moments of speculative instability. A selection of outputs illustrating the fidelity, ambiguity, and limits of the model is available at Synthetic Ornithology evaluation samples [4]. The documentation is organised into three sections:

1. Side-by-side comparisons of model outputs and source recordings, used to compare acoustic features and perceptual fidelity.
2. Notable outputs, including moments of compelling ambiguity or emergent behaviour that, in the context of *Synthetic Ornithology*, create opportunities for audience engagement and reveal the boundaries of the model's adherence to the world schema.
3. Limits and aberrations showing the model's fragility and misalignment with the schema.

[4] https://fred-dev.github.io/Synthetic_ornithology_results/

The case study presented in this research uses worldbuilding that aligns closely with an existing environment. This adds some complexity in evaluating these samples; listeners with knowledge of the real environments represented in the dataset will have specific sonic references against which these samples are evaluated. The goal of including these samples is not to compare them with the experience of a specific real place. Rather, the samples show how the model renders a consistent world, while also making visible the limits of that world's stability.

## Conditioning as Compositional Strategy

The analysis of Synthetic Ornithology demonstrates that conditioning design can function as more than a steering interface layered onto generative systems. When aligned across ontology, dataset structure, and model architecture, conditioning specifies the relational domain within which a generative system operates, within a work of sonic art (Mirza and Osindero 2014; Rombach et al. 2022). The shift proposed in this paper is conceptual: conditioning data can be understood as a world-schema rather than auxiliary metadata. In many generative systems, control parameters are treated as post-hoc selectors that steer a trained model toward desired outputs (Tahiroğlu, Kastemaa, and Koli 2021; Scurto and Postel 2023; Scurto and Chemla–Romeu-Santos 2024). In contrast, world-schema conditioning defines the axes along which variation is permitted to occur. The artist's intervention therefore takes place prior to generation, at the level of variable definition, domain bounding, and relational expectation. The model renders these relations in sound.

Under this perspective, compositional labour shifts toward structural specification. The artist designs:

- ***Schema design:*** defining world variables, domains, and intended relations within an artistic framing.
- ***Dataset realisation:*** constructing or curating material in which those variables have learnable audible consequence.
- ***Conditioning-integrated generation:*** selecting and implementing an architecture in which conditioning participates directly in generative behaviour.

The resulting work is not a fixed composition but a generative domain structured by an authored relational schema.

### Design Principles for World-Schema Alignment

Reflection on Synthetic Ornithology suggests several practical principles for artists working with conditioning-based generative systems. These principles are relevant for generative systems in which artists construct fictional, speculative, or abstract worlds whose behaviour is rule-consistent rather than stylistically implied.

***Variable Minimalism:*** Conditioning schemas should define only those dimensions that can be meaningfully enacted through dataset variation.

***Relational Density:*** Each conditioning variable must correspond to systematic and repeated variation in the dataset. Dataset construction therefore functions as an ontological act rather than neutral collection.

***Cross-Axis Decorrelation:*** Conditioning variables should not be structurally entangled within the dataset. If world dimensions covary perfectly, the model cannot distinguish their effects.

***Renderer–Ontology Separation:*** The generative model does not define the world. Ontological commitment resides in the schema and dataset structure; the model renders these relations statistically.

***Statistical Enactment:*** World-schema alignment aims for probabilistic regularity rather than deterministic mapping. Conditioning states organise regions of behaviour rather than specifying exact outcomes.

### Implications for Generative Art and Evaluation

Understanding conditioning as world-schema also reframes how generative systems may be evaluated in artistic contexts. Conventional model evaluation emphasises fidelity to ground truth, and in sonic arts contexts, perceptual realism or novelty (Evans et al. 2024; Copet et al. 2023). In world-schema systems, an additional criterion emerges: coherence relative to the authored ontology.

Evaluation may therefore consider criteria such as:

- Interpretability of conditioning regions.
- Sensitivity and continuity under parameter change.
- Constraint consistency relative to the schema.

These criteria complement rather than replace conventional metrics. A system may generate convincing sound while remaining ontologically incoherent, or it may enact a coherent relational domain despite imperfect acoustic realism. In speculative artistic contexts, the primary concern may instead be sustaining a stable sense of worldhood.

World-schema conditioning therefore positions generative systems not as simulations of external reality but as rule-governed fictional domains. The generative model acts as a statistical renderer of an authored ontology whose plausibility derives from internal consistency rather than empirical completeness.

## Conclusion

This paper has proposed conditioning design as a compositional strategy grounded in world-schema alignment. Through the analysis of Synthetic Ornithology, it has shown that conditioning can function as an explicit ontological specification rather than an auxiliary control interface. When schema definition, dataset structure, and model architecture are deliberately aligned, a generative system can enact a bounded relational domain whose behaviour is structured yet stochastic.

The contribution of this framework lies not in claiming accurate simulation, but in articulating how structured fictional worlds can be embedded within generative audio systems through conditioning design. This reframes control architecture as a primary site of creative authorship and positions dataset construction as an ontological act within computational creativity. As generative systems increasingly shape sonic art practice, the central challenge shifts from producing convincing sounds to structuring the worlds those sounds inhabit. World-schema conditioning reverses the typical generative workflow: instead of training a model and then adding control, the artist defines a world first and trains the system to render it. World-schema conditioning therefore offers one method for structuring generative systems as coherent sonic worlds.

## References


[Akhtar et al. 2023] Akhtar, M.; Shankarampeta, A.; Gupta, V.; Patil, A.; Cocarascu, O.; and Simperl, E. 2023. Exploring the Numerical Reasoning Capabilities of Language Models: A Comprehensive Analysis on Tabular Data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 15391–15405. Singapore: Association for Computational Linguistics.

[Azimi 2025] Azimi, M. 2025. (un)Stable (dis)Connection. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, 63–64. Zenodo.

[Caillon and Esling 2021] Caillon, A., and Esling, P. 2021. RAVE: A variational autoencoder for fast and high-quality neural audio synthesis. *CoRR* abs/2111.05011. arXiv: 2111.05011.

[Copet et al. 2023] Copet, J.; Kreuk, F.; Gat, I.; Remez, T.; Kant, D.; Synnaeve, G.; Adi, Y.; and Défossez, A. 2023. Simple and controllable music generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.

[Ekman and Taylor 2016] Ekman, S., and Taylor, A. I. 2016. Notes Toward a Critical Approach to Worlds and World-Building. *Fafnir: Nordic Journal of Science Fiction and Fantasy Research* 3(3):7–18.

[Engel et al. 2020] Engel, J.; Lamtharn Hantrakul; Chenjie Gu; and Roberts, A. 2020. DDSP: Differentiable Digital Signal Processing. In *International Conference on Learning Representations*.

[Evans et al. 2024] Evans, Z.; Carr, C.; Taylor, J.; Hawley, S. H.; and Pons, J. 2024. Fast timing-conditioned latent audio diffusion. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

[Flowers et al. 2001] Flowers, J. H.; Whitwer, L. E.; Grafel, D. C.; and Kotan, C. A. 2001. Sonification of Daily Weather Records: Issues of Perception, Attention and Memory in Design Choices. In *Proceedings of the 2001 International Conference on Auditory Display (ICAD)*, 222–226.

[Frigg 2010] Frigg, R. 2010. Fiction and Scientific Representation. In Frigg, R., and Hunter, M., eds., *Beyond Mimesis and Convention*, volume 262. Dordrecht: Springer Netherlands. 97–138. Series Title: Boston Studies in the Philosophy of Science.

[Goodfellow, Bengio, and Courville 2016] Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep learning*. MIT Press.

[Hermann, Hunt, and Neuhoff 2011] Hermann, T.; Hunt, A.; and Neuhoff, J. 2011. *The sonification handbook*. Berlin: Logos Verlag.

[Kumar et al. 2023] Kumar, R.; Seetharaman, P.; Luebs, A.; Kumar, I.; and Kumar, K. 2023. High-fidelity audio compression with improved rvqgan. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.

[Lam et al. 2023] Lam, M. W. Y.; Tian, Q.; Li, T.; Yin, Z.; Feng, S.; Tu, M.; Ji, Y.; Xia, R.; Ma, M.; Song, X.; Chen, J.; Wang, Y.; and Wang, Y. 2023. Efficient neural music generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.

[Lambert-Beatty 2009] Lambert-Beatty, C. 2009. Make-Believe: Parafiction and Plausibility. *October* 129:51–84.

[Martin and Sneegas 2020] Martin, J. V., and Sneegas, G. 2020. Critical Worldbuilding: Toward a Geographical Engagement with Imagined Worlds. *Literary Geographies* 6(1):1–6.

[Mirza and Osindero 2014] Mirza, M., and Osindero, S. 2014. Conditional Generative Adversarial Nets. *CoRR* abs/1411.1784. arXiv: 1411.1784.

[Perez et al. 2018] Perez, E.; Strub, F.; de Vries, H.; Dumoulin, V.; and Courville, A. 2018. FiLM: visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

[Pons et al. 2025] Pons, J.; Zukowski, Z.; Parker, J. D.; Carr, C.; Taylor, J.; and Evans, Z. 2025. Music and Artificial Intelligence: Artistic Trends. *arXiv preprint arXiv:2508.11694*.

[Rodrigues 2025] Rodrigues, F. 2025. Synthetic Ornithology: Machine learning, simulations and hyper-real soundscapes. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, 84–92. Zenodo.

[Rombach et al. 2022] Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10674–10685. New Orleans, LA, USA: IEEE.

[Scurto and Chemla–Romeu-Santos 2024] Scurto, H., and Chemla–Romeu-Santos, A. 2024. Deeply Listening Through/Out the Deepscape. In *ISEA2023 PROCEEDINGS*. Paris, France: Ecole des arts decoratifs - PSL.

[Scurto and Postel 2023] Scurto, H., and Postel, L. 2023. Soundwalking deep latent spaces. In Ortiz, M., and Marquez-Borbon, A., eds., *Proceedings of the international conference on new interfaces for musical expression*, 232–235. Number of pages: 4 tex.articleno: 33 tex.track: Papers.

[Tahiroğlu, Kastemaa, and Koli 2021] Tahiroğlu, K.; Kastemaa, M.; and Koli, O. 2021. AI-terity 2.0: An autonomous NIME featuring GANSpaceSynth deep learning model. In *Proceedings of the international conference on new interfaces for musical expression*. tex.articleno: 80.

[Vaswani et al. 2017] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 6000–6010. Red Hook, NY, USA: Curran Associates Inc.

[Wiggins 2006] Wiggins, G. A. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems* 19(7):449–458.