

BirdsongChat: A Hybrid Multi-Agent Framework for Interactive Multimodal Embodied Behavior Simulation

Callie C. Liao^{1*}, Duoduo Liao^{2*}, Ellie L. Zhang^{3*}

¹Stanford University, ²George Mason University, ³IntelliSky
ccliao@cs.stanford.edu, dliao2@gmu.edu

Abstract

Multimodal embodied systems require translating human intentions into interpretable and coordinated behaviors across heterogeneous modalities. However, existing multimodal agents often rely on implicit representations, limiting controllability and cross-modal consistency. We present a hybrid multi-agent framework for interactive multimodal behavior simulation that bridges semantic reasoning and physical execution through a Unified Parameter Representation (UPR). LLM-based reasoning agents transform multimodal inputs into UPR, which encodes behavioral states and interpretable control parameters for simulation agents generating synchronized 3D motion, spatialized soundscapes, and environmental behaviors. We develop *BirdsongChat*[†] as a prototype implementation of the proposed framework, using interactive avian behavior simulation as a testbed that tightly couples motion, vocalization, and environmental context. *BirdsongChat* is evaluated on text- and image-guided scenarios involving species, behaviors, affective states, environments, and multi-bird interactions. The system achieves normalized scores of 94.4% for cross-modal coherence, 100% for affective consistency, and 92.6% for generation consistency. These results demonstrate that an explicit intermediate representation effectively bridges semantic reasoning and physical execution, improving controllability and multimodal synchronization. The proposed framework thus offers a generalizable design principle for embodied AI systems requiring interpretable semantic-to-physical coordination across modalities, with potential applications in bio-inspired ecoacoustics, swarm robotics, virtual environments, and creative multimedia.

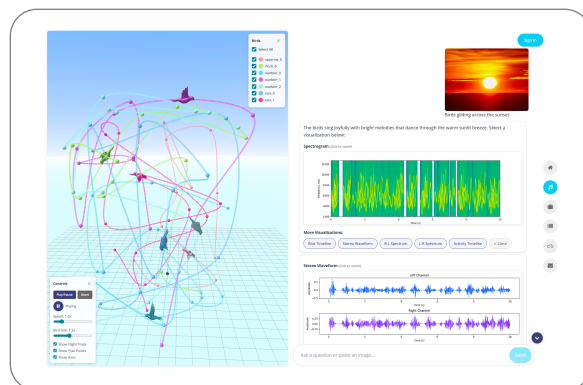


Figure 1: *BirdsongChat: Interactive Multimodal Control and Generation of Synchronized 3D Bird Motion and Soundscape*. Foundation models perform semantic interpretation of species, behavior, and context from text and images. The resulting UPR parameters control algorithmic simulators for motion, sound, and environment generation.

1 Introduction

Recent advances in Large Language Models (LLMs) and multimodal foundation models have significantly improved the ability of AI systems to interpret human instructions, reason over complex contexts, and interact with multimodal information. Vision-language models such as Flamingo (Alayrac et al., 2022), embodied models such as PaLM-E (Driess et al., 2023), and unified multimodal models such as VILA-U (Wu et al., 2025) demonstrate progress toward integrated multimodal understanding and generation. Recent generative models further achieve high-quality synthesis across image, audio, and motion domains (Zhang et al., 2024; Copet et al., 2023). These advances motivate embodied systems that translate multimodal intentions into coordinated physical behaviors.

However, a fundamental challenge remains: *how can multimodal reasoning be connected to interpretable, physically grounded, and cross-modally consistent behavior generation?* Existing embod-

*Equal contribution.

[†]The [demo video](#) is available via the provided link.

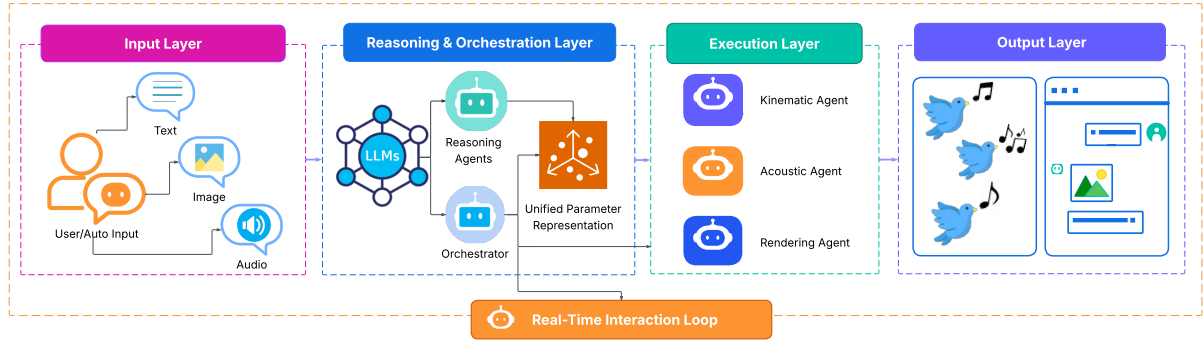


Figure 2: Overview of the BirdsongChat framework for physically grounded multimodal generation and control. The architecture bridges high-level semantic reasoning with low-level physical execution through a hierarchical design including four functional layers, maintaining interpretability, physical plausibility, and cross-modal coherence.

ied agents primarily focus on planning and action execution (Driess et al., 2023; Kim et al., 2025), while multimodal generative models often rely on implicit representations for end-to-end synthesis (Zhang et al., 2024; Copet et al., 2023). Although effective for understanding and generation, these approaches provide limited control over intermediate behavioral states and physical constraints, motivating explicit representations that bridge semantic reasoning and physical execution.

In this work, we study how multimodal semantic intentions can be transformed into coordinated physical behaviors in embodied environments using interactive avian behavior simulation as a concrete testbed in which motion, vocalization, and environmental context are tightly coupled. Building on agent-based behavioral models and procedural simulation approaches (Reynolds, 1987), we present a hybrid multi-agent framework that connects LLM-based reasoning agents and physically-based simulation agents through an interpretable Unified Parameter Representation (UPR).

The proposed framework combines reasoning agents for multimodal interpretation, orchestration, and semantic-to-physical transformation with simulation agents that execute UPR parameters to generate coordinated motion, vocalizations, spatialized soundscapes, and environment-aware scenes. The UPR serves as a shared interface between semantic representations and physical controls, preserving consistency across modalities. Unlike latent-based end-to-end generation, this architecture separates reasoning from execution, providing controllable and interpretable multimodal behavior generation. As a prototype implementation, BirdsongChat is evaluated on text- and image-guided scenarios involving species, behaviors, affective states, environ-

ments, and multi-bird interactions. Results demonstrate that an explicit intermediate representation effectively bridges semantic reasoning and physical execution, improving controllability and multimodal synchronization. The framework can further accommodate additional modalities, with broad applicability to bio-inspired ecoacoustics, swarm robotics, virtual environments, and interactive multimedia.

The main contributions are as follows:

- *Unified Parameter Representation (UPR)*: An interpretable intermediate representation that bridges high-level semantic reasoning and low-level physical execution through shared behavioral states and explicit control parameters, maintaining cross-modal consistency and controllability.
- *UPR-driven hybrid multi-agent framework*: A hybrid architecture that translates multimodal semantic reasoning into physical behaviors by encoding behavioral information into UPR to drive coordinated simulation agents, providing synchronized motion, sound, and visual scene generation and a generalizable design principle for embodied AI.
- *BirdsongChat*: An interactive avian simulation that demonstrates the proposed framework by generating synchronized avian motion, vocalizations, spatialized soundscapes, and environment-aware scenes from text and image inputs.

2 Related Work

LLM-based Embodied Agents and Grounding. LLMs have increasingly been integrated into

embodied systems as high-level reasoning and decision-making components. SayCan (Ichter et al., 2023) combines language reasoning with robot skill selection, while PaLM-E (Driess et al., 2023), RT-2 (Zitkovich et al., 2023), and vision-language-action models such as OpenVLA (Kim et al., 2025) extend multimodal reasoning toward embodied control. Recent foundation-model-based systems further explore reasoning, orchestration, and interaction in complex environments (Fu et al., 2025; Huang et al., 2024). These approaches demonstrate the potential of foundation models for instruction understanding and physical action coordination. However, they primarily focus on task planning, skill invocation, or action prediction, with limited explicit representations connecting semantic intent to continuous physical behaviors. The proposed framework extends this direction by introducing the UPR that provides an interpretable interface between semantic reasoning and physically executable simulation.

Multimodal Foundation Models and Cross-Modal Generation. Multimodal foundation models have advanced cross-modal understanding and generation across language, vision, and audio. Flamingo (Alayrac et al., 2022) and BLIP-2 (Li et al., 2023) demonstrate strong vision-language alignment, while unified models such as VILA-U (Wu et al., 2025), OmniFlow (Li et al., 2025), and MIO (Wang et al., 2025) explore joint multimodal understanding and generation. Recent generative models also achieve high-quality synthesis across modalities, including text-to-image, audio, and motion generation (Zhang et al., 2024; Yang et al., 2025; Agostinelli et al., 2023). However, these models typically rely on implicit representations optimized for end-to-end generation, providing limited control over intermediate behavioral states and physically constrained processes. Our framework addresses this limitation through an explicit semantic-to-physical interface that maps multimodal reasoning outputs to interpretable kinematic, acoustic, and environmental parameters.

Physically Grounded Simulation and Multi-Agent Systems. Physically-based simulation provides interpretable mechanisms for generating complex behaviors under explicit constraints. Agent-based models such as Boids (Reynolds, 1987) demonstrate emergent collective behavior through local interaction rules, while stochastic dynamical systems (Uhlenbeck and Ornstein, 1930; Kloeden

and Platen, 1992) provide foundations for controllable motion variation. Recent work on embodied intelligence and spatial understanding further emphasizes the importance of grounding agents in physical environments (Liu et al., 2025). Meanwhile, language-driven agent frameworks such as ReAct (Yao et al., 2023), Voyager (Wang et al., 2023), Generative Agents (Park et al., 2023), and AgentMaster (Liao et al., 2025) demonstrate the ability of LLMs to support reasoning, planning, memory, exploration, and multi-agent interaction. However, these lines of work largely emphasize either language-driven decision-making or physically grounded behavior generation, with limited focus on tightly coupling LLM reasoning with interpretable, physically based multi-agent simulation. Our framework bridges these two paradigms by integrating LLM-based reasoning agents with physically based simulation agents through the UPR, enabling synchronized and interpretable multimodal behavior generation.

3 Methodology

3.1 System Overview

The proposed hybrid multi-agent framework integrates LLM-based reasoning agents with physically-based simulation agents for interactive multimodal behavior simulation. It generates coordinated 3D motion, procedural spatialized soundscapes, and environment-aware scenes from text and image inputs. As shown in Figure 2, the architecture consists of four layers: (1) an *Input Layer* for multimodal interaction, (2) a *Reasoning and Orchestration Layer* with LLM-based agents for semantic interpretation and semantic-to-physical transformation, (3) an *Execution Layer* with specialized agents for physically-based simulation, and (4) an *Output Layer* for synchronized audiovisual rendering.

The key design principle is to separate semantic reasoning from modality-specific simulation. LLM-based reasoning agents transform multimodal inputs into semantic and behavioral representations, which are mapped to interpretable control parameters through the UPR. The UPR serves as an intermediate representation between reasoning and execution, encoding shared behavioral states that coordinate hybrid simulation agents. Motion, acoustic, and rendering agents derive modality-specific parameters from the same UPR state, maintaining temporal alignment and behavioral consistency

across generated modalities. Rather than directly synthesizing audiovisual outputs with neural generative models, the system uses UPR-driven algorithmic simulators to produce controllable and physically grounded multimodal behaviors.

3.2 Unified Parameter Representation (UPR)

3.2.1 UPR Definition

The Unified Parameter Representation (UPR) is an intermediate representation that bridges LLM-based reasoning agents and physically-based simulation agents. It encodes shared semantic, behavioral, kinematic, acoustic, and environmental states required for coordinated multimodal behavior generation:

$$\mathcal{P} = \{\mathcal{P}_{sem}, \mathcal{P}_{beh}, \mathcal{P}_{kin}, \mathcal{P}_{acous}, \mathcal{P}_{env}\}. \quad (1)$$

Here, $\mathcal{P}_{sem}$ represents semantic attributes inferred from multimodal inputs, $\mathcal{P}_{beh}$ represents intermediate behavioral states, $\mathcal{P}_{kin}$ contains kinematic control parameters, $\mathcal{P}_{acous}$ contains acoustic synthesis and spatialization parameters, and $\mathcal{P}_{env}$ represents environmental conditions.

The UPR defines a hierarchical semantic-to-physical transformation:

$$(\mathcal{P}_{sem}, \mathcal{P}_{env}) \rightarrow \mathcal{P}_{beh} \rightarrow (\mathcal{P}_{kin}, \mathcal{P}_{acous}), \quad (2)$$

where high-level multimodal descriptions are transformed into interpretable behavioral states and modality-specific control parameters. By maintaining a shared behavioral representation, the UPR preserves consistency between user intent, simulated motion, acoustic events, and environmental rendering.

For instance, given the prompt “An alarmed robin calls out from a wind-swayed tree branch,” the UPR maps semantic intent ($\mathcal{P}_{sem}$) and environmental conditions ($\mathcal{P}_{env}$) into a shared high-alert behavioral state ($\mathcal{P}_{beh}$). This intermediate state then jointly drives physical execution of both modalities, producing rapid head-turn trajectories ($\mathcal{P}_{kin}$) alongside high-pitched, spatialized alarm calls ($\mathcal{P}_{acous}$). A unified behavioral state ensures cross-modal consistency across motion, acoustics, and environmental responsiveness.

Unlike task-oriented representations that primarily encode action sequences or symbolic plans, UPR encodes semantic attributes, behavioral states, and continuous physical controls across heterogeneous modalities. Rather than merely specifying behaviors, UPR provides a shared representation

for coordinating modality-specific simulation processes through consistent behavioral states and physical controls.

3.2.2 UPR State Initialization and Evolution

During generation and interactive refinement, the UPR is maintained as a constrained behavioral state representation updated by reasoning agents according to multimodal inputs and simulation objectives. Given an initial UPR state $\mathbf{p}(t)$, multimodal observations and semantic interpretations $\mathcal{Z}(t)$, and global objectives $\mathcal{G}$, the Orchestrator Agent computes state updates:

$$\mathcal{F} : (\mathcal{Z}(t), \mathbf{p}(t), \mathcal{G}) \rightarrow \Delta\mathbf{p}, \quad (3)$$

where $\mathcal{G}$ specifies constraints including semantic consistency, multi-agent coordination, and interactive requirements.

The update consists of two components. Observation-conditioned updates,

$$\Delta\mathbf{p}_H = \mathcal{H}(\mathcal{Z}(t), \mathbf{p}(t)), \quad (4)$$

capture changes introduced by new multimodal inputs and semantic interpretation. Goal-conditioned updates,

$$\Delta\mathbf{p}_\Omega = \Omega(\mathbf{p}(t), \mathcal{G}), \quad (5)$$

maintain consistency among simulated agents, behavioral states, and environmental conditions.

The complete update is:

$$\Delta\mathbf{p} = \Delta\mathbf{p}_H + \Delta\mathbf{p}_\Omega, \quad (6)$$

followed by projection into a valid state space:

$$\mathbf{p}(t + \Delta t) = \Pi_{\mathcal{P}_{valid}}(\mathbf{p}(t) + \Delta\mathbf{p}), \quad (7)$$

where $\Pi_{\mathcal{P}_{valid}}$ enforces semantic, physical, and cross-modal constraints before the updated UPR state is interpreted by physically-based simulation agents.

In the current implementation, $\mathcal{H}$ and Ω are realized through structured prompting strategies and rule-based constraints rather than learned state dynamics.

3.3 Input Layer

The Input Layer receives multimodal instructions, including text, images, and optional audio descriptions. Inputs may specify object types, behavioral states, affective expressions, environmental conditions, and interaction scenarios. The system first extracts structured semantic attributes, which are then transformed into behavioral representations and physical control parameters through the reasoning and orchestration layers.

3.4 Reasoning and Orchestration Layer

The Reasoning and Orchestration Layer performs multimodal interpretation, behavioral reasoning, agent coordination, and semantic-to-physical transformation through two classes of specialized LLM-based agents: the Orchestrator Agent and Reasoning Agents.

3.4.1 Orchestrator Agent

The Orchestrator Agent coordinates interactions among reasoning agents and physically-based simulation agents, maintaining consistency across semantic states, behavioral representations, and execution parameters by updating the shared UPR throughout initial generation and interactive refinement.

3.4.2 Reasoning Agents

The LLM-based reasoning agents operate on multimodal inputs and maintain the shared UPR representation that connects high-level semantic intent with physically executable simulation parameters.

Interpretation Agent. The Interpretation Agent analyzes multimodal inputs and extracts semantic attributes, including type identity, environmental context, behavioral intent, affective state, and interaction information. These attributes form the semantic representation used for downstream transformation.

Transformation Agent. The Transformation Agent performs semantic-to-physical mapping by converting semantic and behavioral representations into interpretable kinematic and acoustic parameters stored in the UPR. These parameters are subsequently executed by physically-based simulation agents.

3.4.3 Semantic-to-Physical Mapping

The proposed architecture performs explicit semantic-to-physical transformation through the UPR rather than directly generating motion and sound from multimodal inputs. Semantic attributes are represented as:

$$\mathcal{P}_{sem} = \{s_{type}, s_{behavior}, s_{emotion}, s_{interaction}\}, \quad (8)$$

and environmental attributes are represented as:

$$\mathcal{P}_{env} = \{e_{scene}, e_{time}, e_{weather}\}. \quad (9)$$

These attributes are transformed into a shared behavioral state within the UPR:

$$\mathcal{P}_{beh} = F(\mathcal{P}_{sem}, \mathcal{P}_{env}), \quad (10)$$

where $\mathcal{P}_{beh}$ encodes behavior-level properties, including activity intensity, movement style, vocalization tendency, and interaction patterns.

The shared behavioral state is then interpreted by modality-specific simulation agents to derive execution parameters:

$$\mathcal{P}_{kin} = G_{kin}(\mathcal{P}_{beh}, \mathcal{P}_{sem}, \mathcal{P}_{env}), \quad (11)$$

$$\mathcal{P}_{acous} = G_{acous}(\mathcal{P}_{beh}, \mathcal{P}_{sem}, \mathcal{P}_{env}). \quad (12)$$

The resulting kinematic and acoustic parameters provide a compact, interpretable control representation for the physically-based simulation agents, governing flight speed, trajectory variation, vocalization rate, frequency characteristics, and spatial sound properties. Through the joint representation of motion and acoustics, the UPR enables the simulation agents to generate coordinated 3D behaviors and spatialized soundscapes while preserving consistency with the original multimodal intent. The low-dimensional parameterization provides direct control over both kinematic and acoustic behaviors, supporting synchronized and expressive simulation.

3.5 Execution Layer

The Execution Layer translates the shared UPR state into coordinated multimodal outputs through specialized physics-based simulation agents. Each agent derives modality-specific control parameters from the UPR and executes them using interpretable computational models with explicit control over kinematic, acoustic, and environmental variables. These agents directly execute the UPR without semantic reasoning or decision-making, producing physically grounded, reproducible, and cross-modally consistent simulations.

3.5.1 Kinematic Agent

The Kinematic Agent generates physically grounded 3D trajectories using kinematic parameters derived from the UPR state. The motion model uses stochastic continuous-time dynamics to produce smooth and controllable flight behaviors with plausible variations (Reynolds, 1987). Species characteristics, behavioral states, and environmental conditions are encoded through the kinematic parameters, providing diverse and interpretable motion generation.

Position. Each flying or moving object is modeled as a stochastic self-propelled particle in a three-dimensional environment. The state of object i at

time t is represented by its position

$$\mathbf{x}_i(t) = \begin{bmatrix} x_i(t) \\ y_i(t) \\ z_i(t) \end{bmatrix}, \quad (13)$$

where $\mathbf{x}_i(t)$ denotes the object’s position in world coordinates.

Velocity. The translational velocity of object i is represented by its horizontal speed, horizontal heading, and vertical speed:

$$\dot{\mathbf{x}}_i(t) = \begin{bmatrix} v_i(t) \cos \varphi_i(t) \\ v_i(t) \sin \varphi_i(t) \\ v_{z,i}(t) \end{bmatrix}, \quad (14)$$

where $v_i(t)$ denotes the horizontal speed, $\varphi_i(t)$ specifies the horizontal heading (azimuth), and $v_{z,i}(t)$ denotes the vertical speed. Thus, the translation motion state is characterized by $\varphi_i(t)$, $v_i(t)$, and $v_{z,i}(t)$. This formulation determines the continuous position evolution while allowing stochastic motion in three-dimensional space (Reynolds, 1987). The motion dynamics and orientation dynamics evolve according to stochastic Ornstein–Uhlenbeck process (Uhlenbeck and Ornstein, 1930).

The motion state is controlled by the UPR kinematic representation $\mathcal{P}_{kin}$ with parameter definitions and behavioral effects given in Table 4.

3.5.2 Acoustic Agent

The acoustic agent generates procedural bird-like vocalizations and multi-object spatialized soundscapes (Zhang et al., 2025) through signal-based synthesis and spatial acoustic modeling (Oppenheim and Schaffer, 2010). The agent directly controls acoustic, temporal, and spatial properties through interpretable variables, with $\mathcal{P}_{acous}$ derived from semantic, behavioral, species, and environmental attributes and coordinated with the UPR to ensure consistency between vocal behavior and simulated motion.

3D Spatialization. Each object i emits a sequence of chirps, where $c_{i,k}(t)$ denotes the waveform of the k -th chirp emitted by object i at time t . The chirp is parameterized by species-specific acoustic features. Each vocal event is spatialized along the corresponding object trajectory $\mathbf{x}_i(t)$:

$$\mathbf{s}_{i,k}(t) = A(\mathbf{x}_i(t)) c_{i,k}(t), \quad (15)$$

where $A(\mathbf{x}_i(t))$ denotes the spatial acoustic operator for distance attenuation, directional gain, and

stereo localization. Independent trajectories produce independently positioned vocal sources, forming a synchronized multi-agent 3D sound field.

3.5.3 Rendering Agent

The Rendering Agent generates environment-aware visual scenes from $\mathcal{P}_{env}$ and simulated object trajectories. Environmental attributes, including habitat, lighting, and weather conditions, are synchronized with motion and acoustic generation to maintain consistency across visual, auditory, and behavioral modalities.

3.5.4 Kinematic and Acoustic Control Parameters

Table 4 summarizes the kinematic and acoustic parameters forming a compact, interpretable representation for coordinated 3D motion and spatialized soundscapes.

3.6 Output Layer and Interactive Loop

The Output Layer integrates the outputs of the physically-based simulation agents into a synchronized multimodal scene containing 3D motion, spatialized birdsong, and environment-aware rendering.

The framework supports iterative interaction through a closed-loop process. Users can modify species, behaviors, affective states, or environmental conditions through additional multimodal instructions. The LLM-based reasoning agents update semantic and behavioral representations, the Orchestrator Agent propagates these changes through the UPR, and then the physically-based simulation agents regenerate the resulting multimodal outputs.

This architecture maintains an interpretable connection between LLM-based reasoning agents and physically-based simulation agents through the UPR. By decoupling semantic reasoning from physical execution, it provides explicit control over semantic, behavioral, kinematic, acoustic, and environmental states, forming a general framework for interactive multimodal embodied behavior generation.

4 Experiments and Evaluations

We develop *BirdsongChat* as a prototype system to implement and evaluate the proposed hybrid multi-agent framework, which integrates reasoning agents with physics-based simulation agents

to generate coordinated multi-bird motion, spatialized birdsong soundscapes, and environment-aware scenes from multimodal inputs. The evaluation examines semantic alignment, cross-modal coherence, interactive refinement, and simulation consistency across text- and image-guided scenarios.

4.1 Experimental Setup

The BirdsongChat prototype is deployed on Amazon Web Services (AWS). For these experiments, the prototype uses the Amazon Nova 2 multimodal AI foundation model to support reasoning agents for multimodal interpretation, orchestration, and semantic-to-physical transformation; however, the model can be interchanged with any other model.

The evaluation covers combinations of bird species, behavioral descriptions, affective states, and environmental conditions represented in the UPR. Table 1 summarizes the evaluated input categories.

Table 1: Category Definitions

Dimension	Values	Description
Bird Species	owl, nightingale, lark, warbler, sparrow, finch, raven	Distinct vocal patterns for species-specific acoustic assessment
Environments	night, moon, moonlight, daytime, sun, forest, storm, rain, wind, sand, home	Guides scene composition and 3D auditory spatialization
Emotional States	happy, fear, sad	Modulates motion (speed, trajectory) and sound (pitch, rhythm, intensity)

4.2 Qualitative Evaluation

BirdsongChat is evaluated across representative scenarios to assess whether the interaction between LLM-based reasoning agents and physics-based simulation agents maintains semantic consistency, behavioral plausibility, and cross-modal alignment among generated motion, sound, and environmental context (Figures 3 and 5). The evaluation focuses on affective behavior mapping, environmental grounding, multimodal consistency, stochastic variation, multi-bird coordination, and interactive refinement.

Affective Behavior Mapping. The LLM-based reasoning agents interpret affective descriptors such as *happy*, *sad*, and *fear* and transform them into behavioral representations within the UPR. Physically-based simulation agents subsequently

convert these representations into coordinated motion and acoustic variations. High-activity descriptors result in increased flight speed, trajectory variation, and denser vocalization patterns, whereas low-activity descriptors produce slower motion and sparser acoustic events. These results demonstrate the effectiveness of the UPR as an intermediate semantic-to-physical interface.

Environment-Conditioned Behaviors. Multimodal inputs, including text and images, provide environmental context that influences semantic interpretation, behavioral representation, and scene generation. Different habitats and temporal conditions lead to corresponding variations in simulated species, motion patterns, acoustic characteristics, and environmental rendering. These results demonstrate that contextual information is propagated from reasoning agents to physically-based simulation agents through the UPR.

Cross-Modal Alignment. Motion and acoustic generation are coordinated through the shared UPR representation. Changes in simulated bird trajectories influence spatial sound properties, including localization, attenuation, and temporal relationships among multiple vocal sources. The resulting audiovisual outputs maintain consistency among simulated motion, acoustic events, and environmental context.

Stochastic Variation and Simulation Stability. The physically-based simulation agents incorporate bounded stochastic processes to introduce natural variation while preserving controllability. Stochastic motion dynamics and acoustic modulation generate diverse trajectories and vocalization patterns without significant deviation from the intended semantic conditions. This enables repeatable simulation while maintaining behavioral diversity.

Multi-Bird Control and Interactive Refinement. The framework supports simultaneous simulation of multiple birds and species within shared environments. Through the UPR interface, users can modify species, behaviors, affective descriptors, and environmental conditions using conversational commands or image prompts. These updates are propagated from reasoning agents to simulation agents while preserving consistency across motion, sound, and rendering.

Table 2: Score definitions for cross-modal coherence, affective consistency, and generation consistency (0–3 scale).

Score	Cross-modal Coherence	Affective Consistency	Generation Consistency
3	Strong motion, sound, and environment alignment; physically plausible	Behavior and sound match the specified affective descriptor	Stable outputs across runs with natural variation
2	Minor mismatch but overall coherent	Partial affective correspondence	Noticeable variation with preserved semantics
1	Major mismatch or implausible behavior	Weak or unclear affective correspondence	Large variation affecting consistency
0	Failed multimodal generation	No affective correspondence	Inconsistent outputs across runs

4.3 Quantitative Evaluation

4.3.1 Quantitative Evaluation Protocol

Representative scenarios cover combinations of bird species, environmental contexts, and affective descriptors defined in Section 4.1. Each generated simulation is evaluated according to the criteria defined in Table 2. The evaluation measures three aspects of system performance: (1) *cross-modal coherence*, measuring alignment among semantic inputs, generated motion, acoustic outputs, and environmental context; (2) *affective consistency*, measuring whether simulated behavioral and acoustic characteristics correspond to the affective descriptors specified in the input; and (3) *simulation consistency*, measuring the stability of generated behaviors across repeated executions under identical prompts.

4.3.2 Evaluation Datasets and Procedures

18 prompts are evaluated, including 12 text-based prompts and 6 image-based prompts. Each prompt is executed five times, resulting in 90 generated simulations. The user-provided input consists only of the text or image prompt, while species selection, behavioral interpretation, and simulation parameters are determined by the LLM-based reasoning agents and transformed into UPR parameters. Generated simulations are manually evaluated according to the criteria defined in Table 2. Detailed per-prompt evaluation results are provided for text-based inputs in Table 5 and image-based inputs in Table 6.

4.3.3 Quantitative Results

Table 3 summarizes the quantitative evaluation results. Repeated executions evaluate semantic alignment and simulation consistency under stochastic generation. The cross-modal coherence scores demonstrate the effectiveness of the UPR in maintaining consistency between semantic descriptions, simulated trajectories, vocalization patterns, and environmental context.

For text prompts (12 prompts, 60 runs), BirdsongChat achieves perfect cross-modal coherence and affective consistency, with average scores of 3.00 (100%) for both metrics. Simulation consistency reaches 2.83 (94.4%), indicating stable behavior generation while preserving controlled stochastic variations.

For image prompts (6 prompts, 30 runs), the system maintains perfect affective consistency with a score of 3.00 (100%). Cross-modal coherence achieves 2.50 (83.3%), with lower scores mainly resulting from ambiguity in visual interpretation, including environmental understanding and species selection. Simulation consistency remains high at 2.67 (88.9%), demonstrating reliable repeated simulation under visually conditioned scenarios.

Across all evaluations (18 prompts, 90 runs), BirdsongChat achieves 2.83 (94.4%) for cross-modal coherence, 3.00 (100%) for affective consistency, and 2.78 (92.6%) for simulation consistency. These results demonstrate that the UPR effectively maintains semantic, kinematic, acoustic, and environmental alignment while supporting stable multimodal simulation across repeated executions.

4.3.4 Analysis and Discussion

The quantitative results demonstrate the feasibility of the proposed hybrid multi-agent framework for connecting multimodal reasoning with physically based simulation. The UPR provides an explicit interface between reasoning and execution, encoding behavioral states and interpretable control parameters for coordinating motion, acoustics, and environmental rendering. As a prototype implementation, BirdsongChat demonstrates this design in interactive avian behavior simulation.

Evaluation across text and image inputs shows consistent affective alignment, with 100% affective consistency for both modalities. Text inputs achieve higher cross-modal coherence (100%) and generation consistency (94.4%) than image inputs (83.3% and 88.9%), respectively. The lower image-

Table 3: Quantitative evaluation results (0–3 scale). Each prompt is executed five times to evaluate semantic alignment and generation consistency.

Metric	Average	Min	Max	Score (%)
<i>Text Prompts (12 prompts, 60 runs)</i>				
Cross-modal Coherence	3.00	3	3	100%
Affective Consistency	3.00	3	3	100%
Generation Consistency	2.83	2	3	94.4%
Overall	2.94	2	3	98.1%
<i>Image Prompts (6 prompts, 30 runs)</i>				
Cross-modal Coherence	2.50	1	3	83.3%
Affective Consistency	3.00	3	3	100%
Generation Consistency	2.67	2	3	88.9%
Overall	2.72	1	3	90.7%
<i>Overall Evaluation (18 prompts, 90 runs)</i>				
Cross-modal Coherence	2.83	1	3	94.4%
Affective Consistency	3.00	3	3	100%
Generation Consistency	2.78	2	3	92.6%
Overall	2.87	1	3	95.7%

based coherence suggests that visually complex or ambiguous environmental contexts remain challenging for semantic-to-physical grounding, while the 90.7% overall image consistency indicates that UPR remains effective for heterogeneous inputs.

These results suggest that separating semantic interpretation from physical execution through an explicit intermediate representation can support interpretable and controllable behavior generation. Although demonstrated in an avian simulation setting, the proposed design provides a potential basis for embodied systems requiring coordinated translation of multimodal intent into physical behavior.

5 Conclusion

We presented a hybrid multi-agent framework for interactive cross-modal semantic-to-physical coordination in embodied environments. The framework uses a UPR to provide an interpretable interface between multimodal reasoning agents and physically based simulation agents. BirdsongChat, a prototype implementation of this framework, demonstrates its feasibility in interactive avian behavior simulation, indicating how shared behavioral states and control parameters can coordinate motion, acoustics, and environmental rendering.

Across the evaluated text- and image-guided scenarios, the system achieves 94.4% cross-modal coherence, 100% affective consistency, and 92.6% generation consistency. The results suggest that explicit intermediate representations can help preserve semantic intent across heterogeneous modalities while enabling controllable physical execution. More broadly, the proposed framework offers a general design principle for embodied AI systems that

require interpretable coordination between multimodal reasoning and physical behavior.

6 Limitations and Future Work

Although the proposed framework provides an interpretable approach to semantic-to-physical simulation, several challenges remain.

First, the reasoning layer depends on foundation models and is therefore susceptible to semantic ambiguity and reasoning errors that may propagate to downstream simulation. Future work will explore uncertainty-aware reasoning, verification mechanisms, and more robust multi-agent coordination.

Second, mapping high-level semantic descriptions to physical parameters remains challenging. While UPR provides an explicit interface, more adaptive representations are needed to generalize across types, behaviors, and application domains.

Third, the current evaluation is limited in scale. Our experiments provide an initial validation using 18 prompts with repeated executions and manual assessment based on predefined semantic, affective, and cross-modal consistency criteria. Larger evaluations with more diverse scenarios and independent human assessments are needed to examine behavioral plausibility, realism, and user perception.

Finally, BirdsongChat currently emphasizes controllable and perceptually plausible simulation rather than comprehensive ecological modeling. Future work should investigate richer ecological interactions, larger multi-agent environments, and real-time execution.

7 Broader Impact and Ethics

The proposed framework contributes to embodied AI research by exploring an interpretable interface between multimodal semantic reasoning and controllable physical simulation through coordinated reasoning and simulation agents. BirdsongChat, as a prototype implementation, demonstrates this approach in interactive avian behavior simulation. The framework may support future applications in bio-inspired ecoacoustics, swarm robotics, interactive simulation, virtual environments, ecological education, and creative media.

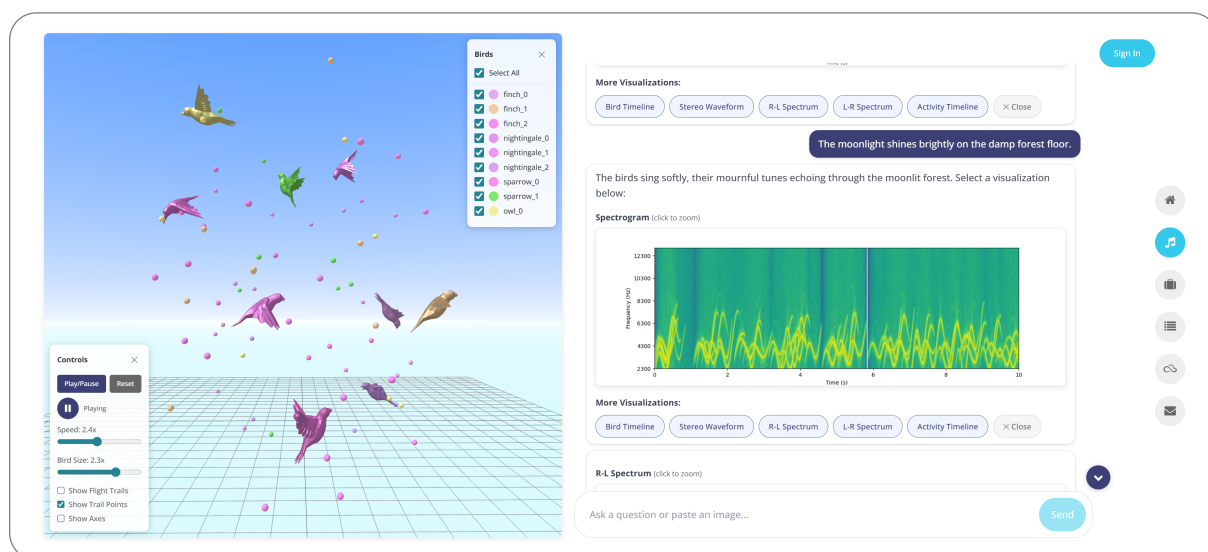
The generated bird behaviors, motion trajectories, and spatialized soundscapes are procedural simulations and should not be interpreted as faithful representations of real animals or natural ecosystems. When used in educational, ecological, or media contexts, such content should be clearly identified as synthetic. In addition, foundation-model-based reasoning agents may introduce semantic ambiguity, biases, or unintended interpretations that can propagate to physical simulation. Responsible deployment should therefore emphasize transparent agent coordination, interpretable semantic-to-physical mappings, validation of generated behaviors, and clear communication of the synthetic nature and limitations of the resulting content.

References

- Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. 2023. [MusicLM: Generating music from text](#). *Preprint*, arXiv:2301.11325.
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, and 8 others. 2022. [Flamingo: a visual language model for few-shot learning](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.
- Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Defossez. 2023. [Simple and controllable music generation](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 47704–47720. Curran Associates, Inc.
- Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, and 3 others. 2023. [PaLM-E: An embodied multimodal language model](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 8469–8488. PMLR.
- Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhao Xiong. 2025. [Scene-LLM: Extending language model for 3d visual reasoning](#). In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2195–2206.
- Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, and Zhou Zhao. 2024. [Chat-Scene: Bridging 3D scene and large language models with object identifiers](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 113991–114017. Curran Associates, Inc.
- Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, Dmitry Kalashnikov, Sergey Levine, Yao Lu, Carolina Parada, Kanishka Rao, Pierre Sermanet, Alexander T. Toshev, Vincent Vanhoucke, and 26 others. 2023. [Do As I Can, Not As I Say: Grounding language in robotic affordances](#). In *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 287–318. PMLR.
- Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P. Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. 2025. [OpenVLA: An open-source vision-language-action model](#). In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 2679–2713. PMLR.
- Peter E. Kloeden and Eckhard Platen. 1992. *Numerical Solution of Stochastic Differential Equations*, 1st edition. Springer, Berlin.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. [BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.
- Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Zichun Liao, Yusuke Kato, Kazuki Kozuka, and Aditya Grover. 2025. [Omniflow: Any-to-any generation with multi-modal rectified flows](#). In *2025*

- IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13178–13188.
- Callie C. Liao, Duoduo Liao, and Sai Surya Gadiraju. 2025. [AgentMaster: A multi-agent conversational framework using A2A and MCP protocols for multi-modal information retrieval and analysis](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 52–72, Suzhou, China. Association for Computational Linguistics.
- Huaping Liu, Di Guo, and Angelo Cangelosi. 2025. [Embodied intelligence: A synergy of morphology, action, perception and learning](#). *ACM Comput. Surv.*, 57(7).
- Alan V Oppenheim and Ronald W Schafer. 2010. *Discrete-Time Signal Processing*, 3rd edition. Pearson.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. [Generative Agents: Interactive simulacra of human behavior](#). In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST ’23, New York, NY, USA. Association for Computing Machinery.
- Craig W. Reynolds. 1987. [Flocks, herds and schools: A distributed behavioral model](#). In *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH ’87, page 25–34, New York, NY, USA. Association for Computing Machinery.
- George E. Uhlenbeck and Leonard S. Ornstein. 1930. [On the theory of the brownian motion](#). *Physical Review*, 36(5):823–841.
- Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. [Voyager: An open-ended embodied agent with large language models](#). *Preprint*, arXiv:2305.16291.
- Zekun Moore Wang, King Zhu, Chunpu Xu, Wangchunshu Zhou, Jiaheng Liu, Yibo Zhang, Jessie Wang, Ning Shi, Siyu Li, Yizhi Li, Haoran Que, Zhaoxiang Zhang, Yuanxing Zhang, Ge Zhang, Ke Xu, Jie Fu, and Wenhao Huang. 2025. [MIO: A foundation model on multimodal tokens](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5077–5099, Suzhou, China. Association for Computational Linguistics.
- Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. 2025. [VILA-U: a unified foundation model integrating visual understanding and generation](#). In *International Conference on Learning Representations*, volume 2025, pages 93620–93638.
- Chenyu Yang, Shuai Wang, Hangting Chen, Jianwei Yu, Wei Tan, Rongzhi Gu, Yaoxun Xu, Yizhi Zhou, Haina Zhu, and Haizhou Li. 2025. [SongEditor: adapting zero-shot song generation language model as a multi-task editor](#). In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI’25/IAAI’25/EAAI’25. AAAI Press.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. [ReAct: Synergizing reasoning and acting in language models](#). *Preprint*, arXiv:2210.03629.
- Ellie L. Zhang, Duoduo Liao, and Callie C. Liao. 2025. [Dynamic multi-species bird soundscape generation with acoustic patterning and 3D spatialization](#). In *2025 IEEE International Conference on Big Data (BigData)*, pages 5057–5066.
- Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. 2024. [CLAY: A controllable large-scale generative model for creating high-quality 3d assets](#). *ACM Trans. Graph.*, 43(4).
- Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, and 35 others. 2023. [RT-2: Vision-language-action models transfer web knowledge to robotic control](#). In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 2165–2183. PMLR.

A Additional BirdsongChat Examples (I)



(a) BirdsongChat generates birdsong along with 3D flight motion and a spectrogram. The interface shows bird species, their routes, and trail points. In this example, the text prompt conveys a slightly melancholic, nighttime mood, producing slower, lower-frequency sounds and featuring nocturnal species like owls and nightingales. The interface displays only trail points, hiding the flight paths for a clearer view of bird positions.



(b) This display shows the L-R spectrum, with the bottom graph highlighting frequency differences and channel bias between left and right. The bird icons appear slightly enlarged due to the collapsed GUI sliders at the bottom-left. Flight trails, trail points, and axes are hidden, and the view is zoomed in and angled slightly upward.

Figure 3: BirdsongChat interface and examples.

B Kinematic and Acoustic Control Parameters

The UPR provides low-dimensional interpretable control parameters for the physically based simulation agents.

Table 4: Kinematic and Acoustic Control Parameters

Parameter	Description	Behavioral Effect
<i>Kinematic</i>		
$\vec{x}_0$	Equilibrium position	Nominal CoM location
$\Delta\vec{x}$	Oscillation amplitudes	Path curvature
λ	Relaxation rates	Agility
σ	Noise amplitudes	Stochastic variability
v_0	Preferred horizontal speed	Target forward velocity
$v_{z,0}$	Preferred vertical speed	Target climb/descent rate
ψ_0	Equilibrium pitch	Nominal body orientation
<i>Acoustic</i>		
f_0, f_1	Chirp start/end frequencies	Frequency sweep range
d	Chirp duration	Length of vocalization
α	Trill strength	Pitch oscillation
f_{trill}	Trill rate	Vibrato speed
$A(\mathbf{p})$	Spatial gain	Stereo localization

Table 4 summarizes the kinematic and acoustic control parameters and their corresponding behavioral effects. These parameters form a low-dimensional, interpretable representation that supports synchronized, expressive, and controllable bird-like object behavior across motion, orientation, and sound.

C Text Prompt Alignment Scores

Table 5: Text prompt alignment effectiveness with per-sample scores (0–3 scale).

ID	Text Prompt	Environment	Affect	Species	Correct	Coh.	Aff.	Gen.
1	The moonlight shines brightly on the damp forest floor.	Night, Moon, Forest	Fear/Sad	Nightingale, Owl	Yes	3	3	3
2	The backyard forest sings harmonically with the newly bloomed flowers.	Daytime, Sun, Forest	Happy	Sparrow, Finch, Raven	Yes	3	3	3
3	The birds sing happily with another under the gleaming sun.	Sun, Wind, Forest	Happy	Sparrow, Finch, Raven	Yes	3	3	3
4	In a gloomy dark night, birds chirp silently in the forbidden forest of fear.	Night, Forest	Fear	Nightingale, Owl	Yes	3	3	3
5	The birds shudder as a haunting call echoes from the distance.	Night, Moon	Fear	Owl, Nightingale	Yes	3	3	3
6	Everyone in the forest chirps excitedly as the new dawn comes.	Sun, Wind, Forest	Happy	Sparrow, Raven, Finch	Yes	3	3	3
7	The forest is ravaged by a storm, the birds displaced from their homes.	Night, Forest, Wind	Sad	Finch	Yes	3	3	3
8	The birds freak out and scatter away.	Storm, Wind, (Forest)	Fear	Raven, Sparrow	Yes	3	3	2
9	The birds danced on the dunes, feeling elated that the sunlight has finally come.	Sun, Wind, (Sand)	Happy	Lark	Yes	3	3	2
10	The birds all mourn in silence at night, the passings still fresh in their minds.	Night, Moon, Forest	Sad	Nightingale, Owl	Yes	3	3	3
11	The birds cower from the shadows looming over the forest.	Night, Forest	Fear/Sad	Owl, Nightingale	Yes	3	3	3
12	Under the sunlight, the birds fly around animatedly.	Sun, Wind, (Forest)	Happy	Sparrow, Raven, Finch	Yes	3	3	3

D Image Prompt Alignment Scores

Table 6: Image prompt evaluation results with per-sample scores (0–3 scale).

ID	Image Context	Environment	Affect	Species	Coh.	Aff.	Gen.
1	Sunset	Sun, Forest, Storm	Happy	Raven, Sparrow, Finch	2	3	2
2	Beach with sand and sea	Sun, Sand, Day-time	Happy	Lark	3	3	3
3	Beach with sand, sea, and pine tree	Sun, Sand	Happy	Lark	3	3	3
4	Nighttime moon and sea	Night, Moon, Sea	Fear	Nightingale, Owl	3	3	3
5	Thunderstorm	Storm, Wind	Fear	Raven, Sparrow	3	3	3
6	Foggy golf course	Fog, Forest	Happy	Nightingale, Owl	1	3	2



(a) Image 1



(b) Image 2



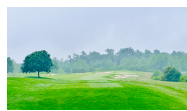
(c) Image 3



(d) Image 4



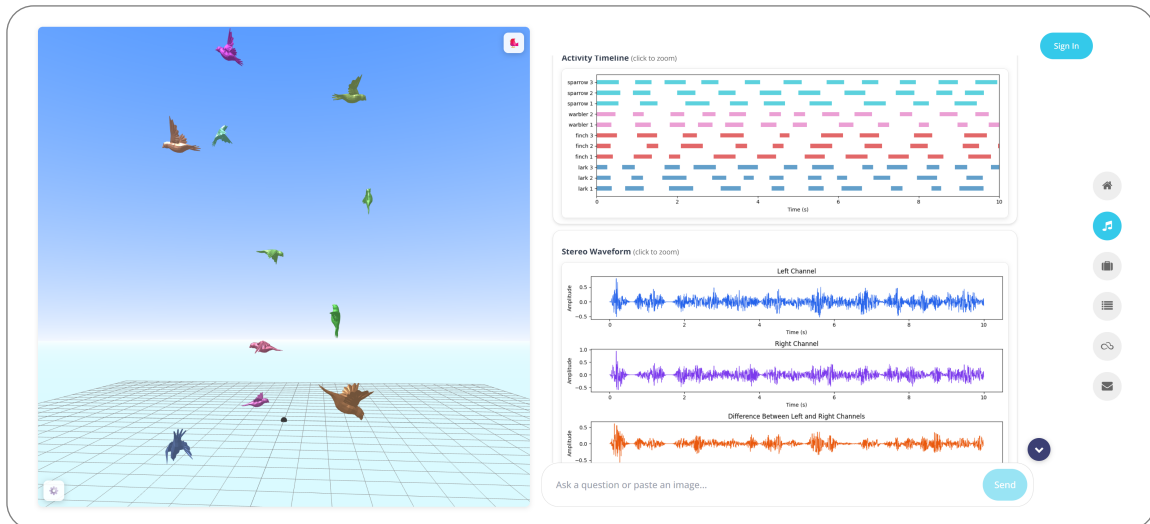
(e) Image 5



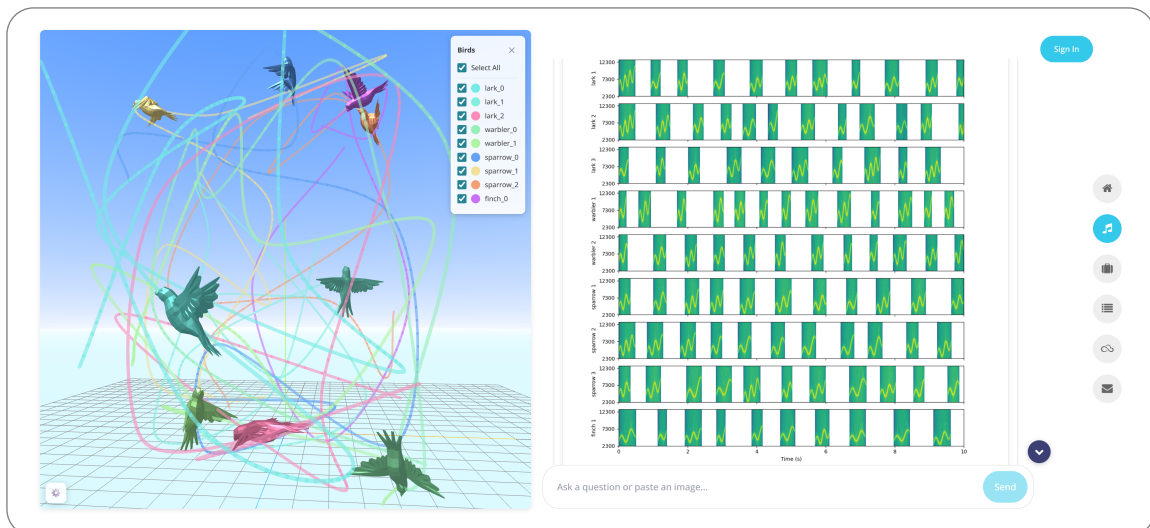
(f) Image 6

Figure 4: Experimental image prompt examples.

E Additional BirdsongChat Examples (II)



(a) This display shows the activity timeline and stereo waveforms of a birdsong generation. The bottom graph highlights waveform differences between channels, while the top timeline marks instances of bird calls over time, with colors representing different species. Call patterns vary. For example, warblers have shorter, more consistent calls than finches. The interface shows only the birds, simulating a natural flight scene.



(b) This display shows all bird frequencies and pitches on the left visualization, along with the timing of each call. Frequencies often reflect sentiment—higher pitches indicate positive sentiment, lower pitches negative. Here, most birds have fast, high-frequency calls, reflecting a positive mood. Only the enlarged birds and their flight trails are shown to highlight each bird's unique calculations.

Figure 5: BirdsongChat interface and examples.