Starchild-1: A real-time multimodal world model

Team Odyssey
research@odyssey.systems

Abstract

World models have emerged as a new form of generative intelligence. In contrast to large language models, which predict the next token in a sequence, world models predict the next state of an environment conditioned on actions and past observations. We present the next step in world modeling: `Starchild-1`, a real-time interactive audio-video world model. `Starchild-1` autoregressively generates synchronized video and audio while responding to streaming open-ended text input. Compared to video-only autoregressive models, we find that multimodal causal generation requires substantially more careful design of the training data, distillation procedure, and rollout adaptation pipeline due to the coupled dynamics of audio and video generation, and we provide insights into these design choices. We demonstrate that `Starchild-1` models in-world conversational dynamics, ambient world exploration with causal sound effects, and streaming human-model dialogue. The model supports long-horizon audio-video rollout with interactive visual and auditory updates while maintaining audio-video coherence and output stability. Running at up to 24 fps on modern hardware, `Starchild-1` enables a new paradigm of real-time interaction with multimodal world models.

1 Introduction

Recent advances in generative modeling have increasingly shifted attention from static visual content generation toward world models: systems that learn to simulate the temporal evolution of environments under external interventions. Early progress was driven by bidirectional video generation models, which rapidly improved in visual fidelity, temporal coherence, and controllability, but remained fundamentally limited to short fixed-horizon offline generation. While such models allow users to control generated content through prompts or conditioning signals, they do not support continuous interaction during rollout. This limitation motivated a transition toward causal and autoregressive formulations capable of long-horizon generation, where future states are produced incrementally conditioned on past observations. Building on this paradigm, recent interactive world models further incorporate action conditioning, enabling closed-loop generation in which user actions influence the future evolution of the generated environment. In parallel, generative modeling has expanded from purely visual generation toward joint audio-video modeling, enabling synchronized multimodal generation across vision and sound. However, existing audio-visual systems remain predominantly bidirectional and offline, producing fixed outputs rather than interactive streams.

We argue that extending causal world models from the visual to the joint audio-visual domain represents a fundamental shift in generative modeling: from static media artifacts toward persistent interactive multimodal environments. Unlike offline synthesis systems, causal audio-visual world models continuously evolve in response to user actions, speech, and environmental feedback while maintaining synchronized multimodal state over time.

To this end, we present `Starchild-1`, a long-horizon real-time interactive causal audio-video generation model that jointly streams synchronized audio and video under continuous user interaction.

Preprint.

`Starchild-1` supports long-run generation autoregressive rollout, mid-generation steering through text, speech, and visual conditioning, and temporally coherent multimodal interaction within a single unified model. More broadly, we view `Starchild-1` as an initial step toward interactive multimodal world models that jointly reason over visual and auditory dynamics. In contrast to language-only systems, multimodal world models can ground generation in multiple sensory streams, enabling richer representations of environments and their temporal evolution. This transition opens a substantially broader downstream application space, including interactive gaming, embodied agents, robotics, adaptive educational systems, conversational accessibility tools, and real-time simulated environments.

Unlike bidirectional offline generators, interactive causal audio-visual world models must maintain coherent multimodal state evolution under continuous user intervention. Extending causal generation from the visual to the joint audio-visual domain introduces additional challenges arising from the differing temporal structure, frequency, and information density of audio and video streams, as well as multimodal error accumulation where artefacts in one modality propagate into the other during long-horizon rollout.

To address these challenges, we introduce *architectural adaptations* that preserve the synchronized multimodal generation capabilities of the original bidirectional teacher during causal post-training. Furthermore, we propose a *causal distillation pipeline* that adapts a bidirectional joint audio-visual teacher into a causal student, where standard video-only distillation strategies fail due to the strong asymmetry between audio and video dynamics. This includes an asynchronous KV-cache design tailored to the differing spatio-temporal characteristics of audio and video during long-horizon autoregressive rollout. We additionally introduce interactive mid-rollout steering mechanisms that preserve previously established audio-visual coherence under changing user inputs. Finally, we discuss the limitations of standard short-horizon offline generation metrics, which only partially capture the capabilities required for evaluating interactive causal audio-visual world models. To summarize, our **contributions** are as follows:

- We present, a **long-horizon real-time interactive causal joint audio-video generation model**, combining causal rollout, synchronized multimodal streaming, and mid-rollout interactive control within a single unified system.
- We introduce a **causal distillation pipeline for bidirectional joint audio-visual foundation models** that preserves multimodal synchronization during causal post-training despite the strong asymmetry between audio and video dynamics.
- We propose an **asynchronous KV-cache and rollout adaptation strategy** tailored to the differing spatio-temporal characteristics of audio and video, enabling stable long-horizon autoregressive generation under multimodal error accumulation.
- We **study interactive causal audio-visual generation** as a distinct evaluation regime beyond offline prompt-conditioned synthesis, demonstrating multiple interaction modes including interactive world exploration, dialogue control, conversational interaction, and narrator-style companionship.

2 Related Work

Bidirectional Video Generation. Inspired by the success of text-to-image diffusion models Rombach et al. [2022], early work on text-to-video (T2V) generation extends diffusion modeling to spatio-temporal data, generating fixed-length video clips conditioned on text. These models operate on full space-time representations and are therefore inherently bidirectional, allowing each frame to condition on both past and future context during denoising. Conditioning is typically introduced via cross-attention between video tokens and encoded text embeddings. To make training tractable, these approaches adopt latent diffusion, compressing video into a lower-dimensional representation using spatio-temporal autoencoders Blattmann et al. [2023], Ho et al. [2022], Gupta et al. [2023]. More recently, flow matching has emerged as a compelling alternative to standard diffusion objectives, offering improved training efficiency and stability while retaining high sample quality Lipman et al. [2023]. Scaling this paradigm in data, model size, and compute has led to rapid improvements in visual fidelity and temporal coherence, as demonstrated by systems such as Singer et al. [2022] and recent large-scale models including OpenAI [2024], DeepMind [2024], Runway [2023], Wan [2025].

Joint Bidirectional Audio-Video Generation. As video foundation models have matured, a natural next step is extending them to jointly generate synchronized audio, the modality most immediately complementary to video. This has given rise to a growing body of work on end-to-end joint audio-video generation, where architectural choices have largely bifurcated into single-tower and multi-tower designs.

The multi-tower line of work couples two parallel diffusion transformers, one per modality, through various cross-modal interaction mechanisms. JavisDiT Liu et al. [2025a] targets fine-grained temporal alignment between modalities via a hierarchical spatio-temporal synchronization prior estimator over separate audio and video branches. UniVerse-1 Wang et al. [2025] addresses the cost of training joint AV models from scratch by stitching a pretrained video backbone and a pretrained music backbone through blockwise cross-attention. Veo 3 Wiedemer et al. [2025] set the closed-source standard for this task at the time of its release, demonstrating remarkable zero-shot reasoning capabilities for bidirectional audiovisual generation. Ovi Low et al. [2025] tackles the architectural mismatch in stitching heterogeneous experts by training a symmetric audio tower from scratch, matched in architecture to a pretrained video DiT, with bidirectional cross-attention and scaled-RoPE. Harmony Hu et al. [2025] focuses on the correspondence drift problem in joint diffusion via a global-local decoupled interaction module, combined with a cross-task synergy training regime and synchronization-focused CFG. LTX-2 HaCohen et al. [2026] scales the dual-stream paradigm via an asymmetric 14B-video / 5B-audio transformer with modality-aware CFG, extending coverage to multilingual lip-synchronized speech alongside general-domain audio. ALIVE Guo et al. [2026b] targets adapting an existing pretrained T2V model to joint AV generation by extending an MMDiT-style "dual stream + single stream" paradigm with temporally-aligned cross-attention and unified audio-visual RoPE. MOVA Yu et al. [2026] scales the multi-tower paradigm to a 32B-parameter mixture-of-experts model unifying speech, sound effects, and music.

In parallel, single-tower approaches argue that shallow cross-attention between separate towers fundamentally caps audio-visual alignment, instead processing both modalities within a unified backbone via shared self-attention. Apollo Wang et al. [2026] concatenates audio, video, and text tokens into a single sequence processed by unified DiT blocks with an Omni-Full Attention module. JavisDiT++ Liu et al. [2026] addresses the converse risk that full parameter sharing degrades unimodal quality via modality-specific feed-forward experts within a shared-attention backbone.

From Causal Video Generation to Interactive World Models. Advancing pre-defined length bidirectional video models, causal generators suitable for streaming deployment have recently gained tremendous attention. Diffusion Forcing Chen et al. [2024] laid important conceptual groundwork by training a diffusion model to denoise token sequences with independent per-token noise levels, unifying next-token prediction with full-sequence diffusion. Introducing the paradigm of causal adaptation, CausVid Yin et al. [2025] established the core paradigm for streaming video diffusion by distilling a bidirectional teacher into a causal student through an asymmetric distribution matching pipeline, achieving the first near-real-time autoregressive generation. Self Forcing Huang et al. [2025] identified and addressed the exposure bias inherent in this distillation setup by performing full autoregressive rollouts with KV-cache unrolling during training, enabling real-time streaming on a single GPU. Advancing these foundations, a range of methods have targeted key bottlenecks in distilling bidirectional diffusion models into stable long-horizon causal generators. Causal Forcing Zhu et al. [2026] and Context Forcing Chen et al. [2026] address the bidirectional-to-causal gap and student–teacher mismatch, respectively, by introducing principled initialization and long-context supervision. Rolling Forcing Liu et al. [2025b], Self-Forcing++ Cui et al. [2025], and PackForcing Mao et al. [2026] focus on improving long-horizon stability and scalability through multi-frame denoising, error mitigation, and KV-cache compression. On the efficiency front, Dummy Forcing Guo et al. [2026a] and Resampling Forcing Guo et al. [2025] reduce computational overhead and remove reliance on pretrained bidirectional teachers. Beyond temporal fidelity, Geometry Forcing Wu et al. [2025], Spatial Forcing Li et al. [2025], and Reward Forcing Lu et al. [2025] incorporate geometric, embodied, and reward-based signals to improve structure and perceptual quality.

While these approaches enable autoregressive rollout over extended horizons, they largely operate in an open-loop setting and do not yet support fully interactive or closed-loop generation. LongLive Yang et al. [2026] complements these efforts by introducing a multi-scale spatiotemporal compression scheme for the history buffer, enabling bounded-memory rollout over extended horizons under external inputs. This design opens the path toward interactive world models that maintain a persistent state and support closed-loop conditioning. Recent systems such as Hunyuan-Gamecrafter-2 Tang et al. [2026], LingBot Team [2025], GAIA-2 Russell et al. [2025], and Genie-3 DeepMind [2025] build on this direction by explicitly incorporating action or feedback signals into the generation process. These models move beyond passive video prediction toward controllable simulation, where future frames are conditioned not only on past observations but also on external interventions. In particular, Genie-3 demonstrates large-scale training of action-conditioned video models capable of real-time environment rollout, while Hunyuan-Gamecrafter-2 and LingBot explore structured interaction interfaces and embodied control. Together, these systems mark a shift from open-loop autoregressive generation to interactive, closed-loop world modeling.

While the above works focus primarily on simulating visual streams, `Starchild-1` extends this paradigm to jointly model audio-visual generation in a streaming, causal setting. By integrating audio alongside video within a unified generative process, it enables temporally aligned multimodal rollout, moving beyond silent video prediction toward richer environment simulation. Concurrent work such as OmniForcing Su et al. [2026] explores similar directions in scaling causal generation to multimodal settings, investigating how unified architectures can support joint audio-video modeling under autoregressive or streaming constraints. They design techniques specifically for their chosen base model and for better global anchoring across modalities for autoregressive rollout, in ways that we found overly constraining to interactive generation in similar experiments.

3 Interactive Causal Audio-Visual World Models

3.1 Interactive Audio-Visual World Models

In contrast to bidirectional models, causal interactive models enable closed-loop audio-visual interaction unfolding in real time. This expands the problem setting beyond offline content synthesis toward interactive environments in which future generation depends continuously on user actions, speech, or other external interventions. Consequently, the core challenge is no longer solely producing visually and acoustically plausible outputs, but maintaining temporally coherent multimodal state evolution under continuous interaction. Such capabilities are particularly relevant for applications including interactive agents, embodied assistants, adaptive educational systems, and real-time simulated environments.

In contrast to causal interactive video generation, causal interactive audio-visual generation must additionally maintain synchronization and coherence across modalities throughout long-horizon rollout. While video-only world models operate over a single visual stream, joint audio-visual models must account for the differing temporal structure, frequency, and information density of audio and video signals, as well as cross-modal error propagation where artifacts in one modality destabilize the other. Furthermore, interactive interventions such as speech, dialogue injection, or prompt switching must remain temporally aligned with the evolving visual state during autoregressive generation. Consequently, the challenge extends beyond stable visual rollout toward maintaining coherent multimodal world dynamics under continuous interaction.

3.2 Problem Formulation

Moving beyond video-only world models toward a more comprehensive representation of environment state, we extend next-state prediction from the purely visual domain to the joint audio-visual domain. It is defined as $o_t = (v_t, \omega_t)$ where v_t is the visual state and ω_t the audio state at timestep t both representing chunks of latent frames. The chunk sizes of both differ due to the different sampling frequencies. Additionally, we define an action at timestep a_t to span visual control signals, speech content and auxiliary audio conditioning such as sound or music descriptions all of which we define in text space. This yields a single text prompt and encode them into a joint embedding space using a pretrained text encoder Raffel et al. [2020]. Given a history of previous observations $O_{t-1} = \{o_i\}_{i=t-L}^{t-1}$ where L is the temporal context window length and $o_i = (v_i, \omega_i)$, the goal of `Starchild-1` Φ is to predict the next visual and auditory state:

$$v_t, \omega_t = \Phi(a_{t-1}, v_0, O_{t-1}) \tag{1}$$

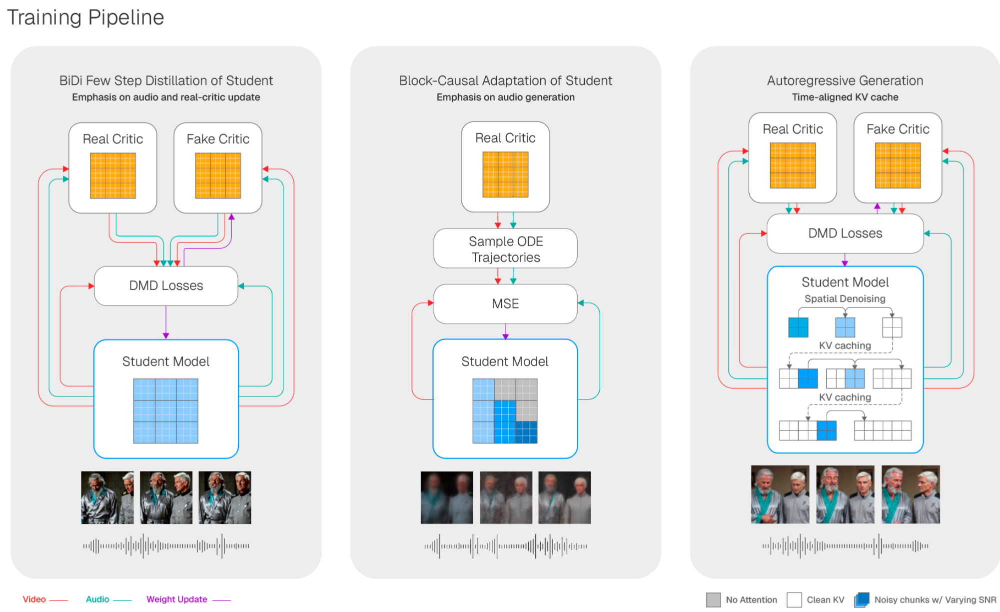

Figure 1: **Training Pipeline of** `Starchild-1`**.** (i) We distill the original 50-step base model Low et al. [2025] into a few-step bidirectional model (left), emphasizing audio optimization and fake-critic adaptation to balance multimodal learning dynamics. (ii) We causally adapt the bidirectional student through block-causal ODE trajectory supervision (middle), enabling autoregressive audio-visual rollout while preserving multimodal synchronization. (iii) We train the model under KV-cache-based autoregressive rollout (right), introducing temporally aligned multimodal caching for stable long-horizon interactive generation.

where v_t denotes the next visual state, ω_t the corresponding audio state, and v_0 is the initial visual state. We represent state at timestep t in our model with modality specific self-attention and crossmodal attention KV-caches.

4 Method

4.1 Method Overview

We build on a bidirectional dual-stream diffusion transformer Low et al. [2025], originally developed for fixed-horizon joint audio-visual generation. The architecture consists of symmetric audio and video processing streams coupled through dedicated cross-attention layers, enabling information exchange across modalities while preserving modality-specific representations. Modality-specific RoPE scaling accounts for the differing temporal resolutions of audio and video signals.

The audio stream operates over latent representations derived from compressed Mel spectrograms. During inference, generated audio latents are decoded into Mel spectrograms through the audio VAE and subsequently converted into raw audio waveforms using a neural vocoder.

To enable long-horizon streaming generation, we extend techniques developed for interactive causal video modeling Huang et al. [2025], Yang et al. [2026] to the joint audio-visual setting. We first construct a balanced multimodal training dataset which we introduce in Section 5. We then distill the original 50-step diffusion model into a low-step generator 4.2 to causally adapt the distilled model into an autoregressive streaming model afterwards while accounting for the differing temporal and structural characteristics of audio and video modalities (Sections 4.3 and 4.4). Finally, we introduce an orchestration layer (see 4.5) that conditions the world model Φ on synchronized multimodal actions, enabling closed-loop audio-visual generation under external control.

4.2 Bidirectional Few-Step Distillation

We first distill the original 50-step bidirectional diffusion model into a low-step generator following the distribution matching distillation (DMD) strategy used in CausVid Yin et al. [2025]. The goal of this stage is to obtain an efficient few-step joint audio-visual generator that preserves the multimodal generation quality of the original bidirectional teacher while enabling subsequent causal adaptation. Joint audio-visual distillation introduces additional challenges compared to video-only settings due to the differing characteristics of audio and video signals. In particular, audio primarily exhibits dense temporal structure, whereas video contains both spatial and temporal dynamics. We observe that naively applying standard DMD training leads to imbalanced optimization behavior in which the visual modality dominates training.

Emphasis on Audio Optimization. Audio is an inherently lower-dimensional and predominantly temporal signal compared to video, which contains both high-dimensional spatial structure and temporal dynamics. As a result, we observe that multimodal optimization naturally biases toward the visual modality during distillation. To better balance multimodal learning dynamics, we therefore explicitly emphasize audio optimization by increasing the learning rate of audio-related parameters.

Emphasis on Fake Critic Adaptation. Additionally, we find that increasing the learning rate of the fake critic further stabilizes training and improves audio fidelity during few-step generation. We hypothesize that this effect arises because joint audio-visual optimization introduces stronger non-stationarity in the student distribution due to the coupled learning of temporally dense audio signals and spatio-temporal video signals. A higher critic learning rate may therefore help the fake critic track the evolving multimodal distribution more effectively.

4.3 Causal Mask Adaptation

In this stage, we adapt the bidirectional teacher, which relies on full-sequence attention, into a causal generator using block-causal attention, where each chunk may only attend to previously generated chunks. Starting from the few-step distilled model, we sample ODE trajectories from the original bidirectional teacher Low et al. [2025] using UniPC Zhao et al. [2023]. Following CausVid Yin et al. [2025], we train the student to reproduce the teacher's clean latent predictions over selected denoising steps. Unlike prior work operating purely in the visual domain, our formulation jointly distills temporally aligned video and audio trajectories. To obtain aligned audio and video next state predictions, we adapt the default setting.

Student Initialization with a Few-Step Bidirectional Model. We observe that causal adaptation produces substantially more stable rollout when initialized from a few-step distilled bidirectional model rather than directly from the original 50-step bidirectional teacher as we visualize in Figure 2. We hypothesize that decomposing the overall optimization problem into two stages of (i) few-step distillation and (ii) causal adaptation simplifies the transition from offline bidirectional generation to stable autoregressive rollout. In the joint audio-visual setting, this decomposition appears particularly important, as mismatches between bidirectional and causal dynamics can accumulate errors across modalities over time. In particular, instability in either the visual or audio stream propagates through multimodal cross-attention, amplifying synchronization errors and degrading long-horizon audio-video alignment.

Accounting for Different Sampling Frequencies. To support causal rollout, we align audio and video chunk boundaries through temporal trimming and modality-specific RoPE scaling. We then construct causal self-attention and cross-attention masks for both modalities to account for differing temporal granularities. Within each modality, chunks attend only to previous chunks, while cross-modal attention has access to previously generated chunks of the other modality.

Optimization Details Matter. We find that seemingly minor optimization choices significantly affect the stability of causal adaptation. In particular, using an MSE loss during ODE initialization together with the AdamW optimizer and disabling weight decay proved important for stable multimodal rollout and successful causal adaptation.

Emphasis on Conditioned Audio Generation. We observe that increasing the classifier-free guidance (CFG) scale for the audio stream during ODE trajectory generation increases conditioning strength, leading to stronger lip synchronization, and better preservation of low-amplitude background audio cues. We hypothesize that stronger guidance increases the saliency of conditioning-dependent

Figure 2: **Rollout Quality After ODE Initialization.** Comparison of causal rollout quality when initializing the student from 4-step distilled bidirectional model (top) and original 50-step bidirectional teacher (bottom) shows a significant higher quality in the prior.

Figure 3: **Emphasis on Audio in the Bidirectional Teacher.** Increasing the classifier-free guidance (CFG) scale for audio produces more pronounced and temporally aligned lip motion, providing a stronger multimodal alignment signal during distillation.

audio-visual structure within the teacher trajectories, thereby facilitating the transfer of subtle multimodal alignment behaviors during distillation. Figure 3 illustrates this, showing more pronounced and temporally aligned lip motion under higher audio guidance scales.

Perceptual Quality and Rollout Stability. Interestingly, we observe that improved perceptual visual quality during ODE trajectory generation does not necessarily translate to improved downstream DMD performance or more stable causal rollout. We hypothesize that visually sharper trajectories may correspond to more locally optimized but less temporally stable latent dynamics, which can become more difficult to distill autoregressively. In the multimodal setting, rollout stability additionally depends on preserving consistent cross-modal temporal alignment rather than maximizing short-horizon visual fidelity alone.

ODE Initialization as a Key Factor for Stable Rollout. We find the distillation stage to be critical for stable long-horizon autoregressive generation. Importantly, the effectiveness of the distilled ODE initialization is not solely determined by short-horizon perceptual quality, but rather by how well the resulting latent trajectories align with the requirements of subsequent causal rollout adaptation. In particular, rollout stability depends strongly on the temporal consistency and autoregressive compatibility of the distilled trajectories, with poorly matched initializations rapidly collapsing during KV-cache-based long-horizon generation. This highlights the importance of matching rollout-time data distributions during distillation, consistent with observations in OmniForcing (Su et al. [2026]) and Krea's realtime 14B model (Millon [2025]). Notably, despite still lacking high-frequency visual and auditory details, models at this stage already exhibit meaningful action-following behavior under prompt-switching interactions.

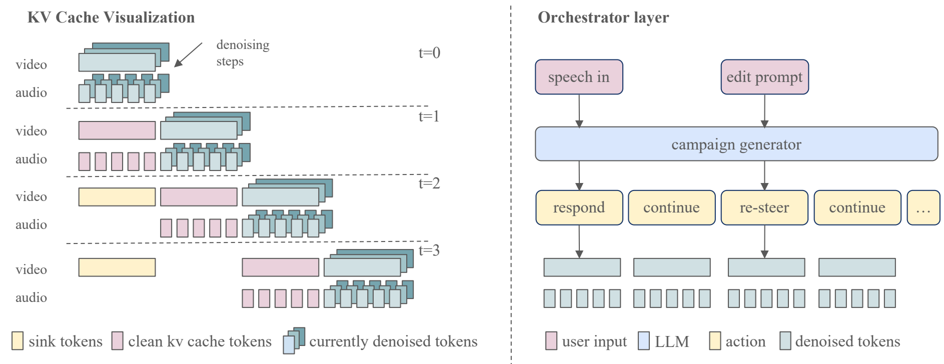

Figure 4: **KV Cache And Orchestrator Layer.** Left: We visualize the audio and video KV-cache for self-attention where the same pattern holds for cross attention. We keep video sink tokens while we evict audio sink tokens. Right: Orchestrator layer processing user input via the campaign generator and transforming it into actions.

4.4 KV-cache dependent rollout adaptation

Autoregressive rollout over extended horizons requires an efficient caching strategy to maintain bounded memory usage during inference. To this end, we employ a KV-cache mechanism that stores self-attention keys and values from previously generated audio and video chunks. However, we observe that models distilled solely from bidirectional trajectories remain unstable during long-horizon autoregressive rollout despite caching. Following Self-Forcing Huang et al. [2025], we therefore further adapt the model on self-generated autoregressive trajectories to better align training and inference dynamics.

Caching Self- and Cross-Attention States. Beyond caching self-attention states, we additionally cache cross-attention keys and values across modalities, which further improves multimodal temporal consistency during streaming generation. We further observe that introducing sink tokens within the video KV-cache stabilizes long-horizon rollout, whereas analogous sink tokens in the audio cache provide limited benefit (see Figure 4). This motivates an asymmetric KV-cache design in which the video kv cache contains an additional chunk of latent frames.

Reduced Need for KV-Cache Re-Caching Under Prompt Transitions. In contrast to prior causal generation work that performs KV-cache re-caching after prompt transitions, we observe that preserving the original video sink tokens leads to more stable long-horizon rollout while still allowing adaptation to evolving visual and audio conditioning signals. We hypothesize that aggressive cache resets introduce autoregressive distribution shifts in the visual latent dynamics, which subsequently propagate into the audio stream through multimodal cross-attention. Because audio and video are jointly rolled out and continuously condition each other, instability in one modality can accumulate and amplify across modalities over time. Since updated conditioning is continuously injected through cross-attention, the model can adapt to prompt changes without fully discarding previously accumulated multimodal world state.

Improving Supervision Quality Through Real-Critic Biasing. We identified that much of the performance gain in SF-style causal adaptation after the ODE init stage came from the stability of the training signal on generation quality gain as controlled primarily by biasing the real critic to have as much of a generation quality advantage as possible over the fake critic. We index on this observation by optimising the inference parameters specifically for the real critic, increasing the audio guidance scale to 7 and optimising the choice of the SLG layer in that frozen Ovi critic.

Causal Audio. We find that the specific techniques proposed in prior work for causal adaptation of bidirectional base video models generalize quite poorly beyond the specific assumptions of T2V base video models themselves. In particular, we identify that much of the demonstrated stability in kv-cache dependent self-rollouts stems from the inherent visual structure and next frame predictiveness of latent representations in the visual domain. Text conditioning via full bidirectional cross-attention can still transfer over well in the visual domain through the relative scene stability and anchoring that comes from a global visual scene context.

Neither of these base assumptions holds for the causal audio and joint causal audiovisual generation setting. When setting boundaries for audiovisual state at the phoneme or subphoneme level, we found that predictiveness of the next aural state to be much less well-conditioned than the next visual state. The assumption of bidirectional text cross-attention inherited from video models further worsened this problem by introducing variable confusion between prior state context and text action context for predicting the next aural state.

We resolve this issue by defining audio state at the level of single words or short utterances, to enable clearer causal predictiveness of the next aural state given prior aural state and text action context. This approach also enables further improvements on streaming text as world model action updates, described further in 4.5.

4.5 Interactive Orchestration Layer

Given our autoregressive formulation in Section 3.2, interactive generation requires an input action sequence $a_{1:t}$ during rollout while maintaining consistency with the previously generated multimodal history O_t. Unlike offline prompt-conditioned generation, causal rollout operates over persistent modality-specific self-attention and cross-attention KV-caches storing previously generated audio and video states. Consequently, new conditioning signals should not overwrite past generations, but instead should evolve coherently relative to the immutable generated history. While the world model Φ predicts audio-visual chunks autoregressively at the token level, interactive generation additionally requires a higher-level orchestration mechanism that manages temporally evolving action sequences over extended rollout horizons. To this end, we introduce an orchestration layer that dynamically constructs structured multi-chunk action trajectories synchronized with the streaming generation process.

Campaign-based Action Scheduling. Rather than conditioning the model using a single static prompt, the orchestrator generates temporally ordered prompt sequences, which we refer to as *Campaigns*. Each Campaign defines a sequence of chunk-level actions $a_{1:t}$ describing evolving scene dynamics, speech content, narration, ambient audio, music, and interaction events over future rollout windows (see Figure 4). Campaigns are generated online by an external language model conditioned on the current interaction state and previously generated rollout history.

Streaming Prompt Injection. Campaign actions are then asynchronously injected into the causal model at chunk boundaries through a rolling prompt-update interface. The orchestration layer supports multiple interaction regimes, including interactive world editing, scripted dialogue control, conversational interaction, narrator-style narration, and speakerless environmental evolution. To preserve multimodal temporal consistency, the system dynamically determines whether prompt updates trigger cross-attention recaching or soft prompt transitions depending on scene continuity and interaction type.

Chunk-Local Speech Constraints. A key constraint in causal audio-visual generation is that speech content must fit within the temporal duration of each generated audio chunk ω_t. We therefore constrain prompt generation through chunk-local syllable budgets derived from the rollout duration of each audio segment. This encourages temporally realizable speech generation and reduces truncation artefacts during long-horizon rollout.

Prompt Delta Conditioning. Finally, we observe that repeatedly conditioning on identical prompts across consecutive chunks can degrade speech quality during autoregressive generation. We hypothesize that this arises from interactions between persistent KV-cache state and redundant cross-attention conditioning. Consequently, the orchestrator structures prompt updates primarily as temporally local action deltas rather than re-conditioning on identical speech content across adjacent rollout steps.

5 Dataset Curation

The bidirectional Ovi base model contains strong priors for text-conditioned joint audio-visual generation. However, adapting the model to interactive causal rollout requires exposing it to prompt distributions that reflect the interaction dynamics encountered during streaming generation.

Change Prompts. To support mid-rollout editing and interactive steering, the model is trained on context-aware prompt transitions spanning both the visual and audio domains. This is particularly important for speech-driven applications, where conversational dynamics, speaker transitions, and evolving environmental audio must remain temporally coherent throughout rollout.

Interaction Dynamics. The model must additionally learn the interaction regimes targeted by the final system, including sound-effect-driven environments, in-world monologue and dialogue interactions, multi-character conversational settings, and narrator-style disembodied speech over evolving scene dynamics. Consequently, dataset construction must account not only for visual scene diversity, but also for temporally evolving multimodal interactions across both in-world and observer-conditioned audio streams.

We therefore design specialized prompt datasets for the various stages of causal adaptation with emphasis on interactive multimodal consistency and controllable rollout behavior. We identify seven distinct categories of soundscape scenarios for which to generate a sufficiently diverse and representative prompt set: (1) SFX only, (2) Monologue in-world character speech, (3) Conversational in-world character dialogue, (4) Character speech with SFX, (5) Monologue disembodied narrator speech, (6) Disembodied narrator dialogue, and (7) Disembodied narrator speech with SFX. The disembodied narrator dialogue is designed in particular such that speech turns correspond to chunk boundaries in rollout, such that speech responses can be learned through cache context and prompt changes at natural generation state transitions. We train these models on 2.5M representative prompt samples across these seven categories.

In practice, we find it important to construct prompts that produce strong and distinguishable supervisory trajectories under the frozen bidirectional teacher at the guidance scales used during distillation. To this end, candidate generations are filtered using a separate VLM-based judge to favor prompts yielding visually and acoustically well-conditioned teacher samples. We further identify the prompt descriptions that produced the best overall audio generations and video-aligned audio syncing: ordering speech prompts always before general ambience prompts, anchoring on continuous broadband sounds when generating sounds from narrowband sources, emphasizing amplitude and amplitude changes in prompt descriptions. Visual prompting followed standard modern best practices.

6 Experiments

6.1 Implementation Details

A *chunk* in our formulation denotes a sequence of latent frames. We use a chunk size of 15 latent frames for video and 75 latent frames for audio, corresponding to approximately 2.5s of temporal context at 24 FPS. During autoregressive rollout, we maintain asymmetric modality-specific KV-cache windows with a video cache size of 45 and an audio cache size of 150 to account for the differing temporal structure and sampling frequencies of the two modalities. Except for the few-step bidirectional distillation stage, we follow the learning rate and batch size settings of prior work Yang et al. [2026]. We train all models using Fully Sharded Data Parallel (FSDP) training.

6.2 Quantitative Analysis

Evaluation metrics only evaluates short-horizon offline rollout quality over $4-5$s clips whereas the primary focus of `Starchild-1` is long-horizon interactive audio-video generation under continuously evolving conditioning signals (see Subsection 6.3). However, we view these metrics as evidence that our causal adaptation preserves meaningful audio-video synchronization and multimodal consistency despite the substantially weaker bidirectional initialization.

Evaluation Metrics. To compare `Starchild-1` to our base model Ovi, we report the VBench metrics Huang et al. [2024] Subject Consistency, Aesthetic Quality, and Imaging Quality to measure perceptual quality. For audio quality, we report DNSMOS Manocha et al. [2022], UTMOS Saeki et al. [2022], CPCER Morris et al. [2004], and CLAP-score Wu et al. [2023]. Finally, we also provide comparison with respect to audio-visual synchronization with LSE-C and LSE-D Chung and Zisserman [2017] score as well as ImageBind Girdhar et al. [2023], Audio Alignment, Audio–visual Event Synchronization, and Audio–visual Coherence. The latter three are metrics we define.

Audio–visual Event Synchronization measures whether audio transients fire at the same moments as visual changes. We extract a per-frame visual-change signal as the perceptual-hash (pHash)

Metric	Dir.	Ovi	Starchild-1
Perceptual Quality			
Subject Consistency Huang et al. [2024]	↑	0.961 ± 0.053	$\mathbf{0.964 \pm 0.04}$
Aesthetic Quality Huang et al. [2024]	↑	0.520 ± 0.108	$\mathbf{0.563 \pm 0.09}$
Imaging Quality Huang et al. [2024]	↑	0.451 ± 0.119	$\mathbf{0.561 \pm 0.08}$
Audio Quality			
DNSMOS Manocha et al. [2022]	↑	$\mathbf{2.330 \pm 0.236}$	2.30 ± 0.24
UTMOS Saeki et al. [2022]	↑	1.566 ± 0.626	$\mathbf{1.78 \pm 0.58}$
CPCER Morris et al. [2004]	↓	$\mathbf{1.219 \pm 0.824}$	1.80 ± 1.42
WER Morris et al. [2004]	↓	$\mathbf{0.0035 \pm 0.043}$	0.007 ± 0.06
Clap Score Wu et al. [2023]	↑	$\mathbf{0.265 \pm 0.145}$	0.182 ± 0.16
Audio-Visual Consistency & Synchronization			
LSE-C Chung and Zisserman [2017]	↑	$\mathbf{0.338}$	0.336
LSE-D Chung and Zisserman [2017]	↓	$\mathbf{1.151}$	1.153
ImageBind Score Girdhar et al. [2023]	↑	$\mathbf{0.188 \pm 0.118}$	0.160 ± 0.11
Audio Alignment	↑	$\mathbf{0.609 \pm 0.151}$	0.474 ± 0.27
AV Event Sync	↑	0.122 ± 0.135	$\mathbf{0.147 \pm 0.15}$
AV Coherence	↑	$\mathbf{0.628 \pm 0.062}$	0.588 ± 0.10

Table 1: **Comparison between Ovi and `Starchild-1`.** We report multimodal quality, synchronization, semantic alignment, and speech-related metrics. Best values are shown in bold. `Starchild-1` remains close to the performance of the initial pretrained model, Ovi. Audio Alignment, AV Event Sync, and AV Coherence metrics are defined in Section 6.2

Hamming distance between consecutive frames, and an audio-energy signal as the RMS envelope of the soundtrack hopped at the video frame rate. After z-scoring both, we take their normalized cross-correlation within a ± 0.5 s lag window, clip the peak to $[0, 1]$, and report it as $\mathrm{av\_event\_sync}$. The metric is content-agnostic and provides a general-purpose temporal-alignment signal in settings where conventional lip-sync metrics (e.g. SyncNet-based LSE) cannot be computed — clips with no visible face, occluded mouths, animated or non-human subjects, or non-speech audio such as foley, music, and ambient. It rewards tight temporal lock-step (door slam, hand clap, gong strike, footstep) and is insensitive to whether the audio matches any specific concept.

Audio Alignment measures whether the generated audio matches what the prompt asked for, semantically. For clips that contain speech (detected by Whisper-base's no-speech probability), we transcribe with Whisper-large-v3 and compute $\max(0,\, 1 - \mathrm{WER})$ against the speech text inside the prompt's `<S>...</S>` tags. For non-speech clips, we take the cosine similarity between LAION-CLAP audio and text embeddings — the text being the prompt's `<AUDCAP>...</ENDAUDCAP>` caption — and remap it from $[-1, 1]$ to $[0, 1]$. The metric thus answers "is the audio about the right thing," independent of whether it is temporally aligned to the video.

The Audio–visual Coherence metric rolls three pre-normalized $[0, 1]$ components into a single score:

$$\text{AV coherence} = 0.40 \cdot \text{Subject Consistency} + 0.35 \cdot \text{Audio Alignment} + 0.25 \cdot \text{AV Event Sync}$$

Video coherence (DINO inter-frame consistency) carries the heaviest weight because it is the broadest indicator of generation stability; audio semantic alignment receives the second-largest weight as the primary signal that the model actually responded to the prompt's audio side; temporal event sync receives the smallest weight, reflecting that it is a noisier measurement, that many prompts (music, ambient) have no sharp transients to align to, and that its meaningfulness varies with prompt content. The weighted sum stays in [0, 1] and is unbounded by any single component, so a model that is strong on video but weak on either audio dimension still tops out below 0.65.

Comparison to Bidirectional Base Model Table 1 compares the bidirectional Ovi base model with `Starchild-1` after causal adaptation. Interestingly, while `Starchild-1` exhibits moderate degradation on several audio and cross-modal alignment metrics, particularly CLAP Score, IB Score, and Audio Alignment, we simultaneously observe improvements on perceptual visual quality metrics including Subject Consistency, Aesthetic Quality, and Imaging Quality. We hypothesize that autoregressive rollout adaptation encourages temporally smoother and visually more coherent latent dynamics, which benefits perceptual image quality despite introducing mild degradation in semantic and cross-modal alignment metrics.

Importantly, synchronization-related metrics such as LSE-C and LSE-D remain nearly unchanged relative to the original Ovi model, indicating that causal adaptation largely preserves low-level lip synchronization and temporal alignment behavior. Similarly, AV Event Sync slightly improves after causal adaptation, suggesting that autoregressive rollout training may strengthen causal temporal coupling between audio and visual events. In contrast, the strongest degradation occurs in CLAP and IB-based semantic alignment metrics, likely reflecting accumulated distribution shift during long-horizon autoregressive rollout. Overall, the results suggest that causal multimodal adaptation primarily affects high-level semantic alignment and perceptual fidelity while preserving much of the underlying temporal synchronization structure between modalities.

6.3 Qualitative Evaluation of Interactive Causal Audio-Visual Generation

Existing joint audio-visual benchmarks primarily evaluate offline prompt-conditioned generation through metrics targeting visual fidelity, audio quality, and synchronization. While these metrics remain necessary, they only partially capture the capabilities required for interactive causal audio-visual generation, where generation evolves continuously under user intervention. We therefore complement standard offline metrics with qualitative evaluation across four interaction regimes: (i) interactive world exploration, (ii) scripted dialogue control, (iii) conversational interaction, and (iv) narrator-style companionship. Since no established benchmarks currently exist for this setting, we evaluate these capabilities qualitatively through curated rollout examples and structured analysis of multimodal consistency, synchronization behavior, and responsiveness under interaction. Importantly, all interaction regimes are supported by the same unified model without task-specific specialization.

Mid Prompt Interactivity. Figure 5 illustrates interactive mid-rollout prompt injection during causal audio-visual generation after 2.5s. In each example, the rollout is conditioned on a prompt transition, after which the model adapts both the visual and audio streams in a temporally coherent manner. We observe that the generated audio remains tightly synchronized with the evolving visual dynamics, including temporally aligned frog croaks, penguin emergence sounds, lightning strikes, and volcanic eruption audio. The accompanying Mel spectrograms further highlight that prompt transitions induce structured changes in the generated acoustic patterns rather than isolated amplitude spikes, suggesting that the model maintains coherent cross-modal temporal dynamics during long-horizon rollout. These results demonstrate that the model can respond to evolving conditioning signals while preserving audio-video synchronization and multimodal consistency under causal generation.

Additional Visual Results. We provide further visualizations of 30s autoregressive rollouts in Figures 6 and 7 and refer the reader to the project page for accompanying audio-visual examples.

7 Limitations.

`Starchild-1` inherits the failure modes of its bidirectional teacher and additionally exhibits failure modes specific to causal rollout. Most notably, extended interactive sessions can still display gradual drift in scene identity and acoustic environment over long horizons, and the conversational interaction mode has not yet been evaluated against the dialogue coherence standards established in the language modeling community. Our evaluation in this work is also deliberately qualitative: we have demonstrated that the four interaction modes are achievable and characterized the qualitative space of model behavior within each, but we have not yet established the quantitative benchmarks that would allow rigorous comparison of future causal AV models against `Starchild-1` or against each other.

8 Conclusion

We present `Starchild-1`, a long-horizon real-time interactive causal joint audio-video generation model. By combining causal distillation of a bidirectional joint audiovisual foundation model with architectural adaptations for stable multimodal autoregressive rollout, `Starchild-1` supports

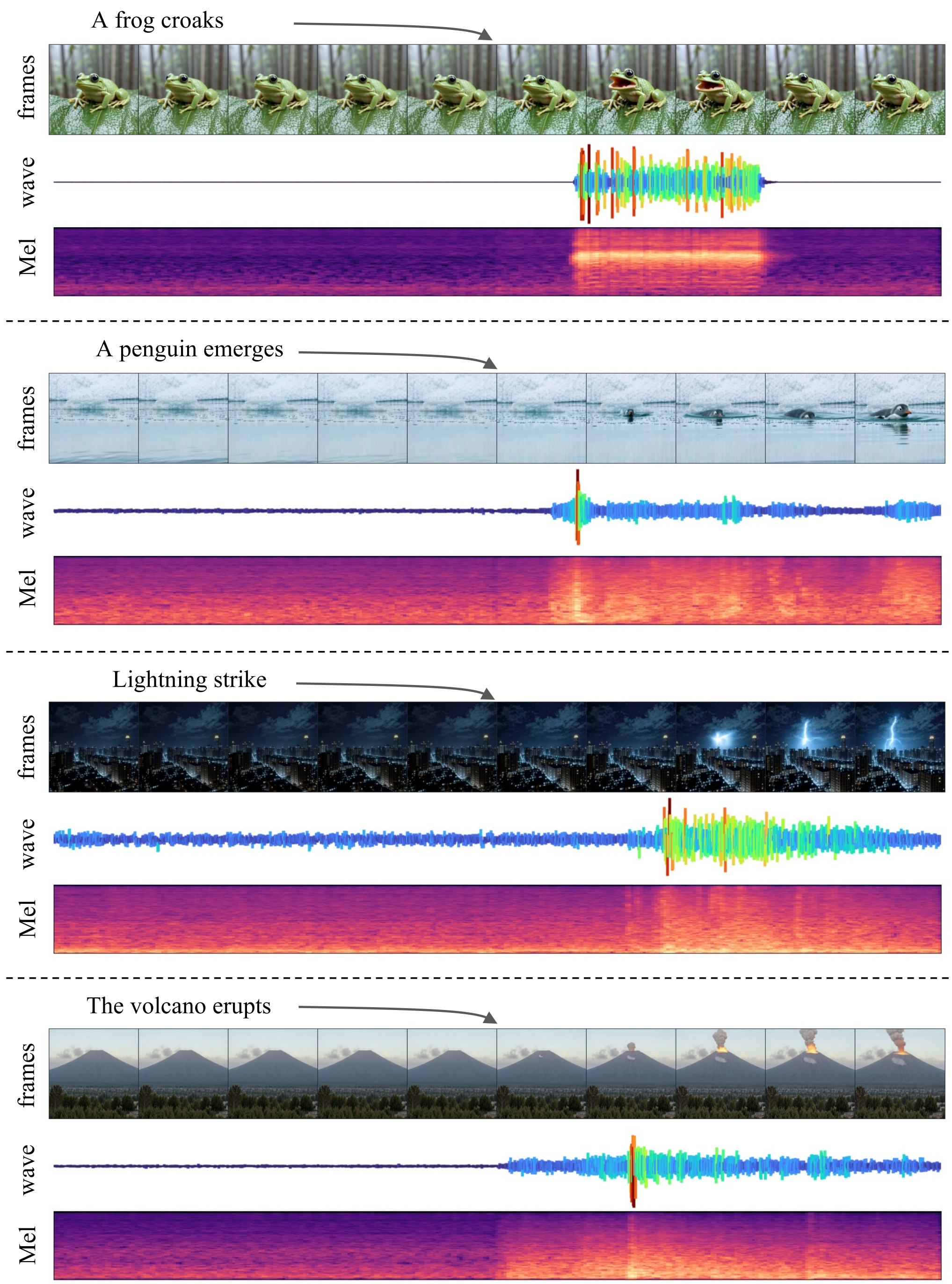

Figure 5: **Interactive Mid-Rollout Prompt Injection.** Examples of prompt transitions streamed into 5 second audio-visual rollouts after 2.5s. We show video frames sampled at regular intervals together with the corresponding generated waveform and Mel spectrogram, illustrating synchronized multimodal adaptation under changing conditioning signals.

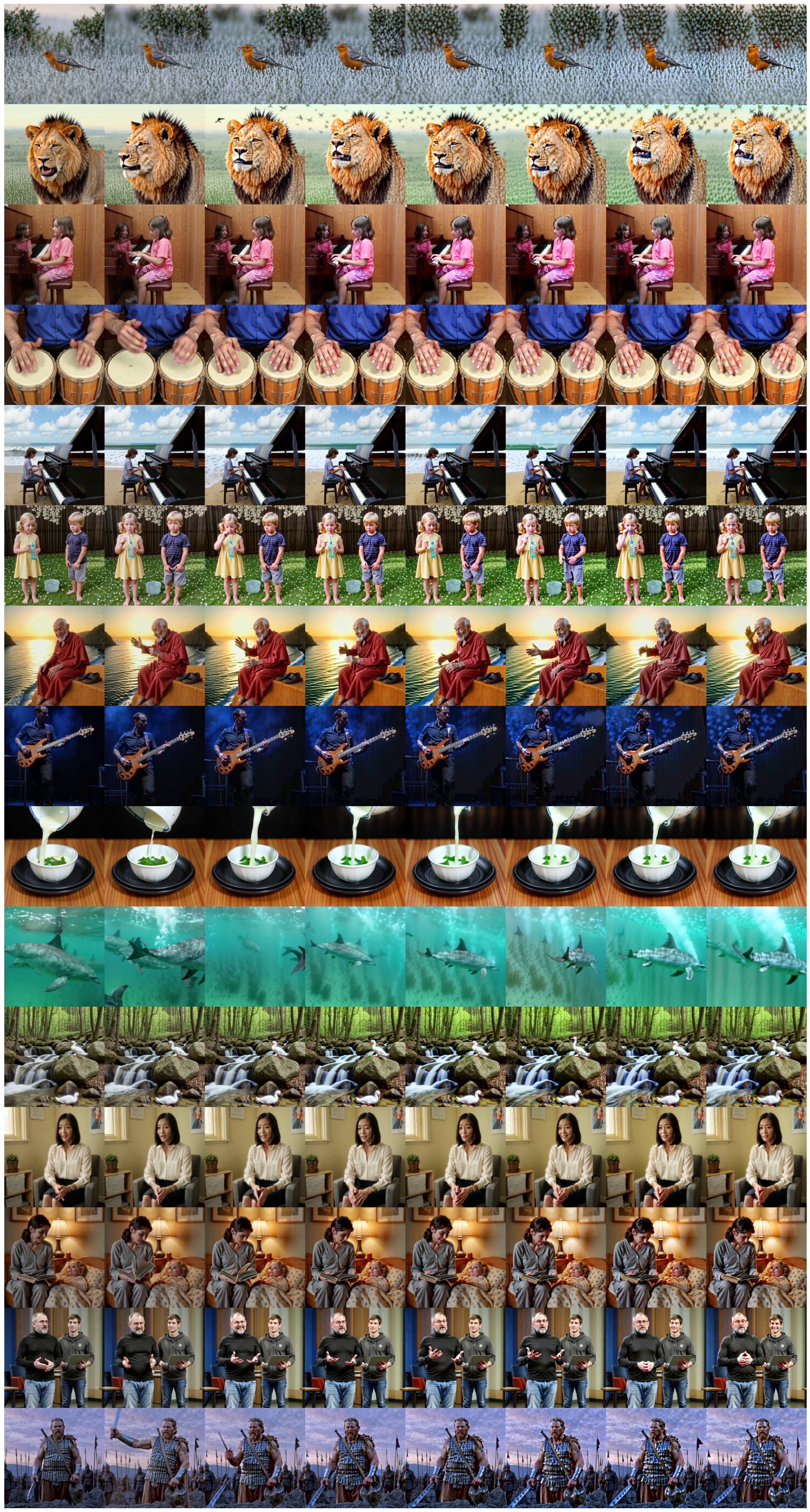

Figure 6: **Additional Visual Results.** Additional examples of 30s autoregressive rollouts generated by `Starchild-1`.

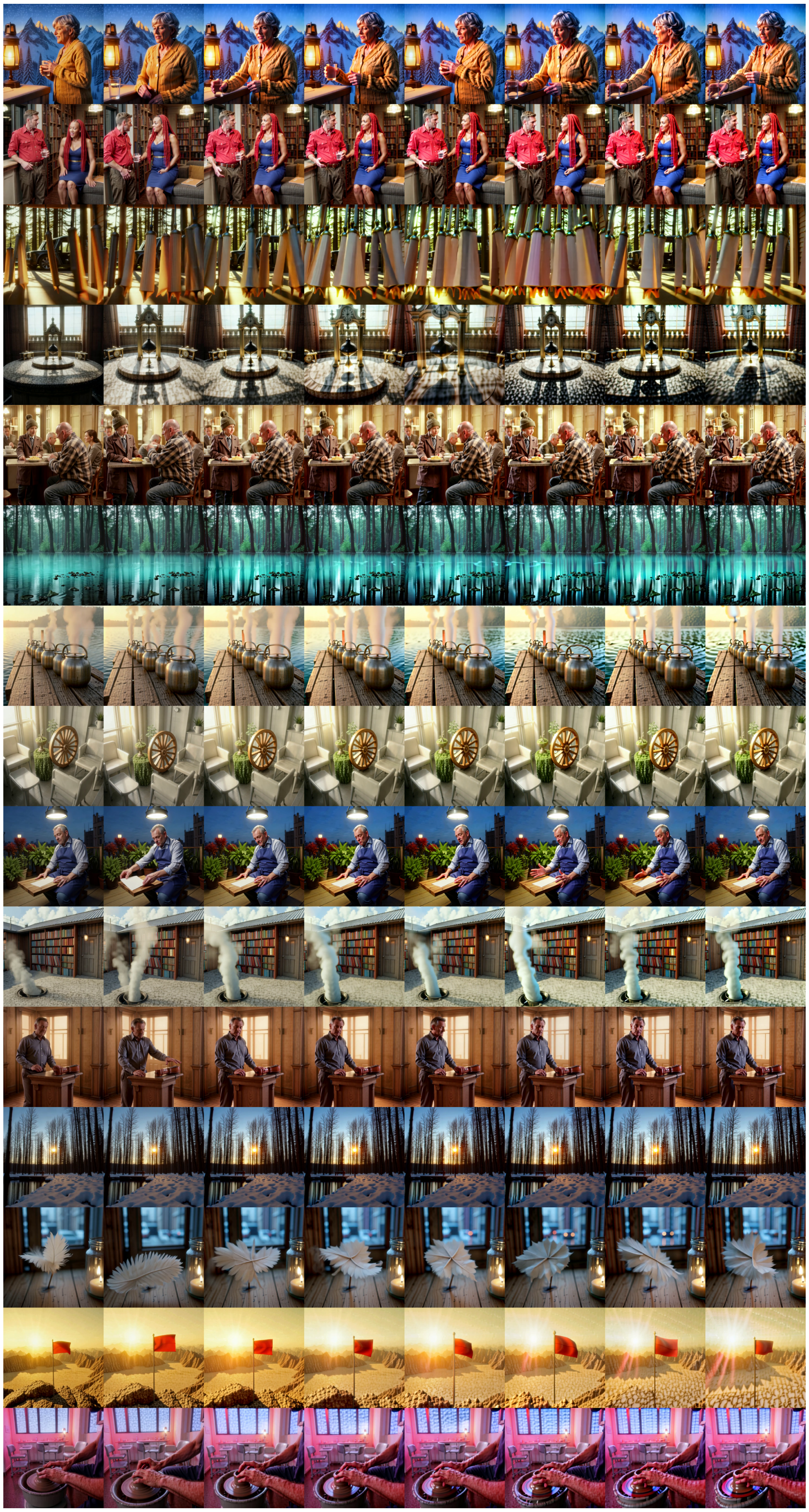

Figure 7: **Additional Visual Results.** Additional examples of 30s autoregressive rollouts generated by `Starchild-1`.

synchronized real-time audio-video streaming and interactive mid-generation control through text, speech, and visual conditioning.

More broadly, we argue that causal joint audio-video generation constitutes a distinct modeling regime from bidirectional offline AV synthesis. Unlike offline generators, causal multimodal world models must maintain coherent audio-visual state evolution under continuous user intervention. Our qualitative results suggest that jointly modeling audio and video causally over time encourages grounded multimodal dynamics, including temporally consistent environmental audio and causal synchronization between visual events and sound.

Several important directions remain for future work, including scaling model capacity and training data, extending stable rollout horizons, and developing quantitative evaluation protocols for interactive causal generation. We believe that interactive causal audio-visual generation opens a promising direction toward multimodal world models capable of persistent real-time interaction within evolving audio-visual environments.

9 Contributors

`Starchild-1` was developed by the Odyssey team, listed alphabetically below.

Core contributors: Ahmad Nazeri, Amogh Adishesha, Jenny Seidenschwarz, Richard Shen, Sarah King, Tobiah Rex, Vighnesh Birodkar.

Full team: Ahmet Hamdi Guzel, Andy Kolkhorst, Aravind Kaimal, Ben Graham, Derek Sarshad, Fabian Güra, Finley Code, James Grieve, Jesse Allardice, Jessica Inman, Jonathan Sadeghi, Kaiwen Guo, Kristy McDonough, Nicolas Griffiths, Nima Rezaeian, Robin Tweedie, Sirish Srinivasan, Vinh-Dieu Lam, Zygmunt Łenyk.

Leadership: Jeffrey Hawke, Oliver Cameron.

References

Andreas Blattmann, Robin Rombach, Huan Ling, and Björn Ommer. Align your latents: High-resolution video synthesis with latent diffusion models. *arXiv preprint arXiv:2304.08818*, 2023.

Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion, 2024. URL `https://arxiv.org/abs/2407.01392`.

Shuo Chen, Cong Wei, Sun Sun, Ping Nie, Kai Zhou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context forcing: Consistent autoregressive video generation with long context, 2026. URL `https://arxiv.org/abs/2602.06028`.

Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *ACCV Workshops*, 2017.

Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation, 2025. URL `https://arxiv.org/abs/2510.02283`.

Google DeepMind. Veo: Generating video with multimodal diffusion models. `https://deepmind.google/discover/blog/veo/`, 2024.

Google DeepMind. Genie 3: Towards real-time interactive world models. `https://deepmind.google/discover/blog/`, 2025.

Rohit Girdhar et al. Imagebind: One embedding space to bind them all. *CVPR*, 2023.

Hang Guo, Zhaoyang Jia, Jiahao Li, Bin Li, Yuanhao Cai, Jiangshan Wang, Yawei Li, and Yan Lu. Efficient autoregressive video diffusion with dummy head, 2026a. URL `https://arxiv.org/abs/2601.20499`.

Ying Guo, Qijun Gan, Yifu Zhang, Jinlai Liu, Yifei Hu, Pan Xie, Dongjun Qian, Yu Zhang, Ruiqi Li, Yuqi Zhang, Ruibiao Lu, Xiaofeng Mei, Bo Han, Xiang Yin, Bingyue Peng, and Zehuan Yuan. Alive: Animate your world with lifelike audio-video generation, 2026b. URL https://arxiv.org/abs/2602.08682.

Yuwei Guo, Ceyuan Yang, Hao He, Yang Zhao, Meng Wei, Zhenheng Yang, Weilin Huang, and Dahua Lin. End-to-end training for autoregressive video diffusion via self-resampling, 2025. URL https://arxiv.org/abs/2512.15702.

Abhishek Gupta, Nan Xue, et al. Photorealistic video generation with diffusion models. *arXiv preprint arXiv:2312.06662*, 2023.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. Ltx-2: Efficient joint audio-visual foundation model, 2026. URL https://arxiv.org/abs/2601.03233.

Jonathan Ho, William Chan, Chitwan Saharia, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Teng Hu, Zhentao Yu, Guozhen Zhang, Zihan Su, Zhengguang Zhou, Youliang Zhang, Yuan Zhou, Qinglin Lu, and Ran Yi. Harmony: Harmonizing audio and video generation through cross-task synergy, 2025. URL https://arxiv.org/abs/2511.21579.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion, 2025. URL https://arxiv.org/abs/2506.08009.

Ziqi Huang et al. Vbench: Comprehensive benchmark suite for video generative models. *CVPR*, 2024.

Fuhao Li, Wenxuan Song, Han Zhao, Jingbo Wang, Pengxiang Ding, Donglin Wang, Long Zeng, and Haoang Li. Spatial forcing: Implicit spatial representation alignment for vision-language-action model, 2025. URL https://arxiv.org/abs/2510.12276.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2023.

Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Jiebo Luo, Ziwei Liu, Hao Fei, and Tat-Seng Chua. Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization, 2025a. URL https://arxiv.org/abs/2503.23377.

Kai Liu, Yanhao Zheng, Kai Wang, Shengqiong Wu, Rongjunchen Zhang, Jiebo Luo, Dimitrios Hatzinakos, Ziwei Liu, Hao Fei, and Tat-Seng Chua. Javisdit++: Unified modeling and optimization for joint audio-video generation, 2026. URL https://arxiv.org/abs/2602.19163.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time, 2025b. URL https://arxiv.org/abs/2509.25161.

Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation, 2025. URL https://arxiv.org/abs/2510.01284.

Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, Yujun Shen, and Min Zhang. Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation, 2025. URL https://arxiv.org/abs/2512.04678.

Pranay Manocha et al. Dnsmos p.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. *ICASSP*, 2022.

Xiaofeng Mao, Shaohao Rui, Kaining Ying, Bo Zheng, Chuanhao Li, Mingmin Chi, and Kaipeng Zhang. Packforcing: Short video training suffices for long video sampling and long context inference, 2026. URL https://arxiv.org/abs/2603.25730.

Erwann Millon. Krea realtime 14b: Real-time video generation, 2025. URL `https://github.com/krea-ai/realtime-video`.

Andrew Morris, Viktoria Maier, and Phil Green. A comparison of traditional and novel metrics for automatic speech recognition evaluation. In *Eurospeech*, 2004.

OpenAI. Video generation models as world simulators. `https://openai.com/research/video-generation-models-as-world-simulators`, 2024.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL `http://jmlr.org/papers/v21/20-074.html`.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Runway. Gen-2: The next step forward for generative video. `https://research.runwayml.com/gen2`, 2023.

Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. Gaia-2: A controllable multi-view generative world model for autonomous driving, 2025. URL `https://arxiv.org/abs/2503.20523`.

Hiroshi Saeki et al. Utmos: Utokyo-sarulab system for voicemos challenge 2022. In *Interspeech*, 2022.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Song Zhang, Qiang Hu, Ming-Hsuan Yang, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Yaofeng Su, Yuming Li, Zeyue Xue, Jie Huang, Siming Fu, Haoran Li, Ying Li, Zezhong Qian, Haoyang Huang, and Nan Duan. Omniforcing: Unleashing real-time joint audio-visual generation, 2026. URL `https://arxiv.org/abs/2603.11647`.

Junshu Tang, Jiacheng Liu, Jiaqi Li, Longhuang Wu, Haoyu Yang, Penghao Zhao, Siruis Gong, Xiang Yuan, Shuai Shao, Linfeng Zhang, and Qinglin Lu. Hunyuan-gamecraft-2: Instruction-following interactive game world model, 2026. URL `https://arxiv.org/abs/2511.23429`.

LingBot Team. Lingbot: Language-conditioned interactive world models. *arXiv preprint arXiv:2502.01234*, 2025.

Wan. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Duomin Wang, Wei Zuo, Aojie Li, Ling-Hao Chen, Xinyao Liao, Deyu Zhou, Zixin Yin, Xili Dai, Daxin Jiang, and Gang Yu. Universe-1: Unified audio-video generation via stitching of experts, 2025. URL `https://arxiv.org/abs/2509.06155`.

Jun Wang, Chunyu Qiang, Yuxin Guo, Yiran Wang, Xijuan Zeng, and Feng Deng. Apollo: Unified multi-task audio-video joint generation, 2026. URL `https://arxiv.org/abs/2601.04151`.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.

Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling, 2025. URL `https://arxiv.org/abs/2507.07982`.

Yusong Wu et al. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, 2023.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Ying-Cong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=nCAODkpsPJ`.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. 2025.

SII-OpenMOSS Team: Donghua Yu, Mingshu Chen, Qi Chen, Qi Luo, Qianyi Wu, Qinyuan Cheng, Ruixiao Li, Tianyi Liang, Wenbo Zhang, Wenming Tu, Xiangyu Peng, Yang Gao, Yanru Huo, Ying Zhu, Yinze Luo, Yiyang Zhang, Yuerong Song, Zhe Xu, Zhiyu Zhang, Chenchen Yang, Cheng Chang, Chushu Zhou, Hanfu Chen, Hongnan Ma, Jiaxi Li, Jingqi Tong, Junxi Liu, Ke Chen, Shimin Li, Shiqi Jiang, Songlin Wang, Wei Jiang, Zhaoye Fei, Zhiyuan Ning, Chunguo Li, Chenhui Li, Ziwei He, Zengfeng Huang, Xie Chen, and Xipeng Qiu. Mova: Towards scalable and synchronized video-audio generation, 2026. URL `https://arxiv.org/abs/2602.08794`.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *NeurIPS*, 2023.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation, 2026. URL `https://arxiv.org/abs/2602.02214`.