Genie: Generative Interactive Environments

Jake Bruce[*,1], **Michael Dennis**[*,1], **Ashley Edwards**[*,1], **Jack Parker-Holder**[*,1], **Yuge (Jimmy) Shi**[*,1], **Edward Hughes**[1], **Matthew Lai**[1], **Aditi Mavalankar**[1], **Richie Steigerwald**[1], **Chris Apps**[1], **Yusuf Aytar**[1], **Sarah Bechtle**[1], **Feryal Behbahani**[1], **Stephanie Chan**[1], **Nicolas Heess**[1], **Lucy Gonzalez**[1], **Simon Osindero**[1], **Sherjil Ozair**[1], **Scott Reed**[1], **Jingwei Zhang**[1], **Konrad Zolna**[1], **Jeff Clune**[1,2], **Nando de Freitas**[1], **Satinder Singh**[1] and **Tim Rocktäschel**[*,1]

[*]Equal contributions, [1]Google DeepMind, [2]University of British Columbia

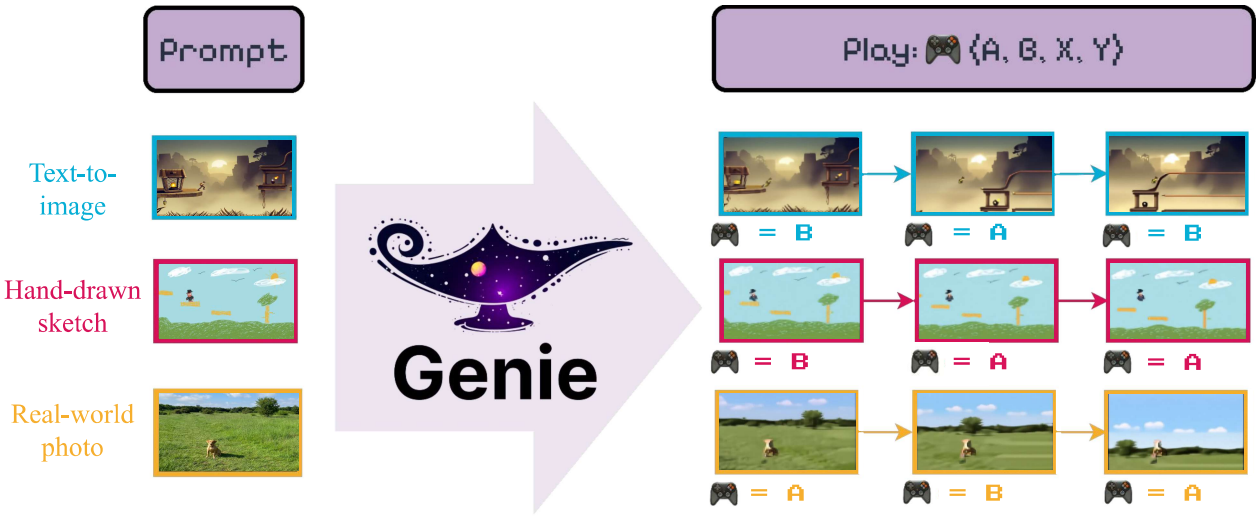

Figure 1 | **A whole new world**: Genie is capable of converting a variety of different prompts into interactive, playable environments that can be easily created, stepped into, and explored. This is made possible via a latent action interface, learned fully unsupervised from Internet videos. On the right we see a few generated steps for taking two latent actions. See more examples on our website.

We introduce Genie, the first *generative interactive environment* trained in an unsupervised manner from unlabelled Internet videos. The model can be prompted to generate an endless variety of action-controllable virtual worlds described through text, synthetic images, photographs, and even sketches. At 11B parameters, Genie can be considered a *foundation world model*. It is comprised of a spatiotemporal video tokenizer, an autoregressive dynamics model, and a simple and scalable latent action model. Genie enables users to act in the generated environments on a frame-by-frame basis *despite training without any ground-truth action labels* or other domain-specific requirements typically found in the world model literature. Further the resulting learned latent action space facilitates training agents to imitate behaviors from unseen videos, opening the path for training generalist agents of the future.

Keywords: Generative AI, Foundation Models, World Models, Video Models, Open-Endedness

Corresponding author(s): Ashley Edwards (edwardsashley@google.com), Jack Parker-Holder (jparkerholder@google.com).

1. Introduction

The last few years have seen an emergence of *generative AI*, with models capable of generating novel and creative content. Driven by breakthroughs in architectures such as transformers (Vaswani et al., 2017), advances in hardware, and a recent focus on scaling models and datasets, we can now generate coherent, conversational language (Brown et al., 2020; Radford et al., 2018, 2019), as well as crisp and aesthetically pleasing images from a text prompt (Ramesh et al., 2021, 2022; Rombach et al., 2022; Saharia et al., 2022). Early signs indicate video generation will be yet another frontier, with recent results suggesting that such models may also benefit from scale (Blattmann et al., 2023a; Esser et al., 2023; Ho et al., 2022a; Hong et al., 2023). Still, there remains a gulf between the level of interactions and engagement of video generative models and language tools such as ChatGPT, let alone more immersive experiences.

What if, given a large corpus of videos from the Internet, we could not only train models capable of generating novel images or videos, but entire interactive experiences? We propose *generative interactive environments*, a new paradigm for generative AI whereby interactive environments can be generated from a single text or image prompt. Our approach, Genie, is trained from a large dataset of over 200,000 hours of publicly available Internet gaming videos and, despite training *without action or text annotations*, is controllable on a frame-by-frame basis via a learned latent action space (see Table 1 for a comparison to other approaches). At 11B parameters, Genie exhibits properties typically seen in foundation models—it can take an unseen image as a prompt making it possible to create and play entirely imagined virtual worlds (e.g Figure 2).

Genie builds on ideas from state-of-the-art video generation models (Gupta et al., 2023; Villegas et al., 2023), with a core design choice being spatiotemporal (ST) transformers (Xu et al., 2020) which are used in all of our model components. Genie utilizes a novel video tokenizer, and extracts latent actions via a causal action model. Both the video tokens and latent actions

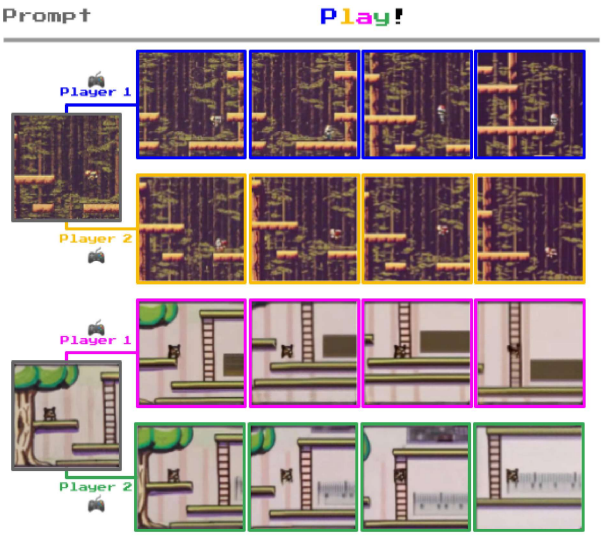

Figure 2 | **Diverse trajectories**: Genie is a generative model that can be used as an interactive environment. The model can be prompted in various ways, either with a generated image (top) or a hand-drawn sketch (bottom). At each time step, the model takes a user-provided latent action to generate the next frame, producing trajectories with interesting and diverse character actions.

are passed to a dynamics model, which autoregressively predicts the next frame using MaskGIT (Chang et al., 2022). We provide a rigorous scaling analysis of our architecture with respect to both batch and model size, which we vary from 40M to 2.7B parameters. The results show that our architecture scales gracefully with additional computational resources, leading to a final 11B parameter model. We train Genie on a filtered set of 30,000 hours of Internet gameplay videos from hundreds of 2D platformer games, producing a foundation world model for this setting.

To demonstrate the generality of our approach, we also train a separate model on action-free robot videos from the RT1 dataset (Brohan et al., 2023), learning a generative environment with consistent latent actions. Finally, we show that latent actions learned from Internet videos can be used for inferring policies from unseen action-free videos of simulated reinforcement learning (RL) environments, indicating that Genie may hold the key to unlocking unlimited data for training the next generation of generalist agents (Bauer et al.,

2023; Clune, 2019; Open Ended Learning Team et al., 2021; Reed et al., 2022).

Table 1 | **A new class of generative model**: Genie is a novel video and world model that is controllable on a frame-by-frame basis, which requires **only video data** at train time.

Model Class	Training Data	Controllability
World Models	Video + Actions	Frame-level
Video Models	Video + Text	Video-level
Genie	Video	Frame-level

2. Methodology

Genie is a generative interactive environment trained from video-only data. In this section we begin with preliminaries before explaining the main components of our model.

Several components in the Genie architecture are based on the Vision Transformer (ViT) (Dosovitskiy et al., 2021; Vaswani et al., 2017). Notably, the quadratic memory cost of transformers poses challenges for videos, which can contain up to $O(10^4)$ tokens. We thus adopt a memory efficient ST-transformer architecture (inspired by Xu et al. (2020), see Figure 4) across all model components, balancing model capacity with computational constraints.

Unlike a traditional transformer where every token attends to all others, an ST-transformer contains L spatiotemporal blocks with interleaved spatial and temporal attention layers, followed by a feed-forward layer (FFW) as standard attention blocks. The self-attention in the spatial layer attends over the $1 \times H \times W$ tokens within each time step, and in the temporal layer attends over $T \times 1 \times 1$ tokens across the T time steps. Similar to sequence transformers, the temporal layer assumes a causal structure with a causal mask. Crucially, the dominating factor of computation complexity (i.e. the spatial attention layer) in our architecture scales linearly with the number of frames rather than quadratically, making it much more efficient for video generation with consistent dynamics over extended interactions. Further, note that in the ST block, we include only one FFW after both spatial and temporal components, omitting the post-spatial FFW to allow for scaling up other components of the model, which we observe to improve results significantly.

2.1. Model Components

As shown in Figure 3, our model contains three key components: 1) a ***latent action model*** that infers the latent action $\boldsymbol{a}$ between each pair of frames and 2) a ***video tokenizer*** that converts raw video frames into discrete tokens $\boldsymbol{z}$ and 3) a ***dynamics model*** that, given a latent action and past frame tokens, predicts the next frame of the video. The model is trained in two phases following a standard autoregressive video generation pipeline: we train the video tokenizer first, which is used for the dynamics model. We then co-train the latent action model (directly from pixels) and the dynamics model (on video tokens).

Latent Action Model (LAM) To achieve controllable video generation, we condition each future frame prediction on the action taken at the previous frame. However, such action labels are rarely available in videos from the Internet and action annotation can be costly to obtain. Instead, we learn *latent actions* in a fully unsupervised manner (see Figure 5).

First, an encoder takes as inputs all previous frames $\boldsymbol{x}_{1:t} = (x_1, \cdots x_t)$ as well as the next frame x_{t+1}, and outputs a corresponding set of continuous latent actions $\tilde{\boldsymbol{a}}_{1:t} = (\tilde{a}_1, \cdots \tilde{a}_t)$. A decoder then takes all previous frames and latent actions as input and predicts the next frame $\hat{x}_{t+1}$.

To train the model, we leverage a VQ-VAE-based objective (van den Oord et al., 2017), which enables us to limit the number of predicted actions to a small discrete set of codes. We limit the vocabulary size $|A|$ of the VQ codebook, i.e. the maximum number of possible latent actions, to a small value to permit human playability and further enforce controllability (we use $|A| = 8$ in our experiments). As the decoder only has access to the history and latent action, $\tilde{a}_t$ should encode the most meaningful changes between the past and the future for the decoder to successfully reconstruct the future frame. Note that this de-