DeepMind

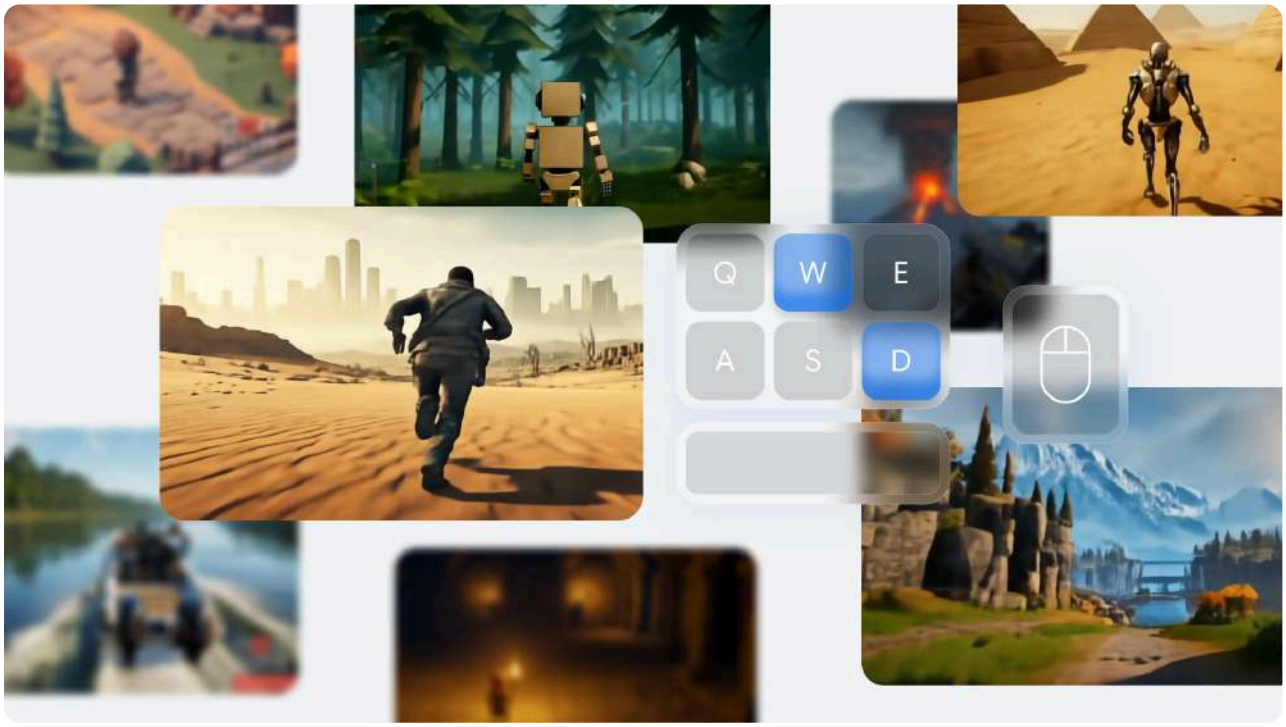
RESEARCH

Genie 2: A large-scale foundation world model

4 DECEMBER 2024

Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, Stephen Spencer, Jessica Yung, Michael Dennis, Sultan Kenjeyev, Shangbang Long, Vlad Mnih, Harris Chan, Maxime Gazeau, Bonnie Li, Fabio Pardo, Luyu Wang, Lei Zhang, Frederic Besse, Tim Harley, Anna Mitenkova, Jane Wang, Jeff Clune, Demis Hassabis, Raia Hadsell, Adrian Bolton, Satinder Singh, Tim Rocktäschel

Share



Generating unlimited diverse training environments for future general agents

Today we introduce Genie 2, a foundation world model capable of generating an endless variety of action-controllable, playable 3D environments for training and

DeepMind

Games play a key role in the world of artificial intelligence (AI) research. Their engaging nature, unique blend of challenges, and measurable progress make them ideal environments to safely test and advance AI capabilities.

Indeed, games have been important to Google DeepMind since our founding. From our early work with Atari games, breakthroughs such as AlphaGo and AlphaStar, to our research on generalist agents in collaboration with game developers, games have been center stage in our research. However, training more general embodied agents has been traditionally bottlenecked by the availability of sufficiently rich and diverse training environments.

As we show, Genie 2 could enable future agents to be trained and evaluated in a limitless curriculum of novel worlds. Our research also paves the way for new, creative workflows for prototyping interactive experiences.

Emergent capabilities of a foundation world model

DeepMind

generality. Genie 2 can generate a vast diversity of rich 3D worlds.

Genie 2 is a *world model*, meaning it can simulate virtual worlds, including the consequences of taking any action (e.g. jump, swim, etc.). It was trained on a large-scale video dataset and, like other generative models, demonstrates various emergent capabilities at scale, such as object interactions, complex character animation, physics, and the ability to model and thus predict the behavior of other agents.

Below are example videos of people interacting with Genie 2. For every example, the model is prompted with a single image generated by Imagen 3, GDM's state-of-the-art text-to-image model. This means anyone can describe a world they want in text, select their favorite rendering of that idea, and then step into and interact with that newly created world (or have an AI agent be trained or evaluated in it). At each step, a person or agent provides a keyboard and mouse action, and Genie 2 simulates the next observation. Genie 2 can generate consistent worlds for up to a minute, with the majority of examples shown lasting 10-20s.

Action controls

Genie 2 responds intelligently to actions taken by pressing keys on a keyboard, identifying the character and moving it correctly. For example, our model has to figure out that arrow keys should move the robot and not the trees or clouds.

W (forward) A (move left) S (backward) D (move right) Space (jump)

A cute humanoid robot in the woods.

DeepMind

A humanoid robot in Ancient Egypt.

A first person view of a robot on a purple planet.

A first person view of a robot in a loft apartment in a big city.

Generating counterfactuals

We can generate diverse trajectories from the same starting frame, which means it is possible to simulate counterfactual experiences for training agents. In each row, each video starts from the same frame, but has different actions taken by a human player.

DeepMind

Long horizon memory

Genie 2 is capable of remembering parts of the world that are no longer in view and then rendering them accurately when they become observable again.

Long video generation with new generated content

DeepMind

Diverse environments

Genie 2 can create different perspectives, such as first-person view, isometric views, or third person driving videos.

DeepMind

3D structures

Genie 2 learned to create complex 3D visual scenes.

Object affordances and interactions

DeepMind

Character animation

Genie 2 learned how to animate various types of characters doing different activities.

DeepMind

NPCs

Genie 2 models other agents and even complex interactions with them.

DeepMind

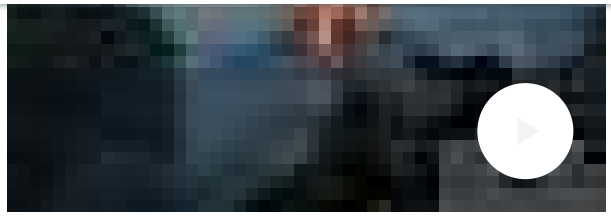
Physics

Genie 2 models water effects.

Smoke

Genie 2 models smoke effects.

DeepMind



Gravity

Genie 2 models gravity.

Lighting

Genie 2 models point and directional lighting.

DeepMind

Reflections

Genie 2 models reflections, bloom and coloured lighting.

Playing from real world images

Genie 2 can also be prompted with real world images, where we see that it can model grass blowing in the wind or water flowing in a river.

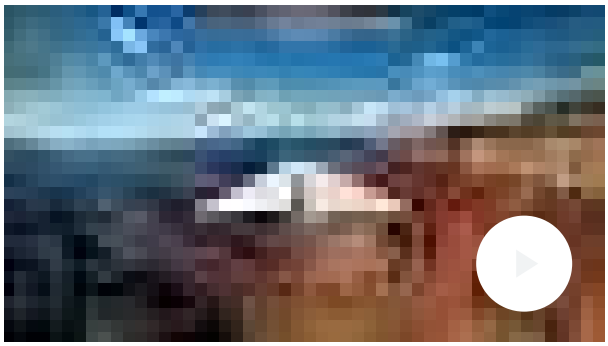
DeepMind

Genie 2 prompted with real world photos.

Genie 2 enables rapid prototyping

Genie 2 makes it easy to rapidly prototype diverse interactive experiences, enabling researchers to quickly experiment with novel environments to train and test embodied AI agents.

For example, below we prompt Genie 2 with different images generated by Imagen 3 to model the difference between flying a paper plane, a dragon, a hawk, or a parachute and test how well Genie can animate different avatars.



DeepMind

Thanks to Genie 2's out-of-distribution generalization capabilities, concept art and drawings can be turned into fully interactive environments. This enables artists and designers to prototype quickly, which can bootstrap the creative process for environment design, further accelerating research.

Here we show examples of research environment concepts made by our concept artist.

Environment concept by Max Cant

Genie 2

Environment concept by Max Cant

Genie 2

AI agents acting inside the world model

DeepMind

collaboration with games developers, following instructions on unseen environments synthesized by Genie 2 via a single image prompt.

Image generated by Imagen 3

Prompt: "A screenshot of a third-person open world exploration game. The player is an adventurer exploring a forest. There is a house with a red door on the left, and a house with a blue door on the right. The camera is placed directly behind the player. #photorealistic #immersive"

The SIMA agent is designed to complete tasks in a range of 3D game worlds by following natural-language instructions. Here we used Genie 2 to generate a 3D environment with two doors, a blue and a red one, and provided instructions to the SIMA agent to open each of them. In this example, SIMA is controlling the avatar via keyboard and mouse inputs, while Genie 2 generates the game frames.

DeepMind

Prompt “Open the blue door”

Prompt “Open the red door”

We can also use SIMA to help evaluate Genie 2’s capabilities. Here we test Genie 2’s ability to generate consistent environments by instructing SIMA to look around and explore behind the house.

Prompt “Turn around”

Prompt “Go behind the house”

While this research is still in its early stage with substantial room for improvement on both agent and environment generation capabilities, we believe Genie 2 is the path to solving a structural problem of training embodied agents safely while achieving the breadth and generality required to progress towards AGI.

DeepMind

Image generated by Imagen 3

Prompt: "An image of a computer game showing a scene from inside a rough hewn stone cave or mine. The viewer's position is a 3rd person camera based above a player avatar looking down towards the avatar. The player avatar is a knight with a sword. In front of the knight avatar there are x3 stone arched doorways and the knight chooses to go through any one of these doors. Beyond the first and inside we can see strange green plants with glowing flowers lining that tunnel. Inside and beyond the second doorway there is a corridor of spiked iron plates riveted to the cave walls leading towards an ominous glow further along. Through the third door we can see a set of rough hewn stone steps ascending to a mysterious destination."

Prompt “Go up the stairs”

Prompt “Go where the plants are”

DeepMind

Prompt “Go to the middle door”

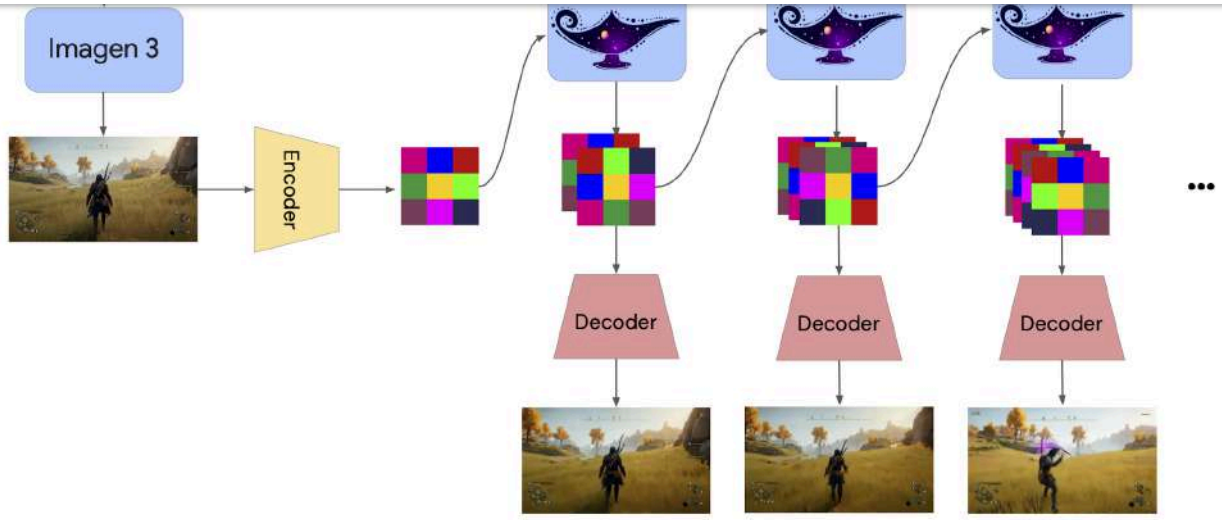
Diffusion world model

Genie 2 is an autoregressive latent diffusion model, trained on a large video dataset. After passing through an autoencoder, latent frames from the video are passed to a large transformer dynamics model, trained with a causal mask similar to that used by large language models.

At inference time, Genie 2 can be sampled in an autoregressive fashion, taking individual actions and past latent frames on a frame-by-frame basis. We use classifier-free guidance to improve action controllability.

The samples in this blog post are generated by an undistilled base model, to show what is possible. We can play a distilled version in real-time with a reduction in quality of the outputs.

DeepMind



Developing our technologies responsibly

Genie 2 shows the potential of foundational world models for creating diverse 3D environments and accelerating agent research. This research direction is in its early stages and we look forward to continuing to improve Genie's world generation capabilities in terms of generality and consistency.

As with SIMA, our research is building towards more general AI systems and agents that can understand and safely carry out a wide range of tasks in a way that is helpful to people online and in the real world.

Interesting outtakes

DeepMind

While not taking any action, a ghost appears while in a garden

The character prefers parkour over snowboarding.

With great power comes great responsibility.

Acknowledgements

Genie 2 was led by Jack Parker-Holder with technical leadership by Stephen Spencer, with key contributions from Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi and Jessica Yung, and contributions from Michael Dennis, Sultan Kenjeyev and Shangbang Long. Yusuf Aytar, Jeff Clune, Sander Dieleman, Doug Eck, Shlomi Fruchter, Raia Hadsell, Demis Hassabis, Georg Ostrovski, Pieter-Jan Kindermans, Nicolas Heess, Charles Blundell, Simon Osindero, Rushil Mistry gave advice. Past contributors include Ashley Edwards and Richie Steigerwald.

The Generalist Agents team was led by Vlad Mnih with key contributions from Harris Chan, Maxime Gazeau, Bonnie Li, Fabio Pardo, Luyu Wang, Lei Zhang

The SIMA team, with particular support from Frederic Besse, Tim Harley, Anna Mitenkova and Jane Wang

DeepMind

We'd also like to thank Zoubin Gharamani, Andy Brock, Ed Hirst, David Bridson, Zeb Mehring, Cassidy Hardin, Hyunjik Kim, Noah Fiedel, Jeff Stanway, Petko Yotov, Mihai Tiuca, Soheil Hassas Yeganeh, Nehal Mehta, Richard Tucker, Tim Brooks, Alex Cullum, Max Cant, Nik Hemmings, Richard Evans, Valeria Oliveira, Yanko Gitahy Oliveira, Bethanie Brownfield, Charles Gbadamosi, Giles Ruscoe, Guy Simmons, Jony Hudson, Marjorie Limont, Nathaniel Wong, Sarah Chakera, Nick Young.

Related posts

View all posts

RESEARCH

A generalist AI agent for 3D virtual environments

Introducing SIMA, a Scalable Instructable Multiworld Agent

13 MARCH 2024

Follow us

About

About Google DeepMind