# Genie: Generative Interactive Environments

Google DeepMind

2024-2-26

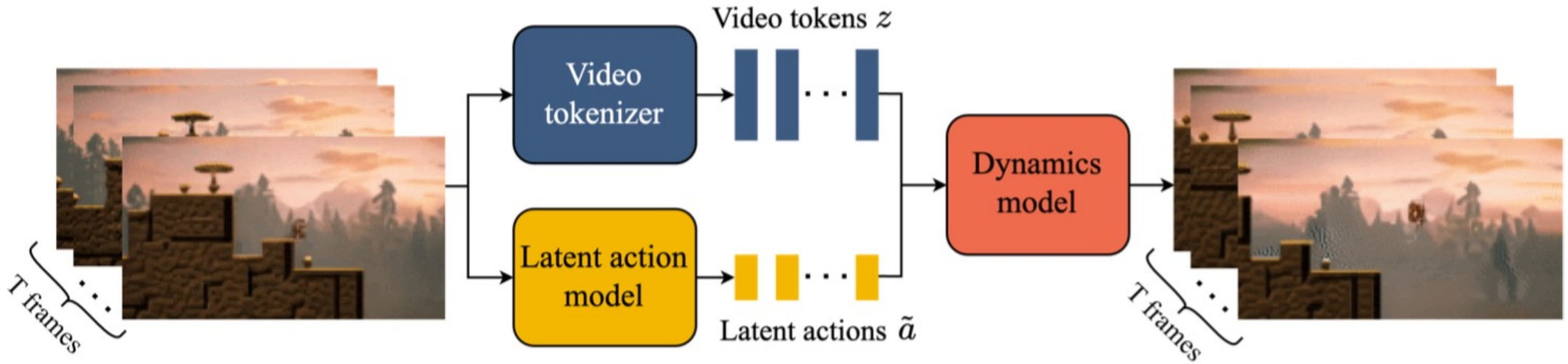
# Genie: Generative Interactive Environments

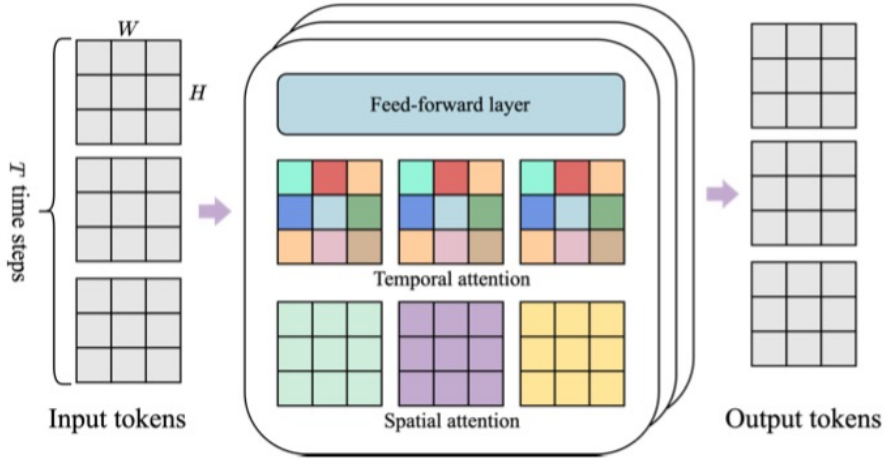
**Jake Bruce**[*,1], **Michael Dennis**[*,1], **Ashley Edwards**[*,1], **Jack Parker-Holder**[*,1], **Yuge (Jimmy) Shi**[*,1], **Edward Hughes**[1], **Matthew Lai**[1], **Aditi Mavalankar**[1], **Richie Steigerwald**[1], **Chris Apps**[1], **Yusuf Aytar**[1], **Sarah Bechtle**[1], **Feryal Behbahani**[1], **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

ICML 2024 Best Paper

# Training Pipeline



# Model

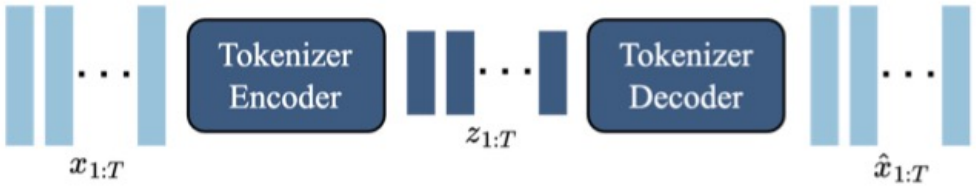



Figure 6 | **Video tokenizer**: a VQ-VAE with ST-transformer.

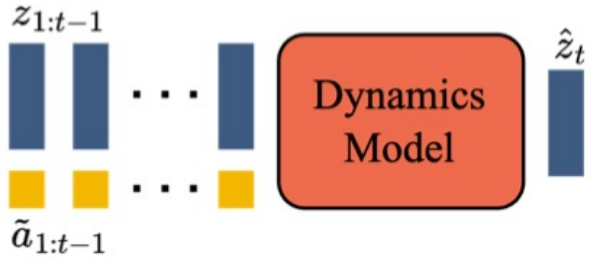



Figure 7 | **Dynamics model**: takes in video tokens and action embeddings, and predicts future masked video tokens.

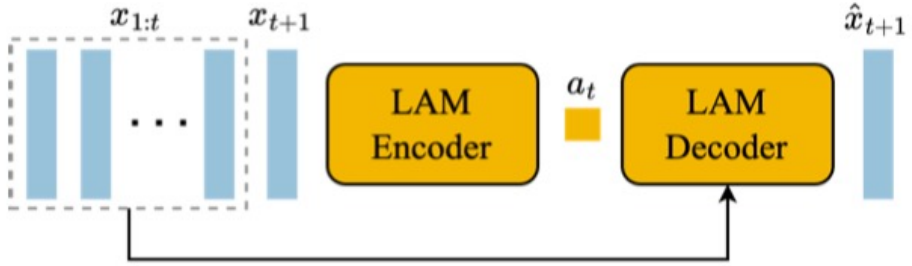
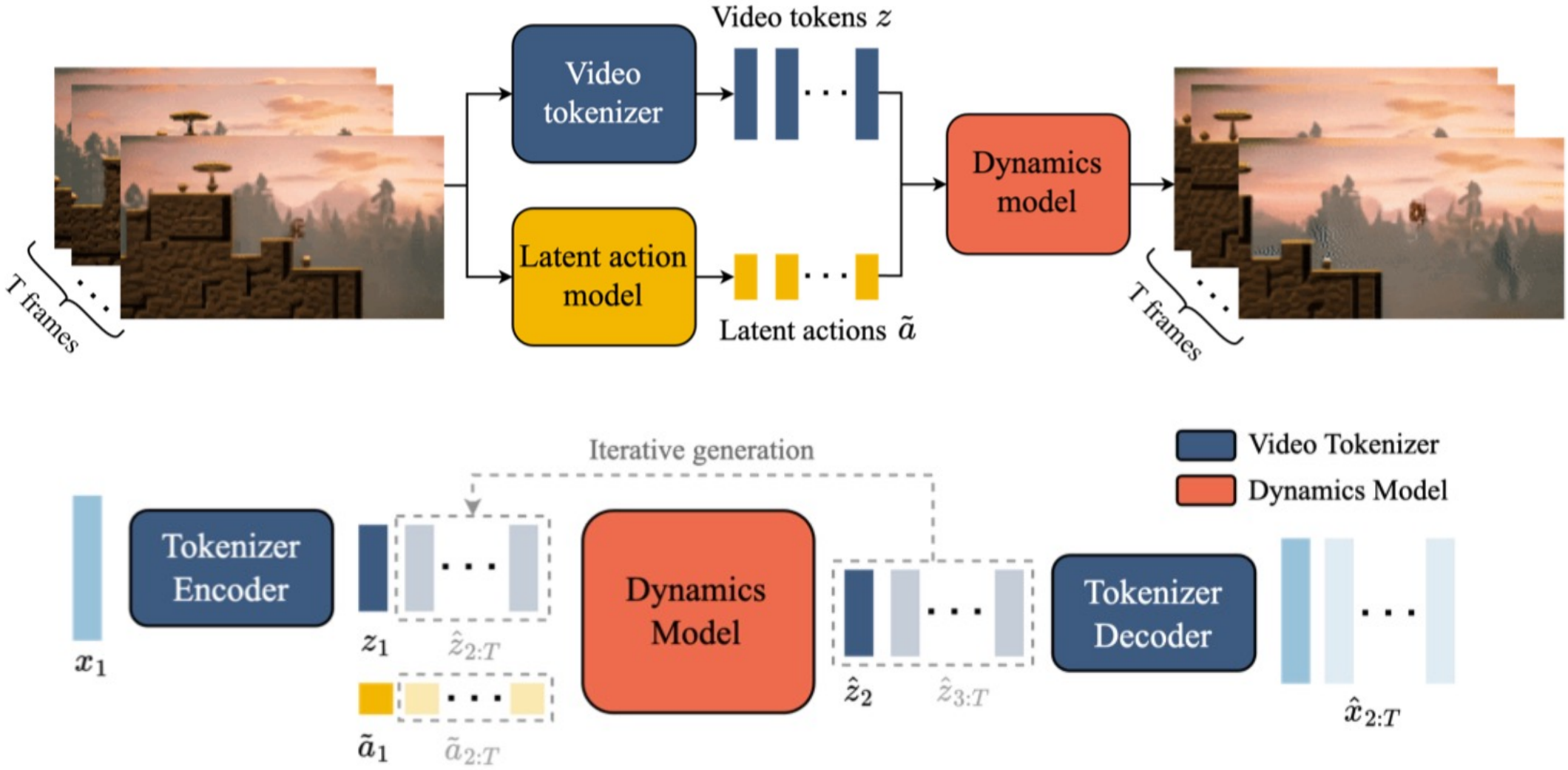



Figure 5 | **Latent action model**: learns actions $a_t$ unsupervised from unlabelled video frames.

# Model


Video tokens $z$
Video tokenizer
Latent action model
Latent actions $\tilde{a}$
Dynamics model
T frames
T frames
Iterative generation
Video Tokenizer
Dynamics Model
Tokenizer Encoder
Dynamics Model
Tokenizer Decoder
$x_1$
$z_1$
$\hat{z}_{2:T}$
$\tilde{a}_1$
$\tilde{a}_{2:T}$
$\hat{z}_2$
$\hat{z}_{3:T}$
$\hat{x}_{2:T}$

# Evaluation

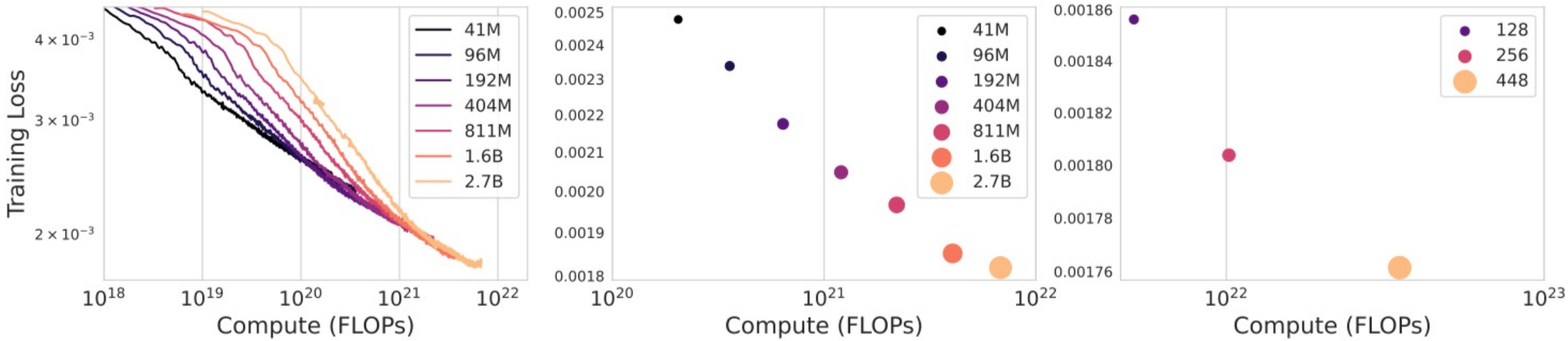



Figure 9 | **Scaling results**. **Left**: Training curves for different model sizes, **Middle**: Final training loss for each model size, averaged over the last 300 updates, **Right**: Final training loss for a 2.3B model with different batch sizes.


Figure 10 | **Playing from Image Prompts**: We can prompt Genie with images generated by text-to-image models, hand-drawn sketches or real-world photos. In each case we show the prompt frame and a second frame after taking one of the latent actions four consecutive times. In each case we see clear character movement, despite some of the images being visually distinct from the dataset.


Figure 13 | **Controllable, consistent latent actions in Robotics**: trajectories beginning from three different starting frames from our Robotics dataset. Each column shows the resulting frame from taking the same latent action five times. Despite training without action labels, the same actions are consistent across varied prompt frames and have semantic meaning: *down*, *up* and *left*.

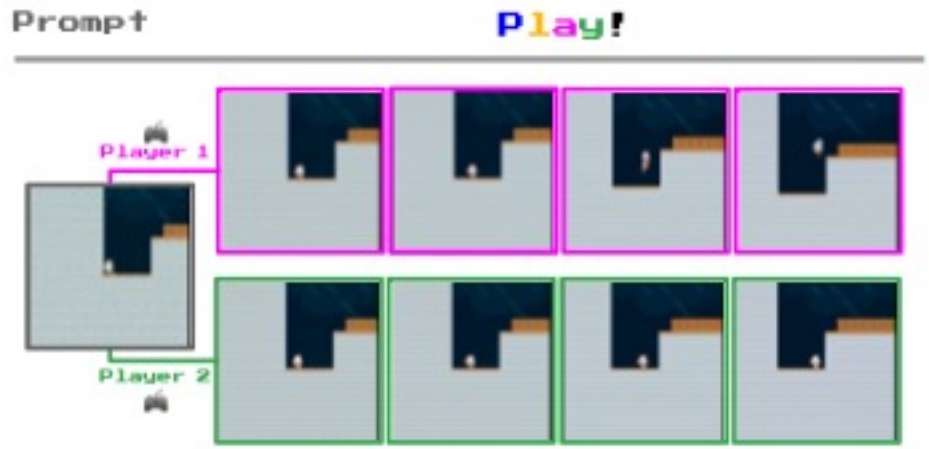



Figure 14 | **Playing from RL environments**: Genie can generate diverse trajectories given an image of an unseen RL environment.

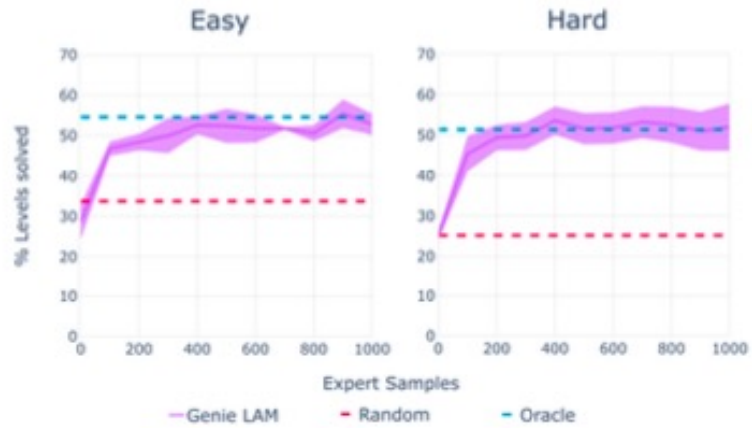



Figure 15 | **BC results**. Mean percentage of levels solved out of 100 samples, averaged over 5 seeds with 95% confidence intervals.

Thank You