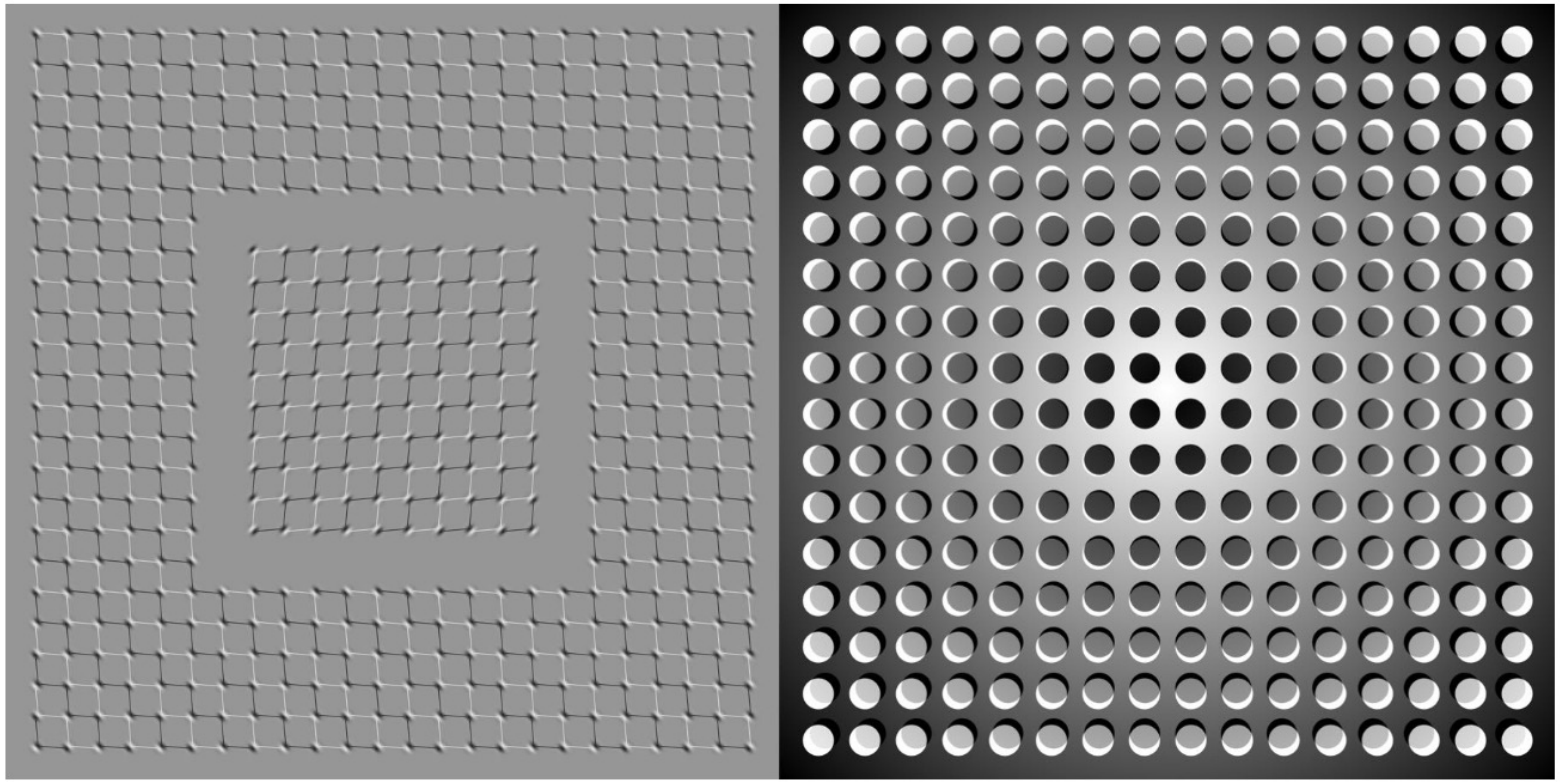
Genie: Generative Interactive Environments

Yuting Hu@XLab-UB

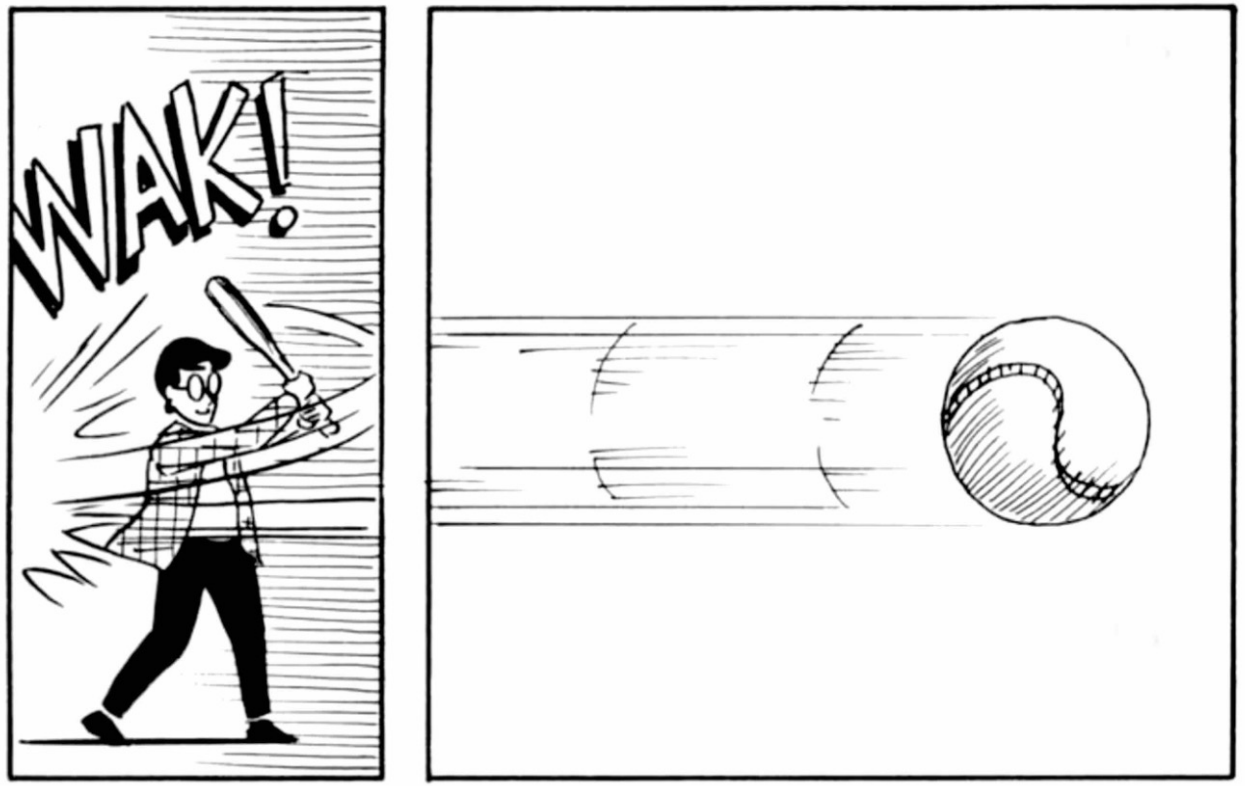
2023.03.01



What we see is based on our brain's prediction of the future.

What is World Model ?

"World Model" generally refer to a computational representation of the physical world, capable of **simulating changes in the world's state in response to various actions**.



Human brain produces mind by modeling.

Existing Distance Between Generative AI and World Model

OpenAI Sora Text-to-Video Model: Generative Virtual Videos

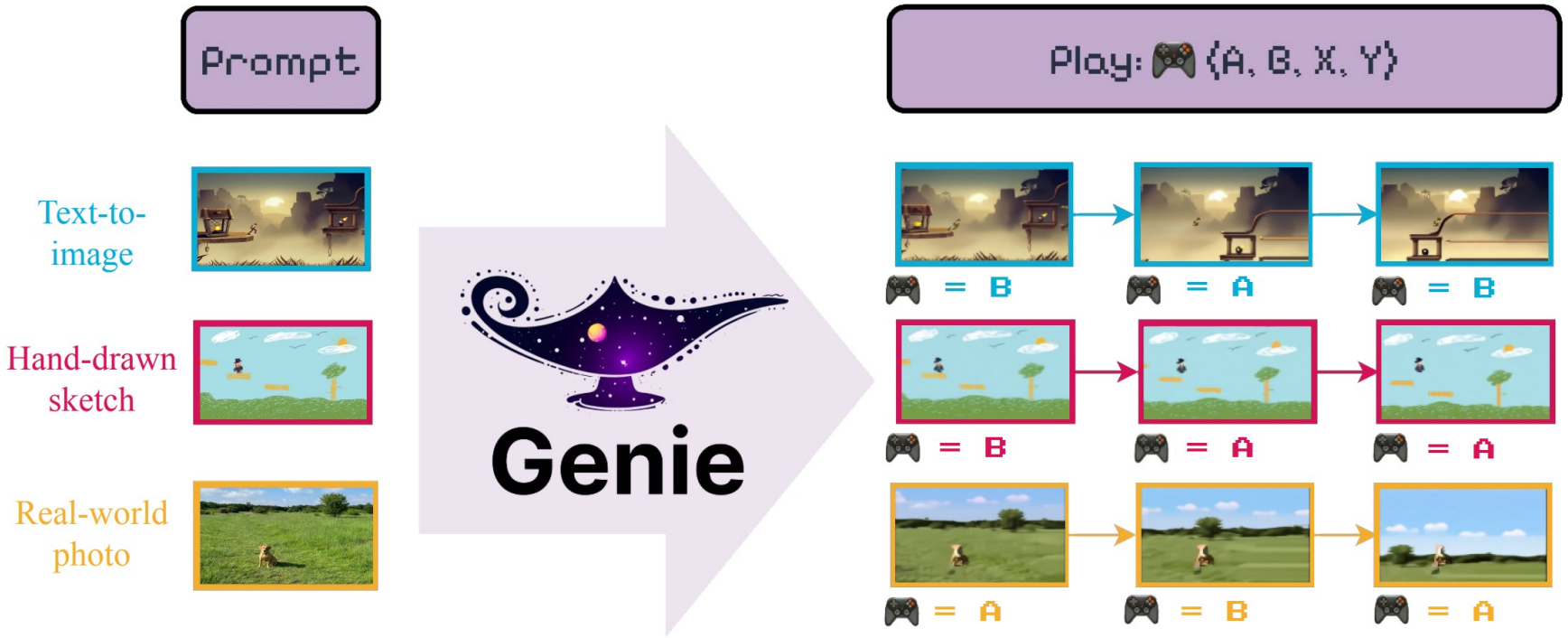
Prompt: Animated scene features a close-up of a short fluffy monster kneeling beside a melting red candle. The art style is 3D and realistic, with a focus on lighting and texture. The mood of the painting is one of wonder and curiosity, as the monster gazes at the flame with wide eyes and open mouth. Its pose and expression convey a sense of innocence and playfulness, as if it is exploring the world around it for the first time. The use of warm colors and dramatic lighting further enhances the cozy atmosphere of the image.

Not Interactive Environment.

Challenges of Developing World Models

- Ground-truth action labels or text annotations of videos are needed.
- Realtime interactions require low computation complexity.
- High-resolution environments for immersive interactive experiences.
- Mitigate hallucinations of model generation.

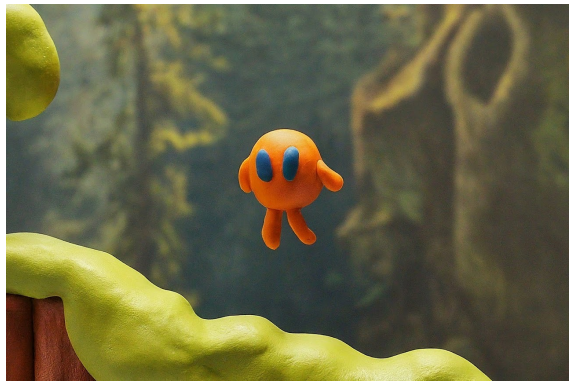
Genie: Making Virtual World be Interactive.

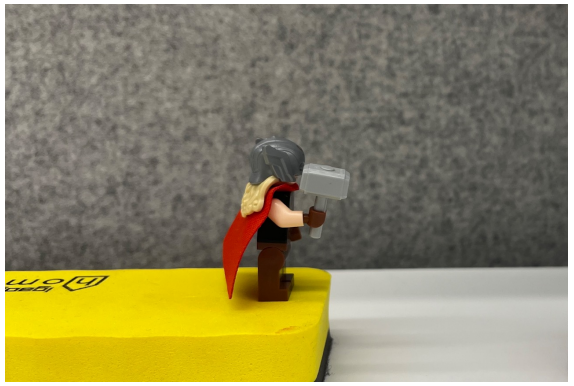


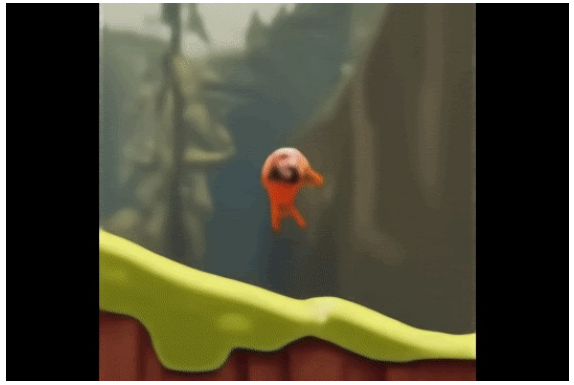
Generate an endless variety of action controllable virtual worlds described through text, synthetic images, photographs, and even sketches.

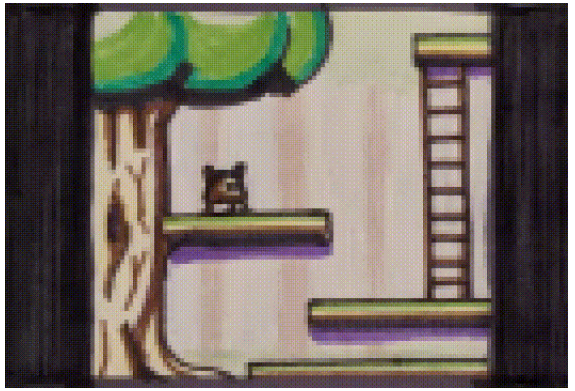
A Foundation World Model.

Genie: Making Virtual World be Interactive.









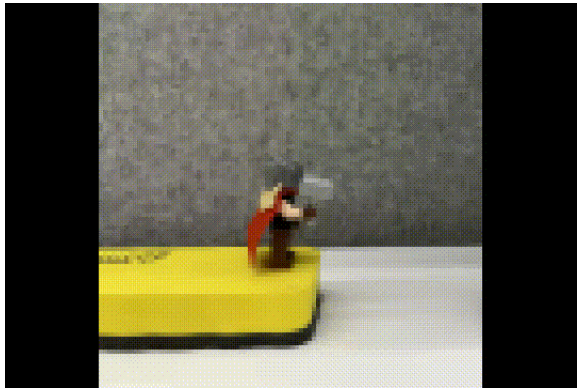
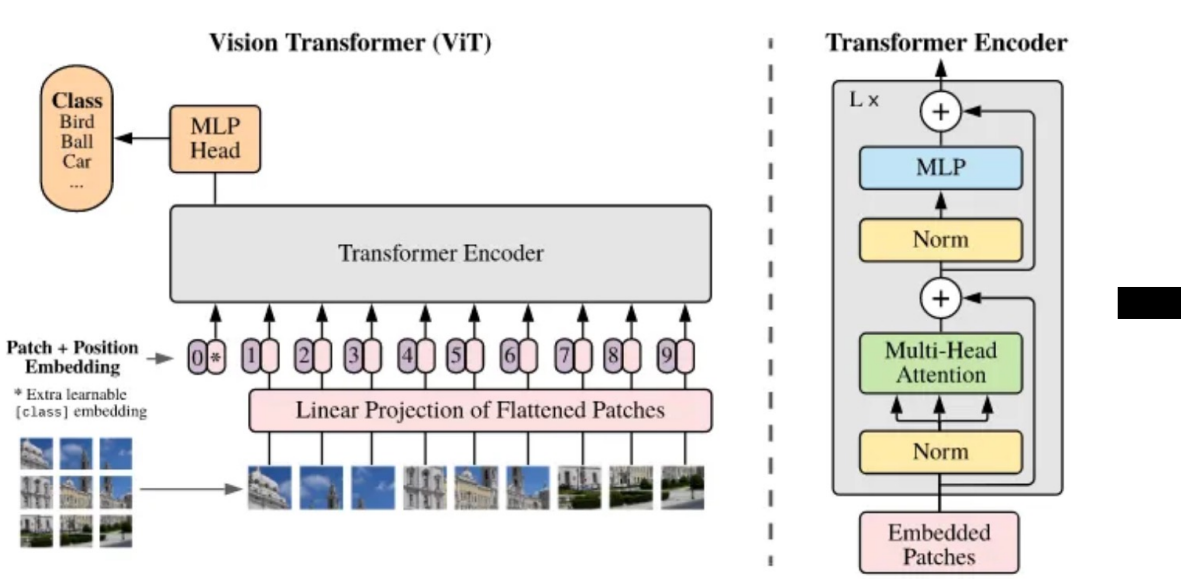


Image Generated with Imagen2

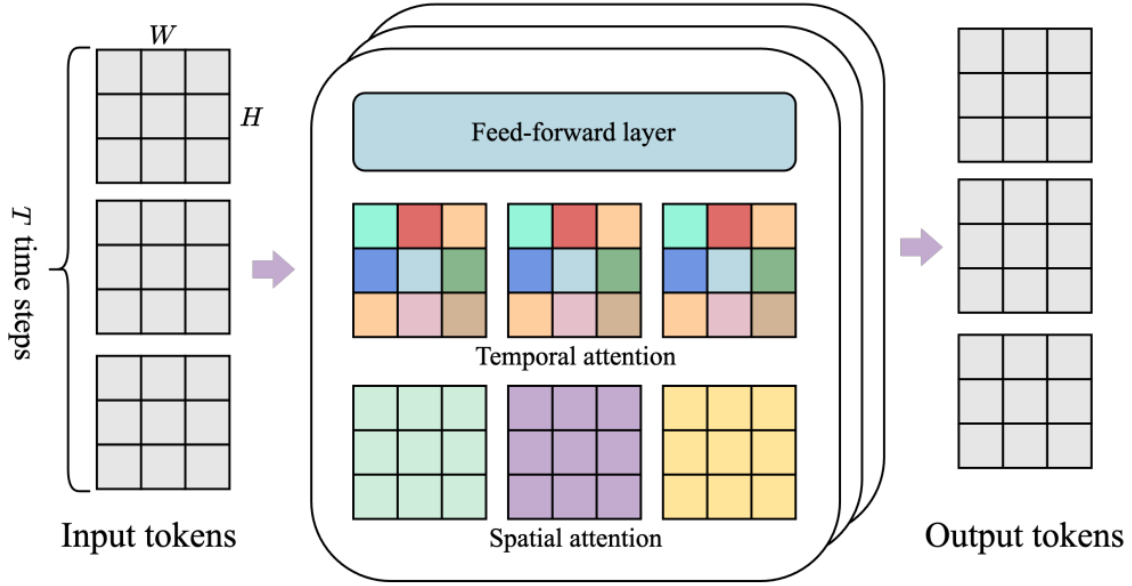
Human Scratch

Real World Image

Genie Model Architecture (1) ST-Transformer



Vision Transformer
(computation, memory are quadratic with input sequence length)

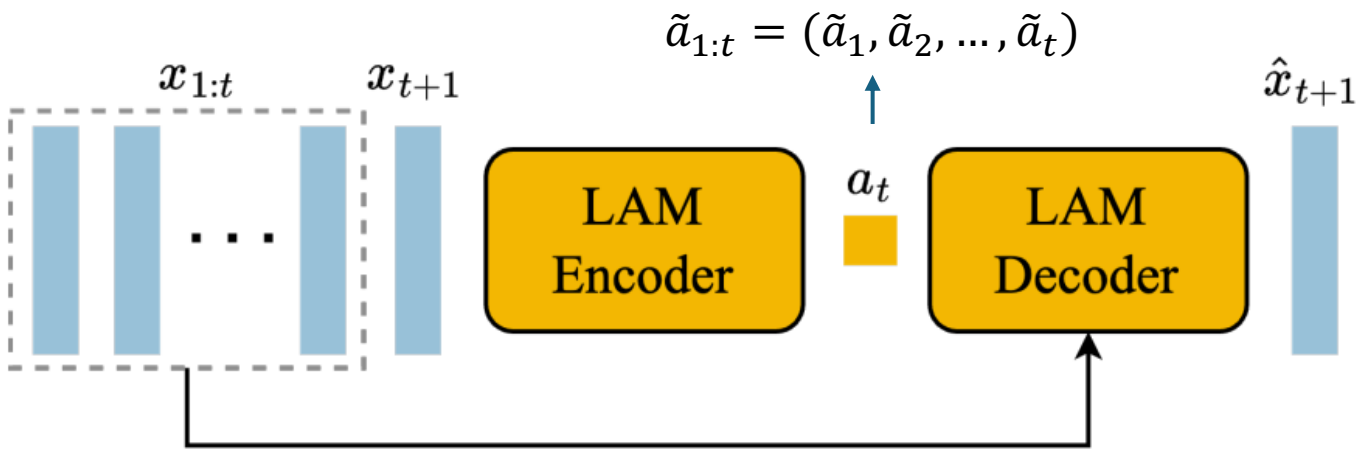


ST-Transformer
(computation, memory scales linearly with number of frames)

ST-Transformer is used in all Genie model components.

Genie Model Architecture (2) LAM Encoder

Condition each future frame is generated by the action taken at the previous frame.
LAM(Latent Action Model) learn latent actions in unsupervised manner.

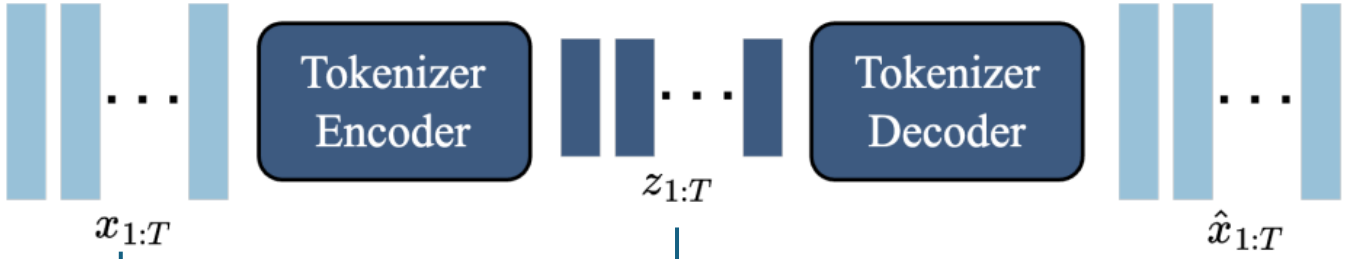


VQ-VAE Based Objective Function

Entire LAM is discarded at inference stage and replace by actions from the user.

Genie Model Architecture (3) Tokenizer

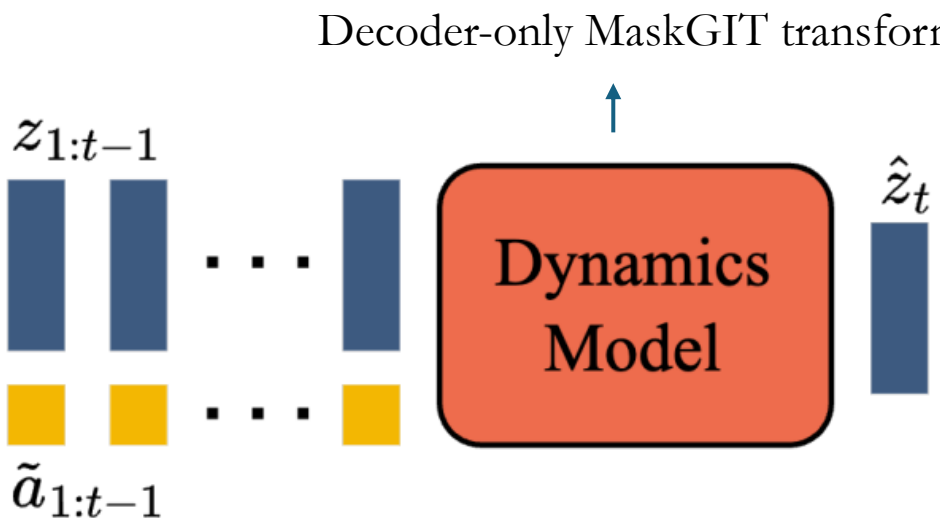
Compress videos into discrete tokens to reduce dimensionality.



$x_{1:T} = (x_1, x_2, \ldots, x_T) \in R^{T \times H \times W \times C}$ $\quad$ $Z_{1:T} = (z_1, z, \ldots, z_T) \in R^{T \times D}$

VQ-VAE Based Objective Function

Genie Model Architecture (4) Dynamics Model



Trained with cross-entropy loss between predicted tokens $\tilde{z}_{2:T}$ and ground-truth tokens $z_{2:T}$

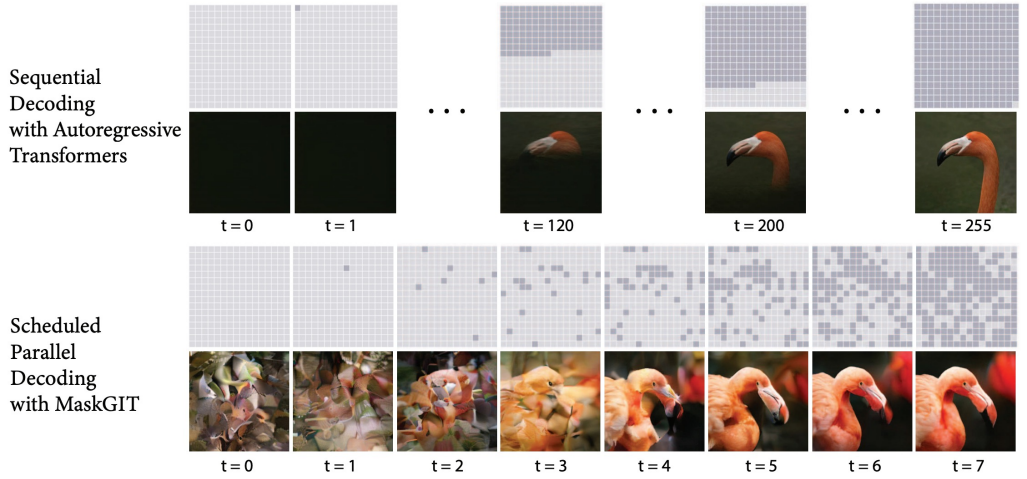
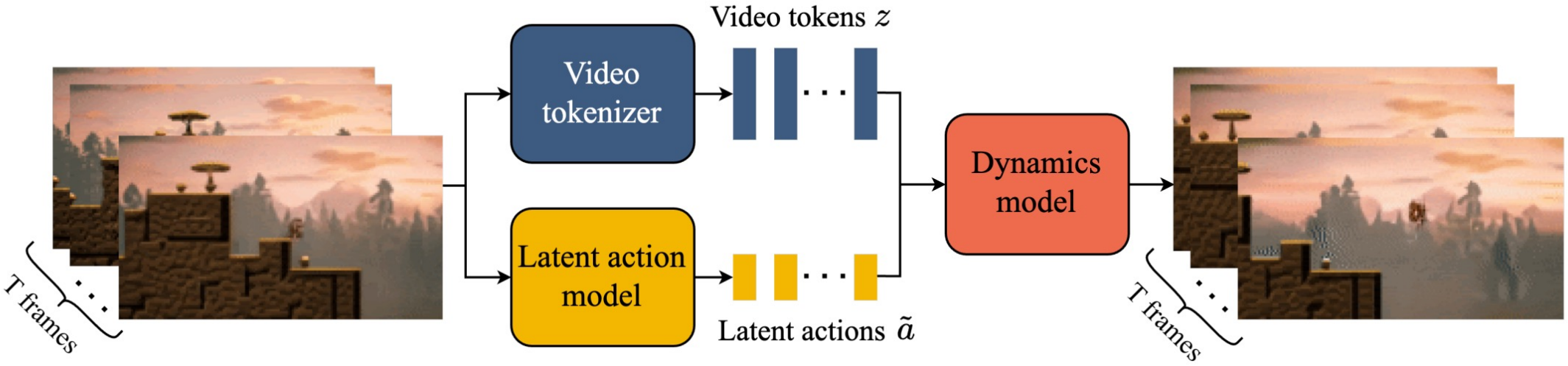


Figure 2. **Comparison between sequential decoding and MaskGIT's scheduled parallel decoding.** Rows 1 and 3 are the input latent masks at each iteration, and rows 2 and 4 are samples generated by each model at that iteration. Our decoding starts with all unknown codes (marked in lighter gray), and gradually fills up the latent representation with more and more scattered predictions in parallel (marked in darker gray), where the number of predicted tokens increases sharply over iterations. MaskGIT finishes its decoding in 8 iterations compared to the 256 rounds the sequential method takes.

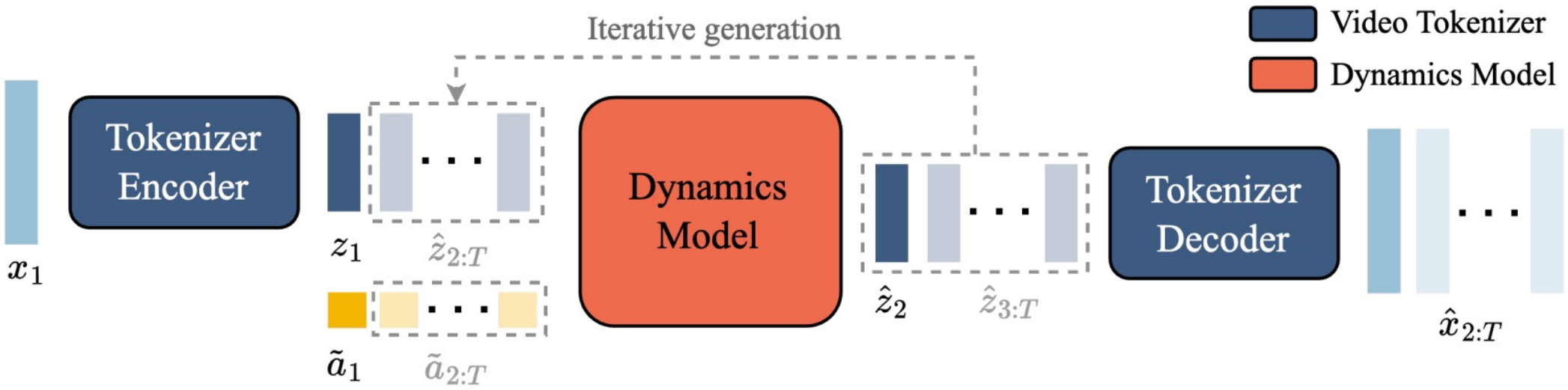
Genie Model Architecture (4) Overview



Trained in two phases following a standard autoregressive video generation pipeline:

1) Train video tokenizer first;
2) Co-train LAM(from pixels) and dynamics model(video tokens);

Genie Model Inference



Experiment Setting

Dataset

1) Platfomers (Internet videos of 2D games): 55M 16s Video clips at 10FPS with 160x90 resolution.
2) Robotics (robot data): ~130k(Simulation Robot) + 209k episodes (Real Robot)

(Simply treat as videos, no action labels are used.)

Metrics

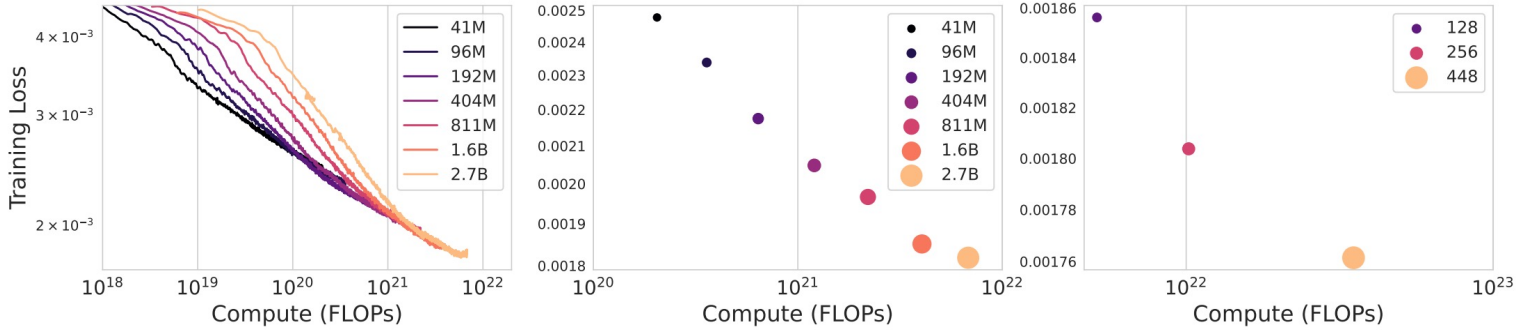
1) Video fidelity: Frechet Video Distance (FVD)
2) Controllability: $\Delta_t\text{PSNR} = \text{PSNR}(x_t, \hat{x}_t) - \text{PSNR}(x_t, \hat{x}'_t)$

where x_t denotes the ground-truth frame at time t, $\tilde{x}_t$ denotes the frame from latent actions $\tilde{a}_{1:t}$ inferred from ground-truth frames, and $\tilde{x}_t'$ the same frame generated from a sequence of latent actions randomly sampled from a categorical distribution. As such, the greater Δ_tPSNR is, the more the video generated from random latent actions differs from ground-truth, which indicates a higher level of controllability from the latent actions.

Released 11B-parameter Gemini model is trained with 6.8M 16s video clips (30k Hours)

Results

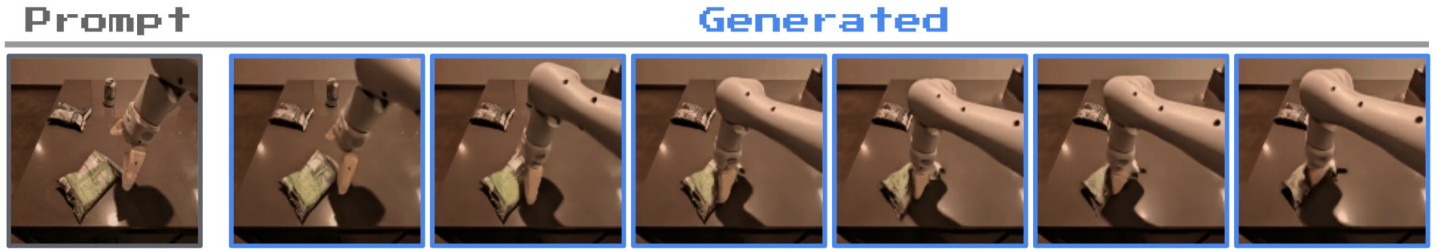
1) Scaling Results



2) Image Prompt Results

Results

3) Learning to simulate deformable objects



4) Controllable, consistent latent actions in robotics