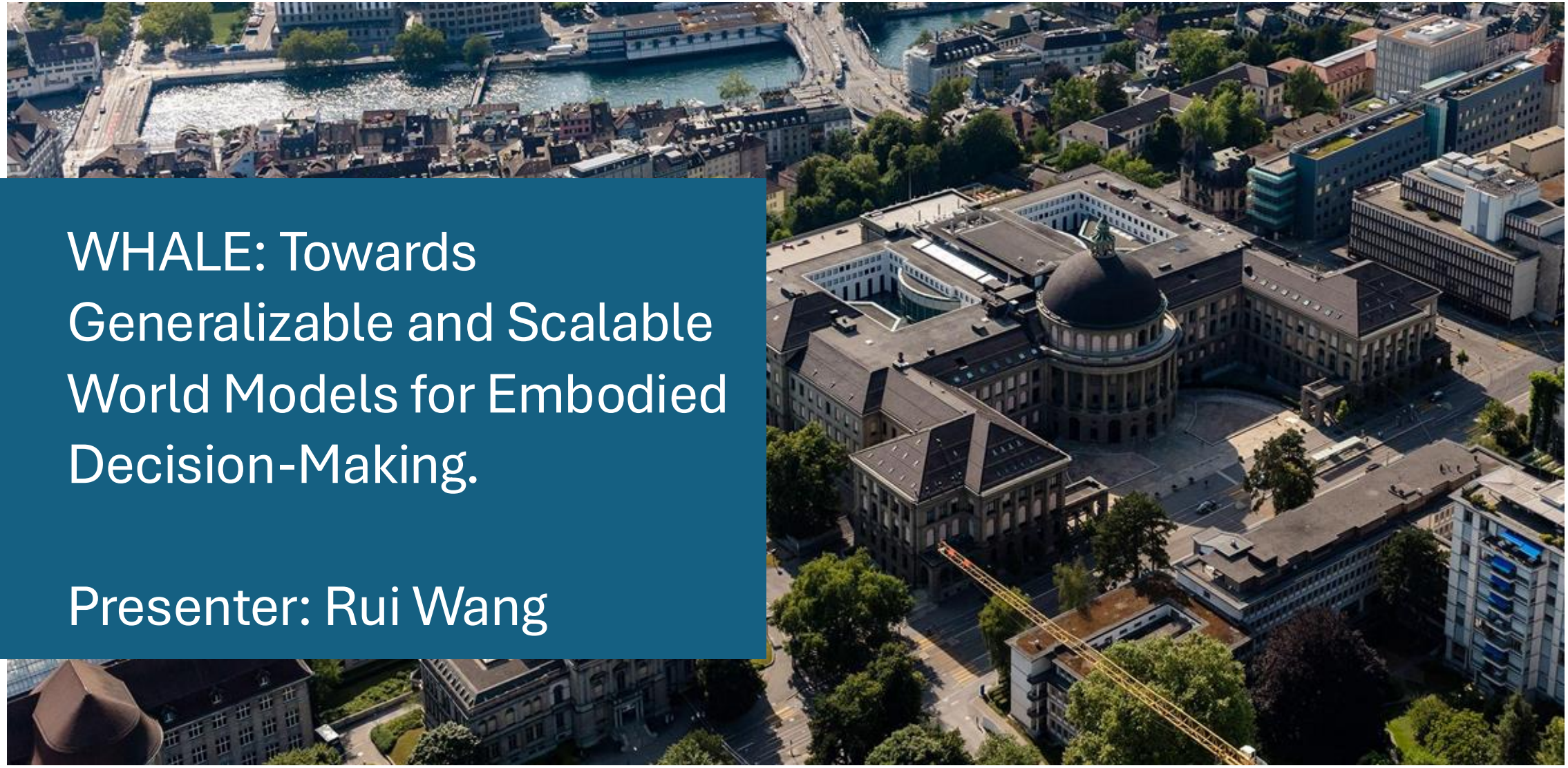
ETH zürich



WHALE: Towards Generalizable and Scalable World Models for Embodied Decision-Making.

Presenter: Rui Wang

What is a World Model?

An AI system that learns an **internal representation of the environment** to predict future states.

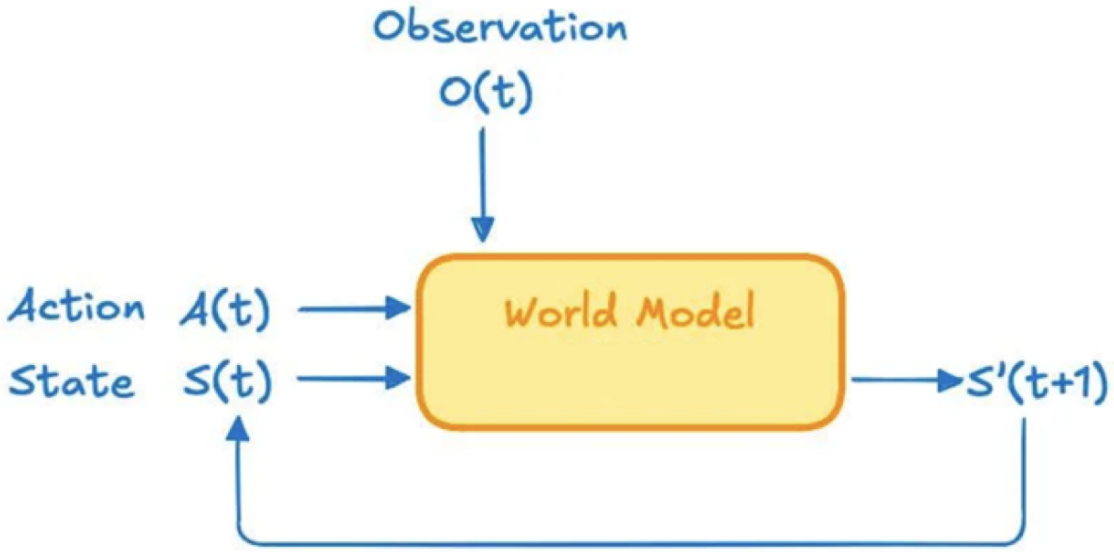
Given: state S(t) + action A(t)
Predict: next state S(t+1)

Captures physics, dynamics, and cause-and-effect — enabling cost-free exploration.

Source: https://github.com/knightnemo/Awesome-World-Models

Source: https://github.com/gxchris95/awesome-world-models-video-games-3d

What is world Model?



Abstract representation

Concrete representation

JEPA / Latent

Predict in latent space
JEPA, LeCun 2022

Simulators

Hand-crafted physics as Isaac Gym

3D/4D Neural Scene

Marble

Autoregressive Video

Token-level next-frame
Genie, WHALE

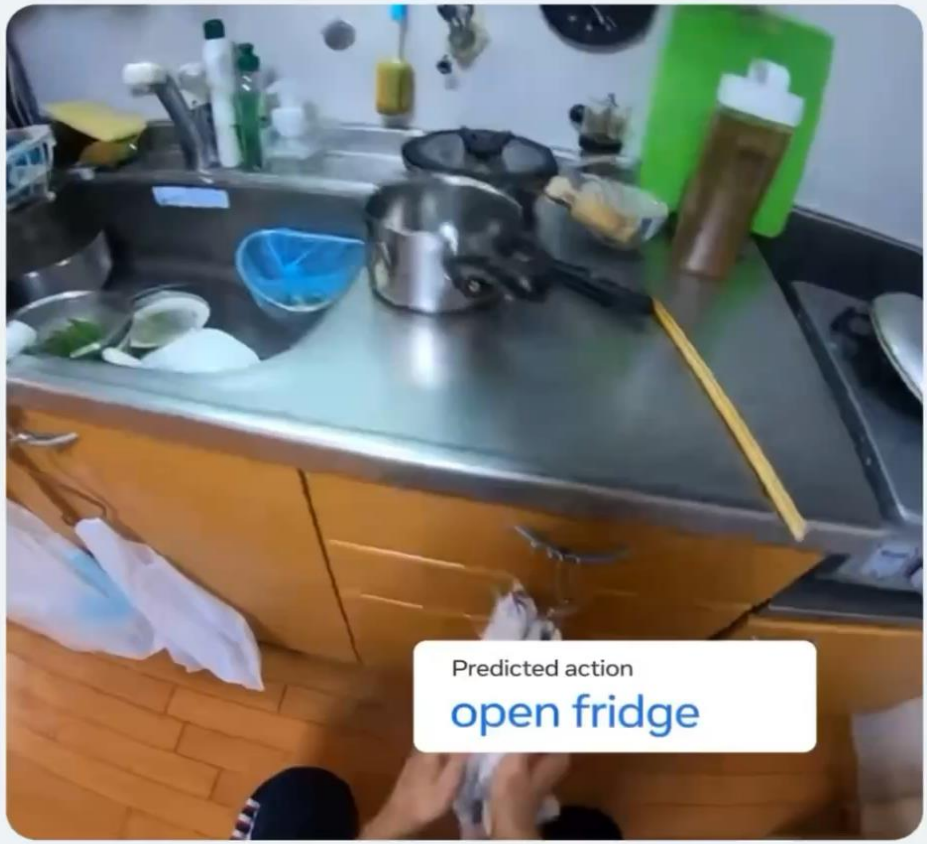
Given the current state S(t) and action A(t), predict the next state S'(t+1).
Usually take in observations/inputs O(t) from the external world, to update its internal state.

JEPA

Joint Embedding Predictive Architecture — Abstract World Understanding

Unlock world understanding

V-JEPA 2 delivers exceptional motion understanding as well as leading visual reasoning capabilities when combined with language modeling.

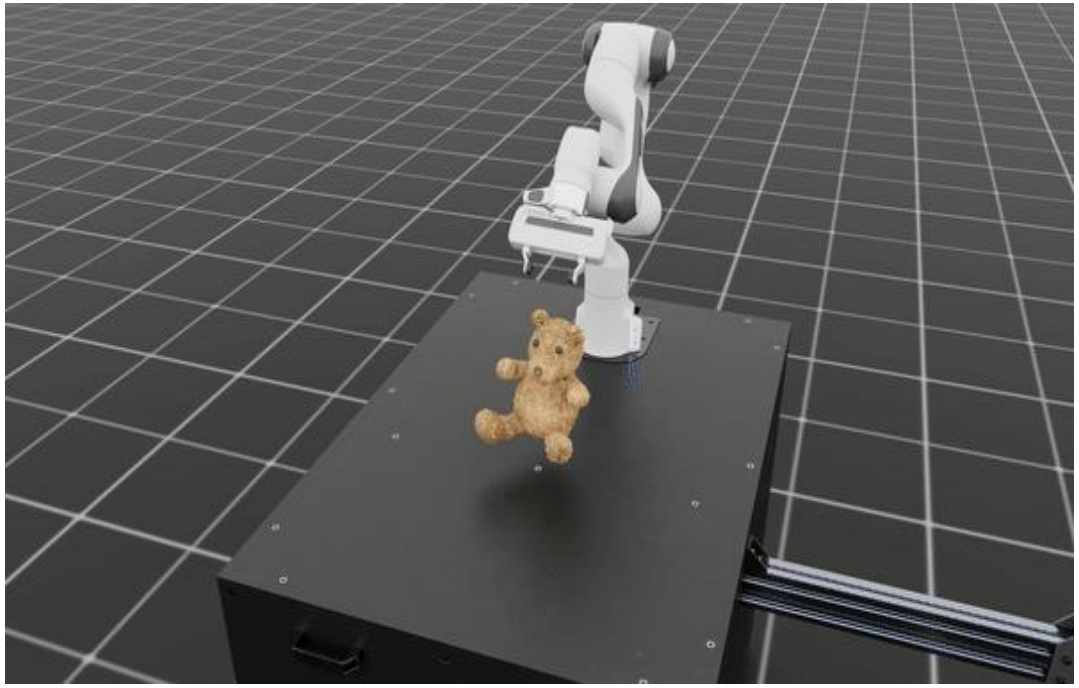


Anticipate what's next

V-JEPA 2 can make predictions about how the world will evolve, setting a new state-of-the-art in anticipating actions from contextual cues.

Physical Simulator

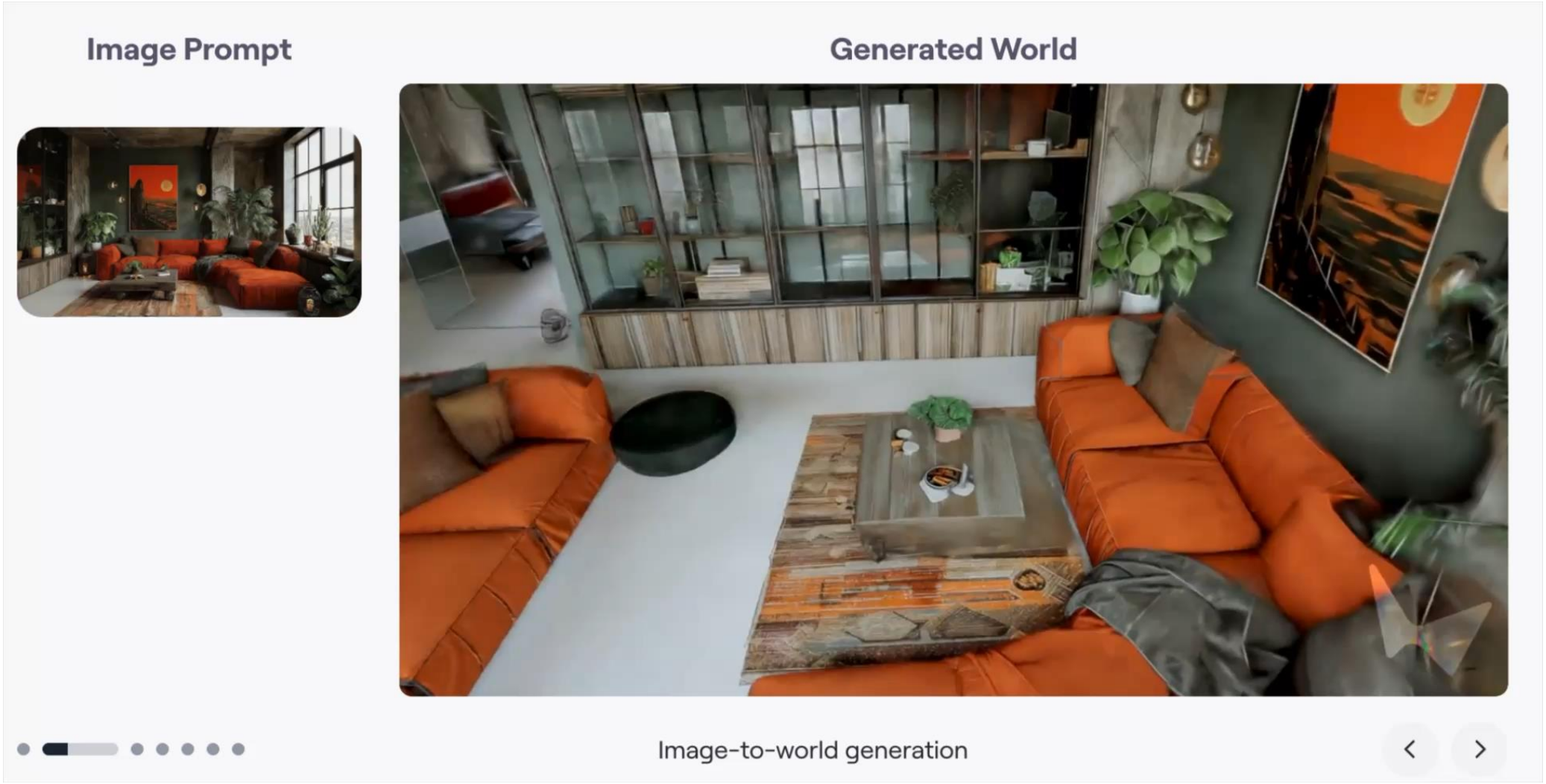
Hard-coded physics engines: powerful but brittle against the real world



Source: developer.nvidia.com/isaac/lab

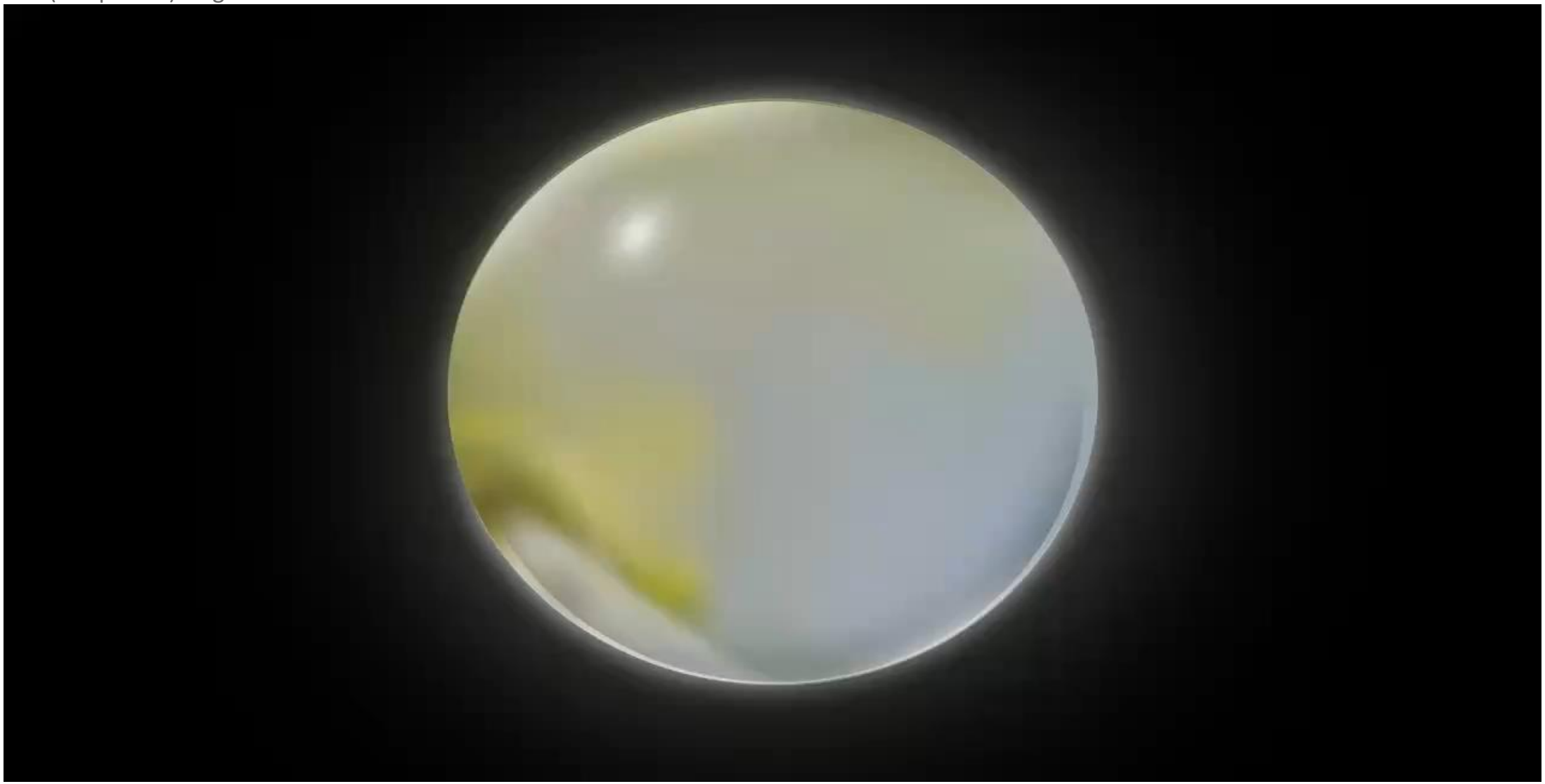
Marble

Neural 3d scene representation learned world models



Video World Model

Genie (DeepMind) — generative interactive environments from internet video



The Bottleneck in Robotics & The World Model Solution

Traditional simulators **can't simulate deformable objects, friction & dynamic lighting** — World Models offer **"learned imagination"**

Zero physical robot Cost

Practice millions of tasks inside neural "imagination" on GPUs — no physical motors needed

100% Safe Exploration

Make catastrophic mistakes in the neural sim — learn what not to do before real deployment

Efficient Data Scaling

Turn a small set of human demos into billions of diverse training scenarios via synthetic generation

Genie: Generative Interactive Environments (Backbone of Whale)

Bruce et al., Google DeepMind, 2024 — 11B Parameters

First generative interactive environment
trained unsupervised from Internet videos

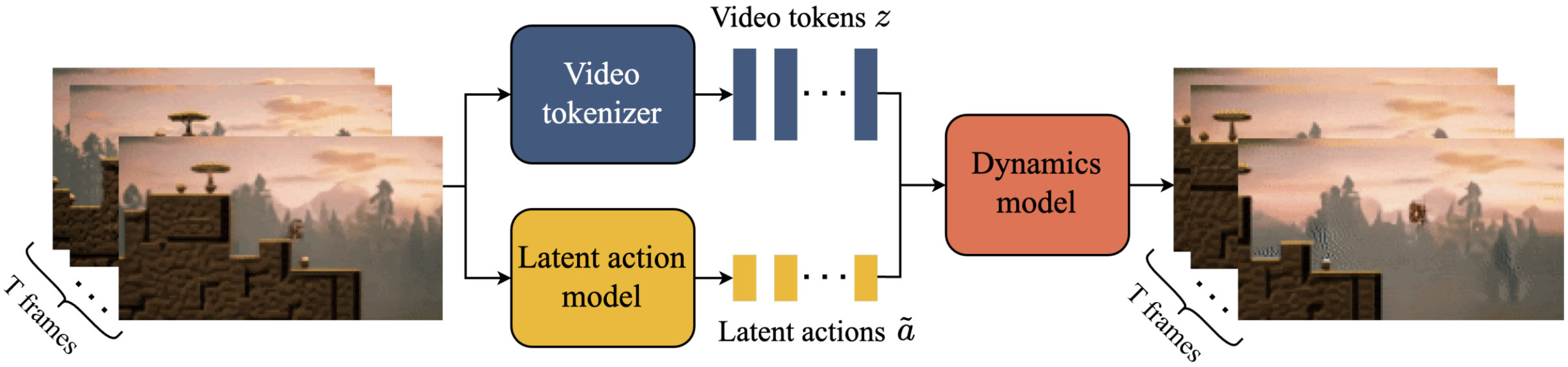
- Converts any image prompt into a playable environment
- Learns latent actions from video — no action labels needed
- Three components: Video Tokenizer + LAM + Dynamics Model

latent actions: 6, 6, 7, 6, 7, 6, 5, 5, 2, 7

Genie: Generative Interactive Environments

Works seamlessly on Human sketches

Genie: Model Architecture



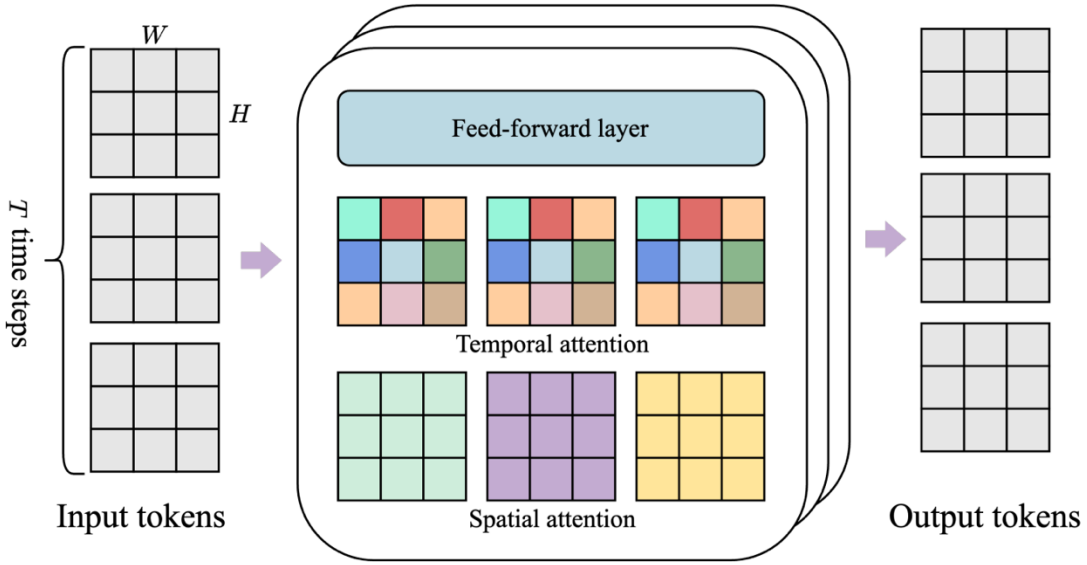
Video Tokenizer
VQ-VAE with ST-transformer
Compresses frames → discrete tokens
200M parameters

Latent Action Model
Infers actions between frame pairs
8 discrete latent actions via VQ
No action labels needed

Dynamics Model
MaskGIT transformer
Predicts next frame autoregressively
From tokens + latent actions

The ST-Transformer: Backbone of Genie

Challenge: Videos have ~10,000 tokens per frame — full attention is too expensive



Each ST-block has 3 layers:

1. Spatial attention
Attends over H×W tokens within one frame

2. Temporal attention
Same token position across T frames

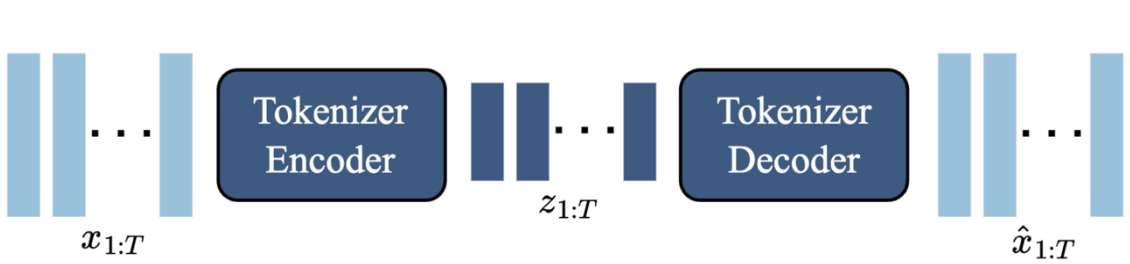
3. Feed-forward layer
(one FFW, not two — saves compute)

Why not full space-time attention?

C-ViViT uses full attention → much higher memory, tends to overfit, needs strong regularization
ST-transformer factorizes into spatial + temporal → linear memory, better FVD and controllability

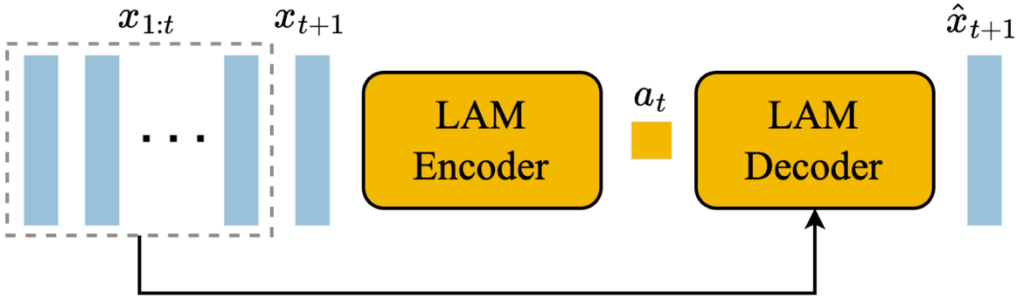
Video Tokenizer + Latent Action Model

Video Tokenizer (Phase 1)



- VQ-VAE with ST-transformer enc/dec
- Inputs T frames → discrete tokens per frame
- Spatiotemporal (not spatial-only) encoding
- Codebook: 1024 codes, dim 32
- 200M parameters, patch size 4

Latent Action Model (Phase 2)

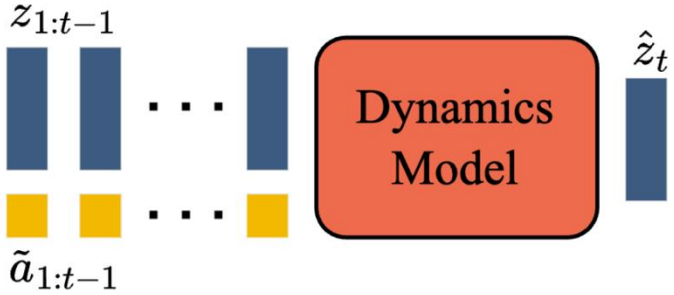


- Separate VQ-VAE on raw pixel frame pairs
- Encodes (x_t, x_{t+1}) → latent action a_t
- VQ bottleneck quantizes to 8 actions
- Only VQ codebook kept after training
- User picks action directly at inference

Trained separately first (tokenizer), then co-trained (LAM + dynamics) — LAM encoder is discarded after training

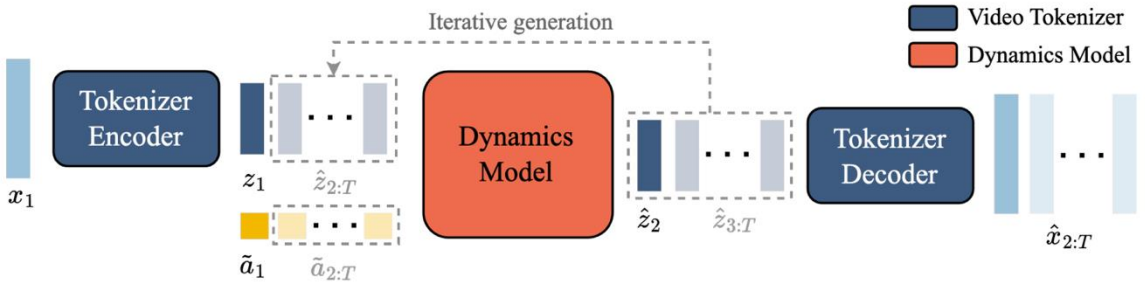
Dynamics Model + Inference

Dynamics Model (Training)



- Decoder-only MaskGIT transformer
- Input: video tokens z_{1:t-1} + latent actions
- Predicts next frame tokens z_t
- Random masking (50-100%) at train time
- Cross-entropy loss on predicted tokens

Interactive Inference

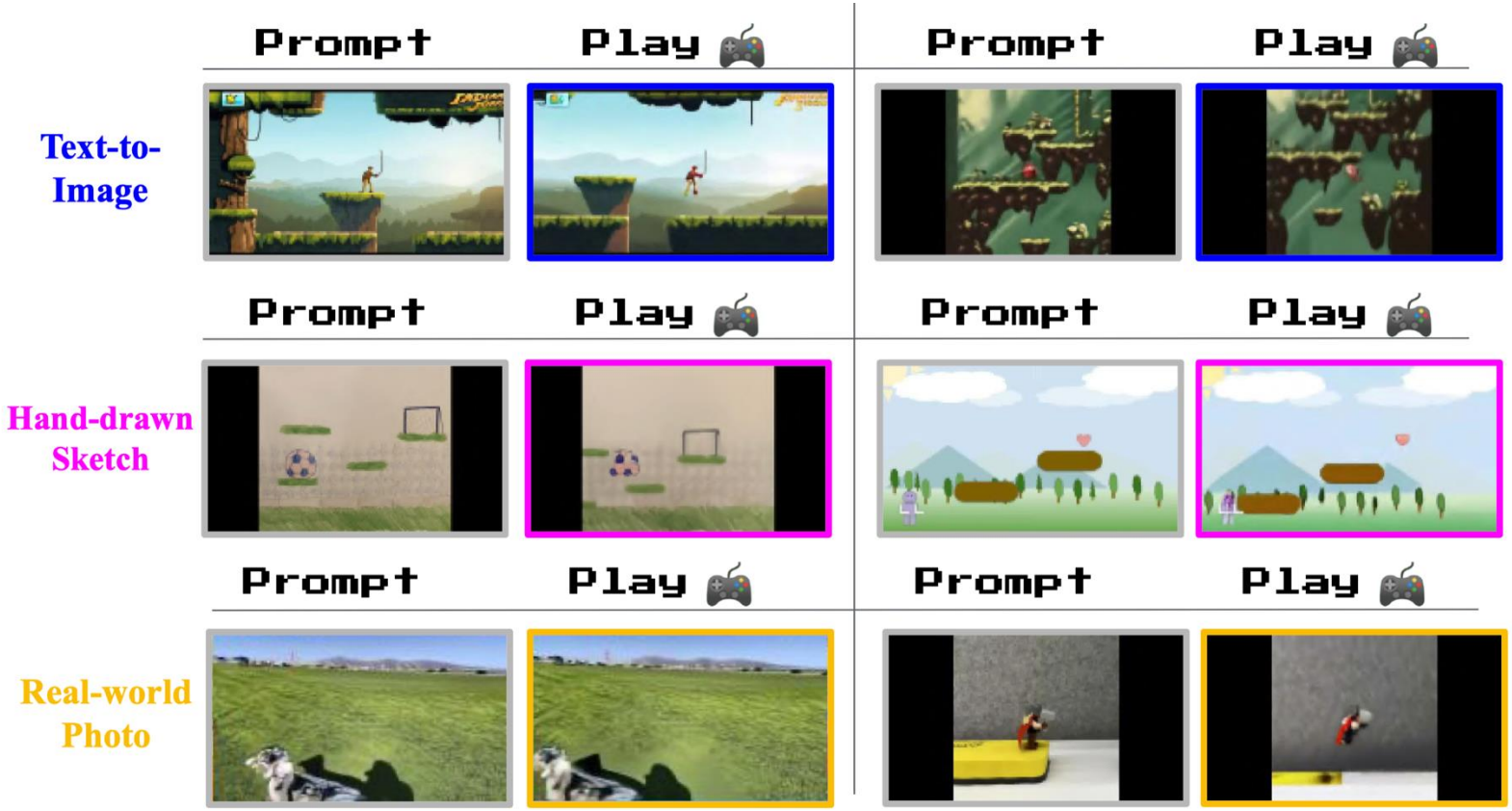


1. Tokenize the prompt image
2. User selects a latent action (1 of 8)
3. Dynamics model predicts next tokens
   (iterative parallel decoding via MaskGIT)
4. Decode tokens → pixels, loop from step 2

MaskGIT enables parallel token generation — much faster than pure autoregressive decoding for each frame

Genie: Key Insight — Unsupervised Action Learning

Internet videos have no action labels — but Genie learns to control environments by discovering latent actions from frame transitions



Training: LAM encoder sees frame pairs → infers discrete action between them

Inference: User directly picks 1 of 8 latent actions → dynamics model generates next frame

Result: Actions are semantically consistent — same action = same effect across different prompts

PART II

From Genie to WHALE: The Generalization Gap

Genie ✓

- Unsupervised action learning
- Scalable ST-transformer
- OOD prompts from T2I, sketches

Genie ✗

- No behavior-conditioning
- No uncertainty estimation
- Not tested on real robots

WHALE adds

- **Behavior-conditioning** for OOD generalization
- **Retracing-rollout** for uncertainty
- Real-robot validation (Meta-World + ARX5)
- Whale-X: 414M params, 970K real demos
- Offline MBRL integration

WHALE: universal plug-and-play techniques for any world model architecture

Open Challenges for World Models

Value Gap Bound (WHALE, Theorem 1):

$$\left|V_T^\pi - V_{T^*}^\pi\right| \leq 2R_{\max} \underbrace{H^2}_{\text{teacher-forcing}} \left(\underbrace{\sqrt{2\, l_{\mathrm{KL}}(T;\Pi)}}_{\text{in distribution error}} + \underbrace{L \cdot W_1(d^\pi, d^\Pi)}_{\text{policy divergence}} \right),$$

1. Generalization

- Policy distribution shift
- Compounding errors ~ H^2
- train a world model on a robot moving slowly, and then ask it to simulate a robot moving fast

2. Uncertainty

- Ensembles too expensive at scale
- When to trust model predictions?
- A mathematical way to detect hallucination

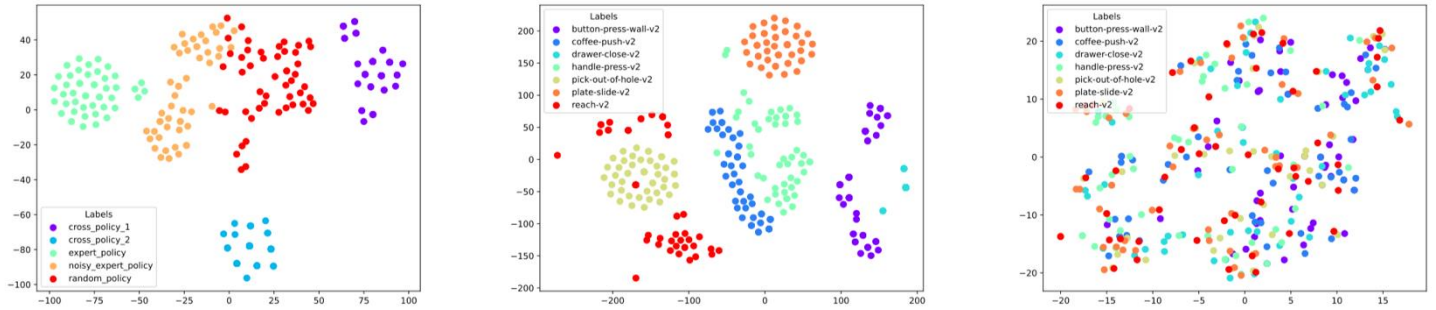
Genie addresses action-free learning | WHALE addresses both terms

Behavior-Conditioning: Formulation

ELBO-based behavior embedding (Eq. 4):

$$\log P(\tau_H|\tau_h) \geq \mathbb{E}_{q_\phi(z|\tau_H)} \sum_{t=h}^{H} \log \pi_w(a_t|o_t, \tau_{t-1}, z) - D_{\mathrm{KL}}(q_\phi(z|\tau_H)||p_\psi(z|\tau_h)) + Const,$$

- **Posterior $q(z|\tau_H)$:** encodes full trajectory into behavior z
- **Prior $p(z|\tau_h)$:** predicts z from partial history (used at inference)
- β-VAE (continuous), KL constraint creates information bottleneck on z
- z captures **policy intent**, not task identity (confirmed via t-SNE)



(a) Same task different policies (b) Different tasks expert policies (c) Different tasks random policies

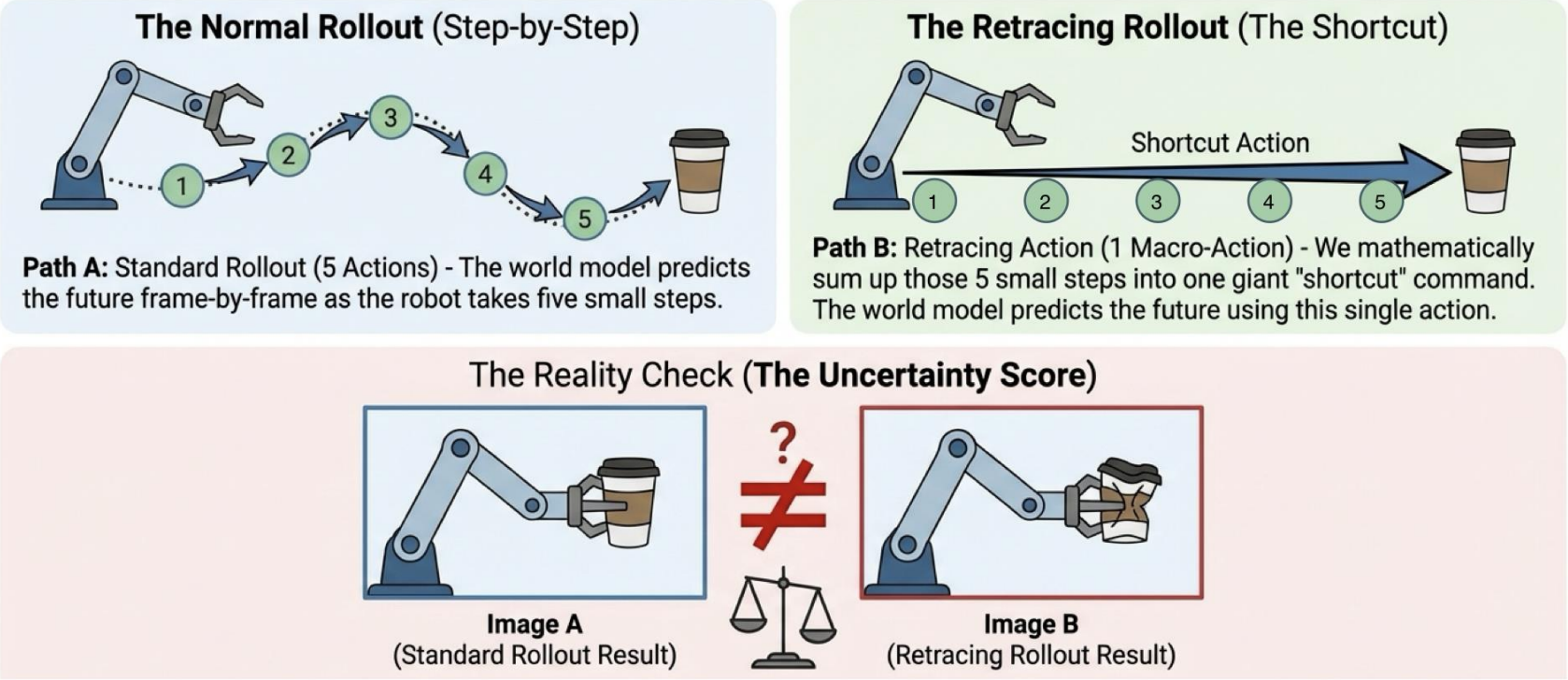
Figure 15: The behavior embedding visualization via t-SNE [99]. The different colors denote different policies in the same task (15a) and expert policies in different tasks (15b) or random policies in different tasks (15c).

Retracing-Rollout: Uncertainty Without Ensembles

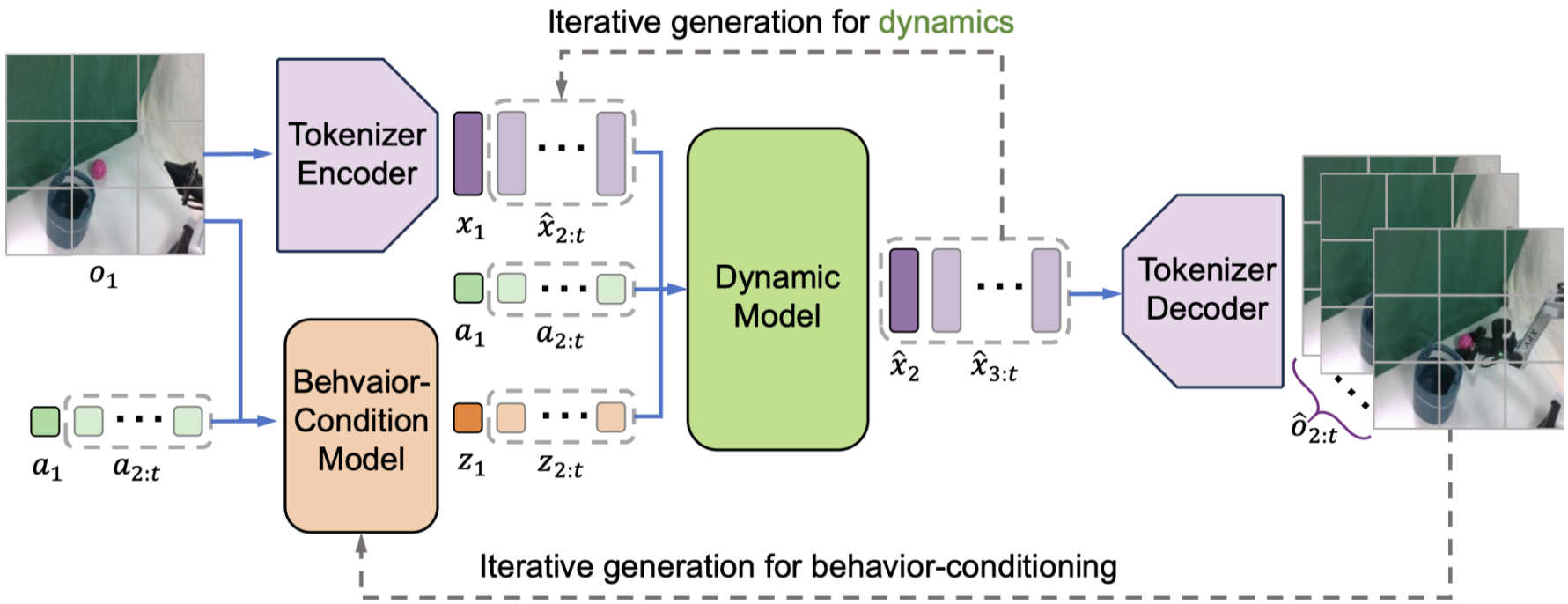
Retracing-action (Eq. 5) — collapse multi-step actions:

$$a_k^r = \left(\sum_{i=1}^{k} a_i^{(0)}, \cdots, \sum_{i=1}^{k} a_i^{(5)}, a_k^{(6)} \right).$$

- Backtrack k steps, execute collapsed action
- Uncertainty = **feature disagreement × predictive entropy**
- No ensemble needed — single model, **no training changes**
- +500% over ensemble on error correlation



Whale-ST: Full Architecture



Video Tokenizer

VQ-VAE encoder/decoder

L_1 + codebook + commit loss

Behavior-Cond.

CNN + ST-Transformer

Posterior + Prior + Decoder

Dynamics Model

ST-Transformer

tokens + actions + z_h → x̂_h+1

Training Pipeline

Phase 1: Train tokenizer (VQ-VAE loss) + behavior model (ELBO)

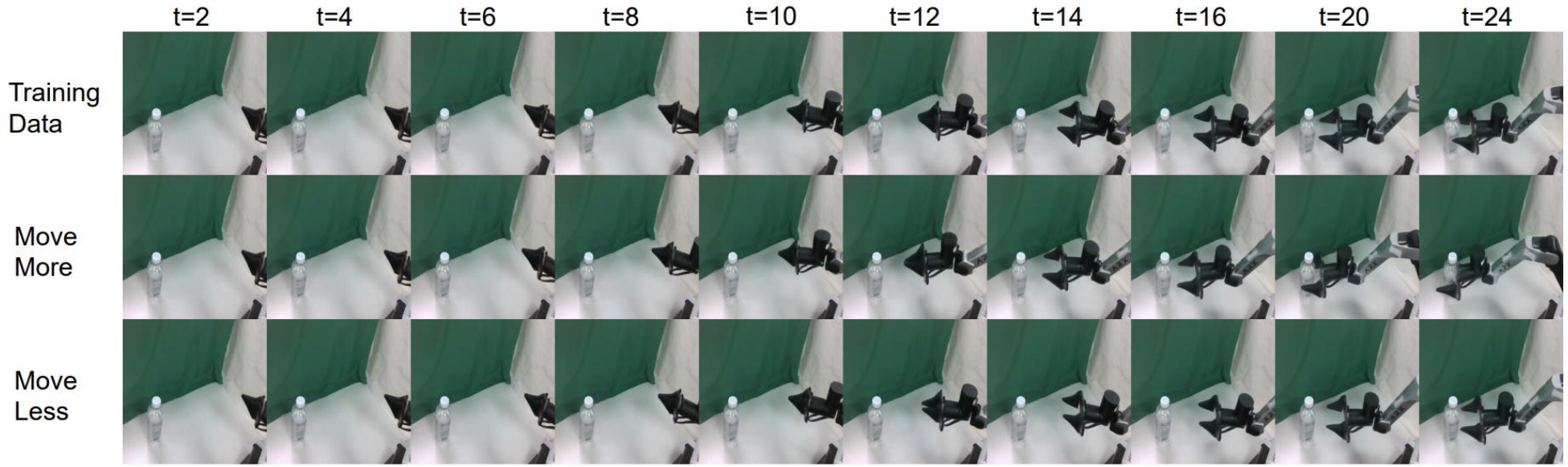
Phase 2: Freeze tokenizer + BC model, train dynamics with z_h conditioning

Inference: prior p_ψ infers z from history → adapts generation on-the-fly

Whale-X: Scaling to Real Robots

414M params | 970K trajectories from OpenX-Embodiment | 27 datasets

- Fine-tune with **only 60 demos** per task on ARX5 robot
- +63% consistency vs. no behavior-conditioning
- 3× consistency gain scaling 77M → 203M dynamics



Results: Whale-ST on Meta-World

Value Prediction (256×256)

Value Gap: **9.7** (DreamerV3: 11.7)
Return Corr: **0.86** (DreamerV3: 0.72)
Regret@5: **3.9** (DreamerV3: 11.5)

Video Fidelity (256×256)

FVD↓: **28.2** (DreamerV3: 112.4)
PSNR↑: **29.2** (DreamerV3: 26.2)
SSIM↑: **95.0** (DreamerV3: 91.7)

Meta-World	#Params	Value Gap↓	Return Correlation↑	Regret@5↓
64×64 resolution				
FitVid	143M	18.2	0.64	22.0
MCVD	53M	20.6	0.72	12.2
DreamerV3	44M	**10.0**	0.70	16.5
iVideoGPT	63M	15.9	0.62	**7.2**
Whale-ST (ours)	51M	**10.3 ± 0.8**	**0.77±0.01**	**7.3±1.2**
256×256 resolution				
DreamerV3	61M	11.7	0.72	11.5
Whale-ST (w/o bc)	63M	11.8±0.4	0.84±0.01	5.0±0.3
Whale-ST (ours)	63M	**9.7±0.1**	**0.86±0.01**	**3.9±0.3**

Meta-World	#Params	FVD↓	PSNR↑	SSIM↑	LPIPS↓
64×64 resolution					
FitVid	143M	154.6	23.7	90.3	6.5
MCVD	53M	272.8	**29.7**	92.3	4.0
DreamerV3	44M	142.7	27.6	92.1	4.3
iVideoGPT	63M	115.7	28.5	92.8	4.5
Whale-ST (ours)	51M	**38.5±2.6**	28.8±0.0	**93.5±0.1**	**3.7±0.1**
256×256 resolution					
DreamerV3	61M	112.4	26.2	91.7	8.5
Whale-ST (w/o bc)	63M	32.0±0.4	28.9±0.2	94.6±0.1	4.6±0.1
Whale-ST (ours)	63M	**28.2±3.6**	**29.2±0.2**	**95.0±0.1**	**4.3±0.1**

Outlook

Follow-up work

→ **Behavior-Conditioning + VLM Reasoning**

Later approaches adopt behavior-conditioning with semantic upgrades (+VLM reasoning)

→ **Diffusion / Flow Matching**

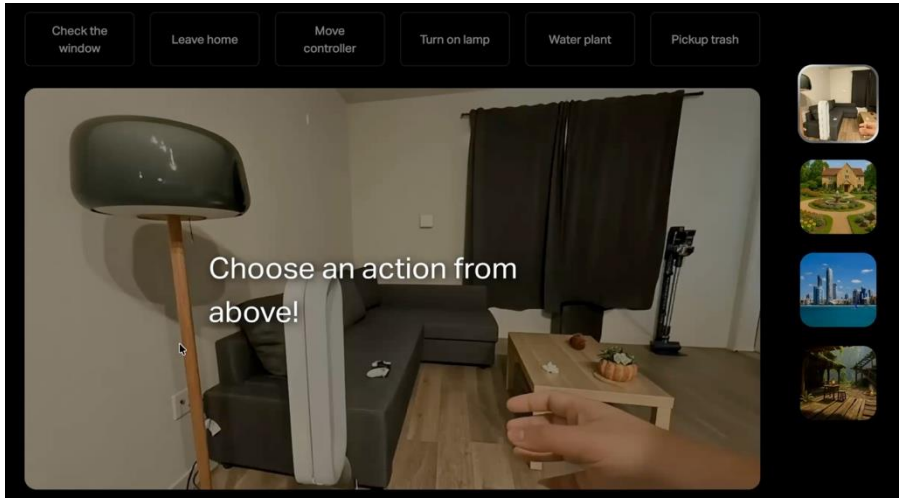
Inherent uncertainty modeling capability via diffusion / flow matching architectures

Outlook

World model via Video Generation

Diamond

Matrix Game 2.0



Real-time 30 FPS Interactive Generation

Offline → real-time interactive world generation

Long-context and multimodal inference