

iVideoGPT: Interactive VideoGPTs are Scalable World Models

<https://thuml.github.io/iVideoGPT>

**Jialong Wu^{1,*}, Shaofeng Yin^{1,2,*}, Ningya Feng¹, Xu He³, Dong Li³, Jianye Hao^{3,4},
Mingsheng Long¹✉**

¹School of Software, BNRist, Tsinghua University, ²Zhili College, Tsinghua University

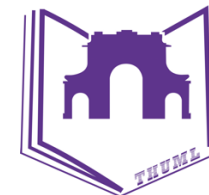
³Huawei Noah's Ark Lab, ⁴College of Intelligence and Computing, Tianjin University
wujialong0229@gmail.com, ysf22@mails.tsinghua.edu.cn, mingsheng@tsinghua.edu.cn



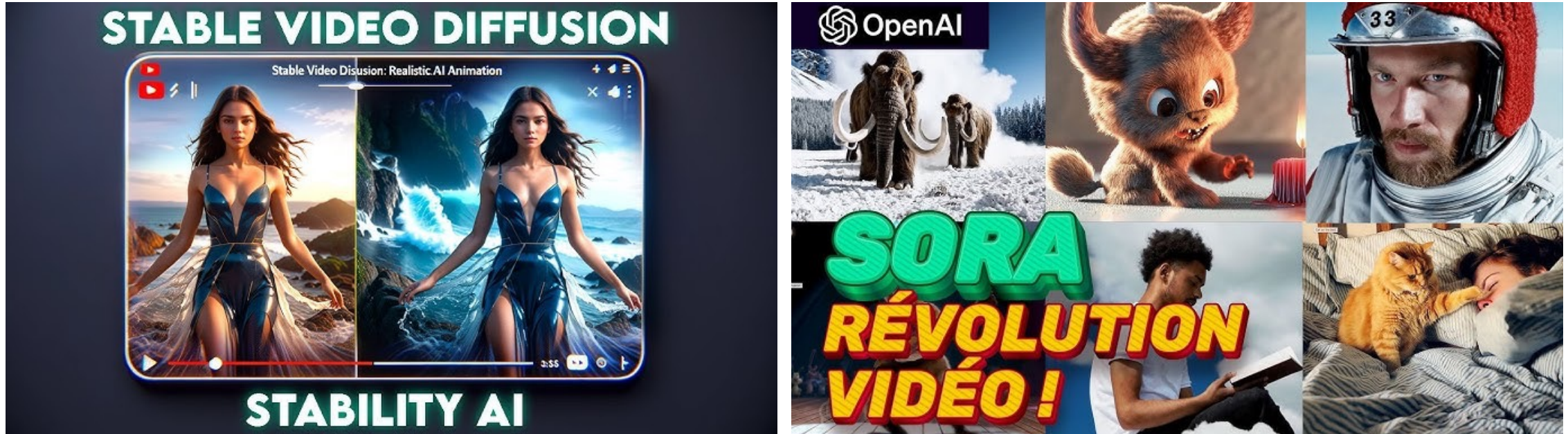
清华大学
Tsinghua University



HUAWEI



Motivation: Video Generation vs. World Models

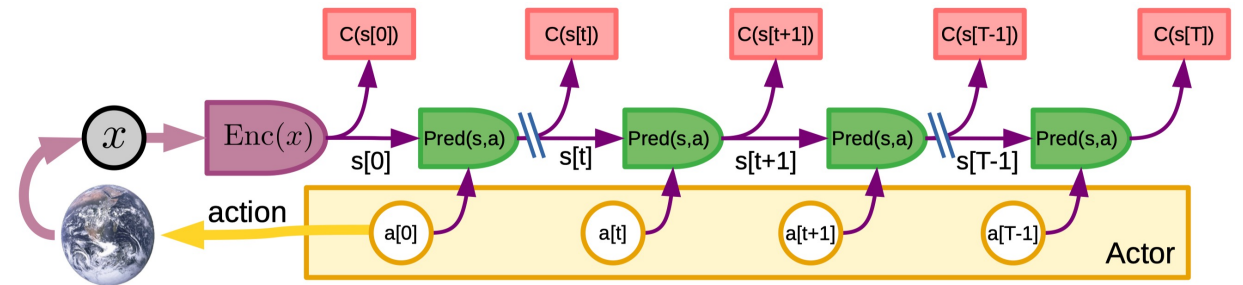
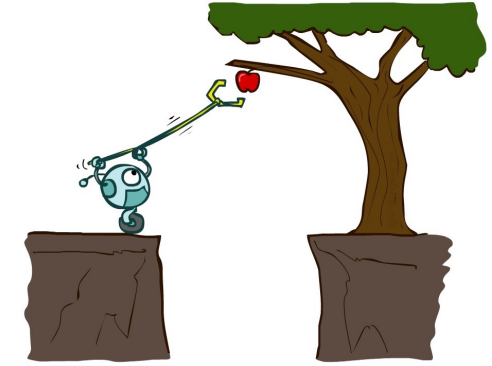
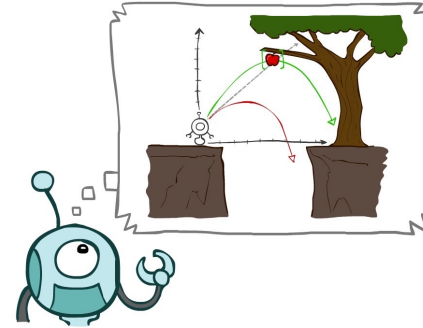
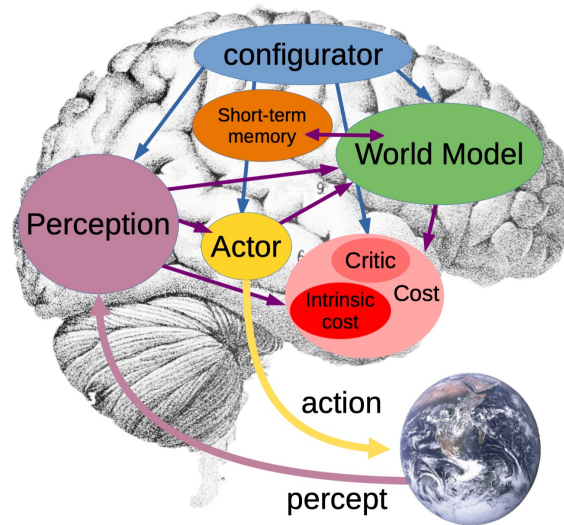


Are Video Generation Models World Worlds?

Not Yet! *(explained later)*

Our work: How can we leverage the advancements in **scalable video generative models** for developing **interactive visual world models**?

World Models: From System-1 to System-2



World Models:

internal models of how the world works

Model-based Agents:

Act through an optimization procedure (**planning**) running the **world model**.

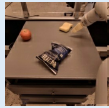
Yann LeCun. A path towards autonomous machine intelligence. 2022.

Dan Klein and Pieter Abbeel. Introduction to Artificial Intelligence.

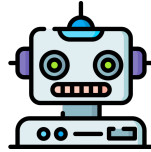
Task: World Models as Interactive Video Prediction



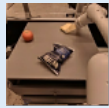
$o_t =$



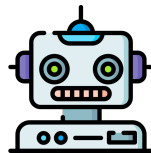
$a_t = (\Delta X, \Delta R)$



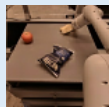
$o_{t+1} =$



$a_{t+1} = (\Delta X, \Delta R)$



$o_{t+2} =$



⋮

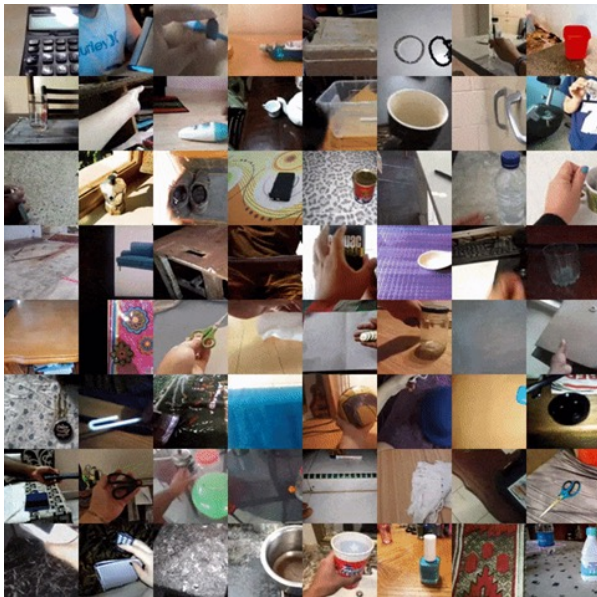
A process of making decisions and imagine outcomes:

$$\begin{aligned}
 & p(o_{T_0+1:T}, a_{T_0:T-1} \mid o_{1:T_0}) \\
 &= \underbrace{p(a_{T_0:T-1} \mid o_{1:t})}_{\text{Agent}} \underbrace{p(o_{T_0+1:T} \mid o_{1:T_0}, a_{T_0:T-1})}_{\text{World model}} \quad \text{Non- (Low-) interactive} \\
 &= \prod_{t=T_0}^{T-1} \underbrace{p(a_t \mid o_{1:t})}_{\text{Agent}} \underbrace{p(o_{t+1} \mid o_{1:t}, a_{T_0:t})}_{\text{World model}} \quad \text{Interactive}
 \end{aligned}$$

A problem with fundamental connection to **video prediction/generation models**, referred to as **interactive video prediction**

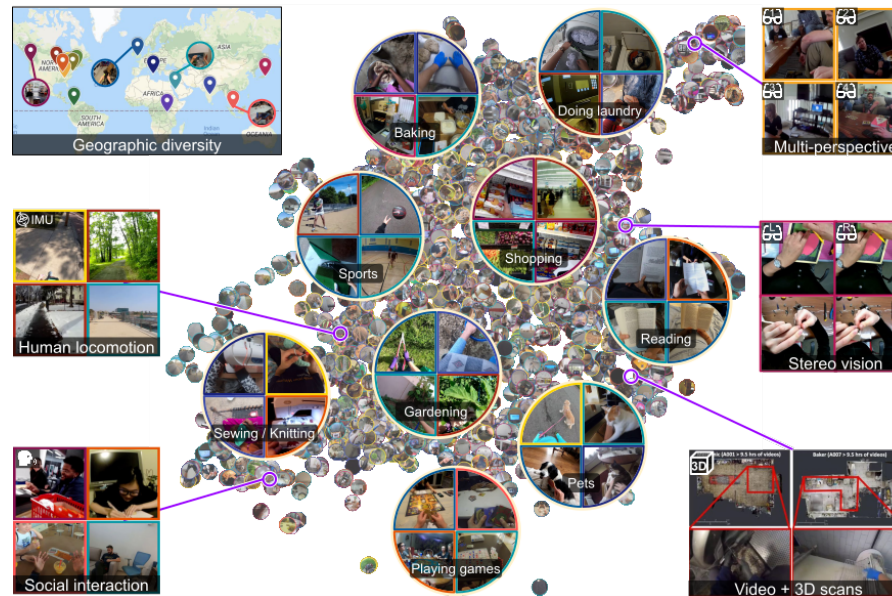
Data: Towards a **General** World Model

General world knowledge for a variety of downstream tasks
from **abundant in-the-wild videos** on the Internet



Something-Something V2

Goyal et al. ICCV 2017



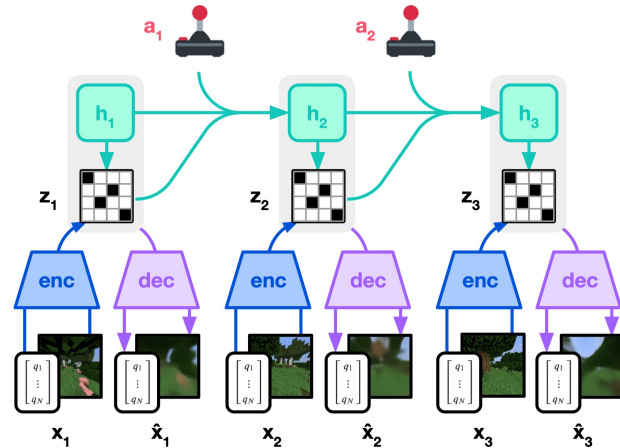
Ego4D

Grauman et al., Facebook AI. CVPR 2022

- ✓ Task-agnostic
- ✓ Widely available
- ✓ Broad Knowledge

Model: Recurrent World Models Have Limited Scalability

DreamerV3: Naturally allows step-by-step transitions but with limited capability

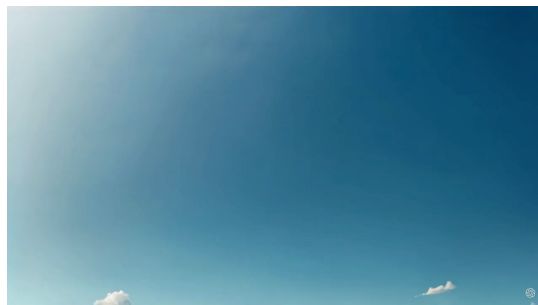


A case study on Minecraft

Ground truth
Prediction (DreamerV3-L)



Sora: Internet-scale video generative models can synthesize realistic long videos



High-fidelity
Minecraft
simulation:



Model: Video Generative Models Have Limited Interactivity

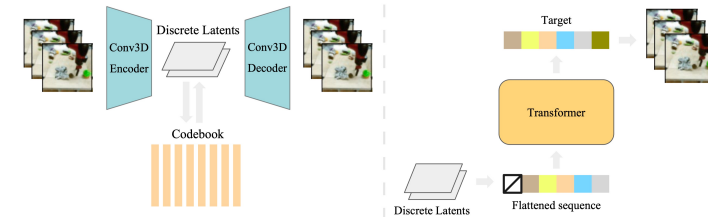
Typically design **non-causal temporal modules**



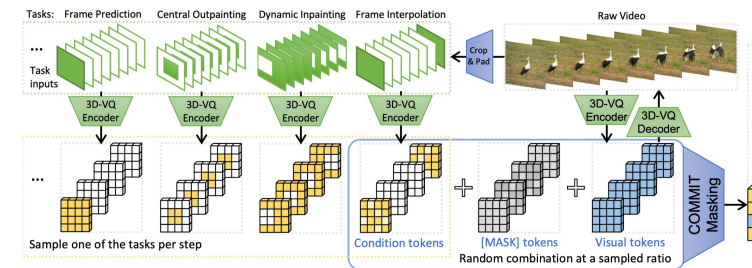
Provide only **trajectory-level interactivity**

- Allow text/action conditions **only at the beginning** of the video
- Lacking the ability for **intervention during simulations**
- Typically produce videos of **a fixed length**

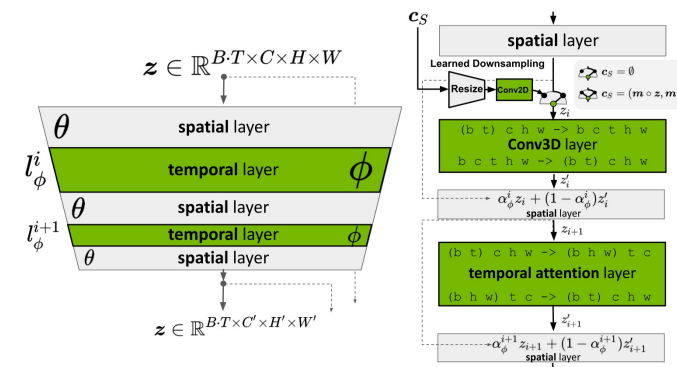
Our work: achieve **step-level interactivity**



Autoregressive model: VideoGPT



Masked model: MAGVIT

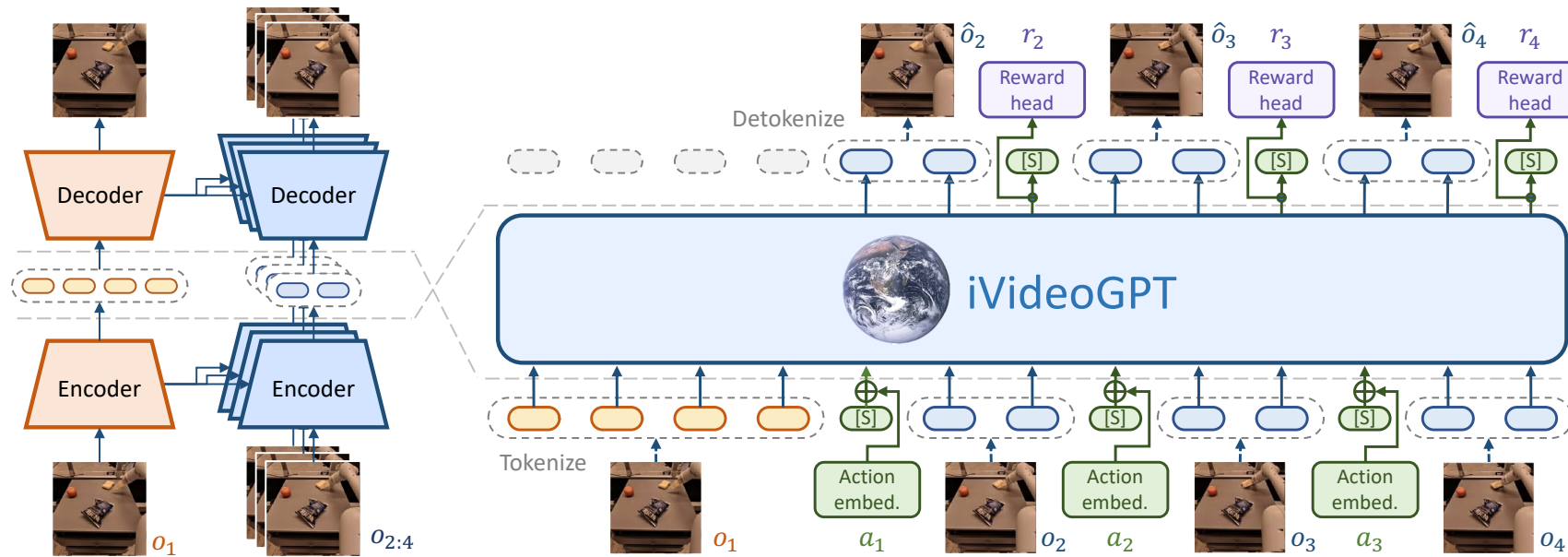


Diffusion model: Stable Video Diffusion

iVideoGPT: Interactive VideoGPT

Overview:

iVideoGPT integrates multimodal signals—visual observations (via **compressive tokenization**), actions, and rewards—into a sequence of tokens, and providing interactive experience via next-token prediction of an **autoregressive transformer**.

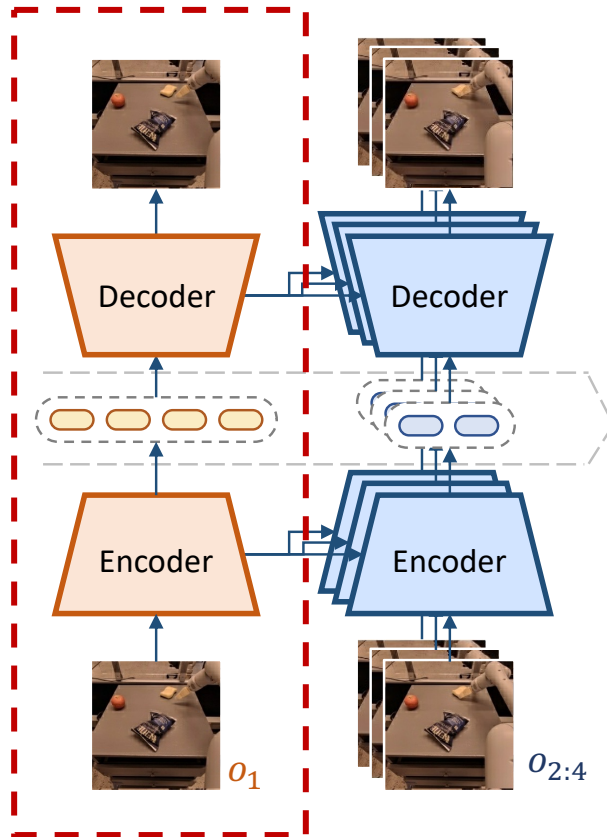


Compressive tokenization

Interactive prediction with Transformers

- ✓ Scalability
- ✓ Interactivity

Compressive Tokenization



($T_0 = 1$ for simplicity)

Transformers particularly shine when operating over sequences of discrete tokens



Commonly used visual tokenizer:

VQGAN

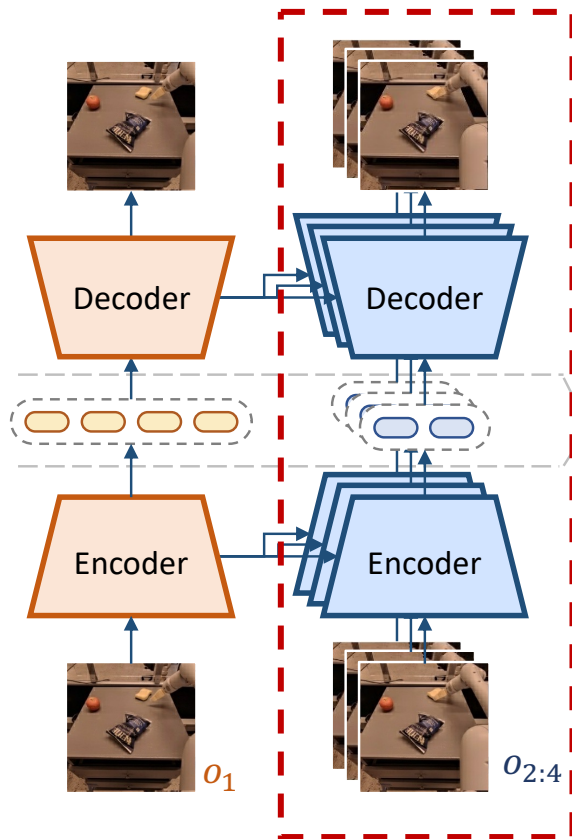
Context frames independently tokenized:

- Rich in contextual information
- Discretized into N tokens each frame:

$$z_t^{(1:N)} = E_c(o_t), \hat{o}_t = D_c(z_t) \text{ for } t = 1, \dots, T_0$$

- To tokenize future frames as well? **Low efficiency!**

Compressive Tokenization



($T_0 = 1$ for simplicity)

Future frames conditionally tokenized:

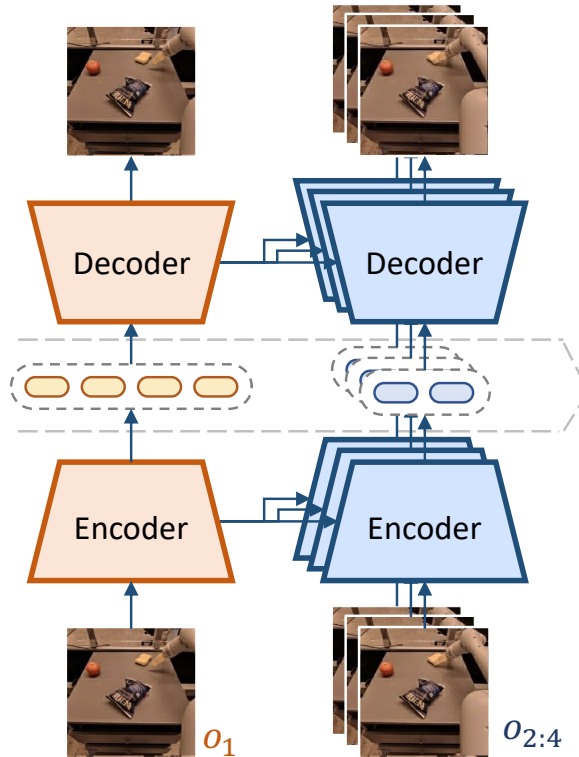
- Temporal redundancy between context and future frames
- Discretized into $n \ll N$ tokens each frame through **conditional VQGAN**:

$$z_t^{(1:n)} = E_p(o_t \mid \underbrace{o_{1:T_0}}_{\text{conditional encoder}}), \hat{o}_t = D_p(z_t \mid \underbrace{o_{1:T_0}}_{\text{conditional decoder}}) \quad \text{for } t = T_0 + 1, \dots, T$$

- Conditioning mechanism using **cross-attention between multi-scale feature maps** (the same as in **ContextWM**)

Wu, Jialong, et al. Pre-training Contextualized World Models with In-the-wild Videos for Reinforcement Learning. NeurIPS 2023.

Compressive Tokenization



($T_0 = 1$ for simplicity)

Overall objective:

$$\mathcal{L}_{\text{tokenizer}} = \sum_{t=1}^{T_0} \underbrace{\mathcal{L}_{\text{VQGAN}}(o_t; E_c(\cdot), D_c(\cdot))}_{\text{context frames}} + \sum_{t=T_0+1}^T \underbrace{\mathcal{L}_{\text{VQGAN}}(o_t; E_p(\cdot | o_{1:T_0}), D_p(\cdot | o_{1:T_0}))}_{\text{future frames}}$$

Benefits:

- ✓ Shorter token sequence, **faster rollouts** for model-based planning and reinforcement learning
- ✓ Maintain **temporal consistency** of the context much easier and focus on **modeling essential dynamics** information

Interactive Prediction with Transformers

A sequence of tokens:

$$x = \left(\underbrace{z_1^{(1)}, \dots, z_1^{(N)}}_{\text{context frame}}, [\text{S}], \underbrace{z_2^{(1)}, \dots, z_2^{(N)}}_{\text{slot token}}, \dots, \underbrace{[\text{S}], z_{T_0+1}^{(1)}, \dots, z_{T_0+1}^{(n)}}_{\text{future frame}} \right)$$

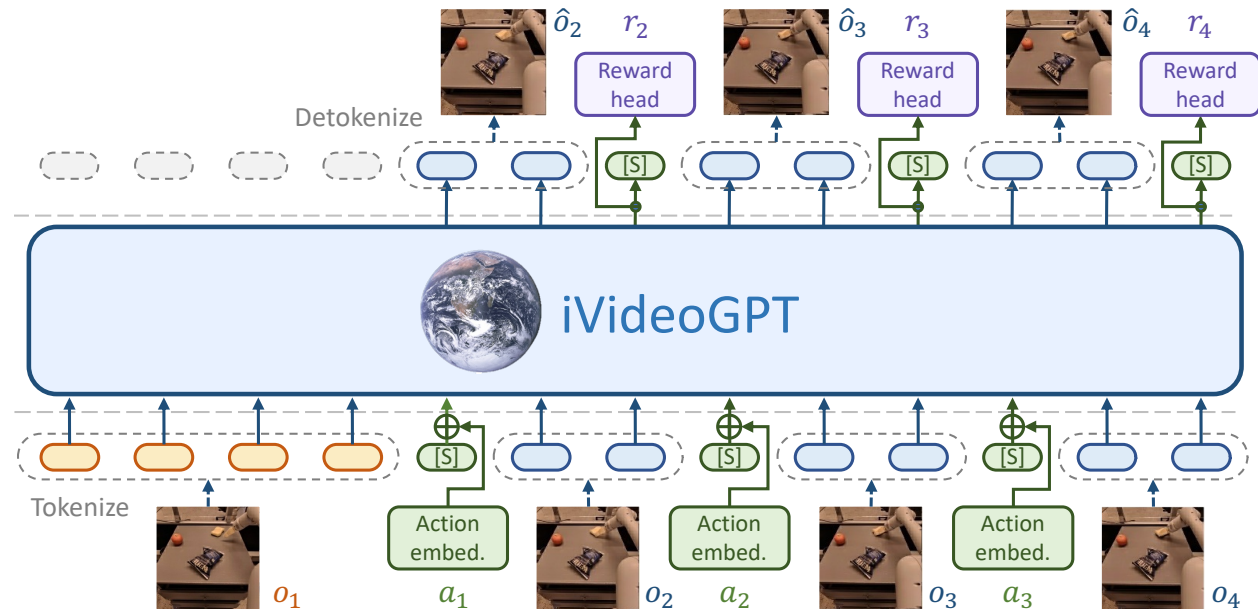
Delineate **frame boundaries** and facilitate optional **action and reward integration**

Total length $L = (N + 1)T_0 + (n + 1)(T - T_0) - 1$ grows linearly with frame numbers but at a much smaller rate ($n \ll N$)

GPT-2 size,

LLaMA architecture:

Embrace the latest innovations for LLM architecture



Pre-Training and Fine-Tuning



Action-free video prediction:

Not trained to generate context frames, focusing on dynamics information

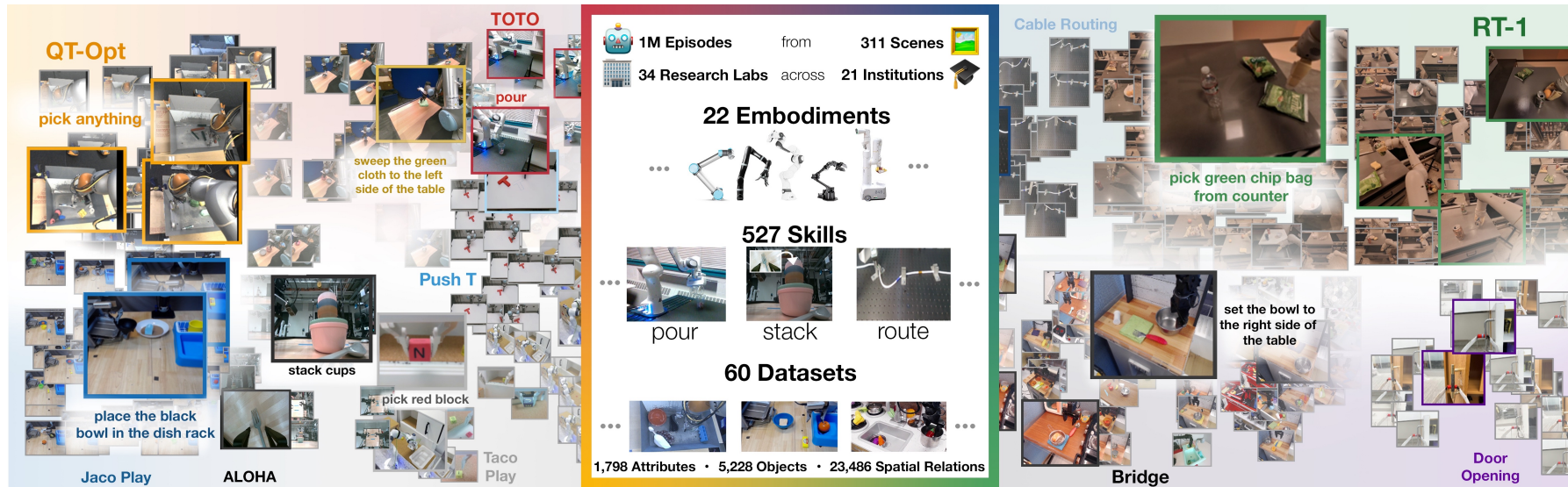
$$\mathcal{L}_{\text{pre-train}} = - \sum_{i=(N+1)T_0+1}^L \log p(x_i | x_{<i})$$

↑
First token index of predicted frames

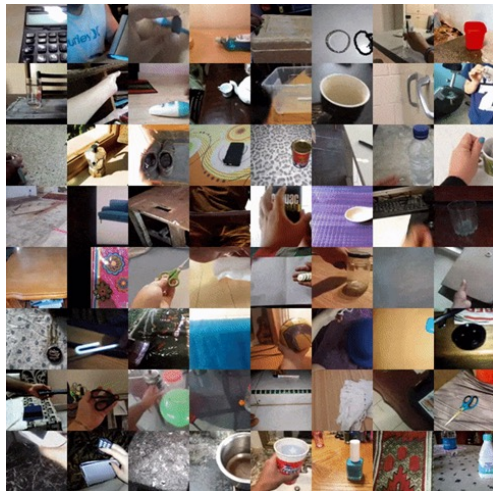
Flexibly incorporate extra modalities:

- **Action conditioning:** linear projection and adding to the slot token embeddings
- **Reward prediction:** linear head to the last token's hidden state of each observation; mean-squared error (MSE) loss

Pre-Training Data



Open X-
Embodiment
Padalkar et al. 2023



Something-
Something V2

Goyal et al. ICCV 2017

Total 1.4 million trajectories:

- Select 35 datasets from OXE, in addition to SSv2, by **excluding** mobile robots, excessive repetition, and low image resolutions
- **Filter out** overlaps with downstream test data
- **Sampling weights** based on sizes and diversity
- Varied **frame step sizes**, based on control frequency

Video Prediction

Per-frame tokenization suffers from temporal inconsistency and flicker artifacts

BAIR [20]	FVD↓	PSNR↑	SSIM↑	LPIPS↓	RoboNet [15]	FVD↓	PSNR↑	SSIM↑	LPIPS↓
<i>action-free & 64×64 resolution</i>					<i>action-conditioned & 64×64 resolution</i>				
VideoGPT [97]	103.3	-	-	-	MaskViT [26]	133.5	23.2	80.5	4.2
MaskViT [26]	93.7	-	-	-	SVG [87]	123.2	23.9	87.8	6.0
FitVid [3]	93.6	-	-	-	GHVAE [94]	95.2	24.7	89.1	3.6
MCVD [89]	89.5	16.9	78.0	-	FitVid [3]	62.5	28.2	89.3	2.4
MAGVIT [100]	62.0	<u>19.3</u>	<u>78.7</u>	<u>12.3</u>	iVideoGPT (ours)	<u>63.2±0.01</u>	<u>27.8±0.01</u>	90.6±0.02	4.9±0.00
iVideoGPT (ours)	<u>75.0±0.20</u>	20.4±0.01	82.3±0.05	9.5±0.01	<i>action-conditioned & 256×256 resolution</i>				
<i>action-conditioned & 64×64 resolution</i>					MaskViT [26]	211.7	20.4	67.1	17.0
MaskViT [26]	70.5	-	-	-	iVideoGPT (ours)	197.9±0.66	23.8±0.00	80.8±0.01	14.7±0.01
iVideoGPT (ours)	60.8±0.08	24.5±0.01	90.2±0.03	5.0±0.01					

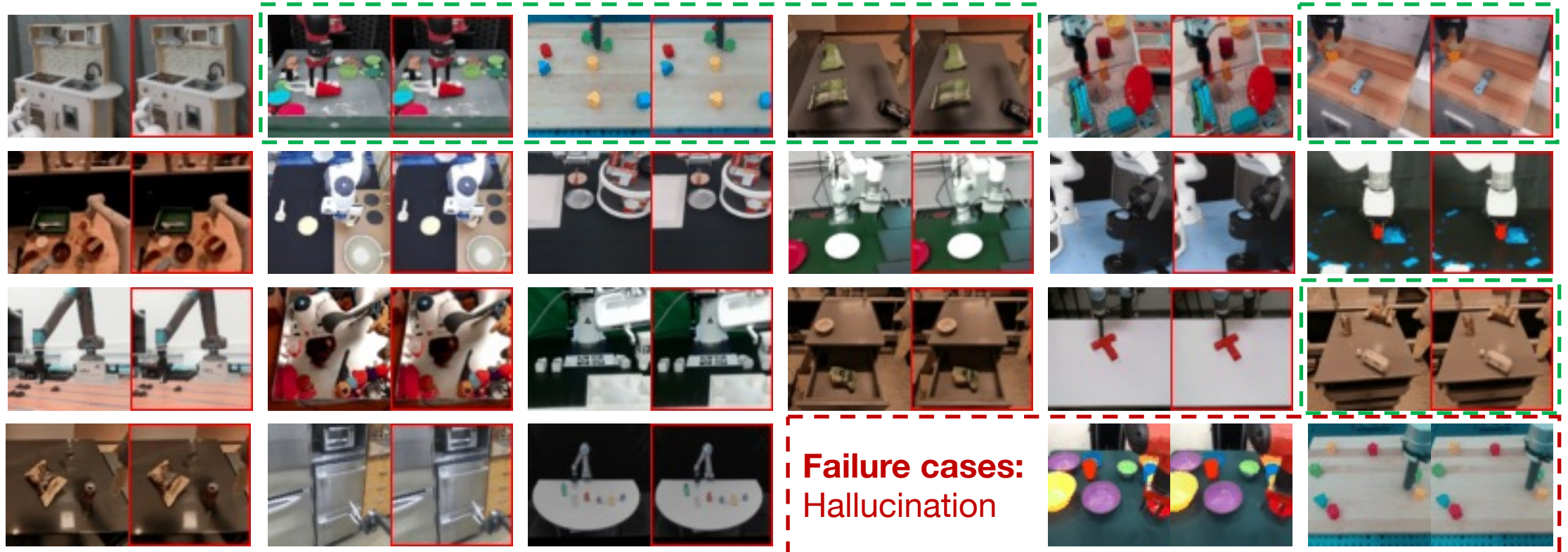
Initially pre-trained action-free,
flexibly allows for **action-conditioning**

Primary experiments at 64×64,
easily extended to **high resolution** 256×256

iVideoGPT provides competitive performance compared to state-of-the-art methods, MAGVIT for BAIR and FitVid for RoboNet

Video Samples: Open X-Embodiment (Action-free)

Natural movement diverging from ground truth, without actions



*Left: ground truth, right: prediction.
Red border: context frames, green border: predicted frames.*

Video Samples: Open X-Embodiment (Goal-conditioned)

Flexibility of sequence modeling

Rearranging the frame sequence

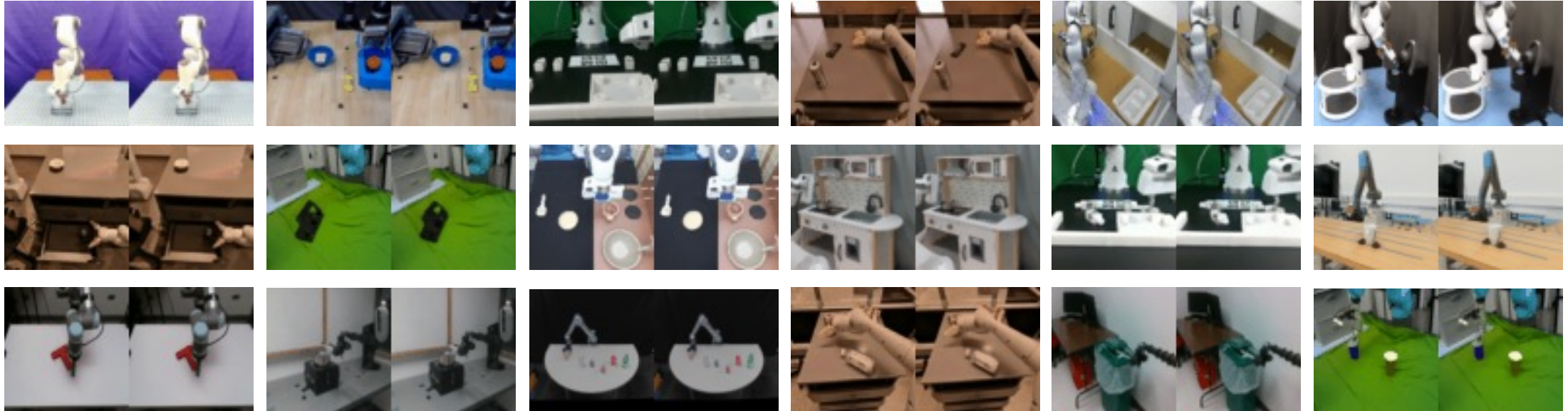


Goal-conditioned video prediction

$$\tilde{o}_{1:T} = (o_T, o_1, o_2, \dots, o_{T-1})$$

$$p(o_{T_0+1:T} \mid o_{1:T_0}, o_T)$$

More accurate paths to reach specified goals



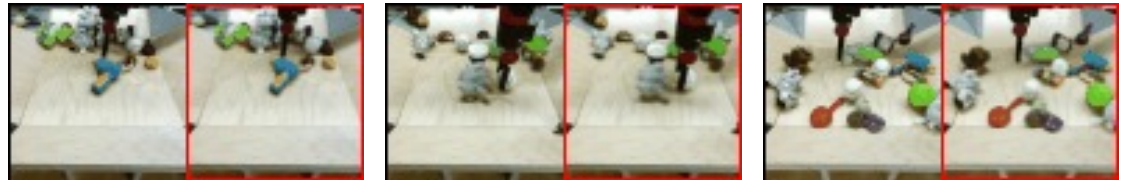
Video Samples: BAIR Robot Pushing & RoboNet

BAIR Robot Pushing Ebert et al. CoRL 2017

Action-free



Action-conditioned



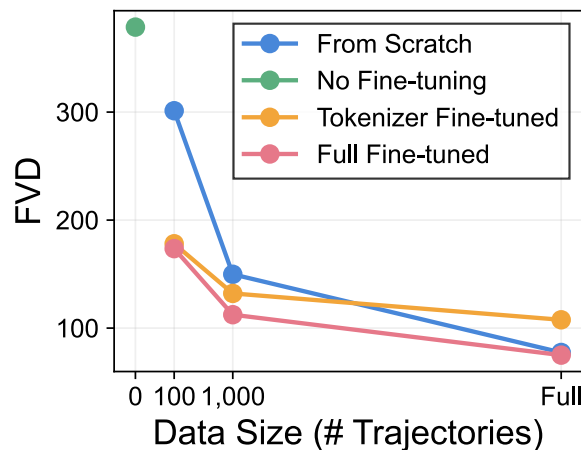
RoboNet (Action-conditioned) Dasari et al. CoRL 2019



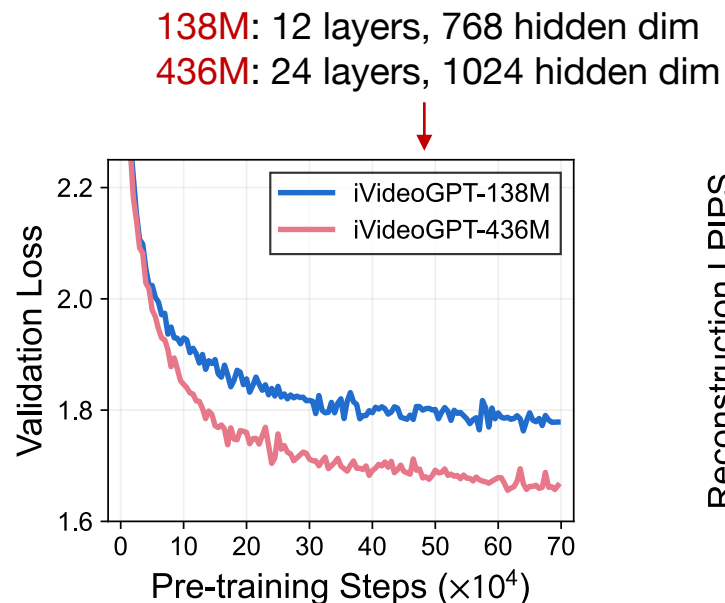
**High
Resolution:
256 × 256**



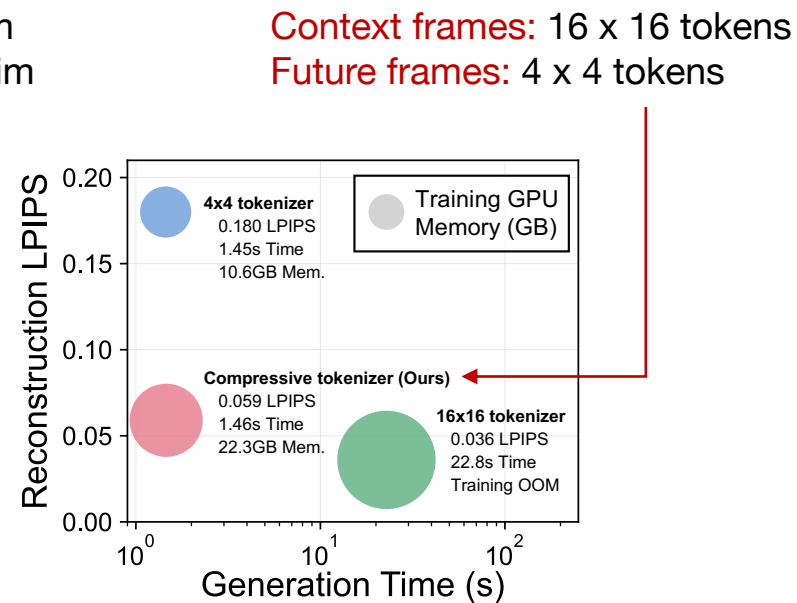
Model Analysis



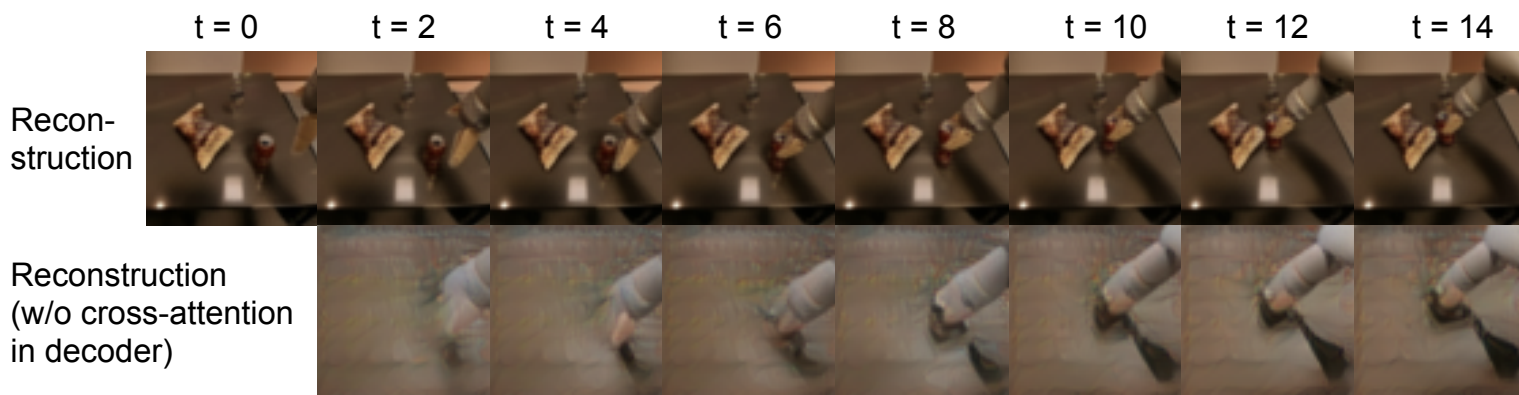
Few-shot adaptation: significant advantages under data scarcity



Model scaling: increased computation can build more powerful iVideoGPTs



Tokenization efficiency: memory savings during training and faster rollouts during generation

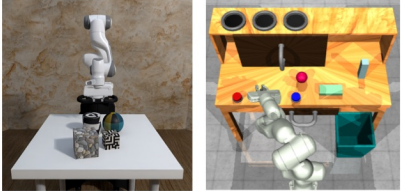
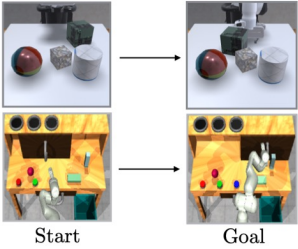
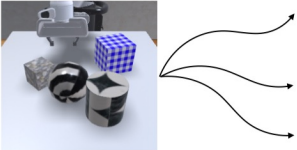
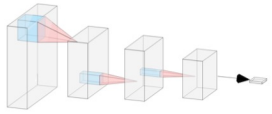



Context-dynamics decoupling: visualizing by removing cross-attention to context frames in the decoder when reconstructing future frames

Visual Planning

Excellent perceptual metrics do not always correlate with effective control performance

VP2: A control-centric benchmark for video prediction

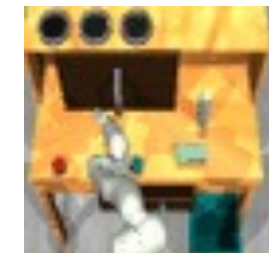
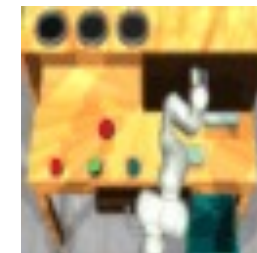
Task Definitions	Planning Implementation
<p>Environments (Robosuite & RoboDesk)</p>  <p>Task instance specifications</p> 	 <p>MPPI/CEM sampling-based optimizers</p>  <p>Pre-trained classifier cost functions</p>
<p>Training Datasets</p>	<p>Video Prediction Interface</p>
<p>Expert scripted interaction datasets</p> 	<pre># Only required to implement one function! def __call__(self, context_frames, action_seq): # Input: 2 context frames & T actions # Output: Predictions for T future frames return model_predictions</pre>

Model-predictive control



Goal observation

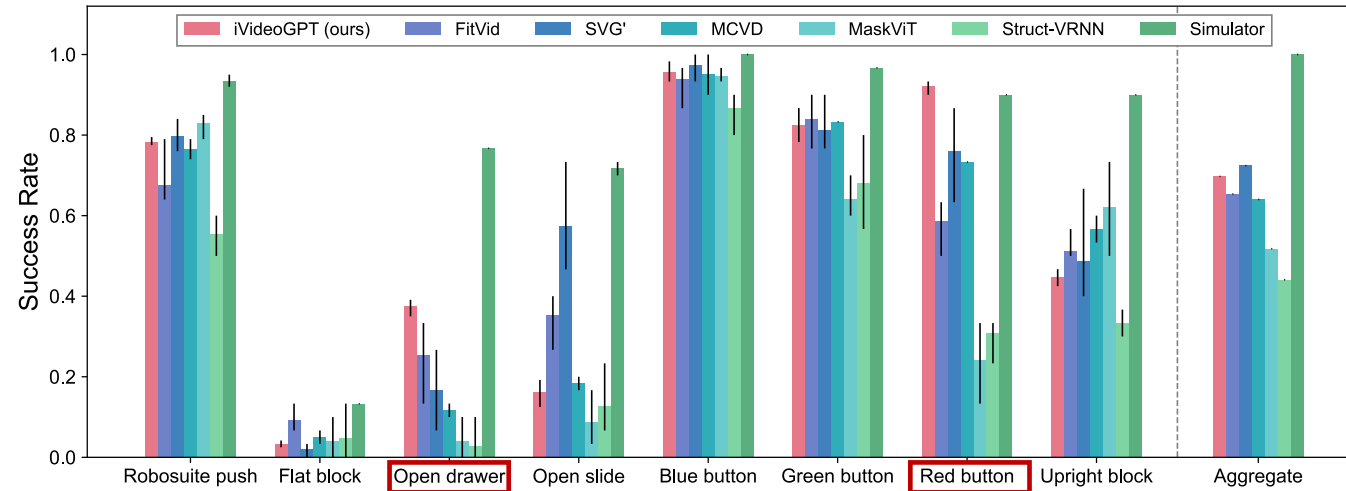
Successful trajectory



Goal observation

Successful trajectory

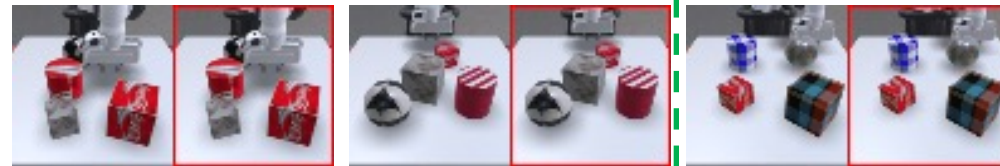
Visual Planning: VP2



iVideoGPT outperforms all baselines in two RoboDesk tasks with a large margin and achieves comparable average performance to the strongest model.

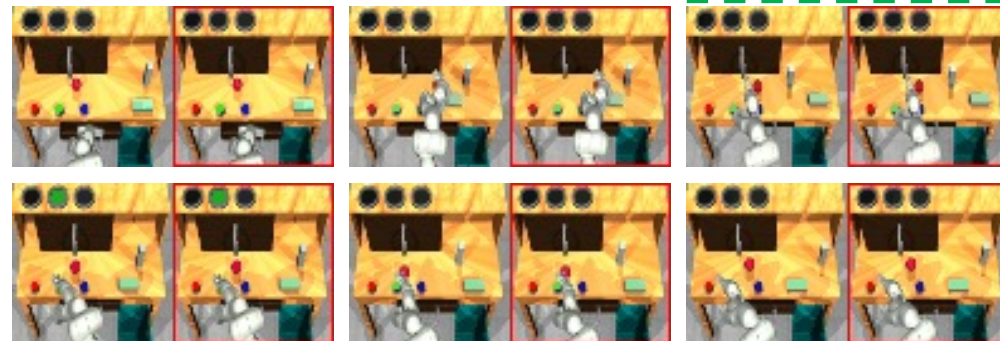
Video Samples:

RoboSuite



Predicted natural collision

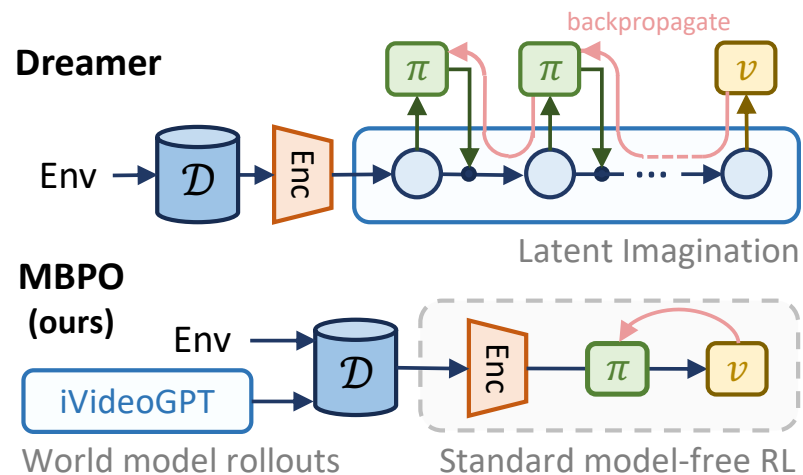
RoboDesk



Visual Model-based RL

Model-based RL with iVideoGPT:

- **Adapted from MBPO:** Augments the replay buffer with **synthetic rollouts** into replay buffer to train a **standard actor-critic RL** algorithm (DrQ-v2)
- **Eliminate latent imagination:** **Decoupling model and policy learning** can substantially simplify the design space, facilitating real-world applications



Algorithm 1 Model-Based Policy Optimization (MBPO), adapted from [40]

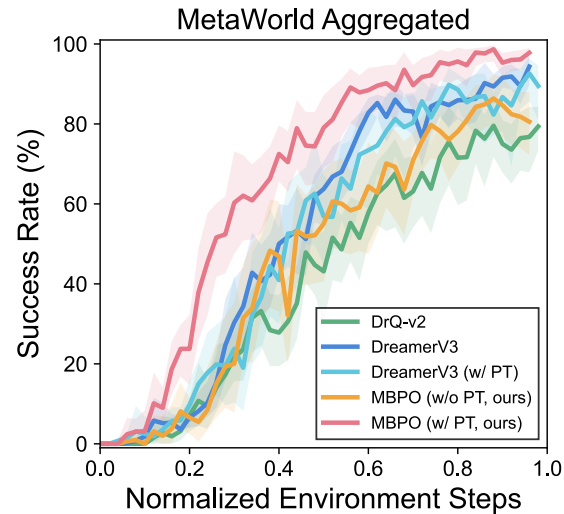
```
1: Initialize actor-critic  $\pi_\phi, v_\psi$ , world model  $p_\theta$ 
2: Initialize real replay buffer  $\mathcal{D}_{\text{real}}$  with random policy
3: Initially train model  $p_\theta$  on  $\mathcal{D}_{\text{real}}$ 
4: Initialize imagined replay buffer  $\mathcal{D}_{\text{imag}}$  with random rollouts using  $p_\theta$ 
5: for  $N$  steps do
6:   // Training
7:   if model update step then
8:     Update world model  $p_\theta$  on a mini-batch from  $\mathcal{D}_{\text{real}}$ 
9:   end if
10:  Update actor-critic  $\pi_\phi, v_\psi$  with model-free objectives on a mini-batch from  $\mathcal{D}_{\text{imag}} \cup \mathcal{D}_{\text{real}}$ 
11:  // Data collection
12:  if model rollout step then
13:    Sample a mini-batch of  $o_t$  uniformly from  $\mathcal{D}_{\text{real}}$ 
14:    Perform  $k$ -step model rollout starting from  $o_t$  using policy  $\pi_\phi$ ; add to  $\mathcal{D}_{\text{imag}}$ 
15:  end if
16:  Take action in environment according to  $\pi_\phi$ ; add to  $\mathcal{D}_{\text{real}}$ 
17: end for
```

Janner, Michael, et al. When to trust your model: Model-based policy optimization. NeurIPS 2019.

Yarats, Denis, et al. Mastering visual continuous control: Improved data-augmented reinforcement learning. ICLR 2022.

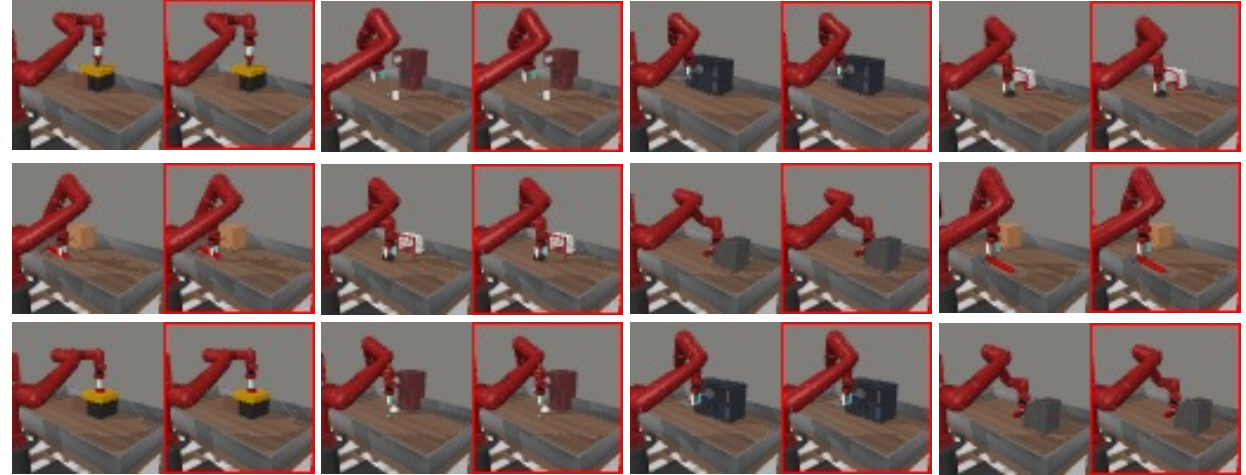
Visual Model-based RL: Meta-world

Six Meta-world manipulation tasks



Video Samples:

True and predicted rewards are labeled at the top left corner.



- Empowered by iVideoGPT:
 - **remarkably improves** over its model-free counterpart; **matches or exceeds DreamerV3**
- Baseline comparison:
 - **iVideoGPT trained from scratch** can **degenerate** the sample efficiency
 - **DreamerV3** does not benefit from **ineffective pre-training**

Summary

- **iVideoGPT**, a generic and efficient world model architecture based on compressive tokenization and autoregressive transformers
- Pre-trained on millions of human and robotic manipulation trajectories
- Adapted to a wide range of downstream tasks, particularly:
 - Accurate and generalizable video prediction
 - Simplified yet performant model-based RL



Open Source

iVideoGPT Public

Edit Pins Unwatch 4 Fork 3 Starred 68

main 2 Branches 0 Tags

Go to file Code

Manchery Update README.md 2413461 · last week 21 Commits

assets	Init commit	6 months ago
configs	fix	last month

About

Official repository for "iVideoGPT: Interactive VideoGPTs are Scalable World Models" (NeurIPS 2024), <https://arxiv.org/abs/2405.15223>

[thuml.github.io/iVideoGPT/](https://github.com/thuml/iVideoGPT/)

thuml/ivideogpt-oxe-64-act-free like 0

Diffusers Safetensors arxiv:2405.15223 License: mit

Model card Files Community Settings Use this model

main ivideogpt-oxe-64-act-free 1 contributor History: 4 commits Add file

manchery	Update README.md	474ab84	VERIFIED	12 days ago
tokenizer	upload models			12 days ago
transformer	upload models			12 days ago

<https://github.com/thuml/iVideoGPT>

Pre-trained model, training & inference code released

Thank You!

Code Available: <https://github.com/thuml/iVideoGPT>

Contact: wujialong0229@gmail.com

Machine Learning Group, School of Software, Tsinghua University

<http://ise.thss.tsinghua.edu.cn/~mlong/>

