iVideoGPT: Interactive VideoGPTs are Scalable World Models

Jialong Wu*, Shaofeng Yin*, Ningya Feng, Xu He, Dong Li, Jianye Hao, Mingsheng Long#

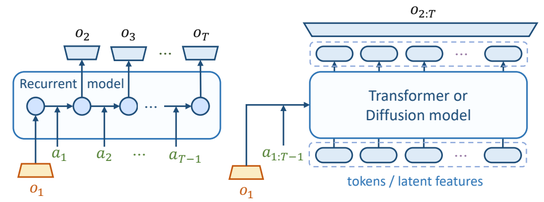
NEURAL INFORMATION PROCESSING SYSTEMS

Motivation: Video Generation vs. World Models

Are Video Generation Models World Worlds?

Not Yet!

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

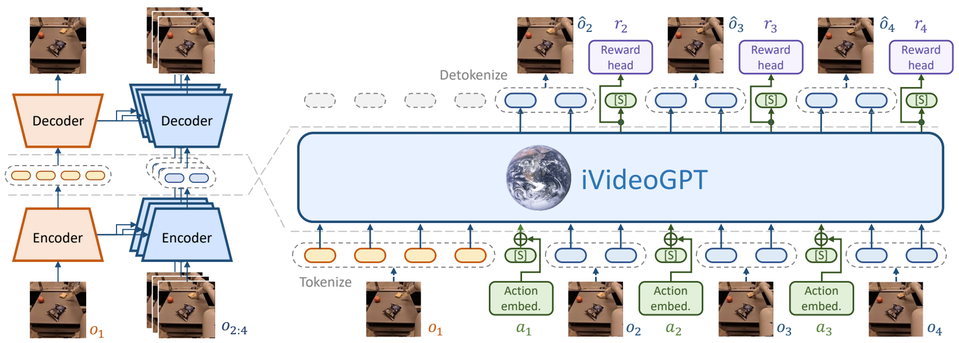


- Recurrent world models have limited **scalability**
- Video generative models have limited **interactivity**

Method: 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

Context frames independently tokenized:

- Rich in contextual information: N tokens each frame:

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

Future frames conditionally tokenized:

- Temporal redundancy: $n \ll N$ tokens each frame:

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

conditional encoder — conditional decoder

Overall objective:

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

context frames — future frames

Benefits:
- ✓ Faster rollouts
- ✓ Temporal consistency

Interactive Prediction with Transformers

A sequence of tokens:

Delineate frame boundaries and facilitate optional action and reward integration

$$x = \left(z_1^{(1)}, \ldots, z_1^{(N)}, \texttt{[S]}, z_2^{(1)}, \ldots, z_2^{(N)}, \ldots, \texttt{[S]}, z_{T_0+1}^{(1)}, \ldots, z_{T_0+1}^{(n)}, \ldots\right)$$

context frame — slot token — future frame

Flexibility of sequence modeling

Rearranging the frame sequence

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

→ Goal-conditioned video prediction

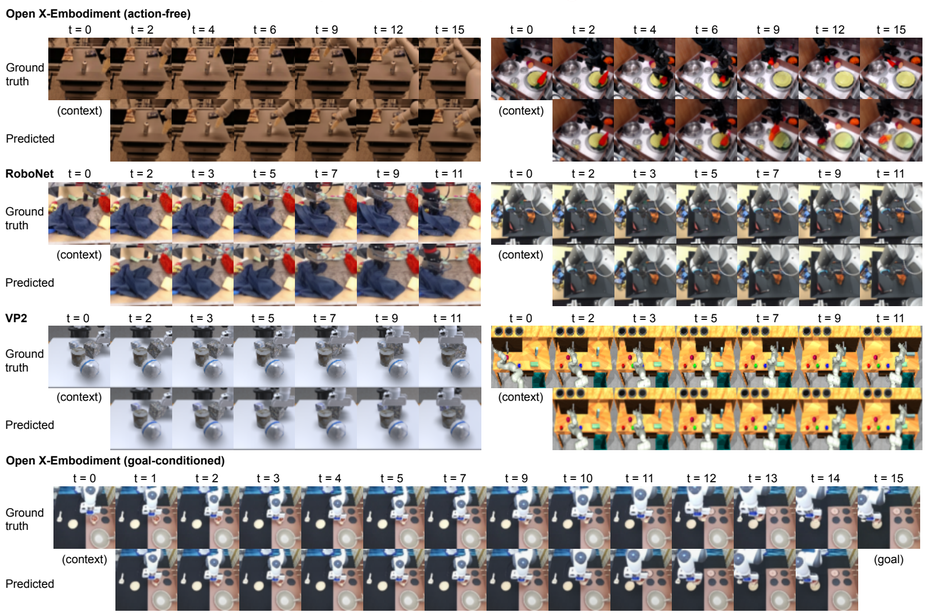
$$p(o_{T_0+1:T} \mid o_{1:T_0}, o_T)$$

Pre-training & Finetuning

Action-free video pre-training on 1.4 million trajectories (Open X-Embodiment + Something-Something V2)

Experiments

Video Samples: https://thuml.github.io/iVideoGPT



Video Prediction

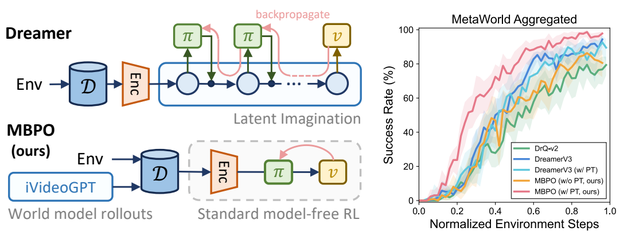
- Open X-Embodiment, BAIR, RoboNet
- Supporting action-conditioning, goal-conditioning and high-resolution

Visual Planning

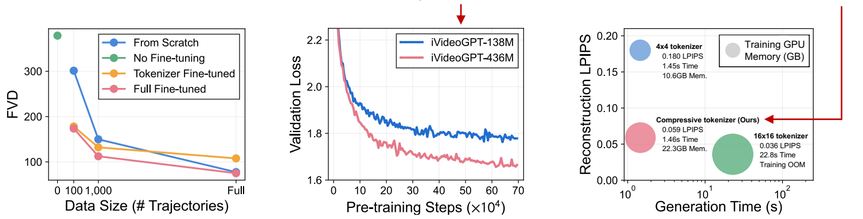
- VP2 benchmark

Visual Model-based RL

- **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



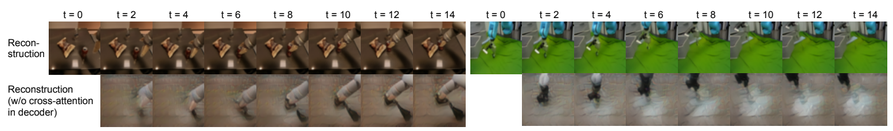
Model Analysis



Few-shot adaptation: Significant advantages under data scarcity

Model scaling: Increased computation can build more powerful iVideoGPTs

Tokenization efficency: 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

School of Software - BNRist - Tsinghua University - China

Code: https://github.com/thuml/iVideoGPT

Mail: wujialong0229@gmail.com