FROM VIDEO GENERATION TO WORLD MODELS

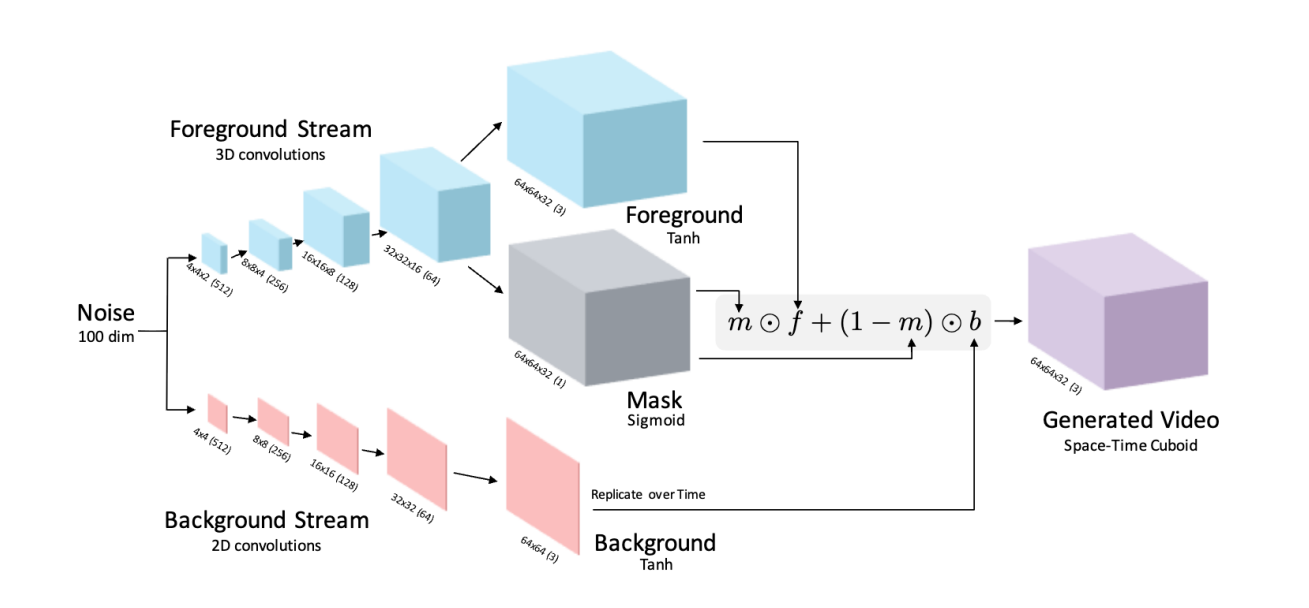
Chenhao Zheng

TABLE OF CONTENTS

HISTORY OF VIDEO GENERATION

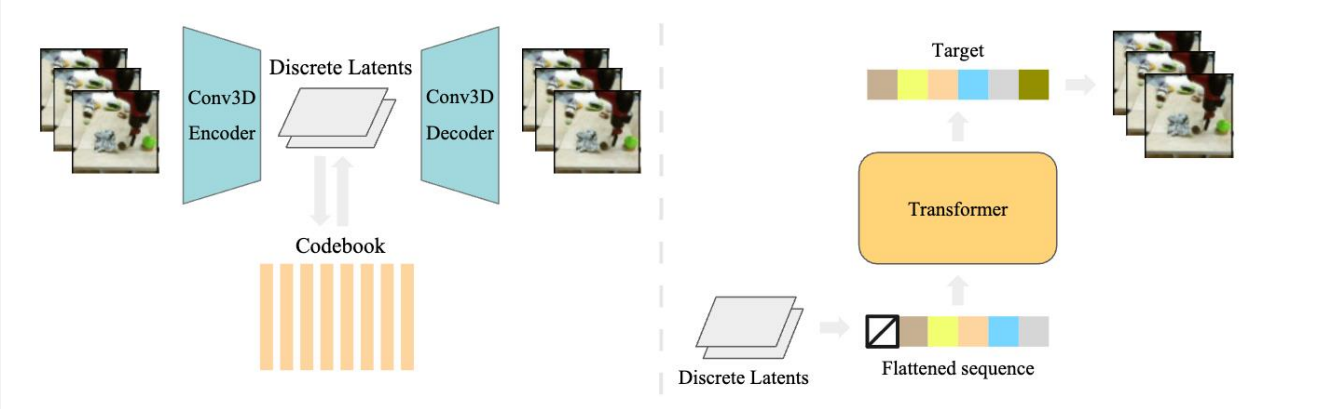
VGAN: Generating Videos with Scene Dynamics, 2016

- GAN with spatio-temporal convolution
- Separately model moving foreground and static background
- Generate very short, low-resolution videos



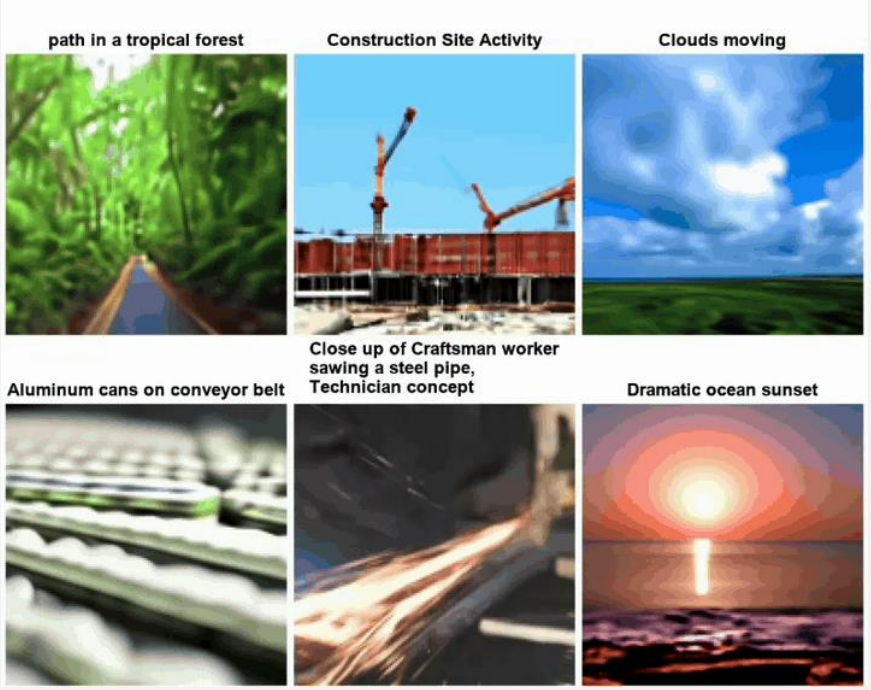
VideoGPT: Video Generation using VQ-VAE and Transformers, 2021

- Represent videos as spacetime patches or latent tokens
  - modeling a compressed representation rather than pixel
- First attempt to use transformer



Diffusion model Era

- Diffusion models were easier to scale and gave better sample quality than previous methods
- At that time CNN is still better: Typically UNet based architecture
- Example paper 1: Video Diffusion Models, 2022: **Pixel-space**
- Example paper 2: Video LDM / latent video diffusion, 2023: **latent space (and is better)**



Video diffusion model, 2022

Video LDM, 2023

Scaling up model and data

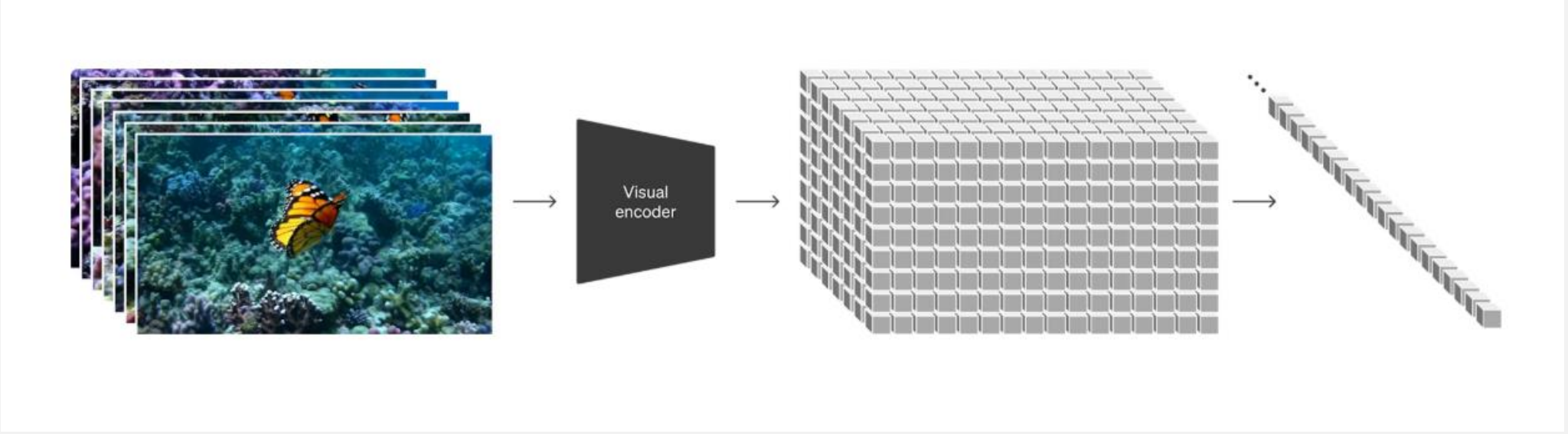
- **Sora From OpenAI**
  - Also a diffusion model
  - Much more photo realistic
  - Good instruction following ability

Prompt: Photorealistic closeup video of two pirate ships battling each other as they sail

CASE STUDY: SORA MODEL

Transform Pixel to latent

- Stage 1: Train an Encoder to compress videos to latent
  - Reduce dimensionality
  - AutoEncoder like architecture and Training
- Stage 2: From compressed latent, extract a sequence of spacetime patches
  - Further reduce dimensionality

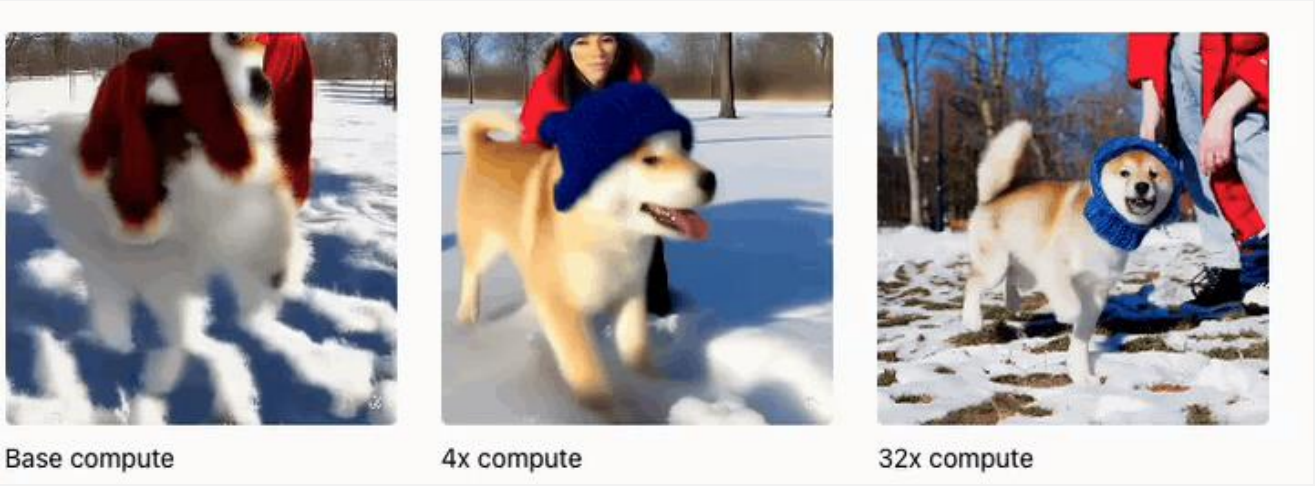


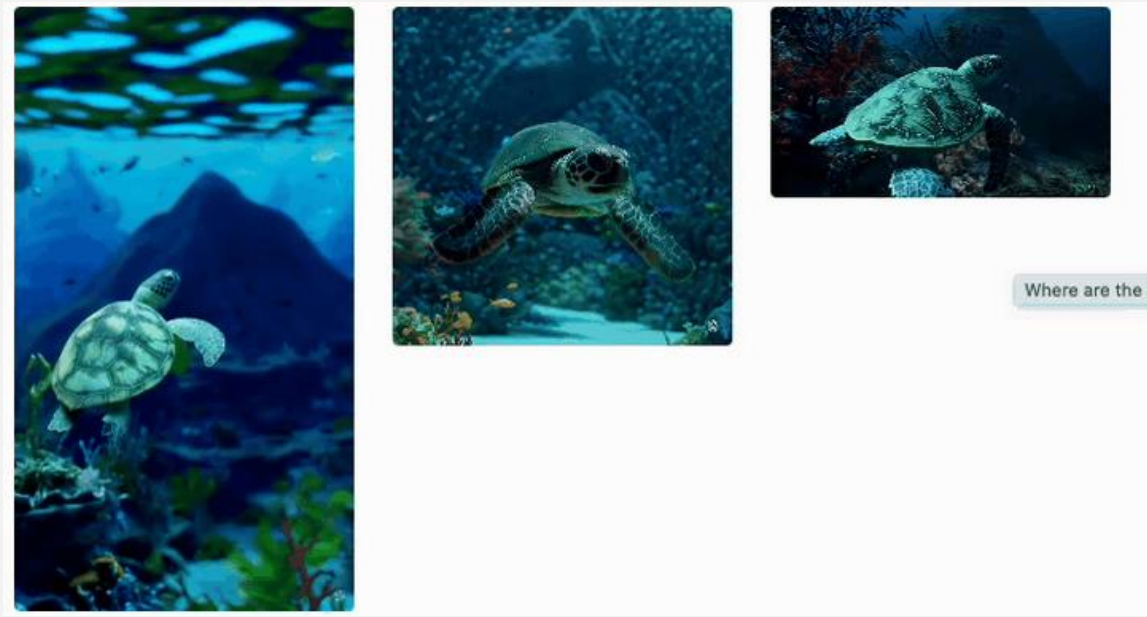
Diffusion transformers

Diffusion objective; Use **Transformer** architecture

- Scales well with model size; better than CNN at large scale
- Allow various aspect ratio and size

Capacity: [website link]





RECENT PROGRESS

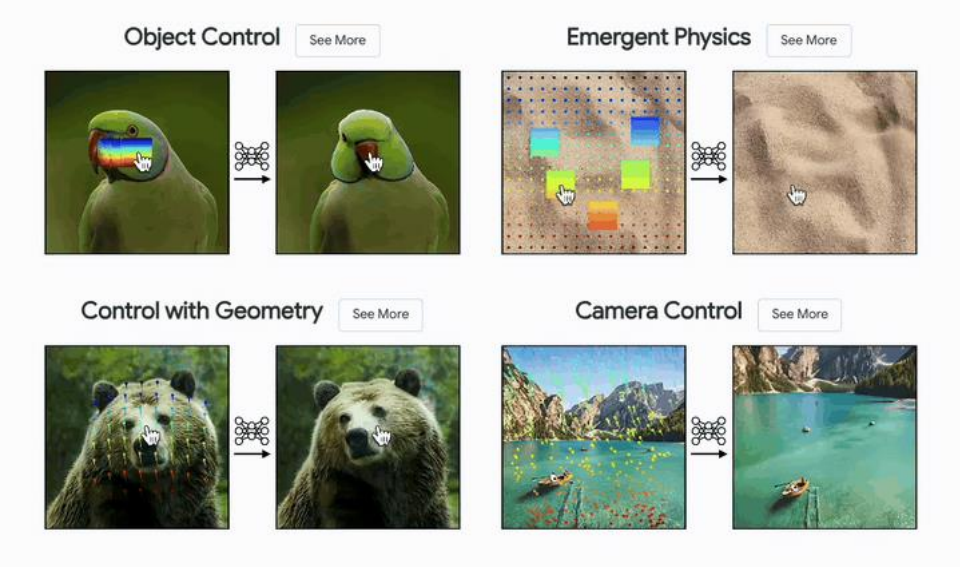
Recent Progress: object and 3D consistency

Sora still face object and 3D inconsistency issue

Recent works solve by

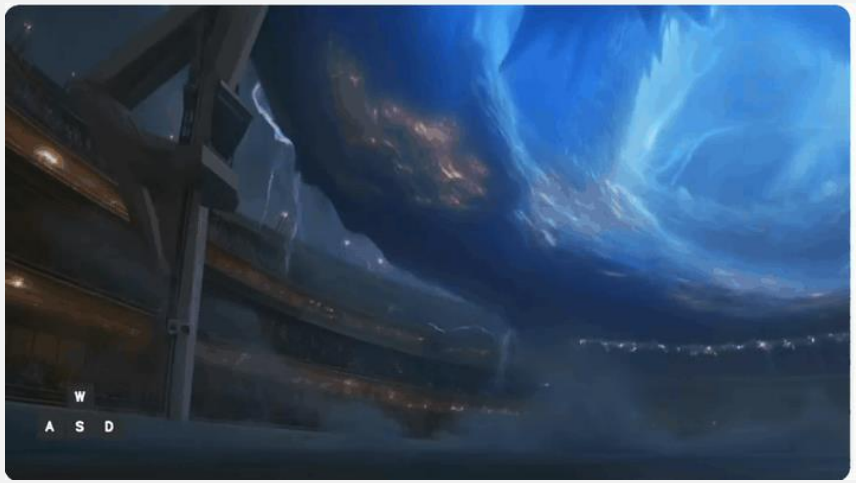
- Scale up data and model further ….
- Use explicit spatial-temporal memory

Recent Progress: more control

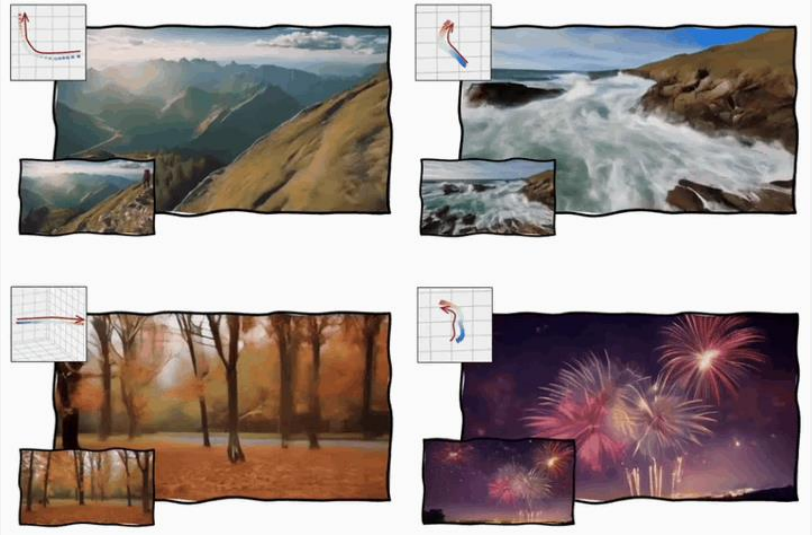


Motion Control

Style Control



Action Control



Camera Control

Recent Progress: Minute-Length Video

- Few second video -> minute level video
- Long video generation is a hard problem:
  - One-forward pass – sequence length too long, long context problem in transformer;
  - Multi-forward roll out -- drifting problem

Popular method: diffusion forcing & test time training

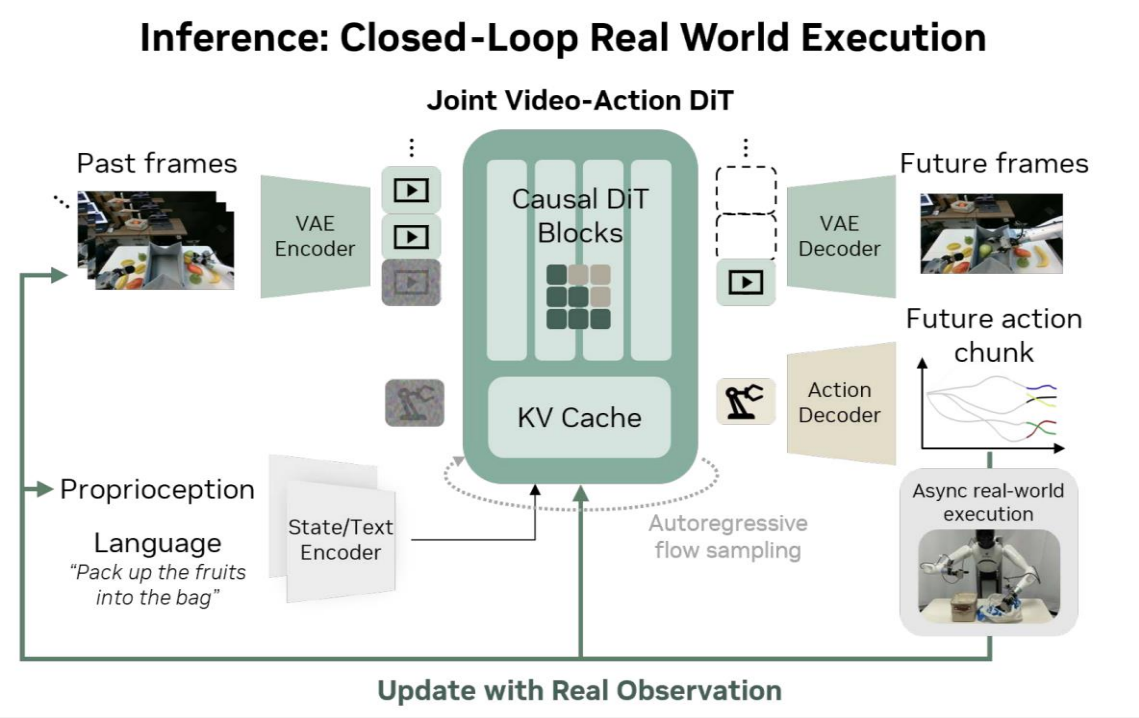
THE ROAD TO WORLD MODELS

Video generation can be "world simulator": predicting what happens next as the state of the world

Video generation for robotics

predict the outcome of robot action before taking action

Two ways to go from generated video to robot action:
Early: Inverse Dynamics
Recent: Joint Video-Action Modeling

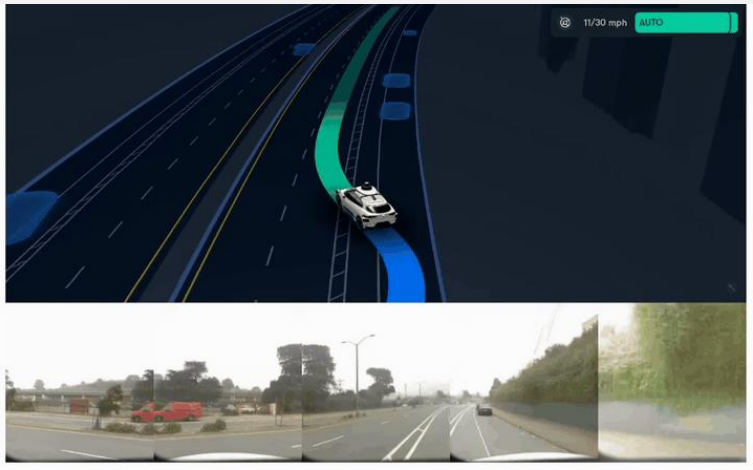


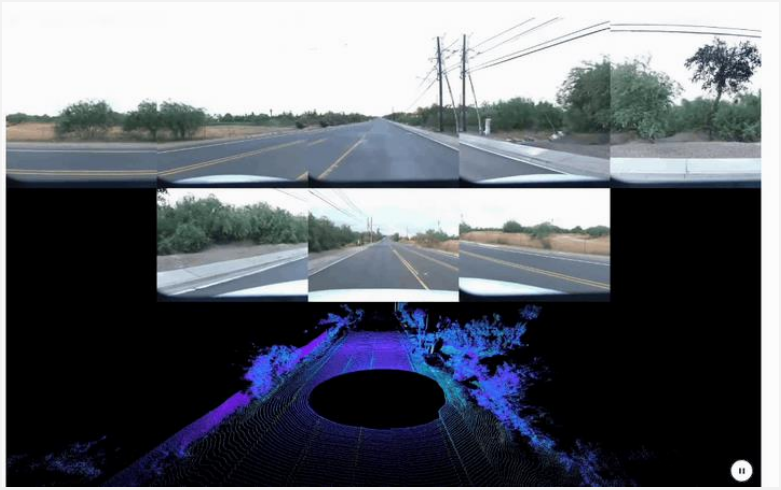
Video generation as realistic simulator

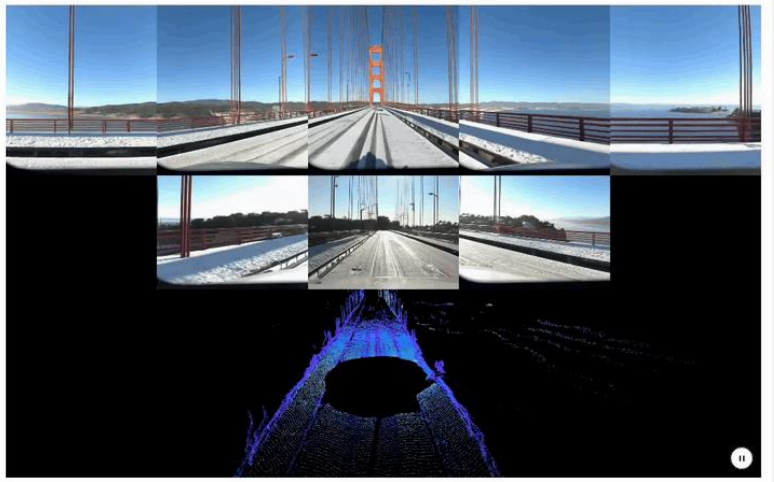
Problem: real-world testing is expensive, slow, and risky

Video world simulator: real log / scene / action / prompt → realistic future sensor video

Example: Waymo world model



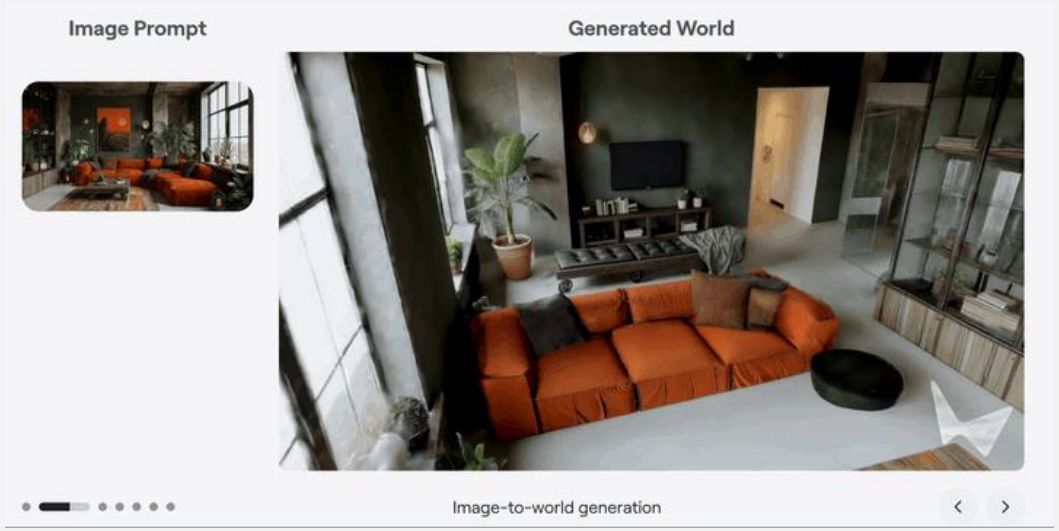




Video generation for interactive environments

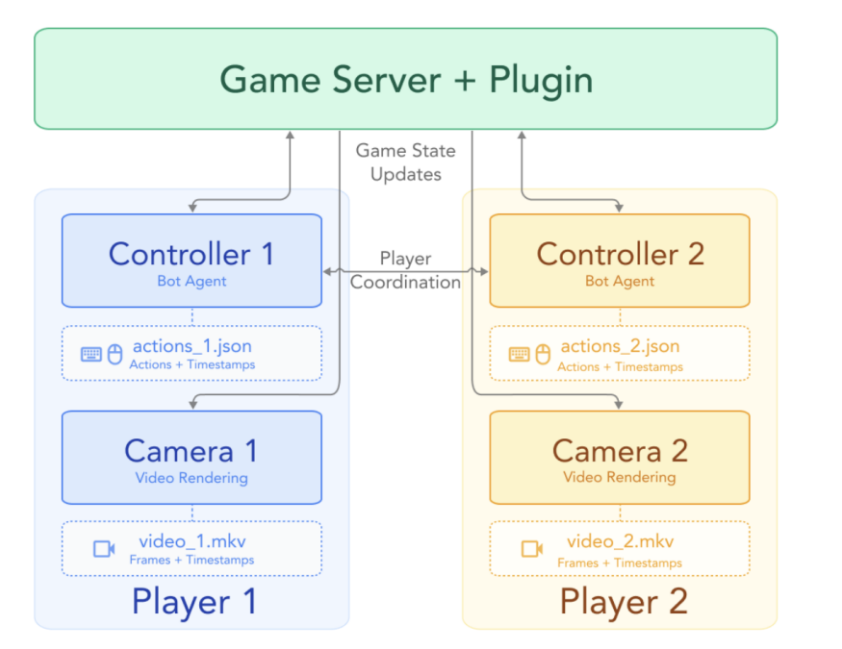
Static / spatial world model:
generate a persistent 3D world that users can move through, edit, and inhabit

Dynamic / game-engine world model:
generate the next visual state in real time as the user takes actions



Multi-agent world model

- The video generation model generates synchronized egocentric observation for all agents while they act on the same evolving environment.
- Challenge: not just 3d consistency, but also it is consistency of **evolving states and dynamics**
- Interesting problem: State / memory representation that efficiently represents a shared evolving world state.



From beautiful video to usable video