Lecture 12:

Finishing up data generation/supervision for 3D Generation

+

Intro to Action-Conditioned Video Models

Visual Computing Systems
Stanford CS348K, Spring 2026

Class discussion: SAM 3D

How do we know if a 3D scene generation is “correct”?

Techniques for automatically (or semi-automatically) validating the output of scene generation

Discussion led by Sharon

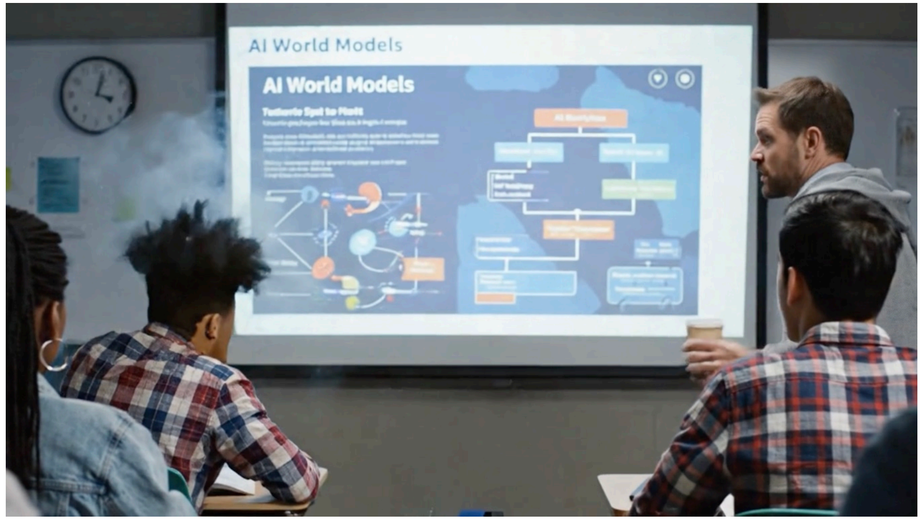
Action-conditioned video model basics

Let’s begin with video generation models

- **Model’s task: text description —> video**

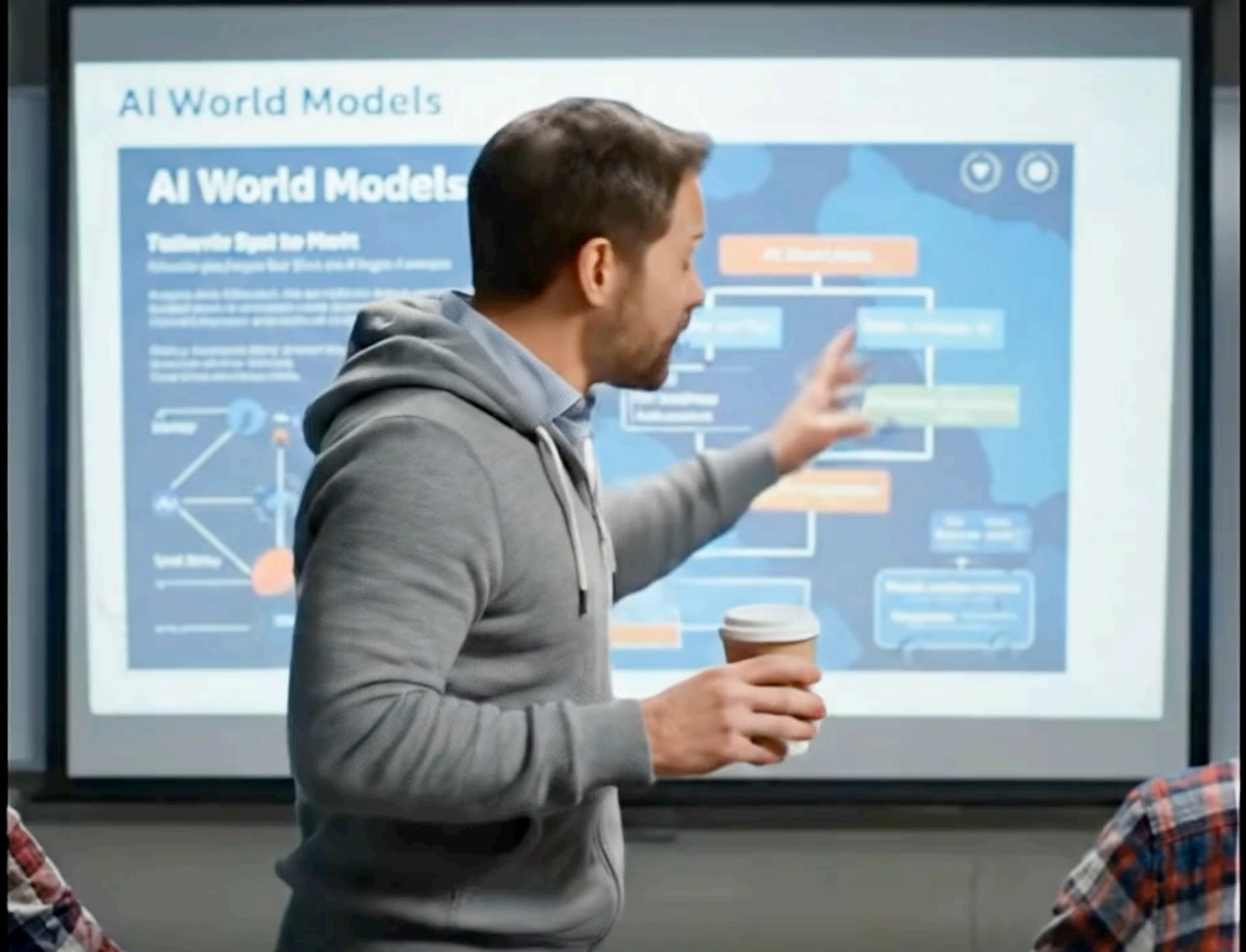
Video of a 40-something male professor with short brown hair wearing a surprisingly trendy gray cashmere hoodie holding a disposable to-go cup of tea in one hand while lecturing passionately and vigorously to a class of college students in a small classroom. Make the professor point wildly at slides that depict AI world models. Make the professor say "The thing I want you to focus on in your projects is doing a thoughtful evaluation. And if you don't, there will be huge consequences.” After he says the word "consequences" make a lightning bolt flash and zap one of the students as an example of consequences.


AI World Models
AI World Models



Start with the frame I'm uploading with this prompt. Now make the professor in the gray hoodie holding the cup of tea say "And if you do, the rewards will be incredible." After he says the word "incredible" make one of the students that was not hit by lightning shout "YES" in triumph, do a fist pump, and then get buried in a pile of cash falling from the sky.


AI World Models
AI World Models


AI World Models
AI World Models

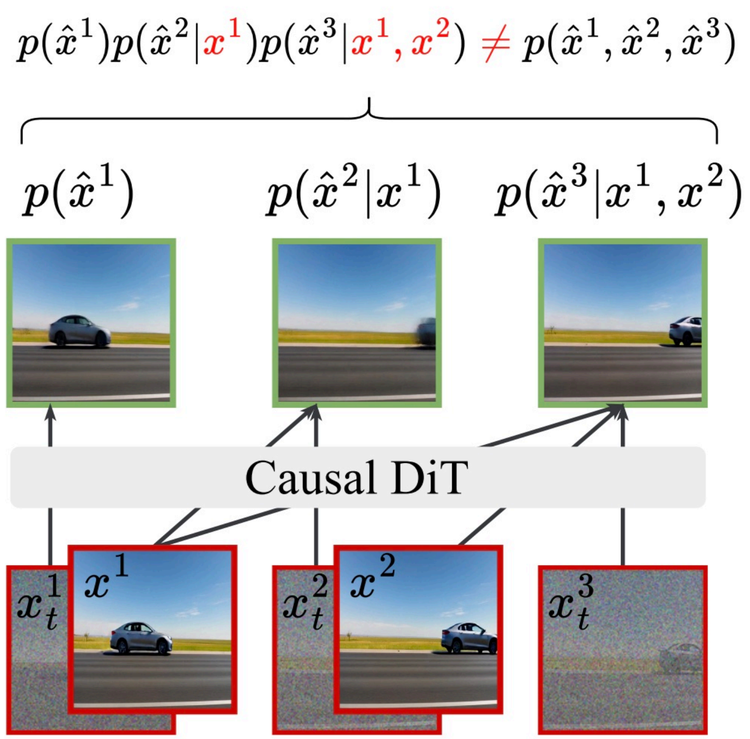
Autoregressive video generation

- Generate frames one at a time. Future frames are conditioned on prior frames
- $(\text{frame}_{(0..t)}) \longrightarrow \text{frame}_{t+1}$

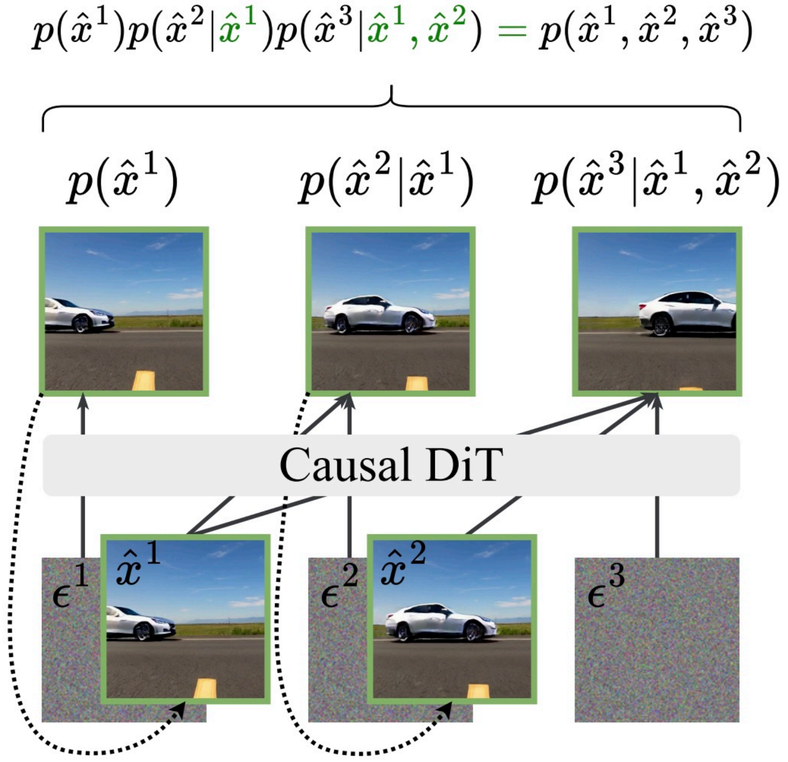
- e.g., supports streaming video if frame generation is real time

The “exposure bias” problem

- Consider training an autoregressive video model to reproduce video training data
  - During training, the model is given “ground truth” inputs from prior frames and asked to predict the next frame
  - But at runtime, the model does not have access to ground truth frames, only the frames it produced!
- Solution (“self forcing”): feed models own outputs back to itself during training, so it trains on the data it produces



(a) Teacher Forcing Training

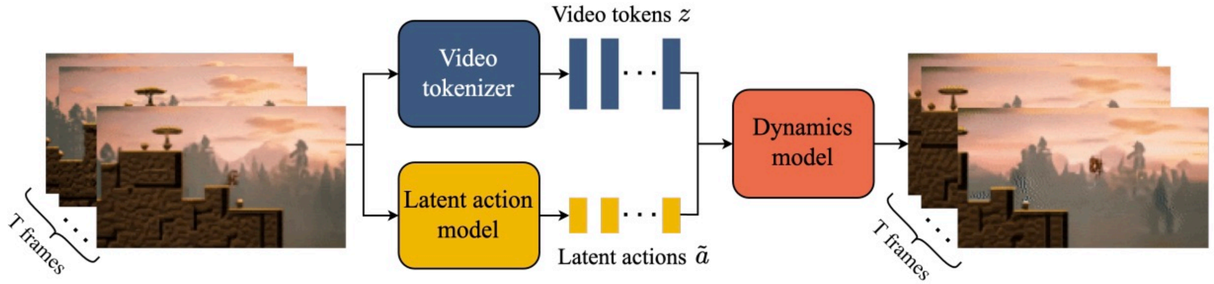


(c) Self Forcing Training (ours)

Action-conditioned video generation models

- $(\text{state}_{(0..t)}, \text{action}) \longrightarrow \text{state}_{t+1}$

Examples: Deepmind's Genie / NVIDIA COSMOS



Input: scene prompt + action at each frame

Output: next frame RGB
(and potentially additional per-frame data: embeddings, attributes)

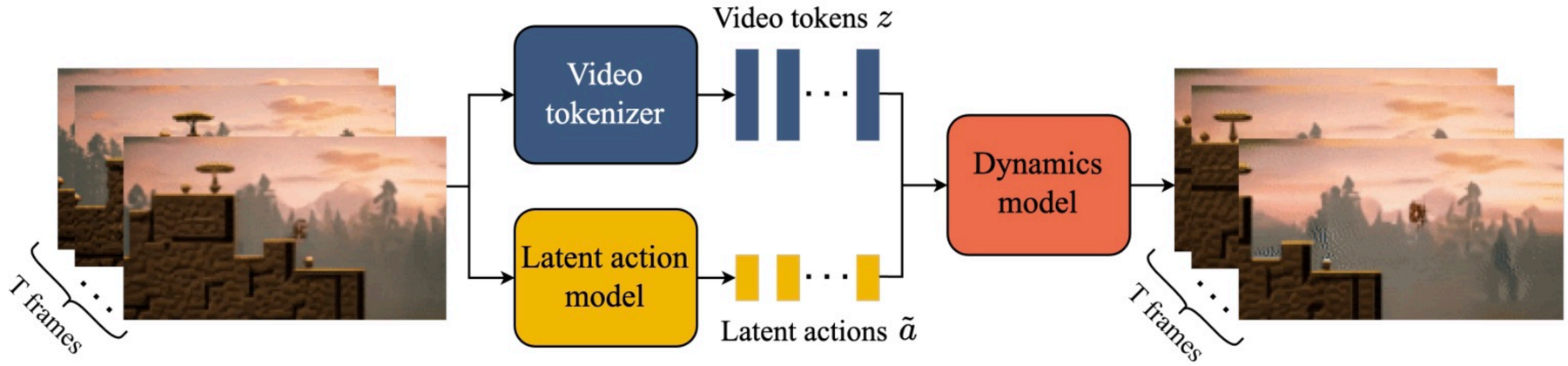
Training data??

- Now what is the training data?

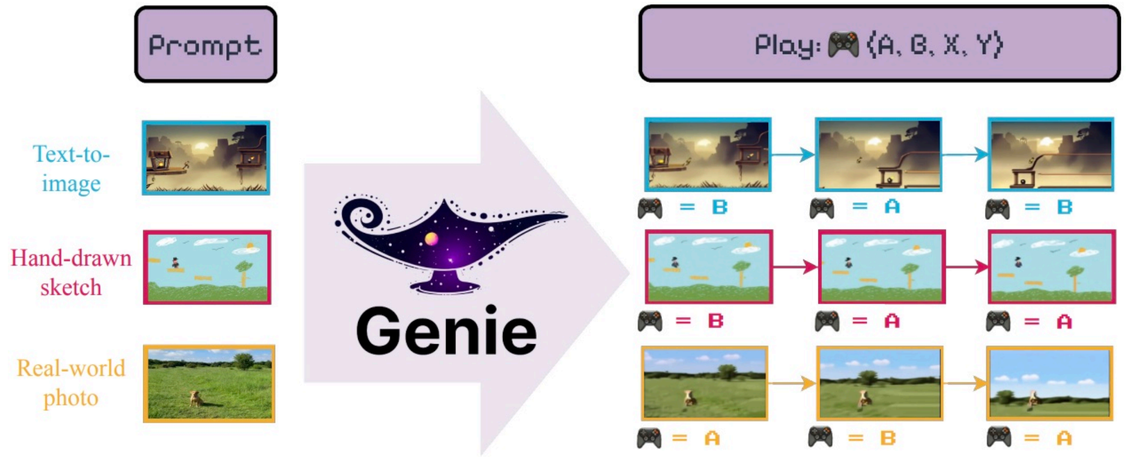
- What questions might you ask when curating this training dataset?

Genie (tonight's reading)

- **Key idea: learn a world simulator from videos of video game play**
  - **From video, learn latent user actions, and dynamics model that steps work given (current state, action)**



  - **Then at "test time" given a novel world state (perhaps one generated from a prompt), and given user input, time step the novel world forward in time**



Modern pattern for action conditioned models today

- Phase 1: Train a great video generation model
  - Huge numbers of videos (billions?)
  - From this data: learn how the world visually "time evolves"

- Acquire a dataset of video/action sequence pairs
  - Typically much smaller (hundreds of hours), because it is expensive to acquire
    - Requires simulated data, or humans playing computer games, or humans tele operating robots, etc.
  - From this data: learn how time evolution responds to stimuli

Challenges of training action-conditioned models

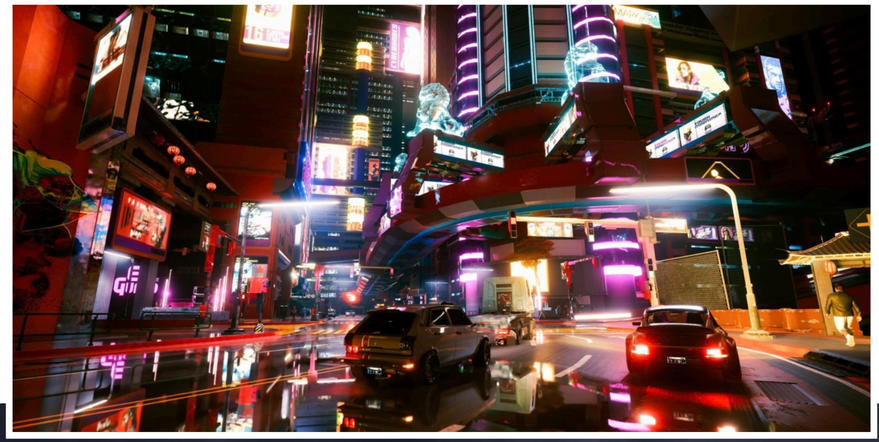
- How to acquire training data?
  - Need to pair actions with frames
  - Need to acquire examples of interactions that the model must learn

- How to ensure model ensures "consistency" over time (aka how to encode the world's state)
  - Subject of next class

The great CS348K debate

Team World Models

Team Traditional Simulation Engine



The scenario

- You now know a bit about how generative AI will help accelerate the development of many "traditional" 3D scene representations: generating 3D objects, producing scene layouts, generating animations, scripts, etc.

- And you've had an early taste of how one might create a simulator by never modeling any of this 3D information: just by learning how the world evolves from annotated video

- Your job is to lead a team for the next 10 years to develop the ultimate virtual world simulator for training AI agents

- Consider technologies we have talked about in class so far for simulating worlds… what technology bets will you make? What is your first step?

Things to consider

- Consider types of tasks you may want to support
- Consider what types of interactions the simulation must accuracy represent
- Consider the fidelity of simulation you want to achieve?
  - What does “high fidelity” even mean?

- Consider various costs
  - Costs to create (or acquire via capture) world content
  - Costs to execute simulations and training
  - Costs to develop the systems or debug agents that you train