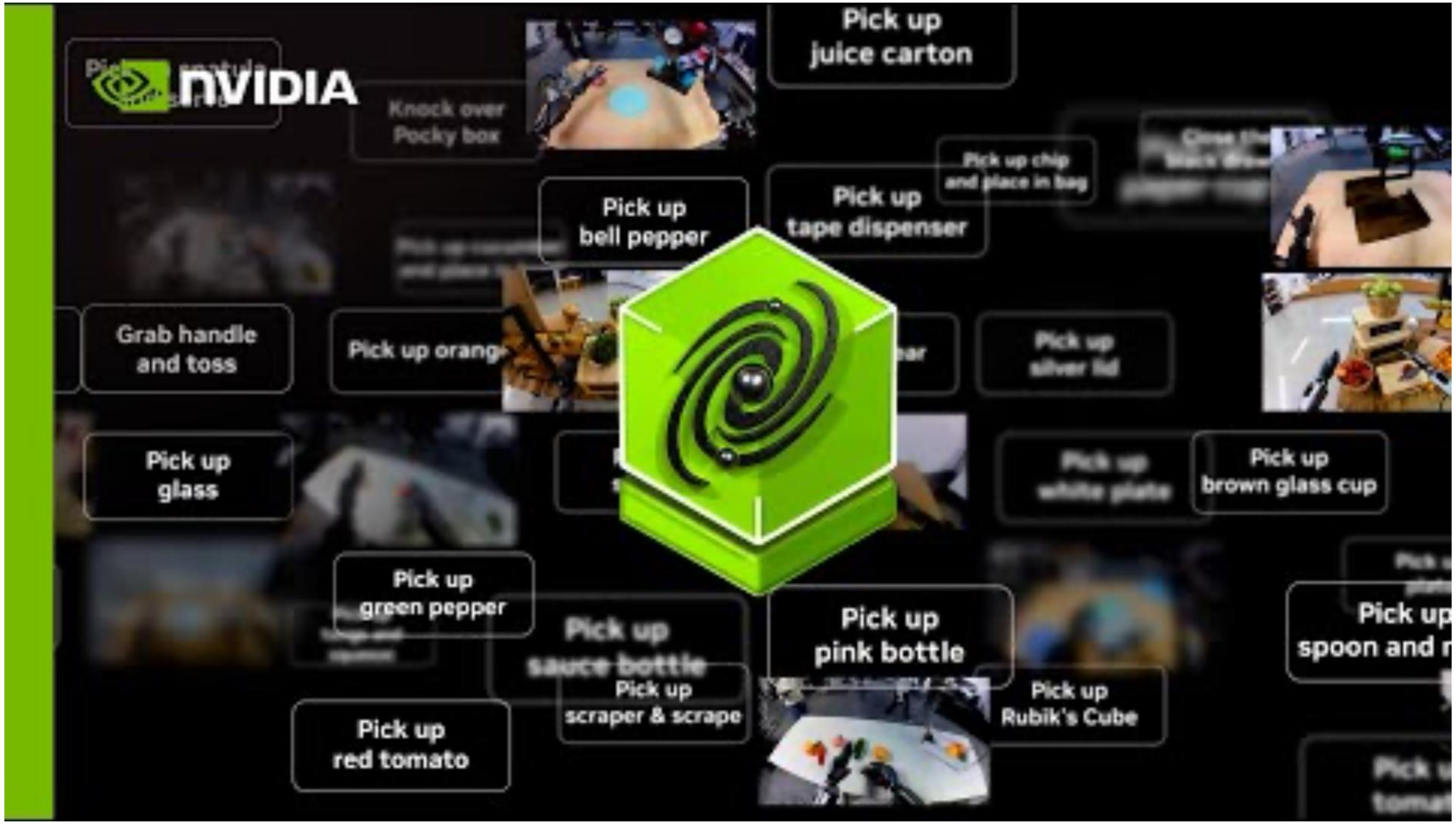
AA203
Optimal and Learning-based Control

World Models for Robotics
(Extra Lecture)

Stanford University

Autonomous Systems Lab

World Models: a ‘new’ robot learning paradigm?



[How Robot Brains Dream and Explore Unseen Worlds](#) (NVIDIA)

World Models: a 'new' robot learning paradigm?

AI pioneer Fei-Fei Li's World Labs raises $1 billion in funding
Reuters

Yann LeCun's AMI Raises $1BN Seed Round – Is the World Model Era Finally Here?
Futurum

'World models' are AI's latest sensation: what are they and what can they do?
nature

Genie 3: A new frontier for world models

1X Launches Humanoid Robot World Model Lab: 'You Can't Fine-Tune Your Way To AGI'
Forbes

Announced yesterday!

What are World Models?

An internal, predictive model of how the world evolves — under your actions.

Given the current state and a planned action sequence, a world model predicts future observations. It lets an agent **imagine outcomes before acting**.

Two intellectual roots converge:

- **Model-based control & RL** — system identification, MPC, Dyna-style planning.
- **Generative sequence models** — learning dynamics directly from high-dimensional data.

A SHORT LINEAGE

1943	Craik: the mind builds "small-scale models" of reality to predict and plan.
1990s	Mental models in cognitive science; learned dynamics in control.
2018	Ha & Schmidhuber, "World Models" — learn latent dynamics, train a policy in the "dream."
2024+	Large generative & VLA world models for robotics.

Key distinction: a world model is ***action-conditioned*** *— it answers "what happens if I do this?", not merely "what comes next?"*

Formalizing a world model

A world model is a conditional distribution over future states:

$$p\left(x_{t+1:t+H} \mid x_t,\ a_{t:t+H-1},\ l\right)$$

- ***x*** — state / observation at time t (images, proprioception, …)
- ***a*** — the planned action sequence over a horizon H
- ***l*** — optional task context (e.g. a language instruction)
- **Output** — a distribution over future states / observations

Narrower than "future prediction"

Because it is conditioned on actions, a world model is **controllable**: it models the consequences of decisions, not just the passive flow of the scene.

Why world models matter for robots

THREE CORE CAPABILITIES

1. **Foresight**
   predict the consequences of actions before executing them.
2. **Imagination-driven planning**
   search / optimize over imagined rollouts (MPC, trajectory optimization).
3. **Data amplification**
   generate synthetic experience to train data-hungry policies.

THE ROBOT POLICY

$$p(a_{t+1:t+k} \mid o_t, l)$$

Two dominant families:

– **Visuomotor policies** — e.g. diffusion / behavior-cloning policies.
– **VLA models** — vision-language-action: RT-2, OpenVLA, π_0.

A world model can ***supervise, evaluate, or generate data*** *for any of these policies.*

One distribution, four queries

Many robot-learning paradigms are just **conditionals or marginals** of a single joint distribution over future observations and actions:

$$p\left(o_{t+1:t+k},\ a_{t+1:t+k}\ \middle|\ o_t,\ l\right)$$

Policy

$$p(a_{t+1:t+k} \mid o_t,\, l)$$

act from observations

Passive world model

$$p(o_{t+1:t+k} \mid o_t,\, l)$$

forecast the scene, action-free

Controllable world model

$$p(o_{t+1:t+k} \mid o_t,\, a_{t+1:t+k},\, l)$$

predict given planned actions

Inverse dynamics (IDM)

$$p(a_{t+1:t+k} \mid o_{t:t+k},\, l)$$

infer actions from transitions

Video generation models

World models that predict directly in pixel space:

$$p\left(v_{t+1:t+H} \mid o_t,\ a_{t:t+H-1},\ l\right)$$

WHY MODEL PIXELS?

- **Internet-scale pretraining** — reuse priors from massive video corpora.
- **General observation space** — one representation across tasks and embodiments.
- **Inspectable rollouts** — you can literally watch the model's imagination.

WHAT ROBOTICS DEMANDS

- Action controllability
- Temporal consistency
- Physical plausibility
- Near-real-time generation

Examples: UniSim, Genie, Cosmos, driving world models.

Architectures: from decoupled to unified

A spectrum of how tightly perception, prediction, and action are coupled:

IDM-style → **single-backbone** → **MoE / MoT** → **unified VLA** → **latent-space**

decoupled & modular *internalized & end-to-end*

Unified backbone

one token stream mixes visual and action latents.

$$x = [\, z^v;\ z^a \,] \qquad \hat{y} = f_\theta(\tilde{x}_\tau,\ o_t,\ l,\ \tau)$$

Mixture-of-Transformers

modality-specific experts share attention.

$$(h^v_{\ell+1},\ h^a_{\ell+1}) = F^{\text{mix}}_\ell(h^v_\ell,\ h^a_\ell;\ o_t,\ l)$$

IDM: predict-then-act

imagine future frames, then infer the action.

$$\pi(a_t \mid o_t, l) = P(a_t \mid E(o_t),\ E(l),\ \Phi(\hat{o}_{t+1:t+H}))$$

Latent space models

JEPA-style: predict in representation space, not in pixels.

$$\min_{\theta} \| s_{\psi}\left(\hat{z}_{t+1:t+H}\right) - \mathrm{sg}[E_{\bar{\theta}}(o_{t+1:t+H})] \|^2$$

Pixel / video world model

- Models every pixel of the future
- Photorealistic, but spends capacity on irrelevant detail
- Reconstruction objective

Latent / JEPA world model

- Predicts abstract features of the future
- Efficient; focuses on decision-relevant structure
- No pixel reconstruction

Methods: V-JEPA, FLARE, VLA-JEPA. *Resonates with control intuition — plan over a compact latent/state space rather than raw observations.*

World models through the lens of this class

"World models" generalize the **learned-dynamics + planning** ideas of this course to high-dimensional, generative, language-conditioned settings.

WORLD-MODEL CONCEPT	THIS COURSE
Controllable world model	the learned dynamics model in optimal control / MPC
Planning in a world model	MPC / trajectory optimization over imagined rollouts
World model as simulator	model-based RL: Dyna, Dreamer-style policy learning
Inverse dynamics (IDM)	system identification — recover actions from transitions

Use cases of world models for robotics

(1) A learned simulator for reinforcement learning

Roll the model forward to produce imagined transitions, and train the policy inside it:

$$(\hat{o}_{t+1}, \hat{r}_t, \hat{d}_t) \sim p_\phi(\,\cdot\,|\, o_{\leq t}, a_{\leq t}, l\,) \qquad\qquad J(\theta) = \mathbb{E}_{\hat{\tau}\sim(\pi_\theta,\, p_\phi)}\Big[\sum_t \gamma^t\, \hat{r}_t\Big]$$

- **Sample-efficient & safe** — train on cheap synthetic rollouts instead of real-robot time.
- **Policy optimization in the model** — policy-gradient / GRPO over imagined returns.
- **Co-evolution** — model and policy improve together — the model sharpens where the policy visits.

Caveat: *policies can exploit model errors; compounding error over long horizons.*

More use cases: evaluation, imagination, driving

(2) World model as evaluator

Score candidate action sequences by their predicted outcome and pick the best — a learned cost / value for planning.

(3) Video imagination

Generate future rollouts for planning and for data. A progression: generate → ground actions → close the loop → run in real time.

(4) Navigation & autonomous driving

Predictive scene / occupancy forecasting; driving world models (e.g. GAIA, Cosmos) for simulation and planning.

Outlook on world models: what’s next?

Open challenges shaping the next few years:

Causal action conditioning
controllability, not mere correlation

Efficiency & real-time inference
fast enough to close the control loop

Multimodal perception
vision, touch, force, proprioception

Bridging classical control
stability and guarantees

Symbolic / structured models
compositional, object-centric state

Evaluation metrics
task & control relevance over pixel fidelity

Key takeaways

1. A world model is an action-conditioned predictive model of dynamics — "imagine before acting."
2. Policy, world model, and inverse dynamics are all queries on one joint distribution over observations and actions.
3. Two families: pixel / video world models (general, internet-scale) vs. latent / JEPA models (efficient, decision-relevant).
4. For robotics, world models act as simulators (RL), evaluators (planning), and data engines.
5. It connects directly to this course: learned dynamics + planning (MPC, model-based RL), at generative, language-conditioned scale.