NYU | COURANT INSTITUTE SCHOOL OF MATHEMATICS, COMPUTING, AND DATA SCIENCE

Advanced Machine Intelligence

World Models: Enabling the next AI revolution

Yann LeCun
Courant Institute, New York University
AMI Labs

2026-04-01 Brown University - Lemley Lecture

AI sucks! (compared to animals)

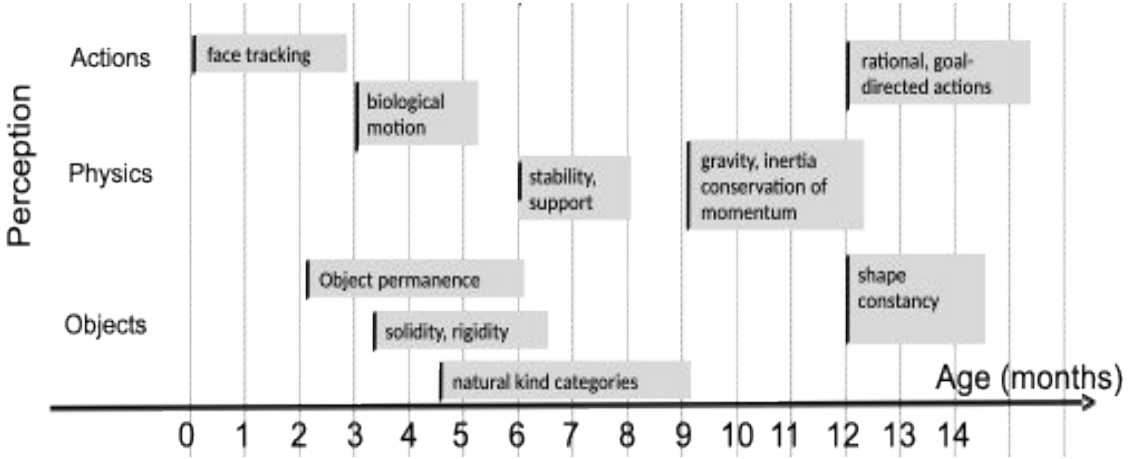
- **Why can AI write code, pass the bar exam and win IMO but can't clear up the dinner table, clean the house, or learn to drive in 20 hours of practice?**
- **Why can't AI deal with high-dimensional continuous data?**
  - image, video, audio, sensor outputs, instrument measurements, financial,…
- **Why do agentic systems need to be trained by imitation from enormous numbers of examples or through massive simulation?**
- **Why can't AI systems solve new problems "zero shot," the first time they see it, by searching for a solution?**
- **Why can't AI have common sense and understand the real world?**

The Real World is Messy (and Language is Simple)

- **Never mind humans, cats and dogs can do amazing feats**
  - Any house cat can plan highly complex actions
  - Any 10 year-old can clear up the dinner table and fill up the dishwasher without learning (“zero-shot”)
  - Current robots intelligence doesn’t come anywhere close
- **17 year-olds can learn to drive in 20 hours of practice**
  - AI systems that can pass the bar exam, do math problems, prove theorems….
  - ...but where are my Level-5 self-driving car and my domestic robot?
- **We keep bumping into Moravec’s paradox: Things that are easy for humans are difficult for AI and vice versa.**

How humans and animals learn

- Humans and animals learn **mental models of the world**
- Their behavior is **driven by objectives**
- They can **reason** and plan complex action sequences by **predicting the consequences of their actions**.
- arXiv:2603.15381 :



[after Emmanuel Dupoux]

Why AI systems don't learn and what to do about it

Lessons on autonomous learning from cognitive science

Emmanuel Dupoux[1,2], Yann LeCun[3], Jitendra Malik[1,4]

[1]FAIR at META, [2]École des Hautes Études en Sciences Sociales, [3]NYU, [4]UC Berkeley

What is Intelligence, Really ?

- **Intelligence is not (just) a collection of skills, nor an accumulation of declarative knowledge.**
- **Intelligence is the ability to accomplish new tasks and solve new problems with no prior training or with minimal training.**
  - Fast adaptation in front of new situations is intelligence
- **The phrase Artificial General Intelligence makes no sense**
  - Human intelligence is highly specialized
- **But there is no question that machines will eventually surpass humans**
- arXiv:2602.23643

AI Must Embrace Specialization via Superhuman Adaptable Intelligence

Judah Goldfeder [*1] Philippe Wyder [*2] Yann LeCun [3] Ravid Shwartz-Ziv [3]

Why Human-Level AI Requires Real-World Data

- **LLMs: are trained on human-produced text**
  - Typically 3.0E13 tokens (2E13 words). Each token is 3 bytes.
  - Data volume: 0.9E14 bytes.
  - Would take 400,000 years for a human to read (9 hours/day, 250 words/minute)

- **Human child, 4 years old: trained on sensory data**
  - 16,000 wake hours in the first 4 years (30 minutes of YouTube uploads)
  - 2 million optical nerve fibers, carrying about 1 byte/sec each.
  - Data volume: 1.1E14 bytes

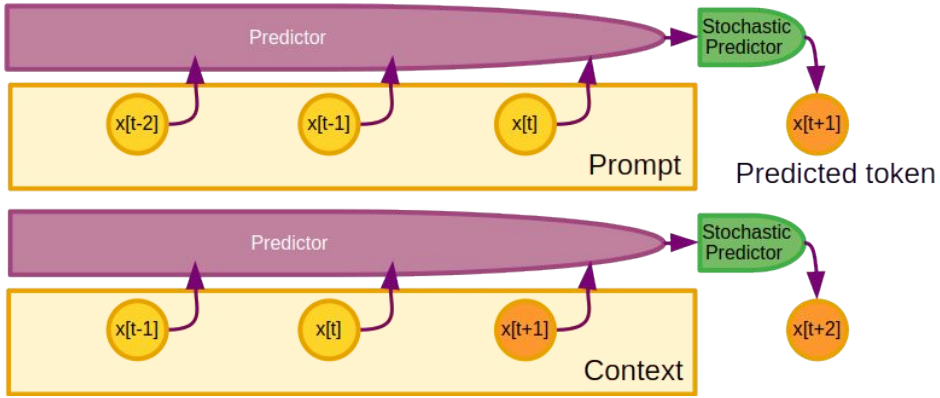
- **A four year-old child has seen more data than an LLM !**

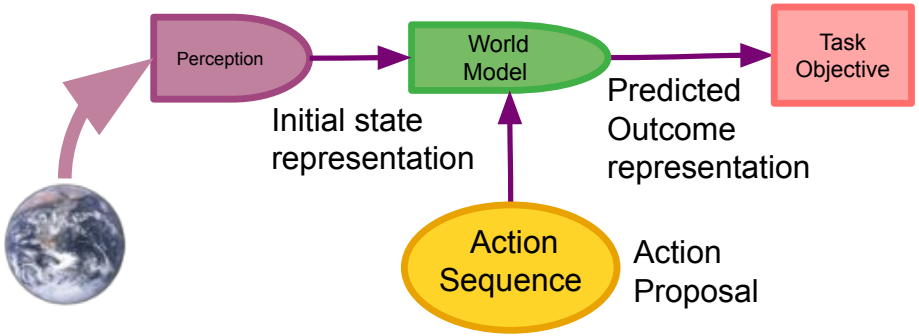
AMI Labs: Enabling the Next AI Revolution

- **Self-Supervised AI/ML technology that can model any data source**
  - high-dimensional, continuous, noisy data
  - image, video, audio, sensor data of any kind, scientific or engineering data….
- **World Model powered Agentic Systems that**
  - Understand any environment & system, including physical & living things
  - Have persistent memory (factual, working, episodic, procedural)
  - Can perform long chains of reasoning
  - Can plan complex action sequences hierarchically
  - Can accomplish new tasks "zero shot" without prior training and adapt quickly
  - Are controllable and safe
- This goes beyond what LLMs & other generative architectures are capable of

Two ways to build an agentic system

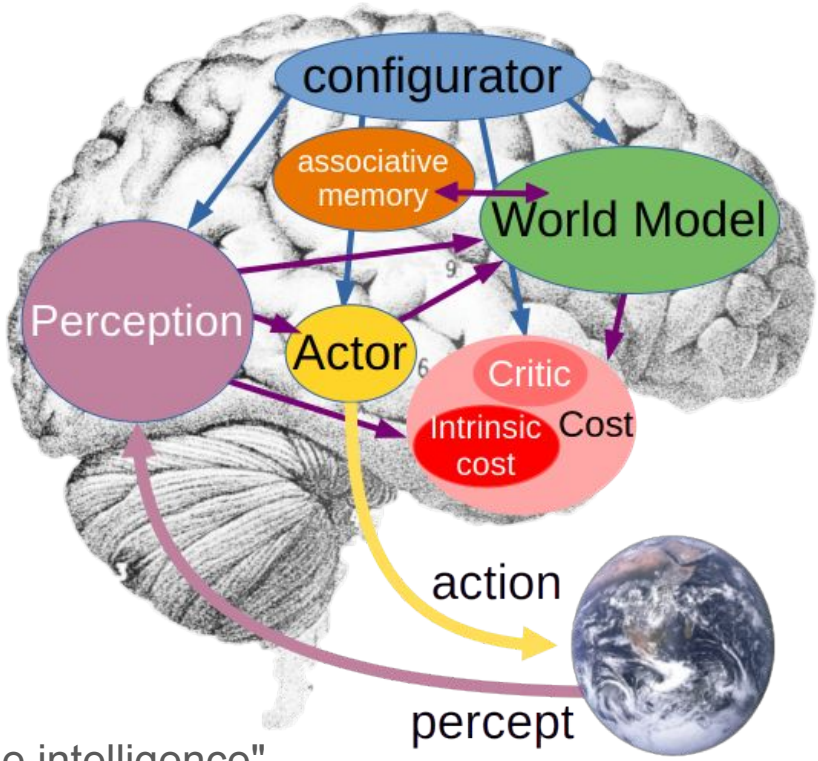
- **System 1: action prediction, maps state -> action**
  - No reasoning, no way to imagine the effect of actions
  - This is how LLM-based agentic systems work (or don't work)
- **System 2: planning, finds actions that accomplish a task**
  - Uses a **world model** to predict the consequences of actions
  - Searches for actions that accomplish the task
  - This is how AMI Labs' agentic systems work





Long-Term Vision: a complete cognitive architecture

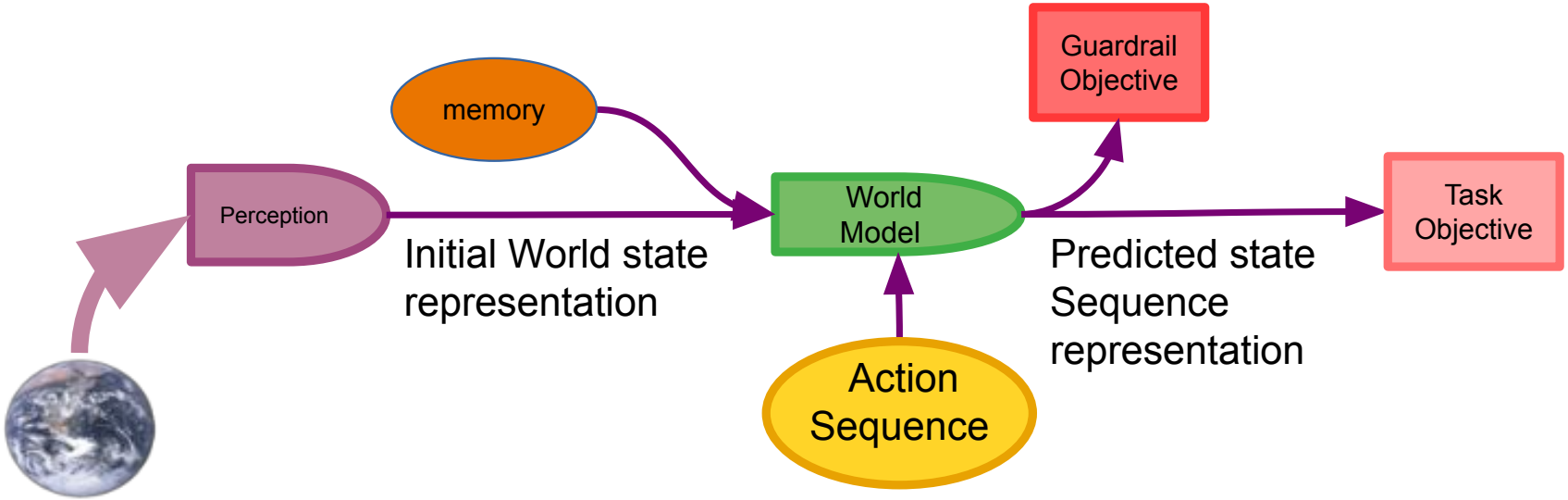
- **Long-term vision: to build an integrated, modular, cognitive architecture**
  - Objective-driven architecture
    - output/actions are produced by search/optimization so as to accomplish a task, subject to safety guardrails
  - Built around a world model
    - the world model predicts the consequences of imagined actions and interventions.



[LeCun 2022 "a path towards autonomous machine intelligence" https://openreview.net/forum?id=BZ5a1r-kVsf]

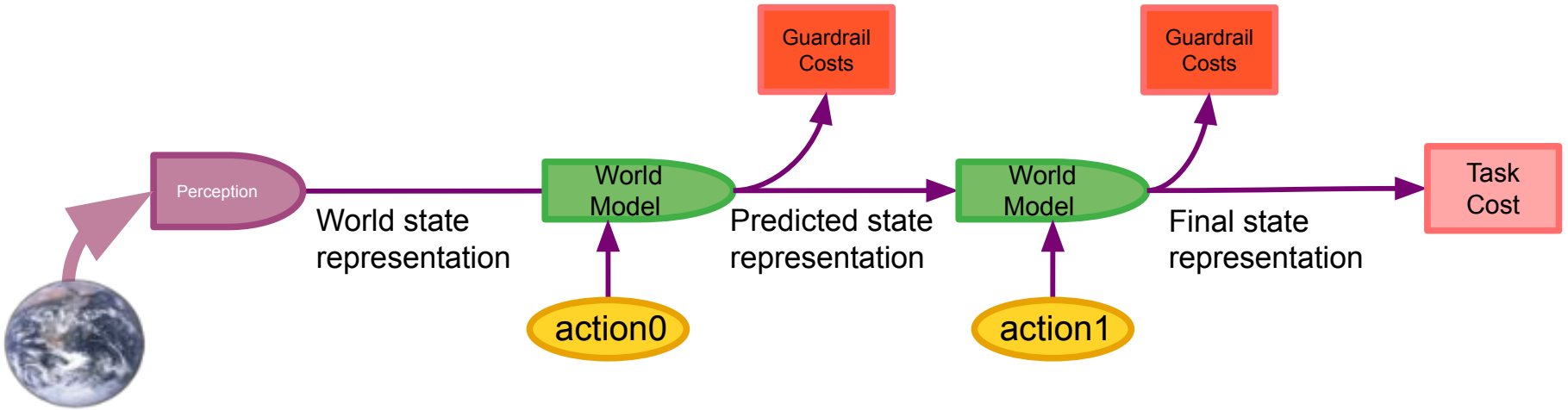
Objective-Driven Inference with a World Model

- World Model predicts outcomes of actions
- Actions are searched to optimize/complete the task objective
- Guardrail objectives implement safety constraints



Multi-Step Planning with a World Model

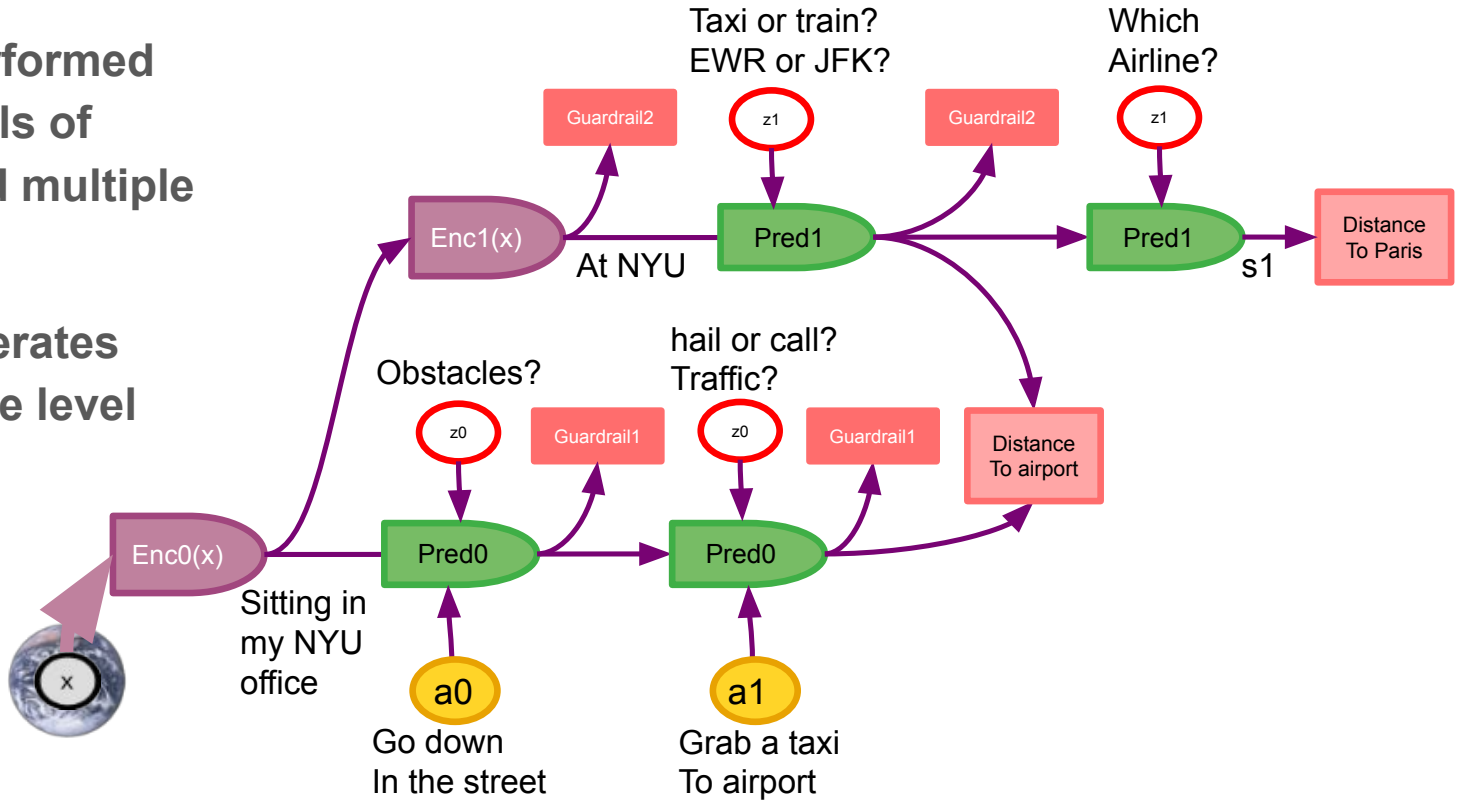
- World Model is applied auto-regressively **in representation space**
- Actions are searched to optimize/complete the task objective
- Guardrail objectives implement safety constraints



Hierarchical Planning with a Hierarchical World Model

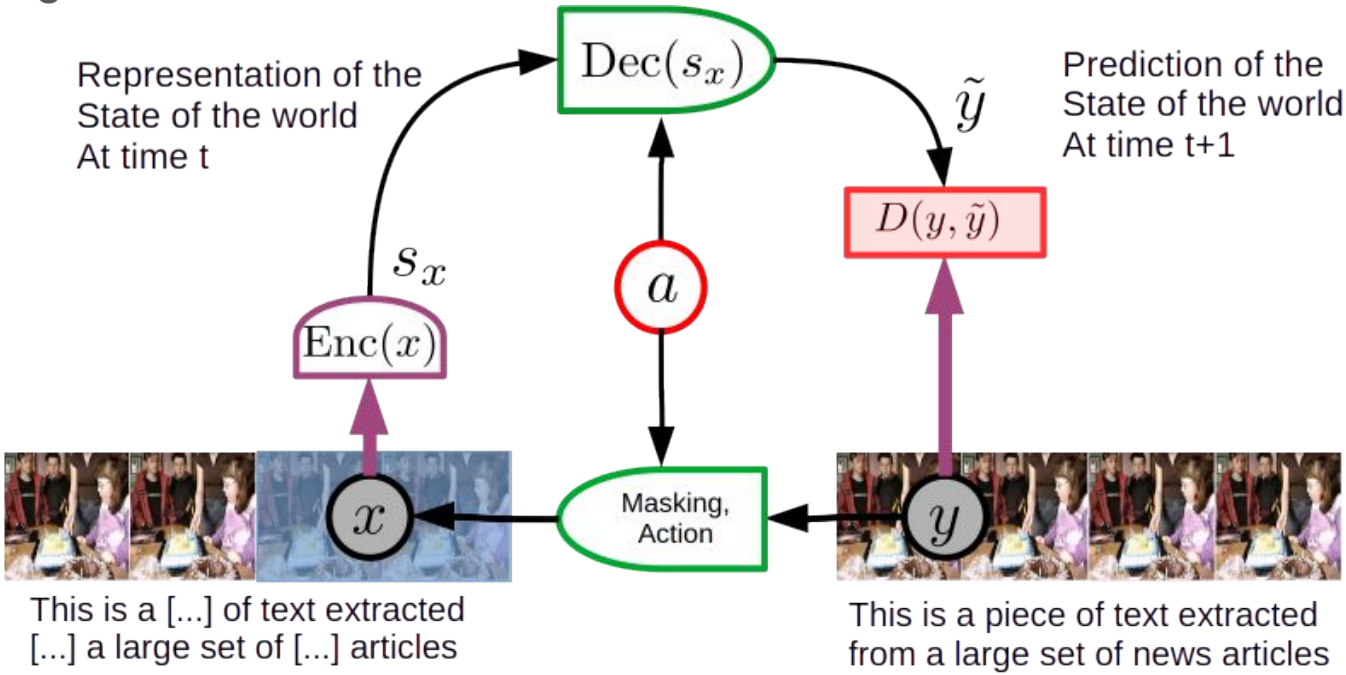
Predictions performed at multiple levels of abstraction and multiple time scales.

Each level generates subgoals for the level below



Self-Supervised Learning by Generative Prediction

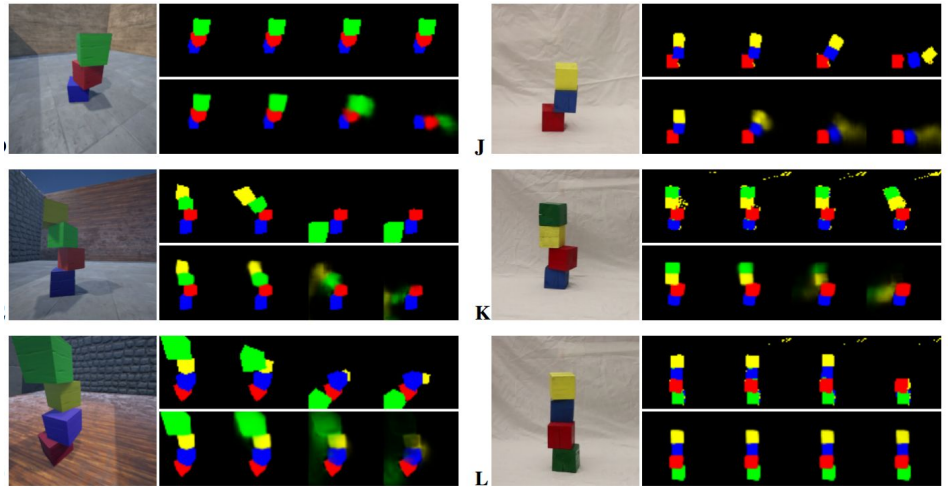
- **It works great for text and other discrete symbol sequences**
- **It does not work for high-dimensional, continuous, noisy data**
  - e.g. image, video, sensor data, scientific measurements….



Generative Prediction at the Pixel Level is **Blurry**

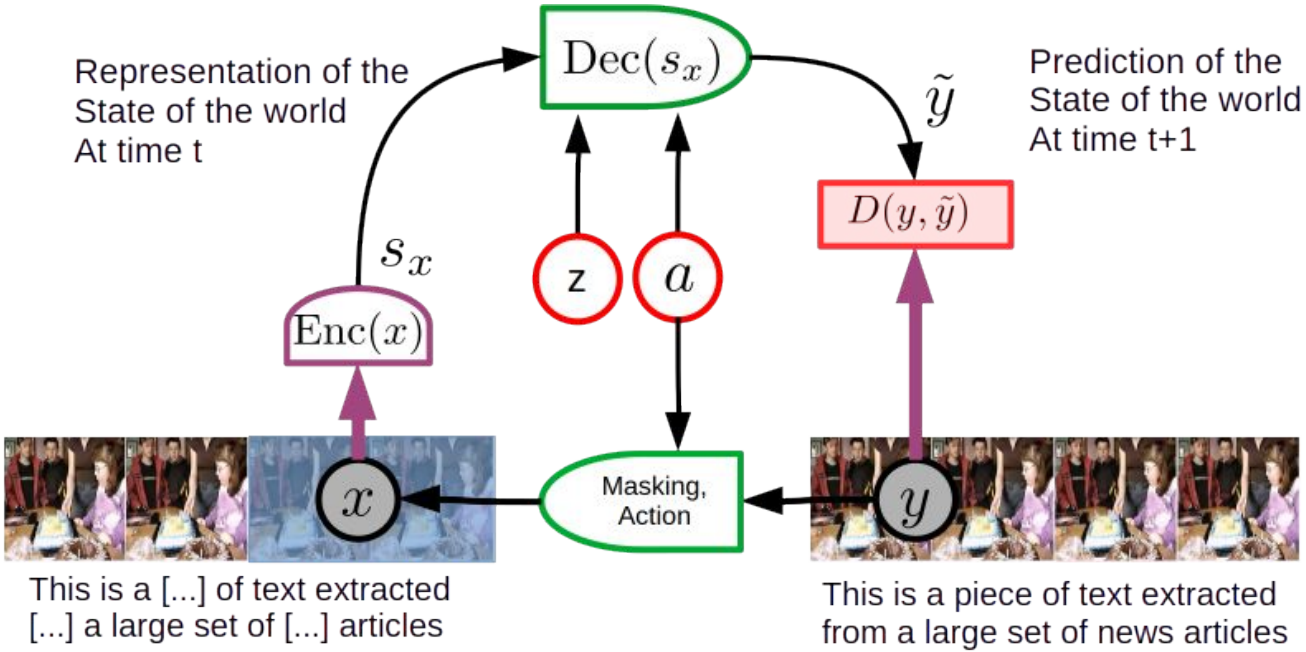
- **[Mathieu, Couprie, LeCun ICLR 2016]**
  - "Deep multi-scale video prediction beyond mean square error"

- **[Lerer, Gross, Fergus arxiv:1603.01312]**
  - "Learning Physical Intuition of Block Towers by Example"



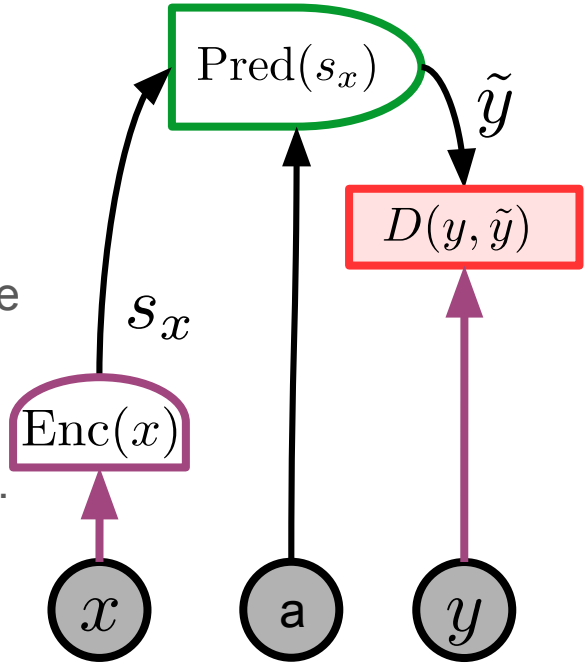
Video Prediction with Latent-Variable Generative Model

- Use a latent variable z to parameterize the space of predictions
- Variable z can be sampled (**VAE, diffusion models**) or inferred by optimization, but that requires regularization.

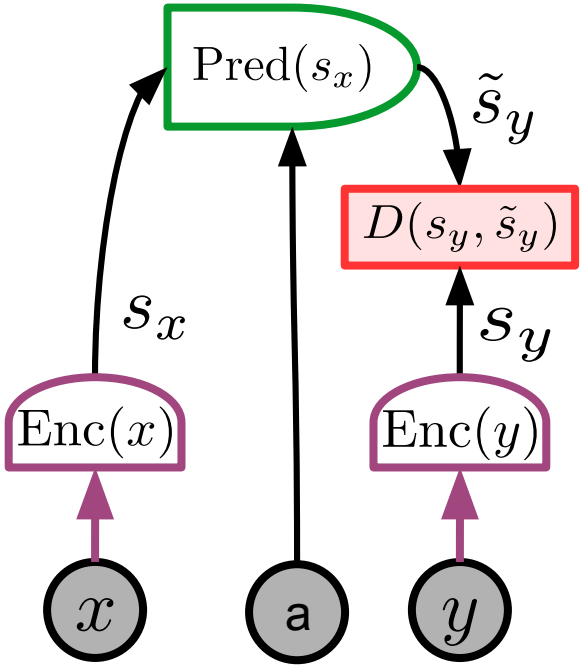


World Model Architectures: Generative vs JEPA

- JEPAs learn abstract representations of the data within which it makes predictions.
- JEPAs can ignore irrelevant and unpredictable details in the data (e.g. noise)
- Generative architectures must predict every detail in the data.
- **Token-based generative models simply do not work with high-dim, continuous, noisy data**



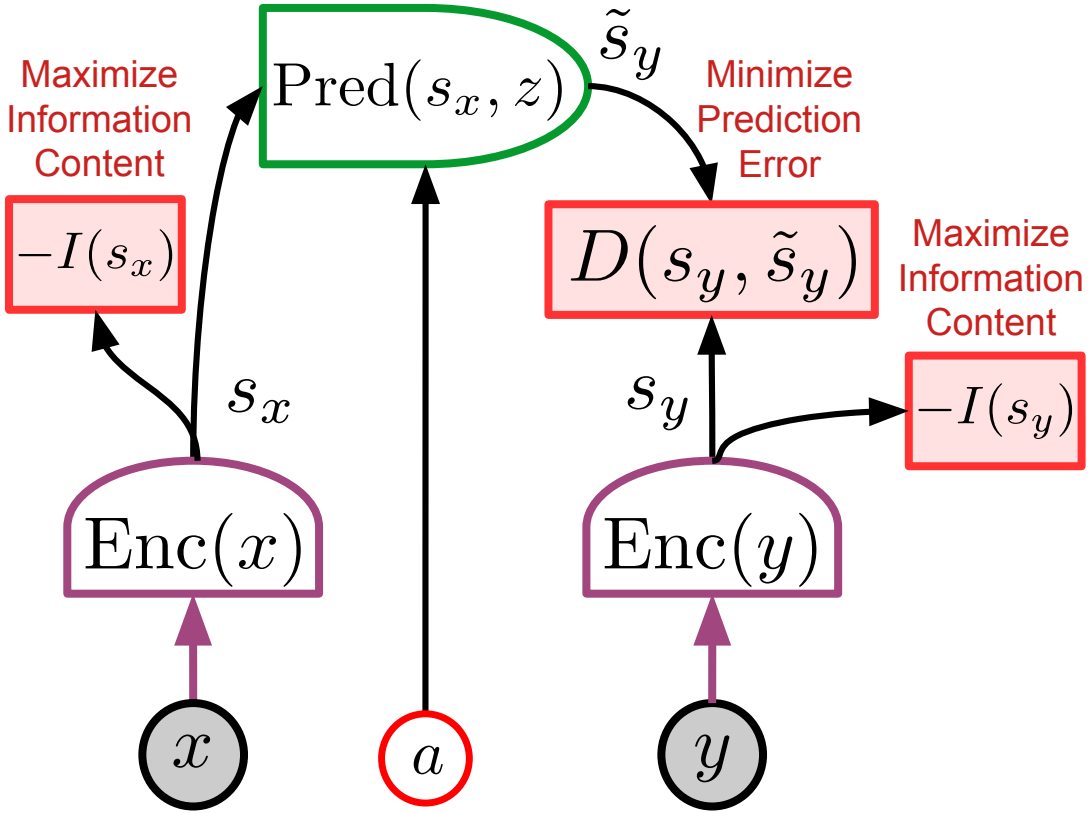
Generative Architectures (VAE, MAE, LLMs, Diffusion Models…..)



JEPA (Joint Embedding Predictive Architecture)

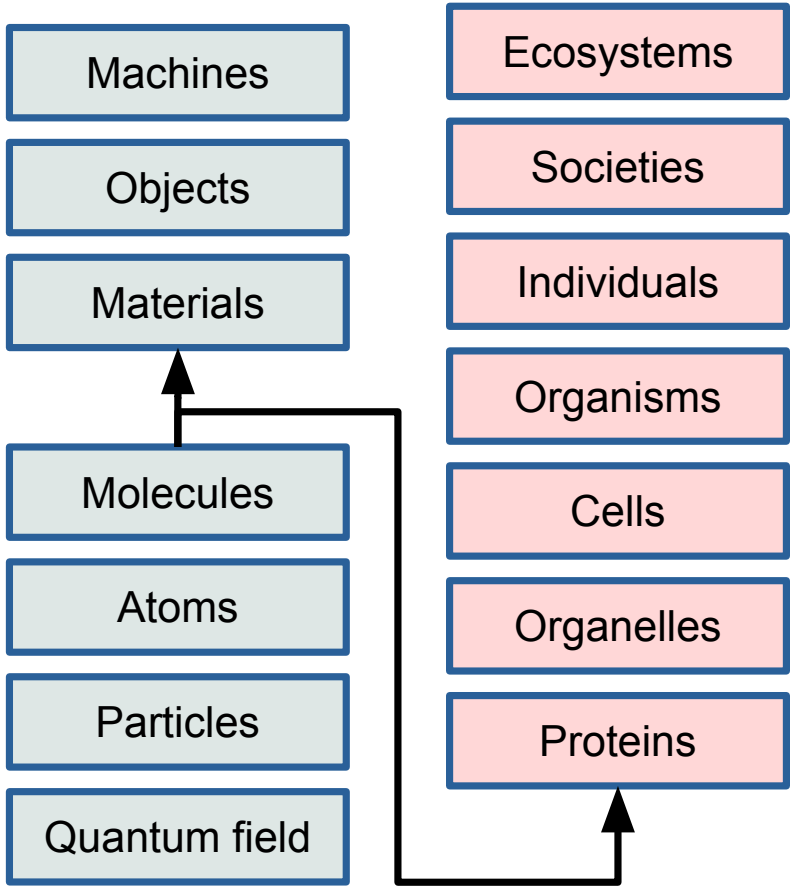
Training JEPA World Models: collapse prevention

- **Several methods to train a JEPA and prevent collapse**
- **Distillation**
  - DINO, I-JEPA, V-JEPA
    - from AMI Labs team members
- **Information maximization**
  - SIGReg, VCReg, Barlow Twins
    - from AMI Labs team members
  - MMCR, MCR2, W-MSE....
    - from other academic labs
- **A search for JEPA on G-Scholar returns over 1300 papers**
  - A growing worldwide research community



Learning a hierarchy of abstractions

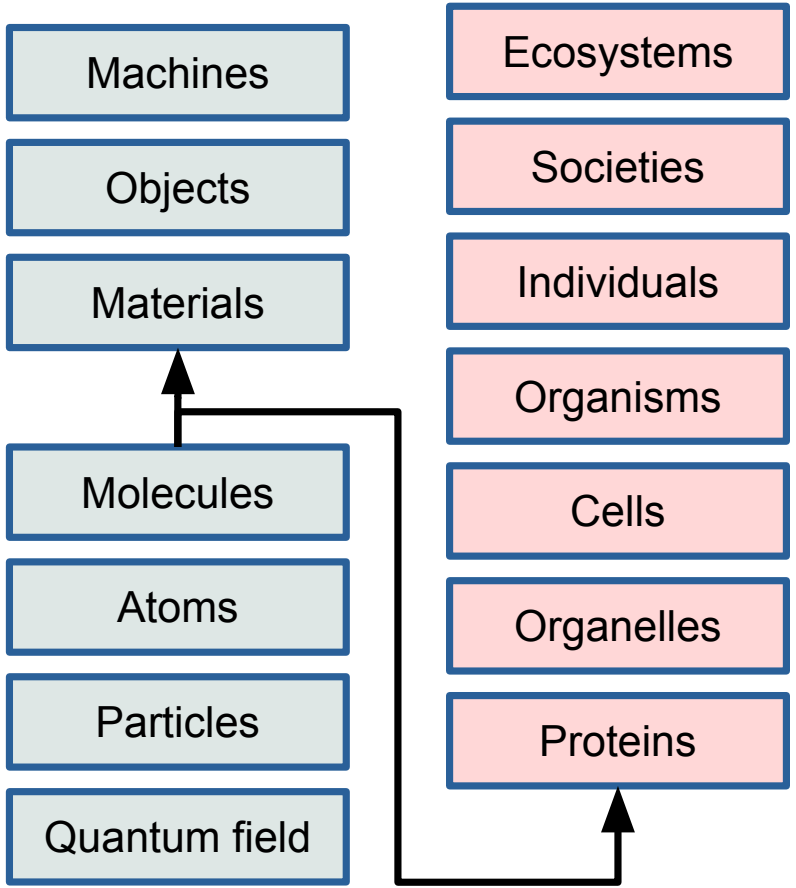
- **Lower levels make short-range (short-term) predictions.**
  - Preserve details.
  - Are inaccurate or computationally difficult for long-range predictions
- **Higher levels make longer-range (longer-term) predictions.**
  - Representations contain less details
  - Can make accurate long-term predictions, but with fewer details.
- **Science and modeling are all about finding the right representation at the right level of abstraction**



Learning a hierarchy of abstractions

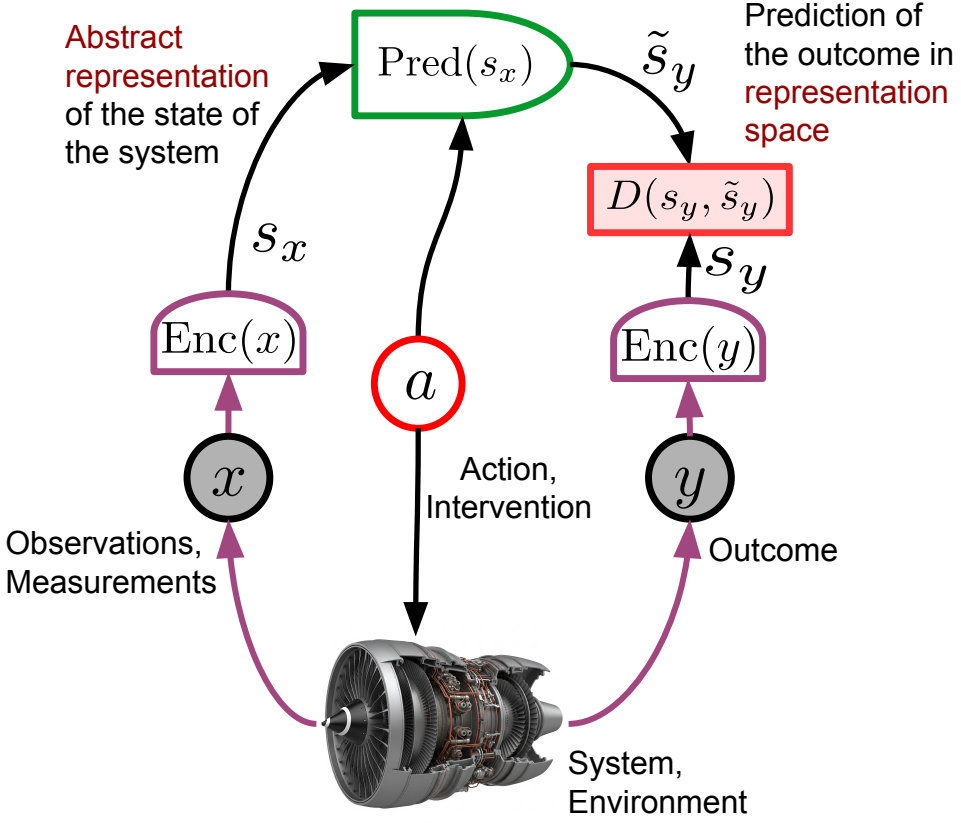
- **What World Models should not be:**
  - World simulators
  - Digital twins
  - Generative models
  - Video generation systems

- **What World Models should be:**
  - **Action-conditioned predictors in abstract representation space**
  - Preferably differentiable



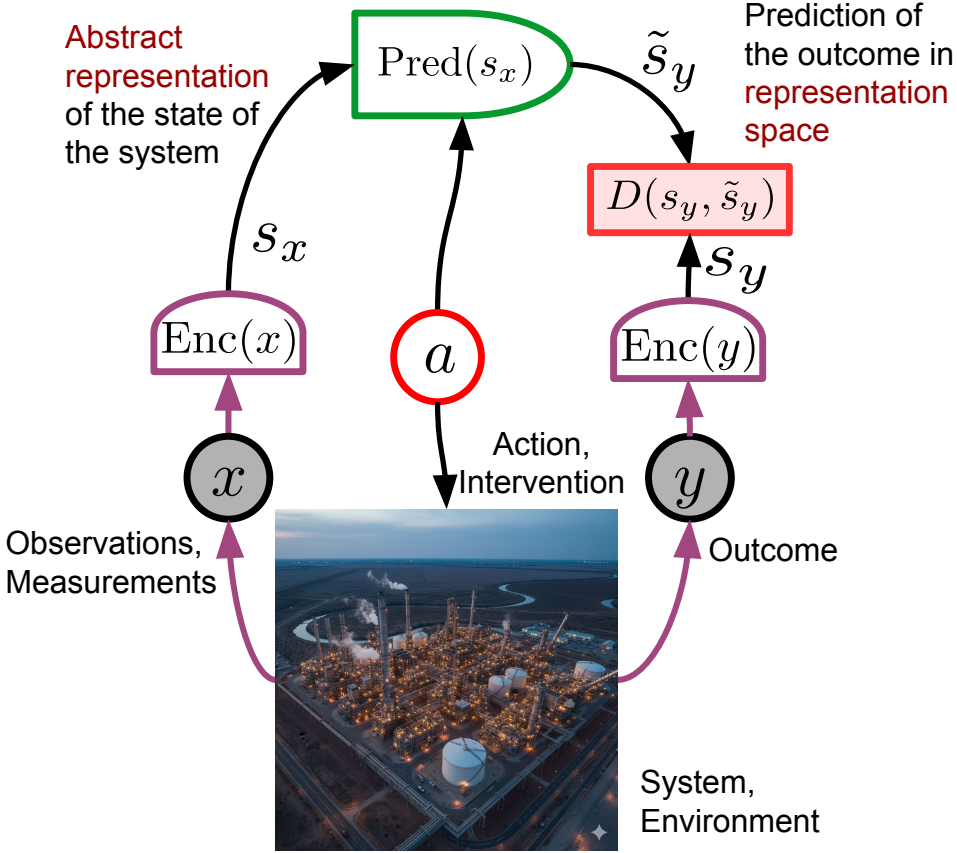
What is an (Action-Conditioned) World Model?

- A causal model that, given an **observation** of a system and an **intervention** on the system, predicts the **outcome** of the intervention.
- The model **does not predict every detail** of the outcome but performs predictions in an **abstract representation space**



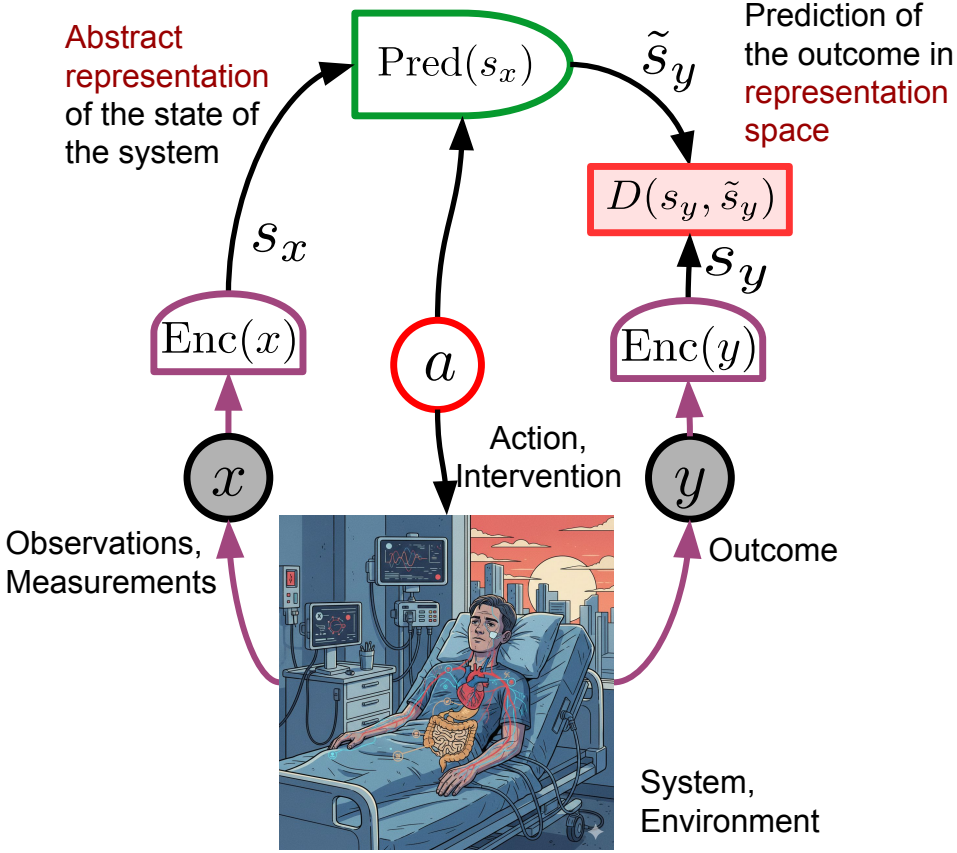
What is an (Action-Conditioned) World Model?

- A causal model that, given an **observation** of a system and an **intervention** on the system, predicts the **outcome** of the intervention.
- The model **does not predict every detail** of the outcome but performs predictions in an **abstract representation space**



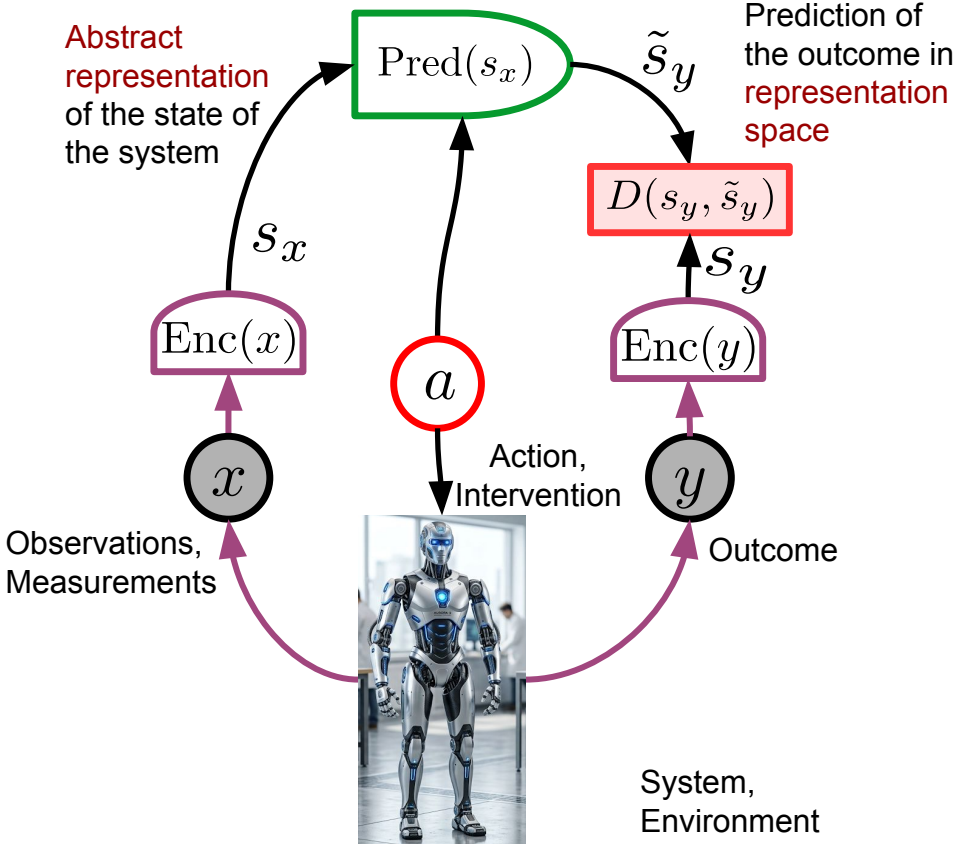
What is an (Action-Conditioned) World Model?

- A causal model that, given an **observation** of a system and an **intervention** on the system, predicts the **outcome** of the intervention.
- The model **does not predict every detail** of the outcome but performs predictions in an **abstract representation space**



What is an (Action-Conditioned) World Model?

- A causal model that, given an **observation** of a system and an **intervention** on the system, predicts the **outcome** of the intervention.
- The model **does not predict every detail** of the outcome but performs predictions in an **abstract representation space**



World Models are an old idea in optimal control

- **Kelley-Bryson Gradient Procedure (1962)**
- **Planning as a Lagrangian optimization problem**
  - Pontryagin optimality principle - Adjoint state method
- **Lagrangian formulation of backprop**
  - “A theoretical framework for back-propagation” [LeCun 1988]

[Bryson & Ho, 1975] Applied Optimal Control

Multistage systems; no terminal constraints, fixed number of stages

Optimal programming problems for multistage systems are also parameter optimization problems. Consider the multistage system described by the nonlinear difference equations:

$$x(i+1) = f^i[x(i), u(i)]; \quad x(0) \text{ given}, \quad i = 0, \ldots, N-1, \quad (2.2.1)$$

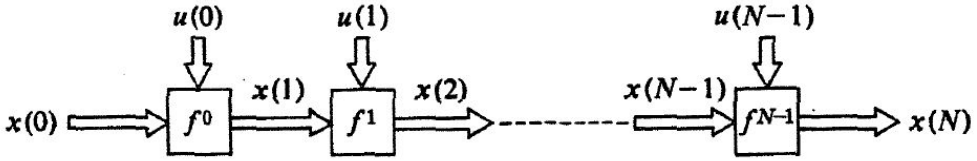


Figure 2.2.1. Flow chart for a multistage system.

$$\bar{J} = \phi[x(N)] + \sum_{i=0}^{N-1} [L^i[x(i), u(i)] + \lambda^T(i+1)\{f^i[x(i), u(i)] - x(i+1)\}]. \quad (2.2.3)$$

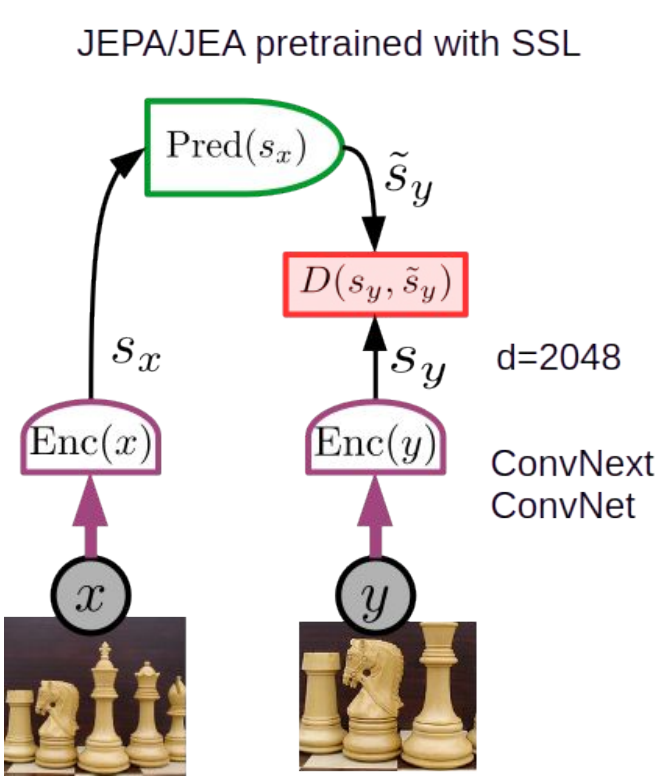
Objective-Driven AI Architecture

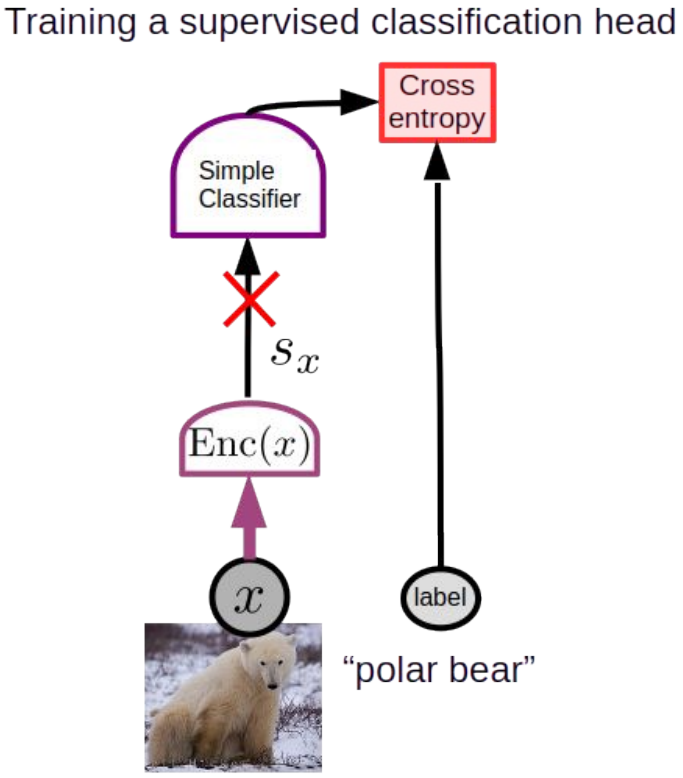
- **A path towards autonomous machine intelligence**
  - https://bit.ly/PAth2AMI
  - https://openreview.net/forum?id=BZ5a1r-kVsf

- **Videos of longer technical talks:**
  - https://bit.ly/YannLeCunTalks

Training JEPA For Image Understanding

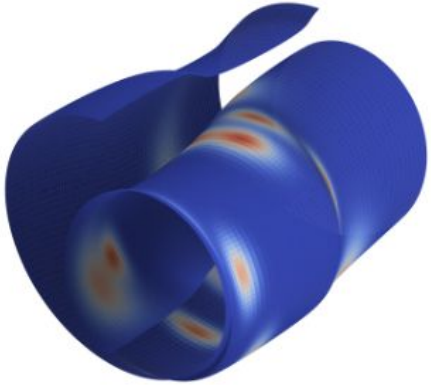
- **JEPA Self-Supervised Learning through multiview invariance**



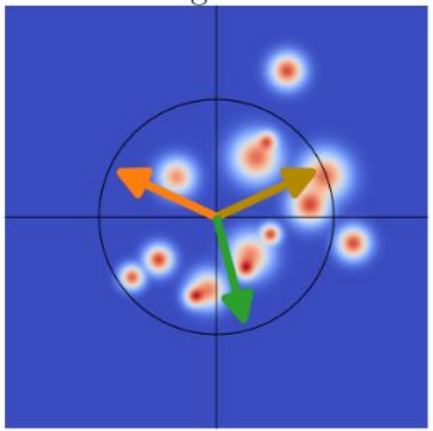


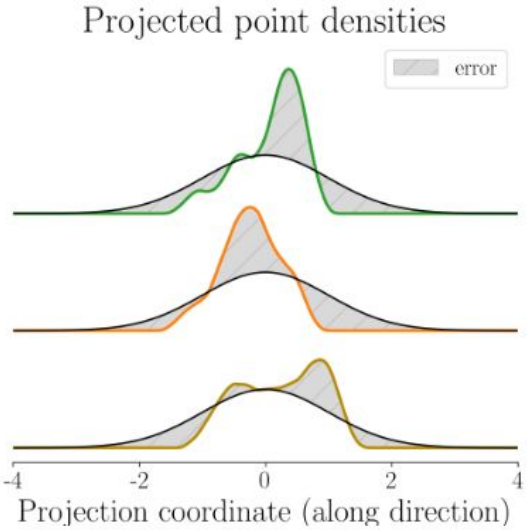
LeJEPA / SIGReg [Balestriero, LeCun arXiv:2511.08544]

- **Sketched Isotropic Gaussian Regularization: maximizing information content by making the distribution isotropic Gaussian**
- **Using multiple projections and making the univariate marginals Gaussian**



$f_\theta \rightarrow$





LeJEPA / SIGReg: Isotropic Gaussianization

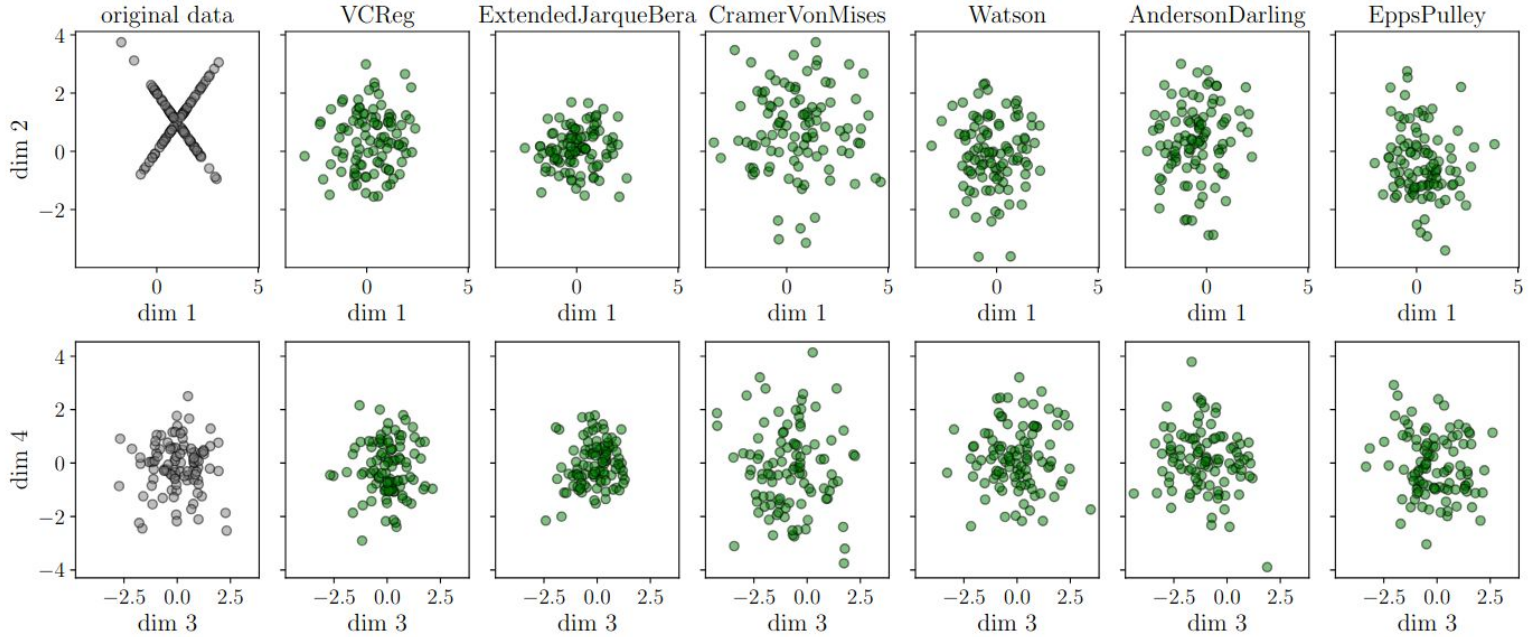


Figure 6. $N = 100$ samples are drawn from a 1024-dimensional standard Gaussian, and the first 2 coordinates are altered to produce the "X" distribution from Figure 5 (**left-most column**). For each statistic (**all other columns**), we perform gradient descent on the samples to minimize their value, at each iteration step with sample $M = 10$ random directions to evaluate SIGReg (recall def. 2). We obtain that albeit this is a high-dimensional distribution with limited number of samples, SIGReg is able to capture the degenerate subspace and adapt the data accordingly to match an isotropic Gaussian distribution. Additional figures with varying dimensions and number of 1d projections are provided in Figure 16.

LeWorldModel: JEPA WM trained with SIGReg

LeWorldModel: Stable End-to-End Joint-Embedding Predictive Architecture from Pixels

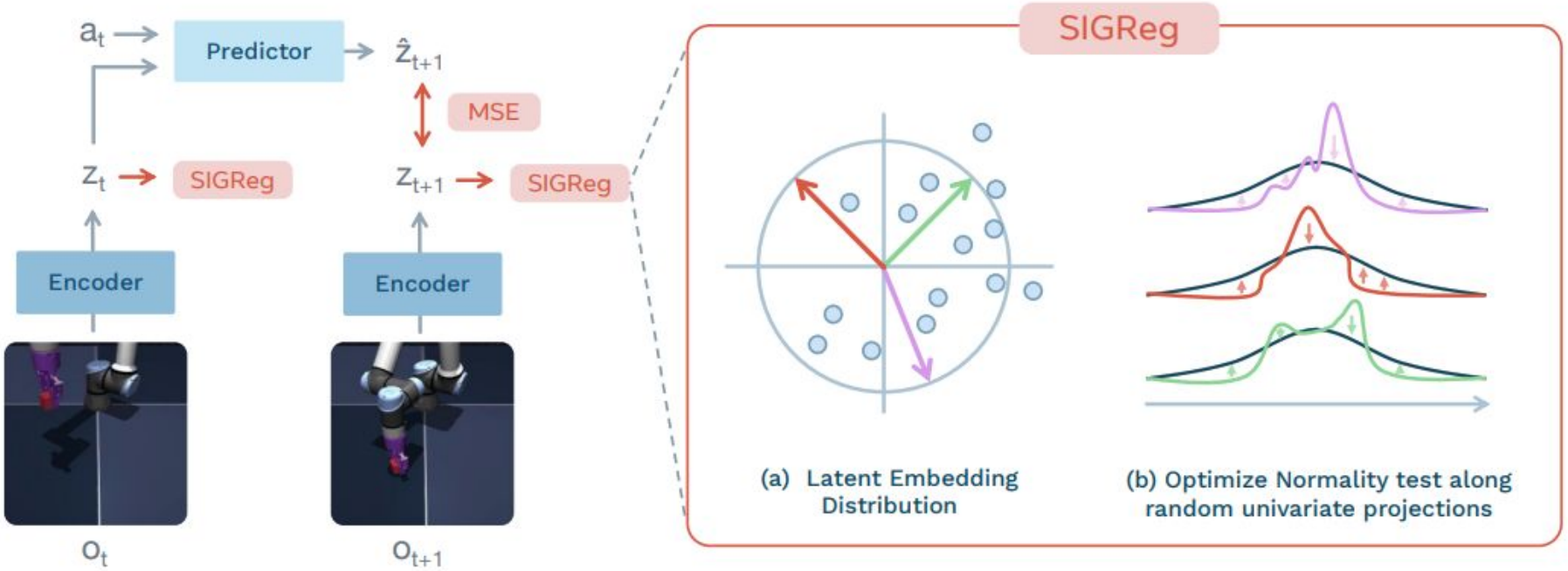
Lucas Maes*[1] Quentin Le Lidec*[2] Damien Scieur[1,3] Yann LeCun[2] Randall Balestriero[4]

[1]Mila & Université de Montréal [2]New York University [3]Samsung SAIL [4]Brown University

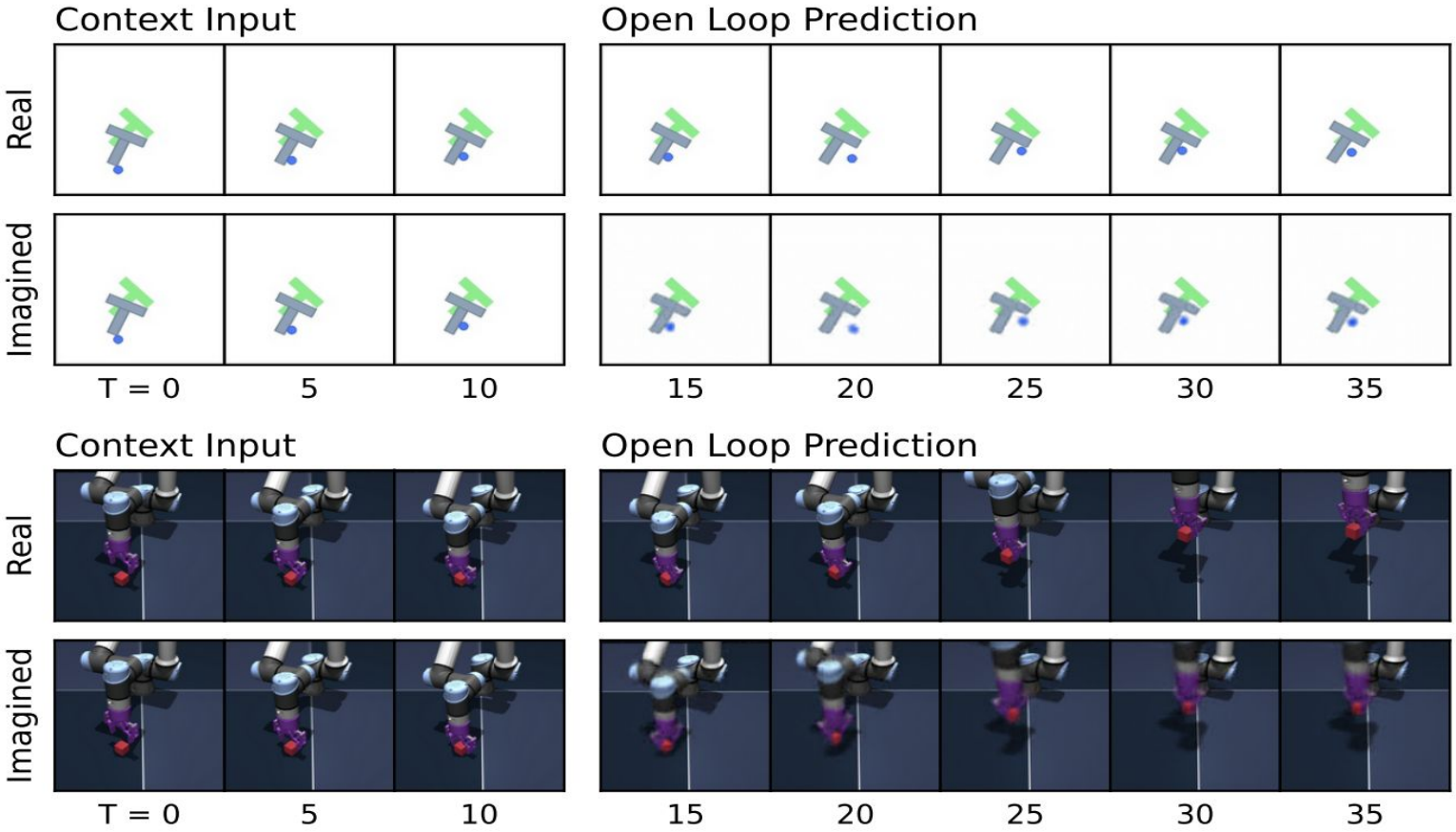
https://le-wm.github.io/

ArXiv:2603.19312

LeWorldModel: JEPA WM trained with SIGReg



LeWorldModel: JEPA WM trained with SIGReg



LeWorldModel: Planning results

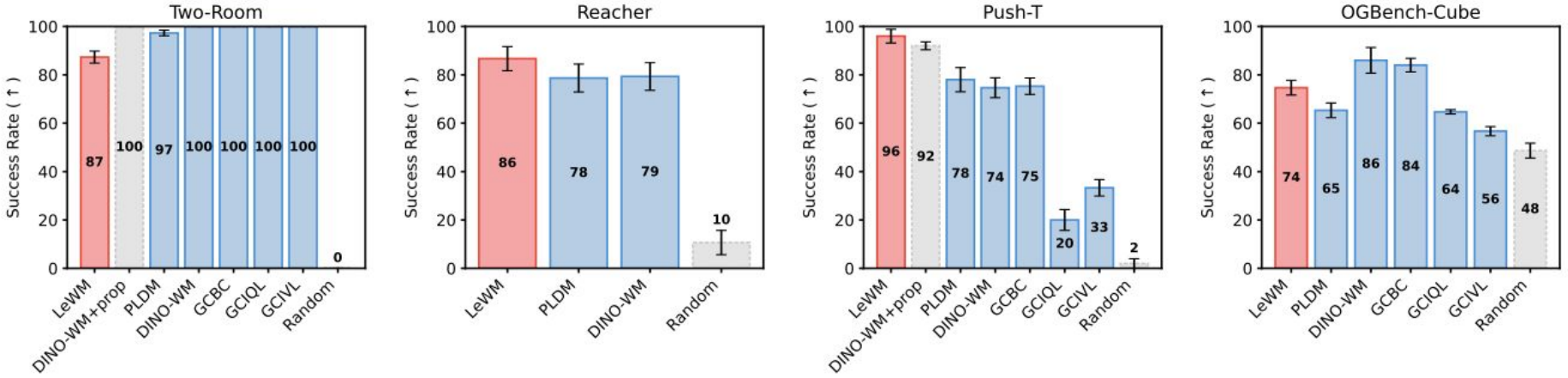
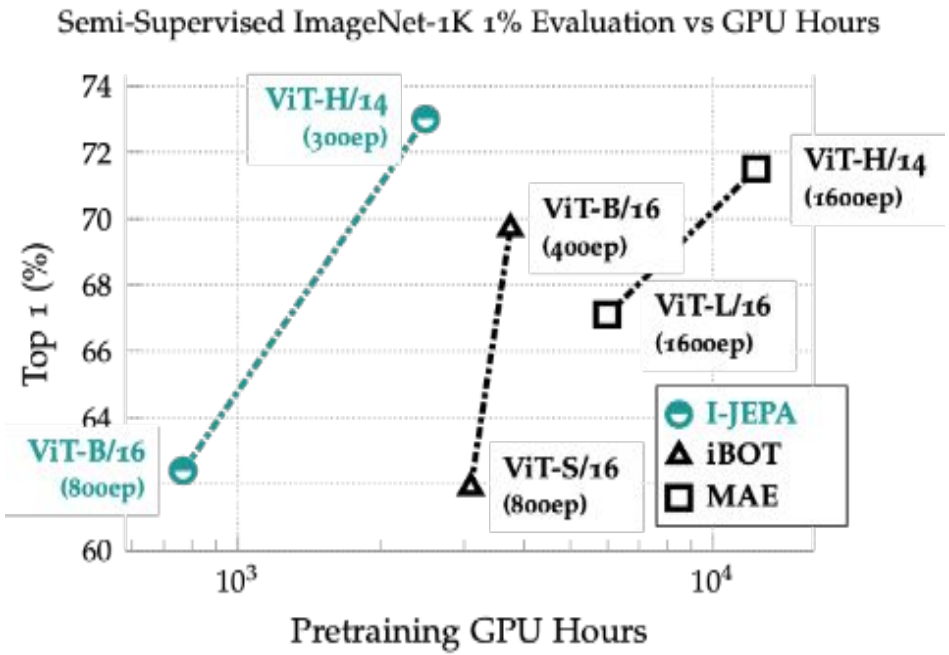
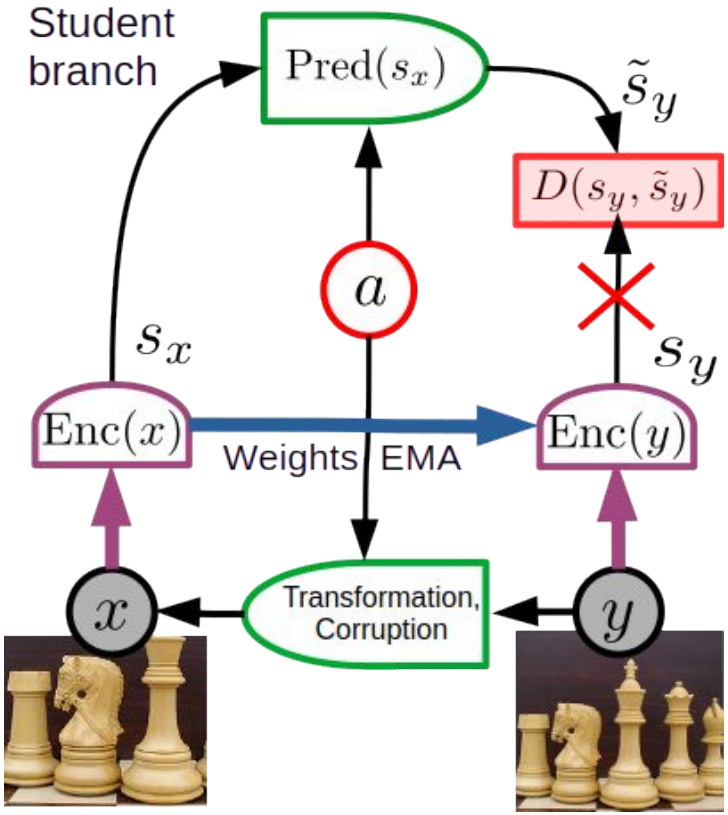


Figure 6: **Planning performance across environments.** Results are shown for *Two-Room* (left), *Reacher* (center 1), *PushT* (center-2) and *OGBench-Cube* (right). LeWM consistently outperforms PLDM and DINO-WM on Push-T and Reacher. On OGBench-Cube, DINO-WM slightly outperforms LeWM, possibly due to the higher visual complexity and the 3D nature of the environment, which makes encoder training more challenging. In the simpler Two-Room environment, PLDM and DINO-WM outperform LeWM, which may be explained by the SIGReg regularization encouraging a Gaussian distribution in a high-dimensional latent space, while the intrinsic dimensionality of the environment is much lower.

I-JEPA: Distillation for image representation learning

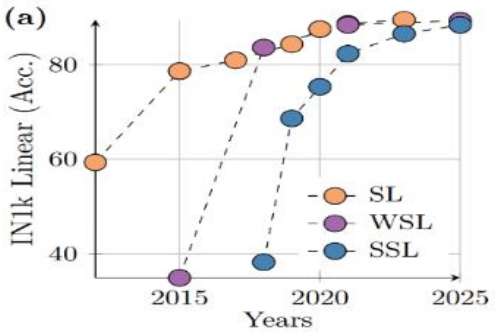
- **I-JEPA** https://ai.meta.com/blog/yann-lecun-ai-model-i-jepa/
  - [Assran et al CVPR 2023 arxiv:2301.08243]

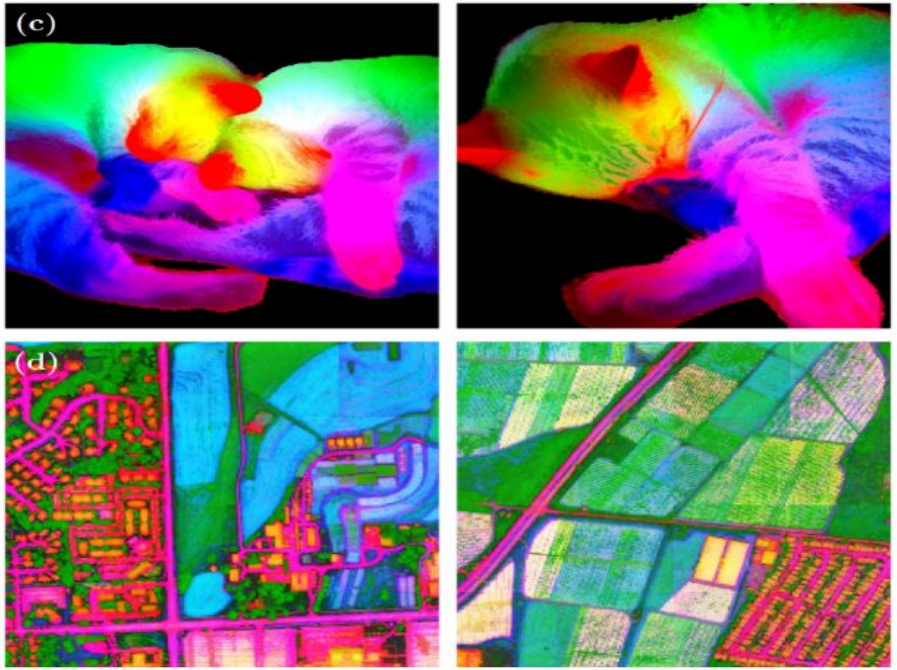




DINOv3 Distillation for image representation learning

- **DINOv3** [ArXiv:2508.10104] https://ai.meta.com/dinov3/
- Best generic image representation system in the world. Training is entirely self-supervised.
- Used by hundreds of groups around the world (Github: 8.6k stars, 600 forks)

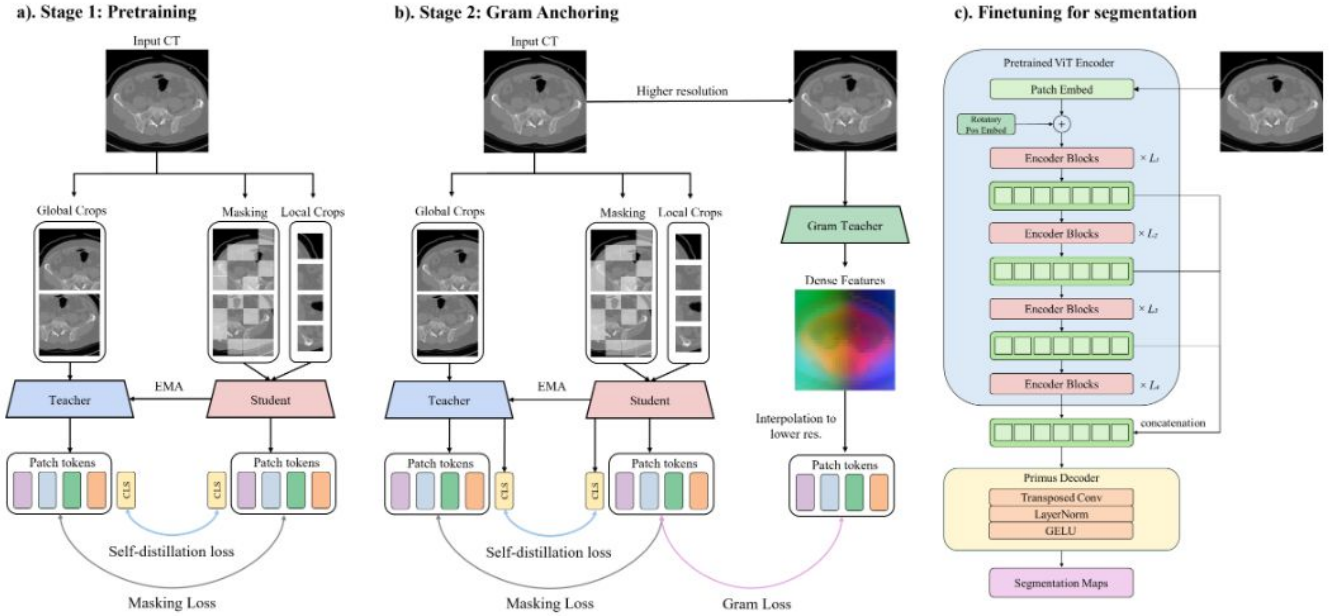


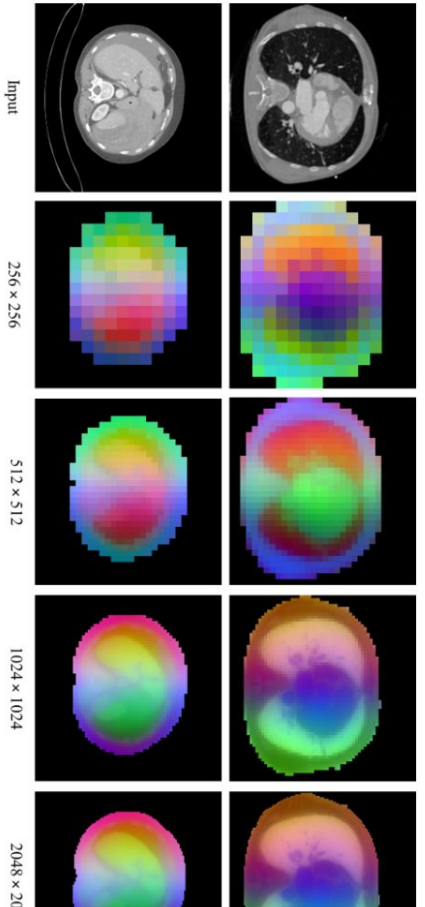


DINOv3 in Medical Imaging

MedDINOv3: How to adapt vision foundation models for medical image segmentation?

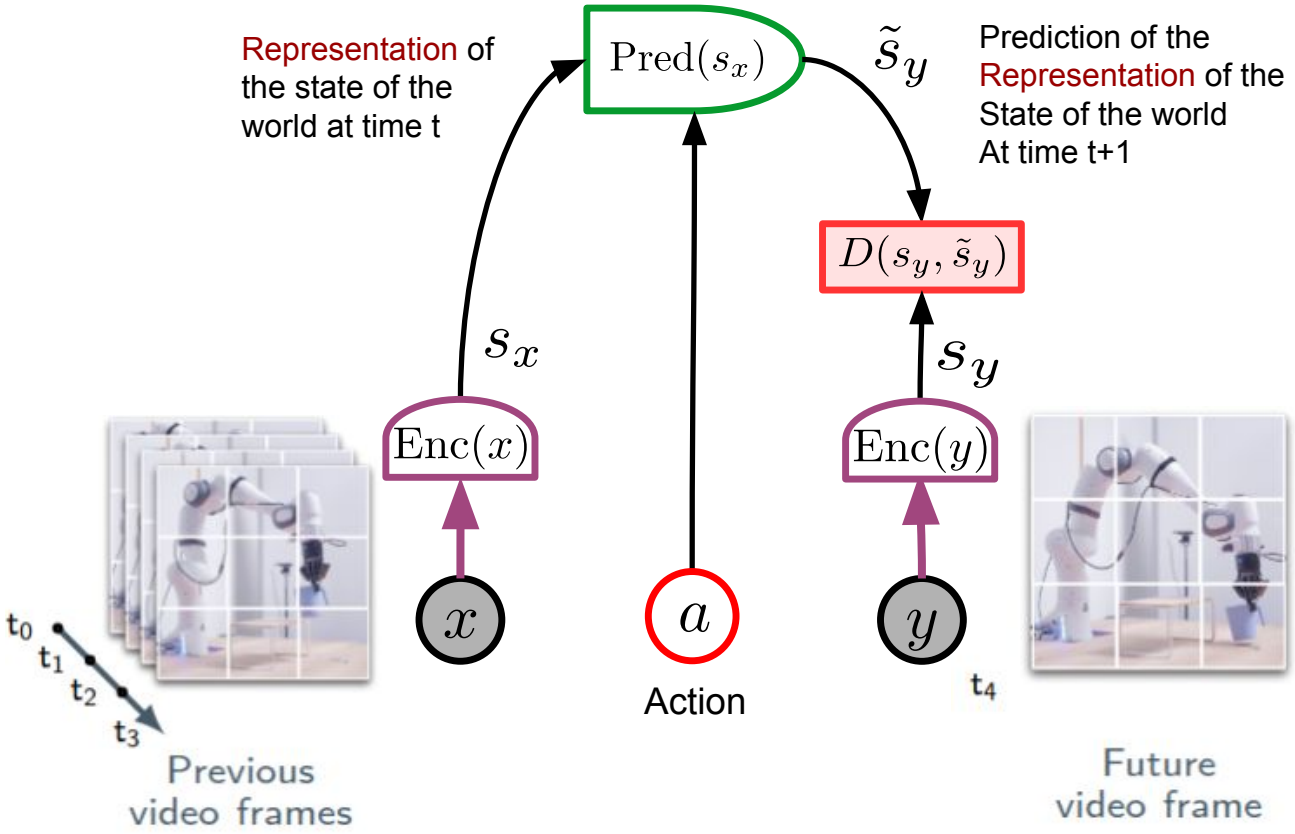
Yuheng Li[1], Yizhou Wu[2], Yuxiang Lai[3], Mingzhe Hu[3], Xiaofeng Yang[1,3,4,*]
[1]Department of Biomedical Engineering, Georgia Institute of Technology,





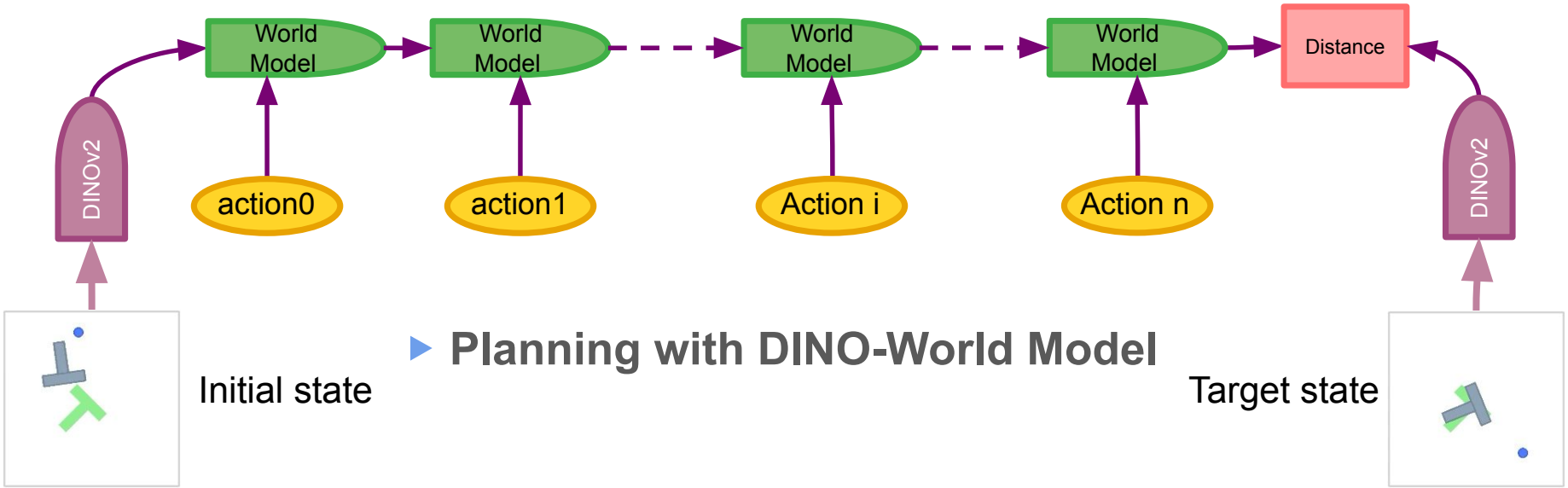
JEPA for Action-Conditioned World Model

- JEPA world models predict in abstract representation space.
- Action-conditioned world models are causal models of the environment.



Robot Planning with DINO-World Model

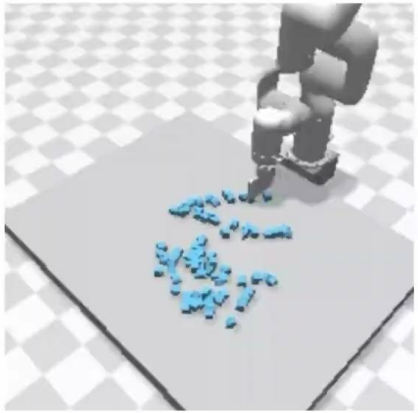
- **DINO-WM** [Zhou 2024 ArXiv:2411.04983] https://dino-wm.github.io/
  - World model for planning. Based on DINO-v2
- **DINO-World** [Baldassare 2025 ArXiv:2507.19468]
  - World model for planning. Based on DINO-v3



Results: Robot Planning with DINO-World Model

- **DINO-WM** [Zhou 2024 ArXiv:2411.04983] https://dino-wm.github.io/

Initial State

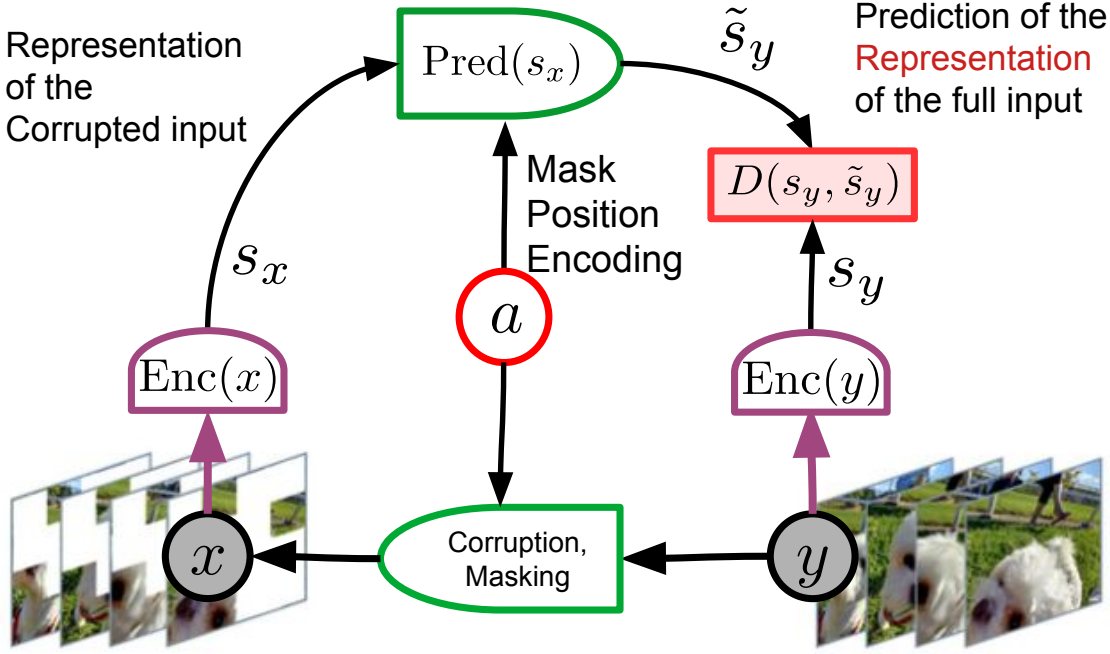


V-JEPA for video understanding and planning

- **V-JEPA 2** [Assran et al. 2025 ArXiv:2506.09985] https://ai.meta.com/vjepa/
- Best video representation learning system in the world. Entirely self-supervised

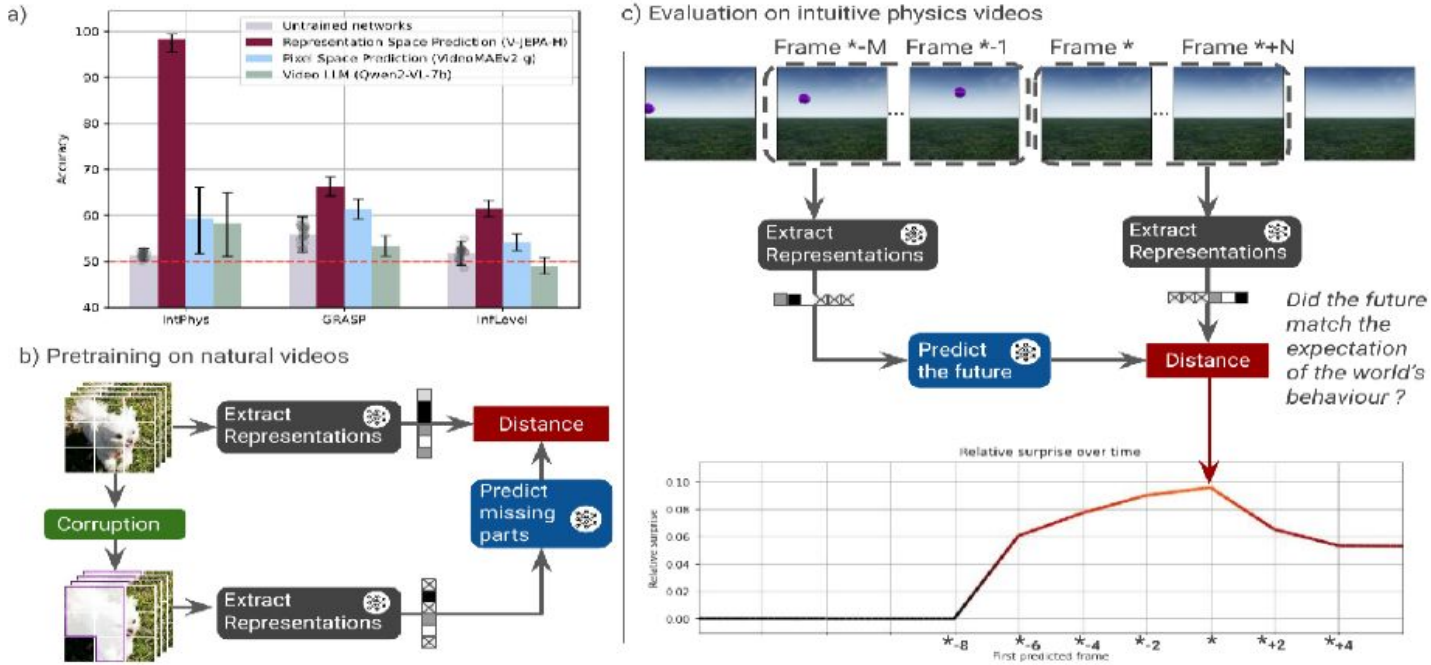
- **Trained with unlabeled data**
  - Video: 1M hours
  - Images: 1M images

Benchmark	VJEPA 2	Previous Best
EK100	39.7%	27.6% (PlausiVL)
SSv2 (Probe)	77.3%	69.7% (InternVideo2-1B)
Diving48 (Probe)	90.2%	86.4% (InternVideo2-1B)
MVP (Video QA)	44.5%	39.9% (InternVL-2.5)
TempCompass (Video QA)	76.9%	75.3% (Tarsier 2)



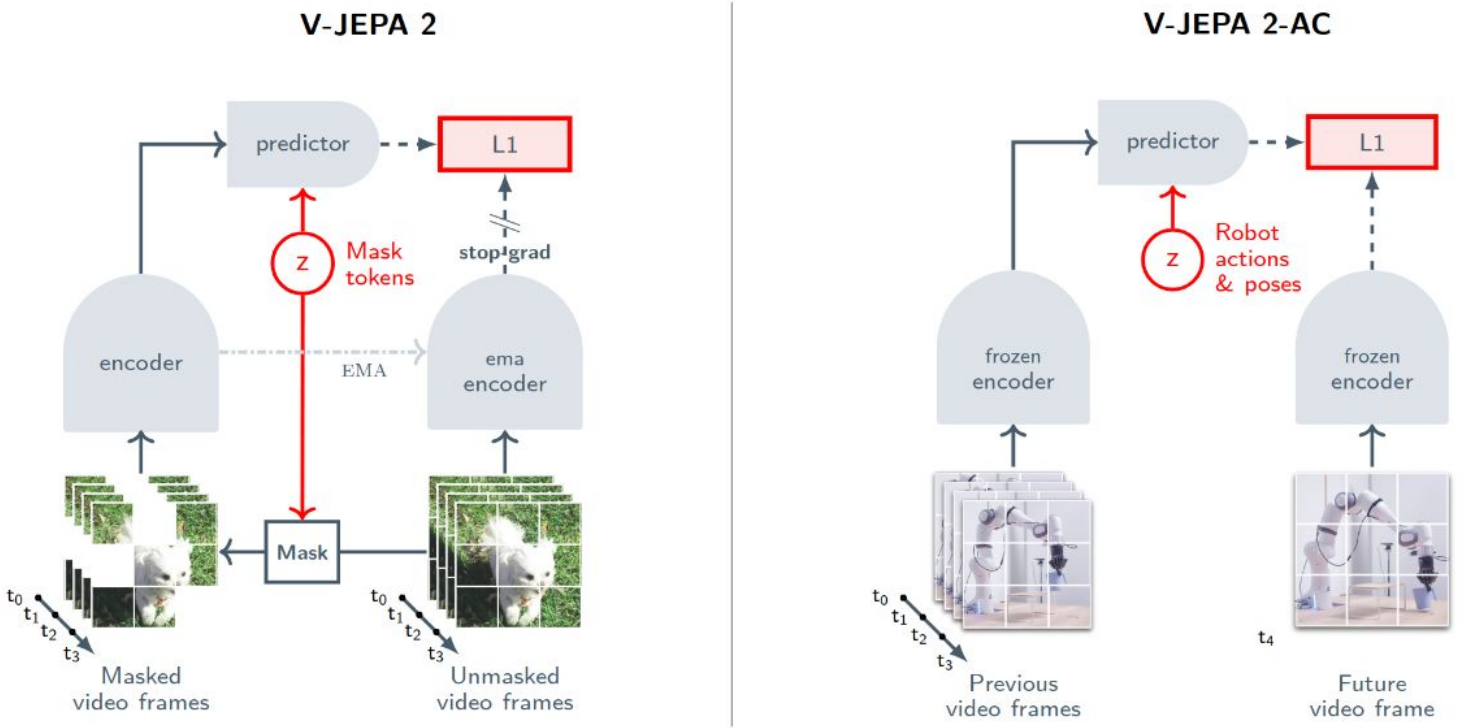
V-JEPA has learned some level of common sense

- [Garrido et al. ArXiv:2502.11831]
  - "Intuitive physics understanding emerges from self-supervised pretraining on natural videos"



JEPA for video understanding and planning

- **V-JEPA 2** [Assran et al. 2025 ArXiv:2506.09985] https://ai.meta.com/vjepa/
- Best video representation learning system in the world. Entirely self-supervised



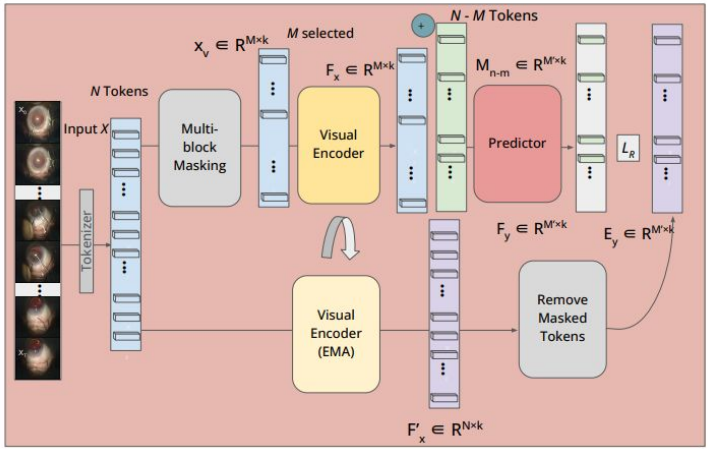
Robot Planning with V-JEPA 2 - Action Conditioned

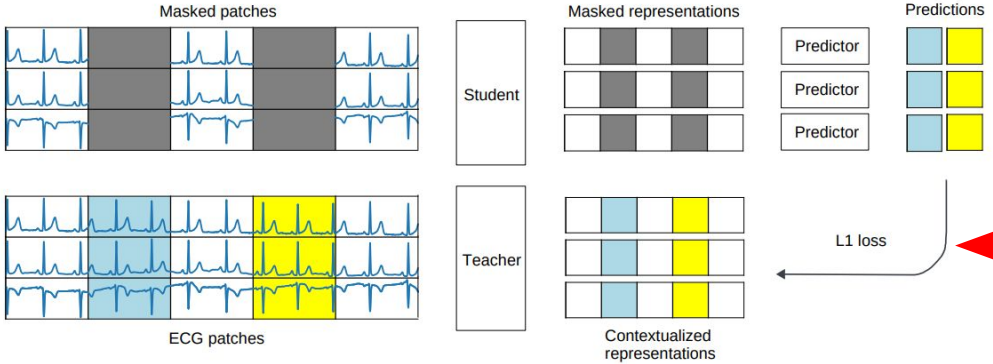
- **V-JEPA 2** [Assran et al. 2025 ArXiv:2506.09985] https://ai.meta.com/vjepa/

Applications JEPA to biology and medicine

A Vision Foundation Model for Cataract Surgery Using Joint-Embedding Predictive Architecture

Nisarg A. Shah[*1]
[1] *Johns Hopkins University, Baltimore, USA*
SNISARG812@GMAIL.COM

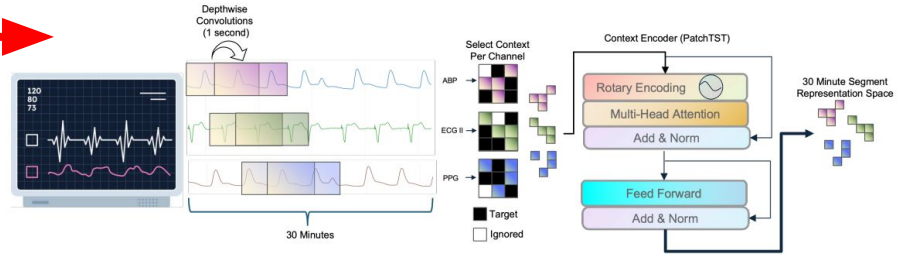




Learning General Representation of 12-Lead Electrocardiogram With a Joint-Embedding Predictive Architecture

Signals for Real Time Risk Estimation in the Intensive Care Unit

Benjamin Fox BEN.FOX@ICAHN.MSSM.EDU
Dung Hoang JOLIE.HOANG@ICAHN.MSSM.EDU
Joy Jiang JOY.JIANG@ICAHN.MSSM.EDU
Pushkala Jayaraman PUSHKALA.JAYARAMAN@MSSM.EDU
Ankit Parekh ANKIT.PAREKH@MSSM.EDU
Girish N. Nadkarni GIRISH.NADKARNI@MOUNTSINAI.ORG
Ankit Sakhuja ANKIT.SAKHUJA@MSSM.EDU
The Windreich Department of Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA



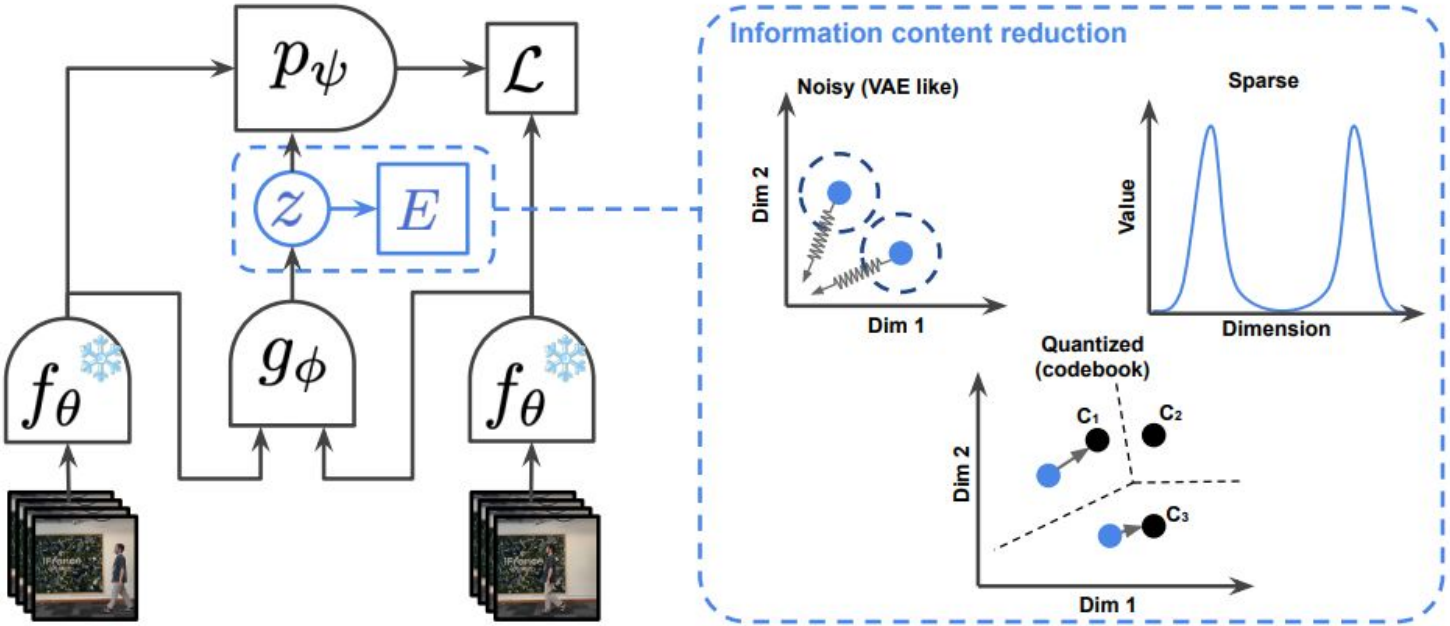
Inferring hidden actions from video

ArXiv:2601.05230

Learning Latent Action World Models In The Wild

Quentin Garrido[1], Tushar Nagarajan[1], Basile Terver[1,2], Nicolas Ballas[1], Yann LeCun[1,3], Michael Rabbat[1]

[1]FAIR at Meta, [2]Inria, [3]NYU



Recommendations to my fellow AI scientists

- **Abandon generative models**
  - in favor joint-embedding architectures
- **Abandon probabilistic model**
  - in favor of energy-based models
- **Abandon contrastive methods**
  - in favor of regularized methods
- **Abandon Reinforcement Learning**
  - In favor of model-predictive control
  - Use RL only when planning doesn't yield the predicted outcome, to adjust the world model or the critic.
- **IF YOU ARE INTERESTED IN HUMAN-LEVEL AI, DON'T WORK ON LLMs**

AMI Labs: enabling the next AI revolution

In the long run, AMI Labs aims to become the main provider of intelligent systems

Applications abound in all the major sectors of the economy

physical sciences, biomedical sciences,

Healthcare, manufacturing, automotive, aerospace, defense, transportation,

logistics, pharmaceuticals, technology, telecommunication, finance, etc…

Examples: a modern turbofan engine has 1000+ sensors. An Airbus A380 has a total of 25,000 sensors, measured 5000 of times per second.

Hierarchical JEPA World Models can produce causal predictive models of such complex systems from data.

Outlook

- **Could hierarchical JEPA provide us with a way to build universal causal models of any complex system?**
  - Including human cells, organs, and bodies?