

Architecture for the world model: JEPA

JEPA: Joint Embedding Predictive Architecture.

▶ x : observed past and present

▶ y : future

▶ a : action

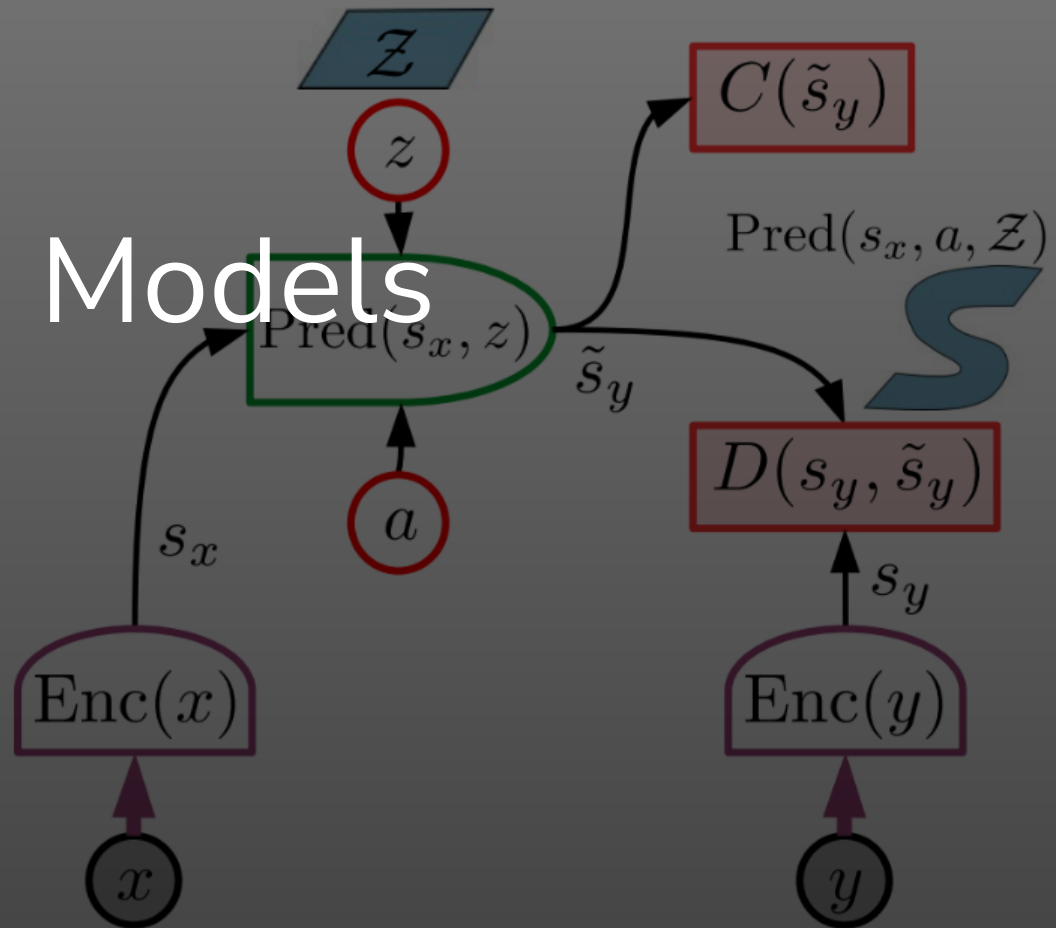
▶ Machine Learning Singapore

▪ Martin Andrews = WorldActionModels @ mdda.net

▶ $D(\cdot)$: prediction cost

▶ 29-April-2026
▶ $C(\cdot)$: surrogate cost

▶ JEPA predicts a representation of the future S_y from a representation of the past and present S_x



About Me

- Machine Intelligence / Startups / Finance
 - Moved from NYC to Singapore in Sep-2013
- 2014 = 'fun' :
 - Machine Learning, Deep Learning, NLP
 - Robots, drones
- Since 2015 = 'serious' :: NLP + deep learning
 - Including Papers...
 - & GDE ML; ML-Singapore co-organiser...
 - & Red Dragon AI...



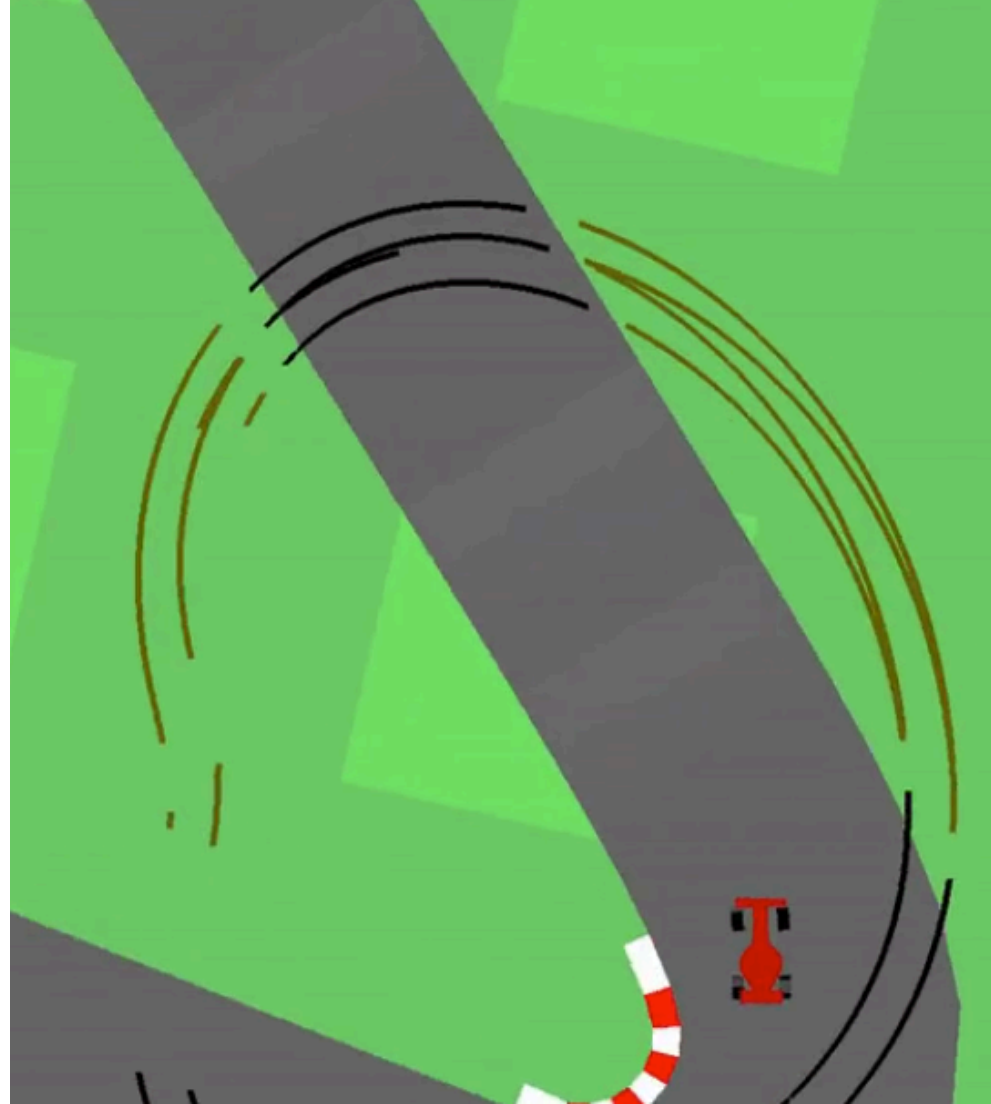
Outline

- What are World Models?
- Three key 'strands'
 - Video / Action modeling
 - JEPA models
 - Robot action "Foundation" models
- Aside: "Undercover Agent"
- Wrap-up & QR-code

What are World Models?

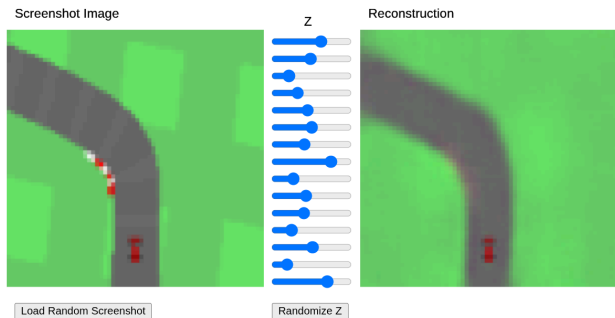
The Original "World Model"

- World Models - Ha & Schmidhuber (2018)
 - David Ha (Sakana AI in Tokyo)
 - Jürgen Schmidhuber (needs no introduction)
- Idea:
 - Learn a representation of the world (VAE)
 - depends on pixels
 - Learn to predict the next representation
 - may include actions taken
 - Then : Learn to act to maximise rewards
 - can learn entirely in 'fake environment'

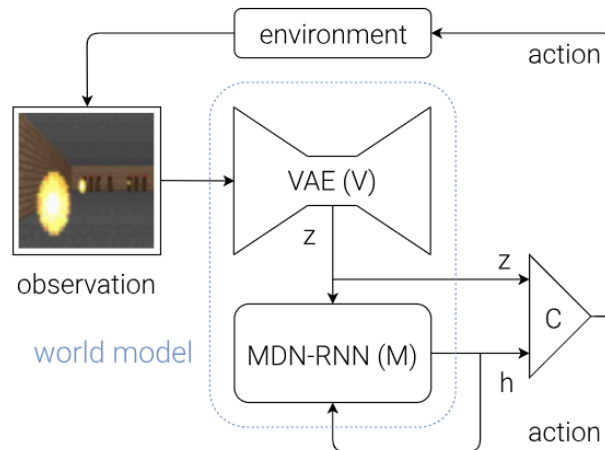


"World Models" : Interactive Project Page

- VAE reconstruction:



- Reinforcement Learning loop:



Video / Action modeling



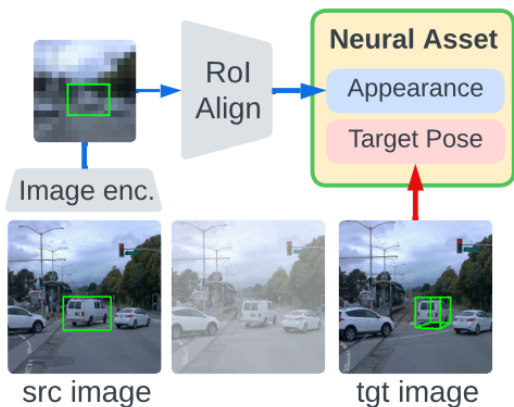
Video Models

- Key examples :
 - Sora-{1, 2}
 - VEO-{1, 2, 3, 3.1, 3.2}
 - WAN 2.2 / ++
- Do these models have a model of the world?
 - they clearly 'understand' some things...

Neural Assets (with Demo)

- Neural Assets: 3D-Aware Multi-Object Scene Synthesis with Image Diffusion Models - Wu *et al* (2024)
- Idea:
 - Use image descriptions on two frames of video
 - Then train model to *transform* one image into the other
 - we already know what the end result should look like

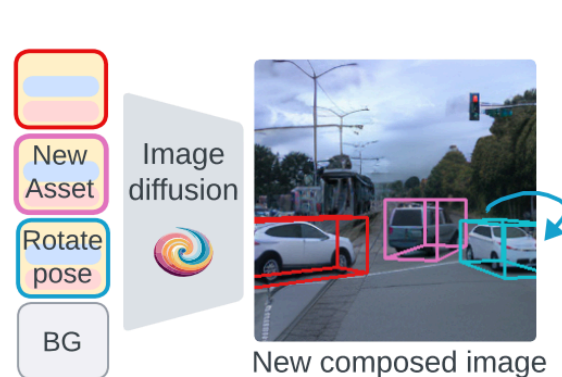
(a) Extracting Neural Assets



(b) Controllable generation



(c) Test-time compositionality



Interactive Video

- DeepMind's GENIE 3
 - Sadly, the demo is not available in Singapore...
- "SpAltial AI" Interactive Demo

SpAltial AI: Announcing Echo — our new frontier

SpAltial AI



Watch on

A Minecraft World Model

- Dreamer 4: Training Agents Inside of Scalable World Models
 - Project Page
- Learn to predict frames from actual game + actions
 - Autoregressive model with diffusion/flows to create next image
- Agent 'within the simulation':
 - obtain diamonds in Minecraft from only *offline* data
 - the agent only learns to play in simulation
 - then it must play the game for real
 - getting a diamond takes around 20,000 'actions'
- Essentially the original "World Models"
 - but with huge models & data & sophisticated RL

Dreamer 4 | Diamonds from Offline Experience

Danijar Hafner



Joint Embedding Predictive Architecture (JEPA) models

JEPA key points

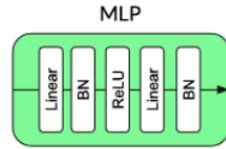
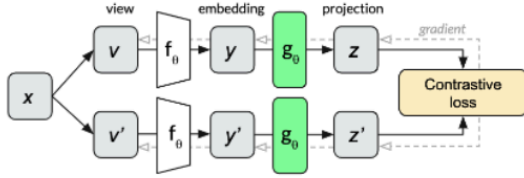
- Pixels = wasteful way to describe the world
 - pixels are ~irrelevant for world state representation
 - clean representation allows for better planning
- How to get a 'clean representation' is tricky...
 - key things :
 - don't predict pixels
 - represent everything that's useful
 - importantly : don't *collapse*
- Yann LeCun has been driving force



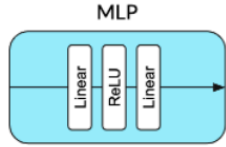
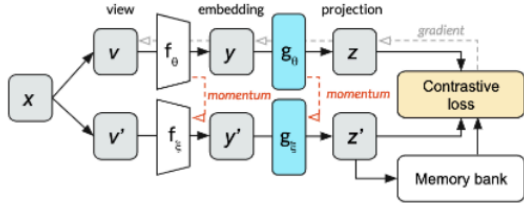
Figure 14. LeJEPa learns rich semantic representations through self-supervised learning. PCA visualization of last-layer features from LeJEPa (ViT-Large, 100 epochs on ImageNet-1K). For each image, fea-

Representation Learning

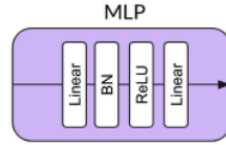
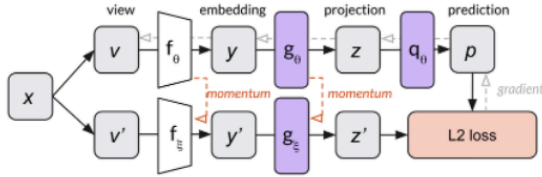
SimCLR



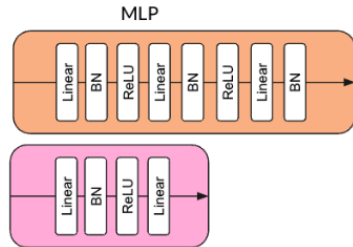
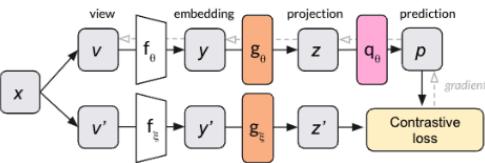
MoCo v2



BYOL



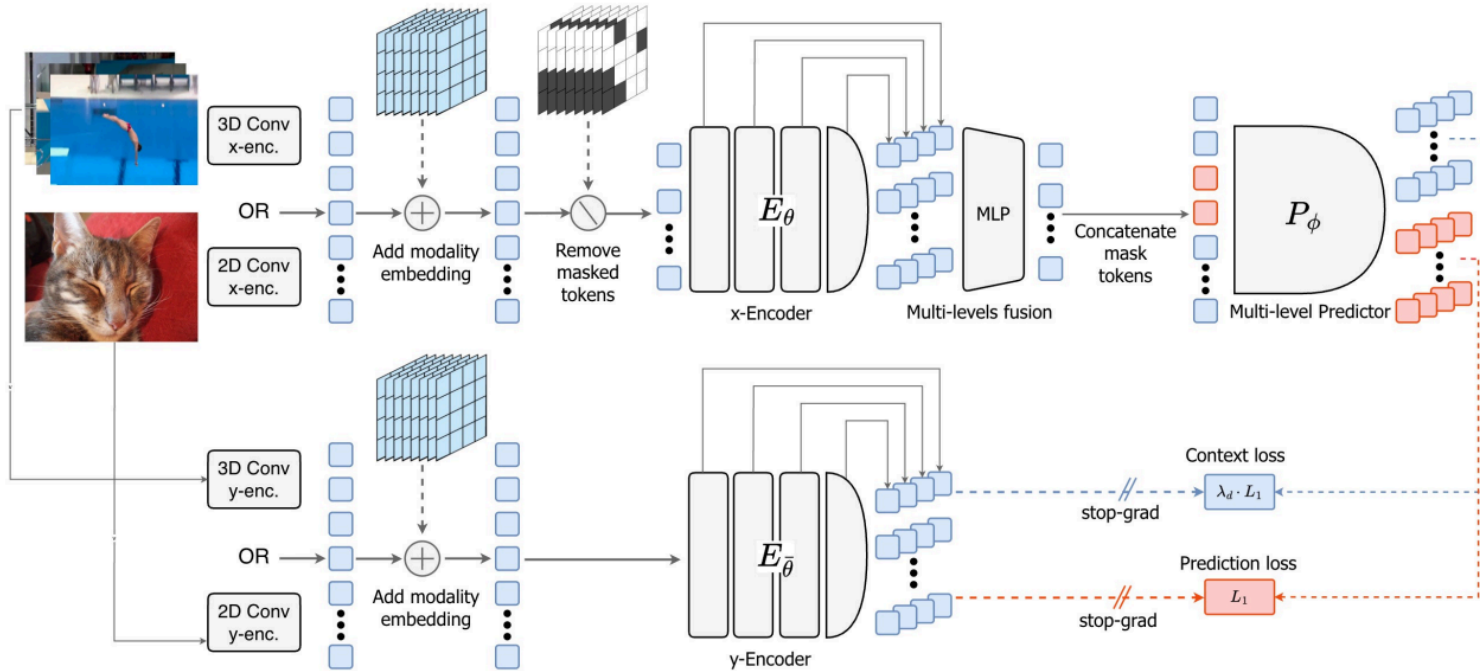
SimSiam



- Key element in ImageNet CNNs:
 - train a model to classify ImageNet images
 - then use the internal representation ...
 - ... to do Transfer Learning
- Other approaches : extract representation g
 - but making one representation *agree* with another
 - and *disagree* with others (contrastive)
 - lot's of ideas of teacher/student
 - hide information (from teacher)
 - graduate student into teacher
 - etc ...
- See this [Nice Blog Post](#) for more...

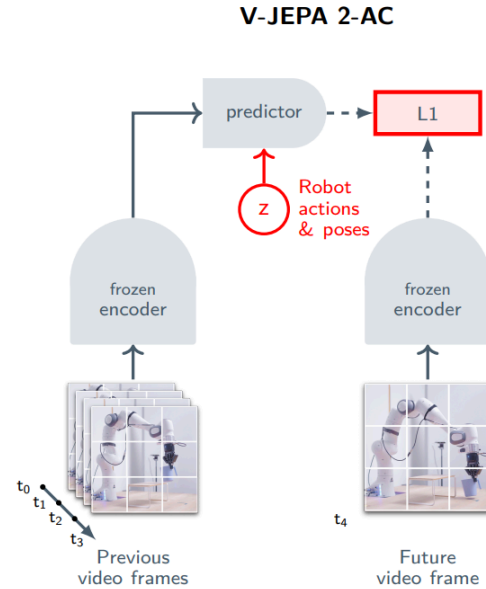
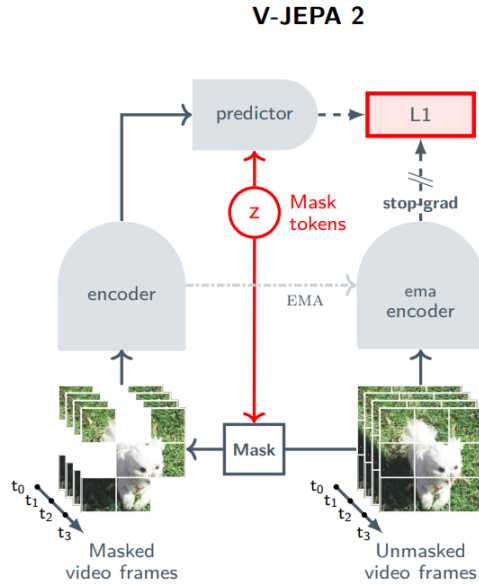


V-JEPA 2.1 : Unlocking Dense Features in Video SSL



Let's add in Actions

- V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning
 - Main training : 1M hours & 1M images
 - Robot training : 62 hours (states + actions)
 - enables basic robot skills like reaching, grasping, and pick-and-place via model-predictive control
 - V-JEPA 2-AC (Action-Conditioned) generalizes zero-shot to new environments





Latest updates...

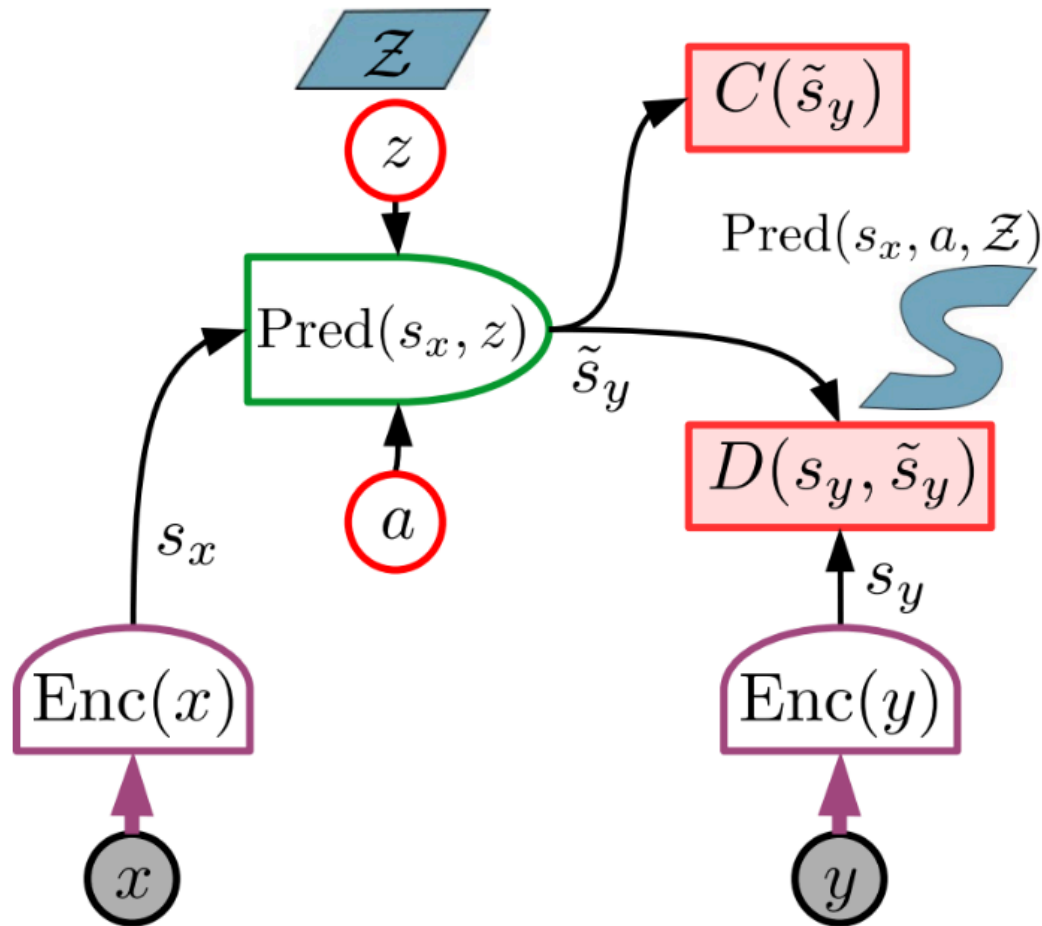
- LeWorldModel: Stable End-to-End JEPA from Pixels
 - ~15M parameters - single GPU training in ~hours
 - only 1 hyperparameter
 - trick to stop representation collapse:
 - force *distribution* of representations to be Normal
- Temporal Straightening for Latent Planning
 - successive representations \implies straighter paths
 - so that planning naturally becomes easier
- Yann LeCun has been driving force
 - and "AMI Labs" has just raised ~\$1Bn to pursue this
 - AMI = "Advanced Machine Intelligence" (and *friend* in French)

Architecture for the world model: JEPA

EVER-GREEN SLIDE!

▶ JEPA: Joint Embedding Predictive Architecture.

- ▶ x : observed past and present
- ▶ y : future
- ▶ a : action
- ▶ z : latent variable (unknown)
- ▶ $D(\cdot)$: prediction cost
- ▶ $C(\cdot)$: surrogate cost
- ▶ JEPA predicts a representation of the future S_y from a representation of the past and present S_x



Robot "Foundation" models

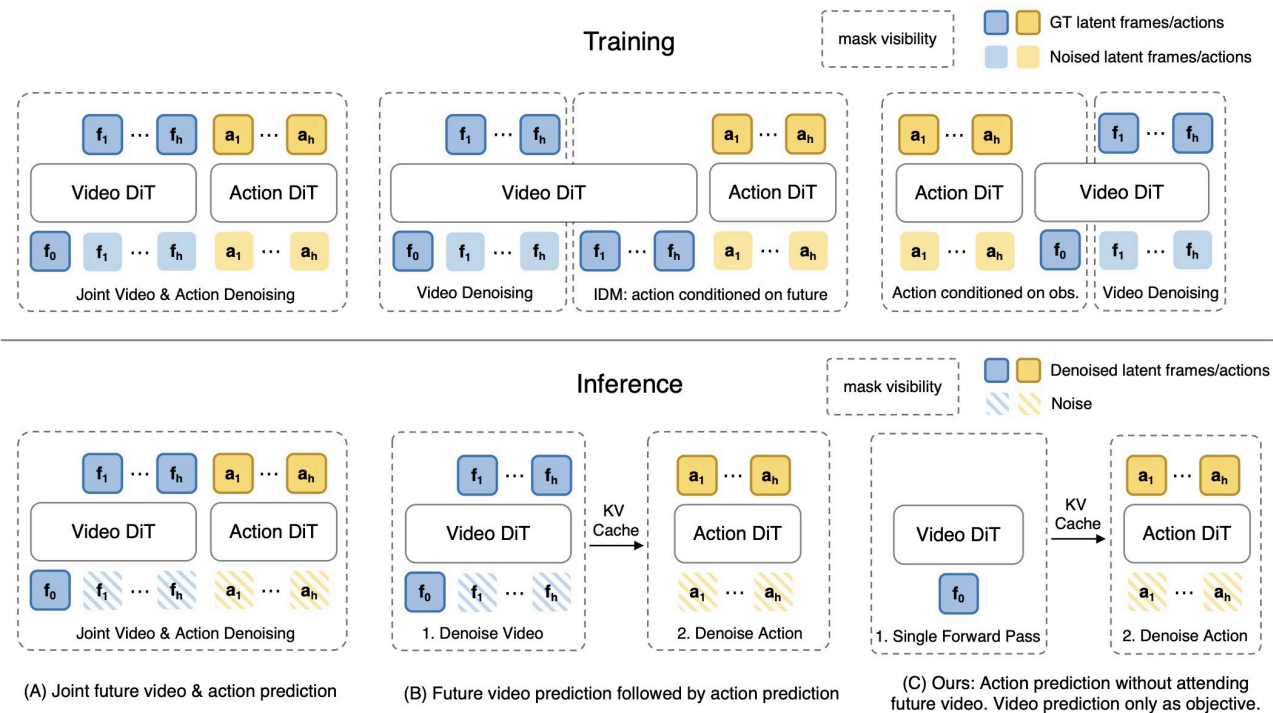
Robot Foundation model goals

- Want to do the same for (real world) Robots
 - as ChatGPT did for language
- Mantra (from Physical Intelligence)
 - one model many tasks
 - one model many environments
 - one model many robots
- Challenges:
 - robots have different ways of
 - sensing world
 - acting in the world
 - difficult to gather data
 - and that data may not *transfer*



Simplified Architecture

- See Fast-WAM Project Page
 - also: Add instructions in Natural Language



f_0

Video DiT

1. Single Forward Pass

KV Cache

$a_1 \dots a_h$

Action DiT

$a_1 \dots a_h$

2. Denoise Action

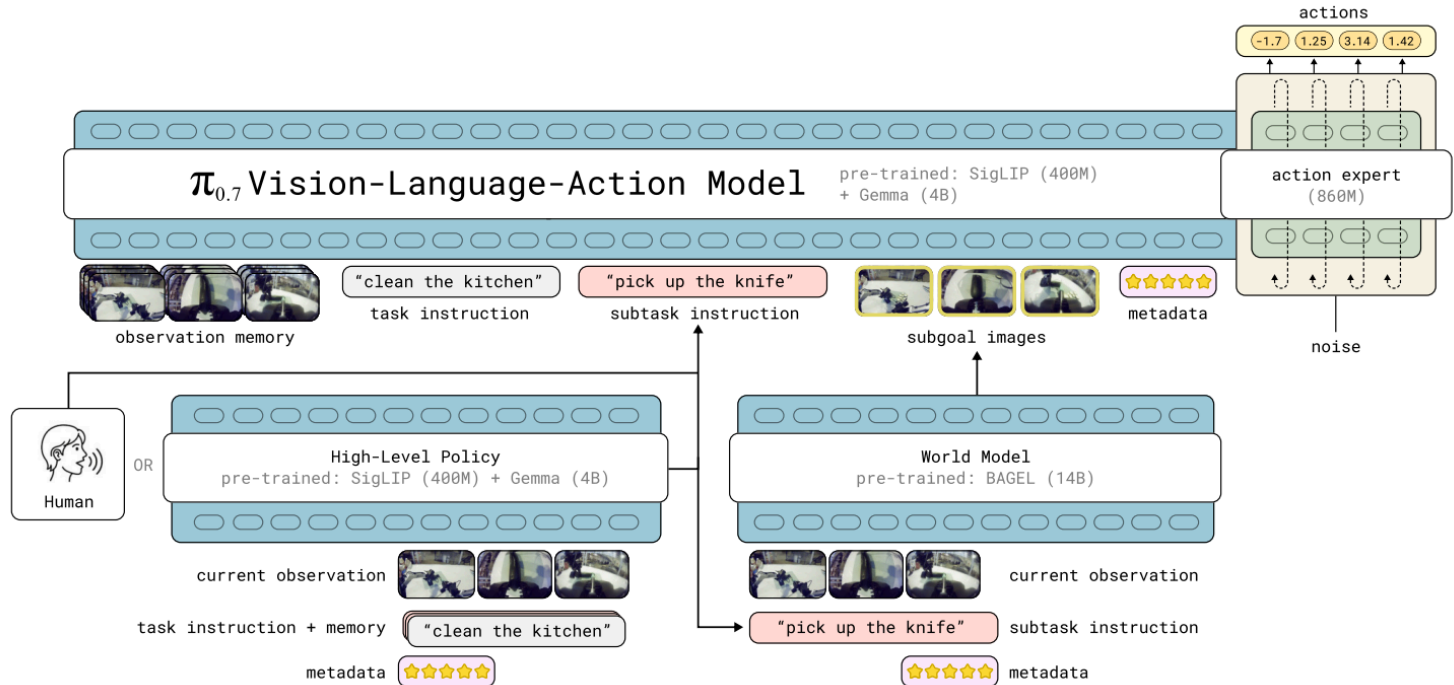
(A) Joint future video & action prediction

(B) Future video prediction followed by action prediction

(C) Ours: Action prediction without attending future video. Video prediction only as objective.

Physical Intelligence " π 0.7"

- Company Blog Post
 - Physical Intelligence - Founder Podcast @ YC
 - Paper : π 0.7 "Vision-Language-Action Model"- a steerable generalist robot model with emergent capabilities

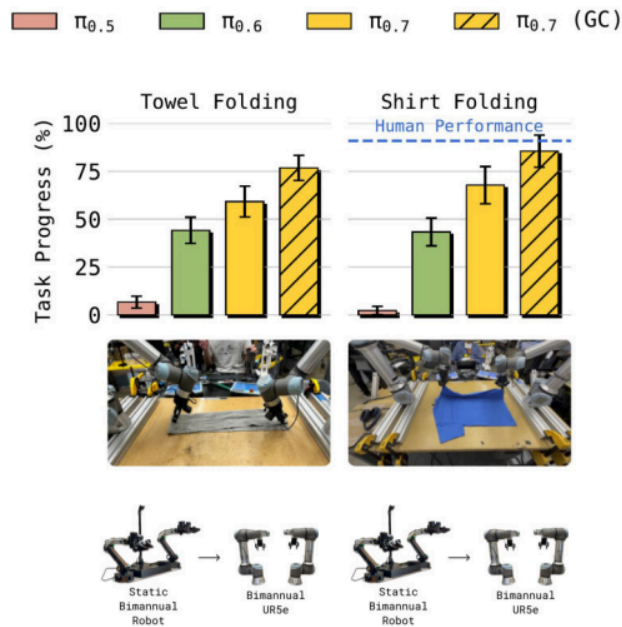
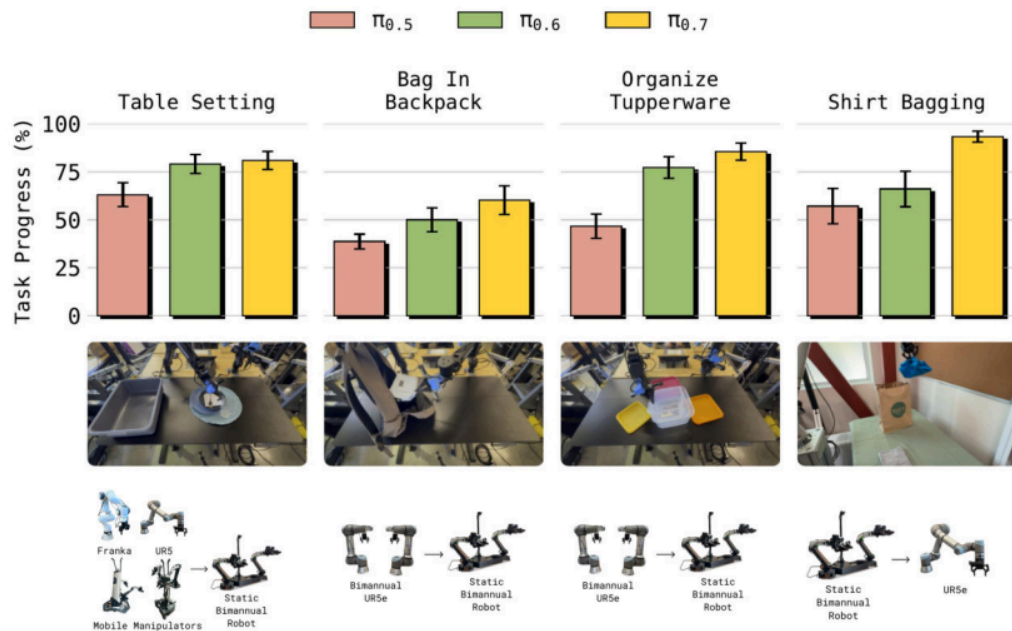


Physical Intelligence Research Engineer Thread

- "bittersweet lessons" Thread (first-person summary):
 - *my* bet was that world models would be the key unlock
 - that they'd dramatically outperform standard VLA approaches for task generalization
 - BUT : Our VLA baseline kept getting stronger as we collected more data
 - until one day it was also showing very promising signs of compositional generalization
 - key insights:
 - diverse data needs diverse context
 - Metadata about speed, quality, and visual subgoals disambiguate the training signal
 - surprise: cross-embodiment transfer on dexterous tasks
- We tested $\pi 0.7$ zero-shot on heavy industrial UR5e arms with imprecise parallel jaw grippers
 - No laundry folding data for that robot at all
 - It folded t-shirts consistently - and
 - applied a completely different folding strategy
 - adapted to the different robot affordances on its own

Cross-embodiment Transfer

- Left are simpler tasks, right are "(sub-)Goal Conditioned" : [Check Folding Video on Website](#)



Wrap-Up

- World Models can refer to many things
 - Alexandre LeBrun, the CEO of AMI Labs:
 - "In six months, every company will call itself a world model to raise funding."
 - of course, it all traces back to Schmidhuber (1992)
- JEPA vs "LLM-style"
 - representation learning techniques learn *stealthily*
 - ... but run counter to the (dominate) LLM narrative
- The robots are coming!



\$50M+ Robotics Rounds

Global · 2026 · top 20 of 38

GLOBAL VENTURE CAPITAL				
1/		Skild AI	\$1.4B	Robot foundation models for physical AI <i>SoftBank, NVIDIA, Lightspeed</i>
2/		Zipline	\$800M	Drone-based medical & logistics delivery <i>TPG, a16z, Sequoia</i>
3/		Apptronik	\$520M	Humanoid robots for real-world tasks <i>B Capital, Samsung NEXT, ARK</i>
4/		TARS	\$513M	Embodied intelligence systems <i>Qiming, Linear Capital, BlueRun</i>
5/		Mind Robotics	\$500M	AI robotics for industrial environments <i>a16z, Accel, Eclipse</i>
6/		Rhoda AI	\$450M	Robot foundation models for general AI <i>Khosla, Mayfield, Temasek</i>
7/		Galaxy Bot	\$364M	Universal service robots <i>Qiming, Matrix China, IDG</i>
8/		X Square	\$293M	Industrial & service robotic systems <i>Sequoia, Alibaba Cloud, Xiaomi</i>
9/		Galaxea AI	\$290M	Humanoid robots, embodied intelligence <i>Baidu, Lenovo Capital, Hillhouse</i>
10/		Bedrock Robotics	\$270M	Robotics for the construction sector <i>Eclipse, Valor, Emergence Capital</i>
11/		Linkerbot	\$217M	Embodied intelligent platforms <i>Sequoia, CDH, DT Capital</i>
12/		ENGINEAI	\$200M	General-purpose intelligent robotics <i>Baidu, 5Y Capital, CATL</i>
13/		Zhuji Dynamics	\$200M	General-purpose footed robots <i>Lenovo Capital, Alibaba, NIO Capital</i>
14/		Sunday	\$165M	AI-powered home robots <i>Coatue, Bain Capital, Tiger Global</i>
15/		D-Robotics	\$150M	Robotic computing solutions <i>5Y Capital, Plum, Prosperity7</i>
16/		Booster Robotics	\$147M	AI multi-modal control robots <i>Source Code, IDG, Tsinghua</i>
17/		PuduTech	\$146M	Autonomous delivery robots <i>Sequoia, Meituan, Tencent</i>
18/		Lightwheel	\$146M	Embodied AI simulation platform <i>Beijing AI Fund, Fresh Capital</i>
19/		Noetix Robotics	\$146M	Humanoid robots, AGI integration <i>CICC, Vertex China, Yunqi</i>
20/		Spirit AI	\$145M	"Universal brain" for real-world robots <i>Shunwei, Highlight, Tsinghua</i>

Source: Crunchbase, company announcements - Data compiled Apr 2026