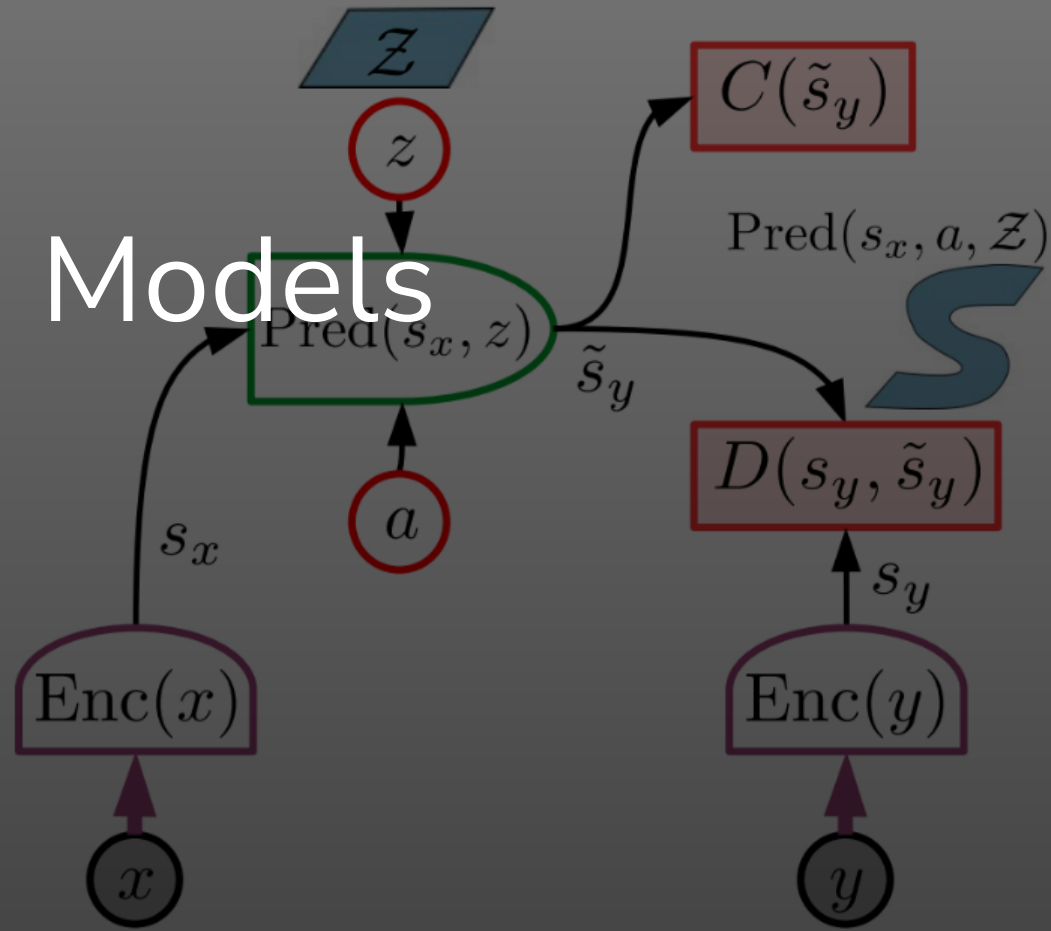
Architecture for the world model: JEPA

- **JEPA: Joint Embedding Predictive Architecture.**
  - x: observed past and present
  - y: future
  - a: action
  - Z: latent variable (unknown)
  - D(): prediction cost
  - C(): surrogate cost
  - JEPA predicts a representation of the future S_y from a representation of the past and present S_x



World [Action] Models

Machine Learning Singapore

- Martin Andrews = WorldActionModels @ mdda.net

29-April-2026

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...

MLSG
MACHINE LEARNING
SINGAPORE

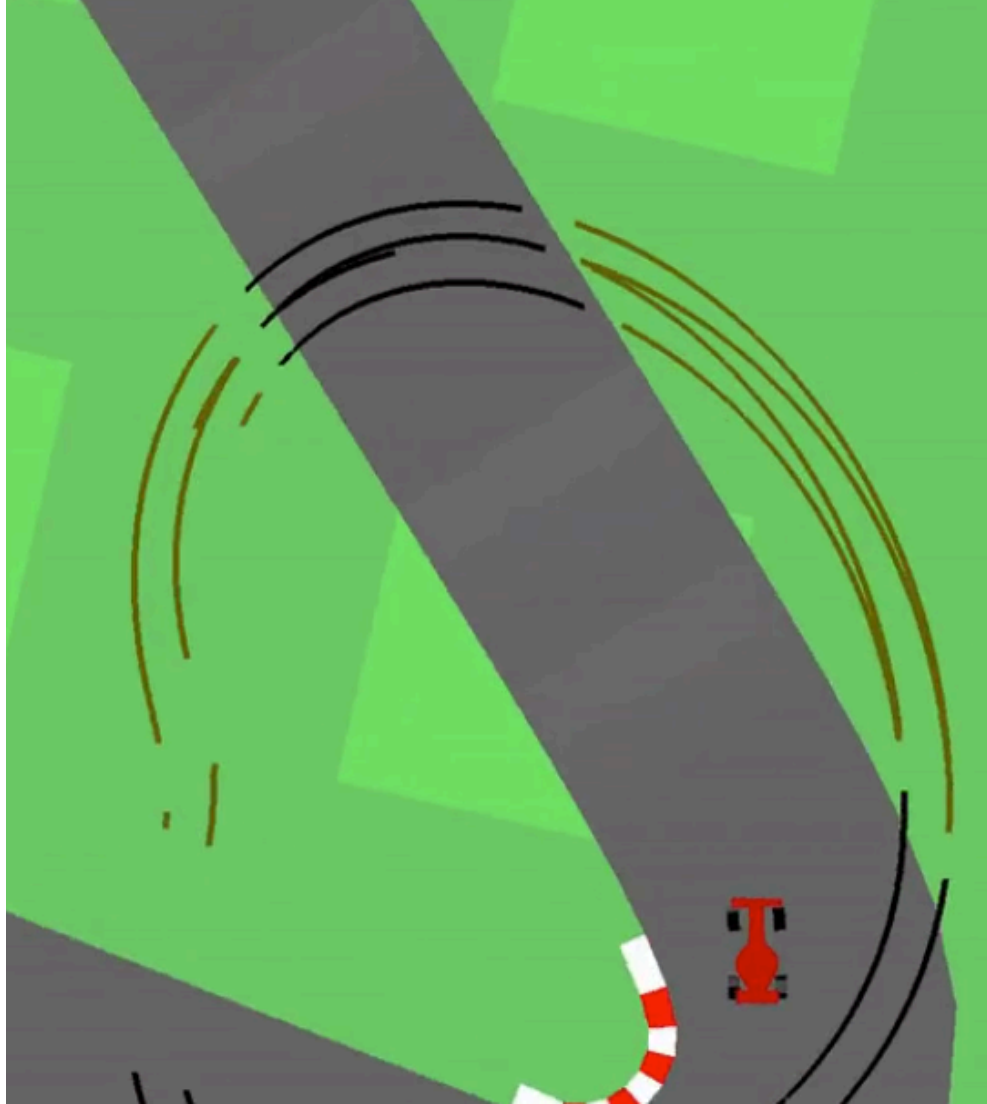
Outline

- 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"

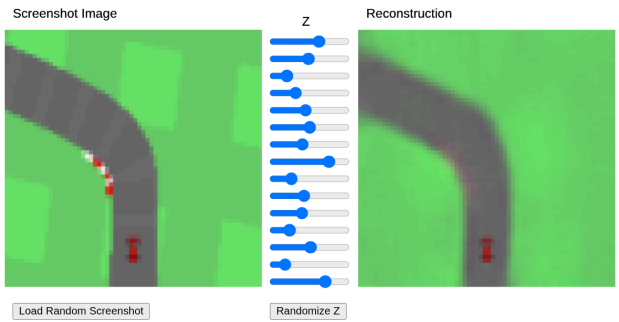
- World Models - Ha & Schmidhuber (2018)
  - David Ha (Sakana AI in Tokyo)
  - Jürgen Schmidhuber (needs no introduction)
- Idea:
  - Learn a representation of the world (VAE)
    - depends on pixels
  - Learn to predict the next representation
    - may include actions taken
  - Then : Learn to act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



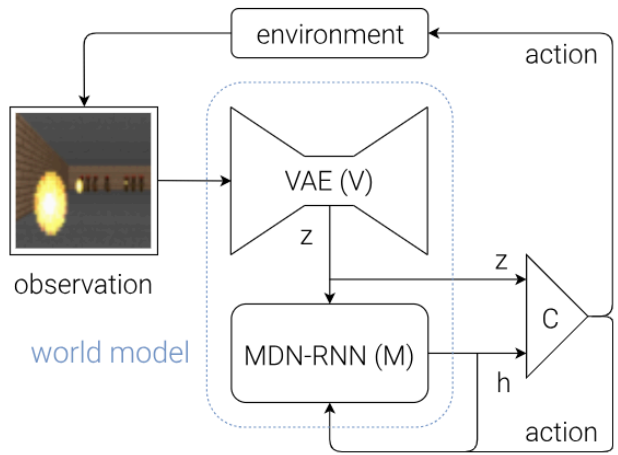
MLSG
MACHINE LEARNING
SINGAPORE

"World Models" : Interactive Project Page

- VAE reconstruction:



- Reinforcement Learning loop:



MLSG
MACHINE LEARNING
SINGAPORE

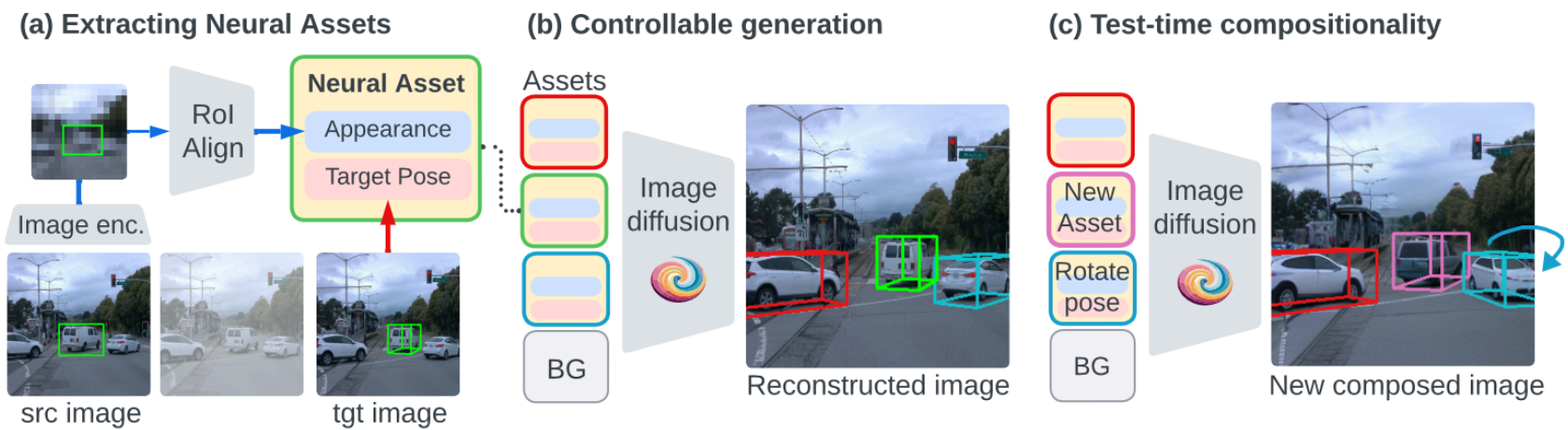
Video / Action modeling

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



MLSG
MACHINE LEARNING
SINGAPORE

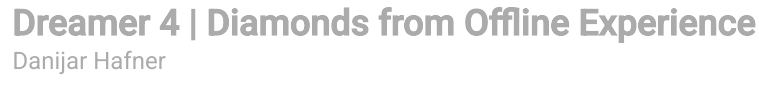
Interactive Video

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

MLSG
MACHINE LEARNING
SINGAPORE

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



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

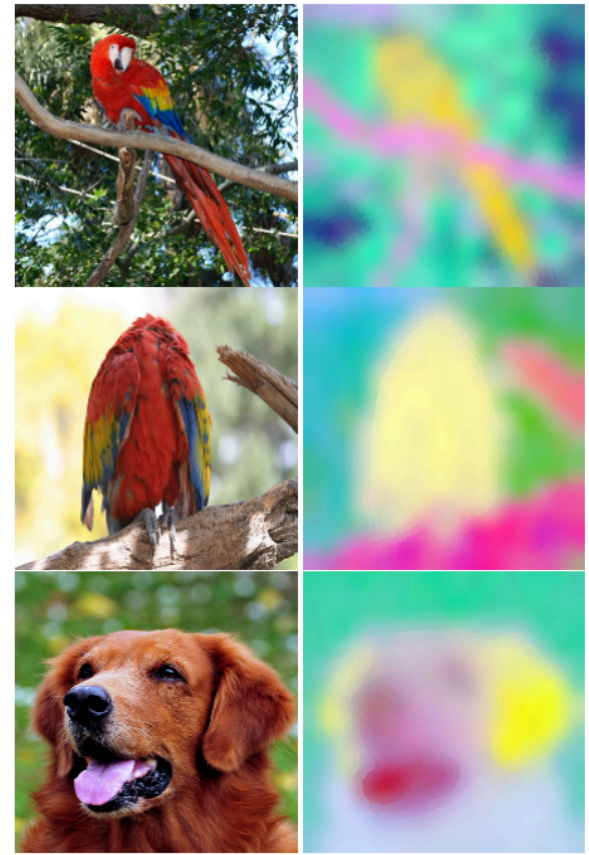
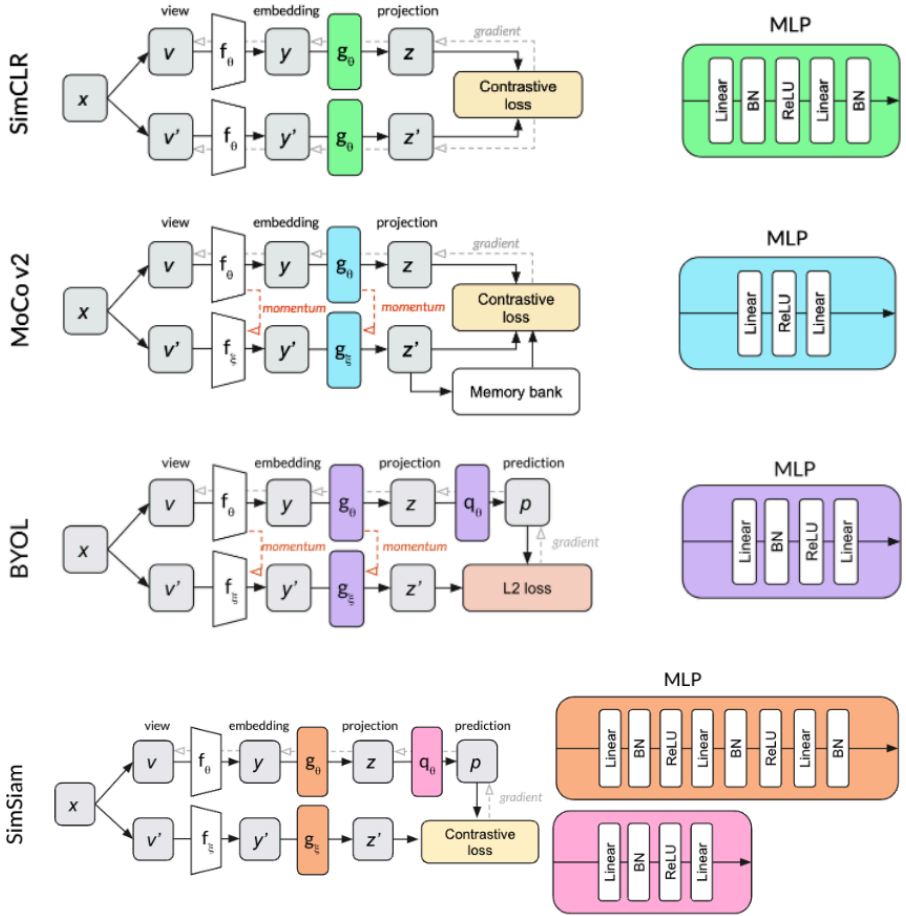


Figure 14. LeJEPA learns rich semantic representations through self-supervised learning. PCA visualization of last-layer features from LeJEPA (ViT-Large, 100 epochs on ImageNet-1K). For each image, fea-

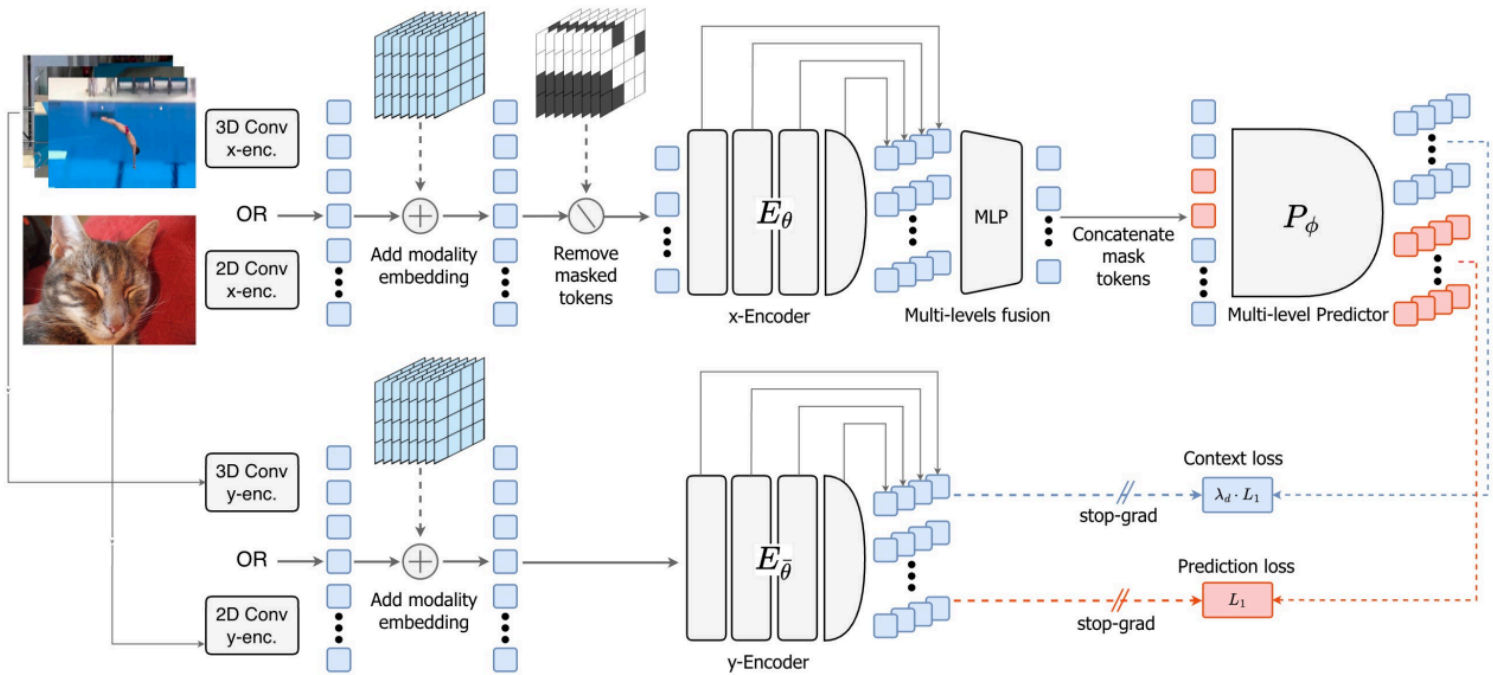
MLSG
MACHINE LEARNING
SINGAPORE



Representation Learning

- 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...

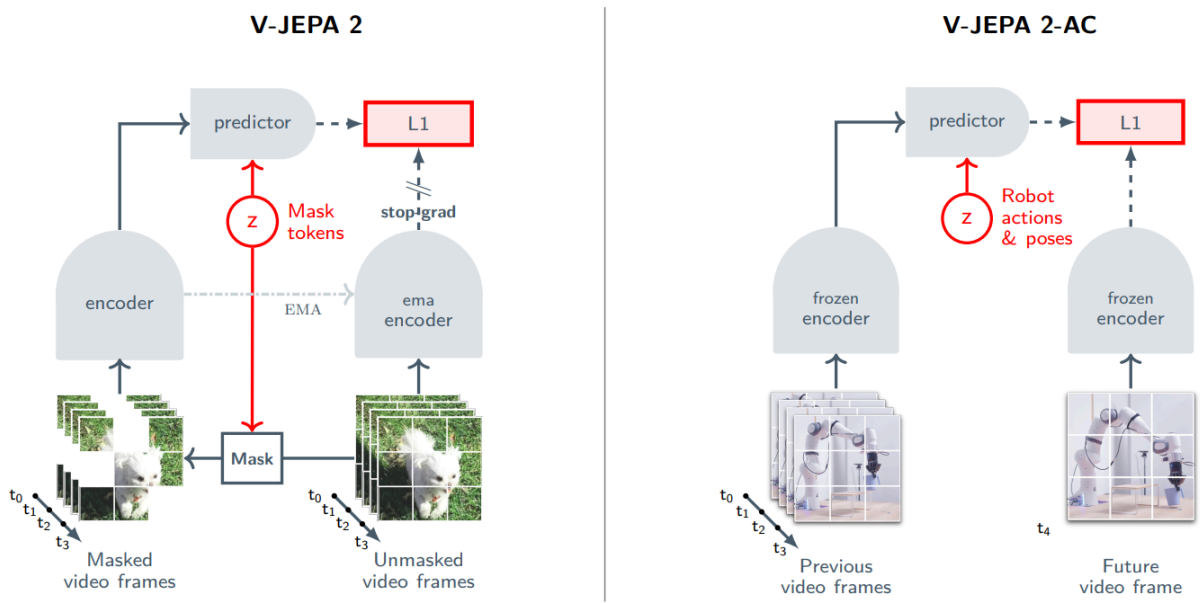
V-JEPA 2.1 : Unlocking Dense Features in Video SSL



MLSG
MACHINE LEARNING
SINGAPORE

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



MLSG
MACHINE LEARNING SINGAPORE

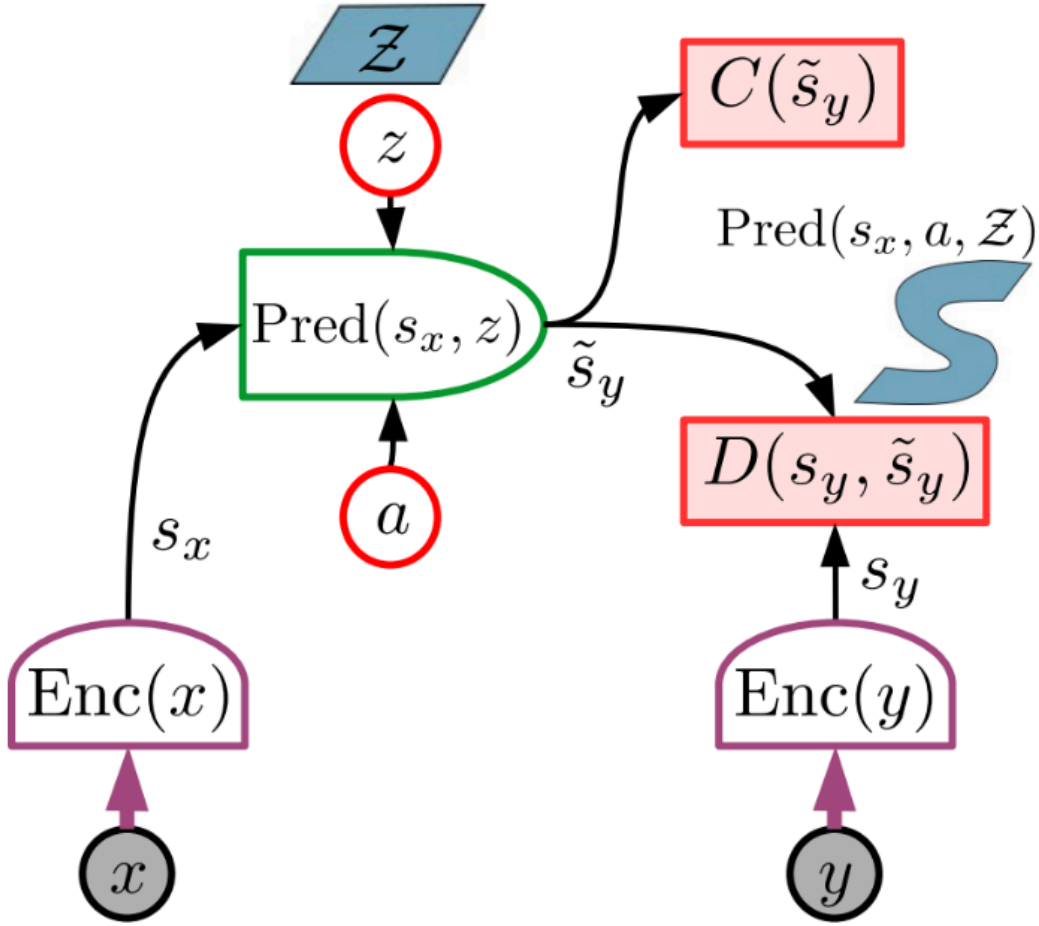
Latest updates...

- 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 $\Longrightarrow$ straighter paths
    - so that planning naturally becomes easier
- Yann LeCun has been driving force
  - and "AMI Labs" has just raised ~$1Bn to persue this
    - AMI = "Advanced Machine Intelligence" (and *friend* in French)

Architecture for the world model: JEPA

EVER-GREEN SLIDE!

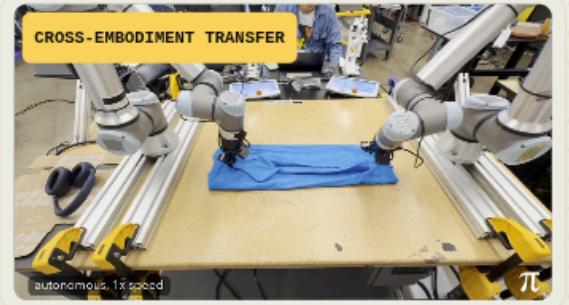
- **JEPA: Joint Embedding Predictive Architecture.**
  - x: observed past and present
  - y: future
  - a: action
  - z: latent variable (unknown)
  - D( ): prediction cost
  - C( ): surrogate cost
  - JEPA predicts a representation of the future S_y from a representation of the past and present S_x



Robot "Foundation" models

Robot Foundation model goals

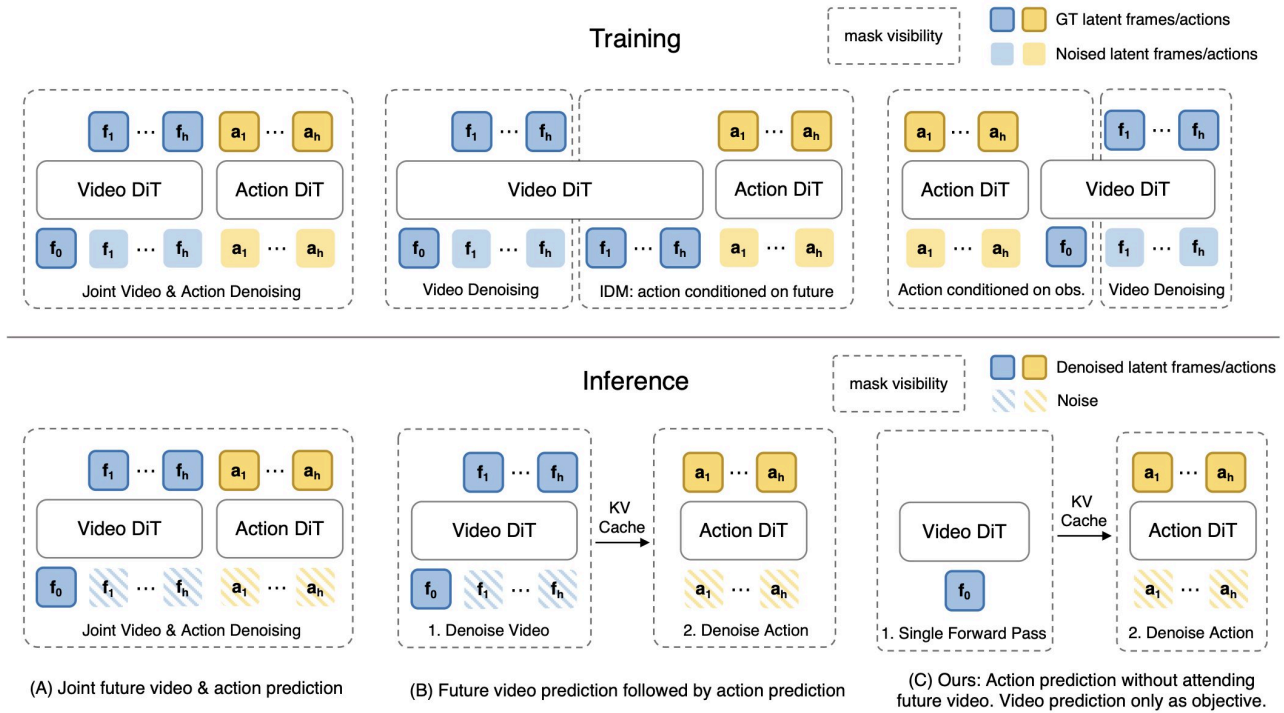
- Want to do the same for (real world) Robots
  - as ChatGPT did for language
- Mantra (from ""Physical Intelligence")
  - one model many tasks
  - one model many environments
  - one model many robots
- Challenges:
  - robots have different ways of
    - sensing world
    - acting in the world
  - difficult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



MLSG
MACHINE LEARNING
SINGAPORE

Simplified Architecture

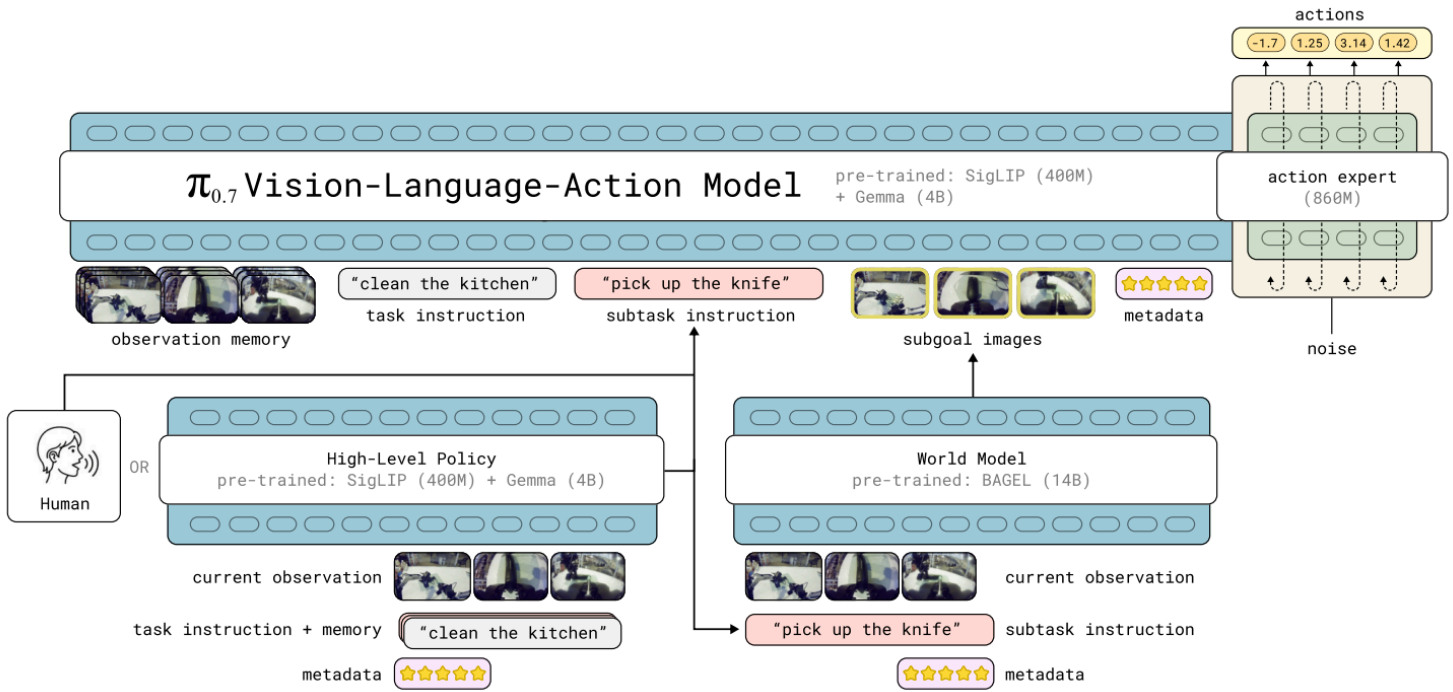
- See Fast-WAM Project Page
  - also: Add instructions in Natural Language



MLSG MACHINE LEARNING SINGAPORE

Physical Intelligence "π0.7"

- Company Blog Post
  - Physical Intelligence - Founder Podcast @ YC
  - Paper : π0.7 "Vision-Language-Action Model"- a steerable generalist robot model with emergent capabilities



MLSG MACHINE LEARNING SINGAPORE

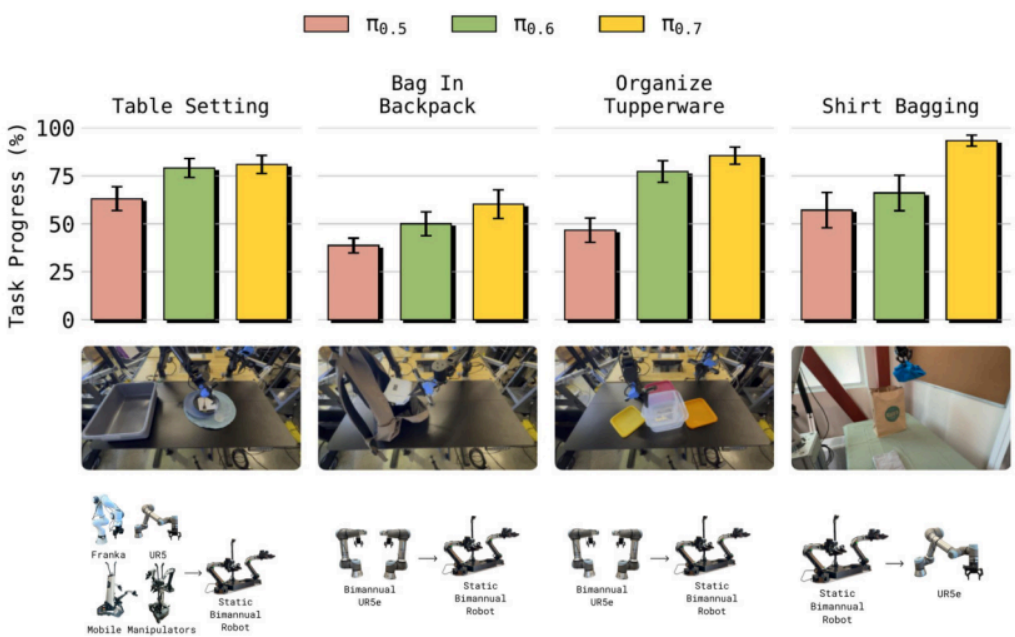
Physical Intelligence Research Engineer Thread

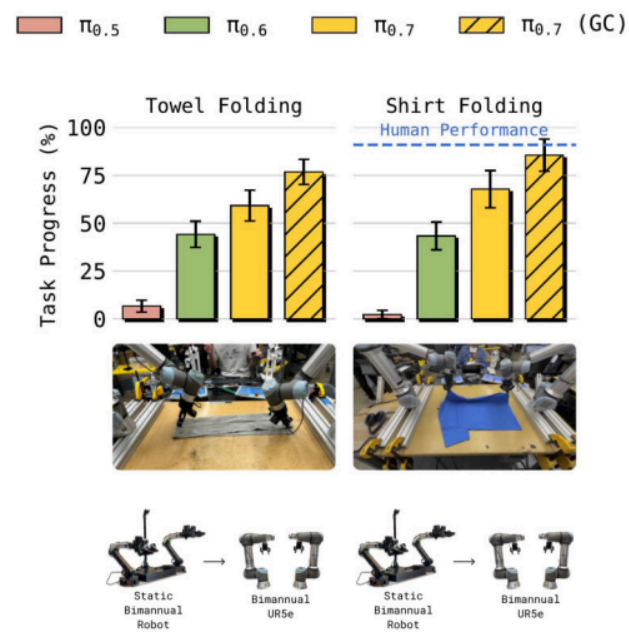
- "bittersweet lessons" Thread (first-person summary):
  - *my* bet was that world models would be the key unlock
    - that they'd dramatically outperform standard VLA approaches for task generalization
  - BUT : Our VLA baseline kept getting stronger as we collected more data
    - until one day it 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 π0.7 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





MLSG
MACHINE LEARNING
SINGAPORE

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!

MLSG
MACHINE LEARNING
SINGAPORE