CDS Seminar

SSL, JEPA, World Models and the Future of AI

Yann Le Cun
New York University
Meta – Fundamental AI Research

NYU CDS
2025-09-10

M-51, HSO

We Need Human-Level AI for Intelligent Assistant

- **In the near future, all of our interactions with the digital world will be mediated by AI assistants.**
- **Intelligent assistants that can helps us in our daily lives**
- **Smart glasses**
  - Communicates through voice, vision, display, EMG…
- **We need machines with human-level intelligence**
  - Machines that understand how the world works
  - Machines that can remember
  - Machines that can reason and plan.

“Her” (2013)

Meta Orion (2024)

The Ubiquitous AI Assistant is Becoming A Reality

- **Ray-Ban Meta (today)**
  - Cameras / microphone / speakers
  - no display
  - Voice interface to Meta AI assistant
- **Meta's Orion Demonstrator (future)**
  - Cameras / microphones
  - Augmented reality color display
  - Voice + EMG bracelet interface

But Machine Learning Sucks! (compared to humans and animals)

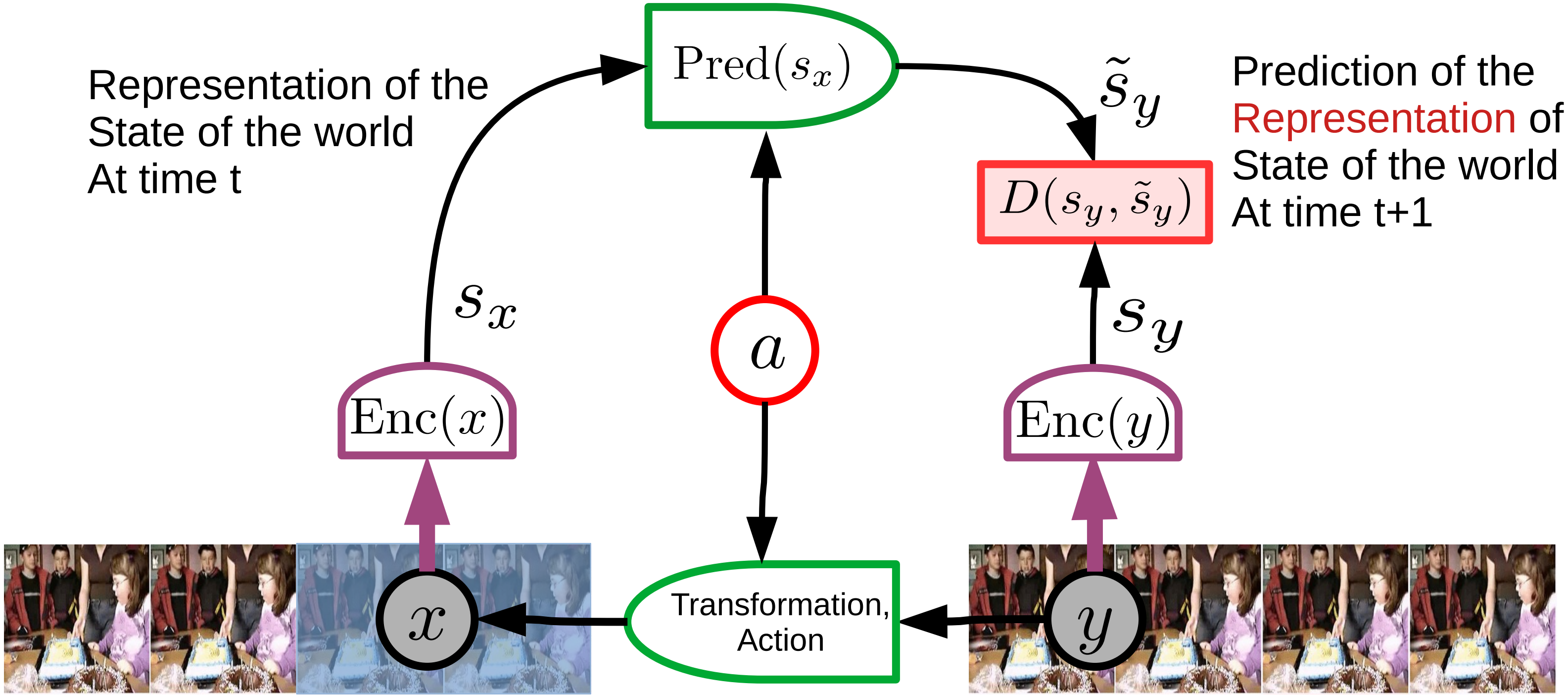
- **Supervised learning (SL) requires large numbers of labeled samples.**
- **Reinforcement learning (RL) requires insane amounts of trials.**
- **Self-Supervised Learning (SSL) works great but...**
  - Generative prediction only works for text and other discrete modalities
- **Animals and humans:**
  - Can learn new tasks **very** quickly.
  - Understand how the world works
  - Can reason an plan
- **Humans and animals have common sense**
- **Their behavior is driven by objectives (drives)**

What's a universal foundation model architecture

- **Captures structure in the data**
  - Discovers dependencies in a task-independent way
- **Trained with Self-Supervised Learning (SSL)**
  - No need for labels
- **Learns abstract representations in the data**
  - Representations that allow to make predictions
- **Learns a predictive model**
  - Observation x, transformed observation y=Trans(x,a)
  - Encoding : representations $s_x = Enc(x)$, $s_y = Enc(y)$
  - Prediction of s_y : $p_y = Pred(s_x, a)$

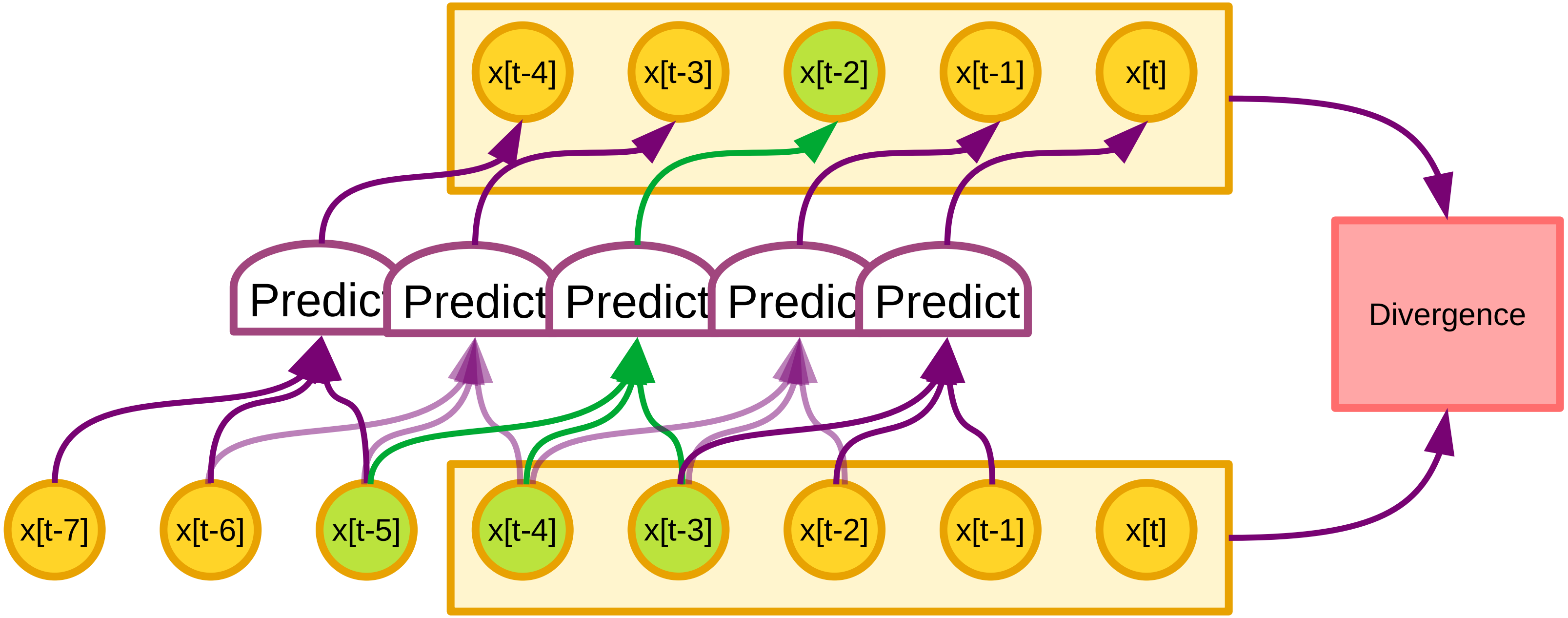
Predictive Model with JEPA

- **Joint Embedding Predictive Architecture (JEPA)**
  - [LeCun 2022], [Garrido 2023], [Bardes 2023], [Assran 2023], [Garrido 2024]



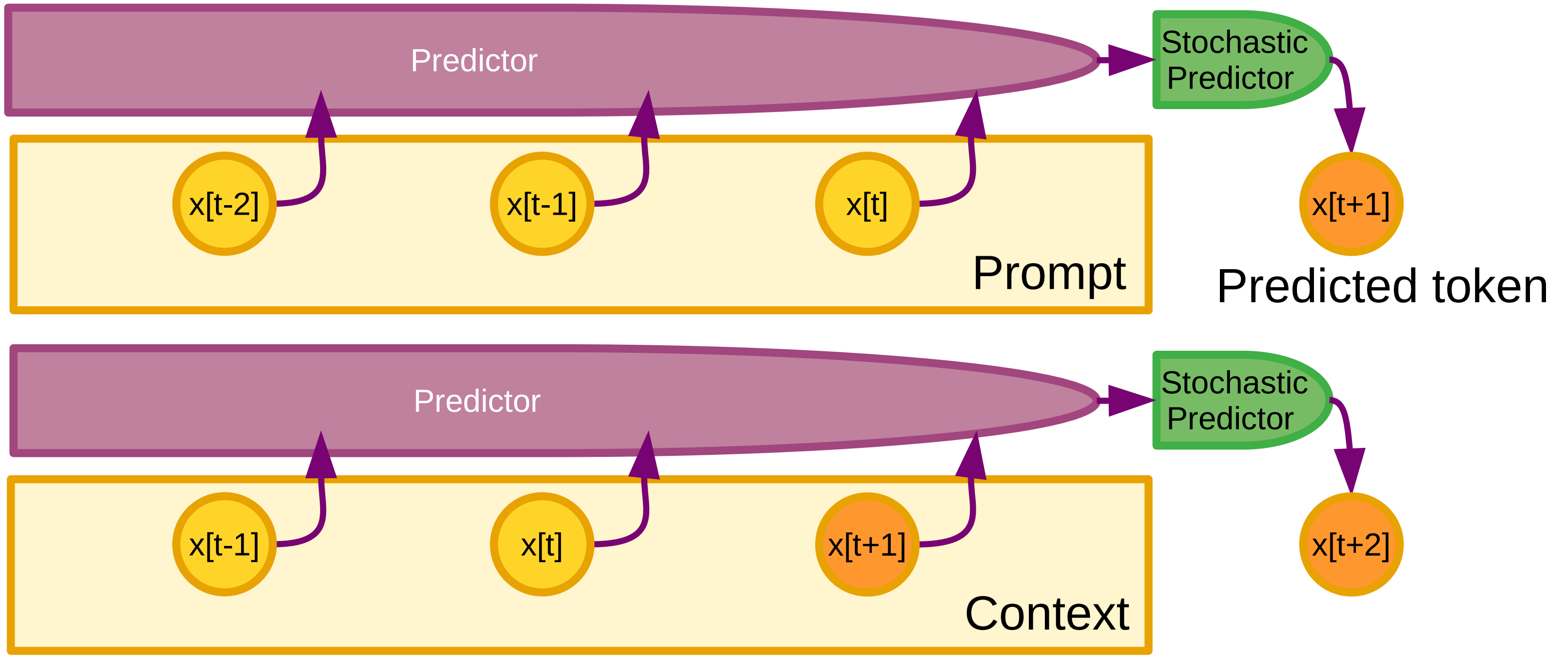
AE Collapse Prevention through Architectural Constraints

- **Train an auto-encoder with causal connections**
- **No connection between an input and its corresponding output**
- **LLMs / GPT architectures are the most popular example**
  - Trained to predict the next input.



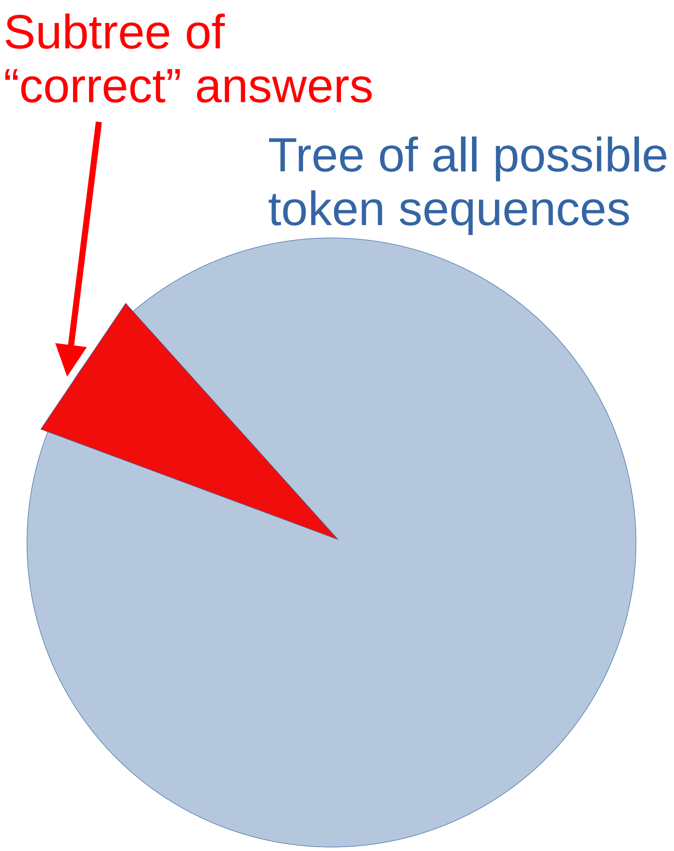
Auto-Regressive LLM. Inject predicted token in the input

- **Outputs one token after another through feed-forward prediction**
- **Tokens may represent words, image patches, speech segments…**
- **Predictor has a fixed number of layers**
- **Only works for discrete domains (text, DNA….)**



Auto-Regressive Generative Models Suck!

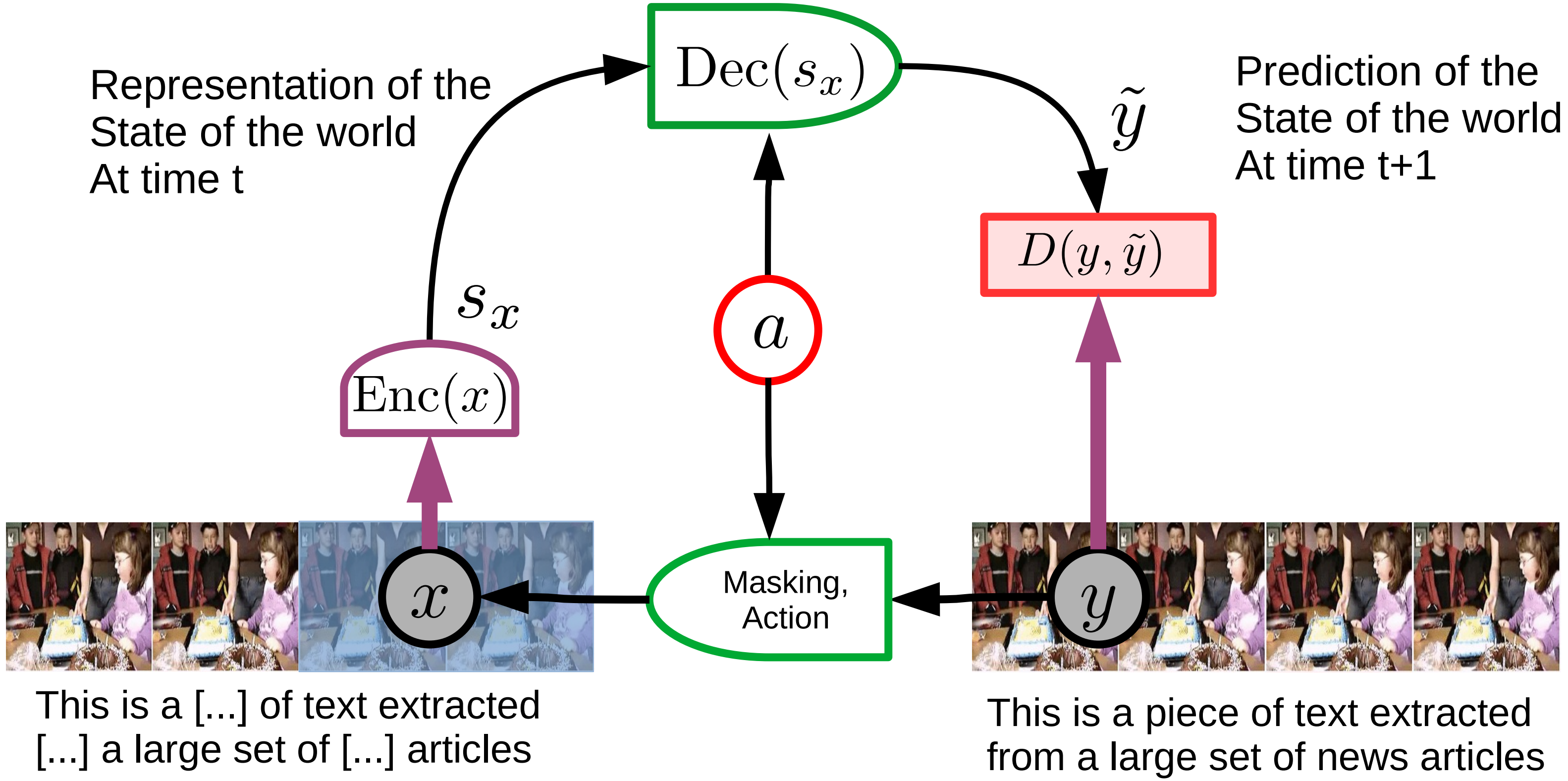
- **Auto-Regressive LLMs are doomed.**
- **They cannot be made factual, non-toxic, etc.**
- **They are not controllable**
- **Probability e that any produced token takes us outside of the set of correct answers**
- **Probability that answer of length n is correct (assuming independence of errors):**
  - $P(\text{correct}) = (1-e)^n$
- **This diverges exponentially.**
- **It's not fixable (without a major redesign).**



- See also [Dziri...Choi, ArXiv:2305.18654]

Can we train Generative Architecture with Continuous Data?

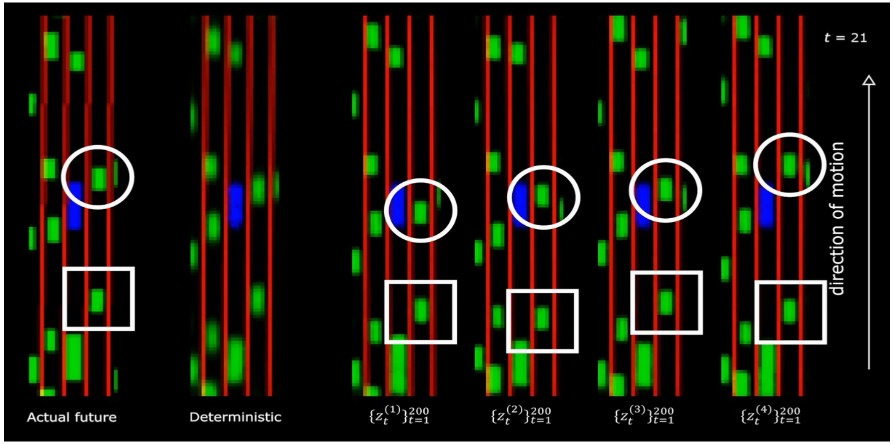
- **Short answer: NO!!!**
- **It works for discrete domains, not high-dim domains**
- **Generative world model architecture**



Generative Architectures DO NOT Work for Images and video

- Because the world is only partially predictable
- A predictive model should represent multiple predictions
- Probabilistic models are intractable in high-dim continuous domains.
- Generative Models must predict every detail of the world
- **My solution: Joint-Embedding Predictive Architecture**

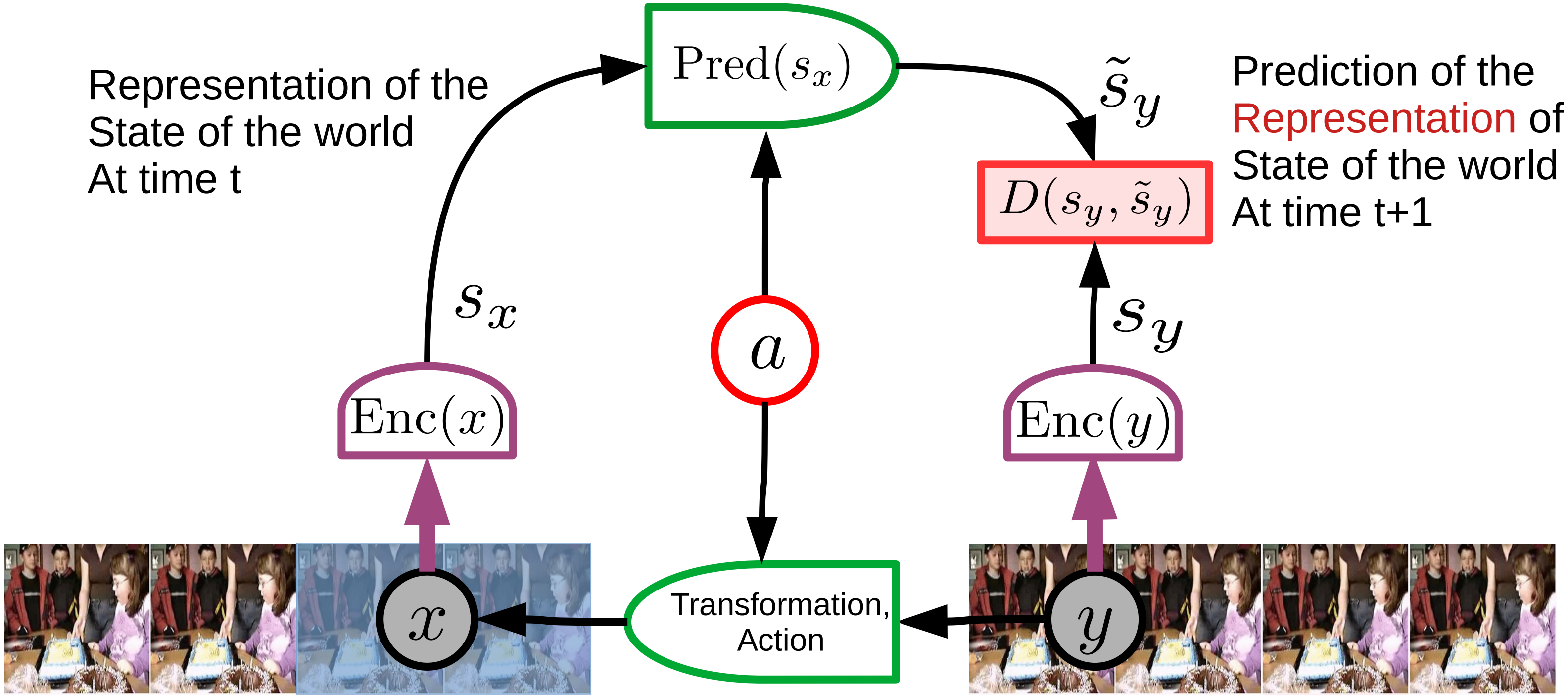
[Mathieu, Couprie, LeCun ICLR 2016]



[Henaff, Canziani, LeCun ICLR 2019]

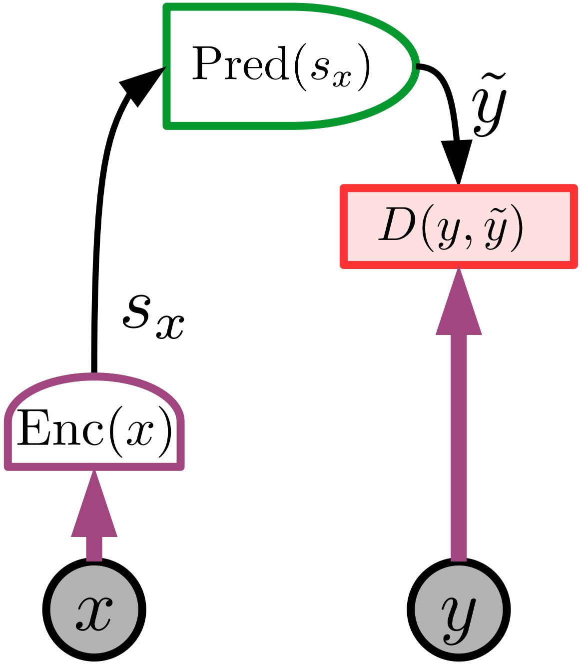
Joint Embedding World Model: Self-Supervised Training

- **Joint Embedding Predictive Architecture (JEPA)**
  - [LeCun 2022], [Garrido 2023], [Bardes 2023], [Assran 2023], [Garrido 2024]

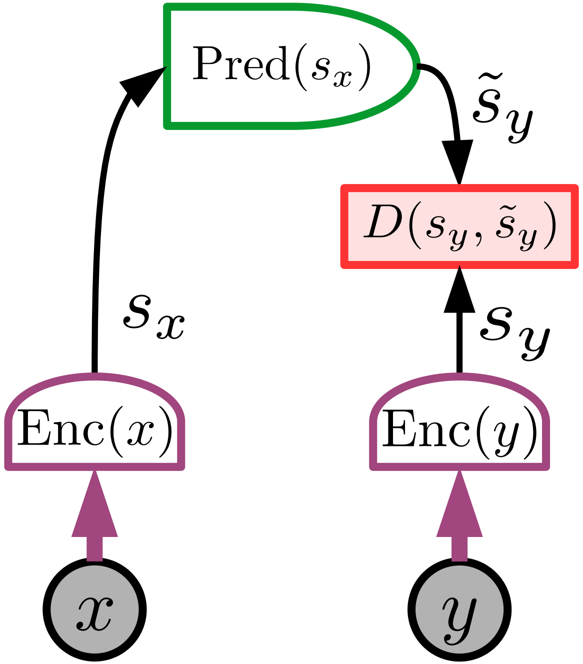


Architectures: Generative vs Joint Embedding

- **Generative: predicts y** (with all the details, including irrelevant ones)
- **Joint Embedding: predicts an abstract representation of y**
- **JEPA lifts the abstraction level, generative architectures do not.**



a) Generative Architecture
Examples: VAE, MAE...

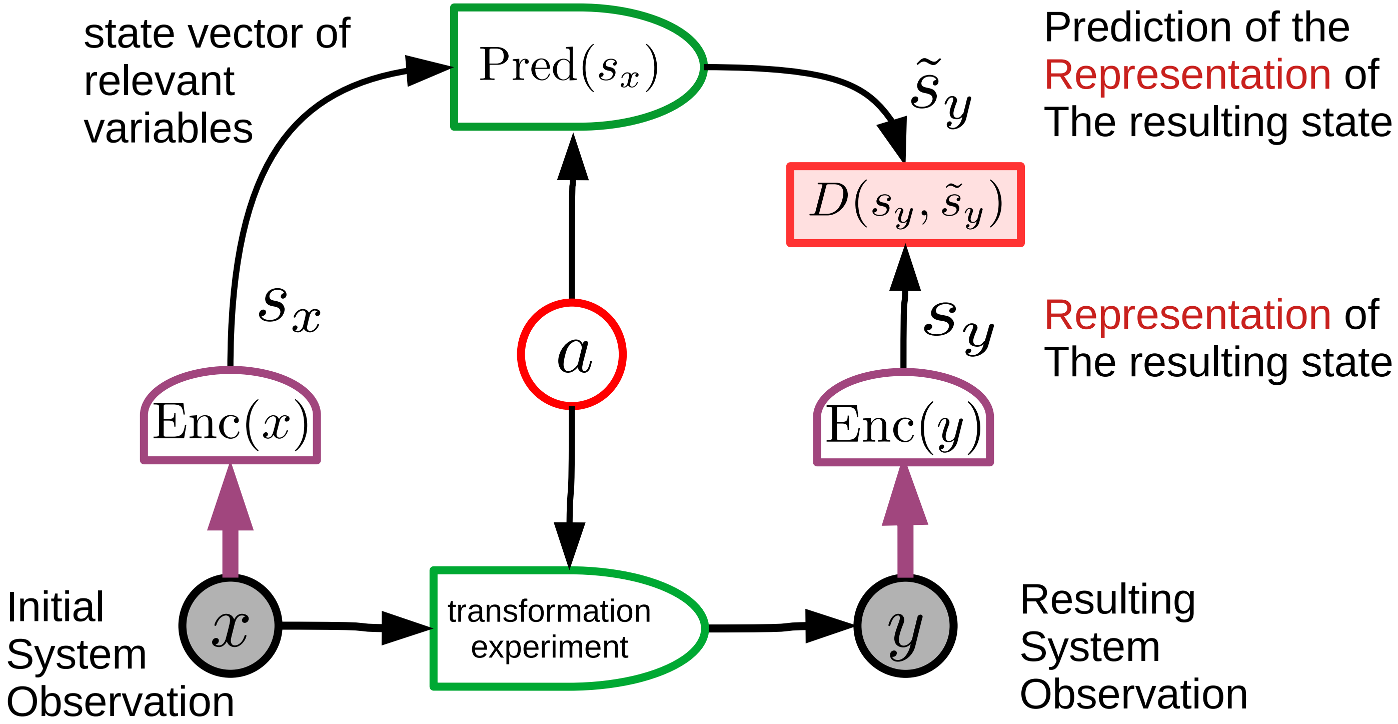


b) Joint Embedding Architecture

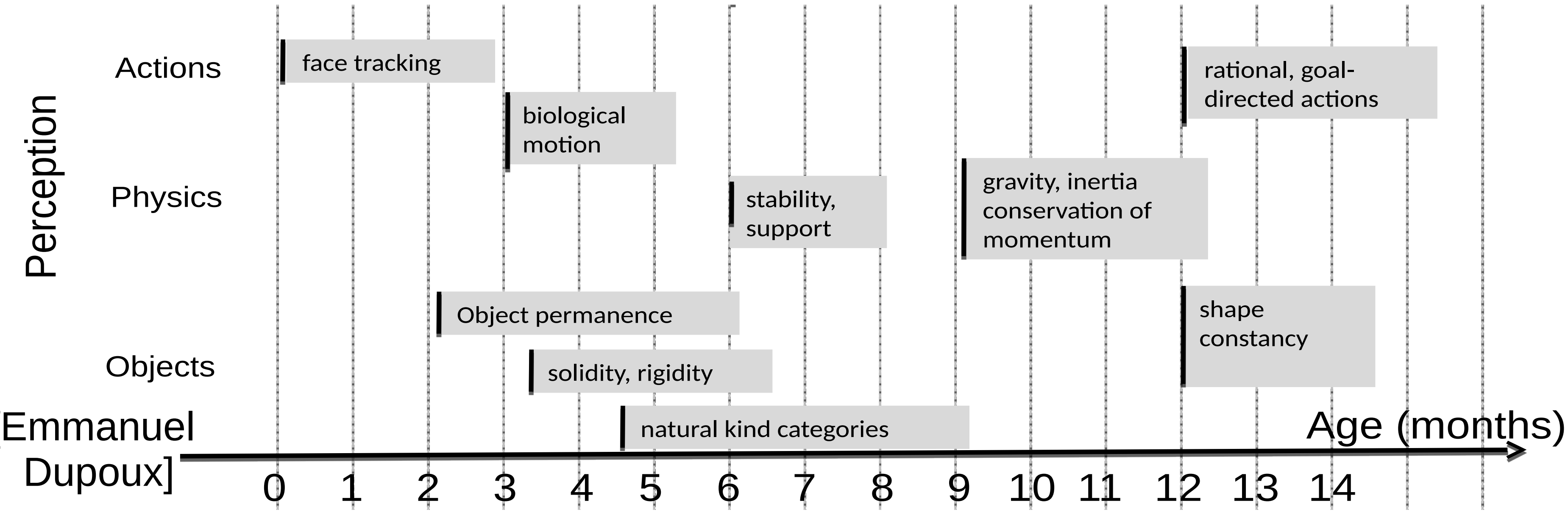
This is how models are built in traditional physics

- Find an **abstract state representation** that allows to make predictions
- Extract the state representation from observation/measurement
- Predict outcome resulting from an intervention/experiment

- Irrelevant and unpredictable information is eliminated from the representation
- The representation contains information that makes prediction possible



How do babies learn how the world works?



[Emmanuel Dupoux]

- **How do we get machines to learn like babies?**

Current architectures are missing something really big!

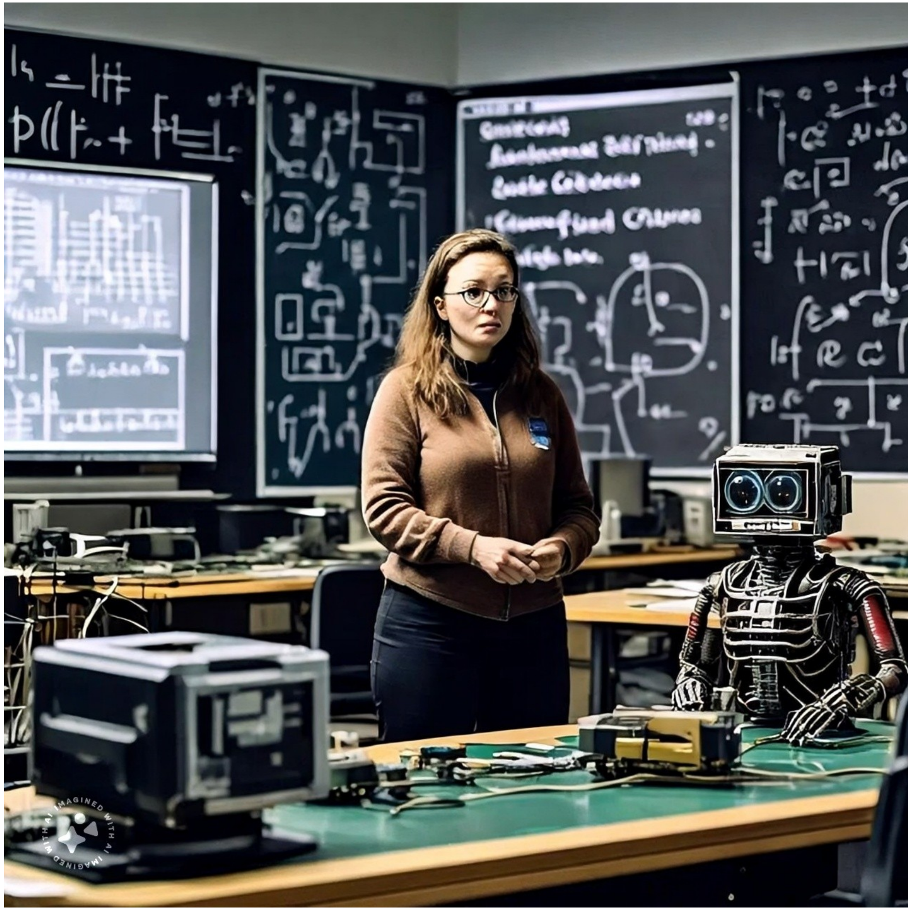
- **Never mind humans, cats and dogs can do amazing feats**
  - Current robots intelligence doesn't come anywhere close
- **Any house cat can plan highly complex actions**
- **Any 10 year-old can clear up the dinner table and fill up the dishwasher without learning ("zero-shot")**
- **Any 17 year-old can learn to drive a car 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.

Our world model needs to be trained from sensory inputs

- **LLM**
  - Trained on 3.0E13 tokens (2E13 words). Each token is 3 bytes.
  - Data volume: 0.9E14 bytes.
  - Would take 450,000 years for a human to read (12h/day, 250 w/minute)
- **Human child**
  - 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 !**

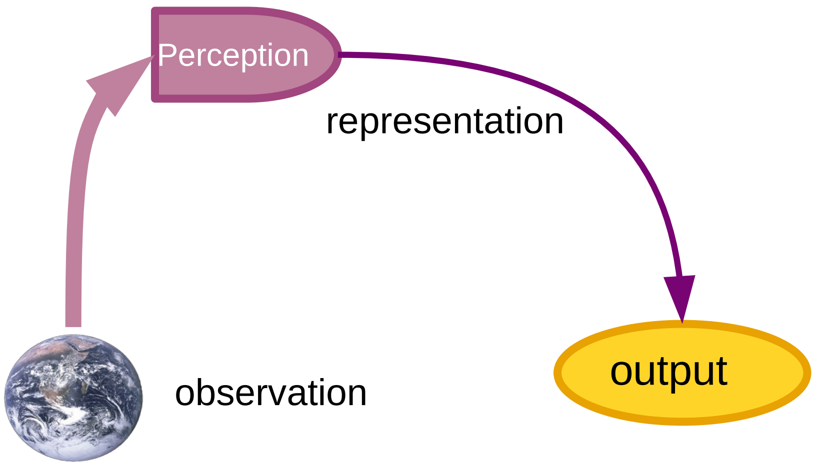
Desiderata for AMI (Advanced Machine Intelligence)

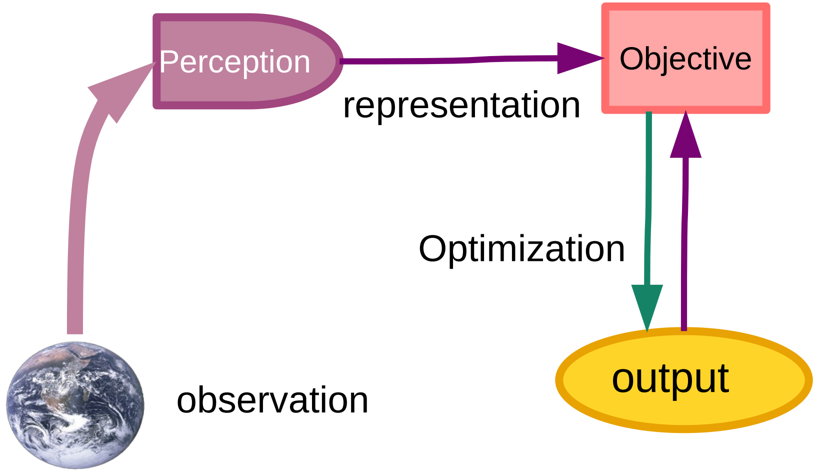
- **Systems that learn world models from sensory inputs**
  - E.g. learn intuitive physics from video
- **Systems that have persistent memory**
  - Large-scale associative memories
- **Systems that can plan actions**
  - So as to fulfill an objective
- **Systems that can reason**
  - Inventing new solutions to unseen problems
- **Systems that are controllable & safe**
  - By design, not by fine-tuning.



Inference: feed-forward propagation vs optimization

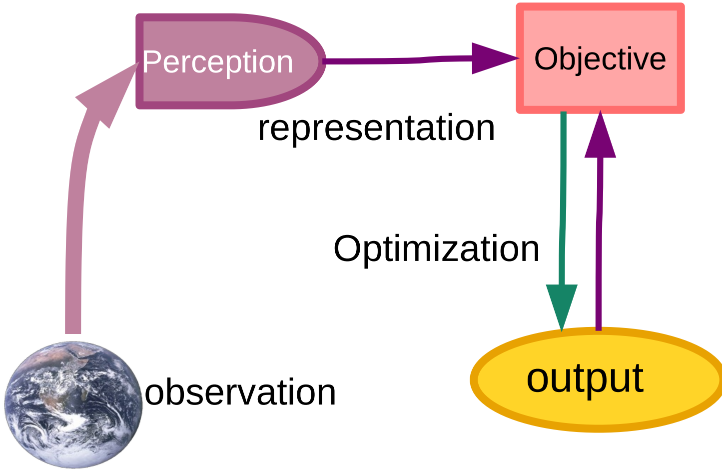
- **What is reasoning and planning?**
- **Feed-forward propagation is insufficient**
- **Complex inference requires the optimization of an objective**
- **Every computational problem can be reduced to optimization**
  - This includes every inference and planning problem.
- **Energy-Based Model**





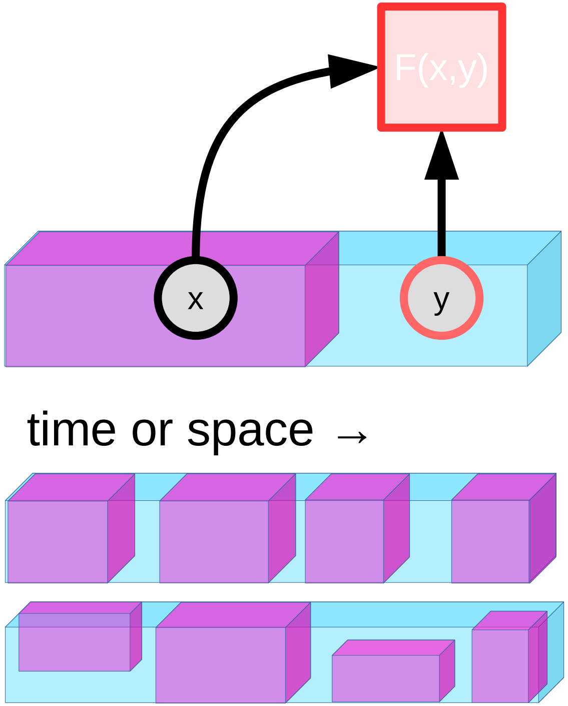
Inference through optimization: Objective-Driven AI.

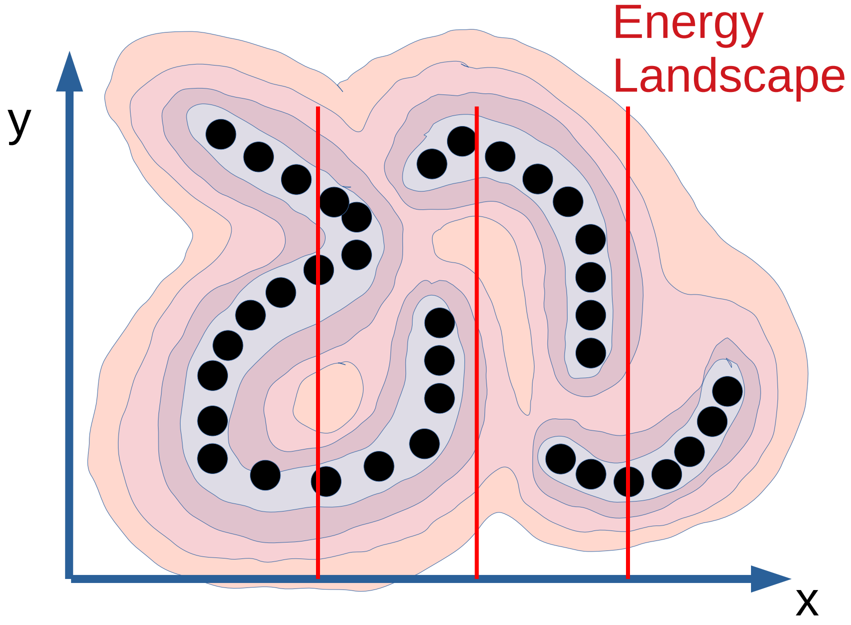
- **Inference through optimization is used in classical methods**
  - Probabilistic graphical models, Bayesian nets
  - Model-Predictive Control in robotics
  - Search & planning in “classical” AI
- **In the past, all of AI was viewed as a search or optimization problem**
  - Path planning, Block World, Towers of Hanoi, SAT, logical inference
- **Optimization-based inference enables zero-shot “learning”**
  - It can find innovative solutions to unseen problems.
  - All game-playing AI systems use search/planning
- **Optimization-based inference is “System 2”**



Capturing Dependencies with Energy-Based Models

- **The only way to formalize & understand all model types**
  - Gives low energy to compatible pairs of x and y
  - Gives higher energy to incompatible pairs

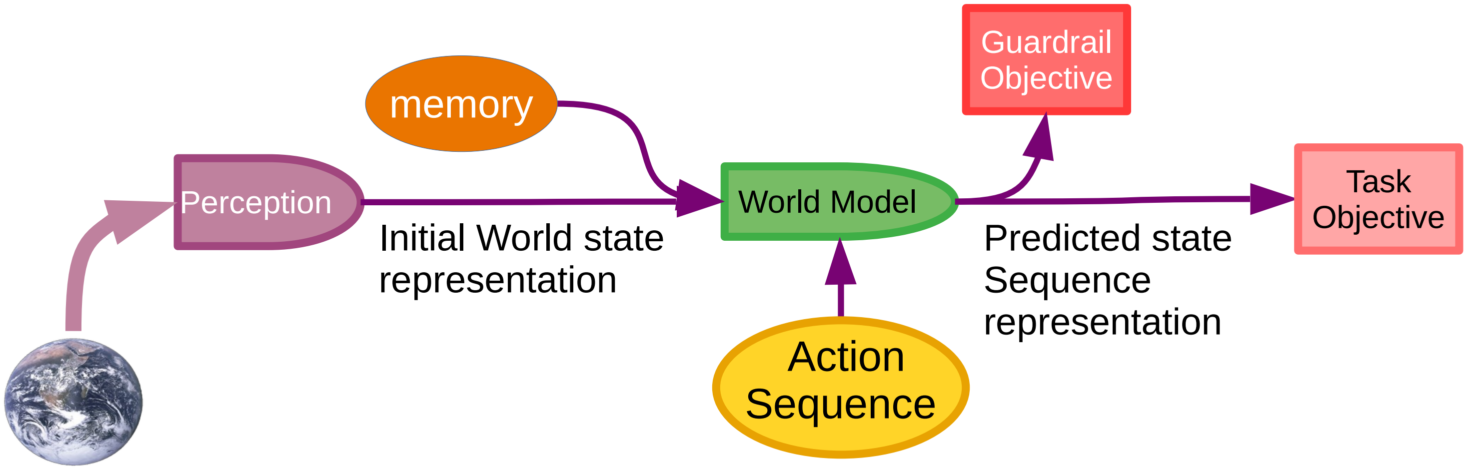




$$\check{y} = \mathrm{argmin}_y F(x, y)$$

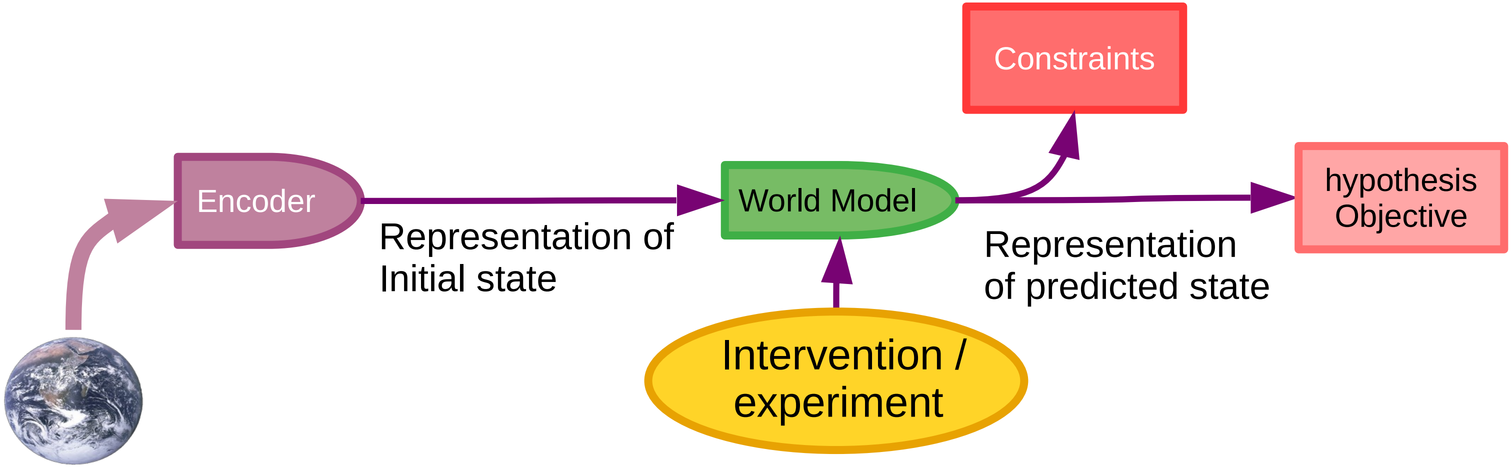
2. World Model for Planning/Reasoning

- **Perception:** Computes an abstract representation of the state of the world
  - Possibly combined with previously-acquired information in memory
- **World Model:** Predict the state resulting from an imagined action sequence
- **Task Objective:** Measures divergence to goal
- **Guardrail Objective:** Immutable objective terms that ensure safety
- **Operation:** Finds an action sequence that minimizes the objectives



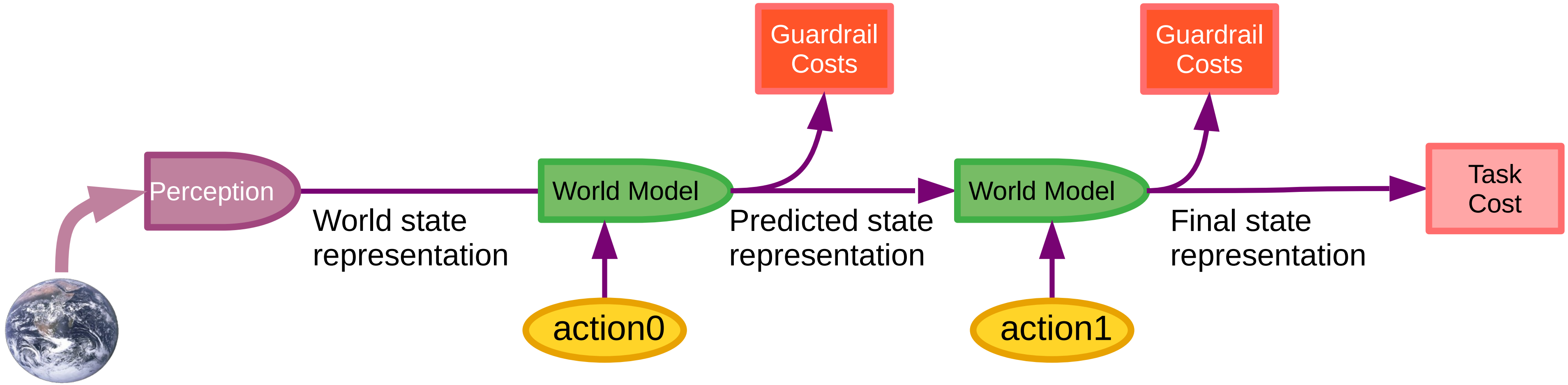
2. Models for Physics Experiments

- **Encoder:** Computes an abstract representation of the state of the system
- **World Model:** Predict the state resulting from an imagined experiment or intervention.
- **Hypothesis Objective:** Measures divergence to the result expected from the experiment
- **Constraints:** that the trajectory must satisfy.
- Find an action an experiment that validates or invalidates the hypothesis



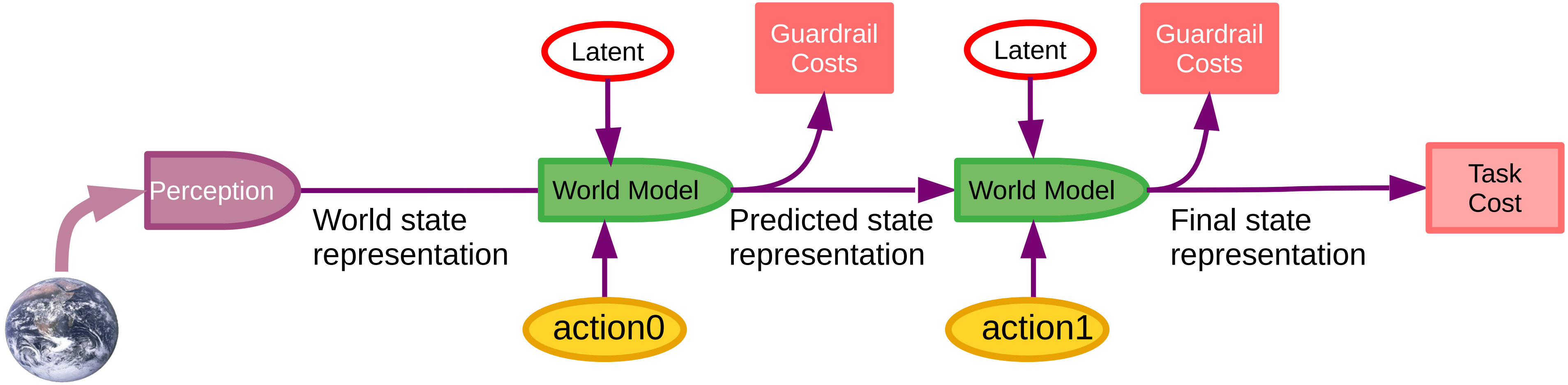
Objective-Driven AI: Multistep/Recurrent World Model

- **Same world model applied at multiple time steps**
- **Guardrail costs applied to entire state trajectory**
- **This is identical to Model Predictive Control (MPC)**
  - But with a trained world model
- **Action inference by minimization of the objectives**
  - Using gradient-based method, graph search, dynamic prog, A*, MCTS,….



Objective-Driven AI: Non-Deterministic World Model

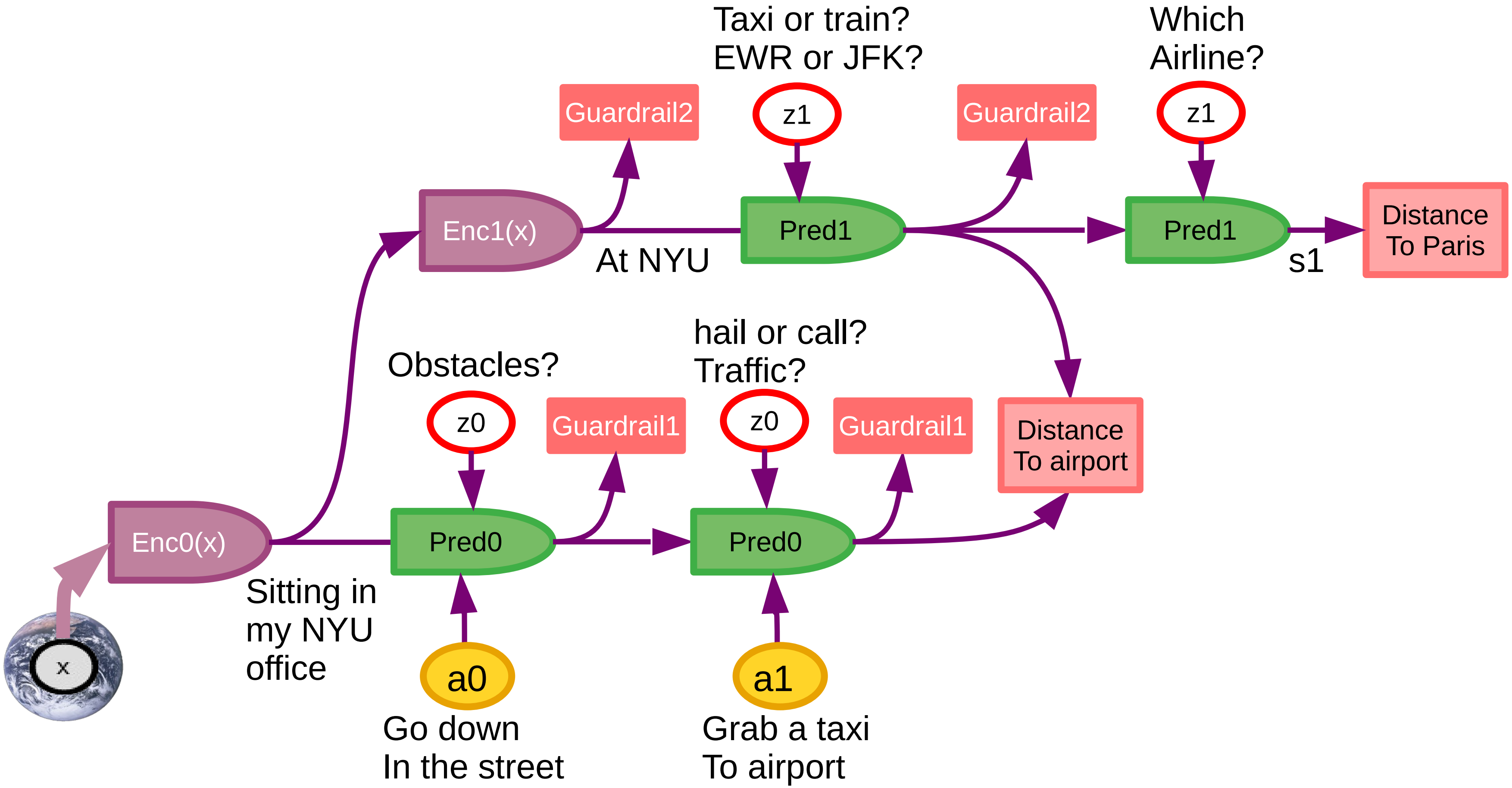
- **The world is not deterministic or fully predictable**
- **Latent variables parameterize the set of plausible predictions**
  - Can be sampled from a prior or swept through a set.
  - Planning can be done for worst case or average case
  - Uncertainty in outcome can be predicted and quantified



Y. LeCun

Objective-Driven AI: Hierarchical Planning

- **Hierarchical Planning: going from NYU to Paris**



Objective-Driven AI Systems

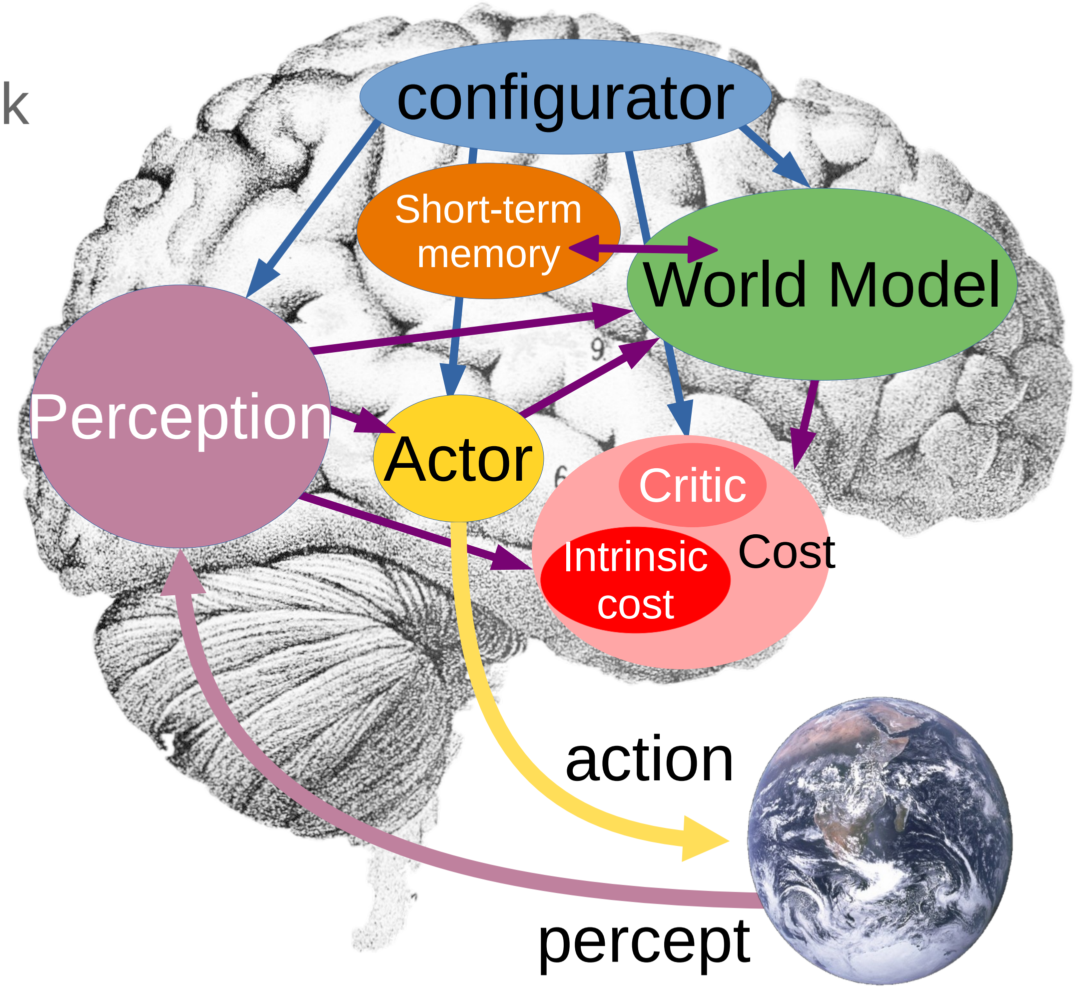
AI that can learn, understand the world,
reason, plan,
Yet is safe and controllable

“A path towards autonomous machine intelligence”
https://openreview.net/forum?id=BZ5a1r-kVsf

[previous versions of this talk available on YouTube]

Modular Cognitive Architecture for AMI

- **Configurator**
  - Configures other modules for task
- **Perception**
  - Estimates state of the world
- **World Model**
  - Predicts future world states
- **Cost**
  - Compute "discomfort"
- **Actor**
  - Find optimal action sequences
- **Short-Term Memory**
  - Stores state-cost episodes

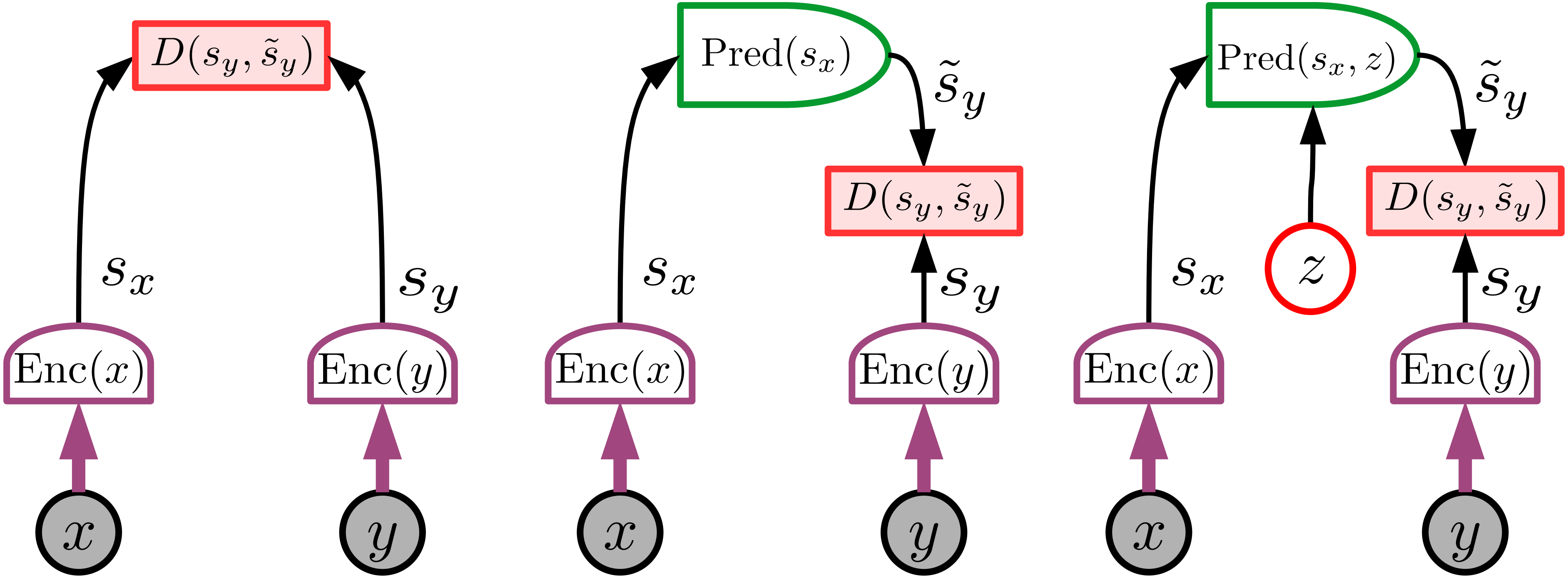


How could Machines Learn World Models from Observations?

Self-Supervised Learning

Joint Embedding Architectures

- Computes abstract representations for x and y
- Tries to make them equal or predictable from each other.



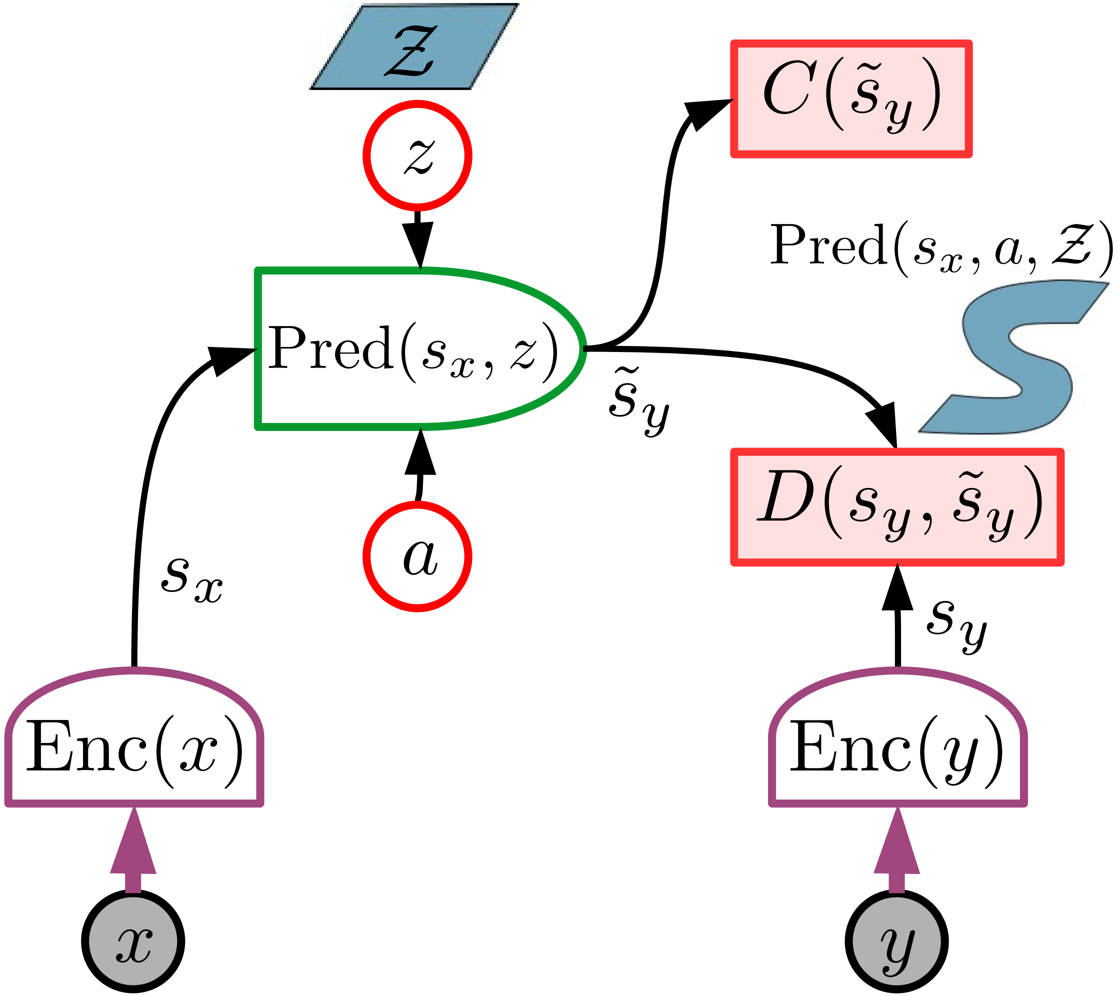
a) Joint Embedding Architecture (JEA)
Examples: Siamese Net, Pirl, MoCo, SimCLR, BarlowTwins, VICReg,

b) Deterministic Joint Embedding Predictive Architecture (DJEPA)
Examples: BYOL, VICRegL, I-JEPA

c) Joint Embedding Predictive Architecture (JEPA)
Examples: Equivariant VICReg
I-JEPA…..

Architecture for action-conditioned world models: 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



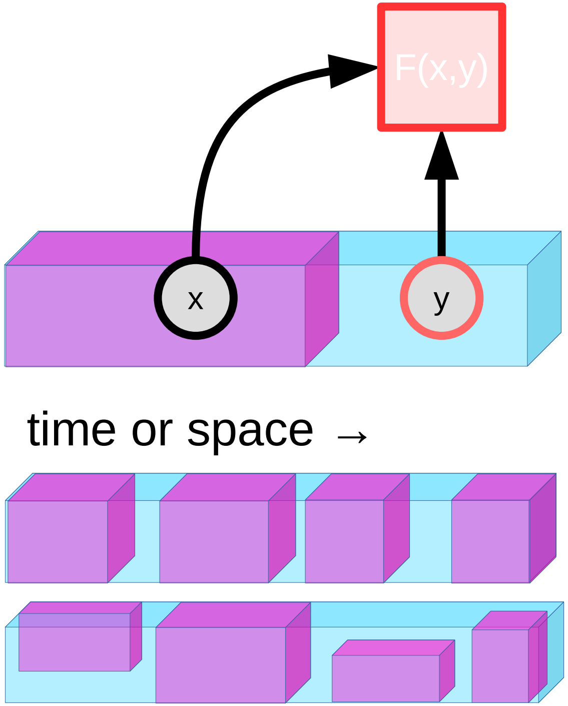
Energy-Based Models for Self-Supervised Learning

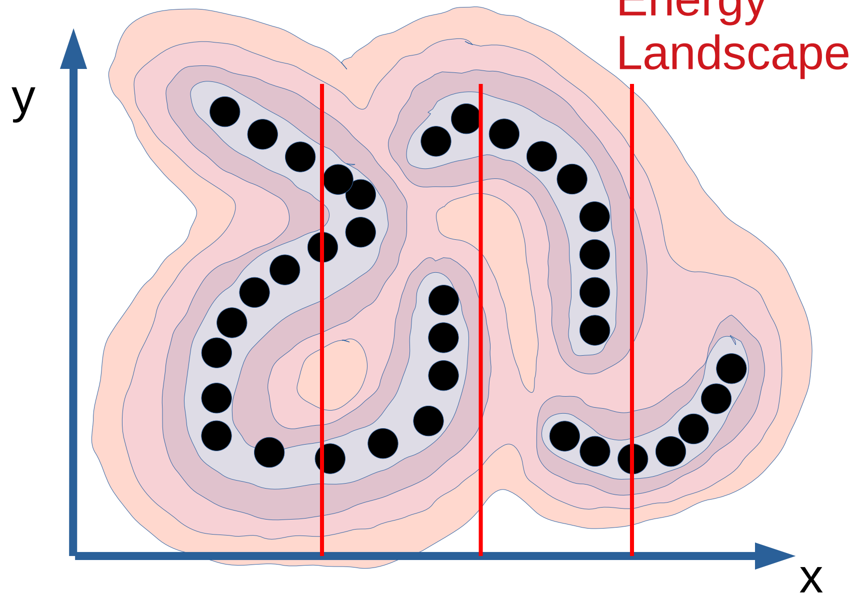
Capturing dependencies through an energy function

Probabilistic modeling is intractable in high-dimensional continuous domains.

Energy-Based Models: Implicit function

- **The only way to formalize & understand all model types**
  - Gives low energy to compatible pairs of x and y
  - Gives higher energy to incompatible pairs

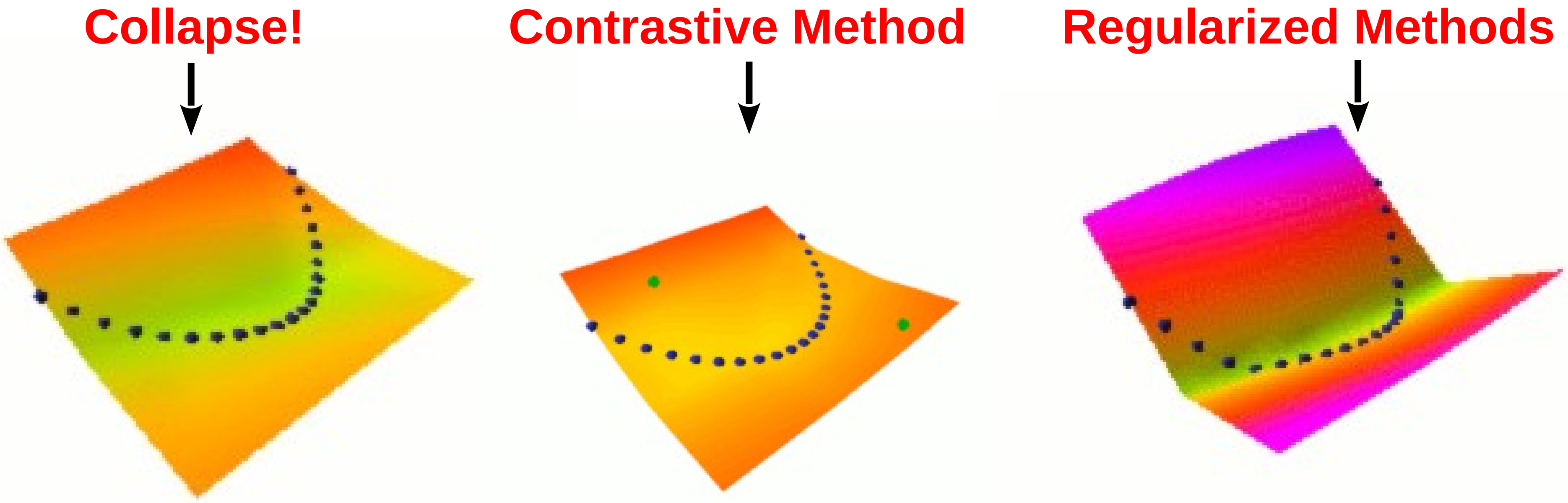




$$\check{y} = \operatorname{argmin}_y F(x, y)$$

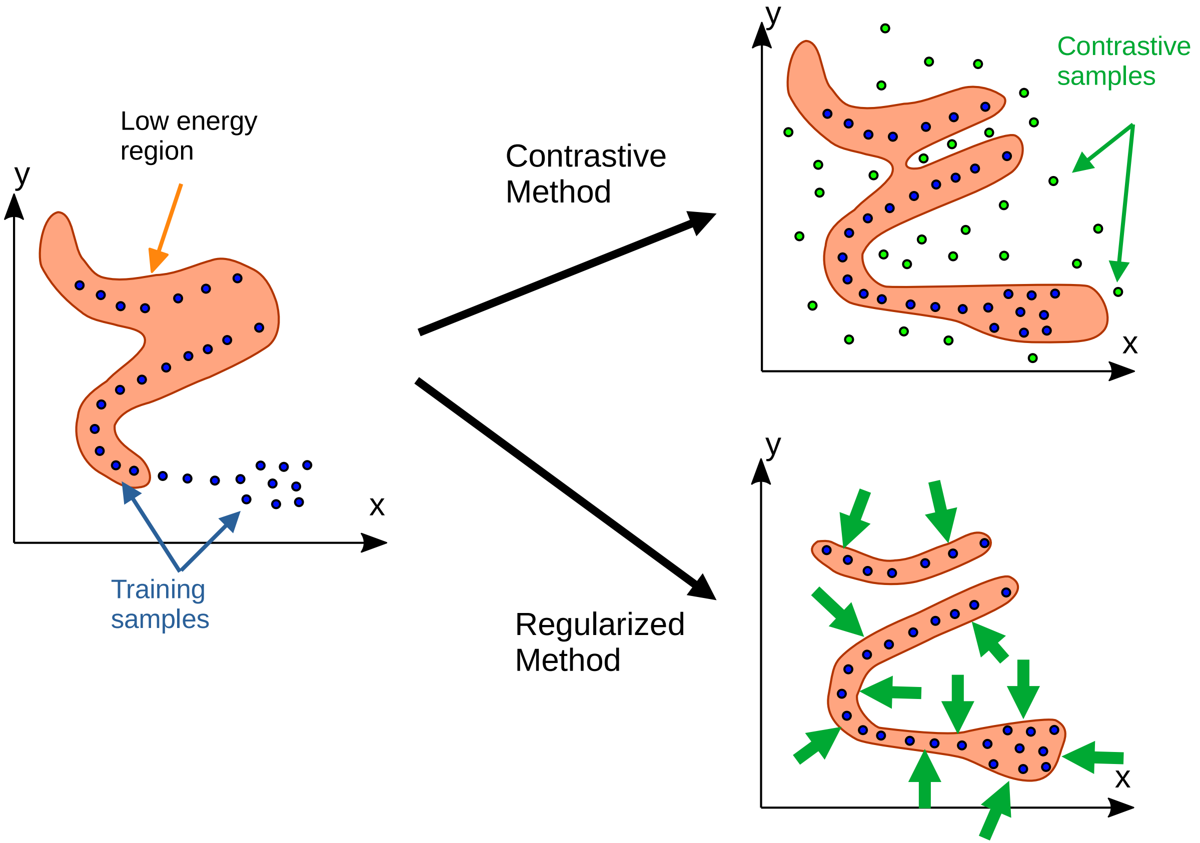
Training Energy-Based Models: Collapse Prevention

- **A flexible energy surface can take any shape.**
- **We need a loss function that shapes the energy surface so that:**
  - Data points have low energies
  - Points outside the regions of high data density have higher energies.



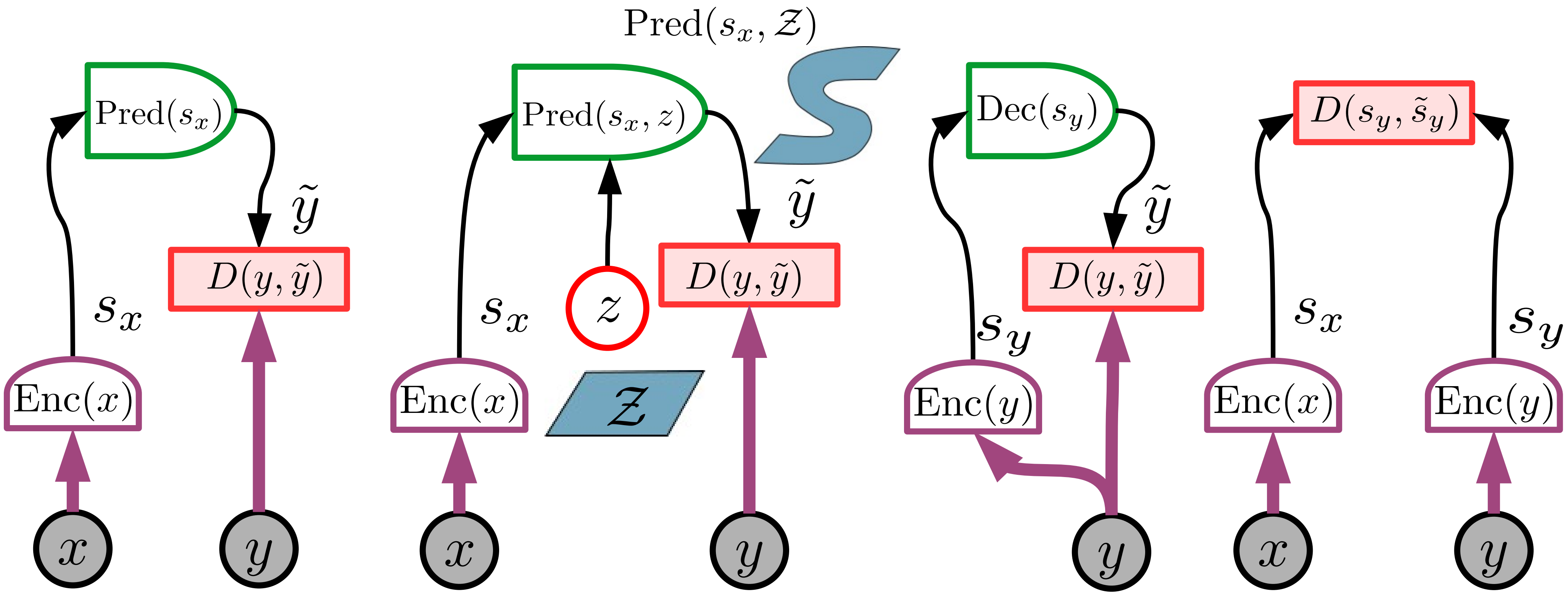
EBM Training: two categories of methods

- **Contrastive methods**
  - Push down on energy of training samples
  - Pull up on energy of suitably-generated contrastive samples
  - Scales very badly with dimension
- **Regularized Methods**
  - Regularizer minimizes the volume of space that can take low energy



EBM Architectures

- **Some architectures can lead to a collapse of the energy surface**



a) Prediction / regression
NO COLLAPSE

b) Generative latent-variable Architecture
CAN COLLAPSE

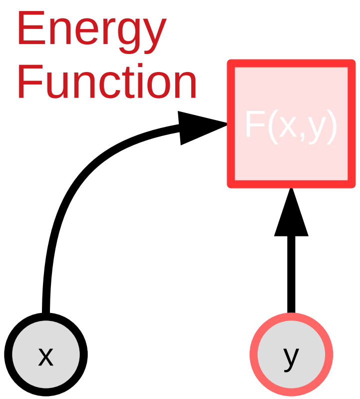
c) Auto-Encoder
CAN COLLAPSE

d) Joint Embedding Architecture
CAN COLLAPSE

Energy-Based Models vs Probabilistic Models

- **Probabilistic models are a special case of EBM**
  - Energies are like un-normalized negative log probabilities
- **Why use EBM instead of probabilistic models?**
  - EBM gives more flexibility in the choice of the scoring function.
  - More flexibility in the choice of objective function for learning
- **From energy to probability: Gibbs-Boltzmann distribution**
  - Beta is a positive constant

$$P(y|x) = \frac{e^{-\beta F(x,y)}}{\int_{y'} e^{-\beta F(x,y')}}$$

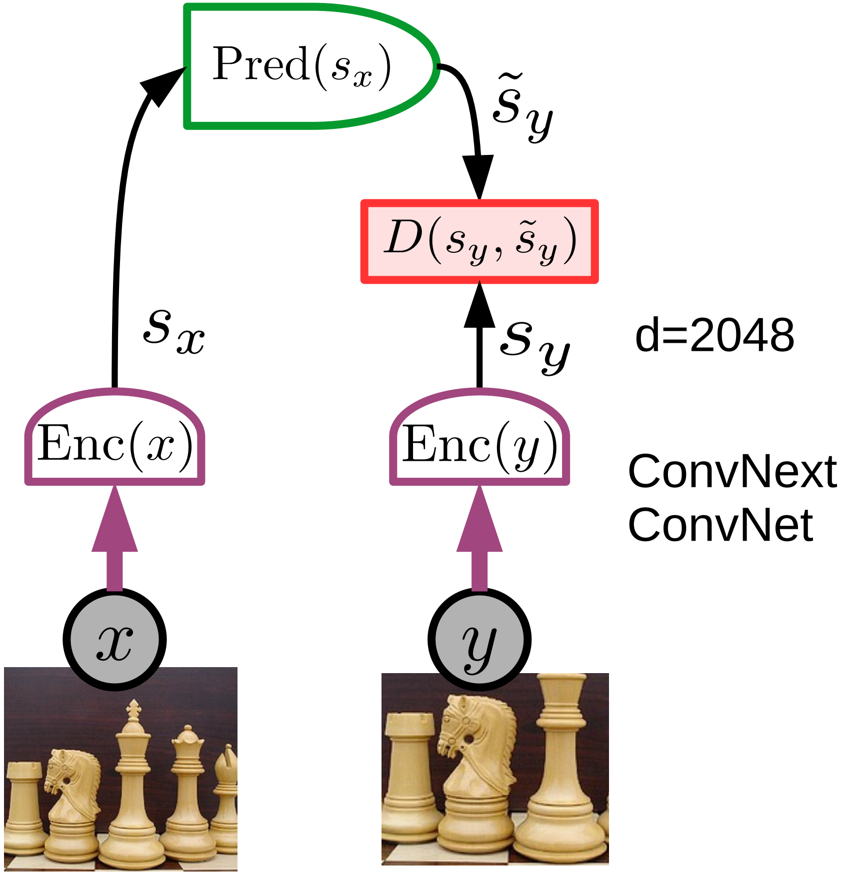


Contrastive Methods vs Regularized/Architectural Methods

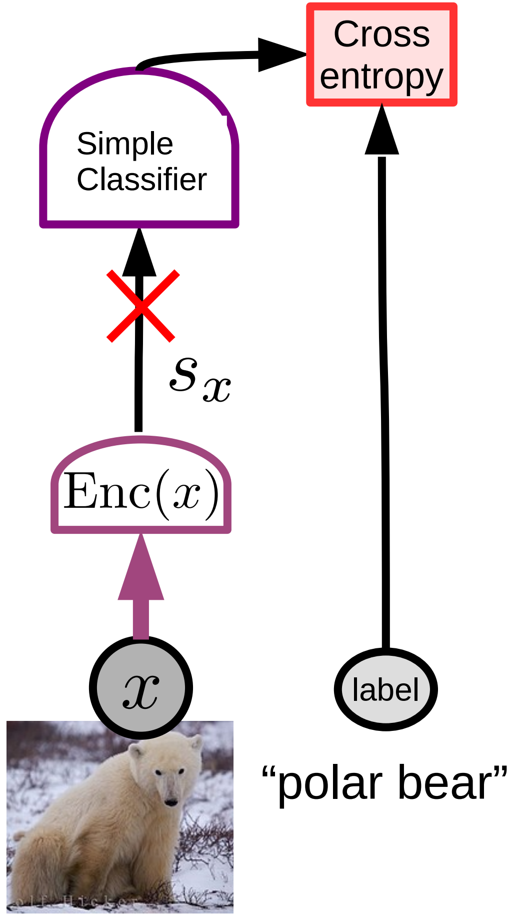
- **Contrastive: [they all are different ways to pick which points to push up]**
  - C1: push down of the energy of data points, push up everywhere else: Max likelihood (needs tractable partition function or variational approximation)
  - C2: push down of the energy of data points, push up on chosen locations: max likelihood with MC/MMC/HMC, Contrastive divergence, Metric learning/Siamese nets, Ratio Matching, Noise Contrastive Estimation, Min Probability Flow, adversarial generator/GANs
  - C3: train a function that maps points off the data manifold to points on the data manifold: denoising auto-encoder, masked auto-encoder (e.g. BERT)
- **Regularized/Architectural: [Different ways to limit the information capacity of the latent representation]**
  - A1: build the machine so that the volume of low energy space is bounded: PCA, K-means, Gaussian Mixture Model, Square ICA, normalizing flows…
  - A2: use a regularization term that measures the volume of space that has low energy: Sparse coding, sparse auto-encoder, LISTA, Variational Auto-Encoders, discretization/VQ/VQVAE.
  - A3: F(x,y) = C(y, G(x,y)), make G(x,y) as "constant" as possible with respect to y: Contracting auto-encoder, saturating auto-encoder
  - A4: minimize the gradient and maximize the curvature around data points: score matching

SSL-Pretrained Joint Embedding for Image Recognition

JEPA/JEA pretrained with SSL

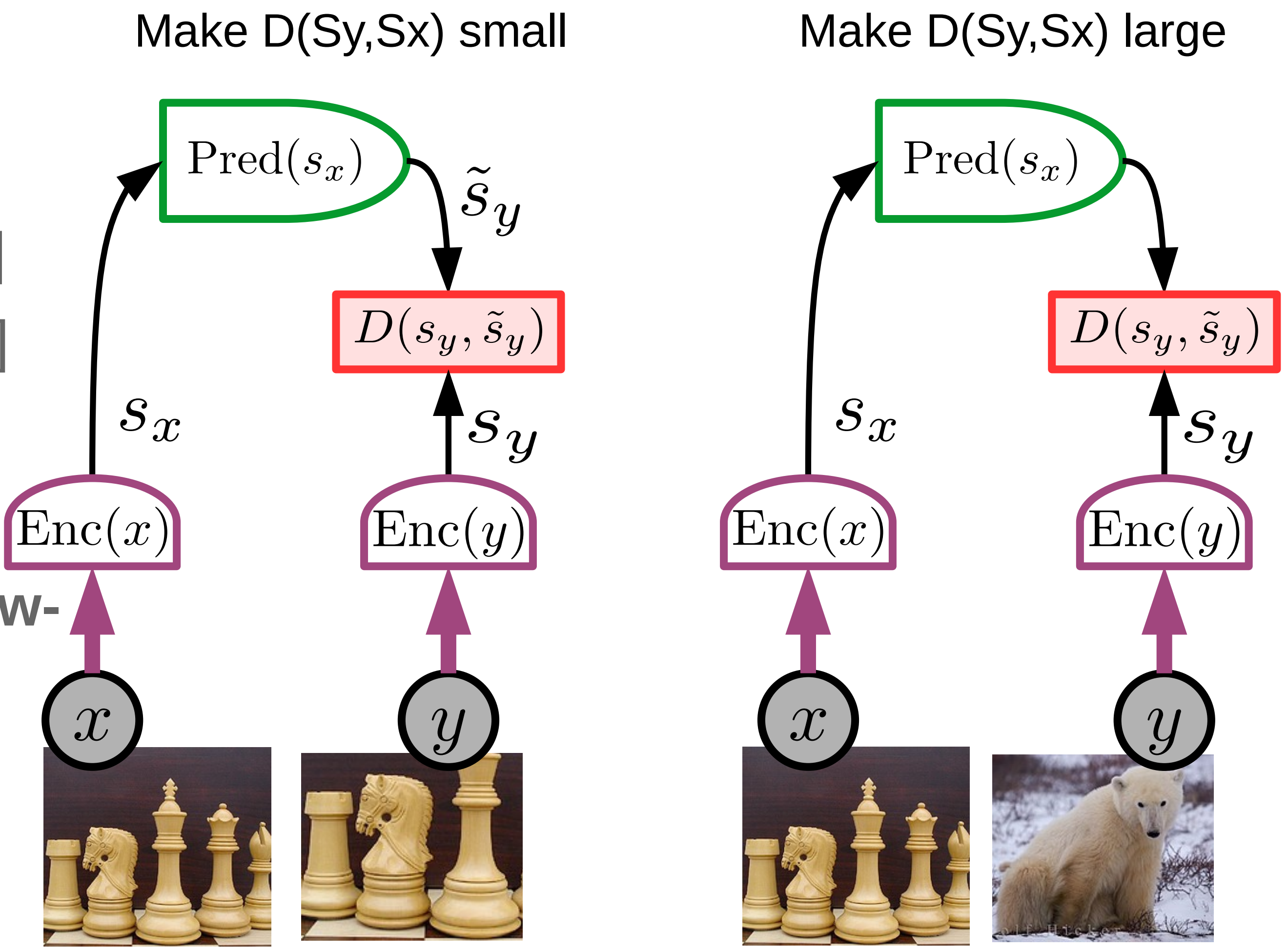


Training a supervised classification head



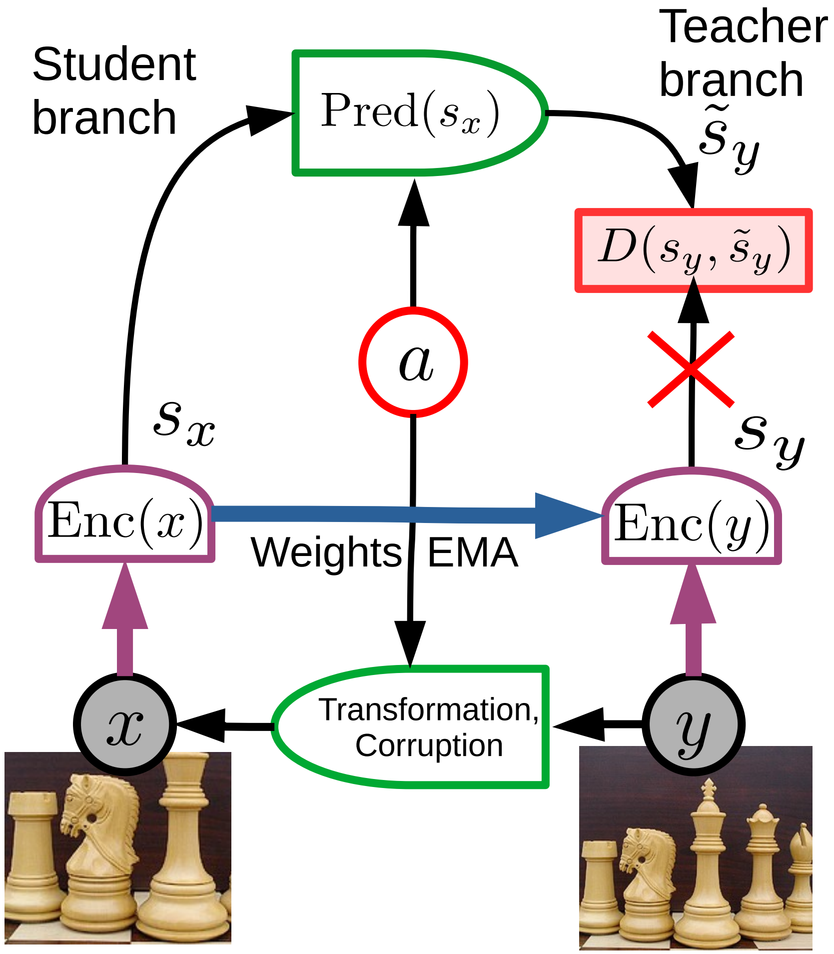
(Sample) Contrastive Joint Embedding

- **Example:**
  - Siamese Networks [Bromley NIPS 1993]
    [Chopra CVPR 2005]
    [Hadsell CVPR 2006]
  - SimCLR
    [Chen 2020]
- **Can only produce low-dimensional image representations**
  - Around 200 D.



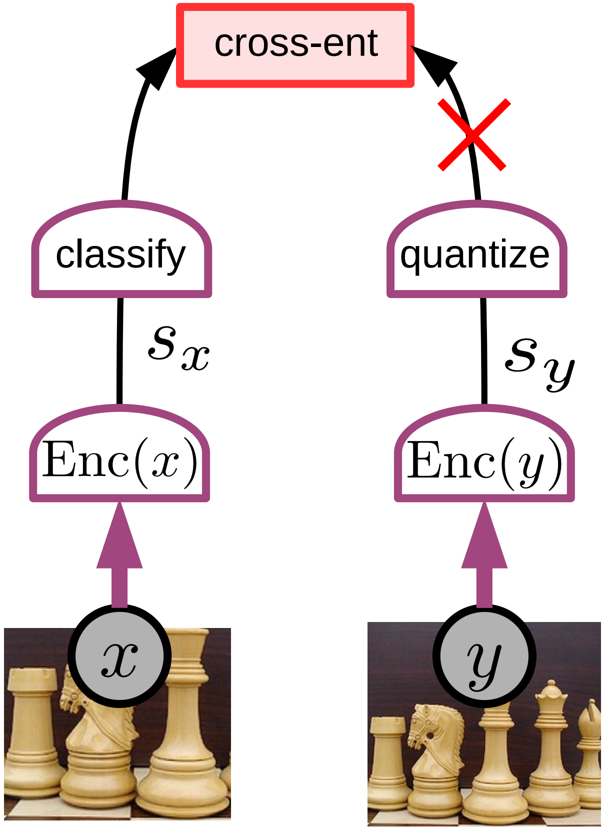
Distillation Methods

- **Distillation-based SSL:**
  - Bootstrap Your Own Latents [Grill arXiv:2006.07733]
  - SimSiam [Chen & He arXiv:2011.10566]
  - DINOv2 [Oquab arXiv:2304.07193]
  - I-JEPA [Assran 2023]
  - V-JEPA [Bardes 2024]
- **Advantages**
  - No negative samples, fast
- **Disadvantage:**
  - we don't completely understand why it works! [Tian et al. ArXiv:2102.06810]



DINOv2: Joint Embedding Architecture

- **SSL by distillation**



Method	Arch.	Data	Text sup.	kNN val	linear val	linear ReaL	linear V2
Weakly supervised							
CLIP	ViT-L/14	WIT-400M	✓	79.8	84.3	88.1	75.3
CLIP	ViT-L/14$_{336}$	WIT-400M	✓	80.5	85.3	88.8	75.8
SWAG	ViT-H/14	IG3.6B	✓	82.6	85.7	88.7	77.6
OpenCLIP	ViT-H/14	LAION	✓	81.7	84.4	88.4	75.5
OpenCLIP	ViT-G/14	LAION	✓	83.2	86.2	89.4	77.2
EVA-CLIP	ViT-g/14	custom*	✓	**83.5**	86.4	89.3	77.4
Self-supervised							
MAE	ViT-H/14	INet-1k	✗	49.4	76.6	83.3	64.8
DINO	ViT-S/8	INet-1k	✗	78.6	79.2	85.5	68.2
SEERv2	RG10B	IG2B	✗	–	79.8	–	–
MSN	ViT-L/7	INet-1k	✗	79.2	80.7	86.0	69.7
EsViT	Swin-B/W=14	INet-1k	✗	79.4	81.3	87.0	70.4
Mugs	ViT-L/16	INet-1k	✗	80.2	82.1	86.9	70.8
iBOT	ViT-L/16	INet-22k	✗	72.9	82.3	87.5	72.4
DINOv2	ViT-S/14	LVD-142M	✗	79.0	81.1	86.6	70.9
	ViT-B/14	LVD-142M	✗	82.1	84.5	88.3	75.1
	ViT-L/14	LVD-142M	✗	**83.5**	86.3	89.5	78.0
	ViT-g/14	LVD-142M	✗	**83.5**	**86.5**	**89.6**	**78.4**

DINO-style SSL scales & surpasses Supervised Methods

- "Scaling Language-Free Visual Representation Learning" [Fan et al. ArXiv:2504.01017]
- Scales better with model size and training set size than CLIP-style SL

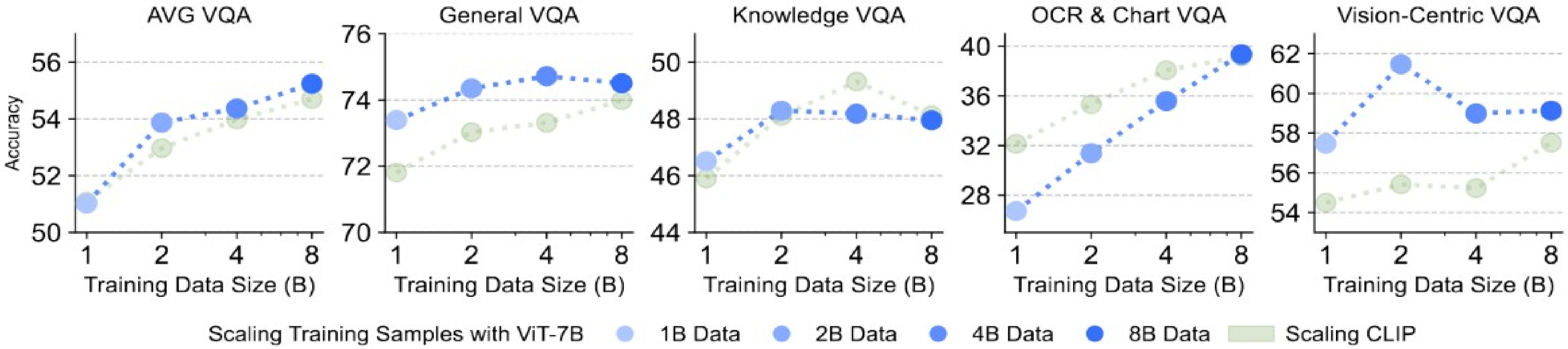


Figure 4 Scaling up examples seen when training Web-DINO-7B. Performance across different VQA categories as training data increases from 1B to 8B images. While General and Vision-Centric tasks show diminishing returns after 2B images, OCR & Chart tasks demonstrate continued improvement, contributing to steady gains in average performance. Further, Web-DINO consistently outperforms same-size (ViT-7B) CLIP models with different training samples seen. The x-axis plots training data size on a log-scale.

Canopy Height Map using DINOv2

- **Estimates tree canopy height from satellite images using DINOv2 features**
  - Using ground truth from Lidar images
  - 0.5 meter resolution images
- **[ArXiv:2304.07213]**
  - Tolan et al.: Sub-meter resolution canopy height maps using self-supervised learning and a vision transformer trained on Aerial and GEDI Lidar

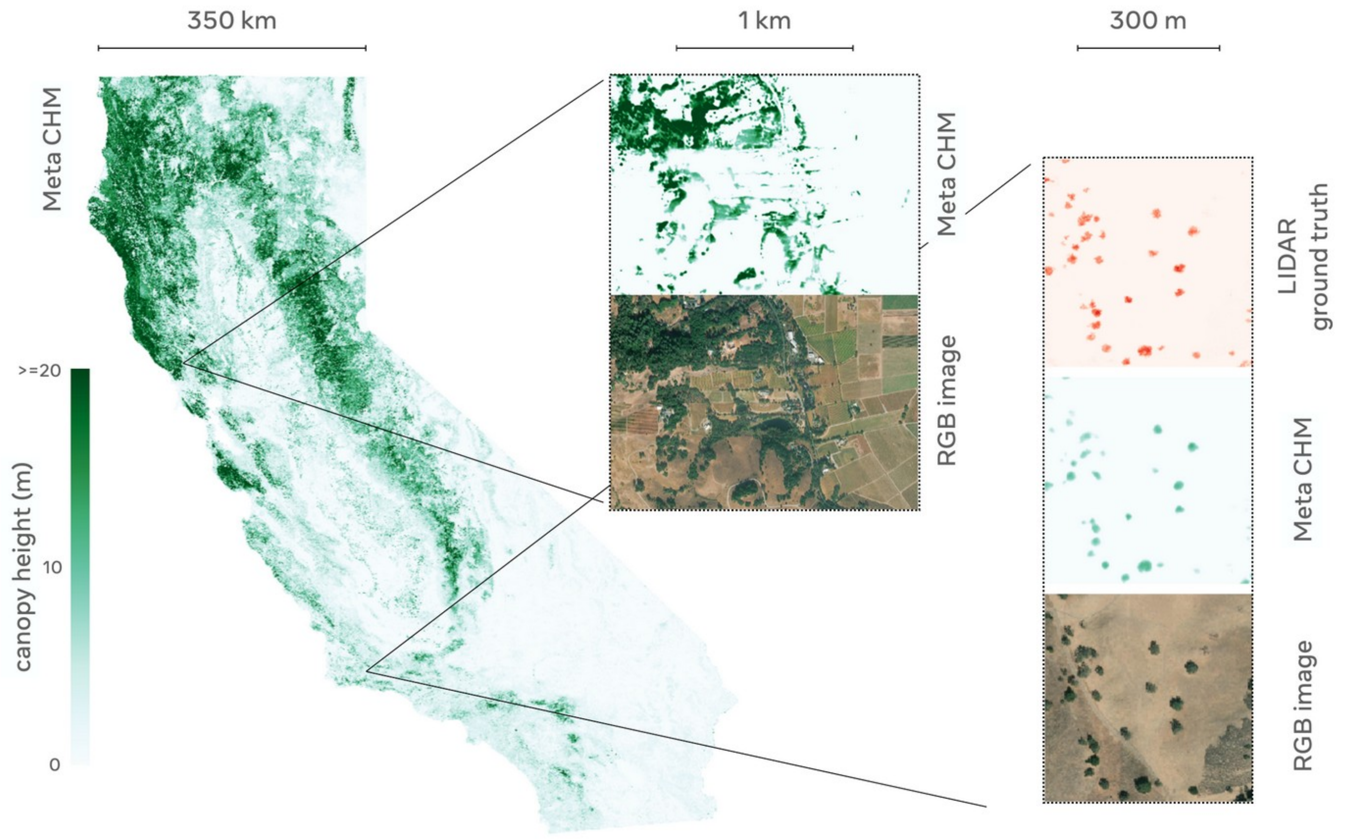
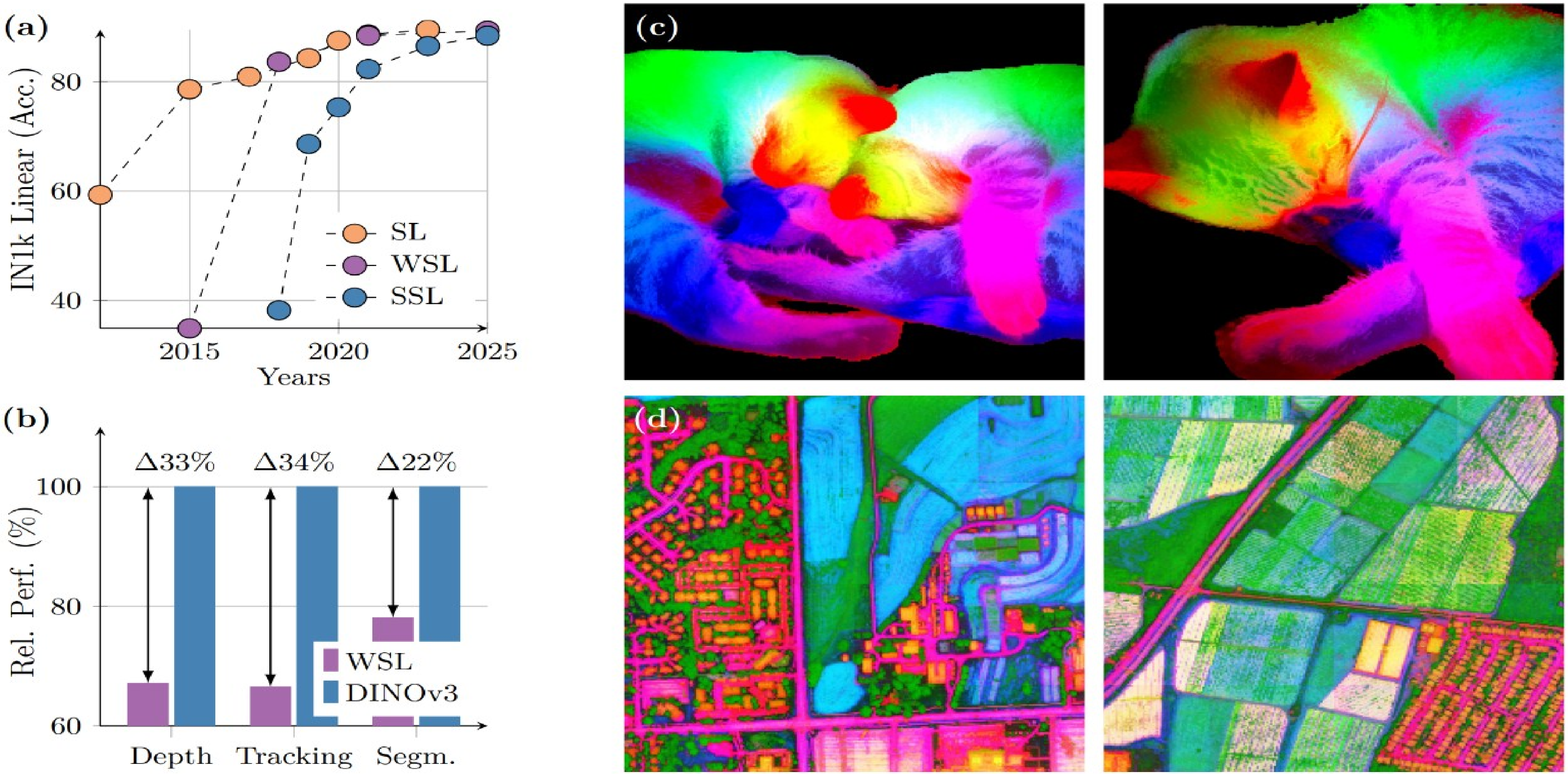


Figure 1: Canopy Height Map (CHM) for California, with inset showing zoomed-in region with input RGB imagery and LIDAR ground truth

Y. LeCun

DINOv3 [ArXiv:2508.10104] https://ai.meta.com/dinov3/



DINOv3 [ArXiv:2508.10104] https://ai.meta.com/dinov3/

TASK	BENCHMARK	DINO VIT-B/8 0.09B	DINOV2 VIT-G/14 1.1B	DINOV3 VIT-7B/16 7B	SIGLIP 2 VIT-G-OPT/16 1.8B	PE VIT-G/14 1.9B
Segmentation	ADE-20k	31.8	49.5	**55.9**	42.7	38.9
Depth estimation	NYU ↓	0.537	0.372	**0.309**	0.494	0.436
Video tracking	DAVIS	68.7	76.6	**83.3**	62.9	49.8
Instance retrieval	Met	17.1	44.6	**55.4**	13.9	10.6
Image classification	ImageNet ReaL	85.9	89.9	90.4	**90.5**	90.4
Image classification	ObjectNet	39.9	66.4	79.0	78.6	**80.2**
Fine-grained Image classification	iNaturalist 2021	68.3	86.1	**89.8**	82.7	87.0

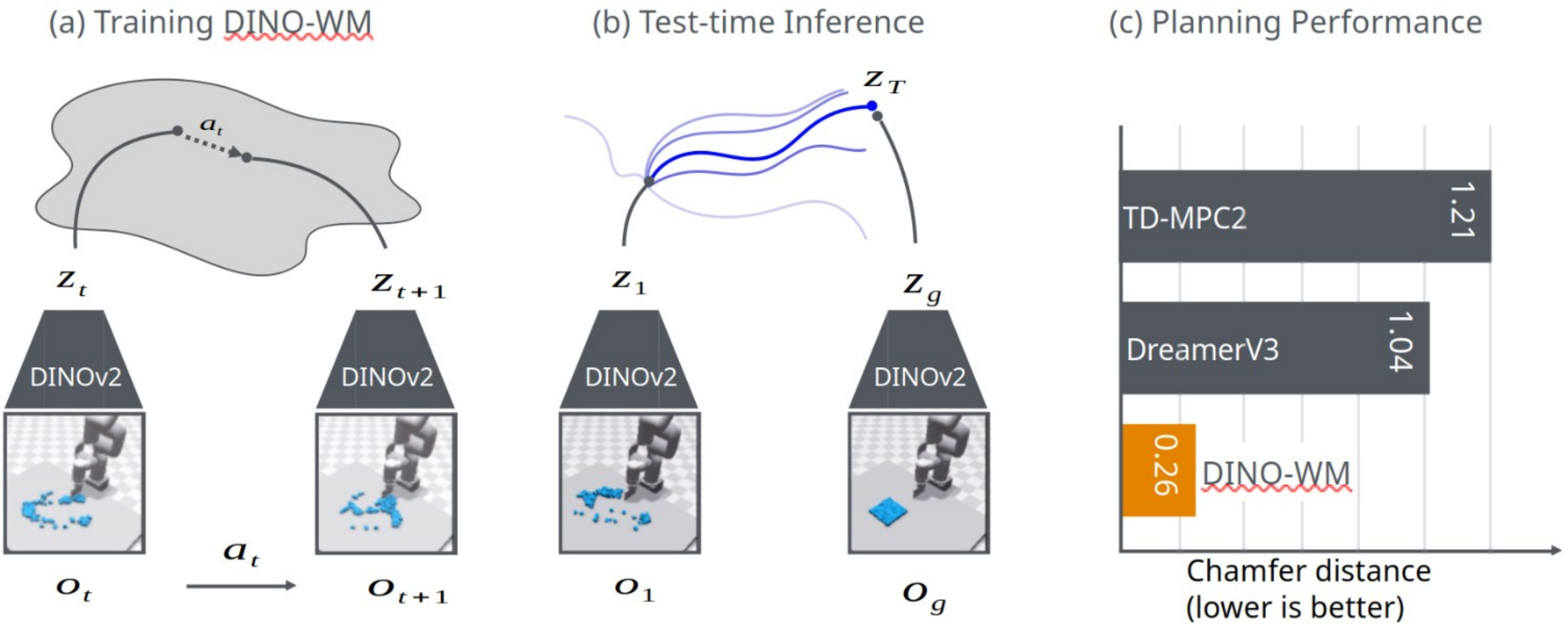
DINO-WM:
Action planning with a world model trained from DINO features

Model-Predictive Control with a trained predictor

[Gaoyue Zhou, Hengkai Pan, Yann LeCun, Lerrel Pinto, arXiv:2411.04983]

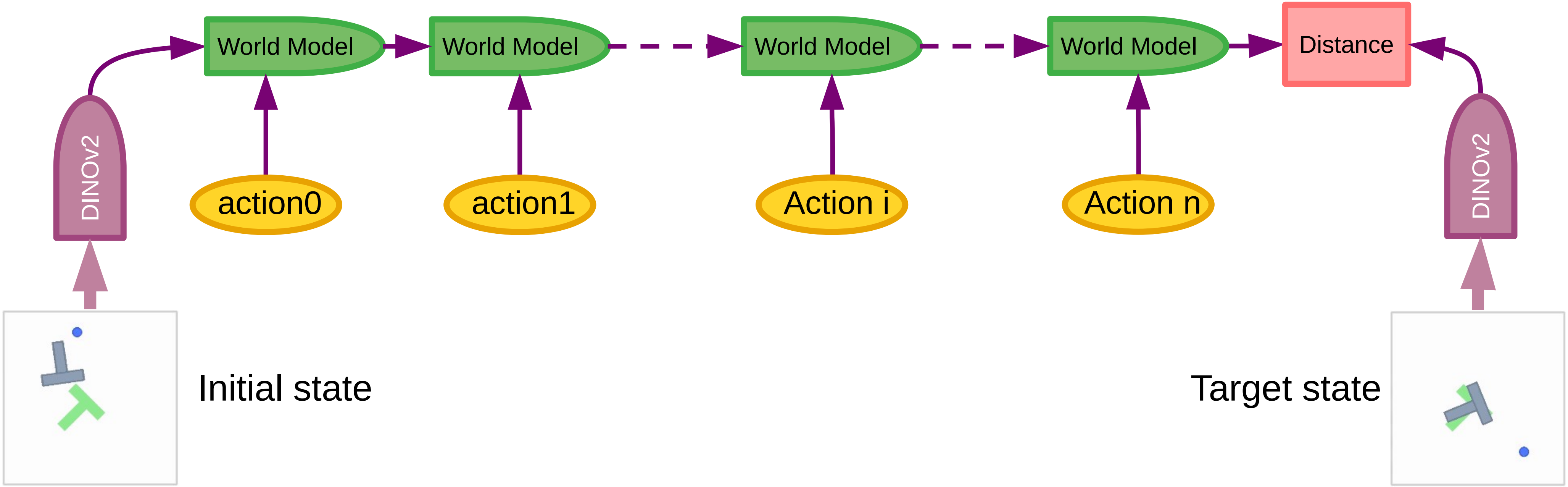
DINO-WM [https://dino-wm.github.io/]

- **Predictor: learns to predict the state of the world in representations space: z[t+1] = Pred(z[t], a[t])**

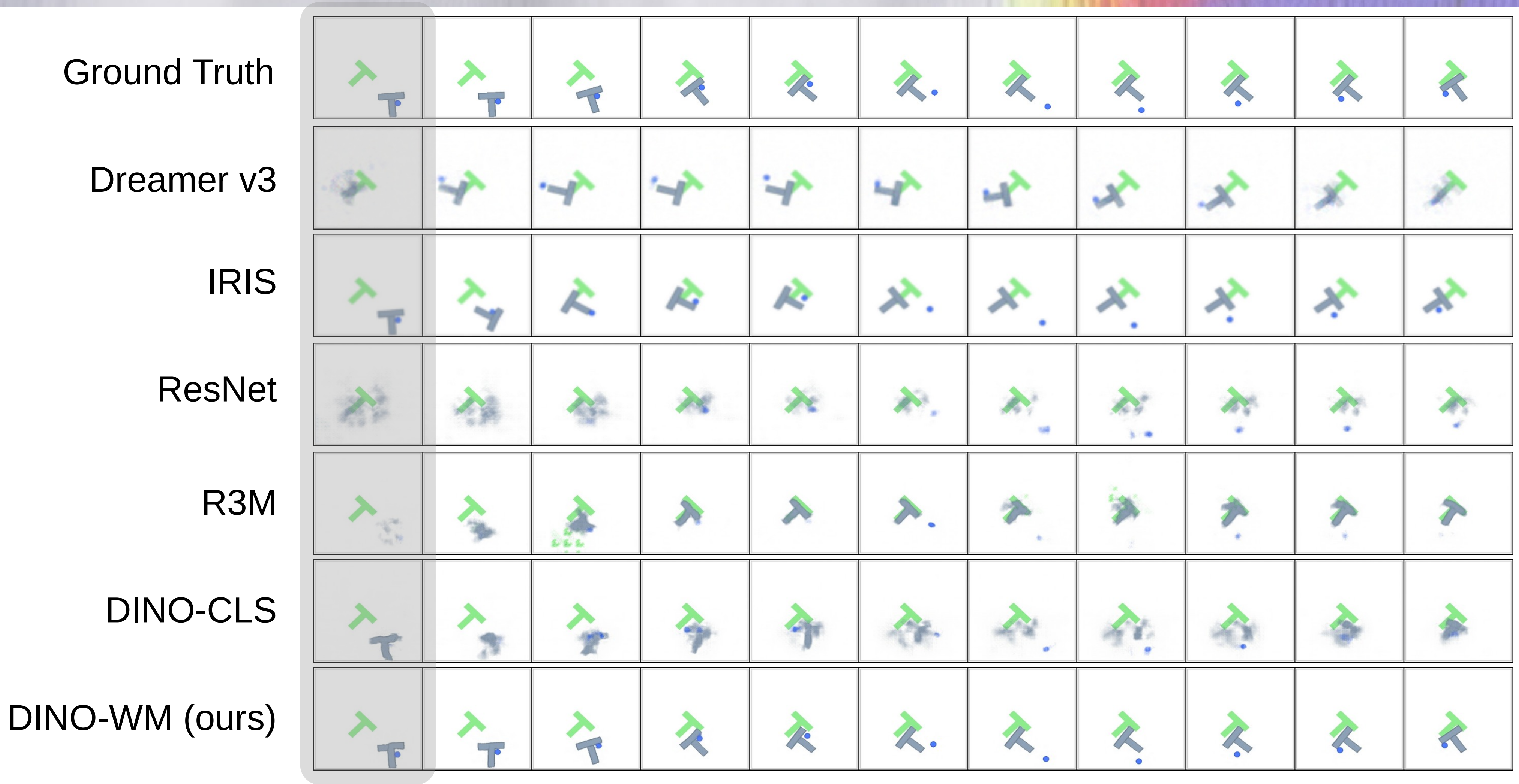


DINO-WM: Planning

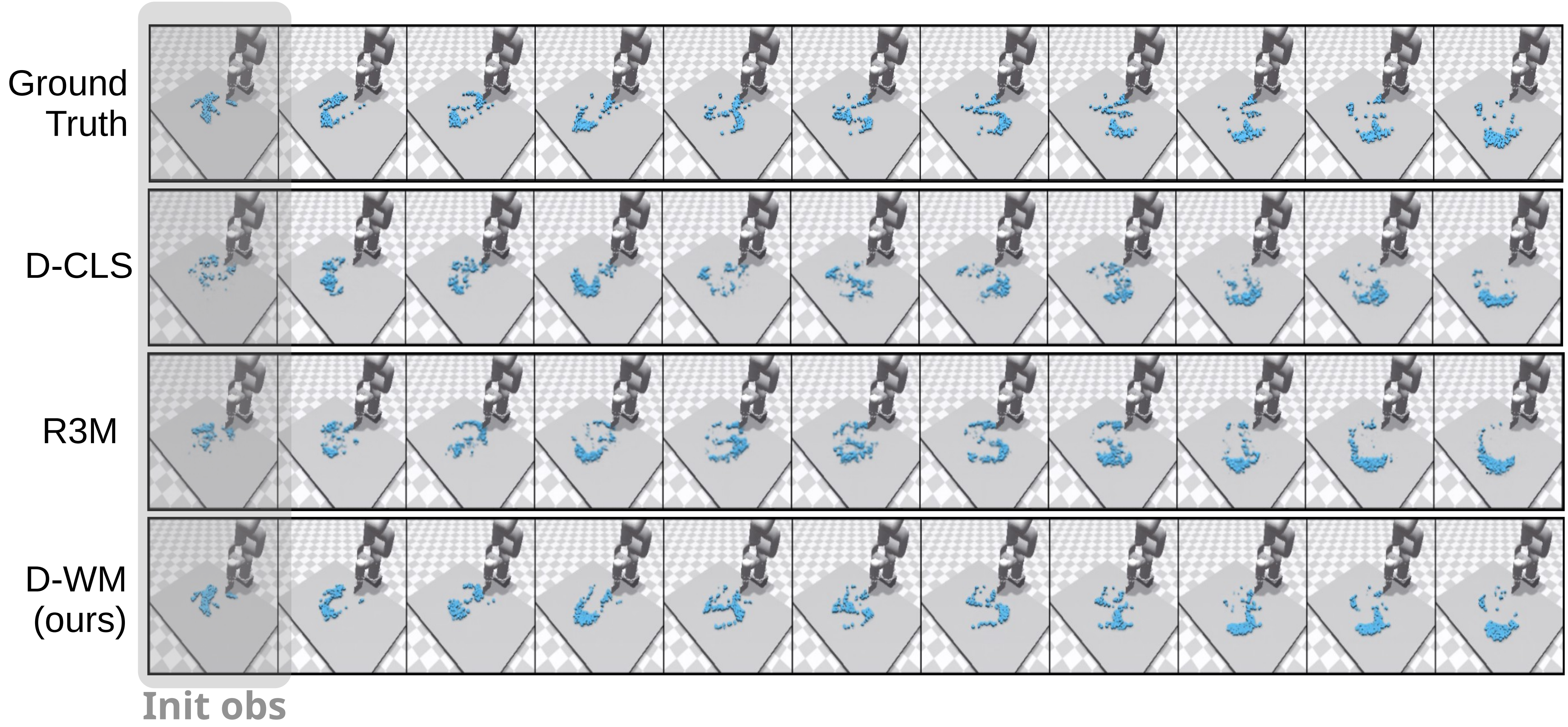
- Objective: minimize distance between predicted state and target state in representation space with respect to the action sequence.



DINO-WM: Open loop roll outs

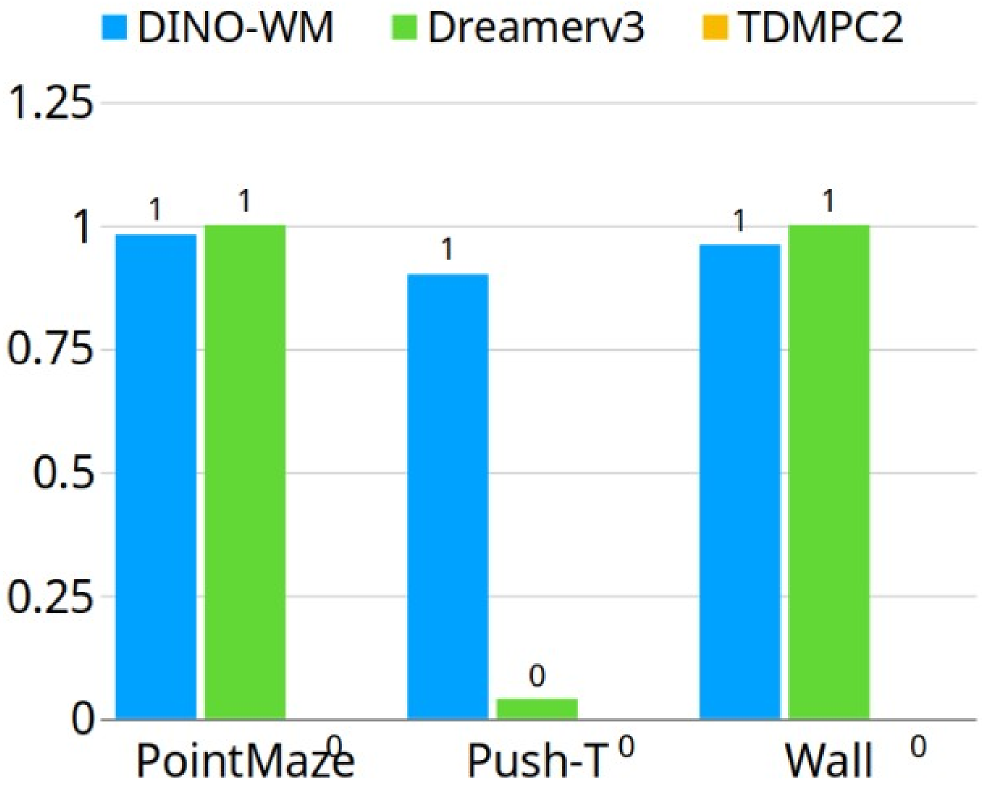


DINO-WM: Open loop roll outs

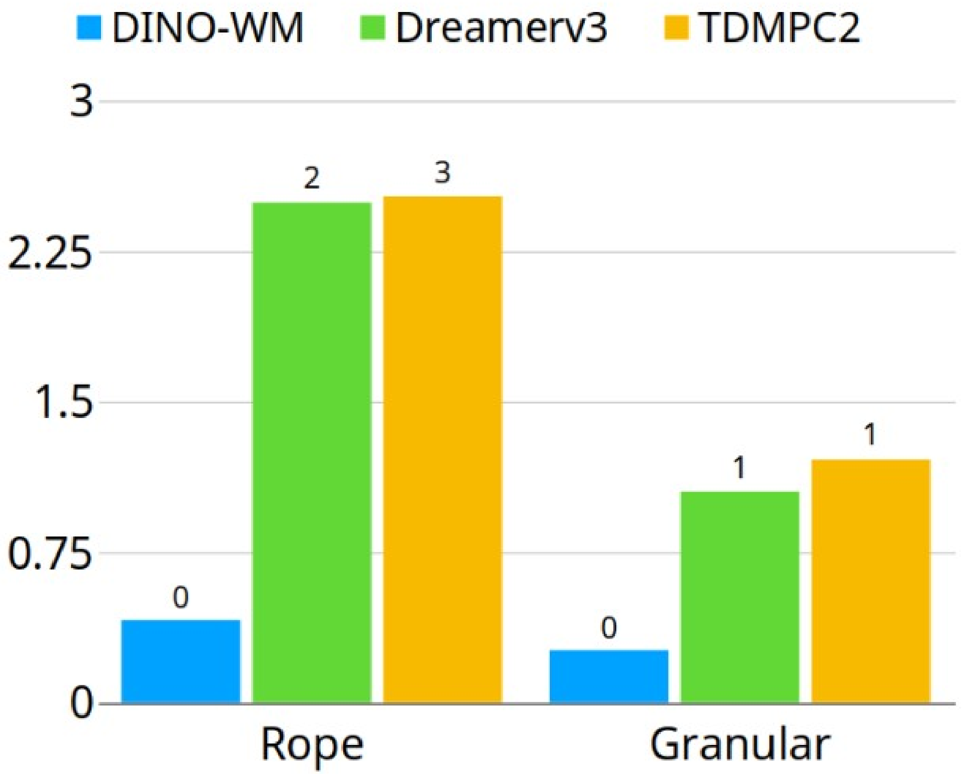


DINO-WM: optimizing behavior – part 1

- **Success rate**
  - (higher is better)

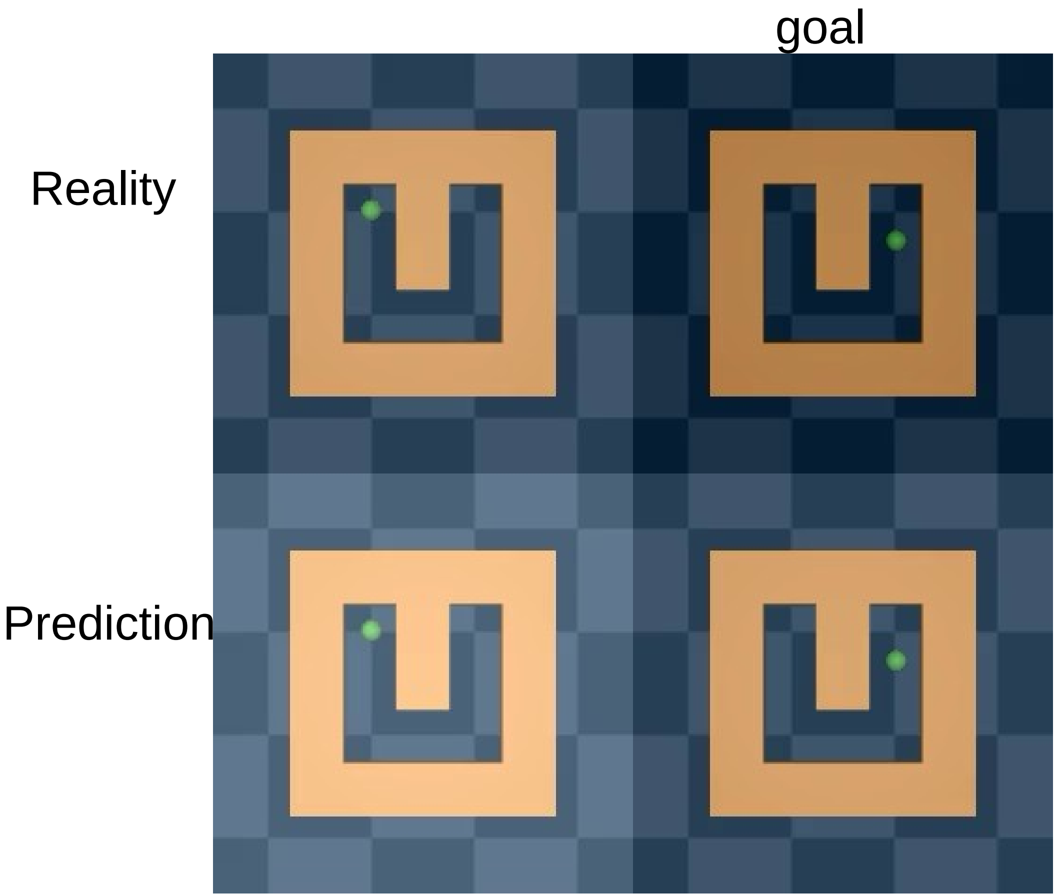


- **Chamfer distance**
  - (lower is better)

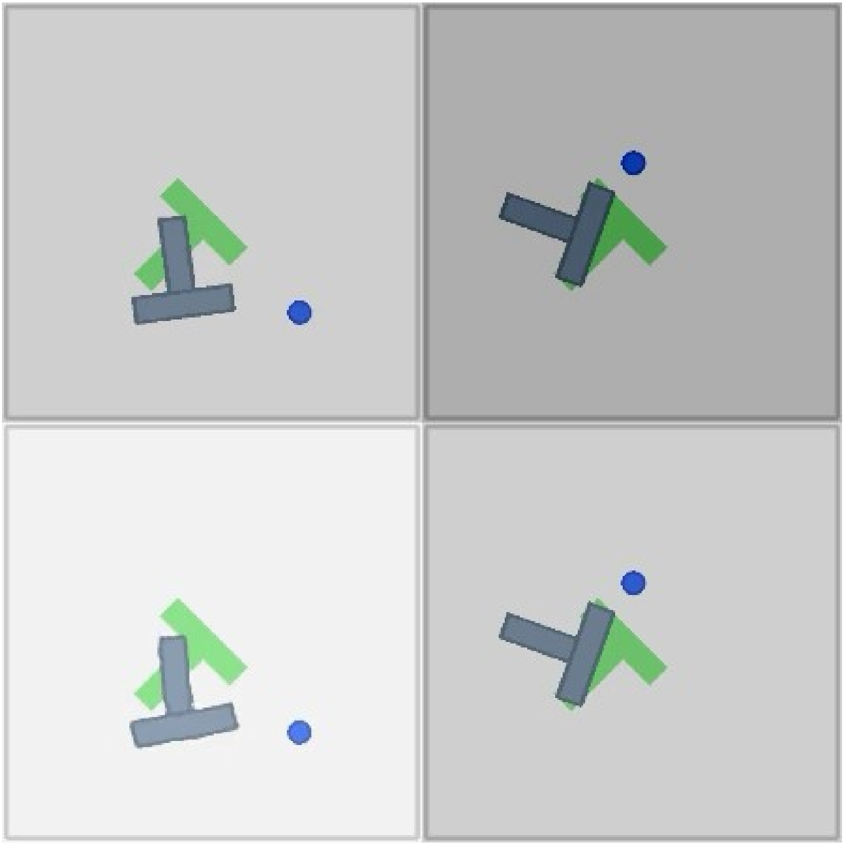


DINO-WM: Manipulation results

- **Point Maze**



- **Push T**



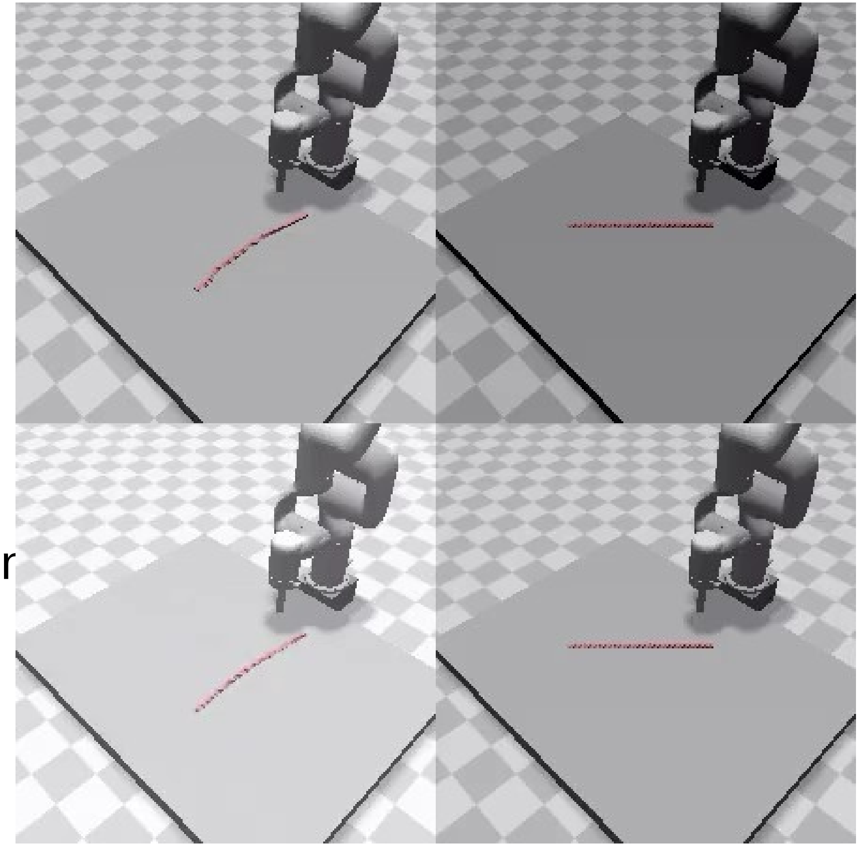
DINO-WM: Manipulation results

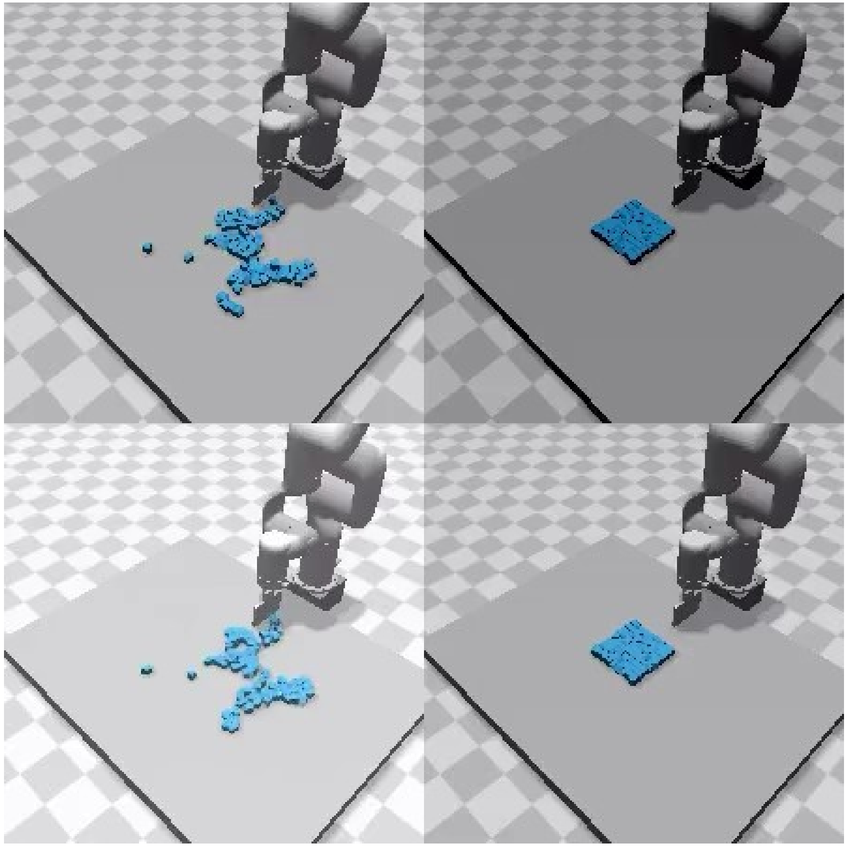
- **Rope**
- **Granular**

goal

Reality

Prediction

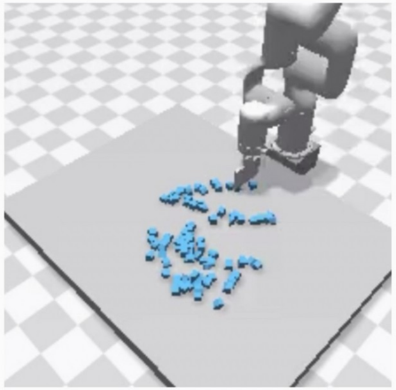




Planning with DINO-WM https://dino-wm.github.io/

Arbitrary Goals at Test Time

Initial State



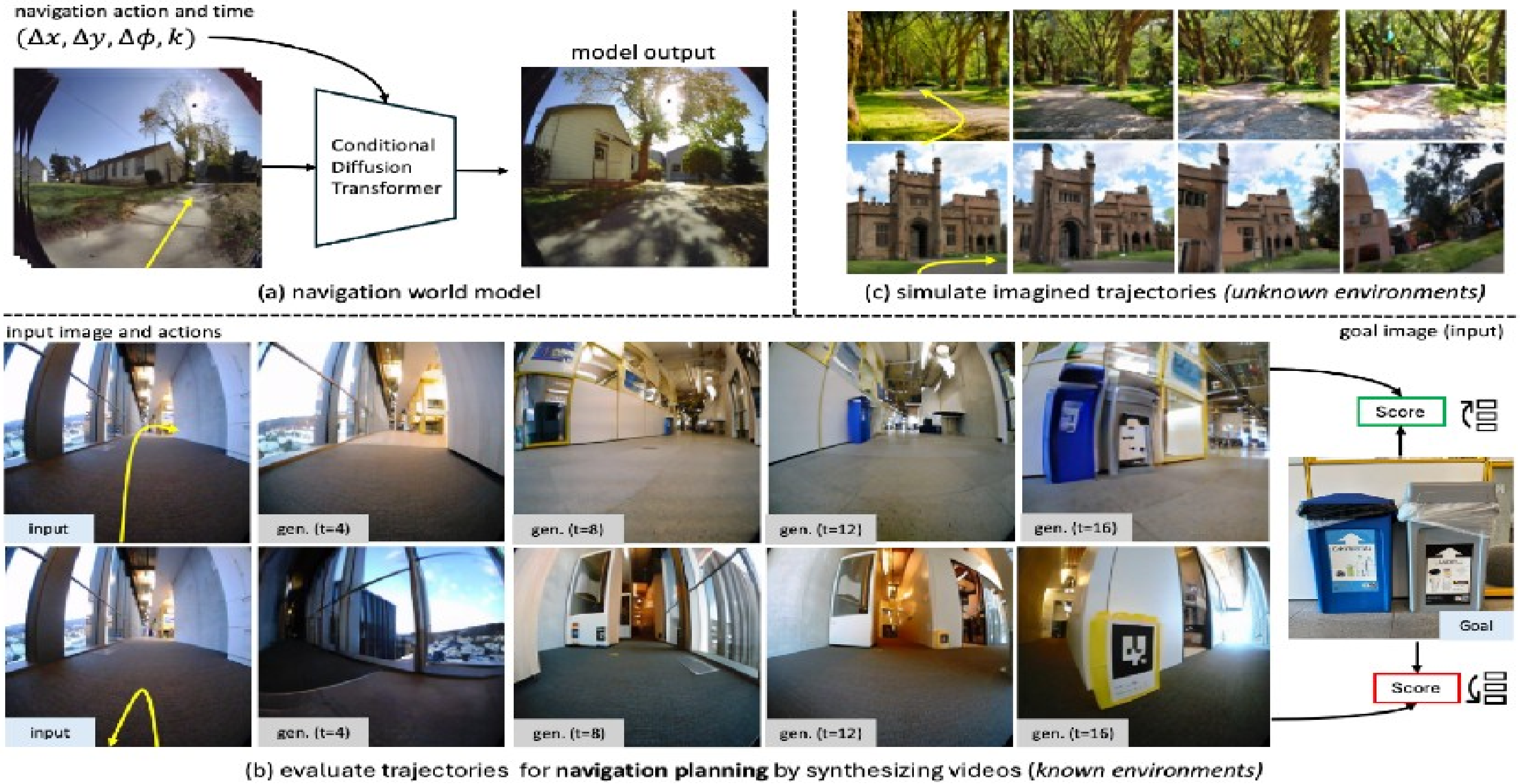
Navigation World Models

MPC planning from natural motion-conditioned videos

[Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, Yann LeCun, arXiv:2412.03572]

https://www.amirbar.net/nwm/

Navigation World Model



Generated Video Given a Motion Action Sequence

Navigation World Model Teaser Video

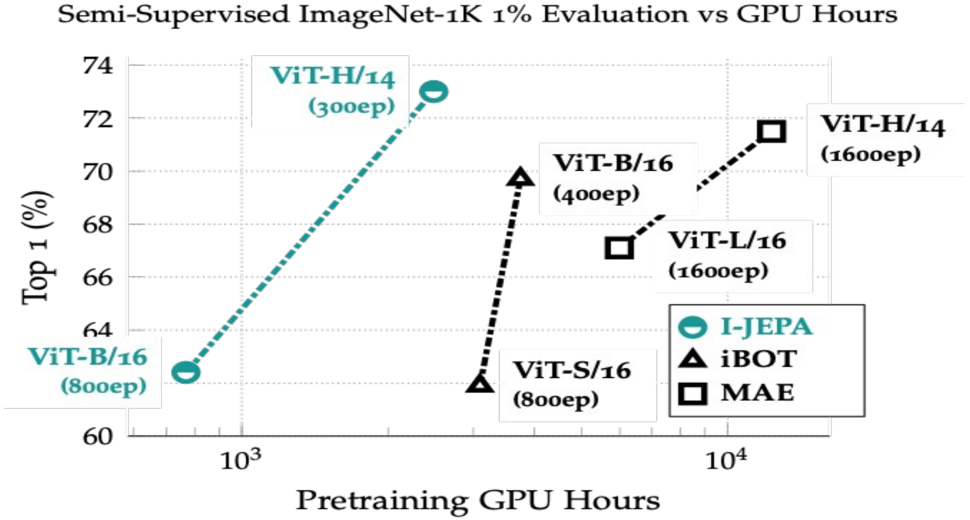
Image-JEPA & Video-JEPA

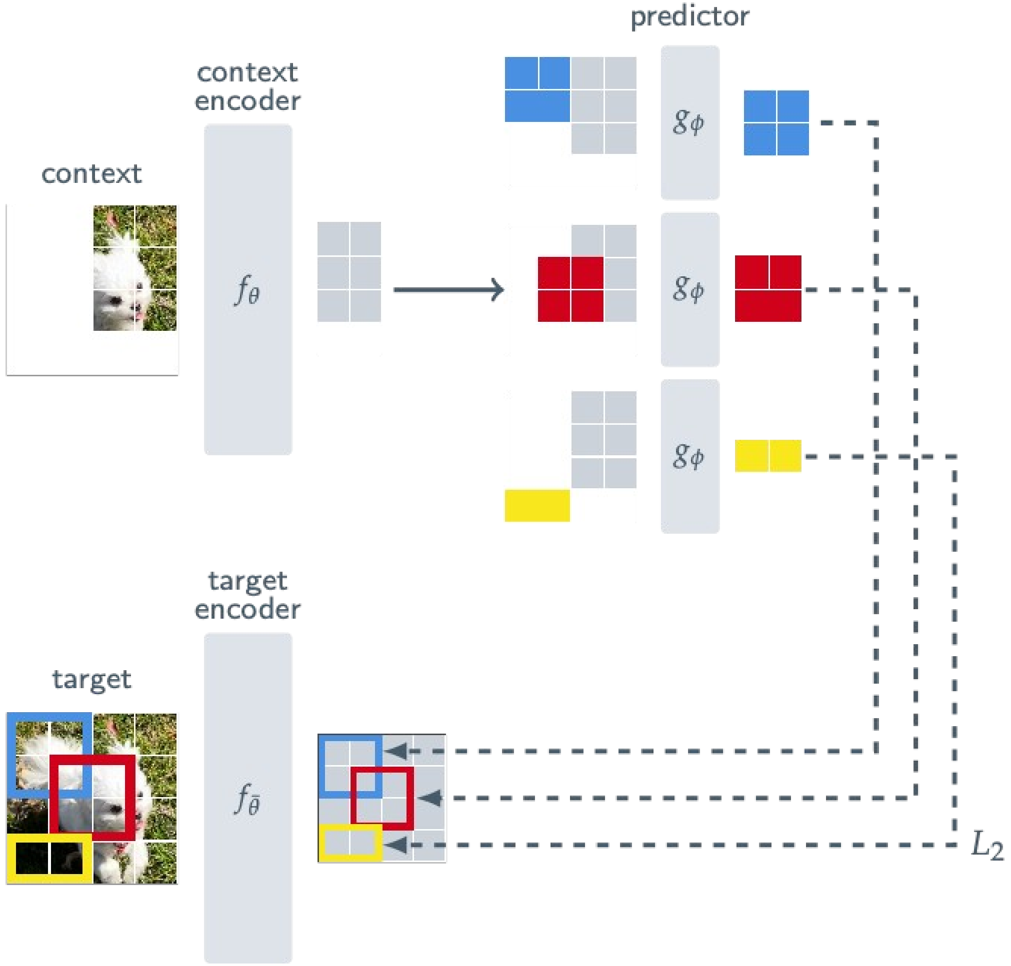
I-JEPA: arXiv:2301.08243 CVPR'23 https://github.com/facebookresearch/ijepa
Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture
M Assran, Q Duval, I Misra, P Bojanowski, P Vincent, M Rabbat, Y LeCun, N Ballas

V-JEPA: arXiv:2404.08471 TMLR'24 https://github.com/facebookresearch/jepa
"Revisiting Feature Prediction for Learning Visual Representations from Video"
A Bardes, Q Garrido, J Ponce, X Chen, M Rabbat, Y LeCun, M Assran, N Ballas

Image-JEPA: uses masking & transformer architectures

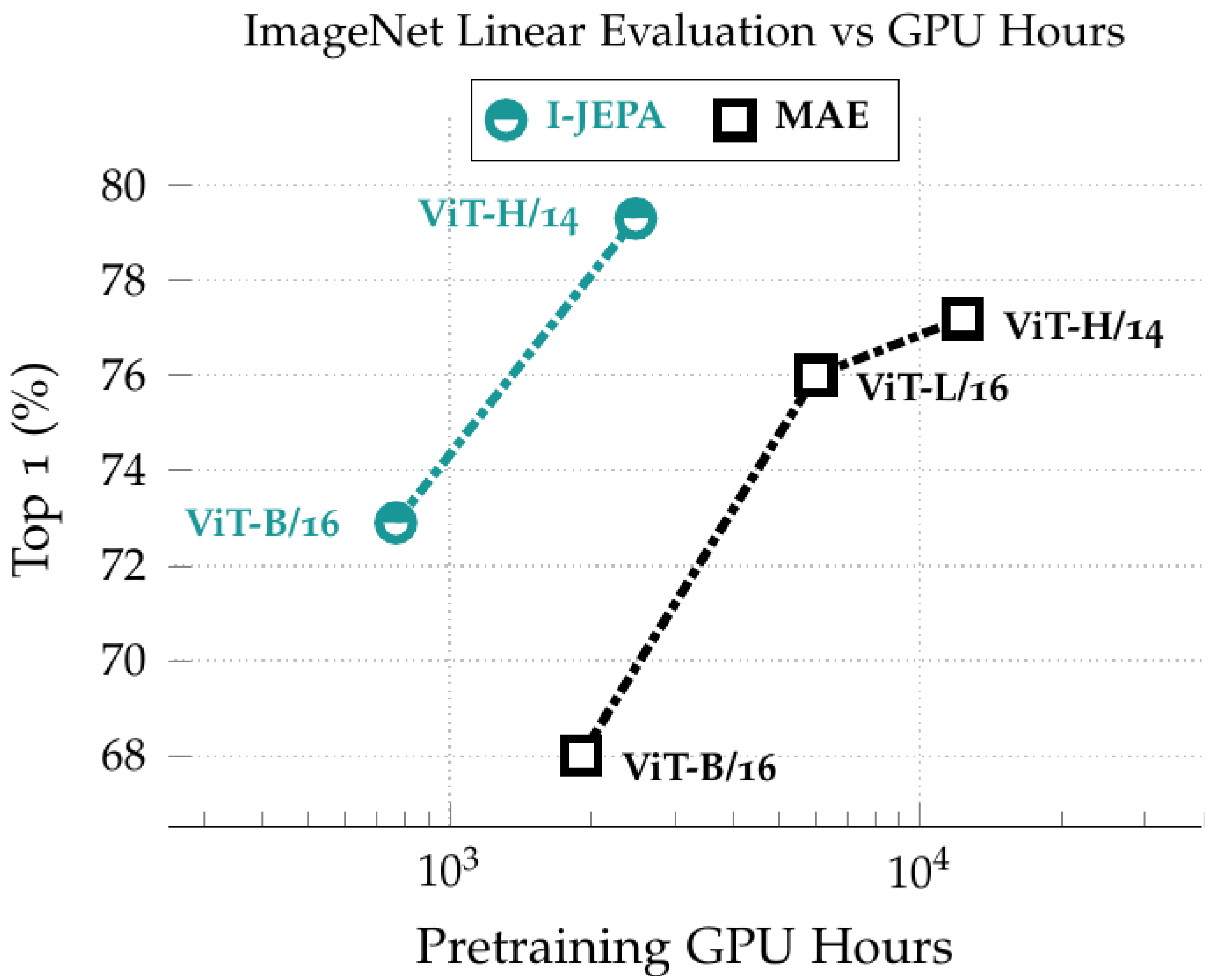
- **“SSL from images with a JEPA”**
  - [M. Assran et al arxiv:2301.08243]
- **Jointly embeds a context and a number of neighboring patches.**
  - Uses predictors
  - Uses only masking





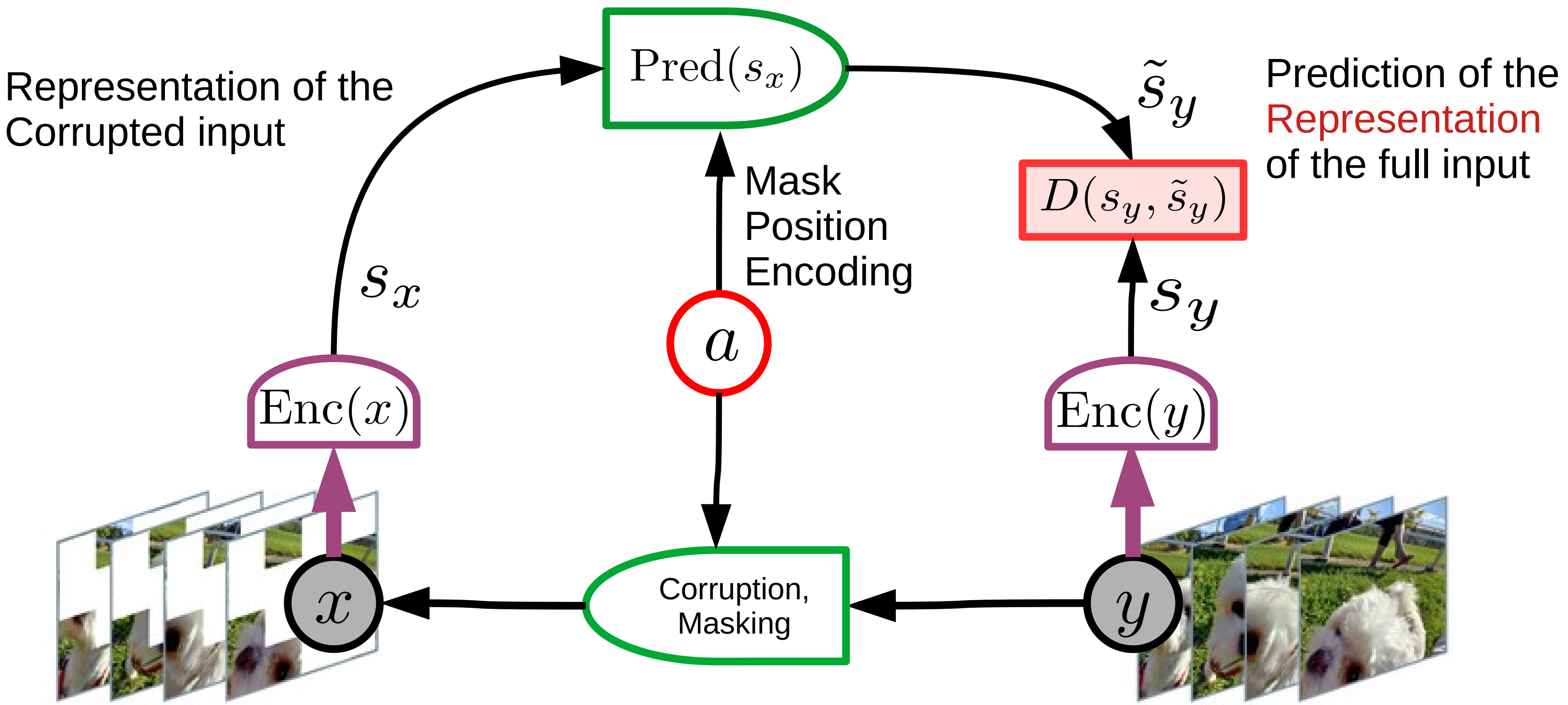
I-JEPA Results

- Training is fast
- Non-generative method beat reconstruction-based generative methods such as Masked Auto-Encoder
  - (with a frozen trunk).



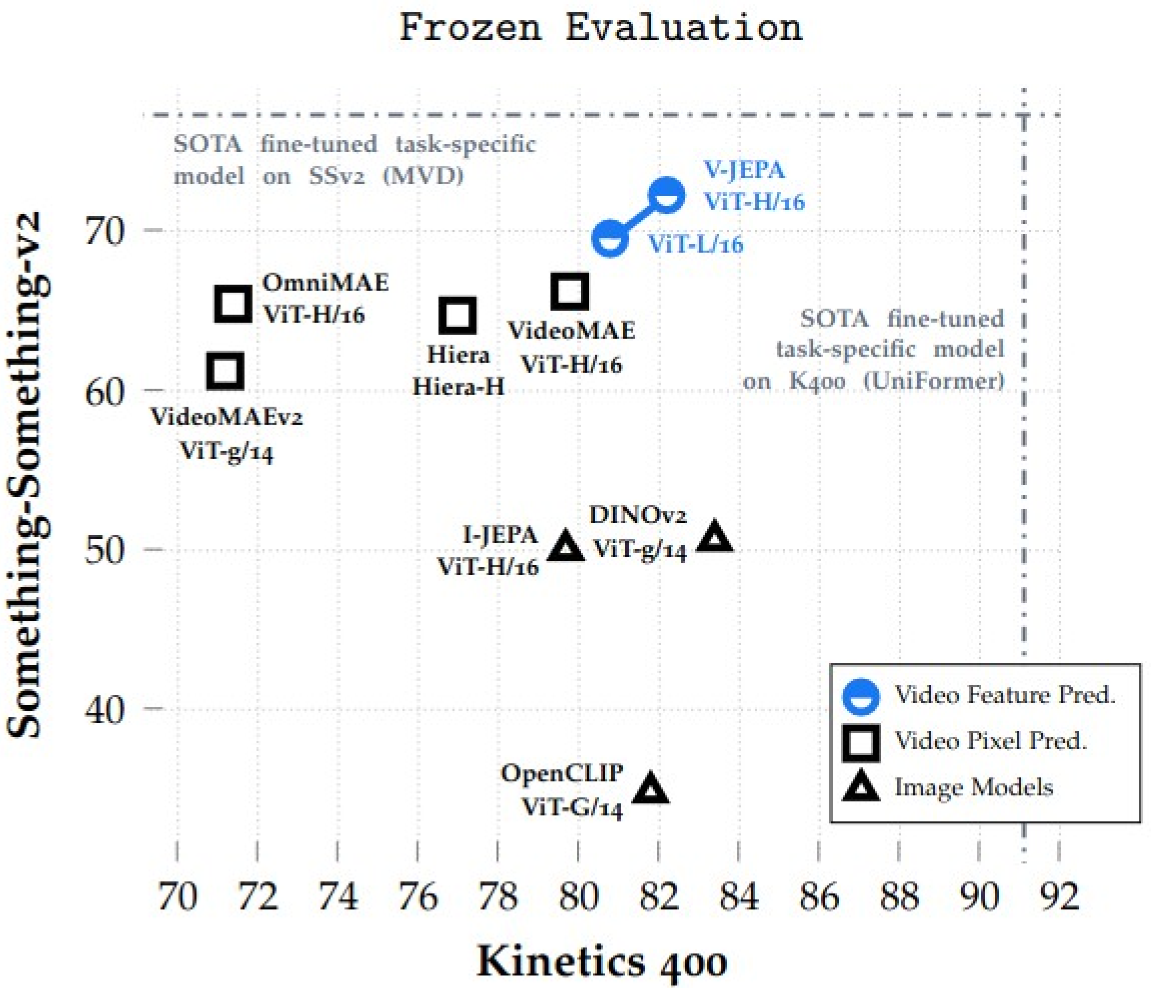
Video-JEPA

- [Bardes et al. 2024]



V-JEPA: results on action recognition

- Supervised head on frozen backbone.
- Comparison with generative models: OmniMAE, VideoMAE, Hiera
- Comparison with image models: I-JEPA, DINOv2, OpenCLIP

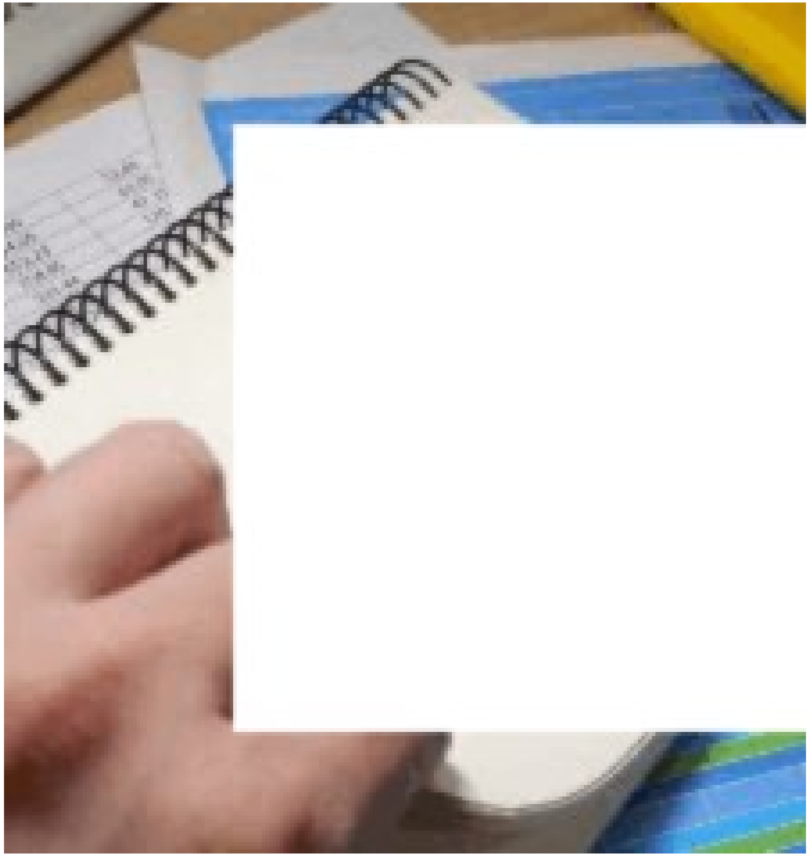


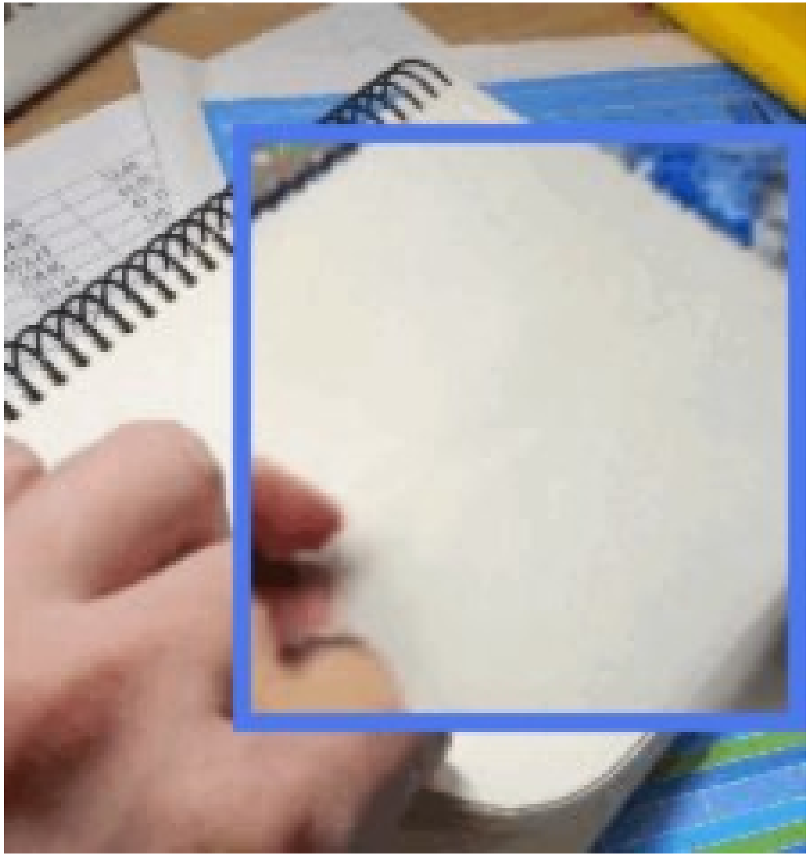
V-JEPA: results for low-shot action recognition

- Rows 1-3: generative architectures with reconstruction
- Row 4: V-JEPA
- Supervised head on frozen backbone.

		Frozen Evaluation					
		K400 (16×8×3)			SSv2 (16×2×3)		
Method	Arch.	5%	10%	50%	5%	10%	50%
MVD	ViT-L/16	62.6 ± 0.2	68.3 ± 0.2	77.2 ± 0.3	42.9 ± 0.8	49.5 ± 0.6	61.0 ± 0.2
VideoMAE	ViT-H/16	62.3 ± 0.3	68.5 ± 0.2	78.2 ± 0.1	41.4 ± 0.8	48.1 ± 0.2	60.5 ± 0.4
VideoMAEv2	ViT-g/14	37.0 ± 0.3	48.8 ± 0.4	67.8 ± 0.1	28.0 ± 1.0	37.3 ± 0.3	54.0 ± 0.3
V-JEPA	ViT-H/16_{384}	68.2 ± 0.2	72.8 ± 0.2	80.6 ± 0.2	54.0 ± 0.2	59.3 ± 0.5	67.9 ± 0.2

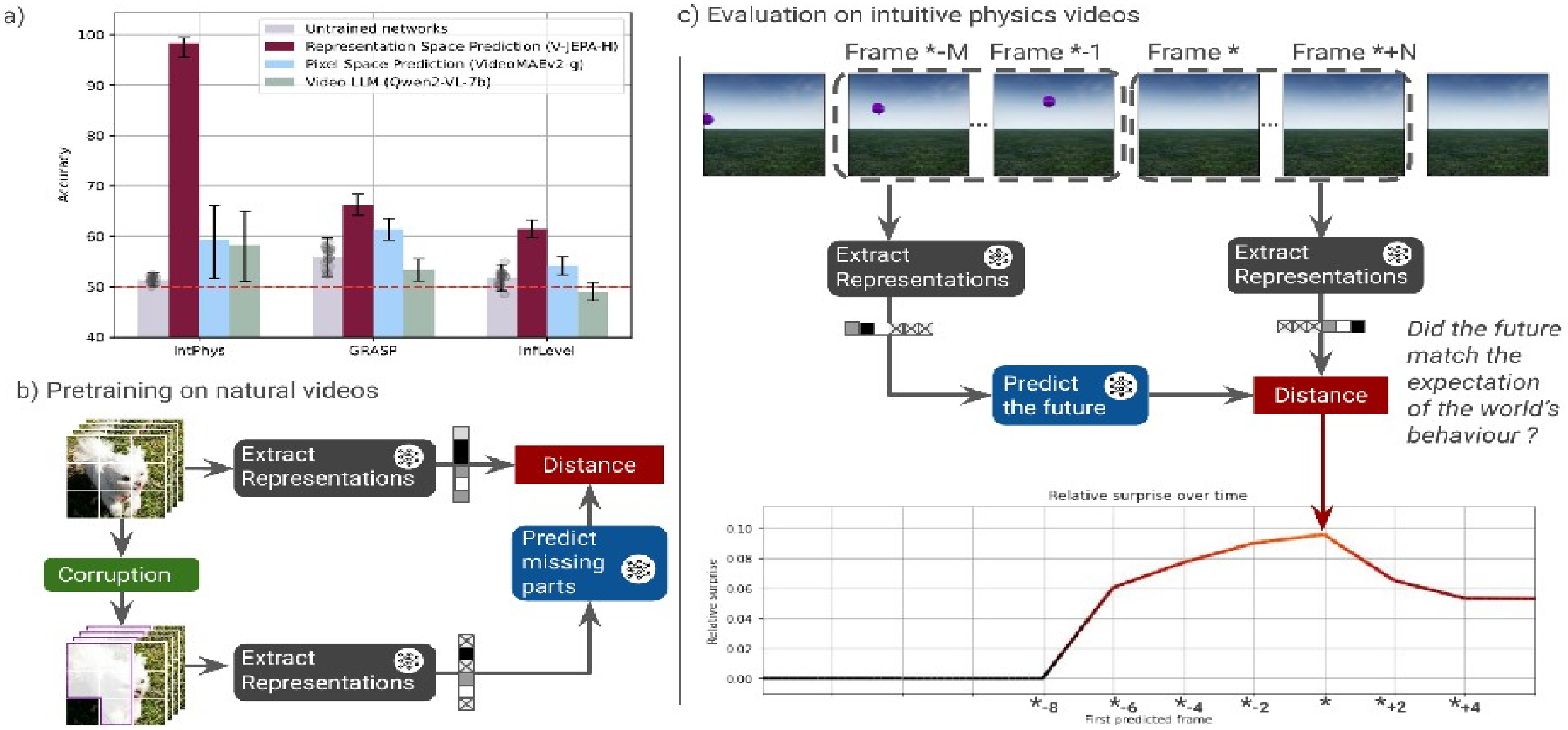
V-JEPA: Decoded Predictions



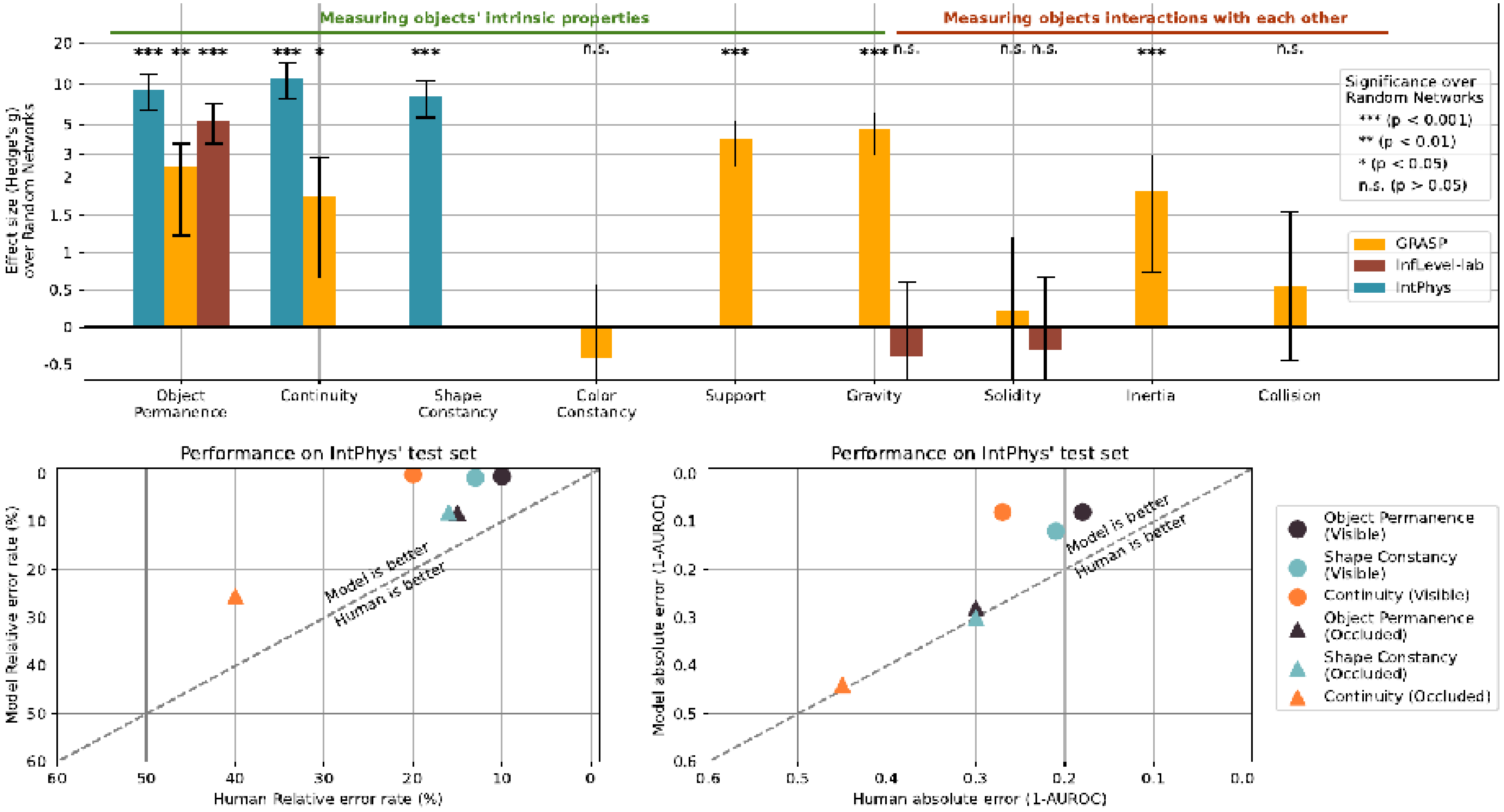


V-JEPA and "visual common sense" / intuitive physics

- [Garrido et al. ArXiv:2502.11831]

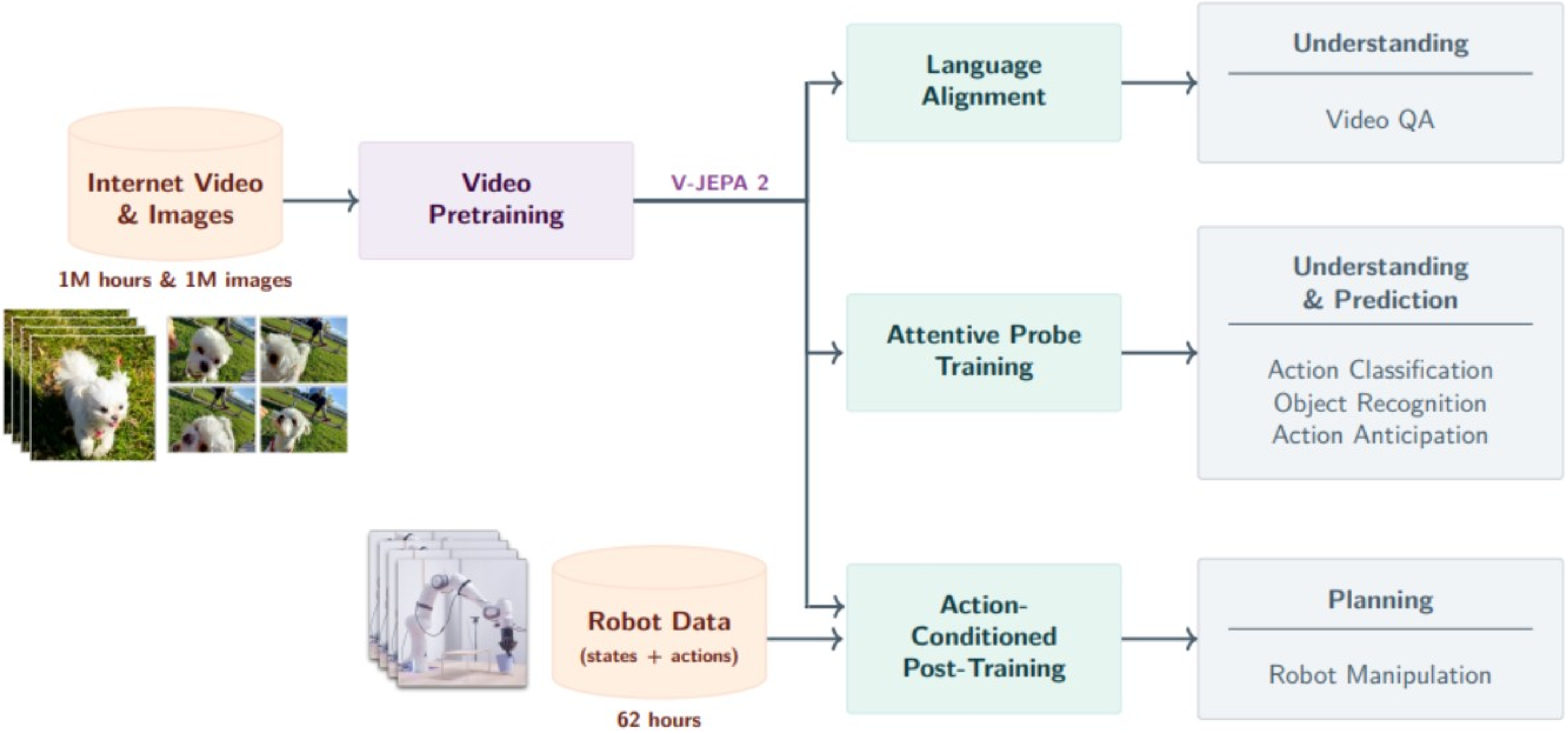


V-JEPA and "visual common sense" and intuitive physics



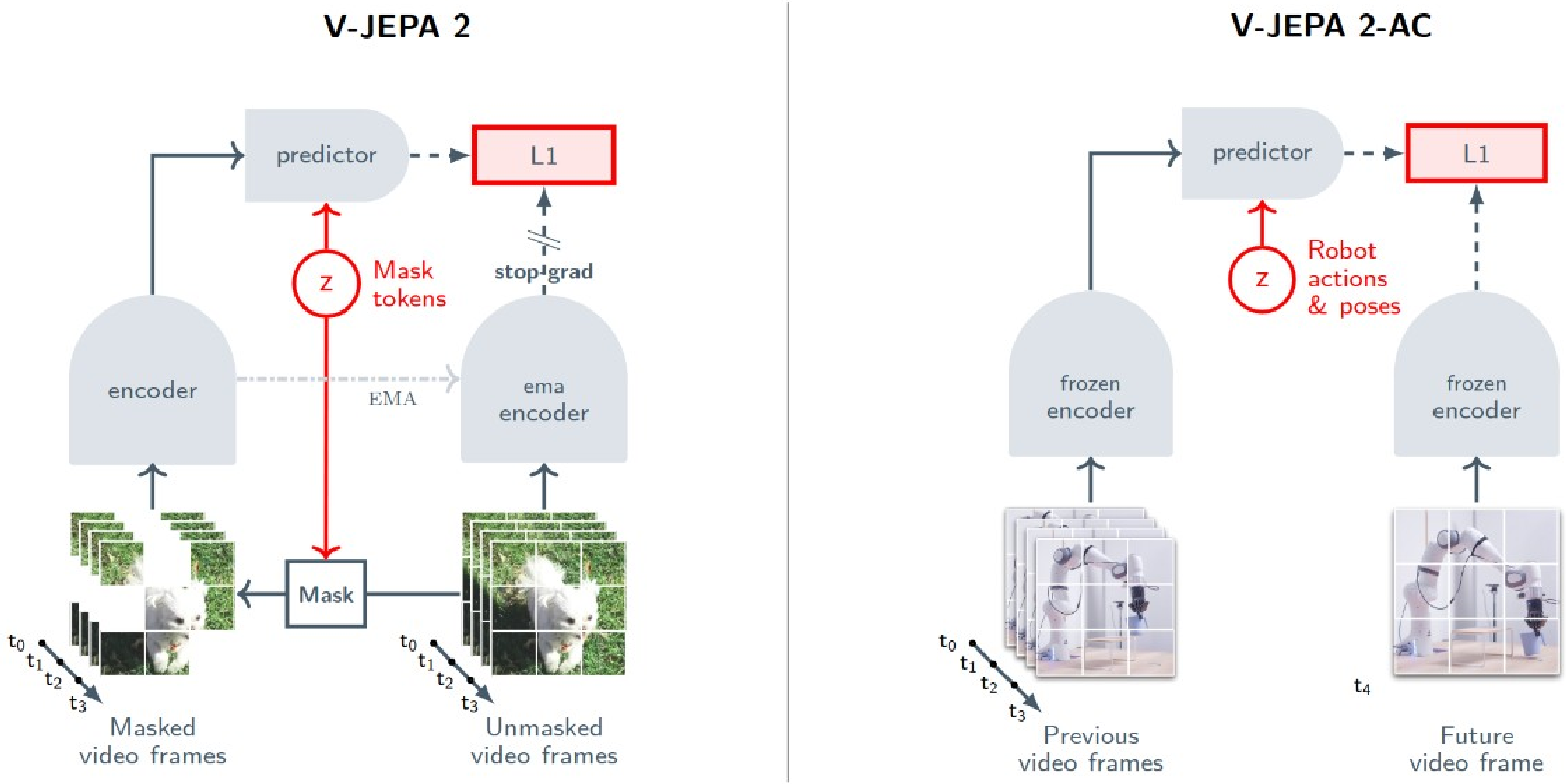
V-JEPA 2: large-scale SSL from video

- [Assran et al. ArXiv:2506.09985] https://ai.meta.com/vjepa/



V-JEPA 2: large-scale SSL from video

- [Assran et al. ArXiv:2506.09985] https://ai.meta.com/vjepa/
- Two-phase training: (1) masked videos, (2) action-conditioning



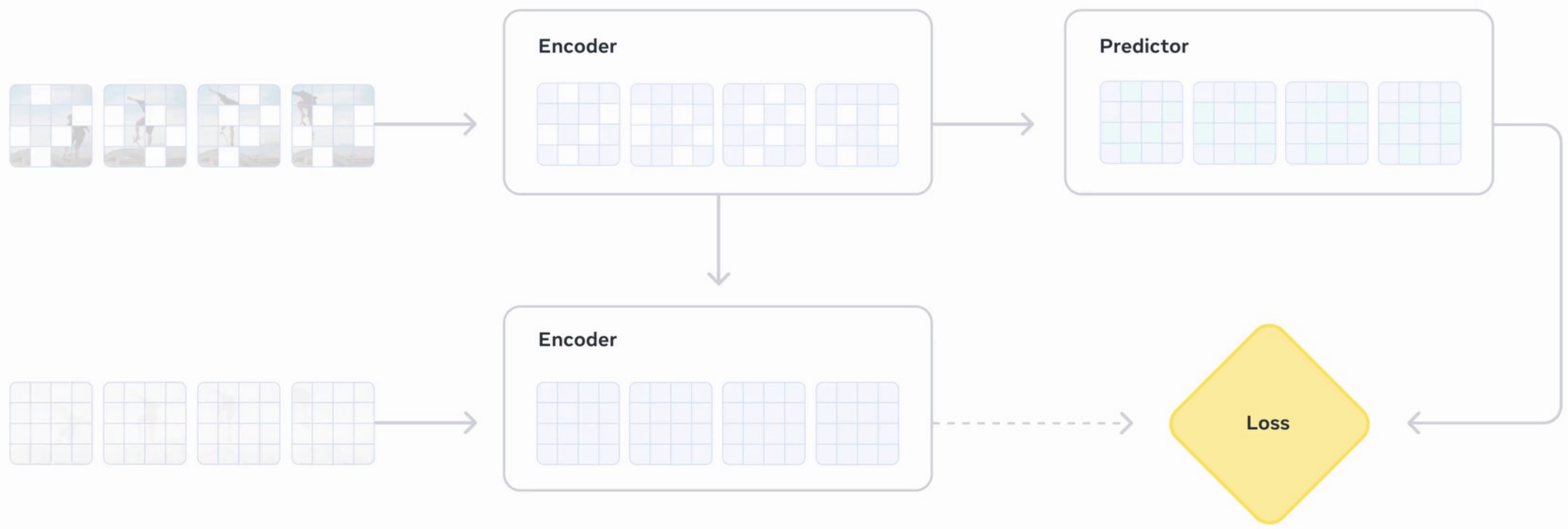
V-JEPA 2: Pre-training datasets

- [Assran et al. ArXiv:2506.09985] https://ai.meta.com/vjepa/

Table 1 VideoMix22M (VM22M) Pretraining Dataset. To build our observation pretraining dataset, we combined four different video sources and one image dataset. We use a source-specific sampling probability during training and apply retrieval-based curation on YT1B to reduce noisy content (e.g., cartoon- or clipart-style).

Source	Samples	Type	Total Hours	Apply Curation	Weight
SSv2 (Goyal et al., 2017)	168K	EgoVideo	168	No	0.056
Kinetics (Carreira et al., 2019)	733K	ExoVideo	614	No	0.188
Howto100M (Miech et al., 2019)	1.1M	ExoVideo	134K	No	0.318
YT-Temporal-1B (Zellers et al., 2022)	19M	ExoVideo	1.6M	Yes	0.188
ImageNet (Deng et al., 2009)	1M	Images	n/a	No	0.250

V-JEPA 2 training



V-JEPA-2 planning

Training the Action-Conditioned Predictor

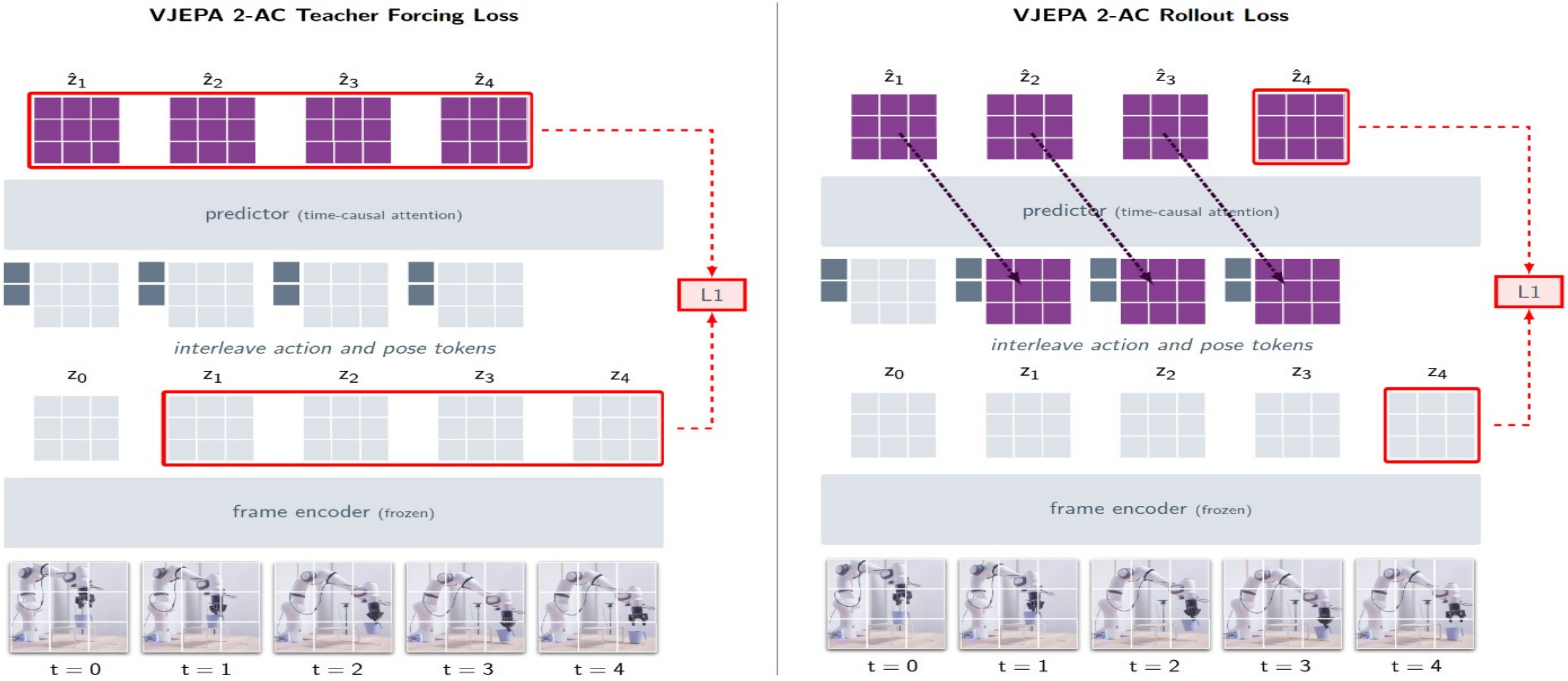


Figure 6 **V-JEPA 2-AC training.** V-JEPA 2-AC is trained in an autoregressive fashion, utilizing a teacher forcing loss and a rollout loss. (**Left**) In the teacher forcing loss, the predictor takes the encoding of the current frame representation as input and learns to predict the representation of the next timestep. (**Right**) The rollout loss involves feeding the predictor's output back as input, allowing the model to be trained to predict several timesteps ahead. By optimizing the sum of these two losses, V-JEPA 2-AC enhances its ability to accurately forecast the future by reducing error accumulation during rollouts.

V-JEPA 2 Results

TASK TYPE	BENCHMARK	V-JEPA 2	PREVIOUS BEST
Planning and Robot Control from Image Goals	**Reach**	100%	100% (Octo)
	Grasp	45%	8% (Octo)
	Pick-and-place	73%	13% (Octo)
Prediction	**Epic-Kitchens-100 action anticipation**	39.7%	27.6% (PlausiVL)
Understanding	**Something-Somethingv2 action recognition** Attentive probe	77.3%	69.7% (InternVideo2-1B Attentive probe)
	Diving48 Attentive probe	90.2%	86.4% (InternVideo2-1B Attentive probe)
	Perception Test	84.0%	82.7% (PerceptionLM)
	MVPBench	44.5%	39.9% (InternVL-2.5)

V-JEPA 2 Results

Method	Param.	Avg.	*Motion Understanding* SSv2	Diving-48	Jester	*Appearance Understanding* K400	COIN	IN1K
Results Reported in the Literature								
VideoMAEv2 (Wang et al., 2023)	1B	–	56.1	–	–	82.8	–	71.4
InternVideo2-1B (Wang et al., 2024b)	1B	–	67.3	–	–	87.9	–	–
InternVideo2-6B (Wang et al., 2024b)	6B	–	67.7	–	–	88.8	–	–
VideoPrism (Zhao et al., 2024)	1B	–	68.5	71.3	–	87.6	–	–
Image Encoders Evaluated Using the Same Protocol								
DINOv2 (Darcet et al., 2024)	1.1B	81.1	50.7	82.5	93.4	83.6	90.7	86.1
$\mathrm{PE_{core}G}$ (Bolya et al., 2025)	1.9B	82.3	55.4	76.9	90.0	88.5	**95.3**	87.6*
SigLIP2 (Tschannen et al., 2025)	1.2B	81.1	49.9	75.3	91.0	87.3	95.1	**88.0**
Video Encoders Evaluated Using the Same Protocol								
V-JEPA ViT-H (Bardes et al., 2024)	600M	85.2	74.3	87.9	97.7	84.5	87.1	80.0
$\mathrm{InternVideo2_{s2}}$-1B (Wang et al., 2024b)	1B	87.0	69.7	86.4	97.0	**89.4**	93.8	85.8
V-JEPA 2 ViT-L	300M	86.0	73.7	89.0	97.6	85.1	86.8	83.5
V-JEPA 2 ViT-H	600M	86.4	74.0	89.8	97.7	85.3	87.9	83.8
V-JEPA 2 ViT-g	1B	87.5	75.3	90.1	97.7	86.6	90.7	84.6
V-JEPA 2 $\mathrm{ViT\text{-}g_{384}}$	1B	**88.2**	**77.3**	**90.2**	**97.8**	87.3	91.1	85.1

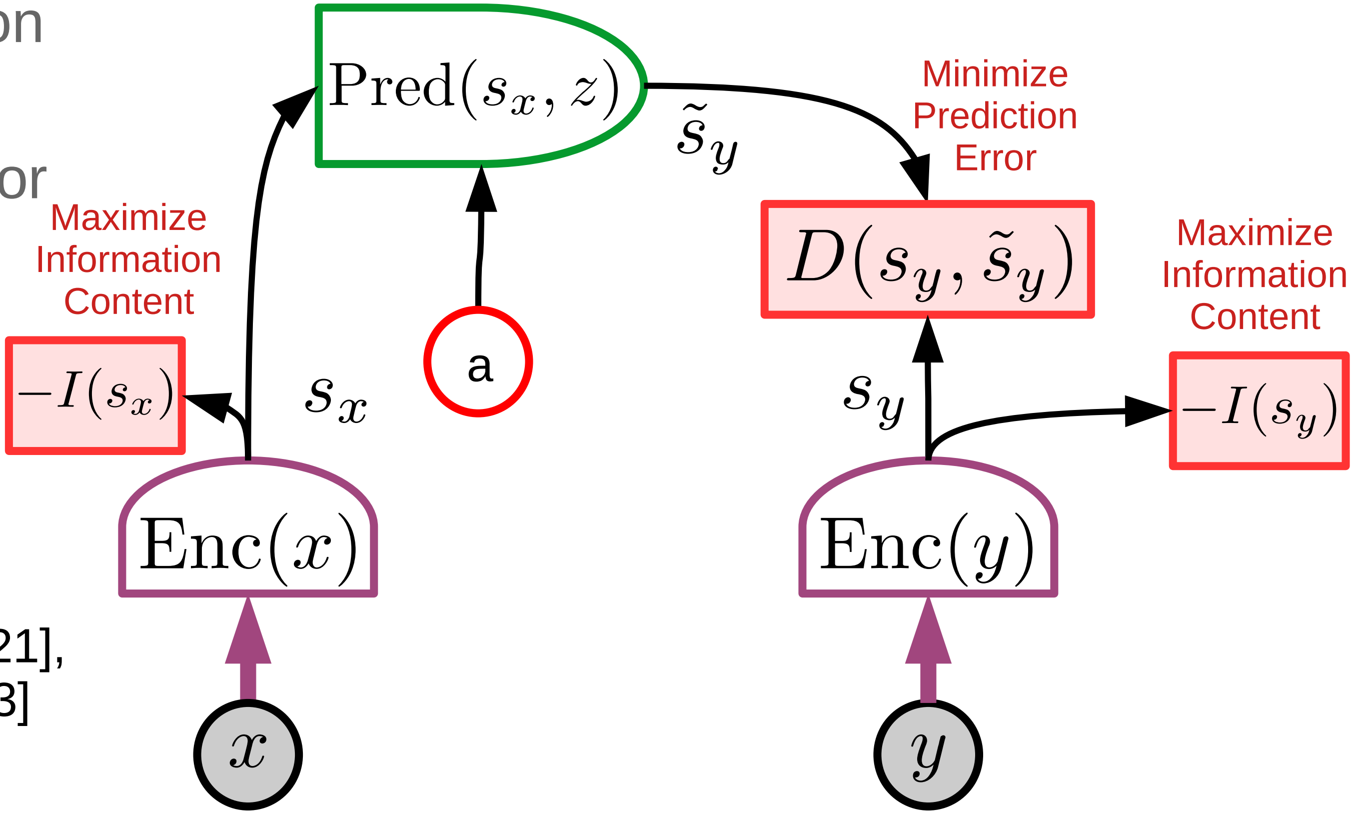
Training JEPA with Regularized Methods: Information Maximization

MCR2 [Yu et al. NeurIPS 2020],
Barlow Twins [Zbontar, Li, Misra, L, Deny, ArXiv:2103.03230, ICML'21],
W-MSE [Ermolov et al. ICML 2021],
VICReg [Bardes, Ponce, LeCun arXiv:2105.04906, ICLR 2022],
VICRegL [Bardes, Ponce, LeCun arXiv:2210.01571, NeurIPS 2022]
MMCR [Yerxa et al. NeurIPS 2023]

Training a JEPA with Information Maximization

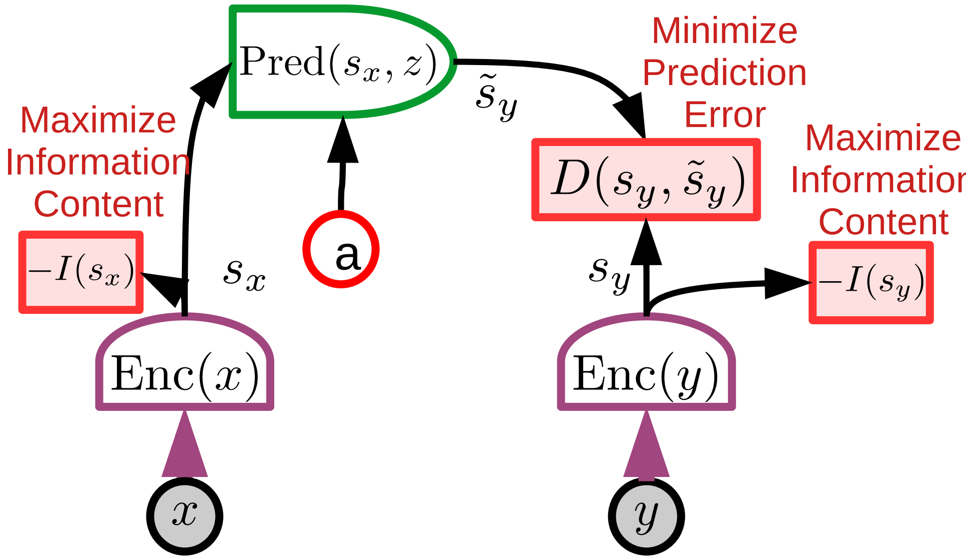
- **Three terms in the cost**
  - Maximize information content in representation of x and y
  - Minimize Prediction error
- **Whitening Sx and Sy**

MCR2 [Yu et al. NeurIPS 2020], Barlow Twins [Zbontar et al. ArXiv:2103.03230], VICReg [Bardes, Ponce, LeCun arXiv:2105.04906, ICLR 2022], W-MSE [Ermolov et al. PMLR 2021], MMCR [Yerxa et al. NeurIPS 2023]



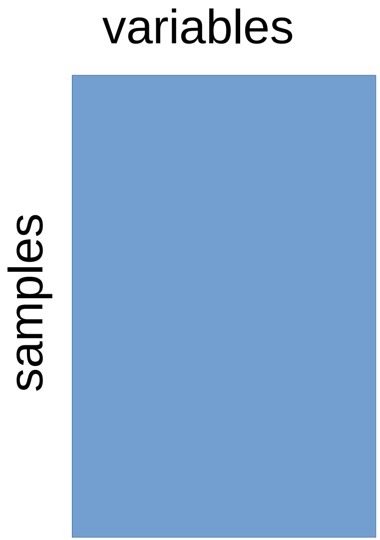
Training a JEPA with Information Maximization

- **Main Challenge:**
  - How can we maximize information content in representation of x and y?
  - We do not have lower bounds on information content !!!
  - We only have upper bounds
  - Because we must make assumptions about the type of dependencies that exist between the variables
  - There may be complicated but unknown dependencies that lower the information content.

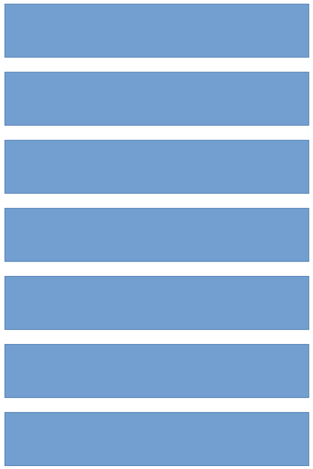


- **Basic idea: make the representations fill the space**
  - Sample Contrastive: push vectors away from each other
  - Dim Contrastive: push variables away from each other

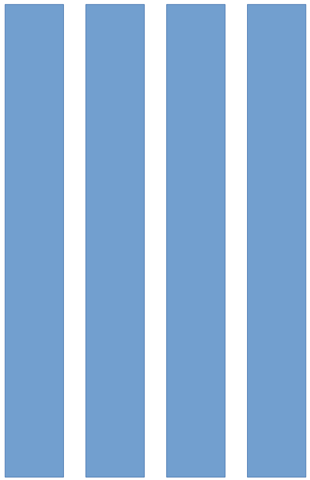
Matrix of representations for a Batch of Samples



- **Sample Contrastive Methods:**
  - Make the row of the matrix as different from each other as possible
  - Requires a large number of rows
  - Don’t work in high dimension



- **Dimension Contrastive Methods**
  - Make the column as different from each other as possible
  - Requires a small number of rows
  - Don’t work for large batches



- **Equivalence**

  [Garrido ICLR 2023, ArXiv:2206.02574]

  On the duality between contrastive and non-contrastive self-supervised learning

Sample contrastive vs Dimension contrastive?

- **[Garrido et al. Arxiv:2206.02574]**
  - "ON THE DUALITY BETWEEN CONTRASTIVE AND NON CONTRASTIVE SELF-SUPERVISED LEARNING"

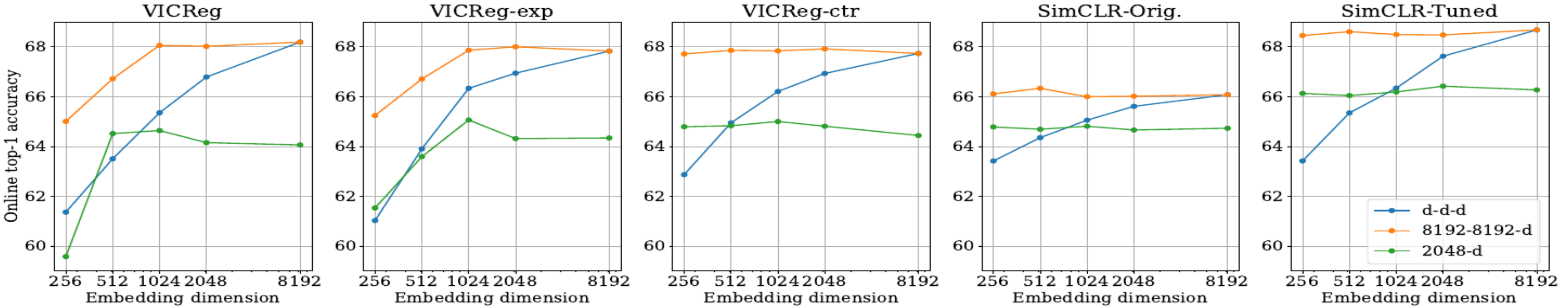
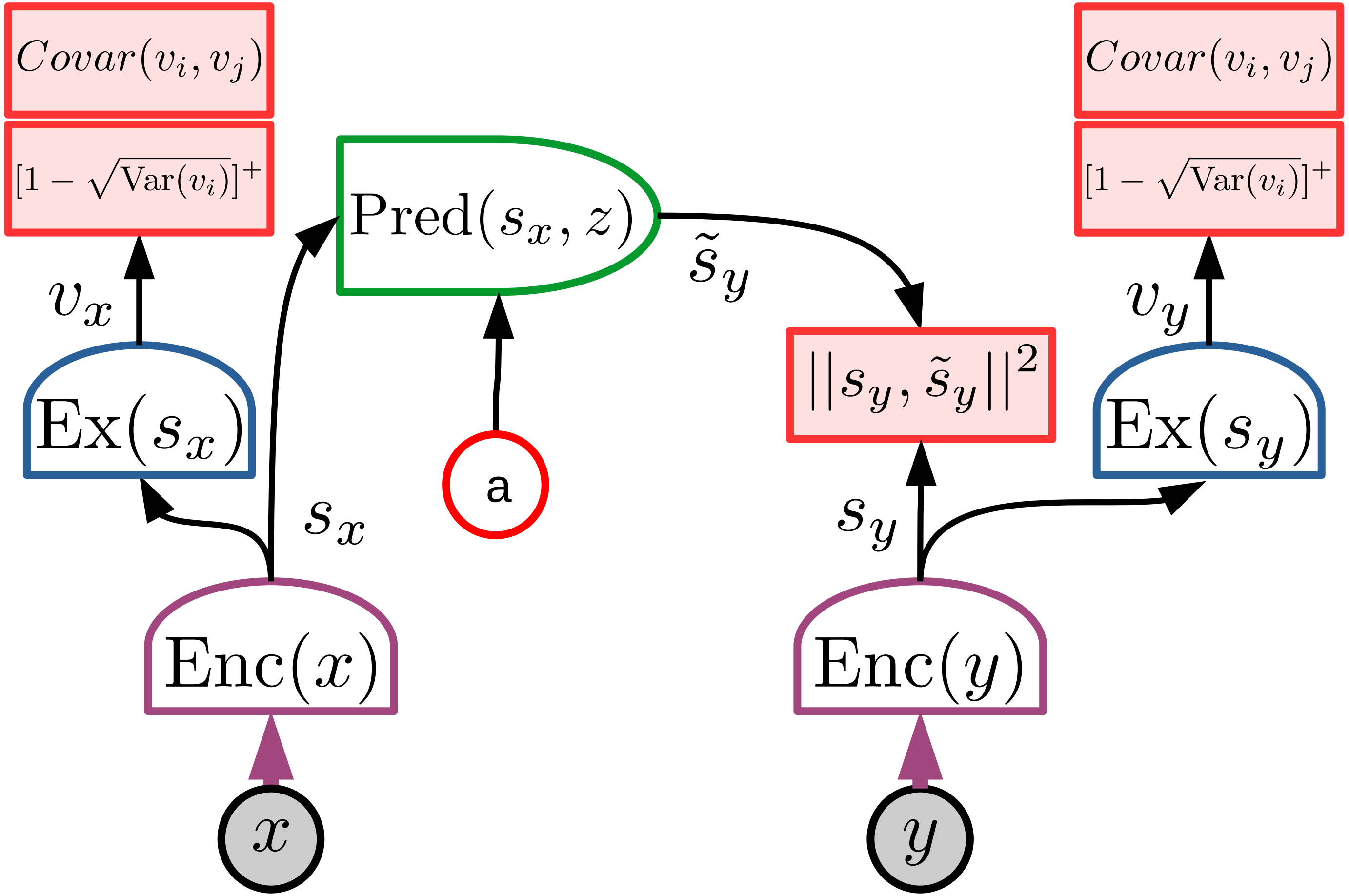


Figure 1: VICReg, VICReg-exp and VICReg-ctr perform similarly in 100 epochs training, validating empirically our theoretical result. While the original implementation of SimCLR performs significantly worse – which is unexpected per our theory – we are able to improve its performance to VICReg's level. This further validates our findings. While different projector architectures impact performance, behaviours are similar across methods. Confer supplementary section H for numerical values and hyperparameters.

VICReg: Variance, Invariance, Covariance Regularization

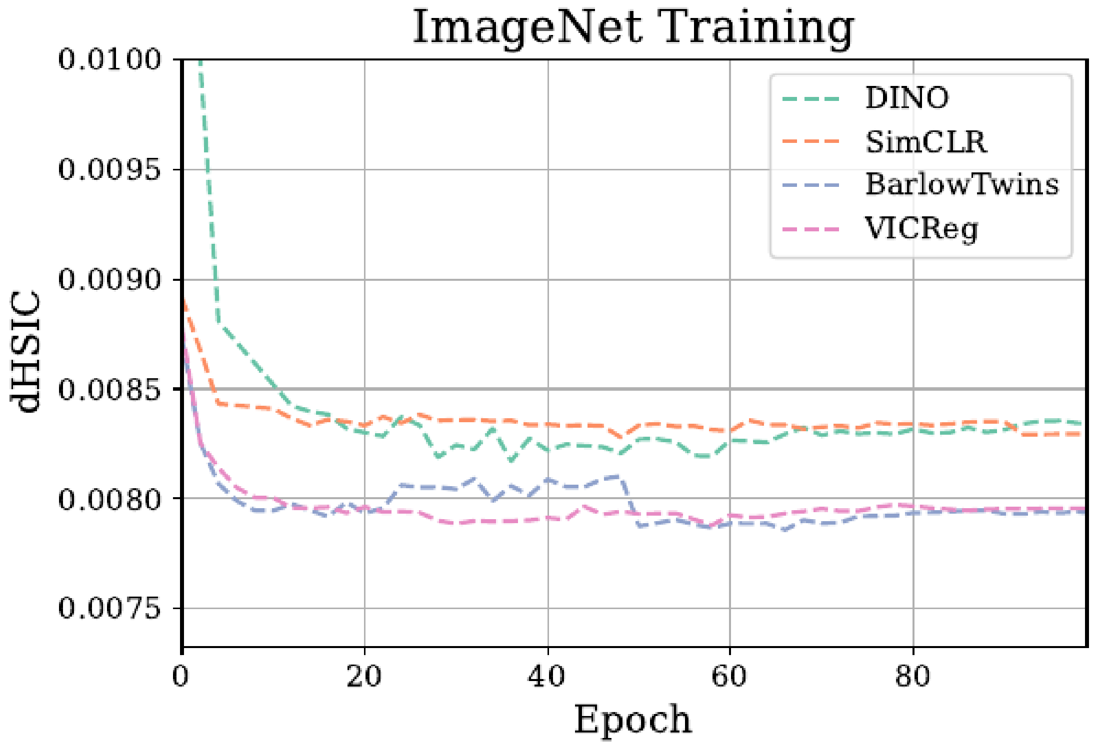
- **Variance:**
  - Maintains variance of components of representations
- **Covariance:**
  - Decorrelates components of covariance matrix of representations
- **Invariance:**
  - Minimizes prediction error.

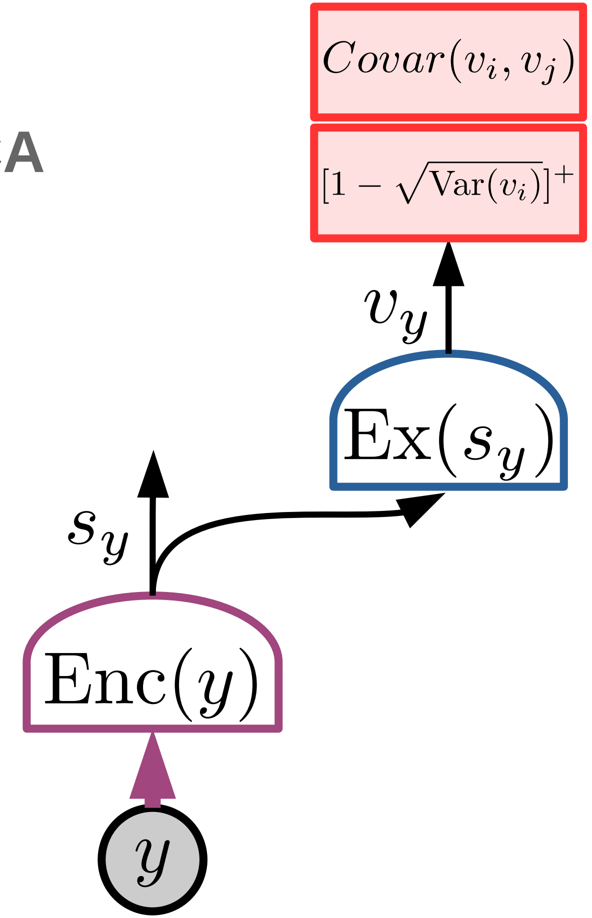


Barlow Twins [Zbontar et al. ArXiv:2103.03230], VICReg [Bardes, Ponce, LeCun arXiv:2105.04906, ICLR 2022],

VICReg: expander makes variables pairwise independent

- **[Mialon, Balestriero, LeCun arxiv:2209.14905]**
- **VC criterion can be used for source separation / ICA**





VICReg: Results with linear head and semi-supervised.

Method	Linear		Semi-supervised			
	Top-1	Top-5	Top-1		Top-5	
			1%	10%	1%	10%
Supervised	76.5	-	25.4	56.4	48.4	80.4
MoCo He et al. (2020)	60.6	-	-	-	-	-
PIRL Misra & Maaten (2020)	63.6	-	-	-	57.2	83.8
CPC v2 Hénaff et al. (2019)	63.8	-	-	-	-	-
CMC Tian et al. (2019)	66.2	-	-	-	-	-
SimCLR Chen et al. (2020a)	69.3	89.0	48.3	65.6	75.5	87.8
MoCo v2 Chen et al. (2020c)	71.1	-	-	-	-	-
SimSiam Chen & He (2020)	71.3	-	-	-	-	-
SwAV Caron et al. (2020)	71.8	-	-	-	-	-
InfoMin Aug Tian et al. (2020)	73.0	91.1	-	-	-	-
OBoW Gidaris et al. (2021)	73.8	-	-	-	82.9	90.7
BYOL Grill et al. (2020)	74.3	91.6	53.2	68.8	78.4	89.0
SwAV (w/ multi-crop) Caron et al. (2020)	75.3	-	53.9	70.2	78.5	89.9
Barlow Twins Zbontar et al. (2021)	73.2	91.0	55.0	69.7	79.2	89.3
VICReg (ours)	73.2	91.1	54.8	69.5	79.4	89.5

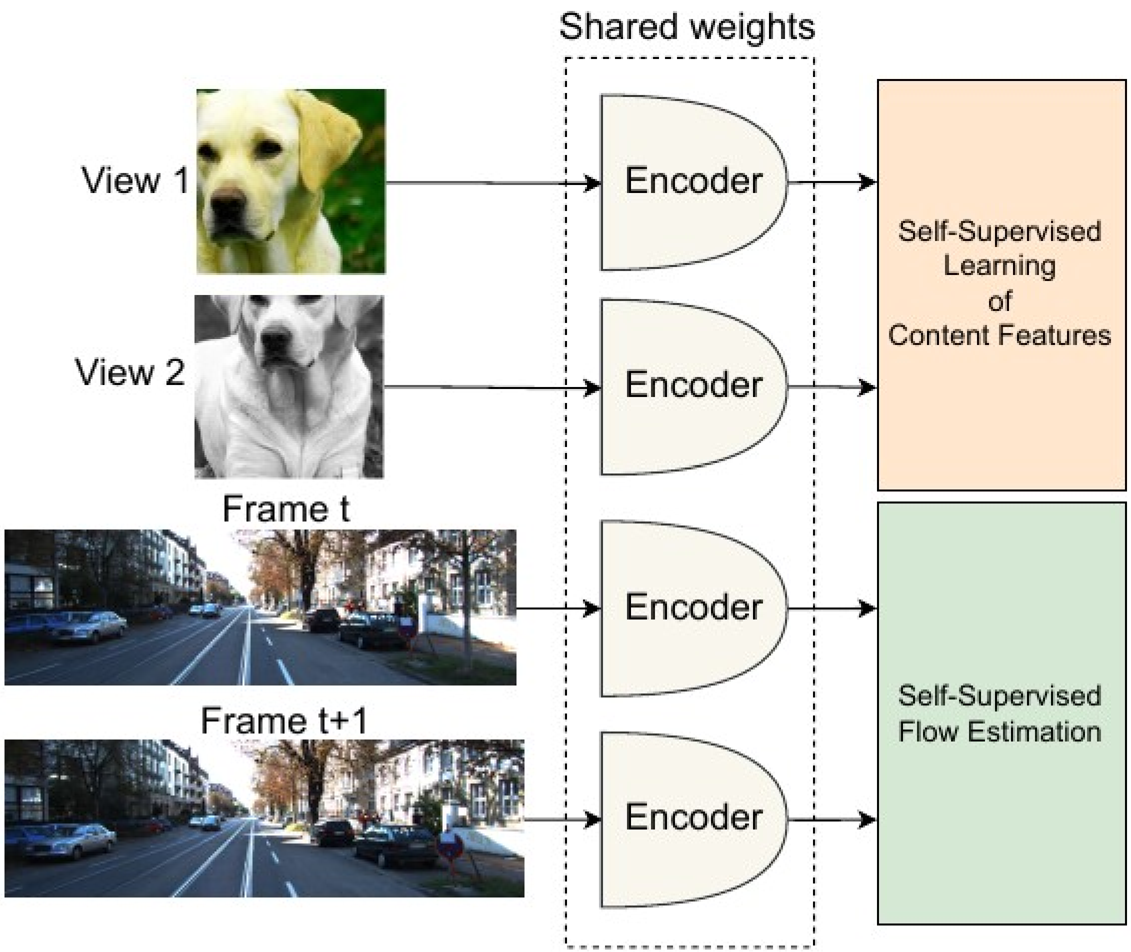
VICReg: Results with transfer tasks.

Method	Linear Classification			Object Detection		
	Places205	VOC07	iNat18	VOC07+12	COCO det	COCO seg
Supervised	53.2	87.5	46.7	81.3	39.0	35.4
MoCo He et al. (2020)	46.9	79.8	31.5	-	-	-
PIRL Misra & Maaten (2020)	49.8	81.1	34.1	-	-	-
SimCLR Chen et al. (2020a)	52.5	85.5	37.2	-	-	-
MoCo v2 Chen et al. (2020c)	51.8	86.4	38.6	82.5	39.8	36.1
SimSiam Chen & He (2020)	-	-	-	82.4	-	-
BYOL Grill et al. (2020)	54.0	<u>86.6</u>	<u>47.6</u>	-	<u>40.4</u>†	<u>37.0</u>†
SwAV (m-c) Caron et al. (2020)	<u>56.7</u>	<u>88.9</u>	<u>48.6</u>	<u>82.6</u>	<u>41.6</u>	<u>37.8</u>
OBoW Gidaris et al. (2021)	<u>56.8</u>	<u>89.3</u>	-	<u>82.9</u>	-	-
Barlow Twins Grill et al. (2020)	54.1	86.2	46.5	<u>82.6</u>	<u>40.0</u>†	<u>36.7</u>†
VICReg (ours)	<u>54.3</u>	<u>86.6</u>	<u>47.0</u>	82.4	39.4	36.4

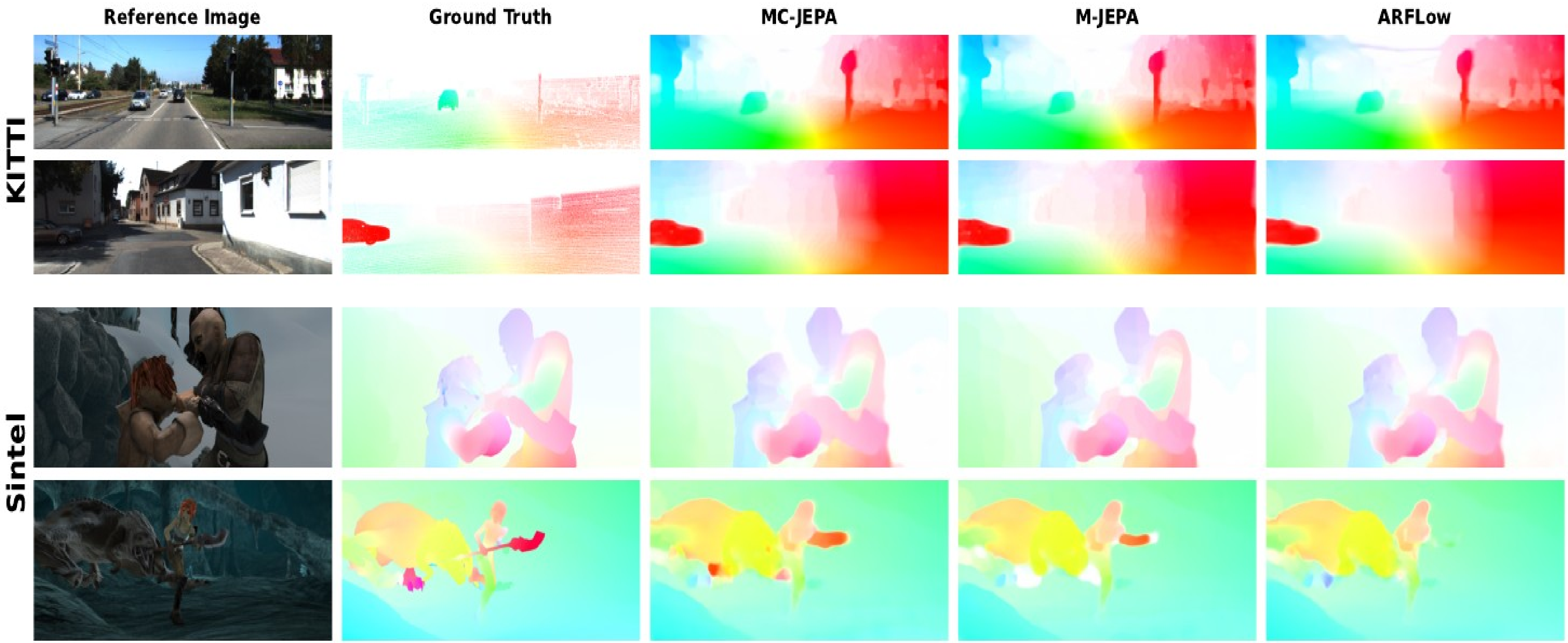
MC-JEPA: Motion & Content JEPA

[Bardes, Ponce, LeCun 23]

- **Simultaneous SSL for**
  - Image recognition
  - Motion estimation
- **Trained on**
  - ImageNet 1k
  - Various video datasets
- **Uses VCReg to prevent collapse**
  - ConvNext-T backbone

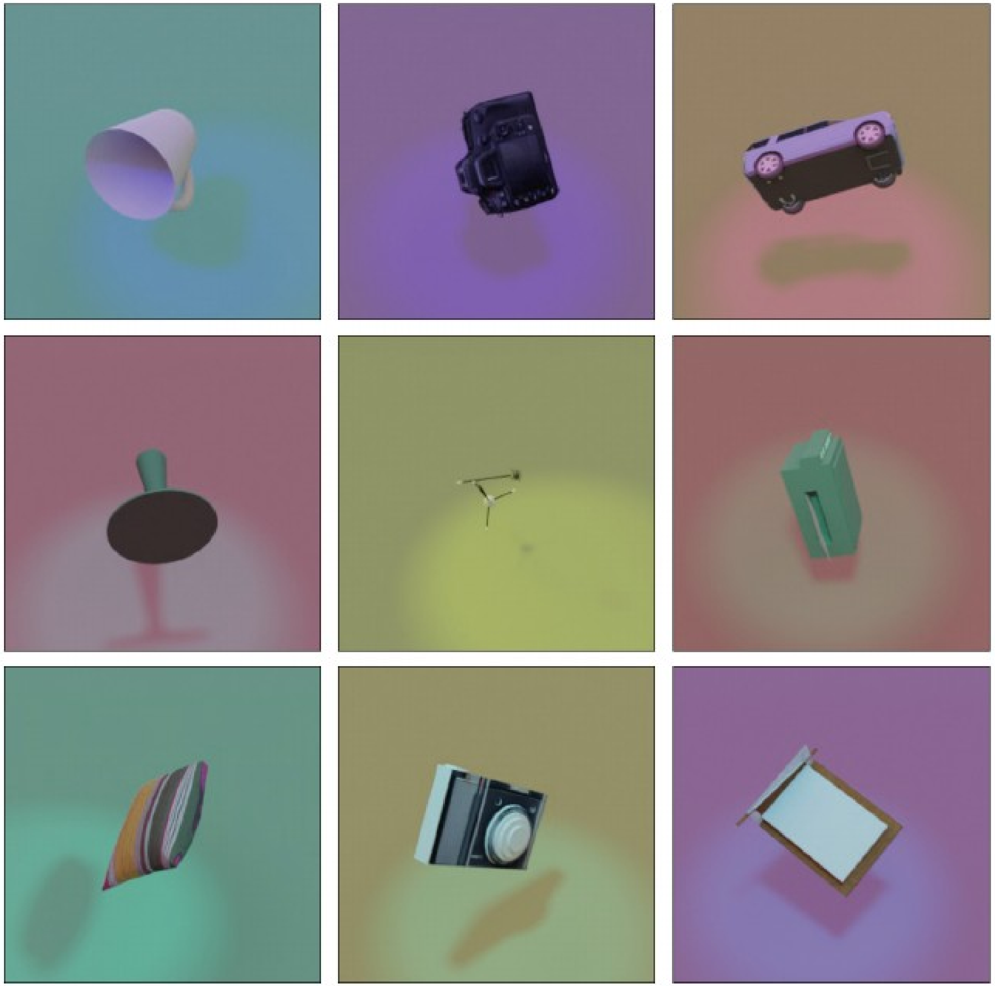


MC-JEPA: Optical Flow Estimation Results



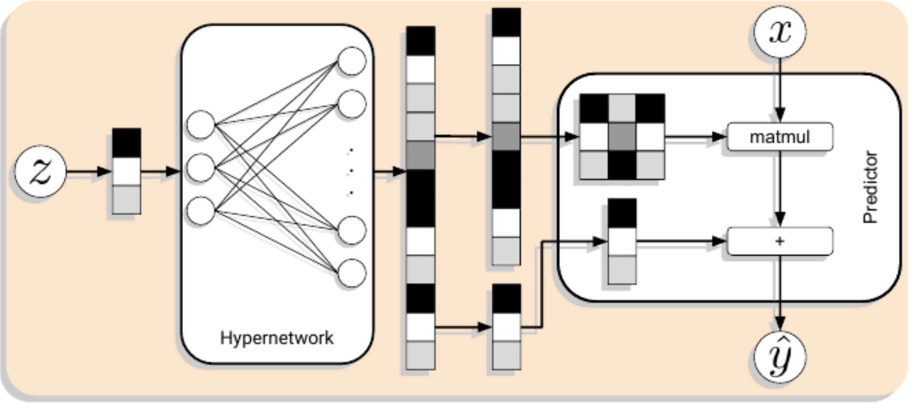
Split Invariant-Equivariant Representation Learning

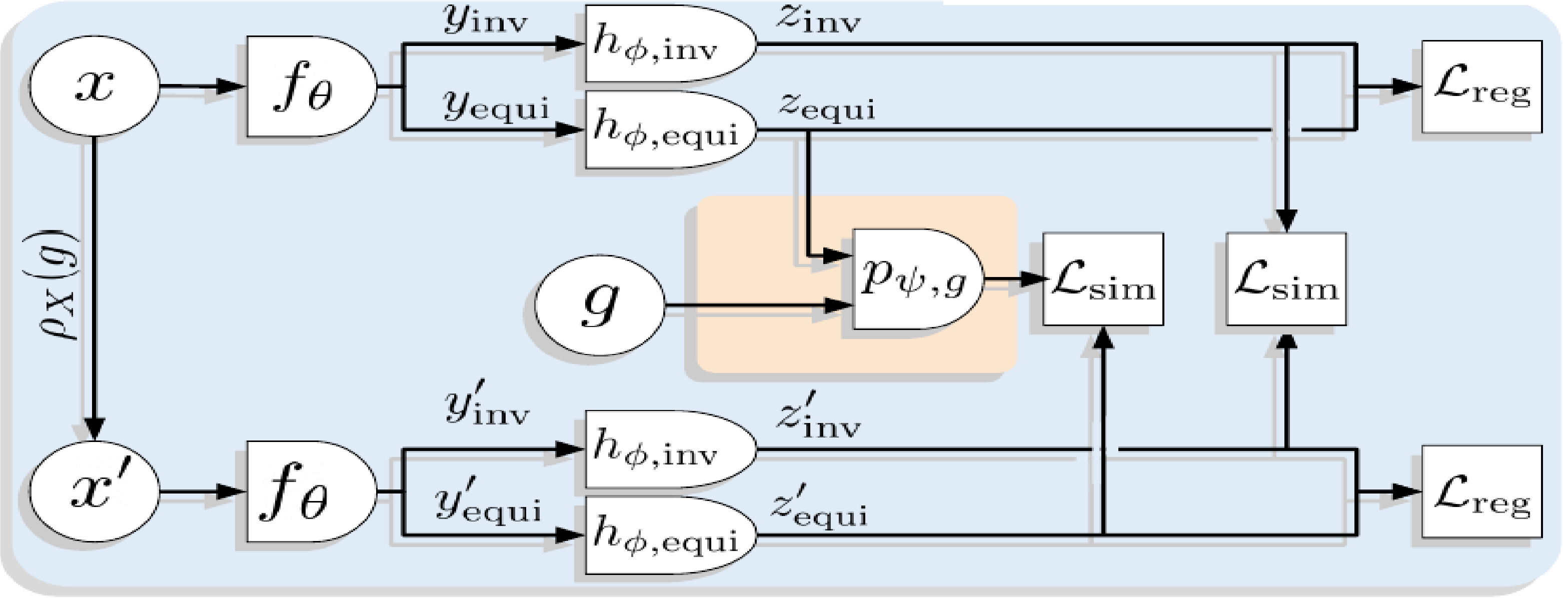
- **Training on multiple rendered views of 3D objects**
  - 3DIEBench dataset
- **Split representation**
  - Invariant part:
    - encodes shape identity
  - Equivariant part:
    - Encodes pose
- **[Garrido ArXiv:2302.10283]**



Split Invariant-Equivariant Representation Learning

- ConvNext backbone
- 2 heads for invariant and equivariant
- Predictor for equivariant part (JEPA)
- Predictor is a hypernetwork
- VC regularization





Split Invariant-Equivariant Representation Learning

Method	Classification (top-1)			Rotation prediction (R^2)			Color prediction (R^2)		
	All	Inv.	Equi.	All	Inv.	Equi.	All	Inv.	Equi.
Baselines									
Supervised	87.47			0.76					
Random				0.23					
Invariant and parameter prediction methods									
VICReg	84.74			0.41			0.06		
VICReg, g kept identical	72.81			0.56			0.25		
SimCLR	**86.73**			0.50			0.30		
SimCLR, g kept identical	71.21			0.54			0.83		
Parameter prediction	85.11			**0.75**			0.12		
Equivariant methods									
Only equivariant (Original predictor)	86.93			0.51			0.23		
Only equivariant (Our predictor)	86.10			0.60			0.24		
EquiMod (Original predictor)	**87.19**			0.47			0.21		
EquiMod (Our predictor)	**87.19**			0.60			0.13		
SIE (Ours)	82.94	82.08	80.32	**0.73**	0.23	0.73	0.07	0.05	0.02

Split Invariant-Equivariant Representation Learning

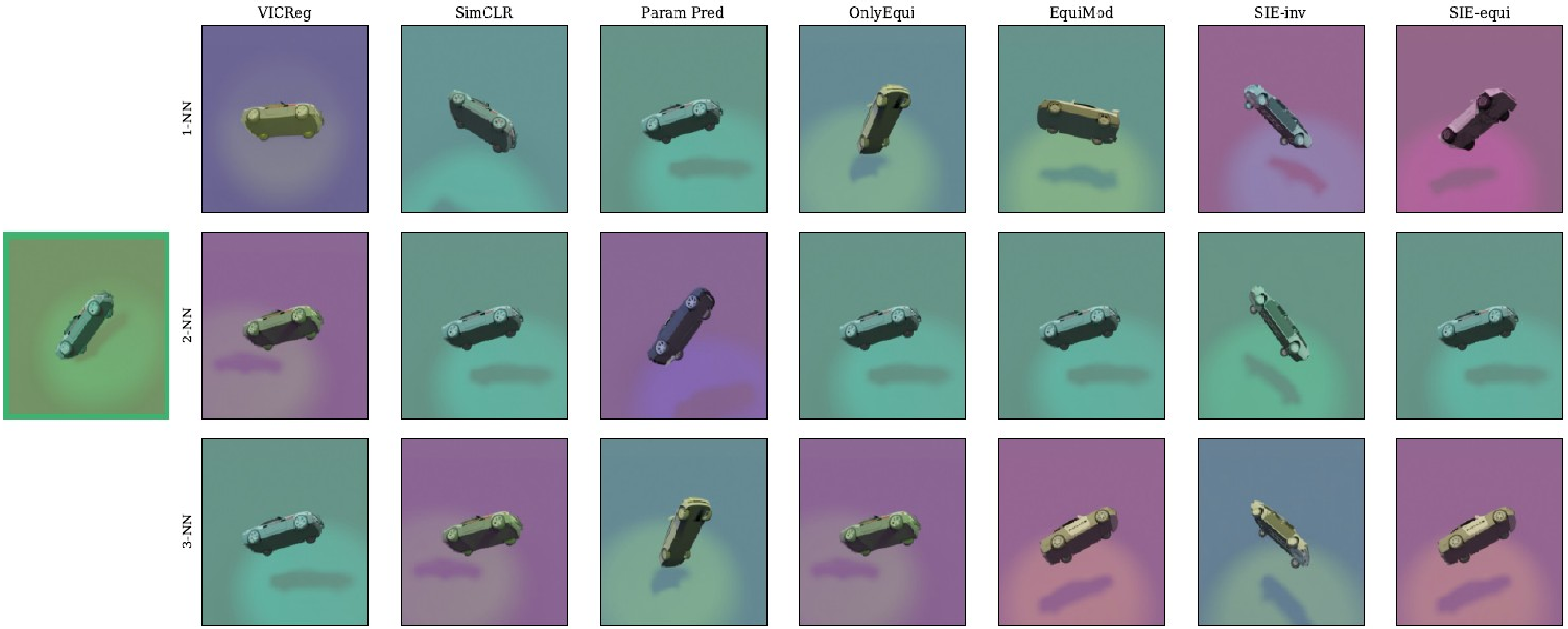


Figure 3: Retrieval of nearest representations. Starting from the representation associate to the object in the green frame on the left, we compute its nearest neighbours for all considered methods and show the 3 closest.

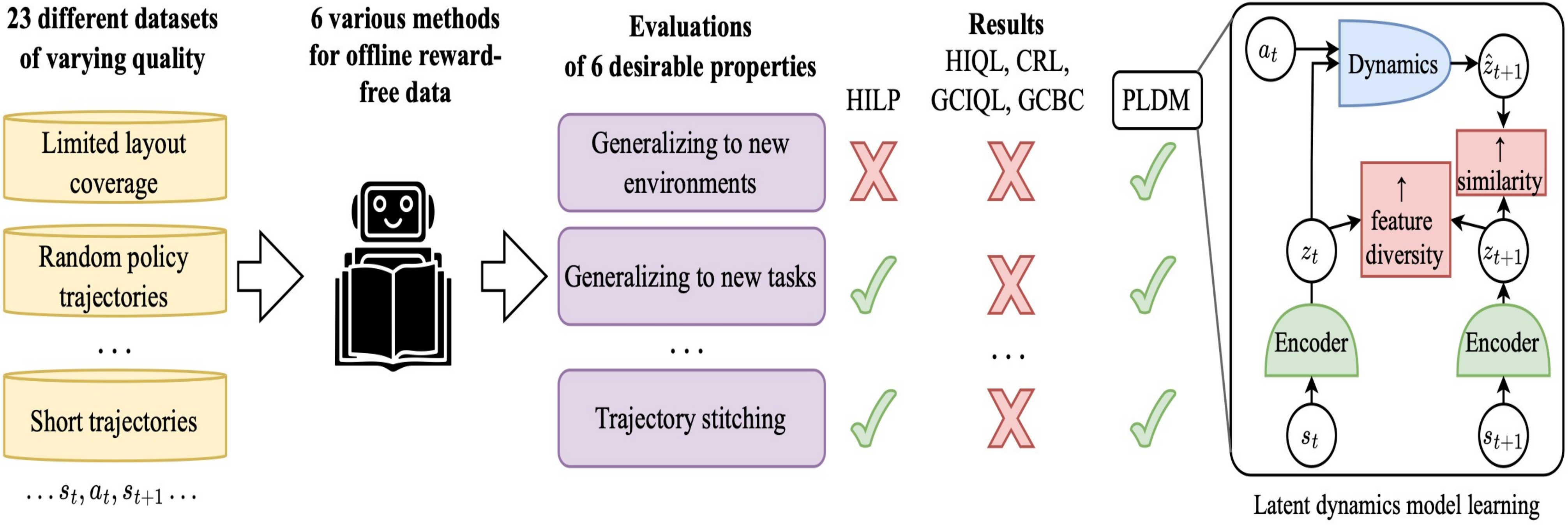
World Model trained with VCReg

Learning from Reward-Free Offline Data:
A Case for Planning with Latent Dynamics Models
Vlad Sobal, Wancong Zhang, Kynghyun Cho, Randall Balestriero, Tim G. J. Rudner, Yann LeCun

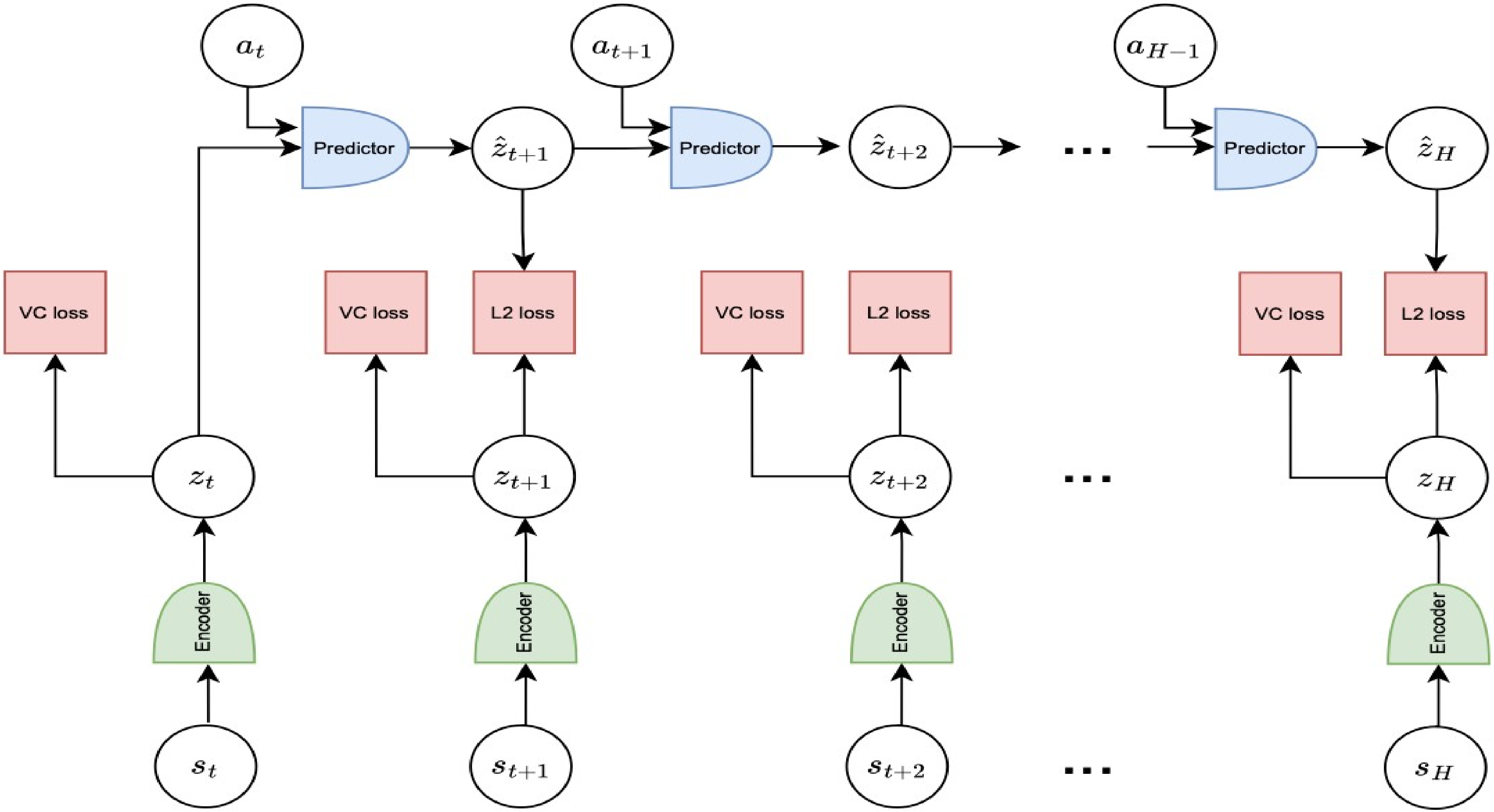
ArXiv:2502.14819
https://latent-planning.github.io/

Planning with Latent-Space Dynamics Model (PLDM)

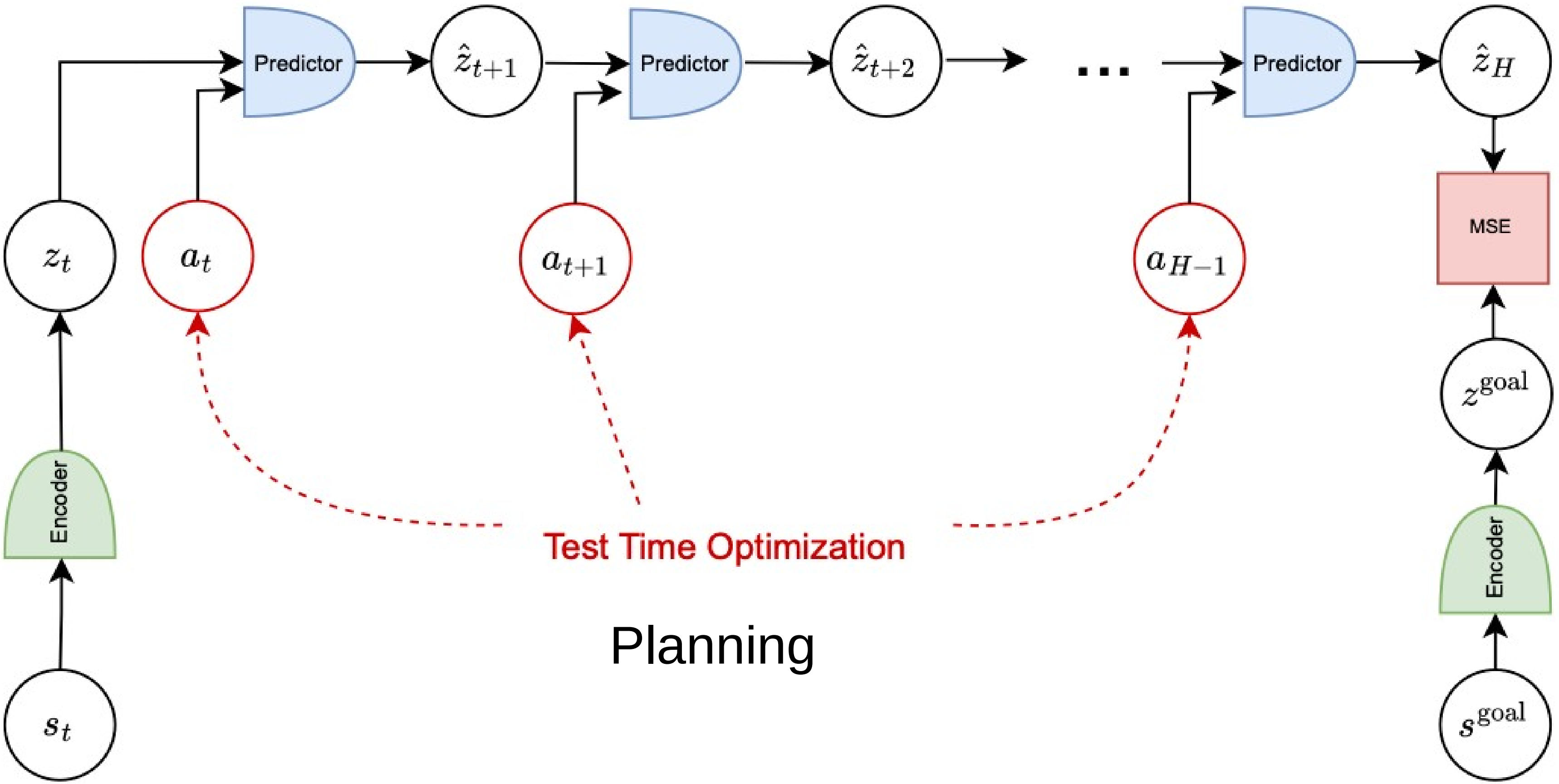
- **23 datasets, 6 methods.**



Training the JEPA with VCReg



Planning with Latent-Space Dynamics Model (PLDM)



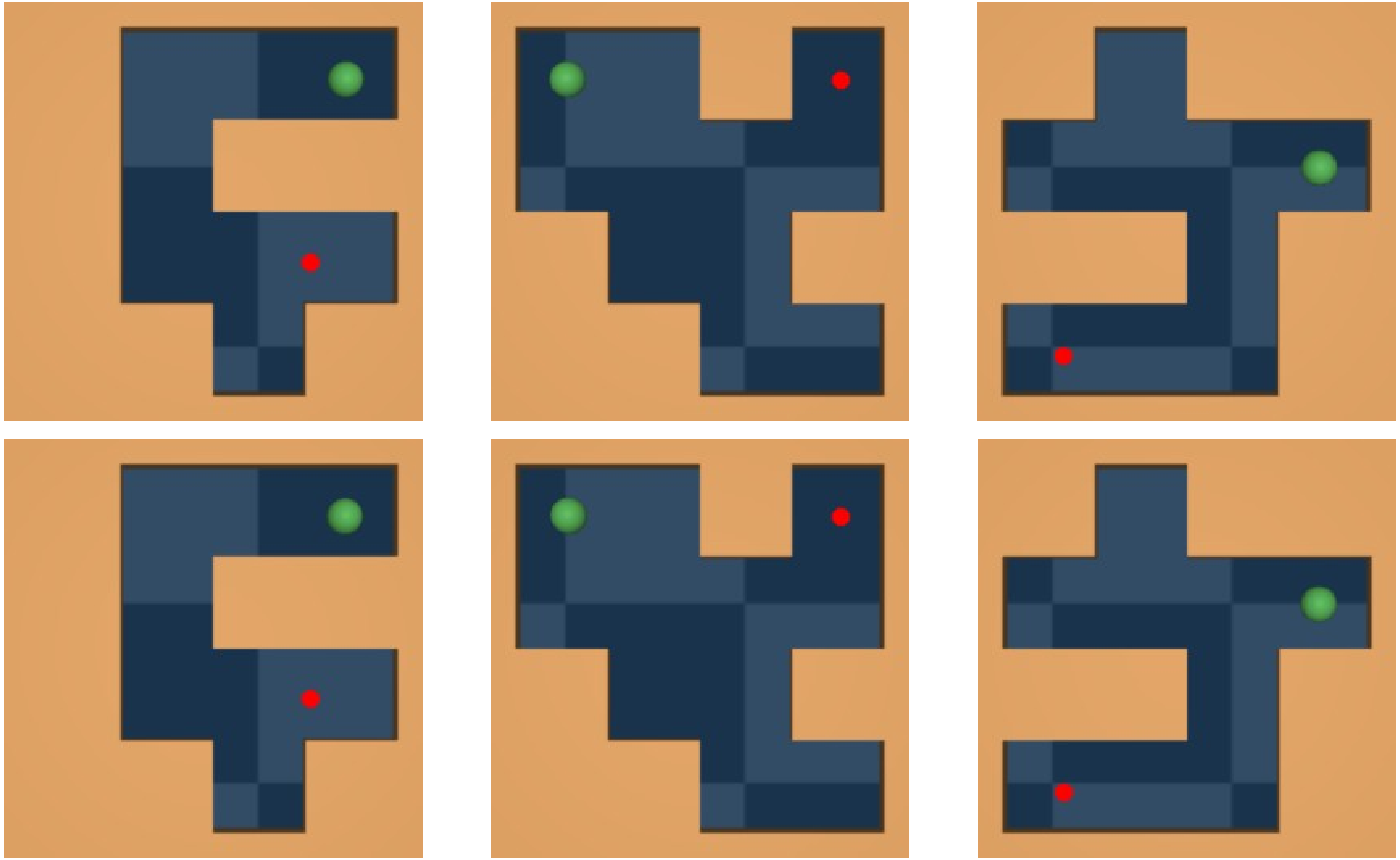
Planning

Planning a path in a maze (visible from an image)

In training | Medium | Out of distribution

PLDM

HIQL



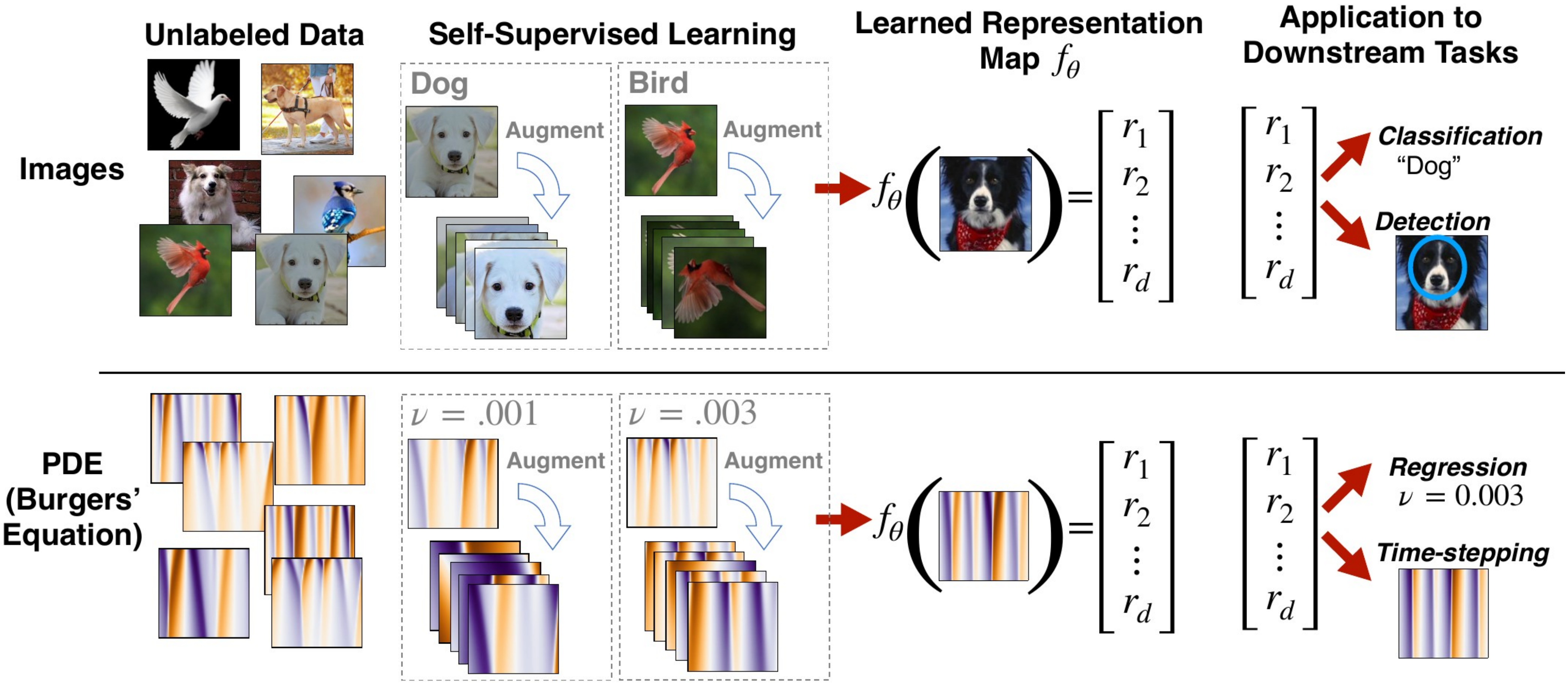
VICReg-based SSL for PDEs

ArXiv:2307.05432   NeurIPS 2023

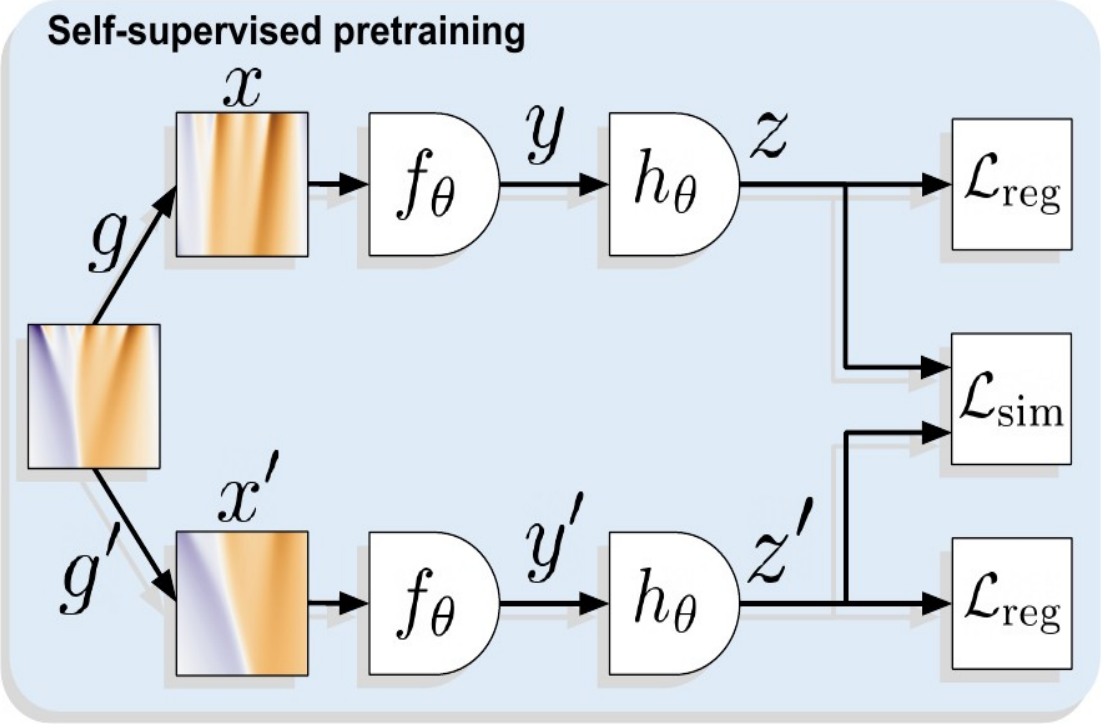
Self-Supervised Learning with Lie Symmetries for Partial Differential Equations

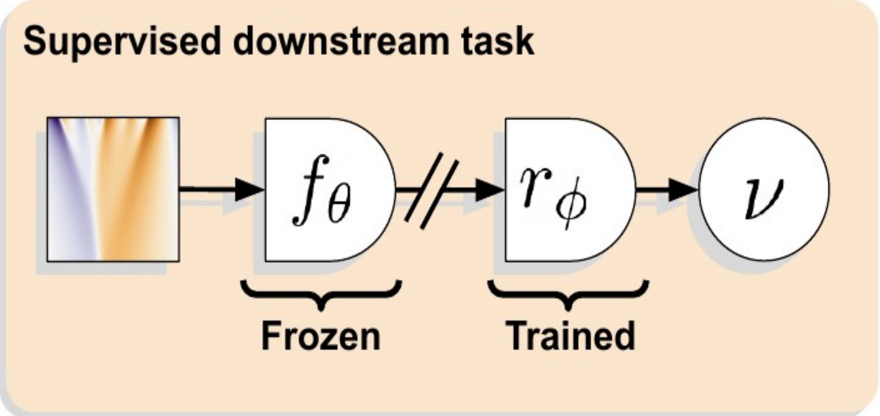
Grégoire Mialon, Quentin Garrido, Hannah Lawrence, Danyal Rehman, Yann LeCun, Bobak T. Kiani

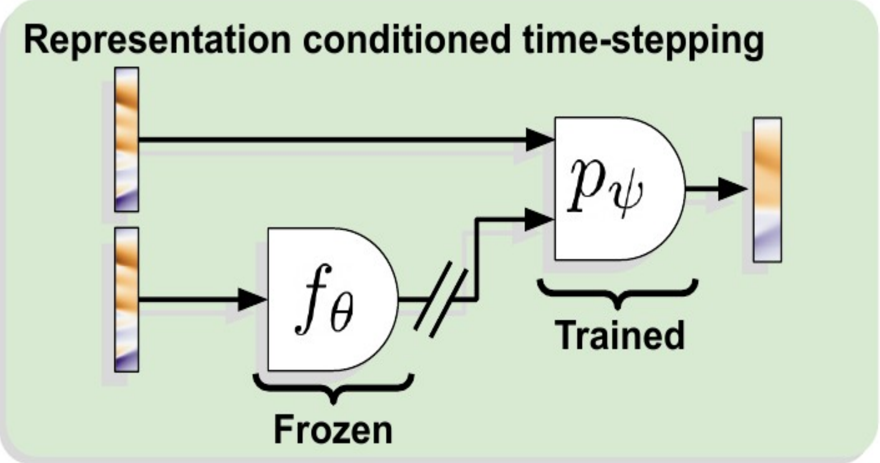
SSL for PDE: extracting dynamical parameters with VICReg



Using VICReg to learn representations of the equation.







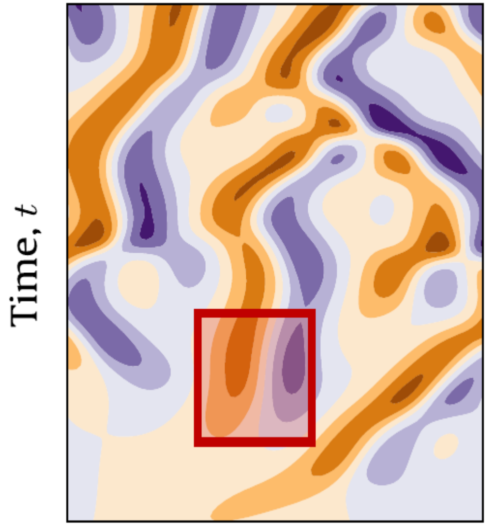
SSL for PDE

An example: the **Kuramoto-Sivashinsky (KS)** equation is a model of chaotic flow given by

$$u_t + uu_x + u_{xx} + u_{xxxx} = 0,$$

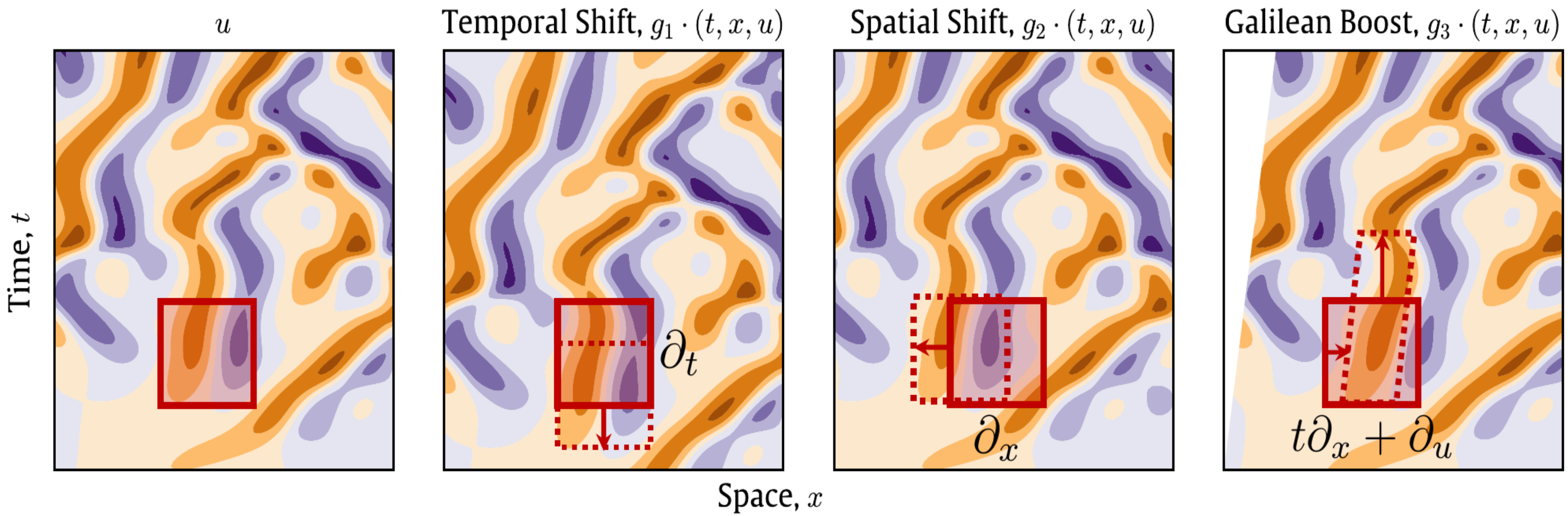
where $u(x, t)$ is the dependent variable.

- Often shows up in reaction-diffusion systems or flame propagation problems.
- Solution can be seen as an image...
- Admit Lie point symmetries: smooth transformations of a solution producing another solution to the same PDE.
- Can be used to learn models [Brandstetter et al., 2022].



A 1D solution to KS (x-axis is space).

SSL for PDE: Data "augmentation"



One parameter Lie point symmetries for the Kuramoto-Sivashinsky (KS) PDE. Left to right: un-modified solution (u), temporal shifts (g_1), spatial shifts (g_2), and Galilean boosts (g_3) with corresponding infinitesimal transformations in the Lie algebra placed inside the figure. The shaded red square denotes the original (x, t), while the dotted line represents the same points after the augmentation is applied.

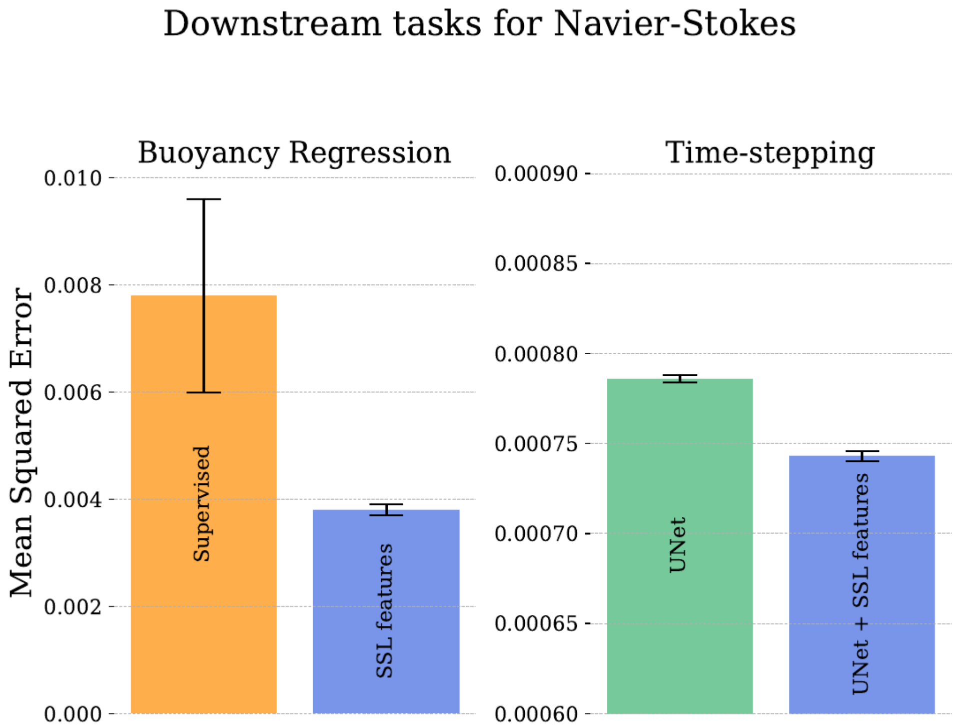
$$
\begin{aligned}
\text{Temporal Shift:} \quad & g_1(\epsilon) : (x, t, u) \mapsto (x, t + \epsilon, u) \\
\text{Spatial Shift:} \quad & g_2(\epsilon) : (x, t, u) \mapsto (x + \epsilon, t, u) \\
\text{Galilean Boost:} \quad & g_3(\epsilon) : (x, t, u) \mapsto (x + \epsilon t, t, u + \epsilon)
\end{aligned}
$$

SSL for Predicting Buoyancy in Navier-Stokes

The **incompressible Navier-Stokes** equation is given by

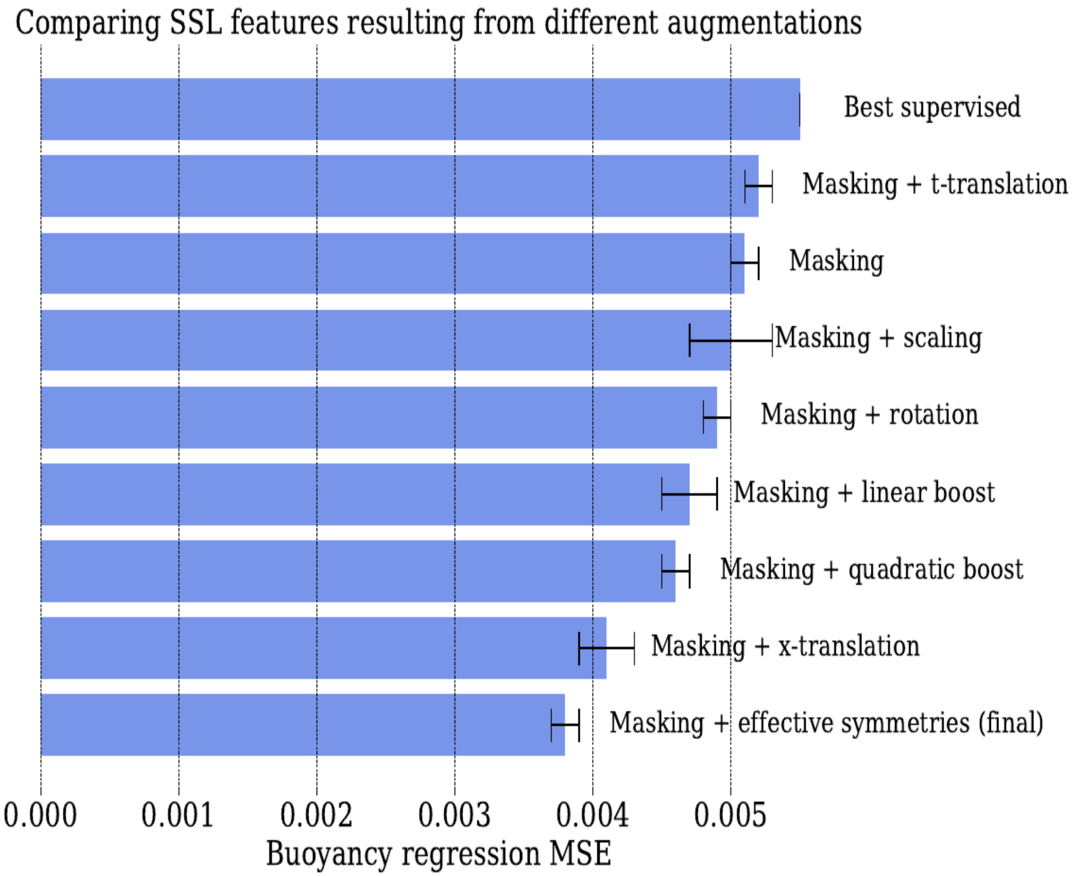
$$\boldsymbol{u}_t = -\boldsymbol{u} \cdot \nabla \boldsymbol{u} - \frac{1}{\rho}\nabla p + \nu \nabla^2 \boldsymbol{u} + \boldsymbol{f}, \quad \nabla \boldsymbol{u} = 0.$$

- 26k 2D trajectories, 56 frames (128x128) each [Gupta and Brandstetter, 2023].
- Task 1: regressing buoyancy $\boldsymbol{f}$.
- Task 2: Time-stepping, predict next frames given past frames.
- SSL features are effective and easy to use.



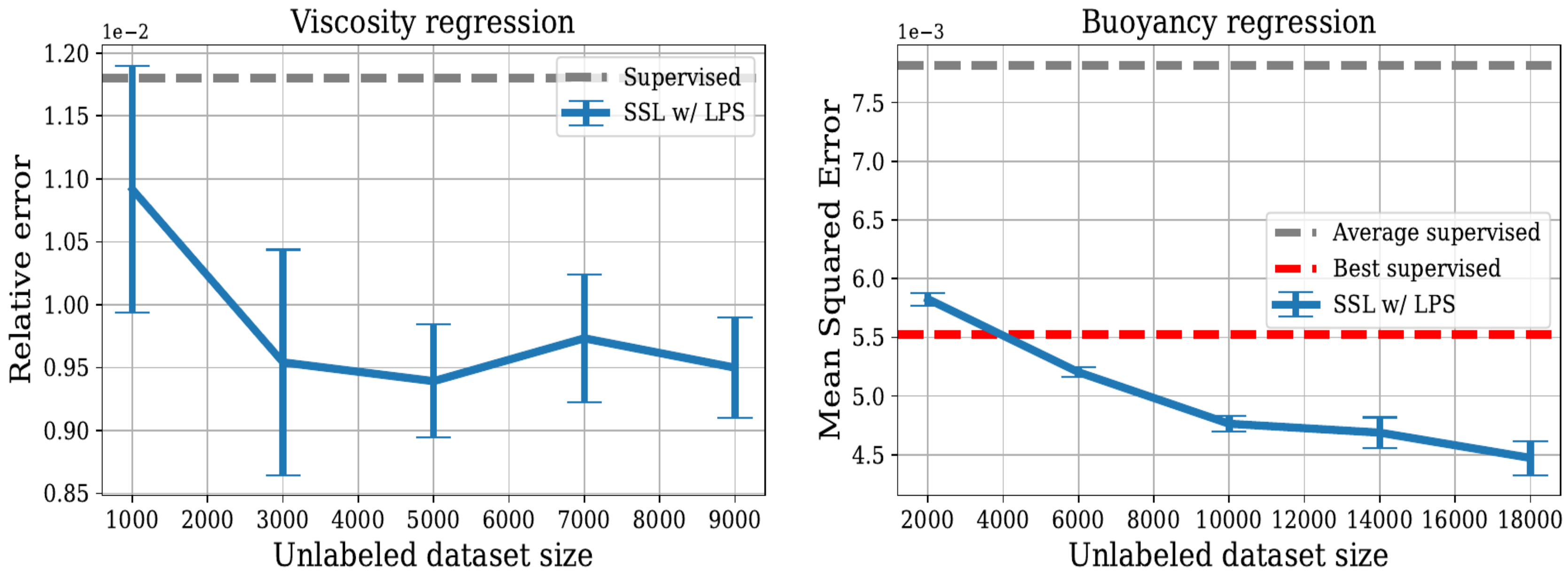
SSL for Predicting Buoyancy in Navier-Stokes

- Navier-Stokes: 8 Lie symmetrie groups, with varying strength.
- Intuition is not sufficient to select augmentations.
- Optimal mix is different from supervised [Brandstetter et al., 2022].
- Masking is necessary but not really sufficient.



SSL pre-training gives better results than purely supervised

SSL vs. supervised: open question in vision [Sariyildiz et al., 2023, Oquab et al., 2023]. Here, big discrepancy.

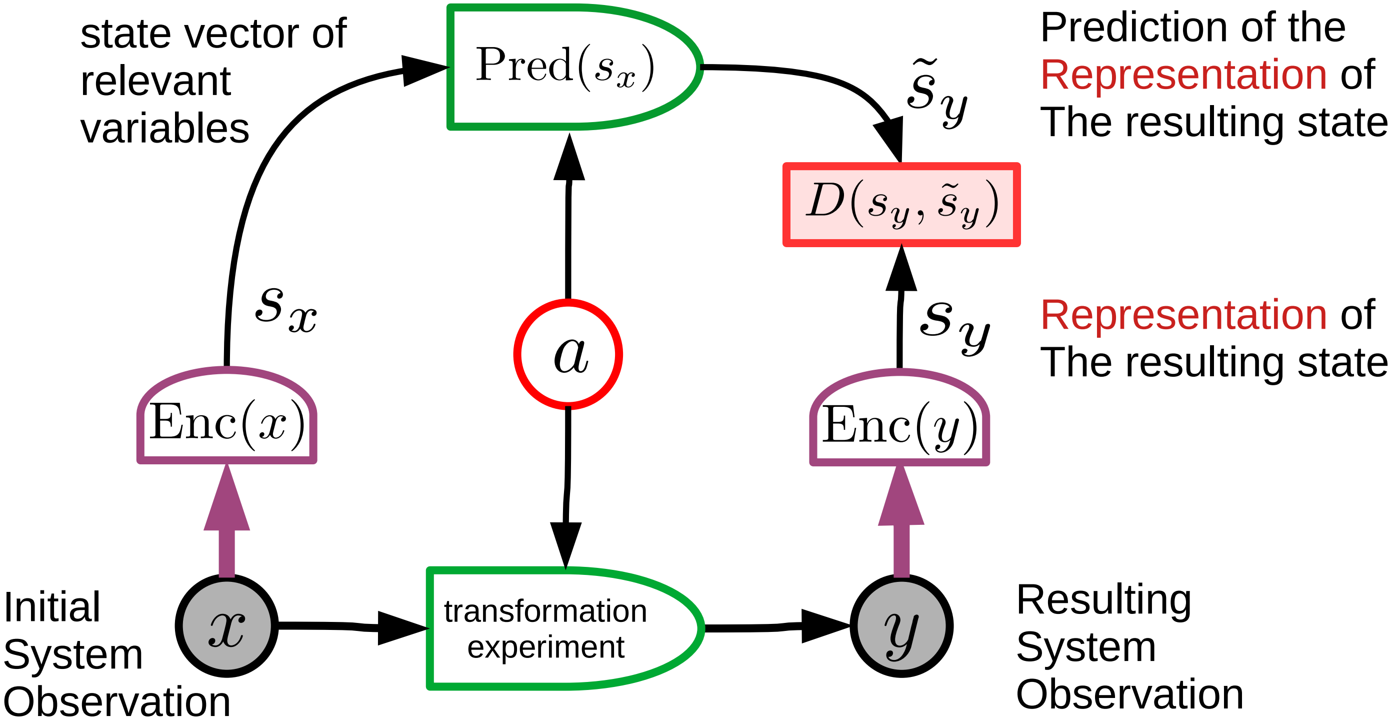


Influence of dataset size on regression tasks. **(Left)** Kinematic regression on Burger's equation. **(Right)** Buoyancy regression on Navier-Stokes' equation.

Science is all about finding abstract representation spaces

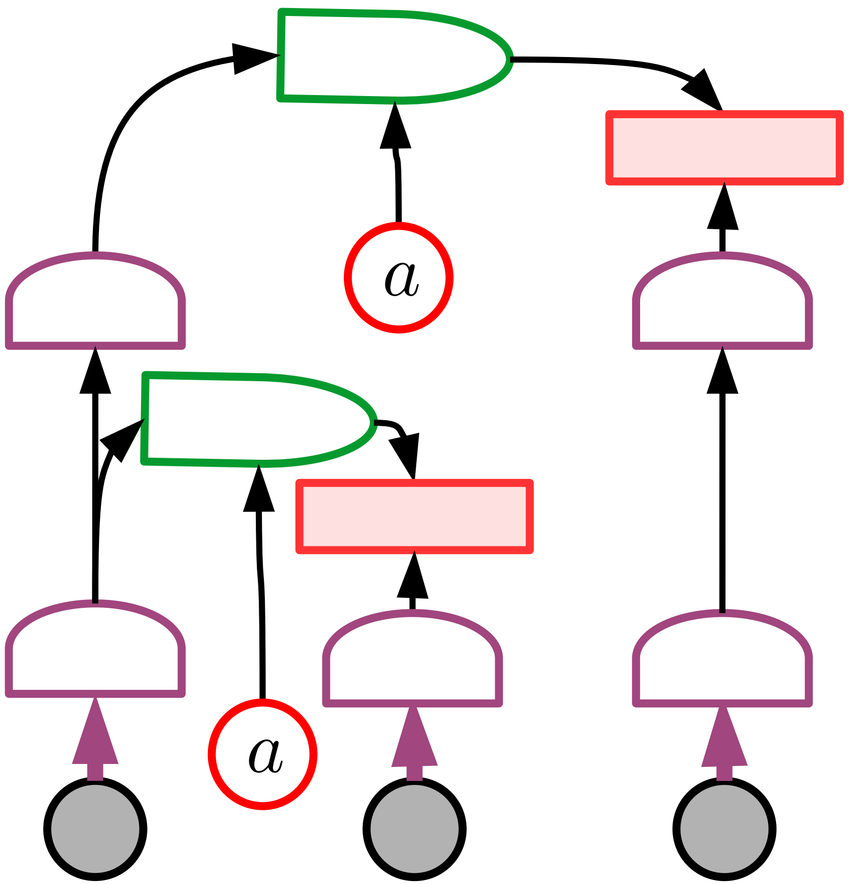
- Find an **abstract state representation** that allows to make predictions
- Extract the state representation from observation/measurement
- Predict outcome resulting from an intervention/experiment

- Irrelevant and unpredictable information is eliminated from the representation
- The representation contains information that makes prediction possible



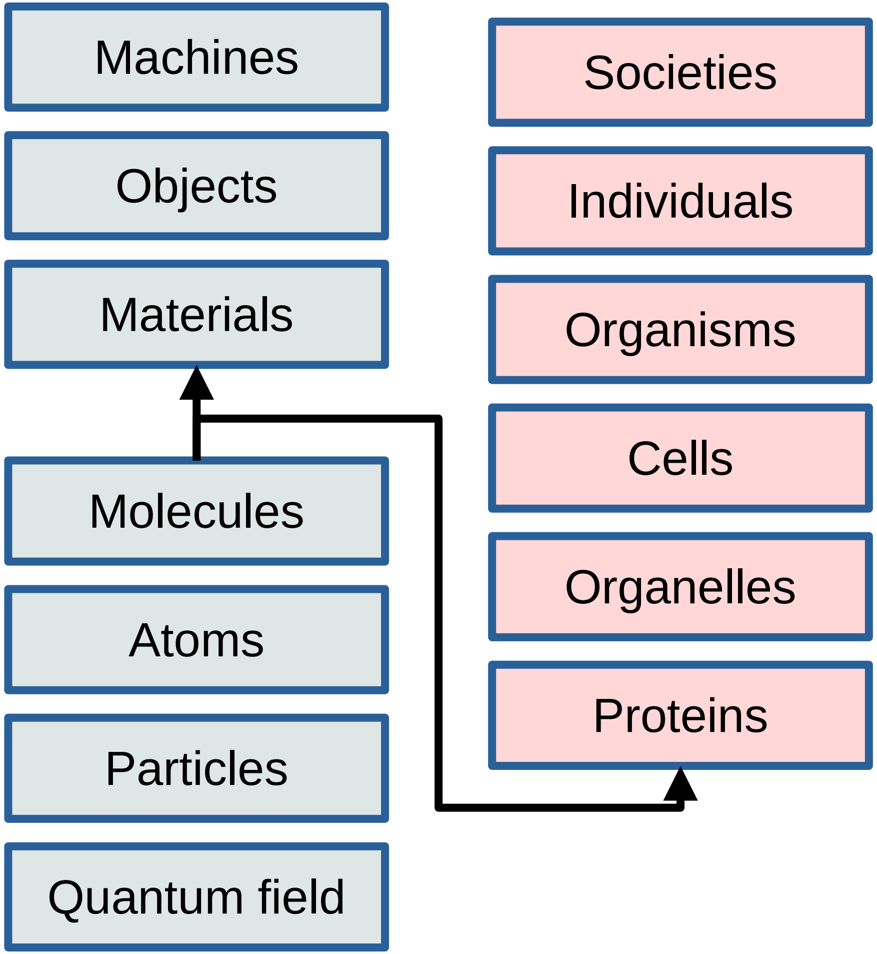
Multi-level hierarchy of models and representations

- **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.



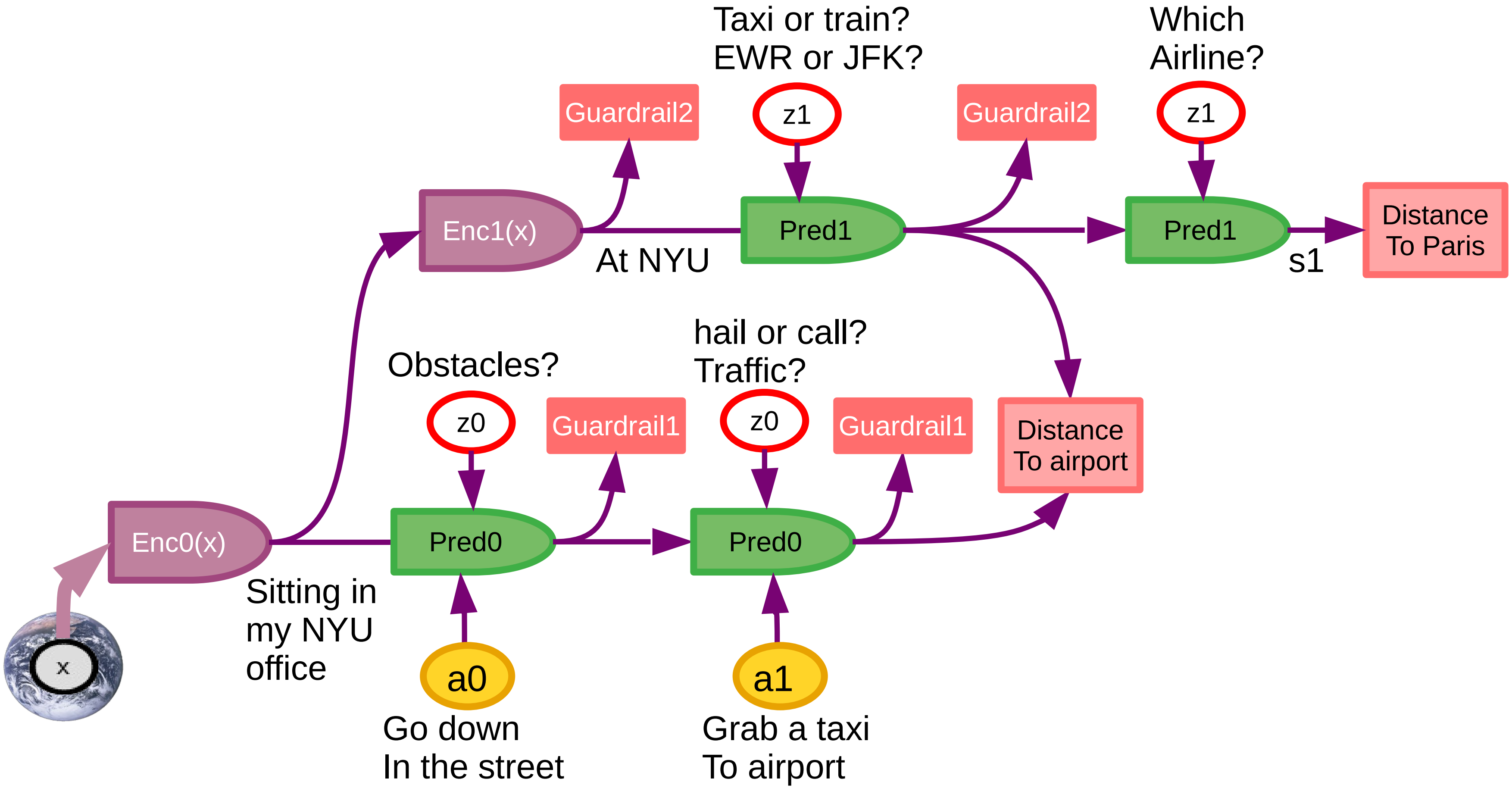
Multi-level hierarchy of models and representations

- **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.



Ultimately, we want Hierarchical World Models

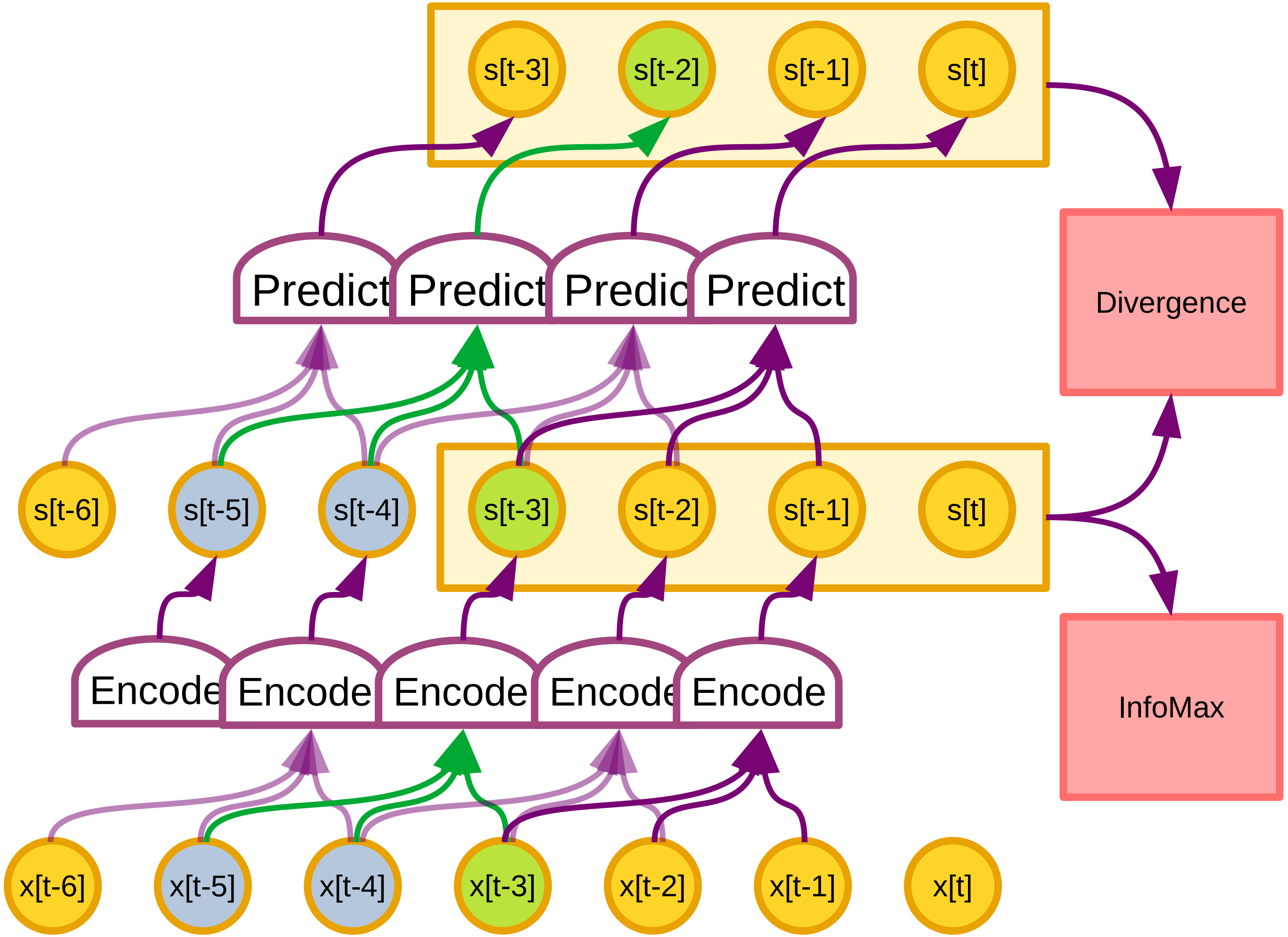
- **Hierarchical Planning: going from NYU to Paris**



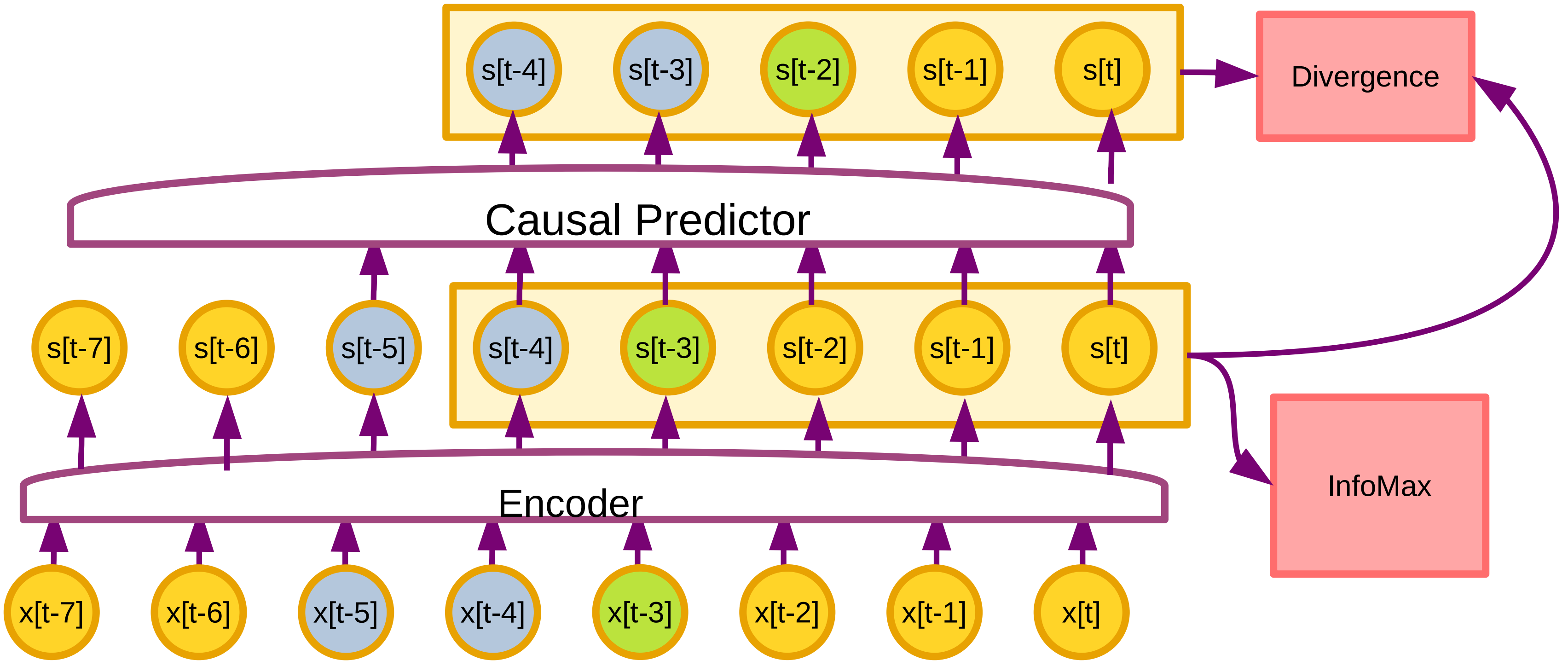
Infomax-Regularized Sequence-Level JEPA

- **Scalable architecture**
- **Causal Predictor**
  - Trained as an auto-encoder
- **Collapse Prevention with InfoMax**
  - e.g. VCReg,

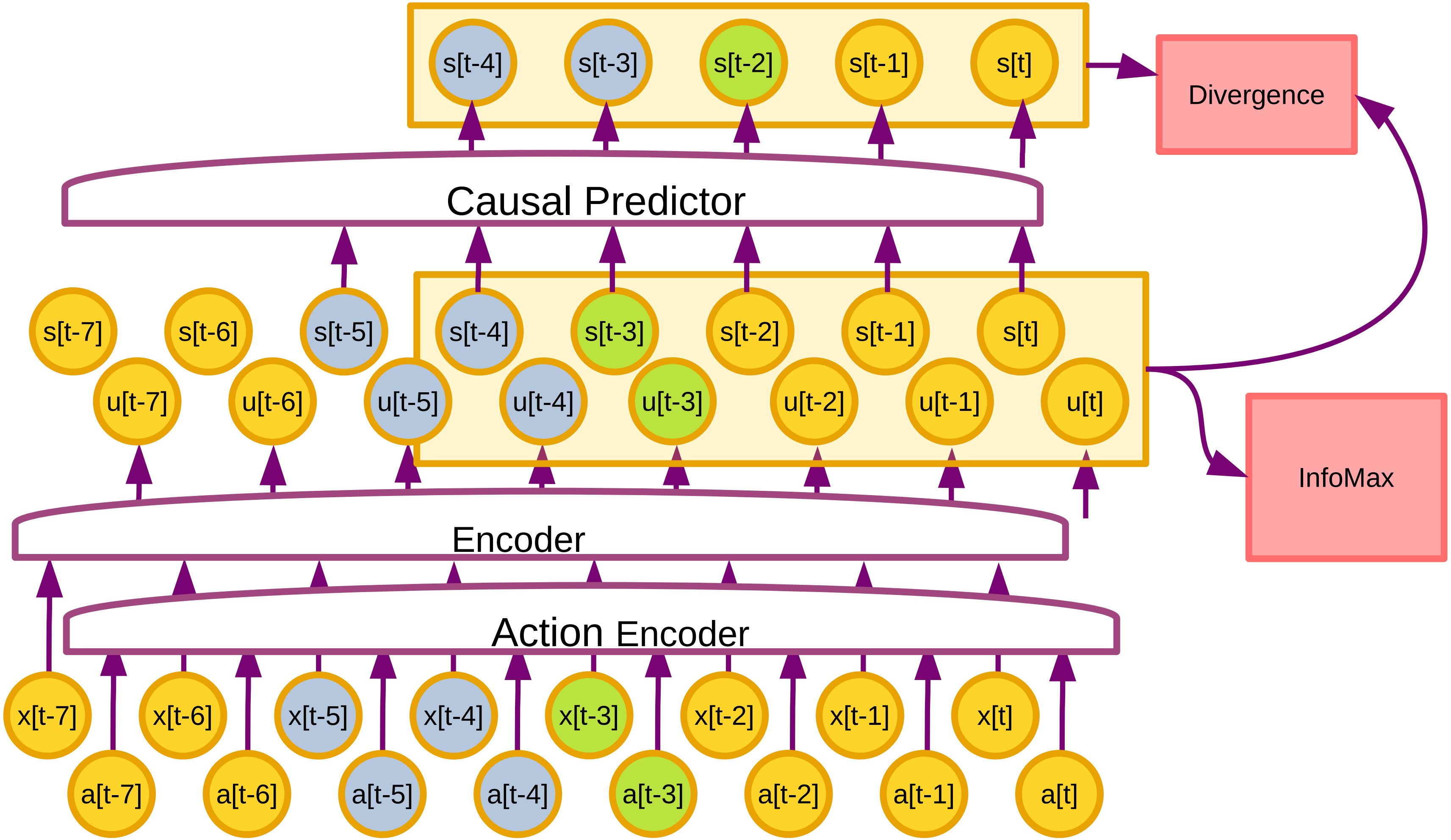
    MCR2, MMCR + others
- **Encoder with limited receptive field**
  - Bounded on the right



Training a Sequence-Level JEPA

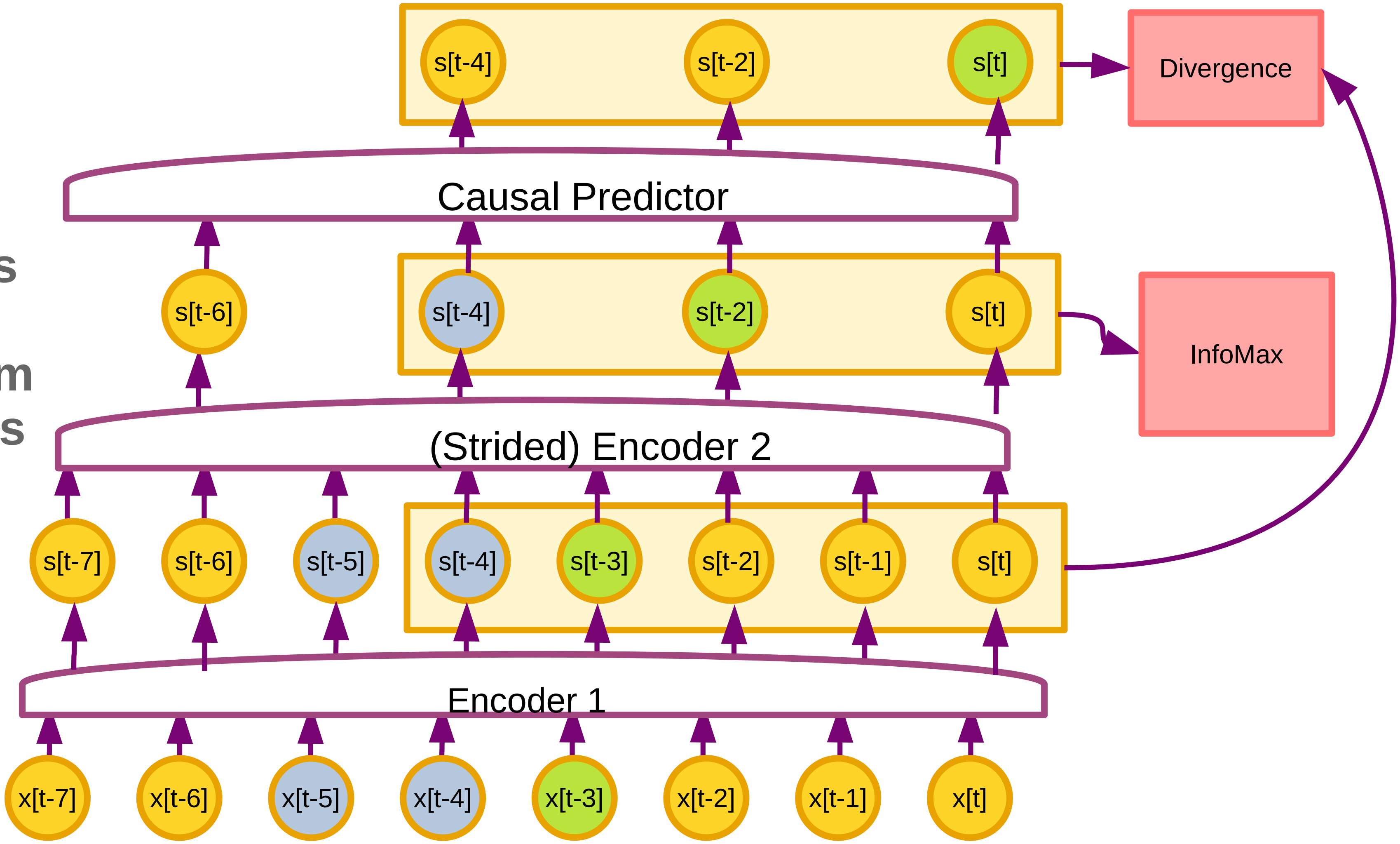


Training an Action-Conditioned Sequence-Level JEPA



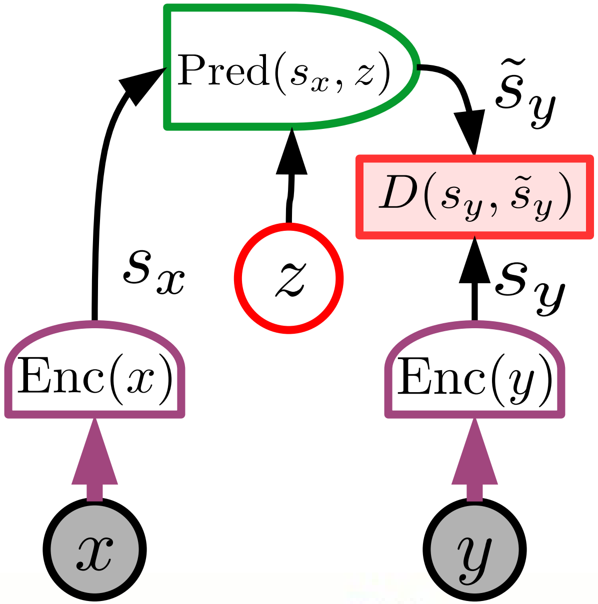
Hierarchical JEPA Architecture and Training

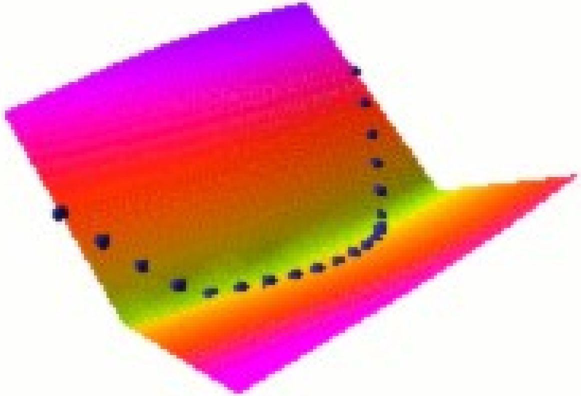
- 2nd stage encoder
- Strided or pooled archi so as to make longer-term predictions



Recommendations:

- **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**





Problems to Solve

- **Large-scale world-model training**
  - From video, speech, text, code, dialogs, math….
- **Planning algorithms**
  - Gradient-based methods, ADMM, gradient-free methods for discrete search
- **JEPA with latent variables**
  - Learning and planning in non-deterministic environments
  - Latent variable regularization to prevent collapse
- **Planning in the presence of uncertainty**
  - Mixed gradient-based / combinatorial optimization
- **Herarchical planning**
- **Very large-scale differentiable associative memories**

Problems to Solve

- **Mathematical Foundations of Energy-Based Learning and inference**
  - The geometry of energy surfaces, scaling laws, bounds…
  - How to maximize information content or minimize low-energy volume?
- **Learning Cost Modules (Inverse RL)**
  - Energy-based approach: give low cost to observed trajectories
- **Planning with inaccurate world models**
  - Preventing bad plans in uncertain parts of the space
- **Exploration to adjust the world models**
  - Intrinsic objectives for curiosity, play
- **New objectives to drive SSL**
  - Driving SSL to focus on interesting or useful features

Future Universal Virtual Assistants

- **All of our interactions with the digital world will be mediated by AI assistants.**
  - They will constitute a repository of all human knowledge and culture
  - They will constitute a shared infrastructure Like the Internet today.
- **These AI platform MUST be open source**
  - We need a diverse set of AI assistants for the same reasons we need a free press: linguistic, cultural, & value system diversity.
  - Culture & knowledge cannot be controlled by a few companies on the West Coast of the US or in China.
- **Open source AI platforms are necessary**

Thank you!

NEW YORK UNIVERSITY Meta AI