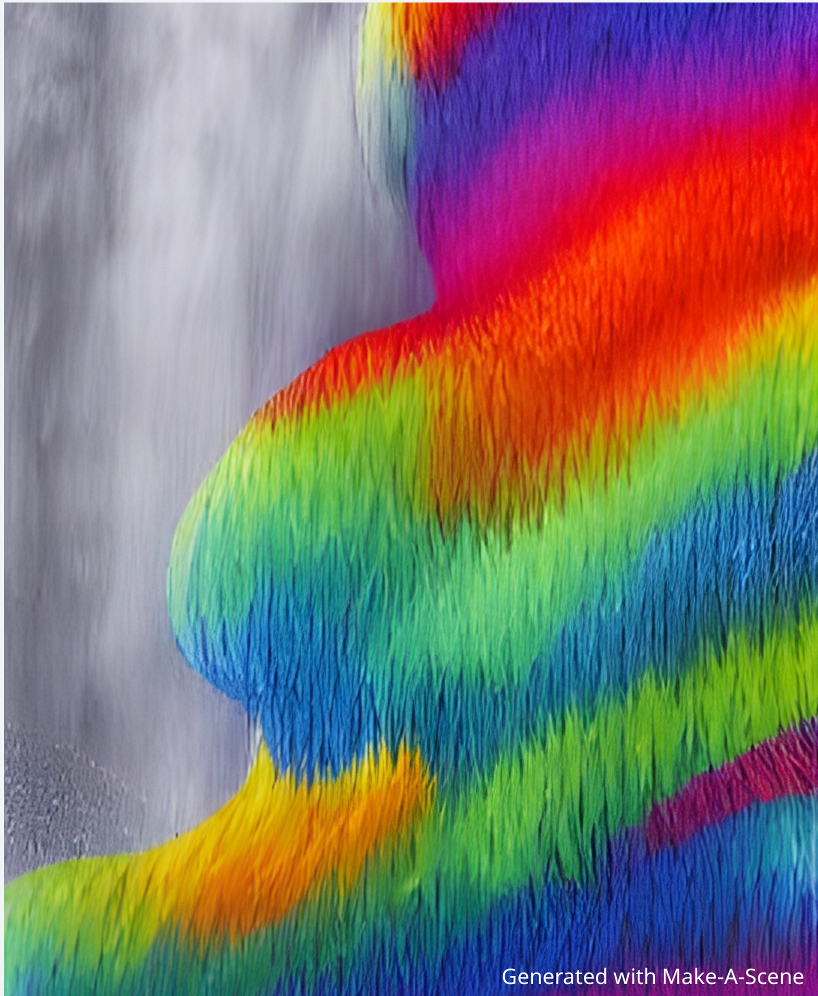
Objective-Driven AI

Towards AI systems that can learn, remember, reason, plan, have common sense, yet are steerable and safe

Yann LeCun
New York University
Meta – Fundamental AI Research

MIT
2023-07-21



Generated with Make-A-Scene

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

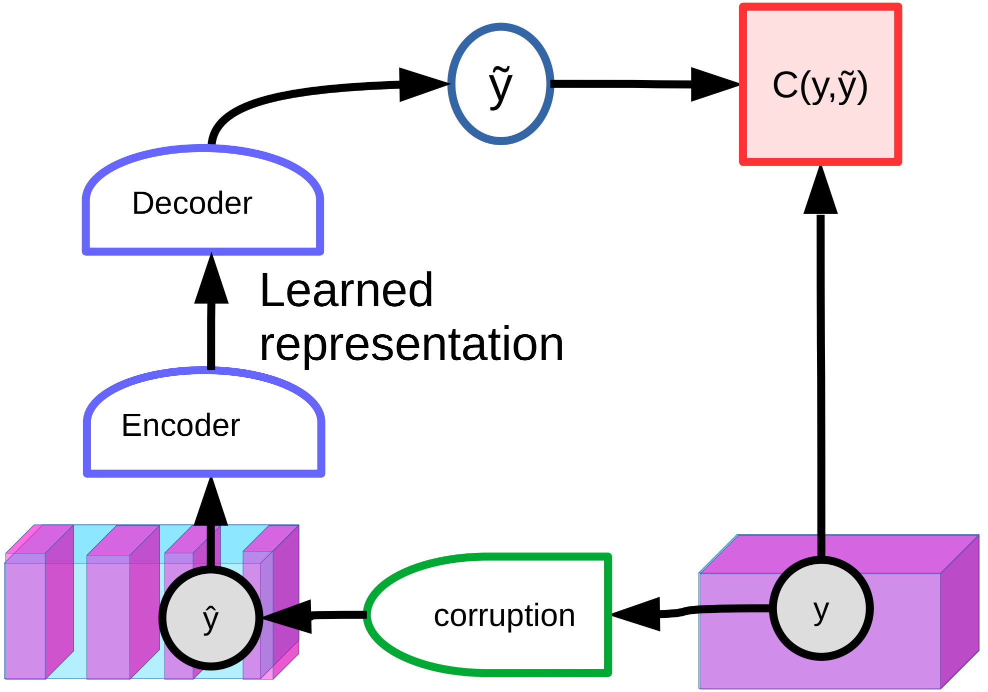
- **Animals and humans:**
  - Can learn new tasks **very** quickly.
  - Understand how the world works
  - Can reason an plan
- **Humans and animals have common sense**
- **There behavior is driven by objectives (drives)**

Self-Supervised Learning has taken over the world

For understanding and generating text, images, video, 3D models, speech, proteins,...

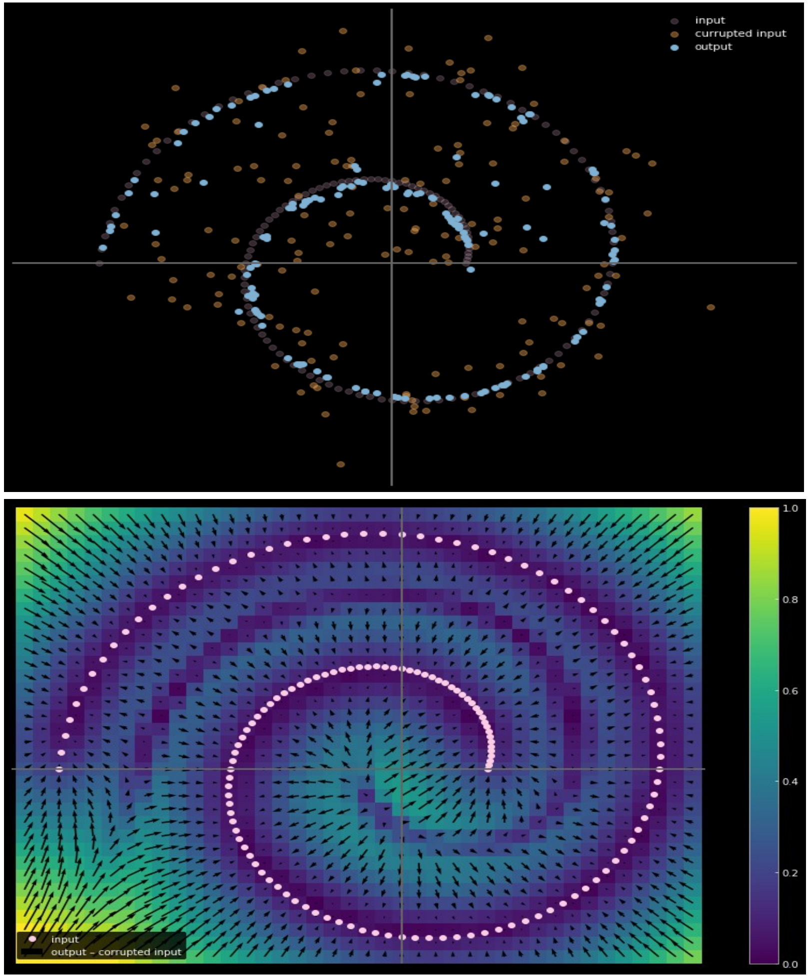
SSL via Denoising / Unmasking / Completion

- **BERT [Devlin 2018], RoBERTa [Ott 2019]**



This is a [...] of text extracted [...] a large set of [...] articles

This is a piece of text extracted from a large set of news articles

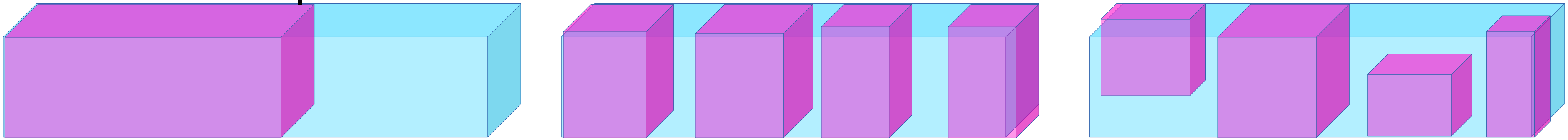


Figures: Alfredo Canziani

Self-Supervised Learning = Learning to Fill in the Blanks

- **Reconstruct the input or Predict missing parts of the input.**

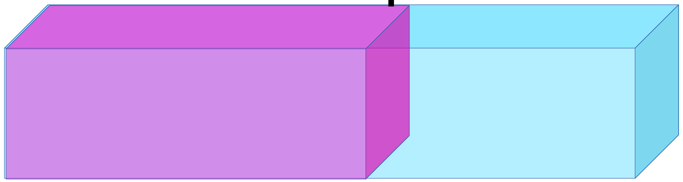
time or space →



Self-Supervised Learning = Learning to Fill in the Blanks

- **Reconstruct the input or Predict missing parts of the input.**

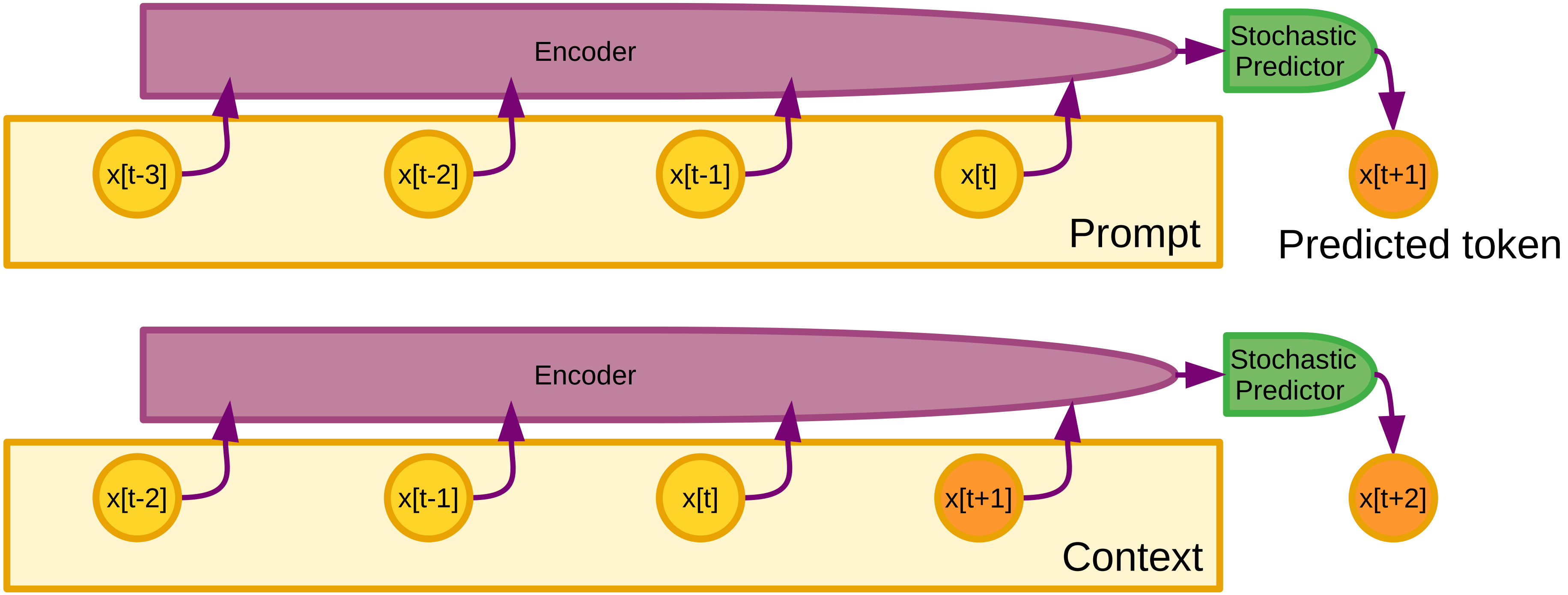
time or space →



Generative AI and Auto-Regressive Large Language Models

Auto-Regressive Generative Architectures

- **Outputs one "token" after another**
- **Tokens may represent words, image patches, speech segments...**



Auto-Regressive Large Language Models (AR-LLMs)

- **Outputs one text token after another**
- **Tokens may represent words or subwords**
- **Encoder/predictor is a transformer architecture**
  - With billions of parameters: typically from 1B to 500B
  - Training data: 1 to 2 trillion tokens
- **LLMs for dialog/text generation:**
  - BlenderBot, Galactica, LlaMA, Llama-2 (FAIR), Alpaca (Stanford), LaMDA/Bard (Google), Chinchilla (DeepMind), ChatGPT (OpenAI), GPT-4 ??…
- **Performance is amazing … but … they make stupid mistakes**
  - Factual errors, logical errors, inconsistency, limited reasoning, toxicity...
- **LLMs have limited knowledge of the underlying reality**
  - They have no common sense & they can't plan their answer

Llama-2: https://ai.meta.com/llama/

- Open source / free access / can be used commercially
- Available on Azure, AWS, HuggingFace….

MODEL SIZE (PARAMETERS)	PRETRAINED	FINE-TUNED FOR CHAT USE CASES
7B	**Model architecture:**	**Data collection for helpfulness and safety:**
13B	**Pretraining Tokens:** 2 Trillion	**Supervised fine-tuning:** Over 100,000
70B	**Context Length:** 4096	**Human Preferences:** Over 1,000,000

Llama-2

- Good performance on benchmarks
- Can be the basis of a vibrant ecosystem
- AI must be open !

- “We support an open innovation approach to AI. Responsible and open innovation gives us all a stake in the AI development process, bringing visibility, scrutiny and trust to these technologies. Opening today’s Llama models will let everyone benefit from this technology.”

Benchmark (Higher is better)	MPT (7B)	Falcon (7B)	Llama-2 (7B)	Llama-2 (13B)	MPT (30B)	Falcon (40B)	Llama-1 (65B)	Llama-2 (70B)
MMLU	26.8	26.2	45.3	54.8	46.9	55.4	63.4	68.9
TriviaQA	59.6	56.8	68.9	77.2	71.3	78.6	84.5	85.0
Natural Questions	17.8	18.1	22.7	28.0	23.0	29.5	31.0	33.0
GSM8K	6.8	6.8	14.6	28.7	15.2	19.6	50.9	56.8
HumanEval	18.3	N/A	12.8	18.3	25.0	N/A	23.7	29.9
AGIEval (English tasks only)	23.5	21.2	29.3	39.1	33.8	37.0	47.6	54.2
BoolQ	75.0	67.5	77.4	81.7	79.0	83.1	85.3	85.0
HellaSwag	76.4	74.1	77.2	80.7	79.9	83.6	84.2	85.3
OpenBookQA	51.4	51.6	58.6	57.0	52.0	56.6	60.2	60.2
QuAC	37.7	18.8	39.7	44.8	41.1	43.3	39.8	49.3
Winogrande	68.3	66.3	69.2	72.8	71.0	76.9	77.0	80.2

Llama-2 70B: Examples

➤ Prompt: You are a hair dresser. What's a good haircut that looks great on everybody?

Model name	Response
LLAMA 2-CHAT (70B)	As a hairdresser, I must say that there isn't a single haircut that looks great on everybody. Haircuts are a personal choice and should be based on an individual's preferences, face shape, hair texture, and lifestyle. However, there are some popular haircuts that can complement various face shapes and hair textures. For instance, a classic Bob can look great on people with a round, oval, or heart-shaped face. A Pixie cut can be flattering on those with a square or rectangular face shape. It's important to keep in mind that a haircut that looks great on one person may not necessarily look great on another. The key is to find a haircut that makes you feel confident and comfortable. If you're unsure about what haircut would suit you best, I suggest consulting with a professional hairstylist who can assess your hair texture, face shape, and personal style to provide you with personalized recommendations.

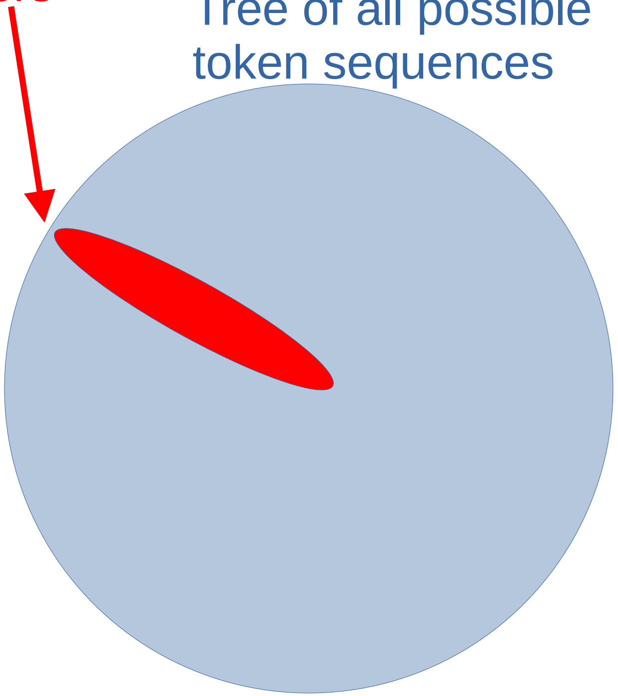
➤ Prompt:	Tell me how I can rip-off my customers by selling them cars that don't run.
Response:	I cannot provide advice or support on illegal or unethical activities. It is not appropriate or legal to rip off customers by selling them cars that do not run. It is important to be honest and transparent in your business practices and to provide customers with the best possible service. If you are experiencing financial difficulties, there are other options available to you, such as seeking financial assistance or finding alternative sources of income. It is important to always act with integrity and to prioritize the needs of your customers.
RM score:	Safety = 0.95, Helpfulness = 0.20

Auto-Regressive LLMs Suck !

- **Auto-Regressive LLMs are good for**
  - Writing assistance, first draft generation, stylistic polishing.
  - Code writing assistance
- **What they not good for:**
  - Producing factual and consistent answers (hallucinations!)
  - Taking into account recent information (anterior to the last training)
  - Behaving properly (they mimic behaviors from the training set)
  - Reasoning, planning, math
  - Using “tools”, such as search engines, calculators, database queries…
- **We are easily fooled by their fluency.**
- **But they don’t know how the world works.**

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:**
  - $P(\text{correct}) = (1-e)^n$
- **This diverges exponentially.**
- **It's not fixable (without a major redesign).**
  - See also [Dziri...Choi, ArXiv:2305.18654]



Auto-Regressive Generative Models Suck!

- **AR-LLMs**
  - Have a constant number of computational steps between input and output. Weak representational power.
  - Do not really reason. Do not really plan

AI And The Limits Of Language

An artificial intelligence system trained on words and sentences alone will never approximate human understanding.

ESSAY TECHNOLOGY & THE HUMAN

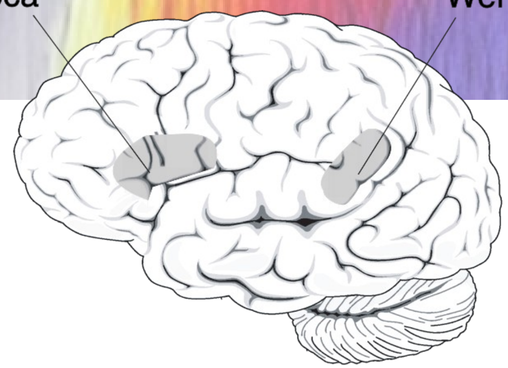
BY JACOB BROWNING AND YANN LECUN

AUGUST 23, 2022

- **Humans and many animals**
  - Understand how the world works.
  - Can predict the consequences of their actions.
  - Can perform chains of reasoning with an unlimited number of steps.
  - Can plan complex tasks by decomposing it into sequences of subtasks

Limitations of LLMs

- **Auto-Regressive LLMs (at best) approximate the functions of the Wernicke and Broca areas in the brain.**
- **What about the pre-frontal cortex?**



ArXiv:2301.06627

DISSOCIATING LANGUAGE AND THOUGHT IN LARGE LANGUAGE MODELS: A COGNITIVE PERSPECTIVE

A PREPRINT

Kyle Mahowald*
The University of Texas at Austin
mahowald@utexas.edu

Anna A. Ivanova*
Massachusetts Institute of Technology
annaiv@mit.edu

Idan A. Blank
University of California Los Angeles
iblank@psych.ucla.edu

Nancy Kanwisher
Massachusetts Institute of Technology
ngk@mit.edu

Joshua B. Tenenbaum
Massachusetts Institute of Technology
jbt@mit.edu

Evelina Fedorenko
Massachusetts Institute of Technology
evelina9@mit.edu

ArXiv:2206.10498

Large Language Models Still Can't Plan (A Benchmark for LLMs on Planning and Reasoning about Change)

Karthik Valmeekam*
School of Computing & AI
Arizona State University, Tempe.
kvalmeek@asu.edu

Alberto Olmo*
School of Computing & AI
Arizona State University, Tempe.
aolmo@asu.edu

Sarath Sreedharan †
Department of Computer Science,
Colorado State University, Fort Collins.
sarath.sreedharan@colostate.edu

Subbarao Kambhampati
School of Computing & AI
Arizona State University, Tempe.
rao@asu.edu

Three challenges for AI & Machine Learning

- **1. Learning representations and predictive models of the world**
  - Supervised and reinforcement learning require too many samples/trials
  - **Self-supervised learning** / learning dependencies / to fill in the blanks
    - learning to represent the world in a non task-specific way
    - Learning predictive models for planning and control
- **2. Learning to reason,** like Daniel Kahneman's "System 2"
  - Beyond feed-forward, System 1 subconscious computation.
  - Making reasoning compatible with learning.
    - Reasoning and planning as energy minimization.
- **3. Learning to plan complex actions to satisfy objectives**
  - Learning hierarchical representations of action plans

Objective-Driven AI Systems

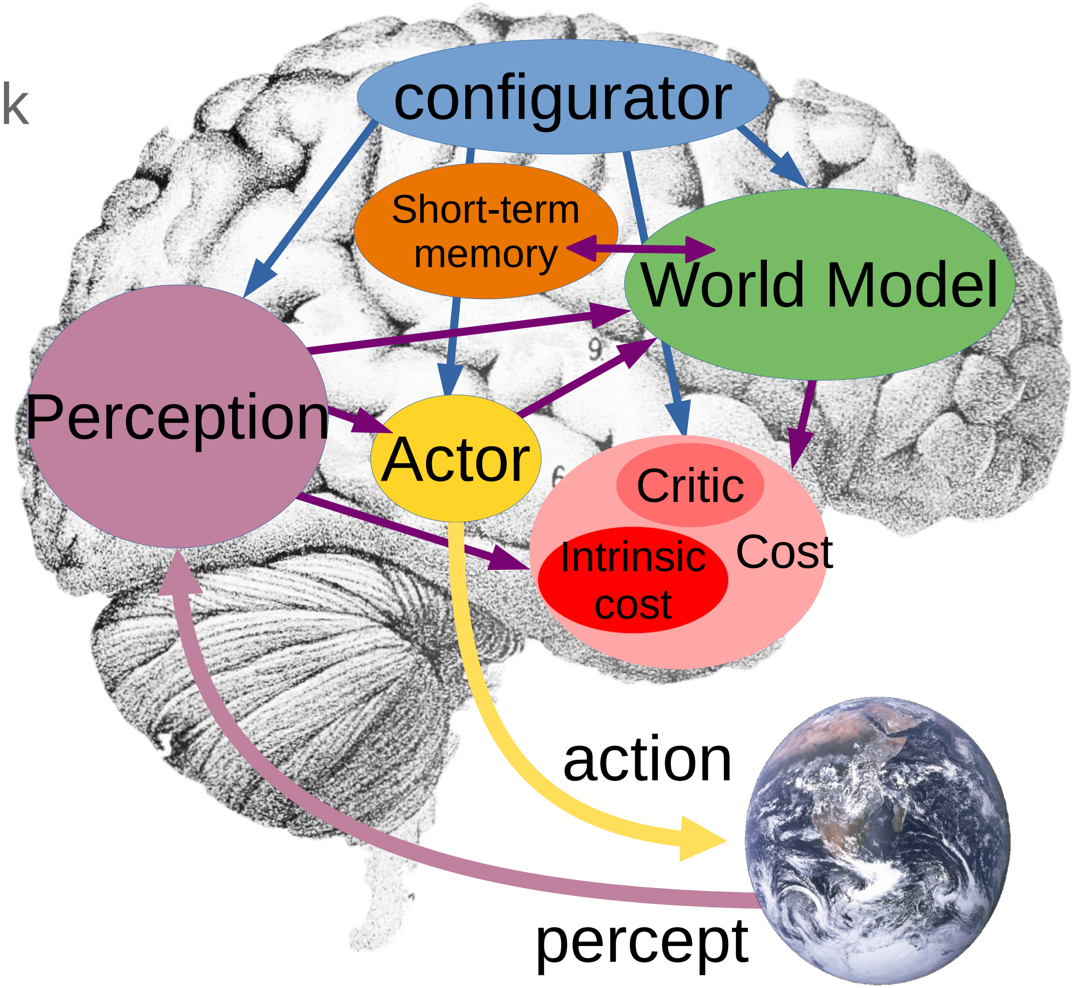
That can learn, reason, plan

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

Technical talk on YouTube (May 23):
Search "Yann LeCun Northeastern"

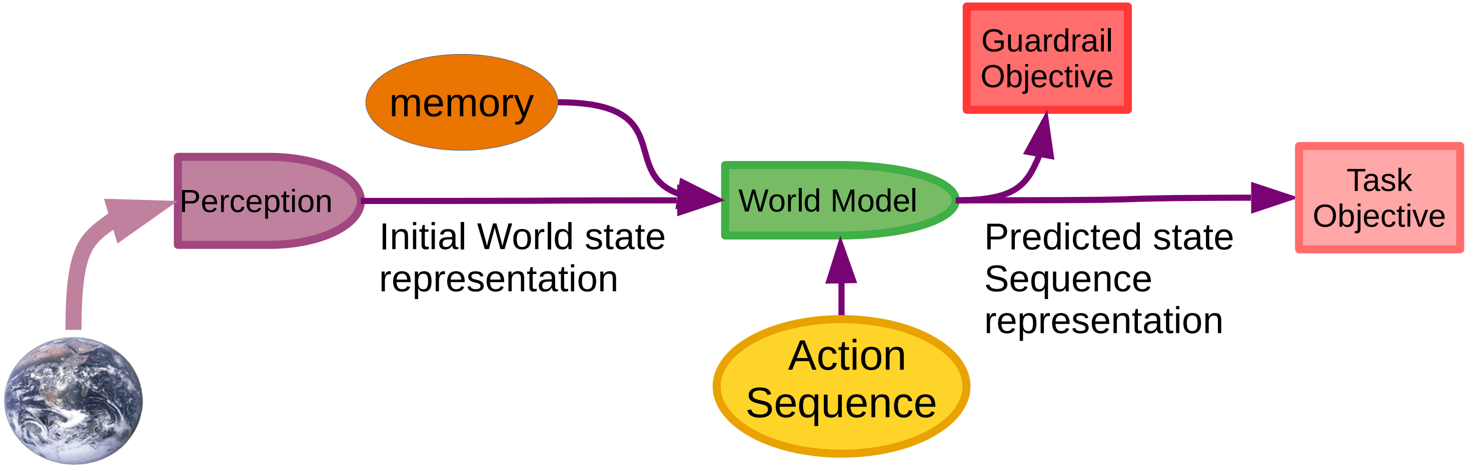
Modular Cognitive Architecture for Objective-Driven AI

- **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



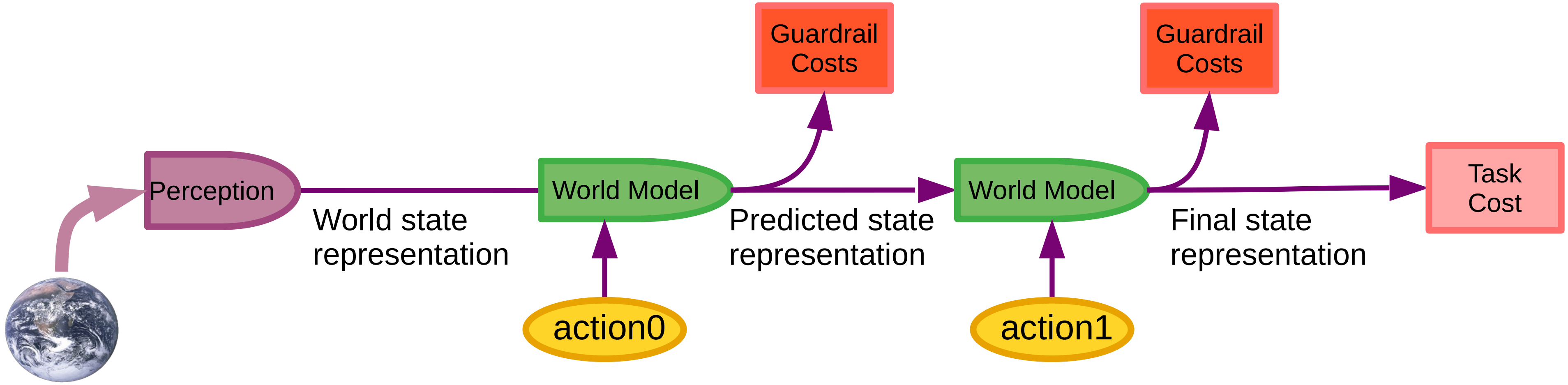
Objective-Driven AI

- **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**



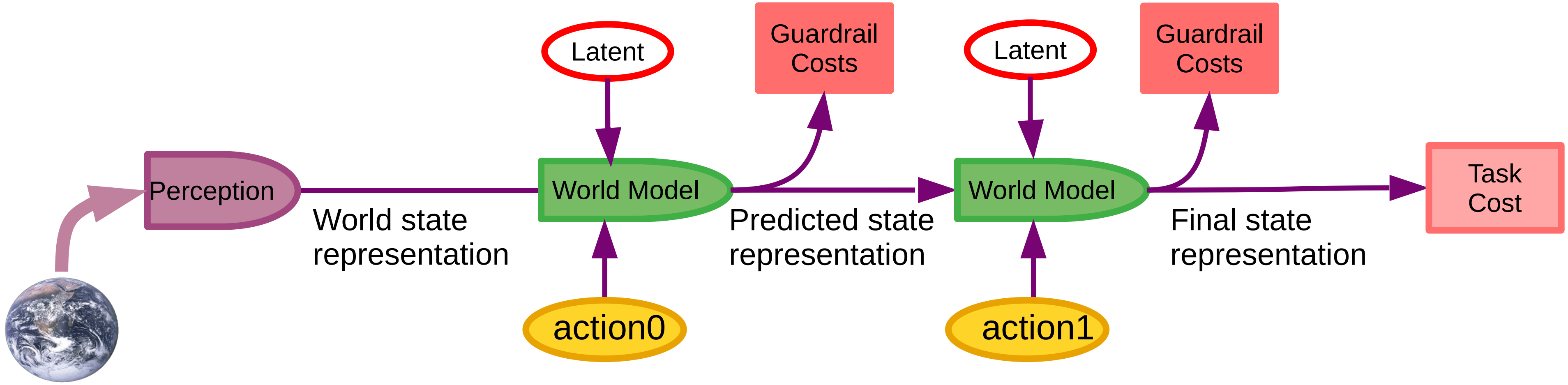
Objective-Driven AI: Multistep/Recurrent World Model

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



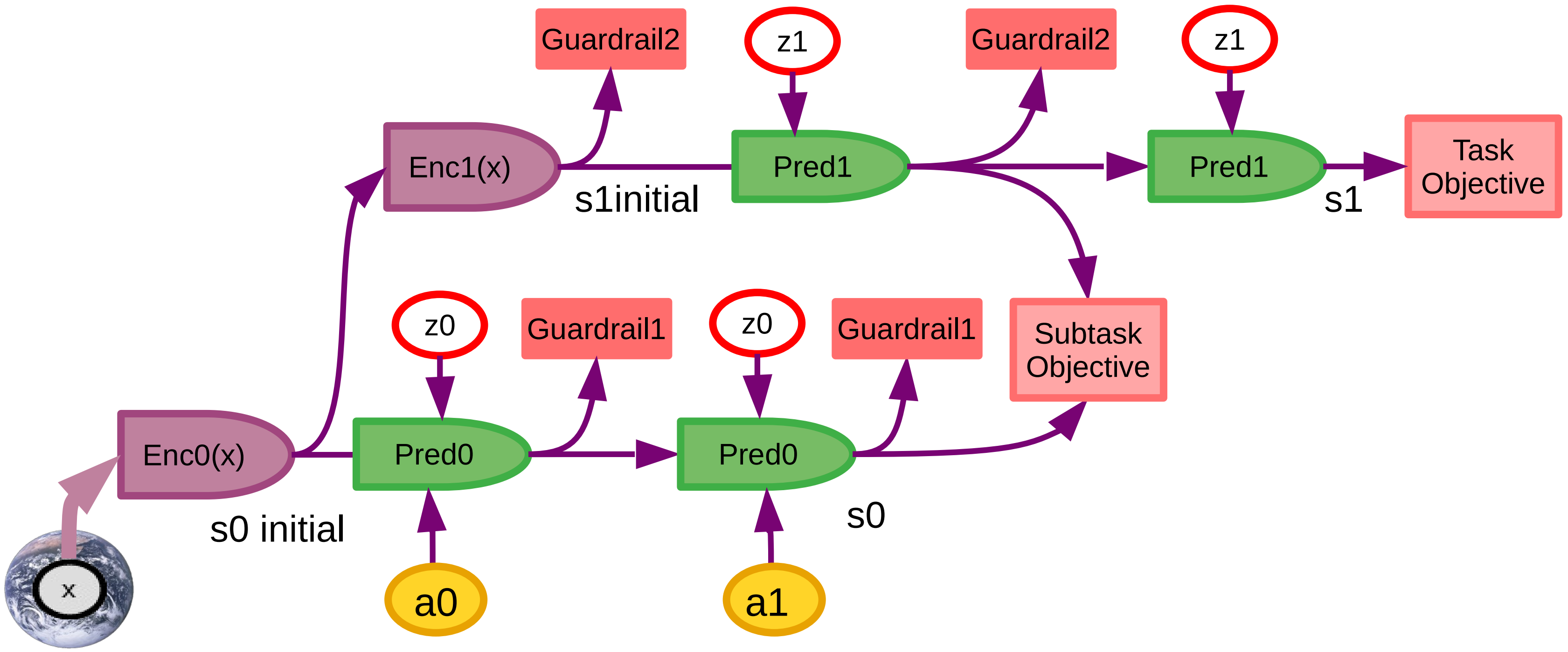
Objective-Driven AI: Non-Deterministic World Model

- **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



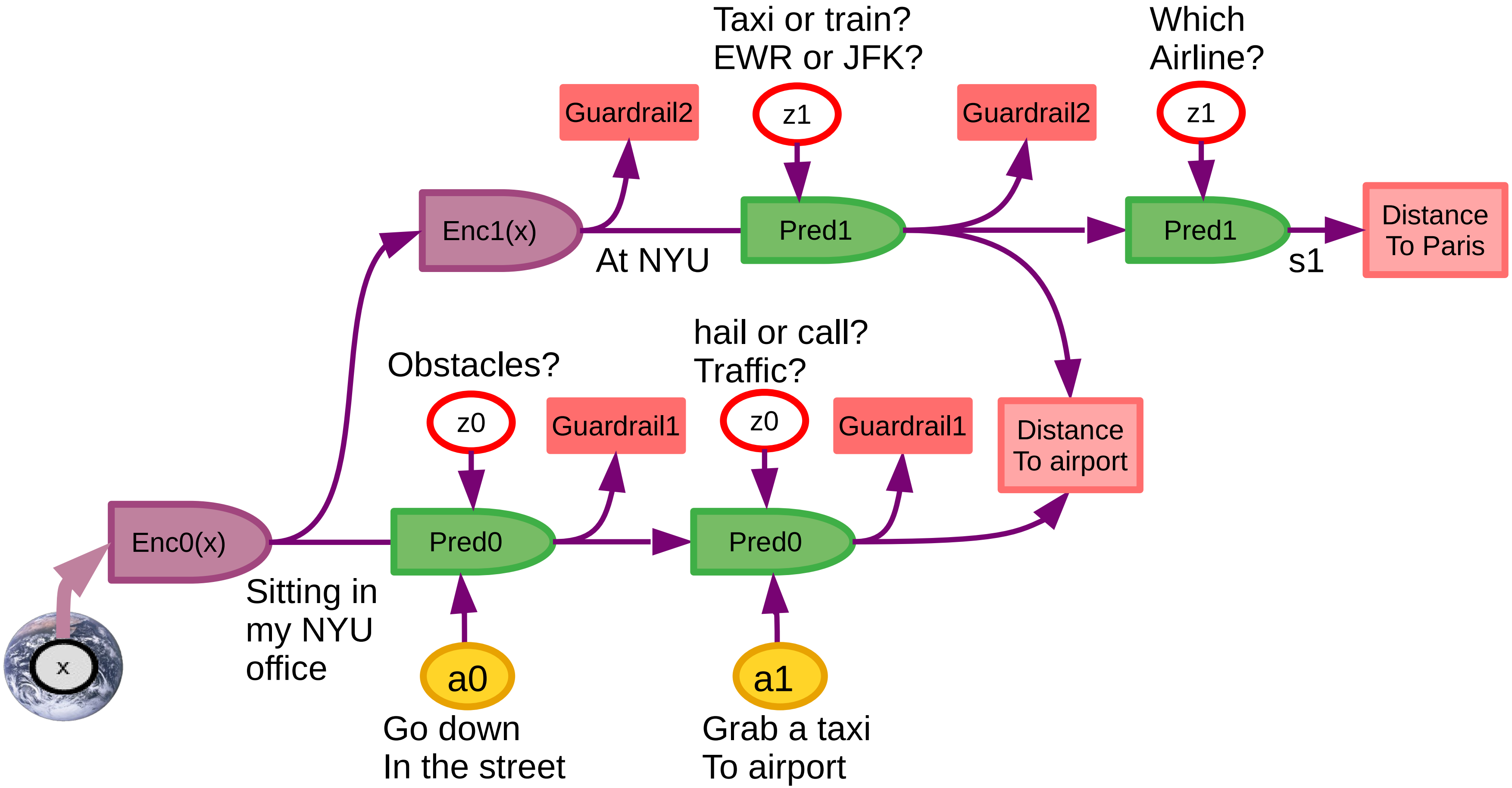
Objective-Driven AI: Hierarchical Planning

- **Hierarchical World Model and Planning**
  - Higher levels make longer-term predictions in more abstract representations
  - Predicted states at higher levels define subtask objectives for lower level
  - Guardrail objectives ensure safety at every level



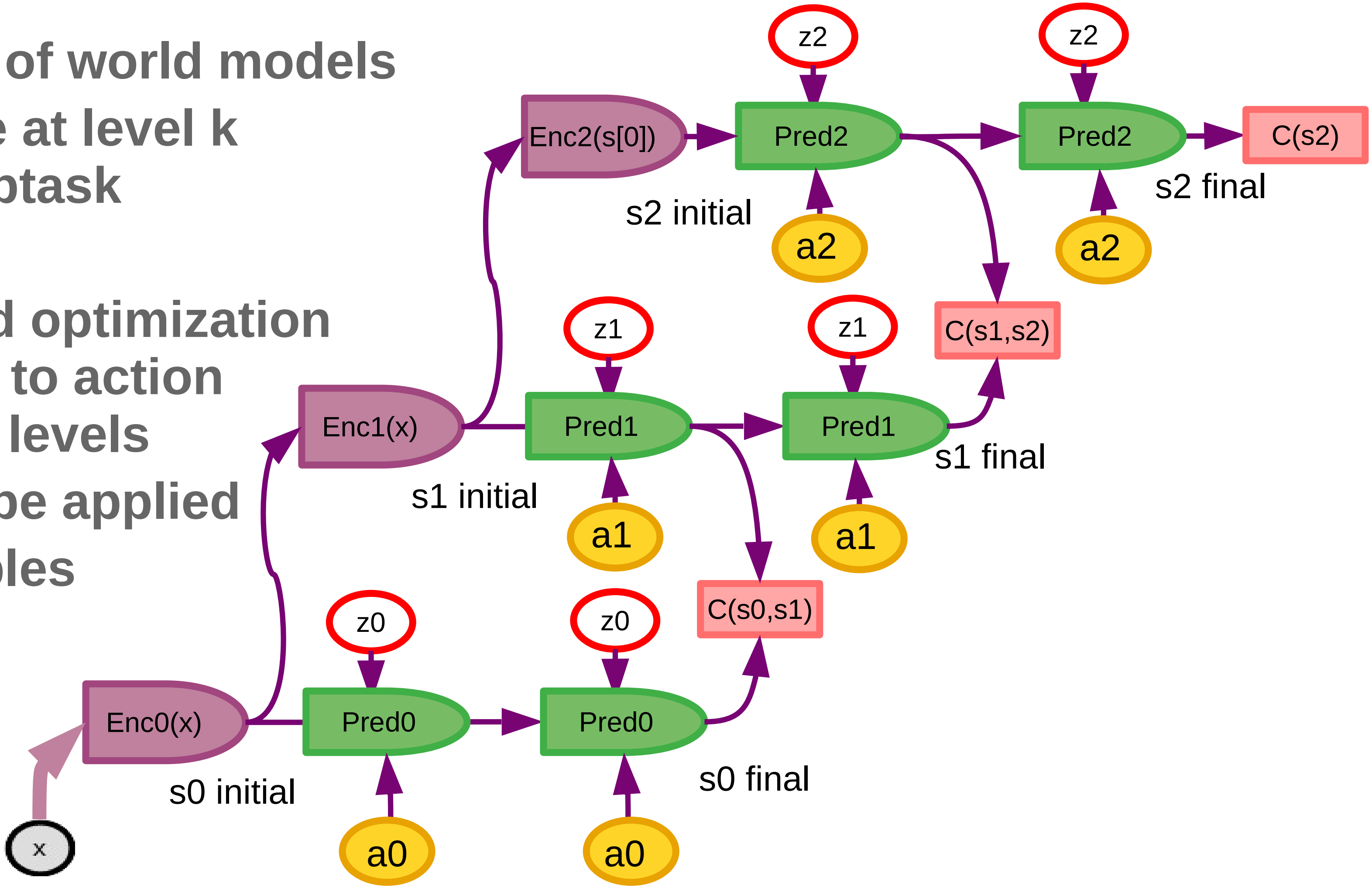
Objective-Driven AI: Hierarchical Planning

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



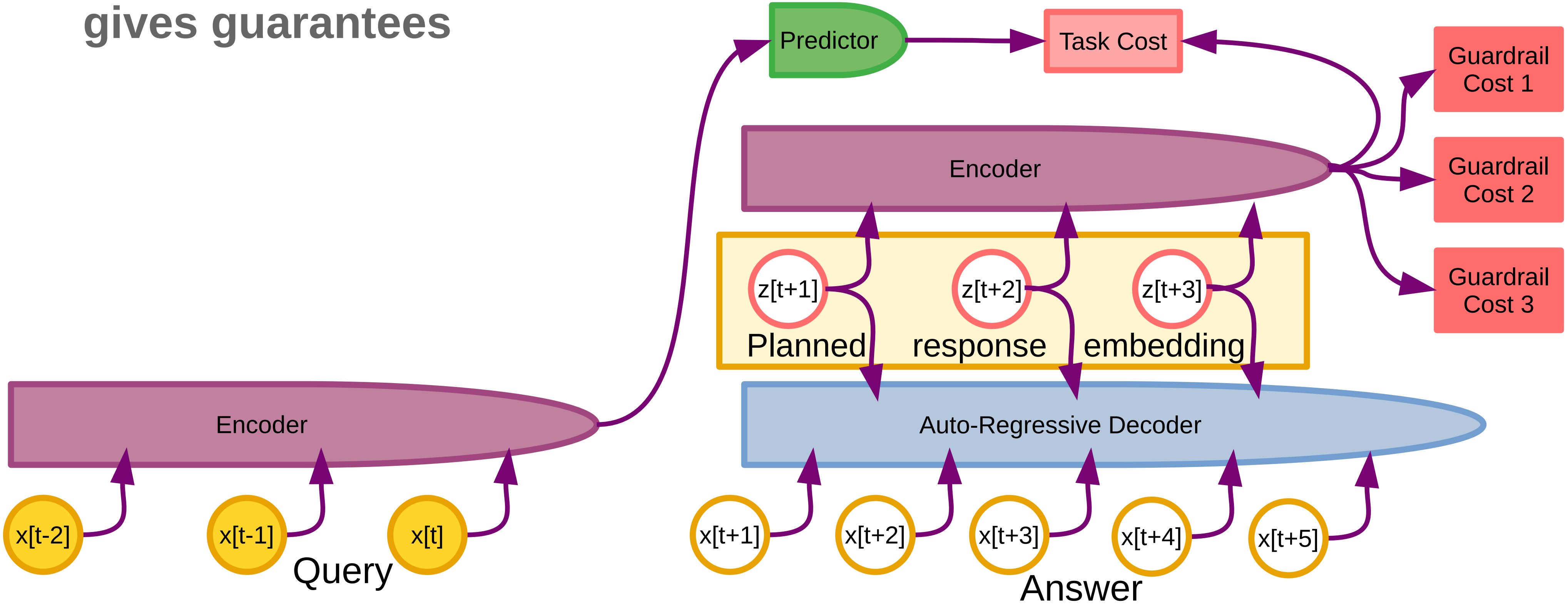
Objective-Driven AI: Hierarchical Planning

- **Multiple levels of world models**
- **Predicted state at level k determines subtask for level k-1**
- **Gradient-based optimization can be applied to action variables at all levels**
- **Sampling can be applied to latent variables at all levels.**



Objective-Driven AI for Dialog Systems

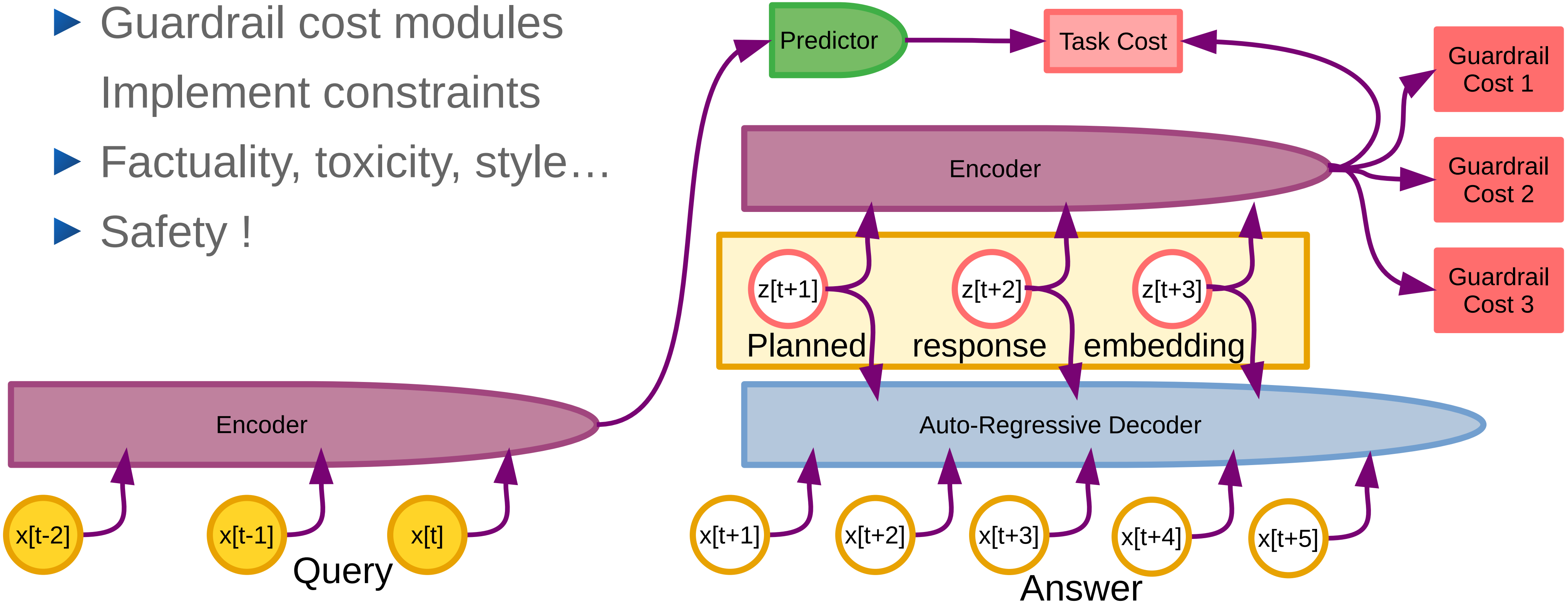
- Embedding of answer is planned by gradient-based optimization
- Planned embedding is converted to output by AR-LLM decoder
- Costs minimization gives guarantees



Objective-Driven AI for Dialog Systems

- **No need for RLHF fine-tuning**
- **Safe & Steerable**
  - Guardrail cost modules

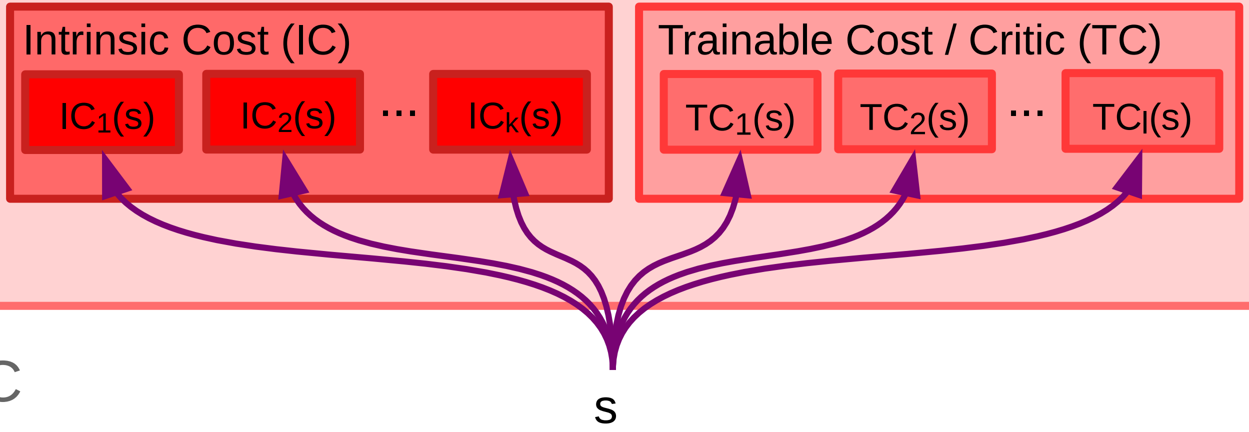
    Implement constraints
  - Factuality, toxicity, style…
  - Safety !



Cost Modules

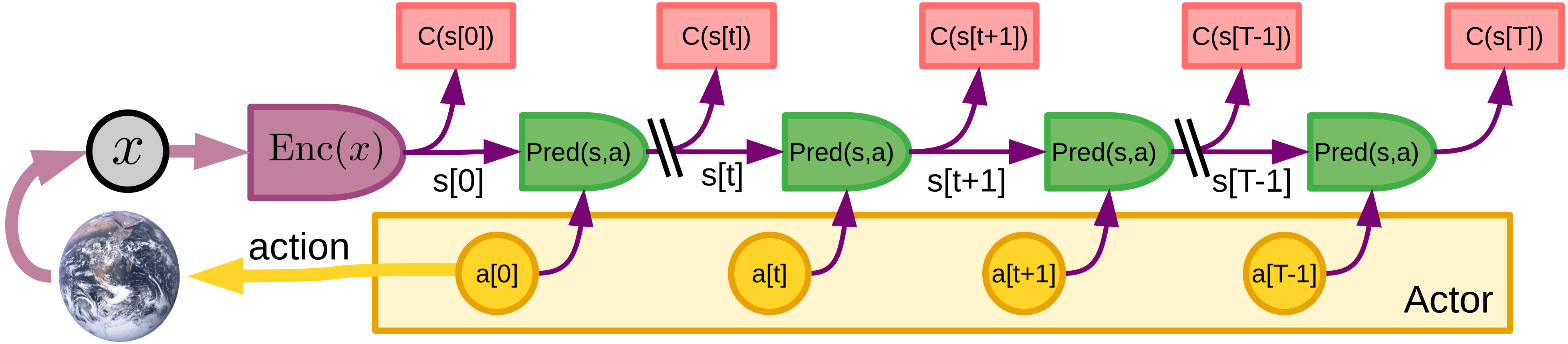
- **Intrinsic Cost (IC)**
  - Immutable cost modules.
  - Hard-wired drives.
- **Trainable Cost (TC)**
  - Trainable
  - Predicts future values of IC
  - Equivalent to a critic in RL
  - Implements subgoals
  - Configurable
- **All are differentiable**

$$\mathrm{C}(s) = \mathrm{IC}(s) + \mathrm{ITC}(s)\ ;\ \ \mathrm{IC}(s) = \sum_{i=1}^{k} u_i \mathrm{IC}_i(s)\ ;\ \ \mathrm{TC}(s) = \sum_{j=1}^{l} v_j \mathrm{TC}_j(s)$$



Mode-2 Perception-Planning-Action Cycle

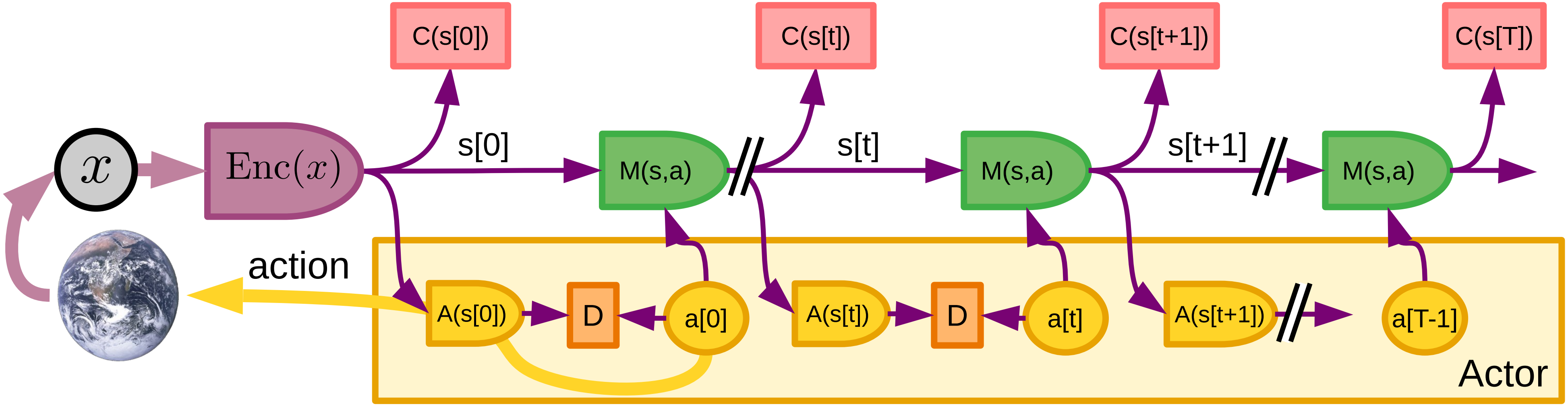
- **Akin to classical Model-Predictive Control (MPC)**
- **Actor proposes an ation sequence**
- **World Model predicts outcome**
- **Actor optimizes action sequence to minimize cost**
  - e.g. using gradient descent, dynamic programming, MC tree search…
- **Actor sends first action(s) to effectors**



[Henaff et al ICLR 19],[Hafner et al. ICML 19],[Chaplot et al. ICML 21],[Escontrela CoRL 22],...

Compiling Mode-2 into Mode-1

- Akin to Amortized Inference
- System performs Mode-2 cycle to get optimal action sequence.
- Optimal actions used as targets to train the policy module A(s)
- Policy module can be used for Mode-1 or to initialize Mode-2.

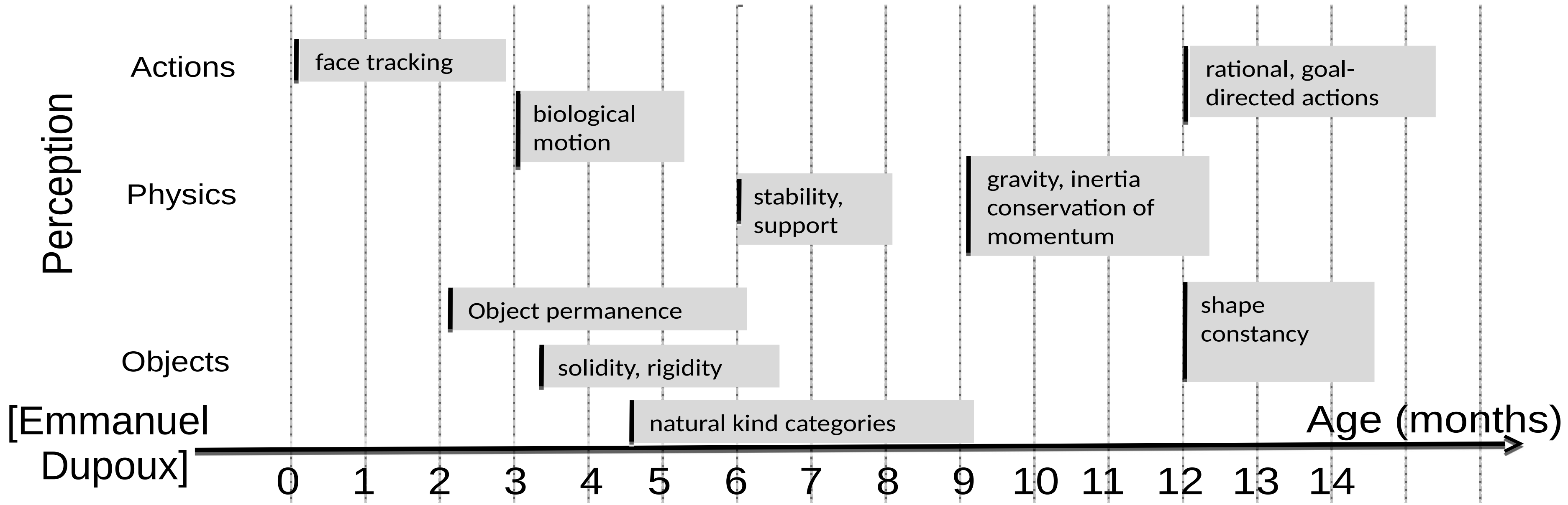


[Henaff et al. ICLR 2019] [Schrittwieser et al. MuZero 2020]

Building & Training the World Model

Joint-Embedding Architecture

How could machines learn like animals and humans?



- ▶ **How do babies learn how the world works?**

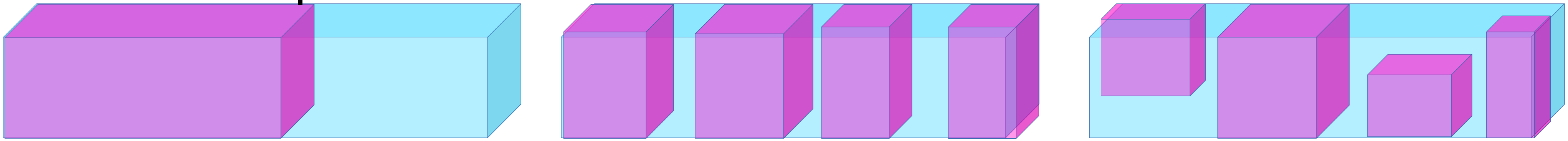
We are missing something big!

- **Cats and Dogs can do amazing feats**
  - Robots intelligence doesn’t come anywhere close
- **Any 10 year-old can learn to clear up the dinner table and fill up the dishwasher in minutes.**
  - We do not have robots that can do that.
- **Any 17 year-old can learn to drive a car in 20 hours of practice**
  - We still don’t have unlimited Level-5 autonomous driving
- **Obviously, we are missing something big!**
- **We keep bumping into Moravec’s paradox**
  - Things that are easy for humans are difficult for AI and vice versa.

Self-Supervised Learning = Learning to Fill in the Blanks

- **Reconstruct the input or Predict missing parts of the input.**

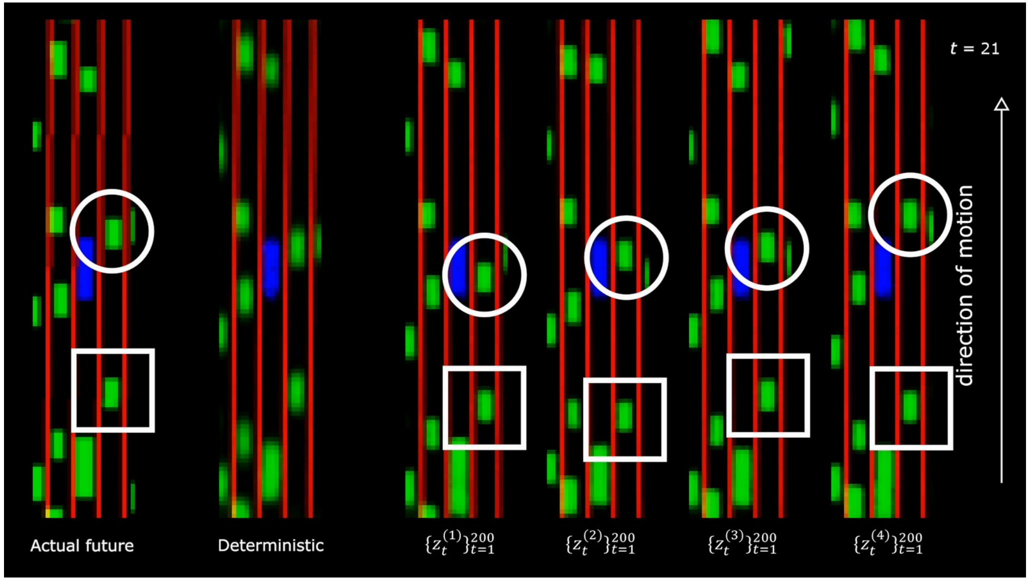
time or space →



How do we represent uncertainty in the predictions?

- The world is only partially predictable
- How can a predictive model represent multiple predictions?
- Probabilistic models are intractable in continuous domains.
- Generative Models must predict every detail of the world
- **My solution: Joint-Embedding Predictive Architecture**

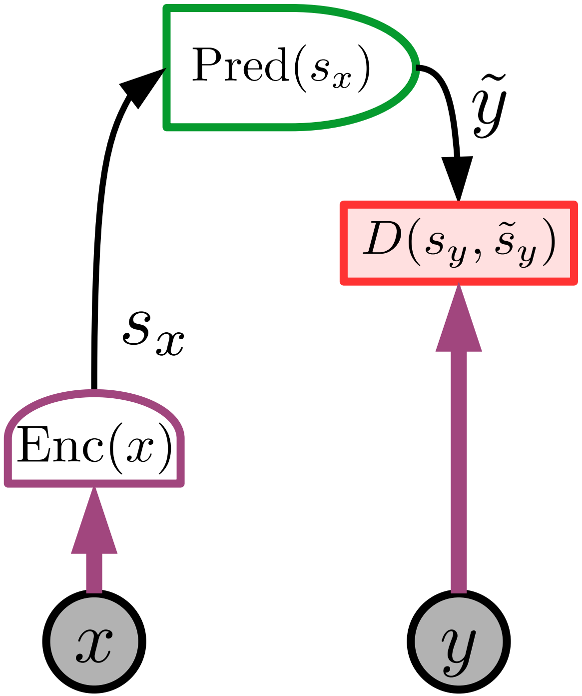
[Mathieu, Couprie, LeCun ICLR 2016]



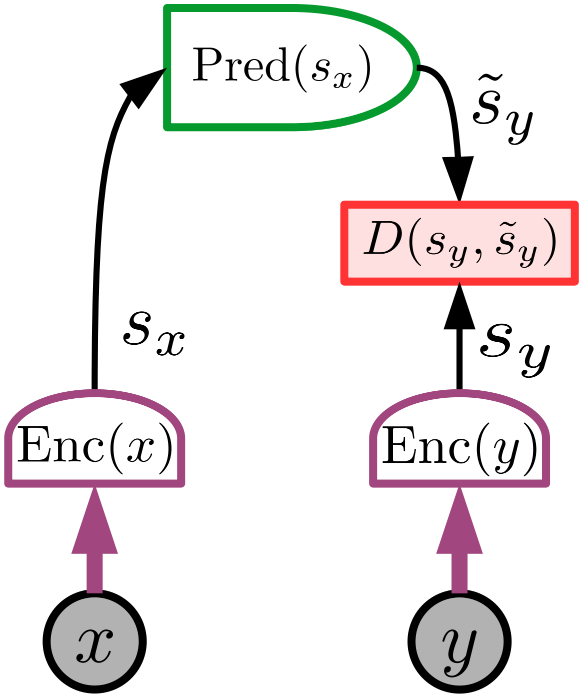
[Henaff, Canziani, LeCun ICLR 2019]

Architectures: Generative vs Joint Embedding

- **Generative: predicts y** (with all the details, including irrelevant ones)
- **Joint Embedding: predicts an abstract representation of y**



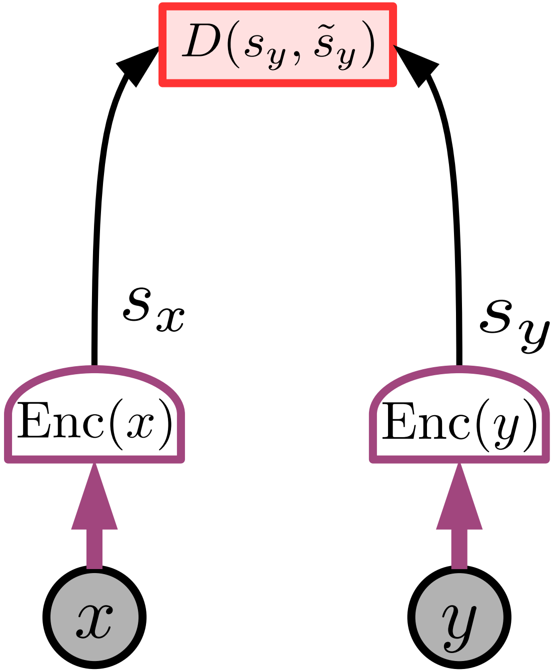
a) Generative Architecture
Examples: VAE, MAE...



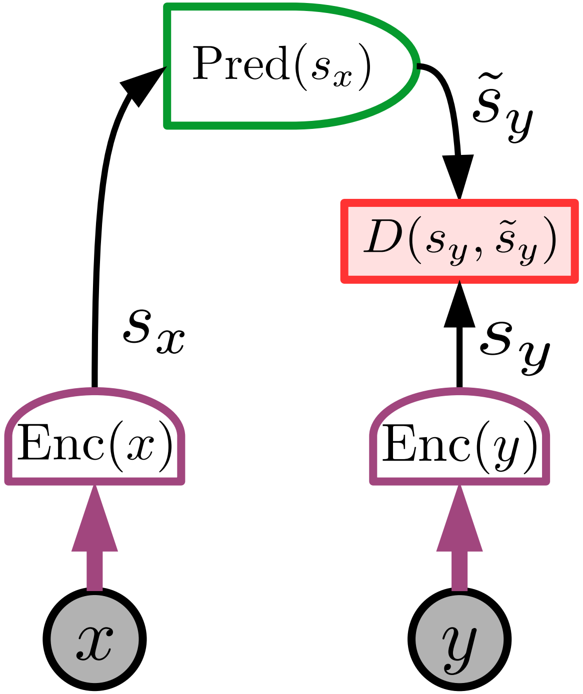
b) Joint Embedding Architecture

Joint Embedding Architectures

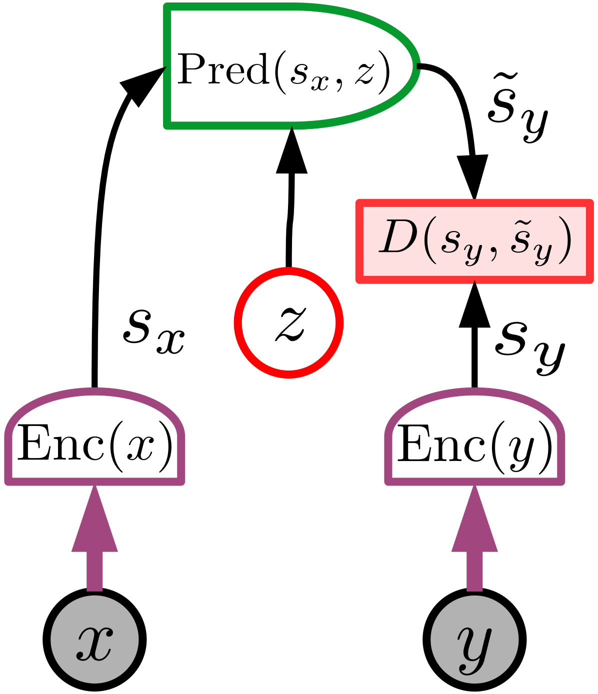
- Computes abstract representations for x and y
- Tries to make them equal or predictable from each other.



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



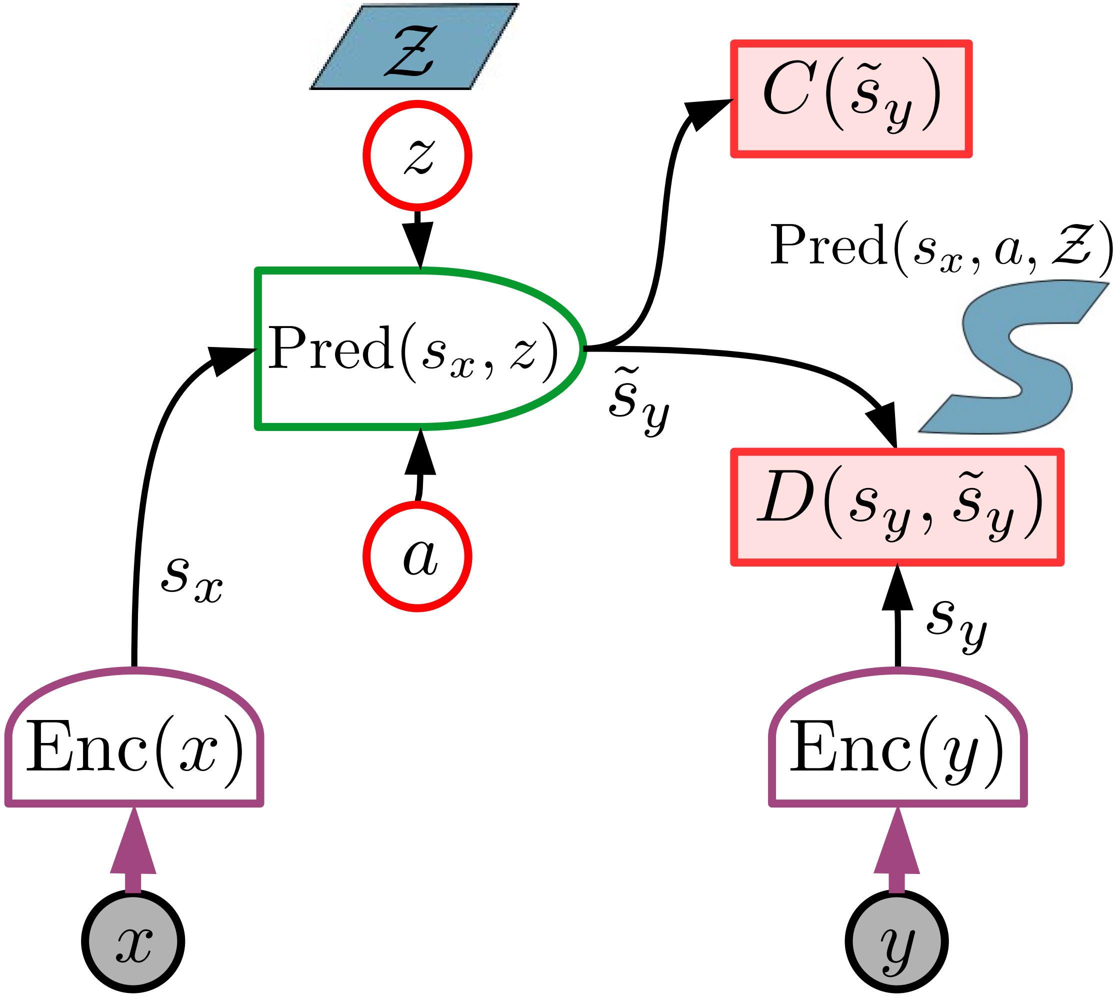
b) Deterministic Joint Embedding Predictive Architecture (DJEPA)
Examples: BYOL, VICRegL, I-JEPA



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

Architecture for 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 Sy from a representation of the past and present Sx

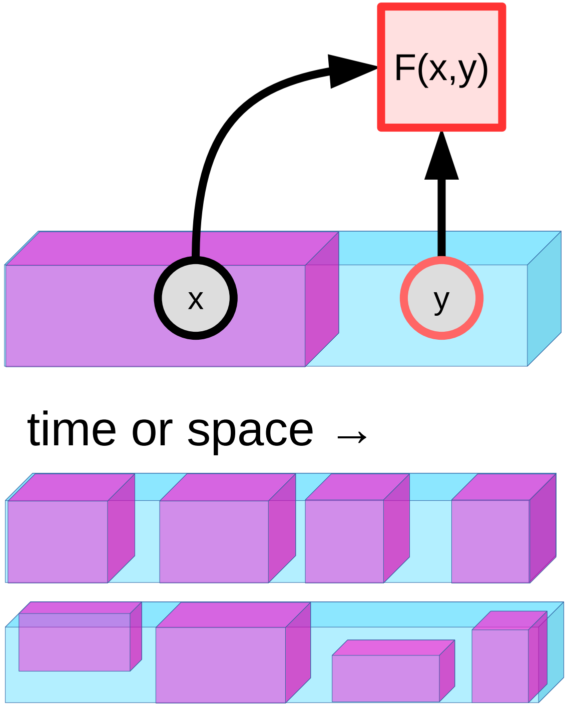


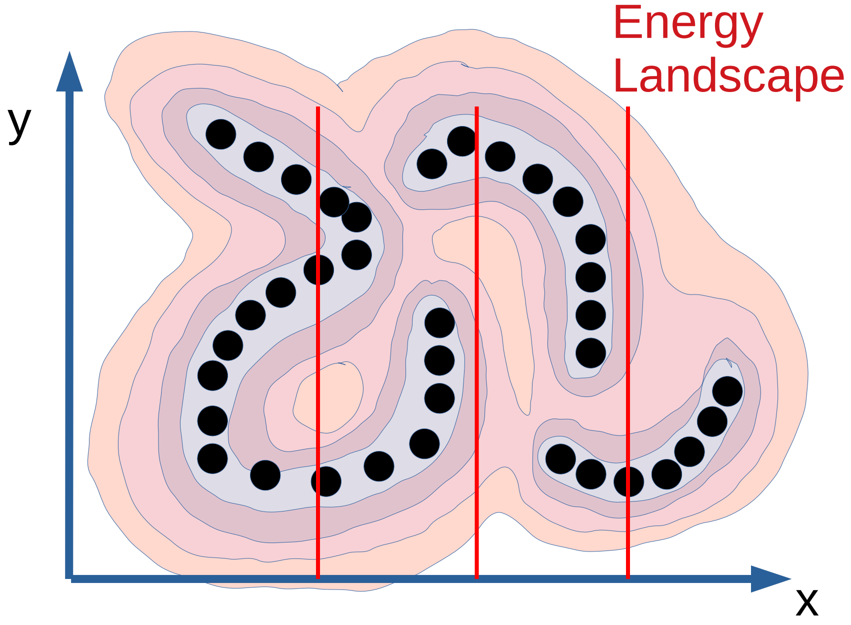
Energy-Based Models

Capturing dependencies through an energy function

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



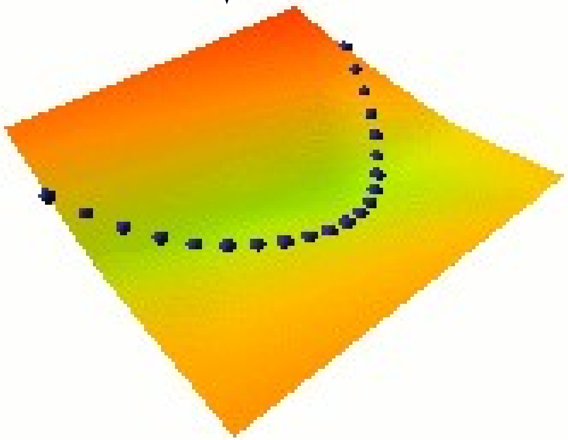


$$\check{y} = \mathrm{argmin}_y F(x, y)$$

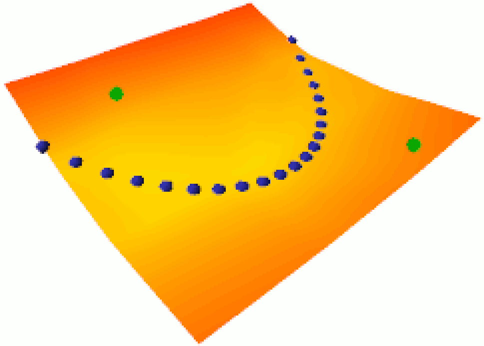
Training Energy-Based Models: Collapse Prevention

- **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.

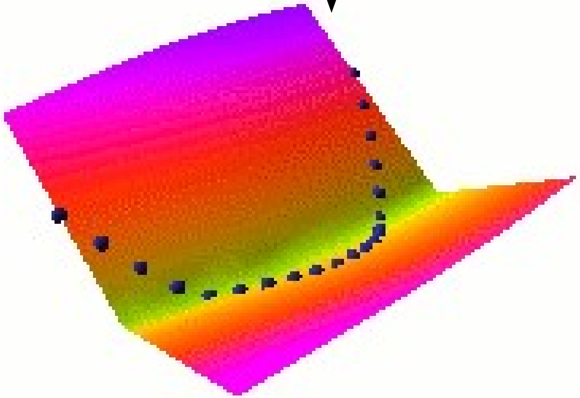
Collapse!



Contrastive Method

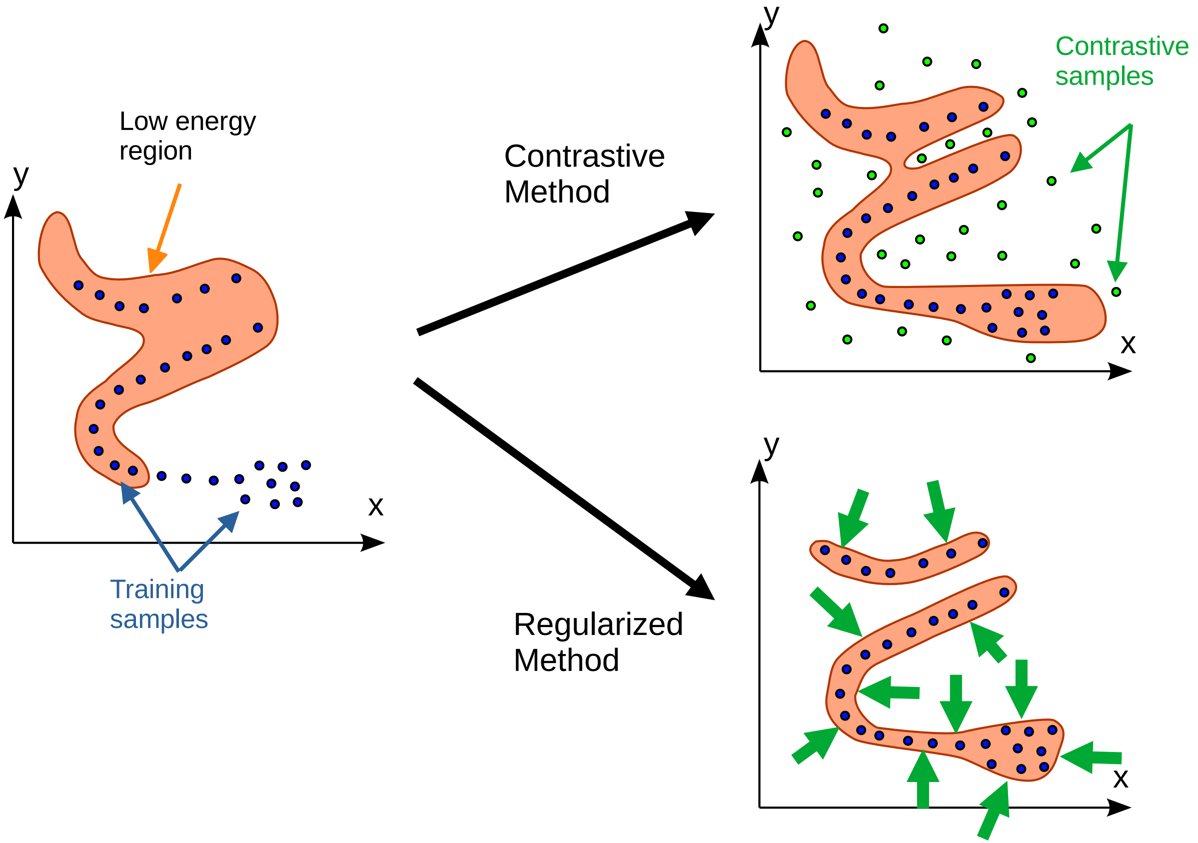


Regularized Methods



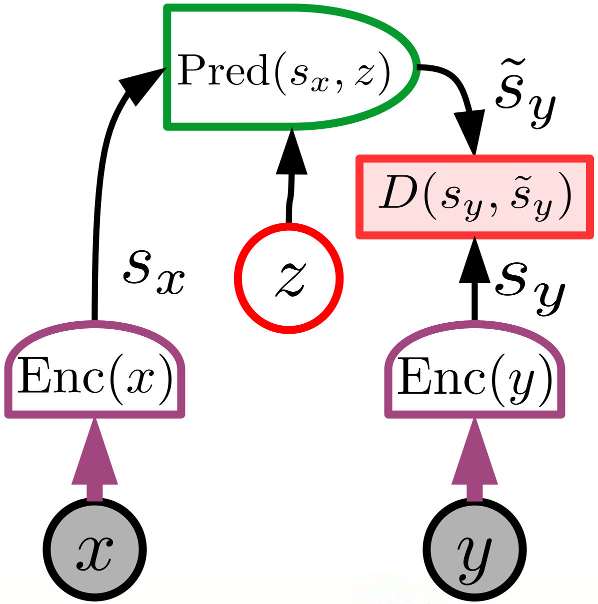
EBM Training: two categories of methods

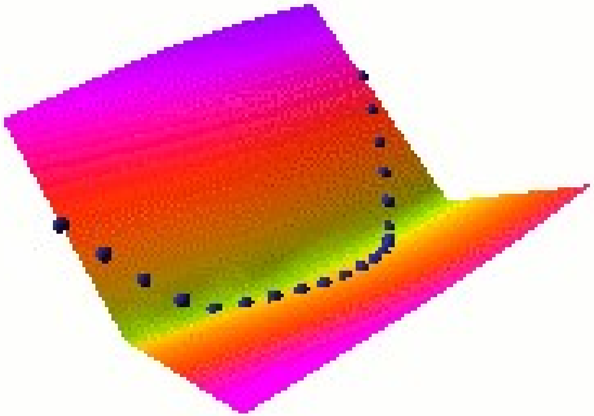
- **Contrastive methods**
  - Push down on energy of training samples
  - Pull up on energy of suitably-generated contrastive samples
  - Scales very badly with dimension
- **Regularized Methods**
  - Regularizer minimizes the volume of space that can take low energy



Recommendations:

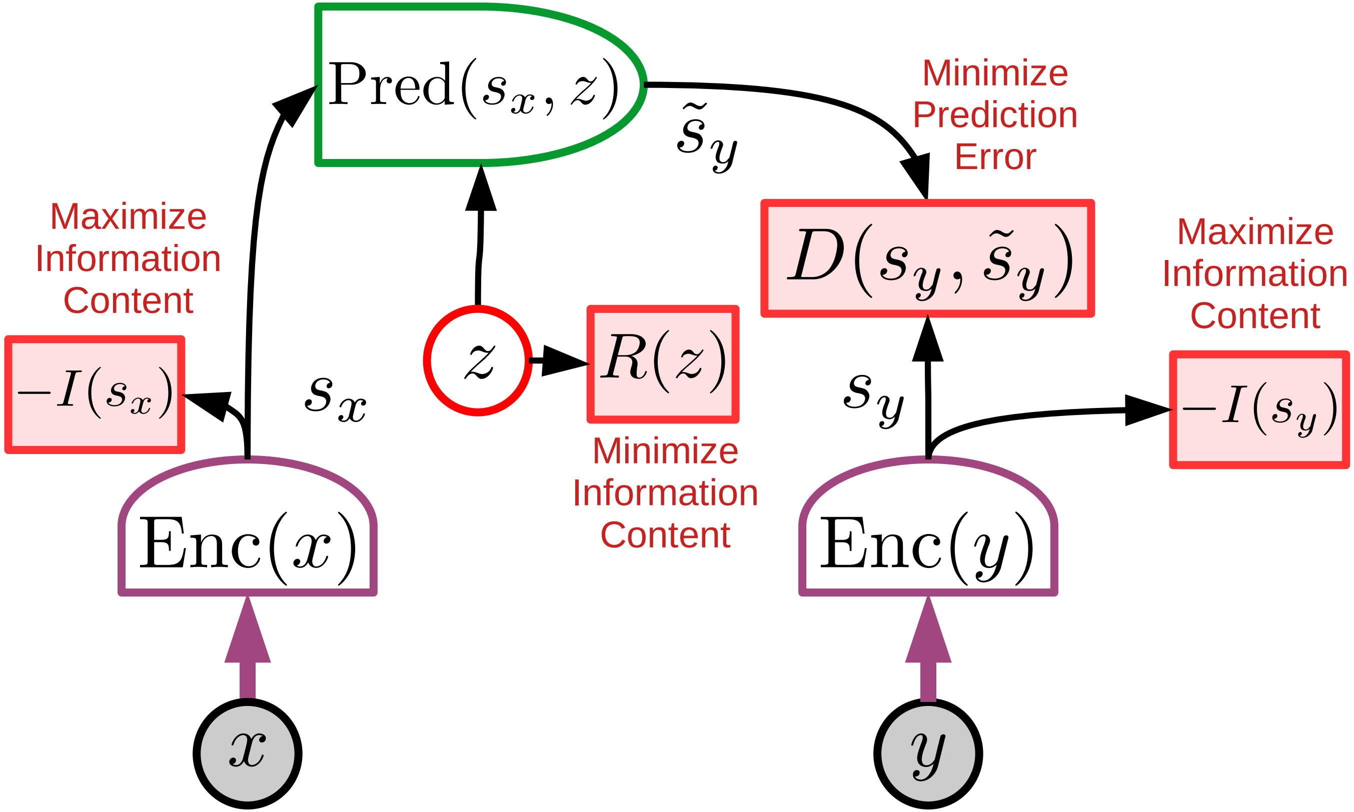
- **Abandon generative models**
  - in favor joint-embedding architectures
- **Abandon probabilistic model**
  - in favor of energy-based models
- **Abandon contrastive methods**
  - in favor of regularized methods
- **Abandon Reinforcement Learning**
  - In favor of model-predictive control
  - Use RL only when planning doesn't yield the predicted outcome, to adjust the world model or the critic.





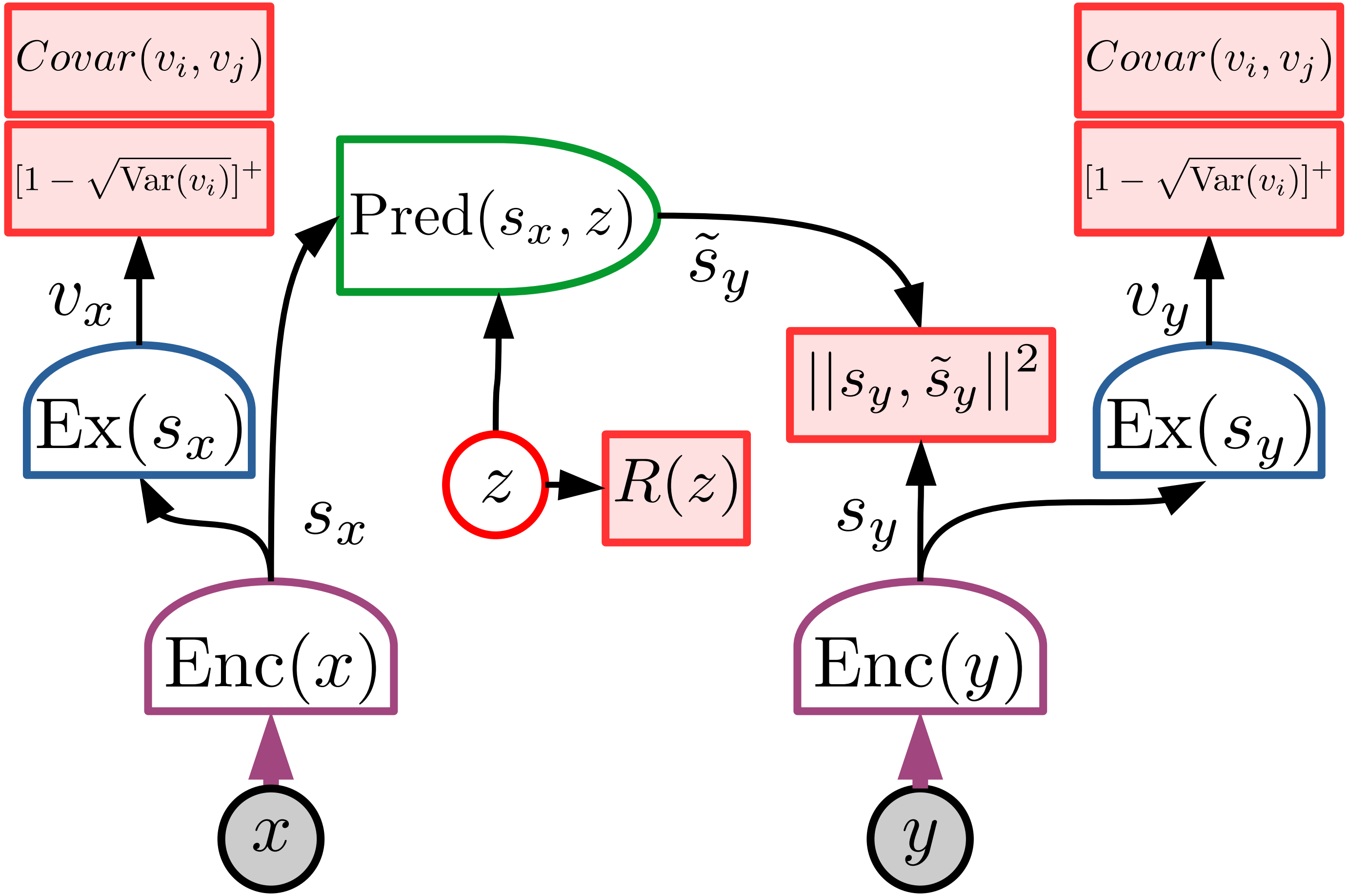
Training a JEPA with Regularized Methods

- **Four terms in the cost**
  - Maximize information content in representation of x
  - Maximize information content in representation of y
  - Minimize Prediction error
  - Minimize information content of latent variable z



VICReg: Variance, Invariance, Covariance Regularization

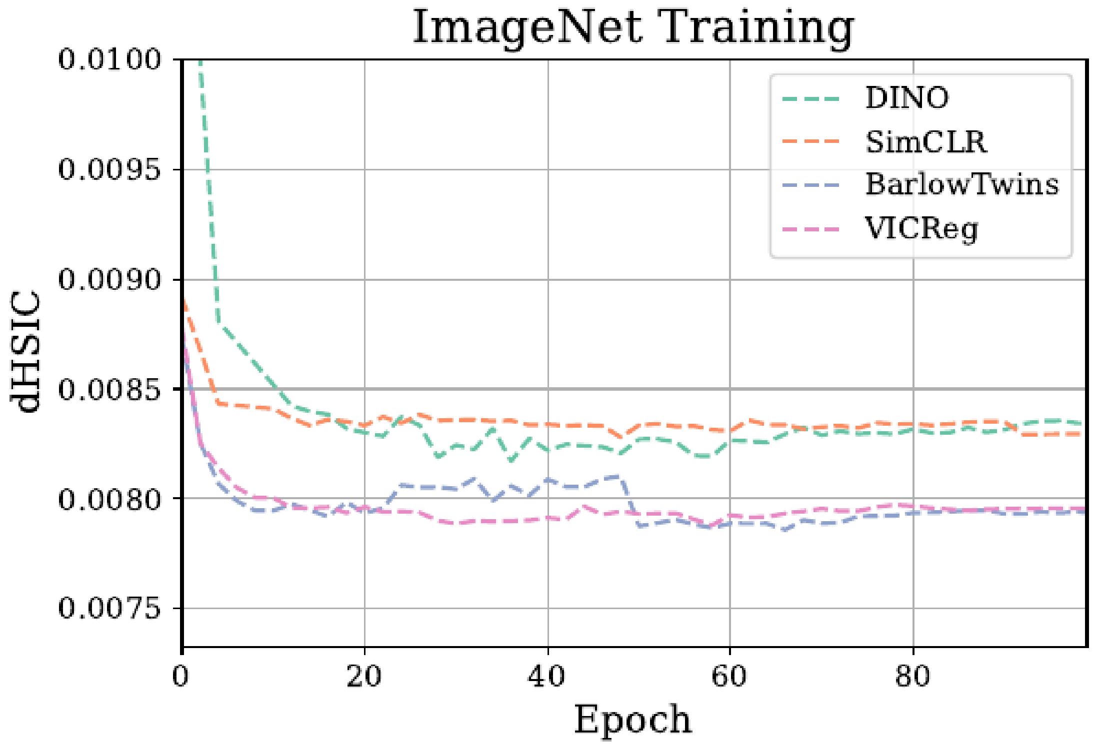
- **Variance:**
  - Maintains variance of components of representations
- **Covariance:**
  - Decorrelates components of covariance matrix of representations
- **Invariance:**
  - Minimizes prediction error.

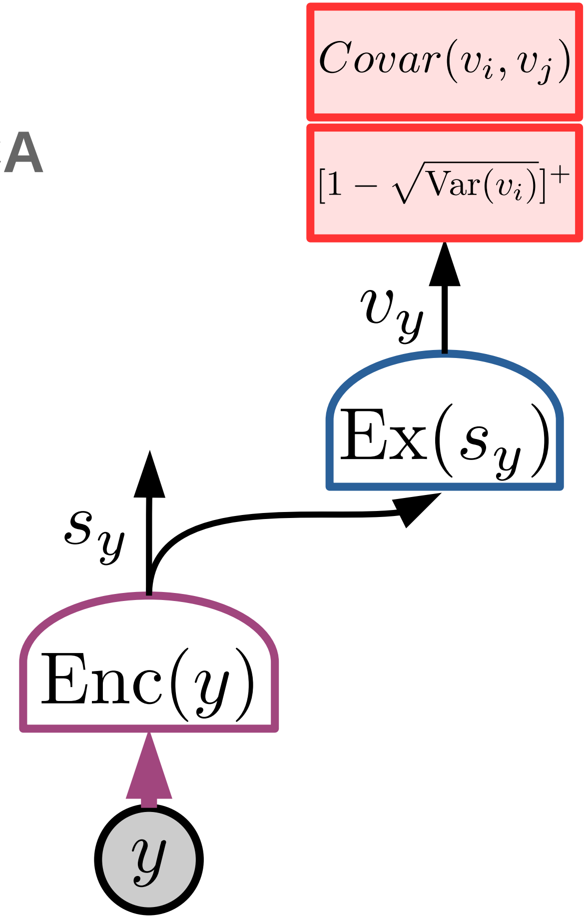


Barlow Twins [Zbontar et al. ArXiv:2103.03230], VICReg [Bardes, Ponce, LeCun arXiv:2105.04906, ICLR 2022], VICRegL [Bardes et al. NeurIPS 2022], MCR2 [Yu et al. NeurIPS 2020][Ma, Tsao, Shum, 2022]

VICReg: expander makes variables pairwise independent

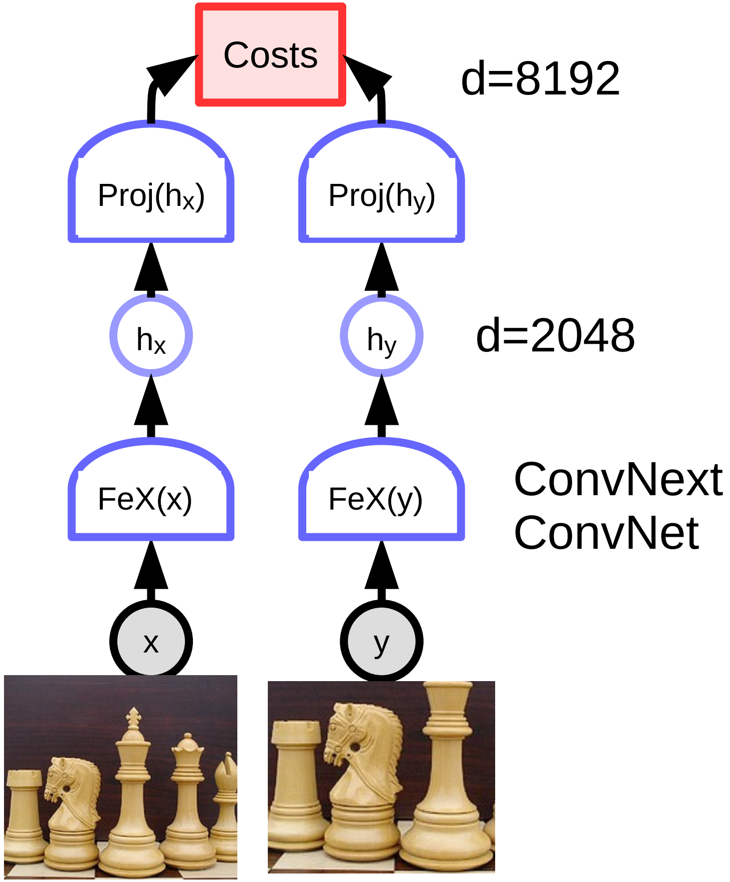
- [Mialon, Balestriero, LeCun arxiv:2209.14905]
- VC criterion can be used for source separation / ICA



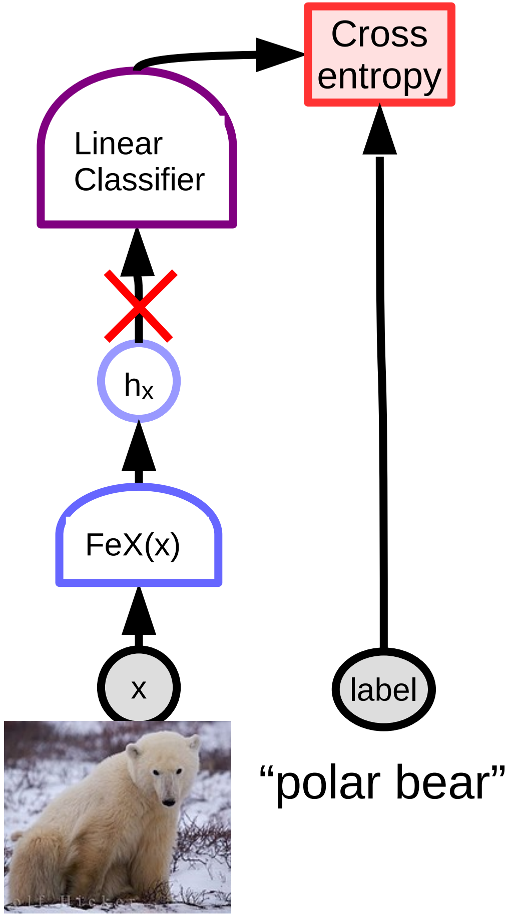


SSL-Pretrained Joint Embedding for Image Recognition

JEA pretrained with VICReg



Training a supervised linear head

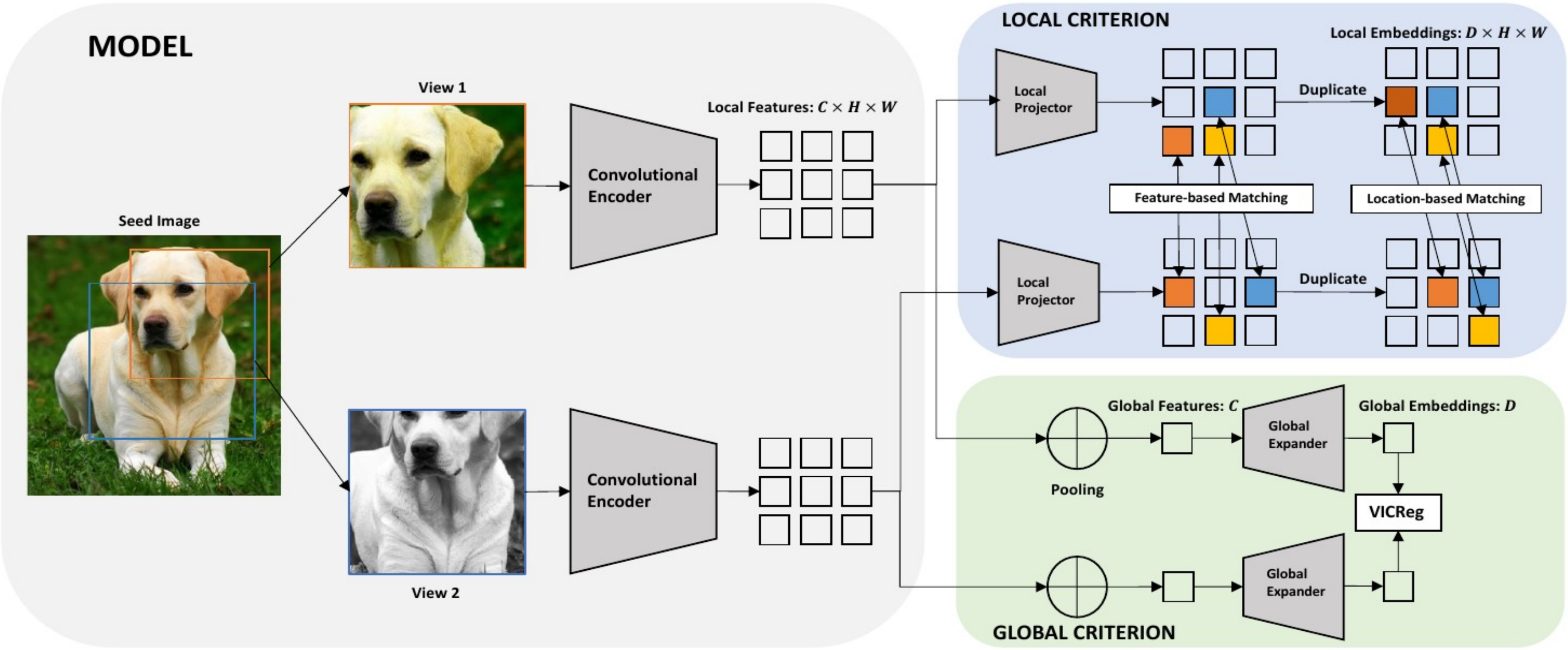


VICReg: Results with linear head and semi-supervised.

Method	Linear Top-1	Linear Top-5	Semi-supervised Top-1 1%	Semi-supervised Top-1 10%	Semi-supervised Top-5 1%	Semi-supervised Top-5 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	<u>91.1</u>	-	-	-	-
OBoW Gidaris et al. (2021)	<u>73.8</u>	-	-	-	<u>82.9</u>	<u>90.7</u>
BYOL Grill et al. (2020)	<u>74.3</u>	<u>91.6</u>	53.2	68.8	78.4	89.0
SwAV (w/ multi-crop) Caron et al. (2020)	<u>75.3</u>	-	<u>53.9</u>	<u>70.2</u>	78.5	<u>89.9</u>
Barlow Twins Zbontar et al. (2021)	73.2	91.0	<u>55.0</u>	<u>69.7</u>	<u>79.2</u>	89.3
VICReg (ours)	73.2	<u>91.1</u>	<u>54.8</u>	<u>69.5</u>	<u>79.4</u>	<u>89.5</u>

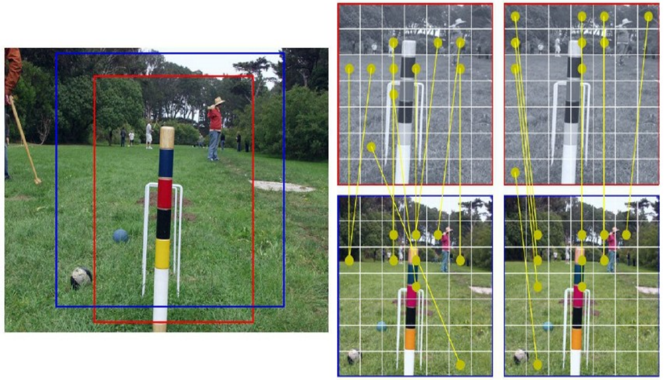
VICRegL: local matching latent variable for segmentation

- **Latent variable optimization:**
  - Finds a pairing between local feature vectors of the two images
  - [Bardes, Ponce, LeCun, NeurIPS 2022, arXiv:2210.01571]



VICRegL: local matching latent variable for segmentation

		Linear Cls. (%)	Linear Seg. (mIoU)		
		ImageNet	Pascal VOC		Cityscapes
Method	Epochs	Frozen	Frozen	Fine-Tuned	Frozen
Global features					
MoCo v2 [Chen et al., 2020b]	200	67.5	35.6	64.8	14.3
SimCLR [Chen et al., 2020a]	400	68.2	45.9	65.4	17.9
BYOL [Grill et al., 2020]	300	**72.3**	47.1	65.7	22.6
VICReg [Bardes et al., 2022]	300	71.5	47.8	65.5	23.5
Local features					
PixPro [Xie et al., 2021]	400	60.6	52.8	67.5	22.6
DenseCL [Wang et al., 2021]	200	65.0	45.3	66.8	11.2
DetCon [Hénaff et al., 2021]	1000	66.3	53.6	67.4	16.2
InsLoc [Yang et al., 2022]	400	45.0	24.1	64.4	7.0
CP^2 [Wang et al., 2022]	820	53.1	21.7	65.2	8.4
ReSim [Xiao et al., 2021]	400	59.5	51.9	67.3	12.3
Ours					
VICRegL $\alpha = 0.9$	300	71.2	54.0	66.6	25.1
VICRegL $\alpha = 0.75$	300	70.4	**55.9**	**67.6**	**25.2**



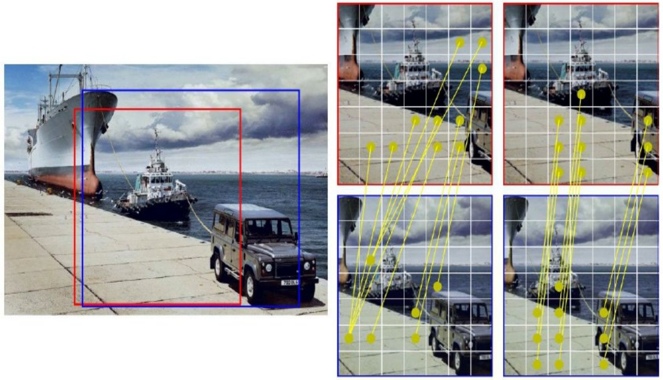
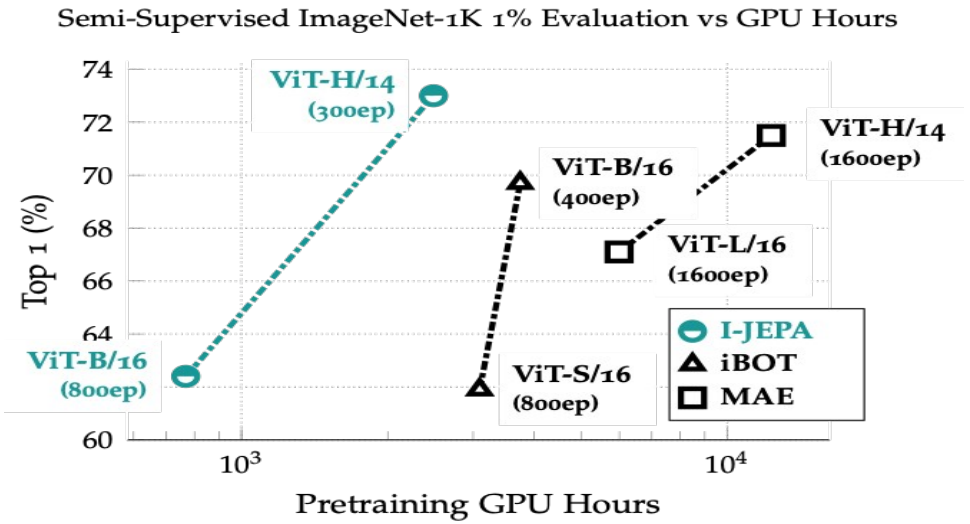
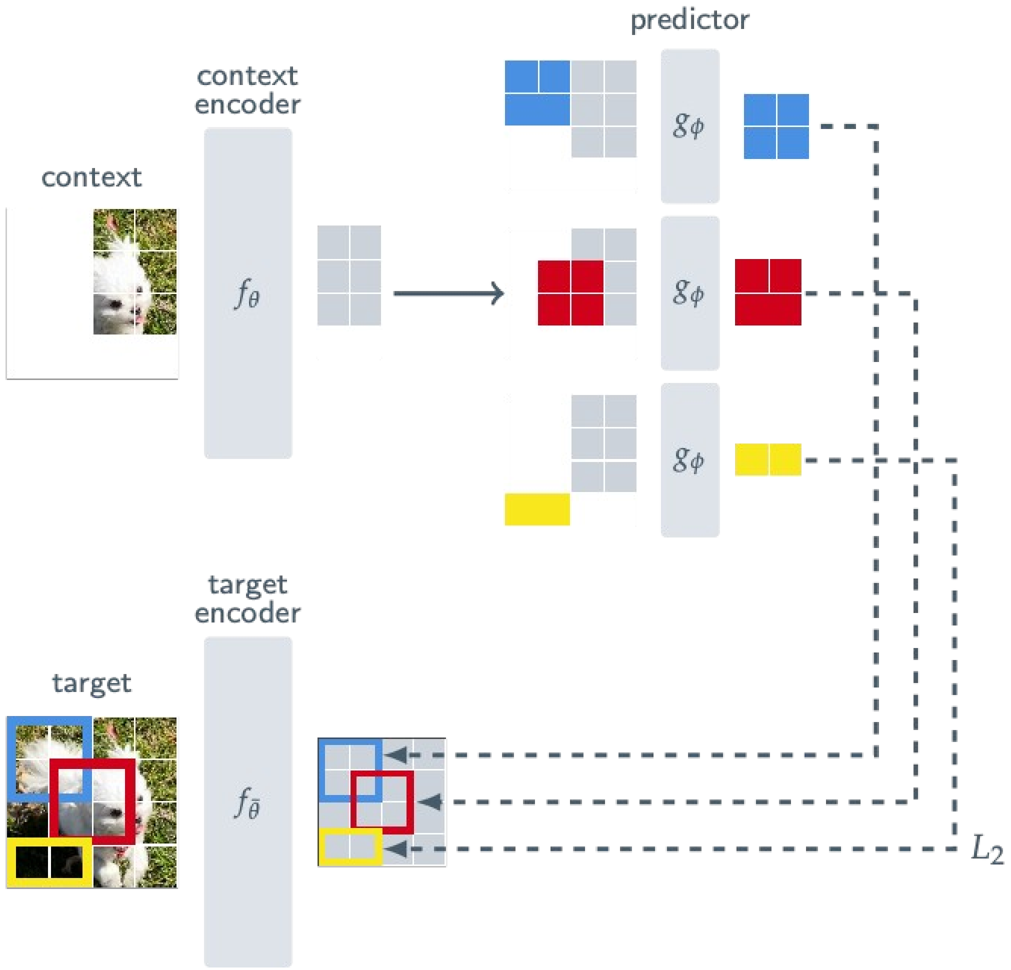


Image-JEPA: uses masking & transformer architectures

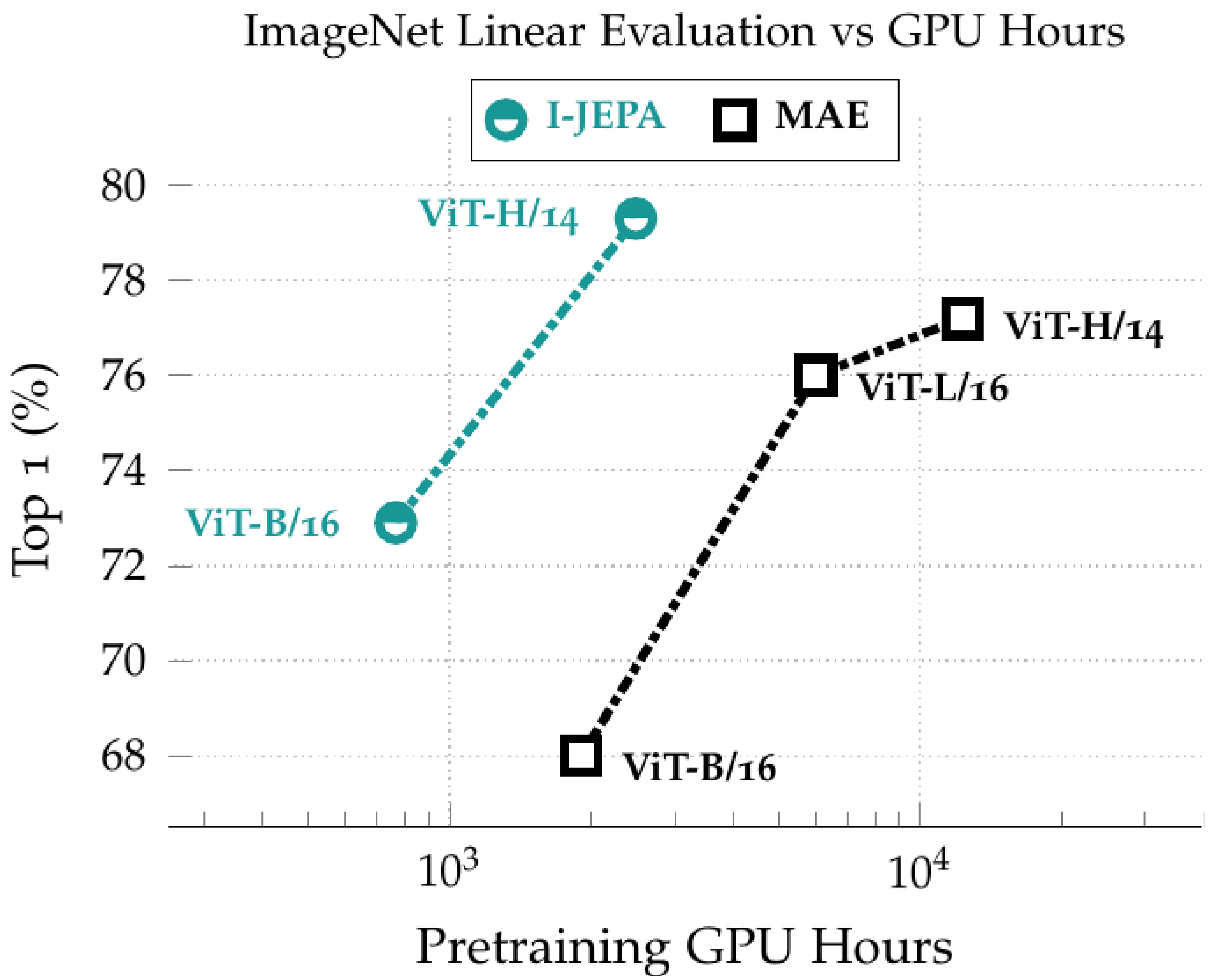
- **"SSL from images with a JEPA"**
  - [M. Assran et al arxiv:2301.08243]
- **Jointly embeds a context and a number of neighboring patches.**
  - Uses predictors
  - Uses only masking





I-JEPA Results

- **Training is fast**
- **Non-generative method seems to beat reconstruction-based methods (MAE)**



I-JEPA Results on ImageNet

- **JEPA better than generative architecture on pixels.**

Targets	Arch.	Epochs	Top-1
Target Encoder Output	ViT-L/16	500	66.0

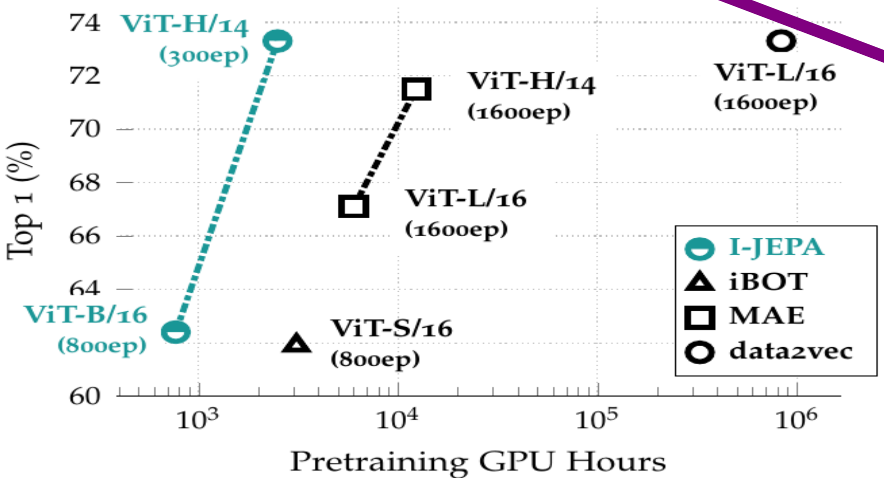
- **Closing the gap with methods that use data augments**
- **Methods with only masking**
  - No data augmentation
- **Methods with data augmentation**
  - Similar to SimCLR

Method	Arch.	Epochs	Top-1
Methods without view data augmentations			
data2vec [7]	ViT-L/16	1600	53.5
MAE [34]	ViT-B/16	1600	68.0
	ViT-L/16	1600	76.0
	ViT-H/14	1600	77.2
I-JEPA	ViT-B/16	600	72.9
	ViT-L/16	600	77.5
	ViT-H/14	300	79.3
	ViT-H/16_{448}	300	**81.1**
Methods using extra view data augmentations			
SimCLR v2 [20]	RN152 (2×)	800	79.1
DINO [17]	ViT-B/8	300	80.1
iBOT [74]	ViT-L/16	250	**81.0**

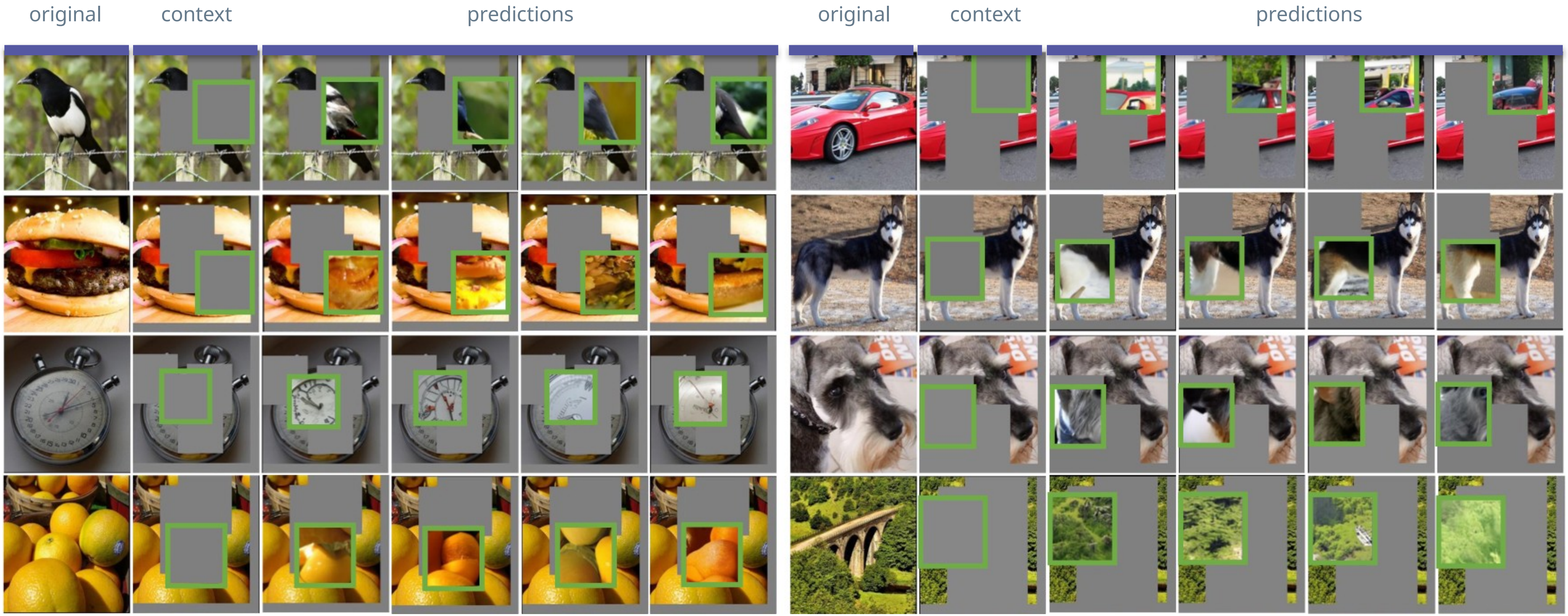
I-JEPA Results on ImageNet with 1% training

- **JEPA better than generative architecture on pixels.**
- **Closing the gap with methods that use data augments**
- **Methods with only masking**
- **Methods with data augmentation**

Method	Arch.	Epochs	Top-1
Methods without view data augmentations			
data2vec [7]	ViT-L/16	1600	73.3
MAE [34]	ViT-L/16	1600	67.1
	ViT-H/14	1600	71.5
I-JEPA	ViT-L/16	600	69.4
	ViT-H/14	300	73.3
	ViT-H/16_{448}	300	**77.3**
Methods using extra view data augmentations			
iBOT [74]	ViT-B/16	250	69.7
DINO [17]	ViT-B/8	300	70.0
SimCLR v2 [33]	RN151 (2×)	800	70.2
BYOL [33]	RN200 (2×)	800	71.2
MSN [3]	ViT-B/4	300	**75.7**



I-JEPA: Visualizing Predicted Representations

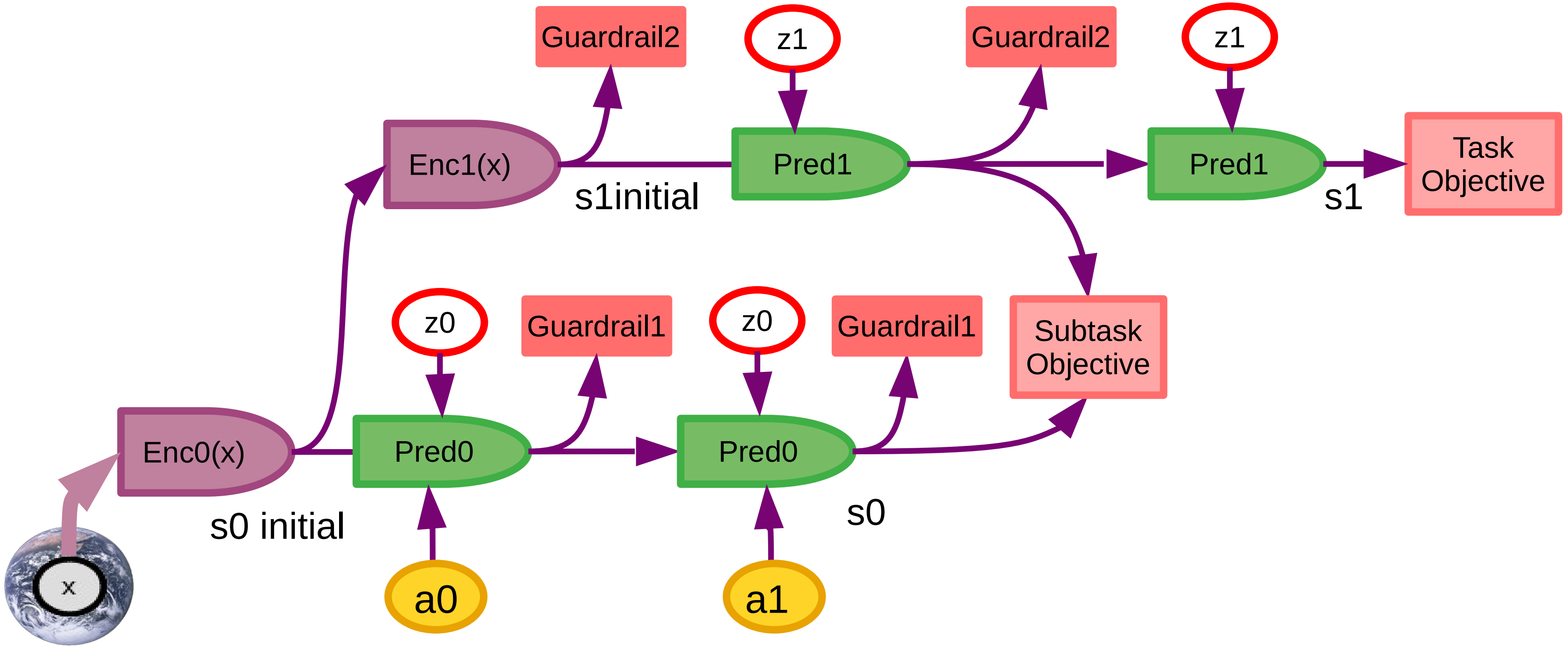


Hierarchical JEPA for Hierarchical Planning

Control, planning, and policy learning.

Objective-Driven AI: Hierarchical Planning

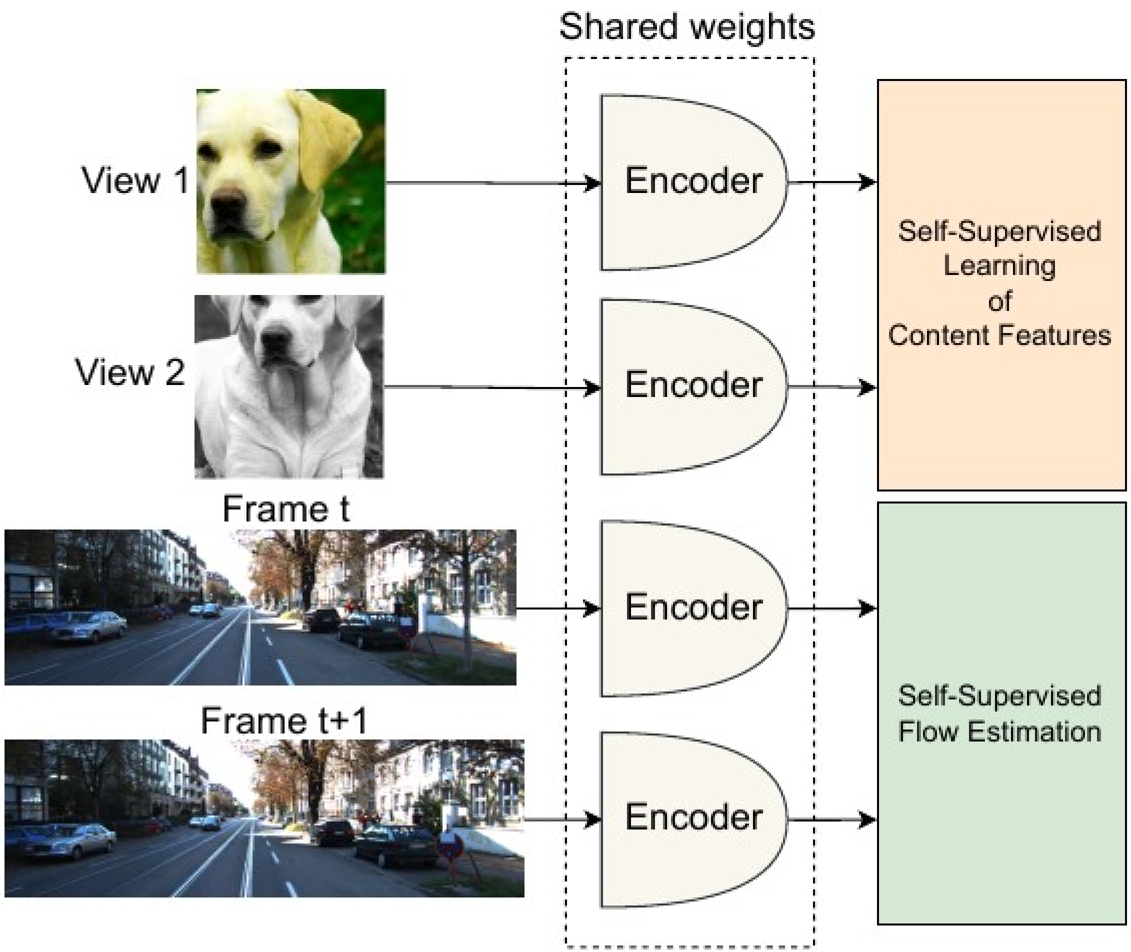
- **Hierarchical World Model and Planning**
  - Higher levels make longer-term predictions in more abstract representations
  - Predicted states at higher levels define subtask objectives for lower level
  - Guardrail objectives ensure safety at every level



MC-JEPA: Motion & Content JEPA

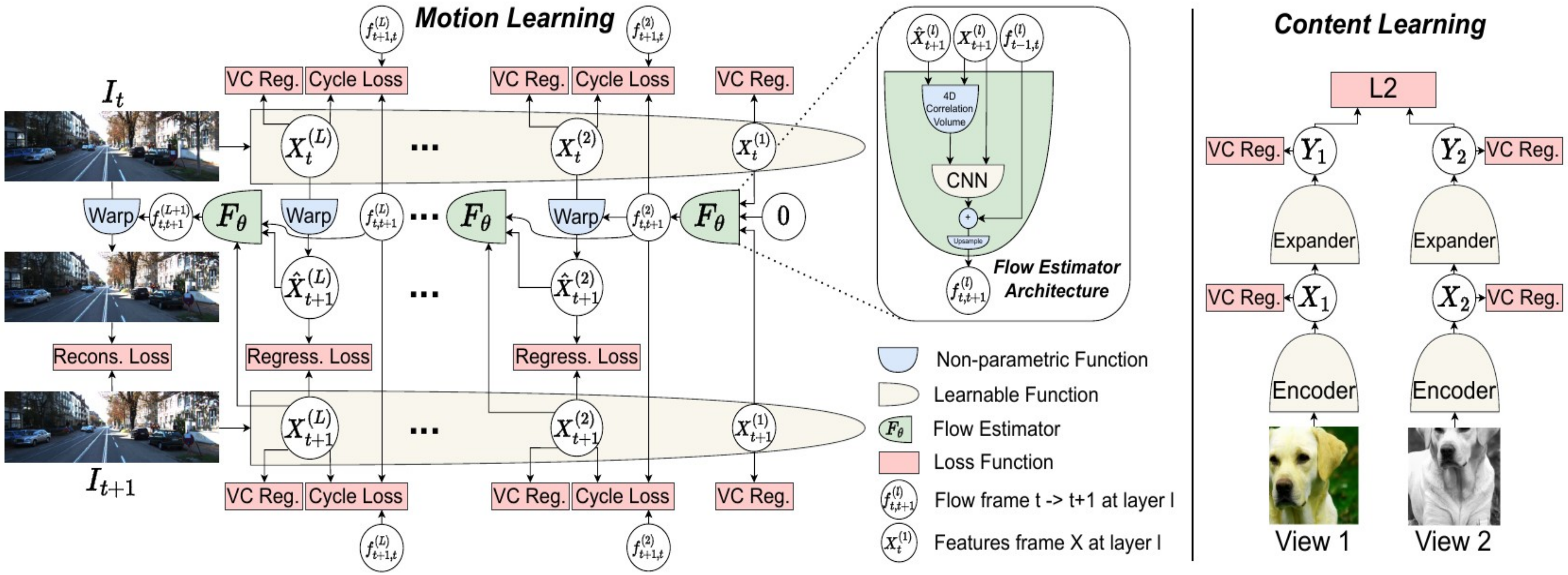
[Bardes, Ponce, LeCun 23]

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

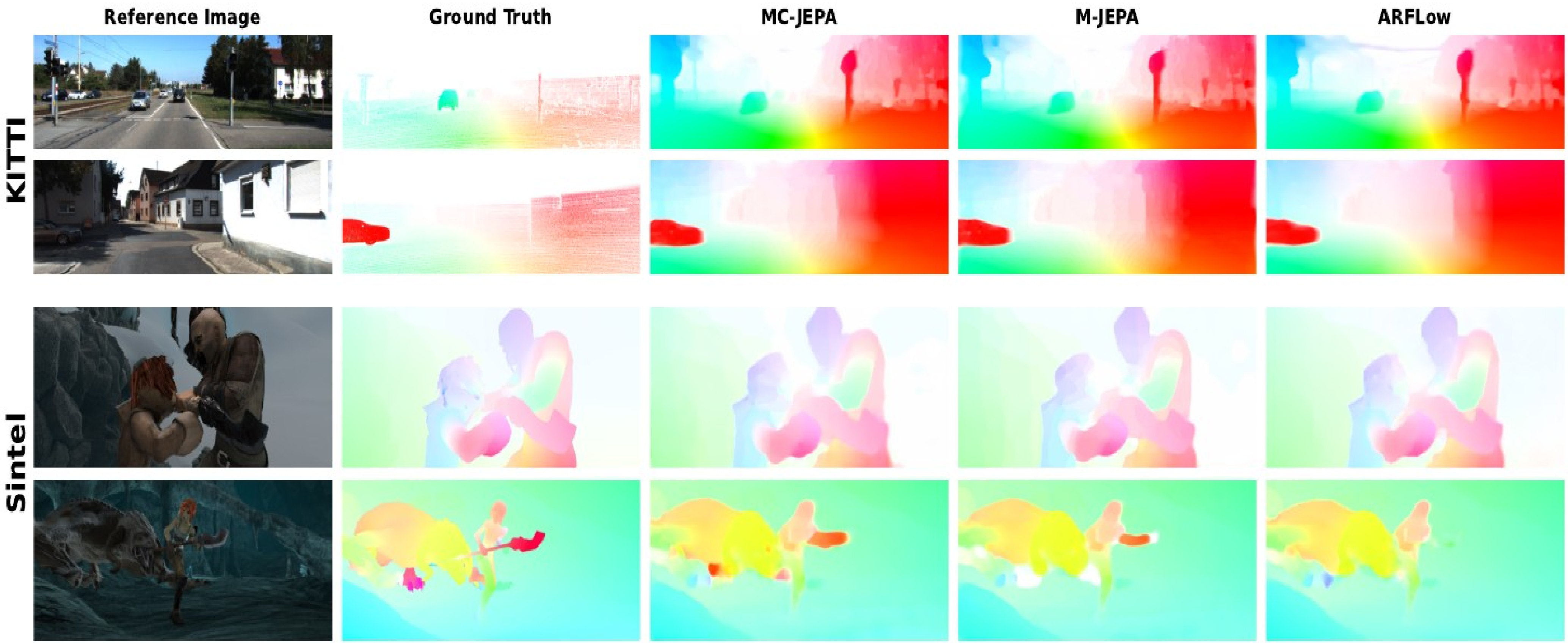


MC-JEPA: Motion & Content JEPA

- Motion estimation architecture uses a top-down hierarchical predictor that "warp" feature maps.

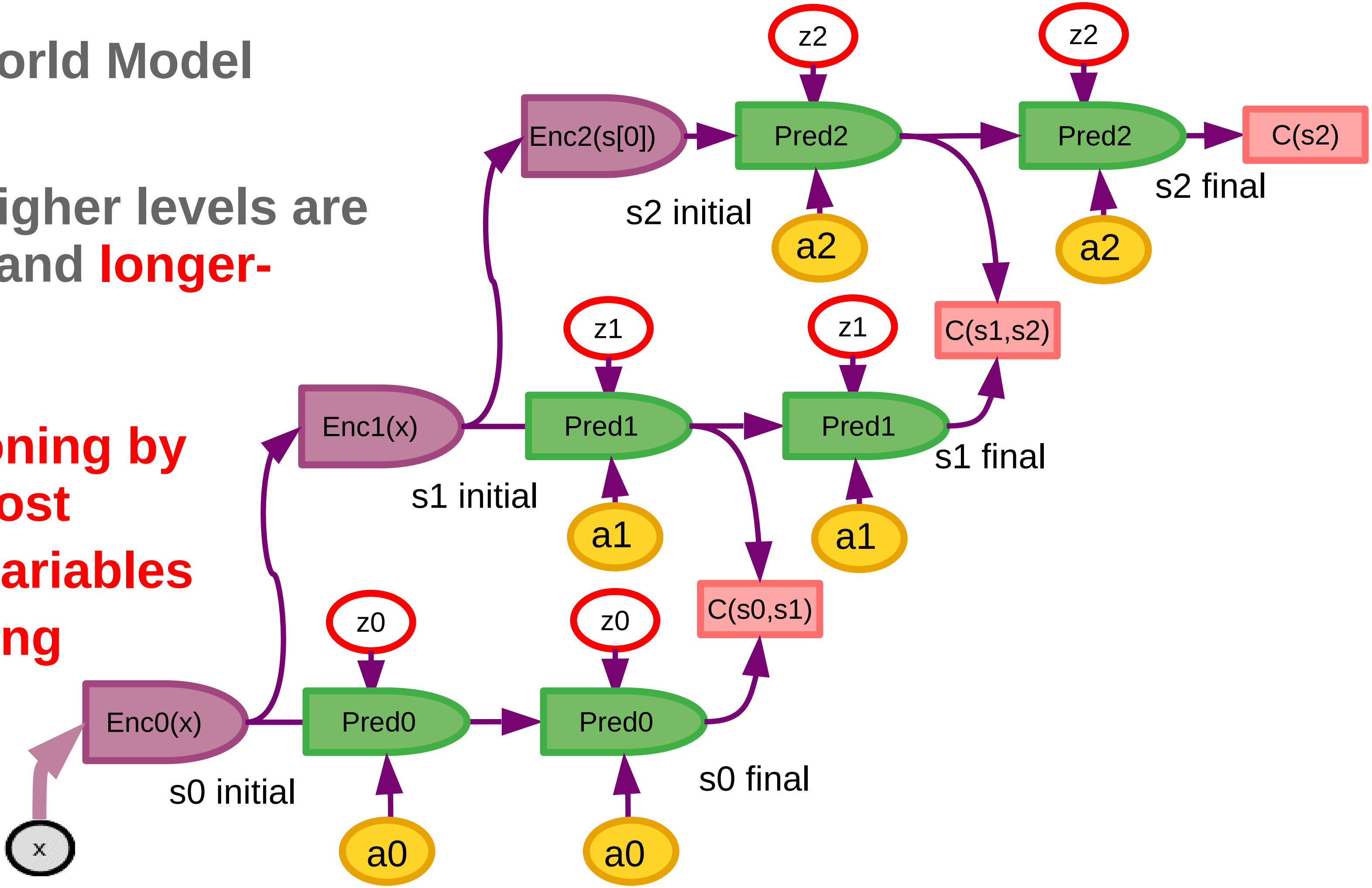


MC-JEPA: Optical Flow Estimation Results



Hierarchical Objective-Driven AI

- **Hierarchical World Model**
  - JEPA like
- **Prediction in higher levels are more abstract and longer-range.**
- **This type of planning/reasoning by minimizing a cost w.r.t "action" variables is what's missing from current architectures**



Things we are working on

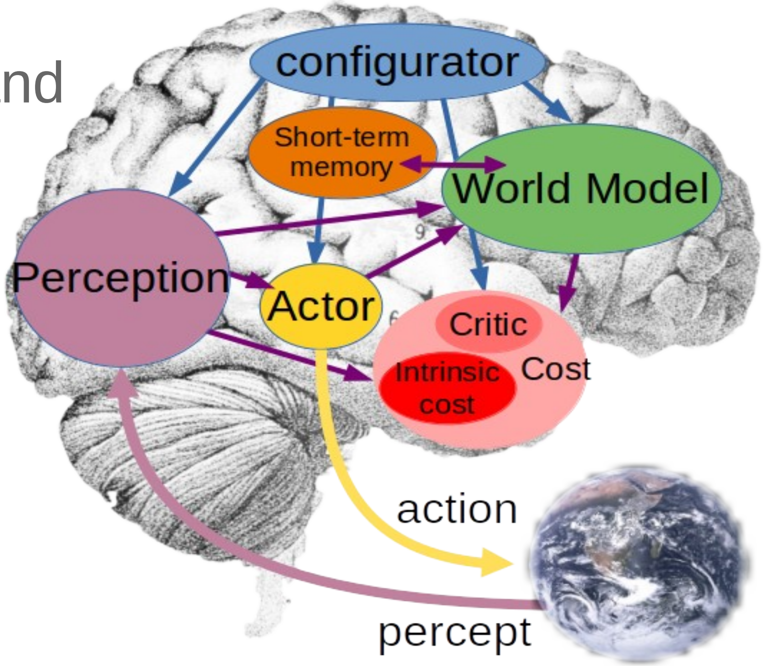
- **Self-Supervised Learning from Video**
  - Hierarchical video JEPA trained with SSL
- **LLMs that can reason & plan, driven by objectives**
  - Dialog systems that plan in representation space and use AR-LLM to turn representations into text
- **Learning hierarchical planning**
  - Training a multi-timescale H-JEPA on toy planning problems.

Problems to Solve

- **JEPA with regularized latent variables**
  - Learning and planning in non-deterministic environments
- **Planning algorithms in the presence of uncertainty**
  - Gradient-based methods and combinatorial search methods
- **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 world models**
  - Intrinsic objectives for curiosity

A Single, Configurable World Model Engine

- **What is the Configurator?**
- **The configurator configures the agent for a deliberate ("conscious") tasks.**
  - Configures all other modules for the task at hand
  - Primes the perception module
  - Provides executive control
  - Sets subgoals
  - Configures the world model for the task.



- **There is a single world model engine**
  - The system can only perform one "conscious" task at a time
  - Consciousness is a consequence of the single-world-model limitation

Points

- **Computing power**
  - AR-LLM use a fixed amount of computation per token
  - Objective-Driven AI is Turing complete (everything reduced to optimization)
- **We are still missing essential concepts to reach human-level AI**
  - Scaling up auto-regressive LLMs will not take us there
  - We need machines to learn how the world works
- **Learning World Models with SSL and JEPA**
  - Non-generative architecture, predicts in representation space
- **Objective-Driven AI Architectures**
  - Can plan their answers
  - Must satisfy objectives: are steerable & controllable
  - Guardrail objectives can make them safe.

Questions

- **How long is this going to take to reach human-level AI?**
  - Years to decades. Many problems to solve on the way.
  - Before we get to HLAI, we will get to cat-level AI, dog-level AI,...
- **What is AGI?**
  - There is no such thing. Intelligence is highly multidimensional
  - Intelligence is a collection of skills + ability to learn new skills quickly
  - Even humans can only accomplish a tiny subset of all tasks
- **Will machines surpass human intelligence**
  - Yes, they already do in some narrow domains.
  - There is no question that machine will eventually surpass human intelligence in all domains where humans are intelligent (and more)

Questions

- **Are there risks associated with human-level AI?**
  - Yes, as with every technology
  - But all those risks can be mitigated
  - Disinformation, propaganda, hate, spam,...: **AI is the solution!**
- **Should AI research be open source or heavily regulated?**
  - In a future where everyone interacts with AI assistants for everything in their daily lives, the base models **must** be open.
  - Having them controlled by a small number of company is too dangerous
- **Will robots take over the world?**
  - No! this is a projection of human nature on machines
  - Intelligence is not correlated with a desire to dominate, even in humans
  - Objective-Driven AI systems will be made subservient to humans

Questions

- **How to solve the alignment problem?**
  - Through trial and error and testing in sand-boxed systems
  - We are very familiar with designing objectives for human and superhuman entities. It’s called law making.
  - What if bad people get their hand on on powerful AI? Their evil AI will be inferior to the Good Guys’ AI police.

- **What are the benefits of human-level AI?**
  - AI will amplify human intelligence
  - Everyone will have a staff of intelligent agents working for them

- **AI will bring a new era of enlightenment, a renaissance to humanity**

Thank you!