World Models and Virtual Cell

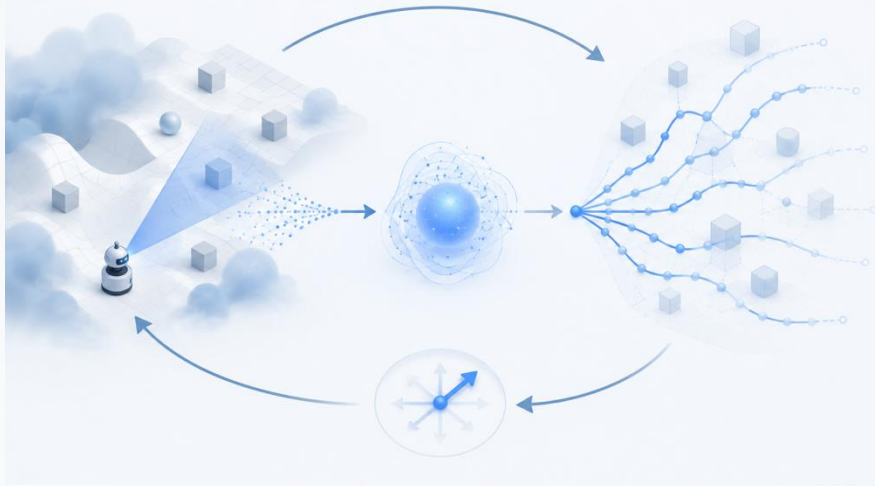
From latent prediction to virtual cells

Junzhe Li · CUHK · 2026.6.24

Content

- Foundations
- A Path Towards Autonomous Machine Intelligence
- The JEPA lineage: I-JEPA, V-JEPA, V-JEPA 2/2.1
- Theory & Control: LeJEPA, LeWorldModel
- World Model in the Biology Domain: Virtual Cell

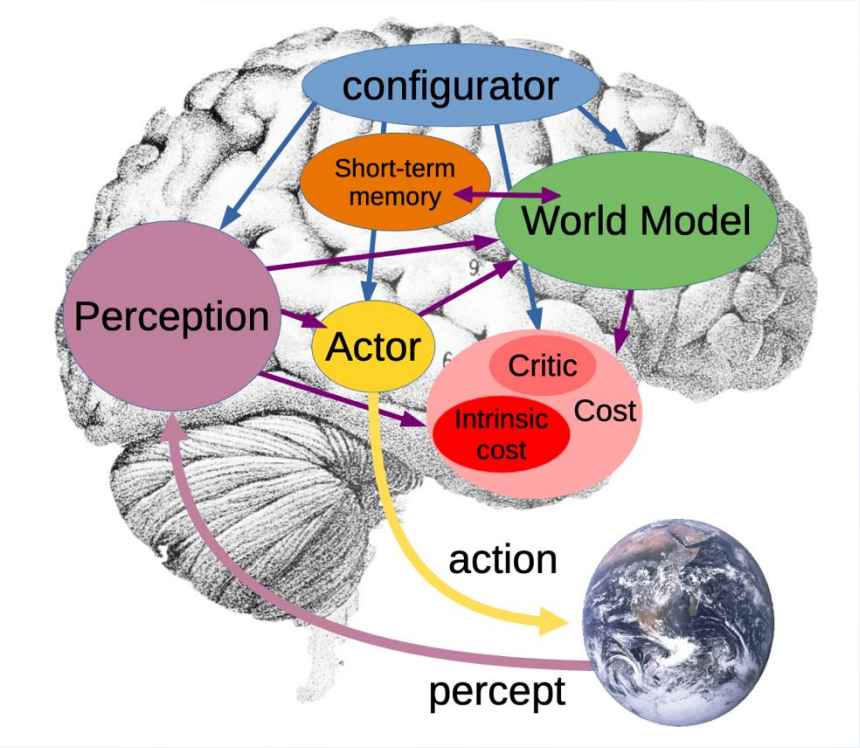
What Is a World Model, and Why Does It Matter?



> World model = an internal predictive model of how the environment changes over time, optionally conditioned on actions.

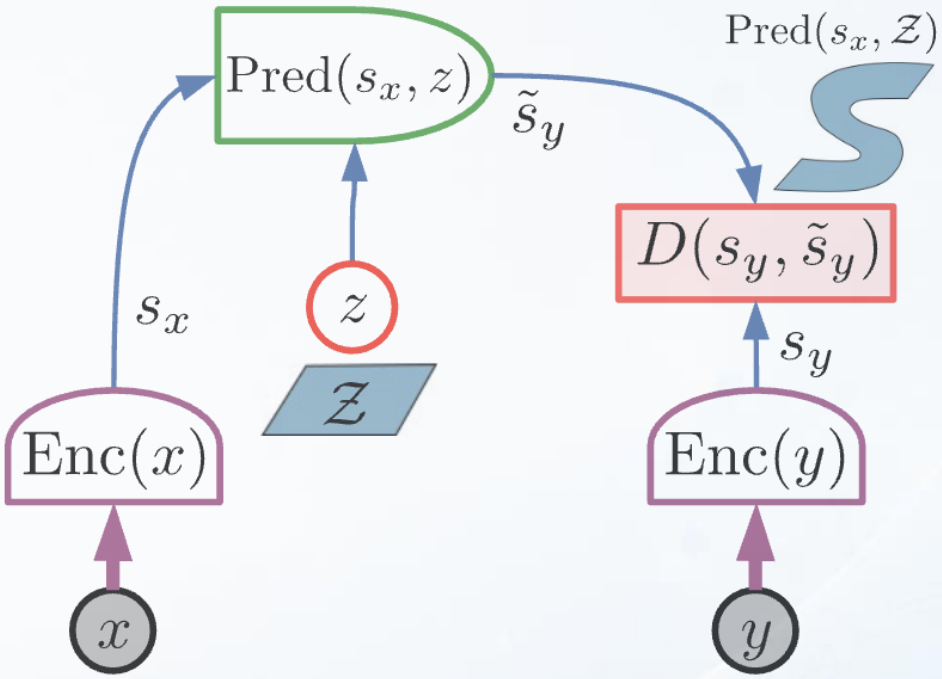
- POMDP (partially observable Markov decision process)
- This loop — agent to action to state to observation and back — is the structure that gave the modern term “world model” its technical meaning.
- Renderer, Simulator and Planner

- Humans learn driving-scale commonsense with limited active practice; many AI agents still need massive trial-and-error or supervised coverage.

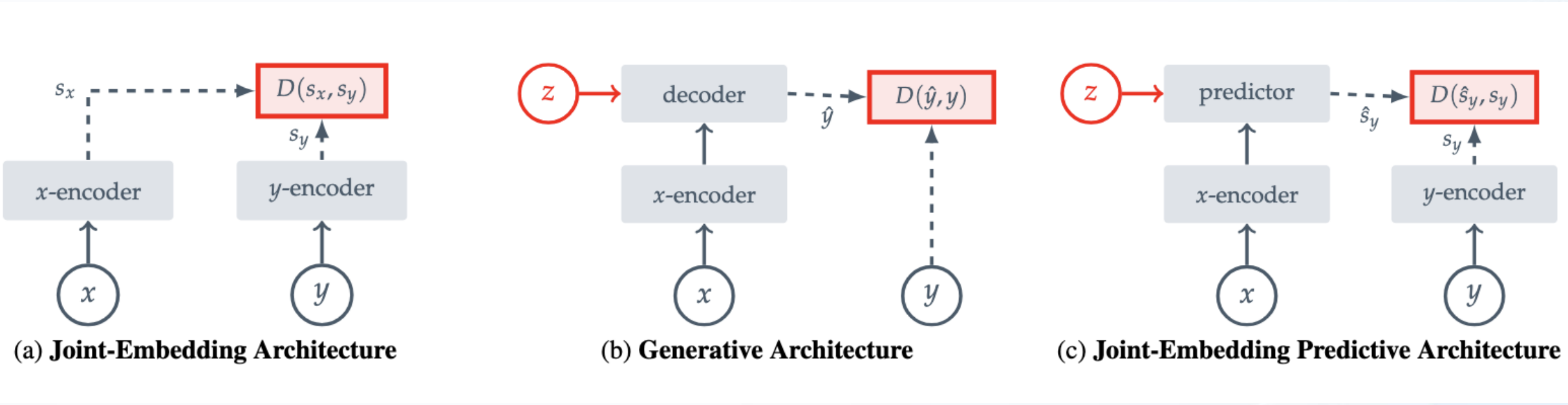


World Model: The JEPA Series

Predict future state in latent space.



The JEPA Architecture: Predictive instead of Generative



The JEPA Lineage

A progression from representation prediction to controllable latent world models

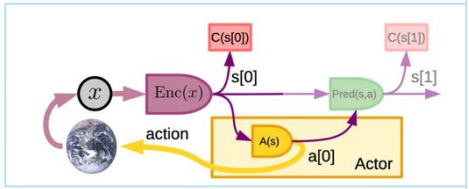
2022 — AMI / H-JEPA

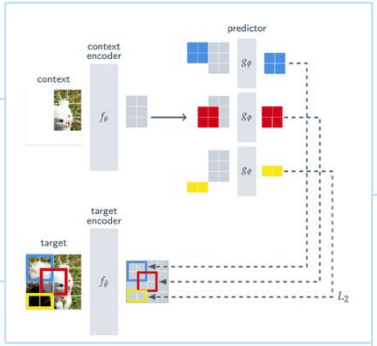
2023 — I-JEPA

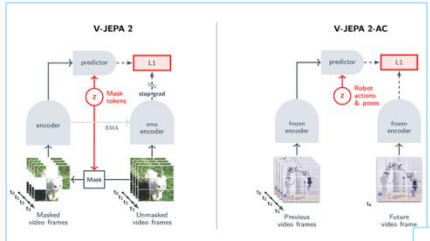
2024 — V-JEPA

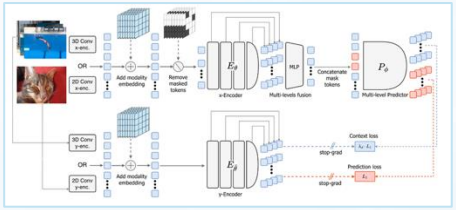
2025 — V-JEPA 2 / LeJEPA

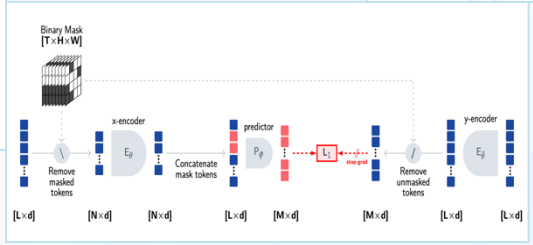
2026 — V-JEPA 2.1 / LeWM / C-JEPA



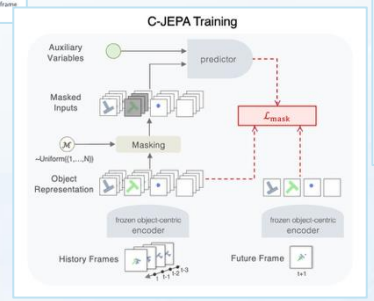


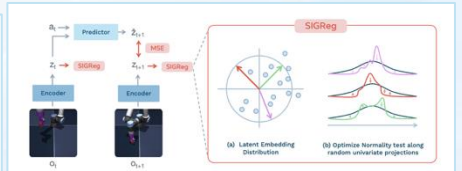






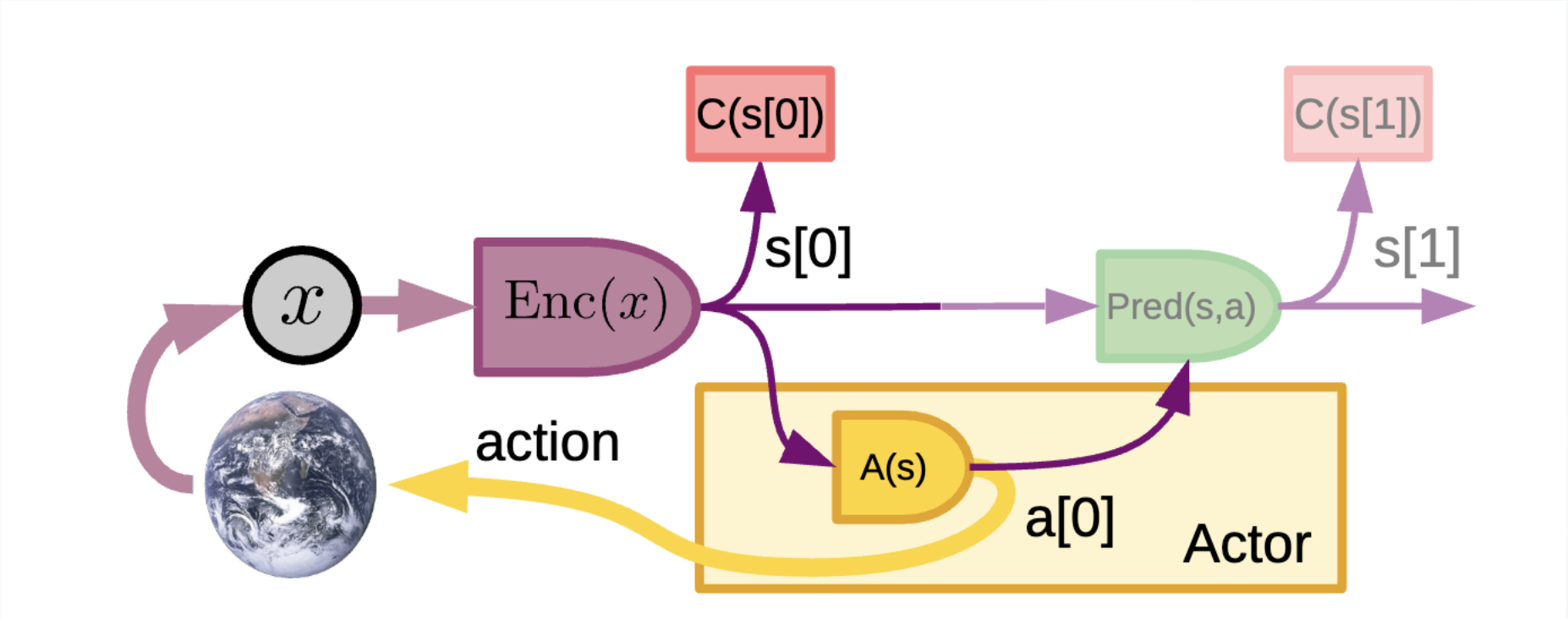
SIGReg





2022: AMI / H-JEPA

The philosophical blueprint: A Path Towards Autonomous Machine Intelligence



- Stop reconstructing pixels
- Learn latent representations in which prediction is easy and irrelevant details can be ignored.

2022: AMI / H-JEPA

Latent prediction needs anti-collapse pressure

Collapse Problem:

Solution1: VICReg (Variance-Invariance-Covariance Regularization):

1. **Variance**: a hinge loss that maintains the standard deviation of each component of s_y and v_y above a threshold over a batch.

2. **Covariance**: a covariance loss in which the covariance between pairs of different components of v_y are pushed towards zero. This has the effect of decorrelating the components of v_y, which will in turn make the components of s_y somewhat independent.

Idea:

- Maximizing Information Capacity
- Minimizing Redundancy

Later evolves into SIGReg in LeJEPA

2022: AMI / H-JEPA

Latent prediction needs anti-collapse pressure

Collapse Problem:

Solution2: stop-gradient+EMA target encoder

$$\theta_{target} = \tau\theta_{target} + (1-\tau)\theta_{context}$$

Used by I-JEPA, V-JEPA, V-JEPA2/2.1

2022: AMI / H-JEPA

Reasoning as continuous optimization, not symbolic tree search

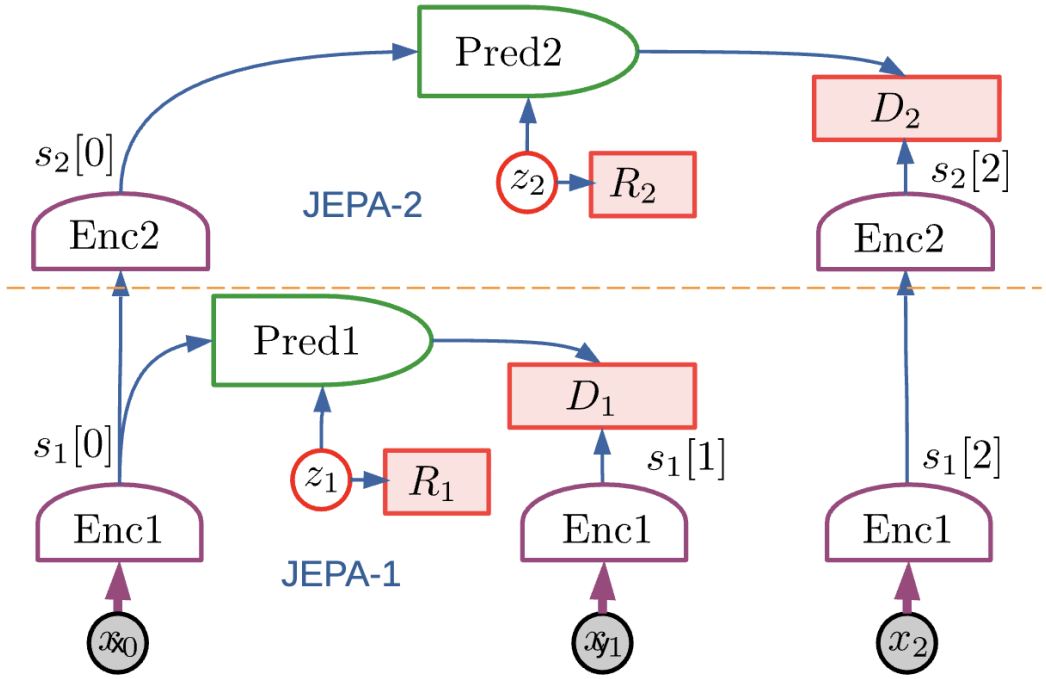


Figure 15: *Hierarchical JEPA (H-JEPA)*
The ability of the JEPA to learn abstract representations in which accurate prediction can be performed allows hierarchical stacking. In this diagram JEPA-1 extracts low-level representations and performs short-term predictions. JEPA-2 takes the representations extracted by JEPA-1 as inputs and extracts higher-level representations with which longer-term predictions can be performed. More abstract representations ignore details of the inputs that are difficult to predict in the long term, enabling them to perform longer-term predictions with coarser descriptions of the world state.

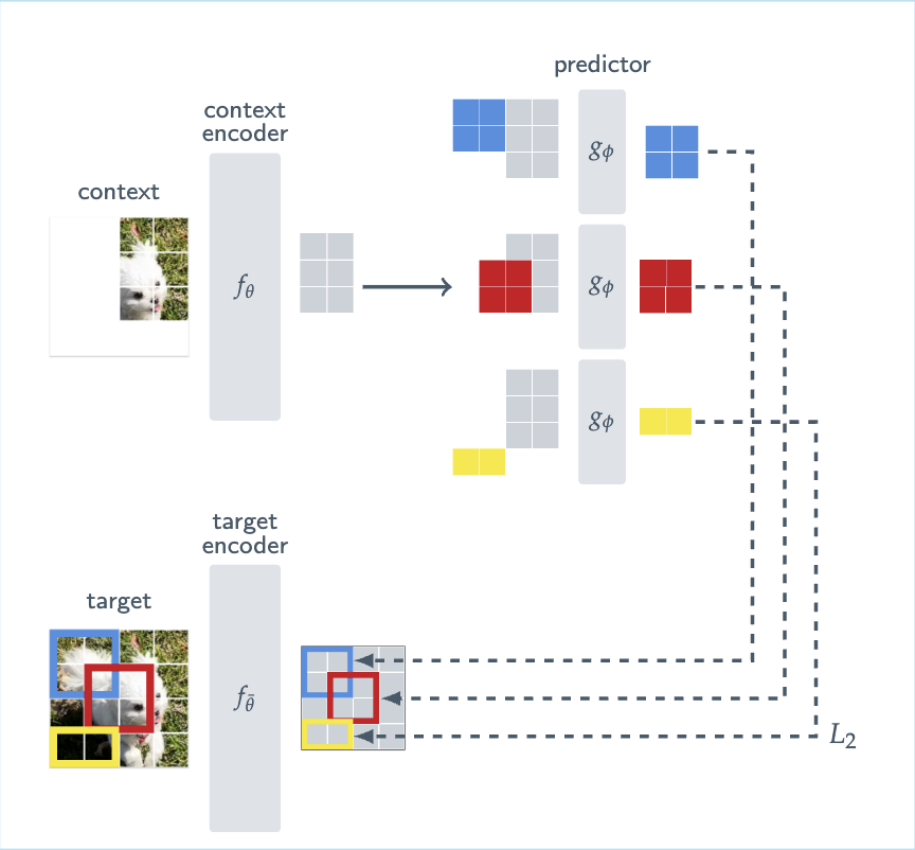
- Hierarchy: high-level "actions" become subgoals/constraints for lower levels, enabling longer-horizon prediction with coarser state descriptions.

A planning episode in latent space

1. **Encode current + goal**
2. **Infer high-level subgoals**
3. **Constrain low-level predictors**
4. **Optimize action variables by gradients**
5. **Execute first action, replan**

2023: I-JEPA

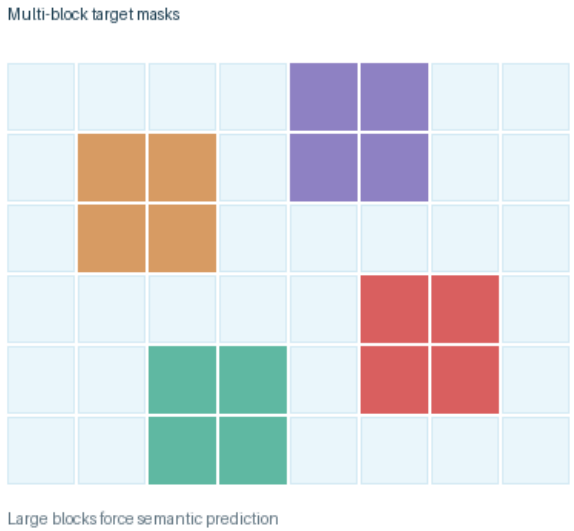
Images: predict missing regions in representation space



- Context encoder sees a large visible region; target encoder sees the original image and supplies target embeddings.
- Predictor uses mask-position tokens to predict target-block embeddings. Loss is in representation space, not pixel space.

I-JEPA Multiblock Masking Tricks

Make the task semantic, not edge interpolation

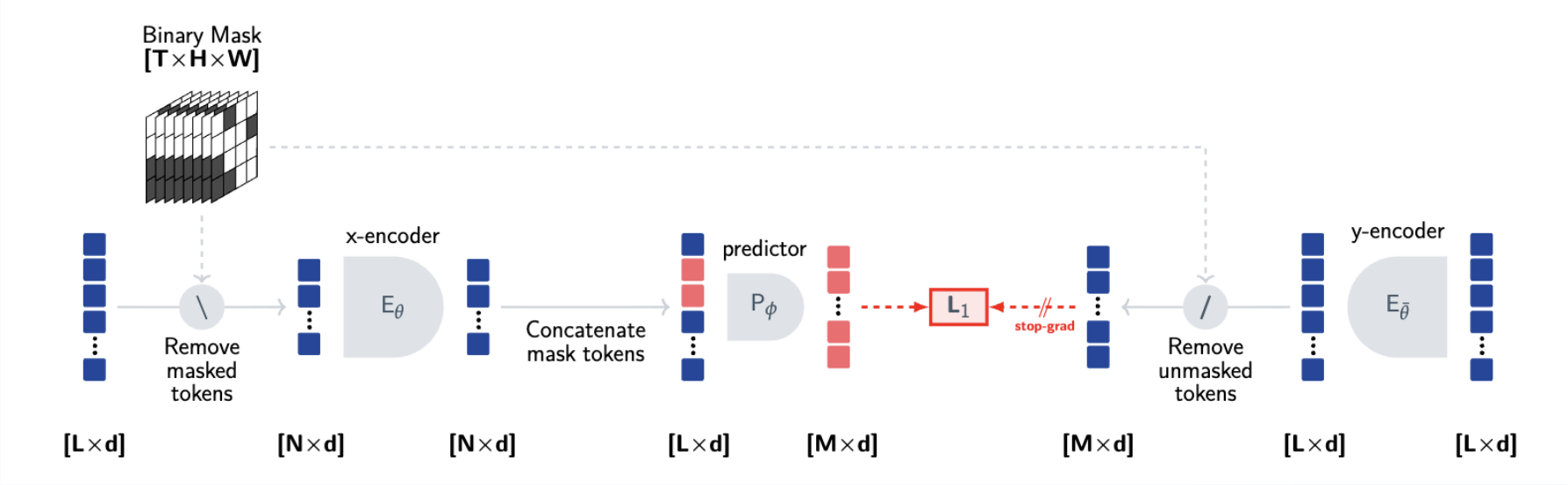


Default recipe: 4 possibly-overlapping target blocks; target scale 0.15-0.20; aspect ratio 0.75-1.5; context scale 0.85-1.0 with target regions removed.

> Why it works: a block mask suppresses local shortcuts and forces object/scene-level inference.

2024: V-JEPA

Video: feature prediction as a standalone objective



- Masking large spatiotemporal regions represented by spatiotemporal tubelets
- Short-range masking + Long range masking → global context+local movement

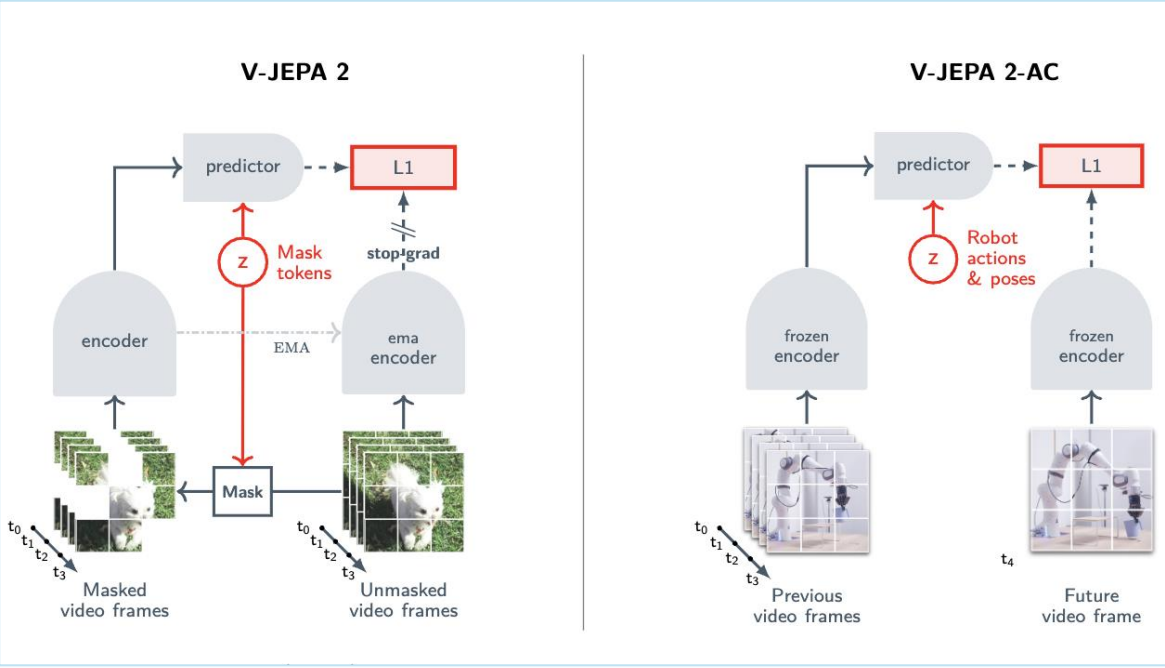
Question: Why the 90% Mask Matters

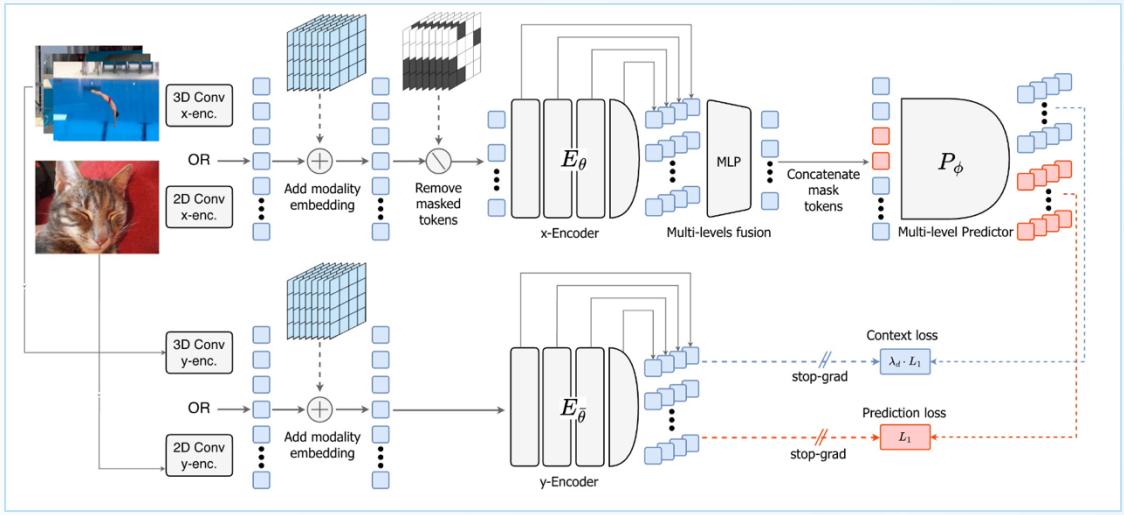
Blocking both spatial interpolation and temporal copy-paste

- A single missing frame can be copied from neighbors. A large tube-like missing region forces the model to use object permanence, motion, and scene dynamics.

2025-2026: V-JEPA 2 & V-JEPA 2-AC & V-JEPA 2.1

Goal-conditioned control without pixel rollout

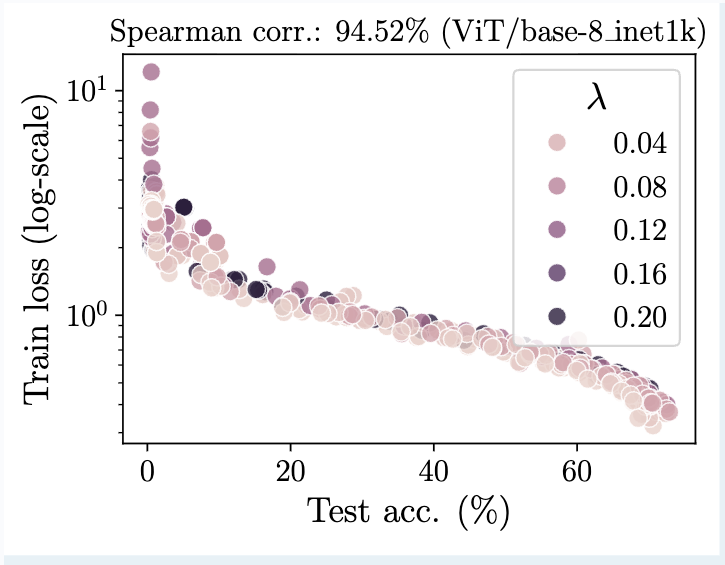


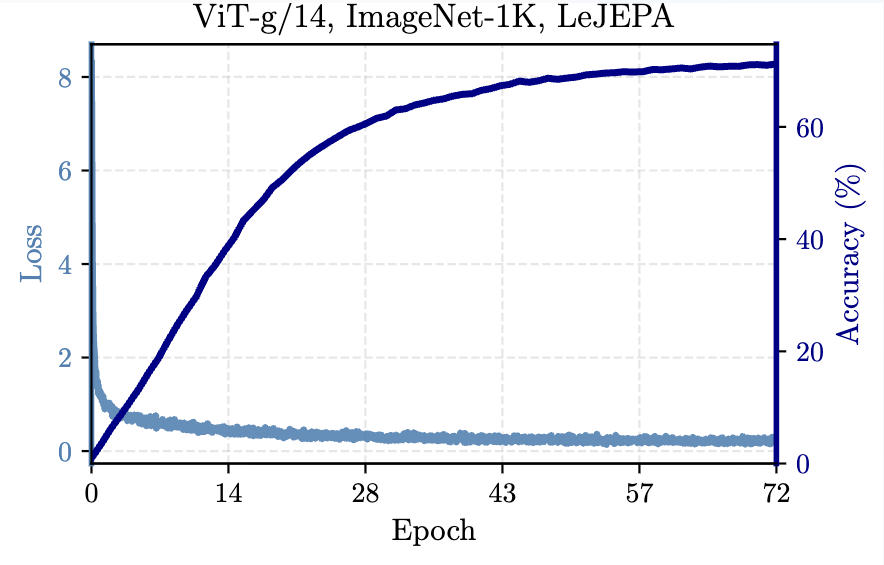


1. Scaling up.
2. Planning minimizes an energy between imagined latent state and goal-image latent state; CEM samples action sequences and executes the first action.
3. Dense Predictive Loss - visible context and masked tokens both contribute to the loss, grounding every token. (Introduce context loss into consideration)

2025: LeJEPA: Sketched Isotropic Gaussian Regularization

Provable and Scalable Self-Supervised Learning Without the Heuristics





- Isotropic Gaussian embeddings minimize expected downstream prediction risk across broad task families.
- Training loss directly translated to downstream tasks accuracy
- Bounded smooth gradients from characteristic-function tests make optimization stable even when features are far from Gaussian.

Sketched Isotropic Gaussian Regularization

Provable and Scalable Self-Supervised Learning Without the Heuristics

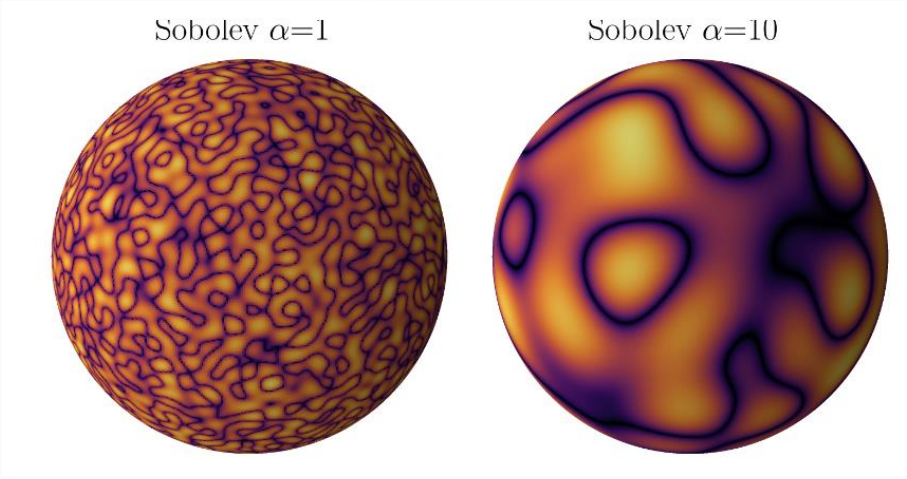
$$T^{(m)} = \int_{-\infty}^{\infty} \left| \hat{\phi}_N^{(m)}(t) - \phi_0(t) \right|^2 w(t) dt$$

1. **Generate Random Slices (Sketching):** Sample M direction vectors $u^{(m)}$ uniformly distributed on the d-dimensional unit hypersphere $\mathbb{S}^{d-1}$. *(Note: Low-discrepancy sequences like Sobol sequences are recommended for better uniform coverage).*
2. **1D Projection:** Project the N high-dimensional embeddings Z onto each of the M direction vectors. This results in M sets of 1D scalar projections: $h^{(m)} = Z \cdot u^{(m)}$.
3. **Compute Empirical Characteristic Function (ECF):** For each 1D projection set m and at each frequency point t_k, compute the ECF $\hat{\phi}_N(t_k)$ using Euler's formula.
4. **Compute Target Characteristic Function (CF):** Compute the analytical Characteristic Function of the standard 1D Gaussian $\mathcal{N}(0, 1)$ at each frequency point t_k, which is $\phi_0(t_k) = e^{-t_k^2/2}$.
5. **Calculate the Epps-Pulley Statistic (1D Loss):** For each direction m, compute the squared difference between the ECF and the Target CF across all frequency points. Integrate these differences over t using the Trapezoidal rule (with a Gaussian weighting function) to obtain the 1D test statistic $T^{(m)}$.
6. **Aggregate (Final Loss):** Average the test statistics $T^{(m)}$ across all M directions to obtain the final $\text{SIGReg}(Z)$ loss.

Interesting Tricks of SIGReg

Provable and Scalable Self-Supervised Learning Without the Heuristics

1. Why can we just sample M directions?



2. Why use trapezoid quadrature to integral?

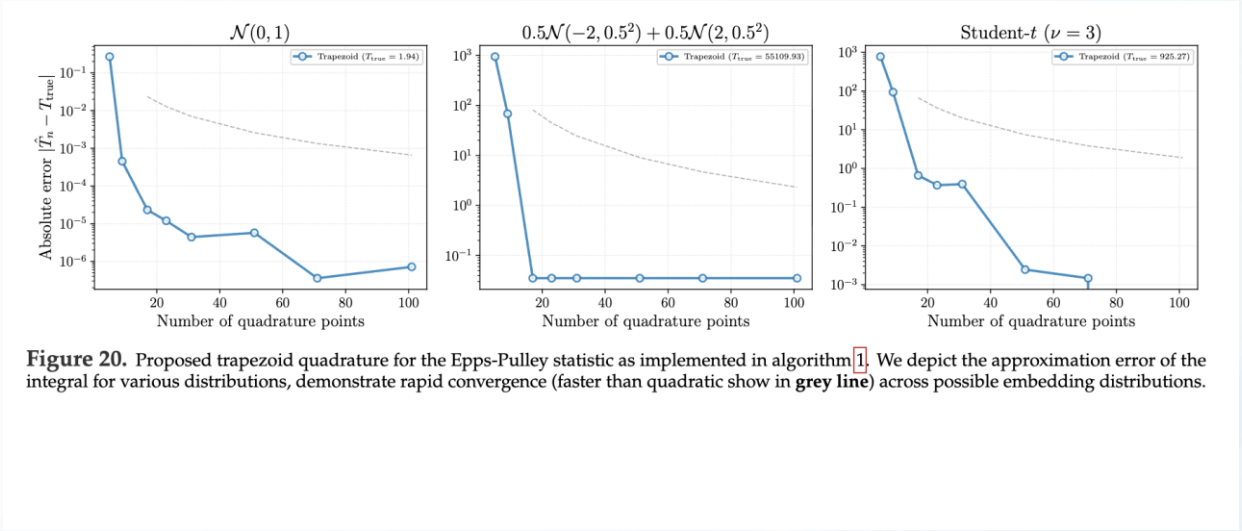


Figure 20. Proposed trapezoid quadrature for the Epps-Pulley statistic as implemented in algorithm 1. We depict the approximation error of the integral for various distributions, demonstrate rapid convergence (faster than quadratic show in **grey line**) across possible embedding distributions.

- Unified Error Bounds (large Sobolev Regularity gives exponentially low error bound)
- SGD's compounding effect of resampling
- QMC - Quasi-Monte Carlo

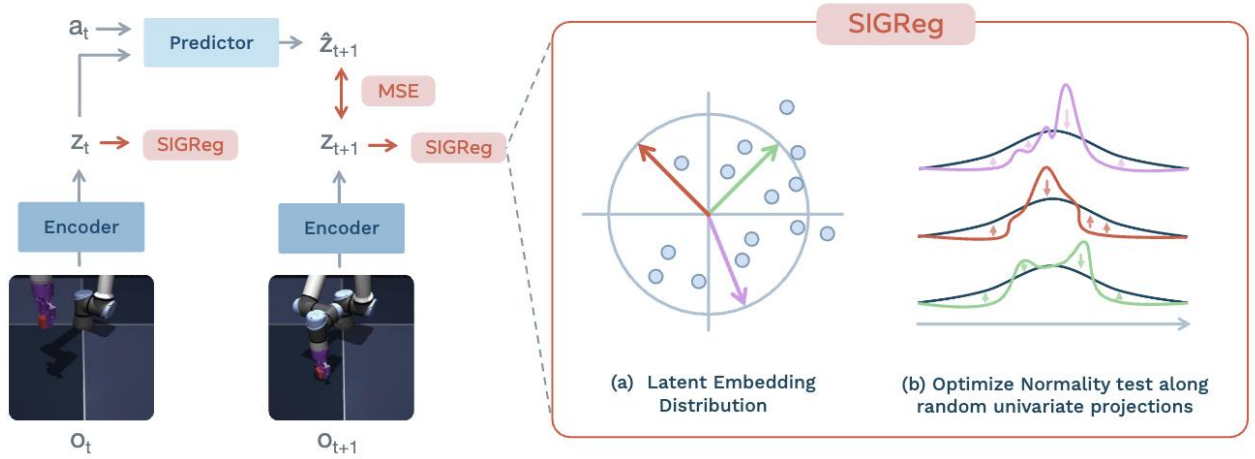
Exponential Convergence

2026: LeWorldModel

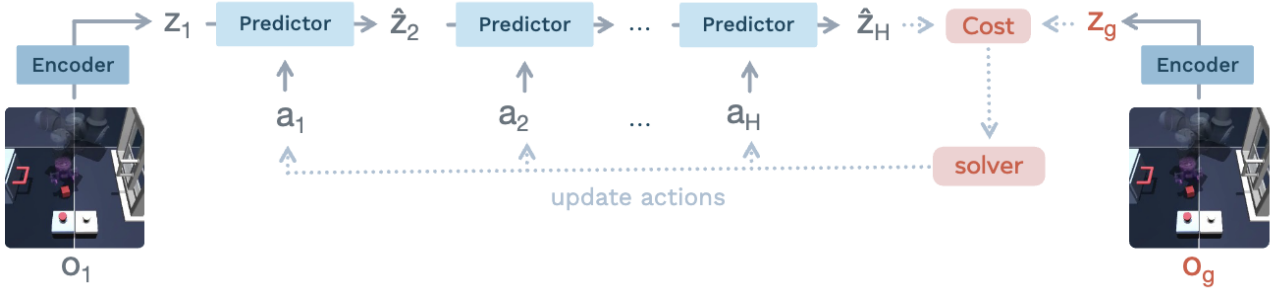
End-to-end latent world model from pixels

$$\mathcal{L}_{\text{LeWM}} \triangleq \mathcal{L}_{\text{pred}} + \lambda\,\text{SIGReg}(\boldsymbol{Z}).$$

Pipeline:



Latent planning:

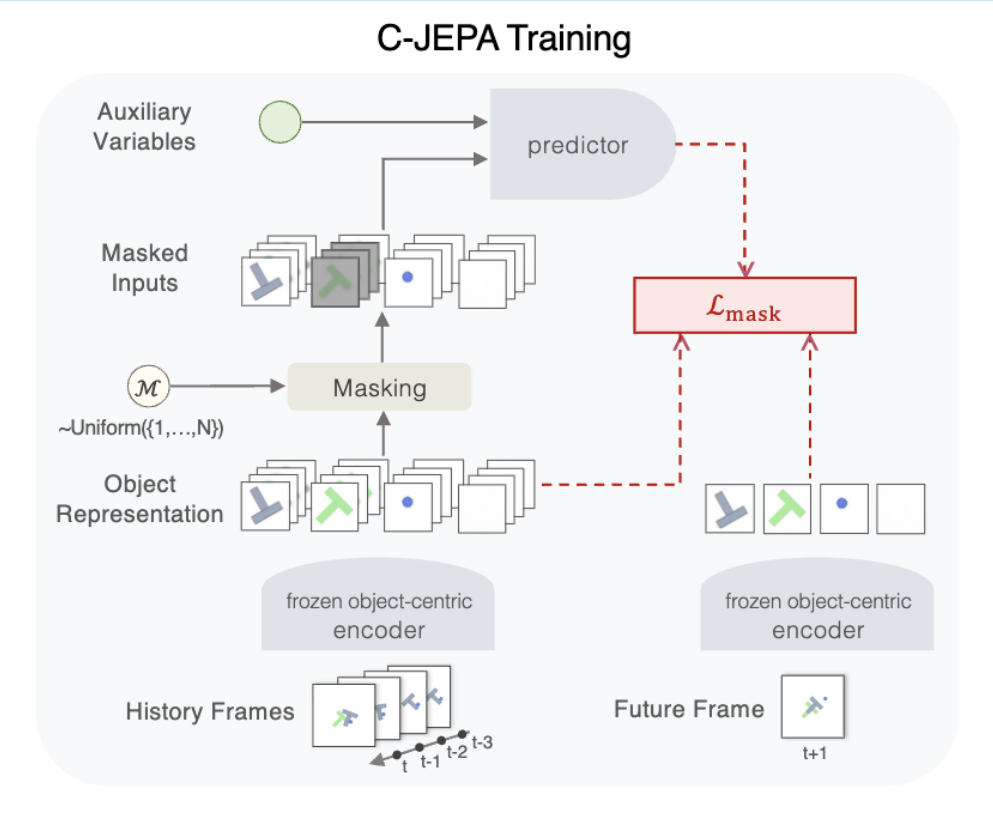


Surprising Results:

- 15M parameters, single-GPU training in a few hours, and planning up to 48x faster than foundation-model-based WMs while competitive on control tasks
- Violation-of-expectation: surprise spikes more for physically implausible teleportation than for visual color changes.
- Temporal straightening: latent velocity directions become more aligned during training, suggesting simpler latent dynamics.

2026: Causal-JEPA

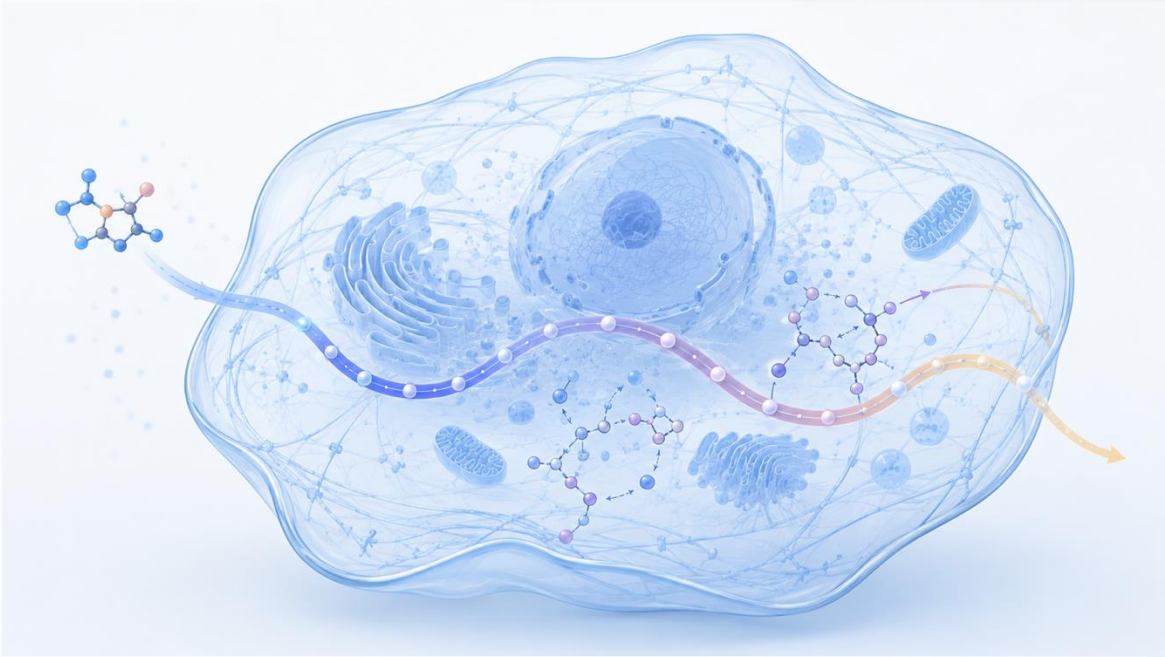
Learning World Models through Object-Level Latent Masking



- Object-level latent masking and Influence Neighborhood (No DAG)

World Models for Virtual Cell

A speculative but coherent extension: from physical scenes to biological state-space dynamics.



What is Virtual Cell

Arc Institute:

- Virtual cell models offer a path to accelerate the study and treatment of complex diseases by deeply understanding the dynamic nature of cellular state across cell contexts. By learning how cellular gene expression and behavior shifts in response to chemical, genetic, or environmental changes across cell types, virtual cell models can begin to predict how we can nudge a cell from a diseased state to a healthy one.

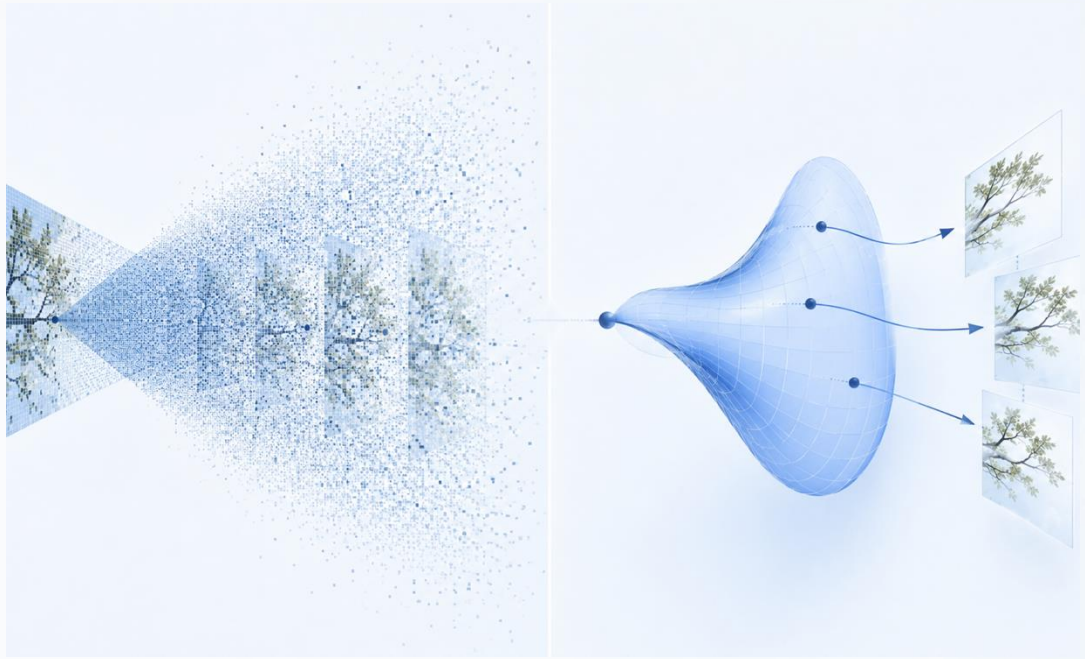
Virtual Cell: a World Model Biological Translation

World Model	***Virtual Cell***	***Example***
State	Cell state	transcriptome、epigenome、proteome、morphology、cell type、disease state
Action	Intervention	CRISPR knockout/ knockdown, drug,temperature, microenvironment
Transition dynamics	Biodynamics	cell-cycle, stress response, differentiation, apoptosis, metabolism
Observation	Partial measurements	scRNA-seq、Perturb-seq、spatial omics、imaging、proteomics
Planning	Intervention search and experimental design	Gene knockout and its consequence

Virtual Cell: why consider JEPA?

Aleatoric noise dominates the microscopic world

- Brownian motion, stochastic binding, random organelle motion, and measurement noise are not the right targets for deterministic long-horizon prediction.
- A model trained to predict all pixels may waste capacity on noise or hallucinate visually plausible but biologically irrelevant detail.
- JEPA hypothesis: learn the predictable invariants - cell-cycle phase, pathway activity, causal response - while suppressing irreducible randomness.



- Spatial mask: hide a cellular region and predict latent features from cytoskeleton / organelle context.
- Temporal mask: hide intermediate frames to force causal event inference, e.g., organelle fusion/fission or mitosis progression.
- Modality mask: predict missing molecular readout from imaging or vice versa.

In Silico Intervention Planning: Possible Paradigm

Drug perturbation as action-conditioned latent rollout

Current cell
z_t

→

Candidate action
a_t = drug / CRISPR

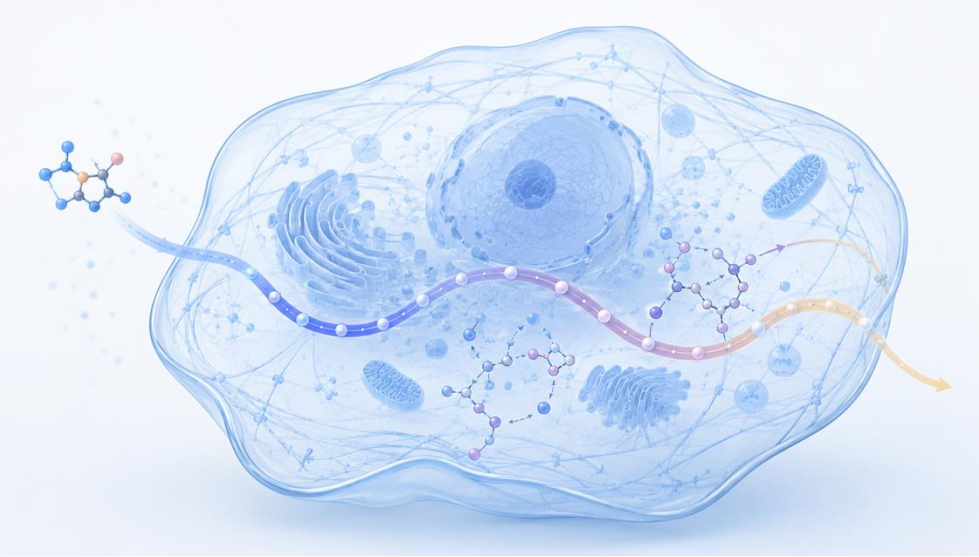
→

Latent rollout
z_{t+1:t+T}

→

Objective
healthy target / kill tumor

Context+Uncertainty



Latent MPC: sample or optimize perturbation sequences in latent space, score by target phenotype, then test top candidates experimentally.

Potential uses: drug screening, target discovery, toxicity prediction, synthetic lethality, and disease-state reversal.

Working hypothesis: intervention-relevant latent predictability may be a useful inductive bias for biological dynamics.

Challenges

1. State identifiability

2. Intervention labels

3. Multi-scale time

4. Causality

5. Evaluation

Takeaways & Discussion

From pixel slavery to latent world engines

World Model

- World models are not just video generators; they are state-estimation, prediction, and planning systems.
- Possibility of integrating LLM +WAM?

JEPA

- JEPA's core bet is to predict abstract embeddings and ignore unpredictable detail.
- Regulation trick: SIGReg
- JEPA itself is a learning paradigm. Use it as a block.

World Models for Virtual Cell

- Virtual cell is a natural frontier if we treat biological dynamics as a partially observed, action-conditioned latent system.
- action-conditioned simulation、counterfactual reasoning 和 long-horizon planning

Thanks for listening

Q&A

References

- LeCun. A Path Towards Autonomous Machine Intelligence. 2022.
- Assran et al. Self-Supervised Learning from Images with a JEPA. CVPR 2023.
- Bardes et al. Revisiting Feature Prediction for Learning Visual Representations from Video. 2024.
- Assran et al. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. 2025.
- Balestriero & LeCun. LeJEPA: Provable and Scalable SSL Without the Heuristics. 2025.
- Mur-Labadia et al. V-JEPA 2.1: Unlocking Dense Features in Video SSL. 2026.
- Maes et al. LeWorldModel: Stable End-to-End JEPA from Pixels. 2026.
- Nam et al. Causal-JEPA: Learning World Models through Object-Level Latent Masking. 2026.
- Fei-Fei Li. A Functional Taxonomy of World Models. 2026.

Appendix: Unified View of the JEPA Family

Model	Observation	Prediction target	Anti-collapse	What it adds
H-JEPA	multi-scale state	future latent state	VICReg idea	hierarchical planning
I-JEPA	image patches	target block embeddings	EMA + stop-grad	semantic 2D reps
V-JEPA	video tubelets	masked video embeddings	EMA + stop-grad	motion/world dynamics
V-JEPA 2	internet video + robot data	action-conditioned future latents	frozen encoder + predictor	real robot planning
V-JEPA 2.1	image/video tokens	dense all-token prediction	EMA + deep SSL	localization + depth
LeJEPA/LeWM	views or pixels + actions	latent agreement / next z	SIGReg	heuristics-free JEPA
C-JEPA	object slots	masked object trajectory	object-level masking	interaction bias