
What Must JEPA Preserve for Tasks It Has Not Seen?

Odysseus Kaloidas
UC Berkeley

Abstract

JEPA learns from images or video before all its later uses are known. How can it learn what to ignore when a future task may need that very detail? Joint-Embedding Predictive Architectures learn numerical representations by predicting parts of those representations from available context, rather than reconstructing every pixel. This can capture relationships useful across many tasks. The question is how far that usefulness extends when the task changes. A simple example explains why predictive success alone cannot answer it, while leaving open what real JEPA models preserve and what further learning can recover.

1 Useful Before We Know the Use

A model may learn from a video long before anyone decides what to ask it. Later, one user might want to predict an object’s motion; another might want to identify the object. Both questions concern the same scene, but they need different details. How can a model learn what matters before those questions have been chosen?

This problem matters for Joint-Embedding Predictive Architectures, or JEPAs, because their ambition extends beyond solving the prediction task used to train them. During pretraining, they learn numerical representations of observations by predicting target representations from available context. A downstream task is a later use of what was learned, such as recognizing an object or planning an action. The hope is that relationships learned through prediction will help with many such uses.

Consider a pendulum’s motion and the serial number printed on it. Position and speed help forecast its motion, while the number identifies the device. These are two possible uses of a broadly pretrained model, rather than a description of its training objective. The interesting question is whether learning from many observations gives the model reasons to preserve both kinds of information, even before either use is selected.

2 Why Predictive Representations Might Transfer

I-JEPA learns from images by predicting representations of hidden regions from visible context [2]. Here, prediction concerns missing parts of an image, rather than necessarily the next moment in time. V-JEPA 2 learns from video and images; further training produces an action-conditioned model, which predicts future representations given proposed actions [3]. Learning a reusable representation and learning how actions change a scene are related, but distinct, steps.

Prediction in latent space, the space of these learned numerical representations, offers a way to learn useful structure without reconstructing every pixel. Which structure emerges depends partly on the training data, the regions hidden during training and the representations used as prediction targets. I-JEPA reports successful transfer to classification, object counting and depth prediction [2]. That breadth gives a concrete reason to take the approach seriously.

The pendulum illustrates why transfer is plausible. A representation of position and motion could help forecast its path, time an intervention or detect an unexpected change. One relationship supports

several uses. Identifying the device by its serial number asks for something different. The question is where this sharing of useful structure continues and where a new task needs a distinction that pretraining did little to encourage.

Temporal models make the positive case more concrete. World Models combines reconstruction, forecasting and control [7]. LeWorldModel predicts the next representation from images [9], while Fast LeWorldModel predicts future representations in parallel using proposed actions [6]. An August 2026 preprint also reports planning with action-conditioned JEPA variants using point clouds, collections of positions in three-dimensional space, with varying resistance to noise [10]. These studies concern particular prediction and control tasks. Transfer to substantially different uses remains a separate question.

3 What Predictive Success Does Not Settle

A simple example isolates the distinction. Imagine an observation containing two yes-or-no facts: one describes motion, and the other identifies the object. Call them S and U , and write the observation as $X = (S, U)$. Suppose the bits are independent, each equally likely to be zero or one. Suppose also that S persists, so the next-step forecasting target is $F = S$. A later identification task asks for $Y = U$.

Compare two summaries. The shorter code, $Z_1 = S$, keeps only the motion fact. The longer code, $Z_2 = (S, U)$, keeps both. Both predict F perfectly. For identification, however, the shorter code leaves only a guess, with best accuracy $1/2$; the longer code gives accuracy 1. The prediction score cannot tell these summaries apart, even though their usefulness for identification differs.

Real JEPA targets are learned, and pretraining spans many examples and contexts. This fixed-target example does not establish what JEPA actually discards. It explains why the prediction result alone cannot establish what else a representation preserves. This concern applies to predictive representation learning more broadly, not uniquely to JEPA.

The same example helps interpret agreement between models. Huh and colleagues' Platonic Representation Hypothesis proposes that learned representations are converging toward a shared statistical reality, while recognizing that different modalities contain different information [8]. In the example, several models could all retain S and omit U . Agreement would then coexist with a shared omission. This possibility does not refute observed convergence; it explains why convergence and coverage of later information needs are different questions.

The information bottleneck gives a formal account of learning a compact code while preserving information about a chosen target [11; 1]. Here the difficulty is that the later target may differ. Let Y be the answer to a downstream question, X the original observation and $Z = E(X)$ its encoded representation. Write $H(Y | Z)$ for the uncertainty about that answer given the representation, and $H(Y | X)$ for the uncertainty given the original observation. For finite variables and deterministic encoding,

$$H(Y | Z) - H(Y | X) = I(Y; X | Z) \geq 0.$$

The right side, conditional mutual information, measures information about the answer still present in the observation but absent from the code [5]. A value of zero means that encoding has lost nothing relevant to that answer. In the example, the shorter code loses all information about identity; the longer code preserves it. This is an exact-information criterion, stronger than merely performing well on a task. Useful transfer can tolerate some loss. The empirical question is which losses matter for the tasks we care about.

4 Keeping Information and Being Able to Use It

Keeping more detail is not automatically the practical solution. World Models describes a reconstruction-trained encoder retaining wall patterns while missing useful road details [7]. In ideal decision theory, extra information cannot worsen an optimal decision when it can simply be ignored [4]. A learner with limited data and computation still has to find the useful information. A serial number might remain recoverable from its representation yet require considerable training to read.

There is also a difference between a fixed representation and an adaptable system. Fine-tuning the encoder with access to original observations can change what it retains; a fresh observation can supply a missing detail. A limitation of the pretrained code need not be a permanent limitation of the system.

JEPA’s appeal lies partly in learning what many tasks share before being told what those tasks are. Its success on existing tasks makes that prospect credible. The open question is how far the shared structure will carry: when does a new question fit what the model has already learned, and when does it make a previously dispensable detail essential?

References

- [1] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. URL <https://arxiv.org/abs/1612.00410>.
- [2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture, 2023. URL <https://arxiv.org/abs/2301.08243>. arXiv:2301.08243.
- [3] Mido Assran, Adrien Bardes, David Fan, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL <https://arxiv.org/abs/2506.09985>. arXiv:2506.09985.
- [4] David Blackwell. Equivalent Comparisons of Experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953. doi: 10.1214/aoms/1177729032.
- [5] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, second edition, 2006. ISBN 9780471241959. URL <https://www.wiley.com/en-us/Elements+of+Information+Theory%2C+2nd+Edition-p-9780471241959>.
- [6] Yuntian Gao and Xiangyu Xu. Fast LeWorldModel, 2026. URL <https://arxiv.org/abs/2606.26217>. arXiv:2606.26217, preprint.
- [7] David Ha and Jürgen Schmidhuber. World models, 2018. URL <https://worldmodels.github.io/>. arXiv:1803.10122.
- [8] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20617–20642. PMLR, 2024. URL <https://proceedings.mlr.press/v235/huh24a.html>.
- [9] Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. LeWorld-Model: Stable end-to-end joint-embedding predictive architecture from pixels, 2026. URL <https://arxiv.org/abs/2603.19312>. arXiv:2603.19312, preprint.
- [10] Fabio F. Oberweger and Michael Schwinghackl. Does latent planning survive point clouds? action-conditioned JEPA world models for geometric observations, 2026. URL <https://arxiv.org/abs/2608.29434>. arXiv:2608.29434, preprint.
- [11] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. arXiv:physics/0004057, 2000. URL <https://arxiv.org/abs/physics/0004057>.