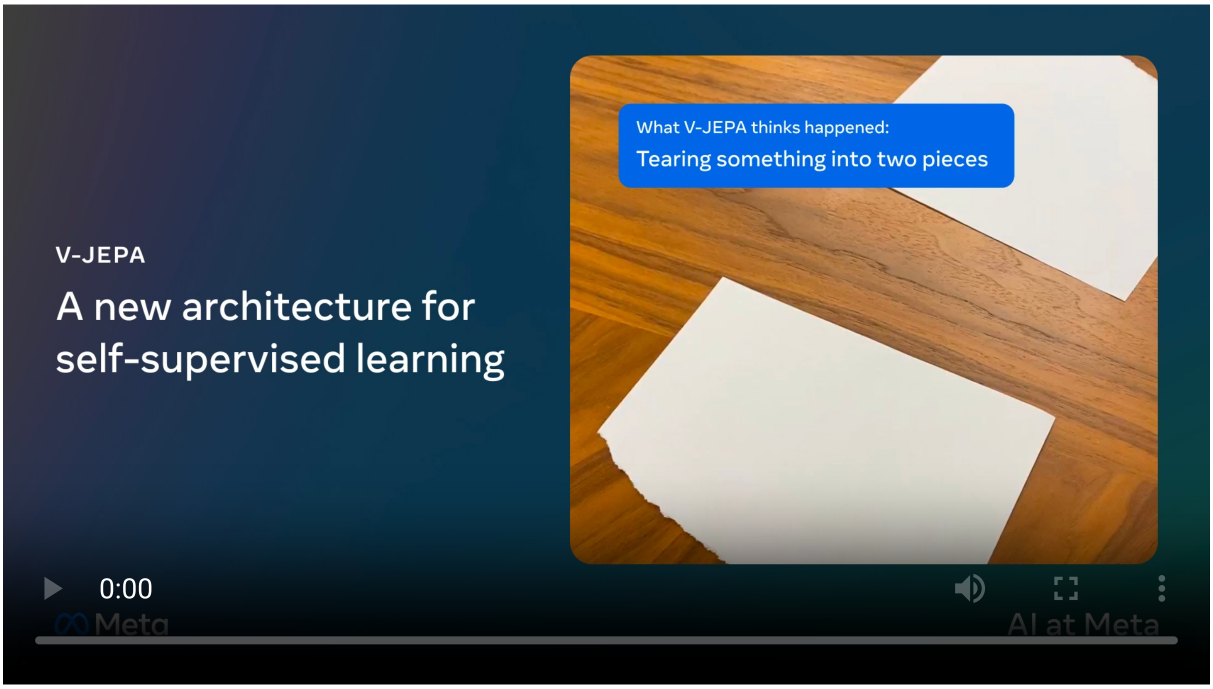
Research

V-JEPA: The next step toward Yann LeCun’s vision of advanced machine intelligence (AMI)

February 15, 2024 • 3 minute read



V-JEPA

A new architecture for self-supervised learning

What V-JEPA thinks happened:

Tearing something into two pieces

Takeaways:

- Today, we’re publicly releasing the Video Joint Embedding Predictive Architecture (V-JEPA) model, a crucial step in advancing machine intelligence with a more grounded understanding of the world.
- This early example of a physical world model excels at detecting and understanding highly detailed interactions between objects.
- In the spirit of responsible open science, we’re releasing this model under a Creative Commons NonCommercial license for researchers to further explore.

As humans, much of what we learn about the world around us—particularly in our early stages of life—is gleaned through observation. Take Newton's third law of motion: Even an infant (or a cat) can intuit, after knocking several items off a table and observing the results, that what goes up must come down. You don't need hours of instruction or to read thousands of books to arrive at that result. Your internal world model—a contextual understanding based on a mental model of the world—predicts these consequences for you, and it's highly efficient.

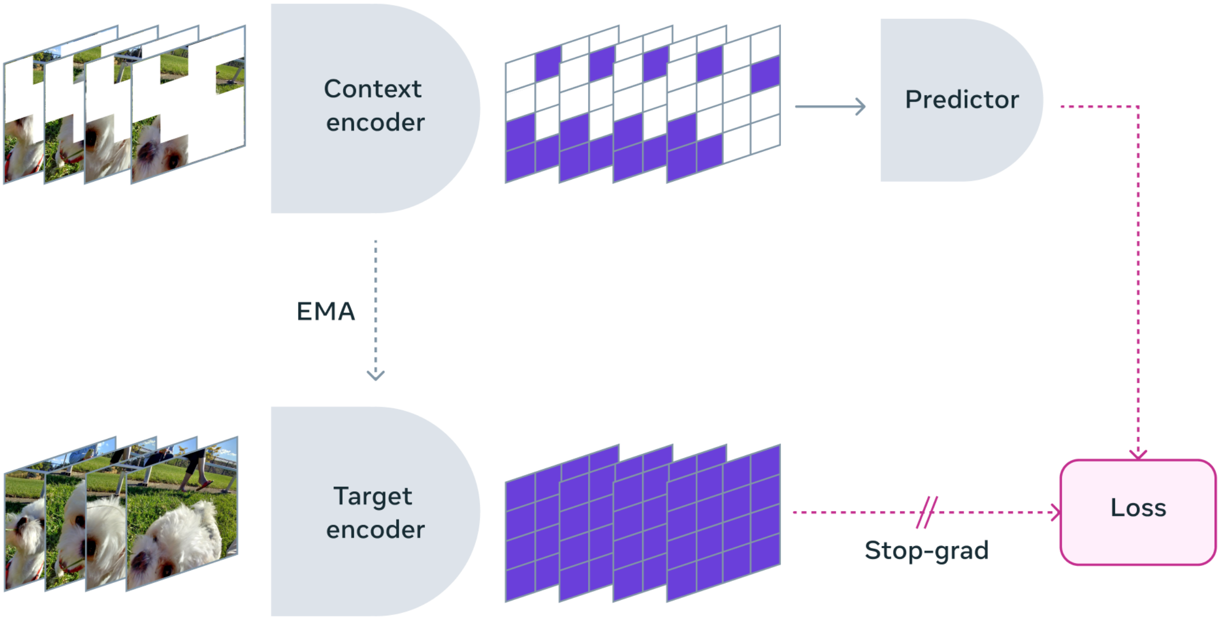
"V-JEPA is a step toward a more grounded understanding of the world so machines can achieve more generalized reasoning and planning," says Meta's VP & Chief AI Scientist Yann LeCun, who proposed the original Joint Embedding Predictive Architectures (JEPA) in 2022. "Our goal is to build advanced machine intelligence that can learn more like humans do, forming internal models of the world around them to learn, adapt, and forge plans efficiently in the service of completing complex tasks."

Video JEPA in focus

V-JEPA is a non-generative model that learns by predicting missing or masked parts of a video in an abstract representation space. This is similar to how our Image Joint Embedding Predictive Architecture (I-JEPA) compares abstract representations of images (rather than comparing the pixels themselves). Unlike generative approaches that try to fill in every missing pixel, V-JEPA has the flexibility to discard unpredictable information, which leads to improved training and sample efficiency by a factor between 1.5x and 6x.

Because it takes a self-supervised learning approach, V-JEPA is pre-trained entirely with unlabeled data. Labels are only used to adapt the model to a particular task after pre-training. This type of architecture proves more efficient than previous models, both in terms of the number of labeled examples needed and the total amount of effort put into learning even the unlabeled data. With V-JEPA, we've seen efficiency boosts on both of these fronts.

With V-JEPA, we mask out a large portion of a video so the model is only shown a little bit of the context. We then ask the predictor to fill in the blanks of what's missing—not in terms of the actual pixels, but rather as a more abstract description in this representation space.



V-JEPA trains a visual encoder by predicting masked spatio-temporal regions in a learned latent space.

Masking methodology

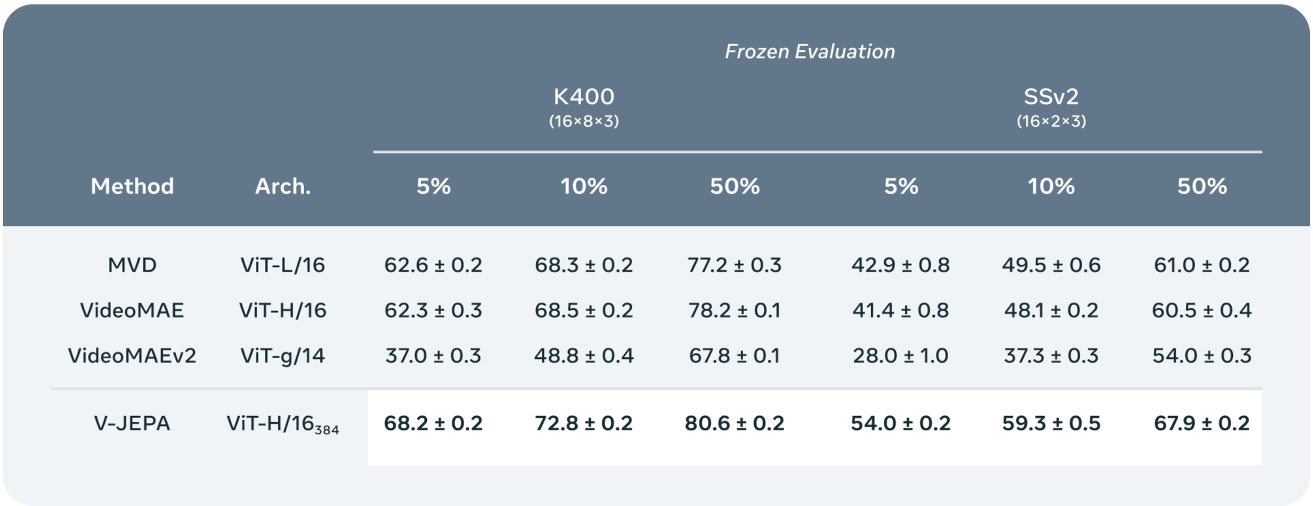
V-JEPA wasn't trained to understand one specific type of action. Instead it used self-supervised training on a range of videos and learned a number of things about how the world works. The team also carefully considered the masking strategy—if you don't block out large regions of the video and instead randomly sample patches here and there, it makes the task too easy and your model doesn't learn anything particularly complicated about the world.

It's also important to note that, in most videos, things evolve somewhat slowly over time. If you mask a portion of the video but only for a specific instant in time and the model can see what came immediately before and/or immediately after, it also makes things too easy and the model almost certainly won't learn anything interesting. As such, the team used an approach where it masked portions of the video in both space and time, which forces the model to learn and develop an understanding of the scene.

Efficient predictions

Making these predictions in the abstract representation space is important because it allows the model to focus on the higher-level conceptual information of what the video contains without worrying about the kind of details that are most often unimportant for downstream tasks. After all, if a video shows a tree, you're likely not concerned about the minute movements of each individual leaf.

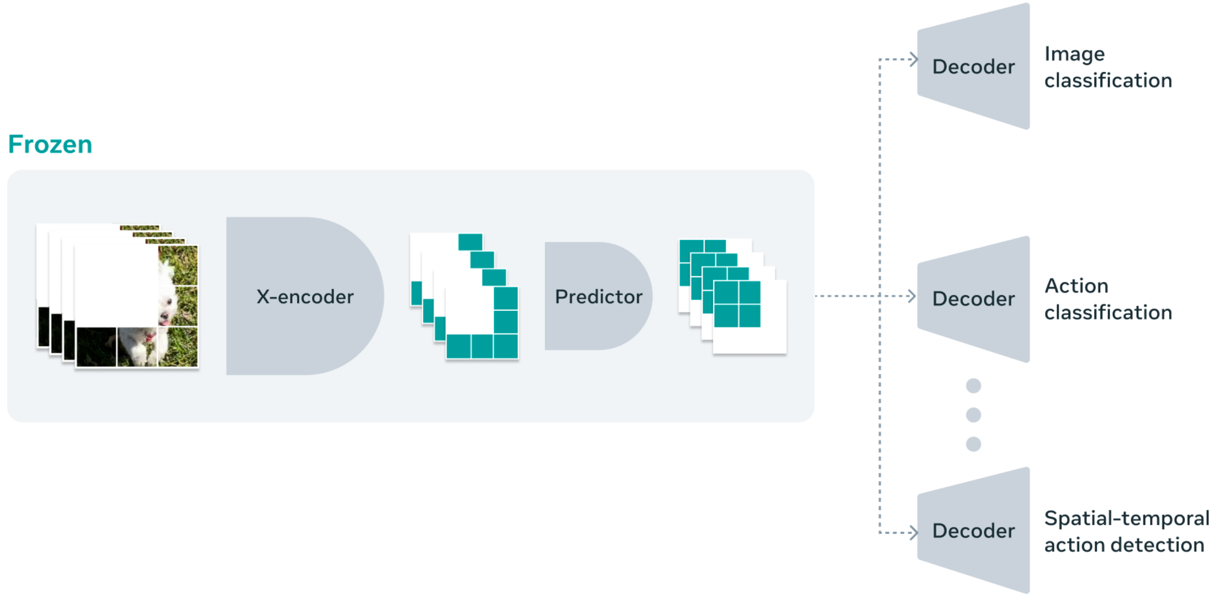
One of the reasons why we're excited about this direction is that V-JEPA is the first model for video that's good at "frozen evaluations," which means we do all of our self-supervised pre-training on the encoder and the predictor, and then we don't touch those parts of the model anymore. When we want to adapt them to learn a new skill, we just train a small lightweight specialized layer or a small network on top of that, which is very efficient and quick.



		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**

Low-shot frozen evaluation: Comparing V-JEPA to other video models in frozen evaluation on Kinetics-400 and Something-Something-v2 as we vary the percentage of labeled examples from each dataset available for training the attentive probe. We train the probes in several low-shot settings: using either 5%, 10%, or 50% of the train set, and take three random splits in each setting to obtain more robust metrics, resulting in nine different evaluation experiments for each model. We report the mean and standard deviation on the official validation K400 and SSv2 validation sets. V-JEPA is more label-efficient than other models—specifically, decreasing the available number of labeled examples from each class increases the performance gap between V-JEPA and the baselines.

Previous work had to do full fine-tuning, which means that after pre-training your model, when you want the model to get really good at fine-grained action recognition while you're adapting your model to take on that task, you have to update the parameters or the weights in all of your model. And then that model overall becomes specialized at doing that one task and it's not going to be good for anything else anymore. If you want to teach the model a different task, you have to use different data, and you have to specialize the entire model for this other task. With V-JEPA, as we've demonstrated in this work, we can pre-train the model once without any labeled data, fix that, and then reuse those same parts of the model for several different tasks, like action classification, recognition of fine-grained object interactions, and activity localization.



A self-supervised approach for learning representations from video, V-JEPA can be applied to various downstream image and video tasks without adaption of the model parameters. V-JEPA outperforms previous video representation learning approaches in frozen evaluation on image classification, action classification, and spatio-temporal action detection tasks.

Avenues for future research...

While the "V" in V-JEPA stands for "video," it only accounts for the visual content of videos thus far. A more multimodal approach is an obvious next step, so we're thinking carefully about incorporating audio along with the visuals.

As a proof of concept, the current V-JEPA model excels at fine-grained object interactions and distinguishing detailed object-to-object interactions that happen over time. For example, if the model needs to be able to distinguish between someone putting down a pen, picking up a pen, and pretending to put down a pen but not actually doing it, V-JEPA is quite good compared to previous methods for that high-grade action recognition task. However, those things work on relatively short time scales. If you show V-JEPA a video clip of a few seconds, maybe up to 10 seconds, it's great for that. So another important step for us is thinking about planning and the model's ability to make predictions over a longer time horizon.

...and the path toward AMI

To date, our work with V-JEPA has been primarily about perception—understanding the contents of various video streams in order to obtain some context about the world immediately surrounding us. The predictor in this Joint Embedding Predictive Architecture serves as an early physical world model: You don't have to see everything that's happening in the frame, and it can tell you conceptually what's happening there. As a next step, we want to show how we can use this kind of a predictor or world model for planning or sequential decision-making.

We know that it's possible to train JEPA models on video data without requiring strong supervision and that they can watch videos in the way an infant might—just observing the world passively, learning a lot of interesting things about how to understand the context of those videos in such a way that, with a small amount of labeled data, you can quickly acquire a new task and ability to recognize different actions.

V-JEPA is a research model, and we're exploring a number of future applications. For example, we expect that the context V-JEPA provides could be useful for our embodied AI work as well as our work to build a contextual AI assistant for future AR glasses. We firmly believe in the value of responsible open science, and that's why we're releasing the V-JEPA model under the CC BY-NC license so other researchers can extend this work.

Read the paper ↗

Get the code ↗

Share:

Our latest updates delivered to your inbox

Subscribe to our newsletter to keep up with Meta AI news, events, research breakthroughs, and more.

Join us in the pursuit of what's possible with AI.

↗ See all open positions

Related Posts

Computer Vision

Introducing Segment Anything: Working toward the first foundation model for image segmentation

April 5, 2023

→ Read post

FEATURED

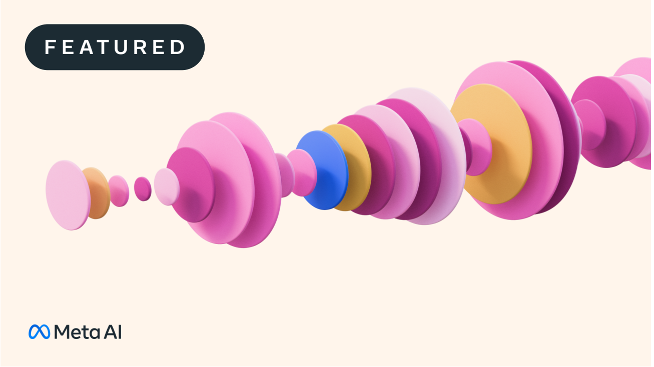
Meta AI

Research

MultiRay: Optimizing efficiency for large-scale AI models

November 18, 2022

Read post



ML Applications

MuAViC: The first audio-video speech translation benchmark

March 8, 2023

Read post

Search AI content

Our approach

Research

Product experiences

Latest news

Foundational models