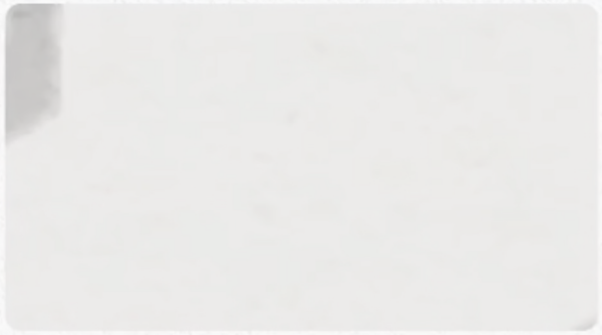
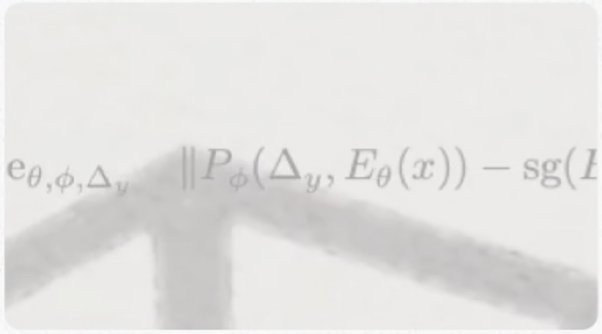
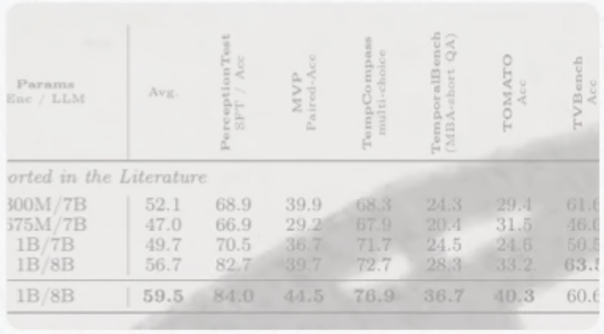


Introduction to World Models

V-JEPA 2

Yan Xu



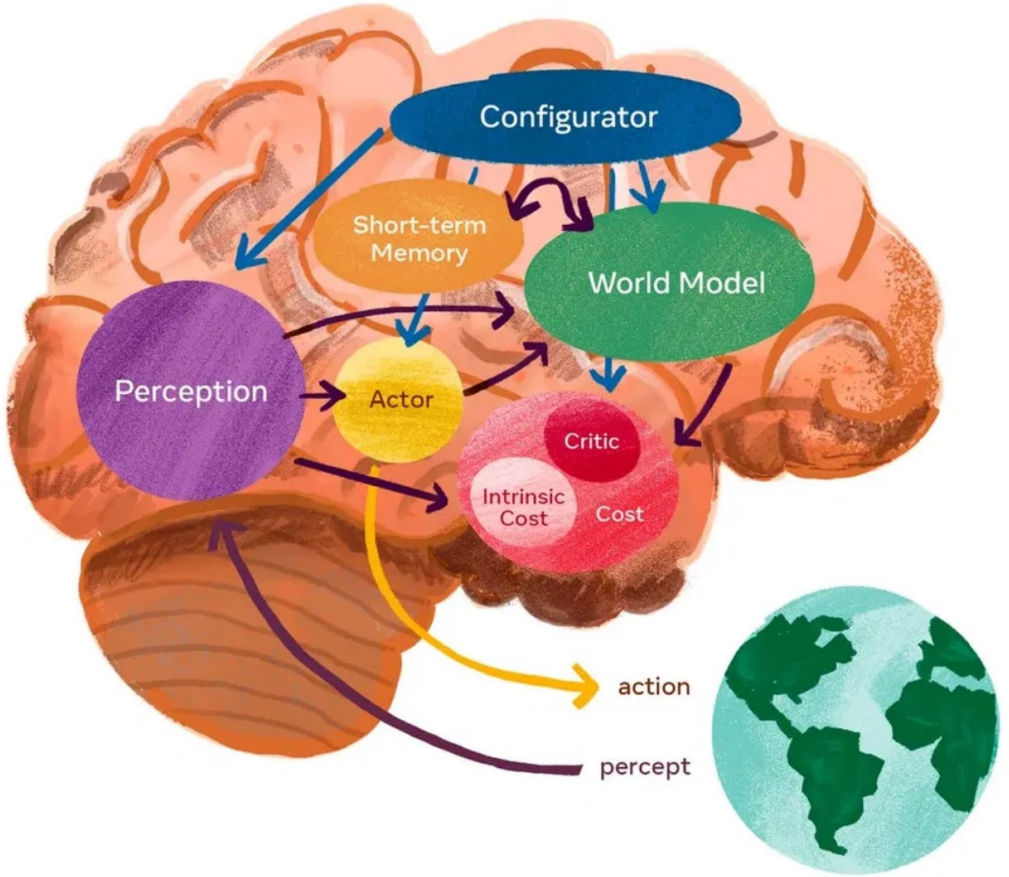


What are World Models

What are World Models: From Yann LeCun

What are World Models

A foundational model that understand physical reality, anticipate outcomes, and plan efficient strategies.



Developments of World Models

- Jan 2023: **Dreamer V3.** The algorithm uses an internal world model to plan and train an agent from simulated "dreams".
- Sep 2024: **World Labs.** World Labs is building "Large World Models" to enable spatial intelligence, allowing AI to perceive, generate, and interact with the 3D world.
- Mar 2025: **NVIDIA COSMOS.** The platform is designed to accelerate training for physical AI like robots and autonomous vehicles.
- Jun 2025: **Meta V-JEPA 2.** It enables "zero-shot" robotic planning, allowing an AI to perform new tasks without specific training.
- Aug 2025: **Google DeepMind Genie 3.** An interactive world model for generating real-time, persistent, and 3D environments from text or images.
- Sep 2025: **Meta Vision Language World Model.** A foundation model trained for language-based world modeling: Perceive the environment through visual observations and predict world evolution using language-based abstraction

Applications: From Image to 3D World

Input Image

3D World

Applications: From Text to 3D World

Prompt: Running by the shores of a glacial lake, exploring branching paths through the forest, crossing flowing mountain streams. Set amidst beautiful snow capped mountains and pine forest. Plentiful wildlife makes the journey a delight.

Prompt: Running by the shores of a glacial lake, exploring branching paths throu...

Applications: Robotics

Can we enable "zero-shot" robotic planning?

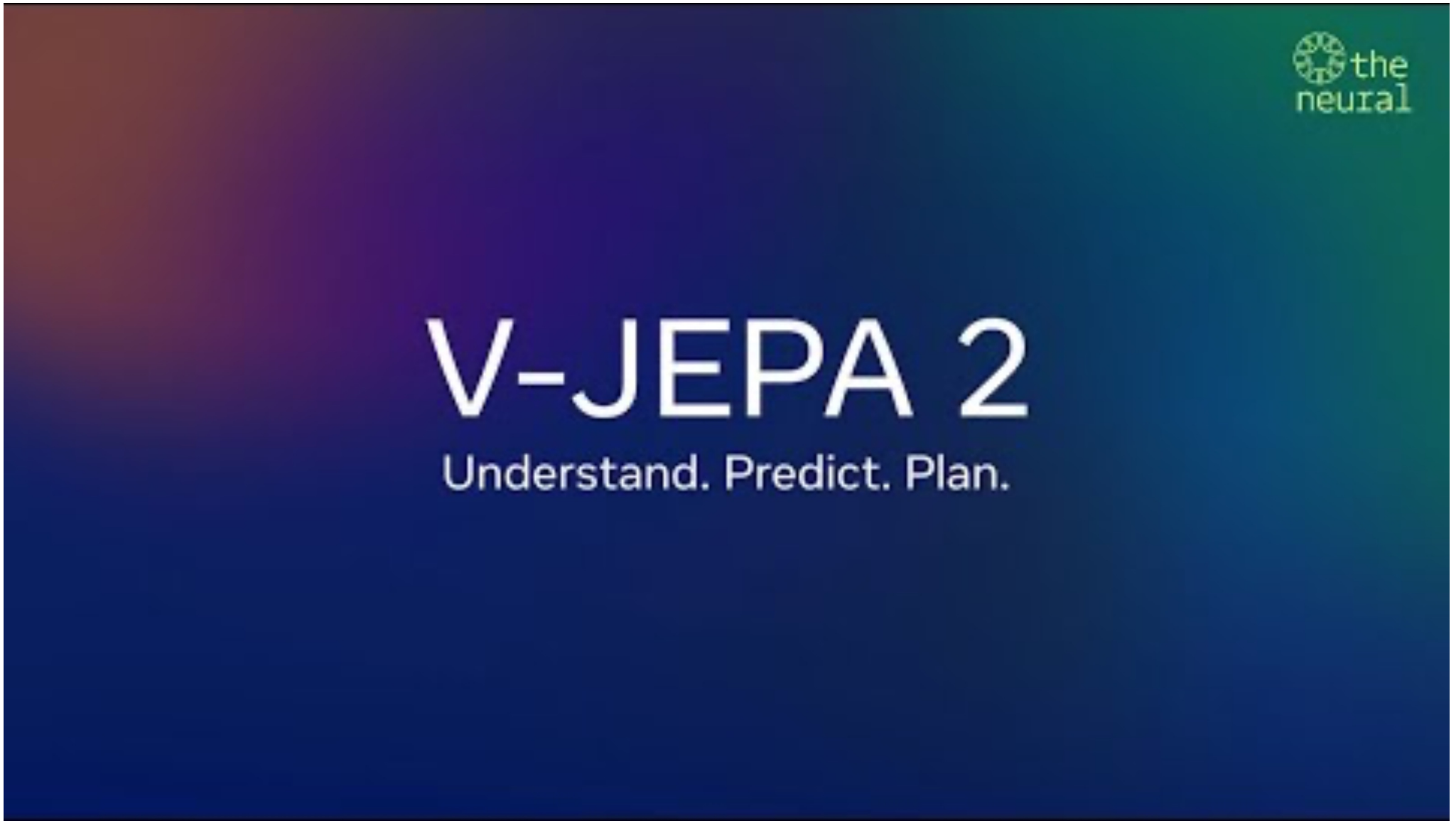
i.e.
plan and perform complex tasks in an unfamiliar environment without needing specific prior training for that task.

World Models vs. Language Models

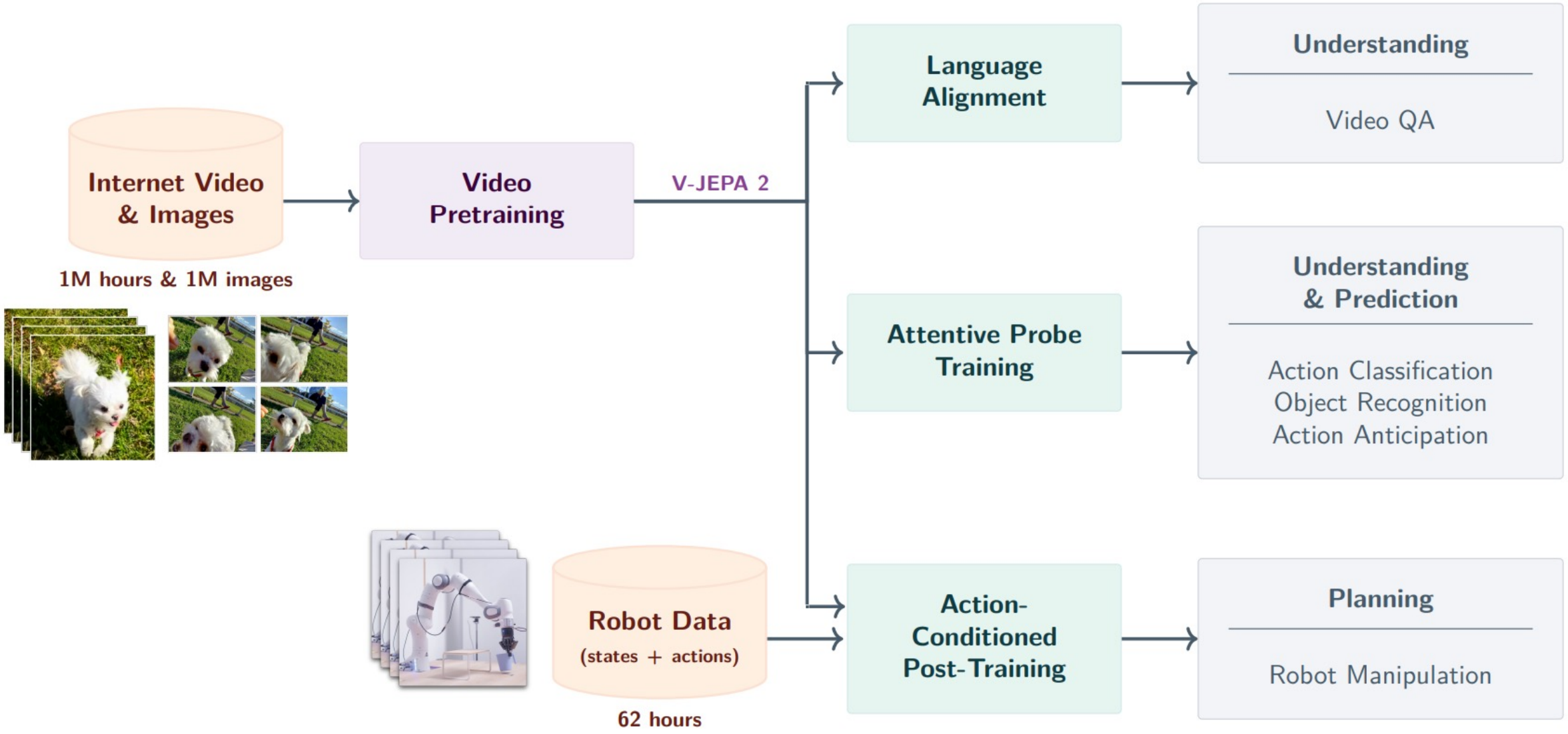
Aspect	Large Language Models (LLMs)	World Models
Primary Data	Textual corpora (e.g., web text, books)	Sensory data, simulations, and telemetry
Architecture	Transformers with self-attention	Hybrid architectures: encoders + latent dynamics
Objective	Predict next tokens	Predict environment states; support decision-making
Training Paradigm	Self-supervised learning on text	Self-supervised or reinforcement learning
Applications	NLP tasks: translation, summarization, QA	Robotics, control, simulation, model-based RL
Grounding	Statistical, linguistic	Physical, causal

Transformers!

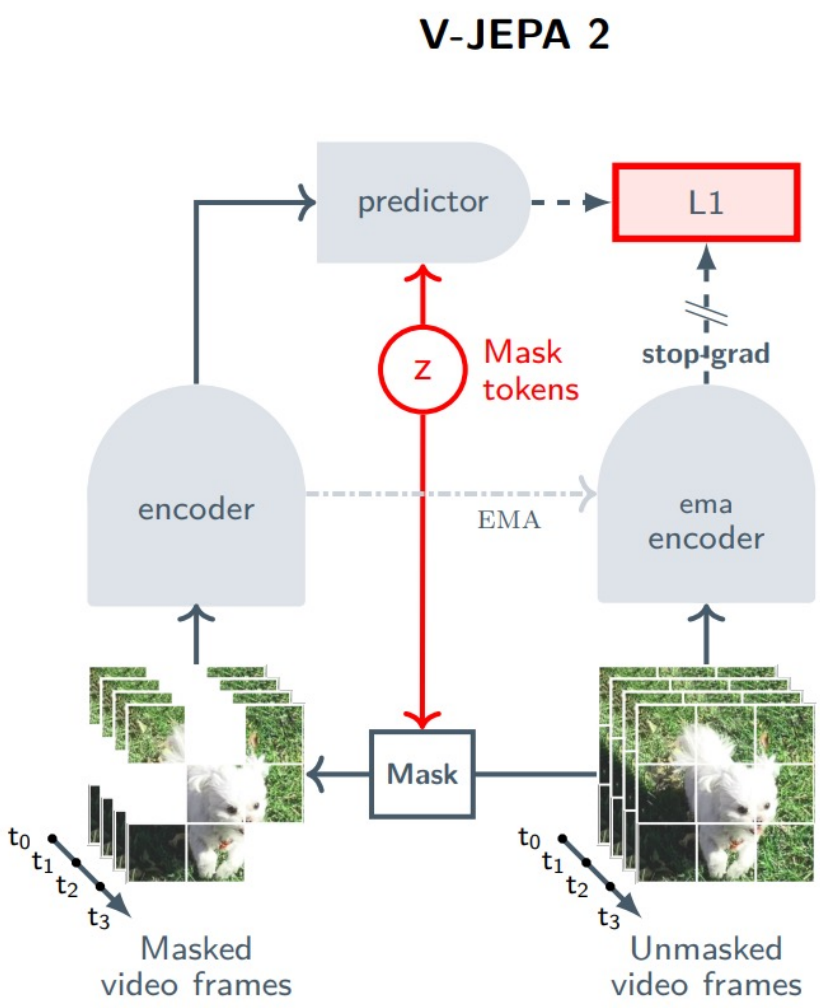
JERA: Joint-Embedding Predictive Architecture



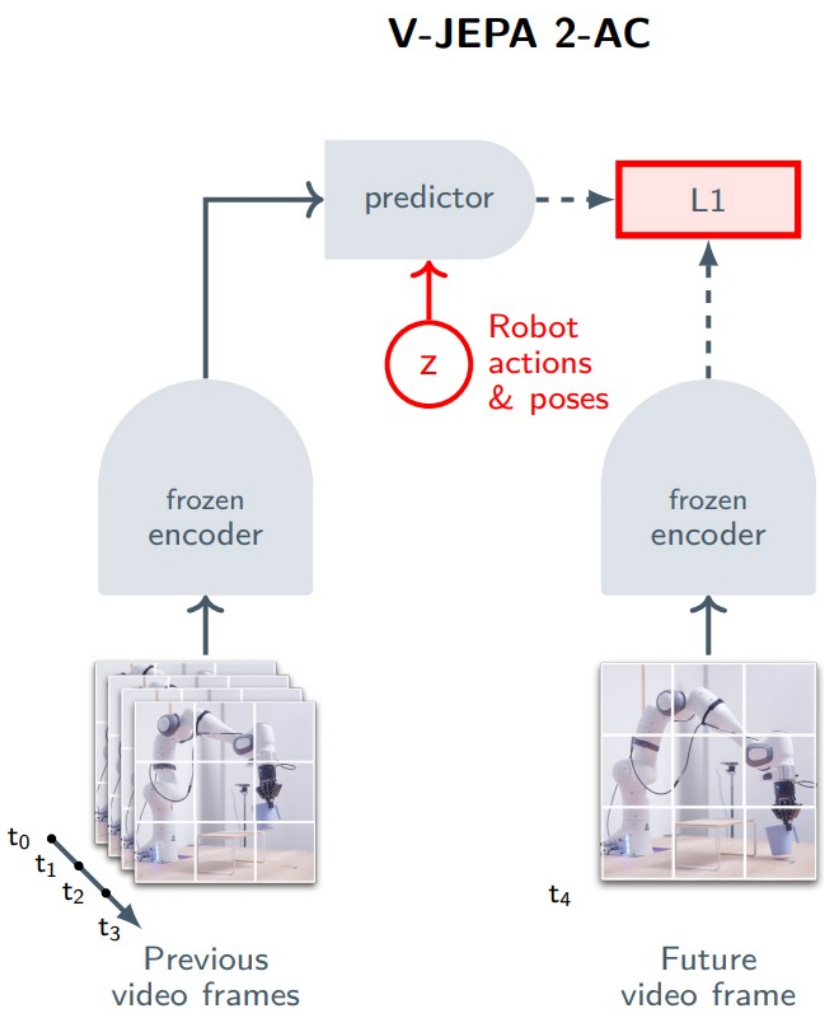
V-JEPA 2 Overview



Multi-stage training: Focus on Semantic Features instead of Pixels

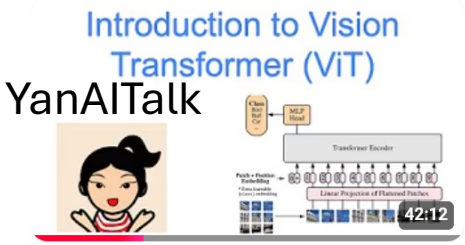


Self-supervised Video Pretraining



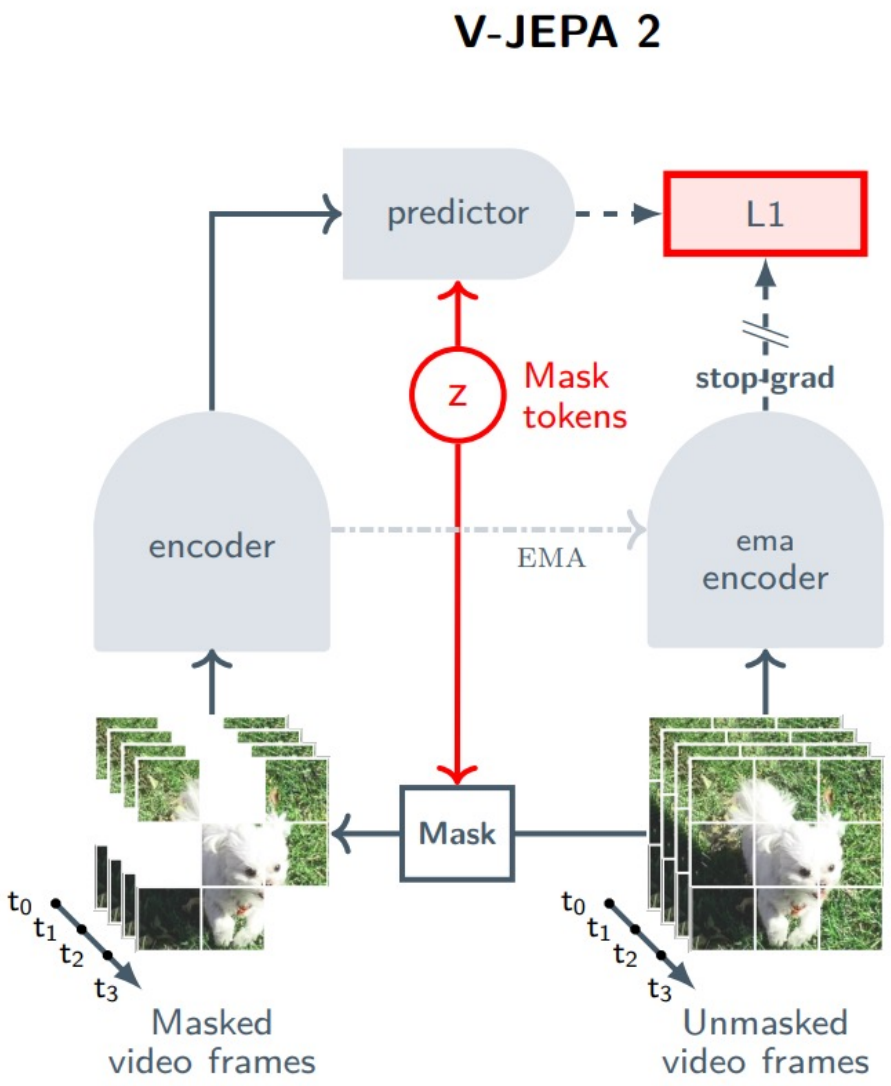
Action-conditioned predictor

Scaling Self-Supervised Video Pretraining



Transformer United: Introduction to Vision Transformer (ViT)

324 views • 1 year ago



- The encoder, and predictor, are each parameterized as a vision transformer (or ViT).
- To encode relative position information in the vision transformer, we leverage 3D-RoPE (Rotary Position Embedding)
- Mask prediction loss

$$\text{minimize}_{\theta,\phi,\Delta_y} \quad \|P_\phi(\Delta_y, E_\theta(x)) - \text{sg}(E_{\bar{\theta}}(y))\|_1,$$

Representation Prediction of masked video

Representation of original video

Scaling Self-Supervised Video Pretraining: Key Scaling Ingredients

1. **Data scaling:** We increase the dataset size from 2 million to 22 million videos

2. **Model scaling:** We scale the encoder architecture from 300 million to over 1 billion parameters

3. **Longer training:** Adopting a warmup-constant-decay learning rate schedule to extend training from 90 thousand up to 252 thousand iterations

4. **Higher resolution:** lower-resolution clips during the warmup and constant phases, and then increasing resolution and/or clip-length during the final decay phase.

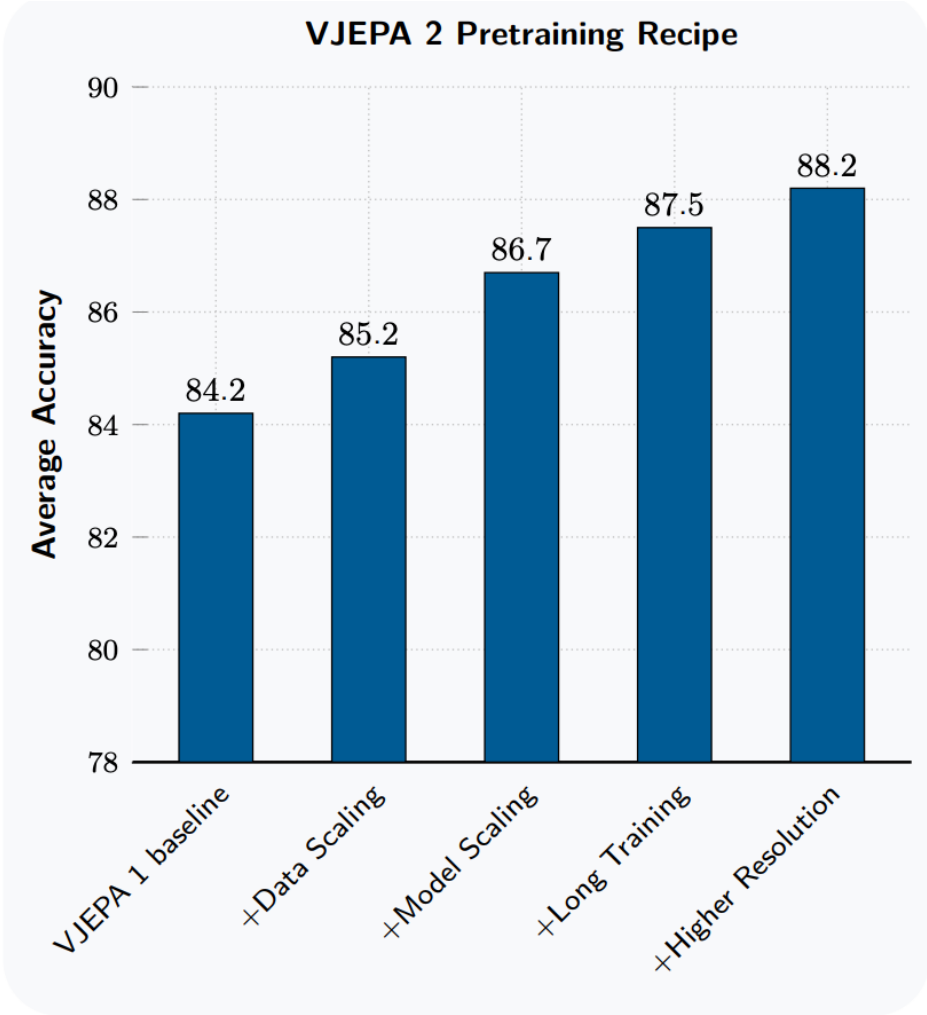


Figure 3 Scaling Ingredients. The effects of

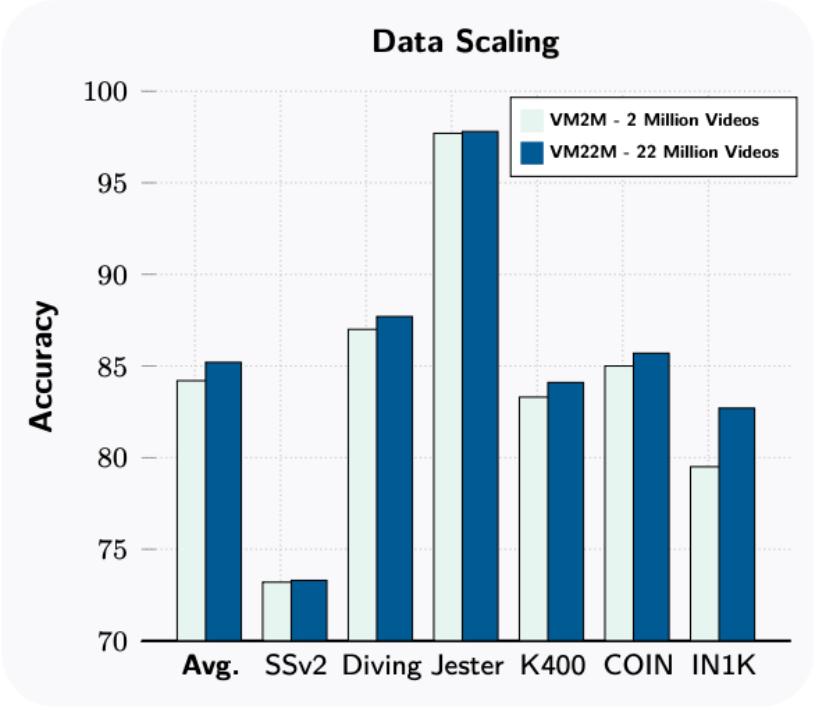
Scaling Self-Supervised Video Pretraining: Pretraining Dataset

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

- Something-Something v2 (SSv2): An ego-centric (first-person perspective) video dataset.
- Kinetics (Kinetics 400, 600, and 700): exo-centric (third-person perspective) action videos.
- HowTo100M: YouTube tutorial videos.
- YT-Temporal-1B (YT1B): General YouTube videos, which underwent a curation process.

Scaling Self-Supervised Video Pretraining: Data Scaling



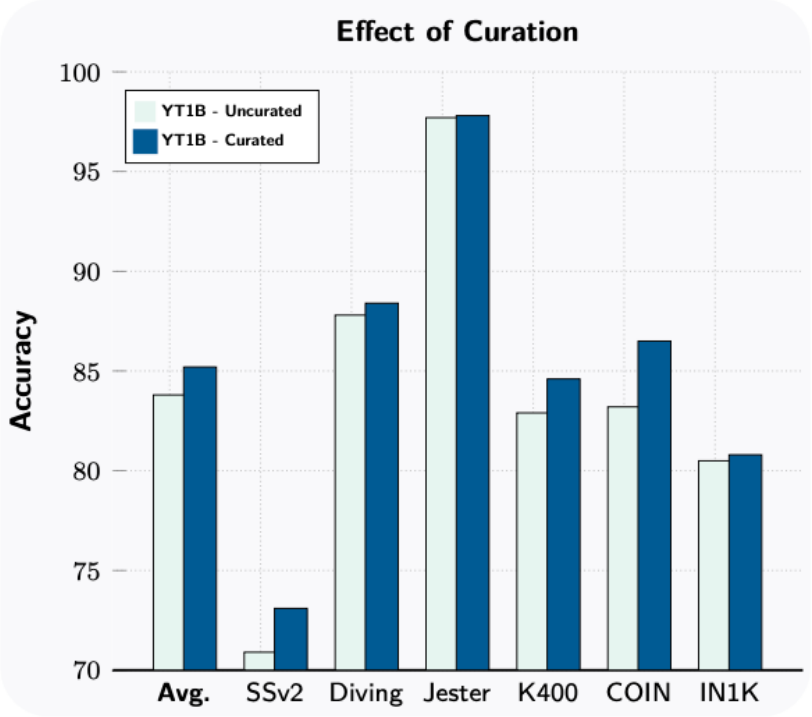
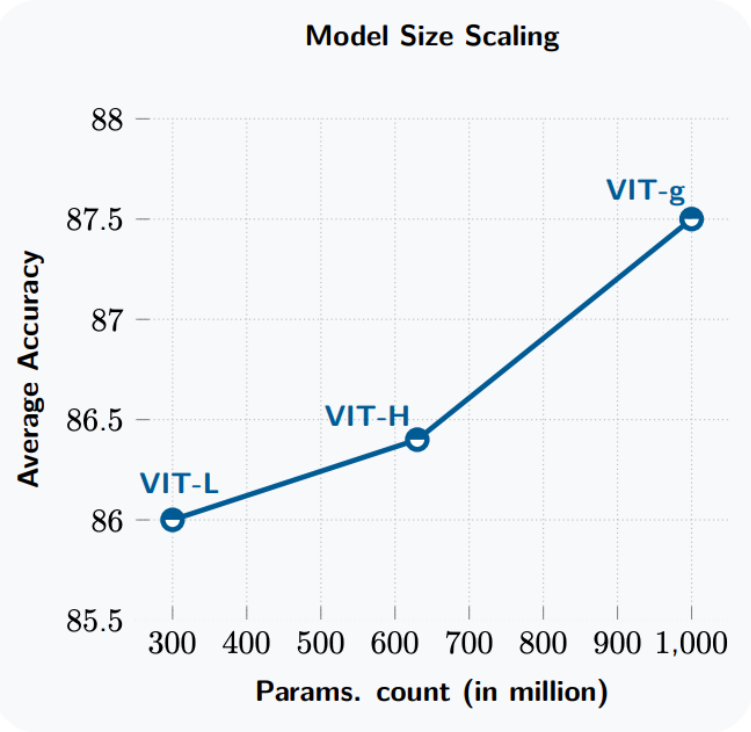
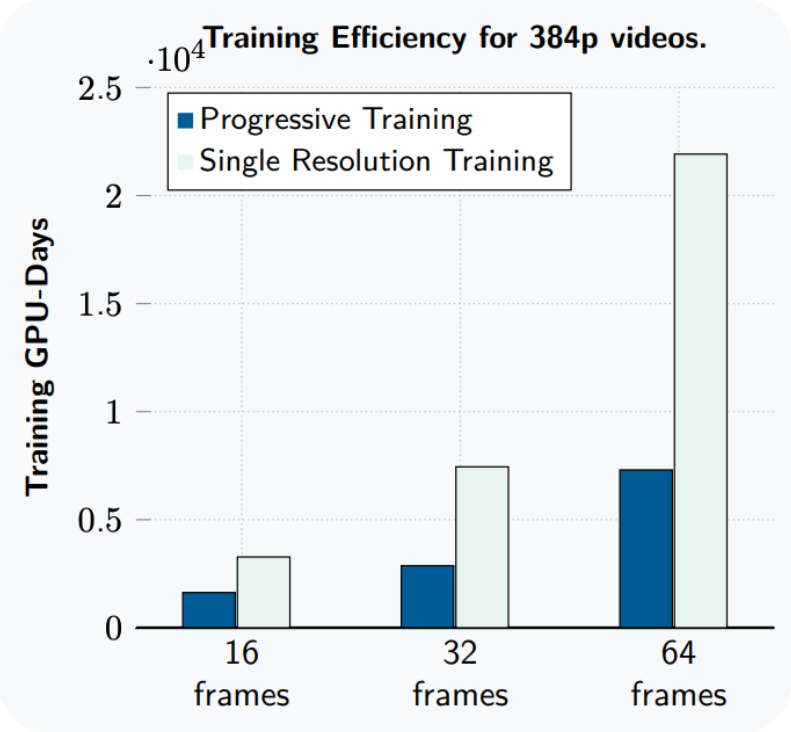
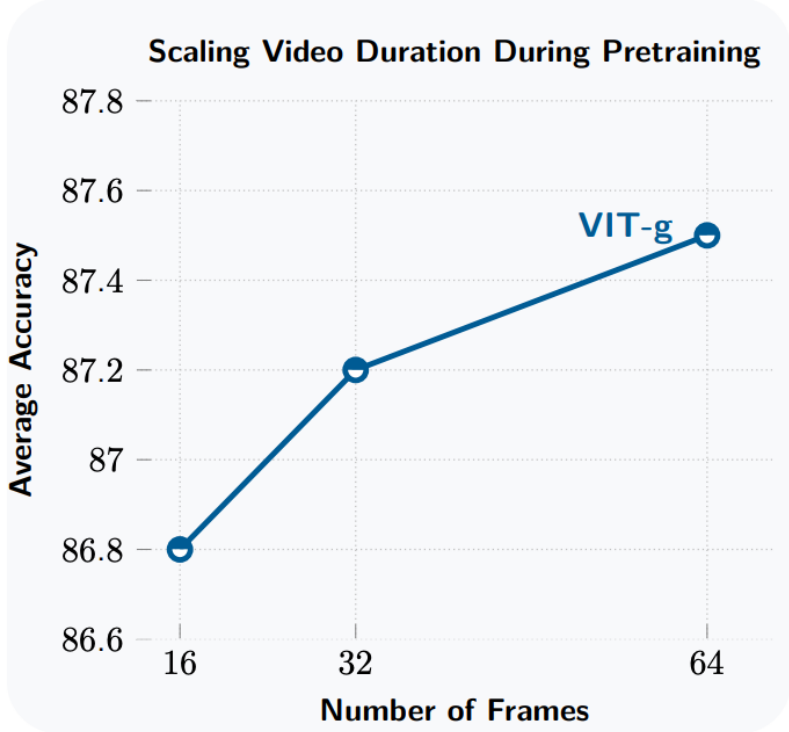


Figure 4 Data Scaling & Curation. We train and compare models on different data-mixes. Models are ViT-L/16 trained for 90K iterations using a cosine learning schedule following Bardes et al. (2024). **(Left)** We compare the

Scaling Self-Supervised Video Pretraining: Model Scaling





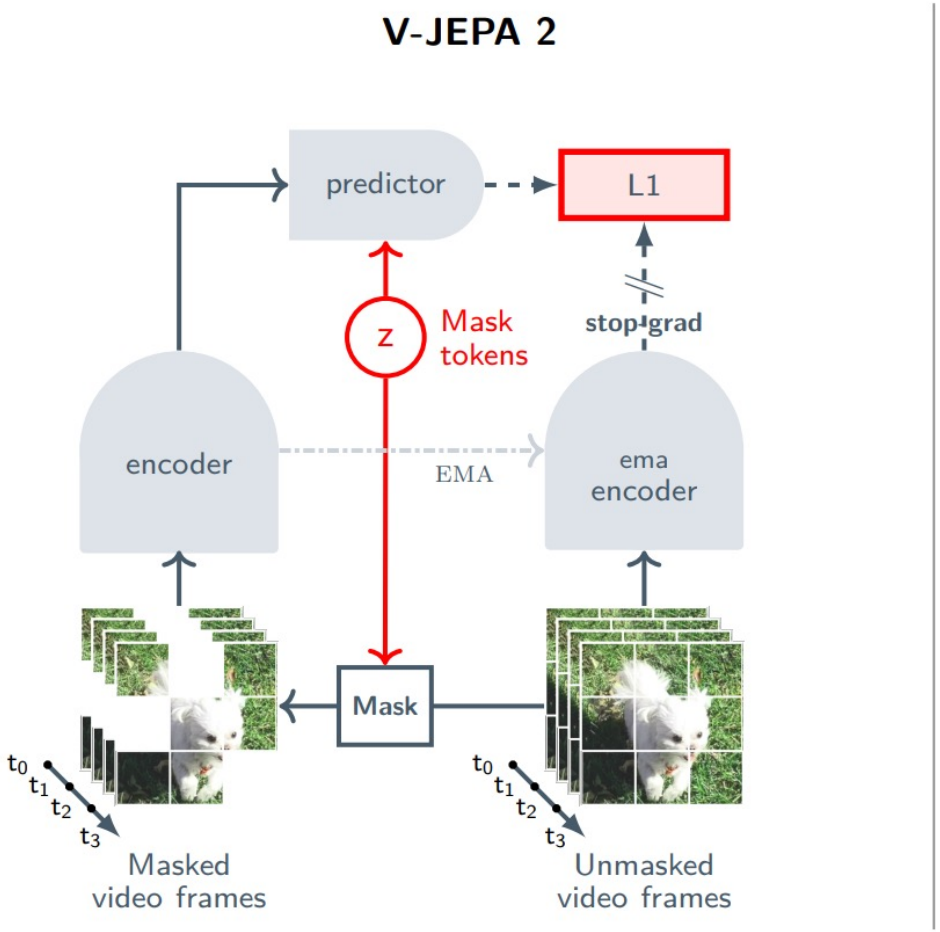


Progressive Training:

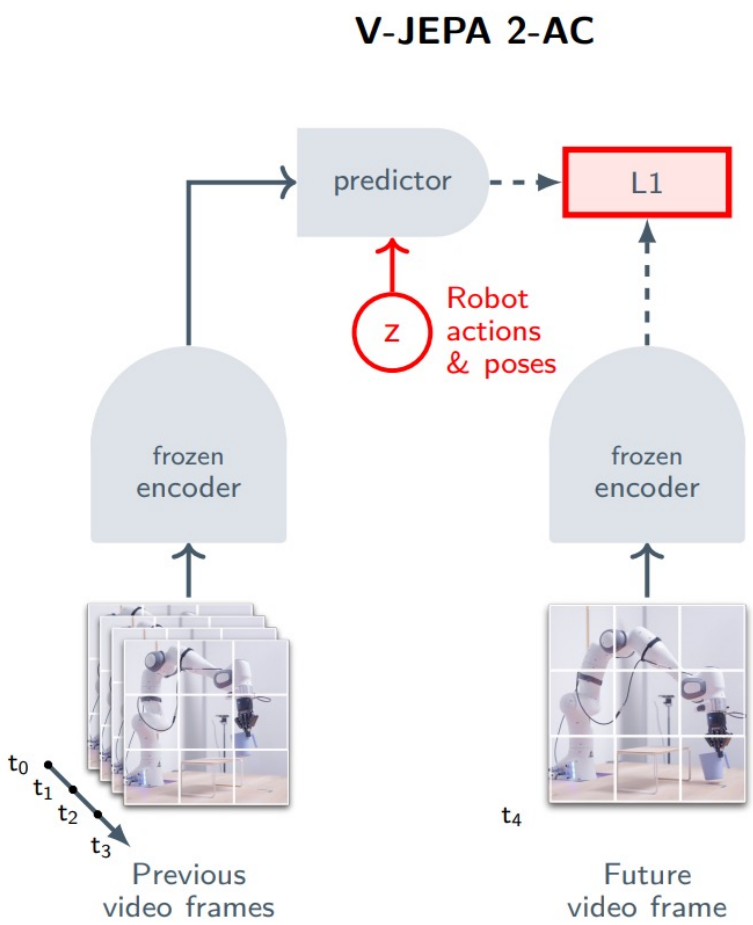
Warmup and Main Training: begins its pre-training with shorter video clips (e.g., 16 frames) at a lower spatial resolution (e.g., 256x256 pixels)

Cooldown Phase: Linear decay of the learning rate and an increase in input video resolution and/or duration.

Multi-stage training



Self-supervised Video Pretraining

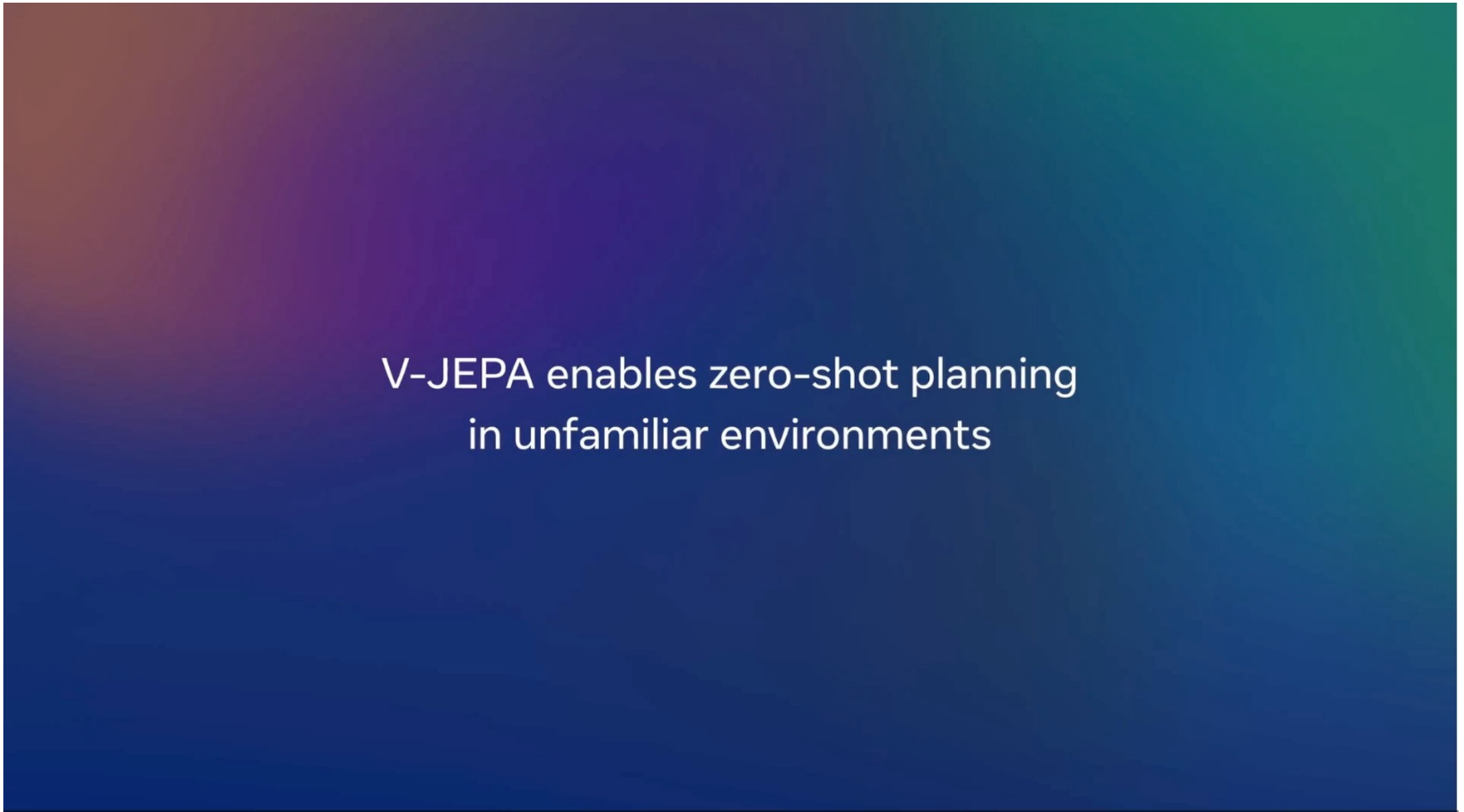


Action-conditioned predictor

Learning an Action-Conditioned World Model

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

Inferring Actions by Planning



Planning: Zero-shot Robot Control

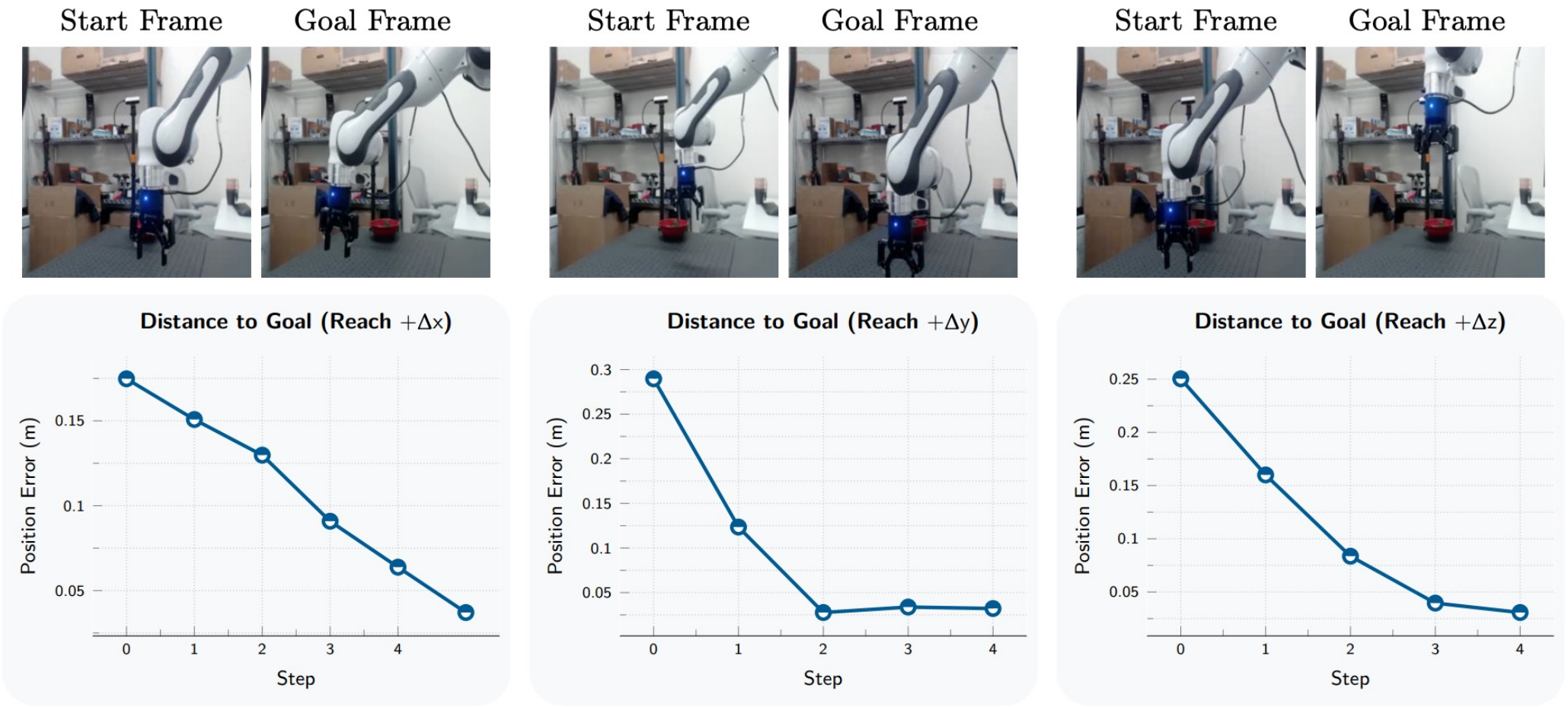


Figure 8 Single-Goal Reaching. Single-goal reaching involves moving the end-effector to a desired location in space based on a single goal image. This task measures for a basic understanding of actions as well as a 3D spatial

Pick & Place

Figure 10 Pick-&-Place. Closed-loop robot execution of V-JEPA 2-AC for multi-goal pick-&-place tasks. Highlighted frames indicate when the model achieves a sub-goal and switches to the next goal. The first goal image shows the

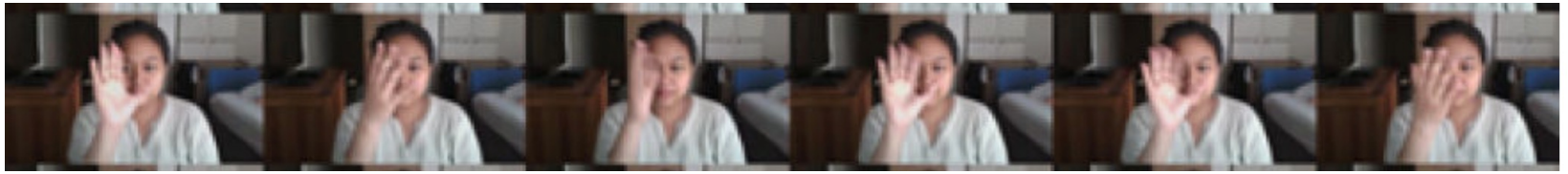
Pick & Place

Table 2 Zero-Shot Robot Manipulation. All models are deployed zero-shot on two Franka arms with RobotiQ grippers located in different labs. Given image-goals for each considered task, all models run closed loop to infer a sequence of actions to achieve the goal. Success rates are reported out of 10 trials with various permutations to the task across trials (e.g., object location, starting pose, etc.).

			Grasp		**Reach w/ Obj.**		**Pick-&-Place**	
Method		**Reach**	Cup	Box	Cup	Box	Cup	Box
Octo (Octo Model Team et al., 2024)	Lab 1	100%	20%	0%	20%	70%	20%	10%
	Lab 2	100%	10%	0%	10%	70%	10%	10%
	Avg	*100%*	*15%*	*0%*	*15%*	*70%*	*15%*	*10%*
V-JEPA 2-AC (ours)	Lab 1	100%	70%	30%	90%	80%	80%	80%
	Lab 2	100%	60%	20%	60%	70%	80%	50%
	Avg	*100%*	*65%*	*25%*	*75%*	*75%*	*80%*	*65%*

Understanding: Action & Object Classification

Jester sample:



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
$PE_{core}G$ (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
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 ViT-g$_{384}$	1B	**88.2**	**77.3**	**90.2**	**97.8**	87.3	91.1	85.1

Prediction: Action Anticipation

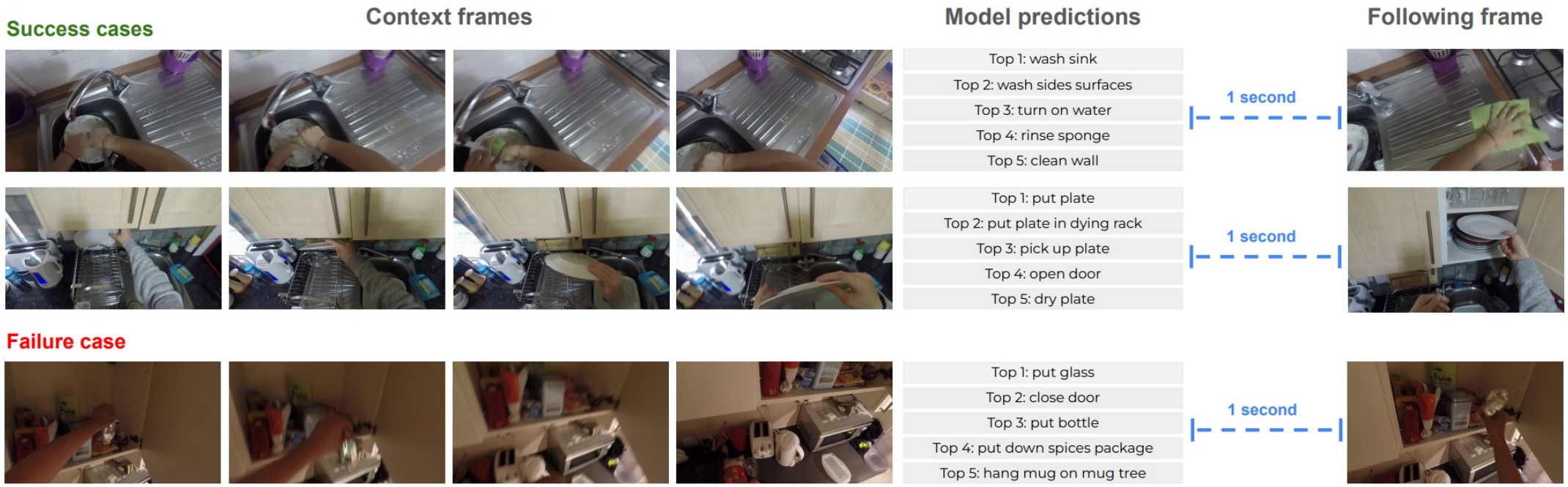


Figure 11 Visualization of EK100 prediction. (Left): four selected frames from the context frames. (Middle): model predictions, ordered by likelihood. (Right): following frame after the 1 second anticipation time. We show two

Understanding: Video Question Answering

MVBench: A Comprehensive Multi-modal Video Understanding Benchmark

Spatial	Temporal	Source	Example
Action	Action Sequence	STAR	*What happened after the person took the food?* (A) Ate the medicine. (B) Tidied up the blanket. (C) Put down the cup/glass/bottle. (D) Took the box.
	Action Prediction	STAR	*What will the person do next?* (A) Put down the pillow. (B) Open the door. (C) Take the book. (D) Open the closet/cabinet.
	Action Antonym	PAXION‡	*Which one of these descriptions correctly matches the actions in the video?* (A) not sure (B) scattering something down (C) piling something up
	Fine-grained Action	MiT V1‡	*What is the action performed by the person in the video?* (A) watering (B) leaking (C) pouring (D) planting
	Unexpected Action	FunQA‡	*What unexpected event contributes to the humor in the video?* (A) The man left without dancing. (B) Two women hugged each other at the end. (C) The man finally danced with the woman. (D) Two men hugged each other unexpectedly.
Object	Object Existence	CLEVRER	*Are there any moving green objects when the video ends?* (A) not sure (B) yes (C) no
	Object Interaction	STAR	*Which object was tidied up by the person?* (A) broom (B) cabinet (C) blanket (D) table
	Object Shuffle	Perception Test	*Where is the hidden object at the end of the game from the person's point of view?* (A) Under the first object from the left. (B) Under the third object from the left. (C) Under the second object from the left.
Position	Moving Direction	CLEVRER‡	*What direction is the cyan sphere moving within the video?* (A) The object is stationary. (B) Up and to the right. (C) Down and to the left. (D) Down and to the right.
	Action Localization	Charades-STA ‡	*During which part of the video does the action 'person sitting on a couch' occur?* (A) In the middle of the video. (B) At the end of the video. (C) Throughout the entire video. (D) At the beginning of the video.
Scene	Scene Transition	MoVQA‡	*What's the right option for how the scenes in the video change?* (A) From the reception desk to the conference room. (B) From the kitchen to the dining area. (C) From the server room to the control center. (D) From the classroom to the library.
Count	Action Count	Perception Test	*How many times did the person launch objects on the table?* (A) 3 (B) 2 (C) 4
	Moving Count	CLEVRER	*How many metal objects exit the scene?* (A) 2 (B) 3 (C) 1 (D) 0

Understanding: Video Question Answering

Table 6 Comparison between off-the-shelf image encoders and V-JEPA 2 in frozen encoder setting. All experiments use the same LLM backbone (Qwen2-7B-Instruct), data, and training setup with a **frozen** vision encoder. PerceptionTest accuracy is reported on the validation set post SFT.

Method	Params Enc / LLM	Avg.	PerceptionTest SFT / Acc	MVP Paired-Acc	TempCompass multi-choice	TemporalBench (MBA-short QA)	TVBench Acc	TOMATO Acc	MVBench Acc
Off-the-shelf image encoders									
DINOv2 ViT-g$_{518}$	1.1B/7B	45.7	67.1	22.4	62.3	26.8	47.6	32.0	61.8
SigLIP2 ViT-g$_{384}$	1.1B/7B	48.1	**72.4**	26.2	66.8	25.7	48.7	33.2	64.0
PE ViT-G/14$_{448}$	1.9B/7B	49.1	72.3	26.7	67.0	27.5	51.6	34.0	64.7
V-JEPA 2 ViT-g$_{512}$	1B/7B	**52.3**	72.0	**31.1**	**69.2**	**33.3**	**55.9**	**37.0**	**67.7**

Understanding: Video Question Answering

Table 8 Comparison with state-of-the-art. We use the full 88.5M-sample alignment dataset and train using the same methodology as PLM 8B Cho et al. (2025), using a Llama 3.1 backbone. We observe significant improvements in downstream evaluations, obtaining state-of-the-art results in the 8B model class. PerceptionTest accuracy is reported on the test set with SFT for V-JEPA 2; all other results are zero-shot.

Method	Params Enc / LLM	Avg.	PerceptionTest Test Acc	MVP Paired-Acc	TempCompass multi-choice	TemporalBench (MBA-short QA)	TOMATO Acc	TVBench Acc	MVBench Acc
$\leq 8B$ Video Language Models Results Reported in the Literature									
InternVL-2.5 (Chen et al., 2024)	300M/7B	52.1	68.9	39.9	68.3	24.3	29.4	61.6	72.6
Qwen2VL (Wang et al., 2024a)	675M/7B	47.0	66.9	29.2	67.9	20.4	31.5	46.0	67.0
Qwen2.5VL (Qwen Team et al., 2025)	1B/7B	49.7	70.5	36.7	71.7	24.5	24.6	50.5	69.6
PLM 8B (Cho et al., 2025)	1B/8B	56.7	82.7	39.7	72.7	28.3	33.2	**63.5**	**77.1**
V-JEPA 2 ViT-g$_{384}$ LLama 3.1 8B	1B/8B	**59.5**	**84.0**	**44.5**	**76.9**	**36.7**	**40.3**	60.6	73.5

Future work

- **Longer-horizon tasks**: Extend this work to longer-horizon tasks such as pick-and-place or even more complex tasks, without requiring sub-goals will require further innovations in modeling

- **Language-based goal**: V-JEPA 2-AC currently relies upon tasks specified as image goals. Although this may be natural for some tasks, there are other situations where language-based goal specification may be preferable.

- **Scaling up models**: V-JEPA 2 models is scaled up to a modest 1B parameters. The results in model scaling demonstrated consistent performance improvements. Previous work has investigated scaling vision encoders to as large as 20B parameters

Conclusion: The Future with World Models

World models represent a fundamental shift in how AI systems understand and interact with their environment:

The path to more capable AI systems may depend on our ability to create increasingly sophisticated world models that can **Understand, Predict, and Plan** about the physical world.

Upcoming meetups – Open to proposals and guest speakers!

- Agentic AI real-world use cases
- Latest trends in AI research
- Build AI agents: Hands-on sessions
- Panel discussions

Slides posted at:
https://github.com/YanXuHappygela/LLM-reading-group

Recordings posted at:

YanAITalk

@yanaitalk · 3.41K subscribers · 70 videos

Make machine learning easy to understand! ...