

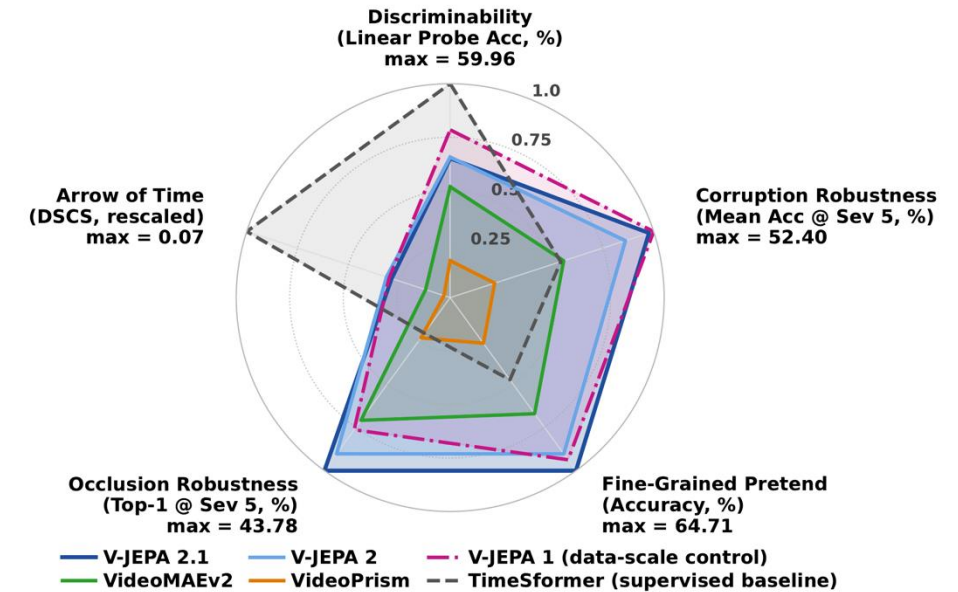
Latent Video Prediction for World Modeling

An Evaluation Uncovering Intriguing Favorable Evidence

Ali J Alrasheed¹ · Aryan Yazdan Parast¹ · Basim Azam¹ · James Bailey² ·
Naveed Akhtar¹

¹The University of Melbourne ²Monash University

ali.alrasheed@student.unimelb.edu.au



Five robustness axes, one consistent profile: latent prediction

Which video encoder makes a robust world model?

Current benchmarks test clean video. A world model must cope with real-world scenarios.

3 objectives · 4 frozen ViT-L models

Pixel reconstruction

rebuilds the raw pixels

VideoMAEv2

Contrastive + masked

pulls similar views together

VideoPrism

Latent prediction

predicts in feature space

V-JEPA 2 · 2.1

6 tests

Discriminability

clean features

Corruption

noise, blur, snow

Occlusion

missing patches,
frames

Pretend vs real

no-contact cues

Arrow of time

video played
backwards

Planning

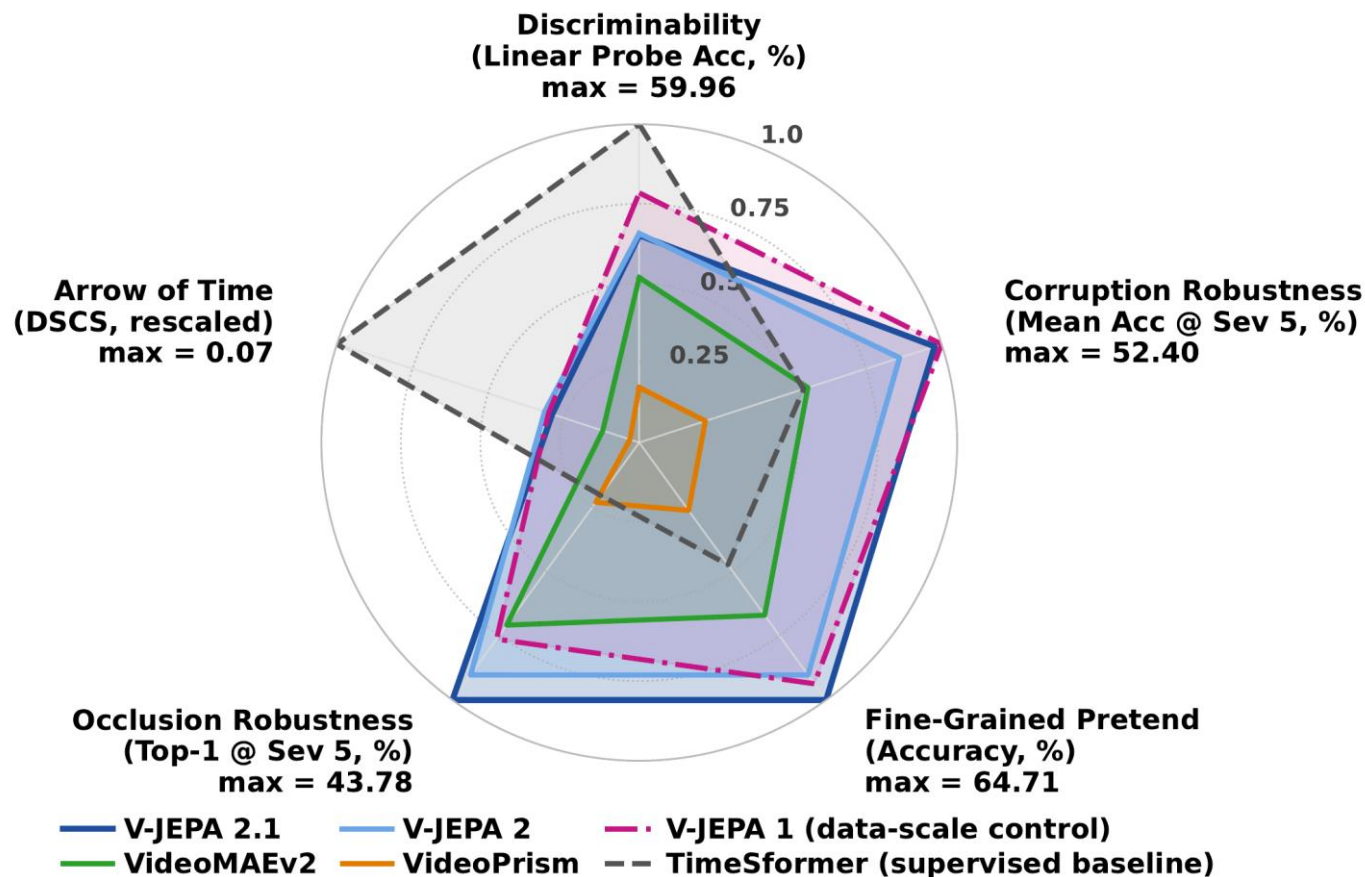
robot tasks, Meta-
World

Something-Something v2, repeated on EGTEA Gaze+

Does the pretraining objective change how robust the world model is?

Latent prediction is the most balanced

Among self-supervised encoders, V-JEPA traces the largest, most even profile

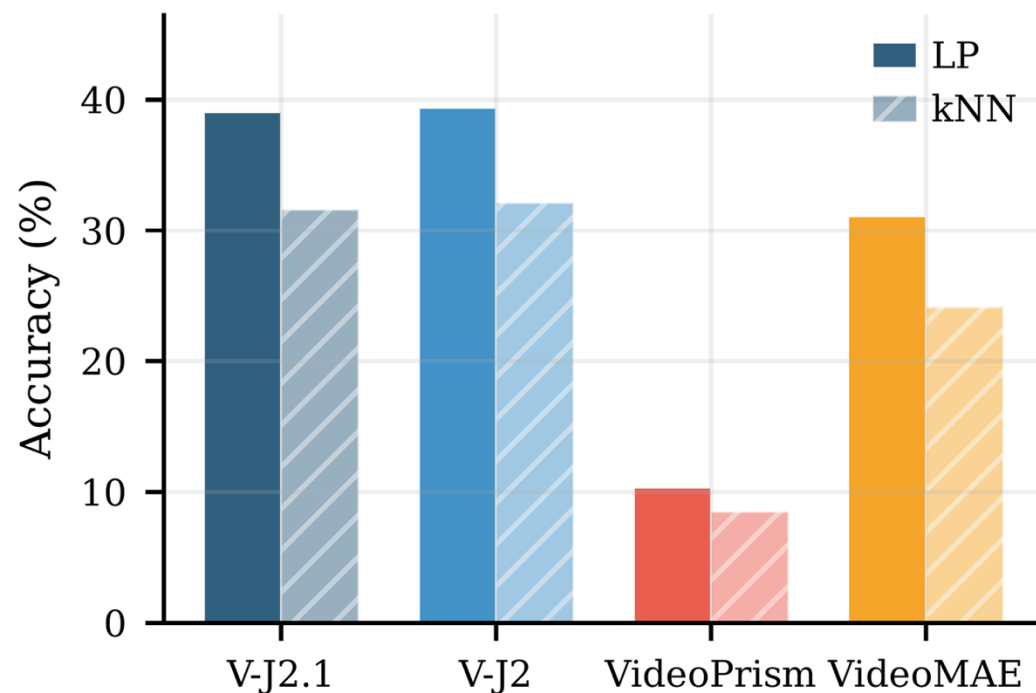


Reading the radar

- Each axis is normalised; raw maxima are annotated
- **V-JEPA 2.1 / 2** trace the largest, most even shape
- **VideoPrism** collapses on almost every robustness axis
- **TimeSformer** (supervised) is strong on clean cues but weak under corruption and occlusion
- **V-JEPA 1** $\approx$ V-JEPA 2 with $\sim 11\times$ less data $\rightarrow$ objective, not scale

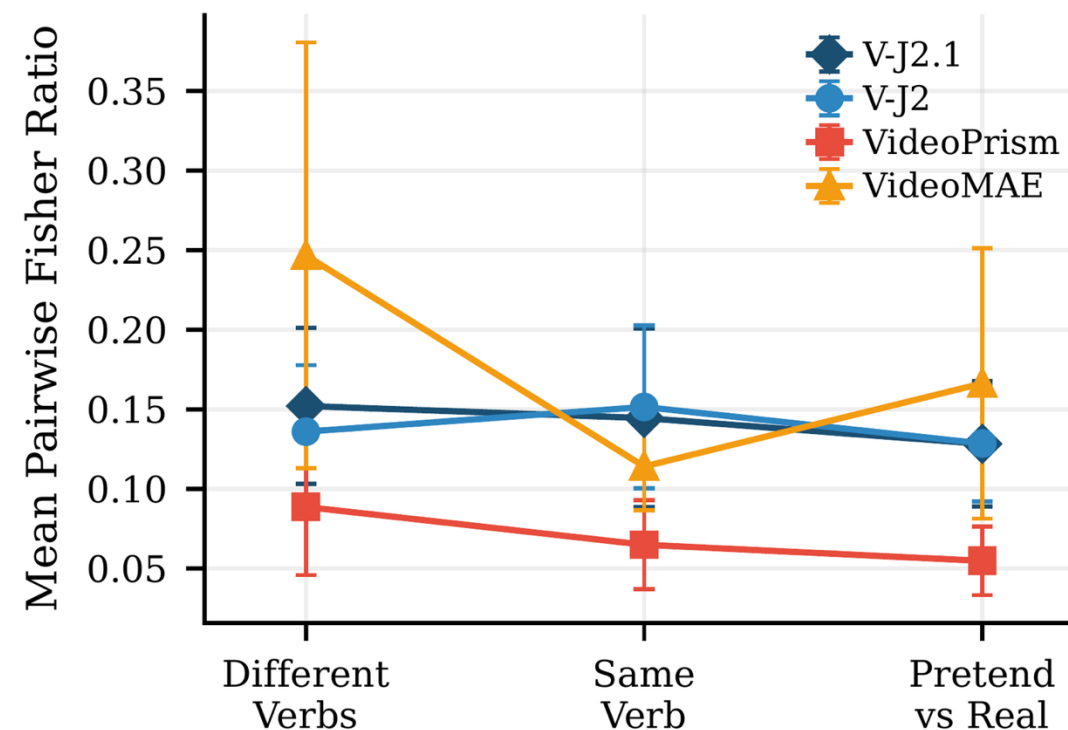
Discriminability: are features separable?

V-JEPA features are the most linearly separable, consistently across difficulty



Higher probe accuracy

- Both V-JEPA variants beat VideoMAEv2 and VideoPrism on 30 stratified classes



Spread $\neq$ separability

- VideoMAEv2 has the highest Fisher ratio but drops on same-verb pairs (Harder actions discrimination)

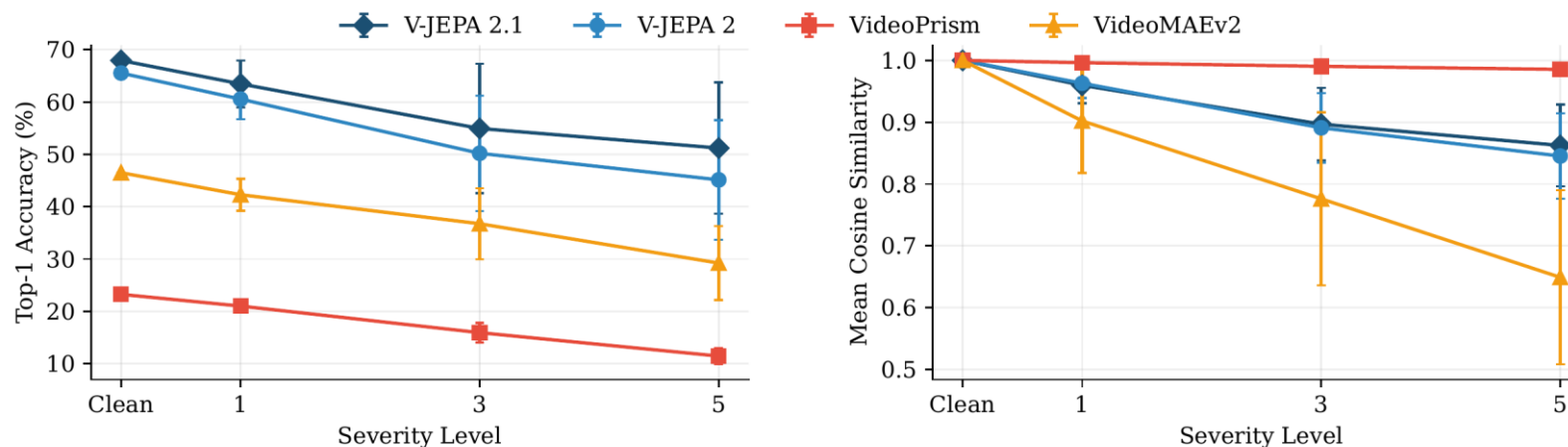
Corruption: meaning over appearance?

Under pixel corruption, V-JEPA 2.1 leads on 5 of 6 types and degrades the slowest



Corrupted examples at the highest severity (5)

Pixel Corruption Robustness by Severity



Why?

- A latent loss gives **no gradient** to preserve noise or brightness
- A pixel loss spends capacity on **texture & lighting**, which corruption destroys
- Right panel: VideoPrism stays most **similar** yet is least **accurate**

Fine-grained: pretend vs. real actions

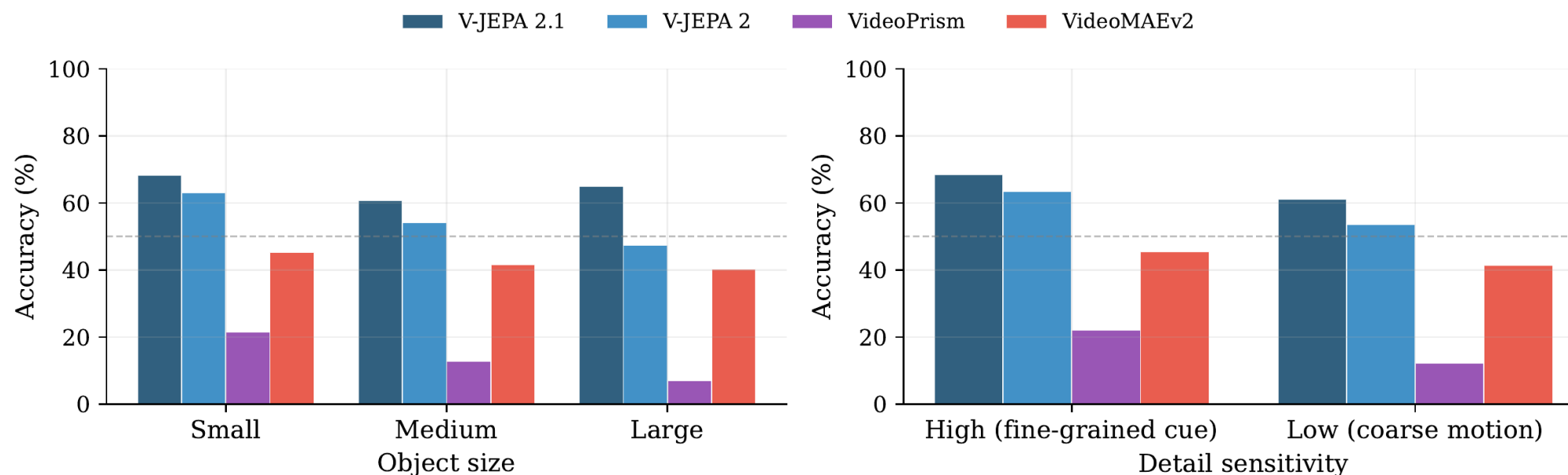
No pixels reconstructed, yet V-JEPA catches the subtle cue: no physical contact

Pouring something into something

vs

Pretending to pour something into something

1,992 videos · 22 classes · same gestures,
viewpoint, objects



Expected: pixel reconstruction should win here

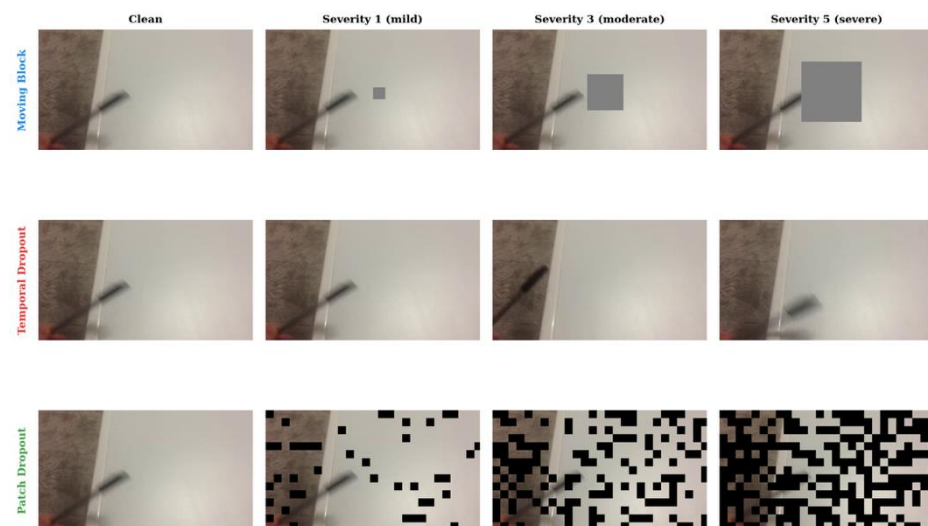
- Fine contact detail is exactly what pixel targets preserve...

Observed: the opposite

- V-JEPA wins nearly every class; largest gap on high-sensitivity ones

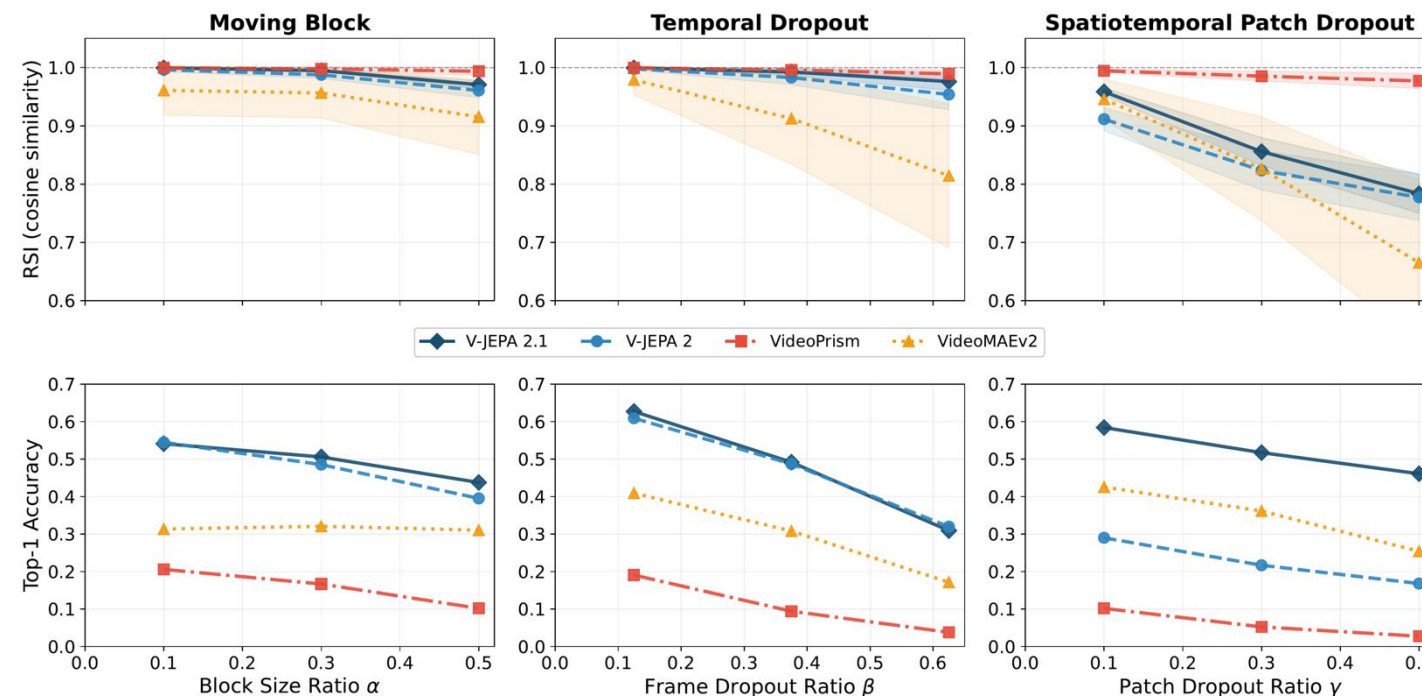
Occlusion: when the input goes missing

Same masking pretext, different target: latent targets generalise to missing input



Three paradigms × three severities

- **Moving block** — spatial occluder
- **Temporal dropout** — frozen frames
- **Patch dropout** — random 3D cuboids, mirroring V-JEPA's own masking



V-JEPA and VideoMAEv2 both pretrain with masking. The difference is the prediction target: **latent** vs. **pixels**. V-JEPA 2.1 has the top accuracy at every severity.

Surprise: stable features $\neq$ usable features

Severe spatiotemporal patch dropout: a stable embedding can still lose the class

VideoPrism

cosine similarity

≈ 0.98

top-1 accuracy

2.7%

Geometrically stable, semantically lost

V-JEPA 2.1

cosine similarity

shifts more

top-1 accuracy

46.1%

Moves, but keeps what matters

- **Cosine similarity alone is a misleading measure of robustness.**
- **Why:** contrastive alignment pulls all views of a concept together, compressing the intra-class detail that separates similar actions. V-JEPA does not force invariance; it keeps whatever helps predict missing content.

Do video models encode the arrow of time?

Play the video backwards: latent predictors are the most likely to flip to the opposite action

Directional Semantic Coherence Score

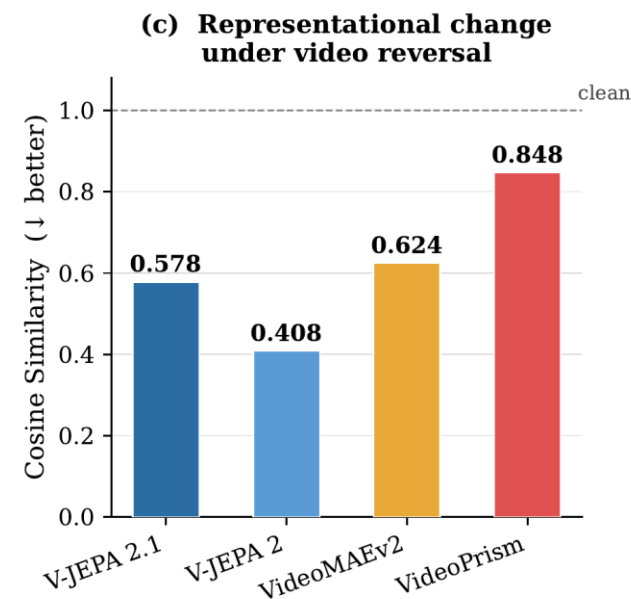
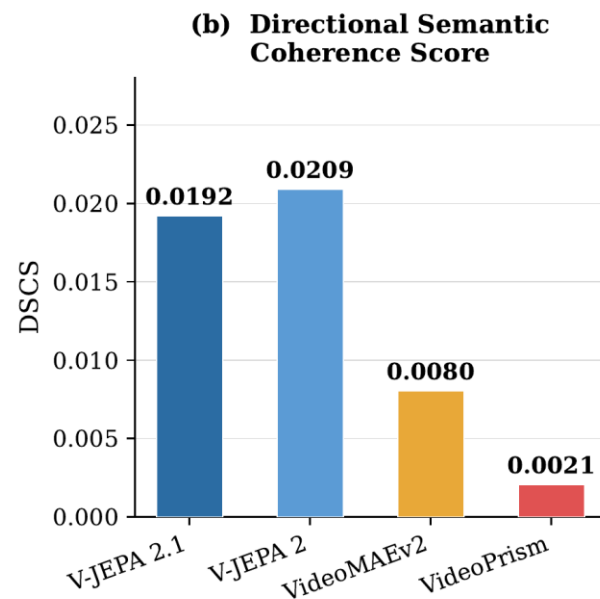
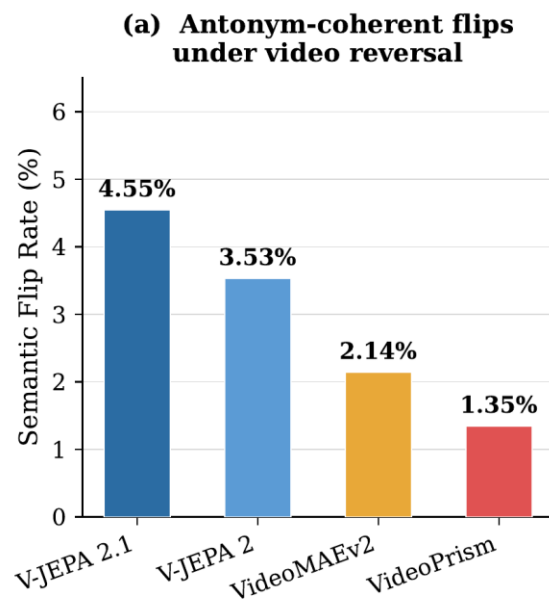
$$\text{DSCS} = r_{\text{sem}} \times (1 - \cos_{\text{rev}})$$

antonym-flip rate $\times$ how far the representation moves

Example of a coherent flip

“pushing something from left to right”

reversed $\rightarrow$ “pulling something from right to left”



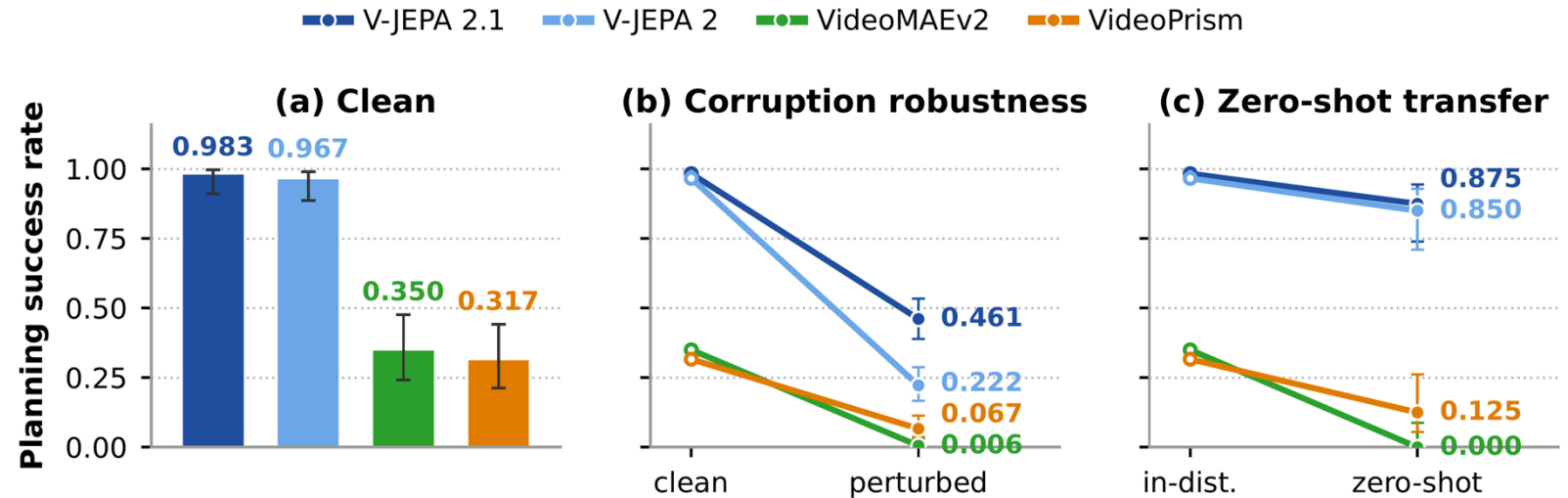
Frozen encoders $\rightarrow$ the direction of time was learned in pretraining: DSCS **2.4 $\times$** VideoMAEv2 and **9 $\times$** VideoPrism

Does it survive downstream planning?

Same predictor, same planner, frozen encoders: only latent prediction plans robustly

Closed-loop setup

- Meta-World manipulation: **drawer-open, door-close**
- Frozen encoder + identical action-conditioned predictor (4-layer causal transformer, $\approx 3.2\text{M}$ params)
- Cross-entropy-method planner
- Perturbed inputs reuse the corruption & occlusion families
- A **held-out task** tests zero-shot transfer



$\geq 97\%$ vs $\leq 35\%$

clean planning success
V-JEPA vs. the others

$83 \cdot 12 \cdot 1$

solved of 180 perturbed episodes
V-JEPA 2.1 · VideoPrism ·
VideoMAEv2

87.5% vs 0%

zero-shot on an unseen task
V-JEPA 2.1 vs. VideoMAEv2

Why? The objective decides what is kept

One mechanism explains all five robustness findings

Pixel reconstruction · VideoMAEv2

- Every pixel error is penalised equally
- Capacity goes to texture, lighting, background

→ **brittle when pixel statistics change**

Contrastive alignment · VideoPrism

- All views of a concept pulled together
- Invariance compresses intra-class detail

→ **stable embeddings, unusable distinctions**

Latent prediction · V-JEPA

- Targets are abstract representations
- Surface changes map to near-identical targets
- Masked future prediction is directional

→ **keeps meaning, contact and time's arrow**

Latent prediction is neither invariant nor pixel-faithful: it preserves exactly the information needed to predict what happens next.

Take-home messages

1

Latent prediction gives the most consistently robust video features

corruption, occlusion, pretend actions, arrow of time

2

Stable $\neq$ useful

judge robustness by accuracy, not only feature similarity

3

The advantage carries through to planning

current benchmarks are not enough for world models

Thank you!
Scan for the paper



Paper: arXiv 2605.15618
ali.alrasheed@student.unimelb.edu.au