Improved Invariant & Equivariant Representations & Visuo-Spatial Abilities via Self-Predictive Autoregressive World Modeling

seq-JEPA: Autoregressive Predictive Learning of Invariant-Equivariant World Models

Hafez Ghaemi Eilif B. Muller* Shahab Bakhtiari*
*Equal Contribution

SUMMARY

- The two-view paradigm in JEPAs ignores sequential actions and observations, which is a disadvantage when facing with tasks that require sequential aggregation of action-observation pairs, e.g., path integration or planning. Furthermore, this paradigm has been shown to suffer from a trade-off between invariant and equivariant downstream performance (Garrido et al., 2024).
- seq-JEPA breaks this paradigm via autoregressive predictive world modeling. We show that our world modeling paradigm addresses both aforementioned problems:

1. Natural separation of invariant and equivariant representational spaces and addressing the trade-off between invariance- and equivariance-related task performances

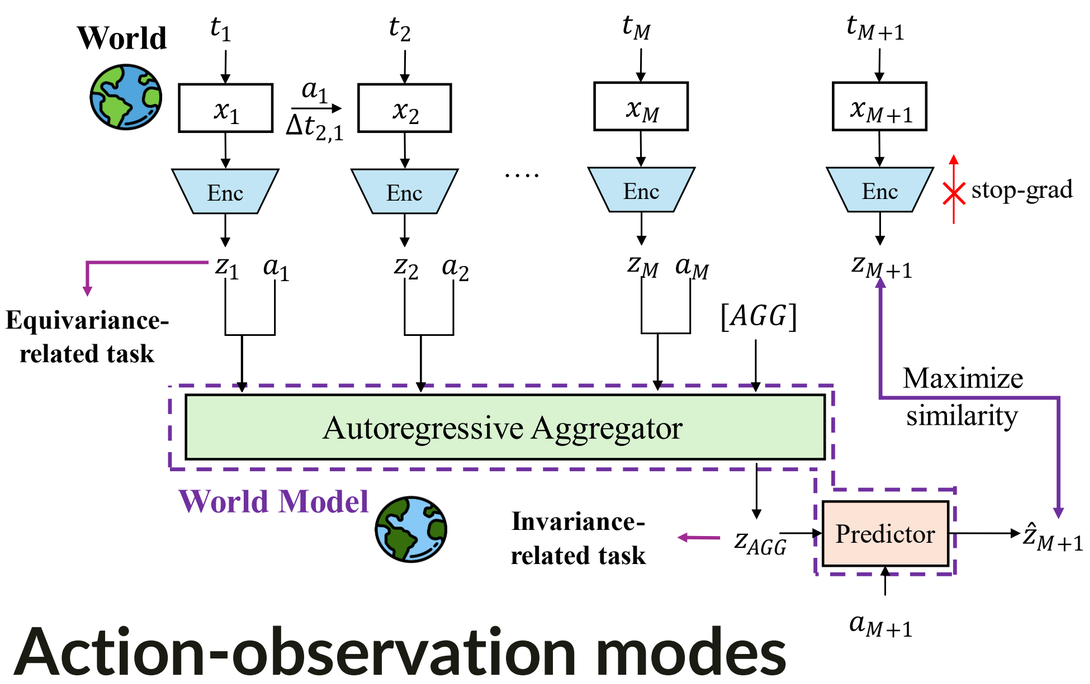
2.a. Aggregating a sequence of partial observations, e.g., low-resolution glances across saccades without using any full-sized image, hand-crafted augmentations, or masking.

2.b. Unlocking visuo-spatial abilities of world models that require aggregating a sequence of action-observation pairs; specifically, we show that seq-JEPA can perform saccade path integration, angular rotation integration, and odd-one-out (anomaly) detection.

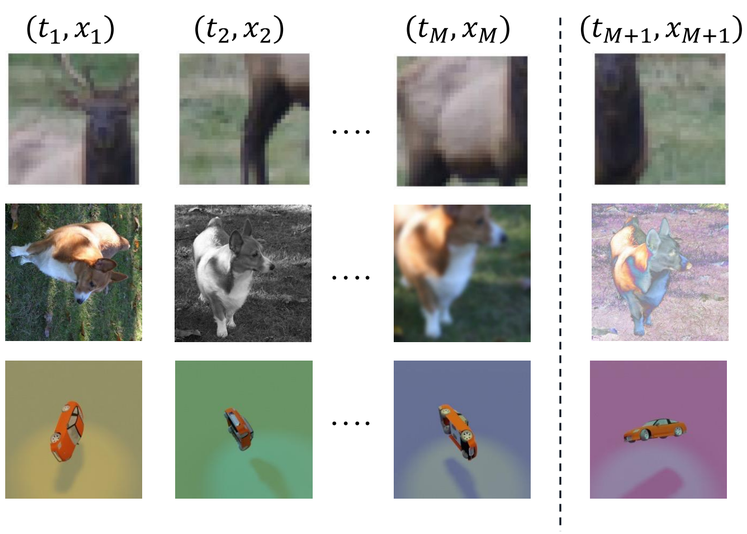
STL-10 object recognition performance; baselines use full-sized images w/ augmentations. seq-JEPA uses a sequence of glances each 1/9th of the image.

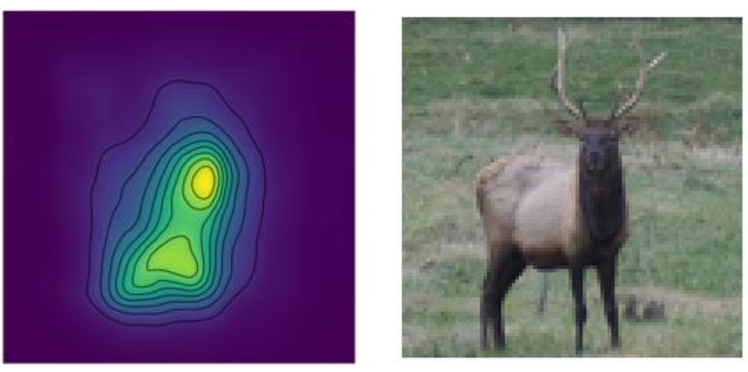
Method	Lin. Probe Acc. (%)
Invariant methods	
SimCLR	79.81
BYOL	78.21
VICReg	77.12
Equivariant methods	
EquiMod	78.40
SEN	77.91
SIE	75.88
seq-JEPA (CIFAR ResNet-18 enc., M=4)	**81.325** (z_{AGG}) 77.83 (z_i)
seq-JEPA ablations w/ M=4 (z_{AGG})	
Full-sized aug. imgs	79.12
No act. conditioning	78.23
ResNet-18 enc.	72.81
No saliency map	78.12
No IoR	79.45

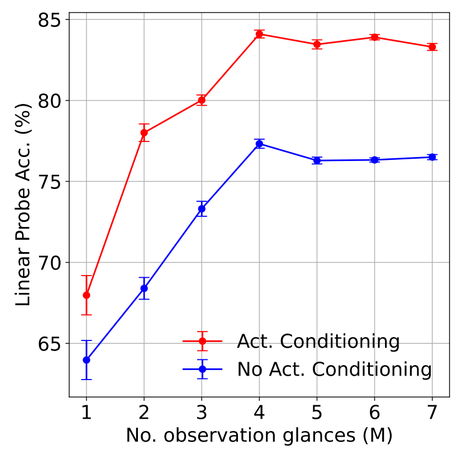
Architecture

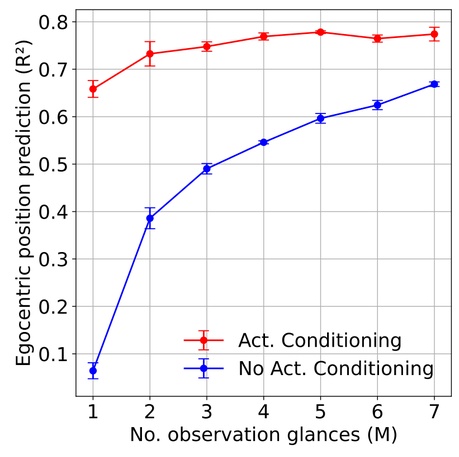


Action-observation modes



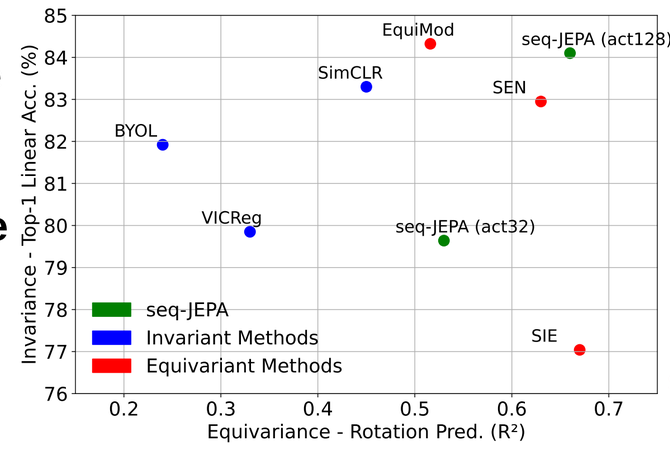


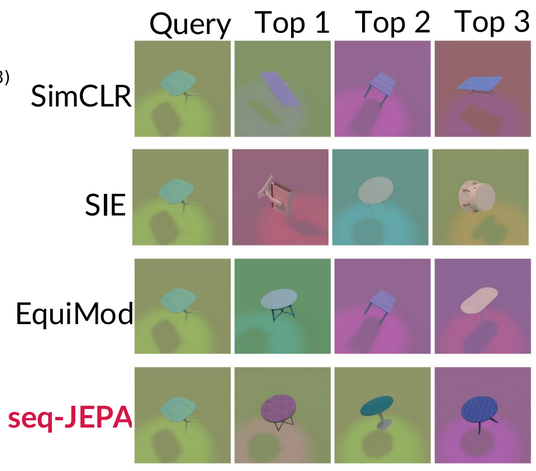




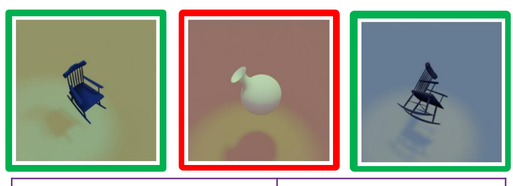
Both invariance and equivariance-related performances gain from longer sequence lengths

Quantitative and qualitative performance comparison on 3DIEBench

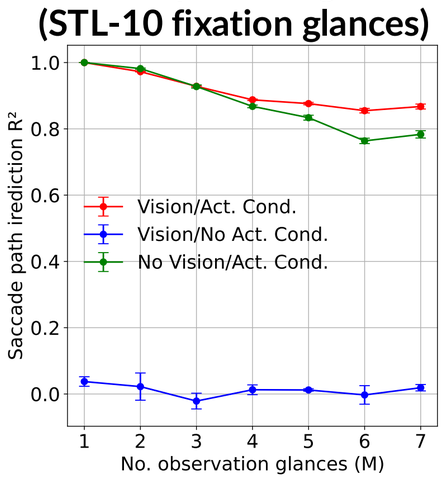




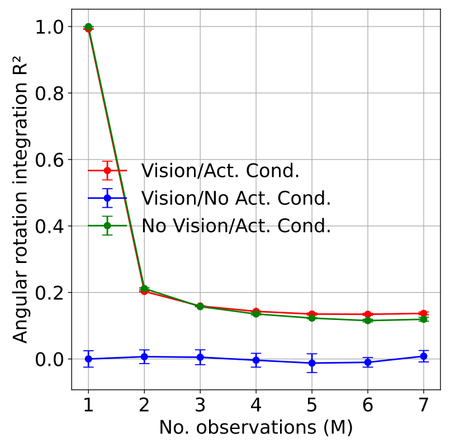
Visuo-spatial tasks requiring sequence aggregation



Saccade path integration (STL-10 fixation glances)



Angular rotation integration (3DIEBench)



Method	Odd-one-out detection Acc. (%)
seq-JEPA (z_{AGG})	52.29
seq-JEPA (z_{AGG}) w/ autoregressor fine-tuning	**66.65**
seq-JEPA (ResNet outputs)	37.56
SimCLR (ResNet outputs)	38.62
SIE (ResNet outputs)	39.16

Mila

Université de Montréal

ABL LAB — Architectures of Biological Learning