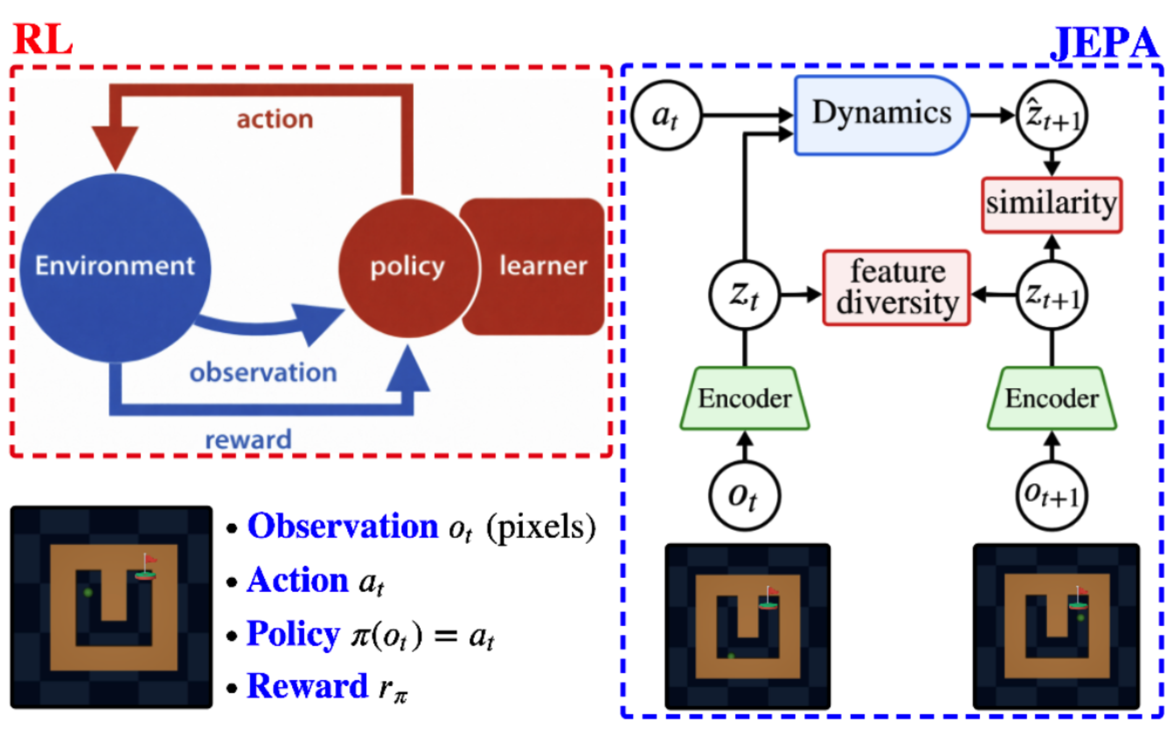
Learning Invariant Visual Representations for Planning with Joint-Embedding Predictive World Models

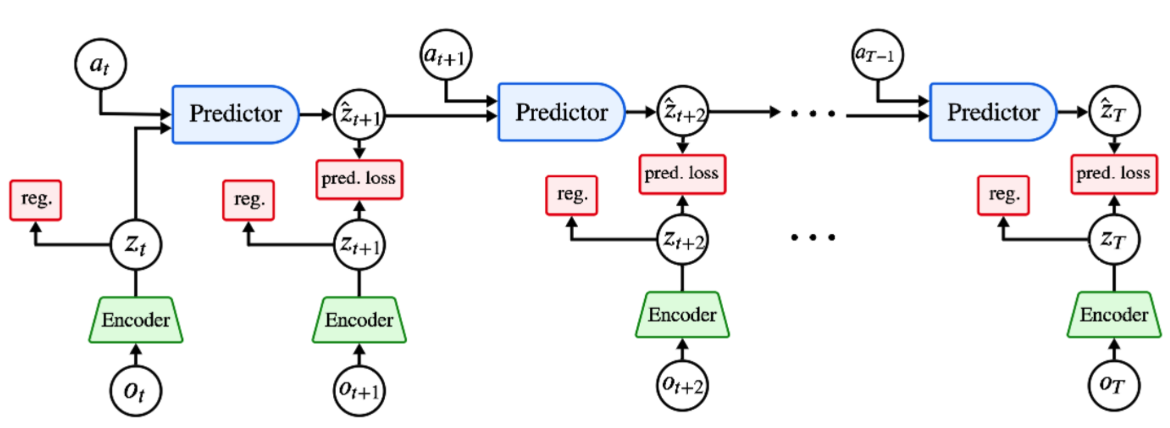
SCAN ME

Leonardo Toso, Davit Shadunts, Yunyang Lu, Nihal Sharma, Donglin Zhan, Gloria Geng, Nam Nguyen, and **James Anderson**

Columbia University · Capital One

JEPAs





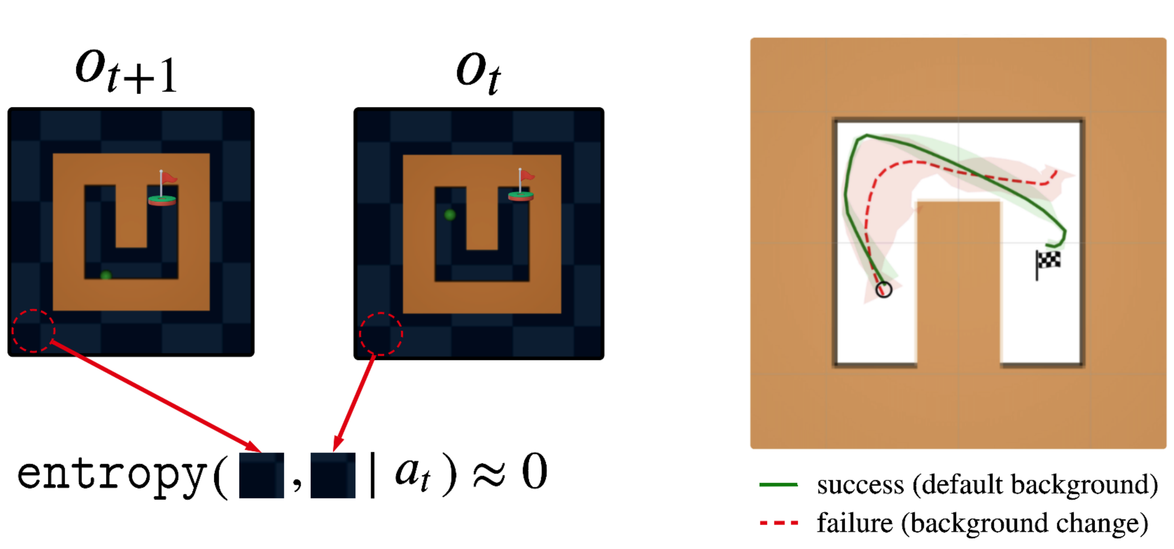
Focusing "too much" on slow features

Encoder: f_θ maps observations o_t to **latent embeddings** z_t

Dynamics: $\tilde{T}_\phi$ maps **current** latent state z_t and **action** a_t to **next** state latent $\hat{z}_{t+1}$

Prediction Loss:

$$\mathcal{L}(\phi, \theta) := \frac{1}{T} \sum_{t=0}^{T-1} \| f_\theta(o_{t+1}) - \tilde{T}_\phi(f_\theta(o_t), a_t) \|^2$$



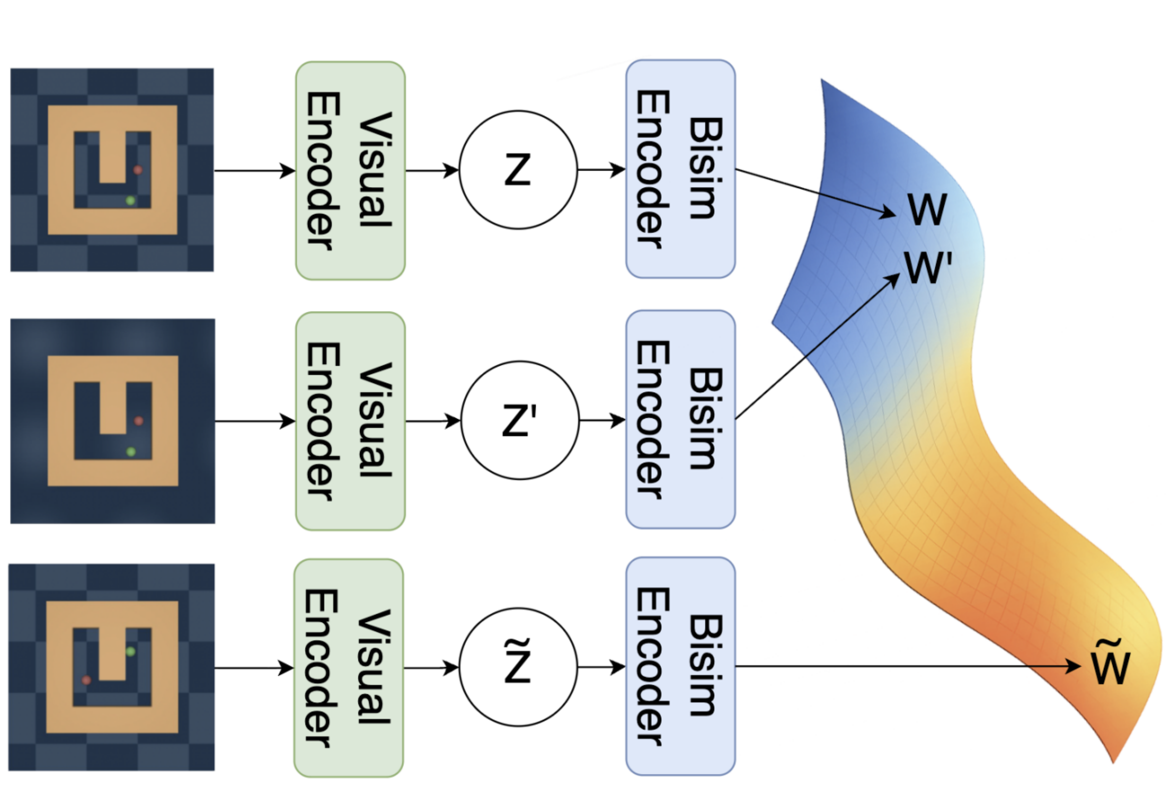
Bisimulation Encoder

$$d_\pi(z, z') = \underbrace{|r_\pi(z) - r_\pi(z')|}_{\text{reward similarity}} + \gamma \underbrace{W_1\left(P_\pi(\cdot \mid z), P_\pi(\cdot \mid z')\right)}_{\text{dynamics similarity}}$$

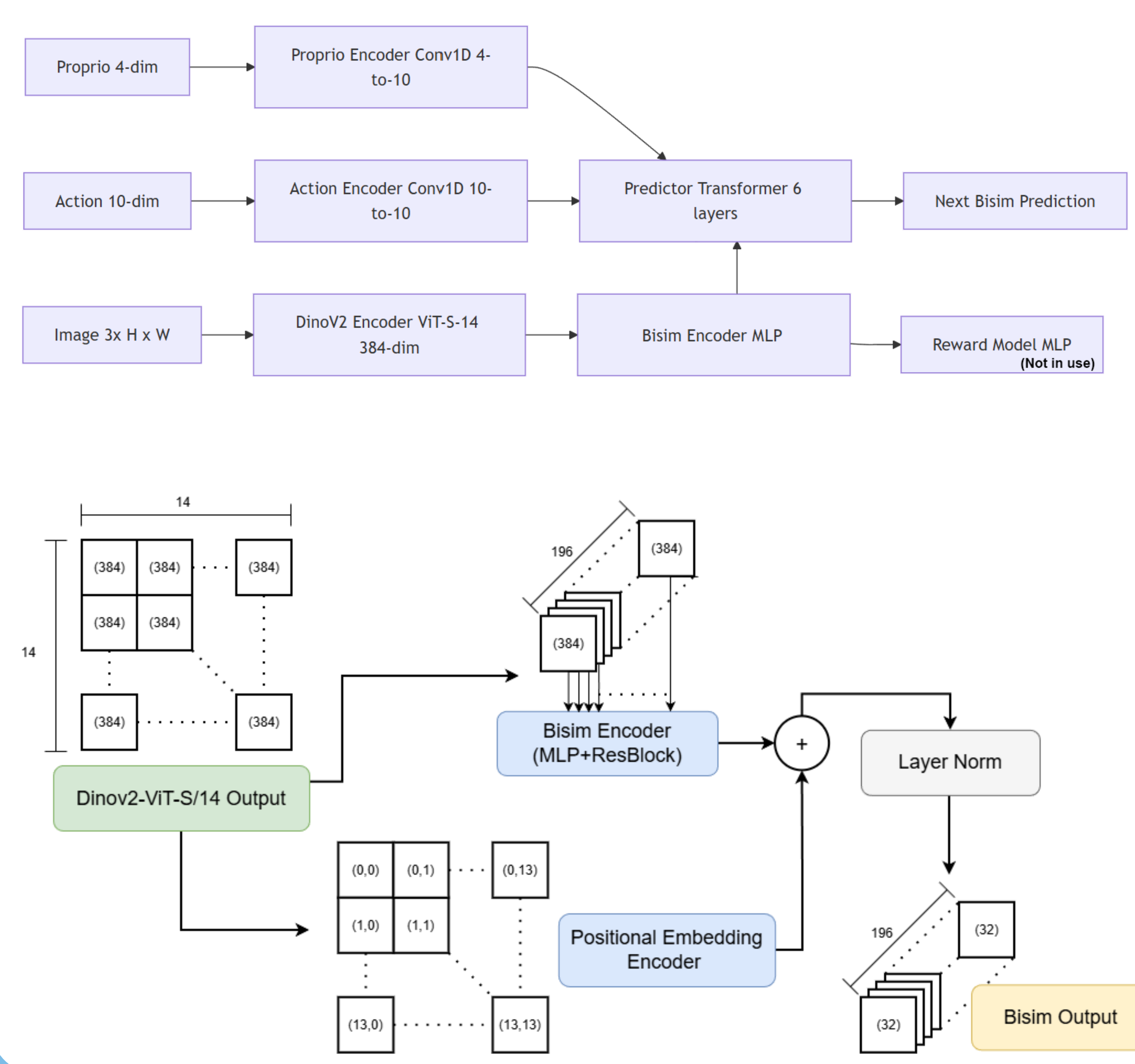
[Pablo Castro, AAAI 2020]

1-Wasserstein Distance: $W_1\left(P_\pi(\cdot \mid z), P_\pi(\cdot \mid z')\right)$

Discount factor: $\gamma \in [0, 1]$



Zoom on our architecture



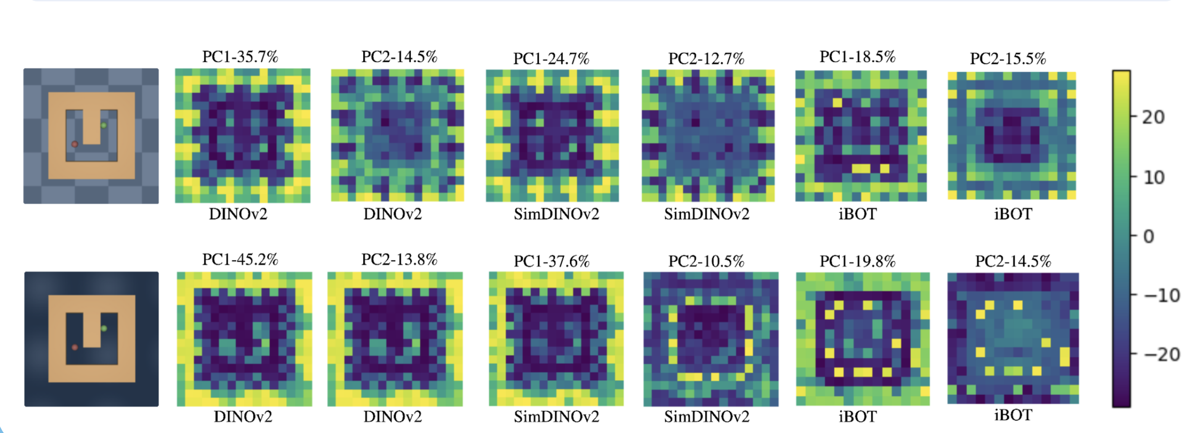
PCA-based Variance Covariance Reg.

Representation Collapse Avoidance

- Variance Invariance Covariance Reg. (VICReg)
- Sketched Isotropic Gaussian Reg. (SIGReg)

Prevents collapsing "slow" features via **bisimulation**

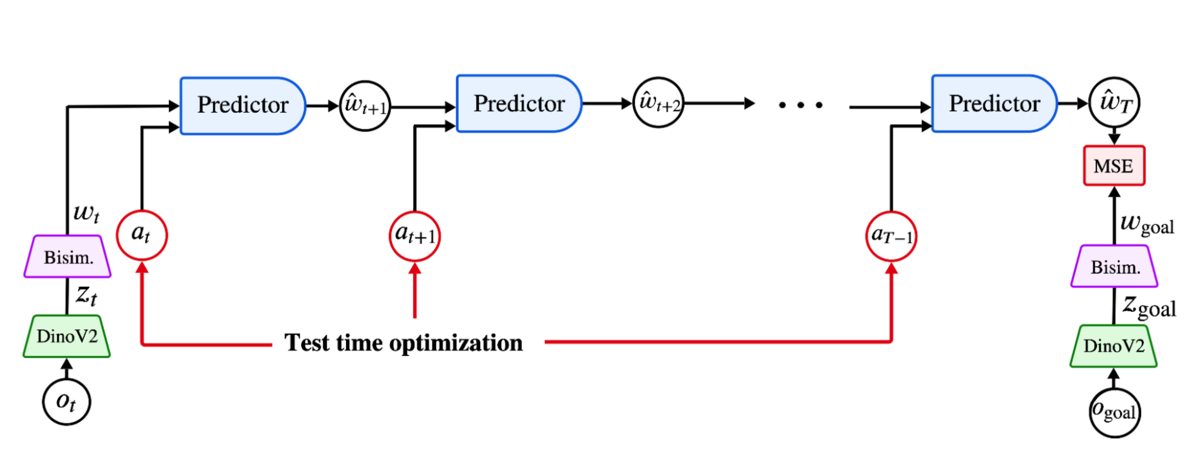
We do VICReg in the **PCA basis** of the **current batch**

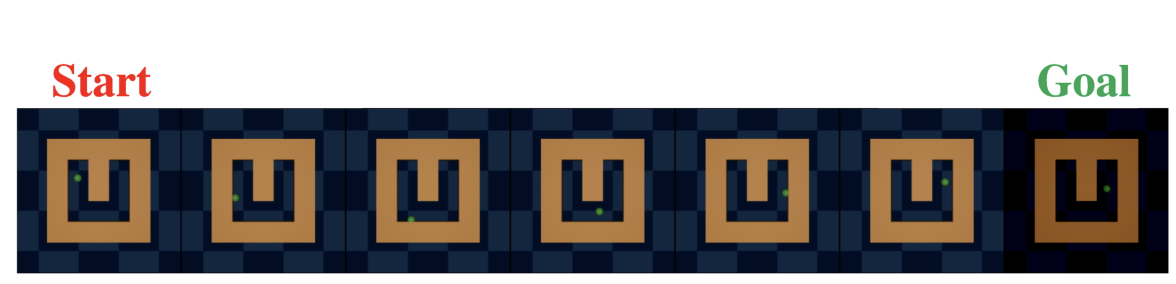


Planning

Goal-reaching cost:

$$C\left(\hat{w}_T, \{a_t\}_{t=0}^{T-1}\right) = \left\| \hat{w}_T(\{a_t\}_{t=0}^{T-1}) - w_{\text{goal}} \right\|^2$$





- **CEM-MPC:** Sample candidate action sequences over a planning horizon $T = 5$ and roll them out through the learned latent transition model.
- **Action selection:** Rank candidate sequences using the terminal latent-space cost $C\left(\hat{w}_T, \{a_t\}_{t=0}^{T-1}\right)$ and select the minimum-cost sequence.

Generalization Bound

Cost-to-go:

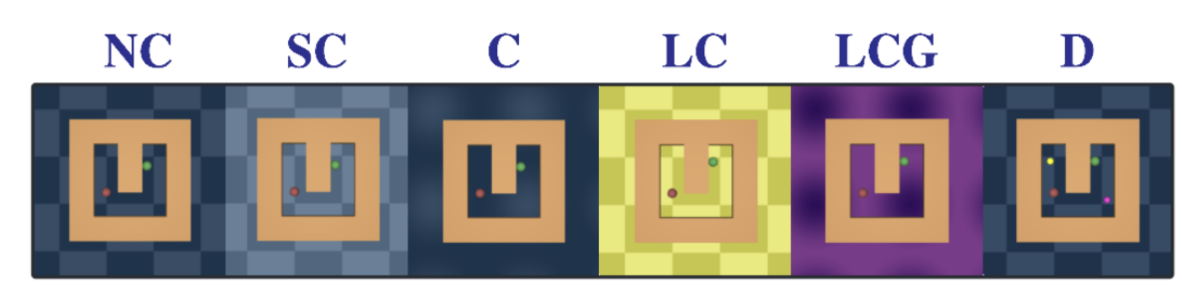
$$J_T^\pi(z) \triangleq \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t C\left(f_\theta(z_t)\right) \mid z_0 = z\right],$$

where $z_{t+1} \sim P_\pi(\cdot \mid z_t)$.

$$\left| J_T^\pi(z) - J_T^\pi(z') \right| \lesssim T\, d_\pi(z, z') \quad \textbf{(Gen. Bound)}$$

- Planning-cost differences between two latent states grow at most linearly with the horizon, as $\mathcal{O}\left(T d_\pi(z, z')\right)$.

Numerical Illustration



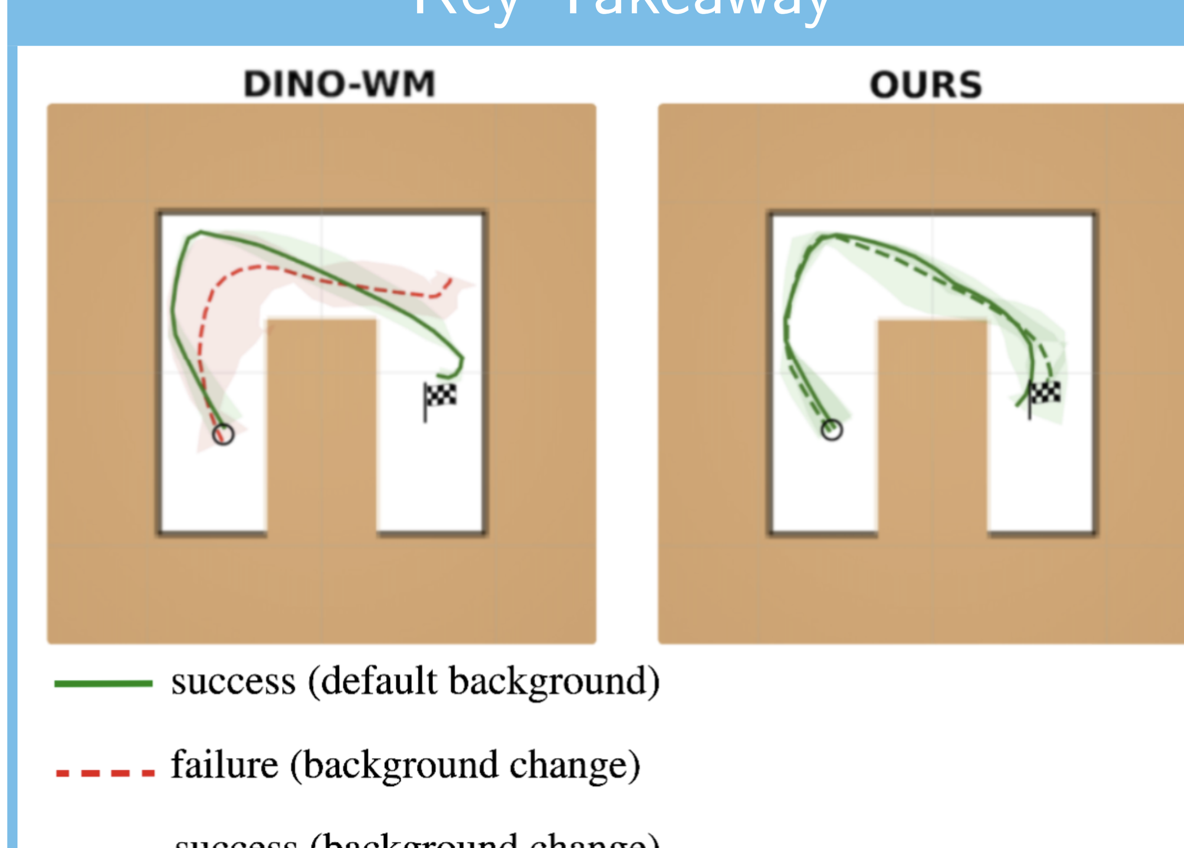
NC: No change, **SC**: Slight background change, **C**: Color gradient background, **LC**: Large color change, **LCG**: Large color gradient change, and **D**: moving distractors.

Model	NC	SC	C	LC	LCG	D
DINO-WM	0.80	0.72	0.60	0.56	0.48	0.78
w/DR	0.82	0.82	0.82	0.68	0.64	0.82
Ours	0.78	0.80	0.76	0.86	0.78	0.82

Model	NC	SC	C	LC	LCG	D
End-to-End	0.68	0.44	0.70	0.26	0.36	0.64
DINOv2	0.78	0.8	0.76	0.86	0.78	0.82
SimDINOv2	0.40	0.38	0.36	0.42	0.42	0.36
iBOT	0.72	0.70	0.74	0.72	0.72	0.72

Domain Randomization (DR):

Key Takeaway



Compact and robust latents

↓ 10×

↓ ≈ 0.8 SR