Latent Actions from Factorized Transition Effects

under Agent Ambiguity

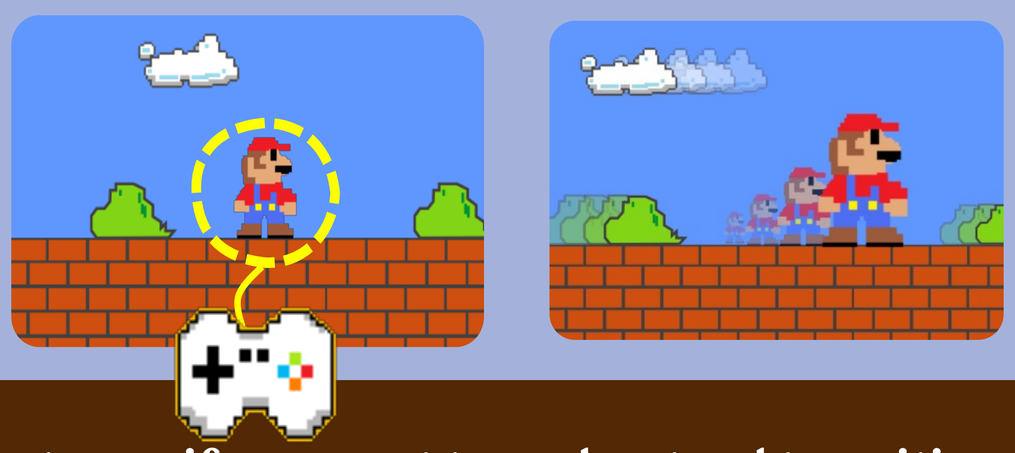
WM@Booth ICML International Conference On Machine Learning BROWN

Heejeong Nam Chandradithya S Jonnalagadda Harshit Aggarwal Eric Xu Randall Balestriero

Paper Project Page

Looking for PhD position

World Model can learn "What is being controlled" by action supervision

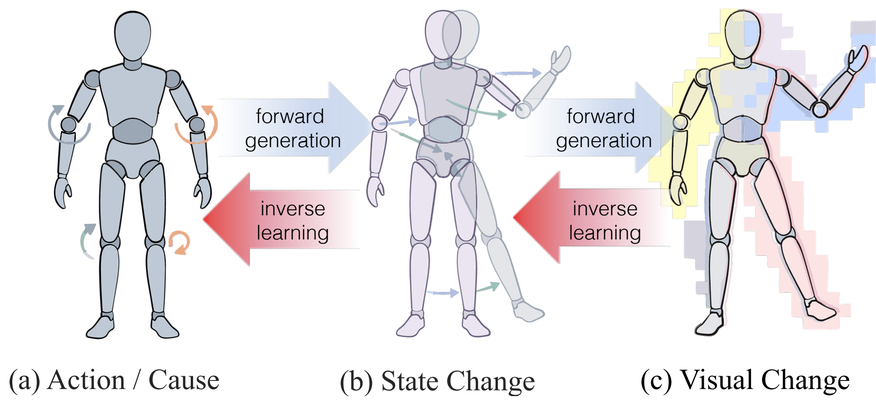


But now, if you want to understand transition without any supervision…

There's no concept of **disctractor or noisy transition** in your Latent Action Model

A more useful objective may be to build a **robust representation of what changed**, under **agent ambiguity**.

Forward Generative Chain / Backward Learning Chain

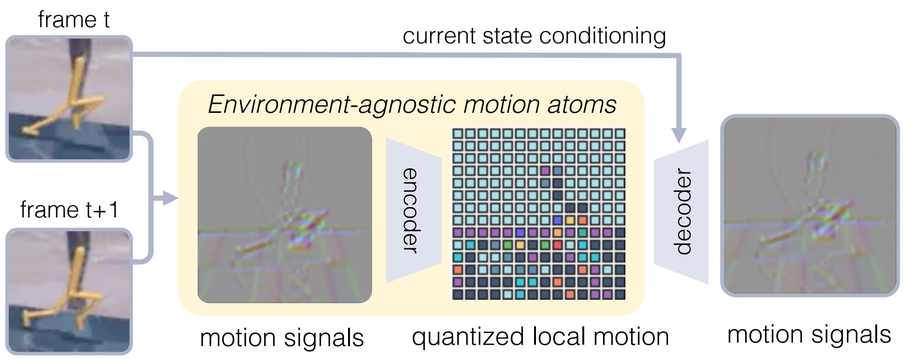


(a) Action / Cause (b) State Change (c) Visual Change

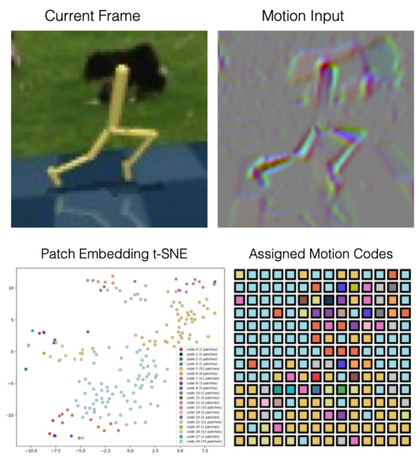
- Actions and uncontrolled dynamics jointly induce state changes and visual transition effects.
- Motion-centered inputs and frame-conditioned decoding suppress appearance and embodiment-specific information.

Step 1. Pretrain Motion Primitives

- OTF learns a shared codebook of reusable local observed-transition primitives.
- Motion-centered inputs and frame-conditioned decoding suppress appearance and embodiment-specific information.

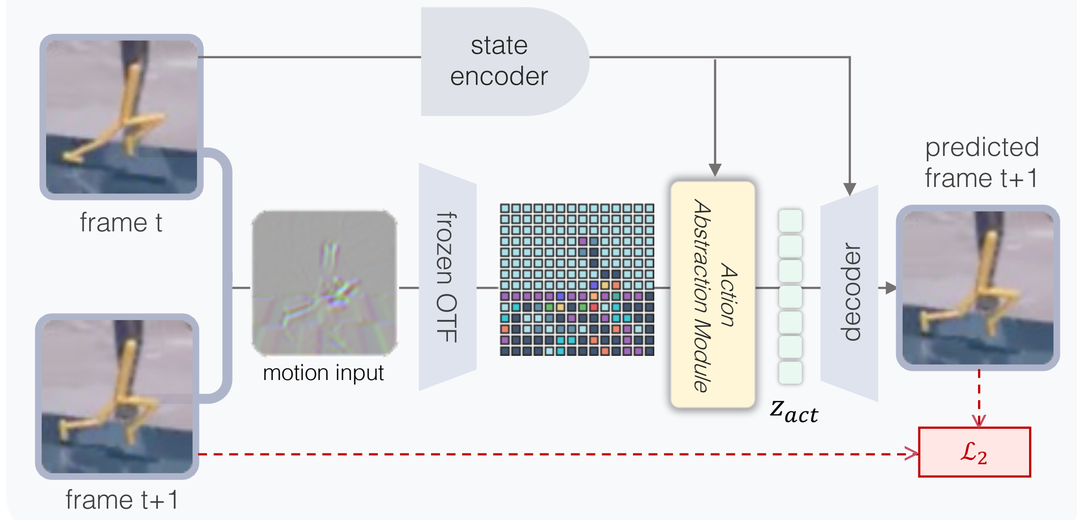


- Example

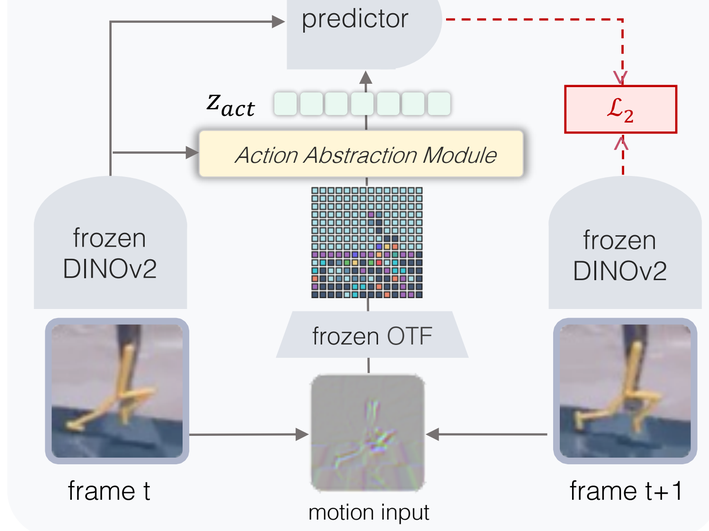


Step 2. Learning Latent Action Models

- Frozen OTF codes are combined with current-state features into state-aware transition tokens.
- Learnable action queries aggregate local transition effects into a compact continuous latent action.
- Pixel and DINO variants learn latent actions through future pixel or feature prediction.

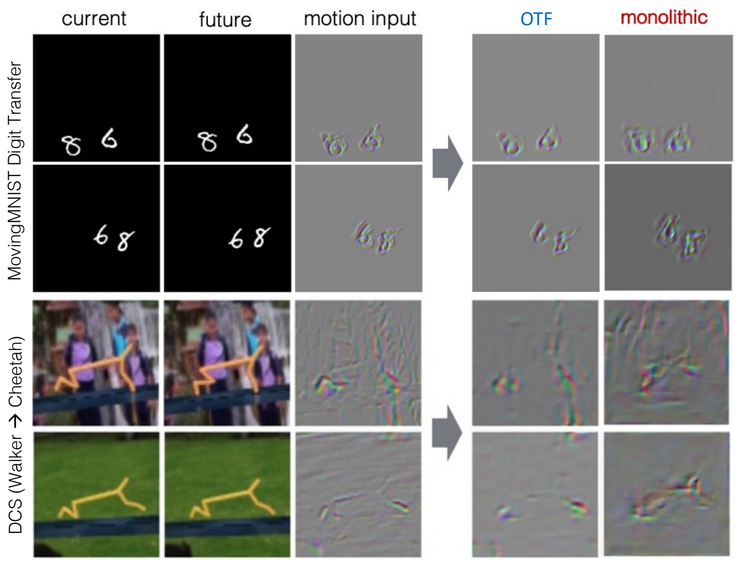


(a) OTF-LAM-Pixel



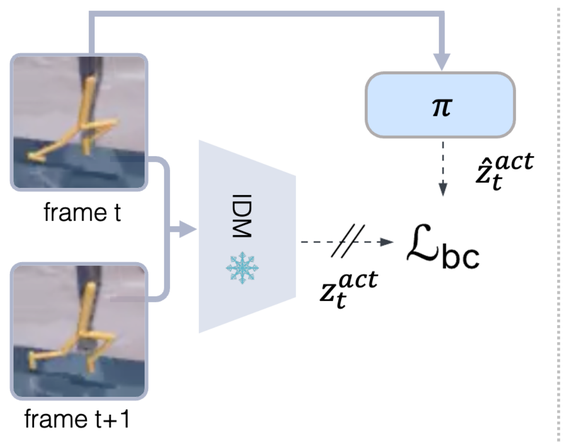
(b) OTF-LAM-Dino

Results

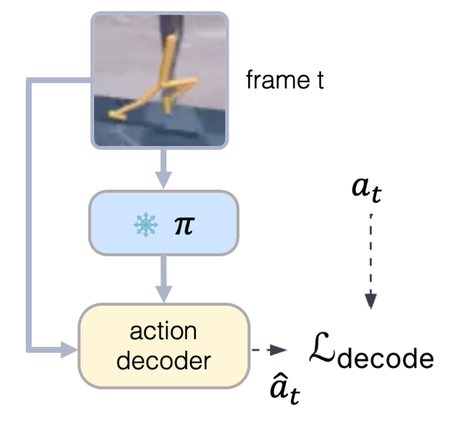


- OTF primitives transfer zero-shot across unseen visual carriers and embodiment morphologies.
- Factorized OTF shows substantially smaller transfer degradation than a monolithic transition bottleneck.
- OTF-LAM-DINO yields substantially stronger action alignment and more sample-efficient action decoding.
- OTF-LAM-DINO achieves competitive or superior downstream policy performance against existing LAMs.

Step 3. Downstream Policy Learning

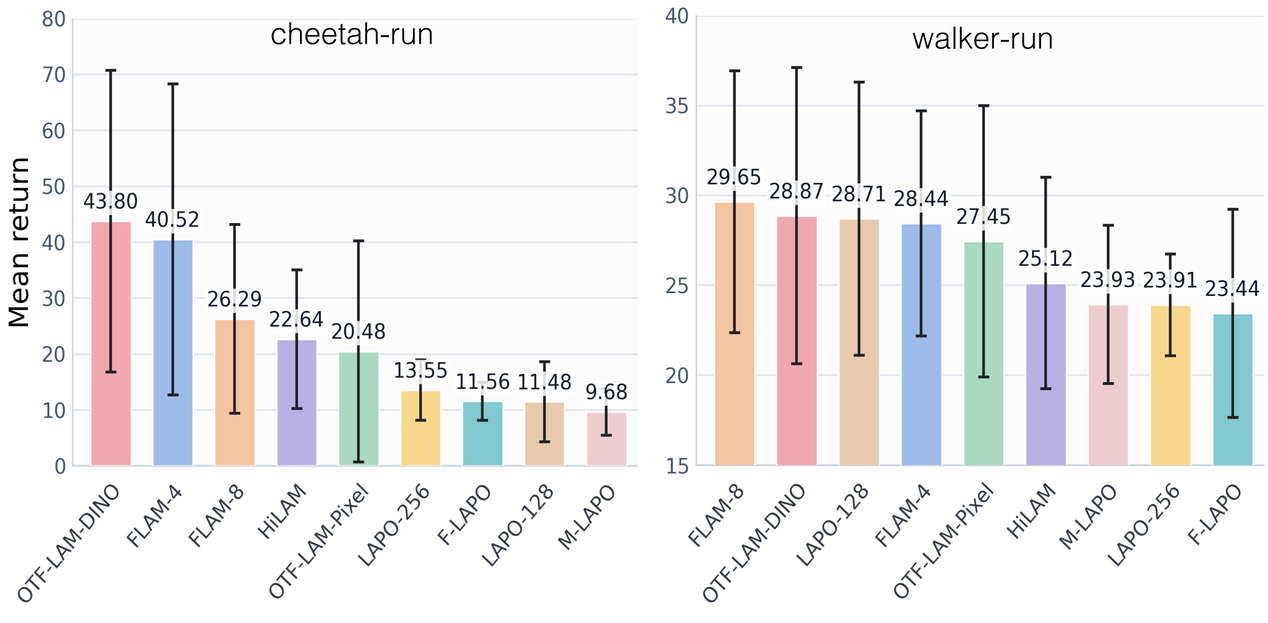


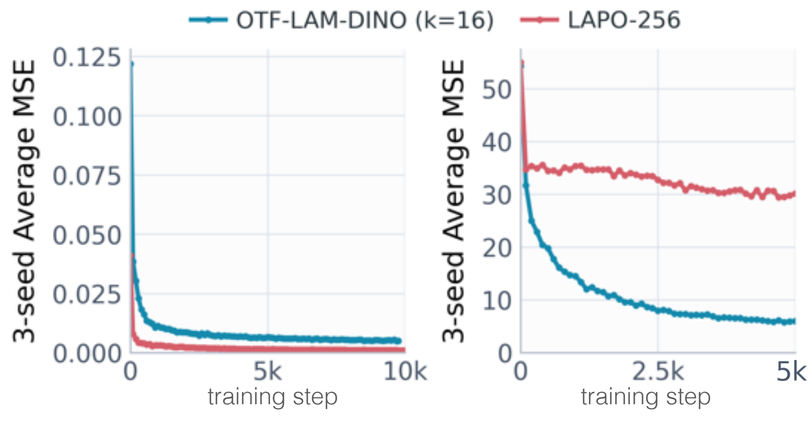
(a) Behavioral Cloning



(b) Action Decoding

- Behavioral cloning predicts latent actions, while limited supervision decodes them into environment actions.





(a) Behavioral Cloning (b) Action Decoding

- OTF-LAM-DINO requires substantially less action supervision to recover environment actions, indicating stronger intrinsic action alignment.