On the use of Pre-trained Visual Representations in Visuomotor Robot Learning

N. Tsagkas, A. Sochopoulos, D. Danier, S. Vijayakumar, C. X. Lu†, O. M. Aodha†

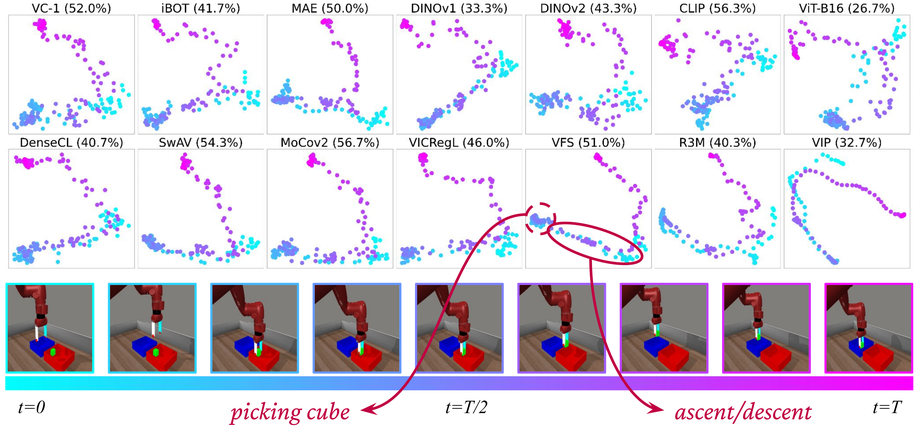
{n.tsagkas, oisin.macaodha}@ed.ac.uk, xiaoxuan.lu@ucl.ac.uk | † indicates equal senior authorship

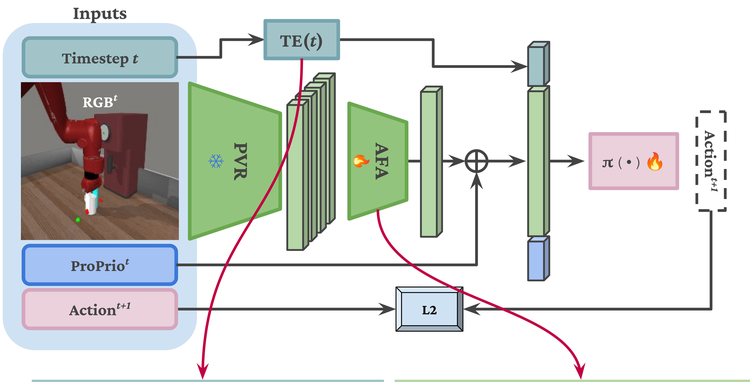
CVPR Nashville JUNE 11-15, 2025

Project Page

Temporal Entanglement

Problem Statement. Policies using frozen PVR features often violate the Markov assumption, as single-frame observations may lack sufficient information to determine the correct action. As shown below, PVR features from a pick-and-place trajectory exhibit temporal entanglement: (i) frames during static grasps cluster due to minimal pixel changes, and (ii) ascent/descent motions yield near identical features, differing only slightly in regions affected by the cube's displacement. This ambiguity hampers learning a consistent mapping from observations to actions.





Temporal Encoding

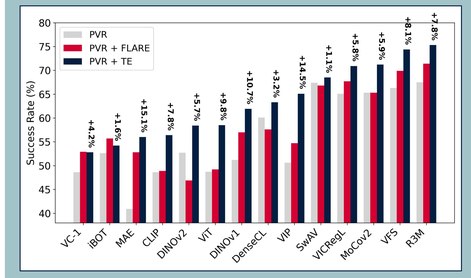
$$\left(\sin\left(\frac{2^k \pi t}{s^k}\right), \cos\left(\frac{2^k \pi t}{s^k}\right)\right)_{k=0}^{T-1}$$

We map each timestep t to a temporal encoding (TE) and append it to the corresponding observation. This simple augmentation injects temporal structure, helping disambiguate visually similar states and improve policy learning.

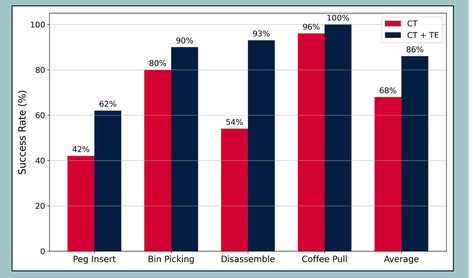
Attentive Feature Aggregation

$$softmax\left(\frac{\boldsymbol{q}\cdot(F\cdot W_K)^\top}{\sqrt{d_k}}\right) F \cdot W_V$$

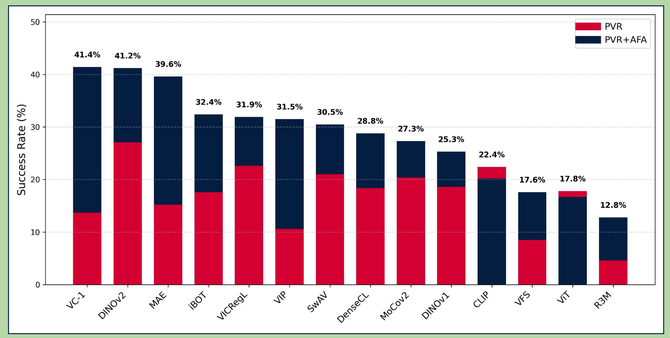
We learn to attend to task-relevant visual cues by training a cross-attention layer, with a trainable query token. This leads to the filtering of scene distractors and visual changes.



*Average policy performance per PVR, trained across 10 tasks. **TE leads to a significant boost** in performance, even when using PVRs that have been trained with a temporal objective and video-data.*



*TE still improves performance even when **introducing a causal-transformer** (CT), with context and action horizon >0, and relative positional encoding in the input.*



***OOD evaluation**: AFA improves policy performance under scene visual changes. MIM-trained PVRs benefit the most.*

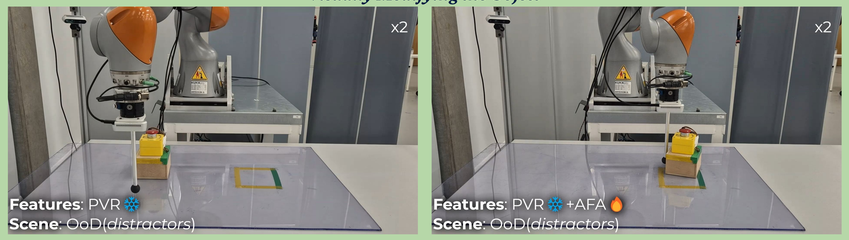
Robustness Under Visual Perturbations

Problem Statement. Training policies using global features from PVRs (i.e., the CLS token in ViTs or average pooled features in CNNs) can lead to overfitting to visually dominant but task-irrelevant scene attributes (e.g., background textures). This dilutes the policy network's ability to focus on features critical for decision-making. Prior work suggests that only specific image regions contribute meaningfully to task success, and recent findings in PVR distillation indicate that local information is especially valuable in robot learning, but this remains underexplored.

Light Perturbation

Scene Distractors

Visually Modifying the Object



***Final step** from policy deployment in the planar pushing task. The PVR+AFA policy has no problem pushing the cube in the goal area, whereas without AFA, the policy fails to generalize OOD.*