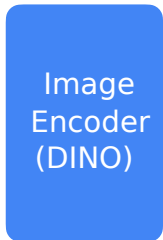


Input (RGB)



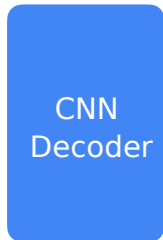
Image
Encoder
(DINO)



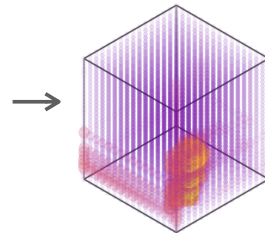
Feature Map



CNN
Decoder



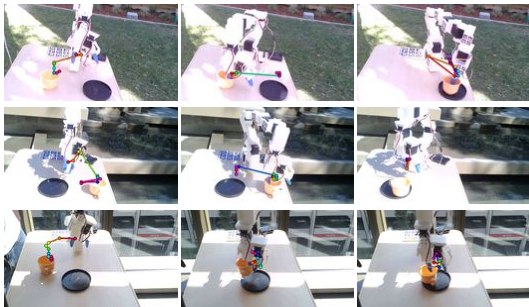
Pixel-Aligned Heatmap Volume
for Future Timestep



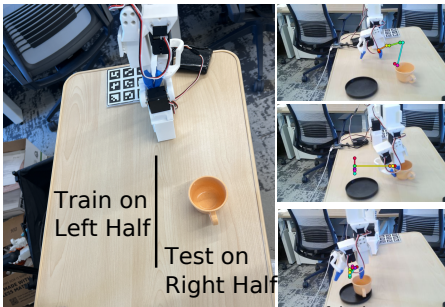
x,y,z

Argmax for
End-Effector
Prediction

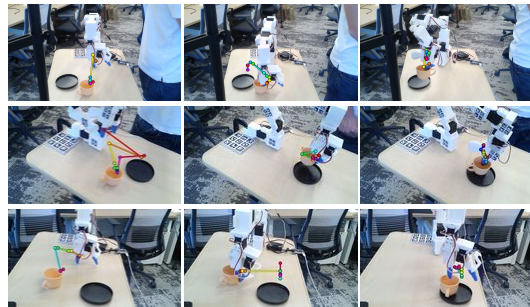
Environment Robustness



Object Position Robustness



Camera Viewpoint Robustness



Note fake images on
this column, todo