VIOLA uses object proposal priors to learn a closed-loop visuomotor policy to make coffee with only 50 raw demonstrations

VIOLA: Imitation Learning for Vision-Based Manipulation with Object Proposal Priors

Yifeng Zhu*, Abhishek Joshi, Peter Stone, Yuke Zhu
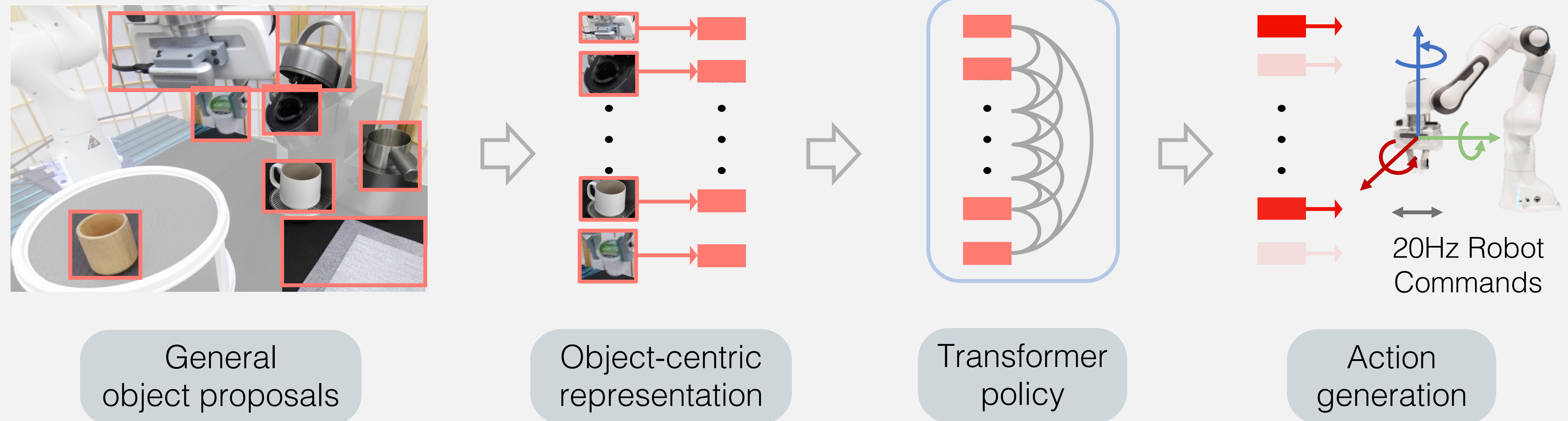
PRESENTER: Yifeng Zhu
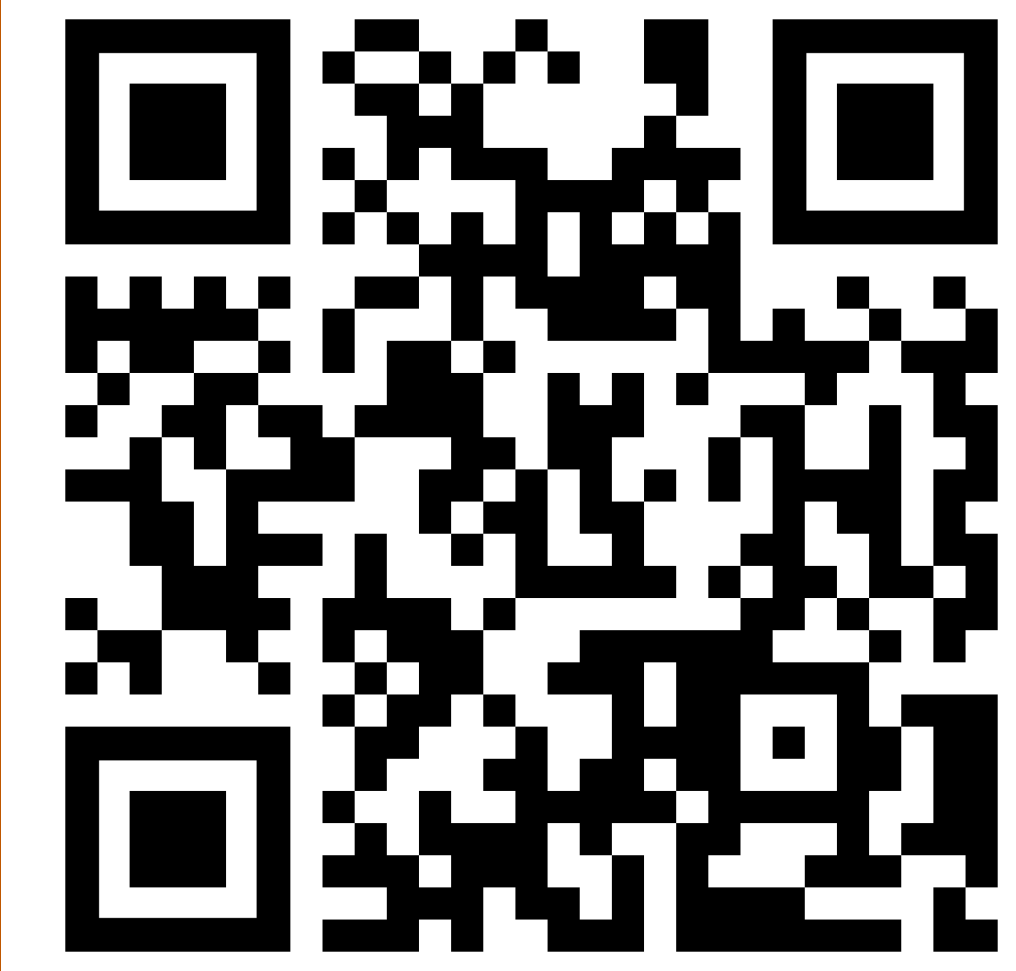
Motivation
- Imitate closed-loop visuomotor policies efficiently for manipulation
- Reduce spurious correlation in visuomotor learning
- Improve policy generalization from a handful of demonstrations

Insights
- Object-centric priors facilitate more efficient and robust policy inference
- Pre-trained RPN from large-scale image datasets captures general object priors
- Transformer self-attention mechanism selects task-relevant object features

General object proposals

Object-centric representation

Transformer policy

Action generation

20Hz Robot Commands

Experiments

Dining-PlateFork

Dining-Bowl

Make-Coffee

Dining-PlateFork — 36.7 / 76.7
Dining-Bowl — 20.0 / 60.0
Make-Coffee — 0.0 / 60.0

Success rates (%)   ■ BC-RNN   ■ VIOLA

Laundry List
- Impedance controllers for behavior cloning
- Suboptimal data actually helps to learn close—loop behaviors
- Generalization is limited to same task setup from training

VIOLA Model

Feature Encoder

Bounding Box Positions

Top K Proposals

Feature Pooling

Region Features

RPN

ResNet-18

Spatial Feature Map

Spatial Softmax

Eye-in-hand Image   ResNet-18

Proprioception   Linear Projection

Context Features

s_t

h_t

Object-Centric Representation

z_t

Feature Encoder

h_{t-H}

s_{t-H}

Temporal Position

0 1 $\cdots$ H

Feature Encoder

h_t

s_t

Transformer

Robot Commands

GMM MLP

Action Token

$\hat{a}_t$