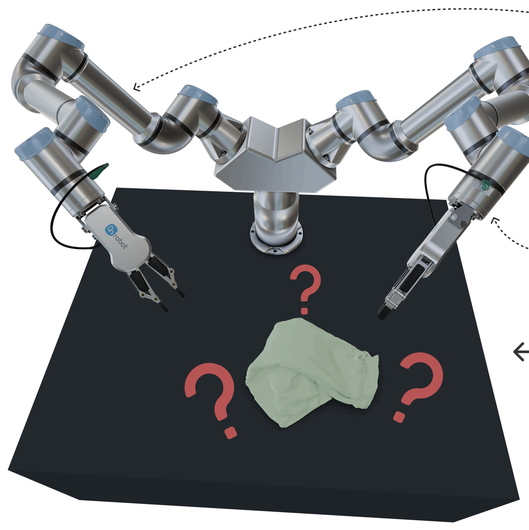
Beyond Static Perception: Integrating **Temporal Context into VLMs** for Cloth Folding

Oriol Barbany, Adrià Colomé, Carme Torras

Motivation



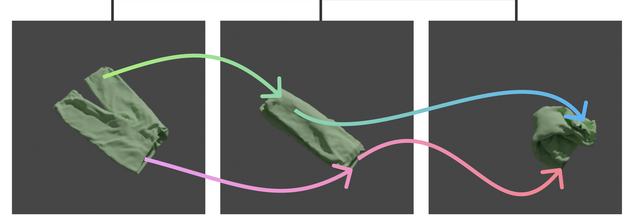
Static Policy

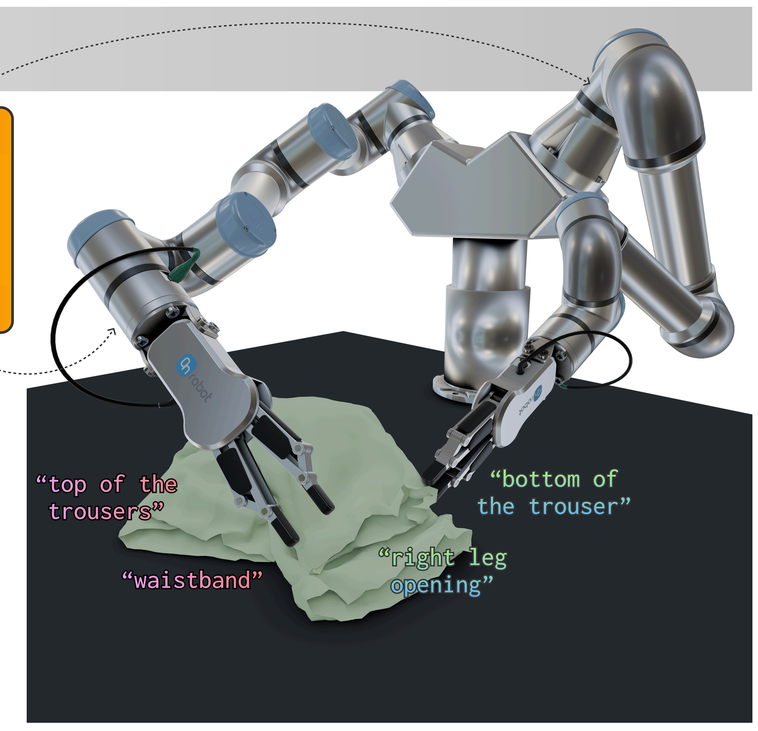
"Grasp the right leg opening"

Policy with temporal context

Challenging to recognize crumpled states

Provide previous observations

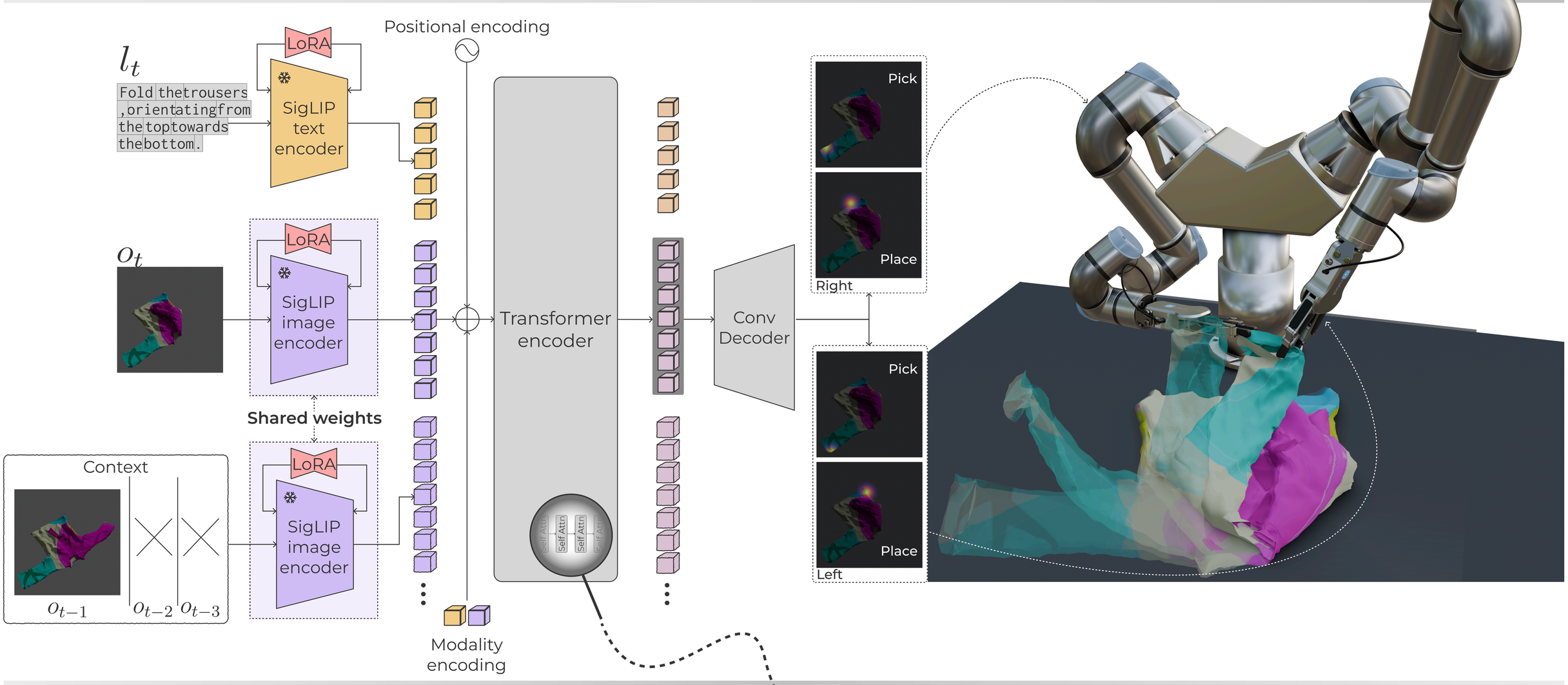




Model	AP_5 (↑)	AP_{10} (↑)	AP_{50} (↑)	Quantile (%) (↑)
BiFold w/o context (static)	40.3	55.7	88.9	97.1
BiFold consecutive frames	34.6	49.8	83.9	97.3
BiFold (***keyframes***)	**75.1**	**92.8**	**98.9**	**98.8**

Frames are redundant
Use *keyframes!*

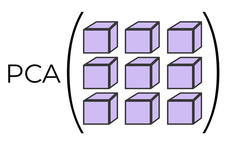
BiFold model architecture

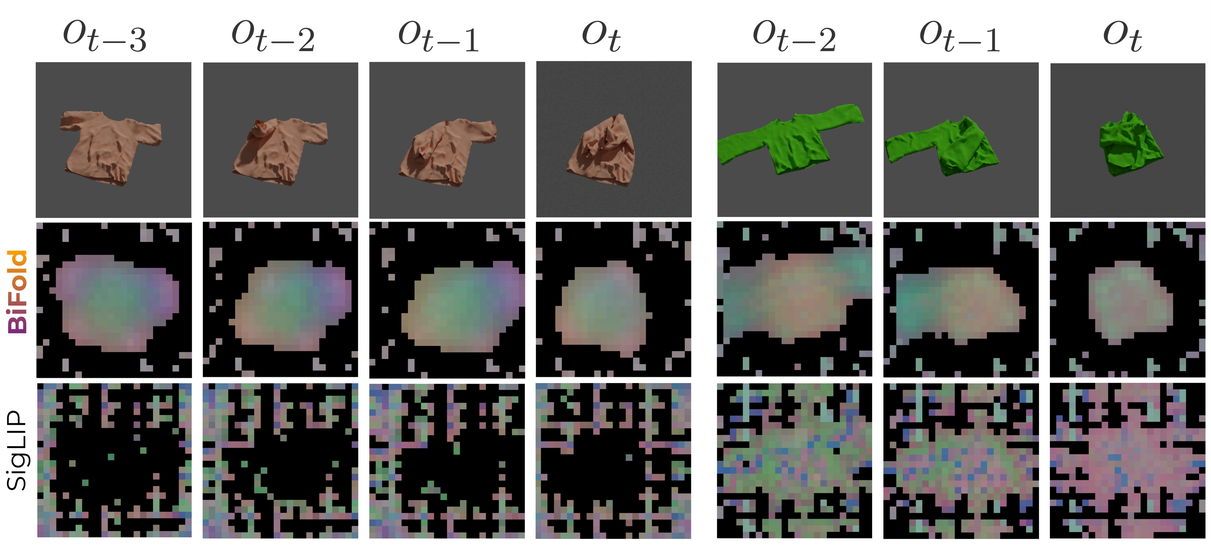


Emergent visual correspondences

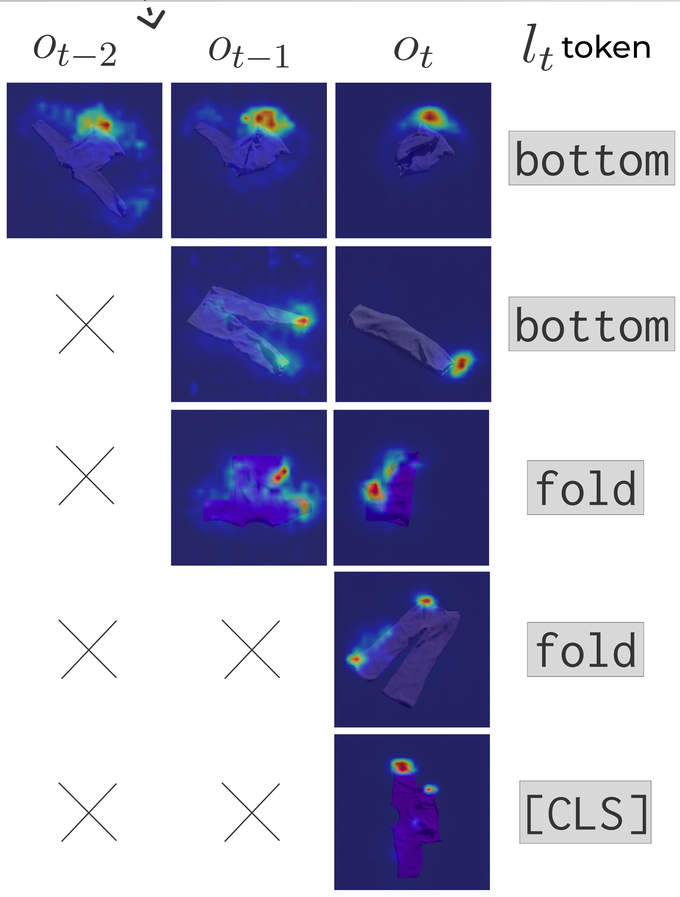
Parameter-Efficient Fine-tuning of VLMs:

- Extract task-specific representations
- Robustness and open-vocabulary understanding from pretraining





Effective modality fusion



Consistent spatial focus across temporal frames

Incorporating ***keyframe*** history improves complex object state recognition

Institut de Robòtica i Informàtica Industrial
UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH
CSIC Consejo Superior de Investigaciones Científicas
ellis European Laboratory for Learning and Intelligent Systems
ICRA IEEE International Conference on Robotics and Automation
ATLANTA 2025