Lift, See, Act: Hierarchical Robot Policy Pretraining with 3D Foundation Models

Yiyuan Ge, Changxing Ding, Ziyu Hao, Zijie Zheng, Xiangmin Xu

South China University of Technology

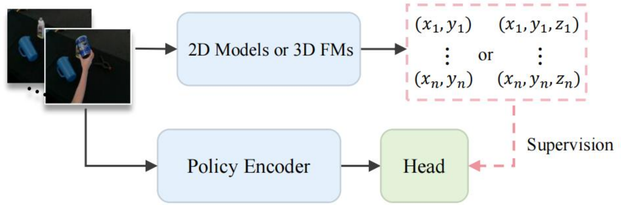
South China University of Technology

NEURAL INFORMATION PROCESSING SYSTEMS

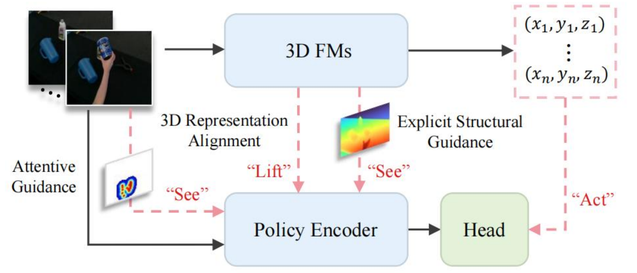
MPILab

Feel free to contact

Motivation



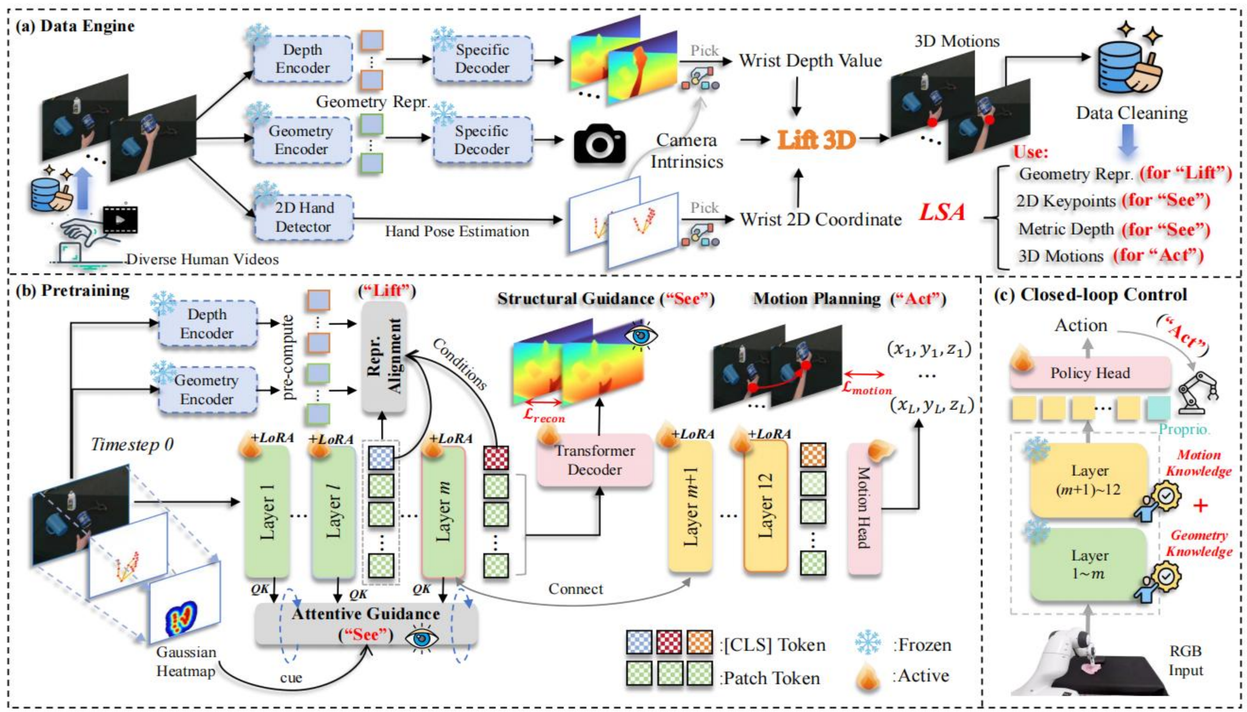
(a) Existing Explicit Methods



(b) Our Lift, See, Act (LSA) Method

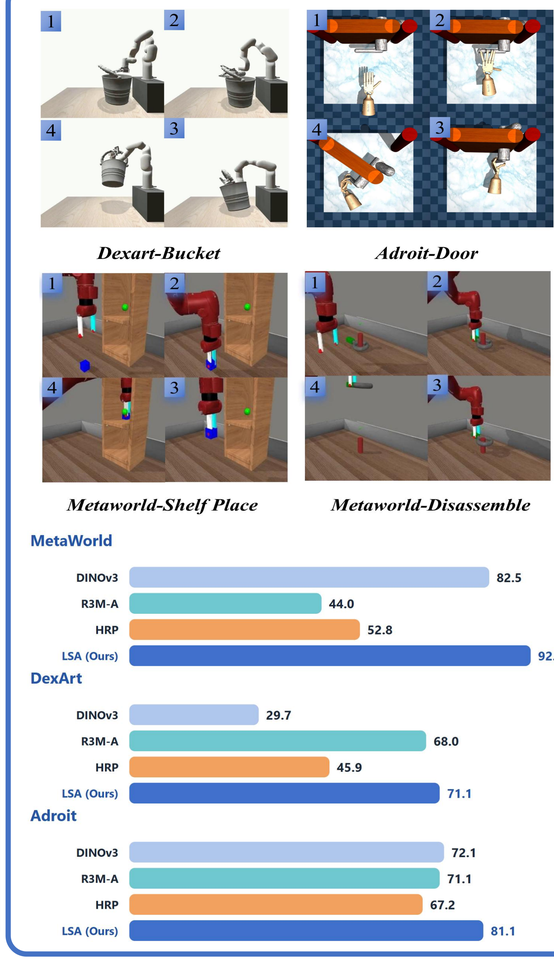
Human videos provide scalable and diverse manipulation knowledge for robot policy pretraining. **However**, **prior 3D-trajectory-based methods** retain only sparse hand trajectories, **discarding the rich geometric features and depth maps** produced by 3D foundation models during trajectory extraction; **LSA** fully exploits this multi-tiered 3D information.

Methods

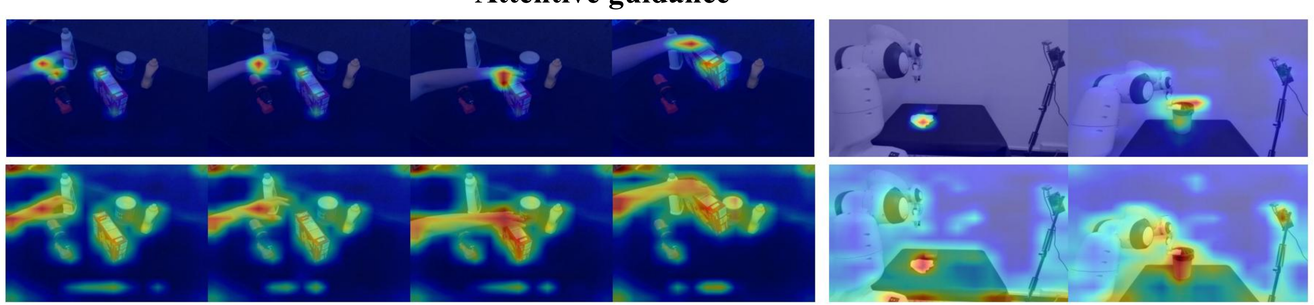


- **Lift:** Align shallow policy features with intermediate representations from multiple 3D foundation models, lifting 2D observations into a rich 3D representation space.
- **See:** Use depth reconstruction and attentive guidance to reveal scene geometry and focus the policy on task-relevant hand-object interactions.
- **Act:** Predict future 3D hand trajectories to learn manipulation dynamics and transfer the pretrained representation to closed-loop robot control.

Simulation



Attentive guidance



Attentive guidance sharpens the model's focus on task-relevant hand–object interactions in human videos and effectively transfers this interaction-aware attention to real-world robot scenes.

Real-World

