Look Only When the Action Could Change

Decision-sensitive wrist-camera queries for language-conditioned manipulation

Abstract

Occlusion makes a robot uncertain about many scene details, although only a subset can change which action it should execute. We propose a wrist-camera policy that queries views according to their expected reduction in action regret. The policy compares affordance hypotheses against a small library of safe actions and stops looking when the expected benefit falls below the cost of another view. We will test whether this criterion saves sensing motions at matched task success on paired, partially observable manipulation scenes.

Keywords / active perception, manipulation, value of information, wrist cameras

Problem and prior work

A robot asked to place a part in a keyed slot may see the slot but miss its orientation. A second view matters if the ambiguity changes the insertion direction, while revealing an irrelevant occluded object offers little value. Task-sufficient active perception [1] already asks when evidence is adequate for a symbolic decision. Our narrower question is whether scoring views by the magnitude of downstream action loss improves on decision sufficiency or uncertainty reduction, especially when equally uncertain hypotheses have unequal failure costs.

Hypothesis and method

Maintain a categorical belief over target identity and the geometry required by an available skill, such as handle orientation or slot axis. A frozen perception model provides likelihoods from calibrated RGB-D observations. For each candidate wrist pose, a lightweight model predicts likely observations and updates the belief. A decision module scores the current and hypothetical post-view beliefs by the minimum expected action loss over feasible skill choices. Query the pose with the largest expected loss reduction net of travel cost. Execute when every pose has non-positive net value, or abstain when the remaining risk exceeds a fixed threshold. Use the same collision-checked pose library and controller across methods.

Discriminating example

Construct two visually matched settings in which uncertainty about the same occluded feature changes the optimal grasp in one setting and leaves it unchanged in the other. Also vary the penalty of selecting the wrong grasp independently of appearance. The policy should selectively query when an alternative action would avoid consequential failure.

Validation plan

Experimental design

Build three task families in ManiSkill3 [2], with hidden target identity, hidden articulation direction, and hidden insertion geometry. Cross each with irrelevant occlusion and vary whether the decisive cue is observable from an admissible pose. Start with 240 held-out paired scenes stratified by task, holding out object meshes and layouts by group. True simulator state labels affordances and outcomes and stays outside the online controller. Compare single-view execution, a fixed two-view policy, entropy-based view selection, a task-sufficiency stop rule, and an oracle value-of-information upper bound. Match sensing budgets by tracing success-cost curves.

Outcomes and rejection criteria

Primary endpoints are manipulation success and extra camera travel at a prespecified noninferiority margin of three percentage points in success. Use scene-paired bootstrap intervals. Report unsafe or failed attempts, abstention coverage, and the interaction between decision relevance and query rate. The hypothesis fails if camera travel does not improve at matched success, or if the policy queries irrelevant occlusions as often as action-changing occlusions under the matched design.

Feasibility and deliverables

Weeks 1-2 build matched scenarios and the pose library. Weeks 3-4 calibrate perception and fit the view model from simulated observations. Weeks 5-6 freeze thresholds and run held-out tests. If view predictions are poorly calibrated, use empirical observation likelihoods and report the gap to the oracle. Release scene pairs, candidate poses, evaluation scripts, and success-cost curves. The first claim is limited to simulation.

References

[1] When to Make a Decision: Task-Sufficient Active Perception for Mobile Manipulator Robots. arXiv:2610.23891.

[2] ManiSkill3: GPU Parallelized Robotics Simulation and Rendering for Generalizable Embodied AI. arXiv:2410.00425.