Time-Consuming

Existing Simulation Benchmarks

☹ : *no user input*

RoboTwin
LIBERO

Fixed Simulation Evaluation Process

Success Rate: 0.63

☹ : *no other explanation*

Efficient

ManipEvalAgent (Ours)

How well does the policy generalize to operated objects?
User

To evaluation the generalize to operated objects, we can first evaluation on tasks about Position generalization

Code Generation 🐍 ......

Execution ......

Based on the observations of 1st round evaluation results, the model do well in Position generalization. Next, we can explore the Appearance generalization

Code Generation 🐍 ......

Execution ......

......

Summary:
Strengths: Good position generalization within ~10–12 cm; robust to color.
Weaknesses: Sensitive to high gloss and ≥1.2× size.
......
Recommendation:......