



Time-Consuming

Existing Simulation Benchmarks



: no user input



RoboTwin
LIGERO

Fixed Simulation Evaluation Process

Success Rate: 0.63



: no other explanation



Efficient

Embodied Evaluation Agent (Ours)



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



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



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




Code Generation 



Execution



Code Generation 



Execution

.....

How well does the policy generalize to operated objects?



User