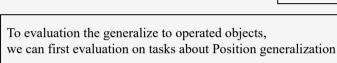
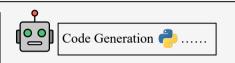


ManipEvalAgent (Ours)

How well does the policy generalize to operated objects?









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



Strengths: Good position generalization within ~10–12 cm; robust to color.

Weaknesses: Sensitive to high gloss and $\geq 1.2 \times$ size.

.....

Recommendation:.....