
(a) Test-time tool ablations
SR (%)
0
20
40
60
80
100
EvoTER (full)
w/o perception
w/o planning
w/o VLA
5/5 5/5 5/5 5/5 4/5 4/5 5/5 5/5 5/5 3/5
92%
3/5 3/5 4/5 4/5 2/5 2/5 4/5 1/5 2/5 0/5
50%
4/5 4/5 2/5 3/5 3/5 1/5 3/5 0/5 0/5 0/5
40%
0/5 0/5 4/5 2/5 1/5 0/5 0/5 4/5 3/5 0/5
28%
PackIdenticalLunches
PortionHotDogs
PrepareCoffee
CoffeeSetupMug
PanTransfer
StirVegetables
StackBowlsCabinet
RecycleBottlesByType
GatherTableware
WeighIngredients
Mean SR
(b) Privileged tools in evolution
w/o privileged
w/ privileged
Success rate (%)
100
50
0
2/5
5/5
3/5
5/5
3/5
5/5
2/5
5/5
0/5
4/5
GatherTableware
SearingMeat
TurnOffStove
PortionHotDogs
PanTransfer