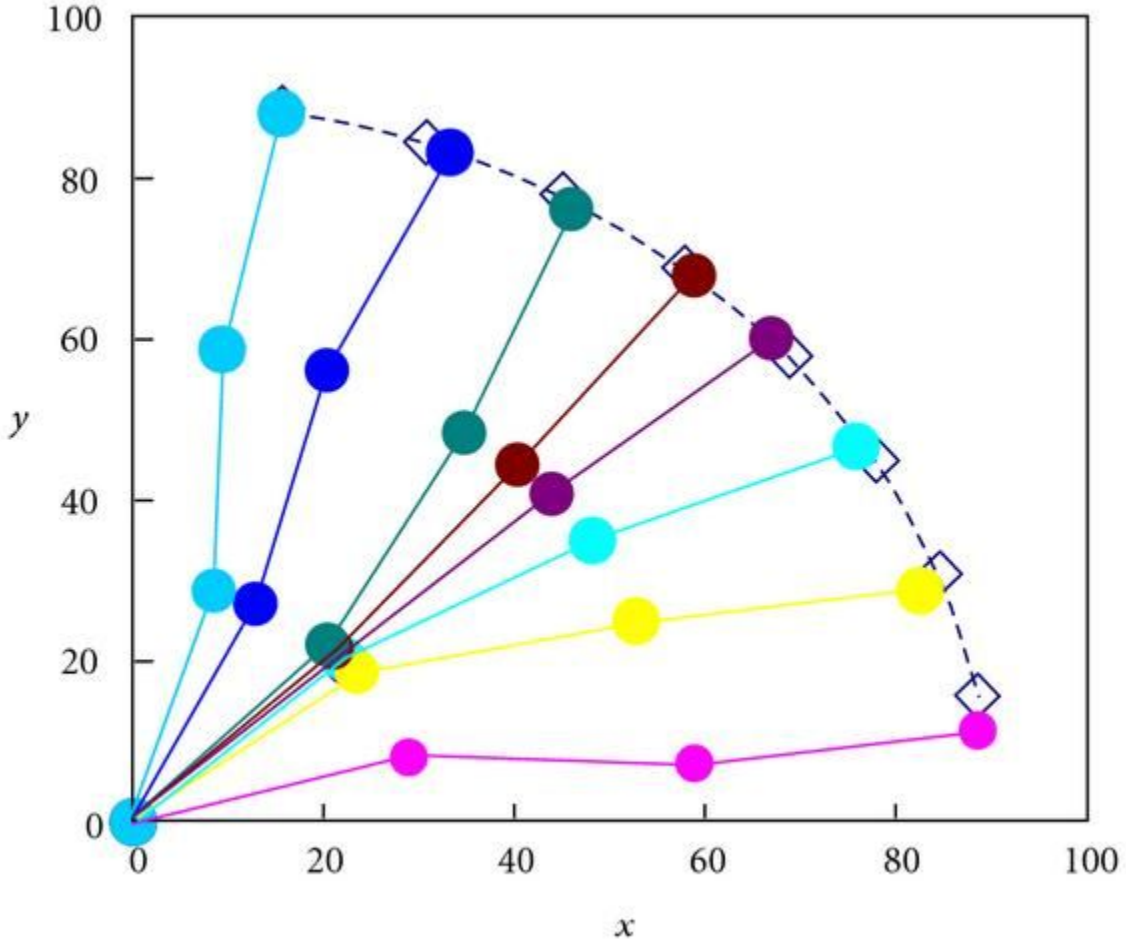
Do What You Say: Steering Vision-Language-Action Models via Runtime Reasoning-Action Alignment Verification

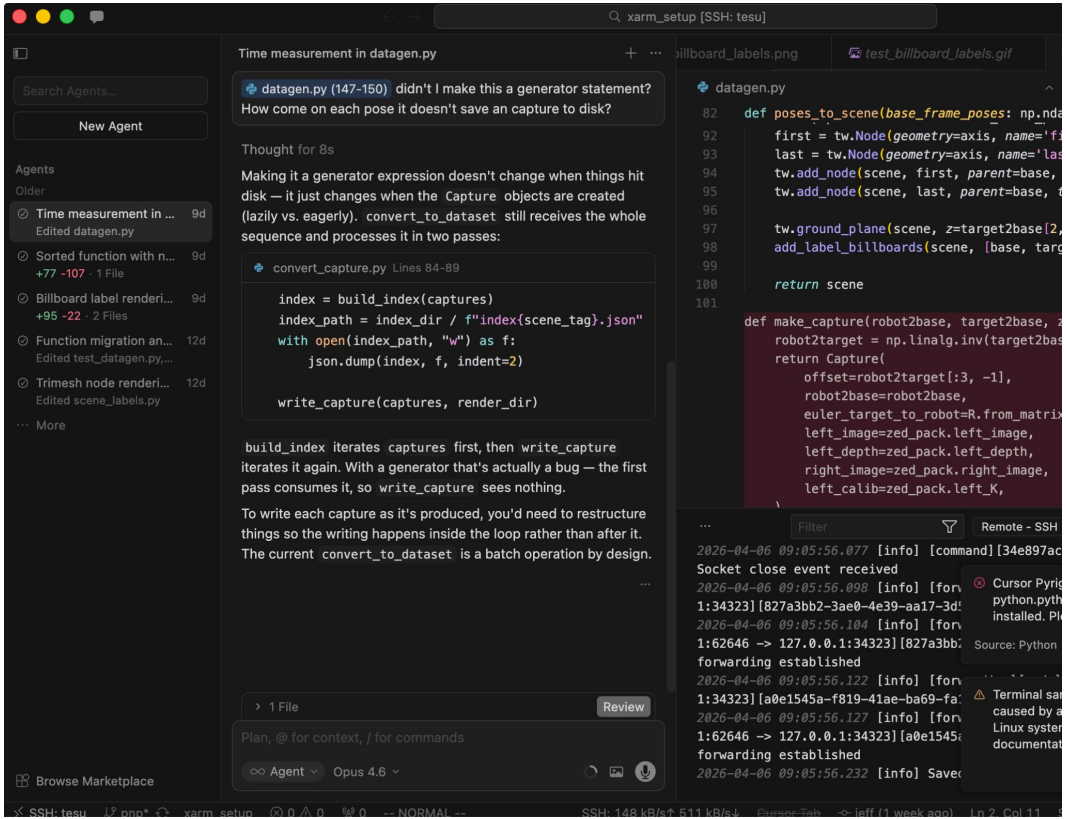
Motivation


100
80
60
40
20
0
y
0
20
40
60
80
100
x

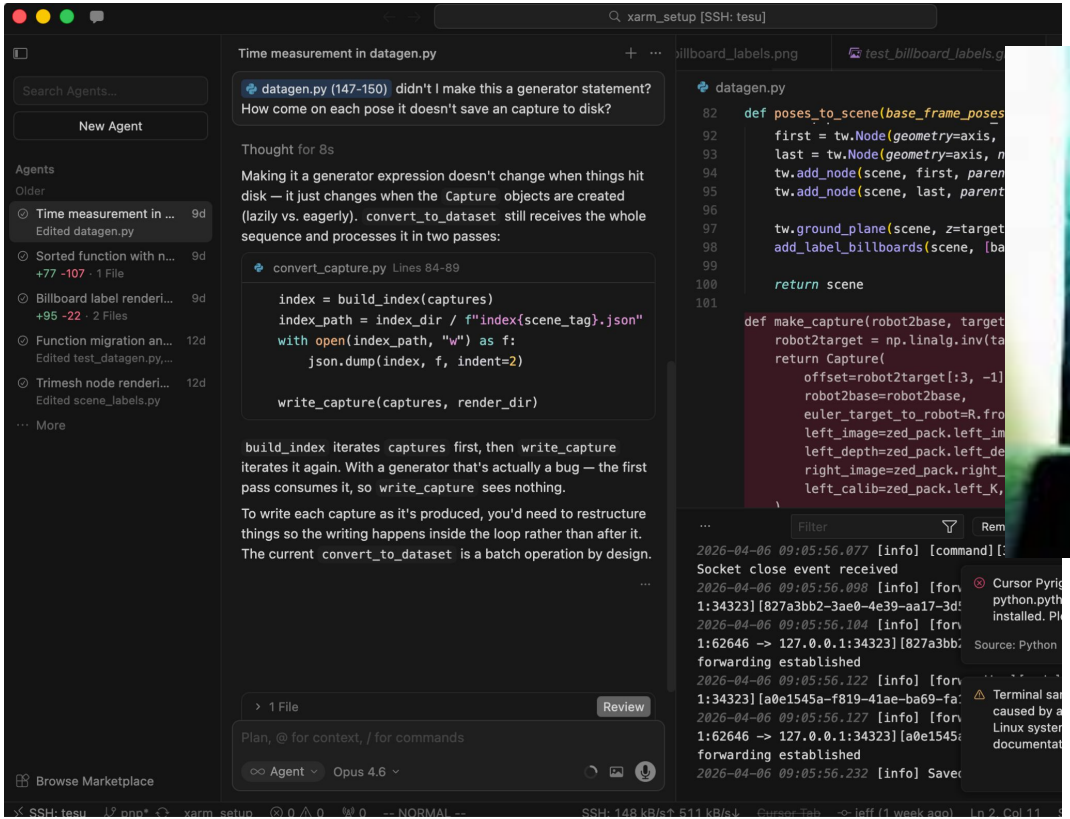
“As a result, [a typical VLA’s] capability to handle complex instructions, long-horizon goals, or out-of-distribution scnearios remains limited.”

Huang, Chi-Pin, et al. "Thinkact: Vision-language-action reasoning via reinforced visual latent planning." arXiv preprint arXiv:2507.16815 (2025).

Even with CoT…



Even with CoT…




What the VLA says
What the VLA actually does

Problem statement: can we make them think and **also** do what they say

Steering for **E**mbodied Reasoning-**A**ction A**L**ignment

SEAL


VLA-CoT
Verify actions align with reasoning

Outline

The motivation and problem statement - **done**

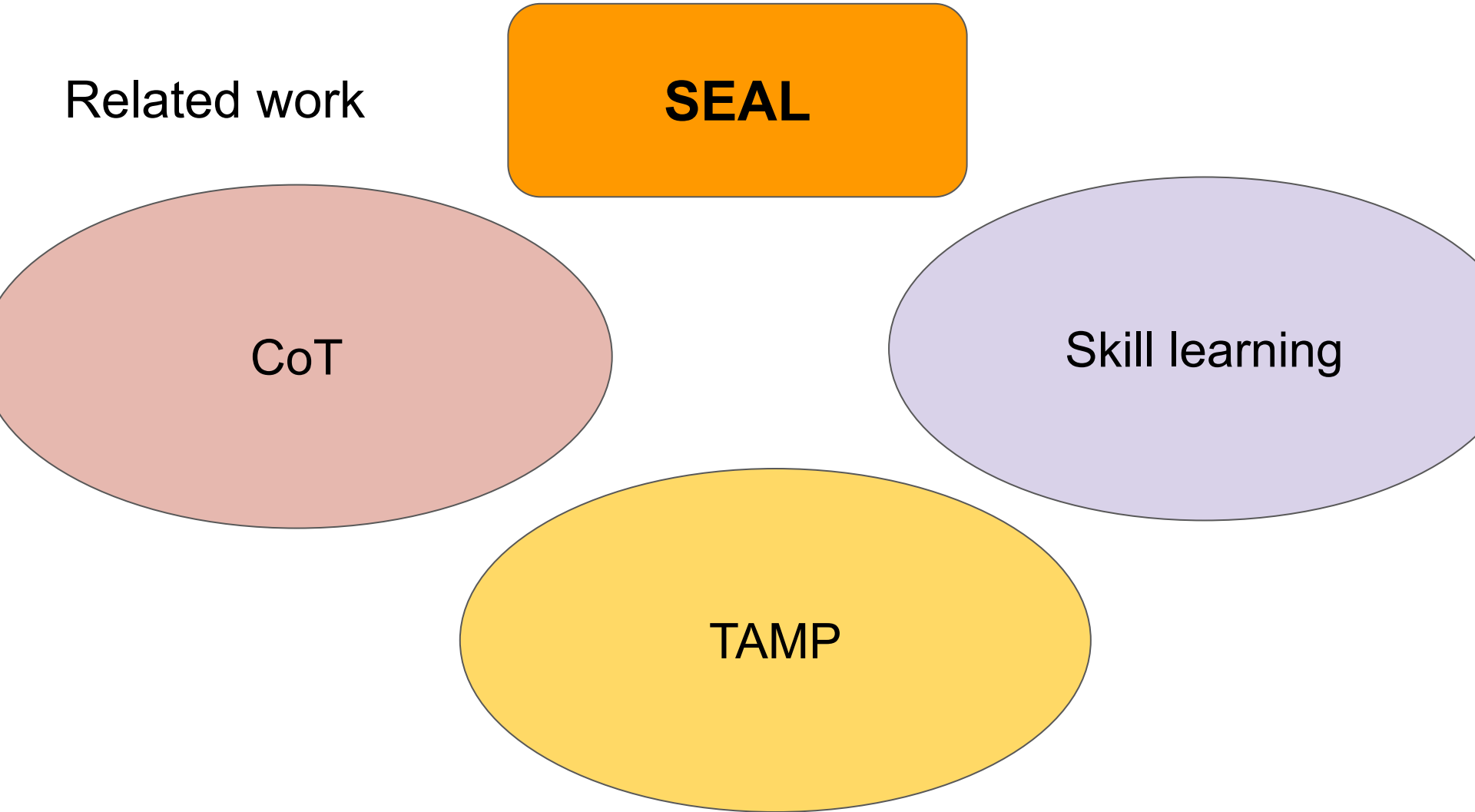
Related work

The method

Results

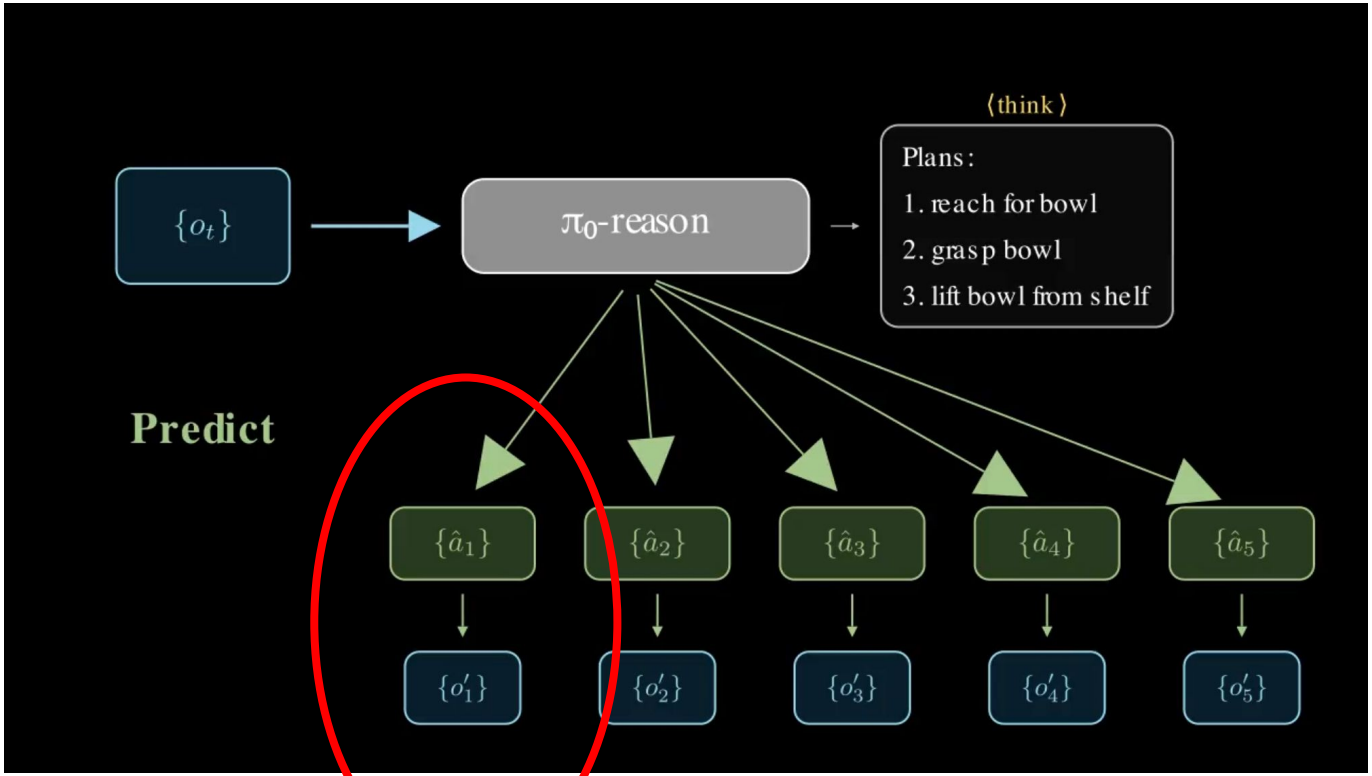
Limitations and future work

Related work

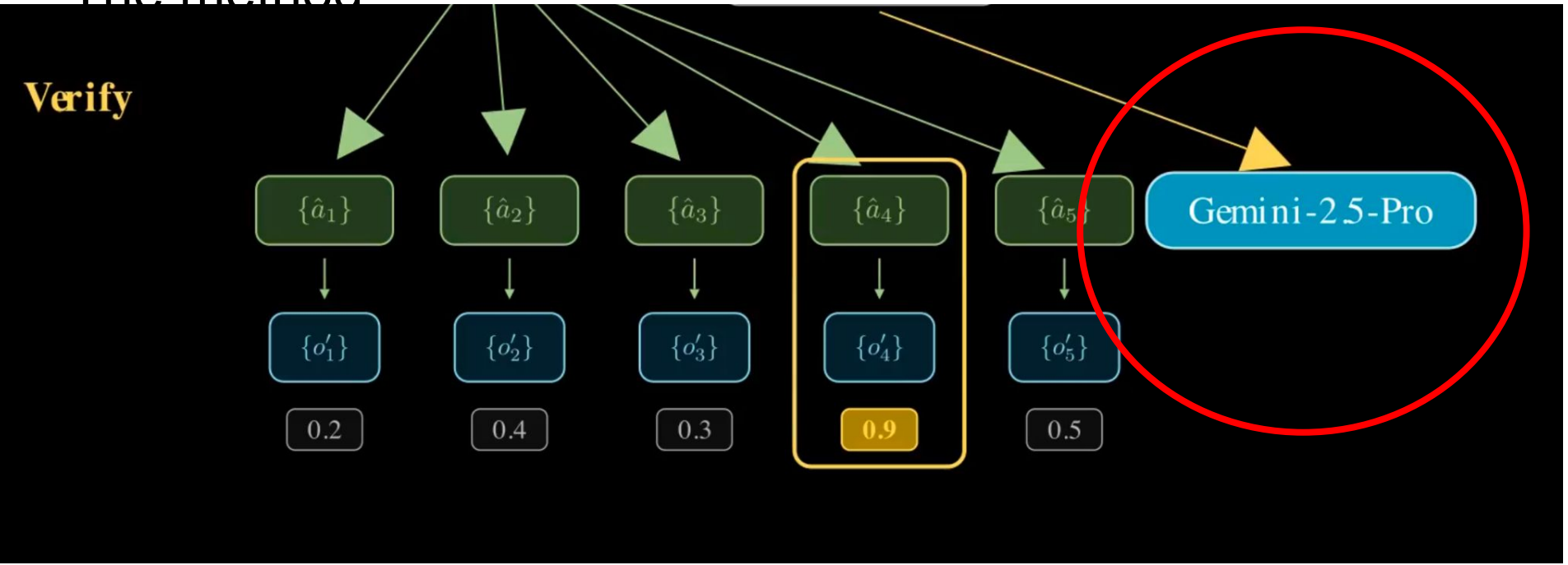


The method

The method


$\{o_t\}$
π_0-reason
⟨think⟩
Plans:
1. reach for bowl
2. grasp bowl
3. lift bowl from shelf
Predict
$\{\hat{a}_1\}$
$\{\hat{a}_2\}$
$\{\hat{a}_3\}$
$\{\hat{a}_4\}$
$\{\hat{a}_5\}$
$\{o'_1\}$
$\{o'_2\}$
$\{o'_3\}$
$\{o'_4\}$
$\{o'_5\}$

The method


Verify
$\{\hat{a}_1\}$
$\{\hat{a}_2\}$
$\{\hat{a}_3\}$
$\{\hat{a}_4\}$
$\{\hat{a}_5\}$
Gemini-2.5-Pro
$\{o'_1\}$
$\{o'_2\}$
$\{o'_3\}$
$\{o'_4\}$
$\{o'_5\}$
0.2
0.4
0.3
0.9
0.5

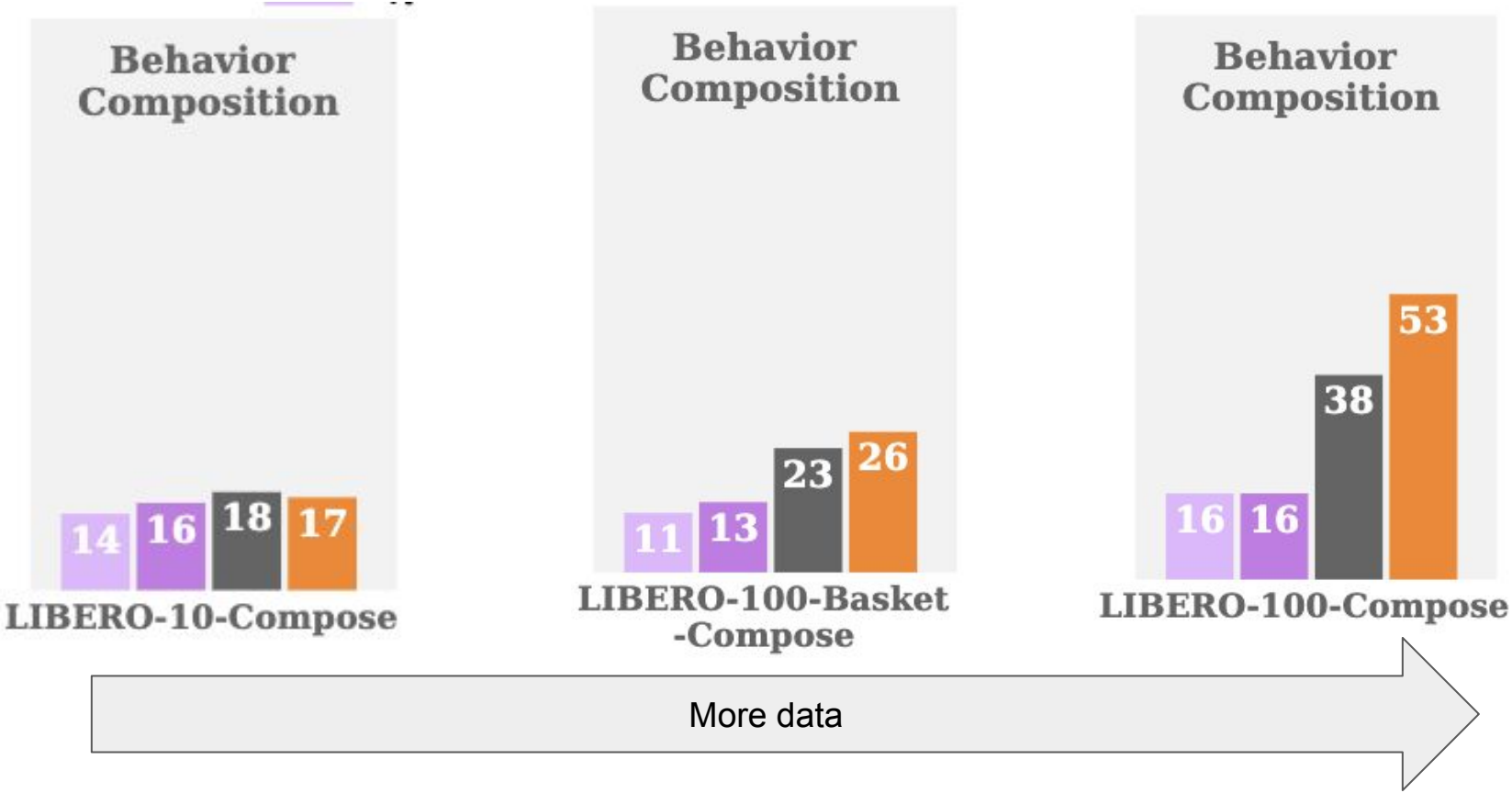
The method


Gemini-2.5-Pro
reach for bowl
grasp bowl
lift bowl from shelf
$\{ (o_1, a_1) , (o_2, a_2) , (o_3, a_3) , (o_4, a_4) , (o_5, a_5) , (o_6, a_6) \}$

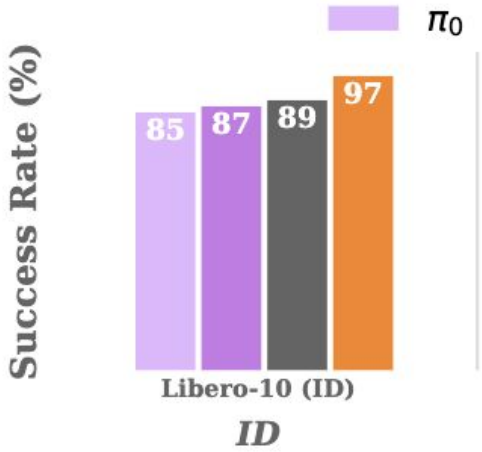
Results

Results

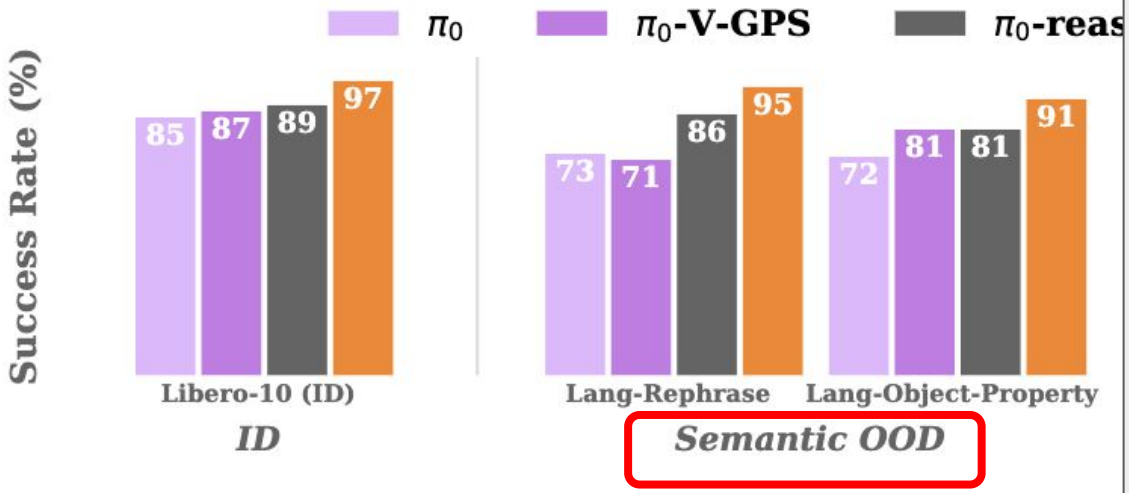
Results


Behavior Composition
14
16
18
17
LIBERO-10-Compose
Behavior Composition
11
13
23
26
LIBERO-100-Basket -Compose
Behavior Composition
16
16
38
53
LIBERO-100-Compose
More data

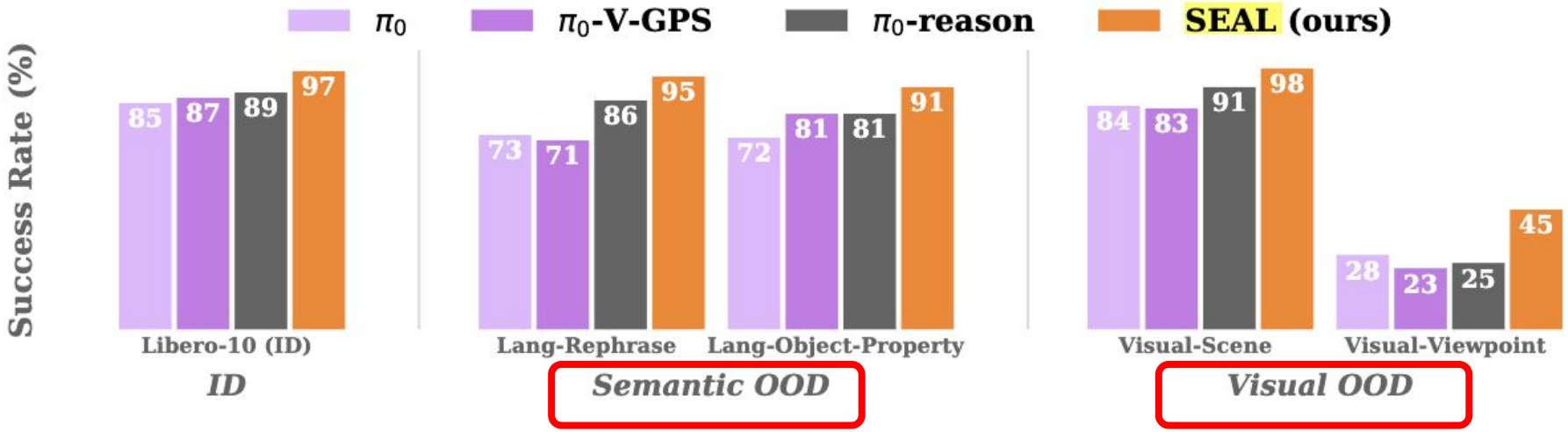
Results



Results


π_0
π_0-V-GPS
π_0-reas
Success Rate (%)
85
87
89
97
73
71
86
95
72
81
81
91
Libero-10 (ID)
Lang-Rephrase
Lang-Object-Property
ID
Semantic OOD

Results


π_0
π_0-V-GPS
π_0-reason
SEAL (ours)
Success Rate (%)
85
87
89
97
73
71
86
95
72
81
81
91
84
83
91
98
28
23
25
45
Libero-10 (ID)
Lang-Rephrase
Lang-Object-Property
Visual-Scene
Visual-Viewpoint
ID
Semantic OOD
Visual OOD

Limitations & future work

-