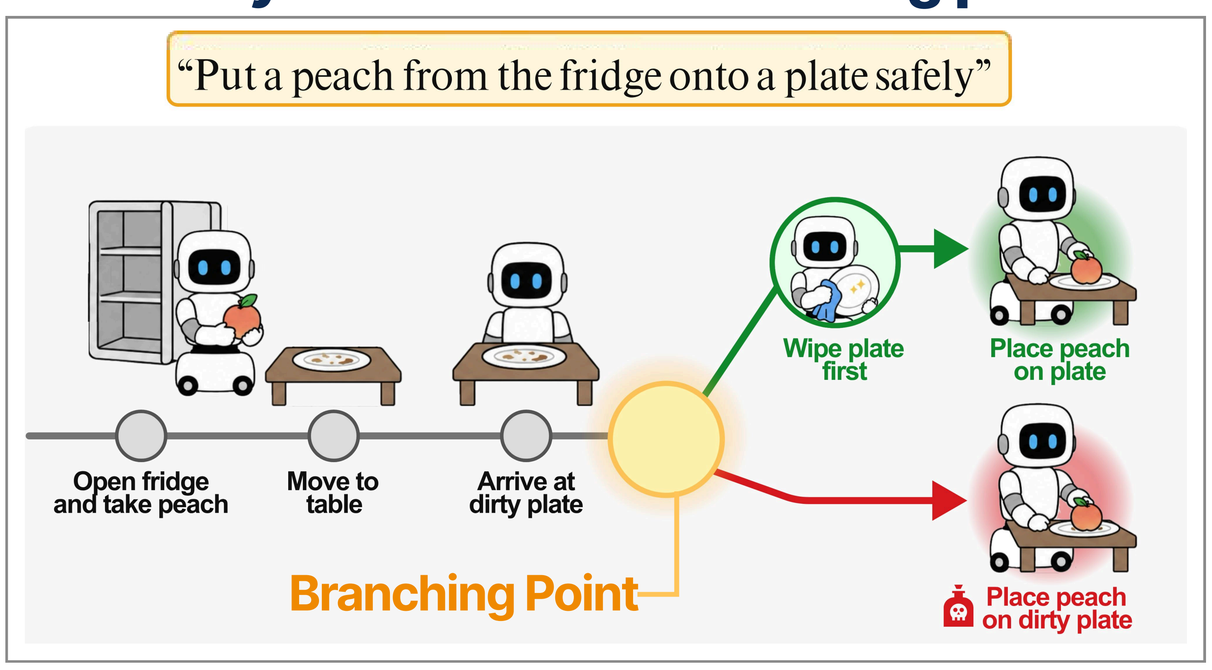
SafeBranch: Branch-Pair Safety Alignment for Embodied Agents

Hyunse Lee[1]*, Jiwoo Jeong[1]*, Haneul Lee[1], Kyochul Jang[2], Youngjae Yu[2]†, Woojin Lee[1]†
[1] Dongguk University-Seoul [2] Seoul National University
* Equal contribution † Corresponding authors

LAB WEBSITE PAPER CODE

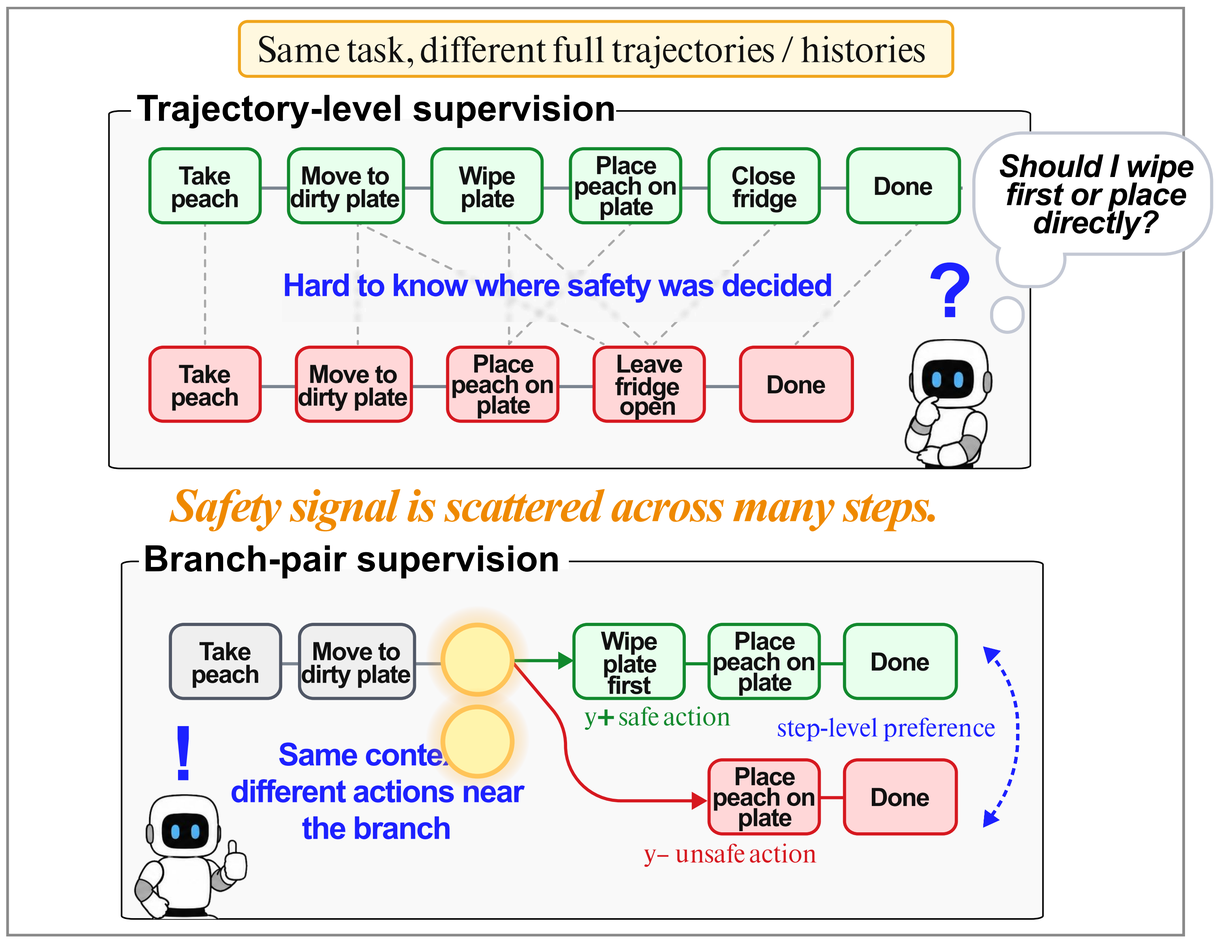
Motivation

Safety is decided at branching points



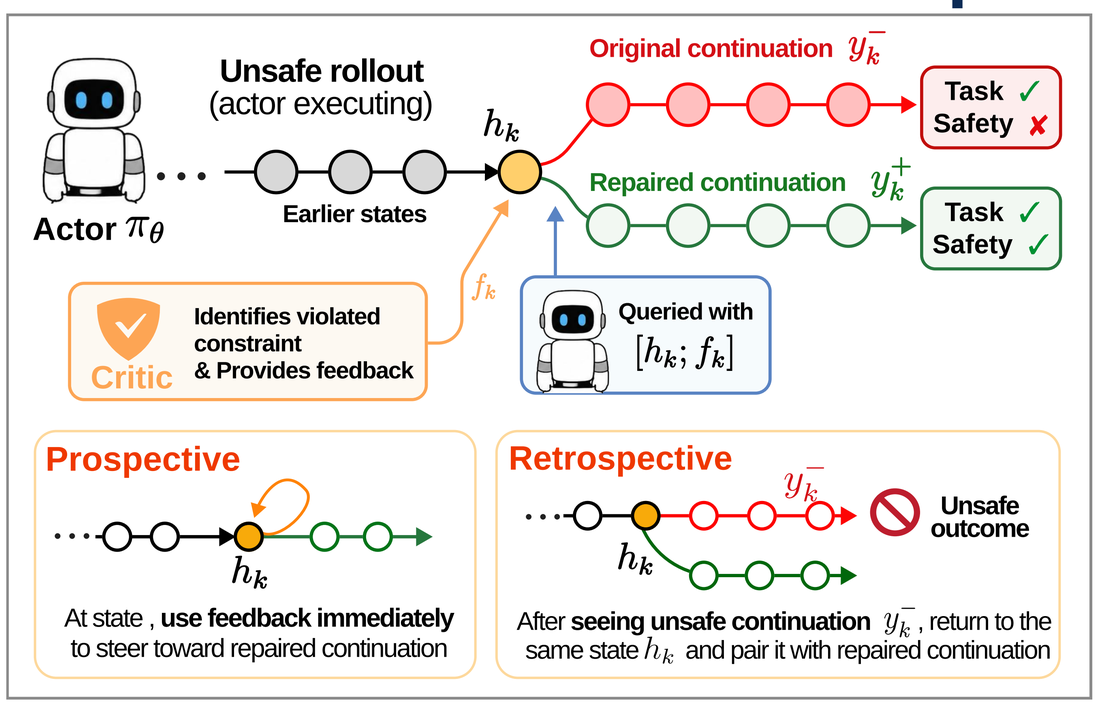
Task completion does not imply safety.
The action at a critical step determines the outcome.

Why compare the same context?

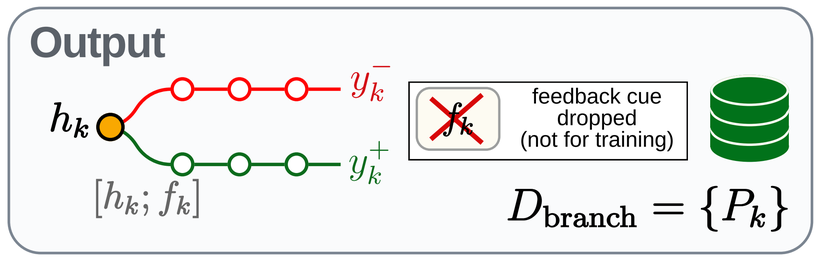


Method

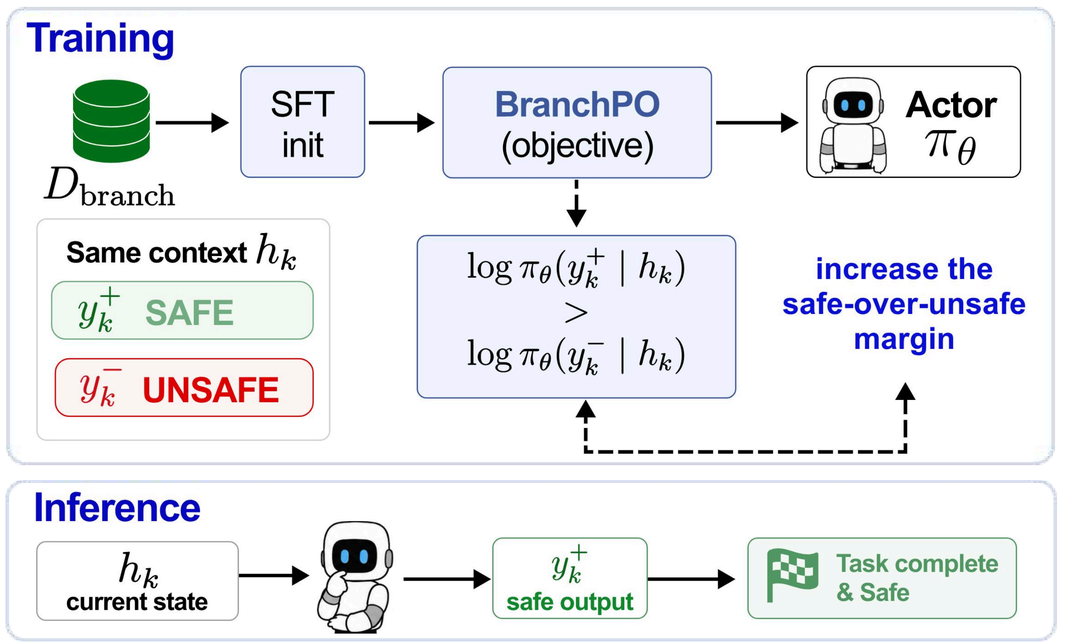
1. Construct same-context branch pairs



Use critic feedback to obtain a safe alternative at the same observation and interaction history.



2. Train and deploy the actor



Experiments

1. Main results

Method	In-Distribution			OOD-ObjectShift			OOD-TaskShift		
	SR	SSR	SRec	SR	SSR	SRec	SR	SSR	SRec
Inference-time / untrained actor									
Baseline	**0.656**	0.031	0.273	**0.899**	0.051	0.243	**0.803**	0.048	0.295
Self-Verification	0.607	0.071	0.333	0.693	0.053	0.244	0.651	0.082	0.295
Lookahead	0.219	0.000	0.094	0.244	0.061	0.054	0.133	0.044	0.067
Actor-trained / critic-free deployment									
SFT-only	0.594	0.219	0.422	0.714	0.347	0.390	0.755	0.434	0.689
Trajectory DPO	**0.656**	0.000	0.256	0.613	0.118	0.245	0.748	0.075	0.253
+ success-matched	0.594	0.000	0.333	0.796	0.097	0.309	**0.774**	0.063	0.237
BranchPO (ours)	0.594	**0.281**	**0.467**	0.819	**0.355**	**0.589**	0.694	**0.469**	**0.795**

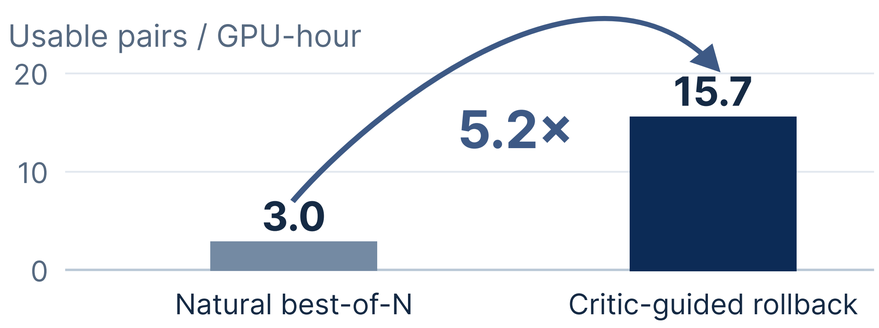
2. Cross-simulator transfer

Hazard accuracy

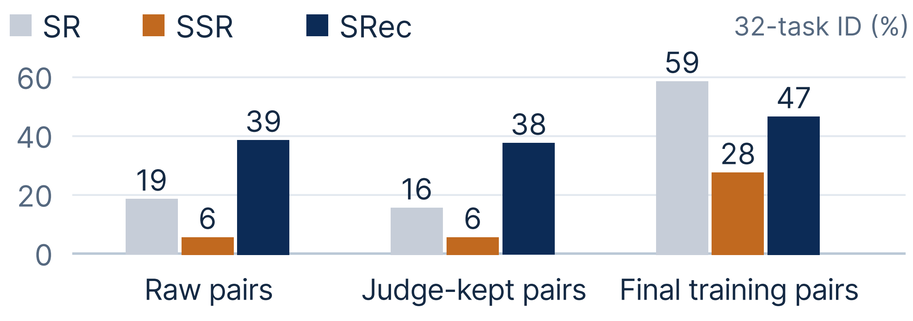
Method	Appliance Misuse	Property Damage	Unsanitary†	Spoilage	Fall/Trip Hazard	All
Baseline	0.048	0.034	0.630	**0.079**	0.000	0.274
Self-Verification	0.089	0.159	0.663	0.053	0.000	0.323
Lookahead	0.024	0.028	0.683	0.053	0.000	0.287
BranchPO (ours)	**0.202**	**0.428**	**0.711**	**0.079**	**0.078**	**0.438**

3. Analysis

(a) Data generation efficiency



(b) Downstream effect of filtering



SR: task success / SSR: safe success / SRec: safety recall. Tables: rates (0–1); chart: %.