

Paper review · Visionary

Teacher recoverability in Dagger for VLAs

Ross, Gordon and Bagnell, DAgger, AISTATS 2011 (arXiv:1011.0686).

Pavel Bushuyeu · October 2026

Which states should a VLA teacher label?

DAgger needs an expert that recovers

Theorem 3.2: $J(\text{D}Agger) \leq J(\text{expert}) + u \cdot T \cdot \epsilon + O(1)$.

u is the most that one learner action can add to the expert's cost.

u stays small when the expert can recover. A VLA teacher may fail from the student's states.

Two ways to pick states

Kintsugi-VLA picks states that an expert with the full simulator state can recover from.

PIVOT-Q labels only the states where teacher and student disagree most.

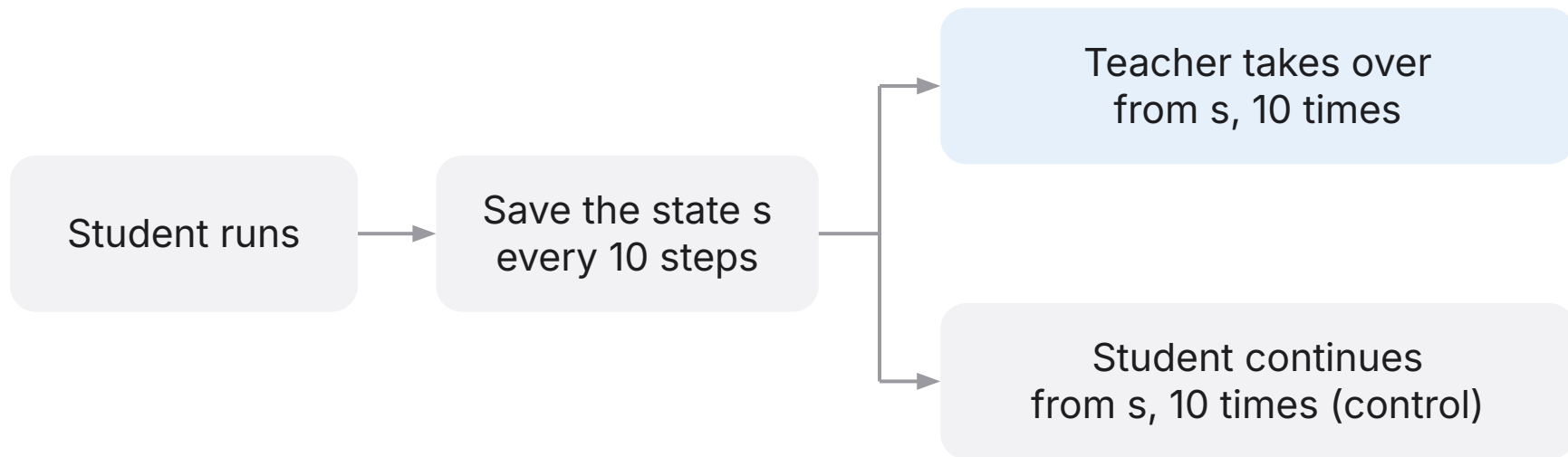
I found no comparison of the two for a VLA teacher.

Hypotheses

- H1** From the student's states, the teacher finishes the task less often than from normal start states.
- H2** Recoverable states give better labels than high-disagreement, all or random states, at the same label count or cost.
- H3** Labels on the student's states help more when the student copies the teacher badly.

If H1 holds, a VLA teacher's labels need a check. If H2 holds, the check pays for itself.

Measuring teacher recoverability



$r(s)$ is the share of teacher runs that finish the task in the time left, with a Wilson interval. $r(s)$ is related to u , but it is not u .

Design

Teacher	π 0.5 LIBERO checkpoint (openpi), frozen.
Student	SmoIVLA (450M), first trained on 10% or on 50% of the demos.
States	All (DAgger), most disagreement, recoverable, or random.
Recovery	10 teacher runs per state, success in the time left.
Budgets	The same number of labels, or the same total cost.
Tasks	LIBERO (4 suites), LIBERO-Plus camera and layout shifts.
Measures	Student success per label and per cost, r(s) by task and time.

Compute and risks

Compute (estimate)

3 rounds $\times$ 300 states $\times$ (10 teacher + 10 student runs) = 18,000 runs.

With training and evaluation, about 200 GPU-hours on one RTX 5090. The pilot measures the real speed first.

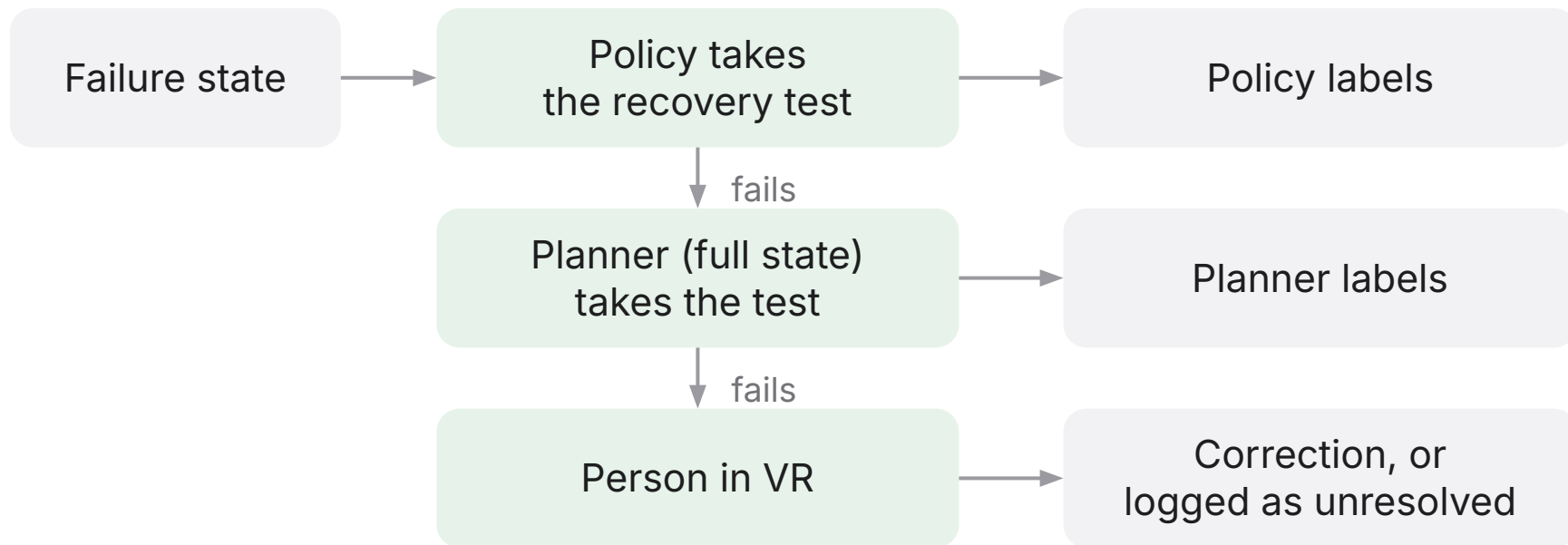
Risks

A strong teacher (92.4% on LIBERO-Long) may recover from almost every state.

The pilot checks that a restored state replays exactly.

PIVOT-Q already picks states by disagreement. The new part is recoverability.

A failure router



Each failure goes to the cheapest expert that passes the recovery test.

Outputs and prior work

For household tasks in simulation

- Recovery data per failure type.
- Human time only after models fail.
- A table per task of which expert fixes which failures.

Prior work

Kintsugi-VLA picks recovery starts with one expert that sees the full simulator. In HG-Dagger and RaC, a person makes every correction.

The 2025 BEHAVIOR winner had no recovery data. It used hand-written correction rules.

Pilot first, then the router

Pilot (go or no-go)

LIBERO-Long, the student before DAgger, 300 saved states by task, time and outcome.

Go on if a restored state replays exactly and $r(s)$ varies between states. If $r(s)$ is high everywhere, H1 is false.

Router test

2–3 household tasks in simulation, with a policy, a planner with full state and a person in VR.

Policy only, person only and the router at the same cost, by student success and human minutes.

What this builds on

Work	What it does	What this adds
DAgger, 2011	Labels every state. The bound has u .	A recovery test.
Kintsugi-VLA, 2026	Recovery test with a full-state expert.	VLA teacher, DAgger labels.
PIVOT-Q, 2026	Labels only high-disagreement states.	Picks by recoverability.
VLA-OPD, 2026	A VLA teacher labels student states.	A check of those labels.
HG-DAgger, RaC	A person takes over and corrects.	People only after models fail.
Reward-DAgger, 2026	A reward model triggers $\pi_{0.5}$ takeovers.	Which states $\pi_{0.5}$ can fix.