Paper review · Empiricist

View, scene and instruction in π0.5-DROID

DROID, Khazatsky et al., RSS 2024 (arXiv:2403.12945).

Pavel Bushuyeu · November 2026

Input sensitivity of π0.5-DROID

The question

DROID argues that diverse scenes, tasks and camera viewpoints make policies generalize.

π0.5-DROID is Physical Intelligence's open π0.5, fine-tuned on DROID.

Without a robot: how much do its actions depend on the camera view, the scene and the instruction?

Hypotheses (set before the run)

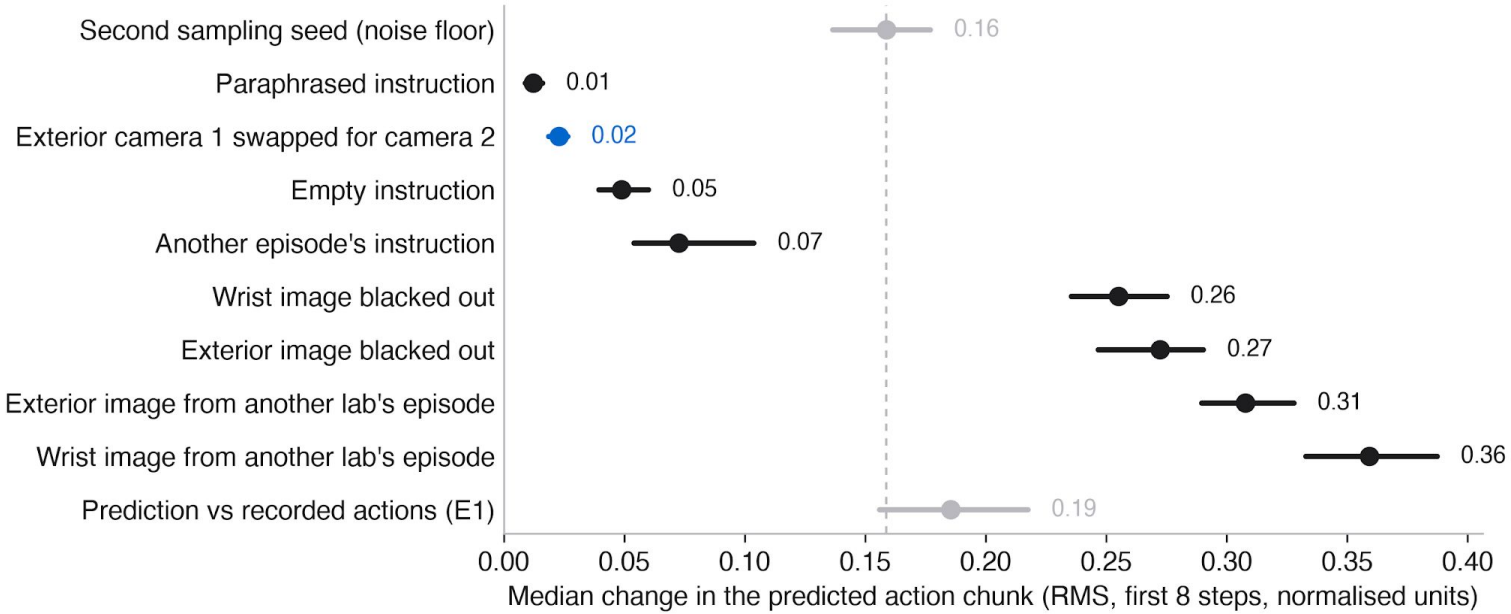
H1: camera diversity buys view invariance. Exterior camera 2 instead of camera 1 changes the actions about as much as a new sampling seed.

H2: language matters early in an episode. A wrong instruction moves the actions well above the noise early, and less late.

Offline probes on 500 DROID states

Model	pi05_droid (openpi): exterior + wrist image, state, instruction → 15 actions.
Data	droid_100: the first 100 DROID episodes, 13 labs; 5 states per episode.
Probes	Other camera, other or empty instruction, paraphrase, black or other-lab images.
Yardstick	Noise floor: mean distance between three sampling seeds on one input. Probes share one seed.
Measure	RMS change over the first 8 actions, normalised units; medians, episode bootstrap.
Compute	RTX 3090 (iolani-5): 5,140 calls in about 28 min (177 ms median inference).

Action change by probe



Instruction probes: 230 states (paraphrase 155), each against its own noise floor.

Camera swap vs another scene's image

Camera 1 → camera 2

Same seed: change 0.023, 0.15× the noise floor [0.12, 0.16].

Above the floor in 1.4% of states.

Small by phase (0.10–0.18×) and by lab (0.06–0.24×).

Error against the recorded actions: median change 0.000.

Control: another lab's scene

Change 0.31: 1.9× the floor, 13× the swap.

Error against the recorded actions rises by 0.14.

H1's pre-set rule expected about 1×; it leaves 0.15× unclassified. The model uses the exterior image, but barely notices which recorded view it gets.

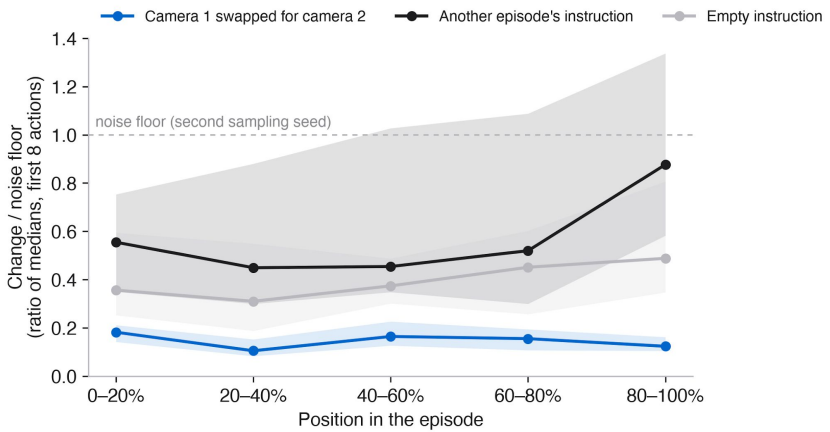
Language effect below sampling noise

Instruction (230 states)	× noise floor
Another episode's	0.58 [0.43, 0.85]
Empty	0.39 [0.32, 0.48]
Paraphrase (155 states)	0.09 [0.07, 0.12]

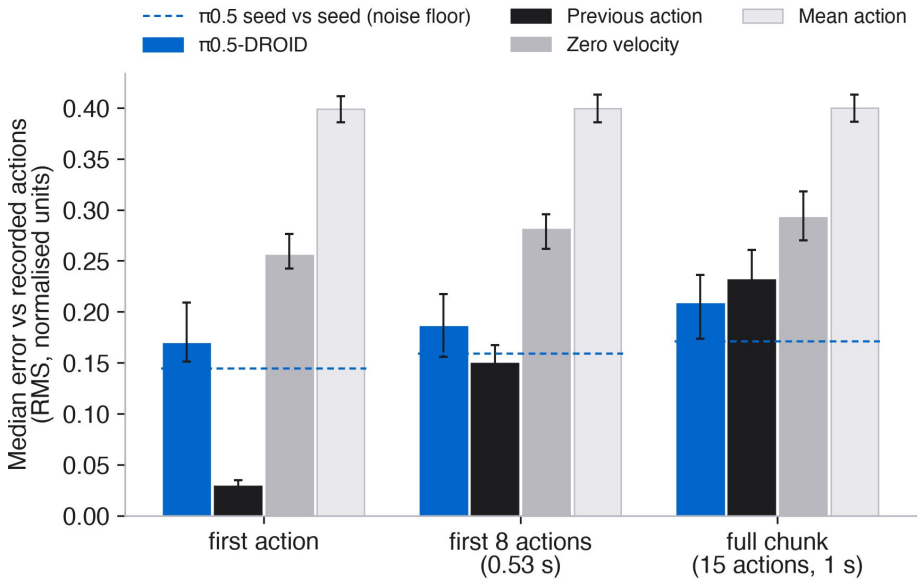
H2 fails: early 0.51, late 0.82 (difference not significant).

19% of states change by more than 2× the floor.

The true instruction still lowers the error.



Agreement with the recorded actions



Error 0.185 = 1.17× the noise floor: open-loop error is close to the model's own sampling spread.

Repeating the previous action wins the first 8 steps; π0.5 wins over the full 1 s chunk.

Failed episodes: error 0.37; successful: 0.15.

Surprises in droid_100

What the sample holds

54 of 100 episodes have no instruction, including all 19 failures.

droid_100 is RLDS v1.0.0: instructions for about 30k episodes; v1.0.1 has about 75k (openpi).

One lab (TRI) has 29 of the 100 episodes.

What it means

Instruction probes use 230 states from the 46 annotated episodes.

Use v1.0.1 for language work.

A 100-episode slice does not show the diversity of the full dataset.

What matches and what does not

Claim	DROID paper	This probe
Viewpoints	Viewpoint diversity helps with new camera angles; policies use both exterior cameras.	Consistent, in-distribution: the two recorded views give nearly the same actions.
Scenes	Scene diversity helps; closed-loop tests after co-training with task demos.	Actions depend on the image (1.9–2.3 ×). Transfer not tested.
Language	86 tasks with instructions; openpi: "good language-following".	Offline, a wrong instruction moves actions less than sampling noise in the median state.

Offline sensitivity, not robot behaviour

Challenges

A shared GPU: the run waited 69 min for one of two TTS jobs.

Disk: 24 GiB at peak on a 95% full disk; all deleted after.

The training recipe is inferred from openpi's open loader.

All states are in DROID: the model may have trained on these frames.

Insights

Compare with the policy's own sampling noise: an error of 0.185 alone looks large, but it is 1.17× the noise.

Offline, a wrong instruction moves the actions less than sampling noise in most states; closed-loop language following is not tested.

Held-out views, same-scene tasks

Hypothesis: the view invariance holds for camera poses outside DROID, and the language effect rises when two tasks are possible in one scene.

Views: new exterior camera poses on a DROID setup, then the same swap probe.

Language: pairs of feasible instructions in one scene (the cup into the bowl, or out of it).

Measure: the same ratios to the noise floor, then success in closed loop.

Sources

1. Khazatsky et al. DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset, RSS 2024.
2. Physical Intelligence. π0.5: a Vision-Language-Action Model with Open-World Generalization, 2025.
3. Driess et al. Knowledge Insulating Vision-Language-Action Models, 2025.
4. Physical Intelligence. openpi: pi05_droid checkpoint and DROID training pipeline (commit 215abfb).
5. DROID team. droid_100 sample, RLDS v1.0.0, gs://gresearch/robotics/droid_100.