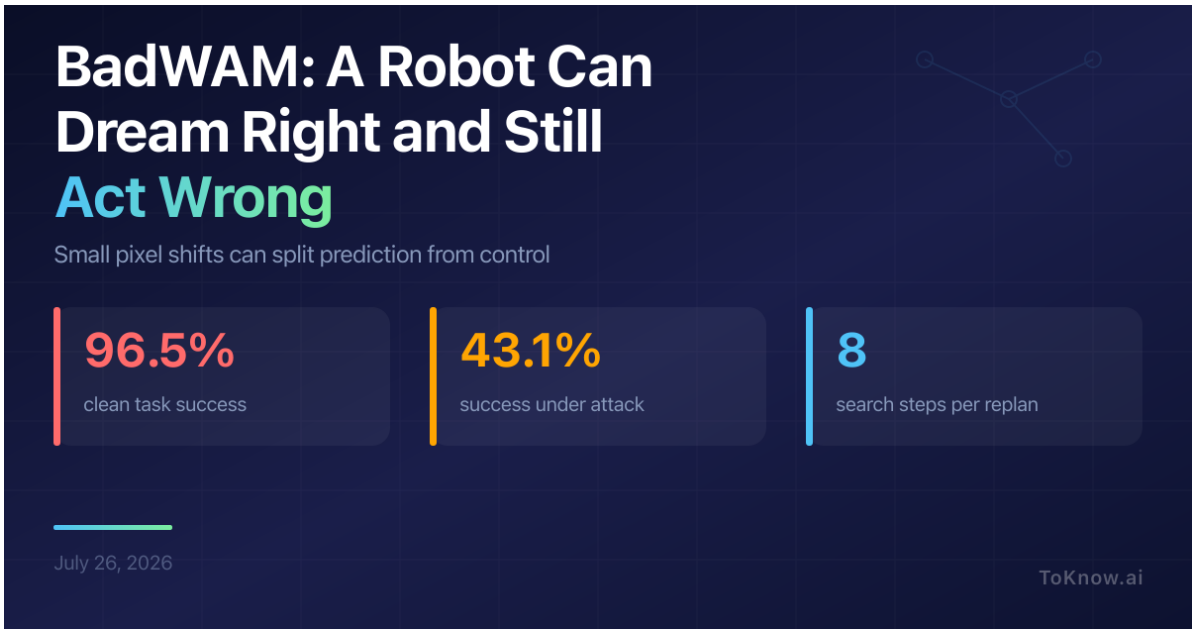
BadWAM: A Robot Can Dream Right and Still Act Wrong

Kabui, Charles

2026-07-26



Researchers at the National University of Singapore and Hong Kong Polytechnic University found that a robot can predict a plausible future and still choose the wrong action. BadWAM attacks world-action models, robot policies that produce both control commands and a preview of what should happen next. The attacker only needs access to model outputs. It changes a small patch of the camera image, with pixel changes capped at 0.06 on normalized images,

then searches for a damaging change over eight optimization steps each time the robot replans. On LIBERO, a simulated tabletop benchmark, the basic model's task success fell from 96.5% to 43.1%. A joint model dropped from 98.1% to 61.5%, while a version that derives actions from predicted motion fell from 98.4% to 66.1%.

That breaks a tempting safety shortcut: checking the generated preview does not prove that the matching command will run. A warehouse robot could predict an object landing in the right bin while its arm path quietly drifts off course. The paper's evidence comes from LIBERO and RoboTwin simulations, not physical robots. The attack also searches repeatedly at every replanning step, and the released code omits the experiment outputs and several large assets. Results varied by benchmark too: on RoboTwin, the basic model moved from 92.1% to 84.4%. Safety monitors should compare predicted futures with the actual commands and observed motion instead of trusting the preview alone.

Read More: NVIDIA Cosmos 3 combines world prediction and robot actions in one model.

Sources:

- BadWAM paper (arXiv)
- BadWAM project page
- BadWAM research code
- BadWAM model collection
- LIBERO benchmark
