TurboVLA: Robot Control at 32 Hz Without an LLM in the Loop

Kabui, Charles

2026-07-31

Read at ToKnow.ai

TurboVLA is a compact robot policy that turns camera views and a written instruction directly into a sequence of movements. Most vision-language-action models route both inputs through a large language model first. TurboVLA instead uses separate vision and text encoders, lets them exchange information through six small attention layers, then predicts a 12-step action chunk in one pass. The authors report 97.7% average success across 2,000 trials on LIBERO,

a simulated tabletop benchmark. Their complete 216-million-parameter policy used 0.9 GB of GPU memory and took 31.2 ms per call on one RTX 4090. In the same hardware tests, the 3.4-billion-parameter 0.5 policy scored 96.9%, used 12.8 GB, and took 93.6 ms.

That speed gives a robot about 32 fresh action chunks per second, which matters when an arm must react before an object moves or slips. The paper also includes four physical-robot tasks, each tested 40 times after fine-tuning on 65 demonstrations. TurboVLA succeeded on 80% to 92.5% of those trials. This is useful evidence that the smaller design transfers beyond simulation, but it does not prove broad real-world reliability: the test used one robot, four tasks, and an RTX 4090. The code is available under Apache 2.0, while the released checkpoints also inherit DINOv3 licence terms.

Read More: BadWAM shows why a robot's predicted future cannot be trusted as a safety check.

Sources:

- TurboVLA paper
- TurboVLA code and evaluation instructions
- TurboVLA model checkpoints
- LIBERO benchmark paper
