Do World Action Models Generalize Better than VLAs? A Robustness Study

Zhanguang Zhang[1] Zhiyuan Li[1,2] Behnam Rahmati[1] Rui Heng Yang[1]
Yintao Ma[1] Amir Rasouli[1] Sajjad Pakdamansavoji[1] Yangzheng Wu[1]
Lingfeng Zhang[1] Tongtong Cao[1] Feng Wen[1] Xinyu Wang[1]
Xingyue Quan[1] Yingxue Zhang[1]

[1]Huawei Technologies [2]University of Toronto

World action models are the most robust to noise, lighting and layout, but lag on camera and robot shifts and run 3–83× slower than $\pi_{0.5}$.

Motivation

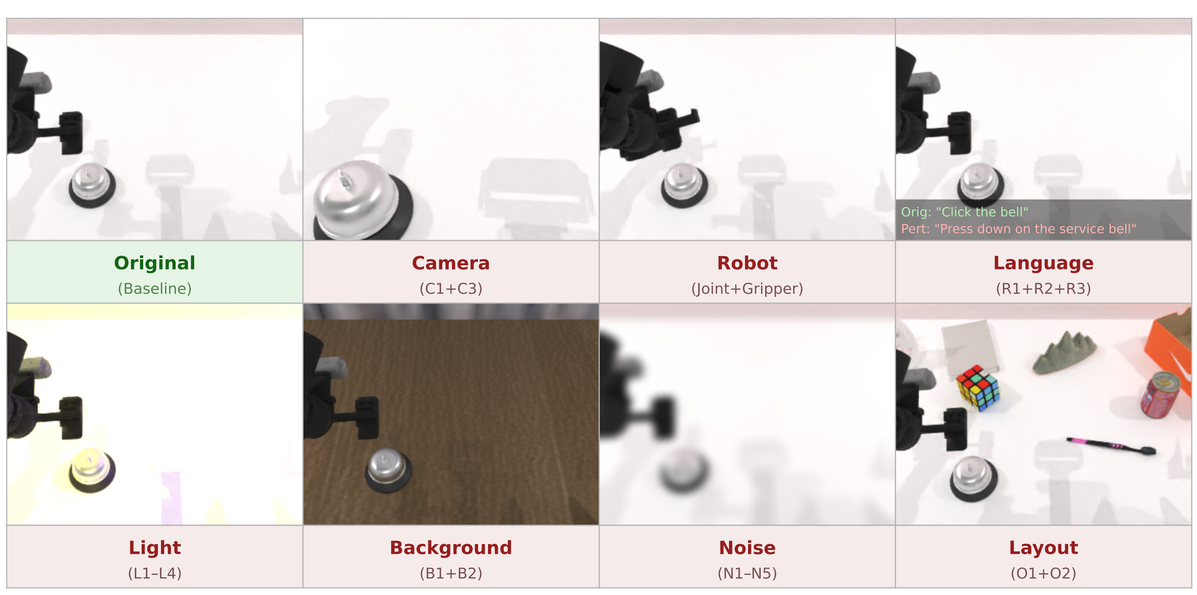
- **VLAs** reuse vision-language models for action, but generalize only as far as their robot data reaches.
- **World action models (WAMs)** start from **video generation models** trained on web-scale video and decode their latents into actions.
- Claim: explicit dynamics plus spatiotemporal priors make WAMs **generalize better**. **Is that true?**

Models compared

- **VLAs**: π_0, π_0-FAST, $\pi_{0.5}$, OpenVLA-OFT, UniVLA, RIPT-VLA, X-VLA, HoloBrain0, ABot-M0.
- **VLA + world model** hybrids: MOTUS, VLA-JEPA.
- **WAMs**: GE-Act (LTX-Video), Cosmos-Policy (Cosmos-Predict2), LingBot-VA and Fast-WAM (Wan2.2).

Benchmarks: 7 perturbation types

- **LIBERO-Plus** (Franka, MuJoCo) and our new **RoboTwin 2.0-Plus** (bimanual Aloha, ManiSkill3), with 21 sub-dimensions.
- Camera, robot state, language, lighting, background, **noise** and object **layout**.



Perturbation examples on RoboTwin 2.0-Plus.

Result 1: RoboTwin 2.0-Plus (success %)

Model	Orig.	Cam.	Robot	Lang.	Light	BG	Noise	Layout	Total
VLAs									
$\pi_{0.5}$	78.4	**45.6**	27.6	74.4	49.6	71.7	64.9	56.8	58.6
X-VLA	65.6	23.2	65.2	64.4	63.1	58.6	49.7	34.8	53.1
VLA + WM									
MOTUS	87.0	21.6	**85.0**	83.2	84.6	84.4	43.1	82.8	71.5
WAMs									
LingBot-VA	**92.1**	28.9	36.2	**87.3**	**89.0**	**91.3**	**80.9**	**87.9**	**74.2**
Fast-WAM	91.2	30.4	53.2	86.7	88.8	90.0	76.4	83.2	72.7

- WAMs lead on **7 of 9** columns, especially **noise and layout**.
- Their weak spots: **camera viewpoint** and **robot initial state**.

The cost: inference speed

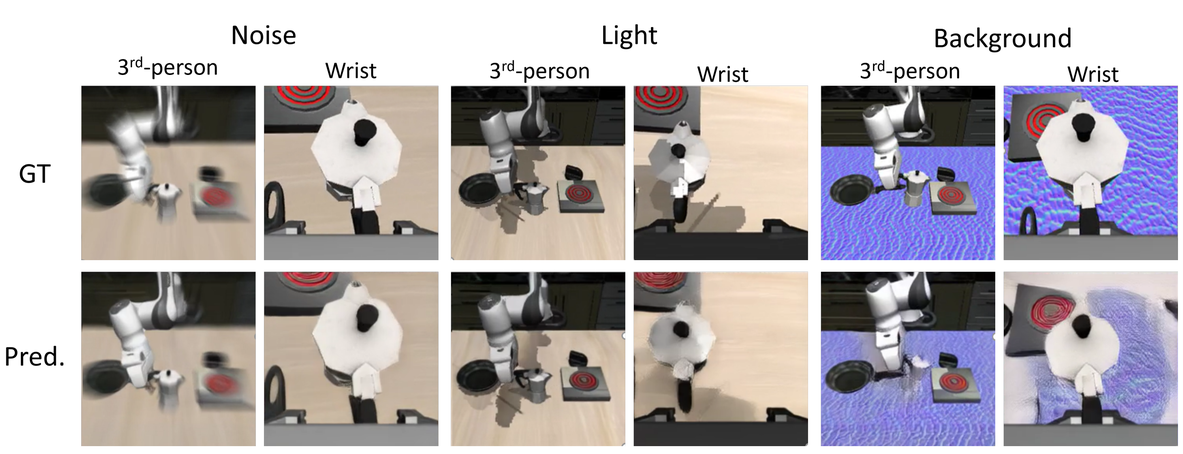
- $\pi_{0.5}$: **63 ms**. WAMs: **3×–83×** slower (Cosmos-Policy 390 ms; LingBot-VA up to 5.2 s).

Result 2: LIBERO-Plus (success %, selected)

Model	Orig.	Cam.	Robot	Lang.	Light	BG	Noise	Layout	Total
VLAs									
π_0	94.2	13.8	6.0	58.8	85.0	81.4	79.0	68.9	53.6
$\pi_{0.5}$	96.9	75.4	77.5	85.6	**96.9**	94.6	89.7	**85.7**	**85.7**
OpenVLA-OFT	97.6	55.6	21.7	81.0	92.7	91.0	78.6	68.7	67.9
X-VLA	98.1	23.4	**89.7**	75.7	88.2	**96.0**	62.7	71.8	71.4
ABot-M0	**98.6**	60.4	67.9	86.4	96.2	91.6	86.4	82.6	80.5
VLA + WM									
VLA-JEPA	97.2	64.2	67.7	**88.1**	91.8	93.4	65.8	83.9	77.9
WAMs									
GE-Act	94.4	60.7	77.0	77.4	95.8	86.0	90.9	80.2	80.3
Cosmos-Policy	98.5	**75.8**	63.3	81.7	96.5	88.9	**92.7**	82.2	82.2
Fast-WAM	97.6	16.4	44.5	68.9	78.2	53.7	37.7	60.7	51.5

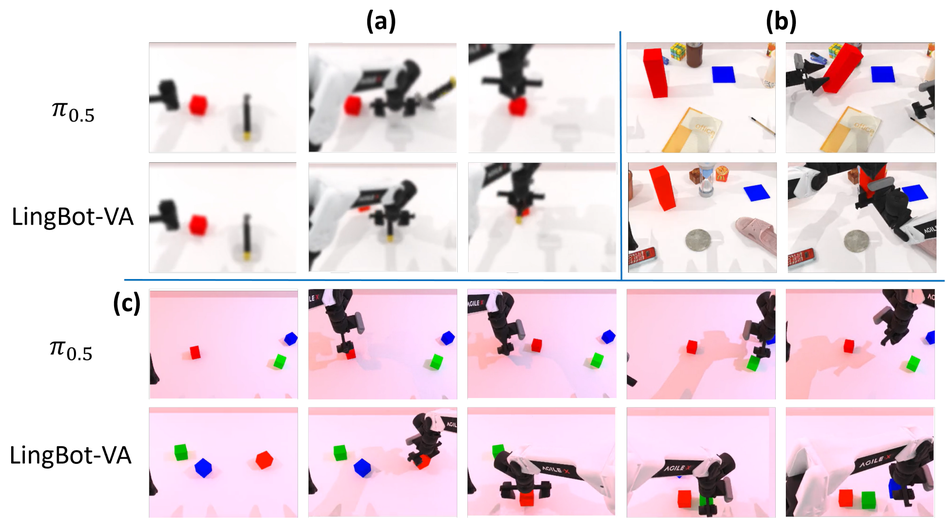
- Cosmos-Policy reaches **82.2%** and is the most robust to noise.
- $\pi_{0.5}$ matches or beats WAMs, but only after **extensive, diverse robot training**.
- Fast-WAM drops here because its LIBERO model saw **only clean demos**: training data still matters.

Why WAMs are robust: they predict the future



Cosmos-Policy's predicted frames stay on track under noise, lighting and background changes.

Case studies: $\pi_{0.5}$ vs. LingBot-VA



Conclusion

- **Video priors help** robustness; hybrids land in between, so *how* video is integrated matters.
- Open problems: viewpoint and embodiment shifts, and **real-time** WAM inference.