THE ACTION GAP / CoRL 2026 WORKSHOP

SmolWAM

A Latent World-Action Model for Robotic Manipulation on Edge Devices

Timi Owolabi · Hikmah Olawore · Frances Adelakun · Ayotomiwa Oyewumi

ML Collective

97%

Of teacher success retained

98.2% → 95.5%

95.5%

LIBERO success
382 / 400 episodes

11×

Lower latency on A100
624 → 57 ms per 16 actions

2.36 GiB

A100 peak GPU memory
3.2× lower than teacher

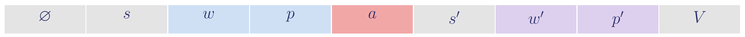
18.5×

Fewer FLOPs
42.6 → 2.30 TFLOP

01 / MODEL DESIGN

Compress the interface. Keep the video prior.

Cosmos Policy 1.96B trunk



1,764 tokens · 28 blocks · 5 steps

SmolWAM 0.88B trunk



16 actions + future latents

1,008 tokens · 12 blocks · 1 step

Current images · Action queries · Future images

w: wrist · p: third-person · s: state · $'$: future

01 Encode each image once.

Single-frame Wan VAE with a trained 0.5M-parameter adapter.

02 Keep the attention sinks.

Cached BERT features, zero-padded to 512 context positions.

03 Distill pruning and compaction.

Keep blocks 0–2 and 19–27; replace the action frame with 28 state-conditioned queries.

3.6 A100 GPU-hours

68,830 observations · 1,700 demos · 5,000 steps

Curriculum: 1,764 tokens (1k steps), then 1,008 tokens (4k steps).

02 / SUCCESS & EFFICIENCY

Most of the success. A fraction of the cost.

Model	Success	TFLOP	ms	GiB
Cosmos 5 steps	98.2†	42.6	624	7.55
Cosmos 1 step[1]	98.0	7.52	126	4.39
SmolWAM	**95.5**	**2.30**	**57**	**2.36**
Action-only[2]	89.3	2.30	57	2.36

Full protocol: 40 tasks × 10 trials.
Success in %; costs per 16-action chunk.

[1]Trained single-frame adapter; teacher weights unchanged.
†Five-step teacher used an earlier evaluation pipeline.
[2]Same student, auxiliary losses removed.

Where performance changes

	Spatial	Object	Goal	Long
Cosmos 5 steps	97	100	99	97
SmolWAM	**99**	**100**	**94**	**89**
Action-only	91	100	93	73

100 episodes per suite. Largest teacher gap: Long-horizon tasks (−8 percentage points).

A100-SXM4-40GB · batch 1 · bf16 · cached text.
Timings include preprocessing and VAE encoding.

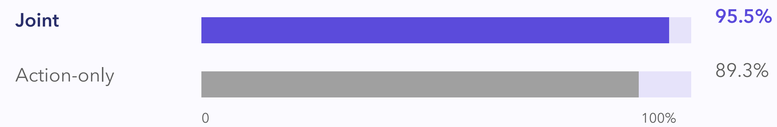
L4 / T4: 83 / 188 ms (edge-class GPU proxies).

03 / WHY PREDICT?

Joint training helps control.

+6.2%

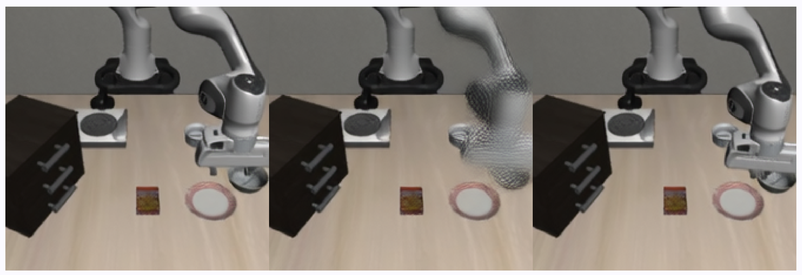
95.5% vs. 89.3%
400-episode protocol



Matched ablation removes visual-future, future-state and value losses; future tokens remain. This supports auxiliary supervision. Predicted futures are never used for planning.

Still predicts scene motion

Current · Predicted $t{+}16$ · Real $t{+}16$



−24%

Latent L1 vs. copy
0.133 → 0.101

1.63 → 1.43 px

Flow endpoint error

96 held-out contexts; decoded only for display. Forecasts remain blurred. LPIPS is worse than copying the current frame (0.128 vs. 0.115).

CONTACT

timiiowolabi@gmail.com

PROJECT + PAPER

smolwam.vercel.app · github.com/iitimii/SmolWAM

PROJECT CODE