Combining Symbolic and Perceptual Memory for Long-Horizon Robot Manipulation

Aaditya Shah [1] Ajay Sridhar [2]

[1]Stanford University, CS 231N [2]Stanford University — PhD student, project advisor (not CS 231N)

Motivation

Long-horizon manipulation is **non-Markovian**: a policy must recall events that left the camera view, track occluded objects, and reproduce demonstrated motion. Vision-language-action (VLA) models such as $\pi_{0.5}$ [3] generalize well but condition on the present frame. Feeding long image histories is expensive and brittle under covariate shift.

Two complementary memory paradigms:

- **Symbolic** (MemER [4]): compress history into grounded language subgoals from a video VLM with keyframe retrieval.
- **Perceptual** (FrameSamp+Modulation [2]): keep dense SigLIP tokens, modulate the action expert via cross-attention and FiLM.

RoboMME [2] shows *no single representation dominates* across 16 memory tasks. **Can both pathways compound in one policy?**

RoboMME Benchmark

16 non-Markovian manipulation tasks in simulation, grouped into four cognitively motivated suites:

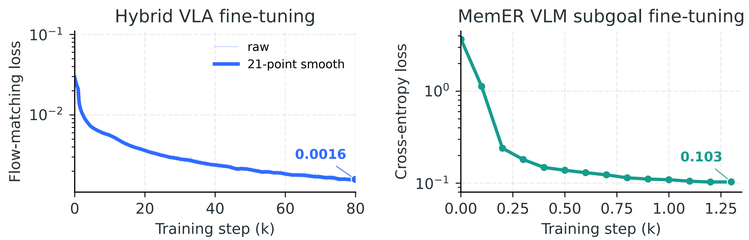
Suite	Tasks (memory demand)
Counting	BinFill, PickXtimes, SwingXtimes, StopCube *(temporal)*
Permanence	Video/Button Unmask & Swap variants *(spatial)*
Reference	PickHighlight, VideoRepick, PlaceButton, PlaceOrder *(object)*
Imitation	MoveCube, InsertPeg, PatternLock, RouteStick *(procedural)*

Data: 1,600 train + 800 test expert trajectories (100/50 episodes per task). **Eval:** 50 rollouts/task, 3 seeds, 1,300-step horizon, 16-frame window.

Training

Fine-tune $\pi_{0.5}$ (OpenPi) with LoRA, action expert, and 512-token perceptual memory (80k steps). MemER VLM trained separately; eval swaps in predicted subgoals. Both trained on the ILIAD GPU cluster—hybrid VLA on 4× A40 (FSDP), MemER VLM on 4× H200 GPUs (DeepSpeed ZeRO-2).

Training losses for the hybrid policy pipeline

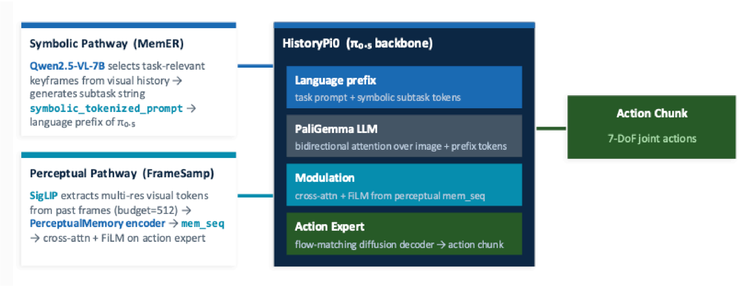


VLA flow-matching and MemER subgoal losses both converge smoothly.

Hybrid Memory Architecture

We build on the $\pi_{0.5}$ VLA architecture [3, 1] to design a hybrid architecture as follows:

$$\pi_\theta\big(A_t \mid o_t,\, q,\, m_t^{\mathrm{sym}},\, m_t^{\mathrm{perc}}\big)\,.$$

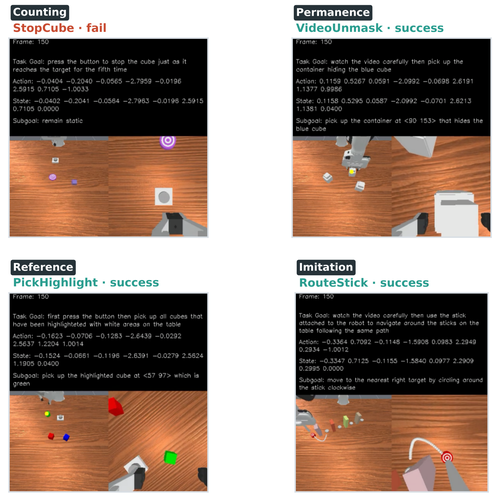


Dual memory pathways feed HistoryPi0 ($\pi_{0.5}$ backbone). Symbolic subgoals enter the language prefix; perceptual tokens modulate the action expert via cross-attention and FiLM.

Key design. m_t^{sym} and m_t^{perc} meet *only* at the action expert—not in a shared VLM token stream.

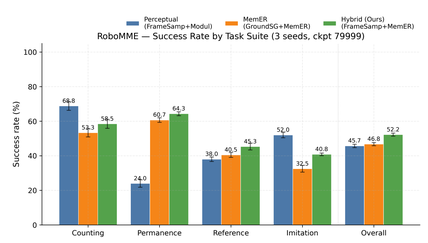
Qualitative Rollouts

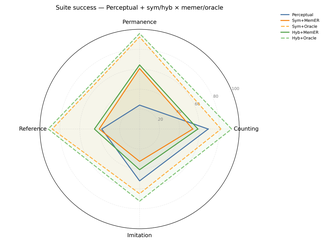
Representative RoboMME rollout stills



Held-out rollouts across four suites. Hybrid supports permanence and reference; StopCube fails under subgoal interference.

Results & Analysis





Left: mean success (%, 3 seeds) by suite for perceptual-only, symbolic-only, and hybrid policies. Hybrid reaches 52.3% overall and leads on Permanence and Reference; perceptual-only leads Counting and Imitation. Right: radar including oracle subgoals (dashed)—oracle hybrid (green dashed) dominates all suites, showing headroom once VLM subgoal quality improves.

On Counting and Imitation, hybrid tracks MemER failures—bad subgoals inject adversarial language into the VLA prefix and override perceptual memory. Re-evaluating the *same* checkpoints with oracle subgoals recovers +35 pp overall, showing the low-level policy can fuse both modalities when subgoal text is correct.

Condition	Count.	Perm.	Ref.	Imit.	**Overall**
Perceptual	68.8	24.0	38.0	52.0	45.7
Sym + MemER	53.3	60.7	40.5	32.5	46.8
Sym + Oracle	81.7	92.3	87.7	64.8	81.6
Hyb + MemER	58.5	64.3	45.3	40.8	52.3
Hyb + Oracle	**92.3**	**95.8**	**91.7**	**72.3**	**88.0**

Takeaways

What we built. A hybrid $\pi_{0.5}$ VLA combining MemER symbolic subgoals with frame-sampled perceptual memory—symbolic cues in the language prefix, dense history via action-expert modulation.

Main result. 52.3% overall (+5.5 pp vs. best single-modality baseline), with largest gains on Permanence (+40 pp) and Reference (+7 pp).

Key insight. Oracle ablation isolates the bottleneck to VLM subgoal quality. With ground-truth subgoals, hybrid reaches **88.0%**—well above perceptual (45.7%) and symbolic+Oracle (81.6%).

References

[1] Kevin Black, Noah Brown, Danny Driess, et al.
π_0: A vision-language-action flow model for general robot control.
arXiv preprint arXiv:2410.24164, 2024.

[2] Yinpei Dai, Hongze Fu, Jayjun Lee, Yuejiang Liu, Haoran Zhang, Jianing Yang, Chelsea Finn, Nima Fazeli, and Joyce Chai.
RoboMME: Benchmarking and understanding memory for robotic generalist policies.
arXiv preprint arXiv:2603.04639, 2026.

[3] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, et al.
$\pi_{0.5}$: A vision-language-action model with open-world generalization.
arXiv preprint arXiv:2504.16054, 2025.

[4] Ajay Sridhar, Jennifer Pan, Satvik Sharma, and Chelsea Finn.
MemER: Scaling up memory for robot control via experience retrieval.
arXiv preprint arXiv:2510.20328, 2025.

aadi2@stanford.edu