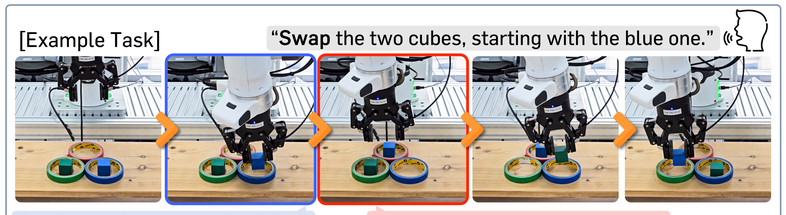
HAMLET: Switch your Vision-Language-Action Model into a History-Aware Policy

Myungkyu Koo*[A], Daewon Choi*[A], Taeyoung Kim[A],
Kyungmin Lee[A], Changyeon Kim[A], Younggyo Seo†[B], Jinwoo Shin†[AC]

[A]KAIST, [B]UC Berkeley, [C]RLWRLD, *Equal Contribution, †Equal Advising

Project Page | Myungkyu's Page | Daewon's Page

Motivation



✓ Most recent VLAs rely on a single-frame assumption, solely depending on current observations.

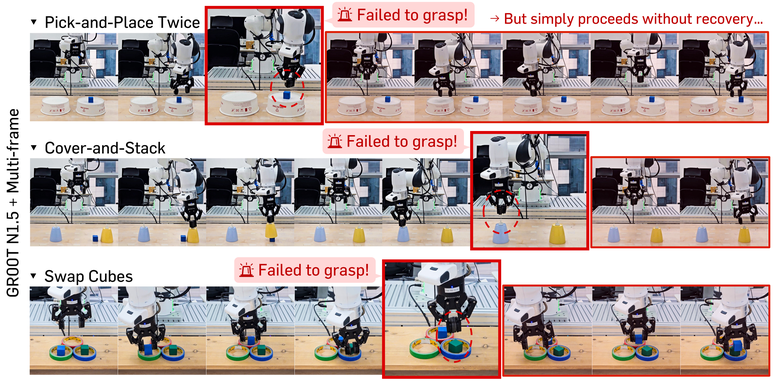
✓ Yet robotic manipulation is inherently history-dependent! (e.g., deciding which cube to grasp next; see example above)

Observation & Challenges

✓ When naïvely using multi-frame inputs for longer-context,

① **Inefficient scaling**: Appending 4 past frames to VLA input results in 1.35× latency, 3.64× peak memory, lacking practicality.

② **Poor generalization**: Easily suffers from causal confusion (e.g., overfits to past proprioceptive states, not current context) and fails to recover from intermediate failures (see below).



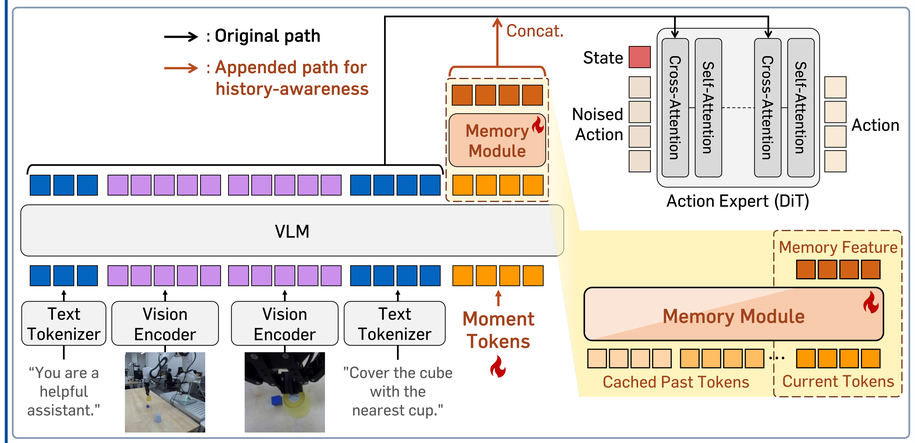
✓ **Our Answer: HAMLET**

① +47.2% on real-world history-dependent tasks

② Consistent gains on public benchmarks (LIBERO, RoboCasa, Simpler)

③ Only 1.02× latency, 1.96× peak memory

Overview of HAMLET



✓ TL;DR: HAMLET is a plug-in framework to incorporate history-awareness into VLAs via i) moment tokens and ii) a memory module.

Method

1. Context compression via **Moment Tokens**

- **Function:** Appends learnable tokens to VLM input to compress per-timestep context into compact representations.

$$[\mathbf{h}_t; \mathbf{m}'_t] = \mathcal{F}_\theta([\mathbf{o}_t, \mathbf{c}; \mathbf{m}_t])$$

- **Initialization:** Utilizes Time-Contrastive Learning to capture task-relevant cues while suppressing static backgrounds.

$$\mathcal{L}_{\text{TCL}}(\mathbf{z}_t, \mathbf{z}_t^+) = -\sum_{t=1}^{B} \log \frac{\exp\big(\text{sim}(\mathbf{z}_t, \mathbf{z}_t^+)/\tau\big)}{\exp\big(\text{sim}(\mathbf{z}_t, \mathbf{z}_t^+)/\tau\big) + \exp\big(\text{sim}(\mathbf{z}_t, \mathbf{z}_t^-)/\tau\big)}, \quad \mathbf{z}_t = g(\mathbf{m}'_t)$$

$\mathbf{z}_t^+$: Same timestep w/ image augmentation, $\mathbf{z}_t^-$: Different timestep

2. Memory consolidation via **Memory Module**

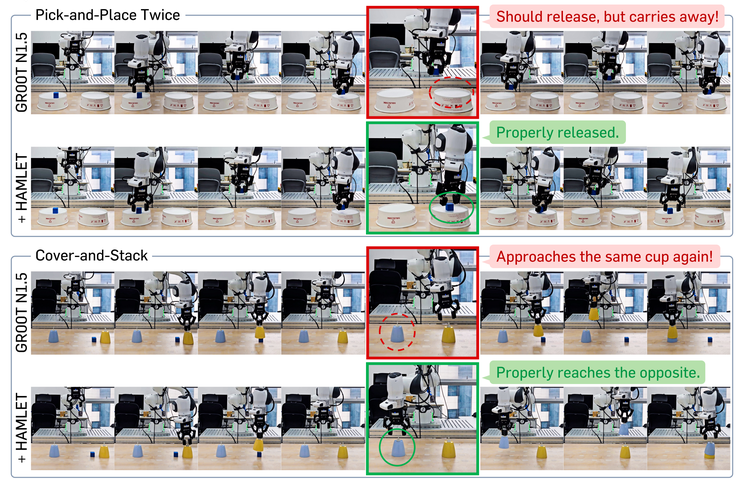
- **Architecture:** A lightweight Transformer that aggregates the cached moment tokens across the history horizon.
- **Mechanism:** Selectively attends to critical past moments via causal self-attention, producing a history-augmented feature.

$$\mathbf{M}' = [\mathbf{m}'_{t-k(T-1)};\ \ldots;\ \mathbf{m}'_{t-k};\ \mathbf{m}'_t] \in \mathbb{R}^{L \times d}$$

$$[\mathbf{a}_t, \mathbf{a}_{t+1}, \ldots, \mathbf{a}_{t+k-1}] = \mathcal{A}_\psi([\mathbf{h}_t; \tilde{\mathbf{m}}'], \mathbf{s}_t), \text{ where } \tilde{\mathbf{m}}' \text{ is the history-augmented feature}$$

Experiments

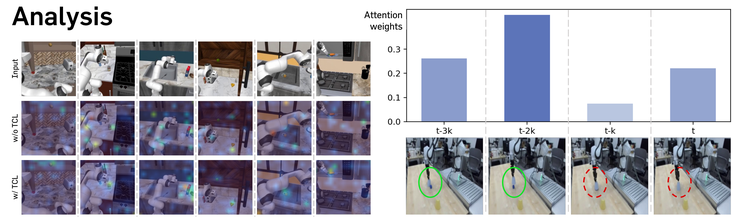
- **Qualitative Results**



- **Quantitative Results**

Method	History?	Pick-and-Place Twice *PnP Once*	Pick-and-Place Twice Success	Cover-and-Stack *Cover Cube*	Cover-and-Stack Success	Swap Cubes *Stage Cube*	Swap Cubes Success	Avg.
π_0	✗	54.2	25.0	87.5	58.3	83.3	12.5	31.9
π_0-FAST	✗	37.5	20.8	54.2	12.5	66.7	4.2	12.5
GR00T N1	✗	54.2	25.0	79.2	33.3	75.0	33.3	30.6
GR00T N1.5	✗	54.2	12.5	62.5	37.5	87.5	37.5	29.2
+ Multi-frame	✓	79.2	45.8	70.8	33.3	91.7	58.3	45.8
+ HAMLET (Ours)	✓	**91.7**	**66.7**	**95.8**	**79.2**	**95.8**	**83.3**	**76.4**

- **Analysis**



Moment tokens attention over vision inputs | Memory feature attention across timesteps

Moment Token	TCL	Memory Module	Avg.
✗	✗	✗	62.6
✓	✗	✗	63.1
✓	✓	✗	63.4
✓	✗	✓	64.8
✓	✓	✓	**65.4**

Method	RoboCasa Kitchen 100 demo	LIBERO Spatial	LIBERO Object	LIBERO Goal	LIBERO Long	LIBERO Avg.
Random initialization	64.8	**99.4**	99.8	98.0	87.8	96.2
SimCLR	63.8	**99.4**	98.6	97.0	87.2	95.5
Single-view TCN	64.8	98.6	99.4	98.2	91.6	96.9
Multi-view TCN	64.9	99.0	99.0	97.8	90.2	96.5
TCL (Ours)	**65.4**	99.0	**100.0**	**99.2**	**92.2**	**97.7**

Token Length	Avg.
1	64.3
4	65.4
8	**66.4**
16	65.9
32	62.7
64	62.5

Method	Avg.
No Memory	62.6
Moment Concat.	62.7
RNN	64.5
LSTM	65.0
GRU	64.3
Transformer	**65.4**

Method	History Length	Latency (ms, ↓)	Peak memory (MB, ↓)
GR00T N1.5	1	80.5 (1.00×)	289 (1.00×)
+ Multi-frame	4	108.5 (1.35×)	1051 (3.64×)
+ HAMLET (Ours)	4	**82.4 (1.02×)**	**566 (1.96×)**
+ Multi-frame	8	193.0 (2.40×)	2023 (7.00×)
+ HAMLET (Ours)	8	**85.8 (1.07×)**	**578 (2.00×)**