What Crosses Time? From Recurrent State to Causal Memory in Vision-Language-Action Models

Hyungjun Nam[1] **Young Kyun Jang**[2] **Byungchul Kim**[1]

[1]Sungkyunkwan University [2]Independent Researcher

Abstract

Vision-language-action (VLA) policies with recurrent memory carry latent states across environment steps, but the presence of past information in these states does not establish which components actually influence later behavior. We study this question in μVLA, whose recurrent state contains 64 memory tokens with 4096 features each. Recurrent-update geometry identifies a rank-one candidate subspace aligned with the uniform token direction. We hold this candidate fixed and test its behavioral role using causal interventions. Past information remains broadly decodable, yet causal influence on memory-dependent behavior is mostly concentrated in this rank-one subspace. Within it, separately fitted rank-16 feature subspaces preserve memory-dependent behavior across two tasks in which a briefly observed cue must guide a later choice. On a qualitatively different task requiring memory of a continuous object position, a separately fitted rank-8 feature subspace also preserves behavior. Thus, although μVLA exposes a 64×4096 recurrent state, with the token and task-specific feature bases fixed, only 16 dynamic coefficients are sufficient in the recall tasks, and 8 are sufficient in the continuous-position task.

1 Introduction

Long-horizon and partially observable manipulation often requires a policy to retain information that is no longer available in the current observation. Recurrent vision-language-action (VLA) models, which have a latent recurrent state that is passed across environment steps, have recently emerged as a promising approach to these settings [Cherepanov et al., 2026b, Xiao et al., 2026]. However, past information can remain decodable from recurrent state without being used for the relevant decision [Elazar et al., 2021, Liao and Cao, 2026]. Identifying the components that actually affect later behavior is therefore important for diagnosing memory failures and may also be useful for building more efficient recurrent architectures. Prior work has recognized this distinction and used causal interventions to study the functional roles of internal representations [Vig et al., 2020, Geiger et al., 2021, Sobotka et al., 2025]; additional related work is discussed in the Appendix A.

We study how recurrent VLA states are organized to carry past information into later behavior, primarily in μVLA [Cherepanov et al., 2026b], a recurrent VLA that carries 64 memory tokens, each with 4096 feature dimensions, across environment steps. By analyzing recurrent update geometry, we first identify a low-dimensional candidate structure and then test its behavioral role using causal interventions. These analyses show that causal influence on memory-dependent behavior is concentrated in a much lower-dimensional token–feature structure than the nominal recurrent state (Fig. 1).

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: 8th Robot Learning Workshop: Is Physical AI Going Zero-Shot?.

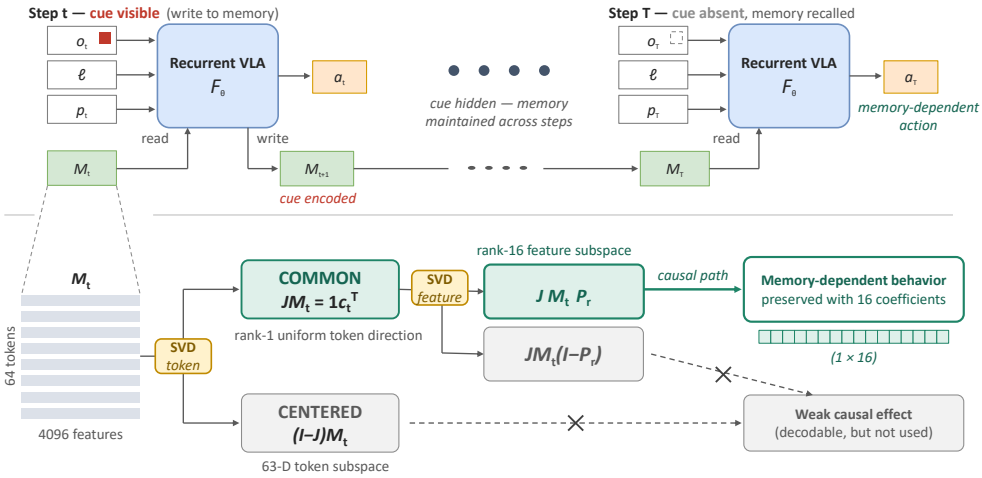


Figure 1: Token- and feature-axis recurrent structure analyzed in μVLA. The token-axis candidate is identified from recurrent updates ΔM_t. The feature basis Q_r is fitted from projected recurrent states JM_t. Causal interventions test the behavioral role of these fixed subspaces.

2 Experimental Setup

2.1 Model and Tasks

We primarily study μVLA. It augments OpenVLA-OFT [Kim et al., 2025] with a bank of 64 recurrent memory tokens, each with 4096 features. Its simple recurrent design exposes a single 64×4096 state that can be directly intercepted and manipulated. The model can be mathematically described as

$$(a_t, M_{t+1}) = F_\theta(o_t, l, p_t, M_t), \tag{1}$$

where o_t, l, p_t and M_t denote the current observation, language instruction, proprioceptive input, and recurrent state, respectively, and a_t is the predicted action.

We evaluate RememberColor5 (RC5) and RememberShapeAndColor3x3 (RSC3x3), which require retaining a briefly observed visual cue—a target color or shape–color identity—for a later choice [Cherepanov et al., 2026a]. We additionally report selected zero-shot results on RememberShapeAndColor3x2 (RSC3x2), and evaluate TakeItBack for continuous-position memory [Cherepanov et al., 2026a].

2.2 Candidate Identification

An SVD of recurrent updates $\Delta M_t = M_{t+1} - M_t$ reveals a rank-one candidate subspace in the 64-dimensional token space whose leading left singular vector remains stable across recurrent timesteps and aligns with $q = \frac{1}{\sqrt{64}}[1, 1, \ldots, 1]^\top \in \mathbb{R}^{64}$ at $|\cos| \approx 0.99$. We therefore define $J = qq^\top$ as the projector onto this direction and decompose recurrent updates into two terms as

$$\Delta M_t = \underbrace{J\Delta M_t}_{\text{COMMON}} + \underbrace{(I - J)\Delta M_t}_{\text{CENTERED}}. \tag{2}$$

Importantly, this alignment does not mean that the recurrent update is confined to COMMON. We measure the fraction of update energy in COMMON as $E_{\text{Common}} = \frac{\|J\Delta M_t\|_F^2}{\|\Delta M_t\|_F^2}$. Across the measured settings, COMMON accounts for only 45–57% of the update energy.

For feature-space candidates, we stack the projected recurrent vectors JM_t across episodes and timesteps and let Q_r contain the top-r right singular vectors. We define the corresponding feature projector as $P_r = Q_r Q_r^\top$. Feature bases are fitted separately for each task using trajectories disjoint from those used for causal evaluation.

2.3 Causal Interventions

For token-space interventions, we compare the COMMON projection, the CENTERED complement, and matched random rank-one projections. Within COMMON, feature-space interventions modify only the selected feature subspace while leaving CENTERED unchanged; for example, retaining P_r (the Q_r intervention) gives

$$M_t \leftarrow (I - J)M_t + JM_tP_r.$$

The joint Q_r restriction is stricter: it applies the token- and feature-space restrictions simultaneously:

$$\widetilde{M}_t = JM_tP_r.$$

Our causal analysis uses four intervention types; which ones apply is determined by the structure of each task. **Read/write sufficiency** tests whether restricting M_t on read or ΔM_t on write preserves memory-dependent behavior. **Cue-absent necessity** removes or resets the tested component at a single timestep after the relevant cue is no longer observable. **Content transfer** replaces the selected recipient component with its donor counterpart and tests whether the choice follows the donor. **Scene rebinding** repeats the transfer after moving the remembered cue to a different location in the recipient scene, testing whether the transferred content follows the cue to its new location rather than the donor's original spatial target or motor plan.

3 Results

3.1 Decodable history is broad, but causal influence is narrow

Past information is not confined to the rank-one subspace. On RC5, episode-disjoint probes for the hidden cue reach balanced accuracy 0.983 from COMMON, 0.979 from CENTERED, and 0.945 from a random rank-one projection (chance 0.20), while a timestep-only probe achieves 0.200 (Fig. 2a; Appendix Table 1). Despite this broad decodability, the **read/write sufficiency** test sharply separates these directions (Fig. 2b; Appendix Table 2). On RC5, COMMON read gives action MAE 0.0031 versus 0.0886 for a random rank-one projection, while COMMON write gives 0.0009. The same separation holds on RSC3x3 and zero-shot RSC3x2 (Appendix Table 2). Thus, past information is broadly decodable while behavior depends on a much narrower token-space structure.

3.2 COMMON: necessity, content transfer, and scene rebinding

Resetting COMMON after the cue disappears reduces RC5 success from $29/30$ to $7/30$ or $8/30$, whereas resetting CENTERED leaves success at $29/30$ (Fig. 2c). Donor COMMON transfers the remembered target in $64/64$ pairs, versus $0/64$ for CENTERED (Fig. 2d). Under scene rebinding, the

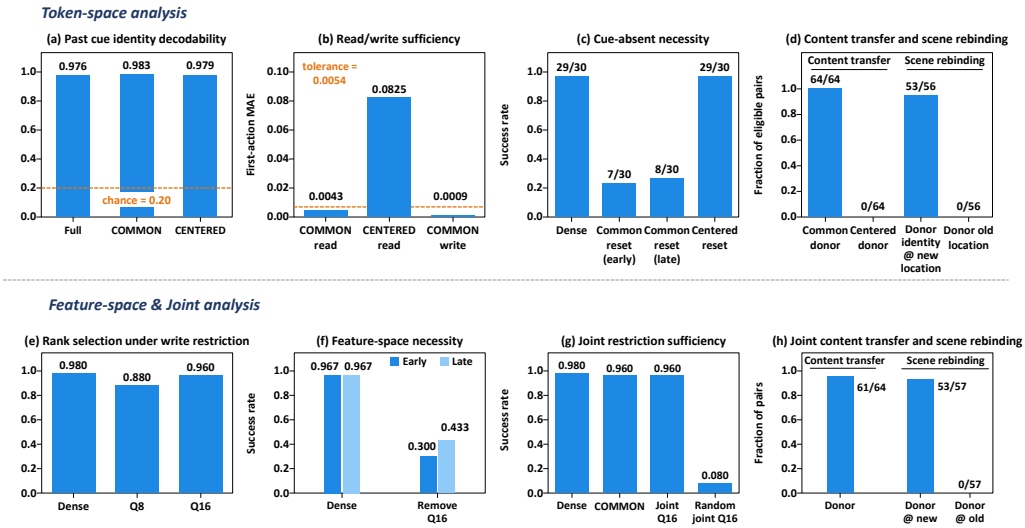


Figure 2: Causal structure of recurrent memory in μVLA on RC5. Token-space interventions identify the COMMON component, while feature-space and joint restrictions further localize causal influence to a compact Q_{16} subspace. Early and late denote two intervention points within the cue-absent interval.

donor cue is selected at its new recipient location in $53/56$ pairs and at the donor's original spatial target in $0/56$ (Fig. 2d).

RSC3x3 shows the same pattern: donor COMMON transfers the remembered identity in $49/49$ pairs and achieves $49/62$ under scene rebinding, the same as full recurrent-state transfer, versus $0/62$ for CENTERED. Content transfer also holds zero-shot on RSC3x2 (Appendix C.1).

3.3 Causal influence is concentrated in a low-dimensional feature subspace

Within COMMON, a rank-8 feature subspace Q_8 preserves read and maintenance on RC5 but fails the write restriction (0.880 success), whereas Q_{16} restores success to 0.960 (Fig. 2e). Q_{16} also passes the cue-absent necessity test (Fig. 2f). Under scene rebinding, Q_{16} succeeds in $53/57$ pairs versus $0/57$ for a matched random subspace (Appendix Table 3).

3.4 The joint restriction preserves memory-dependent behavior

The joint Q_{16} restriction nearly matches Dense on RC5 (0.960 vs. 0.980) (Fig. 2g) and RSC3x3 (0.820 vs. 0.800), whereas matched random restrictions fall to 0.080 and 0.120. On RC5, it also supports content transfer and scene rebinding (Fig. 2h). With the token and feature bases fixed, the full 64×4096 recurrent state under the joint restriction can therefore be represented by only r time-varying coefficients:

$$z_t = Q_r^\top M_t^\top q \in \mathbb{R}^r, \qquad \widetilde{M}_t = q z_t^\top Q_r^\top = J M_t P_r. \tag{3}$$

With $r = 16$, retaining only the 16 coefficients z_t across recurrent steps is sufficient to preserve memory-dependent behavior on RC5 and RSC3x3.

3.5 Low-dimensional sufficiency extends to continuous-position memory

On TakeItBack, COMMON-only recurrence remains sufficient, while CENTERED-only and NoWrite fail. The joint Q_8 restriction reaches 1.000 success, versus 0.060 for a matched random restriction (Appendix Table 4). Thus, retaining only eight dynamic coefficients is sufficient on this continuous-position task, compared with 16 on RC5 and RSC3x3.

4 Discussion and Conclusion

Our results distinguish information that is merely present in recurrent state from the structures through which remembered content affects behavior. In μVLA, causal influence is concentrated in the token-axis COMMON component and compact feature subspaces: rank 16 is sufficient on RC5 and RSC3x3, while a separately fitted rank-8 subspace is sufficient on TakeItBack. Selected token-space interventions show the same pattern zero-shot on RSC3x2. On RC5 and RSC3x3, the rank-16 subspace further supports content transfer and scene rebinding.

One possible explanation for the uniform token direction is the symmetry among memory cells. Permutation-equivariant mappings naturally separate the uniform direction from centered variation [Zaheer et al., 2017], and self-attention has a known bias toward token-uniform representations [Dong et al., 2021]. The prominence of the uniform direction is not simply explained by an inability to distinguish memory cells: attention is uniform across memory cells in early layers but becomes cell-specific in later layers (Appendix C.7). These observations may help explain the geometry of the recurrent state, but not why the uniform direction has causal influence on memory-dependent behavior; that requires the interventions studied here.

These ranks are tested sufficient operating points rather than minimum or intrinsic memory dimensions, and the AVA-VLA comparison (Appendix D) indicates that the exact recurrent geometry is not universal. More broadly, our results suggest that the relevant design target for recurrent memory is not the size of the exposed state, but the low-dimensional structure through which past information actually affects future behavior.

A Related Work

Memory in VLAs. MemoryVLA augments VLA policies with working memory and an external memory bank [Shi et al., 2026], while ReMem-VLA uses recurrent queries across frames and action chunks [Li et al., 2026a]. μVLA carries recurrent memory tokens across environment steps [Cherepanov et al., 2026b], while AVA-VLA maintains a separate history-dependent recurrent state [Xiao et al., 2026].

Causal analysis of representations. Decodability alone does not establish behavioral use. Amnesic Probing tests this distinction through representation removal [Elazar et al., 2021], while interchange interventions and causal mediation intervene directly on internal representations [Geiger et al., 2021, Vig et al., 2020]. *Present but Not Remembered* similarly shows that visual history can remain decodable in VLAs without affecting policy behavior [Liao and Cao, 2026].

Recurrent-state structure. Sobotka et al. identify sparse memory representations and functional circuits in DreamerV3 [Sobotka et al., 2025], while Li et al. relate recurrent-policy hidden-state geometry to behavior [Li et al., 2026b]. Our analysis differs by resolving an explicit recurrent VLA state along both token and feature dimensions, testing the same identified subspaces through causal interventions, and evaluating whether their joint restriction remains sufficient for memory-dependent behavior.

B Experimental Details

B.1 Task Details

The three discrete-cue tasks use delayed recall [Cherepanov et al., 2026a]. A target is initially visible, disappears during a cue-absent interval, and is followed by candidate objects at randomized spatial locations. Cue duration and the cue-absent interval vary across episodes.

RC5. The target color is drawn from five possibilities. Five candidate cubes appear after the cue-absent interval, and success requires touching the cube with the remembered color. The horizon is 25 steps; cue duration and the subsequent empty interval are each sampled independently from 1–5 environment steps.

RSC3x3 and RSC3x2. RSC3x3 uses nine identities formed by three shapes and three colors. RSC3x2 follows the same delayed-recall protocol with six identities formed by three shapes and two colors. RSC3x2 is evaluated zero-shot with the analyzed checkpoint.

TakeItBack. The remembered variable is the continuous initial 2-D position of an object. The policy first moves the object to a target region and must then return it to its initial position. The transition to the return phase is event-triggered rather than tied to a fixed timestep. Because the task has no discrete cue followed by a cue-absent interval, the cue-absent necessity intervention used on RC5 is not defined for TakeItBack.

B.2 Geometry and Feature-Space Basis Fitting

The token-space notation follows Sec. 2.2. Across the audited recurrent updates, the leading token-space direction remains closely aligned with the uniform direction ($|\cos| \approx 0.99$). This does not reflect collapse of the recurrent update: COMMON accounts for only 45–57% of recurrent-update energy across the measured settings.

For feature-space fitting, we compute

$$d_{e,t} = \frac{1}{64}\mathbf{1}^\top M_{e,t} \in \mathbb{R}^{4096}, \tag{4}$$

stack these vectors as rows of D, and compute

$$D = U\Sigma V^\top. \tag{5}$$

Q_r contains the first r right singular vectors and $P_r = Q_r Q_r^\top$. Rows of D are not normalized before the decomposition. Feature bases are fitted separately for each task using trajectories disjoint from causal evaluation. For RC5 and RSC3x3, the feature bases are fitted using 39 episodes (583 recurrent steps) and 92 episodes (1,388 recurrent steps), respectively. The fitting trajectories are disjoint from those used for causal evaluation. The reported feature ranks are selected empirically based on behavioral intervention results across the evaluated ranks, rather than specified a priori. We treat them as tested sufficient operating points, not estimates of the minimum intrinsic memory dimension.

B.3 Decoding and Read/Write Evaluation

Hidden-cue decoding uses episode-disjoint linear probes. Additional RC5 controls are summarized in Table 1.

Probe input	Hidden-cue balanced accuracy
Full state	.976
COMMON	.983
CENTERED	.979
Random rank-one projection	.945
Norm only	.281
Timestep only	.200
Label shuffle	$\sim$.20

Table 1: RC5 hidden-cue decoding controls. Chance is 0.20.

For read/write sufficiency, action distortion is measured on the first executed de-normalized physical action:

$$D_T = \frac{1}{7} \sum_{d=1}^{7} \left| a_d^{\text{int}} - a_d^{\text{Dense}} \right|. \tag{6}$$

The task-specific tolerance is fixed before the corresponding rank comparison at 0.2 times the distortion produced by suppressing one recurrent write at the evaluated state. This tolerance applies only to the action-fidelity test; closed-loop success, cue-absent necessity, content transfer, and scene rebinding are reported directly.

Task	Tolerance	COMMON read	COMMON write	Random R1 read
RC5	.0054	.0031	.0009	.0886
RSC3x3	.0043	.0038	.0010	.1120
RSC3x2	.0034	.0023	.0008	.1099

Table 2: Held-out first-action MAE. RSC3x2 is evaluated zero-shot.

The random token-space control is not a single alternative direction: 32 static random rank-one directions were evaluated. Separately, rank-one directions optimized from random initialization only to preserve Dense actions converge near the same uniform direction ($|\cos| \approx 0.85$–0.91 for read and 0.989–0.995 for write). These controls test orientation rather than rank alone.

B.4 Intervention Details

Token-space read and write interventions use

$$M_t \leftarrow \Pi M_t, \qquad M_{t+1}^{\text{int}} = M_t + \Pi \Delta M_t, \qquad \Pi \in \{J, I - J, R\}, \tag{7}$$

where R is a random rank-one projector.

Feature-space retention and removal use

$$M_t^{\text{retain}} = (I - J)M_t + JM_tP_r, \tag{8}$$

$$M_t^{\text{remove}} = (I - J)M_t + JM_t(I - P_r). \tag{9}$$

The joint restriction retains only JM_tP_r.

For RC5 cue-absent necessity, COMMON is replaced for one recurrent step by its learned initial-state value:

$$M'_t = (I - J)M_t + JM_0, \tag{10}$$

with the symmetric operation used for CENTERED. Normal recurrence resumes immediately afterward. The early and late conditions intervene at different points inside the cue-absent interval; they are not used to infer a temporal trend.

For content transfer from donor M_D to recipient M_R, COMMON transfer uses

$$M'_R = (I - J)M_R + JM_D. \tag{11}$$

Feature-space content transfer uses

$$M'_R = M_R - JM_RP_r + JM_DP_r, \tag{12}$$

while joint content transfer uses

$$M'_R = JM_DP_r. \tag{13}$$

Donor–recipient eligibility is determined solely from unmodified reference rollouts before intervention outcomes are inspected. Each condition starts from the same stored recipient observation, proprioception, simulator state, and recurrent state; only the specified recurrent component is changed. Scene rebinding uses the same operators after placing the donor identity at a different location in the recipient scene from the donor's original spatial target. Random feature-space controls are matched in rank unless explicitly stated otherwise.

B.5 Computational Resources.

All experiments were conducted on a workstation equipped with an AMD Ryzen 7 9800X3D CPU (8 cores, 16 threads), an NVIDIA GeForce RTX 5080 GPU with 16 GB of VRAM (driver version 580.178.04), and 60 GiB of system memory.

C Additional Controls

C.1 Zero-Shot RSC3x2

Across 25 paired seeds, Dense, COMMON-only read, and COMMON-only write obtain success rates 0.64, 0.72, and 0.68, respectively, while CENTERED-only read and NoWrite both obtain 0.12. The point estimate above Dense is not interpreted as a performance improvement.

In a separate content-transfer evaluation, donor COMMON selects the donor target in $21/22$ eligible pairs, whereas donor CENTERED does so in $0/22$. RSC3x2 is used as a selected zero-shot check of the token-space result rather than as a third task for the complete causal analysis.

C.2 Cue-Phase Write Control

We also tested recurrent write during cue presentation on RC5. Across 30 paired seeds, Dense obtains 0.967 success; suppressing COMMON write gives 0.200, while suppressing CENTERED write gives 0.967. However, the pre-specified action-deviation guard is exceeded for the COMMON-write intervention ($0.2115 > 0.05$). A subsequent simulator-state audit does not find consequential physical divergence before the relevant manipulation, but we nevertheless treat this experiment only as supporting evidence and do not use it for the main causal claim.

C.3 Feature-Space Controls

The feature-space scene-rebinding experiment uses a separate 57-pair set from the 56-pair token-space experiment in the main text. The orthogonal complement of Q_{16} has rank 4080 and is therefore not rank matched. The energy-matched control instead uses a random rank-2976 subspace chosen to remove approximately the same feature energy as Q_{16}. In the RSC3x3 control, the donor's 16 coefficients are permuted within the same fitted Q_{16} subspace; the donor's previous spatial target is selected in $0/61$ pairs under scene rebinding.

Test	Condition	Result
RC5 scene rebinding	COMMON	$54/57$
	Q_{16}	$53/57$
	Random rank 16	$0/57$
	$Q_{16}^{\perp}$	$3/57$
RC5 cue-absent necessity	Dense	.967
	Remove Q_{16}	.300 / .433
	Remove random rank 16	.967 / 1.000
	Energy-matched random removal	.867 / .933
RSC3x3 content transfer	Q_{16}	$47/61$
	Permuted coefficients	$1/61$

Table 3: Additional feature-space controls. The two RC5 necessity values correspond to the early and late intervention points.

C.4 Joint Restriction and Coefficient-Only Recurrence

The RC5 joint Q_{16} result is obtained from a separate closed-loop evaluation from the 30-seed cue-absent necessity experiment; the corresponding Dense point estimates should therefore not be read as repeated measurements from the same evaluation set.

The joint state also supports the RC5 content tests. Donor JM_DP_{16} selects the donor identity in $61/64$ eligible content-transfer pairs. Under scene rebinding, it selects the donor identity at its new recipient location in $53/57$ pairs and the donor's original spatial target in $0/57$.

For the coefficient-only realization, only the z_t defined in Sec. 3.4 is retained across environment steps, and JM_tP_r is reconstructed from z_t, q, and Q_r before the next policy evaluation. The fixed bases are not counted as dynamic recurrent state. The RC5 content-transfer and scene-rebinding experiments above are performed on the corresponding joint state; they are not separate repetitions in coefficient coordinates.

C.5 TakeItBack

The TakeItBack feature basis is fitted independently on 30 episodes comprising 960 recurrent steps (trials 200–229), while evaluation uses disjoint trials 400–449. All evaluations use 50 paired seeds.

Condition	N	Success
Dense	50	.980
COMMON-only	50	.980
CENTERED-only	50	.160
NoWrite	50	.000
Feature Q_4 read	50	.860
Random rank-4 read	50	.180
Joint Q_4	50	.820
Joint Q_8	50	1.000
Random joint rank 8	50	.060

Table 4: TakeItBack closed-loop evaluations under the fresh-process protocol.

Rank 4 preserves the feature-space read intervention but does not preserve the stricter joint restriction at the same level. The separately evaluated joint Q_8 restriction reaches 1.000 success, and the coefficient-only Z_8 realization also preserves behavior. We therefore use rank 8 only as a tested sufficient operating point for TakeItBack.

We also tested content transfer between TakeItBack episodes. Donor COMMON does not reliably redirect the recipient toward the donor's remembered initial position, but transferring the full donor recurrent state also fails to reliably induce donor-position behavior. We therefore regard this construction as non-diagnostic for content transfer on TakeItBack and make no scene-rebinding claim from it.

C.6 Memory-Load Scope

The current checkpoint does not provide a usable test of whether the sufficient rank grows with multi-item memory load. In the screened BunchOfColors3 and SeqOfColors3 conditions, both Dense and NoWrite obtain zero success. We therefore exclude these tasks rather than interpreting their results as evidence for or against low-rank memory under higher load.

C.7 Layerwise Memory-Cell Specificity

To test whether the prominence of the uniform token direction could simply reflect an inability of the model to distinguish memory cells, we inspect action-query attention over the 64 recurrent memory cells across layers. At layer 0, attention across the memory cells is bit-identically uniform, with an effective rank of 1.00 for the cell-attention matrix. By layer 31, the attention pattern is cell-specific, with an effective rank of 12.92. Thus, later layers can distinguish individual memory cells even though the recurrent state exhibits a strong token-space COMMON direction.

This observation concerns the geometry of attention, not the causal importance of the cell-specific component. Total attention to memory can also be much smaller in later layers; for example, memory-attention mass is 0.0006 at layer 28 versus 0.0960 at layer 0.

D AVA-VLA

AVA-VLA has a 56×4096 recurrent state. Its leading row-space direction explains 95.9% of recurrent-state variance and aligns with the row mean at $|\cos| \approx 0.997$, but the corresponding rank-one read restriction does not preserve recurrent behavior. Under the separately fixed BF16 action-fidelity tolerance $\epsilon = 0.00740$, ranks 1–3 remain above the criterion across the tested held-out splits. Higher-rank structured projections preserve substantially more of the recurrent read behavior. We therefore use AVA-VLA only as evidence that the exact rank-one token-space geometry observed in μVLA is not universal, not to assign a minimum rank to AVA-VLA.

Content transfer and scene rebinding are not reported for AVA-VLA because the required donor–recipient counterfactuals could not be constructed in native CALVIN [Mees et al., 2022]. Under the pre-specified criterion requiring sufficiently matched current observations with distinct histories, none of $60{,}382$ candidate pairs is eligible. A second construction using matched initialization and subtask depth also yields no eligible pair. We therefore make no content-transfer or scene-rebinding claim for AVA-VLA.

References

Egor Cherepanov, Nikita Kachaev, Alexey Kovalev, and Aleksandr I. Panov. Memory, benchmark & robots: A benchmark for solving complex tasks with reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL `https://openreview.net/forum?id=9cLPurIZMj`.

Egor Cherepanov, Nikita Kachaev, Daniil Zelezetsky, Aydar Bulatov, Artem Pshenitsyn, Yuri Kuratov, Alexey Skrynnik, Aleksandr I. Panov, and Alexey K. Kovalev. μvla: On recurrent memory for partially observable manipulation in vla models. *arXiv preprint arXiv:2606.12497*, 2026b.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2793–2803. PMLR, 2021.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 9574–9586, 2021.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. In *Proceedings of Robotics: Science and Systems*, 2025. doi: 10.15607/RSS.2025.XXI.017.

Hang Li, Fengyi Shen, Dong Chen, Liudi Yang, Xudong Wang, Jinkui Shi, Zhenshan Bing, Ziyuan Liu, and Alois Knoll. Remem-vla: Empowering vision-language-action model with memory via dual-level recurrent queries. *arXiv preprint arXiv:2603.12942*, 2026a.

Jin Li, Yue Wu, Mengsha Huang, Yuhao Sun, Hao He, and Xianyuan Zhan. Unraveling the hidden dynamical structure in recurrent neural policies, 2026b.

Chih-Ting Liao and Xin Cao. Present but not remembered: Auditing how frozen vlas encode, deploy, and steer visual history, 2026.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022. doi: 10.1109/LRA.2022.3180108.

Hao Shi, Bin Xie, Yingfei Liu, Lin Sun, Fengrong Liu, Tiancai Wang, Erjin Zhou, Haoqiang Fan, Xiangyu Zhang, and Gao Huang. Memoryvla: Perceptual-cognitive memory in vision-language-action models for robotic manipulation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Jan Sobotka, Auke Ijspeert, and Guillaume Bellegarda. Reverse-engineering memory in dreamerv3: From sparse representations to functional circuits, 2025. URL `https://openreview.net/forum?id=JmjqTi4FDF`. NeurIPS 2025 CogInterp: Interpreting Cognition in Deep Learning Models Workshop, Spotlight.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401, 2020.

Lei Xiao, Jifeng Li, Juntao Gao, Feiyang Ye, Yan Jin, Jingjing Qian, Jing Zhang, Yong Wu, and Xiaoyuan Yu. Ava-vla: Improving vision-language-action models with active visual attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13453–13463, 2026.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017.