CORSIBOT: A COGNITIVE ROBOTIC MODEL OF SERIAL SPATIAL RECALL

Anonymous authors
Paper under double-blind review

ABSTRACT

Humans can reproduce a recently observed sequence of locations by selecting its items in order. The Corsi block-tapping task is widely used to test this ability and measures visuospatial working memory span. Most studies of this task have explored how span depends on sequence length and spatial layout, and have treated recall as a readout of the stored sequence. However, recall in this task is carried out through action. Whether and how the visual feedback from actions takes part in step-by-step retrieval remains unclear. In this research, we introduce CorsiBot, an embodied cognitive robot model that observes pointing sequences and recalls them through successive selections, allowing the visual feedback after each selection to be systematically varied. CorsiBot reproduces key features of human recall, including a decline in accuracy with increasing sequence length and characteristic error patterns. The results show that the model's internal state carries a stable signal of the next location to be recalled, and that altering only the timing of the visual feedback of a selection, with the network and visual content unchanged, shifts both this signal and the next choice, while the model still retains which location it has just selected. We found that the visual feedback of action do not merely record what has been done but shape which location is retrieved next. These findings suggest that step-by-step retrieval from visuospatial working memory may be coupled to the sensory feedback of human actions. In addition, CorsiBot provides an embodied robotic platform for testing how action shapes sequential memory retrieval.

Humans can reproduce a recently observed sequence of locations by selecting its items in order. The Corsi block-tapping task is widely used to test this ability and measures visuospatial working memory span.Most studies of this task have explored how span depends on sequence length and spatial layout, and have treated recall as a readout of the stored sequence. However, recall in this task is carried out through action. Whether and how the visual feedback from actions takes part in step-by-step retrieval remains unclear. In this research, we introduce CorsiBot, an embodied cognitive robot model that observes pointing sequences and recalls them through successive selections, allowing the visual feedback after each selection to be systematically varied. CorsiBot reproduces key features of human recall, including a decline in accuracy with increasing sequence length and characteristic error patterns. The results show that the model's internal state carries a stable signal of the next location to be recalled, and that altering only the timing of the visual feedback of a selection, with the network and visual content unchanged, shifts both this signal and the next choice, while the model still retains which location it has just selected. We found that the visual feedback of action do not merely record what has been done but shape which location is retrieved next. These findings suggest that step-by-step retrieval from visuospatial working memory may be coupled to the sensory feedback of human actions. In addition, CorsiBot provides an embodied robotic platform for testing how action shapes sequential memory retrieval.

1 INTRODUCTION

Human cognition unfolds through continuous interaction between the body and the environment. Looking, reaching, and moving change what we perceive, while perception and memory guide these actions. For example, reaching for an object that is briefly out of view involves using its remembered location, while watching the approaching hand provides new visual information as the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

movement unfolds. Spatial memory is thus part of embodied cognition, connecting past experience to ongoing bodily action. This relationship can also be studied in embodied cognitive robots, where memory guides action and the sensory consequences of action return as input to the cognitive system. Such models provide a controlled setting for examining how perception, memory, and action jointly organize behavior.

The Corsi block-tapping task is a standard test of visuospatial working memory span (Corsi, 1972; Kessels et al., 2000). Participants observe an experimenter indicate a sequence of locations and then reproduce it by tapping the blocks in the same order (Figure 1A). Studies using spatial sequence tasks have examined how sequence length and spatial organization affect recall (De Lillo et al., 2016; Zhang et al., 2022). During reproduction, however, each tap also generates visual input as the hand reaches the selected block. The task therefore brings memory-guided selection and the sensory consequences of action into the same recall procedure, making it a natural setting for studying their interaction.

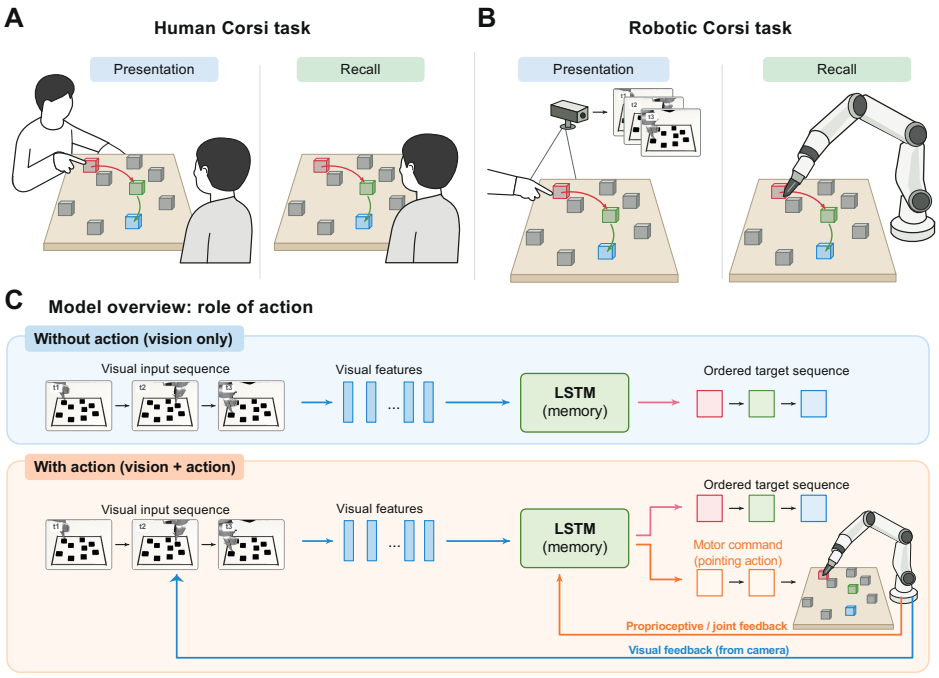


Figure 1: **The human Corsi task and its robotic adaptation. A,** A person observes a sequence of locations and reproduces it in order. **B,** A simulated Franka Panda manipulator points to the locations while a fixed external camera records the presentation. The pictured robotic pointing at recall illustrates the task correspondence; the evaluated model reports location symbols rather than executing a robot policy. **C,** Conceptual visual-only and vision-plus-action routes for serial recall. The motor commands and proprioceptive feedback pictured here describe a broader sensorimotor setting. All three evaluated conditions use joint-state supervision during training and report location symbols at evaluation, with no joint or proprioceptive inputs. S0-null, SF-null, and SF-vision differ in the response content supplied to the recurrent state; SF-vision uses recorded action video. The two conceptual routes do not denote these three conditions.

Previous research links action-related processing to the encoding and retention of spatial sequences (Langner et al., 2014; Pochon et al., 2001; Brunetti et al., 2018; Zhao et al., 2020) (Section 2). Less direct evidence addresses the visual input that accompanies recall itself. We therefore ask whether visual input associated with a recall response contributes to retrieving the next location as recall unfolds through successive actions.

At the recall stage, competitive queuing accounts describe successive choices through competition among activated items, with response suppression reducing an item's activation after it has been recalled (Hurlstone & Hitch, 2015). Responses also generate external visual input: participants can

see their hand reach a block in the physical Corsi task, while touched blocks light up to confirm registered responses in digital versions (Brunetti et al., 2014). We refer to such response-generated visual input as visual feedback. Visual feedback indicates the identity of the selected location, but it can also unfold over time between successive decisions. To examine whether this temporal unfolding contributes to subsequent retrieval, we test whether changing the timing of the same visual content alters the next choice when the selected location is held constant. Unlike human experiments, a computational model allows alternative feedback schedules to be compared from the same memory state and response history.

To address these questions, we introduce CorsiBot, an embodied cognitive robot model of spatial serial recall in a simulated setting (Figure 1B). It observes videos of a Franka Panda arm pointing to sequences of two to nine locations on a nine-location board, then recalls each sequence by successively selecting among the location symbols. A shared two-layer LSTM network processes both observation and recall without resetting its state between phases. Joint-state prediction during training links visual observation to movement configuration, while the model receives no joint measurements during recall. Under a common architecture and training budget, we compare three conditions that differ in what the model receives between selections: no input, symbolic input indicating the selected location, and visual feedback showing the arm moving to that location. The visual-feedback model shows qualitative features of human spatial serial recall, including lower accuracy for longer lists and order errors concentrated around nearby list positions.

We train position-specific offline readouts on natural recall states, in which feedback frames appear in their recorded order, and hold them fixed to track upcoming-location information across time and feedback interventions. With network weights frozen, we start from the same recurrent state after the same selection and compare feedback schedules that share the same frames, including the final frame, but differ in timing. Changing feedback timing shifts both the next choice and the fixed readout, even though the just-selected location remains decodable from the final visual-encoder features. The same final visual input, carrying the same location identity, can therefore precede different next choices. These findings show that visual feedback history contributes to sequential retrieval in CorsiBot and motivate tests of similar dependence in human visuospatial working memory. More broadly, CorsiBot provides an embodied platform for testing how visual feedback shapes sequential memory retrieval.

2 Related Work

Serial order and spatial organization Serial recall requires associating items with their places in a sequence. Positional accounts encode item–position associations (Henson, 1998), while competitive-queuing accounts select responses through competition among activated items (Houghton, 1990). In spatial sequences, relational structure can additionally support chunking (Zhang et al., 2022). Hurlstone & Hitch (2015) compared combinations of order mechanisms and found that joint error and latency data favored a primacy gradient, positional markers, and response suppression among the tested models. Models with similar accuracy and transposition patterns nonetheless predicted different error latencies, showing that behavioral agreement alone cannot identify an order mechanism. We therefore complement behavioral comparisons with analyses of CorsiBot's internal states and controlled changes to its recall input.

Action in visuospatial sequence memory Corsi procedures combine spatial presentation and action in different ways. Digital implementations permit precise stimulus control and timing measurements (Brunetti et al., 2014), while walking versions extend the task to locomotor space (Piccardi et al., 2008; 2013). During presentation, distance-congruent intervals and visible trajectories between successive locations can improve recall even without a visible hand (Brunetti et al., 2018), and active manual tracing during encoding improves backward spatial serial recall (Zhao et al., 2020). Action observation also engages the observer's motor system: it increases motor excitability (Fadiga et al., 1995), and observed, imagined, and executed reaching recruit overlapping dorsal premotor and posterior parietal regions (Filimon et al., 2007). In spatial sequence tasks, motor-related activity during encoding (Langner et al., 2014) and preparation of the intended response during retention (Pochon et al., 2001) link memory to subsequent action. These links motivate CorsiBot's auxiliary joint-state supervision, which associates visual observation states with the demonstrated movement configuration during training (Appendix B). These findings, however, concern processes that precede reproduction. Sensory feedback during reproduction has been examined in memorized music performance, where

altered auditory feedback that matches other events in the sequence disrupts production (Pfordresher & Palmer, 2006). Our question concerns the corresponding visual feedback in spatial working memory, that is, the visual input that reporting actions produce between successive recall responses, and how it influences retrieval of the next location.

Recurrent and robotic models of serial recall Botvinick & Plaut (2006) showed that learned recurrent representations reproduce serial-position effects and transposition gradients, with the representation of each selected item fed back to the recurrent layer; this feedback conveys the identity of the selected item, analogous to our symbolic condition. Wang et al. (2025) found that recurrent networks trained on sequence working memory can develop fixed-slot or dynamic-chain representations, which generalize differently to unseen sequence lengths. Trained recurrent networks have also been compared with neural recordings to investigate cognitive computations (Mante et al., 2013; Yang et al., 2019), and dynamical analyses characterize how such networks implement their computations (Sussillo & Barak, 2013). Robotic models have examined sensorimotor sequence learning (Yamashita & Tani, 2008) and working-memory processes supporting language-guided action (Giorgi et al., 2021). Most closely related, Tekülve et al. (2019) combined visual scene memory with serial-order representations to reproduce learned color sequences through pointing, with action-completion signals advancing the sequence. In that architecture, action completion acts as a designed signal for sequence transitions; in CorsiBot, visual feedback is a temporally extended input whose influence on the next selection is learned and then tested. Linear probes measure information decodable from hidden states (Alain & Bengio, 2016), while interventions on internal activations test hypotheses about its causal role (Geiger et al., 2021). Our interventions instead hold the weights and recurrent state fixed and alter only the timing of the visual feedback that follows a selection, testing whether feedback history, beyond its content, changes the next selection and the decodable upcoming-location information.

3 TASK AND MODEL

Task setup The human Corsi task presents a sequence of locations for later reproduction (Figure 1A). We keep that presentation-and-recall requirement in a simulated robotic setting (Figure 1B). A Franka Panda manipulator points successively to distinct targets on a fixed nine-location board. A stationary external camera records the board and arm throughout each trial. Synchronized seven-joint trajectories describe the demonstrated movement and supply training labels. The model observes RGB frames, not joint measurements, at evaluation.

Each pointing segment contributes 12 sampled frames. Lists contain two to nine targets without repetition. After the last observed target, the robot returns to its home position; a go cue then starts model recall. The model selects a location symbol at each decision and ends its response with an end-of-sequence token. Figure 1B depicts physical pointing as the task analogue of human recall; the measured model response is the symbol sequence. There is no additional blank retention interval after the return movement.

Recurrent architecture and training We model the task with a recurrent visual motor network that receives each trial as a continuous video stream. During presentation, the arm moves from one block to the next, and frames sampled from this stream pass through a visual encoder into two stacked LSTM layers. The recurrent state is never reset: it carries through presentation, return to home, the go cue, and recall. The observed sequence therefore reaches response generation through the same recurrent dynamics. At each recall step, the network selects one of the nine locations or an end token.

We compare three versions with identical trainable architecture. They differ only in what the network receives after each selection. In S0-null, the feedback interval carries no input, so recall relies on the recurrent state alone. In SF-null, the network receives the embedding of the selected location as symbolic feedback. In SF-vision, the network observes recorded video of a pointing action to the selected location, so its response re-enters through the same visual pathway as the presented sequence. In each version, the next decision follows 12 recurrent feedback steps. Figure 1C sketches the relation between visual and action feedback, and Appendix A specifies the inputs and state updates.

4 BEHAVIORAL PERFORMANCE

Recall declines as lists grow longer We first examined how recall changed with list length (Figure 2a; Table 1A). Adult Corsi span is typically around six locations (Kessels et al., 2000). S0-null showed an earlier loss of complete-list recall, falling to 46.67% at five items. SF-null and SF-vision maintained high accuracy through five items and declined from six onward, a pattern qualitatively closer to this human capacity range. SF-vision also showed less between-seed variation in position accuracy across lengths four to nine, as reflected by its narrower shaded range. Whole-list success and stability nevertheless differed: SF-null had the highest mean exact-list accuracy at length nine, with seed values ranging from 16% to 64%.

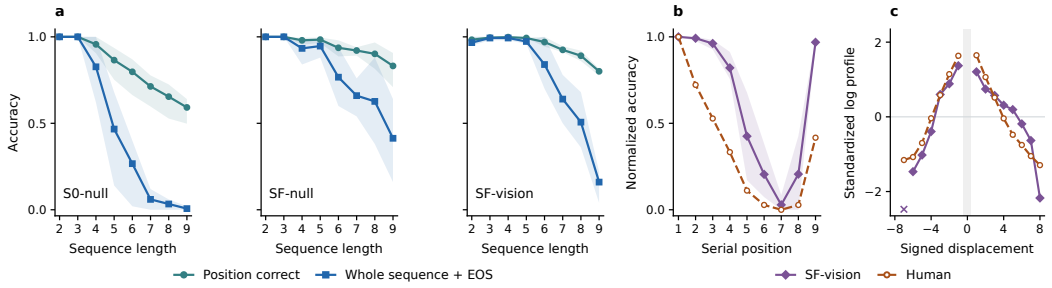


Figure 2: **Length-dependent recall and human-referenced serial-order profiles. a,** Position accuracy (green) and whole-sequence accuracy including EOS (blue), from left to right: S0-null, SF-null, and SF-vision. **b,** Serial-position accuracy for SF-vision and humans, independently min–max normalized. Model curves are normalized within each seed before averaging. **c,** Standardized log frequency of order errors by signed displacement d (output minus presented position), excluding $d = 0$. Model counts are pooled across seeds before log transformation and standardization; gaps and crosses mark zero counts. Model results in b/c use the same held-out nine-item lists. Human curves are digitized from the ungrouped condition of Experiment 3 in Hurlstone & Hitch (2015). Model lines in a/b show seed means and shading spans the minimum to maximum across seeds. Full scoring and the other model–human comparisons appear in Appendix E.

Table 1: **Recall accuracy and error profiles across feedback conditions.** Percentages are mean ± sample SD across three training seeds. All conditions use the same held-out lists. Scoring and denominators are defined in Appendix E.

Measure	S0-null	SF-null	SF-vision
A. Exact-list accuracy, including EOS (↑)			
Length 2	100.00 ± 0.00	100.00 ± 0.00	96.67 ± 2.89
Length 3	100.00 ± 0.00	100.00 ± 0.00	99.33 ± 1.15
Length 4	82.67 ± 19.22	93.33 ± 8.08	99.33 ± 1.15
Length 5	46.67 ± 28.73	94.67 ± 5.77	97.33 ± 1.15
Length 6	26.67 ± 21.39	76.67 ± 17.01	84.00 ± 15.10
Length 7	6.00 ± 6.00	66.00 ± 11.14	64.00 ± 14.00
Length 8	3.33 ± 3.06	62.67 ± 25.01	50.67 ± 16.17
Length 9	0.67 ± 1.15	41.33 ± 24.11	16.00 ± 10.39
All lengths	41.35 ± 10.36	77.66 ± 12.10	74.23 ± 6.63
B. Nine-item serial-order measures			
Strict-position accuracy (↑)	59.19 ± 8.15	83.19 ± 10.73	80.07 ± 2.19
Adjacent-error share	32.30 ± 3.90	52.23 ± 2.93	48.86 ± 7.26
Lists with repetitions (↓)	98.67 ± 2.31	28.00 ± 17.78	71.33 ± 9.45

A: the all-length score weights lists equally. B: nine-item recall. Strict-position accuracy excludes EOS; repetitions count lists with any repeated response. Adjacent-error share is the fraction of within-list wrong responses displaced by one position, including repeated responses.

Complete-list success and response errors distinguish the conditions Table 1B complements exact recall with strict-position accuracy, the locality of within-list order errors, and repeated re-

sponses. These measures draw on the serial-order analyses used in spatial recall (Hurlstone & Hitch, 2015); adjacent-error share is a summary derived from the displacement distribution. At length nine, SF-null and SF-vision have similar mean strict-position accuracy (83.19% and 80.07%), while their exact-list accuracy (41.33% and 16.00%) and the proportion of lists containing repetitions (28.00% and 71.33%) differ more widely. S0-null repeats a location on nearly every nine-item list. Adjacent-error share characterizes the errors conditional on their occurrence; repeated responses also contribute to this statistic. The two measures therefore describe overlapping aspects of recall.

Human-referenced serial-order profiles For nine-item recall, SF-vision showed an asymmetric serial-position profile with high initial accuracy, a late trough, and terminal recovery (Figure 2b). Both its mean normalized curve and the human reference reached their minimum at position seven and recovered thereafter. SF-vision's final-position recovery was stronger, nearly returning to its initial level, whereas the human curve retained a larger primacy advantage. Order errors also shared a locality pattern: they were most frequent at displacements of one position and became less frequent as displacement increased (Figure 2c). These comparisons show common features of spatial serial recall alongside differences in their magnitude. The corresponding S0-null and SF-null profiles are provided in Appendix E, Figure E1.

These profiles characterize how recall accuracy varies across serial positions and how order errors are distributed around the target position. Similar accuracy curves and local displacement gradients can arise from different mechanisms (Hurlstone & Hitch, 2015), so the shared behavioral features leave open how the model represents the sequence and selects successive items. We next examine whether item identity and list position retain the same organization from observation to recall, and how the upcoming target becomes readable between successive choices. Tracking the recurrent state during this interval connects the sequence representations to the visual feedback received after each response.

5 Sequence representations across observation and recall

Target information changes format between observation and recall We first examined whether item identity can be read in the same way as the network moves from observing a sequence to recalling it. We sampled the second-layer hidden state (h_2) after each presented item (O_p) and before each recall choice (R_p). Observation and recall followed different state trajectories (Figure 3a), while full-state analyses showed that both sequence step and item identity contributed to state variation in each phase (Figure 3b). Recall states therefore retained an association with the original-list target. However, an identity decoder trained on observation states performed near chance when applied directly to recall states (9.34%; chance 11.1%). To examine whether identity was accessible through a different set of coordinates, we fitted an offline linear alignment between recall and observation states paired by list and position. After alignment, the same decoder reached 75.39%, compared with 10.10% when pairing preserved position but shuffled lists (Appendix C, Figure C1). Preserving position alone was therefore insufficient to recover the identity readout. These results support a phase-dependent organization of target information: identity remains readable from recall states, but the observation-trained readout does not transfer directly. We therefore use a recall-trained readout to track the upcoming target between successive choices.

Next-target information becomes readable between successive choices We next asked when the target for the next choice becomes readable while the model receives visual feedback from its previous response. In SF-vision, this interval contains 12 frames of the pointing action to the location just selected. For each upcoming position, we trained a linear readout on h_2 immediately before natural recall choices, using the corresponding original-list target as its label. We then froze the readout and followed its accuracy throughout the preceding feedback interval (Figure 3c). This tests when the state makes the next target accessible in the form found at choice time. Accuracy rose from near chance immediately after the preceding choice (11.25%) to 78.31% before the next choice, with most of the rise occurring in the second half of the clip. Over this interval, a separate visual-feature readout came to identify the just-selected location, reaching 100% at the final frame. The feedback thus depicts the preceding selection while the recurrent state increasingly supports readout of the upcoming target. The low initial score concerns this fixed choice-time readout and does not establish that the target is absent in other forms. This temporal association motivates testing whether next-target

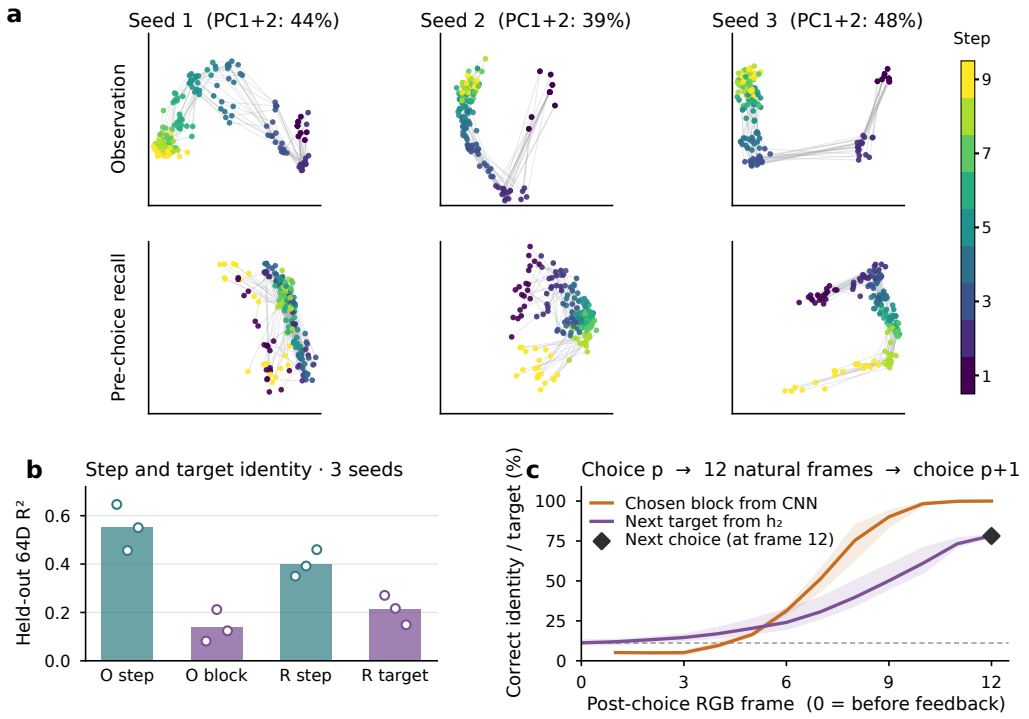


Figure 3: **Recurrent-state organization and the next-target readout under natural feedback.** SF-vision. **a,** Twenty held-out lists per seed during observation (top) and before recall choices (bottom). Within each seed, a single PCA fitted on training states supplies shared axes to both phases; PCA bases are fitted independently across seeds. Dots are O_p or R_p states, gray lines connect successive states within a trial, and color denotes step rather than item identity. **b,** Held-out full-64D state R^2 for step or original-list item identity across three seeds; recall identity is the intended target even on erroneous trials. **c,** After choice p, the model receives 12 natural RGB feedback frames before choice $p+1$. Frozen readouts identify the just-chosen block from CNN features (orange) and the next original-list target from h_2 (purple). The diamond marks actual next-choice accuracy only at frame 12. Lines are seed means and bands seed ranges. The same h_2 probe is reused in Figure 4; methods are in Appendix C.

readability depends on the feedback history. Section 6 holds the number of recurrent updates fixed, rearranges the feedback frames, and measures both this same readout and the actual next choice.

6 Visual Feedback and Next-Item Selection

Next-item selection depends on feedback history The natural-feedback analysis showed that next-target readability rises between successive choices (Section 5). We next asked whether, within SF-vision, this change and the next choice depend on the temporal arrangement of the visual input. For each comparison, we copied the complete recurrent state immediately after the same natural response and held the network weights, selected location, and recorded action clip fixed. Each branch received 12 feedback inputs made from the clip's first (A) and last (B) frames. Both schedules contained six A and six B inputs, with the same first input, final two inputs, and directed-transition counts. They differed in when a five-update B period occurred (Figure 4a; Appendix D). This comparison changes feedback history while matching elapsed updates and the visual input at the end of the interval.

Rearranging these inputs changed the next choice: late B yielded 69.95% correct responses, compared with 52.22% for early B, with the same direction in every seed (Figure 4d). Since both branches began from the same state and ended with the same visual inputs after the same number of updates,

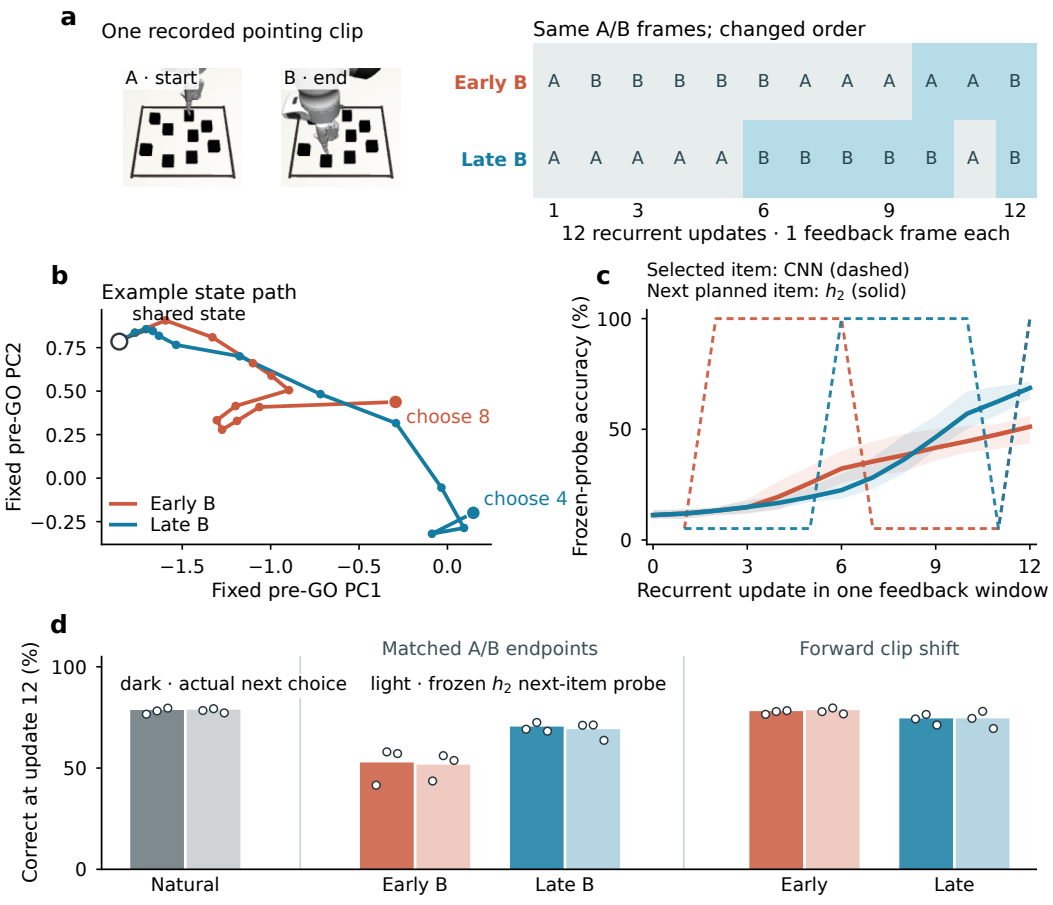


Figure 4: **Feedback timing changes the upcoming-item readout and next choice.** Frozen SF-vision networks across three training seeds. **a,** Recorded RGB endpoints and matched 12-update A/B schedules. **b,** One choice-discordant event in a fixed pre-GO plane; paths are illustrative. **c,** Frozen selected-item visual readout (dashed) and next-target full-h_2 readout (solid); lines are seed means and shading shows seed ranges. Update 0 precedes feedback. **d,** Actual next choice (dark) and the same frozen h_2 probe (light) at update 12 under native, endpoint, and forward-motion feedback; dots are seeds. The forward control differs in endpoint exposure. The interventions affect a single feedback window; see Appendix D.

neither elapsed updates nor the final input alone accounts for their different choices. The network's next selection therefore depends on the feedback history between responses.

The intervention also changed the upcoming target's accessibility to the fixed readout defined under natural feedback. Without refitting, this readout identified the next original-list target at 68.68% with late B and 51.16% with early B (Figure 4c,d). Meanwhile, the frozen visual-feature readout identified the actually selected location at 100% from the final feedback frame in both schedules. Thus, the same final visual input supported equally accurate identification of the preceding selection, while the next-target readout and actual choice differed. Feedback history changes both behavior and the target information accessible through the natural choice-time readout. Their parallel changes identify an accompanying state-readout effect; they do not establish that the probed information mediates the choice effect.

Timing effects depend on the visual trajectory The late-B advantage applies to the matched endpoint schedules. Both were less accurate than native feedback, which yielded 78.13% correct

next choices. When we instead shifted an ordered forward-motion clip earlier or later within the feedback interval, the mean direction reversed: early delivery yielded 77.59%, versus 73.98% for late delivery (Figure 4d). This pair also differed in endpoint exposure, so the comparison does not isolate which aspect of the trajectory produced the reversal (Appendix D). The interventions therefore establish sensitivity to the temporal organization of visual feedback, rather than a general benefit of later feedback. Within a single response interval, feedback history influences which location is selected next and the upcoming target's accessibility to the fixed readout.

7 DISCUSSION

In this work, we trained a small two-layer LSTM network within CorsiBot to learn spatial serial recall from observed pointing actions. We evaluated it as a cognitive model by examining the structure of its successes and errors, using human recall patterns as a reference. Comparing no feedback, symbolic feedback, and visual action feedback highlights why recall accuracy alone is insufficient: models with similar item accuracy can differ in their repeated responses and whole-sequence performance. In the visual-feedback condition, features of human serial-position and transposition profiles appeared without training the network to match human error distributions (Hurlstone & Hitch, 2015). This suggests that a compact recurrent system can develop some characteristic patterns of spatial recall through task learning. Its internal states and responses to feedback manipulations provide a way to investigate how these patterns relate to the interaction between memory and perception.

CorsiBot helps connect psychological descriptions of recall with experiments on the processes that could produce them. The robotic setting brings visual observation, memory, successive responses, and their visual consequences into one continuous task. This allows us to examine how information is represented as the system moves from observing a sequence to recalling it, and how incoming feedback influences the next selection. Accounts in which recalled items update retrieval context (Logan, 2021) offer one possible interpretation: the visual consequences of a response may help change the cues used to retrieve the next item. The small network makes this hypothesis accessible to further testing through analyses and interventions on its state. Such tests can generate predictions for human experiments, including whether changing the visual feedback after a tap changes subsequent recall. The model therefore provides a concrete setting for developing and testing explanations of how perception, memory, and action work together.

The present findings are limited to a fixed board and interventions within a single response interval. Familiar location associations and sensitivity to feedback sequences not encountered during training may contribute to the results. Future work can test new layouts and compare sequences of the same length with different spatial structures, such as path crossings and distances between targets (De Lillo et al., 2016; Zhang et al., 2022). The model could also be extended to study aging and impairment by altering memory retention or introducing local damage to visual or recurrent components. These extensions would ask how changes in memory and perception affect recall errors and the use of visual feedback. Comparisons with human behavioral and neural data would help establish which aspects of the learned model capture human cognitive processes. These studies could inform theories of cognitive architecture and, with human validation, support more precise cognitive assessments by quantifying sequence difficulty, error patterns, and feedback use alongside memory span.

AI USE STATEMENT

AI assistance was used to polish the writing, including language editing and improving clarity, and to help generate parts of the code used in the research. The authors checked all AI-generated code for correctness. The authors are responsible for the research design, analyses, and final text of the paper.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2016. URL https://arxiv.org/abs/1610.01644.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/e995f98d56967d946471af29d7bf99f1-Abstract.html.

Matthew M. Botvinick and David C. Plaut. Short-term memory for serial order: A recurrent neural network model. *Psychological Review*, 113(2):201–233, 2006. doi: 10.1037/0033-295X.113.2.201.

Riccardo Brunetti, Claudia Del Gatto, and Franco Delogu. eCorsi: Implementation and testing of the Corsi block-tapping task for digital tablets. *Frontiers in Psychology*, 5:939, 2014. doi: 10.3389/fpsyg.2014.00939.

Riccardo Brunetti, Claudia Del Gatto, Clarissa Cavallina, Benedetto Farina, and Franco Delogu. Did I see your hand moving? the effect of movement-related information on the Corsi block tapping task. *Psychological Research*, 82(3):459–467, 2018. doi: 10.1007/s00426-016-0834-6.

Caroline Catmur, Vincent Walsh, and Cecilia Heyes. Sensorimotor learning configures the human mirror system. *Current Biology*, 17(17):1527–1531, 2007. doi: 10.1016/j.cub.2007.08.006.

Philip Michael Corsi. *Human Memory and the Medial Temporal Region of the Brain*. PhD thesis, McGill University, 1972. URL https://escholarship.mcgill.ca/concern/theses/05741s554.

Carlo De Lillo, Melissa Kirby, and Daniel Poole. Spatio-temporal structure, path characteristics, and perceptual grouping in immediate serial spatial recall. *Frontiers in Psychology*, 7:1686, 2016. doi: 10.3389/fpsyg.2016.01686.

Luciano Fadiga, Leonardo Fogassi, Giovanni Pavesi, and Giacomo Rizzolatti. Motor facilitation during action observation: A magnetic stimulation study. *Journal of Neurophysiology*, 73(6): 2608–2611, 1995. doi: 10.1152/jn.1995.73.6.2608.

Flavia Filimon, Jonathan D. Nelson, Donald J. Hagler, and Martin I. Sereno. Human cortical representations for reaching: Mirror neurons for execution, observation, and imagery. *NeuroImage*, 37(4):1315–1328, 2007. doi: 10.1016/j.neuroimage.2007.06.008.

Flavia Filimon, Cory A. Rieth, Martin I. Sereno, and Garrison W. Cottrell. Observed, executed, and imagined action representations can be decoded from ventral and dorsal areas. *Cerebral Cortex*, 25(9):3144–3158, 2015. doi: 10.1093/cercor/bhu110.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html.

Ioanna Giorgi, Angelo Cangelosi, and Giovanni Luca Masala. Learning actions from natural language instructions using an ON-world embodied cognitive architecture. *Frontiers in Neurorobotics*, 15: 626380, 2021. doi: 10.3389/fnbot.2021.626380. URL https://www.frontiersin.org/journals/neurorobotics/articles/10.3389/fnbot.2021.626380/full.

Richard N. A. Henson. Short-term memory for serial order: The start-end model. *Cognitive Psychology*, 36:73–137, 1998. doi: 10.1006/cogp.1998.0685. URL https://www.fil.ion.ucl.ac.uk/~rhenson/cp-sem.pdf.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

George Houghton. The problem of serial order: A neural network model of sequence learning and recall. In Robert Dale, Chris Mellish, and Michael Zock (eds.), *Current Research in Natural Language Generation*, pp. 287–319. Academic Press, London, 1990. URL https://www.researchgate.net/publication/334672441_The_problem_of_serial_order_A_neural_network_model_of_sequence_learning_and_recall.

Mark J. Hurlstone and Graham J. Hitch. How is the serial order of a spatial sequence represented? insights from transposition latencies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(2):295–324, 2015. doi: 10.1037/a0038223.

Kristjan Kalm and Dennis Norris. A shared representation of order between encoding and recognition in visual short-term memory. *NeuroImage*, 155:138–146, 2017. doi: 10.1016/j.neuroimage.2017.04.047.

R. P. C. Kessels, M. J. E. van Zandvoort, A. Postma, L. J. Kappelle, and E. H. F. de Haan. The Corsi block-tapping task: Standardization and normative data. *Applied Neuropsychology*, 7(4):252–258, 2000. doi: 10.1207/S15324826AN0704_8.

Robert Langner, Melanie A. Sternkopf, Tanja S. Kellermann, Christian Grefkes, Florian Kurth, Frank Schneider, Karl Zilles, and Simon B. Eickhoff. Translating working memory into action: Behavioral and neural evidence for using motor representations in encoding visuo-spatial sequences. *Human Brain Mapping*, 35(7):3465–3484, 2014. doi: 10.1002/hbm.22415.

Gordon D. Logan. Serial order in perception, memory, and action. *Psychological Review*, 128(1):1–44, 2021. doi: 10.1037/rev0000253.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Valerio Mante, David Sussillo, Krishna V. Shenoy, and William T. Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503(7474):78–84, 2013. doi: 10.1038/nature12742.

Peter Q. Pfordresher and Caroline Palmer. Effects of hearing the past, present, or future during music performance. *Perception & Psychophysics*, 68(3):362–376, 2006. doi: 10.3758/BF03193683. URL https://www.acsu.buffalo.edu/~pqp/pdfs/PfordresherandPalmer2006PandP.pdf.

Laura Piccardi, Giuseppe Iaria, Maura Ricci, Filippo Bianchini, Laura Zompanti, and Cecilia Guariglia. Walking in the Corsi test: Which type of memory do you need? *Neuroscience Letters*, 432(2):127–131, 2008. doi: 10.1016/j.neulet.2007.12.044.

Laura Piccardi, Filippo Bianchini, Ornella Argento, Anna De Nigris, Andrea Maialetti, Liana Palermo, and Cecilia Guariglia. The walking Corsi test (WalCT): Standardization of the topographical memory test in an italian population. *Neurological Sciences*, 34(6):971–978, 2013. doi: 10.1007/s10072-012-1175-x.

Jean-Baptiste Pochon, Richard Levy, Jean-Baptiste Poline, Sophie Crozier, Stéphane Lehéricy, Bernard Pillon, Bernard Deweer, Denis Le Bihan, and Bruno Dubois. The role of dorsolateral prefrontal cortex in the preparation of forthcoming actions: An fMRI study. *Cerebral Cortex*, 11(3):260–266, 2001. doi: 10.1093/cercor/11.3.260.

David Sussillo and Omri Barak. Opening the black box: Low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural Computation*, 25(3):626–649, 2013. doi: 10.1162/NECO_a_00409.

Jan Tekülve, Adrien Fois, Yulia Sandamirskaya, and Gregor Schöner. Autonomous sequence generation for a neural dynamic robot: Scene perception, serial order, and object-oriented movement. *Frontiers in Neurorobotics*, 13:95, 2019. doi: 10.3389/fnbot.2019.00095. URL https://www.frontiersin.org/journals/neurorobotics/articles/10.3389/fnbot.2019.00095/full.

Mingye Wang, Stefano Fusi, and Kimberly Stachenfeld. Brain-like slot representation for sequence working memory in recurrent neural networks. In *ICLR 2025 Workshop on Representational Alignment (Re-Align)*, 2025. URL https://openreview.net/pdf?id=wlJvdLjnhR.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision*, pp. 3–19, 2018. doi: 10.1007/978-3-030-01261-8_1. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Yuxin_Wu_Group_Normalization_ECCV_2018_paper.html.

Yuichi Yamashita and Jun Tani. Emergence of functional hierarchy in a multiple timescale neural network model: A humanoid robot experiment. *PLoS Computational Biology*, 4(11): e1000220, 2008. doi: 10.1371/journal.pcbi.1000220. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000220.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014. doi: 10.1073/pnas.1403112111.

Guangyu Robert Yang, Madhura R. Joglekar, H. Francis Song, William T. Newsome, and Xiao-Jing Wang. Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, 22(2):297–306, 2019. doi: 10.1038/s41593-018-0310-2.

He Zhang, Yanfen Zhen, Shijing Yu, Tenghai Long, Bingqian Zhang, Xinjian Jiang, Junru Li, Wen Fang, Mariano Sigman, Stanislas Dehaene, and Liping Wang. Working memory for spatial sequences: Developmental and evolutionary factors in encoding ordinal and relational structures. *Journal of Neuroscience*, 42(5):850–864, 2022. doi: 10.1523/JNEUROSCI.0603-21.2021.

Yangke Zhao, Chuansheng Chen, and Xiuying Qian. The role of hand movement in spatial serial order memory. *Multisensory Research*, 33(3):313–335, 2020. doi: 10.1163/22134808-20191350.

A Model architecture and training details

A.1 CorsiBot model architecture

A.1.1 Visual encoder

Let $(y_1, \ldots, y_L)$ be a trial's target sequence, with $2 \le L \le 9$ and distinct locations $y_j \in \{0, \ldots, 8\}$. Each pointing segment supplies $K = 12$ RGB images of size 128×128. The response target appends an end-of-sequence (EOS) token, indexed by 9.

We use a custom four-stage convolutional neural network (CNN), trained from scratch, to encode each image independently. For a normalized image $\widetilde{I}_t \in \mathbb{R}^{3\times128\times128}$, stage a computes

$$F_t^{(0)} = \widetilde{I}_t, \qquad F_t^{(a)} = \mathrm{ReLU}\Big(\mathrm{GN}_{G_a}\left(W_a * F_t^{(a-1)} + b_a\right)\Big), \quad a = 1, \ldots, 4. \tag{1}$$

Here $*$ denotes convolution, W_a and b_a are learned parameters, and GN_{G_a} is Group Normalization with G_a channel groups (Wu & He, 2018). Layer dimensions are given in Table A1.

Table A1: Custom CNN encoder. Shapes are channels × height × width. Each convolution includes a bias and is followed by Group Normalization and ReLU.

Stage	Channels	Kernel	Stride	Padding	Groups	Output shape
Input	3	–	–	–	–	$3 \times 128 \times 128$
1	$3 \to 24$	5	2	2	6	$24 \times 64 \times 64$
2	$24 \to 48$	3	2	1	8	$48 \times 32 \times 32$
3	$48 \to 64$	3	2	1	8	$64 \times 16 \times 16$
4	$64 \to 64$	3	2	1	8	$64 \times 8 \times 8$
Average pool	64	–	–	–	–	$64 \times 1 \times 1$

Adaptive average pooling reduces the last feature map to one value per channel:

$$v_{t,d} = \frac{1}{8 \cdot 8} \sum_{u=1}^{8} \sum_{w=1}^{8} F_{t,d,u,w}^{(4)}, \qquad v_t = \phi(\widetilde{I}_t) \in \mathbb{R}^{64}. \tag{2}$$

The same encoder processes presentation, return-to-home, and SF-vision feedback images.

A.1.2 Continuous recurrent module

The recurrent module consists of two stacked long short-term memory (LSTM) cells (Hochreiter & Schmidhuber, 1997). Let $s_t \in \{0, 1\}$ indicate the go cue, and let $r_t \in \mathbb{R}^{32}$ be the response-feedback vector. The inputs to the two layers are

$$x_t^{(1)} = [v_t; s_t; r_t] \in \mathbb{R}^{97}, \qquad x_t^{(2)} = h_t^{(1)} \in \mathbb{R}^{64}. \tag{3}$$

Each layer $a \in \{1, 2\}$ maintains a hidden state $h_t^{(a)} \in \mathbb{R}^{64}$ and a cell state $c_t^{(a)} \in \mathbb{R}^{64}$. Its input, forget, and output gates and candidate cell update are

$$\begin{aligned} i_t^{(a)} &= \sigma(W_i^{(a)} x_t^{(a)} + U_i^{(a)} h_{t-1}^{(a)} + b_i^{(a)}), \\ f_t^{(a)} &= \sigma(W_f^{(a)} x_t^{(a)} + U_f^{(a)} h_{t-1}^{(a)} + b_f^{(a)}), \\ o_t^{(a)} &= \sigma(W_o^{(a)} x_t^{(a)} + U_o^{(a)} h_{t-1}^{(a)} + b_o^{(a)}), \\ \widetilde{c}_t^{(a)} &= \tanh(W_c^{(a)} x_t^{(a)} + U_c^{(a)} h_{t-1}^{(a)} + b_c^{(a)}). \end{aligned} \tag{4}$$

Here σ is the logistic sigmoid, W and U are input and recurrent weights, and b is the combined bias. The states update as

$$\begin{aligned} c_t^{(a)} &= f_t^{(a)} \odot c_{t-1}^{(a)} + i_t^{(a)} \odot \widetilde{c}_t^{(a)}, \\ h_t^{(a)} &= o_t^{(a)} \odot \tanh(c_t^{(a)}), \end{aligned} \tag{5}$$

where $\odot$ denotes elementwise multiplication. The gates f_t, i_t, and o_t control retention, updating, and hidden-state output, respectively.

All four states start at zero and continue through the $12L$ presentation frames, 12 return frames, one go step, and recall. Presentation and return use $s_t = r_t = 0$; the go step uses $v_t = r_t = 0$ and $s_t = 1$. Padding preserves the preceding states. Target length and serial position are not recurrent inputs.

A.1.3 Response and motor-supervision heads

The response head maps the second layer's hidden state at decision time t_j to ten location/EOS logits:

$$\ell_j = W_{\text{out}} h_{t_j}^{(2)} + b_{\text{out}}, \qquad p_{j,k} = \frac{\exp(\ell_{j,k})}{\sum_{a=0}^{9} \exp(\ell_{j,a})}, \qquad \widehat{y}_j = \arg \max_{k \in \{0,\ldots,9\}} p_{j,k}. \tag{6}$$

The head has ten outputs and is shared across decision positions. The first decision follows the go step, and subsequent decisions follow 12 recurrent updates, $t_{j+1} = t_j + 12$. Greedy recall stops at the first EOS or after ten decisions. Previously selected locations remain available.

A second linear head predicts the demonstrated seven-joint configuration from observation states:

$$\widehat{q}_t = W_{\text{joint}} h_t^{(2)} + b_{\text{joint}} \in \mathbb{R}^7, \qquad W_{\text{joint}} \in \mathbb{R}^{7 \times 64}. \tag{7}$$

The prediction $\widehat{q}_t$ is in standardized joint-angle units and is supervised during observation. Recalled outputs are location symbols.

A.1.4 Three response-feedback conditions

The conditions differ in the input between decisions. Let a_j be the feedback location after decision j and $E \in \mathbb{R}^{10 \times 32}$ the response embedding. For a non-EOS selection, feedback step $k = 1, \ldots, 12$ uses

$$(v_{t_j+k}, r_{t_j+k}) = \begin{cases} (0_{64}, 0_{32}), & \text{S0-null,} \\ (0_{64}, E[a_j]), & \text{SF-null,} \\ (\phi(\widetilde{I}_{j,k}), 0_{32}), & \text{SF-vision.} \end{cases} \tag{8}$$

Here $\widetilde{I}_{j,k}$ is the kth feedback frame of the selected pointing movement. The go indicator is zero throughout recall. S0-null continues updating with zero feature input, SF-null repeats one symbol embedding, and SF-vision processes successive movement frames. The zero vectors are supplied directly to the recurrent module.

A.2 Training procedure

A.2.1 Sampling, normalization, and data partition

Each observation segment is sampled at 12 equally spaced timestamps, using the nearest RGB frame and linearly interpolated joint angles. RGB values are scaled to $[0, 1]$. Images and joint targets are standardized using training-set channelwise and jointwise statistics, respectively. All conditions use the sequence partition in Table A2.

Table A2: Sequence partition shared by all three feedback conditions.

Length	Training	Validation	Original test
2	72	10	20
3	300	30	50
4	600	30	50
5–9 (each)	800	30	50
Total	4,972	220	370

A.2.2 SEQUENCE AND JOINT-ANGLE LOSSES

For a minibatch of B trials, let $\mathcal{D}$ contain all valid target decisions, including EOS. Sequence cross-entropy is

$$\mathcal{L}_{\text{seq}} = -\frac{1}{N_{\text{dec}}} \sum_{(b,j)\in\mathcal{D}} \log p_{b,j,y_{b,j}}, \qquad N_{\text{dec}} = \sum_{b=1}^{B} (L_b + 1). \tag{9}$$

Here $p_{b,j,y_{b,j}}$ is the target-token probability under the training feedback history. The loss weights valid decisions equally and excludes padding.

Let $\mathcal{O}$ contain the valid presentation-frame pairs (b, t), with $N_{\text{obs}} = |\mathcal{O}|$ and standardized joint targets $\widetilde{q}_{b,t}$. The auxiliary loss and combined objective are

$$\mathcal{L}_{\text{joint}} = \frac{1}{7N_{\text{obs}}} \sum_{(b,t)\in\mathcal{O}} \sum_{d=1}^{7} \left(\widehat{q}_{b,t,d} - \widetilde{q}_{b,t,d}\right)^2, \tag{10}$$

$$\mathcal{L} = \mathcal{L}_{\text{seq}} + 0.1\mathcal{L}_{\text{joint}}. \tag{11}$$

The joint term averages squared errors over seven standardized coordinates at valid observation frames. Return and recall frames receive no joint supervision.

A.2.3 TEACHER FORCING AND SCHEDULED SAMPLING

Each model is trained for 80 epochs. SF-null and SF-vision transition from target to predicted feedback through scheduled sampling (Bengio et al., 2015). At epoch e, the probability of using a greedy prediction is

$$\rho_e = \begin{cases} 0, & 1 \le e \le 40, \\ (e-41)/39, & 41 \le e \le 80. \end{cases} \tag{12}$$

At each feedback interval, the target token is used with probability $1 - \rho_e$. The chosen token supplies the entire 12-step interval. S0-null uses the same training budget with zero feedback input.

Training evaluates all $L + 1$ target decisions even after an early predicted EOS. Once EOS is chosen as feedback, SF-null repeats its EOS embedding and SF-vision receives zero visual features for the remaining supervised intervals. Evaluation terminates at the first predicted EOS.

A.2.4 OPTIMIZATION

The CNN, recurrent module, and active output heads are trained jointly using AdamW (Loshchilov & Hutter, 2019), with a learning rate of 10^{-3}, weight decay of 10^{-5}, and a batch size of 16. Training uses full backpropagation through time across presentation, return, the go cue, and recall.

B NEURAL MOTIVATION

Joint-angle supervision connects observed movement with an action-related prediction target. Sensorimotor training can alter observation–execution mappings (Catmur et al., 2007), while executed, observed, and imagined reaching recruit overlapping frontoparietal regions with distinguishable activity patterns (Filimon et al., 2007; 2015). These findings motivate learning a relationship between visual movement and arm configuration.

The model's visual and recurrent components also have precedents in task-trained neural models. Hierarchical visual networks predict responses in macaque V4 and inferotemporal cortex (Yamins et al., 2014), and recurrent dynamics can account for sensory selection and integration in prefrontal populations (Mante et al., 2013). Shared position-sensitive activity across encoding and recognition provides a further motivation for studying sequence representations across task phases (Kalm & Norris, 2017). These connections concern computational functions, rather than a one-to-one assignment of CorsiBot units to brain regions.

C REPRESENTATION ANALYSES

State sampling and joint projection For the common epoch-80 checkpoints, O_p is the 64-dimensional second-layer hidden state h_2 at the last update of observed item p, and R_p is the state immediately before recall choice p. The latter includes GO and, for $p > 1$, natural feedback from earlier responses. For each SF-vision seed, one PCA was fitted to the pooled O_1–O_9 and R_1–R_9 states from 800 training lists (14,400 endpoints per seed). Figure 3a shows 20 distinct lists from the previously explored 180-list confirmation set per seed, without refitting or trial averaging. Observation and recall share the PCA basis and axis limits within a seed, while each seed has its own basis. PC1 and PC2 together explain approximately 44%, 39%, and 48% of fitted variance for seeds 1, 2, and 3. These plotted paths illustrate state variation; they are not an identity decoder.

For the full-state analysis, we fitted one-hot linear models with step, the identity of the corresponding original-list item, or both as predictors of the 64-dimensional h_2 state. Observation and recall were fitted separately on 800 training lists (7,200 states per phase) and evaluated on 180 confirmation lists (1,620 states per phase) for each seed. Reported R^2 is relative to the confirmation phase mean and uses the full state. Recall identity is the planned target, not necessarily the actual selected block. The three-seed mean one-factor R^2 values are shown in Table C1. Step and elapsed time are coupled by the fixed presentation and feedback update schedules, so these regressions do not isolate an abstract serial-position variable.

Table C1: **Held-out full-state h_2 variance explained by step and target identity.** Three-seed means at epoch 80, with separate single-factor linear fits for observation (O) and recall (R). The identity is the original-list item at the sampled position. These values describe states and do not identify exclusive functional modules.

Model	O step	O identity	R step	R identity
S0-null	.179	.313	.273	.200
SF-null	.116	.268	.155	.311
SF-vision	.551	.139	.400	.212

Cross-stage identity readout The observation identity decoder uses all positions pooled and is fitted to natural training O_p states; its ridge regularization is selected on validation data. A global regularized affine map from R_p to paired O_p is fitted to the same trial and position in 800 training lists, with regularization selected by validation state-prediction error. The observation decoder is then frozen and applied to mapped recall states. All methods are evaluated on 180 confirmation lists, 1,620 paired states per seed. As controls, we apply the observation decoder directly to recall, use a position-preserving within-training-set pair shuffle before fitting the affine map, and fit an independent recall decoder as a reference. Figure C1 reports all three feedback conditions. The map is fitted offline and does not imply that the recurrent network explicitly computes it. Its decoder is distinct from the position-specific h_2 next-target probe in Figure 3c and Figure 4c.

Position-dependent identity code at a shared time To compare original-list positions without taking their states at different clock times, we analyzed all nine items at the same pre-GO state. For each checkpoint and for h_2 and c_2 separately, a joint position-by-physical-block-identity linear model was fitted on 800 nine-item training lists and evaluated on 180 confirmation lists. Because each list contains each physical block once, position and block main effects are not separately identifiable; we double-centered the position-by-block coefficients and compared the resulting position-dependent identity contrasts. Similarity for a position pair is the cross-validated full-64-dimensional dot product for the same physical block identity, symmetrized across train and confirmation and normalized by within-position cross-set reliability. Figure C2 shows the three-seed SF-vision result. At pre-GO, c_2 similarity averaged .613 for adjacent positions and $-.181$ for positions four or more apart. The same distance gradient occurred at the end of observation and in the other two feedback conditions. The nine original positions still differ in presentation age at this fixed measurement time. A separate residual association between position-pair code similarity and observed wrong-item pair frequency did not have a consistent sign across SF-vision seeds after accounting for position distance, pair endpoints, and mean physical-block distance. The geometry therefore characterizes the state but does not establish a transposition mechanism.

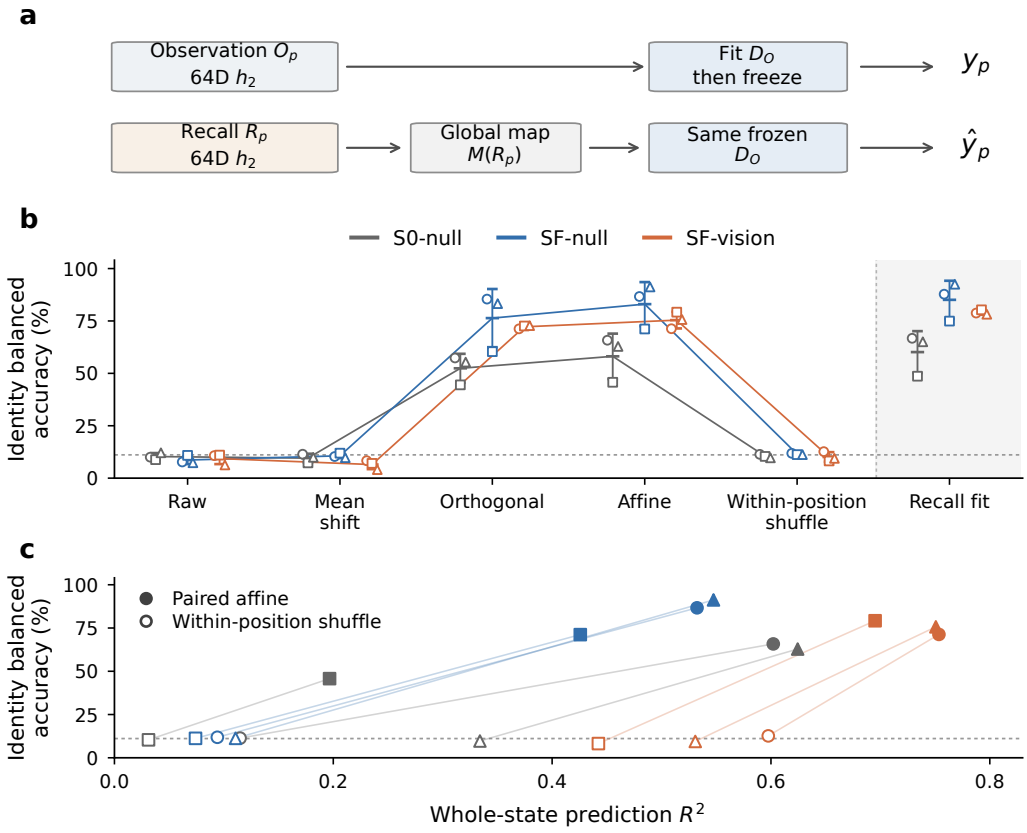


Figure C1: **Paired cross-stage mapping recovers target-identity readout. a,** O_p and R_p come from the same list and target position. An observation identity decoder is fitted across positions and frozen; an offline global map aligns recall to observation states. Labels are original target identities. **b,** Balanced accuracy across three epoch-80 seeds per condition (mean ± sample SD). The first five methods use the frozen observation decoder; "Recall fit" uses a separate recall decoder. Lines join methods, not time points. **c,** Held-out whole-state prediction R^2 versus identity accuracy for paired affine fits (filled) and within-position shuffles (open). Circle, square, and triangle mark seeds 1, 2, and 3; dotted lines mark nine-class chance. Fitting uses 800 lists and 30 validation lists; evaluation uses the previously explored 180-list cohort. The alignment is descriptive, not a model intervention.

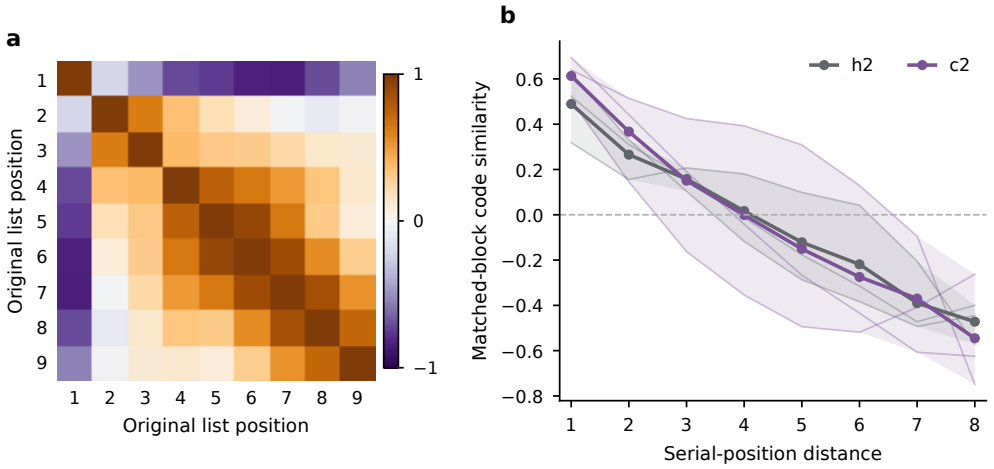


Figure C2: **A position-distance gradient in same-time identity encoding.** SF-vision at epoch 80, evaluated at the pre-GO state before recall feedback. **a,** Mean cross-validated position-dependent identity-code similarity in c_2 across three training seeds. Each cell compares the coding of the same physical block identity when it occupies two different original-list positions. **b,** Similarity by original-list position distance for h_2 and c_2; thick lines are seed means, thin lines individual seeds, and shading their range. Different network coordinate axes are never pooled. Negative values are contrast-code similarities, not evidence of inhibition. All nine positions are measured at one time; original position remains coupled to presentation age.

Natural-feedback next-target readout For each upcoming response position $p = 2, \ldots, 9$, a separate nine-class ridge probe was trained on full h_2 states immediately before the corresponding natural recall decision, with original-list item p as its label. Training used 800 lists; 30 validation lists selected regularization. The feature scaling, weights, and intercept were frozen before evaluating the 180-list confirmation set and all feedback branches. At each event after response $p - 1$, we selected probe p using the known original-list position and applied it at update 0 and after each of 12 native feedback inputs. Each seed contributes eight events on each of 180 lists. The probe is an offline readout, distinct from the network's own ten-class response head; no true position is passed to that head. Feedback-window events share lists and are not independent model replications. A low pre-feedback score under this fixed mapping cannot establish that the next item is absent from the state in some other format.

D FEEDBACK-TIMING INTERVENTION AND READOUTS

Branching protocol We used the common epoch-80 SF-vision checkpoints (three training seeds) and the previously explored confirmation set of 180 nine-item lists per seed. For each natural recall event R1–R8, we copied all four recurrent states (h_1, c_1, h_2, c_2) immediately after that response, preserving the selected location, the corresponding recorded action-clip instance, and the next original-list target. Each branch supplied 12 feedback inputs to this frozen network and read the next output; no branch was continued into an autonomous counterfactual response sequence. The reported 1,440 events per seed are nested within 180 lists. The target remains the next original-list item even when an earlier natural response was erroneous.

The strict endpoint pair repeats the first clip frame A and last clip frame B. Early B uses $[\mathrm{A}, \mathrm{B}^5, \mathrm{A}^5, \mathrm{B}]$ and late B uses $[\mathrm{A}^5, \mathrm{B}^5, \mathrm{A}, \mathrm{B}]$. Each contains six frames of each identity and has the same first and last input, the same final two inputs, two A-to-B transitions, one B-to-A transition, and the same counts of self-transitions. The unequal residence times between those transitions are the intervention. Because these endpoint schedules are repeated frames with abrupt switches, they also change the temporal input distribution relative to the native clip. A separate forward-motion pair compresses the same ordered movement into the 12 updates and shifts it earlier or later, matching the ordered nonzero transitions but not the exposure to the endpoint frames. Table D1 reports this control alongside

the strict pair and native-clip reference. None of these conditions adds or removes visual action observation during presentation.

Fixed readouts and interpretation The current-feedback-identity readout is a frozen probe on the visual encoder features, fitted from natural observation states. Its label is the block that was actually selected and fed back. The next-target readout is a position-specific probe on the full 64-dimensional h_2 state, fitted on natural training/validation recall states; its label is the next original-list item. The probe fits, feature scaling, and model weights were unchanged across feedback branches and update times. The two labels answer different questions, so final-frame current-identity accuracy cannot substitute for next-target accuracy. Figure 4b instead projects one event into a two-dimensional plane fitted on pre-go training states for item 3; this plane is fixed across the two branches, but its shape is only illustrative and is not used for the full-state decoding result.

Table D1: **One-window SF-vision feedback controls at epoch 80.** Mean percentages across three seeds, with 1,440 branched events per seed on the previously explored confirmation lists. Next-choice accuracy is the frozen network's actual output against the next original-list item. The other columns are offline frozen-probe accuracies at feedback update 12. Endpoint early/late schedules match frame and directed-transition counts; the forward-motion pair matches the ordered nonzero transitions but has different endpoint exposure. These are single-window branches, not full-sequence outcomes.

Feedback schedule	Next choice	h_2 next target	Visual current identity
Native video	78.13	78.31	100.00
Endpoint B early	52.22	51.16	100.00
Endpoint B late	69.95	68.68	100.00
Forward motion early	77.59	78.08	100.00
Forward motion late	73.98	74.00	100.00

E Behavioral scoring and human-reference comparisons

E.1 Evaluation and accuracy measures

Table 1 and Figure 2a use the original held-out test set of 370 lists: 20 of length two and 50 at each length from three to nine. The all-length score weights lists equally. Figure 2b/c and Figure E1 use a separate set of 180 nine-item lists, with no complete sequence shared with the original test set. Every condition and seed receives the same lists within each set, using its epoch-80 checkpoint. Table entries report the mean and sample standard deviation across three training seeds; shaded figure bands show the pointwise minimum and maximum across seeds.

Exact-list accuracy requires the complete ordered location sequence followed by EOS at the correct position. Strict-position accuracy divides the number of correctly recalled target positions by the total number of target positions, excluding EOS. Positions missing at or after the first EOS count as errors. The nine-item subset of the original test set contains 50 lists and 450 target positions per seed.

E.2 Order errors and repeated responses

Following Hurlstone & Hitch (2015), signed displacement is $d = r - i$, where a location recalled at output position r appeared at target position i. Targets are distinct within each list, so i is unique. Negative displacements are anticipations and positive displacements are postponements. Each occurrence of a repeated response contributes to the displacement counts. Nine-item lists contain all nine available locations, so location responses cannot be intrusions from outside the list.

Adjacent-error share measures the proportion of within-list wrong responses displaced by one position:

$$A_{\mathrm{adj}} = \frac{n(-1) + n(+1)}{\sum_{d \neq 0} n(d)}, \tag{13}$$

where $n(d)$ counts responses with displacement d. This statistic describes error locality conditional on an error. The proportion of lists with repetitions counts lists containing any repeated location,

divided by the 50 nine-item test lists. Repetitions and transpositions overlap, so these measures do not form an additive error partition.

E.3 HUMAN-REFERENCE TRANSFORMATIONS

The human reference is the ungrouped condition of Experiment 3 in Hurlstone & Hitch (2015), in which 26 participants each recalled 80 nine-item lists. Approximate group means were digitized from their Figure 8A/B. To compare serial-position profile shapes, each model seed's accuracy curve $p_{s,r}$ is min–max normalized before averaging:

$$u_{s,r} = \frac{p_{s,r} - \min_j p_{s,j}}{\max_j p_{s,j} - \min_j p_{s,j}}, \qquad \bar{u}_r = \frac{1}{3}\sum_{s=1}^{3} u_{s,r}. \tag{14}$$

The human group curve is normalized separately by the same rule.

For order-error profiles, model displacement counts are pooled across seeds as $q(d) = \sum_s n_s(d)/(3\times 180 \times 9)$. Counts include every valid response occurrence in the first nine output positions before the first EOS, including repetitions. We standardize $\ln q(d)$ to zero mean and unit population standard deviation over positive-count bins within $S = \{-7, \ldots, -1, 1, \ldots, 8\}$. Zero-count bins appear as gaps and crosses, with no pseudocount. The $d = -8$ count is zero in all three conditions and falls outside this support; $d = 0$ is excluded to focus on errors. The human curve is standardized directly from its 15 published log-frequency coordinates over the same support. These transformations compare profile shapes rather than absolute accuracy or error probability.

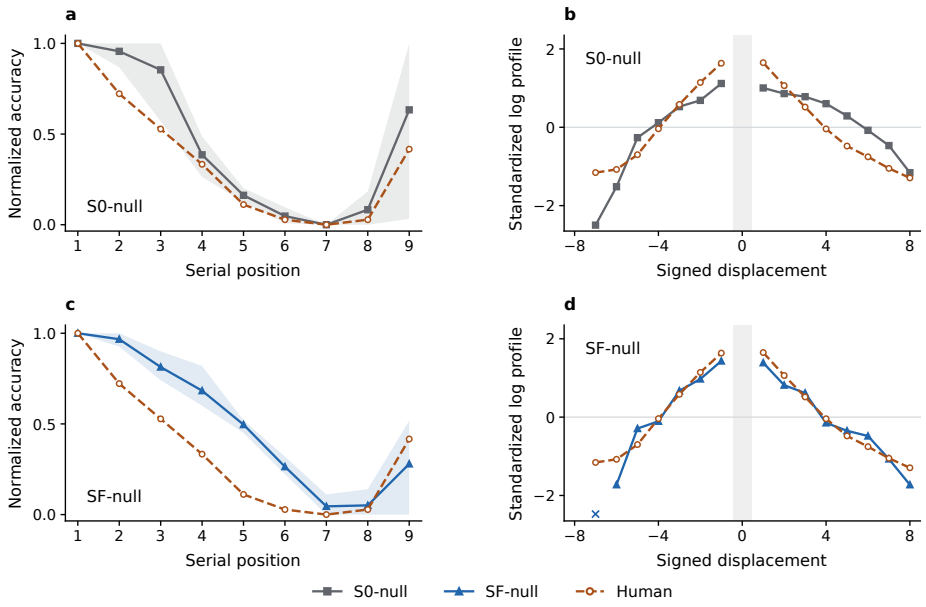


Figure E1: **Serial-position and order-error profiles for S0-null and SF-null. a/b,** S0-null; **c/d,** SF-null. Left: min–max-normalized serial-position accuracy, showing the mean and minimum-to-maximum range across three seeds. Right: pooled order-error frequencies, log-transformed and standardized over positive-count bins; crosses mark zero counts. Orange curves show the digitized human reference. All model results use the same 180 nine-item lists and the transformations described above.