DOUBLE-TAKE: IMPROVING CAPACITY, BINDING, AND SELECTIVITY IN LONG-CONTEXT WITHIN BIDIRECTIONAL MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Processing a long sequence does not guarantee that distant information remains usable. As context grows, a fixed-size memory must retain many associations, connect information across positions, and preserve relevant structure without being overwhelmed by everything it observes. These are distinct requirements: a memory may store more information without representing the relationships needed for retrieval, and it may transport those relationships without preserving the ones that matter. Thus, the challenge is not simply extending context length, but constructing a memory that remains usable as context grows. We study this problem in large-chunk test-time training (TTT) and identify two basic failures: single-step writes accumulate interference, limiting *capacity*, while self-associative writes cannot form cross-position *bindings*. We introduce **Double-Take (DT)**, which repairs these failures with multi-step unpinned updates and cross-position mixing, and adds bidirectional change-sensitive writes for *selective retention*. Capacity alone leaves end-to-end recall at the 0.10 chance floor; adding binding raises it to 0.92. Trained only on $32\,\mathrm{kb}$ genomic contexts, DT reaches 0.927 AUROC at $131\,\mathrm{kb}$ and 0.856 near $1\,\mathrm{Mb}$, while transferring zero-shot from human to mouse. Together, these results identify **capacity, binding, and selectivity** as three requirements for usable long-range memory.

1 INTRODUCTION

Long-context modeling is increasingly measured by how many tokens a model can process, but a large context window does not guarantee that the information inside it remains usable. Long-context language models can lose access to relevant information depending on where it appears in the sequence (Liu et al., 2024), and controlled evaluations show that performance often degrades sharply as context length and retrieval complexity increase (Hsieh et al., 2024). Similar limitations appear in efficient sequence models, where associative recall accounts for a substantial part of the gap with attention-based models (Arora et al., 2024a). Together, these results point to a distinction that context length alone does not capture: *processing distant information and remembering it are different problems*.

This distinction becomes more important as models scale to sequences that cannot remain explicitly addressable. Full attention preserves direct token-to-token interaction but incurs quadratic cost (Vaswani et al., 2017); efficient architectures instead summarize or compress history through structured sequence operators or recurrent states (Gu & Dao, 2023; Nguyen et al., 2023; Schiff et al., 2024). Once information is compressed, three questions become unavoidable. How many distinct associations can survive as new information arrives? Can the memory preserve a relationship between events that occurred at different positions? And, when the context becomes very large, which of those relationships should remain prominent? Genomics makes all three questions concrete: individual nucleotides can matter, their regulatory partners may lie tens of kilobases to megabases away, and only a small fraction of the intervening sequence may be relevant to that interaction (Rao et al., 2014; Dixon et al., 2012; Zhou & Troyanskaya, 2015). The difficulty is therefore not merely reaching farther, but preserving useful structure while doing so.

Test-time training (TTT) (Sun et al., 2024) offers a natural way to revisit this problem because it replaces a fixed recurrent state with an adaptive fast-weight memory that is updated as the sequence

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.

is processed. Large-chunk TTT (LaCT) extends this idea efficiently to long sequences (Zhang et al., 2025). Yet examining what happens when a large chunk is written into that memory exposes the same three questions in a particularly simple form. A single update superposes many associations, so new writes interfere with those already stored. Even if that interference is reduced, same-position key–value writes contain no explicit relationship between different positions. And even if both problems are resolved, transporting every association equally gives the memory no mechanism for distinguishing structure that matters from content that merely occurred. These are three different limitations: *capacity*, *binding*, and *selectivity*.

This observation leads directly to **Double-Take (DT)**. Multi-step unpinned updates allow competing associations to be resolved rather than collapsed into a single write, increasing *capacity*. Local cross-position mixing exposes relationships before they enter memory, enabling *binding*. Finally, bidirectional change-sensitive writes expose differences between opposing contexts, providing a mechanism for *selective retention*. DT therefore follows from a simple view of long-range memory: additional context becomes useful only when a model can *retain enough information, relate information across positions, and preserve the relationships that matter*. Building on this mechanism we present following contributions:

- We identify interference-limited capacity and the inability of self-associative writes to form cross-position bindings as two limitations of large-chunk test-time memory.
- We repair capacity with multi-step unpinned updates, enable cross-position binding through local mixing, and introduce bidirectional memory for selective retention.
- We show that DT remains informative more than $30\times$ beyond its training context, directly recruits information outside the local receptive field through its test-time memory, and transfers zero-shot from human to mouse and to a broader set of species.

2 Background

2.1 Fast-Weight Memory and Test-Time Training

Efficient long-context models cannot keep every past token explicitly available, so they compress the growing sequence into a smaller state that is carried forward. This makes long sequences practical, but recall does not automatically grow with context. Associative recall remains a major source of error in efficient sequence models (Arora et al., 2023), state size trades directly against recall ability (Arora et al., 2024b), and fixed-size latent states face fundamental limitations on copying and retrieval (Jelassi et al., 2024). Related work has therefore explored compressed, external, and persistent memories as ways to preserve information beyond a standard recurrent state (Rae et al., 2019; Wu et al., 2022).

Fast-weight memory takes a different approach: rather than storing the past only in activations, it writes temporary information into parameters that change as the sequence is processed (Ba et al., 2016). Linear attention can itself be interpreted as such a fast-weight memory, where key-value associations are written into a finite state (Schlag et al., 2021). Later delta-rule methods showed that changing the write rule can improve how these associations are corrected and retrieved (Schlag et al., 2021; Yang et al., 2024; 2025). Test-time training (TTT) extends this idea by making the hidden state a small learnable model whose parameters are updated directly from the incoming sequence (Sun et al., 2024). Given keys K and values V, a memory write takes the form

$$W \leftarrow W - \eta \nabla_W \mathcal{L}(f_W(K), V)\,, \tag{1}$$

and a later query Q reads from the updated memory through $f_W(Q)$.

Large-Chunk Test-Time Training (LaCT) makes this process practical at long sequence lengths by updating the fast-weight memory over large blocks of tokens (Zhang et al., 2025). This greatly improves efficiency, but it also makes each update responsible for many associations at once. That immediately raises three questions. Can many associations be written without interfering with one another? Can information from different positions be connected inside the memory? And, as more context arrives, which of those relationships should remain important? We refer to these as *capacity*, *binding*, and *selectivity*. Prior fast-weight and gated-memory work already shows that both the form of the write and the control of memory updates matter for long-range retrieval (Schlag et al., 2021; Yang et al., 2025; Behrouz et al., 2026); our focus is on understanding how these three requirements arise specifically when large chunks are written into test-time memory.

2.2 Long-Context Genomic Modeling

Genomics makes these requirements especially concrete. Regulatory information can depend on individual nucleotides while interactions span tens to hundreds of kilobases through chromatin contacts and domain organization (Dixon et al., 2012; Rao et al., 2014). Akita showed that DNA sequence alone can predict three-dimensional genome folding (Fudenberg et al., 2020), while Enformer demonstrated that integrating long-range sequence improves regulatory prediction (Avsec et al., 2021). More recent models extend this scale further: Borzoi models regulatory activity over large genomic windows (Linder et al., 2025), while HyenaDNA and Caduceus bring long-context modeling to nucleotide resolution (Nguyen et al., 2023; Schiff et al., 2024). These models make distant sequence increasingly accessible; the remaining question is whether a memory can preserve enough of that information, connect the relevant positions, and retain the structure that matters. This is the question we take up next. We examine the large-chunk fast-weight write itself and ask what prevents *capacity*, *binding*, and *selectivity* from being achieved.

3 Double-Take

Double-Take (DT) is a fast-weight memory that processes the sequence in chunks while carrying a persistent memory state across them. DT departs from LaCT in three ways: multiple partial inner steps with free retention instead of per-chunk weight renormalization, causal cross-position mixing on the Q/K/V streams, and a bidirectional change-point write target. For each chunk m, let h_i denote the hidden representation at position i, for $i = 1, \ldots, N$. From each h_i, the model forms three representations: a query used to read from memory, a key used to address a memory write, and a value containing the information to be written. Specifically,

$$q_i = P_Q h_i, \qquad k_i = P_K h_i, \qquad v_i = P_V h_i,$$

where P_Q, P_K, and P_V are learned projection matrices.

The memory itself is a small neural network with fast-changing parameters $W = \{W_0, W_1, W_2\}$:

$$f_W(x) = W_1(\mathrm{SiLU}(W_0 x) \odot W_2 x)\,.$$

The projection matrices are learned during pretraining and remain fixed at test time, while the fast-weight state W is updated as each chunk is processed. The memory is multi-headed, with an independent fast-weight state for each head and $d_k = d_v = d_h = 32$.

DT changes how information is written into this memory in three steps. Multiple partial updates allow more associations to be stored without overwriting the state already carried forward. Local mixing exposes relationships between nearby positions before they are written. Finally, bidirectional views expose changes between opposing contexts so that such structure can be retained preferentially. These three mechanisms address *capacity*, *binding*, and *selectivity*, respectively. Figure 1 summarizes the architecture.

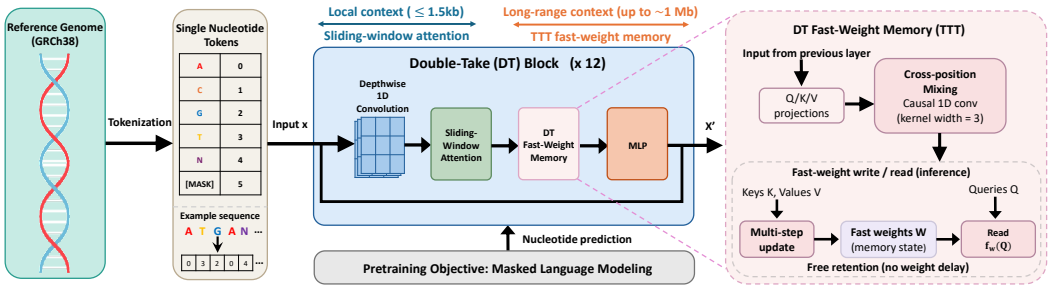

Figure 1: **Double-Take architecture.** Each block combines depthwise local convolution, sliding-window attention, and a persistent fast-weight memory. Local attention handles nearby interactions, while DT carries information across chunks through a fixed-size memory updated during the forward pass.

3.1 Partial Updates Increase Capacity

A large chunk writes many key-value pairs into the same fast-weight state. With a single update, these associations are written simultaneously. In the linear case $f_M(x) = Mx$, $\Delta M \propto \sum_i v_i k_i^\top$, so querying k_j retrieves its desired value together with contributions from other keys that overlap with k_j. As the number of pairs grows, this cross-talk increases and the one-step write becomes capacity-limited. DT therefore does not stop after the first write. Let M_{m-1} denote the memory entering chunk m. Starting from $M_m^{(0)} = M_{m-1}$, DT performs a small number of updates,

$$M_m^{(s+1)} = M_m^{(s)} - \eta_s \, \mathrm{NS}\Big(\nabla_M \mathcal{L}_{\mathrm{mem}}(M_m^{(s)}; K, V)\Big), \qquad s = 0, \dots, S-1, \tag{2}$$

with $\mathcal{L}_{\mathrm{mem}} = -\sum_i \langle f_M(k_i), v_i \rangle$. Here, η_s controls the step size and $\mathrm{NS}(\cdot)$ denotes Muon's Newton-Schulz normalization. Orthogonalizing the update reduces the cross-talk of the Hebbian write, and in the nonlinear memory successive updates are state dependent(Appendix A)

The key is that this refinement remains *partial*. The incoming state already contains information accumulated from earlier chunks, so Under the stock rule, each update is followed by row-wise renormalization, and repeated updates progressively rotate the state toward the current write, erasing what earlier chunks stored. DT therefore limits the number of inner updates and carries the resulting state $M_m = M_m^{(S)}$ directly to the next chunk. The carried state is diluted, and not replaced. Multi-step partial updates thus increase *capacity* while preserving one evolving memory across chunks. Yet capacity alone cannot change the structure of the write: k_i is still associated only with v_i from the same position. A relationship between different positions must therefore enter the memory in a different way.

3.2 Local Mixing Creates Cross-Position Bindings

Increasing capacity lets more associations survive, but it does not change which associations the memory can represent. In the standard write, k_i is paired only with v_i, so the memory contains terms of the form $v_i k_i^\top$. A relationship between two different positions is therefore absent from the write itself. No number of additional inner updates can recover a term that was never presented to the memory.

DT introduces that relationship before writing. We apply a learned depthwise causal convolution of width $c = 3$ to the query, key, and value streams (Poli et al., 2023). For the key stream,

$$\widetilde{k}_i = \sum_{r=0}^{c-1} \alpha_r^{(k)} k_{i-r}, \qquad \sum_i v_i \widetilde{k}_i^\top = \sum_i \sum_{r=0}^{c-1} \alpha_r^{(k)} v_i k_{i-r}^\top. \tag{3}$$

The terms with $r > 0$ now explicitly pair information from different positions. The same mixing is applied to the query and value streams. Thus, rather than asking the memory to infer a missing relationship after the write, DT makes that relationship part of what is written.

Importantly, the width-3 convolution is not the source of long-range reach. It only determines which nearby positions become bound together. Once that relationship enters the persistent fast-weight state, it is carried across subsequent chunks and remains available to distant queries. Local mixing therefore provides the *binding*; the persistent memory provides the *transport*. The memory can now store more associations and carry relationships across distance. But making a relationship available does not make it important: the memory still treats ordinary content and meaningful structural change alike.

3.3 Change-Point Writes Add Selectivity

Binding allows relationships between positions to enter memory, but it does not determine which of them should be emphasized. As more context is written, DT therefore looks for a simple signal of change: whether the context on one side of a position differs from the context on the other. Similar surroundings should contribute little, while a strong difference should remain visible.

To expose this difference, DT views each position from both directions. The forward stream summarizes the $w = 512$ positions preceding i, giving μ_i^L, while the same causal operation on the reversed sequence gives the corresponding right-context summary μ_i^R. After realigning the reverse stream, DT writes their difference,

$$y_i = \mu_i^L - \mu_i^R.$$

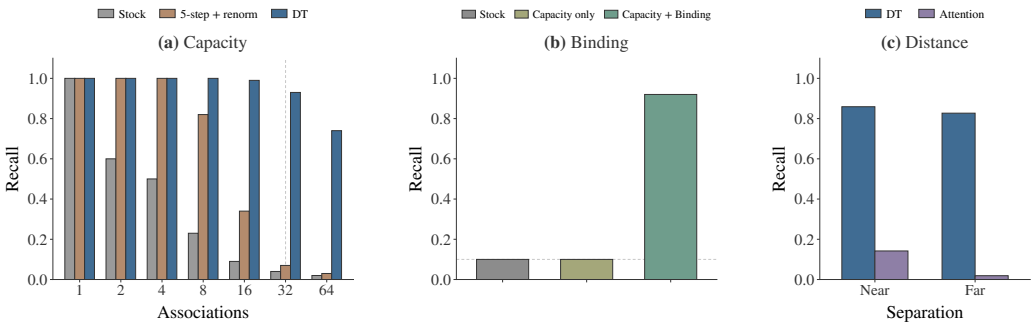

Figure 2: **Capacity and binding are separate requirements for usable memory. a**, Increasing associative load exposes the capacity limit of the stock and renormalized writes, while DT remains effective up to the memory-head dimension. **b**, Greater capacity alone does not enable cross-position retrieval; adding binding raises recall from chance to 0.92. **c**, Once formed, the relationship remains available over distance through the persistent memory, while the local attention path degrades sharply.

This directly produces the desired behavior: when the two sides are similar, their summaries largely cancel; when they differ, the contrast remains. A transition from representation a to b, for example, produces approximately $a - b$, whereas a locally similar region produces a value near zero. The write therefore emphasizes change rather than treating every position equally.

The same distinction is preserved when the memory is read. If o_i^L and o_i^R are the aligned forward and reverse reads, DT retains both their shared component, $(o_i^L + o_i^R)/2$, and their directional difference, $o_i^L - o_i^R$. The forward and reversed sequences are processed together by stacking them along the batch dimension, so the two views do not require a second sequential traversal.

The resulting memory remains causal within each directional pass. For chunk m, DT first reads the state accumulated from earlier chunks, $O_m = f_{M_{m-1}}(\widetilde{Q}_m)$, and only then writes the current chunk to produce M_m. Thus, the current write becomes available only to subsequent chunks. During masked-language pretraining, masking is applied before the query, key, and value streams are constructed, preventing a masked position from writing its own ground-truth nucleotide into memory.

4 Experiments

We evaluate DT at two levels. First, controlled associative-recall experiments isolate the fast-weight memory itself, allowing the capacity and binding mechanisms to be tested independently of genomic pretraining. We then evaluate the complete model on long-context genomic tasks, where the encoder is frozen and only lightweight downstream probes are trained. This second stage tests whether the repaired memory remains useful as context grows beyond the training length, whether its write becomes selective, and whether the learned representation transfers outside the human training distribution.

Our encoder contains 87M parameters in twelve identical blocks, each combining a depthwise convolution, sliding-window attention, DT fast-weight memory, and a position-wise MLP. The embedding width is 768, the memory-head dimension is 32, and memory updates operate on chunks of 2048 tokens. We pretrain at single-nucleotide resolution on GRCh38 using a bidirectional masked-language-modeling objective with $32\,\mathrm{kb}$ sequences; chromosomes 8 and 9 are excluded from pretraining and reserved for downstream human evaluation.

The sliding-window attention path covers 128 positions on either side per layer, giving an attention-only receptive field of approximately $\pm 1.5\,\mathrm{kb}$ across the twelve-block stack. Information used from substantially beyond this range must therefore pass through the fast-weight memory path. Unless stated otherwise, matched comparisons use the same architecture, examples, splits, and downstream protocol; public long-context models are treated as external reference points rather than causal controls. Full dataset construction, memory variants, readout definitions, and hyperparameters are provided in Appendix C.

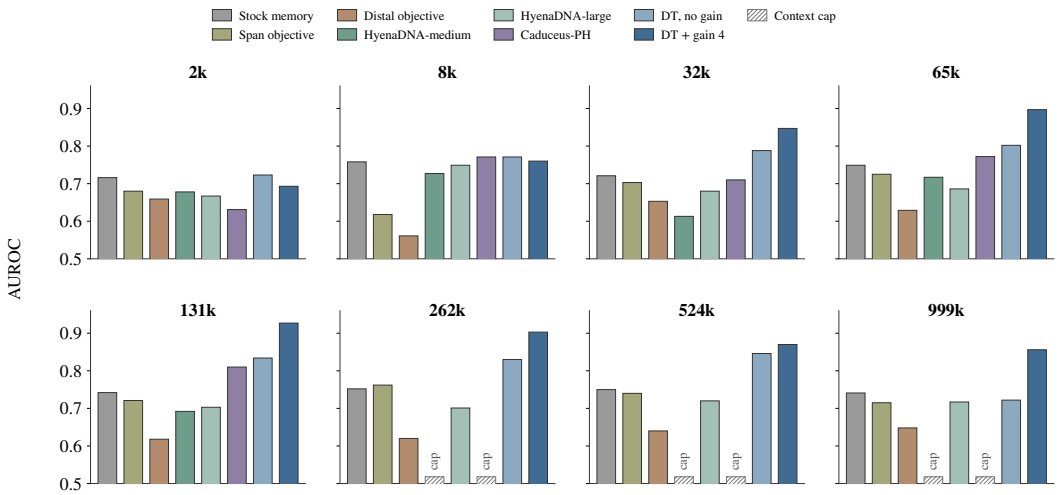

Figure 3: **TAD-boundary context scaling.** AUROC of frozen encoder features under a shared logistic-probe protocol from $2\,\mathrm{kb}$ to approximately $1\,\mathrm{Mb}$. Matched controls use the same encoder architecture, while HyenaDNA and Caduceus serve as external long-context references. "cap" denotes contexts beyond a model's native limit. Values are means over three seeds ($0.014 < \mathrm{std} < 0.043$).

4.1 Capacity and Binding Enable Long-Range Recall

We first isolate DT's memory mechanism using controlled associative recall. We begin with *capacity*: how many key-value associations can a $d = 32$ memory head retain as the number of stored pairs increases? **Stock** is the original single-step large-chunk memory and provides the unrepaired baseline. **5-step + renorm** takes multiple updates while keeping the stock retention rule, separating the benefit of additional optimization from DT's update itself. **DT** uses the partial multi-step update of Sec. 3.1. Capacity, however, only tells us whether more associations survive; it does not tell us whether information at different positions can be related. We therefore place the associated key and value at different positions and progressively compare **Stock**, **Capacity only**, and **Capacity + Binding**, where the final variant adds the cross-position mixing of Sec. 3.2. Finally, successful binding could still be only local. We therefore increase the separation between the associated positions and compare DT with sliding-window attention, testing whether the relationship created locally remains available through the persistent memory.

The results follow the same progression. In Fig. 2a, Stock degrades rapidly as associative load increases, and 5-step + renorm also collapses at higher loads, whereas DT remains near perfect through 16 associations and retains 0.93 recall at 32, matching the head dimension. Greater capacity alone, however, does not solve cross-position retrieval. In Fig. 2b, both Stock and Capacity only remain at the 0.10 chance level, while adding binding raises recall to 0.92. Once that relationship is formed, it also remains available over distance. In Fig. 2c, DT changes only from 0.859 to 0.827 from near to far separation, whereas Attention falls from 0.142 to 0.019. Together, these results show that usable long-range memory requires both capacity and binding: the partial multi-step update determines how much information can survive, cross-position mixing determines whether the required relationship can enter memory, and the persistent state carries that relationship beyond the local receptive field.

4.2 DT Uses Context Beyond Its Training Length

The controlled experiments show that repairing capacity and binding makes information usable inside the fast-weight memory. We next test whether this advantage survives in a real genomic setting and, more importantly, whether the model can continue to benefit from context beyond the $32\,\mathrm{kb}$ length seen during pretraining. We use TAD-boundary discrimination and increase the available sequence from $2\,\mathrm{kb}$ to approximately $1\,\mathrm{Mb}$, while freezing every encoder and training the same logistic probe at each context length.

We organize the eight approaches as a progressively richer sequence of ways to use long context. **Stock memory** provides the closest reference point, using the same 87M-parameter encoder and downstream protocol with the original large-chunk memory. Keeping this architecture fixed, **span** masking removes contiguous regions during pretraining, while the **distal** objective removes local evidence around prediction targets, placing progressively greater pressure on the model to use information beyond its immediate neighborhood. We then move to architectures designed natively for longer genomic sequences. **HyenaDNA-medium** replaces attention with long convolutional sequence mixing and supports $160\,\mathrm{kb}$ context; **HyenaDNA-large** extends the same architectural principle to a native $1\,\mathrm{Mb}$ context. **Caduceus-PH** provides a complementary bidirectional state-space architecture, allowing information from both sequence directions within a $131\,\mathrm{kb}$ context. Finally, **DT, no gain** keeps our encoder fixed but replaces the stock memory with the repaired persistent fast-weight memory, and **DT + gain 4** adds the fixed $4\times$ memory-branch gain used in our strongest inference setting. Contexts beyond a model's native maximum length are marked as capped in Fig. 3.

The results follow the same progression (Fig. 3). With the original **Stock memory**, performance changes little as additional sequence is exposed, remaining around 0.72–0.75 across most context lengths. Changing only the training signal does not substantially alter this behavior: the **distal** and **span** objectives remain comparatively flat as context grows. Moving to architectures designed natively for longer sequences provides stronger long-context references. **HyenaDNA-medium** reaches 0.692 at its $131\,\mathrm{kb}$ supported range, **HyenaDNA-large** extends evaluation to $1\,\mathrm{Mb}$ but remains between 0.680 and 0.720 from 32 to $524\,\mathrm{kb}$, and **Caduceus-PH** rises to 0.810 at its $131\,\mathrm{kb}$ native limit. Replacing the stock memory with DT changes the scaling pattern: **DT, no gain** improves from 0.788 at the $32\,\mathrm{kb}$ training length to 0.834 at $131\,\mathrm{kb}$ and 0.846 at $524\,\mathrm{kb}$. The full **DT + gain 4** configuration strengthens the same trend, increasing from 0.847 at $32\,\mathrm{kb}$ to 0.897 at $65\,\mathrm{kb}$ and 0.927 at $131\,\mathrm{kb}$, while remaining high at $262\,\mathrm{kb}$ (0.903), $524\,\mathrm{kb}$ (0.870), and $999\,\mathrm{kb}$ (0.856). Thus, DT continues to benefit from sequence far beyond the context length seen during training. These results show that distant context becomes useful, but do not identify the path through which that information reaches the representation. We therefore test whether the additional signal is actually recruited through the fast-weight memory.

4.3 DT Retrieves Information Beyond the Local Receptive Field

The context-scaling experiment shows that additional sequence becomes useful, but does not by itself establish how that information reaches the representation. We therefore use distal masking to isolate the memory path directly. Every prediction target is separated from the nearest visible token by at least $3{,}072$ bp, well beyond the approximately $\pm 1.5\,\mathrm{kb}$ receptive field of the local convolution and sliding-window attention path. Predictive information from visible sequence beyond this range must therefore be mediated by the persistent fast-weight state.

With the adapting bidirectional memory enabled, DT reaches a distal validation loss of 1.292. Freezing the same fast-weight state increases the loss to 15.30, showing that the test-time write carries predictive information unavailable to the local path. Replacing the two directed memory passes with a single global-state control reaches 1.326; the bidirectional formulation performs better in all six paired validation evaluations. Because the evaluation pipeline is optimized around the adapting state, we treat the frozen-state comparison as a mechanistic control rather than interpreting the absolute magnitude of the loss gap.

4.4 DT Selectively Preserves Genomic Change

Having established that distant information is both useful and directly recruited through the fast-weight state, we next ask what information that memory preferentially preserves. We begin with *eQTL variant-effect prediction*, where the representation must capture how a sequence change alters regulatory state. We organize the comparison from task-specific prediction to increasingly general long-context encoders. **Task-specific expert** and **CNN** models provide supervised references specialized for the prediction task. **HyenaDNA** introduces long-range convolutional sequence mixing, while **Caduceus-PH** and **Caduceus-PS** add bidirectional state-space processing over long genomic context. We then evaluate **DT** with the encoder frozen and train only a lightweight logistic adapter, asking whether the pretrained memory already exposes the information required for variant-effect prediction. Because the published baselines use their own fine-tuning protocols, we additionally use a protocol-matched frozen **HyenaDNA** comparison on whole blood. Finally, to separate sensitivity to sequence change from generic long-range decodability, we pair eQTL prediction with **enhancer-**

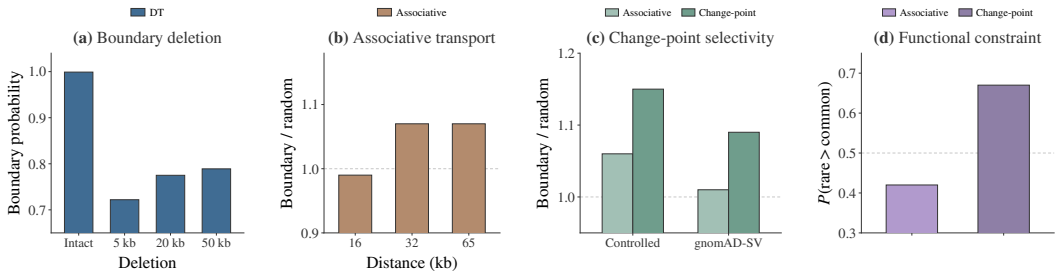

Figure 4: **Selective transport of structural contrast. a**, Boundary deletion changes the readout. **b**, Associative transport is not selective. **c**, The change-point write increases selectivity on controlled and real gnomAD variants. **d**, It also preferentially propagates rare constrained deletions.

Table 1: **eQTL variant-effect prediction across nine GTEx tissues.** AUROC across tissue-specific tasks. DT uses a frozen encoder with a logistic adapter; published baselines use their reported fine-tuning protocols. Bold denotes the best result in each column.

Model	CCF	WB	Thy	SNSE	SSEL	Musc	Nerv	Art	Adip	**Avg**
Task-specific expert	.639	.689	.612	.710	.700	.621	.683	**.741**	**.736**	.681
CNN	.547	.577	.487	.499	.499	.502	.516	.576	.551	.528
HyenaDNA (fine-tuned)	.584	.512	.529	.471	.544	.487	.511	.479	.513	.514
Caduceus-Ph (fine-tuned)	.597	.594	.527	.586	.574	.538	.588	.547	.541	.565
Caduceus-PS (fine-tuned)	.549	.542	.547	.529	.541	.523	.552	.536	.519	.537
DT (ours, frozen)	**.745**	**.747**	**.638**	**.780**	**.829**	**.729**	**.763**	.682	.640	**.728**

target-gene assignment, where the representation must retain the identity of a particular distant element rather than the effect of a local change.

We then move from downstream decoding to increasingly direct tests of what the memory carries. At $262\,\mathrm{kb}$ context, we first delete 5, 20, and $50\,\mathrm{kb}$ regions centered on annotated TAD boundaries and measure the resulting change in the boundary readout (Fig. 4a). We next hold the distal readout fixed and compare boundary-centered perturbations with matched random perturbations over 16-$65\,\mathrm{kb}$. Using the standard **associative write** first establishes how the repaired capacity-and-binding memory transports these distant perturbations (Fig. 4b). We then replace only the write target with the **bidirectional change-point write** and repeat the comparison, first on controlled boundaries and then on real gnomAD structural variants (Fig. 4c). Finally, we compare rare constrained deletions with common tolerated variants, extending the same comparison from structural contrast to functional constraint (Fig. 4d). The nine GTEx tissues are cultured fibroblasts (CCF), whole blood (WB), thyroid (Thy), non-sun-exposed skin (SNSE), sun-exposed lower-leg skin (SSEL), skeletal muscle (Musc), tibial nerve (Nerv), tibial artery (Art), and subcutaneous adipose tissue (Adip).

The results follow this progression. Frozen DT reaches a mean AUROC of 0.728 across nine GTEx tissues and 0.747 in the matched whole-blood comparison versus 0.689 for frozen HyenaDNA; on enhancer–target-gene assignment, the ordering reverses (0.628 versus 0.724) (Table 1). In Fig. 4a, deleting the centered $5\,\mathrm{kb}$ boundary region produces the largest drop in boundary probability (0.277), compared with 0.224 and 0.210 for the 20 and $50\,\mathrm{kb}$ deletions. In Fig. 4b, the associative write remains close to no preference across distance ($0.99, 1.07, 1.07$). Adding the change-point write increases the boundary-to-random ratio from 1.06 to 1.15 on controlled boundaries and from 1.01 to 1.09 on gnomAD-SV (Fig. 4c). Finally, Fig. 4d shows the strongest separation: $P(\mathrm{rare} > \mathrm{common})$ increases from 0.42 with the associative write to 0.67 with the change-point write.

4.5 DT Transfers Beyond Human Genomes

Having established capacity, binding, and selective retention on human genomic sequence, we finally test whether the resulting long-range representation survives distribution shift. We begin with a controlled human-to-mouse transfer setting in which the task and downstream protocol remain fixed while the underlying genome changes. All encoders are pretrained on human sequence and then frozen. Human evaluation uses held-out chromosomes 8 and 9, while mouse evaluation uses

Table 2: **Zero-shot cross-species transfer.** TAD-boundary AUROC from frozen human-pretrained encoders under a matched logistic-probe protocol. Human results use held-out chromosomes 8/9; mouse results use mm10 chromosomes 1-5 without adaptation. "cap" denotes context beyond a model's native limit.

	Encoder	2k	8k	32k	131k	262k	524k	1M
Human	DT, memory on	**.693**	**.760**	**.847**	**.927**	**.903**	**.870**	**.856**
	Same encoder, memory off	.640	.641	.617	.623	.658	.625	.610
	Caduceus-PS	.668	.739	.782	.827	cap	cap	cap
	HyenaDNA-1m	.633	.701	.640	.660	.776	.712	.784
Mouse	DT, memory on	.677	.699	**.786**	**.785**	**.761**	**.783**	**.796**
	Same encoder, memory off	.653	.583	.602	.639	.630	.648	.713
	Caduceus-PS	**.710**	.697	.752	.773	cap	cap	cap
	HyenaDNA-1m	.615	**.711**	.693	.733	.713	.698	.710

Table 3: **Cross-species representation alignment.** Normalized centroid distance from each species mean window embedding to the human embedding centroid, before and after test-time adaptation. Lower is closer; parentheses denote evolutionary distance from human in millions of years(M).

	Human (0)	**Chimp** $(6.4M)$	**Macaque** $(29M)$	**Mouse** $(87M)$	**Chicken** $(319M)$
Frozen	.000	.090	.133	.286	.112
Adapted	.000	.097	.136	.271	.120

mm10 chromosomes 1–5 with no encoder adaptation. We sweep context from $2\,\mathrm{kb}$ to $1\,\mathrm{Mb}$ using the same logistic-probe protocol in both species. **Memory off** uses the same DT encoder with its fast-weight path disabled, providing the closest reference for how much of the transfer comes from long-range memory. **Caduceus-PS** adds bidirectional state-space processing with a native $131\,\mathrm{kb}$ context, while **HyenaDNA-1m** provides a native megabase convolutional long-context model. **DT, memory on** retains the full persistent fast-weight memory, allowing us to compare these increasingly long-range representations under the same human-to-mouse shift (Table 2).

We then increase the difficulty of the transfer itself. Rather than evaluating only one target species, we move progressively farther from the human training distribution through chimpanzee, macaque, mouse, and finally chicken (Tab. 3). At each shift, we compare the human-pretrained representation with its test-time-adapted memory state. This second setting therefore asks not only whether the representation transfers, but how its behavior changes as evolutionary distance from the training genome increases.

The zero-shot results preserve DT's long-context behavior under the human-to-mouse shift. On mouse, DT reaches 0.786 AUROC at $32\,\mathrm{kb}$ and remains 0.785, 0.761, 0.783, and 0.796 from $131\,\mathrm{kb}$ to $1\,\mathrm{Mb}$. Over the same range, the memory-off encoder remains lower, while Caduceus-PS reaches 0.773 at its $131\,\mathrm{kb}$ native limit and HyenaDNA-1m reaches 0.733-0.710 from $131\,\mathrm{kb}$ to $1\,\mathrm{Mb}$ (Table 2). Thus, the long-context representation learned on human sequence remains usable after zero-shot transfer to mouse. Across the broader species sweep, frozen and adapted representations remain similarly aligned to the human embedding space, as measured by normalized centroid distance (Table 3).

5 Conclusion and Limitations

Processing a long sequence is not sufficient to remember it. We identify *capacity*, *binding*, and *selectivity* as three requirements for usable long-range test-time memory, and DT addresses them through multi-step updates, cross-position mixing, and bidirectional change-point writes. Controlled experiments isolate each mechanism, while genomic results show that the resulting memory remains useful far beyond its $32\,\mathrm{kb}$ training context and transfers across species at megabase scale.

Our study is limited to genomic sequence and does not incorporate transcriptomic, epigenomic, or proteomic supervision; comparison with some recent TTT methods is also constrained by the lack of reproducible implementations. Extending these memory principles across modalities, architectures, and domains is an important next step. More broadly, long-context modeling should not only ask how much context a model can process, but how much of that context it can retain, relate, and use.

AI USE

In this work, we used generative AI tools for language editing, improving clarity and organization, drafting and debugging code, and LaTeX formatting. We did not use generative AI to generate experimental measurements, fabricate data, determine reported results, or make final scientific conclusions.

ETHICS STATEMENT

Double-Take (DT) studies long-context test-time memory and introduces mechanisms for improving its capacity, cross-position binding, and selective retention. Our experiments focus on genomic sequence modeling using reference genomes and established genomic benchmarks and involve no direct interaction with human subjects. DT is a general memory mechanism and could in principle be applied to other biological or decision-making settings, where the provenance of the data, the interpretation of retained information, and domain-specific risks would require additional evaluation. The genomic experiments in this work should not be interpreted as establishing clinical or diagnostic validity.

REPRODUCIBILITY STATEMENT

We provide the complete DT formulation in the main paper, with supporting derivations, mechanistic analyses, and detailed experimental protocols in the appendices. We report the model architecture, fast-weight memory configuration, update rules, pretraining setup, context lengths, controlled associative-recall experiments, downstream readout protocols, comparison models, and evaluation settings required to reproduce the study. Matched DT and stock-memory comparisons use the same architecture, pretraining setup, downstream examples, and evaluation protocol unless stated otherwise. We will release the code and configuration files required to reproduce the reported experiments, metrics, and figures.

REFERENCES

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models. *arXiv preprint arXiv:2312.04927*, 2023.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Y Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models. In *International conference on learning representations*, volume 2024, pp. 15664–15730, 2024a.

Simran Arora, Sabri Eyuboglu, Michael Zhang, Aman Timalsina, Silas Alberti, Dylan Zinsley, James Zou, Atri Rudra, and Christopher Ré. Simple linear attention language models balance the recall-throughput tradeoff. *arXiv preprint arXiv:2402.18668*, 2024b.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18 (10):1196–1203, 2021. doi: 10.1038/s41592-021-01252-x.

Jimmy Ba, Geoffrey E Hinton, Volodymyr Mnih, Joel Z Leibo, and Catalin Ionescu. Using fast weights to attend to the recent past. *Advances in neural information processing systems*, 29, 2016.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. *Advances in Neural Information Processing Systems*, 38:113506–113543, 2026.

Jesse R. Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012. doi: 10.1038/nature11082.

Geoff Fudenberg, David R. Kelley, and Katherine S. Pollard. Predicting 3D genome folding from DNA sequence with Akita. *Nature Methods*, 17(11):1111–1117, 2020. doi: 10.1038/s41592-020-0958-x.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

Samy Jelassi, David Brandfonbrener, Sham M Kakade, and Eran Malach. Repeat after me: Transformers are better than state space models at copying. *arXiv preprint arXiv:2402.01032*, 2024.

Johannes Linder, Divyanshi Srivastava, Han Yuan, Vikram Agarwal, and David R Kelley. Predicting rna-seq coverage from dna sequence as a unifying model of gene regulation. *Nature Genetics*, 57 (4):949–961, 2025.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, Stefano Ermon, Christopher Ré, and Stephen Baccus. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. In *Advances in Neural Information Processing Systems*, volume 36, pp. 43177–43201, 2023.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning*, pp. 28043–28078. PMLR, 2023.

Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, and Timothy P Lillicrap. Compressive transformers for long-range sequence modelling. *arXiv preprint arXiv:1911.05507*, 2019.

Suhas S. P. Rao, Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, Ido Machol, Arina D. Omer, Eric S. Lander, and Erez Lieberman Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014. doi: 10.1016/j.cell.2014.11.021.

Yair Schiff, Chia Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-directional equivariant long-range DNA sequence modeling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43632–43648. PMLR, 2024.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *International conference on machine learning*, pp. 9355–9366. PMLR, 2021.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. Learning to (learn at test time): RNNs with expressive hidden states. *arXiv preprint arXiv:2407.04620*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yuhuai Wu, Markus N Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. *arXiv preprint arXiv:2203.08913*, 2022.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. *Advances in neural information processing systems*, 37:115491–115522, 2024.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule. In *International Conference on Learning Representations*, volume 2025, pp. 29687–29707, 2025.

Tianyuan Zhang, Sai Bi, Yicong Hong, Kai Zhang, Fujun Luan, Songlin Yang, Kalyan Sunkavalli, William T. Freeman, and Hao Tan. Test-time training done right. *arXiv preprint arXiv:2505.23884*, 2025.

Jian Zhou and Olga G. Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931–934, 2015. doi: 10.1038/nmeth.3547.

APPENDIX

A DEFECTS IN STOCK TTT

We observe two distinct defects in stock TTT, and show that our implementation of DT addresses these defects.

Defect 1: a single inner step gives Hebbian capacity. Write the fast weights as a linear map W over a chunk of N key–value pairs $K = [k_1, \dots, k_N]$, $V = [v_1, \dots, v_N]$, with unit-norm keys and the dot-product objective

$$\mathcal{L}(W) = -\sum_{i=1}^{N} v_i^\top W k_i, \qquad \nabla_W \mathcal{L} = -V K^\top. \tag{4}$$

A single step from the carried state W_0 gives $W_1 = W_0 + \eta V K^\top$, whose readout at a stored key is

$$W_1 k_j = W_0 k_j + \underbrace{\eta\, v_j}_{\text{signal}} + \underbrace{\eta \sum_{i \neq j} (k_i^\top k_j)\, v_i}_{\text{cross-talk}}. \tag{5}$$

For orthonormal values and keys uniform on the sphere, the expected cross-talk energy relative to the signal is $(N-1)/d$, growing linearly with load. equivalent to one step with rate $S\eta$.

$$P k_j = V (K^\top K)^{1/2} e_j, \tag{6}$$

replacing the Hebbian coefficients $K^\top K = I + E$ with $(I+E)^{1/2} \approx I + E/2$; for weakly correlated keys this reduces cross-talk energy about fourfold at nearly the same signal level.

Retention acts on the carried state. The stock rule renormalizes each row of W after every update; repeated steps in a fixed direction progressively rotate each row toward the new write, erasing W_0 in the limit. Free retention removes this renormalization, so W_0 is diluted rather than replaced. For the nonlinear SwiGLU memory the gradient depends on W, and the benefit of multiple steps there is empirical (Fig. 2a).

Defect 2: the write is self-associative. Capacity is necessary and not sufficient. Both k_i and v_i are projections of the same position i, and every write in Defect 1 takes the form $V\, C\, K^\top$, with $C = I$ for gradient descent and $C = (K^\top K)^{-1/2}$ for the orthogonalized update. ny cross-position coefficient is therefore fixed by the key Gram matrix rather than by which positions should be related: the readout at a stored key is $V\, g(K^\top K)\, e_i$ with $g(x) = x$ or $x^{1/2}$, and for orthonormal keys it returns exactly v_i. No inner rate or step count changes this, so a retrieval requiring cross-token binding is not expressible, whatever the capacity.

A causal convolution of width c applied to the memory's projections before the write removes the restriction. With $\tilde{k}_i = \sum_{s=0}^{c-1} \alpha_s k_{i-s}$, the outer product becomes

$$\sum_i v_i \tilde{k}_i^\top \;=\; \sum_i \sum_{s=0}^{c-1} \alpha_s\, v_i\, k_{i-s}^\top, \tag{7}$$

which contains cross-token terms $v_i k_{i-s}^\top$ for $s \geq 1$. The convolution supplies the local mixing that makes cross-token binding expressible; at $c = 1$ it is the identity and Eq. (7) reduces to the self-associative case.

With a multi-step unpinned inner update—five steps, free retention—recall is 1.00 at sixteen associations and 0.93 at thirty-two, against a head dimension of 32 (Fig. 2a): the memory saturates at approximately the head dimension rather than at two.

B Detailed Experimental Setup

This appendix provides the complete experimental protocol underlying the results in the main paper. We distinguish three components throughout: (i) the pretrained sequence encoder, (ii) the test-time memory configuration, and (iii) the downstream readout. Unless explicitly stated otherwise, downstream genomic evaluations freeze the pretrained encoder and train only a lightweight task-specific probe. Modifications to DT therefore change the memory used during the forward pass, not the pretrained model parameters.

B.1 Model Architecture

DT is evaluated within an 87 M-parameter nucleotide-level encoder composed of twelve pre-normalized blocks. Each block contains a depthwise 1D convolution, bidirectional sliding-window attention, a fast-weight memory module, and a position-wise MLP. The embedding width is 768, with 24 attention and memory heads of dimension 32.

Table 4: **Encoder architecture.** Architecture shared by the matched DT and stock-memory models.

Component	Configuration
Parameters	≈ 87 M
Blocks	12 pre-LN blocks
Embedding width	768
Attention heads	24×32
Memory heads	24×32
Attention window	± 128 nt per block
Attention-only reach	$\pm 1{,}536$ nt across 12 blocks
Memory chunk size	2,048 tokens
Position encoding	RoPE, base 10^4
MLP	$768 \to 3072 \to 768$
MLP activation	GELU
Vocabulary	A, C, G, T, N, `[MASK]`
Input resolution	single nucleotide
Pretraining context	32,768 tokens
Pretraining genome	GRCh38
Held-out chromosomes	chr8, chr9
Objective	bidirectional masked language modeling

The sliding-window attention path can communicate at most 128 positions per layer. Across twelve blocks, its maximum receptive field is therefore $\pm 1{,}536$ nucleotides. Information reaching a central readout from substantially beyond this range must use the fast-weight memory path.

Encoder and output head. Input nucleotides are embedded into 768 dimensions and processed by the twelve-block encoder. During pretraining, the final LayerNorm output is mapped through a linear $768 \to 6$ nucleotide-classification head. This head is used only for masked-language modeling. Downstream experiments extract the encoder representation *before* this vocabulary projection and train a separate task-specific probe. There is therefore no separate autoregressive decoder in the downstream experiments.

Fast-weight network. Each memory head uses a three-matrix SwiGLU fast-weight function,

$$f_M(x) = M_1[\mathrm{SiLU}(M_0 x) \odot M_2 x]\,,$$

with $M_0, M_1, M_2 \in \mathbb{R}^{32\times32}$. Queries and keys are ℓ_2-normalized before entering the memory. The memory state is updated during the forward pass and subsequently read by the query stream.

B.2 DT Memory Configurations

The experiments progressively modify the same underlying fast-weight memory. This design separates the three properties studied in the paper—capacity, cross-position binding, and selective retention—without changing the surrounding encoder.

Table 5: **Memory configurations.** Matched variants share the encoder and differ only in the marked operations. MS: multi-step; FR: free retention; LM: local mixing; CP: change-point write.

Configuration	MS	FR	LM	CP	Role
Stock memory					Matched control
Capacity repair	✓	✓			Capacity test
Capacity + binding	✓	✓	✓		Binding test
DT, no gain	✓	✓	✓		No readout gain
DT + gain 4	✓	✓	✓		$4\times$ memory gain at inference
Selective DT	✓	✓	✓	✓	Selectivity experiments

Partial update. DT starts each chunk from the memory carried by the preceding chunk and performs five Newton–Schulz-normalized inner updates. Unlike the stock renormalized configuration, the resulting state is not projected back to a prescribed norm after each chunk. The finite update budget limits how far a single chunk can move the persistent state while allowing the current associations to be fitted more accurately.

Cross-position binding. Before the memory write, a learned depthwise causal convolution of width 3 is applied to the query, key, and value streams. This operation is deliberately local: it does not itself provide long-range communication. Instead, it makes cross-position relationships available to the write; the persistent fast-weight state subsequently transports those relationships across chunks.

B.3 Pretraining Data and Objective

The model operates at single-nucleotide resolution with the vocabulary $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}, \mathrm{N}, [\mathrm{MASK}]\}$. Any unsupported nucleotide symbol is mapped to N. Pretraining uses GRCh38 sequences of length $32{,}768$ nucleotides, with chromosomes 8 and 9 excluded from training and reserved for downstream human evaluation.

The primary model is pretrained with bidirectional masked-language modeling. Two matched stock-memory controls additionally use long-range masking objectives:

Token MLM. Individual non-N positions are selected with probability 0.15 and use standard masked-token corruption.

Span MLM. Contiguous spans are masked until approximately 15% of the sequence is selected.

Distal MLM. Prediction targets are selected while masking their local neighborhood, forcing the prediction to depend on information outside the attention-only receptive field.

The span and distal models are used only as objective controls. They test whether adding explicitly long-range pretraining supervision is sufficient to recover the long-context behavior of DT without repairing the memory mechanism.

B.4 Comparison Models

We separate *matched controls*, which retain the 87M DT architecture, from *external baselines*, which provide long-context genomic reference points but differ in architecture, scale, and pretraining.

Table 6: **Comparison models.** Matched controls preserve the DT architecture; public models serve as external long-context references.

Model	Params.	Context	Protocol	Purpose
DT	87M	32 kb	frozen	Full repaired fast-weight memory.
Stock memory	87M	32 kb	frozen	Architecture-matched memory control.
Stock + distal	87M	32 kb	frozen	Tests distal supervision alone.
Stock + span	87M	32 kb	frozen	Tests span masking alone.
HyenaDNA-medium	$\sim$ 7M	160 kb	public	Long-context convolutional reference.
HyenaDNA-large	$\sim$ 7M	1 Mb	public	Megabase-context reference.
Caduceus-PH	$\sim$ 8M	131 kb	public	Bidirectional state-space reference.
Caduceus-PS	$\sim$ 8M	131 kb	public	Cross-species state-space reference.

Because the public models are substantially smaller than DT, we do not use their comparison to establish the causal effect of the proposed memory modification. That conclusion instead rests on the matched stock-memory ablation. External models establish the scale of performance achieved by public long-context genomic encoders.

B.5 Downstream Readouts

Unless otherwise stated, genomic evaluations freeze all pretrained encoder parameters. Task performance is measured using lightweight readouts trained on the resulting representations.

Table 7: **Evaluation interfaces.** The encoder remains frozen in the genomic downstream evaluations.

Experiment	Encoder	Readout	Metric
Associative capacity	isolated memory	direct retrieval	recall
Cross-position binding	isolated memory	direct retrieval	recall
TAD boundary	frozen	pooled logistic probe	AUROC
eQTL variant effect	frozen	logistic / linear adapter	AUROC
Enhancer–target gene	frozen	matched frozen adapter	AUROC
Structural perturbation	frozen	boundary readout	probability change
Selective transport	frozen	distal masked readout	boundary/random ratio
Human$\rightarrow$mouse TAD	frozen human encoder	logistic probe	AUROC

B.6 Controlled Memory Experiments

The controlled experiments isolate the fast-weight memory from genomic pretraining and test the two mechanism-level failures identified in the main paper.

Associative capacity. We construct multi-query associative-recall instances containing $N \in \{1, 2, 4, 8, 16, 32, 64\}$ key–value associations. The memory-head dimension is $d = 32$. Each method receives the same associations and queries; recall is measured by recovering the corresponding stored value. We compare the stock single-step write, a five-step write with state renormalization, and DT's partial multi-step update. This experiment tests storage capacity independently of cross-position binding.

Cross-position binding. We next construct an end-to-end associative task in which the required key and value occur at different positions. The task has ten candidate outputs, giving chance recall of 0.10. The stock and capacity-only configurations test whether additional storage alone is sufficient; adding width-3 mixing tests whether the missing cross-position relationship must first be made representable to the memory.

Distance. Finally, we increase the separation between the associated positions while keeping the local mixing operator unchanged. We compare DT with the sliding-window attention path. Because

the convolution itself is local, persistence at the far separation tests whether the resulting association is being transported by the fast-weight state rather than by the mixing operator.

B.7 TAD-Boundary Context Scaling

Annotations. Human TAD boundaries are derived from the GM12878 Arrowhead domain calls of Rao et al. (2014). Domain coordinates are lifted from hg19 to hg38; a domain is retained only when both endpoints map successfully to the same target chromosome and its lifted length remains within 20% of the original. Shared endpoints between adjacent domains are collapsed into a single boundary.

Evaluation chromosomes. Human evaluation uses chromosomes 8 and 9, which are held out from model pretraining. All encoders receive the same sequence windows and downstream labels for a given evaluation condition.

Window construction. For a context length L, positive examples are centered on annotated TAD boundaries. Negative controls are centered at positions at least 100 kb from the nearest annotated boundary. Windows are required to fit inside the source sequence and contain at most 20% N bases. Classes are balanced.

Feature extraction. The frozen encoder produces hidden states in $\mathbb{R}^{L\times 768}$. We extract the central 512 positions and compute both mean and max pooling over the sequence axis, yielding a $1{,}536$-dimensional representation. Public baselines use the same pooling operation on their final hidden states.

Logistic probe. Features are standardized using statistics from the training split. A single linear logit is optimized with binary cross-entropy. The probe is trained for 400 full-batch Adam steps with learning rate 0.1 and weight decay 10^{-3}. Evaluation uses AUROC.

Context sweep. We evaluate contexts from 2 kb to approximately 1 Mb. The pretrained DT parameters are fixed throughout: context length is changed only at inference. Consequently, all cells beyond 32 kb measure extrapolation beyond the context length used during pretraining.

Interpretation. This task measures discrimination between boundary-centered windows and windows sampled far from any boundary. It should therefore not be interpreted as calibrated genome-wide boundary calling. In particular, positions at intermediate distances from an annotated boundary are not represented by either training extreme. This limitation affects the absolute interpretation of the AUROC but not the comparison between encoders evaluated on identical examples.

B.8 DNALongBench Variant-Effect Evaluation

We use paired downstream tasks to test whether the frozen representation favors structural *contrast* over precise long-range *identity*. All encoders use the same examples and downstream adapter within the protocol-matched comparison.

eQTL variant effect. The task asks whether replacing an allele changes regulatory state. DT is evaluated as a frozen encoder using a lightweight logistic adapter. Whole blood also provides a protocol-matched comparison against frozen HyenaDNA. We additionally evaluate the same frozen DT representation across nine GTEx tissues: cultured fibroblasts, whole blood, thyroid, non-sun-exposed skin, sun-exposed lower-leg skin, skeletal muscle, tibial nerve, tibial artery, and subcutaneous adipose.

The task-specific expert, CNN, fine-tuned HyenaDNA, and fine-tuned Caduceus numbers in Table 1 are published DNALongBench references. They use their respective published training protocols and are therefore reported to establish scale rather than as protocol-matched encoder comparisons.

Enhancer–target-gene assignment. This task asks which distant enhancer is associated with a specific gene and therefore requires preserving long-range element identity. We use the same frozen-

encoder evaluation principle as above. The contrast between this task and eQTL effect prediction is used to test the contrast-versus-identity hypothesis in Sec. 4.4.

B.9 Structural Perturbation Experiments

We next test whether the long-range representation responds specifically to boundary-scale genomic structure rather than to arbitrary sequence removal.

Boundary deletion. For 100 annotated boundaries at $262\,\mathrm{kb}$ context, we compare the intact sequence with variants in which 5, 20, or $50\,\mathrm{kb}$ centered regions are deleted. The intact windows have predicted boundary probability 0.999 ± 0.001. We report the change relative to the corresponding intact sequence, yielding drops of 0.277, 0.224, and 0.210 for the three deletion sizes.

Fine-scale motif perturbations. As an additional diagnostic, we perturb individual CTCF motifs using orientation flips, sequence scrambles, and matched controls. These single-motif perturbations are substantially smaller than the boundary-element deletion and are reported in the appendix as a scale control rather than as a primary result.

B.10 Selective Long-Range Transport

Boundary recruitment does not by itself imply selective transport. We therefore measure whether perturbing a functional boundary affects a distant readout more strongly than perturbing a matched random region.

For each sequence, the distal readout location is held fixed while the perturbation is applied outside its local attention receptive field. We compute

$$S_{\mathrm{sel}} = \frac{\Delta_{\mathrm{boundary}}}{\Delta_{\mathrm{random}}},$$

where $S_{\mathrm{sel}} > 1$ indicates preferential transport of the boundary perturbation.

Associative write. With the standard associative target, the ratio remains close to one: 0.99, 1.07, and 1.07 at 16, 32, and $65\,\mathrm{kb}$, respectively. Thus the repaired associative memory can transport distant changes without preferentially preserving boundary structure.

Functional constraint. Finally, we compare constrained rare deletions with common tolerated variants. The probability that the rare variant produces the larger propagated effect is 0.67 for the change-point write and 0.42 for the associative write. This experiment tests whether the selectivity introduced by the change-point target tracks functional constraint rather than perturbation magnitude alone.

C Extended Experimental Details

This section provides the experiment-specific protocols underlying Sec. 4.2. Shared architecture, pretraining, memory configuration, and downstream-readout details are given above; here we focus on how each experiment is constructed, what is held fixed, and what each comparison measures. We follow the same progression as the main paper: controlled capacity and binding, long-context extrapolation, selective transport, and finally cross-species transfer.

C.1 Capacity and Binding Experiments

The controlled experiments isolate the fast-weight memory from genomic pretraining. Their purpose is to separate three questions that are otherwise coupled in a full sequence model: how many associations can be retained, whether information at different positions can be bound, and whether an association created locally remains accessible over distance.

Table 8: **Associative capacity.** Recall as the number of simultaneously stored key–value pairs increases. The memory-head dimension is $d = 32$.

Method	1	2	4	8	16	32	64
Stock	1.00	.60	.50	.23	.09	.04	.02
5-step + renorm	1.00	1.00	1.00	.82	.34	.07	.03
DT	1.00	1.00	1.00	1.00	.99	**.93**	**.74**

Table 9: **Cross-position binding.** End-to-end recall when the associated key and value occur at different positions. Chance performance is 0.10.

Configuration	Recall
Stock	.10
Capacity only	.10
Capacity + Binding	**.92**

Associative capacity. We construct multi-query associative-recall instances containing

$$N \in \{1, 2, 4, 8, 16, 32, 64\}$$

key–value associations. The memory-head dimension is fixed at $d = 32$. Every method receives the same associations and queries, and performance is measured by recall of the corresponding stored value.

We compare three increasingly permissive writes. **Stock** uses the original single-step large-chunk update. **5-step + renorm** performs five updates but retains the stock state-renormalization rule, separating the effect of repeated optimization from persistent retention. **DT** uses five partial updates with free retention, carrying the resulting fast-weight state forward without projecting it back to the stock prescribed norm.

The comparison separates repeated fitting from retention. Repeated updates substantially improve the low-load regime, but the renormalized state again degrades as associative load approaches the head dimension. DT remains near perfect through $N = 16$ and retains 0.93 recall at $N = 32$.

Cross-position binding. Capacity alone does not test whether the memory can represent an association between different sequence positions. We therefore construct an end-to-end retrieval task in which the required key and value occur at different positions. Each example contains ten possible outputs, giving chance recall of 0.10.

We compare **Stock**, **Capacity only**, and **Capacity + Binding**. The first two differ in storage capacity but retain the same-position write. The final configuration additionally applies the width-3 cross-position mixing described in Sec. 3.2.

The capacity repair alone therefore leaves the cross-position task at chance, whereas introducing the relationship before the write raises recall to 0.92.

Distance. The binding operator itself is deliberately local. We therefore increase the separation between the associated positions while leaving the width-3 mixing operator unchanged. DT is compared with the sliding-window attention path under the same retrieval task. Persistence at the farther separation therefore tests whether an association created locally is subsequently carried by the persistent memory rather than by the local mixing operation.

DT changes only modestly from 0.859 to 0.827, whereas local attention falls from 0.142 to 0.019. Thus the local operator creates the cross-position relation, while persistence makes it available beyond the local receptive field.

C.2 Long-Context Genomic Extrapolation

We next evaluate whether the repaired memory remains useful when sequence length grows beyond the $32\,\mathrm{kb}$ context used during pretraining. The experiment uses TAD-boundary discrimination be-

Table 10: **Association transport over distance.** Recall for locally created associations evaluated at near and far separations.

Method	Near	Far
DT	**.859**	**.827**
Sliding-window attention	.142	.019

Table 11: **Full TAD-boundary context sweep.** AUROC of frozen encoder features under the common logistic-probe protocol. "cap" denotes contexts beyond a model's native maximum length.

Encoder	2k	8k	32k	65k	131k	262k	524k	999k
Repaired memory								
DT + gain 4	.693	.760	.847	.897	**.927**	**.903**	**.870**	**.856**
DT, no gain	.723	.771	.788	.802	.834	.830	.846	.722
Matched architecture / objective controls								
Stock memory	.716	.758	.721	.749	.742	.752	.750	.741
Span objective	.680	.618	.703	.725	.721	.762	.740	.715
Distal objective	.659	.561	.653	.629	.618	.620	.640	.648
Public long-context encoders								
HyenaDNA-medium	.678	.727	.613	.717	.692	cap	cap	cap
HyenaDNA-large	.667	.749	.680	.686	.703	.701	.720	.717
Caduceus-PH	.631	.771	.710	.772	.810	cap	cap	cap

cause the same central prediction can be evaluated while progressively revealing more surrounding sequence.

Boundary annotations. Human TAD boundaries are derived from the GM12878 Arrowhead domain calls of Rao et al. (Rao et al., 2014). Domain coordinates are lifted from hg19 to hg38. A domain is retained only when both endpoints map successfully to the same target chromosome and the lifted domain length remains within 20% of its original length. Shared endpoints between adjacent domains are collapsed into a single boundary.

Evaluation chromosomes and windows. Human evaluation is restricted to chromosomes 8 and 9, which are excluded from pretraining. For context length L, positive examples are centered on annotated TAD boundaries. Negative examples are centered at positions at least $100\,\mathrm{kb}$ from the nearest annotated boundary. Windows must fit inside the reference sequence and contain at most 20% N bases. The two classes are balanced.

This construction intentionally measures discrimination between boundary-centered windows and clearly non-boundary controls. It should not be interpreted as calibrated genome-wide boundary calling, since positions at intermediate distances from an annotated boundary are not represented by either extreme.

Frozen representation and probe. All pretrained encoder parameters remain frozen. Given hidden states $H \in \mathbb{R}^{L\times 768}$, we retain the central 512 positions and compute mean and max pooling along the sequence dimension. Concatenating the two summaries gives a 1536-dimensional representation.

Features are standardized using statistics from the training split. A single linear logistic probe is optimized with binary cross-entropy for 400 full-batch Adam steps using learning rate 0.1 and weight decay 10^{-3}. Performance is reported using AUROC. Public long-context models use the same window construction, pooling rule, and downstream probe.

Context sweep. We evaluate sequence lengths from $2\,\mathrm{kb}$ to approximately $1\,\mathrm{Mb}$. DT parameters are not retrained as context grows. Consequently, every evaluation above $32\,\mathrm{kb}$ tests inference-time extrapolation beyond the sequence length observed during pretraining.

C.3 Selective Genomic Transport

The context-scaling experiment establishes that distant sequence becomes decodable. The following experiments ask what information is preferentially represented and transported. We progressively move from downstream variant-effect prediction to direct sequence perturbation, selective long-range transport, real structural variants, and functional constraint.

C.3.1 Variant-Effect and Long-Range Identity Tasks

eQTL variant effect. The eQTL task asks whether replacing one allele changes regulatory state. DT is evaluated as a frozen encoder with a lightweight logistic/linear adapter. Whole blood provides the protocol-matched comparison with frozen HyenaDNA. We additionally evaluate the same frozen DT representation across nine GTEx tissues: cultured fibroblasts (CCF), whole blood (WB), thyroid (Thy), non-sun-exposed skin (SNSE), sun-exposed lower-leg skin (SSEL), skeletal muscle (Musc), tibial nerve (Nerv), tibial artery (Art), and subcutaneous adipose tissue (Adip).

Task-specific expert, CNN, fine-tuned HyenaDNA, and fine-tuned Caduceus results are taken from the corresponding published benchmark protocols. Because those models use their own fine-tuning procedures, they provide external reference points rather than causal comparisons with DT. The whole-blood frozen-DT versus frozen-HyenaDNA result is the protocol-matched encoder comparison.

Enhancer–target-gene assignment. We pair the variant-effect task with enhancer–target-gene assignment. Here, the readout must preserve the identity of a particular distant enhancer associated with a target gene. We use the same frozen-representation evaluation principle and report AUROC. This complementary task separates change-sensitive representation from preservation of precise distant-element identity.

C.3.2 Boundary Deletion

We next replace observational downstream decoding with direct sequence perturbation. For 100 annotated TAD boundaries evaluated at $262\,\mathrm{kb}$ context, we compare the intact sequence with versions in which 5, 20, or $50\,\mathrm{kb}$ centered regions are deleted.

The intact windows have predicted boundary probability 0.999 ± 0.001. For each deletion, we report the change relative to the corresponding intact example.

The largest change occurs for the centered $5\,\mathrm{kb}$ deletion, while larger deletions have similar effects. This provides a direct perturbational test that the long-range readout responds to sequence at the boundary scale.

Fine-scale motif perturbations. As an additional scale control, we perturb individual CTCF motifs using orientation flips, sequence scrambles, and matched controls. These single-motif interventions produce substantially smaller effects than deletion of the broader boundary element and are therefore treated as a diagnostic rather than a primary result.

C.3.3 Selective Long-Range Transport

Responding to a boundary does not imply that the memory preferentially transports boundary information. We therefore fix a distal readout location and perturb sequence outside its local attention receptive field. For each example, a boundary-centered perturbation is paired with a matched random perturbation, and we compute

$$S_{\mathrm{sel}} = \frac{\Delta_{\mathrm{boundary}}}{\Delta_{\mathrm{random}}}, \tag{8}$$

where $S_{\mathrm{sel}} > 1$ indicates that the boundary perturbation produces the larger distal effect.

Associative transport over distance. We first use the standard associative write, retaining the capacity and binding repairs but not the change-point target.

Table 12: **Selectivity of associative transport.** Boundary-to-random effect ratio as perturbation distance increases.

Distance	16 kb	32 kb	65 kb
Boundary / random	.99	1.07	1.07

Table 13: **Selectivity introduced by the change-point write.** Boundary-to-random effect ratios on controlled boundaries and real structural variants.

Write	Controlled	gnomAD-SV
Associative	1.06	1.01
Change-point	**1.15**	**1.09**

The ratio remains close to one across the full range, establishing that the repaired associative memory can carry distant effects without strongly preferring the boundary perturbation.

Change-point write. We then keep the repaired memory fixed and replace the associative target with the bidirectional change-point target defined in Sec. 3.3. The comparison is first performed on controlled boundary examples and then repeated on real gnomAD structural variants.

The change-point write increases the selectivity ratio in both settings, providing the controlled-to-real progression used in Fig. 4c.

C.3.4 Functional Constraint

Finally, we ask whether the same preference extends beyond boundary identity to functional constraint. We compare constrained rare deletions with common tolerated variants and report the probability that the rare deletion produces the larger propagated effect,

$$P(\text{rare} > \text{common}).$$

The probability rises from 0.42 under the associative write to 0.67 under the change-point write.

C.4 Cross-Species Transfer

The final experiments evaluate whether a representation pretrained on human DNA remains useful as the evaluation distribution moves away from the human training genome. We use two settings: a controlled human-to-mouse transfer with the downstream task held fixed, followed by a broader species sweep with increasing evolutionary distance.

C.4.1 Human-to-Mouse Zero-Shot Transfer

All encoders are pretrained on human sequence and remain frozen during evaluation. Human TAD-boundary results use held-out chromosomes 8 and 9. Mouse evaluation uses mm10 chromosomes 1–5 with no encoder adaptation. The same TAD-boundary representation and logistic-probe protocol is applied to both species.

We compare the full DT encoder with the same encoder after disabling its fast-weight memory path. Caduceus-PS provides a bidirectional state-space reference with native context up to $131\,\mathrm{kb}$, while HyenaDNA-1m provides a native megabase-context reference.

The human-to-mouse experiment changes the genome while retaining the task, probe, and human-pretrained encoder, providing the most controlled cross-species test in the paper.

C.4.2 Increasing Evolutionary Shift

We next broaden the shift beyond a single target species. Starting from human, we evaluate chimpanzee, macaque, mouse, and chicken, ordered by increasing evolutionary distance from the human

Table 14: **Selective transport under functional constraint.** Probability that a rare constrained deletion produces a larger distal effect than a common tolerated variant.

Write	$\mathbf{P}(\text{rare} > \text{common})$
Associative	.42
Change-point	**.67**

Table 15: **Zero-shot human-to-mouse transfer.** TAD-boundary AUROC from human-pretrained frozen encoders under the matched logistic-probe protocol. "cap" denotes context beyond a model's native maximum length.

	Encoder	**2k**	**8k**	**32k**	**131k**	**262k**	**524k**	**1M**
Human	DT, memory on	.693	.760	.847	.927	.903	.870	.856
	Same encoder, memory off	.640	.641	.617	.623	.658	.625	.610
	Caduceus-PS	.668	.739	.782	.827	cap	cap	cap
	HyenaDNA-1m	.633	.701	.640	.660	.776	.712	.784
Mouse	DT, memory on	.677	.699	.786	.785	.761	.783	.796
	Same encoder, memory off	.653	.583	.602	.639	.630	.648	.713
	Caduceus-PS	.710	.697	.752	.773	cap	cap	cap
	HyenaDNA-1m	.615	.711	.693	.733	.713	.698	.710

training genome. For each species, we compare the frozen memory state with the corresponding test-time-adapted state.

Frozen and adapted values remain closely aligned across the mammalian species in this sweep, while chicken provides the first non-mammalian evaluation and departs from the mammalian pattern.