S^3-Delta: Spectral-Stratum SpatioTemporal Delta A Scale-Reliable Operator for Long-Horizon Fuzzy Spatial Memory

Dian Yu

Abstract

Coordinate-bearing streams obey two orders: observations arrive in time, whereas memories should be reactivated by physical location. We study fixed-capacity memory for agents that scan two- or three-dimensional spaces and must fuzzily recall previously observed values near the current position. Standard fast weights support one-matrix key–value recall, but uniform forgetting cannot identify reliable spatial scales, while the conventional Delta rule treats all Fourier components as one estimate.

We propose Spectral-Stratum SpatioTemporal Delta (S^3-Delta). Each sine–cosine pair of a fixed Fourier key defines a wavelength stratum of one shared fast-weight matrix. At each event, the aged historical state first produces a causal recall. After the current observation arrives, removing each stratum counterfactually measures whether it reduced or increased the recall error. Accumulated signal and noise evidence forms a reliability gate for later readout, one normalized Delta residual updates the matrix, and wavelength-conditioned continuous-time decay then advances the consolidated state to the next event.

Across variable-density 2-D scans of 16K–262K events, S^3-Delta outperforms tuned dense Delta in all 30 seed-level comparisons over static and slowly changing environments, reducing mean error by 4.73%–8.99%. It also improves by 4.46%–8.65% over an additionally trained causal Gated-Delta control. At 262K, ablations identify counterfactual reliability as the main source of improvement and wavelength-conditioned decay as a smaller, consistent contribution. With 32 Fourier pairs, the spatial kernel attains 93.4% radial monotonicity. The main state remains $O(d_k d_v)$ and is independent of sequence length. These results support S^3-Delta as an interpretable, fixed-capacity operator for long-horizon fuzzy spatial memory.

Keywords: fast weights; Delta rule; Fourier positional encoding; spatial associative memory; counterfactual reliability; continuous-time forgetting

Operational rationale. Activation at the current position is not sufficient evidence that a stratum is accurate. The operator therefore recalls once from the complete matrix and evaluates scale contributions only after the observation arrives. Strata that repeatedly reduce error retain larger readout weights, whereas interference-dominated strata are attenuated. One raw residual fits the complete matrix; wavelength determines the decay prior, and accumulated evidence forms the counterfactual reliability gate.

Terminology. The paper uses only three mechanism names: *wavelength-conditioned decay*, the proposed *counterfactual reliability gate*, and the *instantaneous-energy gate* used only as an ablation control. The last is not part of complete S^3-Delta.

1 Introduction

Mobile agents, continuous canvases, point-cloud streams, and variable-density image scans share a common structure: the input is a temporal sequence, but every observation also carries a physical coordinate. Two observations separated by tens of thousands of events can originate from the same location, whereas consecutive observations can be spatially far apart. A purely temporal representation therefore provides no direct mechanism for retrieving observations previously recorded at the current location or its neighborhood.

A direct solution is to maintain an explicit grid, a table of historical coordinates, or a nearest-neighbor index. Such systems are appropriate when an exact map is required, but their capacity or query work usually grows with the number of visited locations. This paper studies a different target: a fixed-capacity, lossy, continuous state of spatial experience. The operator neither stores every historical point nor performs localization, mapping, or reachability reasoning. The current coordinate is supplied by an external system. Its sole task is to compress position–observation associations into a finite state and return nearby history in one read at the current position. The desired response is kernel-like: strongest at the same address, weaker for nearby positions, and gradually attenuated farther away.

Fast Weight Programming (FWP) provides a compact form for this goal. The outer product of a positional key $\boldsymbol{k}_t$ and a value $\boldsymbol{v}_t$ can be written into a matrix state, and a later query uses $\boldsymbol{M}_t^\top \boldsymbol{k}_q$(Schmidhuber, 1992; Ba et al., 2016; Katharopoulos et al., 2020; Schlag et al., 2021). Additive writing, however, continually accumulates interference. Uniform exponential forgetting controls energy but cannot determine which part of a current recall is useful. The Delta rule replaces blind accumulation with a residual correction at the current key and substantially improves associative memory(Schlag et al., 2021; Yang et al., 2024b). Yet it still treats the total recall at a position as a single, indivisible answer.

The construction follows from the positional key itself. A fixed sine–cosine encoding is not merely a list of anonymous channels: every complete pair corresponds to a wavelength. Long wavelengths describe slowly varying regional relationships, while short wavelengths retain finer local distinctions. The total recall at the current position can therefore be decomposed into scale-specific components. Some scales remain stable and useful, whereas others become unreliable because of environmental change, finite-feature approximation, or cross-location interference. Equal summation preserves these unreliable components. Fully independent writing creates the opposite problem: every stratum fits the same target separately and cross-scale coordination is lost.

The counterfactual gate is motivated by one narrow question. Once the current observation has arrived, would the preceding recall have been better or worse if one wavelength stratum had been absent? Because the raw recall is exactly the sum of its stratum responses, this leave-one-stratum-out error is available algebraically from the same read, without a second memory query or an auxiliary gating network. The current observation thus supplies an online test of which scales deserve trust on later events, while the shared Delta write remains unchanged.

We introduce S^3-Delta to resolve this tension. Its core is not another black-box gating network, but an interpretable online loop:

1. **Stratify by wavelength.** Every sine–cosine row pair forms a minimal spectral view of the complete matrix.

2. **Read before writing.** Only the old state may produce the causal recall before the current truth is revealed.

3. **Evaluate counterfactually.** After the observation arrives, the operator measures whether removing each stratum increases or decreases the error and records the result as signal/noise evidence.

4. **Write one global Delta update.** One raw residual updates the complete matrix without assigning separate write strengths to strata.

5. **Forget in continuous time.** Actual elapsed time rather than event index controls decay, with a longer prior half-life for longer wavelengths.

The main contributions are:

- a new fixed-capacity operator for fuzzy spatial recall with known coordinates, integrating Fourier wavelength strata, a counterfactual reliability gate, and global Delta writing into one closed loop;

- an analytic counterfactual reliability gate that requires neither a separate learned network nor additional supervision, together with a proof that normalized global writing contracts the raw residual at the current key by exactly $1 - \eta$;

- explicit treatment of irregular elapsed time through stratum-wise half-lives that satisfy a continuous-time semigroup property;
- a focused evaluation covering 16K–262K scans, memory-age strata, structural ablations, spatial-kernel behavior, and irregular timing, while reporting the current efficiency limitations directly.

2 Related Work

2.1 Fast weights, linear attention, and the Delta rule

Fast weights store short-term associations in rapidly changing parameters(Schmidhuber, 1992; Ba et al., 2016). Linear attention can be written as a recurrence over a matrix-valued state, replacing a sequence-growing attention cache with a fixed recurrent state(Katharopoulos et al., 2020). Linear Transformers can consequently be interpreted as fast-weight programmers, and the Delta rule corrects the old prediction at the current key to reduce the capacity waste of additive outer-product updates(Schlag et al., 2021). Parallel DeltaNet further supplies a hardware-efficient algorithm that parallelizes the Delta recurrence along the sequence dimension(Yang et al., 2024b).

These methods mainly address how to write an effective key–value mapping. S^3-Delta instead uses the known multiscale structure of the key to gate sine–cosine wavelength strata by counterfactual reliability while preserving the shared Delta residual.

2.2 Online adaptive filtering and expert weighting

The Delta update is closely related to least-mean-square adaptive filtering: both correct a linear predictor with its current residual(Haykin, 2002). Kernel adaptive filters extend this principle to nonlinear feature spaces, and online kernel learning provides a general framework for sequential function estimation(Kivinen et al., 2004; Liu et al., 2010). Under a fixed random Fourier map, dense Delta can therefore be viewed as a compact online linear filter in an approximate kernel feature space. This connection motivates its use as the principal baseline rather than treating it only as a neural-memory heuristic.

Adaptive mixtures of experts and online prediction methods also assign weights from past losses(Jacobs et al., 1991; Cesa-Bianchi and Lugosi, 2006). In S^3-Delta, the components are fixed wavelength strata of one associative state, not separately trained experts. Counterfactual reliability evidence is the marginal error change under removal of a stratum, while storage remains coupled through one Delta residual. This distinction is useful in changing streams, where outdated spatial scales must be separated from a global prediction error(Gama et al., 2014).

2.3 Gated linear memory and selective state-space models

S4 uses a structured state matrix to model long sequences efficiently(Gu et al., 2022). Mamba makes state-space parameters input-dependent so that information can be selectively propagated or forgotten, and combines this selectivity with hardware-aware scanning(Gu and Dao, 2023). GLA introduces data-dependent gates over a matrix-valued state with an efficient training algorithm(Yang et al., 2024a). Gated DeltaNet combines rapid erasure with directed Delta writing(Yang et al., 2025); Kimi Linear uses more fine-grained gating and specialized low-rank transitions(Kimi Team, 2025); and Gated DeltaNet-2 further decouples erasure and writing(Hatamizadeh et al., 2026).

Most of these gates are learned from input features for general-purpose sequence modeling. S^3-Delta does not attempt to replace those backbones or claim stronger general language modeling. It exploits the fixed spectral structure of a known coordinate and derives its gate from *observed recall performance*: a scale receives greater future weight only after repeatedly reducing the true error. The gate is an online statistic with a task-specific meaning, not a learned prediction of how much to forget.

2.4 Fourier features and continuous spatial representations

Random Fourier features approximate shift-invariant kernels with finite-dimensional inner products(Rahimi and Recht, 2007). Multiband positional encodings are also widely used to represent continuous coordinates and high-frequency functions(Tancik et al., 2020). Spatial Semantic Pointers demonstrate how phase structure can represent and manipulate continuous space(Komer et al., 2019). These results explain why the inner product of sine–cosine keys can form a distance-sensitive response, but do not define key–value writing, forgetting, and interference correction for a long streaming memory.

We therefore interpret each Fourier pair as a wavelength stratum. Kernel addressing determines the spatial support of recall, the Delta residual determines the written content, and the counterfactual reliability gate determines each stratum's subsequent readout contribution.

2.5 Explicit spatial memory and the boundary of this work

Classical sparse distributed memory and later differentiable memory systems establish complementary approaches to content-addressed storage(Kanerva, 1988; Graves et al., 2014). Explicit neural maps typically divide space into slots and use pose or geometric transformations to read and write at structured addresses(Parisotto and Salakhutdinov, 2018; Henriques and Vedaldi, 2018). They are well suited to mapping, navigation, and interpretable local access. S^3-Delta exposes a narrower but different interface: coordinates are externally provided, state does not grow with the number of visited points, and the result is lossy neighborhood experience rather than a cell-by-cell recoverable map. It contains no obstacle topology; Euclidean proximity therefore does not imply reachability. It is not a replacement for a nearest-neighbor database or a storage system that requires lossless replay.

3 Problem Definition and Design Principles

Consider an event stream

$$\mathcal{D} = \{(\boldsymbol{p}_t, \boldsymbol{v}_t, \Delta\tau_t)\}_{t=1}^{T}, \tag{1}$$

where $\boldsymbol{p}_t \in \Omega \subset \mathbb{R}^{d_p}$ is a known two- or three-dimensional coordinate, $\boldsymbol{v}_t \in \mathbb{R}^{d_v}$ is the complete observation at that position, and $\Delta\tau_t > 0$ is the true time since the preceding event. Before reading $\boldsymbol{v}_t$, the operator should produce $\hat{\boldsymbol{v}}_t$ from $\boldsymbol{p}_t$ and the old state, approximating lossy history at the current position and within its local kernel neighborhood. The objective is not full historical reconstruction. We require:

- **Fixed capacity:** state size must not grow with T.
- **Spatial locality:** nearby positions should respond more strongly than distant ones.
- **Strict causality:** the current observation must not leak into its own readout.
- **Controlled forgetting:** static information can persist, while changed information can still be corrected by new evidence.
- **Time consistency:** two half-step decays should equal one full-step decay.
- **Few handcrafted rules:** fixed structure and online evidence should determine behavior wherever possible, rather than a growing set of scenario-specific parameters.

Here, *fuzzy recall* denotes an operational property rather than a probabilistic assumption. A finite Fourier basis introduces approximation error and sidelobes, and different positions and values interfere as more events are written. Our goal is to make this finite-capacity loss manageable while preserving a useful spatial bias over long streams.

4 Spectral-Stratum SpatioTemporal Delta

4.1 From coordinates to wavelength strata

Let $\boldsymbol{\omega}_\ell \in \mathbb{R}^{d_p}$ denote the ℓ-th spatial frequency. One complete sine–cosine pair is

$$\boldsymbol{k}_{\ell,t} = \frac{1}{\sqrt{L}} \begin{bmatrix} \cos(\boldsymbol{\omega}_\ell^\top \boldsymbol{p}_t) \\ \sin(\boldsymbol{\omega}_\ell^\top \boldsymbol{p}_t) \end{bmatrix} \in \mathbb{R}^2, \qquad \|\boldsymbol{k}_{\ell,t}\|^2 = \frac{1}{L}. \tag{2}$$

The full key is $\boldsymbol{k}_t = [\boldsymbol{k}_{1,t}; \ldots; \boldsymbol{k}_{L,t}] \in \mathbb{R}^{d_k}$ with $d_k = 2L$. The operator maintains one complete fast-weight matrix and two vectors of scale evidence:

$$\boldsymbol{M}_t \in \mathbb{R}^{d_k \times d_v}, \qquad \mathbf{S}_t, \mathbf{N}_t \in \mathbb{R}^L_{\geq 0}. \tag{3}$$

Equalized key energy is a structural requirement, not a cosmetic convention. The factor $1/\sqrt{L}$ gives every pair the same energy $1/L$ and fixes the complete key at $\boldsymbol{k}_t^\top \boldsymbol{k}_t = 1$ for every position. Differences in stratum response therefore reflect stored content and phase agreement rather than arbitrary feature amplitude. If a learned or approximate positional encoder does not preserve this identity, its pairs must be rescaled to equal energy and the complete key renormalized before use. Otherwise high-energy strata dominate raw recall and counterfactual evidence, and the effective Delta step also changes with key scale; the gate then measures feature energy rather than reliability, which can markedly degrade S^3-Delta. All experiments enforce this unit-total-energy normalization. $\boldsymbol{M}_t$ is always read and written as one matrix. Let $[\boldsymbol{M}_t]_{(\ell)} \in \mathbb{R}^{2 \times d_v}$ denote its ℓ-th sine–cosine row pair. This view is used only by the counterfactual reliability gate, not as a separately maintained Delta state. $S_{\ell,t}$ and $N_{\ell,t}$ accumulate evidence that the corresponding wavelength stratum helped or harmed recall. A bar denotes the already aged state available at the start of an event, whereas the unbarred state denotes the result after that event has been verified and consolidated.

4.2 Recall first: raw retrieval and reliable fusion

The complete memory first produces one global raw recall:

$$\boldsymbol{r}_t^{\mathrm{raw}} = \bar{\boldsymbol{M}}_t^\top \boldsymbol{k}_t. \tag{4}$$

To determine which scales are reliable, the same contraction retains its sine–cosine row-pair contributions:

$$\boldsymbol{r}_{\ell,t} = [\bar{\boldsymbol{M}}_t]_{(\ell)}^\top \boldsymbol{k}_{\ell,t}, \qquad \boldsymbol{R}_t = \begin{bmatrix} \boldsymbol{r}_{1,t}^\top \\ \vdots \\ \boldsymbol{r}_{L,t}^\top \end{bmatrix}, \qquad \boldsymbol{R}_t^\top \mathbf{1} = \boldsymbol{r}_t^{\mathrm{raw}}. \tag{5}$$

This is one batched row-pair contraction, not L independent memory reads. The counterfactual reliability gate is

$$g_{\ell,t} = \frac{\bar{S}_{\ell,t}}{\bar{S}_{\ell,t} + \bar{N}_{\ell,t}}. \tag{6}$$

When the denominator is zero, we set $g_{\ell,t} = 1$. A new stratum therefore starts neutral instead of being permanently silenced before it has evidence. The external output is

$$\hat{\boldsymbol{v}}_t = \boldsymbol{R}_t^\top \boldsymbol{g}_t, \qquad \boldsymbol{g}_t = [g_{1,t}, \ldots, g_{L,t}]^\top. \tag{7}$$

The gate affects readout but does not directly erase the main memory. A stratum suppressed during a noisy period retains a decaying state and can recover if later evidence changes, instead of being irreversibly deleted by one incorrect judgment.

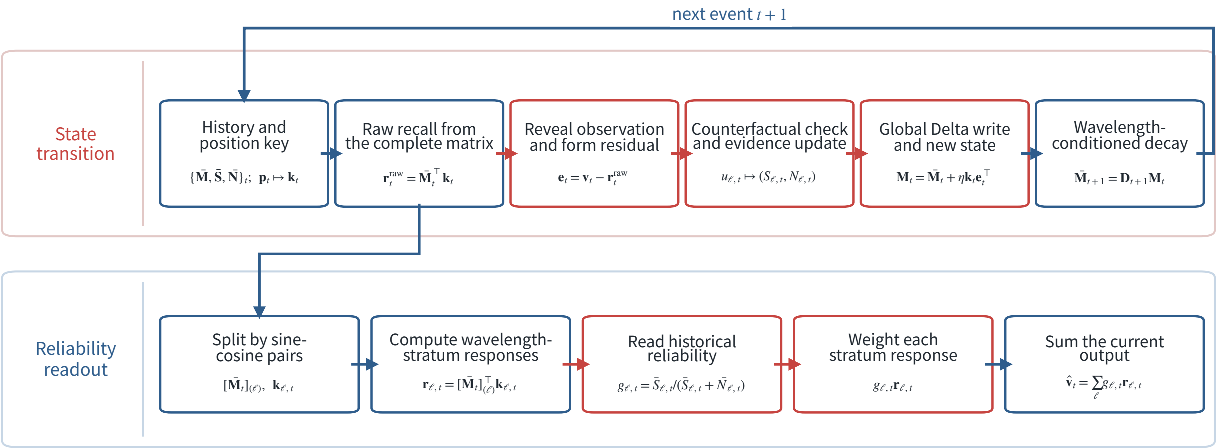

Figure 1: One causal step of S^3-Delta. The upper lane starts from the already decayed historical state, performs raw recall, online verification, and one global Delta write, then wavelength-decays the new state before returning it as history for the next event. The lower lane branches from raw recall, splits the sine–cosine strata, computes their responses, reads historical reliability, weights each response, and sums the current output. The terminal $t \to t+1$ decay is exactly the next-event pre-read operation $\bar{\boldsymbol{M}}_{t+1} = \boldsymbol{D}_{t+1} \boldsymbol{M}_t$ used in the equations.

4.3 Verify afterward: a shared raw residual

After the complete current observation $\boldsymbol{v}_t$ arrives, the residual is computed directly from the global raw recall:

$$\boldsymbol{e}_t = \boldsymbol{v}_t - \bar{\boldsymbol{M}}_t^\top \boldsymbol{k}_t = \boldsymbol{v}_t - \boldsymbol{r}_t^{\text{raw}}. \tag{8}$$

Using raw rather than gated recall is essential. If the write residual also depended on $g_{\ell,t}$, a temporarily suppressed stratum would receive a larger correction simply because it was omitted from the output. Readout weighting would then alter the storage target. Equation (8) avoids this feedback: the main memory has one fitting target, while reliability affects only subsequent outputs.

4.4 Counterfactual reliability evidence

Response magnitude alone does not establish whether a stratum is useful: a strong response can be either a correct recollection or coherent interference. The complete observation, revealed only after the causal output, provides the missing test. We ask whether the already emitted raw recall would have had lower error without stratum ℓ. No second read is needed: removing that stratum changes the error vector algebraically to $\boldsymbol{e}_t + \boldsymbol{r}_{\ell,t}$. This gives the normalized utility

$$u_{\ell,t} = \frac{\|\boldsymbol{e}_t + \boldsymbol{r}_{\ell,t}\|^2 - \|\boldsymbol{e}_t\|^2}{\|\boldsymbol{e}_t\|^2 + \|\boldsymbol{r}_{\ell,t}\|^2 + \epsilon}. \tag{9}$$

If $u_{\ell,t} > 0$, removing the stratum increases error, so the stratum supplied useful information. If $u_{\ell,t} < 0$, removing it improves the result, so it caused interference on this event. Evidence evolves as

$$S_{\ell,t} = \bar{S}_{\ell,t} + [u_{\ell,t}]_+, \qquad N_{\ell,t} = \bar{N}_{\ell,t} + [-u_{\ell,t}]_+. \tag{10}$$

Each wavelength stratum thus accumulates marginal reliability evidence. In this sense, verification acts as post-recall review: a component that helped is trusted more in future recalls, whereas a component that was strongly activated but harmful is attenuated. The mechanism reinforces demonstrated usefulness, not activation magnitude.

4.5 Consolidate: normalized global Delta writing

Consolidation no longer splits the write across strata. Instead, one normalized rank-one Delta update is applied to the complete matrix:

$$\boldsymbol{M}_t = \bar{\boldsymbol{M}}_t + \eta \frac{\boldsymbol{k}_t \boldsymbol{e}_t^\top}{\boldsymbol{k}_t^\top \boldsymbol{k}_t} = \bar{\boldsymbol{M}}_t + \eta \boldsymbol{k}_t \boldsymbol{e}_t^\top, \qquad 0 < \eta \le 1. \tag{11}$$

The second equality follows from $\boldsymbol{k}_t^\top \boldsymbol{k}_t = 1$ in Eq. (2). Frequency differences therefore enter only through wavelength-conditioned decay and the counterfactual reliability gate, not through Delta write strength. The main memory fits one global target; stratification is reserved for temporal prior and reliability estimation.

4.6 Advance the consolidated state to the next event

After consolidation, the elapsed time to event $t+1$ is applied to the new state. Let $s_\ell \in [0, 1]$ be normalized log-frequency, where $s_\ell = 0$ is the longest wavelength and $s_\ell = 1$ the shortest. The stratum half-life is geometrically interpolated:

$$H_\ell = H_{\text{long}} \left(\frac{H_{\text{short}}}{H_{\text{long}}} \right)^{s_\ell}, \qquad H_{\text{long}} \ge H_{\text{short}} > 0. \tag{12}$$

For the true elapsed time $\Delta\tau_{t+1}$, define

$$\gamma_{\ell,t+1} = 2^{-\Delta\tau_{t+1}/H_\ell}, \qquad \boldsymbol{D}_{t+1} = \operatorname{diag}(\gamma_{1,t+1}, \gamma_{1,t+1}, \dots, \gamma_{L,t+1}, \gamma_{L,t+1}). \tag{13}$$

With $\boldsymbol{\gamma}_{t+1} = [\gamma_{1,t+1}, \dots, \gamma_{L,t+1}]^\top$, the terminal transition is

$$\bar{\boldsymbol{M}}_{t+1} = \boldsymbol{D}_{t+1} \boldsymbol{M}_t, \qquad \bar{\mathbf{S}}_{t+1} = \boldsymbol{\gamma}_{t+1} \odot \mathbf{S}_t, \qquad \bar{\mathbf{N}}_{t+1} = \boldsymbol{\gamma}_{t+1} \odot \mathbf{N}_t. \tag{14}$$

This is the terminal $t \to t+1$ arrow in Fig. 1. An implementation may evaluate it when event $t+1$ arrives, immediately before that event's read; the state and output are identical. A longer lifetime for long wavelengths is only a weak prior: broad spatial structure often changes more slowly than fine local detail. Ablations show that the counterfactual reliability gate, not this prior, provides most of the gain.

4.7 Complete step

Algorithm 1: one causal S^3-Delta event. The complete update follows the fixed order below:

1. **Recall:** receive the already aged historical state $(\bar{\boldsymbol{M}}_t, \bar{\mathbf{S}}_t, \bar{\mathbf{N}}_t)$, generate the equalized, unit-total-energy key $\boldsymbol{k}_t$, compute the global raw recall and batched scale contributions $\boldsymbol{R}_t$ once, and emit $\hat{\boldsymbol{v}}_t$ through the historical reliability gate $\boldsymbol{g}_t$.
2. **Verify:** reveal the complete observation $\boldsymbol{v}_t$, compute $\boldsymbol{e}_t$ from the global raw recall, and evaluate every leave-one-stratum-out utility $u_{\ell,t}$.
3. **Consolidate:** update $\mathbf{S}_t, \mathbf{N}_t$ and apply the single normalized Delta write in Eq. (11) to the complete $\boldsymbol{M}_t$.
4. **Advance:** apply the wavelength-conditioned transition in Eq. (14); its result is the aged historical state for event $t+1$.

The recall stage cannot use $\boldsymbol{v}_t$, so the current prediction is strictly historical. Verification and consolidation begin only after the observation becomes available, and forgetting acts on the newly consolidated state only after those operations.

5 Structural Properties and Intuition

5.1 Residual contraction at the current key

Reading the complete update in Eq. (11) again at the same current key gives

$$\left(\eta \frac{\boldsymbol{k}_t \boldsymbol{e}_t^\top}{\boldsymbol{k}_t^\top \boldsymbol{k}_t}\right)^\top \boldsymbol{k}_t = \eta \boldsymbol{e}_t. \tag{15}$$

Therefore, ignoring new decay before the next event, the raw residual at the current key becomes

$$\boldsymbol{e}_t^{\text{after}} = (1-\eta)\boldsymbol{e}_t. \tag{16}$$

The fitting meaning remains as clear as in Delta: $\eta = 1$ removes the current raw residual in one step; $0 < \eta < 1$ performs a stable correction. Spectral stratification and normalization do not destroy this property.

5.2 Bounded and recoverable gates

By Eq. (10), $S, N \geq 0$, so $g \in [0, 1]$ whenever evidence exists. Applying the same continuous-time decay to S and N does not instantly alter their ratio, but reduces the influence of old evidence relative to new observations. The gate therefore retains historical judgment while remaining adaptable after the environment changes.

5.3 Continuous-time semigroup property

For any $a, b \geq 0$,

$$2^{-a/H_\ell} 2^{-b/H_\ell} = 2^{-(a+b)/H_\ell}. \tag{17}$$

Splitting an event-free duration into several short intervals or applying one long interval produces exactly the same decay. This matters for variable-density scanning: different sampler speeds or idle times should not be misinterpreted as equal forgetting durations.

5.4 Why spatial recall is stronger nearby

The key inner product of one stratum at two locations $\boldsymbol{p}$ and $\boldsymbol{p}'$ is

$$\boldsymbol{k}_\ell(\boldsymbol{p})^\top \boldsymbol{k}_\ell(\boldsymbol{p}') = \cos\Big(\boldsymbol{\omega}_\ell^\top (\boldsymbol{p} - \boldsymbol{p}')\Big). \tag{18}$$

A weighted set of frequencies can approximate a local shift-invariant kernel(Rahimi and Recht, 2007). Nearby positions have small phase differences and align across most strata. As distance grows, phases disperse and positive and negative responses cancel. This produces a Gaussian-like center peak without traversing historical coordinates.

5.5 State and computational complexity

The complete $\boldsymbol{M}_t$ contains $d_k d_v$ scalars, equal to one dense fast-weight matrix. The scale-wise S_ℓ, N_ℓ statistics contribute $2L = d_k$ more scalars, for a total state size of

$$d_k d_v + d_k. \tag{19}$$

Read and write cost $O(d_k d_v)$ per event and do not depend on the historical length T. At $d_v = 128$, counterfactual reliability evidence adds only about 0.78% relative to the main matrix.

The complete reference operator nevertheless contains a nonlinear gate and counterfactual update that depend on historical evidence. We do not yet provide a strict associative-prefix formulation

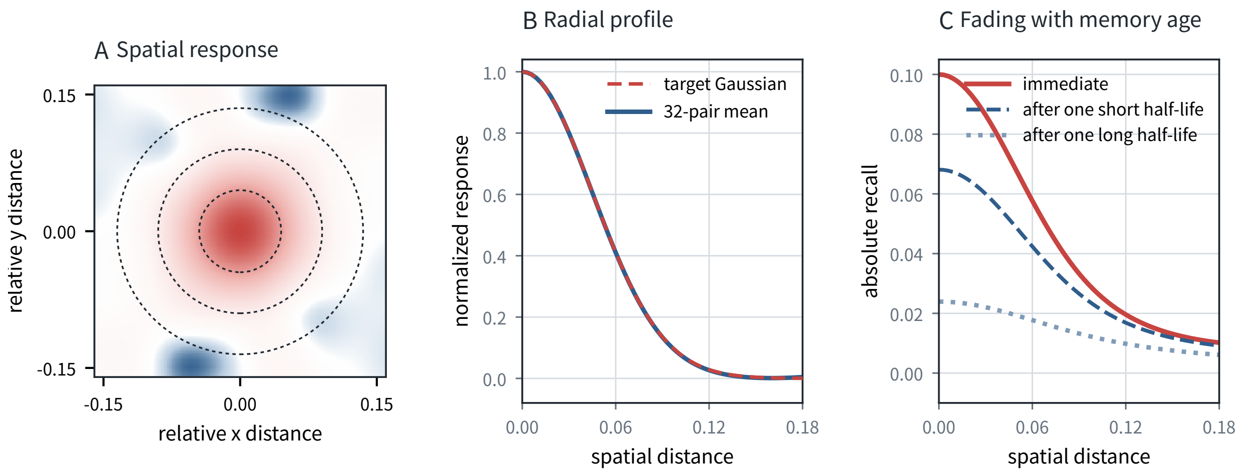

Figure 2: Fuzzy spatial recall induced by fixed sine–cosine keys. The heat map shows a strong central response that weakens with distance; the radial profile validates the local finite-Fourier approximation, while the age profiles show continuous-time fading. With 32 pairs, the near-to-far response ratio is 7.00 and radial monotonicity is 93.4%.

for the entire operator. The Delta part of the main matrix may benefit from chunkwise algorithms, but this does not justify claiming that complete S^3-Delta already has a dense prefix-parallel implementation. The current NumPy prototype is about 3.68× slower in wall-clock time than tuned Delta at $d_v = 128$. The contribution of this work is memory quality and structural closure, not current implementation speed.

6 Experiments

6.1 Task and evaluation protocol

Experiments simulate variable-density scanning in a bounded two-dimensional region. A long trajectory visits coordinates, and every coordinate supplies one complete continuous value vector. Every query is strictly issued before the current value is written and targets the latest history that was *actually observed* at the current position and in its local kernel neighborhood. We use static and slowly changing environments. Sampling intervals are irregular, and the final operator uses the true $\Delta\tau_t$ for forgetting.

6.2 Synthetic data generation and target construction

To make the experiment reproducible without reading the implementation, we specify the complete generator here. Every seed independently generates the world, trajectory, Fourier keys, and observation noise. All methods within a test case receive exactly the same stream and query targets.

1. **Space and value field.** The domain is a 96×96 grid over $[0, 1]^2$, and each observation is a six-dimensional unit vector. The latent field combines 36 random multiscale radial basis functions with 14 random Voronoi regions. RBF radii are sampled log-uniformly from $[0.035, 0.24]$; the regional component enters with weight 0.35. This field generator does not reuse the Fourier frequencies of the memory key. Independent Gaussian noise with standard deviation 0.015 is added to every observation before renormalization.
2. **Static and changing worlds.** The static regime uses one field for the entire stream. The slowly changing test divides the stream into three epochs and applies a local field change at $T/3$ and $2T/3$. Each change selects seven random circular regions of radius 0.045–0.13 and

mixes 15% of the previous value with 85% of a new random direction. This is a controlled piecewise-change stress test, not a continuous physical simulation.

3. **Scanning trajectory.** At every event, the agent performs an eight-neighbor local scan with probability 72%; within these moves, it keeps the previous direction with probability 72%. A random teleport occurs with probability 12%. The remaining 16% triggers a delayed revisit: delays are sampled approximately log-uniformly from 8 events to the full available history, the least recently seen location among eight candidates is preferred, and with probability 45% the trajectory moves one neighboring cell away. The mixture creates contiguous scans, sparse jumps, and revisits across orders of magnitude.

4. **Keys and physical time.** Every method uses the same 32 sine–cosine pairs. As required by Eq. (2), pair energies are equalized to $1/32$ and the complete key is normalized to unit energy at every position. The 2-D frequencies come from randomly shifted Halton points passed through a Gaussian inverse transform. They are sorted by norm into four bands, with kernel scale $\sigma = 0.045$. Inter-event durations are log-normally sampled, clipped to $[0.05, 8]$, and normalized to mean one. A method observes only the current coordinate, its complete observation, and this elapsed duration.

5. **Causal target.** A read is scheduled every eight events and at every forced revisit, always before writing. Within radius 3σ, the evaluator keeps only previously visited cells and their latest ledger observations. Their Gaussian-weighted average, with weight $\exp[-d^2/(2\sigma^2)]$, is the target. A query is classified as *exact* if the current cell was visited and as *near* if only its neighborhood contains history. Completely unseen neighborhoods are excluded from the primary error. The target therefore comes from observed history rather than the unobserved latent field and never includes the current new value.

The primary metric is query MSE,

$$\mathrm{MSE} = \frac{1}{N_q} \sum_{q=1}^{N_q} \left\| \hat{\boldsymbol{v}}_q - \boldsymbol{v}_q^\star \right\|_2^2, \tag{20}$$

where $\boldsymbol{v}_q^\star$ is the historical Gaussian-neighborhood target above. The number is not divided again by the six value channels. Main tables score valid queries only in the second half of each stream to reduce empty-state warm-up effects. The 30 paired evaluations comprise three lengths, two regimes, and five test seeds. We bootstrap 20000 times over the five seed-level paired gains, rather than treating strongly correlated event-level queries as independent samples. A separate 64K scale-mismatch analysis uses ten seeds disjoint from validation, fixes the encoding scale at 0.045, and changes only the evaluator's target radius.

We compare:

- **Scalar-forgetting FWP:** outer-product writing with one exponential forgetting rate for the complete state;
- **Dense Delta:** the conventional Delta residual update over the complete Fourier key;
- **Uniform-lifetime Delta:** dense Delta preceded by a tuned uniform continuous-time decay;
- **Learned causal Gated Delta:** a four-parameter learned gate in the same query-before-write protocol, using the preceding residual to control continuous-time decay and the current residual to control write strength;
- **S^3-Delta:** the complete proposed operator.

Hyperparameters are selected only on a separate validation set with seeds 1709 and 3413, length 8192, and both environments; selection equally averages the final-half errors of the four seed–regime cases. Testing uses five disjoint seeds: 6827, 13649, 27329, 40961, and 54713. The complete search grids are: write rates $\{0.025, 0.05, 0.1, 0.2\}$ for the baselines; half-lives $\{32, 128, 512, 2048\}$ for scalar

Table 1: Paired MSE reduction of S^3-Delta relative to dense Delta under recall-radius mismatch. Brackets are ten-seed bootstrap 95% confidence intervals.

σ_{target}	Static	Slowly changing
.030	1.42% $[0.86, 1.87]$	3.37% $[2.90, 3.95]$
.045	4.02% $[3.13, 4.99]$	8.20% $[7.14, 9.41]$
.060	7.15% $[5.56, 9.00]$	13.40% $[11.57, 15.39]$
.090	13.07% $[10.10, 16.46]$	21.20% $[18.37, 24.19]$

FWP and $\{4096, 16384, 65536\}$ for uniform-lifetime Delta; and write rates $\{0.05, 0.1, 0.2\}$ with long half-lives $\{16384, 32768, 65536\}$ for S^3-Delta. Its short half-life is fixed to one quarter of the long half-life. The selected settings are $\eta = 0.05$ and half-life 512 for scalar FWP; $\eta = 0.05$ for Delta; $\eta = 0.05$ and half-life 65536 for uniform-lifetime Delta; and $\eta = 0.1$, $H_{\text{long}} = 65536$, $H_{\text{short}} = 16384$ for S^3-Delta. Test settings remain frozen across all horizons and are never retuned on test results. Uniform normalized writing in Eq. (11) introduces no additional spectral-tilt hyperparameter.

The learned causal Gated Delta is a task adaptation of the Gated Delta update idea(Yang et al., 2025), not a full language model. Let E_{t-1} be the residual energy before the preceding write. Its causal update is

$$\alpha_t = \exp\left[-\frac{\ln 2}{H}\exp(aE_{t-1})\Delta\tau_t\right], \quad \bar{\boldsymbol{M}}_t = \alpha_t \boldsymbol{M}_{t-1}, \quad \beta_t = \sigma\left(b_0 + b_1 \left\|\boldsymbol{v}_t - \bar{\boldsymbol{M}}_t^{\top}\boldsymbol{k}_t\right\|^2\right), \tag{21}$$

$$\boldsymbol{M}_t = \bar{\boldsymbol{M}}_t + \beta_t \boldsymbol{k}_t(\boldsymbol{v}_t - \bar{\boldsymbol{M}}_t^{\top}\boldsymbol{k}_t)^{\top}. \tag{22}$$

The current query is scored from $\bar{\boldsymbol{M}}_t$ before $\boldsymbol{v}_t$ is revealed, so the current truth cannot leak into the score. The four coefficients are fitted on additional disjoint seeds 911, 1823, and 3659. Each stream has length 4096, both environments are used, and final-half MSE selects among 40 candidate parameter sets generated once. The frozen result is $H = 77303.44$, $a = -0.8275$, $b_0 = -2.9733$, and $b_1 = 0.09268$; it is not retrained for any of the five test seeds or three horizons. This control deliberately receives extra training data and is therefore a strong reference. The appendix exposes the core code of every main-table method.

6.3 Does fuzzy spatial recall hold?

Fig. 2 isolates the positional key from the update rule. With 32 Fourier pairs, the mean-squared approximation error to the target local kernel is 0.011893 with bootstrap interval $[0.011277, 0.012564]$. The mean near-to-far response ratio is 7.00 and radial monotonicity is 93.4%.

As the number of pairs increases from 16 to 32, 64, and 128, kernel approximation error decreases from 0.028457 to 0.011893, 0.004814, and 0.002267. The decrease of response with distance is therefore supported by a finite-kernel approximation that improves consistently with feature count, rather than by the illustrative heat map alone.

6.4 Robustness to recall-radius mismatch

The main benchmark uses the intended recall radius in both the encoder and evaluator. To test dependence on that match, we keep the encoder fixed at $\sigma_{\text{enc}} = 0.045$ and evaluate historical targets with $\sigma_{\text{target}} \in \{0.030, 0.045, 0.060, 0.090\}$. The 64K experiment uses ten validation-disjoint seeds and the already frozen operator settings.

All eight confidence intervals remain positive, including the narrower-than-encoded target at 0.030. We also compared the final uniform write with the earlier low-frequency-tilted write; uniform writing has lower aggregate MSE in every environment–radius condition. The final operator therefore removes that manual preference rather than retaining an unsupported tuning choice.

Table 2: Query MSE in long two-dimensional scans; lower is better. Learned gate denotes learned causal Gated Delta. Gain remains the paired mean improvement of S^3-Delta over tuned dense Delta; brackets show five-seed bootstrap 95% confidence intervals.

Length	Environment	Scalar FWP	Delta	Uniform Delta	Learned gate	S^3-Delta	Gain
16K	Static	.433575	.267890	.268253	.267427	**.254571**	5.28% [3.75, 6.81]
16K	Slow	.354103	.227621	.227168	.228629	**.211926**	7.05% [6.03, 8.09]
64K	Static	.440571	.265917	.266460	.265632	**.254804**	4.73% [3.31, 6.16]
64K	Slow	.349588	.228839	.227342	.228032	**.208750**	8.99% [7.23, 10.85]
262K	Static	.441259	.267126	.267250	.266432	**.255995**	4.82% [3.06, 6.31]
262K	Slow	.349253	.228178	.226383	.226737	**.209418**	8.41% [7.07, 9.74]

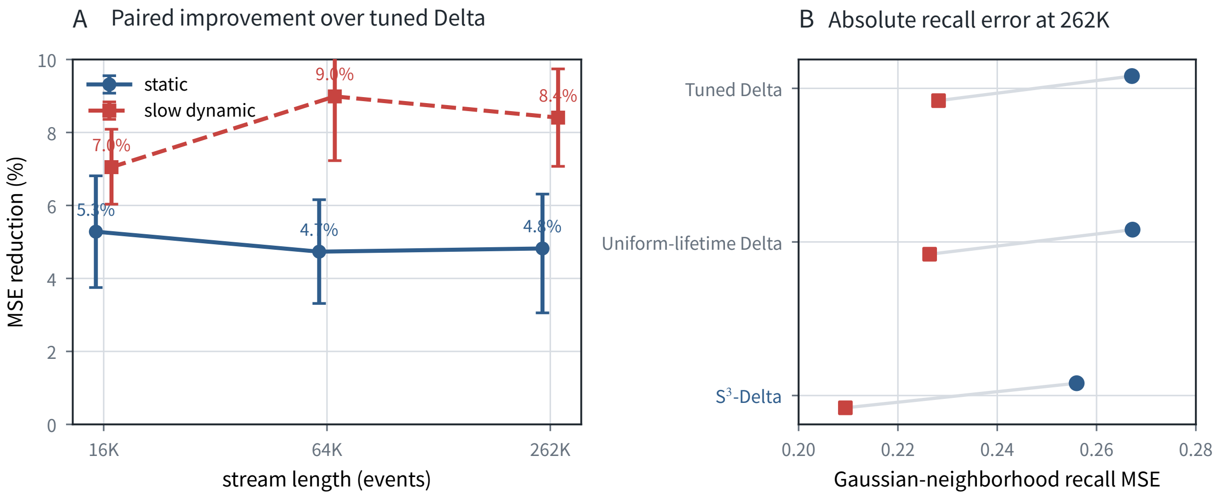

Figure 3: Results at 16K, 64K, and 262K. S^3-Delta obtains the lowest error in both static and slowly changing environments and wins all five paired test seeds, rather than being driven by a few favorable seeds in the mean.

6.5 Main long-horizon results

As shown in Table 2 and Fig. 3, S^3-Delta outperforms Delta in all six horizon–environment combinations. Paired mean improvements range from 4.73% to 8.99%, and the smallest confidence-interval lower bound is 3.06%. Relative to the additionally trained causal Gated Delta, the six mean improvements are 4.46%–8.65%; every five-seed bootstrap lower bound is positive, with a minimum of 3.11%, and all 30 seed-level pairs remain positive. The learned gate stays close to ordinary Delta, indicating that one globally learned data-dependent decay/write gate does not replace explicit wavelength-stratum reliability evaluation. Larger gains in the slowly changing environment indicate that the counterfactual reliability gate suppresses outdated or interference-dominated strata rather than merely extending memory lifetime.

6.6 Performance across memory ages

At 262K, we bucket targets by the number of events since the previous effective visit. Exact queries use the time since the current cell was last visited; near queries use the age of the most recent historical observation in the active kernel neighborhood. In the static environment, the relative gains over Delta for ages 1–64, 65–1024, 1025–16384, and > 16384 are 18.54%, 3.02%, 3.80%, and 4.27%. In the slowly changing environment they are 21.15%, 9.76%, 7.84%, and 8.01%. The gain is not restricted to newly written content; in the dynamic environment it remains consistent across all age ranges.

6.7 Structural ablations and mechanism attribution

To isolate one factor at a time, all six variants except dense Delta use $\eta = 0.1$, the same complete raw residual, and the uniform global update in Eq. (11); only $\boldsymbol{D}_t$ or $\boldsymbol{g}_t$ changes. Dense Delta uses its separately validated $\eta = 0.05$ and has neither decay nor stratum gating. The counterfactual reliability gate is defined by Eqs. (6) and (10). The instantaneous-energy gate stores no $S_{\ell,t}, N_{\ell,t}$: it sets $E_{\ell,t} = \|\boldsymbol{r}_{\ell,t}\|^2 / d_v$ and $g^{\text{energy}}_{\ell,t} = \frac{1}{2} + \frac{1}{2}E_{\ell,t}/[E_{\ell,t} + 10^{-5}\exp(4s_\ell)]$.

Each variant answers one question:

- **Wavelength-conditioned decay, ungated:** are wavelength-specific lifetimes sufficient?
- **Uniform lifetime plus counterfactual reliability gate:** does the gate help without lifetime diversity?
- **Shuffled/reversed lifetimes plus counterfactual reliability gate:** does the lifetime set matter, or its correct wavelength order?
- **Wavelength-conditioned decay plus instantaneous-energy gate:** can current response magnitude replace historical error evidence?
- **Complete S^3-Delta:** are the correct wavelength prior and counterfactual reliability gate complementary?

Table 3: Five-seed structural ablations at 262K with their actual implementations. H_ℓ denotes the wavelength half-life in Eq. (13). Values are query MSE; lower is better.

Method	Stepwise implementation	Static	Slow
Dense Delta	$\boldsymbol{D}_t = I,\ g_{\ell,t} = 1$; read the complete key and apply conventional Delta	.267126	.228178
Wavelength-conditioned decay, ungated	Use H_ℓ in the correct wavelength order; set $g_{\ell,t} = 1$ and do not update evidence	.278108	.242609
Uniform lifetime plus counterfactual reliability gate	Set every $H_\ell = \sqrt{H_{\text{long}}H_{\text{short}}}$; use the counterfactual reliability gate	.257048	.210268
Shuffled lifetimes plus counterfactual reliability gate	Preserve the same $\{H_\ell\}$ but randomly permute its assignment once per test stream	.258017	.210798
Reversed lifetimes plus counterfactual reliability gate	Assign $\{H_\ell\}$ in the exact reverse wavelength order	.260492	.212435
Wavelength-conditioned decay plus instantaneous-energy gate	Decay in the correct order; compute g^{energy}_ℓ from current E_ℓ, without historical evidence	.269160	.231694
Complete S^3-Delta	Decay in the correct wavelength order; use the counterfactual reliability gate	**.255995**	**.209418**

Table 3 and Fig. 4 provide the mechanism evidence. The instantaneous-energy gate improves over the ungated variant by 3.22% and 4.50%, so attenuating weak responses has some value, but magnitude cannot distinguish useful memory from strong interference. The counterfactual reliability gate improves a further 5.17% and 9.75% over the instantaneous-energy gate, and 8.63% and 13.94% over the ungated variant. Lifetime diversity contributes only 0.368% and 0.409% over uniform lifetimes, while correct order contributes 0.701% and 0.674% over shuffled lifetimes. Thus, the counterfactual reliability gate explains most of the gain; wavelength-conditioned decay is a

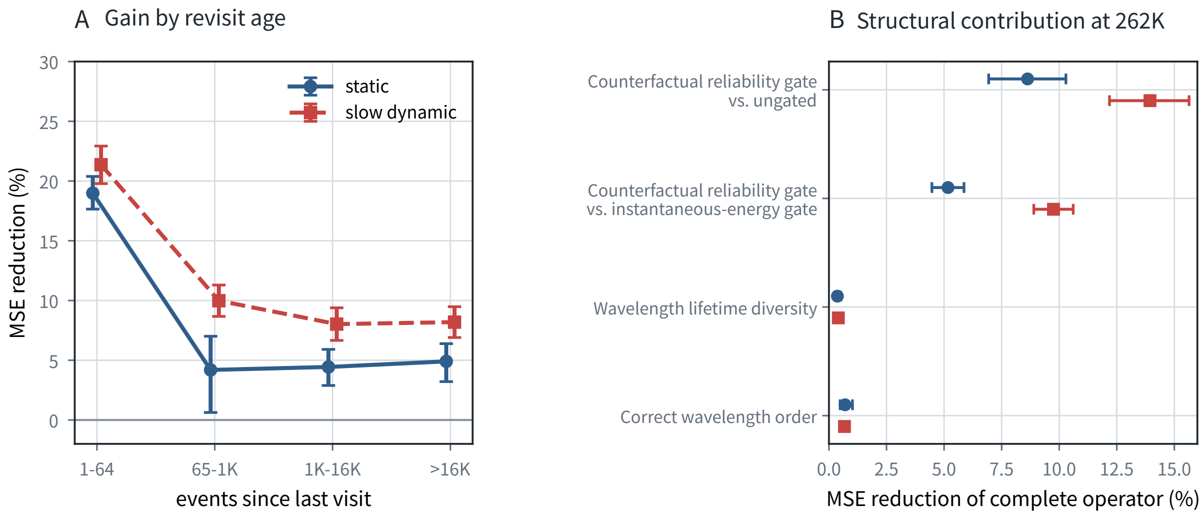

Figure 4: Mechanism and age analysis. The left panel compares gain across revisit intervals; the right panel shows the main structural ablations. The counterfactual reliability gate is the dominant source of improvement; correct wavelength order adds a smaller, consistent benefit.

smaller complement, and the instantaneous-energy gate only excludes a simpler magnitude-gating explanation.

6.8 Consistency under irregular time

We randomly split the same total elapsed duration into different event intervals and compare stepwise decay with merged decay. The largest single-seed metric drift is 0.0061%, while semigroup state and readout errors are about 10^{-14}. Adding $\Delta\tau_t$ therefore preserves the original long-horizon performance and the implementation follows Eq. (13) rather than approximating variable duration with fixed event counts.

6.9 Summary of evidence

The experiments support four conclusions: the Fourier key creates local fuzzy recall; the complete operator outperforms tuned Delta on unseen seeds and streams as long as 262K; most of the gain comes from the counterfactual reliability gate rather than slower forgetting; and continuous-time decay is consistent under irregular sampling. This supports usability for the defined spatial scanning problem, not extrapolation to a universal sequence model or complete robotic system.

7 Discussion and Limitations

Spatial sidecar for a temporal backbone. S^3-Delta is intended to complement, rather than replace, a learned temporal model. Let x_t^- contain information available before the current spatial observation and let $\bar{\mathcal{Z}}_t$ denote the aged S^3-Delta state. A causal adapter can be written as

$$\begin{aligned} m_t &= \mathrm{Read}_{\mathrm{S3}}(\boldsymbol{p}_t; \bar{\mathcal{Z}}_t), \\ h_t &= F_\theta\Big(h_{t-1}, [x_t^-; W_m m_t]\Big), \\ \bar{\mathcal{Z}}_{t+1} &= \mathrm{Step}_{\mathrm{S3}}(\bar{\mathcal{Z}}_t, \boldsymbol{p}_t, E_v(o_t), \Delta\tau_{t+1}), \end{aligned} \tag{23}$$

where F_θ may be an LSTM or Mamba block(Hochreiter and Schmidhuber, 1997; Gu and Dao, 2023). The projected spatial recall can be concatenated with the recurrent input or added through a residual adapter, while the backbone continues to model temporal order. The S^3-Delta state remains a coordinate-indexed, fixed-capacity side memory; its analytic recurrence need not be learned even when W_m and E_v are trained. Equation (23) preserves the paper's causal boundary because o_t enters only after m_t has been produced. This is an architectural interface for future end-to-end evaluation, not an experimentally established Mamba or LSTM result in the present paper.

It stores experience, not a map. S^3-Delta returns a continuous, fuzzy, and forgetful spatial history. It can provide a downstream network with fixed-capacity context about what appeared near the current position. It is unsuitable when a lossless occupancy grid, an exact coordinate list, or fully auditable replay is required.

Euclidean proximity is not reachability. Fourier keys produce similarity from coordinate distance. Points across a wall or on different floors can be geometrically close but mutually unreachable. Tasks requiring topological reachability should modify the positional key using an external graph, an environment representation, or a learned metric.

Complete-observation assumption. This work assumes that each write provides a complete value vector at the current position. If an observation covers only some channels, a full residual would treat unobserved channels as zeros. An observation mask is then required so that updates occur only in the visible value subspace.

Finite spectral approximation. A finite number of sine–cosine pairs causes sidelobes and cross-location interference. More frequencies improve the kernel approximation but increase state and computation linearly. The spectrum should also match the bounded spatial domain and the desired recall radius.

Kernel-scale scope. The main benchmark sets both the positional encoding length scale and the target Gaussian kernel to $\sigma = 0.045$. The additional ten-seed analysis keeps the encoder at 0.045 while varying the target radius from 0.030 to 0.090, and retains positive paired confidence bounds throughout. This reduces concern that the main result requires an exact radius match. Anisotropic neighborhoods, learned spectra, and cross-scene transfer remain untested.

Counterfactual reliability evidence is locally marginal. Counterfactual utility compares full raw recall with removal of one stratum. It is not a Shapley value and does not prove statistical independence. Correlated strata may share or duplicate evidence; stricter decorrelated estimation remains open.

Efficiency remains unfinished. The evidence update is event-recurrent, and no fused GPU kernel is implemented. Future work includes blockwise sufficient statistics, fused kernels, and approximate parallel updates that preserve counterfactual meaning. Until then, theoretical $O(d_k d_v)$ cost does not imply a practical speed advantage over Delta.

External validity. The present evidence comes from controlled two-dimensional synthetic long-horizon scans. This isolates the memory operator but omits real sensor noise, three-dimensional environments, learned value encoders, and end-to-end decision metrics. The main benchmark uses five test seeds and the scale-robustness analysis uses ten; both should still be treated as preliminary rather than large-sample evidence. The next step is evaluation on real or high-fidelity mobile observation streams, with equal-budget comparisons against explicit maps, nearest-neighbor retrieval, and modern gated linear memories.

8 Conclusion

We introduced S^3-Delta, a fixed-capacity fuzzy spatial memory operator for long streams with known coordinates. It interprets sine–cosine positional keys as wavelength strata. At each event it recalls from the already aged history, evaluates stratum contributions after the current truth arrives, consolidates the complete matrix with one global Delta update, and only then wavelength-decays the new state toward the next event. The resulting loop is simple: recall, verify, consolidate, then advance. The spatial prior determines where the state responds; online evidence determines which responses should be trusted.

Across static and slowly changing scans from 16K to 262K, S^3-Delta consistently improves on tuned Delta and on the additionally trained four-parameter causal Gated Delta, while ablations localize most of the gain to the counterfactual reliability gate. This result is specific to the spatial recall protocol and does not establish universal superiority over a full Gated DeltaNet or on general sequence tasks. The operator preserves shared residual fitting and explicitly estimates each wavelength stratum's readout reliability. Its query-before-observation interface also permits use as a spatial sidecar for recurrent backbones such as LSTMs and Mamba, but that end-to-end extension remains to be evaluated.

Statements and Declarations

Funding. No funds, grants, or other support was received for this work.

Competing interests. The author has no relevant financial or non-financial interests to disclose.

Data availability. This study uses synthetic data generated by the protocol specified in Section 6. The generated data and processed results supporting the findings are available from the corresponding author upon reasonable request.

Code availability. Core recurrences for all main-table operators are provided in Section B. A self-contained reproduction package and the generated outputs are included with the submission materials.

Author contributions. The author conceived the study, developed and implemented the method, designed and conducted the experiments, analyzed the results, and wrote and revised the manuscript.

Ethics approval. Not applicable. This study did not involve human participants, animals, or identifiable personal data.

A From Interest Protection to Counterfactual Reliability

The final operator did not emerge directly from its equations. The initial hypothesis was that recalling a memory should temporarily reduce its subsequent forgetting: when the current position reactivates a spatial direction, relevant history should decay more slowly than unrelated background. This is plausible, but activation alone cannot determine correctness. If an interference-contaminated direction is protected, the error survives longer than the background.

We next separated key-related, value-related, and background components into two- or three-spectrum transitions with different forgetting rates, and combined them with directed Delta updates. These variants could describe more cases but exposed two problems. First, several protection and correction mechanisms acted on the same state and pulled against one another. Second, the growing collection of components lacked one decisive criterion and became difficult to explain. Fully atomizing every sine–cosine pair was structurally clean and parallel-friendly, but independent strata repeatedly fitted the same complete observation, lost cross-stratum coordination, and did not outperform dense Delta.

After the shared-residual route was established, an intermediate version still used $w_\ell = \exp(-2s_\ell)$ and $Z_t = \sum_j w_j \|\boldsymbol{k}_{j,t}\|^2$ to distribute the write nonuniformly. Combined with wavelength-conditioned decay and the counterfactual reliability gate, that version improved over Delta by 3.85%–7.23% in the five-seed long-horizon study. We then fixed $w_\ell = 1$. Uniform writing achieved lower aggregate MSE in all eight environment–radius conditions of the ten-seed mismatch test, while the main-table gains increased to 4.73%–8.99%. The final operator therefore confines wavelength dependence to decay and reliability estimation and uses one global Delta update.

Two conclusions were retained. First, each sine–cosine pair has a wavelength interpretation and can be evaluated separately. Second, stratification should not enter the write target: one raw residual should update the complete matrix. The final design therefore replaces activation-based protection with measured marginal error reduction. Reliability controls readout, while global Delta controls storage, so the two mechanisms have distinct roles.

Table 4: Evolution of the design and the insights retained in the final operator.

Stage	Main idea	Problem revealed	Retained insight
Uniform forgetting	Decay the whole state by one scalar	Cannot separate current relevance from background	Continuous-time decay
Interest protection	Slow decay along the current key direction	Incorrect interference can also be protected	Revisits should influence future retrieval
Two/three spectra	Assign rates by key/value agreement	Overlapping mechanisms pull against Delta correction	Decompose, then compare
Independent atomic strata	Run one Delta per sine–cosine pair	No cross-stratum coordination; target is fitted repeatedly	A pair is the minimal scale
Wavelength-conditioned decay	Set lifetimes from wavelength	Alone it does not beat Delta	A weak wavelength-lifetime prior
Nonuniform spectral write	Allocate the shared residual with w_ℓ	Better than Delta with gating, but the extra frequency bias is redundant	Validates the overall route
Final structure	Global uniform Delta, counterfactual reliability gate, wavelength decay	Reference implementation remains recurrent and slower	A clean closed loop

The experiments distinguish two explanations. If lower-frequency lifetimes alone explain the gain, the method is merely a handcrafted decay schedule. If the counterfactual reliability gate explains most of it, post-recall scale evaluation is the central mechanism. The results support the latter interpretation.

B Core Code of the Main-Table Operators

The listing below exposes the recurrences actually used by the five main-table methods, omitting data generation, statistics, and file I/O. Every method receives the same unit-norm Fourier key, and every query is scored before the current observation is written. The four coefficients of the learned gate come from the disjoint training streams described in the main text; this control is not an unmodified full Gated DeltaNet.

```
# M: [key_dim, value_dim], ||k||_2 = 1
def scalar_fwp(M, k, v, dt, H, eta):
    M *= 2.0 ** (-dt / H)
    M += eta * outer(k, v)
    return M
def dense_delta(M, k, v, eta):
    error = v - M.T @ k
    M += eta * outer(k, error)
    return M
def uniform_lifetime_delta(M, k, v, dt, H, eta):
    M *= 2.0 ** (-dt / H)
    error = v - M.T @ k
    M += eta * outer(k, error)
    return M
def learned_causal_gated_delta(M, k, v, dt, previous_E,
                               H, a, b0, b1):
    alpha = exp(-(log(2.0) / H) * exp(a * previous_E) * dt)
    M *= alpha                         # applied before the query
    recall = M.T @ k                   # score this before seeing v
    error = v - recall                 # available only after the score
    E = error @ error
    beta = sigmoid(b0 + b1 * E)
    M += beta * outer(k, error)
    return M, E
# S3-Delta stores M by sine-cosine row pairs: M[l] is [2, value_dim].
def s3_read(M, signal, noise, k):
    atoms = k.reshape(-1, 2)
    contribution = einsum("li,lid->ld", atoms, M)
    evidence = signal + noise
    gate = where(evidence > eps, signal / maximum(evidence, eps), 1.0)
    return sum(gate[:, None] * contribution, axis=0)
def s3_observe(M, signal, noise, k, v, eta):
    atoms = k.reshape(-1, 2)
    contribution = einsum("li,lid->ld", atoms, M)
    error = v - sum(contribution, axis=0)       # ungated raw residual
    base_error = error @ error
    removed_error = error[None, :] + contribution
    utility = (sum(removed_error**2, axis=1) - base_error)
    utility /= maximum(base_error + sum(contribution**2, axis=1), eps)
    signal += maximum(utility, 0.0)
    noise += maximum(-utility, 0.0)
    M += eta * einsum("li,d->lid", atoms, error)  # one global Delta write
def s3_advance(M, signal, noise, next_dt, half_life):
    gamma = 2.0 ** (-next_dt / half_life)
    M *= gamma[:, None, None]
    signal *= gamma
    noise *= gamma
```

For S^3-Delta, the caller uses three ordered calls: (1) `s3_read` and score the historical output; (2) `s3_observe` after the value arrives; and (3) `s3_advance` for the interval to the next event. The listing exposes every state transition; data generation, random-seed management, train/test partitioning, and bootstrap aggregation are omitted.

References

Jimmy Ba, Geoffrey E. Hinton, Volodymyr Mnih, Joel Z. Leibo, and Catalin Ionescu. Using fast weights to attend to the recent past. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games.* Cambridge University Press, 2006. doi: 10.1017/CBO9780511546921.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):44:1–44:37, 2014. doi: 10.1145/2523813.

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

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

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

Ali Hatamizadeh et al. Gated DeltaNet-2: Decoupling erase and write in linear attention. *arXiv preprint arXiv:2605.22791*, 2026.

Simon Haykin. *Adaptive Filter Theory.* Prentice Hall, 4 edition, 2002.

João F. Henriques and Andrea Vedaldi. MapNet: An allocentric spatial memory for mapping environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8476–8484, 2018.

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

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79.

Pentti Kanerva. *Sparse Distributed Memory.* MIT Press, 1988.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5156–5165, 2020.

Kimi Team. Kimi linear: An expressive, efficient attention architecture. *arXiv preprint arXiv:2510.26692*, 2025.

Jyrki Kivinen, Alexander J. Smola, and Robert C. Williamson. Online learning with kernels. *IEEE Transactions on Signal Processing*, 52(8):2165–2176, 2004. doi: 10.1109/TSP.2004.830991.

Brent Komer, Terrence C. Stewart, Aaron R. Voelker, and Chris Eliasmith. A neural representation of continuous space using fractional binding. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*, 2019.

Weifeng Liu, José C. Principe, and Simon Haykin. *Kernel Adaptive Filtering: A Comprehensive Introduction.* Wiley, 2010. doi: 10.1002/9780470608593.

Emilio Parisotto and Ruslan Salakhutdinov. Neural map: Structured memory for deep reinforcement learning. In *International Conference on Learning Representations*, 2018.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *Proceedings of the 38th International Conference on Machine Learning*, pages 9355–9366, 2021.

Jürgen Schmidhuber. Learning to control fast-weight memories: An alternative to dynamic recurrent networks. *Neural Computation*, 4(1):131–139, 1992. doi: 10.1162/neco.1992.4.1.131.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, volume 33, pages 7537–7547, 2020.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. In *Proceedings of the 41st International Conference on Machine Learning*, 2024a.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. In *Advances in Neural Information Processing Systems*, volume 37, 2024b.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *International Conference on Learning Representations*, 2025.