Temporal Memory for Resource-Constrained Agents: Continual Learning via Stochastic Compress–Add–Smooth

Anonymous authors
Paper under double-blind review

Abstract

An agent that operates sequentially must incorporate new experience without forgetting old experience, under a fixed memory budget. We propose a framework in which memory is not a parameter vector but a stochastic process: a Bridge Diffusion on a replay interval $[0, 1]$, whose terminal marginal encodes the present and whose intermediate marginals encode the past. New experience is incorporated via a three-step *Compress–Add–Smooth* (CAS) recursion. We test the framework on the class of models with marginal probability densities modeled via Gaussian mixtures of fixed number of components K in d dimensions; temporal complexity is controlled by a fixed number L of piecewise-linear protocol segments whose nodes store Gaussian-mixture states. The entire recursion costs $O(LKd^2)$ flops per day — no backpropagation, no stored data, no neural networks — making it viable for controller-light hardware.

Forgetting in this framework arises not from parameter interference but from lossy temporal compression: the re-approximation of a finer protocol by a coarser one under a fixed segment budget. We find that the retention half-life scales linearly as $a_{1/2} \approx c\,L$ with a constant $c > 1$ that depends on the dynamics but not on the mixture complexity K, the dimension d, or the geometry of the target family. The constant c admits an information-theoretic interpretation analogous to the Shannon channel capacity. The stochastic process underlying the bridge provides temporally coherent "movie" replay — compressed narratives of the agent's history, demonstrated visually on an MNIST latent-space illustration. The framework provides a fully analytical "Ising model" of continual learning in which the mechanism, rate, and form of forgetting can be studied with mathematical precision.

1 Introduction

The agent problem. Consider an agent — a building controller, a robot, a sensor node — that processes a stream of daily experiences, each represented as a probability distribution over a d-dimensional physical or latent state space. The agent must maintain a fixed-size memory from which it can replay past experiences to inform current decisions: warm-starting a recovery controller from last winter's occupancy pattern, recalling a previously visited room's obstacle layout, or restoring a sensor calibration profile.

The core difficulty — that a network trained sequentially on new data abruptly loses performance on previously learned tasks, a phenomenon termed **catastrophic interference** McCloskey and Cohen (1989) or **catastrophic forgetting** French (1999) – has motivated a large body of work. Standard **Continual Learning** (CL) methods Kirkpatrick et al. (2017); Shin et al. (2017); Schwarz et al. (2018) represent memory as neural network parameters and manage forgetting through regularisation, replay buffers – including recent approaches that use denoising diffusion models as the replay generator Gao and Liu (2023); Jodelet et al. (2023); Meng et al. (2024); Kim et al. (2024); Liang et al. (2024); He et al. (2024); Hu et al. (2025), or architecture expansion Parisi et al. (2019); De Lange et al. (2022); Wang et al. (2024). These approaches require gradient-based training, stored data, and compute budgets that are often unavailable on edge hardware. We propose an alternative: **memory is not a parameter vector but a *stochastic process* whose intermediate-time marginals encode the past**.

The idea. The agent maintains a Bridge Diffusion (BD) description on a fixed replay interval $[0, 1]$. The terminal marginal at $t = 1$ represents the current day. Earlier days are stored as intermediate-time marginals at designated readout times $t_{m|n} \in (0, 1)$. Incorporating a new day is a three-step recursion — **Compress–Add–Smooth** — carried out entirely within a chosen parameterised density class. Fixed memory is enforced by two budgets: a *state budget* K (number of Gaussian-mixture components) and a *temporal budget* L (number of piecewise-linear protocol segments, whose $L + 1$ nodes each store a K-component Gaussian mixture). The total memory footprint is $O(Kd^2L)$ floating-point numbers.

Replaying $0 \to 1$ produces a compressed "movie" of the agent's history, realised on two levels: a smooth-in-time evolution of the marginal probability density, and smooth individual sample paths generated via the drift reconstructed from the density path (Appendix A).

Two stories under one mathematical umbrella. This paper serves two audiences, unified by a common mathematical language rooted in non-equilibrium statistical mechanics, stochastic optimal control, and optimal transport theory. The mathematical backbone is a Bridge Diffusion — a framework for controlled stochastic processes whose marginal density path is prescribed and whose drift is reconstructed from the Fokker–Planck equation. The approach is related to Schrödinger bridges Léonard (2013); Chen et al. (2021); De Bortoli et al. (2021), flow matching Lipman et al. (2023), stochastic interpolants Albergo and Vanden-Eijnden (2023) and Path-Integral Diffusion Behjoo and Chertkov (2025); Chertkov and Behjoo (2025); Chertkov (2025; 2026), but differs in that the density path is specified directly as a piecewise-linear interpolant, rather than optimised or learned. The approach is plug-and-play: the Compress–Add–Smooth recursion works with any parameterised density family for which piecewise-linear interpolation is well defined. We illustrate it here on the simple and analytically transparent-mixture (GM) class, but the same recursion applies, in principle, to richer representations — for instance, normalising flows or score-based models that use neural networks as function approximators.

For the **controls/robotics/edge-AI community** Schwarz et al. (2018); Parisi et al. (2019), the framework is a practical temporal memory for resource-constrained agents: the compress–add–smooth recursion costs $O(LKd^2)$ flops per day (matrix operations, no backpropagation), the replay query costs $O(Kd^2)$ (a single interpolation of GM parameters), and the entire pipeline runs on a microcontroller.

For the **continual learning community** Kirkpatrick et al. (2017); De Lange et al. (2022); Wang et al. (2024), the GM instantiation of the framework is an analytically tractable "Ising model" of forgetting: a minimal, exactly solvable system in which the mechanism (temporal compression), rate (controlled by L), and form (two-regime curve, confusion-dominated) of forgetting can be studied with mathematical precision — questions that are not feasible to answer in neural-network-based and dynamics-absent CL. In the neuroscience-inspired sleep replay literature González et al. (2020); Golden et al. (2022); Tadros et al. (2022); Golden et al. (2025); Vins et al. (2025), off-line replay is shown to prevent catastrophic forgetting by pushing synaptic weights toward joint solution manifolds; our SDE-based replay (Section 9.2) is structurally analogous.

Summary of results. We report experiments for single-Gaussian (K=1), Gaussian-Mixture (GM) (K up to 8), and MNIST LeCun et al. (1998) latent-space GM daily distributions over n=100 days, with the following principal findings:

1. *Two-regime forgetting curve.* The normalized forgetting $\bar{F}(a)$ exhibits a low-error plateau for recent memories, followed by a steep sigmoid transition. The retention half-life $a_{1/2}$ — the age at which $\bar{F}$ crosses 0.5 — is the natural summary statistic (Section 5).
2. *Linear scaling:* $a_{1/2} \approx c\,L$. The half-life scales linearly with the segment budget, from $a_{1/2} = 14$ at $L = 5$ to $a_{1/2} = 74$ at $L = 30$, with $c \approx 2.4$ for the default geometry. Since $c > 1$, the CAS scheme outperforms a naïve First-In-First-Out (FIFO) buffer (which gives $a_{1/2} = L$) by a factor of $\sim$2.4$\times$. We argue that c admits an information-theoretic interpretation as a channel capacity (Section 9.1). This linear scaling is confirmed across all experimental settings (Sections 6–7).
3. *Independence of K, d, and geometry.* The half-life is essentially independent of the mixture complexity K (tested for $K \in \{1, 2, 3, 5, 8\}$), the ambient dimension d (tested for d up to 30), the crowding

geometry, and even topological curriculum changes (split-merge events). Only drift speed has a measurable effect, modulating c from $\sim$2.0 (fast drift) to $\sim$3.6 (slow drift).

4. *Confusion, not destruction.* Old memories collapse toward recent eras ($\bar{F} > 1$) rather than reverting to the prior ($\bar{F} \to 1$). This "confusion" regime is the dominant failure mode.
5. *Adaptive forgetting channel.* The decomposed metric identifies the active information channel: mean-dominated ($\sim$85%) when component means drift (synthetic), covariance-dominated when only weights vary (MNIST). Weight error is negligible for equal-weight mixtures.
6. *Movie replay.* The stochastic process reconstructed from the density path produces temporally coherent replay trajectories — compressed "movies" of the agent's history. On MNIST LeCun et al. (1998), the protocol grid decoded frame-by-frame produces a visual temporal narrative in which digit identities are preserved (Section 8).

Paper outline. Section 2 introduces the CAS recursion. Section 3 identifies forgetting-by-compression. Section 4 defines the forgetting metrics. Sections 5–7 report experiments. Section 9 discusses the capacity law and stochastic replay. Section 8 presents the MNIST illustration. Section 10 concludes. Appendix A derives the drift from the density path; Appendix B describes the software architecture.

2 The Compress–Add–Smooth Framework

The agent maintains a Bridge Diffusion (BD) process on the fixed replay interval $[0, 1]$ whose terminal marginal at $t = 1$ represents the current day, and whose intermediate-time marginals encode the past. Incorporating a new day is a three-step recursion — compress, add, smooth — carried out entirely within a chosen parameterised density class. The approach is generic: it applies to any density family for which piecewise-linear interpolation is well defined. We illustrate it in this paper on the Gaussian-Mixture (GM) class, where all operations reduce to linear algebra on the mixture parameters. The protocol grid is kept *uniform* at all times: after every daily update the node times are $\{0, 1/L, 2/L, \ldots, 1\}$. We achieve this by **compressing** at every time step the domain from $[0, 1]$ to $[0, L/(L+1)]$, then fitting the new day's experience in the newly **added** interval $[L/(L + 1), 1]$, and then **smoothing** the resulting $L + 1$ segments back to L segments.

Fig. 1 illustrates the recursion for $L = 4$.

2.1 Memory representation

At day n, the agent's memory consists of three objects:

(i) a *prior distribution* $q^{(0)} \in \mathcal{G}_K$, where $\mathcal{G}_K$ is the class of K-component Gaussian mixtures;

(ii) a *protocol grid*: $L + 1$ Gaussian-mixture states $\{G_j^{(n)}\}_{j=0}^{L}$, one at each node time $t_j = j/L$. Each $G_j^{(n)} \in \mathcal{G}_K$ is specified by weights $\pi_k^{(j)}$, means $m_k^{(j)} \in \mathbb{R}^d$, and covariances $\Sigma_k^{(j)} \in \mathbb{R}^{d\times d}$, $k = 1, \ldots, K$. Between adjacent nodes the density is defined by piecewise-linear interpolation of the GM parameters (see below);

(iii) a readout-time dictionary $\{t_{m|n}\}_{m=1}^{n}$, mapping each past day m to a query time in $(0, 1]$.

The total memory cost is $O(LKd^2)$ for the protocol grid ($L + 1$ nodes, each storing K means of size d and K covariance matrices of size d^2) plus $O(Kd^2)$ for the prior.

Piecewise-linear interpolation. For any $t \in [t_j, t_{j+1}]$, with $\alpha = (t - t_j)/(t_{j+1} - t_j)$, the marginal density is the Gaussian mixture with linearly interpolated parameters:

$$\pi_k(t) = (1 - \alpha)\, \pi_k^{(j)} + \alpha\, \pi_k^{(j+1)}, \quad m_k(t) = (1 - \alpha)\, m_k^{(j)} + \alpha\, m_k^{(j+1)}, \quad \Sigma_k(t) = (1 - \alpha)\, \Sigma_k^{(j)} + \alpha\, \Sigma_k^{(j+1)}. \quad (1)$$

This interpolation preserves the GM structure: for every t, the marginal is a valid K-component Gaussian mixture (weights sum to 1, covariances are positive definite by convexity). The corresponding SDE drift, needed only when sample paths are required, is reconstructed from the density path via Appendix A.

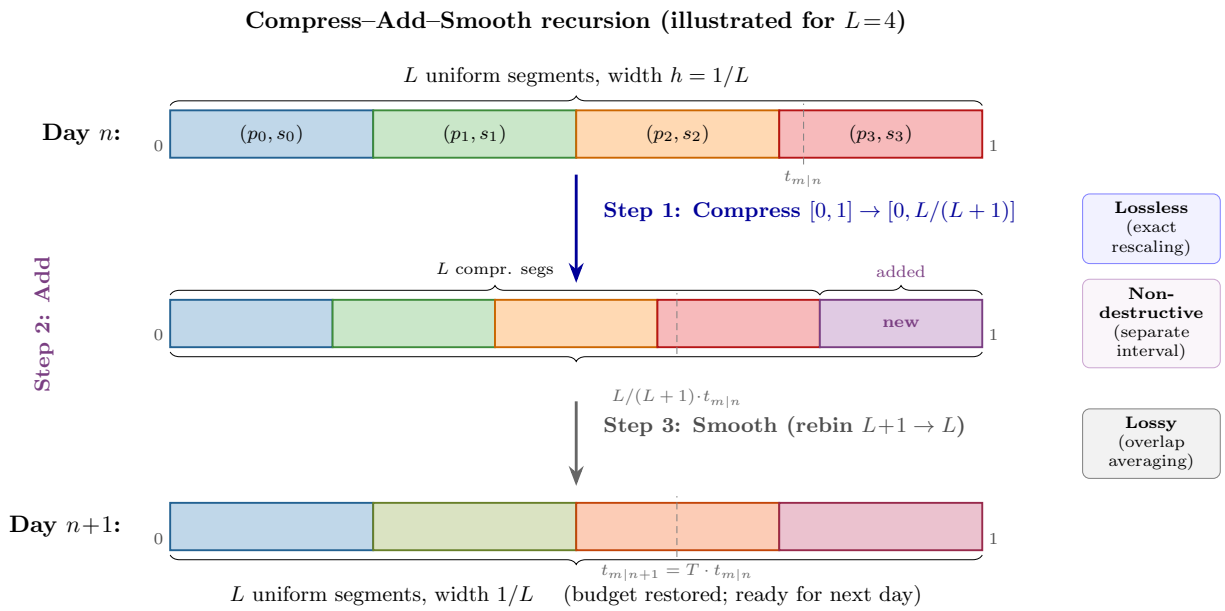

Figure 1: One iteration of the compress–add–smooth recursion, illustrated for $L = 4$ segments. **Top:** the protocol at day n consists of L uniform segments on $[0, 1]$. **Middle:** compression rescales the protocol to $[0, L/(L + 1)]$; the new day is appended on $[L/(L + 1)]$, producing $L+1$ uniform segments. **Bottom:** rebinning averages the $L+1$ segments back onto the L-segment grid. The right-hand labels indicate the information-theoretic role of each step: only smoothing is lossy. Dashed lines track a past-day readout time $t_{m|n}$, which contracts by factor $L/(L + 1)$ every day.

2.2 Initialisation (day 1)

The initial (day 1) protocol is set up by linearly interpolating from the prior distribution $q^{(0)}$ at $t = 0$ to the first day's target $q^{(1)}$ at $t = 1$:

$$p_t^{(1)} = (1 - t)q^{(0)} + tq^{(1)}, \tag{2}$$

where the linear combination acts on the GM parameters as in (1). The $L+1$ initial node states are obtained by evaluating this interpolant at the grid times $t_j = j/L$:

$$G_j^{(1)} = \big(1 - j/L\big)\, q^{(0)} + \big(j/L\big)\, q^{(1)}, \qquad j = 0, \ldots, L. \tag{3}$$

(A nonlinear interpolant can also be used.) The drift corresponding to the density path (2) is reconstructed via Appendix A.

2.3 Step 1: exact compression

The old protocol, defined on $L + 1$ nodes at times $\{0, 1/L, \ldots, 1\}$, is mapped exactly to the subinterval $[0, L/(L + 1)]$ by relabelling the node times:

$$G_j^{(n+1,\mathrm{cmp})} = G_j^{(n)}, \qquad \text{at time } t_j^{\mathrm{cmp}} = \frac{j}{L} \cdot \frac{L}{L+1} = \frac{j}{L+1}, \qquad j = 0, \ldots, L. \tag{4}$$

The GM states at each node are unchanged; only the time labels are rescaled. This is an exact, lossless operation: the compressed protocol defines the same density path, played at $L/(L + 1)$ speed.

2.4 Step 2: addition

A single new node is appended at $t = 1$ with state $q^{(n+1)}$ (the new day's target distribution). The compressed grid already has a node at $t = L/(L + 1)$ with state $G_L^{(n)}$ (the previous day's terminal marginal). Between

these two nodes the density is again defined by linear interpolation:

$$p_t^{(n+1)} = \frac{1-t}{1-L/(L+1)}\, G_L^{(n)} + \frac{t-L/(L+1)}{1-L/(L+1)}\, q^{(n+1)}, \qquad t \in \big[L/(L+1),\, 1\big]. \tag{5}$$

After addition, the augmented protocol has $L+2$ nodes at the uniform grid $\{0, \frac{1}{L+1}, \frac{2}{L+1}, \ldots, \frac{L}{L+1}, 1\}$, constituting $L+1$ segments of width $1/(L+1)$.

2.5 Step 3: smoothing by uniform-grid rebinning

The augmented protocol has $L+2$ nodes (constituting $L+1$ segments), but the budget allows only L segments ($L+1$ nodes). We restore the budget by *rebinning*: evaluating the augmented piecewise-linear density interpolant at the target grid and storing the resulting GM states as the new nodes.

Concretely, the augmented grid has nodes at $t_k^{\text{aug}} = k/(L+1)$, $k = 0, \ldots, L+1$, and the target grid has nodes at $t_j^{\text{new}} = j/L$, $j = 0, \ldots, L$. For each target node, we evaluate the augmented interpolant:

$$G_j^{\text{new}} = p_{t_j^{\text{new}}}^{\text{aug}}, \qquad j = 0, \ldots, L. \tag{6}$$

Since t_j^{new} falls inside some augmented segment $[t_k^{\text{aug}}, t_{k+1}^{\text{aug}}]$, the evaluation is a linear interpolation between two adjacent augmented nodes:

$$G_j^{\text{new}} = (1-\alpha_{jk})\, G_k^{\text{aug}} + \alpha_{jk}\, G_{k+1}^{\text{aug}}, \qquad \alpha_{jk} = \frac{t_j^{\text{new}} - t_k^{\text{aug}}}{t_{k+1}^{\text{aug}} - t_k^{\text{aug}}}, \tag{7}$$

where k is the unique index such that $t_k^{\text{aug}} \le t_j^{\text{new}} < t_{k+1}^{\text{aug}}$. Since the interpolation acts componentwise on the GM parameters (π, m, Σ), the result is a valid K-component Gaussian mixture at every node (weights are convex combinations summing to 1; covariances remain positive definite by convexity of the PSD cone).

Equivalently, the operation can be written as a matrix–vector product using a sparse *rebinning matrix* $W \in \mathbb{R}^{(L+1)\times(L+2)}$ whose rows encode the interpolation weights (7). Each row of W has at most two nonzero entries and sums to 1. The matrix depends only on L and is precomputed once.

The entire smoothing step requires $O(LKd^2)$ operations: for each of $L+1$ target nodes, interpolate the $K \times (d^2+d+1)$ GM parameters. No optimiser, no merge-pair selection, and no policy choice is needed.

2.6 Readout-time evolution

Readout times are updated only in the compression step:

$$t_{m|n+1} = \frac{L}{L+1} \cdot t_{m|n} \qquad \text{for all } m \le n, \qquad t_{n+1|n+1} = 1. \tag{8}$$

The smoothing step does not move readout times; it changes the node states of the protocol grid, so the marginal *at* the readout time changes — that is the forgetting mechanism. After $n-m$ days, the readout time of day m is

$$t_{m|n} = \Big(\frac{L}{L+1}\Big)^{n-m}, \tag{9}$$

which decays geometrically toward 0 with age. For $L = 10$, the readout time of a 20-day-old memory is $(L/(L+1))^{20} \approx 0.12$, placing it in the leftmost 12% of the replay interval.

2.7 Computational cost

Per-day update: $O(LKd^2)$. Compression relabels $L+1$ node times ($O(1)$ per node; the GM states are unchanged). Addition appends one new node and evaluates one interpolation ($O(Kd^2)$). Smoothing evaluates the augmented interpolant at $L+1$ target nodes ($O(Kd^2)$ per node), giving $O(LKd^2)$ total. No backpropagation, no sampling, no optimiser.

Per-replay query: $O(Kd^2)$. The replay marginal at readout time $t_{m|n}$ is obtained by evaluating the piecewise-linear interpolant (1): locate the enclosing segment ($O(\log L)$ or $O(1)$ with a uniform grid), then interpolate K means (Kd operations) and K covariance matrices (Kd^2 operations).

Memory footprint: For $d = 8$, $K = 3$, $L = 20$: the protocol occupies $21 \times 3 \times (64 + 8 + 1) = 4599$ floats $\approx 37\,\text{kB}$ in double precision, plus $\sim$1.8 kB for the prior. No stored data, no replay buffer.

3 Forgetting-by-compression

In standard continual learning, forgetting arises from *parameter interference* McCloskey and Cohen (1989); French (1999): gradient updates on new data overwrite representations needed for old tasks. In our framework, the three steps have distinct information-theoretic roles:

- *Compression* is **lossless** — it is an exact time-rescaling that preserves the marginal flow.
- *Addition* is **non-destructive** — the new day occupies a separate interval $[L/(L+1), 1]$ and does not modify the old protocol on $[0, L/(L+1)]$.
- *Smoothing* is **lossy** — rebinning replaces a finer grid by a coarser one, erasing sub-grid temporal detail.

Forgetting is therefore *localised in a single identifiable step*: the re-approximation of an (L+1)-segment protocol by an L-segment protocol via interpolant evaluation on a coarser grid. The temporal resolution available for old memories shrinks geometrically with age through the readout-time decay (9), making forgetting a consequence of *temporal coarse-graining* rather than parametric interference.

Remark 1 (Temporal blurring of node states)**.** *Each rebinning cycle replaces node states by convex combinations of their neighbors, progressively smoothing the spatial variation along the protocol. Older (leftward) nodes have undergone more rebinning cycles and their GM parameters are therefore more blurred — component means are pulled toward a common average, covariances are inflated, and weight contrasts are reduced. This cumulative blurring is the microscopic mechanism behind the macroscopic forgetting curve. It can serve as a diagnostic: when leftward nodes become nearly indistinguishable, the memory is close to saturation.*

4 Forgetting metrics

We use moment-based metrics as the primary forgetting diagnostics throughout this paper. They are cheap to evaluate, analytically transparent, and sufficient for the GM class. For richer density families (e.g. neural parameterisations), distributional metrics such as KL divergence or Wasserstein-2 distance would be natural alternatives; we leave their systematic study to future work.

4.1 Raw moment mismatch

The replay distribution of past day m at current day n is $\widehat{p}^{(m|n)} = p^{(n)}_{t_{m|n}}$, the marginal of the current protocol evaluated at the readout time via (1). The raw forgetting metric is

$$F_{m\to n} \;=\; \|\mu_{\text{replay}} - \mu_{\text{orig}}\|^2 + \|\Sigma_{\text{replay}} - \Sigma_{\text{orig}}\|_F^2, \tag{10}$$

where (μ, Σ) are the overall mean and covariance of the Gaussian mixture, computed analytically from the GM parameters.

Rather than studying the full (m, n) matrix, we work primarily with the *age variable* $a = n - m$ and define the age-dependent forgetting curve

$$\bar{F}(a) \;=\; \left\langle \bar{F}_{m\to n} \right\rangle_{n-m=a} \tag{11}$$

as the average over all pairs with $n - m = a$.

4.2 Normalised metric

To compare across K, d, and geometric scale, we normalise by the *amnesia baseline*:

$$F_{\text{amnesia}}(m) = \|\mu_0 - \mu_{\text{orig}}^{(m)}\|^2 + \|\Sigma_0 - \Sigma_{\text{orig}}^{(m)}\|_F^2, \tag{12}$$

where (μ_0, Σ_0) are the moments of the starting distribution (for a deterministic start: $\mu_0 = x_0$, $\Sigma_0 = 0$). The normalised forgetting is

$$\bar{F}_{m\to n} = \frac{F_{m\to n}}{F_{\text{amnesia}}(m)} \in [0, \infty). \tag{13}$$

Here $\bar{F} = 0$ is perfect recall, $\bar{F} = 1$ is total amnesia, and $\bar{F} > 1$ indicates *confusion* — the replay is actively worse than having no memory at all. The retention half-life is $a_{1/2} = \min\{a : \bar{F}(a) \geq \theta\}$ with threshold $\theta = 0.5$.

Remark 2 (Confusion: $\bar{F} > 1$)**.** *When $\bar{F} > 1$, old memories have been pulled toward the current day's location rather than decaying toward the prior. We call this regime* confusion *to distinguish it from* destruction *($\bar{F} \to 1$, reversion to the uninformed prior).*

4.3 Decomposed metric for Gaussian mixtures

For $K > 1$, we decompose forgetting into per-component contributions after Hungarian matching:

$$\begin{aligned} F &= F_{\text{mean}} + F_{\text{cov}} + F_{\text{weight}}, \\ F_{\text{mean}} &= \textstyle\sum_k \bar{w}_k \|\Delta m_k\|^2, \quad F_{\text{cov}} = \sum_k \bar{w}_k \|\Delta\Sigma_k\|_F^2, \quad F_{\text{weight}} = \|\Delta\pi\|^2, \end{aligned} \tag{14}$$

where $\bar{w}_k = \max(\pi_k^{\text{replay}}, \pi_k^{\text{orig}})$ and matching is by pairwise mean distance.

5 Experiments: single-Gaussian ($K = 1$, $d = 2$)

5.1 Setup

We consider a stream of $n = 100$ daily Gaussian targets

$$p^{(m)} = \mathcal{N}\big(\mu^{(m)}, \Sigma\big), \qquad m = 1, \ldots, n,$$

in dimension $d = 2$, with fixed covariance

$$\Sigma = 0.5\, I.$$

Unless stated otherwise, the daily means follow a circular drift of radius $R = 2$,

$$\mu^{(m)} = R\big(\cos(2\pi m/P), \sin(2\pi m/P)\big), \qquad P = 50,$$

so that over the 100-day horizon the mean completes two full revolutions. The prior is $q^{(0)} = \mathcal{N}(0, I)$.

The default segment budget is

$$L = 10.$$

The circular drift is a deliberately nontrivial geometry. It is simple enough to visualize, but unlike a monotone linear drift it periodically revisits earlier spatial locations. This makes it possible to separate two effects: genuine temporal forgetting and geometric aliasing caused by revisiting the same region of state space at different times. To assess the role of geometry, we also compare against a linear-drift experiment in which the daily means move along a line at comparable local speed.

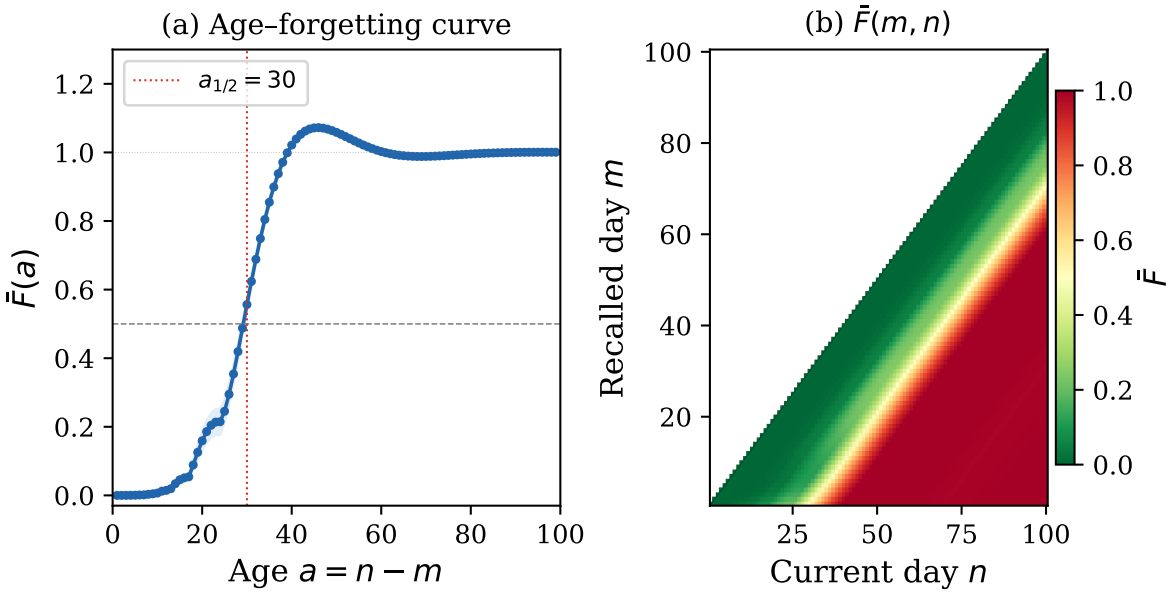

Figure 2: Single-Gaussian experiment ($K = 1$, $d = 2$, $L = 10$) under the default circular-drift setting. (a) Age-averaged normalized forgetting $\bar{F}(a)$, showing retention half-life $a_{1/2} = 30$. The curve exhibits a low-error plateau for ages 0–15, followed by a steep sigmoid transition crossing $\bar{F} = 0.5$ at age 30, a slight overshoot to $\bar{F} \approx 1.08$ around age 50 (the confusion regime), and eventual saturation near $\bar{F} = 1.0$. The curve is weakly non-monotone due to the periodic geometry of the circular drift, which causes geometric recurrence at multiples of the half-period. (b) Full forgetting matrix $\bar{F}(m, n)$ as a function of recalled day m and current day n. The dominant trend is age-controlled forgetting (iso-forgetting contours run parallel to the diagonal), modulated by the periodic geometry of the underlying drift.

5.2 Default behavior: age curve, heatmap, and replay geometry

Fig. 2 shows the basic forgetting diagnostics for the default parameters. Panel (a) reveals a characteristic two-regime structure: recent memories ($a \lesssim 15$) are recalled with near-zero error, while older memories undergo a rapid sigmoid-like degradation. The half-life $a_{1/2} = 30$ means that, with $L = 10$ segments, the agent retains useful recall of the past ∼30 days. The slight overshoot $\bar{F} > 1$ in the age range 40–60 confirms the confusion phenomenon (Remark 2): old replayed means are pulled toward the current day's location rather than decaying to the prior. Panel (b) shows that the dominant structure of the forgetting matrix is age-controlled, with periodic modulation visible as faint stripes at multiples of the half-period $P/2 = 25$.

Fig. 3(a) visualizes the replayed means at the final day $n = 100$. Recent memories are replayed close to their true locations on the right-lower arc of the circle, whereas older memories are displaced toward a compressed cluster near the origin. This spatial collapse is the geometric signature of confusion: old replayed means are attracted toward the time-weighted average of the protocol, which is dominated by recent days.

Fig. 4 makes the confusion mechanism visible: as age increases, the replayed mean migrates inward from the true location on the circle toward the origin (the time-averaged protocol centre), while the replayed covariance inflates dramatically. The arrows connecting original to replayed positions show that the displacement is systematically directed toward the protocol interior, not random.

Fig. 3(b) shows the readout times $t_{m|n}$ versus current day n. They decay geometrically as $(L/(L+1))^{n-m}$, with the theoretical curves (dashed) overlaid for reference. The actual and theoretical curves coincide exactly, confirming the readout-time evolution (9). For $L = 10$, a 30-day-old memory sits at $t \approx (10/11)^{30} \approx 0.047$, deep in the leftward portion of the protocol where rebinning-induced blurring is most severe.

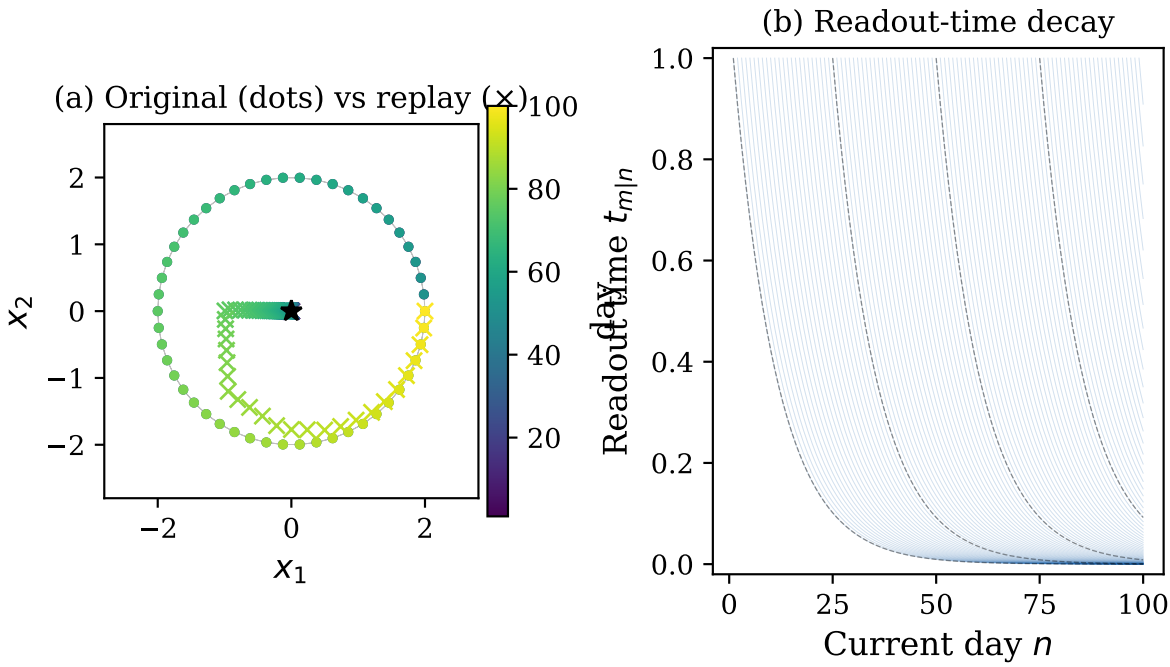

Figure 3: Default single-Gaussian circular-drift experiment ($L = 10$). (a) Original daily means (dots, coloured by day) and replayed means at the final day (crosses). The black star marks the prior mean (the origin), which serves as the protocol's long-time attractor. Confusion is visible as the systematic inward displacement of crosses from the circle toward the star: recent memories (warm colours) are replayed near their true locations on the circle, while older memories (cool colours) are pulled progressively toward the prior mean. This convergence of replay means toward the star — rather than remaining on the circle — is the geometric signature of confusion. (b) Readout times $t_{m|n}$ versus current day n, showing the geometric decay (9).

5.3 Parameter dependence

The segment budget L is the primary determinant of retention. Sweeping $L \in \{5, 8, 10, 15, 20, 30\}$ yields half-lives $a_{1/2} \in \{14, 24, 30, 44, 51, 74\}$, scaling roughly as $a_{1/2} \approx 2.4\,L$ (Fig. 5). This near-linear scaling is consistent with the observation that each CAS cycle degrades the readout time by a factor $L/(L+1)$, so the number of cycles before a memory reaches a fixed resolution threshold is proportional to L.

Drift speed modulates the half-life: faster drift (shorter period P) leads to shorter retention, because larger daily displacements accumulate more error through rebinning. Sweeping $P \in \{25, 50, 100, 200\}$ yields $a_{1/2} \in \{20, 30, 34, 36\}$. The dependence saturates for slow drift ($P \geq 100$), suggesting a floor set by the diffusive contribution of the rebinning itself.

Drift geometry (circular vs. linear) affects the curve shape more than the half-life: linear drift yields a clean monotone sigmoid with $a_{1/2} = 42$ (vs. 30 for circular at the same L), while circular drift introduces non-monotone modulations due to periodic spatial recurrence. The higher linear-drift half-life reflects the absence of geometric aliasing: each recalled location is unique, so the rebinning error is always genuine.

5.4 Takeaway

The $K = 1$ experiments establish three main results. First, the forgetting curve has a universal two-regime shape (plateau + sigmoid) whose transition is controlled by the segment budget L, with $a_{1/2} \approx 2.4\,L$. This is the first observation of the linear retention-capacity law, which we will confirm across progressively more complex settings: multi-component mixtures (Section 6), crowding and dimension scaling (Section 7), and image-derived latent spaces (Section 8). Second, drift speed modulates the half-life but drift geometry

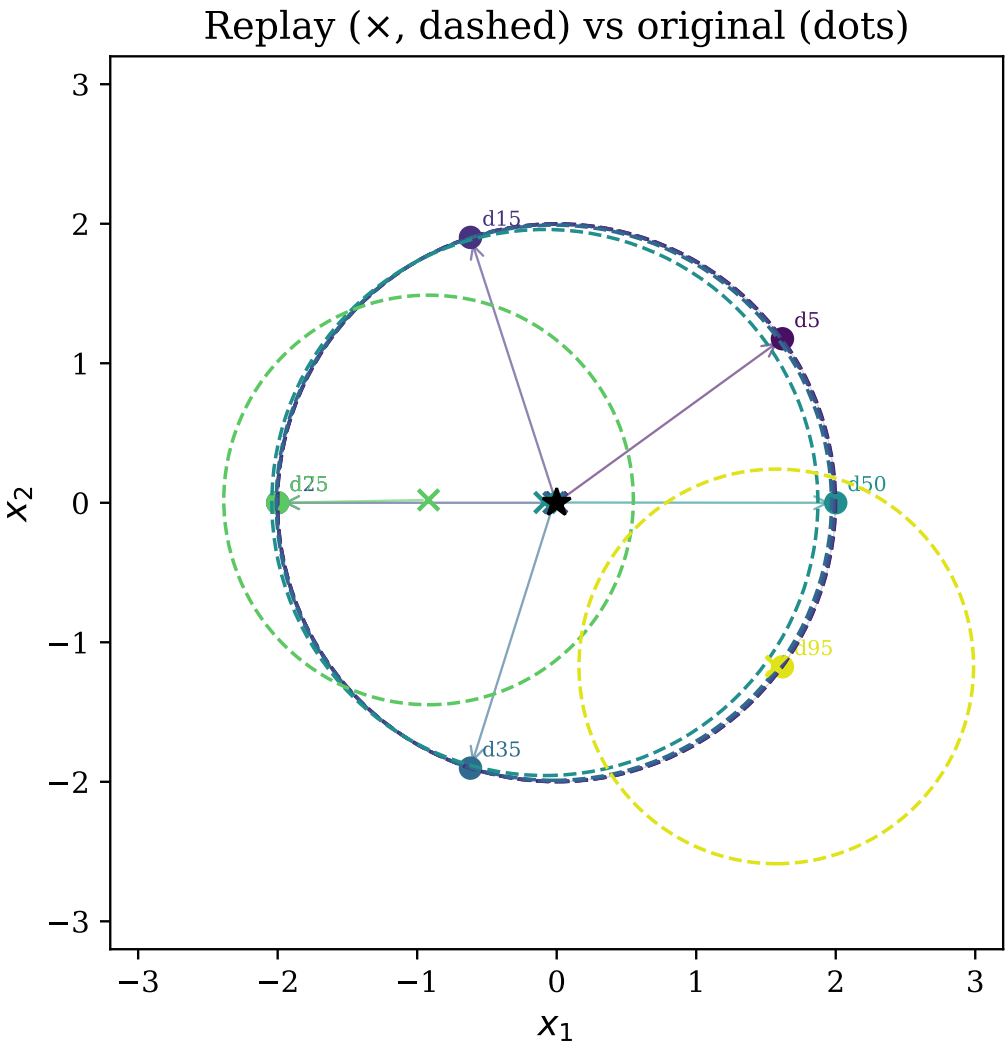

Figure 4: Selected replay ellipses in the default single-Gaussian circular-drift experiment ($L = 10$). For each displayed day, the original target mean is shown by a dot, while the replayed mean is shown by a cross with a dashed ellipse representing the replay covariance. Recent memories (e.g. day 95) are replayed with small displacement and compact ellipses. Intermediate-age memories (e.g. day 35) show large displacement toward the origin and inflated covariances. Very old memories (e.g. day 25) collapse nearly to the origin with very large ellipses. Geometric aliasing is visible for day 5, whose true location on the circle lies close to day 95 (they differ by two full periods), producing an apparently accurate replay that is coincidental rather than genuine recall.

affects only the curve shape. Third, forgetting manifests as confusion (displacement toward recent eras), not destruction (reversion to the prior). These findings motivate the $K > 1$ experiments below, where we test whether state-space complexity affects the retention timescale.

6 Experiments: Gaussian mixtures ($K = 3$, $d = 2$)

We now extend to K-component Gaussian-mixture daily targets. Each day's distribution has $K = 3$ equal-weight components arranged in a rotating equilateral triangle of radius $r = 0.8$ around the drifting circle centre (same circular drift as Section 5, with per-component covariance $0.3\,I$).

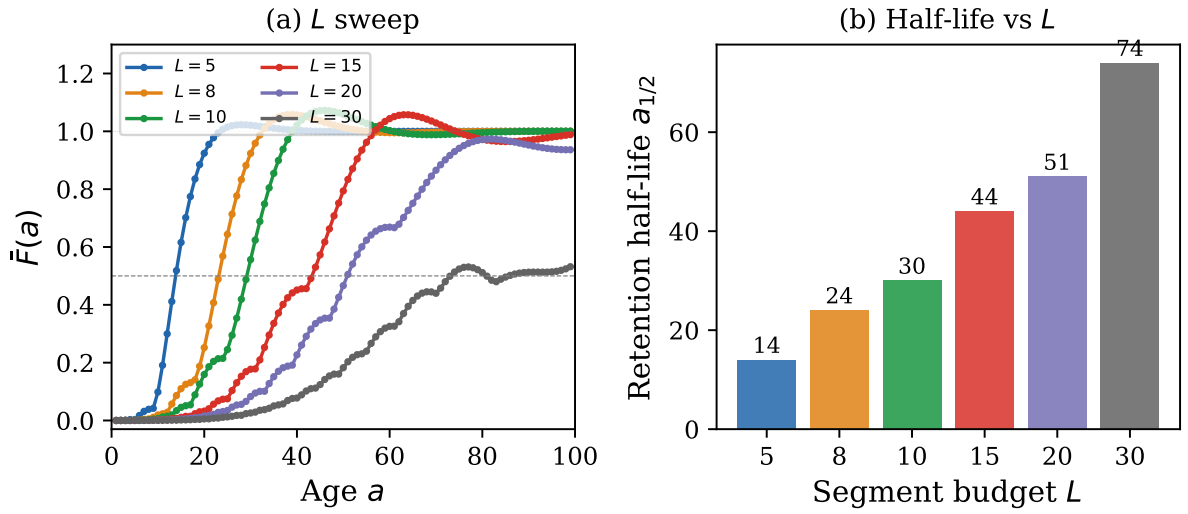

Figure 5: Segment-budget sweep for the single-Gaussian circular-drift experiment. (a) Age–forgetting curves for $L \in \{5, 8, 10, 15, 20, 30\}$. Increasing L shifts the sigmoid transition to higher ages without changing the curve shape qualitatively. (b) Retention half-life $a_{1/2}$ versus L, confirming approximate linear scaling.

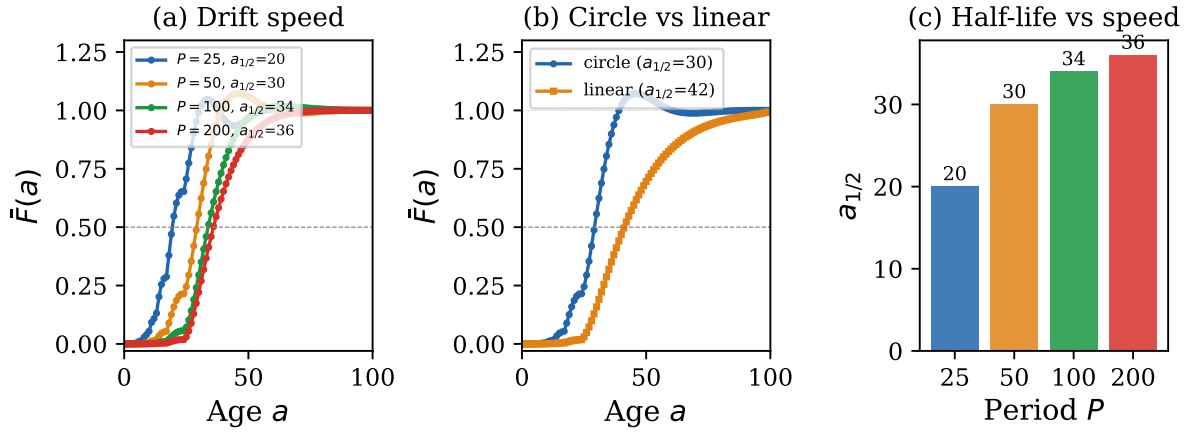

Figure 6: (a) Drift-speed sweep (circle period P). Faster drift shortens the half-life, but the effect saturates for slow drift. (b) Circle vs. linear drift. The circular curve is non-monotone due to periodic spatial recurrence; the linear curve is a clean sigmoid with longer half-life. (c) Half-life versus period P.

6.1 Default run and decomposed forgetting

With $L = 10$, the $K = 3$ experiment yields $a_{1/2} = 30$ — *identical* to the $K = 1$ case (Fig. 7a). This is the first indication that retention is governed by the temporal budget L rather than the state-space complexity K.

The decomposed forgetting (Fig. 7b) reveals that F_{mean} dominates ($\sim$85% of total raw forgetting), F_{cov} contributes $\sim$15%, and F_{weight} is negligible (of order 10^{-17}, i.e. machine precision). The vanishing weight error is a structural consequence of equal-weight mixtures: convex combinations of equal weights remain equal, so the rebinning preserves weights exactly.

6.2 Component-level trajectories

Fig. 8 shows the per-component replayed means (after Hungarian matching) at the final day. Recent days' component means are replayed accurately; older days collapse toward the protocol interior, with all three

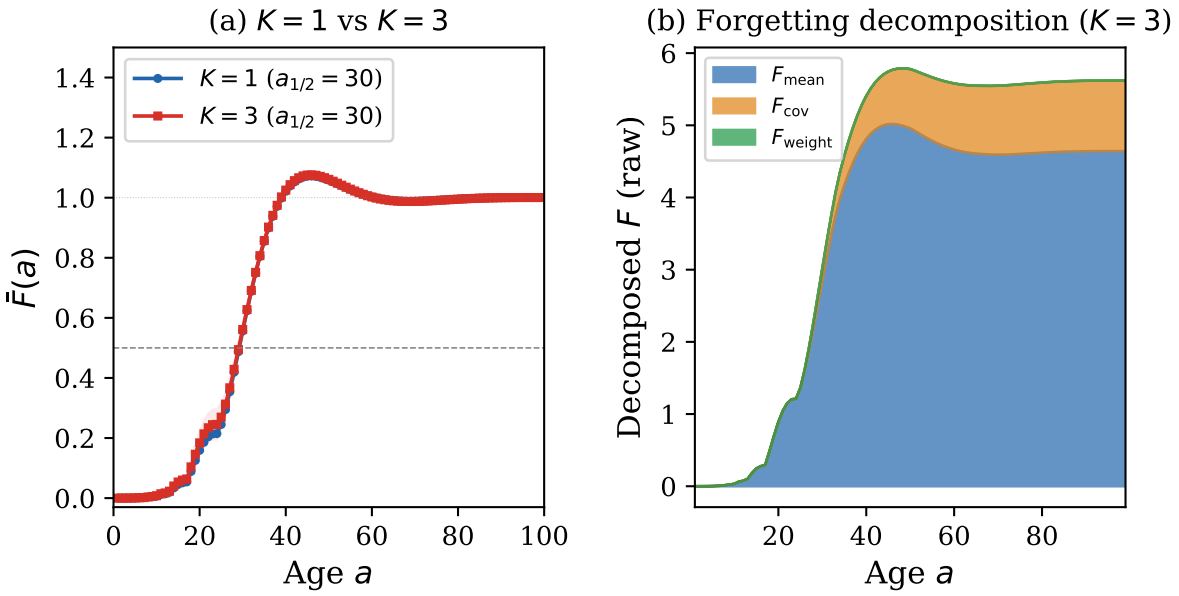

Figure 7: (a) Age–forgetting curves for $K = 1$ (blue) and $K = 3$ (red), both with $L = 10$ and circular drift. The curves nearly coincide, with both yielding $a_{1/2} = 30$. (b) Decomposed raw forgetting for $K = 3$: mean misalignment (blue) dominates, covariance error (orange) is secondary, and weight error (green) is negligible.

components converging toward a common cluster near the origin. This mirrors the $K = 1$ confusion pattern, amplified by the need to simultaneously track three interacting trajectories.

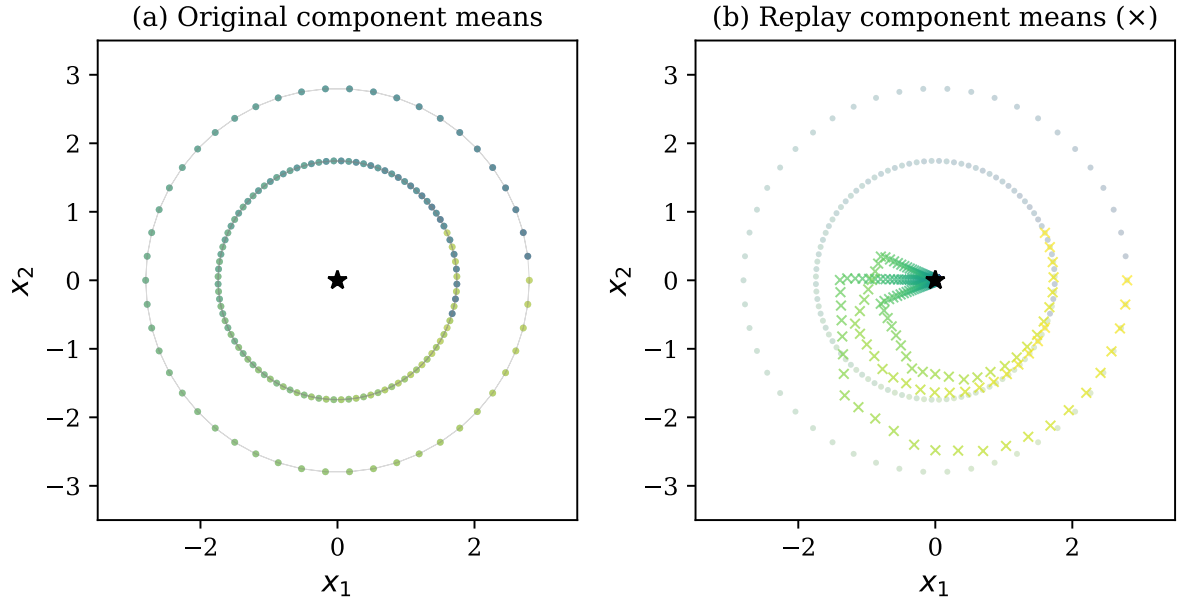

Figure 8: Component-level trajectories for $K = 3$. (a) Original daily component means, coloured by day. The three interleaved helical paths trace the rotating-triangle geometry. (b) Replayed component means (×) at the final day, after Hungarian matching to the originals. Recent components are well-recalled; older ones collapse toward the origin.

6.3 L sweep and K sweep

Sweeping the segment budget at $K = 3$ gives half-lives $a_{1/2} \in \{14, 30, 41, 50, 71\}$ for $L \in \{5, 10, 15, 20, 30\}$ (Fig. 9a), closely matching the $K = 1$ results.

The K sweep at $L = 10$ (Fig. 9b) yields $a_{1/2} \in \{30, 29, 30, 30, 30\}$ for $K \in \{1, 2, 3, 5, 8\}$ — the half-life is essentially *flat* across mixture complexity. This is the paper's central experimental finding: *retention is controlled by the temporal budget L, not by the state-space complexity K*. The K-independence is not approximate: the half-life varies by at most one day across a factor-of-8 range in K.

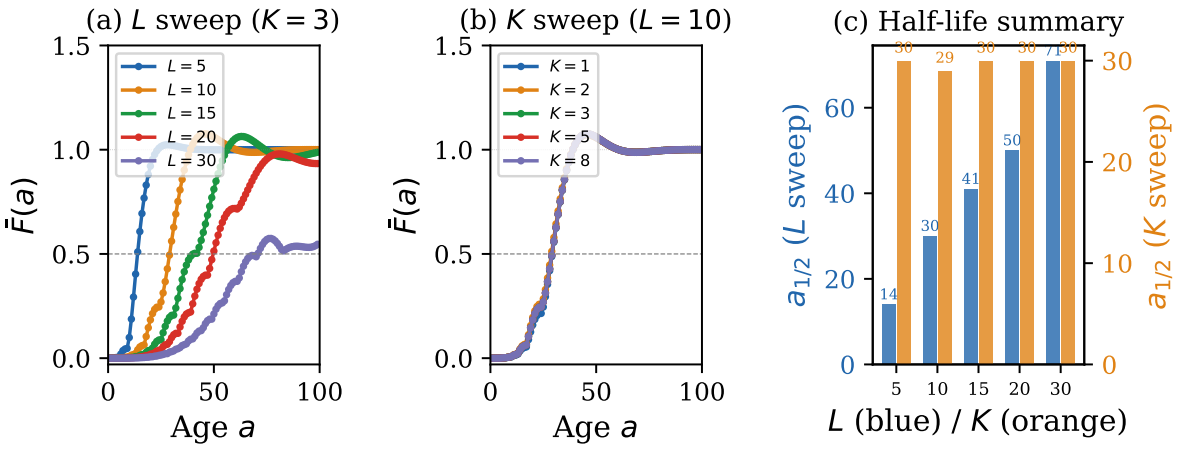

Figure 9: (a) L sweep at $K = 3$: age–forgetting curves for $L \in \{5, 10, 15, 20, 30\}$. The pattern mirrors the $K = 1$ case. (b) K sweep at $L = 10$: age curves for $K \in \{1, 2, 3, 5, 8\}$ nearly coincide. (c) Half-life summary for both sweeps.

6.4 Takeaway

The $K > 1$ experiments establish two main results. First, the half-life is independent of K: adding mixture components does not shorten (or lengthen) retention. This is because the rebinning step treats all GM parameters (weights, means, covariances) uniformly — the interpolation does not "see" how many components there are. Second, forgetting is overwhelmingly driven by mean misalignment; covariance error is secondary and weight error is negligible for equal-weight mixtures. These findings justify using the half-life $a_{1/2}$ as a single scalar summary of retention quality, controlled by L alone.

7 Scaling experiments

We now test how the continual memory mechanism scales when the daily targets become more crowded, when the relevant signal is embedded into a higher-dimensional ambient space, and when the target family undergoes a simple topological curriculum involving split-and-merge events. The goal of this section is not to optimize performance, but to identify which aspects of increasing problem complexity actually shorten retention and which do not.

Based on the K-independence result of Section 6 — specifically, the flat half-life across $K \in \{1, 2, 3, 5, 8\}$ — we conjectured that the same qualitative picture would persist: forgetting is governed primarily by temporal compression under a fixed protocol budget, while many forms of static state-space complexity affect the geometry of replay far more than the retention timescale itself. The experiments below confirm this conjecture under three increasingly challenging scenarios.

7.1 Crowding as a control parameter

We begin with mixtures in $d = 2$, varying the *crowding ratio* $\chi = r/\sigma$, where r is the inter-component offset radius and $\sigma = \sqrt{\text{cov_scale}}$ is the component standard deviation. Small χ corresponds to heavily overlapping components (strong crowding); large χ to well-separated components (weak crowding). We sweep $r \in \{0.15, 0.3, 0.5, 0.8, 1.2, 2.0\}$ at $K = 3$, corresponding to $\chi \in \{0.27, 0.55, 0.91, 1.46, 2.19, 3.65\}$. Fig. 10 summarizes the results.

Panel (a) shows retention half-life $a_{1/2}$ versus crowding ratio for $K = 2, 3, 5, 8$. The first observation is that all curves are *flat at* $a_{1/2} = 30$ for $\chi \lesssim 1.5$: moderate-to-strong crowding has no effect on retention whatsoever. Only at high separation ($\chi > 2$) does the half-life begin to decrease, dropping to $a_{1/2} \approx 20$ at $\chi = 3.65$. The effect is most pronounced for $K = 2$, whose half-life begins declining earlier (at $\chi \approx 1$) than for $K \geq 3$. This decline at large χ is a geometric effect: when components are widely separated, each component's mean displacement under rebinning is larger in absolute terms, accelerating forgetting.

Panel (b) shows the age–forgetting curves at $K = 3$ for six crowding values. The curves for $\chi \leq 1.5$ are nearly indistinguishable, all showing the standard sigmoid with $a_{1/2} = 30$. At $\chi = 2.2$ the half-life shortens slightly to 28, and at $\chi = 3.7$ it drops to 20, with the sigmoid onset shifting leftward and the confusion overshoot ($\bar{F} > 1$) increasing.

Panel (c) reports the average share of raw forgetting attributable to mean misalignment. The mean-error share is ∼90% across all crowding ratios, confirming that even as crowding changes the spatial geometry of forgetting, the dominant error channel remains mean displacement rather than covariance distortion or weight drift.

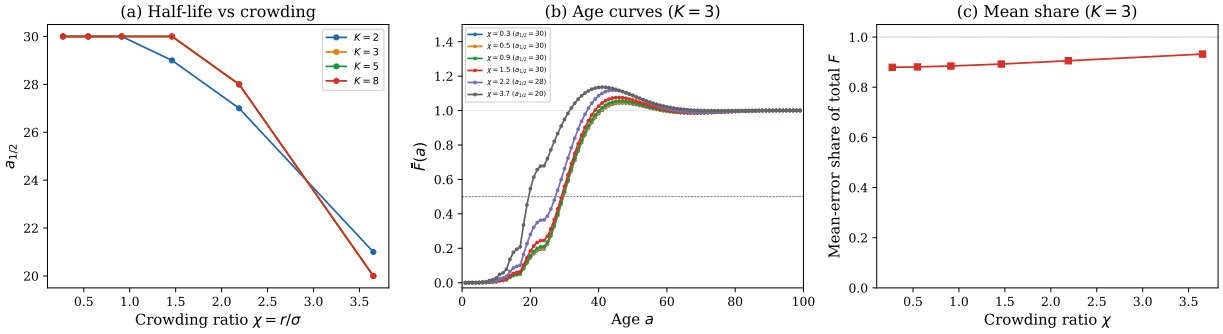

Figure 10: Crowding sweep for Gaussian-mixture memories ($L = 10$). (a) Retention half-life $a_{1/2}$ versus crowding ratio χ for $K = 2, 3, 5, 8$. The half-life is flat at 30 for $\chi \lesssim 1.5$ and declines only for well-separated components. (b) Age–forgetting curves at $K = 3$ for six crowding values. (c) Mean-error share of total forgetting, stable at ∼90% across all χ.

To illustrate the shape of forgetting more directly, Fig. 11 shows three representative age-forgetting curves corresponding to strong ($\chi = 0.3$), medium ($\chi = 0.9$), and weak ($\chi = 2.2$) crowding. The most notable feature is that the strong and medium crowding curves are virtually identical ($a_{1/2} = 30$ for both), while the weakly crowded case shows a slightly earlier sigmoid onset ($a_{1/2} = 28$) and a more pronounced confusion overshoot ($\bar{F} \approx 1.12$ vs. ≈ 1.05). The overshoot amplification at weak crowding is consistent with larger per-component mean displacements when components are far apart.

7.2 Fixed low-dimensional signal in a higher-dimensional ambient space

We next test whether retention degrades when the informative signal remains two-dimensional but is embedded into a higher-dimensional ambient space. Fig. 12 reports the results for ambient dimensions $d = 2, 4, 8, 16$, with $K = 3$ and $L = 10$.

Panel (a) shows the age-forgetting curves when the extra dimensions carry no drift (nuisance coordinates remain at zero). The curves shift *rightward* with increasing d: the half-life increases slightly from $a_{1/2} = 30$

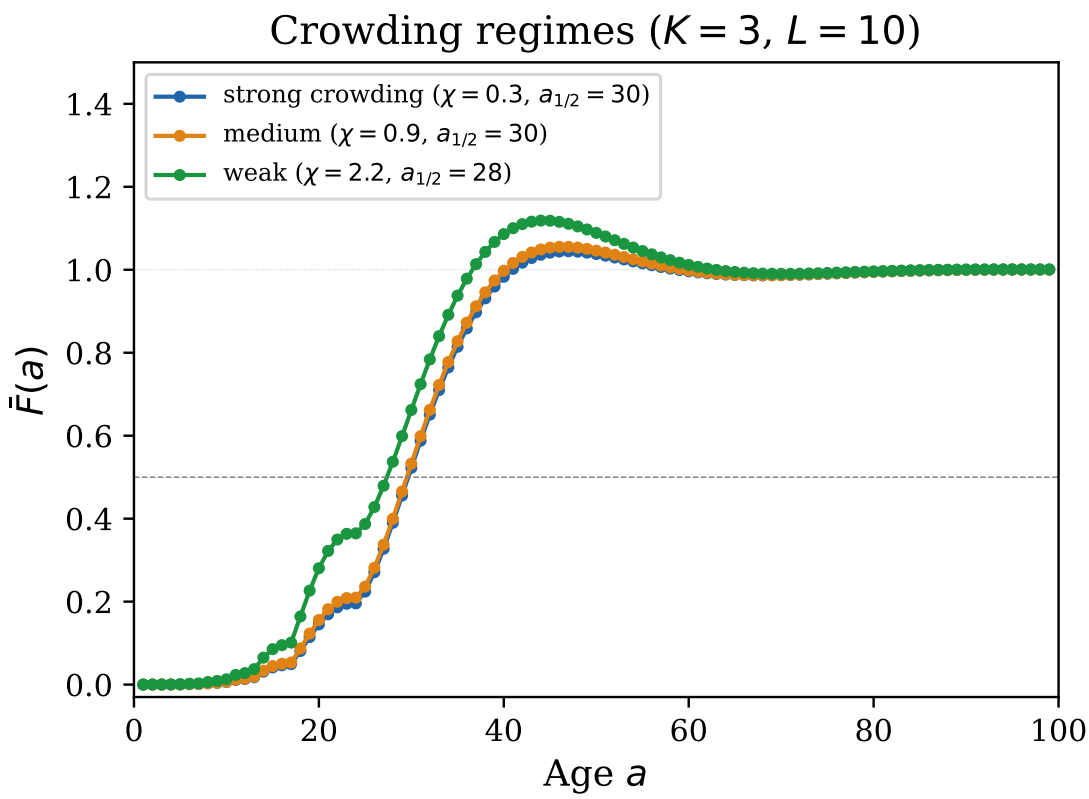

Figure 11: Representative age-averaged forgetting curves across three crowding regimes ($K = 3$, $L = 10$). Strong ($\chi = 0.3$) and medium ($\chi = 0.9$) crowding yield indistinguishable curves ($a_{1/2} = 30$). Weak crowding ($\chi = 2.2$) produces a slightly earlier transition ($a_{1/2} = 28$) and a more pronounced confusion overshoot.

at $d = 2$ to $a_{1/2} = 34$ at $d = 16$. This counter-intuitive improvement occurs because the amnesia baseline F_{amnesia} grows with d (the prior covariance is I_d, contributing more Frobenius-norm distance from each daily target), while the rebinning error in the signal subspace is unchanged. The normalised forgetting is therefore diluted by the larger baseline.

Panel (b) summarizes the half-life as a function of d for two settings. When nuisance dimensions are static, $a_{1/2}$ increases gently from 30 to 34. When nuisance dimensions carry a slow random walk (speed 0.1/day), the half-life follows a similar trend ($a_{1/2} \approx 30$–33), indicating that moderate nuisance drift does not substantially impair retention of the signal.

Panel (c) shows that the mean-error share declines with d, from ∼90% at $d = 2$ to ∼60% at $d = 16$ in the no-nuisance setting. This shift reflects the growing contribution of covariance mismatch in the extra dimensions: as d increases, the $d \times d$ covariance matrices carry more entries that can accumulate rebinning error. With nuisance drift, the mean-error share remains higher (∼70% at $d = 16$) because the drifting nuisance means contribute additional mean-channel error.

7.3 Split-and-merge curriculum

As a final scaling test, we consider a simple curriculum in which the daily mixture geometry changes topologically over time via split-and-merge events. The $K = 3$ mixture undergoes four phases, illustrated schematically in Fig. 13: a normal rotating triangle ($r = 0.8$, days 1–30), a merge phase where two components collapse toward each other ($r_{01} \to 0.05$, days 31–50), a split phase where they separate again ($r \to 0.8$, days 51–80), and a final collapse where all three components converge toward the centre ($r \to 0.1$, days

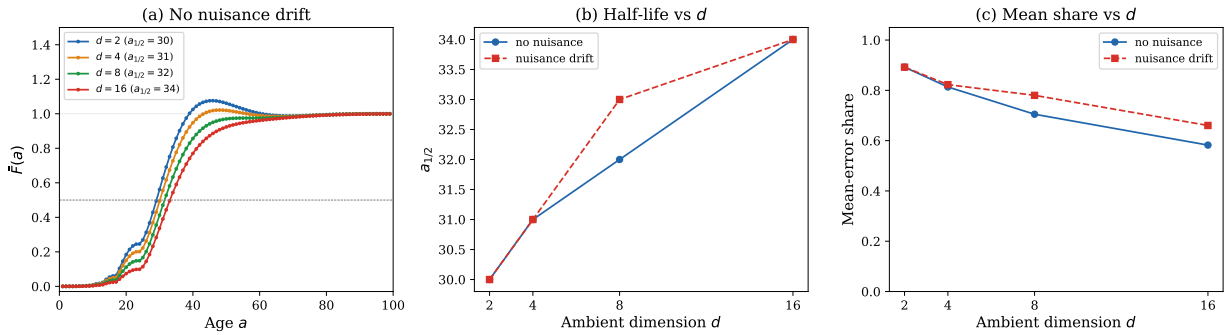

Figure 12: Scaling with ambient dimension for a fixed 2D signal ($K = 3$, $L = 10$). (a) Age-forgetting curves for $d = 2, 4, 8, 16$ without nuisance drift. Higher d slightly *improves* normalised retention due to the larger amnesia baseline. (b) Retention half-life versus ambient dimension under two nuisance settings. Both show gentle increase with d. (c) Mean-error share decreases with d as covariance mismatch grows in the extra dimensions.

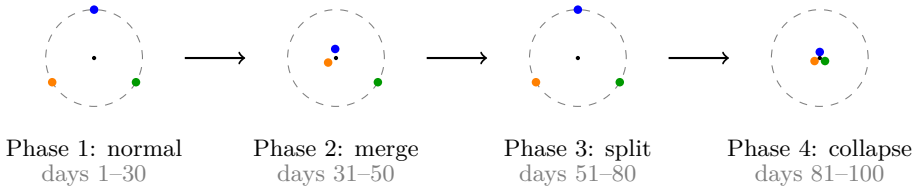

Figure 13: Schematic of the four-phase split-and-merge curriculum for $K = 3$. Coloured dots represent the three mixture component means; the dashed circle indicates the inter-component radius r. Phase 1: normal rotating triangle ($r = 0.8$). Phase 2: two components merge ($r_{01} \to 0.05$). Phase 3: split back to triangle ($r \to 0.8$). Phase 4: all three collapse toward the centre ($r \to 0.1$). Transitions are smoothed over 5-day ramps.

81–100). Transitions are smoothed over 5-day ramps. Fig. 14 shows both the daily component means and the resulting age-forgetting curve.

Panel (a) displays the component centres across the 100 days, with phase-boundary markers (red diamonds) at days 1, 31, 51, and 81. The four phases are clearly visible: the initial rotating triangle, the merged pair, the re-separation, and the final collapse.

Panel (b) shows the corresponding age-forgetting curve. Despite the nontrivial topological evolution, the half-life is $a_{1/2} = 30$ — identical to the stationary-geometry baseline. The curve shape is the standard sigmoid with a mild non-monotone feature around age 10–15, attributable to the interaction between curriculum transitions and the periodic drift geometry.

This result is the strongest evidence that *the retention timescale is set by the temporal budget L alone*: even when the daily target distribution undergoes qualitative structural changes — merging, splitting, and collapsing of mixture components — the half-life is unaffected.

7.4 Overview and interpretation

The three scaling experiments paint a consistent picture. Crowding affects the half-life only at extreme separation ($\chi > 2$, where per-component mean displacements become large), and even then the reduction is modest (from 30 to 20). Ambient dimension either has no effect or slightly *improves* normalised retention (due to the growing amnesia baseline), while shifting the forgetting channel from mean-dominated toward a more even mean/covariance split. A time-varying curriculum with topological changes leaves the half-life entirely unchanged.

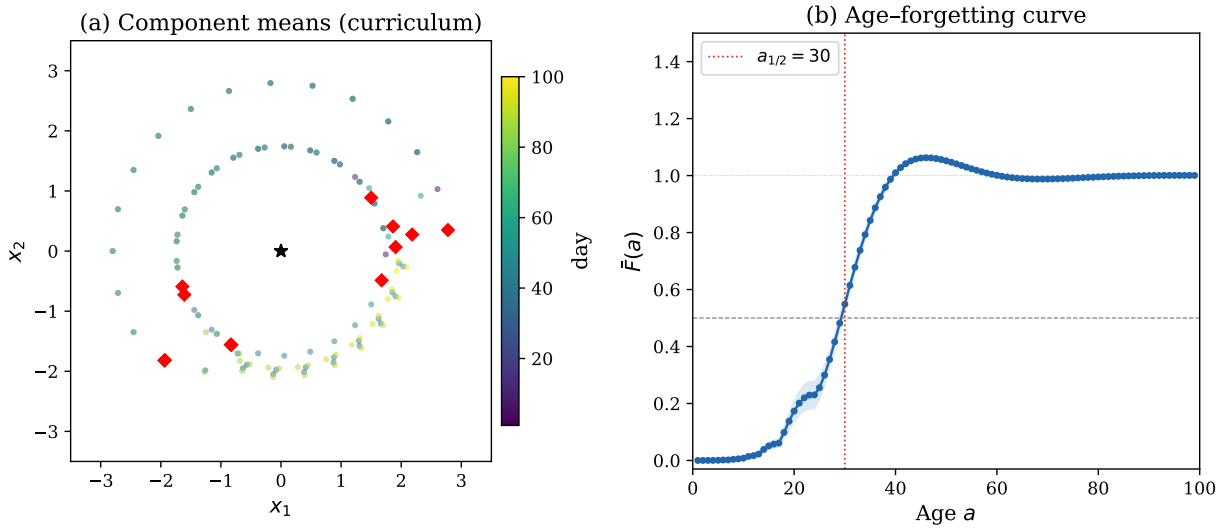

Figure 14: Split-and-merge curriculum experiment ($K = 3$, $L = 10$). (a) Daily component means, coloured by day, with phase-boundary markers (red diamonds) at days 1, 31, 51, and 81. The four phases — normal triangle, merge, split, collapse — are clearly visible. (b) Age-averaged forgetting curve. Despite the nontrivial curriculum, the retention half-life remains $a_{1/2} = 30$, identical to the stationary baseline.

These results confirm the conjecture from Section 6: the retention half-life $a_{1/2} \approx 2.4\,L$ is a robust, universal characteristic of the CAS recursion under uniform-grid rebinning. It depends on the temporal budget L and, to a lesser extent, on the drift speed, but is insensitive to the state-space complexity K, the ambient dimension d, the crowding geometry, and even topological changes in the daily target family. The only avenue for substantially improving retention, within the current framework, is to increase L or to replace the uniform grid with an adaptive one that allocates finer temporal resolution to recent memories.

8 MNIST latent-space illustration

To complement the analytically controlled Gaussian-mixture experiments, we construct an image-based latent-space illustration using MNIST. The purpose is twofold: (i) to test whether the same notions of age-dependent forgetting, confusion, and retention-time control carry over when the GM components represent real image classes; and (ii) to demonstrate the "movie" capability described in Section 9.2 — the protocol grid, decoded frame-by-frame to pixel space, produces a visual temporal narrative of the agent's compressed history.

8.1 Setup: latent embeddings and rotating-dominance curriculum

We select three visually distinct MNIST digit classes — 0, 3, and 8 — and embed the corresponding ∼18,000 training images into a $d = 12$ PCA latent space (57% explained variance). At this dimension, PCA-decoded class centroids are clearly recognisable as their respective digits (Fig. 15).

We fit a single Gaussian per class ($K = 3$ total) and construct a *rotating-dominance curriculum* over $n = 100$ days: the component means and covariances are fixed to their class-conditional fits, while the mixing weights rotate with period $P = 30$:

$$\pi_k^{(m)} = \mathrm{softmax}\big(A\,\cos(2\pi m/P + 2\pi k/3)\big), \qquad A = 2, \tag{15}$$

so that each digit class cycles between dominance ($\pi_k \approx 0.9$) and near-absence ($\pi_k \approx 0.04$). This is the semantic analogue of the synthetic circular drift: the "location" in distribution space rotates through digit classes rather than through spatial coordinates.

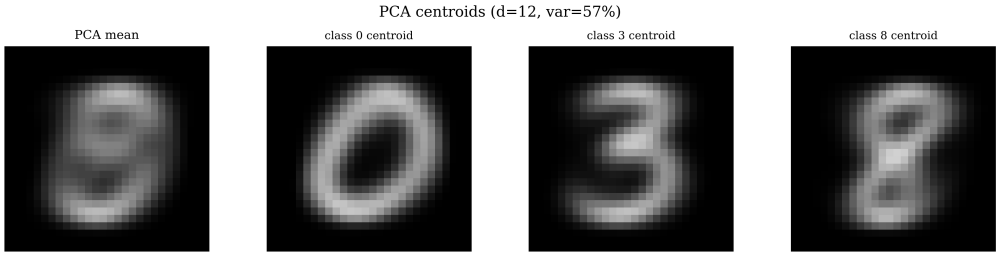

Figure 15: PCA centroids at $d = 12$: the global mean (left) and per-class centroids for digits 0, 3, and 8. Despite capturing only 57% of total variance, the centroids are visually recognisable.

8.2 Forgetting curve and comparison with synthetic experiments

Running CAS with $L = 10$ yields a retention half-life of $a_{1/2} = 37$ (Fig. 16). The age–forgetting curve exhibits the familiar two-regime structure — a low-error plateau followed by a sigmoid transition — with no confusion overshoot ($\bar{F}$ saturates at 1 rather than exceeding it). The absence of overshoot is explained by the nature of the daily variation: since only the weights change (not the component means), the replayed means for old days converge toward a time-averaged centroid rather than being actively displaced past it.

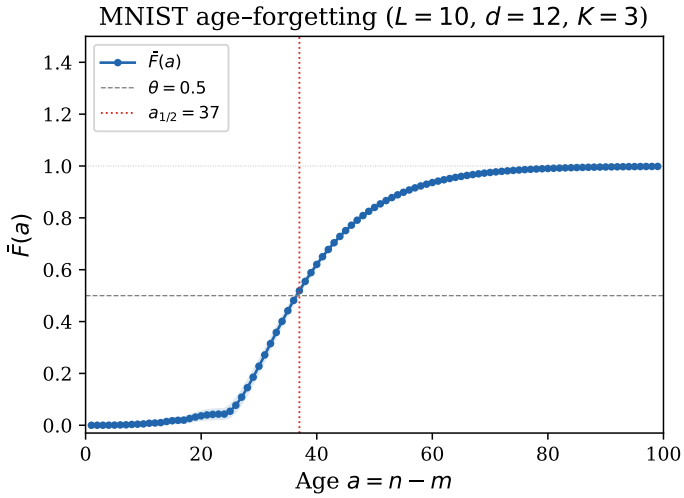

Figure 16: MNIST age–forgetting curve ($L = 10$, $d = 12$, $K = 3$). The half-life $a_{1/2} = 37$ and the curve shape is a clean sigmoid without the confusion overshoot seen in the synthetic mean-drift experiments.

Fig. 17 compares the MNIST and synthetic $K = 3$ forgetting curves at the same $L = 10$ and $P = 30$. The MNIST half-life ($a_{1/2} = 37$) exceeds the synthetic one ($a_{1/2} = 21$). Two effects contribute: (i) the higher latent dimension $d = 12$ inflates the amnesia baseline, diluting the normalised metric; and (ii) the MNIST curriculum perturbs only the weights (a K-dimensional vector), while the synthetic curriculum moves all K component means through $\mathbb{R}^2$, producing larger per-step rebinning error.

8.3 Decomposed forgetting: covariance-dominated regime

The forgetting decomposition (Fig. 18) reveals a qualitative difference from the synthetic experiments. In the MNIST construction, F_{cov} dominates the raw forgetting, accounting for the overwhelming majority of the total, while F_{mean} is comparatively small and F_{weight} contributes visibly but remains secondary. This is the opposite of the synthetic case (where $F_{\text{mean}} \approx 85\%$) and is explained by the design of the curriculum: since component means are fixed, the mean channel accumulates minimal rebinning drift; instead, the $d \times d$ covariance matrices ($12 \times 12 = 144$ entries per component) accumulate Frobenius-norm error as

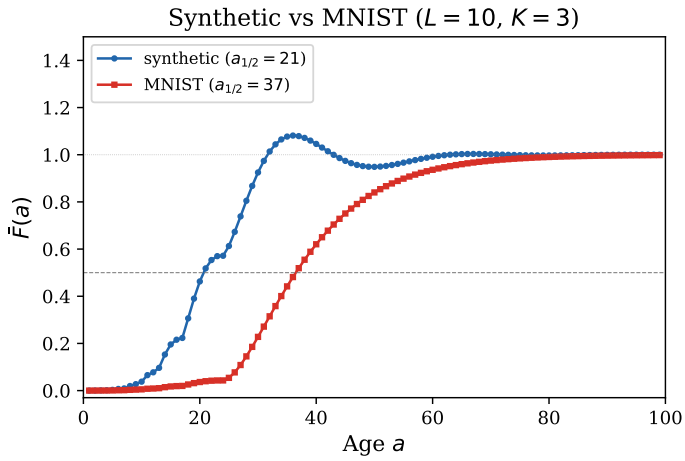

Figure 17: Comparison of synthetic $K = 3$ (blue, $d = 2$, circle drift, $P = 30$) and MNIST (red, $d = 12$, rotating-dominance weights, $P = 30$) age–forgetting curves at $L = 10$. The MNIST curve is shifted rightward due to the higher dimension and milder daily perturbation.

the piecewise-linear interpolation progressively distorts the class-specific covariance structure. The weight error is non-negligible here because the rotating weights are the primary information channel, unlike the synthetic equal-weight setting.

The raw forgetting also exhibits periodic oscillations at large age (visible in Fig. 18), with period $P = 30$ matching the curriculum. The mechanism is straightforward: with $K = 3$ classes cycling with period $P = 30$, days separated by exactly 30 (or 60, 90, ...) share the same dominant digit class. When such a day is recalled, the replayed weight vector happens to be closer to the original (since both have the same class dominant), producing a dip in raw forgetting. Days at half-period offsets ($a = 15, 45, \ldots$) have maximally mismatched weight vectors and produce forgetting peaks. This resonance effect is purely a consequence of the periodic curriculum and has no analogue in the synthetic mean-drift experiments, where the daily dynamics are continuous rather than weight-modulated.

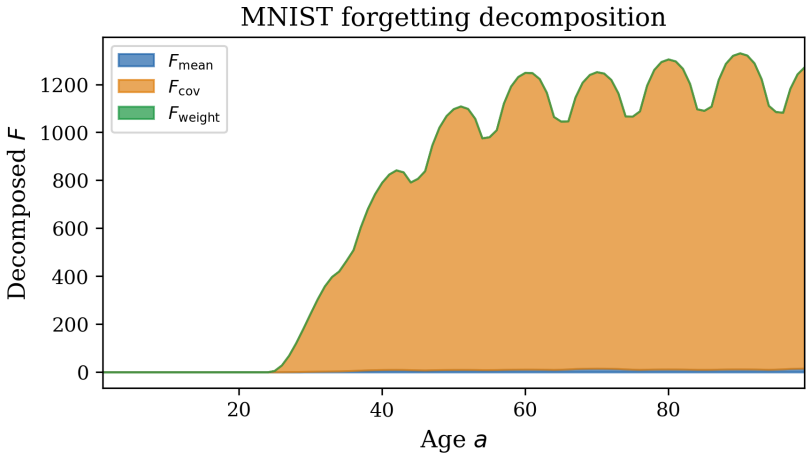

Figure 18: Decomposed raw forgetting for MNIST. Unlike the synthetic experiments where F_{mean} dominates, the MNIST construction is covariance-dominated because component means are fixed and only the weights rotate. The periodic oscillations at large age have period $P = 30$, matching the curriculum: dips occur at ages that are multiples of P (where the recalled and current days share the same dominant digit class), while peaks occur at half-period offsets.

8.4 Visual forgetting and the temporal movie

The key diagnostic of the MNIST experiment is visual: decoded images reveal how forgetting manifests in pixel space.

Fig. 19 shows the per-component replayed means (decoded to 28×28 pixels) for eight selected past days. For recent days (d90, d99), the three components decode to clearly distinct, recognisable digits (0, 3, 8). As age increases, the components blur and converge: by day 25 and earlier, all three components decode to a similar ambiguous shape resembling the PCA grand mean. This is the visual manifestation of confusion — not semantic collapse (which would mean instant failure), but progressive loss of class identity through cumulative rebinning.

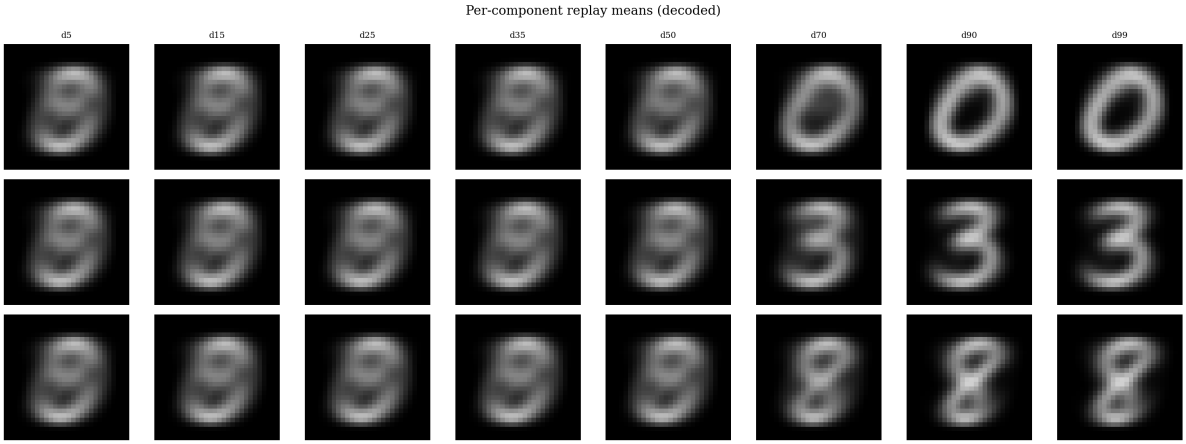

Figure 19: Per-component replayed means decoded to pixel space, for selected past days. Each row corresponds to one digit class (0, 3, 8 from top to bottom). Recent days show distinct digit identities; older days progressively converge toward a common blurred average.

The protocol grid, evaluated at uniformly spaced times $t \in [0, 1]$ and decoded frame-by-frame, produces a *temporal movie* of the agent's compressed history. Fig. 20 shows the per-component movie strip: each row tracks one digit class's mean through the full protocol. Remarkably, all three digit identities are maintained across the entire $0 \to 1$ interval — digit 0 remains recognisably 0, digit 3 remains 3, digit 8 remains 8 — even at the oldest portion ($t \approx 0$) of the protocol. The visual degradation is primarily in sharpness and contrast rather than class identity.

Figure 20: Component movie: per-class decoded means from $t = 0$ (oldest memories, left) to $t = 1$ (current day, right). Each row is one digit class. Digit identities are preserved across the full protocol, with only gradual loss of sharpness at the oldest end.

The protocol weight evolution (Fig. 21) shows the compressed temporal history of the rotating curriculum. At $t = 1$ (current day), digit 8 dominates ($\pi_8 \approx 0.9$). Moving leftward: digit 0 peaked around $t \approx 0.4$, digit 3 around $t \approx 0.2$, and the oldest memories ($t \approx 0$) have roughly equal weights. This weight trajectory is a lossy but recognisable compression of the full 100-day curriculum.

8.5 Dimension sweep

Sweeping the PCA dimension $d \in \{4, 8, 12, 20, 30\}$ yields half-lives $a_{1/2} \in \{36, 37, 37, 37, 37\}$ (Fig. 22) — essentially flat across all dimensions tested. No semantic collapse occurs at any d. This confirms the

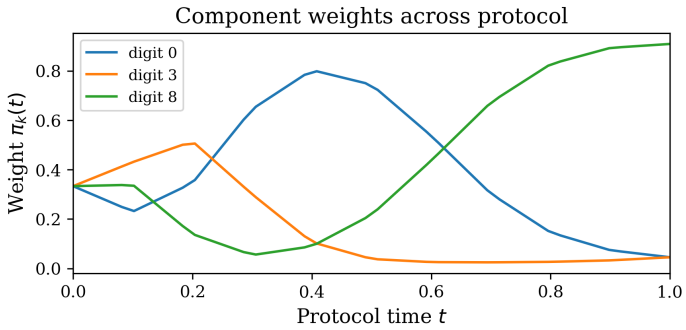

Figure 21: Component weights $\pi_k(t)$ across the protocol grid. The weight trajectory encodes a compressed history of the 100-day rotating-dominance curriculum: recent dominance of digit 8 (green at $t = 1$), earlier dominance of digit 0 (blue peak at $t \approx 0.4$), and convergence toward uniform weights at the oldest end.

dimension-independence observed in the synthetic scaling experiments (Section 7) and demonstrates that the CAS framework handles real image-derived latent spaces as robustly as synthetic ones.

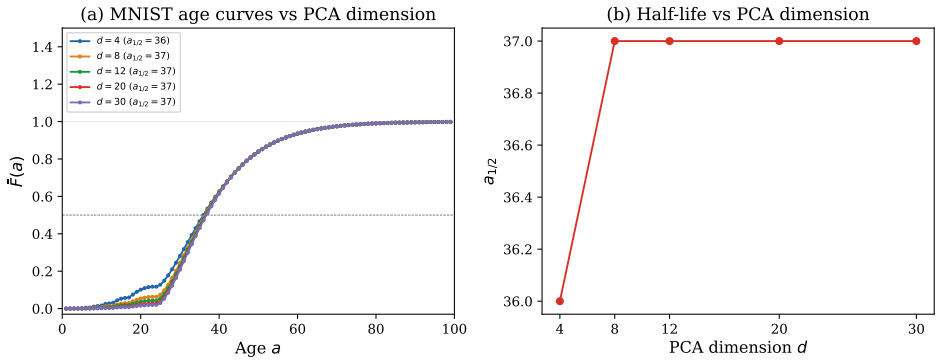

Figure 22: MNIST PCA dimension sweep. (a) Age–forgetting curves for $d \in \{4, 8, 12, 20, 30\}$ nearly coincide. (b) Half-life is flat at $a_{1/2} \approx 37$ across all dimensions.

8.6 Takeaway

The MNIST experiment demonstrates four results. First, the CAS framework transfers successfully from synthetic to image-derived latent spaces: the two-regime forgetting curve and $a_{1/2} \approx c\,L$ scaling persist. Second, the dominant forgetting channel shifts from mean-dominated (synthetic, where means drift) to covariance-dominated (MNIST, where means are fixed and only weights rotate) — the framework correctly identifies the active information channel in each case. Third, the protocol grid serves as a genuine temporal movie: decoded frame-by-frame, it produces a visual narrative in which digit identities are preserved while mixing proportions evolve smoothly from the agent's oldest memories to its most recent experience. Fourth, there is no critical dimension for semantic collapse: the half-life is flat from $d = 4$ to $d = 30$, confirming that temporal compression — not representational capacity — is the binding constraint on retention.

9 Discussion

We now discuss two cross-cutting themes that emerge from the full suite of experiments: the information-theoretic structure of the $a_{1/2} \approx c\,L$ law, and the role of the stochastic process underlying the bridge as a mechanism for temporally coherent replay.

9.1 Retention capacity and the $a_{1/2} \approx c\,L$ law

The empirical law $a_{1/2} \approx 2.4\,L$, first observed in the $K = 1$ experiments (Section 5) and confirmed unchanged across mixture complexity (Section 6), crowding, dimension, curriculum (Section 7), and MNIST (Section 8), deserves closer examination.

Why $c > 1$ matters. A First-In-First-Out (FIFO) buffer — the simplest baseline, which stores the last L daily distributions verbatim and discards the oldest upon each new arrival — provides perfect recall for L days and instant amnesia thereafter, giving $a_{1/2} = L$ exactly. The CAS scheme achieves $a_{1/2} \approx 2.4\,L$ — a factor $c \approx 2.4$ improvement — despite using the same $O(LKd^2)$ storage. The gain arises because the piecewise-linear interpolant between GM nodes implicitly encodes information about *intermediate* days that do not sit on any grid node. A readout at time $t_{m|n}$ between two nodes returns a meaningful blend of the flanking node states, carrying compressed but non-trivial information about the original day-m target. The bridge is performing lossy compression, but it is a *smooth* compression whose interpolation structure extracts more than one "effective day" of retention per grid node.

Where the factor c comes from. The origin of c can be understood from the readout-time geometry. For large L, the compression ratio per step is $L/(L+1) \approx e^{-1/L}$, so after a days a memory's readout time has decayed to $t_{m|n} \approx e^{-a/L}$. Forgetting sets in when $t_{m|n}$ drops below a critical threshold t_* at which the cumulative rebinning error crosses the forgetting criterion $\theta = 0.5$. This gives $a_{1/2} \approx L\,\ln(1/t_*)$, identifying $c = \ln(1/t_*)$. For $c \approx 2.4$, we get $t_* \approx 0.09$ — consistent with the observation that a 30-day-old memory at $L = 10$ sits at $t \approx 0.047$ (well past the threshold), while a 20-day-old memory sits at $t \approx 0.13$ (just above it).

The threshold t_* depends on the drift speed: faster drift raises t_* and lowers c (the speed sweep gives $c \approx 2.0$ at $P = 25$ to $c \approx 3.6$ at $P = 200$). It does not depend on K, d, or the geometry of the target family — explaining the remarkable universality of the linear law across all our experiments.

Information-theoretic interpretation. The linear law $a_{1/2} = c\,L$ has a natural information-theoretic reading. The protocol grid with L nodes is a fixed-capacity "channel" of $O(LKd^2)$ real numbers; each daily incorporation injects $O(Kd^2)$ numbers. The maximum retention is therefore $O(L)$, establishing that linear scaling is fundamental. The constant c quantifies how efficiently the encoding utilises the available capacity — analogous to the capacity constant in Shannon's noisy-channel coding theorem[1]. The current uniform-grid scheme achieves $c \approx 2.4$; the question is how close an optimised scheme can get to the theoretical maximum c_{opt}.

Three concrete optimisation avenues are:

- *Non-uniform (logarithmic) grids.* Placing nodes at times $t_j = e^{-j\alpha}$ instead of j/L would allocate finer resolution to recent memories and directly increase c.
- *Variational rebinning.* Optimising the node placement at each step to minimise KL divergence from the augmented protocol would define c_{opt} operationally.
- *Non-linear interpolation.* Wasserstein geodesics between nodes could preserve more geometric structure through rebinning.

We conjecture that for any CAS-type scheme with L grid nodes and a stationary source with characteristic drift rate v, the retention half-life satisfies

$$a_{1/2} \;\leq\; c_{\text{opt}}(v) \cdot L, \tag{16}$$

where $c_{\text{opt}}(v)$ is a source-dependent constant. Determining c_{opt} is an open problem connecting the CAS framework to rate-distortion theory.

[1]The noisy-channel coding theorem Richardson and Urbanke (2008) establishes that reliable communication over a noisy channel is possible at any rate below the channel capacity C, but not above. In our setting, the "channel" is the L-node protocol grid corrupted by rebinning noise, and c plays the role of C: it is the maximum number of effective retention days per grid node achievable by any CAS-type encoding. The connection to rate-distortion theory Richardson and Urbanke (2008) is even more direct: the CAS recursion trades off temporal resolution (rate) against forgetting quality (distortion) under a fixed memory budget.

9.2 The role of stochastic replay

The protocol grid stores a density path $p_t(x)$ for $t \in [0, 1]$. Appendix A reconstructs a drift $s_t(x)$ such that the SDE

$$dX_t = s_t(X_t)\, dt + dW_t \tag{17}$$

has marginal density exactly p_t. This construction is never needed during the daily CAS update — it is invoked only at "inference time" when sample paths are requested — but it provides qualitative capabilities that go beyond evaluating marginal densities at readout times.

Replay as a movie. Sampling $X_0 \sim p_0$ and integrating (17) forward to $t = 1$ generates a continuous stochastic trajectory that visits, in temporal order, compressed representations of older memories (small t), progresses through intermediate eras, and arrives at the current day ($t = 1$). Each realisation is a different plausible "narrative" connecting the agent's past to its present. The MNIST experiment (Section 8) makes this literal: the protocol grid decoded frame-by-frame produces an actual visual movie of the agent's compressed digit-class history.

Temporal coherence across readout times. Two independent evaluations of p_t at different readout times yield statistically independent marginal samples. In contrast, a single sample path $\{X_t\}_{0 \le t \le 1}$ produces *correlated* samples at multiple readout times — the replay at day m and day m' come from the same trajectory and are therefore dynamically consistent. This temporal coherence is essential for downstream tasks that require more than pointwise recall.

Connection to sleep replay. The SDE generates compressed temporal sequences with stochastic variation — structurally analogous to hippocampal replay during sleep, where the brain replays compressed experience sequences to consolidate memory González et al. (2020); Golden et al. (2022). In this analogy, the protocol grid is the memory substrate, the SDE integration is the replay episode, and the diffusion noise dW_t corresponds to the variability across replay episodes.

Cost separation. A key design feature is the separation between the cheap CAS loop ($O(LKd^2)$ per day, no sampling) and the expensive SDE integration (needed only on demand). Evaluating $s_t(x)$ requires computing $\nabla \log p_t(x)$ and, for time-varying weights, the Poisson correction $\nabla \psi_t(x)$ from (26) — a cost of $O(Kd^2)$ per evaluation point per time step. For microcontroller-class hardware, the daily update runs in real time while movie generation can be deferred to periods of low computational load — a natural parallel to the sleep/wake dichotomy in biological memory consolidation.

9.3 Relation to prior work

Catastrophic interference McCloskey and Cohen (1989) – or catastrophic forgetting French (1999) – in sequentially trained networks has motivated four main CL paradigms: regularization (EWC Kirkpatrick et al. (2017), SI Zenke et al. (2017)), replay (deep generative replay Shin et al. (2017), brain-inspired replay van de Ven et al. (2020)), architecture expansion (progressive nets Rusu et al. (2016)), and compression (Progress & Compress Schwarz et al. (2018)) — all address forgetting-by-interference in shared-parameter models. Our framework is fundamentally different: forgetting arises from temporal coarse-graining rather than parameter overwriting, and the forgetting mechanism is localised in a single identifiable step (rebinning) rather than distributed across gradient updates. Progress & Compress Schwarz et al. (2018) is closest in spirit (it also separates a "knowledge base" from an "active column"), but relies on neural distillation rather than analytical density operations. Variational Continual Learning Nguyen et al. (2018) maintains a sequential posterior, formally similar to our compress–add step; however, it requires gradient-based updates and does not provide a closed-form forgetting analysis. For surveys of the CL landscape, see Parisi et al. (2019); De Lange et al. (2022); Wang et al. (2024).

Within the replay paradigm, recent work replaces the VAE/GAN generator of Shin et al. (2017) with a denoising diffusion model, achieving higher-fidelity replay samples for class-incremental classification Gao and Liu (2023); Jodelet et al. (2023); Meng et al. (2024), object detection Kim et al. (2024), federated

learning Liang et al. (2024), industrial streaming data He et al. (2024), and anomaly detection Hu et al. (2025). Our approach is structurally distinct from all of these: in diffusion-based generative replay, the diffusion model is a generator of past data samples, while the classifier (or RL agent) that actually does the learning is a separate network whose parameters are still updated by gradient descent and still subject to forgetting-by-interference. In the CAS framework, by contrast, the bridge diffusion is the memory — there is no separate generator and no gradient-based forgetting. The SDE protocol replaces the replay buffer entirely.

Our bridge diffusion is constructed by prescribing a density path and recovering the SDE drift from the Fokker–Planck equation (Appendix A). This approach is related to Schrödinger bridges Léonard (2013); Chen et al. (2021); De Bortoli et al. (2021), flow matching Lipman et al. (2023), stochastic interpolants Albergo and Vanden-Eijnden (2023) and Path-Integral Diffusion Behjoo and Chertkov (2025); Chertkov and Behjoo (2025); Chertkov (2025; 2026), but differs in that the density path is specified directly as a piecewise-linear interpolant, rather than optimised or learned.

In the neuroscience literature, Bazhenov and collaborators González et al. (2020); Golden et al. (2022); Tadros et al. (2022); Golden et al. (2025); Vins et al. (2025) show that sleep-like off-line replay prevents catastrophic forgetting by pushing synaptic weights toward joint solution manifolds. Our SDE-based replay (Section 9.2) is structurally analogous, with the CAS protocol playing the role of the synaptic substrate and the SDE integration playing the role of the replay episode.

10 Conclusions and Path Forward

We introduced the Compress–Add–Smooth (CAS) framework for continual learning, in which an agent's temporal memory is encoded as a Bridge Diffusion process on a fixed replay interval $[0, 1]$. The framework is parameterised by two budgets: a state budget K (mixture complexity) and a temporal budget L (protocol segments). Incorporating a new day costs $O(LKd^2)$ flops with no backpropagation, no stored data, and no neural networks.

The key experimental findings, for the Gaussian-mixture instantiation, are:

1. *Two-regime forgetting curve.* The normalised forgetting $\bar{F}(a)$ exhibits a low-error plateau for recent memories followed by a steep sigmoid transition. The retention half-life $a_{1/2}$ — the age at which $\bar{F}$ crosses 0.5 — is the natural summary statistic.
2. *Linear scaling with L and the capacity constant c.* The half-life scales as $a_{1/2} \approx c\,L$ with $c \approx 2.4$ for the default circular-drift geometry, from $a_{1/2} = 14$ at $L = 5$ to $a_{1/2} = 74$ at $L = 30$. The fact that $c > 1$ means the CAS scheme extracts more than one effective day of retention per grid node, outperforming a naïve FIFO buffer by a factor of $\sim$2.4. We derived an analytical expression $a_{1/2} \approx L \ln(1/t_*)$ linking c to a readout-time resolution threshold (Section 9.1) and argued that c plays a role analogous to the Shannon channel capacity.
3. *Independence of K.* Sweeping $K \in \{1, 2, 3, 5, 8\}$ at fixed $L = 10$ yields $a_{1/2} \approx 30$ for all K. Temporal compression — not state-space complexity — controls the forgetting rate.
4. *Drift speed matters, geometry less so.* Faster drift (shorter period P) reduces the half-life (equivalently, reduces c), while the choice between circular and linear drift geometry affects the curve shape but not dramatically the timescale.
5. *Confusion, not destruction.* Old memories collapse toward recent eras ($\bar{F} > 1$) rather than reverting to the prior ($\bar{F} \to 1$). This is visible both in the normalised metric and in the spatial displacement of replayed means toward the protocol interior.
6. *Adaptive forgetting channel.* The decomposed metric correctly identifies the active information channel: mean-dominated ($\sim$85%) when component means drift (synthetic experiments), covariance-dominated when only weights vary (MNIST experiment). Weight error is negligible for equal-weight mixtures.

The stochastic process reconstructed from the density path (Appendix A, Section 9.2) provides temporally coherent replay trajectories — compressed "movies" of the agent's history — that are structurally analogous to hippocampal sleep replay in biological memory systems. The MNIST experiment (Section 8) made this

literal: the protocol grid, decoded frame-by-frame to pixel space, produced a visual temporal narrative in which digit identities were preserved while mixing proportions evolved smoothly from oldest to most recent memories.

Extensions and applications. Several directions are immediate.

- *Optimising the retention constant c.* Non-uniform (logarithmic) grids, variational rebinning, and non-linear interpolants (e.g. Wasserstein geodesics) could increase c beyond 2.4. Determining the theoretical maximum c_{opt} connects the CAS framework to rate-distortion theory (Section 9.1).
- *Neural density families.* The CAS recursion applies to any density class admitting interpolation. Extending it to normalising flows or score-based models would enable high-dimensional, structured data beyond the GM class.
- *Power systems.* A dynamic memory agent could maintain a temporally compressed history of generation/consumption probability densities over a 24-hour cycle, providing input for on-the-fly operational optimisation.
- *Lagrangian turbulence.* Continual learning of particle-tracking statistics could carry information from small-scale to large-scale dynamics via progressively coarsened temporal representations.
- *Sleep-replay applications.* The SDE-based replay (Section 9.2) could be used for off-line trajectory generation in model-based reinforcement learning, where temporally coherent experience replay is known to improve sample efficiency.

Acknowledgments

The author is grateful to Maxim Bazhenov for many inspiring discussions. The author thanks the University of Arizona start-up programme for financial support. Large language models (Claude, Anthropic; ChatGPT, OpenAI) assisted with text editing and code refactoring; all mathematical derivations, scientific claims, and code were independently verified by the author.

A Density Interpolants

Assume that the density path $p_t(x)$, $t \in [0,1]$, $x \in \mathbb{R}^d$, is known. We seek a unit-diffusion Itô process

$$dX_t = s_t(X_t)\,dt + dW_t, \tag{18}$$

whose density is exactly $p_t(x)$. This construction — recovering an SDE drift from a prescribed density path via the Fokker–Planck equation — follows the stochastic interpolant framework of Albergo and Vanden-Eijnden (2023), specialized to a piecewise-linear GM interpolant with unit diffusion coefficient.

The drift s_t must satisfy the Fokker–Planck equation

$$\partial_t p_t(x) + \nabla \cdot J_t(x) = 0, \qquad J_t(x) \doteq s_t(x)\,p_t(x) - \frac{1}{2}\nabla p_t(x), \tag{19}$$

where the first relation is the continuity equation and the second defines the probability current $J_t(x)$. Once the continuity equation is solved – that is the current is expressed via the density – a valid drift is recovered as

$$s_t(x) = \frac{J_t(x)}{p_t(x)} + \frac{1}{2}\nabla \log p_t(x). \tag{20}$$

A.1 Densities which are Gaussian Mixtures

Consider now the case when the density is a Gaussian mixture of degree K:

$$p_t(x) = \sum_{k=1}^{K} \pi_k(t)\,g_k(x,t), \qquad g_k(x,t) \doteq \mathcal{N}(x; m_k(t), \Sigma_k(t)), \tag{21}$$

where $\pi_k(t) > 0$, $\sum_{k=1}^{K} \pi_k(t) = 1$, $\Sigma_k(t) \succ 0$, and $\pi_k(t)$, $m_k(t)$, $\Sigma_k(t)$ are assumed known.

Constant weights. If $\pi_k(t) \equiv \pi_k$, $k = 1, \dots, K$, then the continuity equation can be integrated explicitly, resulting in the Gaussian-mixture expression

$$J_t(x) = \sum_{k=1}^{K} \pi_k \, J_{k,t}(x), \qquad J_{k,t}(x) \doteq g_k(x,t) \left[\dot{m}_k(t) + \frac{1}{2} \dot{\Sigma}_k(t) \Sigma_k^{-1}(t) \big(x - m_k(t)\big) \right]. \tag{22}$$

Time-varying weights. In general, when the weights vary in time, we decompose the current into two parts:

$$J_t(x) = J_t^{\text{shape}}(x) + J_t^{\text{wt}}(x). \tag{23}$$

The first term accounts for the motion and deformation of the Gaussian components:

$$J_t^{\text{shape}}(x) = \sum_{k=1}^{K} \pi_k(t) \, g_k(x,t) \left(\dot{m}_k(t) + \frac{1}{2} \dot{\Sigma}_k(t) \Sigma_k^{-1}(t) \big(x - m_k(t)\big) \right). \tag{24}$$

The correction current associated with the time dependence of the weights satisfies

$$\nabla \cdot J_t^{\text{wt}}(x) = - \sum_{k=1}^{K} \dot{\pi}_k(t) \, g_k(x,t). \tag{25}$$

Since $\sum_{k=1}^{K} \pi_k(t) = 1$, we also have $\sum_{k=1}^{K} \dot{\pi}_k(t) = 0$, and therefore the right-hand side of (25) has zero total mass, which is the compatibility condition for a decaying solution on $\mathbb{R}^d$.

Looking for the correction current in gradient form,

$$J_t^{\text{wt}}(x) = -\nabla \psi_t(x),$$

and decomposing

$$\psi_t(x) = \sum_{k=1}^{K} \dot{\pi}_k(t) \, \psi_{k,t}(x),$$

we obtain for each component the Poisson equation

$$\Delta \psi_{k,t}(x) = g_k(x,t).$$

Its solution can be written as the following one-dimensional integral:

$$\psi_t(x) = \frac{1}{2(2\pi)^{d/2}} \sum_{k=1}^{K} \dot{\pi}_k(t) \int_0^\infty \frac{\exp\big(-\frac{1}{2}(x - m_k(t))^T (\Sigma_k(t) + 2sI)^{-1} (x - m_k(t))\big)}{\sqrt{\det(\Sigma_k(t) + 2sI)}} \, ds. \tag{26}$$

Therefore,

$$J_t^{\text{wt}}(x) = -\nabla \psi_t(x), \qquad J_t(x) = J_t^{\text{shape}}(x) - \nabla \psi_t(x), \tag{27}$$

and the resulting drift is

$$s_t(x) = \frac{J_t^{\text{shape}}(x) - \nabla \psi_t(x)}{p_t(x)} + \frac{1}{2} \nabla \log p_t(x). \tag{28}$$

Equivalently, writing everything out,

$$s_t(x) = \frac{\sum_{k=1}^{K} \pi_k(t) \, g_k(x,t) \left(\dot{m}_k(t) + \frac{1}{2} \dot{\Sigma}_k(t) \Sigma_k^{-1}(t) \big(x - m_k(t)\big) \right) - \nabla \psi_t(x)}{\sum_{k=1}^{K} \pi_k(t) \, g_k(x,t)} + \frac{1}{2} \nabla \log p_t(x). \tag{29}$$

B Software Design and Experimental Protocol

This appendix describes the software architecture underlying the experiments in Sections 5–7 and the rationale for the experimental design choices. Code is available at https://github.com/mchertkov/CAS-Bridge-Diffusion.

B.1 Core API: `bridge_cas.py`

The entire continual-learning pipeline is implemented in a single Python module, `bridge_cas.py`, built on PyTorch to ensure full compatibility with automatic differentiation (autograd). This enables sensitivity analysis — e.g. $\partial a_{1/2}/\partial(\text{drift parameters})$ — and GPU acceleration for scaling experiments. The only non-differentiable component is the Hungarian matching used in the decomposed metric (Section 4.3), which is a discrete assignment computed via `scipy`; it lies outside the main CAS loop and does not affect gradient flow. The main classes are:

- `GaussianMixture`: stores weights $\pi \in \mathbb{R}^K$, means $m \in \mathbb{R}^{K\times d}$, covariances $\Sigma \in \mathbb{R}^{K\times d\times d}$; provides methods for overall moments, density evaluation, and sampling.
- `ProtocolGrid`: stores $L+1$ `GaussianMixture` node states at uniform times $\{0, 1/L, \ldots, 1\}$; implements the three CAS operations (compress, add, smooth) and piecewise-linear interpolation (1) for replay queries.
- `ContinualMemory`: orchestrates the daily incorporate loop; maintains the readout-time dictionary (8) and the history of original daily targets for metric computation.
- `ForgetMetrics`: implements the raw (10), normalised (13), and decomposed (14) forgetting metrics, including Hungarian matching for $K > 1$.

B.2 Data format and storage

The protocol state at any point in time is fully described by a list of $L+1$ Gaussian mixtures. Each mixture is stored as three arrays (π, m, Σ) of shapes $(K,)$, (K, d), (K, d, d). The total storage per protocol snapshot is $(L+1) \times K \times (1 + d + d^2)$ floating-point numbers. For diagnostic purposes, the full CAS history (all intermediate protocol states) can optionally be logged; in production, only the current protocol and readout-time dictionary are retained.

B.3 Experimental protocol

Each experiment follows a common workflow:

1. *Generate daily targets.* A stream of n daily Gaussian-mixture distributions is generated according to a specified drift model (circular, linear, random walk, or curriculum-based).
2. *Run CAS loop.* The `ContinualMemory` object is initialised with a prior $q^{(0)}$ and segment budget L. Each daily target is incorporated via one compress–add–smooth cycle. After each day, forgetting metrics are computed for all stored past days.
3. *Compute diagnostics.* The age-averaged forgetting curve $\bar{F}(a)$, retention half-life $a_{1/2}$, full forgetting matrix $\bar{F}(m, n)$, and (for $K > 1$) the decomposed metric are computed and stored.

B.4 Design principles

1. *Density-level storage.* The protocol stores GM states (not SDE drift coefficients or Hamiltonian parameters). This makes the representation interpretable, cheap to query, and independent of the drift-reconstruction step (Appendix A), which is only needed when sample paths are required.
2. *Modular density class.* The API is designed so that the `GaussianMixture` class can be replaced by any density family supporting: (a) linear interpolation of parameters, (b) moment computation, and (c) density evaluation. This enables future extensions to neural density estimators.
3. *Stochastic generation deferred.* Sample-path generation (via the drift from Appendix A) is not needed during the CAS recursion; it is only invoked at evaluation time for visualisation or downstream tasks. This saves compute during the daily update loop.
4. *Sweep-friendly.* All design parameters (L, K, d, drift geometry, prior) are passed as constructor arguments, enabling clean parameter-sweep loops in experiment notebooks.

References

M. S. Albergo and E. Vanden-Eijnden. Building Normalizing Flows with Stochastic Interpolants, Mar. 2023. URL http://arxiv.org/abs/2209.15571. arXiv:2209.15571 [cs, stat].

H. Behjoo and M. Chertkov. Harmonic Path Integral Diffusion. *IEEE Access*, 13:42196–42213, 2025. ISSN 2169-3536. doi: 10.1109/ACCESS.2025.3548396. URL https://ieeexplore.ieee.org/document/10910146/.

Y. Chen, T. T. Georgiou, and M. Pavon. Stochastic Control Liaisons: Richard Sinkhorn Meets Gaspard Monge on a Schrödinger Bridge. *SIAM Review*, 63(2):249–313, Jan. 2021. ISSN 0036-1445, 1095-7200. doi: 10.1137/20M1339982. URL https://epubs.siam.org/doi/10.1137/20M1339982.

M. Chertkov. Generative Stochastic Optimal Transport: Guided Harmonic Path-Integral Diffusion, Dec. 2025. URL http://arxiv.org/abs/2512.11859. arXiv:2512.11859 [cs].

M. Chertkov. Mean-field path-integral diffusion: From samples to interacting agents, 2026. URL https://github.com/mchertkov/MeanFieldPID.

M. Chertkov and H. Behjoo. Adaptive Path Integral Diffusion: AdaPID, Dec. 2025. URL http://arxiv.org/abs/2512.11858. arXiv:2512.11858 [cs].

V. De Bortoli, J. Thornton, J. Heng, and A. Doucet. Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 17695–17709. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/940392f5f32a7ade1cc201767cf83e31-Paper.pdf.

M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2022.

R. M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999.

R. Gao and W. Liu. DDGR: Continual Learning with Deep Diffusion-based Generative Replay. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10744–10763. PMLR, 2023. URL https://proceedings.mlr.press/v202/gao23e.html.

R. Golden, J. E. Delanois, P. Sanda, and M. Bazhenov. Sleep prevents catastrophic forgetting in spiking neural networks by forming a joint synaptic weight representation. *PLOS Computational Biology*, 18(11): e1010628, 2022.

R. Golden, R. Saxena, O. C. González, J. E. Delanois, S. Kilianski, B. L. McNaughton, and M. Bazhenov. Interleaved replay of novel and familiar memory traces during slow-wave sleep prevents catastrophic forgetting, 2025. Published: bioRxiv 2025.06.25.661579.

O. C. González, Y. Sokolov, G. P. Krishnan, J. E. Delanois, and M. Bazhenov. Can sleep protect memories from catastrophic forgetting? *eLife*, 9:e51005, 2020.

J. He, J. Chen, Q. Liu, S. Dai, J. Tang, and D. Liu. Continual Learning with Diffusion-based Generative Replay for Industrial Streaming Data. In *2024 IEEE/CIC International Conference on Communications in China (ICCC)*. IEEE, 2024. doi: 10.48550/arXiv.2406.15766. URL https://arxiv.org/abs/2406.15766.

L. Hu, Z. Gan, L. Deng, L. Liang, J. Liang, S. Huang, and T. Chen. ReplayCAD: Generative Diffusion Replay for Continual Anomaly Detection. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI)*, 2025. URL https://www.ijcai.org/proceedings/2025/328.

Q. Jodelet, X. Liu, Y. J. Phua, and T. Murata. Class-Incremental Learning using Diffusion Model for Distillation and Replay. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3417–3425. IEEE, 2023. doi: 10.1109/ICCVW60793.2023.00367. URL https://arxiv.org/abs/2306.17560.

J. Kim, H. Cho, J. Kim, Y. Y. Tiruneh, and S. Baek. SDDGR: Stable Diffusion-based Deep Generative Replay for Class Incremental Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024. URL https://arxiv.org/abs/2402.17323.

J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13): 3521–3526, 2017.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

J. Liang, J. Zhong, H. Gu, Z. Lu, X. Tang, G. Dai, S. Huang, L. Fan, and Q. Yang. Diffusion-Driven Data Replay: A Novel Approach to Combat Forgetting in Federated Class Continual Learning. In *Computer Vision – ECCV 2024*, Lecture Notes in Computer Science. Springer, 2024. doi: 10.1007/978-3-031-73404-5_18. URL https://arxiv.org/abs/2409.01128.

Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow Matching for Generative Modeling, Feb. 2023. URL http://arxiv.org/abs/2210.02747. arXiv:2210.02747 [cs, stat].

C. Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport, Aug. 2013. URL http://arxiv.org/abs/1308.0215. arXiv:1308.0215 [math].

M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. Academic Press, 1989.

Z. Meng, J. Zhang, C. Yang, Z. Zhan, P. Zhao, and Y. Wang. DiffClass: Diffusion-Based Class Incremental Learning. In *Computer Vision – ECCV 2024*, volume 15145 of *Lecture Notes in Computer Science*, pages 142–159. Springer, 2024. doi: 10.1007/978-3-031-73021-4_9. URL https://link.springer.com/chapter/10.1007/978-3-031-73021-4_9.

C. V. Nguyen, Y. Li, T. D. Bui, and R. E. Turner. Variational continual learning. In *International Conference on Learning Representations (ICLR)*, 2018.

G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

T. Richardson and R. Urbanke. *Modern Coding Theory*. Cambridge University Press, Cambridge, 2008. doi: 10.1017/CBO9780511791338.

A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell. Progressive neural networks, 2016. Published: arXiv:1606.04671.

J. Schwarz, W. Czarnecki, J. Luketina, A. Grabska-Barwinska, Y. W. Teh, R. Pascanu, and R. Hadsell. Progress & Compress: A scalable framework for continual learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 4535–4544, 2018.

H. Shin, J. K. Lee, J. Kim, and J. Kim. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

T. Tadros, G. P. Krishnan, R. Ramyaa, and M. Bazhenov. Sleep-like unsupervised replay reduces catastrophic forgetting in artificial neural networks. *Nature Communications*, 13:7742, 2022.

G. M. van de Ven, H. T. Siegelmann, and A. S. Tolias. Brain-inspired replay for continual learning with artificial neural networks. *Nature Communications*, 11:4069, 2020.

D. Vins, J. E. Delanois, and M. Bazhenov. Optimal stopping for continual learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

L. Wang, X. Zhang, H. Su, and J. Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5362–5383, 2024.

F. Zenke, B. Poole, and S. Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 3987–3995, 2017.