Continual Learning in Modern Hopfield Networks with an Application to Diffusion Models

Ken Takeda[1,2] **Masafumi Oizumi**[1] **Ryo Karakida**[2,3,*]

Abstract

Generative models, including diffusion models, are increasingly used as foundation models and adapted through sequential fine-tuning, making continual learning an essential problem setting. However, continual learning in such generative models remains poorly understood: after a task change, what aspects of the learned distribution are most easily lost, and what replay samples should be prioritized? We address these questions through the modern Hopfield energy. Recent links between modern Hopfield networks (MHNs) and diffusion models allow analyses in MHNs to be transferred to diffusion models. We introduce intrinsic forgetting as an increase in Hopfield energy after the task change. In tractable settings in an MHN, we prove that high-energy, outlier-like samples undergo a larger energy increase than cluster-like samples, implying that samples located in sharp, isolated basins are more forgettable. We further analyze memory replay and show that replay is particularly effective for high-energy samples, enabling an energy-based selection of replay samples. We validate these predictions in experiments on MHNs and two diffusion models under continual-learning settings: Stable Diffusion and a pixel-space denoising diffusion probabilistic model (DDPM). In these diffusion models, Hopfield energy tracks reconstruction-based forgetting, and replay experiments reveal energy-dependent mitigation of forgetting that is consistent with the MHN analysis.

1 Introduction

Generative models form a foundation of modern AI, and in particular, diffusion models have become a default backbone for image generation [1–4]. In recent years, generative models have increasingly been used not only after training from scratch, but also undergoing continual knowledge transfer, such as continual pretraining and fine-tuning in different domains. This brings continual learning (CL) of diffusion models to the foreground as a practical engineering problem and a fast-growing empirical literature [5–8], Continual learning for diffusion models remains largely empirical: the mechanism of catastrophic forgetting is still poorly understood, and even basic questions that have been extensively studied in supervised models remain open– "*which* information of Task-1 will this model forget after training on Task 2?" and "*which* Task-1 samples should we replay to prevent it?"

Our approach to such fundamental questions is through Hopfield networks, which have played a central role in advancing the understanding and development of generative models [9, 10]. A distinctive feature of Hopfield networks is that they admit an explicitly computable energy function. This enables systematic architecture design with the energy function serving as a guiding principle [11, 12], as well as the development of machine learning methods [13–15]. In particular, modern Hopfield networks (MHNs) form a family of models that has been actively studied since the rise of deep

[1]Graduate School of Arts and Science, The University of Tokyo, Tokyo, Japan
[2]Artificial Intelligence Research Center, AIST, Tokyo, Japan
[3]IIR, Institute of Science Tokyo, Tokyo, Japan
[*]Corresponding author, Email: karakida.ryo@aist.go.jp

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

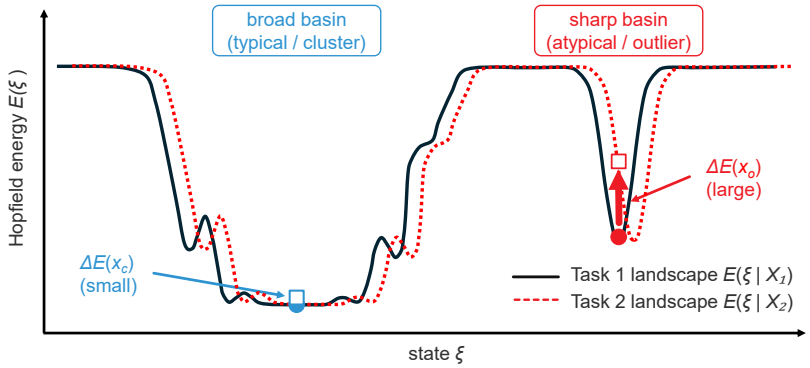


Figure 1: **Conceptual picture.** The Hopfield energy $E(\xi \mid X)$ of associative memory / diffusion models carves two qualitatively different basins: *broad and low-energy basins* around typical samples, and *sharp and high-energy basins* around outliers. A task change shifts the landscape (red dashed). The resulting energy rise at the Task-1 memory, $\Delta E(x) = E(x \mid X_2) - E(x \mid X_1)$, referred to as intrinsic forgetting. For small rotations in the stated regime, we prove $\Delta E_o^{\mathrm{fp}} > \Delta E_{cl}^{\mathrm{fp}}$ at Task-1 fixed points: memories in high-energy basins are more susceptible to forgetting.

learning and has been extended to deep learning architectures [10, 16]. Crucially, a growing literature identifies an underlying connection between MHNs and diffusion models: the score function of a Gaussian kernel-density estimator is exactly a Hopfield update [17, 18], and backward diffusion dynamics can be read as gradient flows on an implicit log-sum-exp energy [19, 20].

In this work, we reveal that the Hopfield energy serves as a theoretically grounded measure for quantitatively characterizing forgetting and prioritizing which samples to replay. Figure 1 illustrates the core idea of quantifying forgetting in generative models admitting the Hopfield energy, whose details we explain in Section 4. For such models, changes in generation probability can be tracked by observing the energy. By characterizing the forgetting of knowledge as a decrease in its generation probability, we can reduce the problem of quantifying forgetting to measuring an increase in the corresponding energy. Regions with densely clustered memory patterns form low energy, less sharp landscapes. Therefore, even if memory vectors shift from Task 1 to Task 2, the energy increase remains relatively small, so the corresponding states are robust to forgetting. By contrast, a Task-1 memory lying in a region with sparse or isolated memories undergoes a much larger energy increase and forgetting. This picture is analogous to the view that associates flat minima in parameter space with better generalization to test samples and out-of-distribution (OOD) samples [21, 22].

Main contributions are summarized as follows:

- **Hopfield energy view of forgetting supported by theoretical analysis (Theorem 4.1).** We introduce intrinsic forgetting as the increase in Hopfield energy caused by a task transition. This provides an internal measure to quantify how the learned generative landscape changes, without relying solely on external classifiers or distributional metrics such as Fréchet inception distance (FID). To formalize the intrinsic forgetting shown in Figure 1, we analyze a modern Hopfield network with structured memories. The analysis shows that outlier-like memories, corresponding to the high-energy state and sharp basin, undergo a larger energy increase under a task perturbation than typical cluster memories.
- **Analysis of energy-dependent replay (Theorem 5.1).** We analyze how Hopfield energy controls the effect of memory replay. The analysis shows that high-energy samples can produce larger replay gains than low-energy samples, revealing an energy dependence in the effect of replay. This suggests that Hopfield energy may provide a useful signal for replay-buffer design. Controlled experiments with an MHN and its Boltzmann-machine counterpart support this energy-dependent replay picture.
- **Empirical examination in diffusion models (Section 4.3, 5.3)**. We empirically confirm these theoretical predictions using two diffusion models in continual-learning settings: Stable Diffusion v1.5 fine-tuned on split CIFAR-10 and a pixel-space DDPM trained on the same one. In both settings, Hopfield energy tracks reconstruction-based forgetting, and

replay experiments show energy-dependent changes in reconstruction error consistent with the MHN analysis.

Thus, we expect that this work will serve as a theoretical and empirical foundation for exploring forgetting and continual learning in modern generative models.

2 Related Work

Modern Hopfield networks. Recent work on MHNs has developed various extensions of the classical Hopfield network [9], including energy functions beyond pairwise correlations [23, 24] and architectural generalizations [10]. In particular, Ramsauer et al. [16] showed that a Hopfield network with a log-sum-exp energy function corresponds to softmax attention in modern deep learning architectures. The term MHN is also commonly used to refer specifically to the model of [16], and we follow this convention. A distinctive feature of MHNs is that they admit an explicit energy function of states. Recent studies have increasingly exploited this energy as an intrinsic model quantity, for example, to systematically design architectures from its gradient [10, 11], to detect out-of-distribution (OOD) samples [13, 14], and to derive biologically plausible algorithms [15]. In this work, we reveal that the energy also works as a useful indicator of forgetting in generative models.

Connection between MHNs and diffusion models. Raya and Ambrogioni [19] showed that diffusion trajectories can be viewed as dynamics on an implicit potential landscape, with attractor-like behavior toward the data manifold. Ambrogioni [17] further made the connection to MHNs explicit by showing that, for finite normalized patterns, the low-noise diffusion energy is asymptotically identical to the MHN energy. Recent work has more clarified this view [20] and extended it to memorization/generalization behavior in diffusion models [18, 20]. Our work uses this established MHN–diffusion correspondence in a new setting: continual learning of generative models.

Continual learning in generative models. Continual learning in generative models has mainly been studied from a practical perspective, with the aim of mitigating catastrophic forgetting, rather than from the viewpoint of theoretically understanding its mechanism. Early studies considered replay algorithms in Hopfield networks, while much of the recent work has focused on VAEs and GANs, including teacher–student generative modeling, memory-replay GANs, latent space consolidation, and distillation for conditional generation [25–28]. More recently, continual learning has also been studied for diffusion models, including experimental demonstrations of forgetting and its mitigation by replay [5], a regularization method tailored to diffusion models [29], and parameter-efficient adaptation and distillation in text-to-image diffusion models [6, 8]. In contrast to these application-oriented studies, our work uses MHNs as a theoretical bridge to understand the mechanism of forgetting and to provide a basis for developing principled methods to mitigate forgetting.

3 Preliminaries

3.1 Modern Hopfield Energy

Let $X = \{x_i\}_{i=1}^N$ be a set of stored memories and let $\xi \in \mathbb{R}^d$ denote a state at which the memory landscape is evaluated. The continuous modern Hopfield energy is defined as

$$E(\xi \mid X; \beta) = -\frac{1}{\beta} \log \sum_{i=1}^{N} \exp(\beta x_i^\top \xi) + \frac{1}{2}\|\xi\|^2, \tag{1}$$

where $\beta > 0$ is an inverse-temperature parameter controlling the sharpness of the energy landscape [10, 16, 30]. For $\nabla_\xi E(\xi \mid X; \beta) = 0$, its fixed points satisfy

$$\xi = \sum_{i=1}^{N} p_i(\xi \mid X) x_i, \quad p_i(\xi \mid X) = \frac{\exp(\beta x_i^\top \xi)}{\sum_{j=1}^{N} \exp(\beta x_j^\top \xi)}. \tag{2}$$

Throughout this paper, $E(\xi \mid X)$ denotes the Hopfield energy evaluated with respect to the memory set X, omitting β when it is clear from context. The energy is low when the state ξ lies near a cluster of memories, and high when it lies near an isolated memory. It is the scalar quantity we use below to characterize how the generative landscape changes across tasks. Appendix A summarizes the notation.

3.2 Boltzmann Machine Formulation

Define the Boltzmann distribution of the MHN by

$$p_\beta(\xi \mid X) = \exp(-\beta E(\xi \mid X; \beta))/Z, \tag{3}$$

where Z is the partition function. One can easily see that, for $\|x_i\| = 1$, Z is explicitly given by $Z = N(2\pi/\beta)^{d/2}e^{\beta/2}$, and this Boltzmann distribution is a Gaussian mixture $p_\beta(\xi \mid X) = e^{\beta/2}\sum_i \exp\left(-\frac{\beta}{2}\|\xi - x_i\|^2\right)/Z$. In the following section, as one of our synthetic experiments, we perform unsupervised learning of an MHN by maximum-likelihood estimation of this Boltzmann distribution. Specifically, we regard this distribution as a Boltzmann machine (BM) with a trainable weight matrix X, and refer to it as MHN-BM. Since Z is analytically tractable, the model can be trained easily by gradient descent for maximum-likelihood learning. Note that a similar Boltzmann machine counterpart of MHNs was investigated by [31], but their treatment of the inverse temperature differs from our focus.

3.3 Diffusion Models and Implicit Energy

Diffusion models are not usually written as explicit energy-based models, but their score can be viewed as the negative gradient of an implicit, time-dependent energy. Let $p_s(x)$ be the noisy data marginal at diffusion time s, and write

$$E_{\mathrm{DM},s}(x) := -\log p_s(x), \qquad \nabla_x \log p_s(x) = -\nabla_x E_{\mathrm{DM},s}(x). \tag{4}$$

Thus, roughly speaking, reverse sampling moves a noisy sample x_s by noisy descent on $E_{\mathrm{DM},s}$; under conservative-drift assumptions, this energy also includes the drift potential [19]. Its stable equilibria therefore act as attractors for generation and memory recall [17, 18]. For a finite set of patterns corrupted by additive Gaussian noise,

$$p_s(x) \propto \sum_i \exp\!\left(-\|x - y_i\|^2/2\tau_s^2\right). \tag{5}$$

Here $\tau_s > 0$ denotes the standard deviation of the Gaussian corruption at diffusion time s. Expanding the square shows that, for normalized memories, $E_{\mathrm{DM},s}$ reduces up to constants and positive rescaling to

$$E_{DM,s}(x) \propto -\frac{1}{\beta_s}\log\sum_{i=1}^{N}\exp(\beta_s y_i^\top x) + \frac{1}{2}\|x\|^2, \qquad \beta_s = \tau_s^{-2}, \tag{6}$$

which is the modern Hopfield energy in Eq. (1). We therefore use Hopfield energy as a tractable proxy for the implicit diffusion energy landscape. We note that this perspective, which involves analysing the reverse (generative) process in terms of attractor dynamics on a Hopfield-like energy landscape, is not unique to this work and has been adopted in prior work [17–20]; our contribution is to leverage this established viewpoint in the new setting of continual learning of generative models. A more explicit derivation, including the variance-preserving case, is given in Appendix B.

4 Energy as an Indicator of Forgetting

We first introduce forgetting with the internal Hopfield energy. Defining "forgetting" in an unsupervised generative model is subtle: label-based metrics via external classifiers are heuristic measures [5, 6], but the generative objective of our interest is essentially unsupervised and class information is not necessarily relevant to the input data. We need an internal, model-native measure applicable to pure unsupervised settings.

Intrinsic forgetting via Hopfield energy. Let $\mathcal{X}_{\mathrm{old}}$ and $\mathcal{X}_{\mathrm{new}}$ denote the stored memory set at task 1 and task 2. For any state x, we define the *intrinsic forgetting* at x as

$$\Delta E(x) = E(x \mid \mathcal{X}_{\mathrm{new}}) - E(x \mid \mathcal{X}_{\mathrm{old}}). \tag{7}$$

Large positive $\Delta E(x)$ means that the energy landscape at x has risen, so the model assigns a smaller probability to that state after the task transition. Under the assumption that the score function has sufficient expressive capacity to approximate the data distribution of each task nearly perfectly,

this quantity captures forgetting induced by the intrinsic and unavoidable distributional differences between tasks. It is noteworthy that, for a fixed number of unit-norm memories and common β, $p_\beta(\xi \mid \mathcal{X}_{\text{new}})/p_\beta(\xi \mid \mathcal{X}_{\text{old}}) = \exp[-\beta\Delta E(\xi)]$. Thus, intrinsic forgetting naturally arises when distributional changes between tasks are measured in terms of probability density ratios.

4.1 Theoretical analysis: equal-angular cluster plus outlier under small rotation

We derived an analytical formulation of intrinsic forgetting within a toy setting.

Setup (Fig. 2). Consider memory vectors (corresponding to training samples) $x_1, \ldots, x_N, x_o \in \mathbb{S}^{d-1}$ satisfying

$$x_i^\top x_j = c \; (i \neq j;\; i,j \leq N), \qquad x_i^\top x_o = 0 \; (i \leq N), \qquad 0 < c < 1, \tag{8}$$

so $\{x_i\}_{i\leq N}$ is an *equal-angular cluster* and x_o is the *orthogonal outlier*. The exact geometry requires $N+1 \leq d$; Appendix G gives a construction and discusses approximate configurations for larger N. Task 1 stores $X = \{x_1, \ldots, x_N, x_o\}$; Task 2 is provided by its rational domain shift:

$$X_2 = VX, \quad \text{with } V = \exp(\varepsilon\Omega), \quad \Omega = (W - W^\top)/2. \tag{9}$$

Note that the rotation identity $E(\xi \mid VX) = E(V^\top\xi \mid X)$ converts a domain shift into a change of argument at the original landscape. This transformation facilitates our analysis. In the analysis, we assume a sufficiently small random rotation with $W_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $\varepsilon \ll 1$. We fix $N \geq 2, d, c$, take β sufficiently large, and then take ε sufficiently small for that β [30, 32].

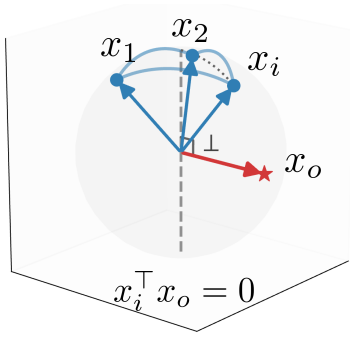


Figure 2: **Memory configuration.**

Fixed points nearby each memory vector. Let ξ_{cl}^*, ξ_o^* denote the cluster-side and outlier-side Task-1 fixed points. By the symmetry of (8), the softmax weights at each Task-1 fixed point reduce to scalar self-consistency conditions whose large-β asymptotics (proved in Appendix C) give the closed-form expansions

$$\xi_o^* \sim x_o + e^{-\beta}(s - Nx_o), \quad \xi_{cl}^* \sim x_{cl} + e^{-\beta(1-c)}\big(\textstyle\sum_{j\geq 2} x_j - (N-1)x_{cl}\big), \tag{10}$$

with $s := \sum_{i\leq N} x_i$. For clarity, here and below we use the informal shorthand "$\sim$" to show only the leading-order term; the precise lower-order terms and error bounds are given in Appendix C (Eqs. (73), (75) for the fixed points and Eqs. (97)–(98) for the energy rises).

The dominant off-memory weights are therefore $e^{-\beta(1-c)}$ for the cluster fixed point and $e^{-\beta}$ for the outlier fixed point, with $e^{-\beta(1-c)} \gg e^{-\beta}$ for $0 < c < 1$. Geometrically, the cluster fixed point is supported by $N-1$ near-neighbors at cosine c, while the outlier sees only orthogonal competitors. Appendix D shows that these same off-memory weights also control the local energy level and the local sharpness $\mathcal{S}$ of the basin. This justifies the shorthand used below: the cluster fixed point is low-energy and broad, whereas the outlier fixed point is high-energy and sharp.

Intrinsic forgetting at the fixed points. For a Task-1 fixed point ξ_*, write $\langle\cdot\rangle_V := \mathbb{E}_\Omega[\cdot]$ for the expectation over the small-rotation ensemble of (9). We absorb this average into the definition and express the (rotation-averaged) intrinsic forgetting at ξ_* as

$$\Delta E_*^{\text{fp}} \;:=\; \big\langle E(\xi_* \mid VX) - E(\xi_* \mid X)\big\rangle_V. \tag{11}$$

Since ξ_* is stationary, Taylor expansion combined with the rotation identity gives, at leading order,

$$\Delta E_*^{\text{fp}} \;\sim\; \tfrac{\varepsilon^2}{4}\big(\|\xi_*\|^2 \operatorname{tr} H_* - \xi_*^\top H_* \xi_*\big), \qquad H_* = I - \beta\Big(\textstyle\sum_i p_i^* x_i x_i^\top - \xi_*\xi_*^\top\Big). \tag{12}$$

Writing the off-memory weights as $p_k^* = 1 - \sum_{j\neq k}\eta_j$, $p_j^* = \eta_j \ll 1$ (with $\alpha_j := x_j^\top x_k$) gives the following:

$$\Delta E_*^{\text{fp}} \;\sim\; \tfrac{\varepsilon^2}{4}\Big[(d-1) - \sum_{j\neq k}\eta_j\,\Lambda(\alpha_j)\Big], \qquad \Lambda(\alpha) := 2(d-1)(1-\alpha) + \beta(1-\alpha^2). \tag{13}$$

Specializing (13) to the two representative Task-1 fixed points, namely the outlier (N competitors at $\alpha = 0, \eta = e^{-\beta}$) and the cluster ($N-1$ competitors at $\alpha = c, \eta = e^{-\beta(1-c)}$ plus one competitor at $\alpha = 0, \eta = e^{-\beta}$), we obtain the following.

Theorem 4.1 (High-energy patterns are forgotten more)**.** *For sufficiently large β with $\beta(1-c) \geq (1+\delta)N$ (fixed $\delta > 0$), all stable fixed points lie near either the outlier or a cluster memory. For every cluster fixed point,*

$$E_o^* > E_{cl}^*, \qquad \Delta E_o^{\mathrm{fp}} > \Delta E_{cl}^{\mathrm{fp}}, \qquad \mathcal{S}_o > \mathcal{S}_{cl}. \tag{14}$$

Proof sketch. The inequalities follow from the off-memory weight expansions in Appendices C and D; Appendix E shows that these cover all stable fixed points. At $c = 0$, the cluster and outlier fixed points are symmetry-equivalent. □

Appendix F compares energy and curvature at the cluster prototype and the outlier reference point, without assuming that they are fixed points.

4.2 Synthetic verification in MHN and Boltzmann-machine learning

We test the qualitative ordering of Theorem 4.1 in finite synthetic cluster-plus-outlier systems outside its asymptotic regime. The closed-form MHN uses an equal-angular cluster and an orthogonal outlier with $N = 12, d = 50, c = 0.35, \beta = 8, \varepsilon = 0.05$, averaged over 2000 random rotations. The outlier fixed point has a larger rotation-induced energy rise than the cluster fixed points, as predicted by Theorem 4.1; details are in Appendix I.1. The same ordering also appears after likelihood training in the MHN-BM with $N = 12, d = 50, c = 0.35, \beta = 16, \varepsilon_{\mathrm{rot}} = 0.02$, shown in Fig. 3a.

4.3 Application to Diffusion models

We next examine whether the prediction of Theorem 4.1 extends beyond the tractable MHN setting to continual learning in *diffusion models*. The main text focuses on *incremental class learning with split CIFAR-10*, which is a standard benchmark for continual learning. We also report results for *domain-incremental learning with rotated MNIST* in Appendix J. This setting most closely matches the rotation-based analysis in Sections 4.1 and 5.1, and yields the same qualitative conclusions.

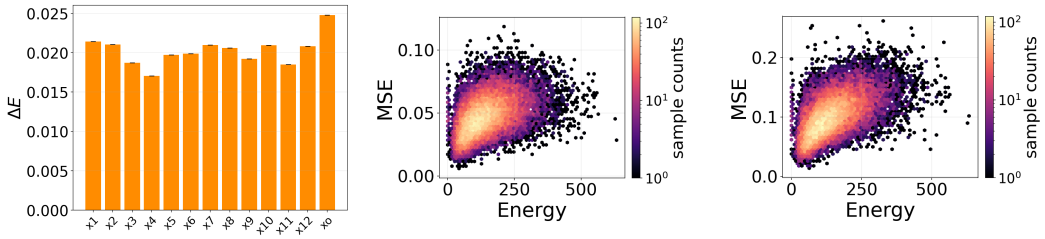
We evaluate two practically relevant diffusion-model systems on split CIFAR-10: latent-space **Stable Diffusion v1.5** fine-tuned on the task split (Fig. 3b), and a **pixel-space DDPM** trained from scratch on the same split (Fig. 3c). Following prior work [17–19], we define the per-sample Hopfield energy of a diffusion model using Eq. (1), with the stored memories $\{x_i\}$ given by the Task-1 training images.

Reconstruction metric for forgetting. In the diffusion-model experiments, we additionally use a reconstruction-based measure of forgetting, which evaluates how well a Task-1 sample can still be recovered by the model after Task-2 training. Specifically, we corrupt a Task-1 sample x to a fixed diffusion timestep $t^\star$, run the post-Task-2 reverse process from $t^\star$ back to 0, and measure

$$\mathcal{F}_{t^\star}(x) = \frac{1}{d_x} \|x - \hat{x}_{0,t^\star}\|_2^2. \tag{15}$$

Here d_x is the number of pixel coordinates. A larger $\mathcal{F}_{t^\star}(x)$ indicates that the post-Task-2 reverse dynamics is less able to reconstruct the original Task-1 sample. We use this heuristic metric to examine whether the energy-based tendency predicted by intrinsic forgetting is also reflected in the behavior of trained diffusion models. Details of this metric and robustness checks across other timesteps are given in Appendix K. We also report a perturbation-based sharpness analysis on CIFAR-10 in Appendix D.4, showing how local sharpness relates to both Hopfield energy and reconstruction-based forgetting.

Stable Diffusion on split CIFAR-10 (Fig. 3b). We fine-tune Stable Diffusion v1.5 (`runwayml/stable-diffusion-v1-5`, VAE frozen, only the UNet trained) on a split CIFAR-10 protocol where each task is composed of different 5 classes. After Task 1 we assign every Task-1 image a scalar Hopfield energy $E_\theta(x)$ computed via the log-sum-exp expression (1) over all training set of Task-1 images. This is our operationalization of the diffusion model's intrinsic energy at x. After Task 2 training, we measure the forgetting by the per-sample reconstruction MSE at noise level $t = 0.8$. This is the main diffusion-model instantiation of Theorem 4.1.



(a) MHN-BM. (b) Stable Diffusion. (c) Pixel diffusion.

Figure 3: Energy predicts per-sample forgetting across diffusion models and a trainable synthetic model. (a) In the trainable MHN-BM, the same ordering emerges after likelihood training on the rotated two-task setup. The MHN with synthetic clustering memory is reported in Appendix I.1. (b) Stable Diffusion v1.5 on split CIFAR-10: per-sample reconstruction MSE $\mathcal{F}_{t^\star}(x)$ after Task 2 training vs. internal Hopfield energy measured on Task 1. Higher-energy images suffer larger reconstruction forgetting. (c) Pixel-space DDPM on the same split reproduces the same monotone relation. Bin colors show sample counts on a logarithmic scale.

Pixel-space diffusion (Fig. 3c). To check that the relation is not specific to the latent space learned by Stable Diffusion, we repeat the same protocol with a DDPM trained directly in pixel space on 32×32 split CIFAR-10 from scratch. The same energy-vs-forgetting ordering is reproduced: per-sample reconstruction MSE after Task 2 is a monotone-increasing function of the Task-1 Hopfield energy. Together, Fig. 3b and Fig. 3c show that Hopfield energy is a model-agnostic, sample-level forgetting indicator for diffusion CL.

5 Energy-guided replay prioritization

Given that high-energy memories are forgotten stronger, the natural question is whether they should be replayed more. We answer this affirmatively with a second theorem, then turn it into a selection rule of replay samples.

5.1 Theoretical analysis: baseline-subtracted replay gain

We add one Task-1 replay sample x_r to the rotated Task-2 dataset, $X_2^{+r}(V) := VX \cup \{x_r\}$, with the replayed sample counted as an additional term in the log-sum-exp. Define the raw replay gain at the Task-1 fixed point ξ_k^* as

$$\mathcal{R}_{r\to k}(V) \;:=\; E(\xi_k^* \mid X_2(V)) - E(\xi_k^* \mid X_2^{+r}(V)). \tag{16}$$

Writing $\Delta E_k^{\mathrm{fp}}(V) := E(\xi_k^* \mid X_2(V)) - E(\xi_k^* \mid X)$ for the intrinsic forgetting (11), the gain decomposes as

$$\mathcal{R}_{r\to k}(V) = \underbrace{\Delta E_k^{\mathrm{fp}}(V)}_{\text{forgetting w/o replay}} - \underbrace{\big[E(\xi_k^* \mid X_2^{+r}(V)) - E(\xi_k^* \mid X)\big]}_{\text{forgetting w/ replay}}, \tag{17}$$

so $\mathcal{R}_{r\to k}(V)$ measures the reduction of intrinsic forgetting due to replay. With the softmax weight $p_{r|k} = e^{\beta x_r^\top \xi_k^*} / \sum_{\mu\in\mathcal{I}} e^{\beta x_\mu^\top \xi_k^*}$, the log-sum-exp form of E gives $\mathcal{R}_{r\to k}(V) = \frac{1}{\beta}\log(1 + p_{r|k}\, e^{\beta \Delta E_k^{\mathrm{fp}}(V)})$, of which $\mathcal{R}_{r\to k}(I) = \frac{1}{\beta}\log(1+p_{r|k})$ is the trivial duplicate-memory baseline. Subtracting this baseline and Taylor-expanding in $\Delta E_k^{\mathrm{fp}}(V) = O(\varepsilon^2)$ gives, with $\rho_{r|k} := p_{r|k}/(1+p_{r|k})$ (full derivation in Appendix H),

$$\Delta_{r\to k} \;:=\; \big\langle \mathcal{R}_{r\to k}(V) - \mathcal{R}_{r\to k}(I) \big\rangle_V \;\sim\; \rho_{r|k}\, \Delta E_k^{\mathrm{fp}}. \tag{18}$$

We call $\rho_{r|k}$ a *replay susceptibility* because it is the linear response of the baseline-subtracted replay gain to the energy rise ΔE_k^{fp}. *Replay gain = replay susceptibility × rotation-induced energy rise.* The first factor asks how much softmax weight the replay point already has at the target fixed point (geometry); the second is exactly the rotation-averaged rise ΔE_k^{fp} produced by Theorem 4.1 (drift).

Let x_{cl_1} be a cluster memory and let $x_{cl_2} \neq x_{cl_1}$ be another member of the same cluster. The replay susceptibilities satisfy

$$\rho_{cl_1|cl_1},\, \rho_{o|o} \sim \tfrac{1}{2}, \qquad \rho_{cl_1|cl_2} \sim e^{-\beta(1-c)}, \qquad \rho_{cl_1|o},\, \rho_{o|cl_1} \sim e^{-\beta}. \tag{19}$$

The precise lower-order corrections are in Appendix H (Eqs. (156), (160)). Combining these with the inequality of ΔE_k^{fp} in Theorem 4.1, we obtain the following.

Theorem 5.1 (Replay ordering)**.** *Under the assumptions of Theorem 4.1, the baseline-subtracted replay gain factorises as (18). The five canonical gains obey*

$$\Delta_{o\to o} > \Delta_{cl_1\to cl_1} \gg \Delta_{cl_1\to cl_2} \gg \{\Delta_{cl_1\to o}, \Delta_{o\to cl_1}\}, \qquad \Delta_{cl_1\to o} \asymp \Delta_{o\to cl_1}. \tag{20}$$

Intuition. Equation (18) says that a replay sample reduces forgetting in proportion to its softmax weight at the target fixed point. The first inequality, $\Delta_{o\to o} > \Delta_{cl_1\to cl_1}$, compares self-replay: both have susceptibility $\rho \simeq 1/2$, but the outlier target has the larger energy rise by Theorem 4.1, so high-energy self-replay gives the largest per-sample rescue. The gap $\Delta_{cl_1\to cl_1} \gg \Delta_{cl_1\to cl_2}$ says that replacing self-replay by another cluster member loses the small overlap factor $e^{-\beta(1-c)}$, although it can still help nearby low-energy states. Finally, $\Delta_{cl_1\to cl_2} \gg \{\Delta_{cl_1\to o}, \Delta_{o\to cl_1}\}$ and $\Delta_{cl_1\to o} \asymp \Delta_{o\to cl_1}$ mean that cross-type replay between the cluster and the outlier is exponentially weaker, because their softmax overlap is only $e^{-\beta}$. Thus high-energy replay gives a tall-but-narrow rescue, while low-energy cluster replay gives a shallower-but-broader rescue over neighboring low-energy states.

5.2 Synthetic verification of replay ordering

We test the qualitative ordering of Theorem 5.1 on the same finite synthetic systems. In the closed-form MHN with $N = 12, d = 50, c = 0.35, \beta = 8, \varepsilon = 0.05$, the replay-gain shows the predicted hierarchy: outlier self-replay is largest, cluster self-replay is next, within-cluster cross replay is smaller, and cross-type replay between the cluster and the outlier is weakest. The full result is reported in Appendix I.2. The trained MHN-BM with $N = 12, d = 50, c = 0.35, \beta = 16, \varepsilon_{\mathrm{rot}} = 0.02$ shows the almost same qualitative ordering in Fig. 4.

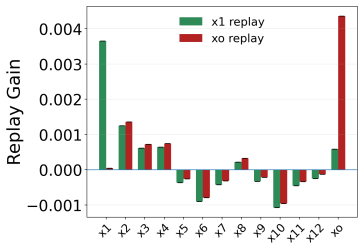


Figure 4: MHN-BM corroboration of the replay ordering. Ticks identify targets; legend entries identify replay sources.

5.3 Application to Diffusion models

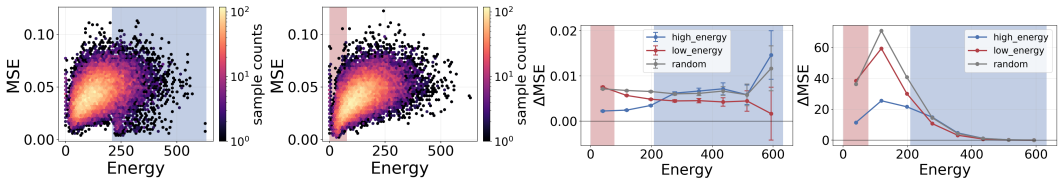
Theorem 5.1 delivers a concrete selection rule of replay samples: among Task-1 samples, replay those that maximize their own energy (thereby inheriting a large ΔE_k^{fp}) and that self-replay (so that $\rho_{k|k} \approx 1/2$). The simplest realization is *energy-guided top-K replay*.

Replay buffer. We pick the top-K highest-energy samples of Task 1 as the buffer, $\mathcal{B} = \mathrm{TopK}_{x\in\mathcal{D}_1} E_\theta(x)$, with $K = 5000$ in all experiments. During Task-2 fine-tuning we augment each mini-batch with replay samples drawn from $\mathcal{B}$, so that the effective training set is $\mathcal{D}_2 \cup \mathcal{B}$ (25,000 new + 5,000 replay images per epoch). We compare against *random* replay (uniform top-K) and *no replay* baselines.

We again focus on the two diffusion-model systems of Section 4.3 as the main empirical test.

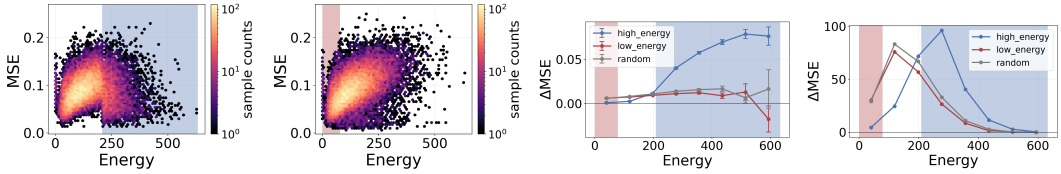
Stable Diffusion (Fig. 5). On split CIFAR-10, we compare three replay policies with identical buffer size $K = 5000$: *energy-guided top-K* (the K highest-energy Task-1 images), *random*, and *energy-guided bottom-K* (the K lowest-energy images). After Task-2 fine-tuning, per-sample reconstruction MSE at $t = 0.8$ is measured for every Task-1 image and binned by energy.

Eq. (18) predicts a mean–range tradeoff in the replay effect: $\Delta_{r\to k} \sim \rho_{r|k}\, \Delta E_k^{\mathrm{fp}}$. The energy-rise term ΔE_k^{fp} makes the per-sample replay gain larger for high-energy targets, whereas the relevance factor $\rho_{r|k}$ determines how broadly a replay sample affects other targets. Thus, high-energy replay is expected to produce a large but localized reduction in the high-energy tail, while low-energy replay is expected to produce a smaller reduction spread over a broader set of low-energy samples. This is what we observe: the high-energy buffer (Fig. 5a) yields a sharply peaked bin-mean reduction



(a) High-energy replay. (b) Low-energy replay. (c) Mean reduction. (d) Sum reduction.

Figure 5: Energy-guided replay on Stable Diffusion / split CIFAR-10. (a) Per-sample reconstruction MSE after Task 2 with the top-K *high-energy* replay buffer: MSE drops sharply for high-energy images (strong but narrow rescue). (b) With the bottom-K *low-energy* buffer the per-sample drop is smaller but spread across a wider energy range (weaker but broader rescue). (c,d) Per-bin ΔMSE (no-replay MSE minus replay MSE) versus energy, averaged (c) and summed (d), with within-bin standard errors in (c) (Appendix L): the high-energy buffer concentrates its rescue at high energy, the low-energy buffer distributes a smaller per-sample rescue over many more samples, exactly as Theorem 5.1 predicts via $\Delta_{r\to k} \approx \rho_{r|k}\, \Delta E_k^{\mathrm{fp}}$.



(a) High-energy replay. (b) Low-energy replay. (c) Mean reduction. (d) Sum reduction.

Figure 6: The same energy-guided replay analysis on a pixel-space DDPM (trained from scratch on split CIFAR-10) reproduces every qualitative conclusion of Fig. 5: high-energy replay rescues high-energy samples, low-energy replay rescues only low-energy samples, and the total reduction is dominated by the high-energy tail. The finding is not tied to the Stable Diffusion latent space.

at high energy (Fig. 5c), but its bin-sum reduction is confined to a narrow energy range (Fig. 5d). In contrast, the low-energy buffer (Fig. 5b) gives a smaller bin-mean reduction at low energy, but the affected range is broader, so the bin-sum reduction extends across a wider energy interval. This complementary pattern is consistent with the replay ordering in Eq. (20).

Pixel-space diffusion (Fig. 6). We repeat the three-policy comparison with the pixel-space DDPM of Section 4.3. Every qualitative conclusion from Stable Diffusion is reproduced: (i) high-energy replay protects high-energy images, (ii) low-energy replay only helps low-energy images, (iii) aggregated reductions place energy-guided top-K above random above low-energy. The replay benefit is again concentrated in the high-energy tail, as predicted. Together, Fig. 5 and Fig. 6 constitute our main empirical case that Hopfield energy is a practical, diffusion-agnostic replay criterion.

6 Conclusion and Limitations

Through theoretical analyses of modern Hopfield networks and empirical confirmation on diffusion models, we find that Hopfield energy serves as an indicator of forgetting and also provides a criterion for selecting replay samples. These results suggest that forgetting in unsupervised generative models can be evaluated not only through external evaluation metrics, but also through the geometry of the learned state landscape. We expect that this work will serve as a theoretical and practical foundation for future studies of knowledge forgetting, replay analysis, and energy-based diagnostics of generative models.

Limitations. (i) Forgetting through the learning of score functions. We consider intrinsic forgetting, that is, an unavoidable factor that arises when a generative model perfectly approximates task-dependent input distributions. In practical diffusion models, however, the score function is represented by a neural network, and additional forgetting may appear even for similar inputs. In supervised learning, theoretical analyses of such forgetting have been developed [33], and extending the analyses

to the score function is an important direction for future work. (ii) Finite sample sizes. Although our theory can handle structured data, it is currently restricted to the case of finitely many memories. Quantifying forgetting in regimes where the number of memories grows with the input dimension, and where the memories encode richer data manifolds [32], remains an open problem. (iii) Broader experimental validation. For future practical development, it would also be useful to examine scenarios not covered in this study, such as domains beyond images and online continual learning without explicit task boundaries. A more elaborated indicator of internal states (e.g., probability-flow ODE marginals) could tighten the correlation of Fig. 3b, 3c.

Acknowledgments and Disclosure of Funding

We thank the reviewers for their insightful and helpful feedback on the manuscript. The authors acknowledge the funding support from JST FOREST (Grant No. JPMJFR226Q). KT is supported by Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (KAKENHI) Grant No. 26KJ1029. M.O. is supported by JST Moonshot R&D Grant No. JPMJMS2012. RK is also supported by JSPS KAKENHI (Grant Nos. 22H05116, 23K16965), JST BOOST (Grant No. JPMJBY25B0) and FRONTia, a Japanese national program led by METI and NEDO.

References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

[2] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[3] Prafulla Dhariwal and Alex Nichol. Diffusion models beat GANs on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021.

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv:2112.10752*, 2021.

[5] Michał Zając, Kamil Deja, Anna Kuzina, Jakub M. Tomczak, Tomasz Trzciński, Florian Shkurti, and Piotr Miłoś. Exploring continual learning of diffusion models. *arXiv preprint arXiv:2303.15342*, 2023.

[6] James Seale Smith, Yen-Chang Hsu, Lingyu Zhang, Ting Hua, Zsolt Kira, Yilin Shen, and Hongxia Jin. Continual diffusion: Continual customization of text-to-image diffusion with C-LoRA. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

[7] Rui Gao and Weiwei Liu. DDGR: Continual learning with deep diffusion-based generative replay. In *International Conference on Machine Learning*, volume 202, pages 10744–10763. PMLR, 2023.

[8] Sergi Masip, Pau Rodriguez, Tinne Tuytelaars, and Gido M van de Ven. Continual learning of diffusion models with generative distillation. In *Third Conference on Lifelong Learning Agents (CoLLAs)*, 2024.

[9] John J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

[10] Dmitry Krotov and John J Hopfield. Large associative memory problem in neurobiology and machine learning. In *International Conference on Learning Representations*, 2021.

[11] Benjamin Hoover, Hendrik Strobelt, Dmitry Krotov, Judy Hoffman, Zsolt Kira, and Duen Horng Chau. Memory in plain sight: Surveying the uncanny resemblances of associative memories and diffusion models. In *NeurIPS 2023 Workshop on Associative Memory and Hopfield Networks*, 2023.

[12] Nima Dehmamy, Benjamin Hoover, Bishwajit Saha, Leo Kozachkov, Jean-Jacques Slotine, and Dmitry Krotov. NRGPT: An energy-based alternative for GPT. In *International Conference on Learning Representations*, 2026.

[13] Jinsong Zhang, Qiang Fu, Xu Chen, Lun Du, Zelin Li, Gang Wang, Xiaoguang Liu, Shi Han, and Dongmei Zhang. Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy. In *International Conference on Learning Representations*, September 2022.

[14] Claus Hofmann, Simon Schmid, Bernhard Lehner, Daniel Klotz, and Sepp Hochreiter. Energy-based hopfield boosting for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[15] Benjamin Scellier and Yoshua Bengio. Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in computational neuroscience*, 11:24, 2017.

[16] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021.

[17] Luca Ambrogioni. In search of dispersed memories: Generative diffusion models are associative memory networks. *Entropy*, 26(5):381, 2024. doi: 10.3390/e26050381.

[18] Bao Pham, Gabriel Raya, Matteo Negri, Mohammed J Zaki, Luca Ambrogioni, and Dmitry Krotov. Memorization to generalization: Emergence of diffusion models from associative memory. *arXiv preprint arXiv:2505.21777*, 2025.

[19] Gabriel Raya and Luca Ambrogioni. Spontaneous symmetry breaking in generative diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pages 66377–66389, 2023.

[20] Luca Ambrogioni. The statistical thermodynamics of generative diffusion models: Phase transitions, symmetry breaking, and critical instability. *Entropy*, 27(3):291, 2025. doi: 10.3390/e27030291.

[21] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

[22] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. SWAD: Domain generalization by seeking flat minima. In *Advances in Neural Information Processing Systems*, volume 34, pages 22405–22418, 2021.

[23] Dmitry Krotov and John J Hopfield. Dense associative memory for pattern recognition. In *Advances in Neural Information Processing Systems*, June 2016.

[24] Mete Demircigil, Judith Heusel, Matthias Löwe, Sven Upgang, and Franck Vermet. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168:288–299, 2017.

[25] Jason Ramapuram, Magda Gregorova, and Alexandros Kalousis. Lifelong generative modeling. *Neurocomputing*, 404:381–400, 2020. doi: 10.1016/j.neucom.2020.02.115.

[26] Chenshen Wu, Luis Herranz, Xialei Liu, Yaxing Wang, Joost van de Weijer, and Bogdan Raducanu. Memory replay GANs: Learning to generate new categories without forgetting. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[27] Mengyao Zhai, Lei Chen, Frederick Tung, Jiawei He, Megha Nawhal, and Greg Mori. Lifelong GAN: Continual learning for conditional image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2759–2768, 2019.

[28] Kamil Deja, Paweł Wawrzyński, Wojciech Masarczyk, Daniel Marczak, and Tomasz Trzciński. Multiband VAE: Latent space alignment for knowledge consolidation in continual learning. *arXiv preprint arXiv:2106.12196*, 2021.

[29] Zekun Wang, Anant Gupta, Zihan Dong, and Christopher J MacLellan. Avoid catastrophic forgetting with rank-1 fisher from diffusion models. In *International Conference on Learning Representations*, 2026.

[30] Carlo Lucibello and Marc Mézard. Exponential capacity of dense associative memories. *Phys. Rev. Lett.*, 132(7):077301, February 2024.

[31] Toshihiro Ota and Ryo Karakida. Attention in a family of Boltzmann machines emerging from modern Hopfield networks. *Neural Computation*, 35:1463–1480, 2023. doi: 10.1162/neco_a_01597.

[32] Beatrice Achilli, Luca Ambrogioni, Carlo Lucibello, Marc Mézard, and Enrico Ventura. The capacity of modern hopfield networks under the data manifold hypothesis. *arXiv preprint arXiv:2503.09518*, 2025.

[33] Sebastian Lee, Stefano Sarao Mannelli, Claudia Clopath, Sebastian Goldt, and Andrew Saxe. Maslow's hammer in catastrophic forgetting: Node re-use vs. node activation. In *International Conference on Machine Learning*, pages 12455–12477. PMLR, 2022.

[34] Felix Koulischer, Cédric Goemaere, Tom van der Meersch, Johannes Deleu, and Thomas Demeester. Exploring the temperature-dependent phase transition in modern hopfield networks. arXiv:2311.18434, 2023. URL `https://arxiv.org/abs/2311.18434`.

Appendices

A Notation

In Section 3.1, N is the total number of stored memories. In the structured model of Section 4.1 and its subsequent analysis, N counts cluster memories, with $N+1$ memories including the outlier. In Appendix B, M is the number of empirical data patterns $\{y_i\}_{i=1}^{M}$. The ambient dimension is d.

Table S.1: Main symbols and their definitions.

Symbol	Meaning	Definition
$X,\ x_i,\ \xi$	Memory set (or column matrix), individual memory, and MHN state.	Section 3.1
$x,\ y_i$	Data/diffusion state and reference data pattern.	Section 3.3
$s,\ \tau_s,\ \beta_s$	Diffusion time, Gaussian noise scale, and inverse temperature τ_s^{-2}.	Section 3.3
$E,\ \beta,\ p_i$	Hopfield energy, inverse temperature, and softmax weight of memory i.	Eqs. (1), (2)
$p_\beta,\ Z$	Boltzmann density and partition function; p_β is distinct from p_i.	Eq. (3)
$c,\ x_o,\ x_{cl}$	Cluster inner product, outlier, and representative cluster memory.	Eq. (8)
$V,\ \varepsilon,\ \Omega$	Task rotation, its amplitude, and skew-symmetric generator.	Eq. (9)
$\xi_k^*,\ \Delta E_k^{\mathrm{fp}}$	Task-1 fixed point near x_k and its rotation-averaged energy rise; (V) denotes one rotation.	Eqs. (10), (11)
$H_*,\ \mathcal{S}_*$	Fixed-point Hessian and rotation-sensitive tangential-curvature factor.	Eq. (12), Theorem 4.1
$x_r,\ p_{r\|k},\ \rho_{r\|k}$	Replay source, its softmax weight at target ξ_k^*, and susceptibility $p_{r\|k}/(1+p_{r\|k})$.	Section 5.1
$\mathcal{R}_{r\to k},\ \Delta_{r\to k}$	Raw replay gain and baseline-subtracted, rotation-averaged replay gain.	Eqs. (16), (18)
$\mathcal{F}_{t^\star},\ K,\ \mathcal{B}$	Reconstruction MSE at noise timestep $t^\star$, replay-buffer size, and selected buffer.	Eq. (15), Section 5.3
$\Delta\mathrm{MSE}_i^{(r)}$	No-replay MSE minus replay MSE for sample i under policy r.	Appendix L

B Diffusion–Hopfield Correspondence

We summarize the correspondence between diffusion energies and modern Hopfield energies. Consider a forward noising SDE

$$dY_s = f(Y_s, s)\, ds + g(s)\, d\widehat{W}_s, \tag{21}$$

and let $p_s(x)$ be the marginal density of Y_s. For additive noise and conservative drift, the reverse-time SDE can be written as

$$dX_t = \left[g^2(T-t)\nabla_x \log p_{T-t}(X_t) - f(X_t, T-t)\right] dt + g(T-t)dW_t. \tag{22}$$

Equivalently,

$$dX_t = -\nabla_x E_{T-t}(X_t)\,dt + g(T-t)dW_t, \tag{23}$$

where the implicit energy is

$$E_s(x) = -g^2(s)\log p_s(x) + \int_0^x f(z,s)\cdot dz, \tag{24}$$

up to an additive constant [19]. To examine the attractor structure at a fixed diffusion time s, we freeze the time-dependent coefficients $g(s)$ and $f(\cdot,s)$ and study the deterministic drift part of the reverse SDE. That is, we drop only the Brownian increment $g(s)dW_t$, while keeping the $g^2(s)\nabla_x \log p_s$ term in the drift. This frozen-time flow is a landscape diagnostic: $\tau_s \to 0$ does not by itself imply $g(s) \to 0$ or exact convergence of the full stochastic reverse process. Under the conservative-drift assumption above, $f(\cdot,s) = \nabla\Phi_s(\cdot)$ for a scalar potential Φ_s, so the line integral in E_s is path independent. Along the deterministic flow

$$\frac{dx_\lambda}{d\lambda} = -\nabla_x E_s(x_\lambda), \tag{25}$$

the energy decreases as

$$\frac{d}{d\lambda}E_s(x_\lambda) = -\|\nabla_x E_s(x_\lambda)\|^2 \le 0, \tag{26}$$

so stable equilibria satisfying $\nabla_x E_s(x) = 0$ define the attractors of this frozen-time flow.

Now assume that the data distribution is the empirical distribution over a finite set of patterns,

$$p_0(y) = \frac{1}{M}\sum_{i=1}^{M}\delta(y-y_i). \tag{27}$$

In the variance-exploding case,

$$dY_s = g(s)d\widehat{W}_s, \qquad f=0, \qquad \tau_s^2 = \int_0^s g^2(r)\,dr. \tag{28}$$

Thus

$$Y_s \mid y_i \sim \mathcal{N}(y_i, \tau_s^2 I), \tag{29}$$

and the noisy marginal is the Gaussian mixture

$$p_s(x) \propto \sum_i \exp\left(-\frac{\|x-y_i\|^2}{2\tau_s^2}\right). \tag{30}$$

Expanding the square gives

$$p_s(x) \propto \exp\left(-\frac{\|x\|^2}{2\tau_s^2}\right)\sum_i \exp\left(\tau_s^{-2}y_i^\top x - \frac{\|y_i\|^2}{2\tau_s^2}\right). \tag{31}$$

When the memories are normalized, $\|y_i\| = r$, the last term is independent of i and can be absorbed into the normalization. Since $f=0$, we have $E_s(x) = -g^2(s)\log p_s(x)$. Multiplying by the positive factor $\tau_s^2/g^2(s)$, which does not change the equilibria, yields

$$-\tau_s^2 \log p_s(x) = -\tau_s^2 \log\sum_i \exp(\tau_s^{-2}y_i^\top x) + \frac{1}{2}\|x\|^2 + \text{const}. \tag{32}$$

With $\beta_s = \tau_s^{-2}$, this becomes

$$E_{\text{MHN}}(x;\beta_s) = -\frac{1}{\beta_s}\log\sum_{i=1}^{M}\exp(\beta_s y_i^\top x) + \frac{1}{2}\|x\|^2 + \text{const}, \tag{33}$$

which is the continuous modern Hopfield energy [17]. Thus, for discrete normalized patterns, the Gaussian-mixture marginal induced by the forward diffusion defines the same log-sum-exp energy landscape as a modern Hopfield network.

For variance-preserving diffusions, the forward SDE is typically

$$dY_s = -\frac{1}{2}b(s)Y_s\,ds + \sqrt{b(s)}\,d\widehat{W}_s. \tag{34}$$

The transition kernel is still Gaussian,

$$Y_s \mid y_i \sim \mathcal{N}(\theta_s y_i, \tau_s^2 I), \qquad \theta_s = \exp\left(-\frac{1}{2}\int_0^s b(r)\,dr\right), \qquad \tau_s^2 = 1-\theta_s^2. \tag{35}$$

Therefore p_s is again a Gaussian mixture, now centered at the shrunk memories $\theta_s y_i$. Substituting this marginal into the implicit energy gives, for normalized memories,

$$\frac{E_s(x)}{b(s)} = -\log\sum_i \exp\left(\frac{\theta_s}{\tau_s^2}y_i^\top x\right) + \left(\frac{1}{2\tau_s^2}-\frac{1}{4}\right)\|x\|^2 + \text{const.} \tag{36}$$

Thus the variance-preserving case differs at finite noise by pattern shrinkage and time-dependent coefficients. In the low-noise regime, $\theta_s \to 1$ and $\tau_s \to 0$, this reduces to the same log-sum-exp plus quadratic structure, with effective inverse temperature proportional to θ_s/τ_s^2. More generally, for additive SDEs with conservative drift, the low-noise transition kernel is locally Gaussian around each pattern, so the same Gaussian-mixture argument gives the Hopfield fixed-point structure asymptotically [17].

C Full derivation of Theorem 4.1

We organize the proof as follows. Subsection C.1 states the equal-angular cluster-plus-outlier model and the rotation identity. Subsection C.2 reduces the fixed-point equations by symmetry. Subsection C.3 separates the existence of pattern-specific fixed points from the stronger localization condition used to control leakage corrections. Subsection C.4 derives the large-β fixed-point asymptotics. Subsection C.5 collects the Hessian calculation, the rotation average, and the first-order expansion in the off-memory weights that produce the formula for ΔE_*^{fp}. Finally, Subsection C.6 specializes that formula to the outlier and cluster fixed points and proves Theorem 4.1.

C.1 Problem setup and rotation identity

Let $x_1, \dots, x_N, x_o \in \mathbb{S}^{d-1}$ satisfy (8); we assume $N \geq 2$, $0 < c < 1$, and $d \geq N+1$, as required by the exact Gram matrix. We fix N, d, c, take β sufficiently large, and then ε sufficiently small for that β. Random normalized approximations are treated separately in Appendix G; they are not exact realizations of these assumptions. The MHN energy and update rule are

$$E(\xi \mid X) = -\tfrac{1}{\beta}\log\sum_{i=1}^{N+1} e^{\beta x_i^\top \xi} + \tfrac{1}{2}\|\xi\|^2, \qquad T(\xi) = \sum_i p_i(\xi)x_i, \qquad p_i(\xi) = \frac{e^{\beta x_i^\top \xi}}{\sum_k e^{\beta x_k^\top \xi}}, \tag{37}$$

so that a fixed point $\xi_* = T(\xi_*)$ satisfies $\xi_* = Xp_*$ with $p_* = \mathrm{softmax}(\beta X^\top \xi_*)$. The rotation $X_2 = VX$ induces

$$E(\xi \mid VX) = E(V^\top \xi \mid X), \tag{38}$$

reducing Task-2 evaluation at ξ to Task-1 evaluation at $V^\top \xi$.

C.2 Symmetry reduction of the fixed-point equations

Define

$$s := \sum_{i=1}^N x_i, \qquad A := 1 + (N-1)c. \tag{39}$$

A direct calculation using (8) gives

$$x_i^\top s = A \quad (i \leq N), \qquad x_o^\top s = 0, \qquad \|s\|^2 = NA. \tag{40}$$

Outlier-like fixed point. The outlier-like fixed point is invariant under permutations of the cluster members, so its softmax weights are

$$p^{\text{out}} = (u, \dots, u, v), \qquad Nu + v = 1, \tag{41}$$

and $\xi_{\text{out}} = u\, s + v\, x_o$. The logits read $x_i^\top \xi_{\text{out}} = Au$ for $i \le N$ and $x_o^\top \xi_{\text{out}} = v$. Substituting into the softmax,

$$u = \frac{e^{\beta A u}}{N e^{\beta A u} + e^{\beta v}}, \qquad v = \frac{e^{\beta v}}{N e^{\beta A u} + e^{\beta v}}, \tag{42}$$

and taking the log-ratio together with $u = (1 - v)/N$ yields the scalar self-consistency

$$\log \frac{Nv}{1 - v} = \beta\Big(v - \frac{A}{N}(1 - v)\Big). \tag{43}$$

Cluster-like fixed point near x_1. Permuting $x_2, \dots, x_N$ leaves the cluster-like fixed point invariant, so

$$p^{cl} = (a, b, \dots, b, q), \qquad a + (N - 1)b + q = 1, \tag{44}$$

and $\xi_{cl} = a\, x_1 + b \sum_{j \ge 2} x_j + q\, x_o$. The three distinct logits are

$$z_1 = a + (N - 1)cb, \qquad z_- = ca + \{1 + (N - 2)c\}b, \qquad z_o = q. \tag{45}$$

The softmax equations and the identity $z_1 - z_- = (1 - c)(a - b)$ give the two pattern-selection equations

$$\log \frac{a}{b} = \beta(1 - c)(a - b), \qquad \log \frac{a}{q} = \beta(z_1 - q). \tag{46}$$

The first is a Curie–Weiss-type fixed-point equation with effective inverse temperature

$$\beta_{\text{eff}} := \beta(1 - c), \tag{47}$$

matching the effective inverse temperature in the equal-angular model of Koulischer et al. [34].

C.3 Existence threshold and localization condition

We separate two requirements: (i) *existence* of pattern-specific cluster fixed points, and (ii) the stronger *localization condition* that keeps the leakage corrections small in the Hessian expansion below.

C.3.1 Existence threshold

Drop the outlier coordinate temporarily and consider the symmetric cluster sector $p^{cl} = (a, b, \dots, b)$ with $a + (N - 1)b = 1$. Introduce the order parameter

$$m := a - b, \qquad a = \frac{1 + (N - 1)m}{N}, \qquad b = \frac{1 - m}{N}. \tag{48}$$

The pattern-selection equation (46) becomes

$$\log \frac{1 + (N - 1)m}{1 - m} = \beta_{\text{eff}}\, m, \qquad \beta_{\text{eff}} = \beta(1 - c). \tag{49}$$

For this auxiliary cluster-only problem, the onset is the infimum of the slope from the origin to the graph:

$$\beta_c(N) := \inf_{m \in (0,1)} \frac{1}{m} \log \frac{1 + (N - 1)m}{1 - m}, \tag{50}$$

For $N > 2$, the infimum is attained at an interior point m_c satisfying

$$\log \frac{1 + (N - 1)m_c}{1 - m_c} = \frac{N\, m_c}{(1 - m_c)\{1 + (N - 1)m_c\}}, \qquad \beta_c(N) = \frac{N}{(1 - m_c)\{1 + (N - 1)m_c\}}. \tag{51}$$

Equivalently, with $t_c := \{1 + (N - 1)m_c\}/(1 - m_c)$,

$$N\, t_c \log t_c = (t_c - 1)(t_c + N - 1), \qquad \beta_c(N) = \frac{(t_c + N - 1)^2}{N\, t_c}. \tag{52}$$

For $N = 2$, the infimum is $\beta_c(2) = 2$ at $m \downarrow 0$, and nonzero fixed points emerge for $\beta_{\text{eff}} > 2$. For $N > 2$, the threshold is implicit; elementary bounds on $\log t_c$ yield

$$\log N + 1 \le \beta_c(N) \le \log N + \log(1 + \log N) + 1 + \frac{\log(1 + \log N)}{\log N}, \tag{53}$$

so $\beta_c(N) = \Theta(\log N)$, and as $N \to \infty$,

$$\beta_c(N) = \log N + \log\log N + 1 + o(1). \tag{54}$$

Pattern-specific fixed points of this cluster-only problem exist above the auxiliary threshold

$$\beta(1 - c) > \beta_c(N). \tag{55}$$

For $N > 2$, a marginal nonzero root also exists at equality. This is not an exact existence criterion for the full cluster-plus-outlier model, whose outlier weight is positive at finite β. Existence and exclusion of other stable fixed points in the full model follow from Proposition E.1.

C.3.2 Leakage and the localization condition

Existence does not by itself bound the size of corrections in the Hessian expansion below. We use a single auxiliary quantity, the *leakage*: the total softmax mass assigned to memories other than the dominant one. For the outlier and cluster fixed points, define

$$\eta_{\text{out}} := 1 - v = Nu, \qquad \eta_{cl} := 1 - a = (N - 1)b + q. \tag{56}$$

Using $s \perp x_o$ for the outlier and the explicit form of the cluster fixed point,

$$\xi_{\text{out}} - x_o = u\, s - (1 - v)\, x_o = \eta_{\text{out}}\Big(\tfrac{s}{N} - x_o\Big), \qquad \xi_{cl} - x_1 = b\sum_{j=2}^{N}(x_j - x_1) + q(x_o - x_1), \tag{57}$$

so $\|\xi_{\text{out}} - x_o\| \le C_{\text{out}}\eta_{\text{out}}$ and $\|\xi_{cl} - x_1\| \le C_{cl}\eta_{cl}$ with c-dependent constants. The condition must control $\beta\eta$, not only η, because the leakage enters the Hessian multiplied by β. Indeed, if $p_k^* = 1 - \sum_{j\neq k}\eta_j$ and $p_j^* = \eta_j$, then

$$\sum_i p_i^* x_i x_i^\top - \xi_*\xi_*^\top = \sum_{j\neq k}\eta_j(x_j - x_k)(x_j - x_k)^\top + O(\eta^2), \tag{58}$$

and hence $H_* = I - \beta\sum_{j\neq k}\eta_j(x_j - x_k)(x_j - x_k)^\top + O(\beta\eta^2)$. Thus the working localization condition is

$$\beta\,\eta_{\text{out}} \ll 1, \qquad \beta\,\eta_{cl} \ll 1. \tag{59}$$

The next two paragraphs derive the corresponding outlier and cluster forms of this same condition.

Outlier-like fixed point. Set $\delta_{\text{out}} := 1 - v = \eta_{\text{out}}$ so $u = \delta_{\text{out}}/N$. Substituting into (43),

$$\log\frac{N(1 - \delta_{\text{out}})}{\delta_{\text{out}}} = \beta\Big[1 - \big(1 + \tfrac{A}{N}\big)\delta_{\text{out}}\Big], \tag{60}$$

or in fixed-point form

$$\frac{\delta_{\text{out}}}{1 - \delta_{\text{out}}} = N\,e^{-\beta}\exp\big[\beta(1 + \tfrac{A}{N})\,\delta_{\text{out}}\big]. \tag{61}$$

If $\beta N e^{-\beta} \ll 1$, the right-hand side is the contraction $Ne^{-\beta}\,(1 + O(\beta\delta_{\text{out}}))$ and the leading balance reads $\delta_{\text{out}} = Ne^{-\beta}(1 + o(1))$. Hence the outlier form of the localization condition is

$$\beta N\, e^{-\beta} \ll 1, \tag{62}$$

or, for $N \to \infty$, $\beta - \log N - \log\beta \to +\infty$. Under (62),

$$v = 1 - N\,e^{-\beta}(1 + o(1)), \qquad u = e^{-\beta}(1 + o(1)). \tag{63}$$

Cluster-like fixed point. Equation (46) gives $b/a = \exp[-\beta(1-c)(a-b)]$. Writing $a = 1-\eta_{cl}$ and $a-b = 1-\eta_{cl}-b$,

$$b = a\, e^{-\beta(1-c)} \exp\big[\beta(1-c)(\eta_{cl}+b)\big]. \tag{64}$$

Similarly, from $\log(a/q) = \beta(z_1 - q)$ with $z_1 = a + (N-1)cb$,

$$q = a\, e^{-\beta} \exp\big[\beta\{(1-c)(N-1)b + 2q\}\big]. \tag{65}$$

The leading values of these off-memory weights are therefore $(N-1)\, e^{-\beta(1-c)}$ from the other cluster memories and $e^{-\beta}$ from the outlier. The exponential feedbacks in (64)–(65) are negligible iff the cluster form of the localization condition holds:

$$\beta(1-c)(N-1)\, e^{-\beta(1-c)} \ll 1, \qquad \beta\, e^{-\beta} \ll 1. \tag{66}$$

Under (66),

$$\begin{aligned} &b = e^{-\beta(1-c)}(1+o(1)), \qquad q = e^{-\beta}(1+o(1)), \\ &a = 1 - (N-1)\, e^{-\beta(1-c)} - e^{-\beta} + o\big((N-1)\, e^{-\beta(1-c)} + e^{-\beta}\big). \end{aligned} \tag{67}$$

Equations (62) and (66) are simply the outlier and cluster forms of the same localization condition $\beta\eta \ll 1$. Since $e^{-\beta(1-c)} \gg e^{-\beta}$ for $0 < c < 1$, the binding cluster constraint is $\beta(1-c)(N-1)\, e^{-\beta(1-c)} \ll 1$, equivalently $\beta(1-c) - \log N - \log\beta(1-c) \to +\infty$, i.e.

$$\beta(1-c) \geq \log N + \log\log N + \omega(1). \tag{68}$$

C.3.3 Working assumption $\beta \gg \log N$

The capacity threshold $\beta_c(N) = \log N + \log\log N + 1 + o(1)$ obtained in (54) is the order at which the pattern-specific phase transition occurs. The boundary scale $\beta \sim \log N$ is mathematically interesting and has been the focus of single-task associative-memory analyses, which study how patterns emerge as β crosses this threshold. The present paper, however, is about CL — a more involved setting in which we want to study forgetting and replay at a Task-1 fixed point that is already cleanly localized. We therefore simplify by working in the strictly stronger regime

$$\beta \gg \log N, \tag{69}$$

which directly implies the working condition

$$\beta\, N\, e^{-\beta} \ll 1. \tag{70}$$

For fixed $0 < c < 1$, $\beta \gg \log N$ also implies $\beta(1-c) \gg \log N$, hence $\beta(1-c)(N-1)\, e^{-\beta(1-c)} \ll 1$, and we obtain the inclusion chain

$$\big\{\beta \gg \log N\big\} \;\subset\; \big\{\beta(1-c) \geq \log N + \log\log N + \omega(1)\big\} \;\subset\; \big\{\beta(1-c) > \beta_c(N)\big\}. \tag{71}$$

The strict inclusions are useful: the outermost set is the existence threshold; the middle set enforces the localization condition; the innermost set is what we adopt here. Under (69), every Task-1 pattern-specific fixed point is localized around its corresponding memory and the Hessian expansion in the leakage variables is valid.

Assumption C.1 (Working regime)**.** For the strict orderings, fix $N \geq 2$, $d \geq N+1$, $0 < c < 1$, and $\delta > 0$, take β sufficiently large with $\beta(1-c) \geq (1+\delta)N$, and then take ε sufficiently small for each such β. The exclusion of other stable fixed points is proved in Appendix E; localization alone does not exclude stable mixed states in arbitrary joint limits.

C.4 Large- β asymptotics of the Task-1 fixed points

We collect the closed-form expansions used in the body.

Proposition C.2 (Outlier-like fixed point)**.** *Under (62),*

$$v = 1 - N\, e^{-\beta} + O(\beta\, e^{-2\beta}), \qquad u = e^{-\beta} + O(\beta\, e^{-2\beta}), \tag{72}$$

hence

$$\xi_{\text{out}} = x_o + e^{-\beta}(s - N\, x_o) + O(\beta\, e^{-2\beta}). \tag{73}$$

Proof. Set $v = 1 - \delta$ and substitute into (43): $\log\{N(1-\delta)/\delta\} = \beta\{1 - \delta - A\delta/N\}$. As $\beta \to \infty$, $\delta \to 0$, so the left-hand side equals $\log(N/\delta) + O(\delta)$ and the leading balance reads $\log(N/\delta) = \beta(1 - O(\delta))$, giving $\delta = N\,e^{-\beta}(1 + o(1))$. The next-order correction is $O(\beta\delta^2) = O(\beta\,e^{-2\beta})$. Substituting $u = \delta/N$ into $\xi_{\text{out}} = u\,s + v\,x_o$ yields (73). □

Proposition C.3 (Cluster-like fixed point)**.** *Under Condition C.1,*

$$b = e^{-\beta(1-c)}(1+o(1)), \qquad q = e^{-\beta}(1+o(1)), \qquad a = 1-(N-1)\,e^{-\beta(1-c)}-e^{-\beta}+o\big(e^{-\beta(1-c)}\big), \tag{74}$$

hence

$$\xi_{cl} = x_1 + e^{-\beta(1-c)}\Big(\sum_{j=2}^{N} x_j - (N-1)x_1\Big) + e^{-\beta}\,(x_o - x_1) + o\big(e^{-\beta(1-c)}\big). \tag{75}$$

Proof. Since $\beta\eta_{cl} \to 0$ under the localization condition, the first equation of (46) gives $\log(a/b) = \beta(1-c)(a-b) = \beta(1-c) + o(1)$, so $b = a\,e^{-\beta(1-c)}(1 + o(1)) = e^{-\beta(1-c)}(1 + o(1))$. The second gives $\log(a/q) = \beta(a + (N-1)cb - q) = \beta + o(1)$, so $q = e^{-\beta}(1 + o(1))$. The sum rule $a+(N-1)b+q = 1$ closes the argument. Substituting into $\xi_{cl} = a\,x_1 + b\sum_{j\geq 2} x_j + q\,x_o$ yields (75). The same scalar equations give $b = e^{-\beta(1-c)} + O(\beta e^{-2\beta(1-c)})$ and $q = e^{-\beta} + O(\beta e^{-\beta(2-c)})$ at fixed N, c. □

Remark C.4. The two leading off-memory weights differ exponentially: the outlier sees $u \sim e^{-\beta}$, while the cluster sees $b \sim e^{-\beta(1-c)}$, and $e^{-\beta(1-c)}/e^{-\beta} = e^{\beta c} \to \infty$. Geometrically, *the cluster fixed point is exponentially broader than the outlier fixed point*; this is the geometric origin of Theorem 4.1.

C.5 Hessian expansion for the rotation-averaged energy rise

This subsection collects the Hessian-related calculations used in the proof: the Taylor expansion at a fixed point, the average over small random rotations, and the first-order expansion in the off-memory weights.

Taylor expansion at a Task-1 fixed point. Because ξ_* is stationary ($\nabla E(\xi_* \mid X) = 0$), Taylor expansion gives

$$\Delta E_*^{\text{fp}}(V) = \tfrac{1}{2}(V^\top\xi_* - \xi_*)^\top H_*(V^\top\xi_* - \xi_*) + O(\varepsilon^3), \qquad H_* := \nabla^2 E(\xi_* \mid X). \tag{76}$$

Direct differentiation of (37) yields

$$H_* = I - \beta\Big(\sum_{i=1}^{N+1} p_i^* x_i x_i^\top - \xi_*\xi_*^\top\Big). \tag{77}$$

Using $V^\top\xi_* - \xi_* = -\varepsilon\Omega\xi_* + O(\varepsilon^2)$,

$$\Delta E_*^{\text{fp}}(V) = \tfrac{\varepsilon^2}{2}(\Omega\xi_*)^\top H_*(\Omega\xi_*) + O(\varepsilon^3). \tag{78}$$

The covariance term in Eq. (77) also explains the $\beta\eta$ scale used in the localization condition. Near a pattern-specific fixed point with $p_k^* = 1 - \sum_{j\neq k}\eta_j$ and $p_j^* = \eta_j$, it has the expansion

$$\sum_i p_i^* x_i x_i^\top - \xi_*\xi_*^\top = \sum_{j\neq k}\eta_j(x_j - x_k)(x_j - x_k)^\top + O(\eta^2). \tag{79}$$

Thus leakage contributes to the Hessian at order $\beta\eta$, which is why Eq. (59) controls $\beta\eta$ rather than merely η.

Rotation average.

Lemma C.5. *For any $u \in \mathbb{R}^d$, $\langle(\Omega u)(\Omega u)^\top\rangle_V = \frac{1}{2}(\|u\|^2 I - uu^\top)$.*

Proof. From $\Omega = \frac{1}{2}(W - W^\top)$ with $W_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, $\mathbb{E}[\Omega_{ab}\Omega_{cd}] = \frac{1}{2}(\delta_{ac}\delta_{bd} - \delta_{ad}\delta_{bc})$. Hence $\langle(\Omega u)_i(\Omega u)_j\rangle_V = \frac{1}{2}(\delta_{ij}\|u\|^2 - u_i u_j)$. □

Substituting into (78) and using the main-text convention $\Delta E_*^{\text{fp}} := \langle\Delta E_*^{\text{fp}}(V)\rangle_V$ from (11),

$$\Delta E_*^{\text{fp}} = \tfrac{\varepsilon^2}{4}(\|\xi_*\|^2\,\mathrm{tr}\,H_* - \xi_*^\top H_*\xi_*) + O(\varepsilon^3). \tag{80}$$

First-order expansion near a pattern. Suppose ξ_* is concentrated at a single stored pattern x_k, with weights

$$p_k^* = 1 - \sum_{j \neq k} \eta_j, \qquad p_j^* = \eta_j \ll 1, \qquad \alpha_j := x_j^\top x_k. \tag{81}$$

The fixed-point identity $\xi_* = X p_*$ then reads

$$\xi_* = x_k + \sum_{j \neq k} \eta_j (x_j - x_k). \tag{82}$$

We compute the three quantities entering (80) step by step, retaining all $O(\eta)$ terms and dropping $O(\eta^2)$. Here $\eta := \sum_{j \neq k} \eta_j$; constants in the leakage remainders may depend on fixed N, d, c, while averaged rotation remainders are $O_{\beta,N,d,c}(\varepsilon^3)$ as $\varepsilon \to 0$ at fixed β. The subscripts mean that the implicit constant may depend on β, N, d, c, but not on ε.

(i) $\|\xi_*\|^2$**.** Using $\|x_k\|^2 = 1$, $x_k^\top (x_j - x_k) = \alpha_j - 1$, and $(x_i - x_k)^\top (x_j - x_k) = O(1)$,

$$\|\xi_*\|^2 = 1 + 2 \sum_{j \neq k} \eta_j (\alpha_j - 1) + O(\eta^2) = 1 - 2 \sum_{j \neq k} \eta_j (1 - \alpha_j) + O(\eta^2). \tag{83}$$

(ii) $\operatorname{tr} H_*$**.** From (77), using $\|x_i\|^2 = 1$ and $\sum_i p_i^* = 1$,

$$\operatorname{tr} H_* = d - \beta \Big(\sum_{i=1}^{N+1} p_i^* \, \|x_i\|^2 - \|\xi_*\|^2 \Big) = d - \beta \big(1 - \|\xi_*\|^2 \big), \tag{84}$$

and substituting (83),

$$\operatorname{tr} H_* = d - 2\beta \sum_{j \neq k} \eta_j (1 - \alpha_j) + O(\beta \eta^2). \tag{85}$$

(iii) $\xi_*^\top H_* \xi_*$**.** From (77),

$$\xi_*^\top H_* \xi_* = \|\xi_*\|^2 - \beta \Big(\sum_i p_i^* \, (x_i^\top \xi_*)^2 - \|\xi_*\|^4 \Big). \tag{86}$$

Using (82), $x_k^\top \xi_* = 1 - \sum_{j \neq k} \eta_j (1 - \alpha_j)$, and $x_j^\top \xi_* = \alpha_j + O(\eta)$ for $j \neq k$. Therefore

$$\begin{aligned} \sum_i p_i^* (x_i^\top \xi_*)^2 &= \Big(1 - \sum_{j \neq k} \eta_j \Big) \Big(1 - \sum_{j \neq k} \eta_j (1 - \alpha_j) \Big)^2 + \sum_{j \neq k} \eta_j \, \alpha_j^2 + O(\eta^2) \\ &= 1 - 2 \sum_{j \neq k} \eta_j (1 - \alpha_j) - \sum_{j \neq k} \eta_j (1 - \alpha_j^2) + O(\eta^2), \end{aligned} \tag{87}$$

where in the last step we used $-\sum_j \eta_j + \sum_j \eta_j \alpha_j^2 = -\sum_j \eta_j (1 - \alpha_j^2)$ and absorbed cross-products of η's into $O(\eta^2)$. Combining (87) with $\|\xi_*\|^4 = 1 - 4 \sum_{j \neq k} \eta_j (1 - \alpha_j) + O(\eta^2)$ from (83) gives

$$\sum_i p_i^* (x_i^\top \xi_*)^2 - \|\xi_*\|^4 = \sum_{j \neq k} \eta_j \big\{ 2(1 - \alpha_j) - (1 - \alpha_j^2) \big\} + O(\eta^2) = \sum_{j \neq k} \eta_j (1 - \alpha_j)^2 + O(\eta^2), \tag{88}$$

using the algebraic identity $2(1 - \alpha) - (1 - \alpha^2) = (1 - \alpha)^2$. Substituting (88) into (86),

$$\xi_*^\top H_* \xi_* = 1 - 2 \sum_{j \neq k} \eta_j (1 - \alpha_j) - \beta \sum_{j \neq k} \eta_j (1 - \alpha_j)^2 + O((1 + \beta) \eta^2). \tag{89}$$

Substituting (83), (85), and (89) into (80), at order η,

$$\|\xi_*\|^2 \operatorname{tr} H_* - \xi_*^\top H_* \xi_* = (d-1) - 2(\beta + d - 1) \sum_{j \neq k} \eta_j (1 - \alpha_j) + \beta \sum_{j \neq k} \eta_j (1 - \alpha_j)^2 \tag{90}$$

$$+ O((1+\beta)\eta^2) \tag{91}$$

$$= (d-1) - \sum_{j \neq k} \eta_j \Big\{ 2(d-1)(1-\alpha_j) + \beta(1-\alpha_j)\big[2 - (1-\alpha_j)\big] \Big\} \tag{92}$$

$$+ O((1+\beta)\eta^2) \tag{93}$$

$$= (d-1) - \sum_{j \neq k} \eta_j \big\{ 2(d-1)(1-\alpha_j) + \beta(1-\alpha_j^2) \big\} \tag{94}$$

$$+ O((1+\beta)\eta^2), \tag{95}$$

where the last line uses the identity $(1-\alpha)\left[2-(1-\alpha)\right] = (1-\alpha)(1+\alpha) = 1-\alpha^2$. Therefore

$$\begin{aligned} \Delta E_*^{\mathrm{fp}} &= \tfrac{\varepsilon^2}{4} \Big[(d-1) - \sum_{j \neq k} \eta_j \, \Lambda(\alpha_j) \Big] \\ &\quad + O(\varepsilon^2 (1+\beta)\eta^2) + O_{\beta,N,d,c}(\varepsilon^3), \\ \Lambda(\alpha) &:= 2(d-1)(1-\alpha) + \beta(1-\alpha^2). \end{aligned} \tag{96}$$

Each competitor at cosine α_j shaves $\eta_j \Lambda(\alpha_j)$ off the bare $(d-1)$. The geometric weight $\Lambda(\alpha)$ vanishes at $\alpha = 1$ (perfect alignment, no help) and is maximal at $\alpha = 0$ (orthogonal competitor, maximum reduction).

C.6 Specialization and proof of Theorem 4.1

This subsection applies the Hessian formula to the two Task-1 fixed points and then compares the resulting energy rises.

Outlier-like fixed point. For the outlier-like fixed point all N competitors are orthogonal: $\alpha_j = 0$ and $\eta_j = u = e^{-\beta}(1 + o(1))$ by Proposition C.2. Substituting into (96),

$$\Delta E_o^{\mathrm{fp}} = \tfrac{\varepsilon^2}{4} \big[(d-1) - N \, e^{-\beta} \{ 2(d-1) + \beta \} \big] + O(\varepsilon^2 \beta^2 e^{-2\beta}) + O_{\beta,N,d,c}(\varepsilon^3). \tag{97}$$

Cluster-like fixed point. For the cluster-like fixed point near x_1 there are two competitor types (Proposition C.3): the other $N-1$ cluster members with $\alpha = c$, $\eta = b = e^{-\beta(1-c)}(1 + o(1))$, and the outlier with $\alpha = 0$, $\eta = q = e^{-\beta}(1 + o(1))$. Substituting into (96),

$$\begin{aligned} \Delta E_{cl}^{\mathrm{fp}} = \tfrac{\varepsilon^2}{4} \Big[&(d-1) - (N-1) \, e^{-\beta(1-c)} \{ 2(d-1)(1-c) + \beta(1-c^2) \} \\ &- e^{-\beta} \{ 2(d-1) + \beta \} \Big] + O(\varepsilon^2 \beta^2 e^{-2\beta(1-c)}) + O_{\beta,N,d,c}(\varepsilon^3). \end{aligned} \tag{98}$$

Comparison. Subtracting (98) from (97) with $\Lambda(\alpha) = 2(d-1)(1-\alpha) + \beta(1-\alpha^2)$,

$$\Delta E_o^{\mathrm{fp}} - \Delta E_{cl}^{\mathrm{fp}} = \tfrac{\varepsilon^2}{4} (N-1) \big[e^{-\beta(1-c)} \, \Lambda(c) - e^{-\beta} \, \Lambda(0) \big] + O(\varepsilon^2 \beta^2 e^{-2\beta(1-c)}) + O_{\beta,N,d,c}(\varepsilon^3). \tag{99}$$

At fixed N, d, c with $0 < c < 1$, the leading coefficient is positive for sufficiently large β, and its relative leakage error is $O(\beta e^{-\beta(1-c)}) \to 0$. For each such β, take ε sufficiently small to control the rotation remainder. Proposition E.1 and cluster symmetry then extend the comparison to all stable fixed points. □

Remark C.6 (Edge cases). At $c = 0$, all $N+1$ memories are mutually orthogonal. Their memory-localized fixed points remain distinct but are symmetry-equivalent, so their energies and rotation-averaged forgetting are equal. The strict theorem concerns $0 < c < 1$; negative correlations are outside its scope.

D Energy–Sharpness Relation

In the main text, we describe isolated memories as forming high-energy and sharp basins, and cluster-supported memories as forming low-energy and broad basins. This appendix makes that statement precise in the equal-angular cluster-plus-outlier model. The key point is that both the local energy level and the local sharpness of a fixed point are controlled by the same off-memory softmax weights.

D.1 Definition of local sharpness

Let ξ_* be a Task-1 fixed point of the Hopfield energy

$$E(\xi \mid X) = -\frac{1}{\beta} \log \sum_i \exp(\beta x_i^\top \xi) + \frac{1}{2}\|\xi\|^2. \tag{100}$$

The Hessian at ξ_* is

$$H_* = \nabla_\xi^2 E(\xi_* \mid X) = I - \beta \left(\sum_i p_i^* x_i x_i^\top - \xi_* \xi_*^\top \right), \tag{101}$$

where $p_i^* = p_i(\xi_* \mid X)$. We define the local sharpness of the basin by the average Hessian trace in random tangential directions induced by an infinitesimal rotation,

$$\mathcal{S}_* := \|\xi_*\|^2 \operatorname{tr} H_* - \xi_*^\top H_* \xi_*. \tag{102}$$

Indeed, for $V = \exp(\varepsilon\Omega)$ with $\Omega = (W - W^\top)/2$ and $W_{ij} \sim \mathcal{N}(0,1)$, one has

$$\mathbb{E}_\Omega \left[(\Omega\xi_*)^\top H_* (\Omega\xi_*) \right] = \frac{1}{2} \left(\|\xi_*\|^2 \operatorname{tr} H_* - \xi_*^\top H_* \xi_* \right) = \frac{1}{2}\mathcal{S}_*. \tag{103}$$

Thus $\mathcal{S}_*$ is the curvature factor that appears in the rotation-induced energy increase:

$$\mathbb{E}_\Omega \left[E(V^\top \xi_* \mid X) - E(\xi_* \mid X) \right] = \frac{\varepsilon^2}{4}\mathcal{S}_* + O(\varepsilon^3). \tag{104}$$

Larger $\mathcal{S}_*$ means that the basin rises more rapidly under small displacements, and we therefore use it as the local sharpness of the Hopfield basin.

D.2 First-order expansion of energy and sharpness

Consider a fixed point concentrated near a stored memory x_k. Write the softmax weights as

$$p_k^* = 1 - \sum_{j \neq k} \eta_j, \qquad p_j^* = \eta_j, \qquad \eta_j \ll 1, \tag{105}$$

and denote the cosine similarity between the competitor x_j and the dominant memory x_k by

$$\alpha_j := x_j^\top x_k. \tag{106}$$

The fixed-point condition gives

$$\xi_* = x_k + \sum_{j \neq k} \eta_j (x_j - x_k) + O(\eta^2). \tag{107}$$

First, the energy at the fixed point has the expansion

$$E(\xi_* \mid X) = -\frac{1}{2} - \frac{1}{\beta} \sum_{j \neq k} \eta_j + O(\eta^2). \tag{108}$$

Thus a fixed point supported by more nearby memories has lower energy. In this sense, the total leakage $\sum_{j \neq k} \eta_j$ measures how much additional support the dominant memory receives from neighboring memories.

Second, the sharpness defined in Eq. (102) satisfies

$$\mathcal{S}_* = (d-1) - \sum_{j \neq k} \eta_j \Lambda(\alpha_j) + O((1+\beta)\eta^2), \tag{109}$$

where

$$\Lambda(\alpha) := 2(d-1)(1-\alpha) + \beta(1-\alpha^2). \tag{110}$$

Since $\Lambda(\alpha) > 0$ for $\alpha < 1$, leakage from neighboring memories lowers the local curvature. Therefore the same off-memory softmax mass that lowers the energy also flattens the basin. This is the precise local sense in which low-energy basins are broad, whereas high-energy basins are sharp.

D.3 Cluster-plus-outlier specialization

We now apply Eqs. (108) and (109) to the equal-angular configuration of Section 4.1. For the outlier fixed point, all N competitors are orthogonal, so $\alpha = 0$ and $\eta = e^{-\beta}(1 + o(1))$. Therefore,

$$E_o^* = -\frac{1}{2} - \frac{N}{\beta} e^{-\beta} + o(e^{-\beta}/\beta), \tag{111}$$

and

$$\mathcal{S}_o = (d-1) - N e^{-\beta} \Lambda(0) + o(e^{-\beta}). \tag{112}$$

For the cluster fixed point near x_1, the $N-1$ other cluster memories have cosine c and leakage $e^{-\beta(1-c)}$, while the outlier has cosine 0 and leakage $e^{-\beta}$. Thus,

$$E_{cl}^* = -\frac{1}{2} - \frac{1}{\beta}\left\{(N-1)e^{-\beta(1-c)} + e^{-\beta}\right\} + o(e^{-\beta(1-c)}/\beta), \tag{113}$$

and

$$\mathcal{S}_{cl} = (d-1) - (N-1)e^{-\beta(1-c)}\Lambda(c) - e^{-\beta}\Lambda(0) + o(e^{-\beta(1-c)}). \tag{114}$$

Subtracting Eq. (113) from Eq. (111) gives

$$E_o^* - E_{cl}^* = \frac{N-1}{\beta}\left(e^{-\beta(1-c)} - e^{-\beta}\right) + o(e^{-\beta(1-c)}/\beta) > 0, \tag{115}$$

because $0 < c < 1$ implies $e^{-\beta(1-c)} \gg e^{-\beta}$ at large β. Hence the outlier fixed point has higher Hopfield energy than the cluster fixed point.

Likewise, subtracting Eq. (114) from Eq. (112) gives

$$\mathcal{S}_o - \mathcal{S}_{cl} = (N-1)\left[e^{-\beta(1-c)}\Lambda(c) - e^{-\beta}\Lambda(0)\right] + o(e^{-\beta(1-c)}) > 0. \tag{116}$$

Thus the same outlier fixed point is also sharper. Equations (115) and (116) justify the shorthand used in the main text: in this model, high-energy memories correspond to sharp isolated basins, whereas low-energy memories correspond to broad cluster-supported basins.

D.4 Sharpness in diffusion-model experiments

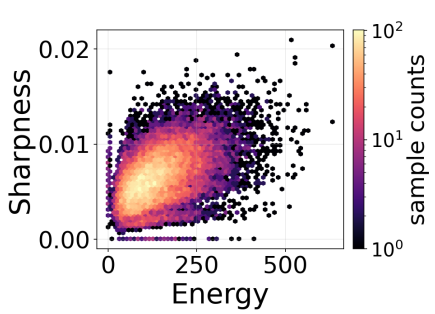
We further examine the relation between Hopfield energy, local sharpness, and reconstruction-based forgetting in the diffusion-model experiments. Here we focus on CIFAR-10 and compute all quantities sample-wise for Task-1 images.

For each Task-1 sample x, we first compute its Hopfield energy with respect to the Task-1 reference set $\mathcal{D}_1$,

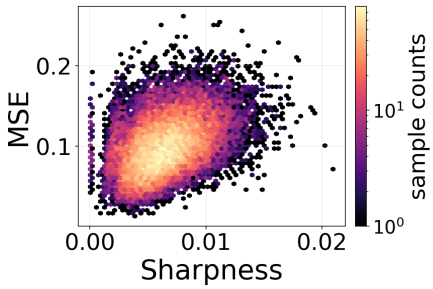
$$E_t(x \mid \mathcal{D}_1), \tag{117}$$

using the same energy computation as in the main diffusion-model experiments. To quantify local sharpness, we do not explicitly form the Hessian. Instead, we use a perturbation-based finite-difference estimate: for each sample x, we draw M small random perturbations $\{\delta_m\}_{m=1}^M$ and measure how much the Hopfield energy increases around x,

$$\widehat{\mathcal{S}}_t(x) = \frac{1}{M}\sum_{m=1}^{M}\left[E_t(x+\delta_m \mid \mathcal{D}_1) - E_t(x \mid \mathcal{D}_1)\right], \tag{118}$$



(a) Energy vs. perturbation sharpness.



(b) Perturbation sharpness vs. reconstruction forgetting.

Figure S.1: Perturbation-based sharpness in the diffusion-model experiment on CIFAR-10. (a) Higher-Hopfield-energy samples tend to have larger perturbation sharpness $\widehat{\mathcal{S}}_t(x)$. (b) Samples with larger perturbation sharpness tend to have larger reconstruction-based forgetting $\mathcal{F}_{t^\star}(x)$ after Task-2 training.

up to the fixed perturbation-scale normalization used in the implementation. In the experiments, we use $M = 100$ random perturbations for each sample. Thus, $\widehat{\mathcal{S}}_t(x)$ measures how steeply the Hopfield energy changes under local random perturbations: larger values indicate that the sample lies in a sharper local region of the energy landscape.

Figure S.1 summarizes the result. First, we compare the perturbation-based sharpness $\widehat{\mathcal{S}}_t(x)$ with the Hopfield energy $E_t(x \mid \mathcal{D}_1)$. Higher-energy samples tend to have larger perturbation sharpness, consistent with the theoretical picture that weakly supported samples lie in sharper regions of the landscape. Second, we compare $\widehat{\mathcal{S}}_t(x)$ with the reconstruction-based forgetting metric $\mathcal{F}_{t^\star}(x)$ defined in Eq. (15). Samples with larger perturbation sharpness also tend to show larger reconstruction error after Task-2 training. These results suggest that the energy-based ordering observed in the main text is accompanied by a corresponding difference in local sharpness, and that this difference is reflected in reconstruction behavior after continual learning.

E Excluding stable spurious fixed points

Proposition E.1 (No additional stable fixed points)**.** *For the configuration in Eq.* (8)*, fix $N \geq 2$, $d \geq N + 1$, and $0 < c < 1$. For sufficiently large β with $\beta(1 - c) \geq (1 + \delta)N$ (fixed $\delta > 0$), there are exactly $N + 1$ stable fixed points, one near each stored memory. Every other critical point has a negative Hessian eigenvalue.*

Proof. **The fixed-point equation narrows down the candidates.** Put $g = 1 - c$, $q = p_o$, and $r = \sum_{i=1}^{N} p_i = 1 - q$. At any fixed point, $x_i^\top \xi_* = g p_i + cr$ for a cluster memory, so the softmax equation gives

$$\log p_i - \beta g p_i = \beta c r - \log Z, \qquad Z = \sum_{j=1}^{N} e^{\beta x_j^\top \xi_*} + e^{\beta x_o^\top \xi_*}. \tag{119}$$

All cluster weights therefore solve the same scalar equation $f(t) =$ constant, where $f(t) = \log t - \beta g t$. Since $f'(t) = 1/t - \beta g$, this function increases up to $t = 1/(\beta g)$ and then decreases. It takes any value at most twice. Thus the cluster weights are either all equal or take two values $a > b$, with

$$b < \frac{1}{\beta g} < a. \tag{120}$$

If two or more memories have the larger weight a, choose two of them, $p_i = p_j = a$, and set $v = x_i - x_j$. The equal-angular geometry gives $x_i^\top v = g$, $x_j^\top v = -g$, and $x_k^\top v = 0$ for every other memory, including the outlier. Since $\xi_* = \sum_k p_k x_k$ at a fixed point, we also have $\xi_*^\top v =$

$g(p_i - p_j) = 0$. The Hessian formula in Eq. (77) therefore yields

$$H_* v = v - \beta \left(\sum_k p_k x_k (x_k^\top v) - \xi_* (\xi_*^\top v) \right) = v - \beta a g (x_i - x_j) = (1 - \beta g a) v. \tag{121}$$

Thus $x_i - x_j$ is a Hessian eigenvector with eigenvalue $1 - \beta g a < 0$, where the inequality follows from $a > 1/(\beta g)$. Such a mixed state is unstable: a bias toward one of these memories grows. Consequently, every potentially stable fixed point has one of two forms, up to a permutation of the cluster memories:

$$\begin{aligned} \text{Symmetric:} \quad & (p_1, \ldots, p_N, q) = (r/N, \ldots, r/N, 1 - r), \\ \text{One preferred memory:} \quad & (p_1, \ldots, p_N, q) = (a, b, \ldots, b, q), \quad a > b. \end{aligned} \tag{122}$$

This classification is exact at finite β; it leaves no other form that could be stable.

Spurious candidates are unstable in the stated regime. For the symmetric form, let $u = (x_i - x_j)/\sqrt{2g}$, $i \neq j$, be the unit direction from memory x_j to memory x_i in the same cluster. Here, directional curvature means $u^\top H_* u$. Using the Hessian formula in Appendix C.5, Eq. (77), the same calculation as Eq. (121) with $p_i = p_j = r/N$ gives $H_* u = (1 - \beta g r/N) u$. Thus $1 - \beta g r/N$ is the Hessian eigenvalue along u, not a maximum over directions. A Hessian with no negative eigenvalues therefore requires

$$r \leq \frac{N}{\beta g}. \tag{123}$$

For fixed N, c and sufficiently large β, this places the point near the outlier. The spurious cluster representative instead has $r \to 1$; this Hessian eigenvalue $1 - \beta g r/N$ is negative in this regime.

For the second form, suppose a is the weight of x_k. The low weights already satisfy $b < 1/(\beta g)$. Stability also requires a small outlier weight. To see this, take the unit direction $v = (x_k - x_o)/\sqrt{2}$. The covariance C_p in $H_* = I - \beta C_p$ satisfies

$$v^\top C_p v \geq \frac{2aq}{a+q}, \qquad H_* \succeq 0 \implies \frac{1}{\beta a} + \frac{1}{\beta q} \geq 2. \tag{124}$$

The variance bound follows by retaining the two memories x_k, x_o, whose projections onto v differ by $\sqrt{2}$. Since $1/(\beta a) < g$, it follows that

$$q < \frac{1}{\beta(1+c)}, \qquad 1 - a = (N-1)b + q < \frac{N}{\beta g}. \tag{125}$$

Hence this candidate can be stable only near its preferred memory; a substantial mixture with the outlier is unstable.

Exactly one stable point remains near each memory. Both cases place every fixed point with no negative Hessian eigenvalues within distance $2N/(\beta g)$ of a stored memory. In the ball $\|\xi - x_k\| \leq g/4$, memory k has an overlap advantage of at least $g/2$ over every competitor. Thus the update map $T(\xi) = \sum_i p_i(\xi) x_i$ satisfies

$$\|T(\xi) - x_k\| \leq 2N e^{-\beta g/2}, \qquad \|DT(\xi)\|_{\text{op}} \leq 4\beta N e^{-\beta g/2}. \tag{126}$$

For sufficiently large β, it maps this ball into itself and is a contraction, giving a unique fixed point with positive definite Hessian. There are $N + 1$ disjoint memory balls, and every potentially stable fixed point lies in one of them. Therefore there are exactly $N + 1$ stable fixed points and no stable spurious points. □

F Energy and curvature at the cluster prototype

We evaluate the energy and tangential curvature directly at the mixed state represented by the cluster prototype and at the outlier. Neither evaluation point is assumed to be a fixed point. Assume Eq. (8), with $N \geq 2$, $0 < c < 1$, and $d \geq N + 1$, and define

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i, \qquad \alpha = \|\bar{x}\|^2 = c + \frac{1 - c}{N}. \tag{127}$$

For the MHN energy E_β, define the curvature factor

$$G_\beta(\xi) = \|\xi\|^2 \operatorname{tr} \nabla^2 E_\beta(\xi) - \xi^\top \nabla^2 E_\beta(\xi) \xi. \tag{128}$$

Proposition F.1 (Comparison at the reference points)**.** *Put* $a = \alpha(1-\alpha)$ *and*

$$\beta_G = \log \frac{N(1-a)}{a}, \qquad \beta_E = \frac{2\log N}{1-\alpha}. \tag{129}$$

If $\beta_G \le \beta < \beta_E$*, then*

$$E_\beta(\bar{x}) < E_\beta(x_o), \qquad G_\beta(\bar{x}) < G_\beta(x_o). \tag{130}$$

Proof. **Energy.** Direct evaluation gives

$$E_\beta(\bar{x}) = \frac{\alpha}{2} - \frac{1}{\beta}\log(Ne^{\beta\alpha}+1), \qquad E_\beta(x_o) = \frac{1}{2} - \frac{1}{\beta}\log(N+e^{\beta}). \tag{131}$$

Thus, for $\beta > 0$,

$$E_\beta(\bar{x}) < E_\beta(x_o) \iff \big(e^{\beta(1+\alpha)/2}-1\big)\big(N - e^{\beta(1-\alpha)/2}\big) > 0 \tag{132}$$

$$\iff \beta < \beta_E. \tag{133}$$

Curvature. Let $S = N^{-1}\sum_{i=1}^{N}(x_i-\bar{x})(x_i-\bar{x})^\top$. Then $\operatorname{tr} S = 1-\alpha$ and $S\bar{x} = Sx_o = 0$. At either reference point the cluster memories have equal softmax weights. Writing their total weight as w, the covariance is

$$C_w = wS + w(1-w)(\bar{x}-x_o)(\bar{x}-x_o)^\top. \tag{134}$$

The total cluster weights at $\bar{x}$ and x_o are, respectively,

$$u = \frac{Ne^{\beta\alpha}}{Ne^{\beta\alpha}+1}, \qquad v = \frac{N}{N+e^{\beta}}. \tag{135}$$

Substituting $\nabla^2 E_\beta = I - \beta C_w$ gives

$$\begin{aligned} G_\beta(\bar{x}) &= (d-1)\alpha - \beta\alpha u(2-\alpha-u), \\ G_\beta(x_o) &= (d-1) - \beta v(1-\alpha v). \end{aligned} \tag{136}$$

Hence

$$G_\beta(\bar{x}) - G_\beta(x_o) = -(d-1)(1-\alpha) - \beta\big[\alpha u(2-\alpha-u) - v(1-\alpha v)\big]. \tag{137}$$

Put $\eta = 1-u$. Since $\beta > 0$ and $c > 0$, $\eta < 1/(N+1) < 1/N < \alpha$, and therefore

$$\alpha u(2-\alpha-u) = a + \alpha\eta(\alpha-\eta) > a. \tag{138}$$

Moreover, $\beta \ge \beta_G$ is equivalent to $v \le a$, so

$$\alpha u(2-\alpha-u) > a \ge v > v(1-\alpha v). \tag{139}$$

Equation (137) is consequently negative, proving the curvature inequality. □

Temperature range. For fixed $c \in (0,1)$, the thresholds satisfy

$$\begin{aligned} \beta_G &= \log N + \log\frac{1-c+c^2}{c(1-c)} + O(N^{-1}), \\ \beta_E &= \frac{2}{1-c}\log N + O\!\left(\frac{\log N}{N}\right). \end{aligned} \tag{140}$$

Thus the sufficient interval $[\beta_G, \beta_E)$ is nonempty for sufficiently large N. In particular, any fixed $1 < \gamma < 2/(1-c)$ and $\beta = \gamma \log N$ satisfy the two reference-point inequalities for sufficiently large N.

G Dimension constraints and approximately equal-angular memories

For $0 < c < 1$, the exact cluster Gram matrix

$$G_{\mathrm{cl}} = (1-c)I_N + c\mathbf{1}\mathbf{1}^\top \tag{141}$$

has rank N. Including an orthogonal unit outlier raises the rank to $N+1$, so the exact configuration requires $N+1 \le d$. Conversely, factoring this positive-definite block Gram matrix gives a realization whenever that condition holds. Thus exact equal-angularity allows $N/d \to \rho < 1$, but not a fixed $\rho > 1$.

An approximate construction permits more memories than dimensions. For $d \ge 3$, choose orthonormal unit vectors $\mu, x_o \in \mathbb{R}^d$ and independent uniform unit vectors z_i in their common orthogonal complement, and set

$$x_i = \sqrt{c}\,\mu + \sqrt{1-c}\,z_i. \tag{142}$$

Then $\|x_i\| = 1$ and $x_i^\top x_o = 0$ exactly. For fixed $c \in (0,1)$ and $N/d \to \rho \in (0,\infty)$, spherical concentration gives

$$\Delta_{N,d} := \max_{i \ne j} |x_i^\top x_j - c| = O_p\left(\sqrt{\frac{\log d}{d}}\right), \tag{143}$$

where O_p denotes order in probability. If $\beta = \gamma \log d$ with fixed $\gamma(1-c) > 1$, both $\beta\Delta_{N,d}$ and $Ne^{-\beta(1-c)}$ vanish in probability. At a stored cluster memory, the total competing cluster weight relative to the self weight is therefore $(N-1)e^{-\beta(1-c)}(1+o_p(1))$, whereas at the outlier the corresponding ratio is $Ne^{-\beta}$. This preserves the leading softmax leakage hierarchy and motivates a similar local ordering for approximately equal-angular data. It does not prove the forgetting ordering or exclude other stable states: entrywise concentration is not exact symmetry or operator-norm Gram convergence, and uniform fixed-point and remainder bounds are still needed.

The proportional high-dimensional limit is a first step in which memory count and representation dimension grow together. In a latent representation, N may describe effective memory points rather than the raw number of training examples. Holding d fixed while increasing N is a separate regime, for which the overlap errors above need not vanish. A full treatment of that limit and more general N–d scaling remains for subsequent work.

H Full derivation of Theorem 5.1

We derive the exact replay identity, isolate the duplicate-memory baseline, expand the baseline-subtracted gain to leading order in the small rotation, evaluate the five canonical gains, and prove the ordering of Theorem 5.1.

H.1 Exact replay identity

Let $\mathcal{I} = \{1, \dots, N, o\}$, $X = \{x_\mu : \mu \in \mathcal{I}\}$, and $X_2(V) = VX$. When a replay point $x_r \in X$ is added, write $X_2^{+r}(V) := X_2(V) \cup \{x_r\}$, with the replayed sample counted as an additional term in the log-sum-exp. For a Task-1 fixed point ξ_k^*, the energy rise without replay was already defined in Section C.5 as

$$\Delta E_k^{\mathrm{fp}}(V) = E(\xi_k^* \mid VX) - E(\xi_k^* \mid X), \qquad \Delta E_k^{\mathrm{fp}}(I) = 0. \tag{144}$$

We additionally need the Task-1 softmax weight of the replay point at ξ_k^*,

$$p_{r|k} := \frac{e^{\beta x_r^\top \xi_k^*}}{\sum_{\mu \in \mathcal{I}} e^{\beta x_\mu^\top \xi_k^*}}. \tag{145}$$

The raw replay-induced energy decrease is

$$\mathcal{R}_{r\to k}(V) := E(\xi_k^* \mid X_2(V)) - E(\xi_k^* \mid X_2^{+r}(V)). \tag{146}$$

Using the log-sum-exp form (37) of E, the difference of the two log-sum-exp's is

$$\mathcal{R}_{r\to k}(V) = \tfrac{1}{\beta} \log\left(1 + \frac{e^{\beta x_r^\top \xi_k^*}}{\sum_{\mu \in \mathcal{I}} e^{\beta (Vx_\mu)^\top \xi_k^*}}\right). \tag{147}$$

Now observe that, by the definition (144) and the log-sum-exp form,

$$e^{\beta \Delta E_k^{\mathrm{fp}}(V)} = \frac{\sum_\mu e^{\beta x_\mu^\top \xi_k^*}}{\sum_\mu e^{\beta (V x_\mu)^\top \xi_k^*}}. \tag{148}$$

Multiplying numerator and denominator inside (147) by $e^{\beta \Delta E_k^{\mathrm{fp}}(V)}$ converts the rotated denominator back into the Task-1 partition function:

$$\frac{e^{\beta x_r^\top \xi_k^*}}{\sum_\mu e^{\beta (V x_\mu)^\top \xi_k^*}} = \frac{e^{\beta x_r^\top \xi_k^*}\, e^{\beta \Delta E_k^{\mathrm{fp}}(V)}}{\sum_\mu e^{\beta x_\mu^\top \xi_k^*}} = p_{r|k}\, e^{\beta \Delta E_k^{\mathrm{fp}}(V)}. \tag{149}$$

Hence the exact identity

$$\mathcal{R}_{r\to k}(V) = \tfrac{1}{\beta} \log\bigl(1 + p_{r|k}\, e^{\beta \Delta E_k^{\mathrm{fp}}(V)}\bigr). \tag{150}$$

At $V = I$, $\Delta E_k^{\mathrm{fp}}(I) = 0$ and (150) reduces to the trivial duplicate-memory baseline

$$\mathcal{R}_{r\to k}(I) = \tfrac{1}{\beta} \log(1 + p_{r|k}), \tag{151}$$

which is the energy decrease one obtains merely from replicating an already-stored Task-1 memory. Subtracting this baseline and using $\log(a) - \log(b) = \log(a/b)$,

$$\widehat{\Delta}_{r\to k}(V) := \mathcal{R}_{r\to k}(V) - \mathcal{R}_{r\to k}(I) = \tfrac{1}{\beta} \log \frac{1 + p_{r|k}\, e^{\beta \Delta E_k^{\mathrm{fp}}(V)}}{1 + p_{r|k}}. \tag{152}$$

This separates the two effects: $p_{r|k}$ measures the relevance of the replay point (geometry), and $\Delta E_k^{\mathrm{fp}}(V)$ measures how much the rotated Task-2 dataset raises the energy at ξ_k^* (drift) — exactly the same object that drives Theorem 4.1.

H.2 Small-rotation expansion and master formula

Define

$$\rho(p) := \frac{p}{1+p}. \tag{153}$$

This factor determines the response of forgetting mitigation to the energy rise, because

$$\frac{\partial}{\partial \Delta E} \left\{ \frac{1}{\beta} \log \frac{1 + p\, e^{\beta \Delta E}}{1+p} \right\} \Bigg|_{\Delta E = 0} = \rho(p). \tag{154}$$

We therefore call $\rho(p)$ the *replay susceptibility*. This name refers to the response to an energy change in Eq. (154). Taylor-expanding (152) in $\Delta E_k^{\mathrm{fp}}(V) = O(\varepsilon^2)$,

$$\widehat{\Delta}_{r\to k}(V) = \rho(p_{r|k})\, \Delta E_k^{\mathrm{fp}}(V) + \tfrac{\beta}{2} \rho(p_{r|k})(1 - \rho(p_{r|k}))\, \Delta E_k^{\mathrm{fp}}(V)^2 + O\bigl(\beta^2 \rho(p_{r|k})\, |\Delta E_k^{\mathrm{fp}}(V)|^3\bigr). \tag{155}$$

Averaging over V and setting $\Delta_{r\to k} := \langle \widehat{\Delta}_{r\to k}(V) \rangle_V$, and recalling $\Delta E_k^{\mathrm{fp}} := \langle \Delta E_k^{\mathrm{fp}}(V) \rangle_V$ from (11),

$$\Delta_{r\to k} = \rho(p_{r|k})\, \Delta E_k^{\mathrm{fp}} + O\bigl(\beta \rho(p_{r|k}) \varepsilon^4\bigr). \tag{156}$$

The factor ΔE_k^{fp} is exactly the rotation-averaged energy rise computed in Section C.6 (Eqs. (97),(98)).

H.3 Susceptibility asymptotics

From Appendix C,

$$v = p_{o|o} = 1 - N e^{-\beta} + O(\beta e^{-2\beta}), \tag{157}$$

$$a = p_{1|1} = 1 - (N-1) e^{-\beta(1-c)} - e^{-\beta} + o(e^{-\beta(1-c)}), \tag{158}$$

$$b = p_{1|2} = e^{-\beta(1-c)}(1 + o(1)), \qquad u = p_{1|o} = e^{-\beta}(1 + o(1)), \qquad q = p_{o|1} = e^{-\beta}(1 + o(1)). \tag{159}$$

Applying $\rho(p) = p/(1+p)$,

$$\begin{aligned}
&\rho(v) = \tfrac{1}{2} - \tfrac{N\,e^{-\beta}}{4} + O(\beta\, e^{-2\beta}),\\
&\rho(a) = \tfrac{1}{2} - \tfrac{(N-1)\,e^{-\beta(1-c)}+e^{-\beta}}{4} + o(e^{-\beta(1-c)}),\\
&\rho(b) = e^{-\beta(1-c)}(1+o(1)), \qquad \rho(u) = e^{-\beta}(1+o(1)), \qquad \rho(q) = e^{-\beta}(1+o(1)).
\end{aligned} \tag{160}$$

Hence

$$\rho(v) > \rho(a) \gg \rho(b) \gg \rho(u) \asymp \rho(q). \tag{161}$$

H.4 Five replay gains

Recall from (97)–(98), with $\Lambda(\alpha) = 2(d-1)(1-\alpha) + \beta(1-\alpha^2)$,

$$\Delta E_o^{\mathrm{fp}} = \tfrac{\varepsilon^2}{4}\big[(d-1) - N\,e^{-\beta}\,\Lambda(0)\big] \tag{162}$$

$$+O(\varepsilon^2\beta^2 e^{-2\beta}) + O_{\beta,N,d,c}(\varepsilon^3), \tag{163}$$

$$\Delta E_{cl}^{\mathrm{fp}} = \tfrac{\varepsilon^2}{4}\big[(d-1) - (N-1)\,e^{-\beta(1-c)}\,\Lambda(c) - e^{-\beta}\,\Lambda(0)\big] \tag{164}$$

$$+O(\varepsilon^2\beta^2 e^{-2\beta(1-c)}) + O_{\beta,N,d,c}(\varepsilon^3), \tag{165}$$

$$\Delta E_o^{\mathrm{fp}} - \Delta E_{cl}^{\mathrm{fp}} = \tfrac{\varepsilon^2}{4}(N-1)\big[e^{-\beta(1-c)}\,\Lambda(c) - e^{-\beta}\,\Lambda(0)\big] \tag{166}$$

$$+O(\varepsilon^2\beta^2 e^{-2\beta(1-c)}) + O_{\beta,N,d,c}(\varepsilon^3) > 0. \tag{167}$$

The inequality holds for sufficiently large fixed β and then sufficiently small ε. The large-β gain asymptotics below compare the coefficients obtained after the small-rotation expansion; equivalently, choose $\varepsilon = \varepsilon(\beta)$ sufficiently small to control the rotation remainders. Multiplying (160) by ΔE_o^{fp} or $\Delta E_{cl}^{\mathrm{fp}}$ via the master formula (156) yields

$$\Delta_{o\to o} = \tfrac{1}{2}\,\Delta E_o^{\mathrm{fp}}\,(1+o(1)) = \tfrac{\varepsilon^2}{8}(d-1) + O(\varepsilon^2\beta e^{-\beta}) + O_{\beta,N,d,c}(\varepsilon^3), \tag{168}$$

$$\Delta_{cl_1\to cl_1} = \tfrac{1}{2}\,\Delta E_{cl}^{\mathrm{fp}}\,(1+o(1)) = \tfrac{\varepsilon^2}{8}(d-1) + O(\varepsilon^2\beta e^{-\beta(1-c)}) + O_{\beta,N,d,c}(\varepsilon^3), \tag{169}$$

$$\Delta_{cl_1\to cl_2} = e^{-\beta(1-c)}\,\Delta E_{cl}^{\mathrm{fp}}\,(1+o(1)) = \tfrac{\varepsilon^2}{4}(d-1)e^{-\beta(1-c)}(1+o(1)), \tag{170}$$

$$\Delta_{cl_1\to o} = e^{-\beta}\,\Delta E_o^{\mathrm{fp}}\,(1+o(1)) = \tfrac{\varepsilon^2}{4}(d-1)e^{-\beta}(1+o(1)), \tag{171}$$

$$\Delta_{o\to cl_1} = e^{-\beta}\,\Delta E_{cl}^{\mathrm{fp}}\,(1+o(1)) = \tfrac{\varepsilon^2}{4}(d-1)e^{-\beta}(1+o(1)). \tag{172}$$

H.5 Ordering

1. $\Delta_{o\to o} > \Delta_{cl_1\to cl_1}$.

$$\Delta_{o\to o} - \Delta_{cl_1\to cl_1} = \{\rho(v)-\rho(a)\}\Delta E_{cl}^{\mathrm{fp}} + \rho(v)\{\Delta E_o^{\mathrm{fp}} - \Delta E_{cl}^{\mathrm{fp}}\} + O_{\beta,N,d,c}(\beta\varepsilon^4). \tag{173}$$

   From (160), $\rho(v) - \rho(a) = \frac{(N-1)(e^{-\beta(1-c)}-e^{-\beta})}{4} + o(e^{-\beta(1-c)}) > 0$, and Theorem 4.1 gives $\Delta E_o^{\mathrm{fp}} > \Delta E_{cl}^{\mathrm{fp}} > 0$. For each sufficiently large fixed β, shrinking ε makes the $O_{\beta,N,d,c}(\beta\varepsilon^4)$ remainder smaller than the positive order-ε^2 terms. Thus $\Delta_{o\to o} > \Delta_{cl_1\to cl_1}$.

2. $\Delta_{cl_1\to cl_1} \gg \Delta_{cl_1\to cl_2}$.

$$\frac{\Delta_{cl_1\to cl_2}}{\Delta_{cl_1\to cl_1}} = \frac{\rho(b)}{\rho(a)}(1+o(1)) = 2e^{-\beta(1-c)}(1+o(1)) \to 0. \tag{174}$$

3. $\Delta_{cl_1\to cl_2} \gg \Delta_{cl_1\to o}$.

$$\frac{\Delta_{cl_1\to o}}{\Delta_{cl_1\to cl_2}} = \frac{\rho(u)}{\rho(b)}\frac{\Delta E_o^{\mathrm{fp}}}{\Delta E_{cl}^{\mathrm{fp}}}(1+o(1)) = e^{-\beta c}(1+o(1)) \to 0. \tag{175}$$

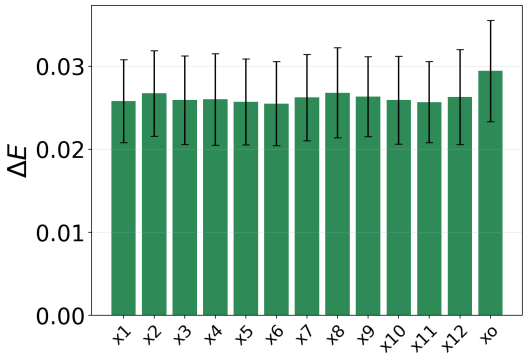


Figure S.2: Closed-form MHN corroboration of Theorem 4.1. The outlier fixed point exhibits a larger rotation-induced energy rise than the cluster fixed points.

4. $\Delta_{cl_1\to cl_2} \gg \Delta_{o\to cl_1}$.

$$\frac{\Delta_{o\to cl_1}}{\Delta_{cl_1\to cl_2}} = \frac{\rho(q)}{\rho(b)}(1+o(1)) = e^{-\beta c}(1+o(1)) \to 0. \tag{176}$$

Moreover,

$$\frac{\Delta_{o\to cl_1}}{\Delta_{cl_1\to o}} = \frac{\rho(q)}{\rho(u)}\frac{\Delta E^{\mathrm{fp}}_{cl}}{\Delta E^{\mathrm{fp}}_{o}}(1+o(1)) = 1+o(1), \tag{177}$$

so $\Delta_{o\to cl_1}$ and $\Delta_{cl_1\to o}$ have the same exponential order. Combining the four comparisons gives

$$\Delta_{o\to o} > \Delta_{cl_1\to cl_1} \gg \Delta_{cl_1\to cl_2} \gg \{\Delta_{cl_1\to o}, \Delta_{o\to cl_1}\}, \qquad \Delta_{cl_1\to o} \asymp \Delta_{o\to cl_1}, \tag{178}$$

which proves Theorem 5.1. □

Remark H.1. The subtraction of $\mathcal{R}_{r\to k}(I) = \frac{1}{\beta}\log(1+p_{r|k})$ removes the trivial gain from duplicating an already-stored memory. What is left is the fraction $\rho(p_{r|k}) = p_{r|k}/(1+p_{r|k})$ of the rotation-induced rise $\Delta E^{\mathrm{fp}}_k(V)$ that the replay point can recover. Self-replay ($p \to 1$) recovers $\frac{1}{2}$; within-cluster cross-replay recovers $e^{-\beta(1-c)}$; cross-type replay between cluster and outlier recovers $e^{-\beta}$. The ordering then follows from $\frac{1}{2} \gg e^{-\beta(1-c)} \gg e^{-\beta}$ and $\Delta E^{\mathrm{fp}}_o > \Delta E^{\mathrm{fp}}_{cl}$.

I Synthetic experiments on MHNs

I.1 Intrinsic forgetting in synthetic MHNs

To complement the theory in Section 4.1, we instantiate the equal-angular MHN at moderate finite values ($N = 12$, $d = 50$, $c = 0.35$, $\beta = 8$, $\varepsilon = 0.02$, 2000 random rotations) – well outside the asymptotic regime $\beta \gg \log N$, $\varepsilon \ll 1$ used in the proof – and compute ΔE^{fp} in closed form. Even in this finite-parameter regime the outlier fixed point has a strictly larger rotation-induced energy rise than the cluster fixed points, matching the qualitative ordering predicted by Theorem 4.1. The corresponding learned MHN-BM result is shown in the main text in Fig. 3a.

I.2 Replay ordering in synthetic MHNs

For completeness, we test the qualitative ordering of Theorem 5.1 in the exact equal-angular geometry at finite temperature. In the closed-form MHN, the full replay-gain $\{\Delta_{r\to k}\}_{r,k}$ shows the predicted ordering: the outlier self-entry $\Delta_{o\to o}$ is largest, cluster self-gains $\Delta_{cl_1\to cl_1}$ are next, within-cluster cross entries $\Delta_{cl_1\to cl_2}$ are substantially smaller, and cross-type entries $\Delta_{cl_1\to o}, \Delta_{o\to cl_1}$ are smallest. The corresponding learned MHN-BM corroboration is shown in the main text in Fig. 4.

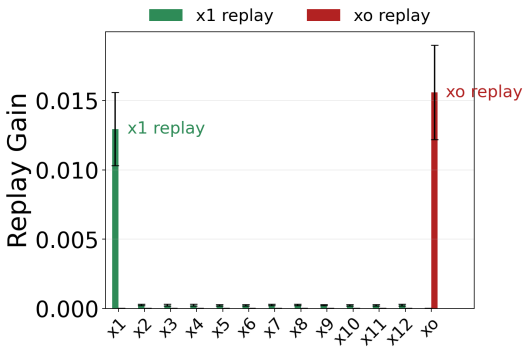


Figure S.3: Closed-form MHN corroboration of the replay ordering in Theorem 5.1. The replay-gain reproduces the ranking $\Delta_{o\to o} > \Delta_{cl_1\to cl_1} \gg \Delta_{cl_1\to cl_2} \gg \{\Delta_{cl_1\to o}, \Delta_{o\to cl_1}\}$.

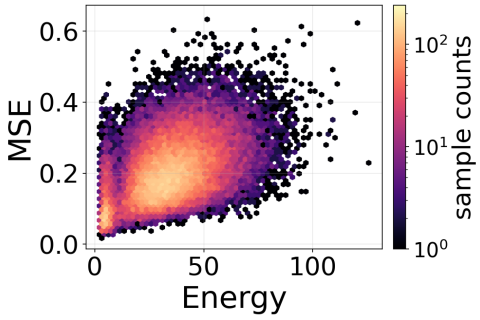


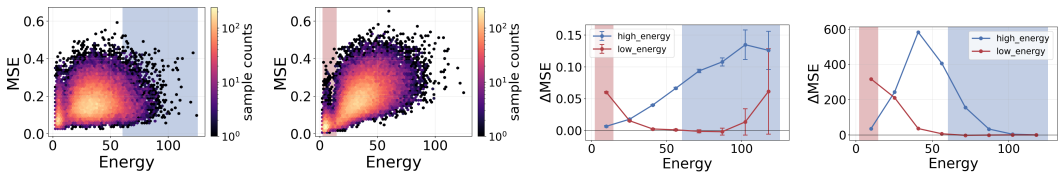
Figure S.4: Rotated MNIST (domain-incremental learning): per-sample reconstruction MSE $\mathcal{F}_{t^\star}(x)$ after Task 2 training versus internal Hopfield energy measured on Task 1, without replay. The same monotone energy–forgetting relation observed for split CIFAR-10 in Fig. 3 is recovered.

J Domain-incremental learning experiment: Rotated MNIST

The theoretical analysis of Sections 4.1 and 5.1 models continual learning as a small rotation V of the dataset, which corresponds most directly to a *domain-incremental* setting in which the same content is presented under a perturbation of the input distribution. The main-text experiments focus on the more widely used *incremental-class-learning* benchmark with split CIFAR-10. To complement them, we report here the corresponding domain-incremental experiment on rotated MNIST – the setup that most literally matches the rotated-dataset assumption underlying our theorems.

Setup. Task 1 is standard MNIST. Task 2 is the same MNIST images rotated by $30°$, instantiating the rotated-dataset construction of Section 4.1 on real images. We train a pixel-space DDPM from scratch on Task 1 and continue training on Task 2. After Task 1 we assign every Task-1 image a scalar Hopfield energy $E_\theta(x)$ via the log-sum-exp expression (1), using all Task-1 training images as stored memories. As in the main-text diffusion-model experiments, forgetting is measured by the per-sample reconstruction MSE $\mathcal{F}_{t^\star}(x)$ defined in Eq. (15). For the replay analysis we use the same three policies as in Section 5.3 – *no replay* (baseline), energy-guided *top-K* (high-energy buffer), and energy-guided *bottom-K* (low-energy buffer) – with an identical buffer size $K = 5000$. Full hyperparameters are listed in Appendix M.

Forgetting indicator (Fig. S.4). After Task 2 fine-tuning, per-sample reconstruction MSE is a monotone-increasing function of the Task-1 Hopfield energy, reproducing the same qualitative relationship observed for split CIFAR-10 in Fig. 3b and Fig. 3c. Higher-energy Task-1 images suffer larger reconstruction forgetting, exactly as predicted by Theorem 4.1.



(a) High-energy replay. (b) Low-energy replay. (c) Mean reduction. (d) Sum reduction.

Figure S.5: Energy-guided replay on rotated MNIST (domain-incremental learning). The same pattern as Fig. 5 and Fig. 6 is reproduced: (a) high-energy replay produces a sharply peaked, narrow rescue at high energy, (b) low-energy replay produces a smaller but broader rescue at low energy, and (c,d) the per-bin reductions follow the theoretical replay ordering in Eq. (20). The replay benefit is again dominated by the high-energy tail.

Energy-guided replay (Fig. S.5). The mean–range tradeoff predicted by Eq. (18) – $\Delta_{r\to k} \sim \rho_{r|k}\, \Delta E_k^{\mathrm{fp}}$ – is reproduced on rotated MNIST. The high-energy buffer (Fig. S.5a) yields a sharply peaked bin-mean reduction concentrated in the high-energy tail (Fig. S.5c), whereas the low-energy buffer (Fig. S.5b) gives a smaller per-sample reduction spread over a wider low-energy range. The bin-summed reduction (Fig. S.5d) is dominated by the high-energy tail under the high-energy buffer, exactly as in the CIFAR-10 results (Fig. 5, 6). Together with the main-text experiments, this confirms that the energy-guided replay rule predicted by Theorem 5.1 is robust across both incremental-class and domain-incremental continual learning of diffusion models.

K Reconstruction metric for diffusion-model forgetting

The theory in Section 4 defines forgetting as an intrinsic change in the Hopfield energy,

$$\Delta E(x) = E(x \mid \mathcal{X}_{\mathrm{new}}) - E(x \mid \mathcal{X}_{\mathrm{old}}). \tag{179}$$

For an explicit modern Hopfield network this quantity can be evaluated directly. For a trained diffusion model, however, the corresponding implicit energy is not available as a normalized scalar function for each sample and each training stage. Although the score determines the local reverse dynamics, measuring the exact sample-wise energy increase before and after Task-2 training would require reconstructing or integrating the implicit energy landscape itself. We therefore use a reconstruction-based quantity as an operational proxy for intrinsic forgetting.

For each Task-1 image x, we first corrupt it to a fixed diffusion timestep $t^\star$ and then run the reverse process of the model obtained after Task-2 training. Let $\hat{x}_{0,t^\star}$ denote the resulting reconstruction. We define the reconstruction forgetting score as

$$\mathcal{F}_{t^\star}(x) = \frac{1}{d_x}\|x - \hat{x}_{0,t^\star}\|_2^2. \tag{180}$$

In Stable Diffusion experiments, the image is first encoded into the VAE latent space, noise is added to the latent, the UNet denoises the latent trajectory, and the final latent is decoded back to pixel space. We report the pixel-space reconstruction MSE in the main figures. In the pixel-space diffusion experiments, the same procedure is performed directly in pixel space. When multiple stochastic reconstructions are used, we average the MSE over generations.

This metric should be interpreted as a proxy rather than an exact estimator of $\Delta E(x)$. If Task-2 training raises or deforms the implicit basin supporting a Task-1 sample, then a noisy version of that sample is less likely to be mapped back to the original point by the post-Task-2 reverse dynamics. This leads to a larger reconstruction error. Conversely, samples that remain well supported by the post-Task-2 landscape should remain easier to reconstruct. Thus, while $\mathcal{F}_{t^\star}(x)$ is not itself an energy difference, it probes the same local loss of support that intrinsic forgetting is intended to capture.

L Figure axes and replay aggregation

In Figs. 3a, 4, S.2, and S.3, $x_1, \ldots, x_{12}$ denote cluster memories and x_o the outlier. Each tick identifies the target Task-1 fixed point ξ_k^*; replay legends identify the source x_r. In diffusion density

plots, colorbars show sample counts per hexagonal bin on a logarithmic scale (e.g., $10^0, 10^1, 10^2$ are counts).

For sample i and policy r, let $e_i^{(0)}$ and $e_i^{(r)}$ be the no-replay and replay reconstruction MSEs under the same evaluation protocol. Define

$$D_i^{(r)} := \Delta\mathrm{MSE}_i^{(r)} = e_i^{(0)} - e_i^{(r)}, \tag{181}$$

so positive values indicate improvement. For energy bin B_b with $n_b = |B_b| \geq 2$, the mean and sum are

$$\overline{D}_b^{(r)} = \frac{1}{n_b} \sum_{i \in B_b} D_i^{(r)}, \qquad S_b^{(r)} = \sum_{i \in B_b} D_i^{(r)}. \tag{182}$$

In Figs. 5, 6, and S.5, panels (c,d) show the bin mean and sum; panels (a,b) show post-replay MSE itself. Bins with fewer than two valid samples are omitted. Error bars in (c) are within-bin standard errors,

$$\mathrm{SE}_b^{(r)} = \frac{1}{\sqrt{n_b}} \sqrt{\frac{1}{n_b} \sum_{i \in B_b} \left(D_i^{(r)} - \overline{D}_b^{(r)} \right)^2}. \tag{183}$$

This is the population-normalized standard deviation (`ddof=0`) divided by $\sqrt{n_b}$, not uncertainty across independent training runs.

M Hyperparameters

Synthetic MHN experiment. $N = 12$ cluster memories, $d = 50$, cosine $c = 0.35$, inverse temperature $\beta = 8$, rotation size $\varepsilon = 0.02$, $n_{\text{rot}} = 2000$ random antisymmetric Ω, fixed-point tolerance 10^{-13}.

Synthetic MHN-BM experiment. $N = 12$ cluster memories, $d = 50$, $c = 0.35$, $\beta = 16$ (consistent in loss, update, and energy), data scale 2.0, rotation $\varepsilon_{\text{rot}} = 0.02$, $n_{\text{hidden}} = N + 1$, Adam with learning rate 0.02, batch size 64.

Stable Diffusion split CIFAR-10. Backbone: `runwayml/stable-diffusion-v1-5`; only the UNet is fine-tuned, VAE frozen. Task 1 classes: bird, cat, deer, dog, horse. Task 2 classes: airplane, automobile, frog, ship, truck. 25,000 samples per task, 32×32 images. Replay buffer $K = 5000$; replay selection $\in$ {no replay, random, high_energy}. Forgetting metric: reconstruction MSE at $t = 0.8$.

Pixel-space diffusion split CIFAR-10. PixelCNN-UNet trained from scratch in pixel space (32×32 RGB), "standard" channels $(128, 256, 512)$ ($\sim$45M params). Same living/non-living split as the Stable Diffusion setup, 25,000 samples per task. AdamW (lr 2×10^{-4}, weight decay 10^{-4}, 500 warmup steps), fp16, batch 512, 2000 epochs/task, EMA decay 0.9999. Linear noise schedule ($T = 1000$, $\beta \in [10^{-4}, 2 \times 10^{-2}]$), 50 inference steps. Replay buffer $K = 5000$; replay selection $\in$ {no replay, random, high_energy, low_energy}. Forgetting metric: reconstruction MSE at $t = 400$.

Pixel-space diffusion rotated MNIST. Same PixelCNN-UNet adapted to 28×28 grayscale. Two tasks defined by `rotation_angles` $= [0°, 30°]$. Optimizer, EMA, and noise schedule identical to the pixel-space CIFAR-10 setup; batch 256 for 200 epochs/task, no augmentation. Replay buffer $K = 5000$; replay selection $\in$ {no replay, random, high_energy, low_energy}. Forgetting metric: reconstruction MSE at $t = 400$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

Justification: The abstract and Section 1 explicitly enumerate the three contributions of the paper: (i) defining intrinsic forgetting as a rise in Hopfield energy and proving Theorem 4.1 (Section 4.1); (ii) the energy-dependent replay analysis stated as Theorem 5.1 (Section 5.1); and (iii) empirical validation on diffusion models in Sections 4.3 and 5.3 (Stable Diffusion v1.5 and a pixel-space DDPM on split CIFAR-10, with corresponding rotated-MNIST results in Appendix J). The scope is restricted to the equal-angular MHN regime in theory and to the listed diffusion-model setups in experiments.

Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: A dedicated *Limitations* paragraph in Section 6 discusses three concrete limitations: (i) we analyze *intrinsic* forgetting in MHN-equivalent score functions, while neural-network score models can exhibit additional forgetting that our theory does not cover; (ii) the theory is restricted to a finite number of memories rather than the data-manifold regime [32]; and (iii) the empirical validation is limited to images and to settings with explicit task boundaries, leaving online or non-image continual learning to future work.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions for Theorems 4.1 and 5.1 are stated in Sections 4.1 and 5.1 -- equal-angular cluster-plus-outlier configuration, small rotation $\varepsilon \ll 1$, and the sufficiently-low-temperature regime at fixed $N \geq 2, d, c$, followed by sufficiently small ε for those parameters. Proof sketches are given inline in the main text and the full derivations, including the symmetry reduction, localization conditions, off-memory expansion, and ordering arguments , are provided in Appendix C (forgetting) and Appendix H (replay), with additional stable-state exclusion in Appendix E. Appendices F and G state separately what their partial extensions do and do not establish.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The experimental setup is fully described: dataset splits, model backbones, replay policies, and evaluation metric in Sections 4.3 and 5.3 (with the rotated-MNIST replication in Appendix J); reconstruction-based forgetting is defined in Eq. (15) and discussed in Appendix K; and the full hyperparameter list (architecture, optimizer, learning rate, schedule, EMA, batch size, replay buffer size, evaluation timesteps) is given in Appendix M for every system reported (MHN, MHN-BM, Stable Diffusion split CIFAR-10, pixel-space split CIFAR-10, pixel-space rotated MNIST).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: All datasets used (CIFAR-10, MNIST) and the Stable Diffusion v1.5 backbone (`runwayml/stable-diffusion-v1-5`, [4]) are publicly available. The code is not released alongside the submission to preserve double-blind anonymity, but Appendix M documents every hyperparameter, dataset split, and replay configuration needed to reproduce the experiments. We will release the code publicly upon acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix M lists, for each experiment, the architecture, dataset split (CIFAR-10 living/non-living, rotated-MNIST $[0°, 30°]$), optimizer (AdamW with learning rate, weight decay, warmup steps, gradient clipping), batch size, number of epochs, mixed-precision setting, EMA decay, noise schedule ($T = 1000$, $\beta_{\text{start}}/\beta_{\text{end}}$), sampling steps, replay buffer size K, and the four replay policies. The reconstruction-based evaluation metric and its timestep choice are stated in Eq. (15) and analyzed in Appendix K.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: For the closed-form MHN and MHN-BM toy experiments we average over 2000 random rotations, and the diffusion-model figures aggregate over thousands of Task-1 samples per energy bin, giving stable mean estimates (Sections 4.3, 5.3, Figs. 5, 6, S.5). However, explicit error bars or confidence intervals are not displayed on the diffusion-model figures. We rely instead on the qualitative consistency of the trend across two independent diffusion-model systems (latent and pixel space) and across CIFAR-10 and rotated MNIST. Bootstrap CIs on the bin-mean reductions are a natural extension we leave to a camera-ready version.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Appendix M reports the per-experiment scale (mixed-precision fp16, batch sizes, number of epochs per task, sample counts for forgetting evaluation), but specific GPU type, total wall-clock time, and total compute budget are not reported in the current version. We will add a complete compute-resource accounting (GPU model, memory, training time per run, total project compute) in the camera-ready version.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses only publicly available standard benchmarks (CIFAR-10, MNIST) and a publicly released model (Stable Diffusion v1.5), does not involve human subjects or personally identifiable data, and does not introduce new high-risk generative capabilities. The research conforms to the NeurIPS Code of Ethics in every respect we are aware of.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is foundational research that provides a theoretical analysis and a diagnostic / replay-selection criterion for continual learning of generative models, rather than a deployable system or a new generative capability. We use only standard benchmarks (CIFAR-10, MNIST) and an existing publicly released backbone (Stable Diffusion v1.5). We are not aware of direct societal impacts beyond those already inherent in generic continual-learning research on these standard benchmarks.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release any new high-risk model, scraped dataset, or generative system. All experiments use a publicly released backbone (Stable Diffusion v1.5) and standard benchmarks (CIFAR-10, MNIST), so no additional safeguards are required.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All third-party assets are publicly available and are credited to their original authors: Stable Diffusion v1.5 (`runwayml/stable-diffusion-v1-5`, [4]), the standard CIFAR-10 split protocol, and MNIST. Citations appear in Section 4.3 and Appendix M. We use these assets in accordance with their public licenses and terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce or release any new dataset, pre-trained model, or other reusable asset. The contributions are theoretical analyses and an empirical evaluation procedure based on existing public assets.

    Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work does not involve any crowdsourcing or research with human subjects. All experiments are run on publicly available datasets (CIFAR-10, MNIST) and pre-trained models.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work does not involve human subjects or study participants, so no IRB review or equivalent approval is required.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

Justification: LLMs are not part of the core methodology, scientific rigor, or originality of this research. They were used only as writing/editing aids, which under the NeurIPS LLM policy does not require declaration.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.