LEARNING BY DISRUPTION: ACCELERATING GROKKING WITH MONTE CARLO PARAMETER SWAPPING

Anonymous authors
Paper under double-blind review

ABSTRACT

Grokking is a striking phenomenon in neural network training, where a model can undergo a prolonged period of pure memorization before abrupt generalization. While previous works have attempted to interpret it through classical machine learning mechanisms like weight norm, recent research draws an analogy from statistical physics, framing grokking as a form of computational glass relaxation. This theory defines the initial memorization as a result of 'fast cooling' where the training loss is reduced so quickly that a glass state is formed, followed by a 'slow relaxation' towards final generalization. Although providing a unifying framework for representative grokking theories, this perspective has remained largely at the theoretical macroscopic level without direct empirical validation on training dynamics. Here we introduce a three-component framework to directly characterize the training dynamics via parameter mobility (PM), and two representative measurements from glassy dynamics: replica correlation (RC) and fractal dimension (FD). We demonstrate that standard optimization presents clear signatures of glass dynamics and inherently traps the grokking network in a kinetically arrested memorization state with collapsed mobility, strong history dependence, and channel-like motions. This quantitative agreement motivates us to introduce State-Aware Monte Carlo Parameter Swapping (SAM-Swap), an optimization plug-in that can accelerate generalization, inspired by the swap Monte Carlo algorithm widely used in glass dynamics. Comparing SAM-Swap, weight decay, and Gaussian gradient noise, we find that accelerated generalization is consistently associated with random exploration in the parameter space, similar to diffusion in physics.

1 INTRODUCTION

Modern overparameterized neural networks (NNs) can perfectly memorize even randomly labeled finite training sets, yet they often generalize with surprising accuracy to unseen test data Zhang et al. (2016). Explaining this apparent contradiction remains one of the most fundamental open problems in AI Zhang et al. (2016); Belkin et al. (2019). Grokking exemplifies this paradox: models achieve near-perfect training accuracy while remaining at chance-level on test data (memorization), then abruptly generalize after prolonged training. First observed in transformer-based models trained on algorithmic datasets Power et al. (2022), grokking has later been reported across a broad range of architectures and domains including vision, language, and molecular learning tasks Liu et al. (2023); Humayun et al. (2024). The pronounced timescale separation between memorization and generalization makes grokking a useful testbed for mechanistic theories of generalization.

Existing theories of grokking can be broadly organized by the level where they characterize the memorization-to-generalization transition. Weight norm and landscape-based theories describe the process macroscopically, relating the delayed generalization to the evolution of the overall parameter norm through distinct training and test loss regimes, where test performance is favorable only within a narrower "Goldilocks" zone Liu et al. (2023). Circuit-based explanations offer a possible microscopic refinement of this picture, describing competitions between rapidly learned but parameter-inefficient memorizing computations and slowly formed, norm-efficient generalizing circuits Nanda et al. (2023); Merrill et al. (2023); Varma et al. (2023). Representation-level theories

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

instead emphasize the delayed feature learning and the gradual alignment of hidden representations with the underlying task structure Kumar et al. (2023); Liu et al. (2022). Optimization-centered studies further investigate the dynamics driving these changes, including the instability of adaptive optimizers Thilak et al. (2022), slowly varying gradients Lee et al. (2024), effective learning rates Prieto et al. (2025), and noise Ersoy & Wiesner (2026).

While these explanations illuminate different perspectives of grokking and are not mutually exclusive, they are formulated primarily in terms of machine learning-specific quantities. A recent study introduced a complementary physics-inspired perspective and proposed that grokking can be interpreted as a form of *computational glass relaxation*, which slowly evolves to a final equilibrium state Zhang et al. (2025). The key question is whether delayed generalization occurs because the generalizing solutions are intrinsically rare or separated from memorizing solutions by hard-to-access regimes of parameter space. By conceptualizing each parameter configuration as a possible state of a statistical-mechanical system, training loss becomes its effective energy, and the Boltzmann entropy measures the size of solutions that exist at different levels of training loss and test accuracy. Results show that generalization states occupy a much larger solution space than the memorization one, which also supports the *volume hypothesis* Chiang et al. (2023); Peleg & Hein (2024); Pakman et al. (2026). Moreover, number of accessible states does not collapse along the transition between memorization and generalization, where training loss decreases monotonically Zhang et al. (2025).

In physical terms, these suggest the absence of both macroscopic entropic and energetic barriers, respectively, making grokking a process free of macroscopic thermodynamic barriers. However, typical training trajectories via AdamW remain in a low-entropy regime with poor generalization for a long period, despite the transition to high-entropy generalizing states being both accessible and statistically favored. This motivates the interpretation of grokking as computational glass relaxation to an equilibrium state with max entropy.

While entropy analysis established the macroscopic accessibility of generalizing solutions, it does not characterize the dynamics directly but simply attributes grokking to slow parameter mobility caused by weak gradients after rapid loss reduction. Here, we focus directly on the evolution of optimization trajectories in parameter space. Specifically, we introduce a three-component framework combining parameter mobility (PM) with two representative measures from glassy dynamics: replica correlation (RC) Berthier & Biroli (2011); Mézard & Parisi (2012) and trajectory fractal dimension (FD) Datseris et al. (2023). PM quantifies the magnitude of stepwise parameter motion during training. RC measures inter-trajectory similarity between independently evolving replicas branched from the same initial state. FD characterizes intra-trajectory dynamics, indicating whether optimization paths remain persistent and channel-like or become tortuous. Thus, the three components respectively ask i) whether the network changes, ii) whether the same state can evolve into alternative paths, and iii) how tortuous each trajectory is.

With this framework, we find that for AdamW-optimized model, PM collapses once the network reaches memorization, despite the poor test performance. RC exhibits clear two-step relaxation coinciding with the memorization period and progressively slower decorrelation at longer waiting times, revealing strong history dependence and aging, which is a representative sign of glassy dynamics Kob & Andersen (1995); Kob & Barrat (1997); Li et al. (2026). Meanwhile, FD decays to approximately 1 after memorization, suggesting the slowed parameter motion becomes persistent and channel-like. This together provides a clear dynamical diagnosis for grokking, supporting the interpretation of grokking as glass-like kinetic arrest in parameter space.

To further test our kinetic arrest theory, we asked: if grokking mimics glassy dynamics, could interventions employed in glassy systems research also shorten the generalization delay in grokking systems? This motivates the development of State-Aware Monte Carlo Parameter Swapping (SAM-Swap), inspired by the swap Monte Carlo method widely used to accelerate equilibration in glass-forming systems by switching particle identities Grigera & Parisi (2001); Berthier et al. (2016). Following the same concept, SAM-Swap treats each parameter in NNs as a 'particle' and randomly exchanges their values within each layer. At first glance, such swaps might appear to destroy learned features and obstruct representation learning, as trained parameters are generally assumed to encode an organized computation Ainsworth et al. (2022) Yet, surprisingly, SAM-Swap substantially shortens the generalization delay. Analysis via our three-component framework further shows that SAM-Swap can restore PM, reduce RC, and increase FD, indicating a transition from arrested dynamics to accelerated relaxation toward equilibrium. We also compare this struc-

tured intervention with large-amplitude additive Gaussian noise and find it can produce a similar acceleration, showing how different forms of disruption alter the dynamical route to generalization.

In this work, we extend the emerging connection between grokking and glassy dynamics through a unified parameter-space framework based on PM, RC, and FD. Using these observables, we find long memorization is associated collapsed mobility, history-dependency, and channel-like trajectories, which provides direct dynamical evidence for interpreting grokking as glass-like kinetic arrest. Motivated by these observations, we introduce SAM-Swap, a low-FD-triggered perturbation inspired by swap Monte Carlo methods in glass physics, and find it can substantially accelerate grokking time, indicating counterintuitive benefits of disruption. We also systematically compare it against other perturbation strategies and show how they alter training dynamics and thus affect the grokking time from the microscopic level. Together, our results support the view of grokking as a shift between dynamically distinct training regimes, and suggest that stability- (RC) and geometry-based (FD) observables can inform targeted interventions for accelerating generalization.

2 METHODS

Here we follow the previously reported setup in the original glass theory work and experiment on a 1-layer transformer on the arithmetic task of $(x^2 + y) \bmod 67$ Zhang et al. (2025). Results are averaged over 20 independent runs, each trained for 10,000 epochs(t). All models use the same training dataset, with 30% of the available samples reserved for evaluating test accuracy. For ease of presentation, we define the onset of generalization as the point where the test accuracy first reaches 99%. Unless otherwise stated, all references to grokking time correspond to the period after training accuracy reaches 99%, and before test accuracy reaches 99%.

2.1 REPLICA-CORRELATION FUNCTIONS

To quantify the similarity between independently evolving training trajectories, we define a replica-correlation (RC) function based on the Pearson correlation between model parameters. We first train a model up to a waiting time t_{w}, after which independent training processes (replicas) are spawned and evolved separately under perturbations. The correlation of pairs of replicas at each t is measured.

Let $\mathbf{S}_l^a(t)$ denote the flattened parameter vector of layer l for replica a at training time t. We define the replica-correlation function as

$$q_l^{ab}(t, t_{\mathrm{w}}) = \mathrm{corr}\big(\mathbf{S}_l^a(t + t_{\mathrm{w}}), \mathbf{S}_l^b(t + t_{\mathrm{w}})\big), \tag{1}$$

where $\mathrm{corr}(\cdot, \cdot)$ denotes the Pearson correlation coefficient computed over all parameters in layer l. This ensures that the correlation reflects the scale-invariant alignment between parameter configurations. Here, $a, b = 1, 2, \ldots$ index different replicas. The reported correlation is averaged over all replica pairs, layers, and independent runs. High correlation indicates that nearby trajectories remain stable under perturbations, whereas a reduction in correlation signals that perturbations can lead to divergent evolution.

2.2 FRACTAL DIMENSION OF THE TRAINING TRAJECTORY

To characterize the geometric structure of neural network training dynamics, we estimate the fractal dimension (FD) of the parameter trajectory in weight space. Let $\mathbf{R}_t$ denote the flattened vector of all model parameters at training step t. We define the one-step increment as

$$\Delta\mathbf{R}_t = \mathbf{R}_{t+1} - \mathbf{R}_t,$$

and the cumulative arc length of the trajectory up to time t as

$$s_t = \sum_{t' \le t} \|\Delta\mathbf{R}_{t'}\|_2.$$

For a given time lag τ, we define the lag-dependent quantities

$$\Delta s(t, \tau) = s_{t+\tau} - s_t, \qquad \Delta R^2(t, \tau) = \|\mathbf{R}_{t+\tau} - \mathbf{R}_t\|_2^2.$$

We evaluate τ in the range of $\tau = 1, ...[t/2]$, ensuring that each lag fits entirely within the fractal window size (T). To reduce sensitivity to non-stationarity and large fluctuations in parameter updates, we aggregate these quantities using the median, obtaining $\Delta s(\tau)$ and $\Delta R^2(\tau)$. We then assume a scaling relation of the form

$$\Delta R^2(\tau) \propto \Delta s(\tau)^\alpha,$$

and estimate the exponent α via linear regression in log-log scale over a selected range of τ, yielding a time-dependent exponent $\alpha(t)$ and the corresponding FD

$$D_f(t) = \frac{2}{\alpha(t)}.$$

The fractal dimension quantifies the relation between the accumulated path length and the net displacement of the trajectory. Values of $D_f \approx 1$ indicate that the trajectory evolves in a relatively straight or ballistic manner, where displacement scales proportionally with path length. In contrast, larger values of D_f correspond to increasingly indirect or tortuous motion, where the trajectory covers more path length relative to its net displacement.

In our experiments, D_f is estimated within a sliding window of $T = 20$ epochs to capture local changes in trajectory geometry. We choose this value empirically, so that the window is sufficiently short to resolve the dynamical features emerging on the 10^2 epoch timescale (beginning of memorization), yet long enough to suppress excessive fluctuations in the FD estimate.

2.3 Fractal Window Displacement

To quantify parameter mobility (PM) within a training window, we measure the *total displacement* (dR) of each layer's weights. For a given window, this displacement is defined as the Euclidean distance between the parameter vector at the end of the window and the parameter vector at its beginning:

$$dR = \frac{1}{|l|} \sum_l \left\| \mathbf{R}_{t+T} - \mathbf{R}_t \right\|_2 ,$$

This measure captures the net change in parameter space over the interval of the FD estimation window, rather than the cumulative path length. For network-level analysis, the displacement values are averaged across all trainable layers.

2.4 Small Additive Gaussian Noise as Probe

We use full-batch Adam or AdamW without dropout to eliminate intrinsic sources of randomness, so that the noise level in the system is explicitly controlled. Replicas branched from the same state remain identical with $RC = 1$, and FD is trivially 1 in the absence of perturbations. Above this deterministic baseline, we inject zero-mean Gaussian noise into all parameters with $\sigma = 10^{-5}$. The value is chosen from control experiments where injecting the noise show no change in macroscopic training performance, yet enabling meaningful evaluation of RC and FD. Models subjected to this protocol are marked as probed.

3 Results

3.1 Parameter Mobility and Slow Relaxation

As shown in Figure. 1b, PM exhibits a clear transition from learning to memorization: dR drops by orders of magnitude, from ~ 1 to $\sim 10^{-2}$ between $t \approx 0 - 10^2$, and continues to decrease toward $\sim 10^{-4}$ by the end of training. This progressive stagnation explains the slow learning dynamics, with mobility becoming vanishingly small around $t \approx 10^4$. Such behavior parallels slow relaxation in glassy systems, where dynamics progressively freeze and escape from the arrested state becomes increasingly difficult. The PM curves for different waiting times largely overlap, except for the initial datapoint, which reflects a larger first update due to untracked stochastic effects, while subsequent updates remain stable.

3.2 Two-step Relaxation of Replica Correlation

As shown in Figure. 1c, RC exhibits a characteristic two-stage behavior for the model with $t_w = 0$. During the initial phase of training, RC begins to decorrelate immediately, as expected since replicas are initialized with different random perturbations and therefore follow distinct optimization paths. As training progresses and test accuracy diverges from training accuracy (Figure 1a), marking the onset of grokking, RC settles into a plateau. This plateau suggests that the trajectory is confined within a restricted region of parameter space and remains deterministic to small perturbations. At later times, coinciding with the onset of generalization, RC decreases markedly for the perturbed model, reflecting reduced inter-trajectory similarity. This indicates that the small perturbation have effectively separated replicas' trajectories, suggesting that there are alternative solutions accessible in the proximity at this time of the optimization path. Together these observations suggest that during the memorization phase the optimization trajectory remains constrained, with replicas confined to a small region of parameter space and converging to only a few similar solutions, whereas at generalization perturbations separate replicas and reveal nearby alternatives, enabling broader exploration of parameter space.

3.3 Aging Behavior in Training Dynamics

A signature of glass dynamics is observed in the dependence of replica correlation on the waiting time t_w, as shown in Figure 1c. For replicas branched early ($t_w = 0$), the correlation exhibits a two-step decay, consisting of an initial drop followed by a plateau and a subsequent relaxation, as we detailed above. In contrast, for larger waiting times (e.g. $t_w = 10^2, 10^3$), the initial decay is strongly suppressed, and the correlation remains close to $q \approx 1$ over an extended period before exhibiting only a minor decrease at later times.

This dependence on t_w indicates that the system exhibits glass-like aging behavior: the response to perturbations depends on how long the system has already evolved. In particular, trajectories that have evolved longer become increasingly stable to perturbations and fail to exhibit rapid relaxation. The coexistence of two-step relaxation at $t_w = 0$ and aging provides compelling evidence for glass-like dynamics, where the system enters a slow, history-dependent regime with constrained exploration Li et al. (2026). Within our framework, this behavior suggests the emergence of a confined, perturbation-stable phase during grokking, which is eventually destabilized as the system regains dynamical flexibility and resumes exploration of parameter space.

3.4 Dynamical Transition in Trajectory Geometry

As shown in Figure 1d, the evolution of the FD exhibits a clear correspondence to different training stages. Initially, D_f fluctuates moderately above 1, reflecting nontrivial exploration of parameter space as the model fits the training data, consistent with relatively high PM. Once training accuracy reaches 99%, D_f collapses toward 1, indicating that the optimization trajectory becomes effectively low-dimensional and ballistic-like. Even under perturbations, the model is not allowed to stray far from a narrow, channel-like paths, reflecting strong confinement of the dynamics. This interpretation is further supported by the temporal alignment between the RC plateau and the collapse of D_f, consistent with restriction to a limited set of directions in parameter space and suppressed PM.

After a prolonged period of $D_f = 1$, we observe a gradual increase in FD, signaling an expansion of the explored parameter space. This increase occurs on a similar timescale as the observed RC decrease. At this time, the model has already generalized; this behavior then reflects the presence of multiple nearby minima. Perturbations induce diffusion in the proximity of the solutions, while the system remains strongly confined to a single basin with low PM, analogous to being trapped in a sharp minimum.

The joint evolution of PM, RC and FD supports a two-stage dynamical picture: an initial relaxation followed by a confined, perturbation-stable regime, and a subsequent increase in sensitivity where perturbations can induce more exploratory dynamics. This supports the view that grokking corresponds to a phase in which the training trajectory is intrinsically low-dimensional, with stable and slow dynamics. We note, however, that alternative accounts based on slow norm-minimizing drift under weight decay have also been proposed, and the present measurements do not uniquely distinguish between these explanations Boursier et al. (2025). Our measurements do not uniquely

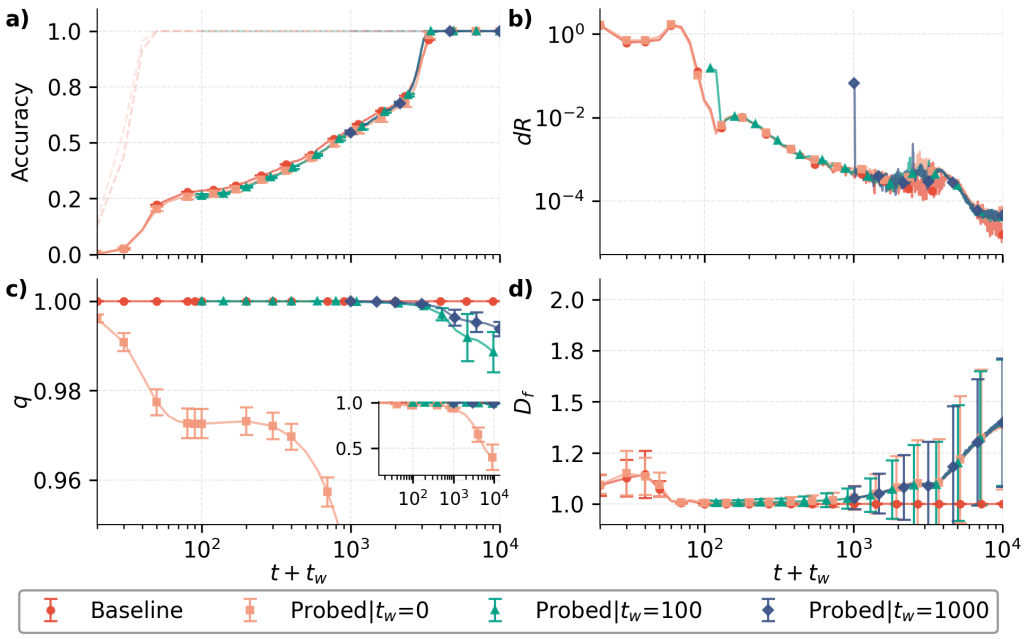

Figure 1: **Comparison across waiting times**. All data are from AdamW with learning rate 10^{-2} and weight decay 10^{-1}, optimized for shortest grokking time. Baseline uses AdamW; probed model adds small Gaussian noise to test trajectory sensitivity. The x-axis is aligned to $t + t_w$. (a) Test accuracy (solid lines show Test accuracy; translucent line shows training accuracy).(b) Parameter mobility, error bars were omitted to declutter the plot and provide a clearer view of the decreasing ΔR. Notable fluctuations are observed in the range from 0 to 10^2. (c) RC average across layers, with the insert showing the full range of replica decorrelation. (d) FD of Path average across layers. Occasionally, the regression yields FD estimates slightly below 1. This arises from fluctuations introduced by added noise, which perturb the slope in the log–log fit. As such, values below 1 are artifacts of estimation instability rather than evidence of sub-linear trajectories.

distinguish between these explanations; instead, they provide a complementary dynamical characterization of the optimization trajectory and the evolution of parameter-space exploration throughout the delayed generalization phase.

3.5 State-Aware Monte Carlo Parameter Swapping (SAM-Swap)

To further examine our interpretation of grokking as a glass-like dynamical phenomenon, we introduce an intervention designed to disrupt the training trajectory. Our core hypothesis is: disrupting constrained trajectories facilitates exploration and accelerates generalization, similar to real glassy systems Berthier & Biroli (2011). Empirically, grokking onset coincides with FD dropping toward 1, indicating low-dimensional dynamics. We use a threshold on D_f to indicate entry into the constrained regime, selected empirically as $d_{f,\text{glass}} = 1.1$ to align with the observed onset of arrested dynamics. Our experiments indicate that the exact threshold value is not critical as long as $d_f > 1$; moderate adjustments yield comparable behavior, whereas extreme settings (e.g., $d_{f,\text{glass}} \approx 1.01$) produce similar overall trends but lead to highly fluctuating PM.

When the FD threshold is reached, we apply a swap-based perturbation by randomly exchanging parameter pairs within the same layer. Grigera & Parisi (2001); Berthier et al. (2016) The amount of parameters swapped depends on the swap ratio r. We find that $r = 10^{-2}$ yields the most stable training and fastest generalization, while smaller ratios recover baseline grokking and larger ratios tend to induce excessive fluctuations. For details of the experiment regarding glass limit and swap ratio, please refer to the technical supplement. Such an exchange operation preserves the overall distribution of parameters while redistributing the configuration in parameter space. To prevent

Algorithm 1 State-Aware Swap Perturbation with Fractal Dimension Trigger

Require: Model parameters θ, training data $\mathcal{D}$, swap ratio r, glass threshold df_{glass}
Save training loss at $t = 0$: L_{cap}
for each transformer layer ℓ **do**
  Compute fractal dimension df_ℓ
  if $df_\ell < df_{\text{glass}}$ **then**
    Store backup $\theta_{\text{backup}} \leftarrow \theta$
    Sample $2 \cdot \lfloor rN_\ell \rfloor$ random indices $\{i_k\}$ in layer ℓ
    for $k = 1$ to $\lfloor rN_\ell \rfloor$ **do**
      Swap $\mathbf{v}_\ell[i_{2k-1}]$ and $\mathbf{v}_\ell[i_{2k}]$
    end for
    Compute new training loss L_{new}
    if $L_{\text{cap}} < L_{\text{new}}$ **then**
      $\theta \leftarrow \theta_{\text{backup}}$ {reject}
    end if
  end if
end for
return updated θ

disruptive updates, we impose a loss cap L_{cap} equal to the initial training loss, rejecting perturbations that lead to significantly large error. This safeguard improves robustness without altering overall dynamics.

We select FD as the trigger for swap interventions because alternatives are less practical: RC requires multiple replicas, so tracking it would be computationally expensive, and PM's magnitude depends heavily on the optimizer, initialization, and hyperparameters. In contrast, FD can be estimated directly from a single training trajectory and provides measurement of local trajectory complexity. Rather than perturbing the optimization process continuously, the FD trigger enables interventions to be applied only when trajectory dynamics become sufficiently confined. This allows SAM-Swap to achieve acceleration comparable to strong Gaussian perturbations while requiring substantially fewer perturbation events.

As shown in Figure 2a, the swap intervention substantially reduces the grokking time, with models generalizing at approximately $t \approx 650$ epochs compared to $t \approx 3000$ epochs for the AdamW model. This result supports the view that the grokking regime corresponds to a dynamically stable phase in which the training trajectory is constrained and resistant to perturbations. The swap intervention instead disrupts this stability, allowing the optimization process to follow alternative trajectories that reach generalization more efficiently. More broadly, these results show that the timing of generalization depends not only on the structure of the loss landscape, but also on the specific trajectory taken during optimization, and how it is influenced by perturbations.

3.6 Comparing perturbation methods

Recent studies demonstrate that stochastic perturbations can strongly affect grokking dynamics, often accelerating or even eliminating delayed generalization Agrawal & Kumar (2025); Ersoy & Wiesner (2026). To investigate how perturbations influence grokking dynamics, we consider four training settings: (i) Adam probe, (ii) AdamW with probe, (iii) Adam with large Gaussian noise, and (iv) Adam with the SAM-Swap introduced in the previous section.

It is important to distinguish between perturbations used for measurement and those used as interventions. Probing refers to the small Gaussian noise described in Methods. As a genuine intervention, we instead use larger noise applied in the same way but with $\sigma = 10^{-2}$, deliberately altering the optimization dynamics to promote escape from the constrained low-FD regime. The proposed swap method serves a similar purpose, but induces structured non-local rearrangements of parameters rather than additive stochastic fluctuations.

As shown in Figure 2a, training setting (ii),(iii) and (iv) all reaches generalization earlier than the baseline setting (i), though with varying effectiveness. Across all interventions examined, acceler-

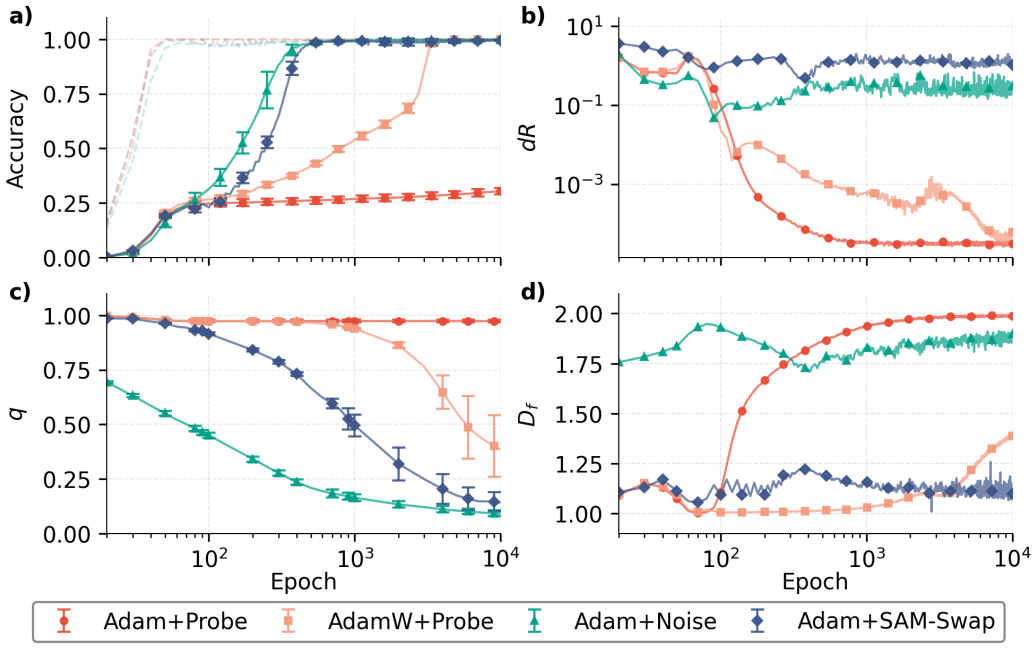

Figure 2: **Comparison across perturbation methods.** Adam uses a learning rate of 10^{-2}, while AdamW additionally applies weight decay of 10^{-1}. Both models are probed with weak Gaussian noise ($\sigma = 10^{-5}$), whereas Adam+Noise employs stronger perturbations ($\sigma = 10^{-2}$) and Adam+Swap uses the proposed SAM-Swap. Hyperparameters are chosen to minimize grokking time. (a) Test accuracy (solid lines) and training accuracy (translucent lines). (b) Parameter mobility. (c) Replica correlation from $t_w = 0$. (d) Fractal dimension of path. Error bars are omitted for clarity; Adam+Noise and Adam+Swap exhibit large fluctuations due to perturbations, so we present average trends.

ated generalization is consistently associated with increased parameter-space exploration, indicating that exploration itself may be the controlling variable for escaping the memorization regime.

3.7 The three component framework

Drawing together the evidence from PM, RC, and FD, we can now articulate a three-component framework for understanding the dynamics of grokking. Within this framework, the prolonged memorization stage of grokking emerges as a glass-like kinetic arrest in parameter space: the network is prevented from generalization as its dynamics become slow, history-dependent, and geometrically constrained. Although the generalizing solutions are abundant and accessible in principle Zhang et al. (2025), the optimization process has difficulty exploring alternative parameter configurations. This glass-like arrest can be systematically characterized through the three components of mobility, history-dependence, and trajectory geometry.

Mobility, measured by PM, quantifies the magnitude of parameter changes during training. Prior work Zhang et al. (2025) shows that in grokking systems, the training loss reduces too quickly, gradients shrink while parameters stagnate. Our experiments also confirm this point as the parameters become nearly immobile after memorization, even though the test performance remains poor, as shown in Figure 2b. However, low mobility alone does not necessarily impose delay in generalization, just like slowing dynamics itself does not establish glass phenomena in Dyre (2006); Berthier & Biroli (2011). A model could move slowly just because of some specific optimization settings, for example, a decaying learning rate.

History dependence, probed by RC, is a key feature of glass systems in physics Berthier & Biroli (2011); Mézard & Parisi (2012). While our baseline dynamics show a clear two-step relaxation pat-

tern, the slower decorrelation between replicas created after long waiting times further supports the analogy between grokking and glass relaxation. This mimics glass aging: the longer the arrest, the harder escaping to an alternative trajectory becomes. In other words, a high RC suggests that even when nearby copies of the model are given an opportunity to evolve independently, they continue to follow highly similar trajectories. The optimizer therefore has limited access to alternative routes in parameter space. When high RC coincides with low mobility, it indicates that slow dynamics are not merely small updates but may stem from limited alternative paths available to the optimizer, as shown in Figure 2c. Nevertheless, simply reducing the RC alone does not guarantee accelerated generalization. For an extreme example, RC can be simply reduced to zero by resetting all parameters after memorization, which destroys all the learned representations and cannot accelerate generalization.

Trajectory geometry, measured by FD, compares net displacement to path length. During grokking, FD decays to 1 after memorization, indicating straight, channel-like paths. This suggests that the optimization follows a highly constrained motion: although the model could accumulate small updates, those updates cannot freely explore the parameter space. It further explains why low mobility itself is insufficient, as two models could move by a similar total distance, with one following the same narrow direction while the other randomly explores many alternative directions before reaching the same level of generalization. Additionally, FD alone is not sufficient, as shown in the Adam+Probe experiment in Figure2d. Without considering PM or RC, a high FD only indicates diffusive motion, but does not tell whether the exploration remains local or not.

In summary, these three observables characterize complementary levels of the optimization dynamics. PM is a local, within-trajectory quantity that measures the magnitude of ongoing parameter motion. FD is also evaluated within individual trajectories but characterizes their larger-scale geometry, distinguishing persistent channel-like motion from more tortuous exploration. RC is instead an inter-trajectory quantity, which compares an ensemble of independently perturbed trajectories originating from the same network state and measures whether the system can access distinct future evolutions. Their combination distinguishes general slow motion from a state where parameter update is simultaneously suppressed, geometrically channelized, and restricted from accessing alternative solutions. Only when taken together do the three components provide a full picture: low PM may reveal that the system is not truly moving away from its initial solution despite apparent motion, while high RC indicates that replicas remain correlated, implying there are no alternative directions being explored; a high FD, on its own, simply reflects that the trajectory is tortuous, but without PM and RC it cannot distinguish whether such motion corresponds to genuine diffusion across minima or merely local wandering, where RC provide the answer for whether there are other minima to diffuse to. Therefore, PM, RC and FD jointly distinguish true diffusion and exploration of nearby solutions from motion that is geometrically complex but dynamically confined.

Our framework also explains why grokking can be accelerated by structured perturbations like SAM-Swap. Accelerating grokking via random exploration is similar to diffusion in physics, as evident by increased mobility, decaying RC, and high FD, which enables the system to escape kinetic arrest and thereby shortening the generalization delay. This is also consistent with previous work from the physics lens: diffusion causes an increase in entropy, and entropy is positively correlated with generalization in a variety of neural networks Zhang et al. (2025); Yang et al. (2026).

4 Conclusion and Future Directions

Overall, our findings suggest that the timing of generalization is determined not only by the structure of the loss landscape, but also by the regions of parameter space accessible during optimization. By jointly measuring PM, RC and FD, we provide a practical framework for identifying when training becomes dynamically constrained and when it regains exploratory behavior capable of reaching generalizable solutions. Our experiments focus on modular arithmetic grokking and employ a local FD estimator based on a fixed temporal window. Future work should investigate multiscale trajectory estimators and evaluate whether the observed dynamics and FD-triggered interventions generalize to other grokking tasks, overparameterized memorization settings, and vision and language models.

ETHICS STATEMENT

This work is a methodological study of training dynamics in controlled algorithmic tasks, aimed at understanding memorization, delayed generalization, and perturbation sensitivity. The main risk lies in interpreting models trained under perturbative dynamics: they may reach similar test accuracy to conventional ones yet differ in representation, robustness, or calibration. In this paper, models are trained only to minimize grokking time for studying dynamics. Another risk is interpretability: FD and RC are indirect diagnostics of dynamics. They are not guarantees of robustness, reliability, fairness, or safety. Overstating them may mislead conclusions, especially in larger or realistic systems.

AI USE STATEMENT

In this work, large language models were used for assistance in improving writing clarity, grammar, formatting, and programming support. We have not used AI tools to generate scientific hypotheses, design experiments, interpret results, or formulate conclusions. All research ideas, theoretical developments, experimental methodologies, data analysis, and scientific interpretations presented in this work were conceived and conducted by the authors.

REFERENCES

Tathagat Agrawal and Manoj Kumar. Grokking in neural networks: a review. *SN Computer Science*, 6(3):4182, 2025.

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries, 2022.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. doi: 10.1073/pnas.1903070116.

Ludovic Berthier and Giulio Biroli. Theoretical perspective on the glass transition and amorphous materials. *Reviews of Modern Physics*, 83(2):587–645, 2011. doi: 10.1103/RevModPhys.83.587.

Ludovic Berthier, Joseph Coslovich, Andrea Ninarello, and Misaki Ozawa. Efficient equilibration of hard spheres up to the jamming density and beyond with swap monte carlo. *Physical Review Letters*, 116(23):238002, 2016. doi: 10.1103/PhysRevLett.116.238002.

Etienne Boursier, Scott Pesme, and Radu-Alexandru Dragomir. A theoretical framework for grokking: Interpolation followed by riemannian norm minimisation. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, 2025.

Ping-Yeh Chiang, Renkun Ni, David Yu Miller, Arpit Bansal, Jonas Geiping, Micah Goldblum, and Tom Goldstein. Loss landscapes are all you need: Neural network generalization can be explained without the implicit bias of gradient descent. In *International Conference on Learning Representations (ICLR)*, 2023.

George Datseris, Inga Kottlarz, Anton P. Braun, and Ulrich Parlitz. Estimating fractal dimensions: A comparative review and open source implementations. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 33(10):103123, 2023. doi: 10.1063/5.0160702.

Jeppe C. Dyre. The glass transition and elastic models of glass-forming liquids. *Reviews of Modern Physics*, 78(3):953–972, 2006. doi: 10.1103/RevModPhys.78.953.

Ibrahim Talha Ersoy and Karoline Wiesner. Noise-driven escape from metastable phases explains grokking in deep neural networks, 2026.

T. S. Grigera and G. Parisi. Fast monte carlo algorithm for supercooled soft spheres. *Physical Review E*, 63(1):011101, 2001. doi: 10.1103/PhysRevE.63.011101.

Ahmed Imtiaz Humayun, Randall Balestriero, and Richard Baraniuk. Deep networks always grok and here is why. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20722–20745. PMLR, 2024.

Walter Kob and Hans C Andersen. Testing mode-coupling theory for a supercooled binary lennard-jones mixture. ii. intermediate scattering function and dynamic susceptibility. *Physical Review E*, 52(4):4134, 1995.

Walter Kob and Jean-Louis Barrat. Aging effects in a lennard-jones glass. *Physical review letters*, 78(24):4581, 1997.

Tanishq Kumar, Blake Bordelon, Samuel J. Gershman, and Cengiz Pehlevan. Grokking as the transition from lazy to rich training dynamics, 2023.

Jaerin Lee, Bong Gyun Kang, Kihoon Kim, and Kyoung Mu Lee. Grokfast: Accelerated grokking by amplifying slow gradients, 2024.

Bin Li, Deng Pan, Ting Qu, and Yuliang Jin. Universal activated aging and weak ergodicity breaking in spin and structural glasses. *Science Advances*, 12(7):eaec4416, 2026. doi: 10.1126/sciadv.aec4416.

Ziming Liu, Ouail Kitouni, Niklas S. Nolte, Eric Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: an effective theory of representation learning. *Advances in Neural Information Processing Systems*, 35:34651–34663, 2022.

Ziming Liu, Eric Michaud, and Max Tegmark. Omnigrok: grokking beyond algorithmic data. In *Proceedings of the International Conference on Learning Representations*, 2023.

William Merrill, Nikolaos Tsilivis, and Aman Shukla. A tale of two circuits: Grokking as competition of sparse and dense subnetworks, 2023.

Marc Mézard and Giorgio Parisi. Glasses and replicas. In Peter G. Wolynes and Vassiliy Lubchenko (eds.), *Structural Glasses and Supercooled Liquids: Theory, Experiment, and Applications*, chapter 4, pp. 151–191. Wiley, 2012. doi: 10.1002/9781118202470.ch4.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *Proceedings of the International Conference on Learning Representations*, 2023. URL https://mlanthology.org/iclr/2023/nanda2023iclr-progress/.

Ari Pakman, Lior Kreimer, and Yakir Berchenko. Revisiting the volume hypothesis, 2026.

Amit Peleg and Matthias Hein. Bias of stochastic gradient descent or the architecture: Disentangling the effects of overparameterization of neural networks. In *International Conference on Machine Learning*, pp. 40154–40184. PMLR, 2024.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets, 2022.

Lucas Prieto, Melih Barsbey, Pedro Mediano, and Tolga Birdal. Grokking at the edge of numerical stability. In *International Conference on Learning Representations*, 2025.

Vimal Thilak, Etai Littwin, Shuangfei Zhai, Omid Saremi, Roni Paiss, and Joshua M. Susskind. The slingshot mechanism: An empirical study of adaptive optimizers and the grokking phenomenon. In *NeurIPS 2022 Workshops: Human Interpretability of Machine Learning Systems (HITY)*, 2022. URL https://mlanthology.org/neuripsw/2022/thilak2022neuripsw-slingshot/.

Vikrant Varma, Rohin Shah, Zachary Kenton, János Kramár, and Ramana Kumar. Explaining grokking through circuit efficiency, 2023.

Entao Yang, Xiaotian Zhang, Yue Shang, and Ge Zhang. High-entropy advantage in neural networks' generalizability. *npj Artificial Intelligence*, 2(1):44, 2026.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization, 2016.

Xiaotian Zhang, Yue Shang, Entao Yang, and Ge Zhang. Is grokking a computational glass relaxation? In *Advances in Neural Information Processing Systems*, volume 38, 2025.

A Appendix

A.1 Alternative perturbation mechanisms

In the main text, we focus primarily on the pair-swap intervention, where randomly selected pairs of parameters within a layer are exchanged. However, pair swapping is not the only perturbation mechanism we tested, as shown in Figure.3. We also evaluated several alternative parameter-space perturbations. In the *random replacement* method, a fraction $x\%$ of parameters is replaced by values sampled from a Gaussian distribution whose scale is chosen according to the range and standard deviation of the existing parameters, so that the injected values remain comparable to the current parameter scale. In the *zero-out* method, a fraction $x\%$ of parameters is set directly to zero. In the *replacement* method, instead of swapping two parameters, one parameter is replaced by the value of another, i.e. $\theta_{i_1}, \theta_{i_2} \rightarrow \theta_{i_1}, \theta_{i_1}$. Finally, in the *max-replacement* method, we use the same replacement operation but choose the larger-magnitude parameter from the selected pair as the replacement value.

Across these perturbation mechanisms, we find a consistent qualitative effect: all tested methods reduce the grokking time relative to the AdamW baseline and produce changes in replica correlation and fractal dimension similar to those discussed in the main text. This suggests that the acceleration of generalization is not unique to pair swapping itself. These results are consistent with the hypothesis that parameter perturbations help the optimization trajectory escape slow, glass-like dynamics and enter a more exploratory regime, thereby helping the model reach generalization earlier. The different perturbation methods may nevertheless differ in their stability, effect on training loss, and degree of disruption, which motivates further systematic comparison in future work.

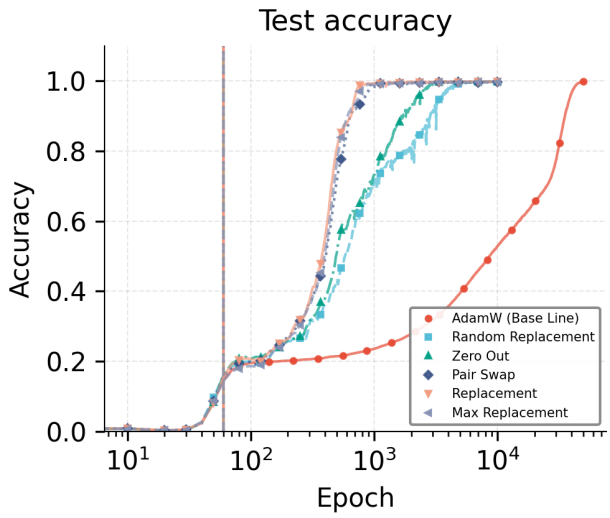

Figure 3: **Comparison of alternative parameter-space perturbation mechanisms:** The vertical dashed line denotes the epoch at which all models first reach 100% training accuracy. Despite operating through different perturbation mechanisms, all methods substantially reduce grokking time relative to the AdamW baseline.

A.2 Sensitivity of Fractal Dimension to Window Size

To evaluate the robustness of the fractal dimension (FD) measurement, we computed the sliding-window FD using window sizes of 10, 20, 50, 100, 200, and 500 epochs on the AdamW+probe model (Figure. 4). Across all window sizes below 200 epochs, we observe the same qualitative behavior: an initially elevated FD, followed by a collapse toward $D_f \approx 1$ after memorization, and a subsequent increase coinciding with generalization.

The primary effect of increasing the window size is temporal smoothing. Smaller windows (10 and 20 epochs) capture finer dynamical features and more accurately resolve the onset of the low-FD regime, which closely coincides with the point at which training accuracy reaches 100% (approximately epoch 60). Larger windows (50–200 epochs) exhibit the same overall trend but respond more slowly, as each estimate contains a substantial fraction of earlier trajectory segments from the more fluctuating training phase. Consequently, the transition into the low-FD regime appears delayed and less sharply defined.

For a window size of 500 epochs, the estimate is dominated by the later plateau phase of training. As a result, the early transition is largely obscured and the first reported FD values already correspond to a trajectory that has substantially relaxed.

Based on these observations, we use a window size of 20 epochs throughout the paper. This choice provides sufficient temporal resolution to resolve the early dynamical transition while remaining stable enough to suppress excessive fluctuations in the FD estimate. Excessively short windows produce noisier FD measurements that may repeatedly trigger SAM-Swap as the estimated FD fluctuates around the threshold. Although the impact of such frequent triggering remains an open question, it would complicate interpretation of the resulting dynamics. We therefore adopt a 20-epoch window as a practical compromise between sensitivity and robustness.

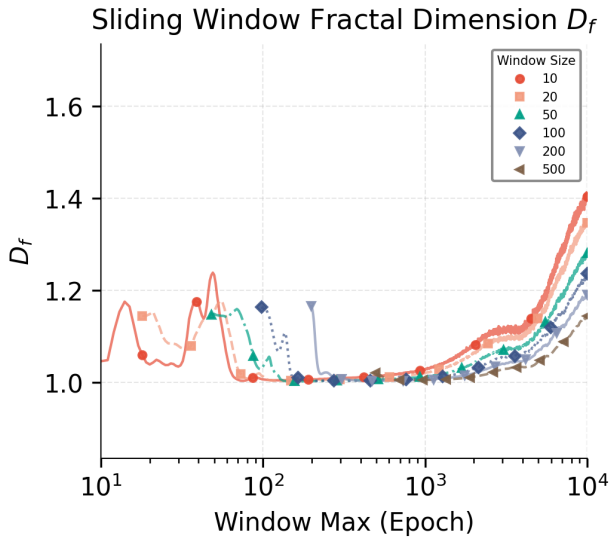

Figure 4: **Sensitivity of the sliding-window fractal-dimension estimate to window size:** FD is computed using window sizes of 10, 20, 50, 100, 200, and 500 epochs for the AdamW+probe model.

A.3 Performance of Swap mechanism without Loss Cap

We compare the proposed SAM-Swap with and without the acceptance criterion described in the main text. As shown in Figure. 5, both variants successfully accelerate generalization to nearly the same extent, indicating that the reduction in grokking time is primarily driven by the swap intervention itself. The primary difference appears after generalization has been achieved. Without the loss cap, the test accuracy exhibits relatively larger fluctuations and a slight downward trend during extended training. Although this degradation remains small within the 10,000-epoch training window, with test accuracy remaining above approximately 98%, the observed behavior suggests that occasional disruptive swap events can disturb an already-generalized solution. In contrast, the loss-cap version displays more stable post-generalization behavior.

The loss cap was therefore introduced as a practical safeguard rather than as a mechanism for accelerating generalization. Its purpose is to reject unusually destructive perturbations and reduce the impact of rare disruptive updates. While we do not observe catastrophic failures within our experiments, the mild downward trend observed in the unguarded variant motivates the use of the loss cap as a conservative measure to improve robustness. Since the primary goal of this work is to evaluate whether SAM-Swap perturbations can shorten grokking time, we adopt the safeguard to ensure that post-generalization behavior remains stable while preserving the benefits of the SAM-Swap.

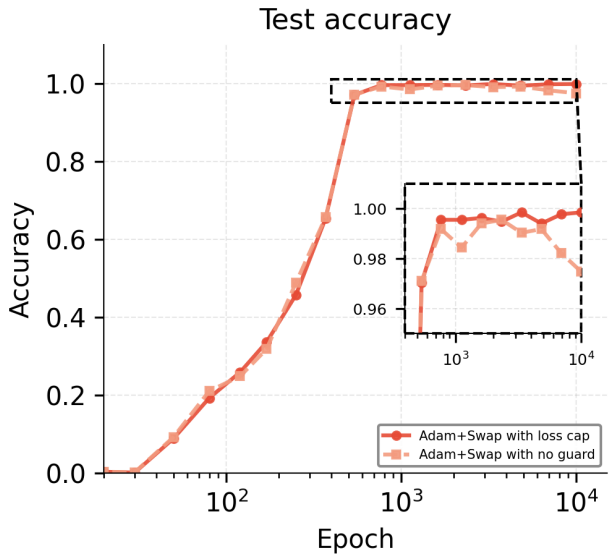

Figure 5: **Representative test-accuracy trajectories for the swap intervention with and without the loss-cap safeguard:** Both methods reach generalization substantially earlier than the baseline. Following generalization, the unguarded variant exhibits slightly larger fluctuations and a gradual downward tendency in test accuracy. The insert provides a magnified view of this region.

A.4 Swap Ratio and Training Dynamics

To assess how the fraction of swapped parameters influences optimization under SAM-Swap, we systematically varied the swap ratio over several orders of magnitude. The results reveal a non-linear dependence of the training dynamics on swap strength (Figure 6). Among all tested values, a swap ratio of 10^{-2} provided the most favorable balance between exploration and stability, consistently accelerating generalization while preserving smooth and stable training trajectories.

Increasing the swap ratio to 10^{-1} resulted in substantially more disruptive dynamics. Although generalization could still occur, the training trajectory exhibited pronounced oscillations in both loss and accuracy, indicating that excessively large perturbations interfere with the optimizer's ability to converge smoothly. At the opposite extreme, a swap ratio of 10^{-4} produced behavior nearly indistinguishable from the baseline. This is expected, as the largest weight matrices contain only $\sim 6 \times 10^4$ parameters, while many other layers are considerably smaller; at such low swap ratios, only a handful of parameter pairs are exchanged at each intervention, making the perturbation too weak to alter the optimization trajectory meaningfully.

An unexpected observation emerged at the intermediate ratio of 10^{-3}. Contrary to the intuition that training instability should increase monotonically with perturbation strength, models trained with 10^{-3} exhibited larger fluctuations than those trained with 10^{-1}. This suggests that the relationship between swap strength and optimization dynamics is not governed solely by the magnitude of the perturbation, but also by how the perturbation interacts with the underlying training trajectory and inter-layer dynamics.

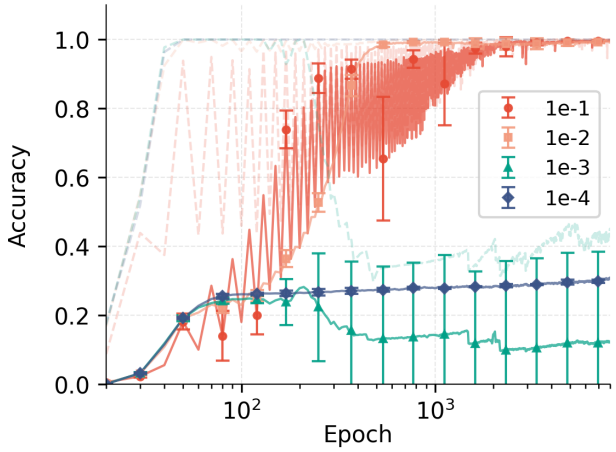

Figure 6: **Test Accuracy under different swap ratios**. Each curve corresponds to a swap ratio (10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}), as indicated in the legend.

A.5 TEST LOSS UNDER ACCELERATED GENERALIZATION.

As shown on the left side of Figure.7, in several perturbation-based runs, we observe that the test loss may increase after the model has already entered the generalization regime and achieved strong predictive performance. Importantly, this increase in loss does not necessarily indicate a degradation of generalization quality, as the corresponding test accuracy remains stable and high. Instead, the behavior resembles a form of post-generalization divergence, where continued optimization substantially alters prediction confidence while leaving the learned decision boundaries largely unchanged.

To investigate this effect, we decomposed the test cross-entropy into contributions from correctly and incorrectly classified samples (Right side of Figure 7). The loss from correctly classified examples remains near zero throughout training, reflecting consistently high confidence. In contrast, the loss from misclassified examples grows sharply at later stages, showing that the rising test loss stems from increasingly confident errors rather than reduced confidence in correct predictions.

This explains the apparent decoupling between accuracy and cross-entropy: accuracy depends only on correctness, while cross-entropy penalizes confidence. Thus, a few highly confident mistakes can disproportionately inflate loss without affecting accuracy. Since our focus is on whether perturbative dynamics shorten the memorization-to-generalization delay, this late-stage divergence is secondary and can be mitigated with early stopping, which halts training once optimal generalization is reached.

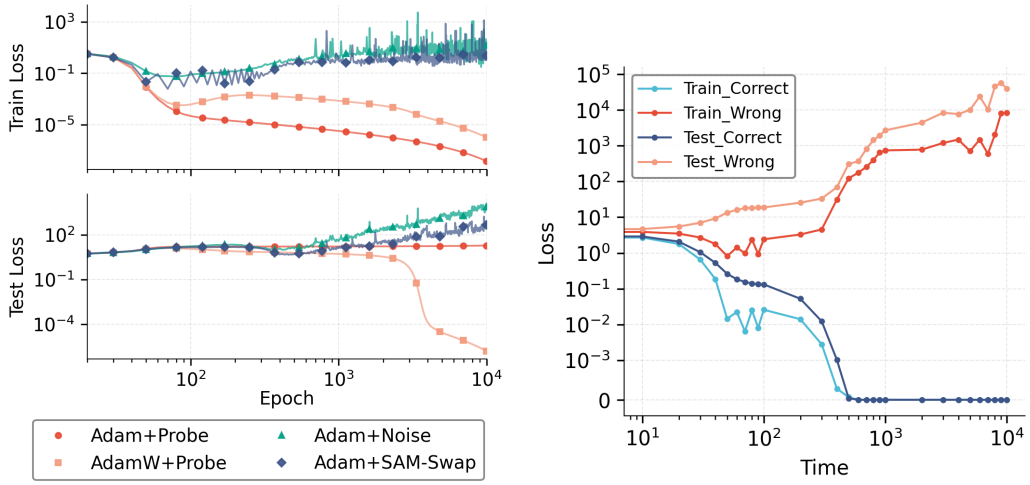

Figure 7:
LEFT: Average Test-loss evolution under perturbation-based interventions: Both Adam+Large Noise and Adam+Swap achieve rapid generalization, reaching 100% test accuracy at approximately 600 epoches as mention in the main text. Following generalization, the test loss gradually increases despite test accuracy remaining high. The onset and magnitude of this increase vary between runs due to the stochastic nature of the perturbations.

RIGHT: Decomposition of training and test cross-entropy losses. Shown for a representative Adam+SAM-Swap run. The losses associated with correctly classified samples rapidly converge toward zero after the onset of generalization and remain negligible throughout training. In contrast, the losses associated with misclassified samples continue to increase and eventually dominate the total loss despite little change in classification accuracy. The late-stage rise in test loss therefore originates from increasingly confident errors rather than from a degradation of predictive performance.

A.6 GLASS LIMIT SENSITIVITY

As shown in Figure.8, we examined the sensitivity of SAM-Swap to the glass-limit threshold df_{glass}, which determines when SAM-Swap are triggered. Across the range $1.01 \leq df_{\text{glass}} \leq 1.5$, both test accuracy and FD exhibited only modest variations, indicating that the overall performance of SAM-Swap is relatively insensitive to the precise threshold value. This robustness suggests that the critical requirement is simply that the threshold remains above the ballistic limit ($D_f = 1$), allowing swaps to be activated once the optimization trajectory enters a sufficiently constrained regime.

Replica correlation (q) showed a consistent downward trend across all thresholds, although smaller values of df_{glass} produced noticeably weaker decorrelation. This behavior indicates that reducing the threshold decreases the extent to which SAM-Swap perturb the trajectory, while preserving the overall dynamical pattern observed for larger thresholds.

Among the measured observables, parameter mobility (dR) displayed the strongest sensitivity to the choice of df_{glass}. In particular, the lowest thresholds, $df_{\text{glass}} = 1.01$ and 1.05, produced unexpectedly large fluctuations in dR. This is somewhat counterintuitive because smaller thresholds activate swaps less frequently, given that interventions are only triggered when $df_{\text{layer}} < df_{\text{glass}}$. Rather than stabilizing the dynamics, infrequent interventions appear to amplify fluctuations in the recovery process following each swap event.

Based on these observations, we selected $df_{\text{glass}} = 1.1$ as the operational threshold used throughout the main text. This value was the smallest threshold that produced relatively stable PM trajectories while maintaining the acceleration effect of SAM-Swap. Although larger thresholds such as $df_{\text{glass}} = 1.5$ also yielded comparable performance, they increased the frequency of SAM-Swap and therefore the risk of excessively disrupting the optimization process. The chosen value provides a practical balance between stability and robustness, ensuring that perturbations occur frequently enough to escape arrested dynamics without becoming overly disruptive.

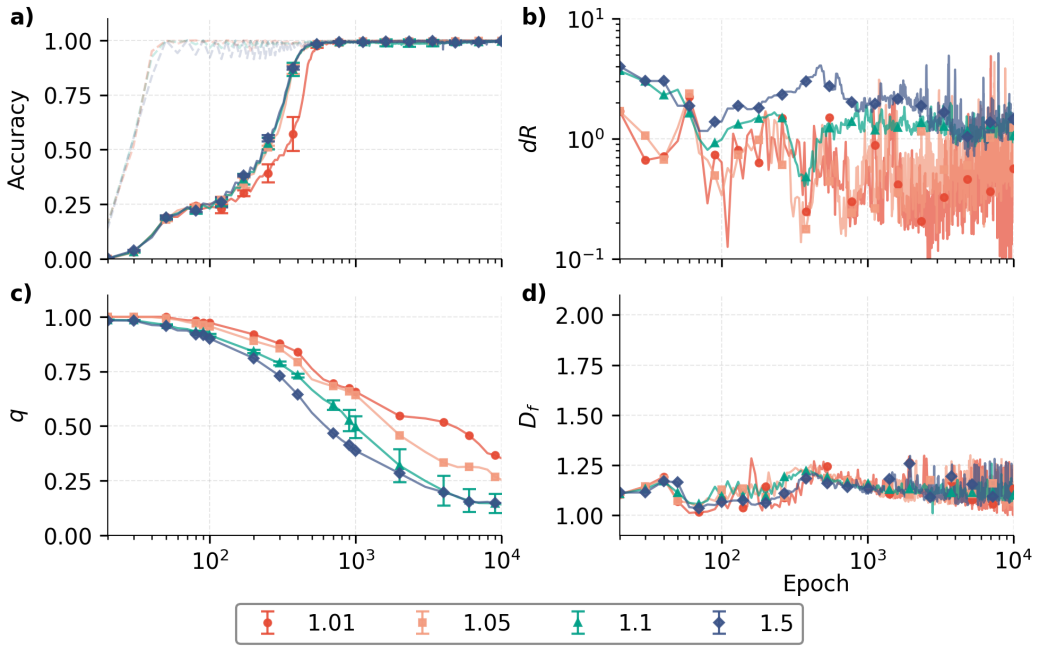

Figure 8: **Training dynamics under different glass limit** . Each curve corresponds to a glass limit of 1.01,1.05,1.1,1.5, as indicated in the legend. (a) Test accuracy,(b) parameter mobility, (c) Replica correlation (RC), (d) fractal dimension (FD). The legend is shared across all subplots.