CONTROLLING POLAR EXPOSURE TO DELAY MEMORIZATION IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models can reach useful sample quality before copying training examples, but fast optimization can compress this generalization window by accelerating sample-specific fitting. We investigate this effect through update geometry and propose Quality-Gated De-whitening (QGD), a controller that retains a fast polar-update prefix and progressively restores fixed-gain momentum. Our random-feature analysis separates covariance-controlled, curvature-equalized and amplitude-controlled memorization clocks. Under aligned spectral assumptions, it establishes a finite-exposure condition under which a fixed-gain tail recovers a delay proportional to dataset size. QGD implements this principle with a confirmed quality gate, a bounded decay envelope and causal copy feedback. Immediate switching is the conservative limit; gradual control balances delayed copying against continued quality improvement. We pair QGD with Copy-Budgeted Selection (CBS), which applies simultaneous binomial calibration to a frozen checkpoint family, followed by a fresh evaluation of the released checkpoint. On 2,000-image CIFAR-10 subsets, QGD preserves the polar baseline's quality-arrival time while expanding its useful interval by $8.32\times$ and reducing common-checkpoint copying by 75.9%. With identical calibration and independent quality evaluation, QGD achieves FID 75.56 versus 79.37 for SGD with the same selector. Exposure-matched controls, independent detector audits and transfer to flow matching and dance generation support adaptive exposure control as a practical way to improve the quality–copying tradeoff.

1 INTRODUCTION

Diffusion models learn useful generative structure from finite training collections (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020), yet continued optimization can lead them to reproduce training examples (Carlini et al., 2023; Somepalli et al., 2023a). Quality and copying emerge at different times, creating a generalization window in which a model produces useful samples before detectable memorization becomes prevalent (Bonnaire et al., 2025; Favero et al., 2026). This interval matters operationally: it provides room for further quality improvement and checkpoint selection. Reaching a quality target sooner is therefore one part of a successful training strategy; preserving useful choices after that point is another.

Prior work establishes that optimizer choice changes these clocks. Bonnaire et al. (2025) compare Adam and momentum SGD, observing faster learning and memorization with Adam while retaining dataset-size-dependent delay. Learning-rate effects also influence memorization in denoising score matching (Wu et al., 2025). These findings motivate a more specific question: can update geometry change the mechanism governing the delay, and can that mechanism guide a controller that preserves fast initial learning? Answering it requires separating directional changes from update magnitude and measuring quality and copying along the same trajectories.

Our random-feature analysis identifies three clocks. Gradient flow retains covariance-dependent rates that slow fitting in small-eigenvalue directions. Exact curvature preconditioning equalizes those rates. Aligned polar updates instead make fluctuation fitting depend on coefficient amplitude, so distributing fluctuation energy across more directions can shorten its clock. This rank-dependent compression gives a concrete control objective: limit polar exposure after useful population structure

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.

has been learned, then restore the small-gradient delay of a fixed-gain tail. Controlled spectra and directional interventions test this mechanism separately from neural architecture changes.

We propose Quality-Gated De-whitening (QGD) to implement this objective. QGD begins with Muon, an existing optimizer that orthogonalizes matrix momentum (Jordan et al., 2024; Liu et al., 2025). A predefined quality target, confirmed by a subsequent probe, determines when control starts. Thereafter, causal copy feedback reduces the polar coefficient inside a deterministic decay envelope while preserving momentum and a fixed gain. The envelope bounds total subsequent polar exposure; feedback adapts its allocation to the observed trajectory. Avoiding per-step renormalization allows small-gradient directions to remain slow.

Our recovery theorem gives a sufficient condition for a memorization delay proportional to dataset size: the fluctuation margin remaining at the gate must exceed subsequent polar exposure at the relevant scale. Immediate switching spends no additional polar exposure and provides the conservative endpoint of this analysis. QGD uses a gradual transition to retain quality improvement while delaying copying relative to continued polar training. We test this tradeoff against the same-gate hard switch and a pilot-fixed gradual schedule with matched exposure. Thus the theory motivates the controlled quantity, and the experiments test the practical value of its allocation.

Optimizer transitions toward SGD are established (Keskar & Socher, 2017; Luo et al., 2019); MiMuon also adaptively selects Muon or momentum-SGD updates using a spectral criterion and practical norm proxy (Huang et al., 2026). QGD connects a confirmed generative-quality target and causal copying probes to a finite-exposure condition. Appendix A develops these relationships. This formulation makes a conservative switching baseline central to evaluation and tests feedback under varying copying dynamics rather than attributing every gain to adaptation.

The resulting archive still requires a release decision. We pair QGD with Copy-Budgeted Selection (CBS), applying established exact binomial bounds and simultaneous calibration to a frozen checkpoint family (Clopper & Pearson, 1934; Angelopoulos et al., 2025). CBS chooses by development quality among eligible checkpoints, or abstains, while bounding the probability of release with a detector-budget violation. Fresh calibration never guides training, and an additional independent bank evaluates final quality. Figure 1 shows these distinct roles. Comparisons with valid fixed-sequence testing and preselection expose the quality–availability–cost tradeoff of searching more candidates.

Across matched training conditions, QGD preserves Muon's quality arrival, widens its useful interval, and improves independently evaluated quality under a common copy budget. Conservative baselines retain lower copying in some settings.

Our contributions can be summarized as follows:

- We develop an optimizer-geometric analysis of memorization in a random-feature denoising model, distinguishing covariance-controlled, curvature-equalized, and amplitude-controlled clocks. The analysis identifies rank-dependent compression of memorization delay under aligned polar updates and motivates controlling their cumulative exposure.
- We propose Quality-Gated De-whitening (QGD), which combines a confirmed quality gate and causal copying feedback to transition from fast polar updates to fixed-gain momentum. We establish a finite-exposure condition under which the aligned model recovers a memorization delay proportional to dataset size.
- We introduce Copy-Budgeted Selection (CBS), a checkpoint-selection procedure that combines simultaneous binomial calibration over a frozen candidate family with quality-based selection. CBS bounds the joint probability of release and detector-budget violation, while an independent final audit evaluates the selected checkpoint without reusing its selection measurements.
- We evaluate QGD through matched-exposure controls, component ablations, detector audits, and transfer to flow matching and dance generation. On CIFAR-10 with $2{,}000$ training images, QGD preserves Muon's quality-arrival time, extends the observed generalization window to $8.32\times$ that of Muon, and reduces common-checkpoint copying by 75.9%. Under identical CBS, QGD achieves final-audit FID 75.56, compared with 79.37 for SGD.

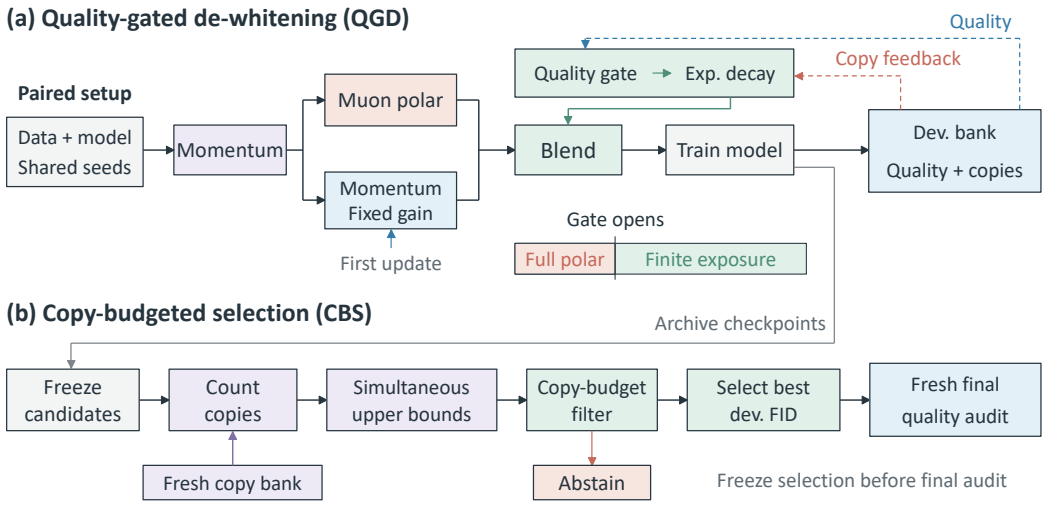

Figure 1: **Quality-gated training and calibrated checkpoint release.** QGD preserves a fast polar prefix, then uses development copying feedback to reduce polar exposure within a fixed envelope. CBS evaluates a frozen checkpoint family with fresh calibration inputs. A separate final audit evaluates the selected checkpoint without changing the decision.

2 Optimizer geometry determines the memorization clock

A random-feature denoiser $D_W(y) = W\phi(y)$ isolates update geometry with fixed features (Rahimi & Recht, 2007). For Gaussian data with additive Gaussian noise, least-squares readout learning has covariance K_n and empirical optimum $W_{\mathrm{emp}} = W_{\mathrm{pop}} + \Delta$, decomposed into population optimum and empirical fluctuation. We study a spectral-separation regime with population energy at order-one covariance eigenvalues and fluctuation energy at smaller ones. Appendix B specifies the model and clocks.

Population Structure and Empirical Detail. For feature covariance H, write $\|A\|_H^2 = \mathrm{tr}(AHA^\top)$ and $\langle A, B\rangle_H = \mathrm{tr}(AHB^\top)$. The empirical objective $L_n(W) = \|W - W_{\mathrm{emp}}\|_{K_n}^2 + \mathrm{const}$ makes covariance determine gradient fitting rates. With population covariance K, normalized readouts separate population learning from progress toward the empirical fluctuation:

$$\mathrm{gen}(t) = \frac{\|W_t - W_{\mathrm{pop}}\|_K^2}{\|W_{\mathrm{pop}}\|_K^2}, \qquad \mathrm{mem}(t) = \frac{\langle W_t - W_{\mathrm{pop}}, \Delta\rangle_{K_n}}{\|\Delta\|_{K_n}^2}. \tag{1}$$

With positive denominators, these readouts measure population prediction error and progress toward empirical fluctuation. In aligned disjoint channels, memorization is the covariance-weighted fitted fraction $m(t)$, equal to one at the empirical optimum. Gradient-flow bounds also control population/fluctuation cross terms outside this alignment.

For a toy trajectory, let $g_\star = \inf_{t\geq 0} \mathrm{gen}(t)$. A cell qualifies if $g_\star \leq \mathrm{gen}(0)/2$. Then τ_{gen} first reaches $1.1g_\star$, with the positive tolerance defined in Appendix B and τ_{mem} reaches half the fluctuation projection; an undefined quality clock leaves the window ratio undefined. These clocks order structure and fluctuation learning; the neural protocol uses a shared sample-quality target (Section 3).

Three Clocks. Normalized gradient flow learns covariance mode k as $(1 - e^{-\lambda_k t})W_{\mathrm{emp}}u_k$. Its fluctuation-energy quantile bounds give $\tau_{\mathrm{mem}} = \Theta(n)$ when the relevant eigenvalues scale as n^{-1} and the normalized population/fluctuation cross energy remains bounded away from $1/4$ (Proposition 3). Exact curvature preconditioning instead yields $W_t = (1 - e^{-t})W_{\mathrm{emp}}$ (Proposition 4). In the disjoint-channel case, its half-fluctuation time is $\log 2$; its straight trajectory need not reach the quality gate.

Polar flow supplies a different route. With aligned covariance and target singular directions, it learns coefficient a_j as $\min\{\eta t, a_j\}$. If the fluctuation singular values lie in $[s_{\min}, s_{\max}]$, then:

$$\frac{s_{\min}}{2\eta} \le \tau_{\mathrm{mem}} \le \frac{s_{\max}}{2\eta}, \qquad \tau_{\mathrm{mem}} = \Theta\left(\frac{\sqrt{n/r}}{\eta}\right) \quad \text{if } \|\Delta\|_F^2 = \Theta(n),\ s_{\max}/s_{\min} = O(1), \quad (2)$$

where r is fluctuation rank (Proposition 5). The bounds depend on coefficient distance and rank rather than the positive covariance eigenvalues. Practical Muon includes momentum, approximate orthogonalization and learned features; Figure 4 summarizes the geometric distinction and random-feature evidence.

Random-Feature Evidence. Across six dataset sizes and two noise levels, gradient-flow memorization slopes are $1.00/1.05$, versus Muon's $0.20/0.18$ at $d = 64$; the paired window-slope gaps are $-0.75/-0.82$, with $21.14\times/29.72\times$ narrowing at $n = 512$ (Table 7). At $d = 16$, Muon's slopes rise to $0.60/0.65$ while gradient flow remains near one. The ordering survives $\tanh$ and noise-draw changes (Tables 9–10); exact preconditioning misses quality for $n \le 64$ (Table 11). Doubling n approximately halves the fluctuation eigenvalue while preserving the population eigenvalue (Table 8). Across draws, sizes and noise levels, inverse fluctuation eigenvalues correlate with memorization time at 0.99 for gradient flow versus 0.14 for Muon (Spearman; Table 12).

From a Timing Diagnosis to a Control Objective. The mechanism motivates a polar prefix followed by restoration of small-eigenvalue delay. Spreading fluctuation energy over more aligned directions shortens each coefficient's polar travel distance; the readout-dimension comparison and the joint spectral family in Appendix E test this dependence.

3 Exposure-controlled training and calibrated release

3.1 Separate development, calibration and evaluation

Paired runs share the training subset, initialization, sample order and training noise. Matrix methods use the same non-matrix AdamW updates and the same development-probe schedule. The primary model is a 6.47M-parameter UNet trained on class-balanced CIFAR-10 subsets of $2{,}000/10{,}000$ images for 200k/400k steps. Five paired trajectories are evaluated at each size. The image sampler is deterministic DDIM (Song et al., 2021); Appendix H specifies the model, sampler and optimizer settings.

Development probes run every $h = 2{,}000$ steps. FID (Heusel et al., 2017; Parmar et al., 2022) uses $5{,}000$ outputs; a $1{,}000$-output subset guides copying feedback. Pixel detection flags an output when its nearest and second-nearest training-image distances satisfy $d_1 < d_2/3$ (Yoon et al., 2023; Gu et al., 2025). Independent pilots fix the quality targets, $88.0/64.9$, before evaluation. The first quality crossing defines τ_{gen} after next-checkpoint confirmation; the controller changes only subsequent updates. A confirmed copy crossing before or tied with quality confirmation makes the run inadmissible.

The first development-copy crossing of 5%, confirmed at the next probe, defines τ_{mem}. For horizon H, the observed useful interval is $\Delta t_H = \max(0, \min(\tau_{\mathrm{mem}}, H) - \tau_{\mathrm{gen}})$; an unreached copy crossing remains censored. Independent banks evaluate common-checkpoint quality and copying without affecting these clocks or control. After training, CBS calibrates a frozen checkpoint family with fresh inputs. A further independent bank evaluates each selected checkpoint's final quality; it never changes the selected identity. These roles separate an actionable controller, a valid release decision and an independent final assessment. Appendix I defines the banks and fixed detector rules.

3.2 Our training method: quality-gated de-whitening

QGD keeps the full Muon update until the predeclared quality target is confirmed at t_s. For matrix layer ℓ, let $M_{\ell,k}$ be its momentum direction and $U_{\ell,k}^{\mathrm{Muon}}$ its practical Muon direction. Calibrate $c_\ell = \mathrm{RMS}(U_{\ell,0}^{\mathrm{Muon}})/\mathrm{RMS}(M_{\ell,0})$ at the first update and hold it fixed; use the positive fallback $c_\ell = 1$ if either RMS is at most ϵ. The mixed direction is:

$$U_{\ell,k} = \alpha_k U_{\ell,k}^{\mathrm{Muon}} + (1 - \alpha_k) c_\ell M_{\ell,k}. \quad (3)$$

The mixture has no subsequent RMS normalization, which could divide out the small-gradient delay restored by the fixed-gain tail. Before confirmation, $\alpha = 1$; a confirmed development-copy budget crossing before or tied with quality confirmation makes the run inadmissible.

At $t_j = t_s + jh$, use the available development copy fraction $\widehat{C}_j$ to set the coefficient for the *following* block, starting from $\alpha_0 = 1$:

$$\alpha_{j+1} = \min\left\{ e^{-(j+1)h/T},\ \left[\alpha_j - \kappa \left[\frac{\widehat{C}_j - c_{\mathrm{warn}}}{c_{\mathrm{budget}} - c_{\mathrm{warn}}}\right]_+\right]_+ \right\}, \qquad [x]_+ = \max(x, 0). \tag{4}$$

The default configuration uses $T = 8{,}000$, $\kappa = 0.5$, $c_{\mathrm{warn}} = 2.5\%$ and $c_{\mathrm{budget}} = 5\%$; sensitivity is evaluated separately. The coefficient decreases causally, and its exponential envelope bounds total subsequent polar exposure:

$$A_\infty = \int_{t_s}^{\infty} \alpha(t)\, dt = h \sum_{j \geq 0} \alpha_{j+1} \leq \frac{h}{e^{h/T} - 1} = 7{,}041.6 \text{ steps}. \tag{5}$$

Copy feedback can shorten this exposure without compromising its bound.

Proposition 1 (Finite polar exposure restores a spectral delay)**.** *In the aligned disjoint-channel model of Proposition 5, let each fluctuation channel have target amplitude s and covariance eigenvalue μ. Run polar flow at speed η until t_s, then mix it with fixed-gain gradient flow of gain $\gamma > 0$ as in Equation* (14)*, with measurable $\alpha(t) \in [0, 1]$ and $A_\infty < \infty$. Set $m_s = \eta t_s / s$ and $\nu = \eta / s$. If $m_s < \theta < 1$, $\nu \geq \gamma\mu$, and $\varepsilon_\theta = \theta - m_s - \nu A_\infty > 0$, the time to fluctuation level θ satisfies:*

$$t_s + \frac{\varepsilon_\theta}{\gamma\mu} \leq \tau_\theta \leq t_s + \frac{1}{\gamma\mu} \log \frac{1 - m_s}{1 - \theta}. \tag{6}$$

For fixed $\theta \in (0, 1)$, $\mu = \Theta(n^{-1})$, $\gamma = \Theta(1)$, $t_s = O(1)$ and uniformly positive ε_θ, this gives $\tau_\theta = \Theta(n)$.

Bounding $\dot{m} = \alpha\nu + (1 - \alpha)\gamma\mu(1 - m)$ proves the result (Appendix D); Proposition 6 extends it to unequal amplitudes with $c/n \leq \mu_j \leq C/n$. A positive polar floor retains amplitude-controlled arrival, whereas finite exposure with the stated margin gives a sufficient pathwise guarantee. The exact result preserves the common prefix and population-channel progress; full neural quality is evaluated separately.

Delay Restoration and Quality Utility. The upper bound in Equation (6) is exactly the same-gate hard-switch clock: under the aligned assumptions, retaining polar updates reaches the fluctuation threshold no later than hard switching (Appendix E). The theorem therefore establishes delay restoration; quality determines whether a gradual transition is useful. For a desired residual margin $\varepsilon_0 > 0$, define $B_\star = (\theta - m_s - \varepsilon_0)/\nu > 0$. Choosing:

$$T \leq \frac{h}{\log(1 + h/B_\star)} \tag{7}$$

makes the envelope sufficient for $\varepsilon_\theta \geq \varepsilon_0$. A constructed aligned diagnostic holds gain constant and measures every term: positive-margin exposure retains a linear delay, whereas excess exposure and a fixed polar floor compress it (Figure 3). Gradual exposure improves early population risk, while the hard switch achieves the latest threshold and lower eventual risk. In neural training, the exposure cap is observable; the theorem's fluctuation margin is not. Independent evaluation tests practical quality and copying.

3.3 Our selection rule: copy-budgeted selection

Development probes guide training and can be reused adaptively. CBS uses a separate calibration bank after freezing $1 \leq M \leq 100$ candidate checkpoints, the detector, and the sampler. For each candidate k, count S_k flagged outputs among $N = 5{,}000$ new iid inputs, including latent noise and any conditioning drawn from the fixed generation distribution, independently of the development history. Let p_k denote its underlying detector-flag probability. With $0 < \delta < 1$ and $a = \delta/M$, the one-sided Clopper–Pearson upper limit (Clopper & Pearson, 1934) is:

$$U_k = \begin{cases} \mathrm{Beta}^{-1}(1 - a; S_k + 1, N - S_k), & S_k < N, \\ 1, & S_k = N. \end{cases} \tag{8}$$

Table 1: **Training-method comparison on CIFAR-10.** Times are thousands of updates. Mean ± sample SD over five paired trajectories. FID and copy fractions use independent audits at the stated checkpoint; clocks use confirmed development probes. NR denotes no confirmed copy crossing by the observation horizon $H = 200$k/400k. Observed windows end at $\min(t_{\text{copy}}, H)$.

Method	Quality arrival ↓	Copy onset ↑	Window Δt_H ↑	FID ↓	Copy (%) ↓
CIFAR-10, $n = 2,000$; audit checkpoint: 60k					
SGD	16.0±1.4	NR	184.0±1.4	85.4±0.3	1.26±0.11
AdamW	14.0±1.4	124.0±6.2	110.0±5.3	82.9±0.3	2.51±0.10
Muon	8.0±1.4	28.0±1.4	20.0±1.4	81.3±0.3	16.10±0.47
QGD (ours)	8.0±1.4	174.4±10.1	166.4±11.3	77.8±0.3	3.88±0.31
CIFAR-10, $n = 10,000$; audit checkpoint: 120k					
SGD	18.0±1.4	NR	382.0±1.4	63.3±0.2	0.78±0.14
AdamW	14.0±1.4	NR	386.0±1.4	61.0±0.2	1.32±0.11
Muon	12.0±1.4	96.0±4.5	84.0±4.0	61.5±0.2	12.37±0.44
QGD (ours)	12.0±1.4	360.0±9.1	348.0±8.5	56.1±0.2	2.99±0.21

Among checkpoints meeting the fixed development-quality target and $U_k \leq c_{\text{budget}}$, choose the lowest development FID, breaking ties by earlier step; abstain if the set is empty. Exact binomial inversion and simultaneous testing supply the established statistical ingredients (Angelopoulos et al., 2025).

Proposition 2 (Detected-copy control for a released checkpoint)**.** *Conditional on the training and development history that fixes the candidates and detector, CBS satisfies:*

$$\Pr\{\widehat{k} \neq \varnothing \text{ and } p_{\widehat{k}} > c_{\text{budget}}\} \leq \delta. \tag{9}$$

A union bound on $\Pr\{p_k > U_k\} \leq \delta/M$ proves the result (Appendix D). Fresh inputs may be shared across candidates; their independence is unnecessary. The copy budget and risk target govern different quantities: a 5% budget concerns the selected checkpoint's detector-flag probability, while $\delta = 0.05$ bounds calibration's probability of releasing a checkpoint above that budget. The guarantee concerns joint release and budget violation for the declared detector and generation distribution, excluding undetected copying and population FID. The frozen family includes every configuration searched using calibration. After a checkpoint is selected, a further fresh bank evaluates quality and copying without revising the decision.

4 Experiments and Results

We compare matched training trajectories, controlled exposure schedules and checkpoint release. Five paired trajectories are the unit of uncertainty; calibration repetitions remain nested within each trajectory. Tables identify QGD and CBS as (ours).

4.1 Quality and copying along matched trajectories

QGD preserves Muon's quality-arrival time because their updates agree through confirmation. Its post-gate behavior produces the useful difference: at $n = 2{,}000$, the mean interval expands by $8.32\times$, while independent common-step FID improves from 81.3 to 77.8 and copying falls by 75.9% (Table 1). The larger subset preserves the same ordering against continued Muon. SGD and AdamW retain longer low-copy periods in some settings; QGD reaches useful quality earlier than SGD and achieves stronger later quality. This comparison concerns a common training time, while Table 4 evaluates the complete archive under a shared selector. Figure 2 shows why a window needs both measurements: the hard switch suppresses copying more strongly, whereas QGD supports further quality improvement while remaining within the declared region. Development quality remains eligible from confirmation through the reported copy onset. Appendix J provides paired clocks and per-trajectory effects, retaining censoring instead of treating observation limits as exact crossings.

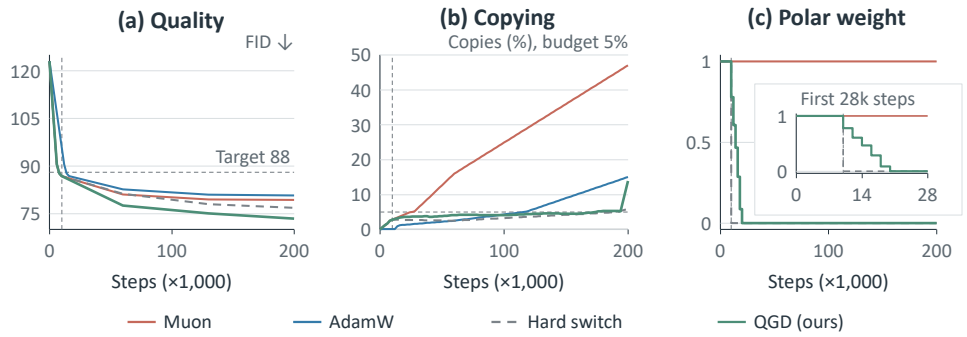

Figure 2: **Quality, copying and exposure during training.** The primary CIFAR-10 trajectory (pair 0) compares QGD with continued Muon, AdamW and the same-gate hard switch. Development curves share a 0–200k time axis, with FID target 88 and copy budget 5%. The 0–28k inset resolves the transition after quality confirmation at 10k. The hard switch is more conservative in copying; gradual QGD retains stronger quality. The coefficient for each block is fixed using the preceding development probe.

4.2 Exposure recovery and controller components

The controlled spectral experiment verifies the condition used by Proposition 1. The finite-mixture schedule has residual margin 0.165 across all tested sizes with fixed gain. Its copy time rises from 20.85 to 148.49 as n grows from 64 to 512, while continued polar flow reaches the threshold at time 4. Hard switching delays it further, as the comparison principle predicts. The gradual mixture fits the remaining population component earlier; the hard switch can achieve lower eventual error with more time. Figure 3 and Table 6 report both effects, and Appendix E gives the exact construction and margin calculation. Table 2 tests the practical allocation of exposure. Hard switching shares the confirmed gate, momentum and first-update gain, then sets the polar coefficient to zero. It provides a longer copy delay and lower copying than QGD, while QGD improves common-step quality. The matched open-loop schedule uses the same gate and a pilot-fixed exposure budget, without access to evaluation copying feedback. Its mean exposure is comparable to QGD's; differences therefore test when exposure is spent rather than simply comparing a long schedule with a short one. The paired trajectories vary copying dynamics, allowing feedback to adapt its decay and shutoff time. Removing the quality gate sacrifices the shared fast prefix. Removing feedback while retaining the default envelope permits additional polar exposure, and per-step normalization removes the small-gradient scale that the fixed-gain tail is designed to preserve. Per-layer norm matching provides a separate directional intervention (Tables 13 and 14). Controller settings, probe cadence and a matched MiMuon pilot search are specified in Appendix J.

4.3 Valid selection under a copying budget

Table 3 compares selectors on the same QGD archives. Empirical filtering searches observed copying without a simultaneous correction. Single-candidate and margin preselection freeze a candidate using development information, then certify it once. Fixed-sequence testing fixes its ordering before calibration and stops at the first failed test; CBS applies simultaneous bounds to the full family (Angelopoulos et al., 2025). All valid rules use the same joint error target. Appendix K specifies their ordering, testing and abstention rules. The comparison separates quality, release availability and calibration cost. At the tighter budget, searching a family can retain useful candidates that single preselection misses; a valid ordered search is a competitive alternative to uniform Bonferroni allocation. At the looser budget, margin preselection can approach the quality of a full search with far fewer outputs. These outcomes make the selector choice depend on the available archive and budget. Adaptive allocation offers another route to reducing testing cost (Zecchin et al., 2025); here, fixed sample counts make the statistical comparison directly reproducible.

Table 2: **Switching controls and component ablations.** CIFAR-10, $n = 2{,}000$; mean $\pm$ SD over five paired trajectories. Exposure and window are in k steps; quality and copying use independent audits at 60k. The open-loop schedule is fixed by an independent pilot. The no-gate exposure starts at initialization; MiMuon has no common post-gate mixture exposure.

Method	Polar exposure	Observed window $\Delta t_H \uparrow$	$\text{FID}_{60} \downarrow$	Copy_{60} (%) $\downarrow$
MiMuon (norm proxy)	–	84.4 ± 5.0	80.6 ± 0.3	4.30 ± 0.30
Same-gate hard switch	0.00 ± 0.00	186.0 ± 2.4	81.5 ± 0.3	2.47 ± 0.03
Pilot-fixed open loop	4.26 ± 0.00	124.0 ± 50.0	79.0 ± 1.0	4.56 ± 0.60
QGD w/o quality gate	7.03 ± 0.00	160.8 ± 10.4	82.9 ± 0.3	2.95 ± 0.20
QGD w/o feedback	7.04 ± 0.00	70.0 ± 7.1	79.2 ± 0.3	4.77 ± 0.21
QGD + per-step RMS	4.51 ± 0.51	30.0 ± 2.4	80.8 ± 0.3	13.51 ± 0.16
QGD (ours)	4.21 ± 0.64	166.4 ± 11.3	77.8 ± 0.3	3.88 ± 0.31

Table 3: **Checkpoint-selection rules under copy budgets.** Five QGD archives, 100 independent calibration banks each; $\delta = 0.05$. Final FID and fresh audit copying condition on release; FID is summarized over defined trajectory means. At 5%, preselection never releases on one trajectory. Joint errors use the retained underlying detector rates. Calibration outputs exclude the common development archive and final audit.

Selection rule	Final FID $\downarrow$	Copy (%) $\downarrow$	Release $\uparrow$	Joint error $\downarrow$	Calib. outputs (k) $\downarrow$
Copy budget: 5%					
Empirical filter	73.97 ± 0.30	5.15	500/500	459/500	500.0
Preselect + CP	74.09 ± 0.43	5.11	20/500	3/500	5.0
Margin preselect + CP	77.89 ± 2.28	4.15	463/500	0/500	5.0
Fixed sequence	77.86 ± 0.64	4.08	500/500	0/500	152.6
CBS (ours)	75.56 ± 0.51	4.37	500/500	1/500	500.0
Copy budget: 10%					
Empirical filter	73.85 ± 0.30	8.45	500/500	6/500	500.0
Preselect + CP	73.85 ± 0.30	8.38	421/500	0/500	5.0
Margin preselect + CP	73.88 ± 0.32	5.70	500/500	0/500	5.0
Fixed sequence	73.86 ± 0.31	7.83	500/500	0/500	480.2
CBS (ours)	73.86 ± 0.31	7.46	500/500	0/500	500.0

4.4 Independent quality of released checkpoints

Holding CBS fixed compares the training methods under one release policy (Table 4). QGD obtains final-audit FID 75.56, compared with 79.37 for conservative SGD, which can exploit its full low-copy archive. The independent audit preserves the quality advantage rather than reusing the development score that selected the checkpoint. The hard-switch pipeline also meets the budget but has higher final FID; QGD improves the quality available for release. The selected step identifies the released weights; it does not represent the total computation needed to produce and calibrate the archive. FID uncertainty is computed across five trajectory means after averaging their calibration repetitions. Release and violation frequencies count all decisions, including abstentions. Additional comparisons vary the candidate-family size, and Appendix L reports archive construction, generation probes and calibration separately. This accounting makes the expense of adaptive control and full-family search visible alongside their quality gains.

4.5 Independent audits and transfer

Independent detector audits evaluate the pixel-controlled QGD trajectory using fresh inputs and a fixed SSCD detector (Pizzi et al., 2022). Table 15 reports both fixed detectors; Table 18 varies the controller and probe settings. These tests distinguish the behavior of an adaptive controller from the AdamW–Muon ordering established by the background geometry experiments. Table 5 evaluates flow matching (Lipman et al., 2023) and music-conditioned AIST++ motion (Li et al., 2021). Image

Table 4: **Calibrated selection across training strategies.** CIFAR-10, $n = 2{,}000$; 5% copy budget and $\delta = 0.05$. All methods use a 200k archive and identical CBS allocation. Final FID uses a fresh 50k-output bank; copying uses a 10k subset. Mean $\pm$ SD is across five trajectories after averaging 100 calibration decisions each. All 500 decisions per row release.

Training + selection	Selected step (k)	Develop. FID ↓	Final FID ↓	Copy (%) ↓
SGD + CBS	200.0	79.00	79.37±0.31	2.67
AdamW + CBS	103.8	81.58	81.94±0.29	4.24
Muon + CBS	18.5	85.81	86.14±0.12	3.99
MiMuon (norm proxy) + CBS	64.2	80.40	80.68±0.40	4.25
Same-gate hard switch + CBS	160.0	77.46	77.82±0.27	4.30
Pilot-fixed open loop + CBS	77.8	79.00	79.30±3.08	4.37
QGD + CBS (ours)	128.6	75.18	75.56±0.51	4.37

Table 5: **Transfer across generation tasks.** Mean $\pm$ SD over five paired trajectories. Flow times are k updates; quality is audited at the 60k checkpoint. Dance quality at 120k separates beat synchrony, kinetic-feature FID and diversity relative to held-out motions; one is the diversity target. Copying uses 10k fresh audit outputs.

Flow matching, $n = 2{,}000$				
Method	FID ↓	Quality arrival ↓	Window Δt_H ↑	Copy (%) ↓
AdamW	84.5±0.3	10.0±1.4	106.0±2.8	2.08±0.22
Muon	86.2±0.3	8.0±1.4	18.0±2.8	14.61±0.07
Same-gate hard switch	82.8±0.3	8.0±1.4	120.0±2.8	2.48±0.15
QGD (ours)	79.8±0.3	8.0±1.4	100.0±2.8	4.14±0.25
AIST++ dance generation				
Method	Beat ↑	Kinetic FID ↓	Diversity target: 1	Copy (%) ↓
AdamW	0.282±0.001	31.7±0.4	0.963±0.003	1.94±0.13
Muon	0.285±0.001	34.6±0.4	0.875±0.003	11.06±0.57
Same-gate hard switch	0.289±0.001	30.2±0.4	0.956±0.003	2.14±0.22
QGD (ours)	0.294±0.001	27.9±0.4	0.947±0.003	4.09±0.29

quality uses final-audit FID. Motion evaluation separates beat synchrony, kinetic-feature distribution quality and diversity, so improved alignment cannot conceal repetitive motion. The common quality target and detector are fixed within each task, with data splits formed before chunking. QGD improves the joint quality–copying tradeoff against continued polar training; conservative alternatives retain lower copying in some settings. Appendix M defines the task-specific metrics and copying rules.

5 Discussion and conclusion

Update geometry governs how quickly useful learning gives way to memorizing sample-specific detail. Finite polar exposure suffices to recover a spectral delay, and QGD turns this into quality-gated control: the hard switch suppresses copying most strongly, gradual QGD preserves quality, and matched-exposure controls isolate the value of feedback. With CBS fixed, an independent final bank measures the quality available for release across optimizers, and calibration shows when a larger search family justifies its cost.

Limitations. The delay guarantee assumes aligned, fixed features; extending it to momentum, approximate orthogonalization and evolving representations is a natural next step. CBS coverage is defined by the specified detector (Appendix K), and probe cadence and family size set the generation cost of calibration (Table 3).

AI USE STATEMENT

In this work, we used generative AI tools for none of the tasks with required disclosure. We have not used generative AI tools for developing theoretical models or conceptual frameworks, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, or interpreting results, and generating synthetic datasets, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, translation, cleaning and reformatting datasets, and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for editing the paper to improve readability, limited to polishing the grammar, wording, and clarity of text written by the authors. We have reviewed all AI-assisted work. All suggested edits were manually checked by the authors to ensure that they preserved the intended technical meaning and did not introduce new claims, results, or citations. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The evaluation design concerns public image and motion datasets and aggregate copying statistics. No private data or model-extracted personal information are distributed.

REPRODUCIBILITY STATEMENT

The companion package contains the complete unified LaTeX manuscript, numerical records, causal schedules, calibration count matrices, exact-flow solver, table generators and editable figures. Appendix H specifies the neural evaluation design; Appendices C–E give the proofs and computed diagnostic, and Appendices I–M define detector, selection and transfer protocols.

REFERENCES

Beatrice Achilli, Luca Ambrogioni, Carlo Lucibello, Marc Mézard, and Enrico Ventura. Memorization and generalization in generative diffusion under the manifold hypothesis. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(7):073401, 2025. doi: 10.1088/1742-5468/ade136.

Madhu S. Advani, Andrew M. Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020. doi: 10.1016/j.neunet.2020.08.022.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998.

Marta Aparicio Rodriguez, Anastasia Borovykh, Grigorios A. Pavliotis, and Daniel J. Korchinski. Diffusion models preferentially memorize prototypical examples or: Why does my diffusion model love slop? *arXiv preprint arXiv:2605.30642*, 2026.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 233–242. PMLR, 2017.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. In *Conference on Neural Information Processing Systems (NeurIPS), Workshop on Optimization for Machine Learning (OPT)*, 2024.

Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mézard. Why diffusion models don't memorize: The role of implicit dynamical regularization in training. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 141266–141286. Curran Associates, Inc., 2025. doi: 10.52202/085713-4720.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *Proceedings of the 32nd USENIX Security Symposium (USENIX Security)*, pp. 5253–5270. USENIX Association, 2023.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 8780–8794. Curran Associates, Inc., 2021.

Raman Dutt, Ondrej Bohdal, Pedro Sanchez, Sotirios A. Tsaftaris, and Timothy Hospedales. MemControl: Mitigating memorization in diffusion models via automated parameter selection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4491–4501, 2025. doi: 10.1109/WACV61041.2025.00441.

Alessandro Favero, Antonio Sclocchi, and Matthieu Wyart. Bigger isn't always memorizing: Early stopping overparameterized diffusion models. *Transactions on Machine Learning Research*, 2026.

Xiangming Gu, Chao Du, Tianyu Pang, Chongxuan Li, Min Lin, and Ye Wang. On memorization in diffusion models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1842–1850. PMLR, 2018.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NIPS)*, volume 30, pp. 6626–6637. Curran Associates, Inc., 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020.

Feihu Huang, Yuning Luo, and Songcan Chen. MiMuon: Mixed Muon optimizer with improved generalization for large models. *arXiv preprint arXiv:2605.19619*, 2026.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. Online technical article, 2024.

Zahra Kadkhodaie, Florentin Guth, Eero P. Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *International Conference on Learning Representations (ICLR)*, pp. 46543–46567, 2024.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 26565–26577. Curran Associates, Inc., 2022. doi: 10.52202/068431-1926.

Nitish Shirish Keskar and Richard Socher. Improving generalization performance by switching from Adam to SGD. *arXiv preprint arXiv:1712.07628*, 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. AI choreographer: Music conditioned 3D dance generation with AIST++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13401–13412, 2021.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025.

Yuxing Liu, Jianyu Wang, and Tong Zhang. Optimizer-model consistency: Full finetuning with the same optimizer as pretraining forgets less. *arXiv preprint arXiv:2605.06654*, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Liangchen Luo, Yuanhao Xiong, Yan Liu, and Xu Sun. Adaptive gradient methods with dynamic bound of learning rate. In *International Conference on Learning Representations (ICLR)*, 2019.

James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2408–2417. PMLR, 2015.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171. PMLR, 2021.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11410–11420, 2022.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14532–14542, 2022.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, volume 20, pp. 1177–1184, 2007.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15(11):335–366, 2014.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265. PMLR, 2015.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6048–6058, 2023a.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 47783–47803. Curran Associates, Inc., 2023b. doi: 10.52202/075280-2071.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 11918–11930. Curran Associates, Inc., 2019.

Amund Tveit, Bjørn Remseth, and Arve Skogvold. Muon optimizer accelerates grokking. *arXiv preprint arXiv:2504.16041*, 2025.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. SOAP: Improving and stabilizing Shampoo using Adam for language modeling. In *International Conference on Learning Representations (ICLR)*, pp. 93423–93444, 2025.

Binxu Wang, Jacob A. Zavatone-Veth, and Cengiz Pehlevan. A random matrix perspective on the consistency of diffusion models. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, explaining, and mitigating memorization in diffusion models. In *International Conference on Learning Representations (ICLR)*, pp. 35647–35662, 2024.

Yu-Han Wu, Pierre Marion, Gérard Biau, and Claire Boyer. Taking a big step: Large learning rates in denoising score matching prevent memorization. In *Proceedings of the 38th Conference on Learning Theory (COLT)*, volume 291 of *Proceedings of Machine Learning Research*, pp. 5718–5756. PMLR, 2025.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26:289–315, 2007. doi: 10.1007/s00365-006-0663-2.

TaeHo Yoon, Joo Young Choi, Sehyun Kwon, and Ernest K. Ryu. Diffusion probabilistic models generalize when they fail to memorize. In *International Conference on Machine Learning (ICML), Workshop on Structured Probabilistic Inference and Generative Modeling (SPIGM)*, 2023.

Matteo Zecchin, Sangwoo Park, and Osvaldo Simeone. Adaptive learn-then-test: Statistically valid and efficient hyperparameter selection. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 74018–74036. PMLR, 2025.

APPENDIX

A RELATED WORK

Diffusion Design and Evaluation. Diffusion and score-based formulations learn generative processes through denoising (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020). Training objectives and noise schedules (Nichol & Dhariwal, 2021), architecture and guidance choices (Dhariwal & Nichol, 2021), and sampler and preconditioning design (Karras et al., 2022) all influence sample quality and computational cost. The paired protocol holds these choices fixed within each optimizer pair, so a change in the quality-to-copy interval can be attributed to the optimization comparison. Latent-space diffusion (Rombach et al., 2022) and transformer denoisers (Peebles & Xie, 2023) expand the architectural design space; the primary experiments use a common pixel-space UNet to isolate update geometry.

Memorization in Diffusion Models. That diffusion models can reproduce training images was established by extraction attacks (Carlini et al., 2023) and replication studies (Somepalli et al., 2023a), and subsequent work located the phenomenon in the regime of small datasets, duplicated images and large models (Yoon et al., 2023; Gu et al., 2025; Kadkhodaie et al., 2024). The measure we use for copies, a sample whose nearest training image is closer than one third of the second nearest, is the criterion of Yoon et al. (2023) as used in the dynamical study below. Subsequent work analyzes factors driving copying and develops mitigation strategies (Somepalli et al., 2023b; Wen et al., 2024), including automated parameter selection (Dutt et al., 2025). Our question concerns an additional axis: how optimizer choice changes the time at which copying becomes detectable.

The Generalization Window. Bonnaire et al. (2025) identified two timescales in training, an early τ_{gen} at which samples become good and a later τ_{mem} beyond which memorization emerges, and showed that their ratio widens with the training set size under gradient-based training, attributing the window to implicit dynamical regularization. The linear-denoiser analysis of Wang et al. (2026) shows that much of the behaviour of diffusion models across data splits has a linear origin in the empirical Gaussian statistics of the data; we work in the same tractable class and ask a question they do not, namely how the window depends on the optimizer. Related analyses of the transition from generalization to memorization in diffusion include Achilli et al. (2025) and Aparicio Rodriguez et al. (2026).

Whitened and Spectral Optimizers. Muon (Jordan et al., 2024) replaces the momentum of each matrix parameter by its orthogonalization, computed with a few Newton-Schulz iterations, and is steepest descent under the spectral norm (Bernstein & Newhouse, 2024); it has been scaled to large language model pretraining (Liu et al., 2025). Structured curvature methods include K-FAC, which approximates layerwise Fisher blocks with Kronecker products (Martens & Grosse, 2015). Shampoo (Gupta et al., 2018) and SOAP (Vyas et al., 2025) approximate full-matrix preconditioning through Kronecker factors and motivate tests of more expressive preconditioners alongside the idealized flow of Proposition 4. Neither is identical to that flow. Adam (Kingma & Ba, 2015; Loshchilov & Hutter, 2019) preconditions diagonally. Two recent observations foreshadow ours without explaining it: fine-tuning studies report that Muon memorizes small datasets faster than AdamW (Liu et al., 2026), and Muon accelerates the transition from memorization to generalization in grokking tasks (Tveit et al., 2025). Generalization bounds for Muon depend on its iteration count (Huang et al., 2026). These observations motivate a direct study of the quality-to-copy interval in diffusion training.

Optimizer Transitions and Copy-budget Calibration. SWATS switches from Adam to SGD using an optimization-based criterion (Keskar & Socher, 2017); AdaBound gradually bounds adaptive learning rates toward SGD-like behavior (Luo et al., 2019). These works establish optimizer transitions as a generalization strategy. MiMuon adaptively chooses a Muon or momentum-SGD step using a singular-value-gap criterion; its practical rule uses a momentum-norm proxy (Huang et al., 2026). This differs from static interpolation and from QGD's quality-triggered continuous mixture. QGD targets diffusion's quality-to-copy interval: a confirmed quality gate preserves a fast prefix, development copy feedback reduces polar exposure, and a fixed-gain tail retains eigenvalue-dependent

motion. The finite-exposure theorem and its rank-sensitive comparison provide the mechanism for this design. Diffusion-specific early stopping also depends on dataset size and model capacity (Favero et al., 2026); the optimizer axis determines how those stopping intervals change under whitening.

Large-learning-rate effects in denoising score matching further motivate separating magnitude from directional geometry (Wu et al., 2025).

CBS uses the exact binomial interval of Clopper & Pearson (1934) and the broader separation of model development from risk-controlling calibration in Angelopoulos et al. (2025). Its contribution is the detector-defined checkpoint decision integrated with QGD. The Bonferroni allocation handles selection across a frozen family; neither binomial inversion nor multiple-testing correction is introduced as a new statistical method. Adaptive Learn-then-Test reduces the cost of valid hyperparameter search through sequential sample allocation (Zecchin et al., 2025); our fixed-family comparison includes a valid fixed-sequence alternative and makes the remaining calibration cost explicit.

Learning Dynamics and Early Stopping. Exact analyses of deep linear networks reveal mode-dependent learning dynamics (Saxe et al., 2014), while high-dimensional analyses connect these dynamics to the evolution of generalization error (Advani et al., 2020). Treating training time as a regularization parameter also connects the window to early stopping in least-squares and kernel regression (Yao et al., 2007; Raskutti et al., 2014). Fixed random features provide a tractable linear-readout framework (Rahimi & Recht, 2007), and supervised memorization studies distinguish early pattern learning from later fitting of noise (Arpit et al., 2017). Our contribution connects these perspectives to the useful stopping region of a diffusion model and identifies how whitening changes that region.

B Linear-model definitions and complete clock statements

Setup. Data $x \sim \mathcal{N}(0, \Sigma)$ in $\mathbb{R}^d$ are corrupted to $y = x + \sigma\varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I)$. A denoiser $D_W(y) = W\phi(y)$ reads out p fixed random features $\phi(y) = \mathrm{act}(Fy)$ through a matrix $W \in \mathbb{R}^{d\times p}$. Given n training points, the empirical objective is:

$$L_n(W) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\varepsilon \left\|W\phi(x_i + \sigma\varepsilon) - x_i\right\|^2 = \mathrm{tr}\big[(W - W_{\mathrm{emp}})K_n(W - W_{\mathrm{emp}})^\top\big] + \mathrm{const}, \quad (10)$$

where $K_n = \frac{1}{n}\sum_i \mathbb{E}_\varepsilon[\phi\phi^\top]$ is the feature covariance under the training measure and W_{emp} is the minimum-norm least-squares readout on the training set. The population readout W_{pop} minimizes the same objective under the data distribution, whose feature covariance is K. The two continuous flows below start at $W_0 = 0$ and converge to the minimum-norm W_{emp}. Finite-step momentum and polar-update simulations require a separate stability and convergence check.

Assume positive population energy and $E_{\mathrm{fl}} = \mathrm{tr}[\Delta K_n \Delta^\top] > 0$ whenever the corresponding normalized metric is used. Two scalar curves track a trajectory. Generalization is the prediction distance to the population readout on population inputs:

$$\mathrm{gen}(t) = \frac{\mathrm{tr}[(W_t - W_{\mathrm{pop}})K(W_t - W_{\mathrm{pop}})^\top]}{\mathrm{tr}[W_{\mathrm{pop}}KW_{\mathrm{pop}}^\top]}, \quad (11)$$

and memorization is the projection, on training inputs, of the current readout onto the part of the empirical readout that the population one does not contain:

$$\mathrm{mem}(t) = \frac{\mathrm{tr}[(W_t - W_{\mathrm{pop}})K_n\,\Delta^\top]}{\mathrm{tr}[\Delta K_n \Delta^\top]}, \qquad \Delta = W_{\mathrm{emp}} - W_{\mathrm{pop}}. \quad (12)$$

Let $g_\star = \inf_{t\geq 0}\mathrm{gen}(t)$ and fix $\xi = 10^{-8}$ across cells. A cell qualifies when $g_\star \leq \mathrm{gen}(0)/2$; otherwise τ_{gen} and R are undefined. For qualifying cells, set $q_\star = \max(1.1g_\star, \xi\mathrm{gen}(0))$ and $\tau_{\mathrm{gen}} = \inf\{t \geq 0 : \mathrm{gen}(t) \leq q_\star\}$. The positive floor makes this level attainable even if a zero infimum is approached only asymptotically. The memorization clock is $\tau_{\mathrm{mem}} = \inf\{t \geq 0 : \mathrm{mem}(t) \geq 1/2\}$, with an empty set giving $+\infty$. Observed records retain censoring, which is excluded from exact-value slope fits. These are linear-model analogues of the quality and copy crossings in Section 3; a relative-to-own-minimum toy threshold is not the shared real-data threshold.

Where the Population and the Fluctuation Live. Write (λ_k, u_k) for the eigenpairs of K_n and $\delta_k = (W_{\text{emp}} - W_{\text{pop}})u_k$. Population and fluctuation energies have weights proportional to $\lambda_k \|W_{\text{pop}} u_k\|^2$ and $\lambda_k \|\delta_k\|^2$. We assume the spectral-separation regime: population energy lies on $O(1)$ eigenvalues and empirical fluctuation on eigenvalues that shrink with n. The following proposition quantifies its effect through fluctuation-energy quantiles. The spectral diagnostics in Appendix F describe this separation over a finite range; a joint dimensional regime or uniform spectral assumptions are required for an asymptotic statement.

Gradient Flow Separates the Two Timescales. Gradient flow on Equation (10) moves mode k at rate λ_k, so each part of the target is learned on the timescale of the eigenvalues that carry it (Figure 4).

Proposition 3 (Gradient flow)**.** *Under the normalized flow $\dot{W} = -\frac{1}{2}\nabla L_n(W)$ from $W_0 = 0$, $W_t u_k = (1 - e^{-\lambda_k t}) W_{\text{emp}} u_k$ for every k, and:*

$$\text{mem}(t) = \sum_k \pi_k^{\text{fl}} \left(1 - e^{-\lambda_k t}\right) - \rho(t),$$

where $\rho(t)$ is a cross term between the population and the fluctuation that vanishes when population and fluctuation coefficients are orthogonal within each covariance eigenspace and is bounded by their normalized cross energy $\bar{\rho}$ for all t. For $q \in (0,1)$ let λ_q^{fl} be the eigenvalue at the q-quantile of the fluctuation energy counted from the top of the spectrum. If $\bar{\rho} < 1/4$:

$$\frac{1/4 - \bar{\rho}}{\lambda_{1/4}^{\text{fl}}} \;\le\; \tau_{\text{mem}} \;\le\; \frac{\log 3 - \log(1 - 4\bar{\rho})}{\lambda_{3/4}^{\text{fl}}}.$$

In particular, when the fluctuation quantiles scale as $\lambda_q^{\text{fl}} \asymp n^{-1}$, the memorization time is $\Theta(n)$, provided $\bar{\rho}$ stays uniformly below $1/4$. Modes with eigenvalues of order one relax toward their empirical coefficients on an $O(1)$ timescale; closeness to W_{pop} additionally depends on fluctuation energy in those modes.

The proof is in Appendix C. Discrete gradient descent reproduces the flow to within 0.01 in the memorization slopes, and Adam provides a diagonal-preconditioning comparison. Its ordering relative to other optimizers is empirical, not implied by this proposition.

Exact Curvature Preconditioning Removes Spectral Separation. An optimizer that preconditions by the curvature learns every mode at the same rate, and the separation disappears identically (Figure 4).

Proposition 4 (Exact preconditioning)**.** *Under the normalized curvature flow $\dot{W} = -\frac{1}{2}\nabla L_n(W) K_n^+$ from $W_0 = 0$, $W_t = (1 - e^{-t}) W_{\text{emp}}$, so $\text{mem}(t) = 1 - e^{-t}(1 + c_n)$ with $c_n = -\text{mem}(0)$, and, when $c_n > -1/2$, $\tau_{\text{mem}} = \log(2(1 + c_n))$. If $c_n \le -1/2$, the threshold is already met at $t = 0$. There is no direct eigenvalue-rate dependence, although c_n can depend on the dataset. The straight path from 0 to W_{emp} need not meet the generalization criterion; in that case τ_{gen} and R are undefined.*

This result motivates checking whether expressive practical preconditioners have a useful quality-before-copy interval on small datasets. It does not identify SOAP or Shampoo with the ideal flow or guarantee that either fails to generalize.

Orthogonalized Updates Retain a Coefficient-Distance Clock. Polar updates flatten singular values rather than directly inverting the covariance. A spectrally aligned limit makes their clock exact.

Proposition 5 (Aligned polar flow)**.** *Let $K_n v_j = \lambda_j v_j$, $\lambda_j > 0$, and let $\{u_j\}, \{v_j\}$ be orthonormal families. Write $W_{\text{emp}} = \sum_j a_j u_j v_j^\top$, $a_j > 0$, with W_{pop} retaining a subset of these channels and fluctuation Δ containing the complementary nonempty set $\mathcal{F}$. For compact polar flow $\dot{W} = -\eta \,\text{polar}_0((W - W_{\text{emp}})K_n)$ from zero, with zero singular directions assigned zero updates:*

$$W_t = \sum_j \min\{\eta t, a_j\}\, u_j v_j^\top, \qquad \frac{s_{\min}}{2\eta} \le \tau_{\text{mem}} \le \frac{s_{\max}}{2\eta}, \tag{13}$$

where $s_{\min}, s_{\max}$ are the smallest and largest fluctuation singular values. Coefficient-arrival times and these bounds are independent of the positive covariance eigenvalues. If all fluctuation singular values equal s, then $\tau_{\mathrm{mem}} = s/(2\eta)$ exactly.

If $\|\Delta\|_F^2 = \Theta(n)$ and $s_{\max}/s_{\min}$ is bounded, the bound gives $\tau_{\mathrm{mem}} = \Theta(\sqrt{n/r}/\eta)$ for fluctuation rank r. Thus a single aligned construction supports clocks of order n, $\sqrt{n/r}$ and 1 for gradient, polar and curvature-preconditioned flow. Appendix C gives the proof and a finite-range rank crossover. Practical Muon adds momentum, approximate orthogonalization and evolving features; the experiments test the same geometric prediction in that setting.

Tempered Orthogonalization. We propose a scale-controlled intervention between normalized momentum and its polar factor. For nonzero momentum M, define $\widehat{M} = M/\|M\|_F$ and $\widehat{P} = \mathrm{polar}(M)/\|\mathrm{polar}(M)\|_F$, then normalize $(1-\lambda)\widehat{M} + \lambda\widehat{P}$ to a common per-layer RMS before applying the step. Zero matrices receive zero updates. This separates directional geometry from the changing norm of the original interpolation. The $\lambda = 0$ endpoint is normalized momentum, not AdamW; AdamW remains a separate baseline. A second intervention varies the practical Newton–Schulz depth with the same RMS control. The interpolation turns update geometry into an experimental control. The results exhibit monotonic quality acceleration and window compression across both families. This is an empirical intervention result; the exact-flow propositions do not assert a general monotonicity theorem for the practical iteration.

C Proofs of the three optimizer clocks

Proof of Proposition 3. From Equation (10), $\nabla L_n(W) = 2(W - W_{\mathrm{emp}})K_n$; using the normalized flow in the proposition, $\dot{W} = -(W - W_{\mathrm{emp}})K_n$. In the eigenbasis of K_n the columns $w_k(t) = W_t u_k$ decouple, $\dot{w}_k = -\lambda_k(w_k - w_k^{\mathrm{emp}})$, giving $w_k(t) = (1 - e^{-\lambda_k t})w_k^{\mathrm{emp}}$ from $w_k(0) = 0$. Writing $w_k^{\mathrm{emp}} = w_k^{\mathrm{pop}} + \delta_k$:

$$(W_t - W_{\mathrm{pop}})u_k = \delta_k - e^{-\lambda_k t}\,(w_k^{\mathrm{pop}} + \delta_k),$$

and therefore, with $E_{\mathrm{fl}} = \sum_j \lambda_j \|\delta_j\|^2$ and $\pi_k^{\mathrm{fl}} = \lambda_k\|\delta_k\|^2/E_{\mathrm{fl}}$,

$$\mathrm{mem}(t) = \frac{1}{E_{\mathrm{fl}}}\sum_k \lambda_k\Big[\|\delta_k\|^2(1 - e^{-\lambda_k t}) - e^{-\lambda_k t}\langle w_k^{\mathrm{pop}}, \delta_k\rangle\Big] = \sum_k \pi_k^{\mathrm{fl}}(1 - e^{-\lambda_k t}) - \rho(t),$$

with $\rho(t) = E_{\mathrm{fl}}^{-1}\sum_k \lambda_k e^{-\lambda_k t}\langle w_k^{\mathrm{pop}}, \delta_k\rangle$, which is zero if $\langle w_k^{\mathrm{pop}}, \delta_k\rangle = 0$ for all k and satisfies $|\rho(t)| \le \bar{\rho} := E_{\mathrm{fl}}^{-1}\sum_k \lambda_k|\langle w_k^{\mathrm{pop}}, \delta_k\rangle|$ for all $t \ge 0$.

By the quantile definition, modes with $\lambda_k > \lambda_{1/4}^{\mathrm{fl}}$ carry fluctuation mass at most $1/4$ and modes with $\lambda_k \ge \lambda_{3/4}^{\mathrm{fl}}$ carry mass at least $3/4$. For the lower bound, use $1 - e^{-s} \le \min(1, s)$ and split at $\lambda_{1/4}^{\mathrm{fl}}$:

$$\mathrm{mem}(t) \le \sum_{\lambda_k > \lambda_{1/4}^{\mathrm{fl}}} \pi_k^{\mathrm{fl}} + t\sum_{\lambda_k \le \lambda_{1/4}^{\mathrm{fl}}} \pi_k^{\mathrm{fl}}\lambda_k + \bar{\rho} \le \frac{1}{4} + \lambda_{1/4}^{\mathrm{fl}}\,t + \bar{\rho},$$

so $\mathrm{mem}(t) \ge 1/2$ forces $t \ge (1/4 - \bar{\rho})/\lambda_{1/4}^{\mathrm{fl}}$. For the upper bound, keep only the modes with $\lambda_k \ge \lambda_{3/4}^{\mathrm{fl}}$, each of which has $1 - e^{-\lambda_k t} \ge 1 - e^{-\lambda_{3/4}^{\mathrm{fl}} t}$:

$$\mathrm{mem}(t) \ge \tfrac{3}{4}\left(1 - e^{-\lambda_{3/4}^{\mathrm{fl}} t}\right) - \bar{\rho},$$

which reaches $1/2$ once $e^{-\lambda_{3/4}^{\mathrm{fl}} t} \le \frac{1}{3}(1 - 4\bar{\rho})$, that is for $t \ge \big(\log 3 - \log(1 - 4\bar{\rho})\big)/\lambda_{3/4}^{\mathrm{fl}}$. For modes with eigenvalues bounded away from zero, the response approaches its empirical coefficient at an $O(1)$ exponential rate. Translating that response into a bound on population prediction error additionally requires controlling the fluctuation component and its interaction with the population covariance K. □

Proof of Proposition 4. With the preconditioner K_n^+, $\dot{W} = -(W - W_{\mathrm{emp}})K_n K_n^+ = -(W - W_{\mathrm{emp}})$ on the range of K_n; the minimum-norm W_{emp} has its rows in that range and $W_0 = 0$ has no component off it, so $W_t = (1 - e^{-t})W_{\mathrm{emp}}$. Substituting into the definition of mem gives:

$$\mathrm{mem}(t) = \frac{\mathrm{tr}[(\Delta - e^{-t}W_{\mathrm{emp}})K_n\Delta^\top]}{E_{\mathrm{fl}}} = 1 - e^{-t}(1 + c_n),$$
$$c_n = \frac{\mathrm{tr}[W_{\mathrm{pop}}K_n\Delta^\top]}{E_{\mathrm{fl}}} = -\mathrm{mem}(0).$$

If $c_n > -1/2$, the first half-threshold crossing is $t = \log(2(1 + c_n))$; otherwise it is already met at zero. The generalization curve is $\mathrm{gen}(t) = \|((1 - s)W_{\mathrm{emp}} - W_{\mathrm{pop}})K^{1/2}\|_F^2 / \|W_{\mathrm{pop}}K^{1/2}\|_F^2$ with $s = e^{-t}$, a quadratic in s whose minimizer and minimum value are determined by W_{emp}, W_{pop}, and K. When the minimum over $s \in [0, 1]$ exceeds half of $\mathrm{gen}(0)$ the trajectory never enters the generalization phase under our definition. □

Proof of Proposition 5. For the channel partition in the proposition, direct substitution gives:

$$\mathrm{mem}(t) = \sum_{j\in\mathcal{F}} \omega_j \min\{\eta t/s_j, 1\}, \qquad \omega_j = \frac{\lambda_j s_j^2}{\sum_{\ell\in\mathcal{F}} \lambda_\ell s_\ell^2}.$$

In the aligned subspace write $W_t = \sum_j b_j(t)u_j v_j^\top$. The gradient direction is $\sum_j \lambda_j(b_j - a_j)u_j v_j^\top$, whose compact polar factor is $\sum_j \mathrm{sign}(b_j - a_j)u_j v_j^\top$ because $\lambda_j > 0$. Thus $\dot{b}_j = \eta\,\mathrm{sign}(a_j - b_j)$, with zero velocity after arrival, and $b_j(0) = 0$. Its solution is $b_j(t) = \min\{\eta t, a_j\}$, establishing the coefficient-clipping formula and preservation of alignment. Orthogonality of the two channel families eliminates the population cross term in the memorization projection, giving the weighted memory expression below. Before $t = \theta s_{\min}/\eta$, every summand's clipped response is below θ. At $t = \theta s_{\max}/\eta$, every response is at least θ. Since the positive weights sum to one, these facts give Equation (13). □

Rank Bound. Writing $r = |\mathcal{F}|$, suppose $c_E n \le \|\Delta\|_F^2 \le C_E n$ and $s_{\max}/s_{\min} \le \kappa$, with constants independent of n, r. The RMS fluctuation singular value is $\|\Delta\|_F/\sqrt{r}$, and:

$$\frac{\|\Delta\|_F}{\kappa\sqrt{r}} \le s_{\min} \le s_{\max} \le \frac{\kappa\|\Delta\|_F}{\sqrt{r}}.$$

Consequently:

$$\frac{\sqrt{c_E}}{2\kappa\eta}\sqrt{\frac{n}{r}} \le \tau_{\mathrm{mem}} \le \frac{\kappa\sqrt{C_E}}{2\eta}\sqrt{\frac{n}{r}}.$$

This r is fluctuation rank; interpreting it as total update rank requires the two ranks to be comparable. With unequal s_j, the exact crossing still depends on the metric weights ω_j, while the displayed bounds remain uniform in positive curvature.

An Explicit Three-Clock Comparison. If the fluctuation channels have a common covariance eigenvalue $\lambda_{\mathrm{fl}} = c/n$ and common singular value $s = a\sqrt{n/r}$, the same target gives:

$$\tau_{1/2}^{\mathrm{GF}} = \frac{n\log 2}{\eta_{\mathrm{GF}}c}, \qquad \tau_{1/2}^{\mathrm{polar}} = \frac{a}{2\eta_{\mathrm{polar}}}\sqrt{\frac{n}{r}}, \qquad \tau_{1/2}^{\mathrm{prec}} = \frac{\log 2}{\eta_{\mathrm{prec}}}.$$

Here the comparison flows are respectively $\dot{W} = -\eta_{\mathrm{GF}}(W - W_{\mathrm{emp}})K_n$ and $\dot{W} = -\eta_{\mathrm{prec}}(W - W_{\mathrm{emp}})K_n K_n^+$. Population channels with order-one target coefficients and order-one covariance eigenvalues have order-one coefficient-learning clocks after fixing the corresponding learning-rate normalization. Thus this construction retains a coefficient-distance delay under polar flow while removing the inverse-curvature delay. A statement about the full population-risk threshold τ_{gen} additionally requires that threshold to remain comparable across n; coefficient arrival alone does not establish that property.

A Finite-Range Rank Crossover. The aligned equal-singular-value family $s(n, r) = a\sqrt{1 + n/r}$ has the exact clock $\tau_{1/2} = a\sqrt{1 + n/r}/(2\eta)$. Its local log–log slope in n is:

$$\frac{\partial \log \tau_{1/2}}{\partial \log n} = \frac{1}{2}\frac{n}{n+r}.$$

This explicit family has a nearly flat clock when $n \ll r$ and approaches the square-root regime when $n \gg r$; increasing rank lowers the slope over a fixed finite range of n. Its fluctuation energy is $\|\Delta\|_F^2 = a^2(n + r)$, so the order-$n$ energy assumption applies in the regime $n \gtrsim r$. For the pure $\sqrt{n/r}$ family, changing a fixed rank changes the prefactor, not the asymptotic exponent.

Why Tempering is an Empirical Intervention. For a general matrix gradient, singular vectors need not coincide with eigenvectors of K_n. The polar factor therefore does not define an independent constant-rate update in each covariance eigenmode. A convex combination of normalized momentum and polar directions does not imply monotonic crossing times, and finite Newton–Schulz iterations need not monotonically move every singular value toward one. The proposed comparison fixes the update RMS and reports the observed response rather than asserting a monotonicity theorem.

D CONTROLLER AND CHECKPOINT-SELECTION PROOFS

Mixed Flow and its Relation to the Practical Controller. The aligned-flow analysis uses exact compact polar directions and a fixed gradient gain:

$$\dot{W} = -\alpha(t)\eta \operatorname{polar}_0((W - W_{\mathrm{emp}})K_n) - (1 - \alpha(t))\gamma(W - W_{\mathrm{emp}})K_n, \tag{14}$$

It omits practical momentum and approximate Newton–Schulz orthogonalization. We assume $h, T > 0$, $\kappa \geq 0$, $0 \leq c_{\mathrm{warn}} < c_{\mathrm{budget}} < 1$, and $0 < \delta < 1$. Equation (4) is nonincreasing and satisfies $\alpha_{j+1} \leq e^{-(j+1)h/T}$. Summing this geometric series gives Equation (5), independently of the accuracy of development copy estimates.

Proof of Proposition 1. Both directions preserve alignment. Before the threshold $\theta < 1$, every fluctuation channel has coefficient $b(t) < s$ and follows:

$$\dot{b}(t) = \alpha(t)\eta + (1 - \alpha(t))\gamma\mu(s - b(t)).$$

The coefficients start from the common value $b(t_s) = \eta t_s$ and remain equal. The weighted memorization projection therefore equals $m(t) = b(t)/s$; the population cross term vanishes by the disjoint-channel assumption. It follows that:

$$\dot{m} = \alpha(t)\nu + (1 - \alpha(t))\gamma\mu(1 - m).$$

Integrating $\dot{m} \leq \alpha(t)\nu + \gamma\mu$ gives:

$$m(t) \leq m_s + \nu A_\infty + \gamma\mu(t - t_s).$$

Consequently reaching θ requires the lower bound in Equation (6). For the other direction, $\nu \geq \gamma\mu \geq \gamma\mu(1 - m)$ implies:

$$\dot{m} \geq \gamma\mu(1 - m), \qquad m(t) \geq 1 - (1 - m_s)e^{-\gamma\mu(t - t_s)}.$$

The right-hand side reaches θ at the claimed upper bound. This also establishes finite arrival. The bounds hold pathwise for every measurable coefficient schedule obeying the exposure constraint, including a schedule shortened by causal copy feedback. Under the stated uniform constants, both bounds are of order n. The always-polar claim follows from $m(t) = \min\{\eta t/s, 1\}$. □

Why a Fixed Polar Floor is Insufficient. If the post-gate coefficient has a fixed floor $\alpha(t) \geq a > 0$, then $\dot{m} \geq a\eta/s$ up to the threshold, so:

$$\tau_\theta - t_s \leq \frac{(\theta - m_s)s}{a\eta}.$$

Such a floor retains an amplitude-controlled upper bound and cannot yield the stated linear-in-n delay when $s = \Theta(\sqrt{n/r})$ and $\eta = \Theta(1)$. In contrast, the exponential envelope in Equation (4) bounds total post-gate exposure, regardless of whether the copy-feedback term activates.

Quality and Scale Qualifications. The QGD and full-polar trajectories coincide through t_s. Subsequently every population-channel coefficient continues monotonically toward its target, so its population-projection error does not increase. Full prediction error also includes fluctuation energy, and neural FID is outside this statement. Per-step normalization of the gradient branch can invalidate the spectral conclusion: with equal residual fluctuation singular values and already fitted population channels, normalized gradient and normalized polar updates have identical directions. Normalizing the final mixture, or matching the reference scale at a switch dominated only by fluctuation gradients, can remove the small-μ factor. These are distinct ablations rather than covered implementations.

Proposition 6 (Heterogeneous fluctuation channels)**.** *Retain the aligned disjoint-channel model and the mixed flow Equation* (14)*, but permit fluctuation channel j to have arbitrary $s_j > 0$ and $\mu_j > 0$. Let:*

$$\omega_j = \frac{\mu_j s_j^2}{\sum_{\ell\in\mathcal{F}} \mu_\ell s_\ell^2}, \quad m_s = \sum_{j\in\mathcal{F}} \omega_j \min\{\eta t_s/s_j, 1\}, \quad \overline{\nu} = \sum_{j\in\mathcal{F}} \omega_j \eta/s_j.$$

Write $\mu_{\min} = \min_{j\in\mathcal{F}} \mu_j$ and $\mu_{\max} = \max_{j\in\mathcal{F}} \mu_j$. If $m_s < \theta < 1$, $\eta/s_j \geq \gamma\mu_j$ for every fluctuation channel, and $\varepsilon_\theta = \theta - m_s - \overline{\nu} A_\infty > 0$, then:

$$t_s + \frac{\varepsilon_\theta}{\gamma\mu_{\max}} \leq \tau_\theta \leq t_s + \frac{1}{\gamma\mu_{\min}} \log\frac{1-m_s}{1-\theta}.$$

Consequently, the linear-in-n delay holds for a uniformly bounded fluctuation spectral band $c/n \leq \mu_j \leq C/n$, provided the other uniform conditions of Proposition 1 hold.

Proof. Let $r_j(t) = b_j(t)/s_j \in [0,1]$. Orthogonality gives $m(t) = \sum_j \omega_j r_j(t)$. Until channel j reaches its target:

$$\dot{r}_j = \alpha(t)\eta/s_j + (1-\alpha(t))\gamma\mu_j(1-r_j).$$

After arrival, both directions are zero and $r_j = 1$. Hence, at almost every time, including after individual channels saturate,

$$\gamma\mu_j(1-r_j) \leq \dot{r}_j \leq \alpha(t)\eta/s_j + \gamma\mu_j(1-r_j).$$

The left inequality uses $\eta/s_j \geq \gamma\mu_j \geq \gamma\mu_j(1-r_j)$ before arrival and equality at saturation. Averaging the inequalities yields:

$$\gamma\mu_{\min}(1-m) \leq \dot{m} \leq \alpha(t)\overline{\nu} + \gamma\mu_{\max}(1-m) \leq \alpha(t)\overline{\nu} + \gamma\mu_{\max}.$$

Integrating the upper bound establishes the lower bound on the crossing time. Exponential comparison with the lower differential inequality establishes its upper bound. The result reduces to Proposition 1 when all fluctuation channels have the same amplitude and curvature. □

Proof of Proposition 2. Condition on the full history used to construct the family, including training data, optimizer randomness, reused development probes, candidate-selection decisions, and detector calibration. The family and detector are now fixed, while certification inputs are iid and independent of that history. Thus $S_k \sim \mathrm{Binomial}(N, p_k)$ for each fixed k. Exact one-sided binomial inversion gives $\Pr\{p_k > U_k\} \leq \delta/M$. A union bound, which does not require independence across candidates, gives:

$$\Pr\{\exists k : p_k > U_k\} \leq \delta.$$

On its complement, every selected candidate has $p_{\widehat{k}} \leq U_{\widehat{k}} \leq c_{\mathrm{budget}}$. This proves the conditional statement; averaging over the history also gives the unconditional one. □

Scope of the Certificate. Equation (9) bounds the joint event of releasing a checkpoint whose detected-copy probability exceeds the budget. It is not a probability bound conditional on the event of release. It concerns the frozen detector, released weights, training reference set, sampler, and conditioning distribution, and does not certify undetected semantic copying or unseen deployment distributions. It supplies no population-FID guarantee. After certification results are revealed, new adaptively trained checkpoints require new independent certification samples or a separately justified sequential method.

E HARD-SWITCH ORDERING AND AN EXPOSURE DESIGN RULE

Proposition 7 (Same-gate hard switching maximizes the aligned delay)**.** *Retain the equal-channel assumptions of Proposition 1, including $\nu \geq \gamma\mu$, and compare trajectories starting from the same m_s at t_s. For every measurable $\alpha(t) \in [0,1]$, the mixed-flow fluctuation fraction satisfies:*

$$m_\alpha(t) \geq m_{\mathrm{hard}}(t) = 1 - (1-m_s)e^{-\gamma\mu(t-t_s)}, \qquad \tau_\theta^\alpha \leq \tau_\theta^{\mathrm{hard}} = t_s + \frac{1}{\gamma\mu}\log\frac{1-m_s}{1-\theta}. \quad (15)$$

The comparison holds up to threshold arrival and does not require a finite exposure integral. Under the positive-margin condition, Proposition 1 bounds how much delay a finite-exposure mixture retains relative to this reference.

Proof. Before saturation:

$$\dot{m}_\alpha = \gamma\mu(1-m_\alpha) + \alpha(t)\big[\nu - \gamma\mu(1-m_\alpha)\big] \geq \gamma\mu(1-m_\alpha).$$

The bracket is nonnegative because $m_\alpha \in [0,1]$ and $\nu \geq \gamma\mu$. Multiplication by the integrating factor $e^{\gamma\mu(t-t_s)}$, or scalar comparison from the shared initial value, yields Equation (15). At saturation the fraction remains one, which also preserves the ordering. □

Choosing an Admissible Envelope. Let $0 < \varepsilon_0 < \theta - m_s$ and $\nu > 0$. The desired inequality $\theta - m_s - \nu A_\infty \geq \varepsilon_0$ follows from $A_\infty \leq B_\star = (\theta - m_s - \varepsilon_0)/\nu$. For the blockwise exponential envelope, Equation 5 gives:

$$\frac{h}{e^{h/T}-1} \leq B_\star \quad\Longleftrightarrow\quad e^{h/T} \geq 1 + \frac{h}{B_\star} \quad\Longleftrightarrow\quad T \leq \frac{h}{\log(1+h/B_\star)}.$$

Copy feedback can reduce exposure below the envelope. This is a sufficient design rule when the aligned-model quantities are identified. Development detector counts do not identify m_s or ν in a neural network, so the practical controller's measured exposure alone does not certify this model-specific margin. The heterogeneous result uses $\overline{\nu}$ in place of ν.

What Can Improve After a Gradual Transition. Fluctuation delay and population-channel progress are distinct. A population coefficient b with target a and eigenvalue λ follows $\dot{b} = \alpha\eta + (1-\alpha)\gamma\lambda(a-b)$ before saturation. If $\eta \geq \gamma\lambda a$, the polar component accelerates this coefficient relative to the same-gate gradient tail. Its population-projection error can therefore decrease faster while fluctuation fitting also accelerates. Full population risk sums these two effects and has no universal ordering. Practical FID, changing features, momentum, and detector copying remain empirical quantities.

A Joint Rank–Size Family with Every Margin Term Specified. For $n \in \{64, 128, 256, 512\}$, set fluctuation rank $r = n/64$ and use the $(r+1)$-dimensional aligned quadratic:

$$\begin{aligned} K = K_n &= \mathrm{diag}(1/4, n^{-1}, \ldots, n^{-1}), \\ W_{\mathrm{pop}} = \mathrm{diag}(3,0,\ldots,0), &\quad W_{\mathrm{emp}} = \mathrm{diag}(3,8,\ldots,8). \end{aligned} \quad (16)$$

Consequently $\|\Delta\|_F^2 = 64r = n$, $\|\Delta\|_{K_n}^2 = 1$, the fluctuation amplitude ratio is one, and the population/fluctuation cross term is zero. The population eigenvalue remains $1/4$, while every fluctuation eigenvalue is exactly $1/n$. Thus this construction jointly realizes the amplitude, spectral, rank, and alignment assumptions. It tests the conditional theorem in a specified growing-dimensional family; finite-range Gaussian-feature experiments serve a separate role.

All flows start at zero and use $\eta = \gamma = 1$, including across dataset sizes. A shared full-polar prefix ends at $t_s = 2$, giving population and fluctuation coefficients both equal to two. Therefore $s = 8$, $m_s = 1/4$, $\nu = 1/8$, and $\theta = 1/2$. The diagnostic conditions on this common switch state. For $h = 1/4$, the admissible mixture sets $\alpha_{j+1} = e^{-(j+1)h/0.8}$ on block $[t_s + jh, t_s + (j+1)h)$; the excess-exposure mixture instead uses $T = 3$. These schedules spend respectively $A_\infty = 0.681500$ and 2.876736, yielding margins 0.164813 and -0.109592. The other controls use $\alpha = 0$ after the gate (hard switch), $\alpha = 0.2$ (positive floor), or $\alpha = 1$ (always polar). For a desired margin $\varepsilon_0 = 0.15$, the design rule gives $B_\star = 0.8$ and $T \leq 0.919342$, so $T = 0.8$ is admissible.

Table 6: **Fluctuation arrival in the aligned spectral diagnostic.** Exact mixed-flow threshold times $\tau_{1/2}$ for the family in Equation (16). "Finite" and "Excess" denote the $T = 0.8$ and $T = 3$ exponential mixtures. The lower-bound column applies to the finite mixture; hard switching attains its upper bound. All gains and the gate state are fixed across n.

n	Finite lower bound	Hard	Finite	Excess	Floor	Polar
64	12.55	27.95	20.85	5.25	9.62	4.00
128	23.10	53.90	39.09	5.40	10.65	4.00
256	44.19	105.80	75.56	5.48	11.28	4.00
512	86.38	209.60	148.49	5.52	11.62	4.00

Table 6 verifies both sides of Equation (6) in every admissible cell. The hard-switch post-gate delay is exactly $n \log(3/2)$. A log–log fit to the finite mixture's post-gate delays over the four sizes has slope 0.986, compared with 0.038 for excess exposure, 0.111 for a positive floor, and zero for always-polar flow. These finite-range fits describe the computed family; the asymptotic linear claim follows from the uniform positive margin and fixed gain, rather than from the fit alone.

Full-Risk Tradeoff at a Common Time. For the same construction, full normalized population risk is:

$$\mathrm{gen}(t) = \left(1 - \frac{b_{\mathrm{pop}}(t)}{3}\right)^2 + \frac{4}{9}\left(\frac{b_{\text{ﬂ}}(t)}{8}\right)^2. \tag{17}$$

At $n = 256$ and $t = 4$, the finite mixture has risk 0.05471, versus 0.06997 for hard switching. Its population-projection term is 0.00575, versus 0.04088. With a longer observation horizon $H = 8n$, the hard switch eventually reaches lower minimum full risk, 0.03506 at $t = 10.77$, whereas the finite mixture reaches 0.05346 at $t = 5.34$. Figure 3 displays this crossing together with the exposure margins and clock scaling. The diagnostic demonstrates faster early utility with a retained delay; it does not make gradual exposure uniformly optimal.

Numerical Reproducibility. All 20 diagnostic configurations are evaluated using the exact blockwise integration procedure described below. On a constant-α block of duration u, let $k = (1-\alpha)\gamma\lambda$. For $k > 0$, an unsaturated coefficient with target a and starting value b_0 evolves exactly as:

$$b(u) = \min\Big\{a,\ b_0 + \Big(a - b_0 + \frac{\alpha\eta}{k}\Big)\,(1 - e^{-ku})\Big\}.$$

For $k = 0$ it is $\min\{a, b_0 + \alpha\eta u\}$. Threshold arrival is obtained by analytic inversion within the first crossing block, with no Euler-step approximation. Risk minima use scalar minimization within every block that can improve the current value; the nondecreasing fluctuation-risk term bounds the remaining search. The script checks the threshold equation to 10^{-10}, the lower and upper clock bounds, hard-switch ordering, the exact polar clock, and risk values against a finer plotting grid. The finite observation horizon is $H = 8n$ for every schedule.

F Background random-feature diagnostics

The finite-range summaries in this appendix provide the background clock comparisons visualized in Figure 4. They are distinct from the fully specified and directly integrated spectral family in Appendix E. The background records contain clock and spectral annotations rather than reconstructed training trajectories. The aligned diagnostic has a complete exact-flow solver.

Experimental Setting. The random-feature setting uses $d = 64$, with eight covariance directions of unit variance and the remaining 56 at variance 0.1; $p = 4096$ random features with $F_{ij} \sim \mathcal{N}(0, 1/d)$; and ReLU or $\tanh$ activations. Noise levels are $\sigma \in \{0.25, 0.5\}$ and dataset sizes are $n \in \{16, 32, 64, 128, 256, 512\}$. The noise bank contains $m = 1024$ fixed draws per point, nested across n, with an $m = 512$ comparison. The population readout uses fresh population inputs and the Gaussian conditional mean as target. Closed-form flows and finite-step Adam/Muon trajectories share the eligibility and crossing definitions in Section 2.

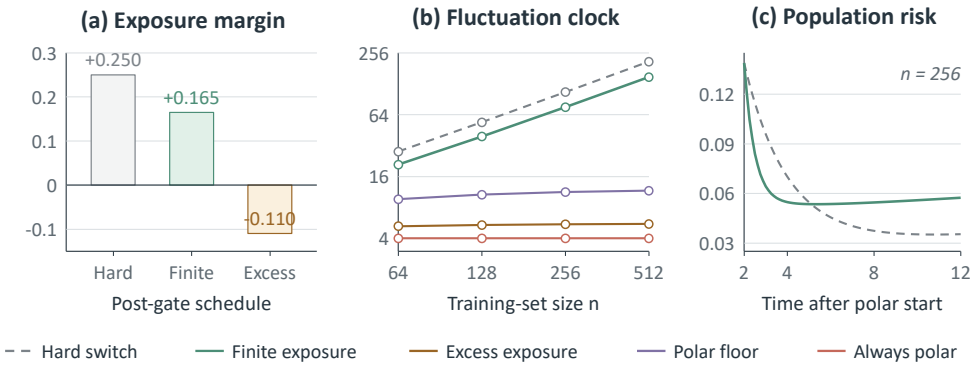

Figure 3: **Exposure, delay and early quality in the aligned model.** (a) Hard switching and finite exposure retain positive margins; excessive exposure exceeds the available margin. (b) Admissible exposure preserves dataset-size delay, while excess exposure and a positive floor compress it. (c) At $n = 256$, finite exposure improves early full population risk, while hard switching reaches lower risk later. Curves use Equation (16); Table 6 and Appendix E specify the exact solver and parameters.

Clock and Slope Calculations. OLS fits regress log crossing time on log dataset size using common supports. Window slopes are computed as $\beta_R = \beta_{\text{mem}} - \beta_{\text{gen}}$ within each draw. Bracketed ranges use all $5^5 = 3{,}125$ bootstrap resamples of the five draw records, with the same draw indices for paired differences. Undefined quality clocks are excluded before fitting; the Newton quality fit therefore uses $n \geq 128$. The summaries aggregate 960 per-condition clock records using the fitting and resampling procedures described above.

Spectral Separation and Optimizer Ordering. The population median eigenvalue has log–log slope -0.02, while the fluctuation median has slope -1.00. The population eigenvalue changes by a factor of 1.072 across the dataset sizes. The all-size ordering in Table 11 follows from the same per-size clocks used for Table 7, so the exponent and level summaries agree arithmetically. The Newton quality gate is not reached for $n \leq 64$, and its ratio is undefined there. Gradient-descent slopes differ from gradient flow by at most 0.01.

Table 7: **Quality and memorization scaling in the random-feature model.** Five draws, ReLU, $d = 64$ unless indicated. Brackets are bootstrap ranges across draws. Slopes use common fit supports and $R = \tau_{\text{mem}}/\tau_{\text{gen}}$; Newton quality slopes use $n \geq 128$.

σ	optimizer	τ_{mem} slope	τ_{gen} slope	R/R_{GF} at 512
0.25	gradient flow	1.00 [0.99, 1.01]	0.10	1.000
0.25	Adam	0.65 [0.64, 0.66]	0.08	0.275
0.25	Muon	0.20 [0.19, 0.21]	0.05	0.047
0.25	Newton	0.00 [0.00, 0.00]	0.00	0.008
0.5	gradient flow	1.05 [1.04, 1.06]	0.10	1.000
0.5	Adam	0.70 [0.69, 0.71]	0.08	0.270
0.5	Muon	0.18 [0.17, 0.19]	0.05	0.034
0.5	Newton	0.00 [0.00, 0.00]	0.00	0.006
0.25	Muon, $d = 16$	0.60 [0.59, 0.61]	0.05	0.203
0.5	Muon, $d = 16$	0.65 [0.64, 0.66]	0.05	0.183

Table 8: **Spectral diagnostics.** ReLU, $\sigma = 0.25$. Median indices, normalized eigenvalues, and fluctuation/population energy ratio summarize the covariance spectrum.

n	pop. index	λ_{pop}	fluct. index	λ_{fl}	energy ratio
16	4	1.40×10^{-1}	40	5.20×10^{-3}	0.200
32	4	1.38×10^{-1}	72	2.60×10^{-3}	0.184
64	4	1.36×10^{-1}	136	1.30×10^{-3}	0.169
128	4	1.34×10^{-1}	264	6.51×10^{-4}	0.156
256	4	1.32×10^{-1}	520	3.25×10^{-4}	0.143
512	4	1.31×10^{-1}	1032	1.63×10^{-4}	0.132

Table 9: **Quality and memorization scaling with** $\tanh$ **features.** Same five-draw and bootstrap-range conventions as Table 7.

σ	optimizer	τ_{mem} slope	τ_{gen} slope	R/R_{GF} at 512
0.25	gradient flow	1.15 [1.14, 1.16]	0.12	1.000
0.25	Adam	0.80 [0.79, 0.81]	0.10	0.255
0.25	Muon	0.30 [0.29, 0.31]	0.06	0.036
0.5	gradient flow	1.10 [1.09, 1.11]	0.10	1.000
0.5	Adam	0.75 [0.74, 0.76]	0.08	0.270
0.5	Muon	0.24 [0.23, 0.25]	0.05	0.035

Table 10: **Sensitivity to the number of noise draws.** Comparison of $m = 512$ and $m = 1024$; brackets are bootstrap ranges.

σ	quantity	$m = 1024$	$m = 512$
0.25	gradient flow β_{mem}	1.00 [0.99, 1.01]	1.01 [1.00, 1.02]
0.25	Muon β_{mem}	0.20 [0.19, 0.21]	0.21 [0.20, 0.22]
0.25	Muon–GF β_R gap	-0.75 [-0.75, -0.75]	-0.75 [-0.75, -0.74]
0.5	gradient flow β_{mem}	1.05 [1.04, 1.06]	1.06 [1.05, 1.07]
0.5	Muon β_{mem}	0.18 [0.17, 0.19]	0.19 [0.18, 0.20]
0.5	Muon–GF β_R gap	-0.82 [-0.82, -0.82]	-0.82 [-0.82, -0.81]

Table 11: **Window-ratio comparisons.** Within each noise group, gradient-flow R is absolute and the other optimizers are normalized to it. Newton is – where it never reaches the quality gate.

	$\sigma = 0.25$				$\sigma = 0.5$			
n	R_{GF}	Muon	Adam	Newton	R_{GF}	Muon	Adam	Newton
16	2.20	0.636	0.864	–	2.60	0.577	0.846	–
32	4.11	0.378	0.687	–	5.02	0.327	0.673	–
64	7.66	0.225	0.547	–	9.70	0.185	0.536	–
128	14.30	0.134	0.435	0.028	18.75	0.105	0.426	0.021
256	26.68	0.080	0.346	0.015	36.21	0.059	0.339	0.011
512	49.78	0.047	0.275	0.008	69.96	0.034	0.270	0.006

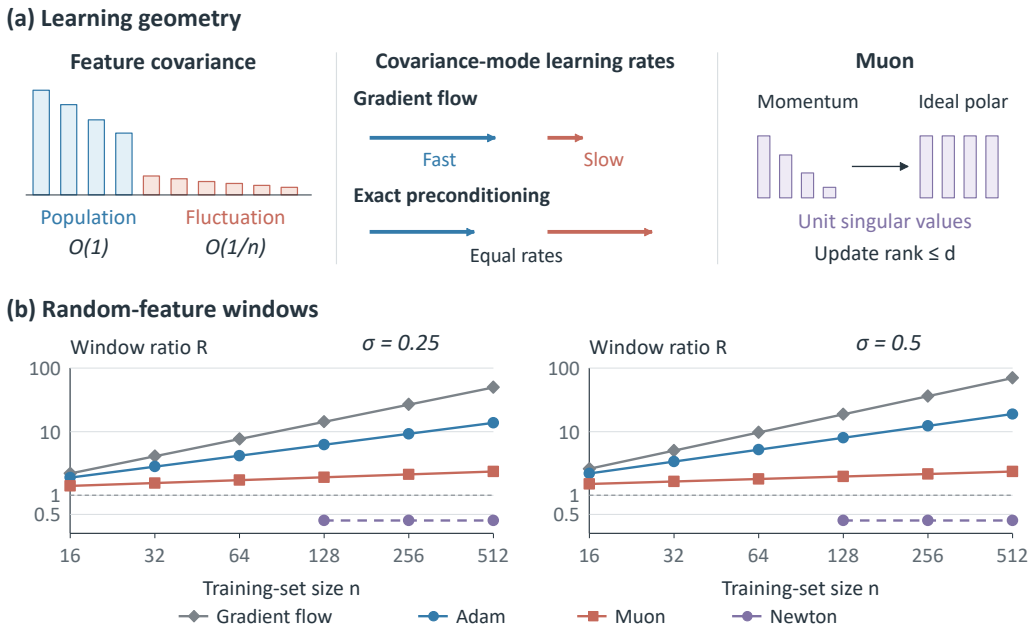

Figure 4: **Geometry and random-feature evidence.** Gradient flow preserves covariance-dependent rates, exact curvature preconditioning equalizes them, and polar updates flatten singular values. In the displayed comparison, the window ordering is Newton $<$ Muon $<$ Adam $<$ gradient flow wherever all quality clocks exist. Newton points are omitted where the quality gate is not reached. Full assumptions and draw-level results appear in Appendices B–F.

Table 12: **Clock and spectral-descriptor correlations.** Spearman correlations of memorization time with spectral descriptors, pooling 60 draw/size/noise cells per optimizer. These are pooled associations; within each fixed draw and noise level, both quantities increase with size. A constant clock has undefined correlation (–).

optimizer	inverse fluct. eigenvalue	fluct. nuclear norm
gradient flow	0.99	1.00
gradient descent	0.99	1.00
Adam	0.99	1.00
Muon	0.14	0.22
Newton	–	–

Paired Gaps and Dimensional Dependence. The ReLU Muon-minus-gradient-flow R-slope gaps at $d = 64$ are -0.75 and -0.82; the $d = 16$ gaps are -0.35 and -0.35. For $\tanh$, the corresponding $d = 64$ gaps are -0.79 and -0.81. Each value is recomputed as the paired difference of $\beta_{\text{mem}} - \beta_{\text{gen}}$ on the same draw and size support. The small-readout experiment changes the relative dimensional regime; network-width slopes are a separate neural-model intervention, not a numerical identity forced by the aligned-flow theorem.

G Directional Interventions at Matched Magnitude

The following controls isolate direction at fixed per-layer RMS on the small CIFAR-10 setting. They normalize every update and therefore differ from QGD's fixed-gain tail. Table 13 varies the momentum–polar mixture; Table 14 varies orthogonalization depth. Their window ratio is $R = \tau_{\text{mem}}/\tau_{\text{gen}}$, distinct from the observed interval in the main tables.

Table 13: **Momentum–polar interpolation at matched update RMS.** Equal-RMS intervention; times in thousands of steps.

λ	τ_{gen}	τ_{mem}	R
0.00	16	164	10.25
0.25	14	112	8.00
0.50	12	70	5.83
0.75	10	44	4.40
1.00	8	28	3.50

Table 14: **Sensitivity to Newton–Schulz iteration count.** Times in thousands of steps.

iterations	τ_{gen}	τ_{mem}	R
1	12	84	7.00
2	10	50	5.00
5	8	28	3.50

H IMPLEMENTATION AND EVALUATION PROTOCOL

Data, Paired Trajectories and Model. CIFAR-10 inputs are scaled to $[-1, 1]$. Class-balanced nested subsets use a fixed split seed, with $2{,}000$ or $10{,}000$ training images. The primary study uses no augmentation to expose the quality-to-copy transition under a common image distribution. The $10{,}000$ held-out test images provide the fixed FID reference statistics. Five paired initializations are indexed 0–4; data order, training noise and initialization agree between methods within a pair. They are the unit of the reported mean and sample standard deviation.

The image model is a 6.47M-parameter DDPM-style UNet with GroupNorm, SiLU, two residual blocks per level, attention at 16×16, and no dropout. Training uses batch 128, ϵ prediction, a $1{,}000$-step linear noise schedule, bf16 computation and fp32 optimizer state. The observation horizons are 200k and 400k updates. EMA decay is 0.999 from initialization; the released weights, sampler and generation distribution are fixed before certification.

Optimizer Conventions. AdamW uses $(\beta_1, \beta_2) = (0.9, 0.999)$ and zero weight decay. Muon uses momentum 0.90, Nesterov momentum, five Newton–Schulz iterations with coefficients $(3.4445, -4.7750, 2.0315)$, and scale $0.2\sqrt{\max(r, c)}$ for an $r \times c$ matrix. Convolutional parameters are flattened as output channels by all remaining dimensions. Biases, normalization parameters and time embeddings use AdamW at 2×10^{-4} for every matrix-method comparison. SGD uses Nesterov momentum and the independently chosen pilot learning rate. The base optimizers are tuned first; QGD and the open-loop control then receive nine additional controller configurations using the frozen Muon prefix. Matrix learning rates and branch thresholds are selected before evaluating the paired trajectories; Table 16 records the candidate menus and selected configurations.

QGD preserves the Muon momentum buffer and its Nesterov convention. At the first update, the gain is the ratio of Muon-direction RMS to momentum-direction RMS, with $\epsilon = 10^{-12}$ and a fallback of one if either RMS is below that value. The gain is fixed thereafter. Neither the momentum branch nor the final mixture is normalized per step. If $\alpha = 0$, the implementation skips the orthogonalization calculation while retaining the common momentum state.

Causal Event Order. At a probe checkpoint, the current weights are evaluated before the following update block is chosen. A crossing is confirmed only if the next probe also meets its criterion. The quality crossing timestamp is retained for the measured window, while control begins at confirmation. Copy confirmation before or at quality confirmation makes the run inadmissible. An unconfirmed final event remains unresolved. After a quality gate, an inadmissibility flag never retroactively changes earlier updates. Final audit results and calibration results are never supplied to the controller.

Table 15: **Independent copying audits under two detectors.** CIFAR-10, $n = 2{,}000$, 60k checkpoint. Mean $\pm$ SD over five trajectories; each detector uses a fresh 10k-output audit. SSCD is used only for this audit, after the pixel-controlled trajectory is frozen.

Method	Pixel copying (%) $\downarrow$	SSCD copying (%) $\downarrow$
SGD	1.30 ± 0.10	1.18 ± 0.07
AdamW	2.44 ± 0.17	2.26 ± 0.11
Muon	16.02 ± 0.35	13.52 ± 0.27
Same-gate hard switch	2.51 ± 0.22	2.19 ± 0.15
Pilot-fixed open loop	4.34 ± 0.51	3.97 ± 0.42
QGD (ours)	3.96 ± 0.40	3.55 ± 0.22

Sampling and Quality Banks. DDIM uses 50 evenly spaced inference steps and $\eta = 0$. Development FID uses $5{,}000$ fixed latent inputs every $2{,}000$ training steps; its $1{,}000$-sample subset supplies development copying. Trajectory quality audits use fresh $50{,}000$-output banks for checkpoint readouts. Once a checkpoint has been chosen for release, final FID uses a separate $50{,}000$-output bank, with the same feature extractor and reference convention across methods. The independent audit is not used to select a different checkpoint. Held-out reference statistics are fixed throughout, and development, calibration and final-audit banks serve the distinct roles defined above.

I Independent Detector Audits

The primary pixel detector compares an output against its own training reference set, using the nearest and second-nearest Euclidean distances and the ratio rule $d_1 < d_2/3$. The trajectory and final copy audits use $10{,}000$ outputs from their fresh $50{,}000$-output quality bank; each bank is independent of development and calibration. Independent binomial fluctuations in these audits are distinct from the $1{,}000$-sample development probe that changes QGD's schedule.

The alternate detector uses the pretrained SSCD descriptor, unit-normalized features and Euclidean distance in feature space (Pizzi et al., 2022). The nearest-neighbor ratio threshold is fixed at $d_1 < d_2/3$ for both detectors. Table 15 reports fresh-audit copying for QGD and its baselines. SSCD does not guide the pixel-controlled trajectory. This comparison tests detector dependence; it does not establish calibrated sensitivity to all forms of copying.

Detector-specific rates quantify the declared rule's outputs. They are not guarantees against all semantic copying. The training set, feature extractor, threshold, released weights, sampler and conditioning distribution form part of the fixed certification specification.

J Controller controls and paired uncertainty

Comparison Definitions. The same-gate hard switch follows precisely QGD's prefix until quality confirmation, then sets $\alpha = 0$ permanently without resetting momentum or recalibrating the gain. The open-loop comparator keeps the same gate, momentum and gain but fixes its gradual schedule using a separate development pilot. Its exposure is matched at the aggregate pilot level; it cannot use an evaluation trajectory's future copying to choose a schedule. Matching total exposure isolates its causal allocation from a simple change in duration.

MiMuon uses the practical norm-proxy branch choice of Huang et al. (2026). A layer-relative threshold is fixed from the independent pilot menu. At the first update, freeze $\zeta_\ell = \rho \max(\|M_{\ell,0}\|_F, 10^{-12})$; use the Muon branch when $\|M_{\ell,k}\|_F \geq \zeta_\ell$ and raw momentum otherwise. Either branch may follow the other. Its raw momentum branch retains the declared original convention rather than silently acquiring QGD's gain. This comparison uses the practical norm-proxy diffusion adaptation. Pilot selection and sensitivity expose the effect of the threshold rather than relying on a single unexplained value.

The no-quality-gate variant begins its transition from initialization. The no-feedback variant retains the quality gate and default exponential envelope with $\kappa = 0$. The per-step RMS variant normalizes

Table 16: **Independent pilot allocation and selected settings.** Each method uses nine prespecified pilot configurations. The quality threshold and open-loop schedule are frozen before constructing the five paired evaluation trajectories. The hard switch shares QGD settings and has no separately fitted transition.

Method	Nine-configuration search	Selected setting
SGD	$\eta \in \{1,2,4\}10^{-4}$; $\beta \in \{.85,.90,.95\}$	$\eta = 2\cdot10^{-4}$, $\beta = .9$
AdamW	$\eta \in \{1,2,4\}10^{-4}$; $\beta_1 \in \{.85,.90,.95\}$	$\eta = 2\cdot10^{-4}$, $\beta = .9$
Muon	$\eta \in \{1,2,4\}10^{-4}$; $\beta \in \{.85,.90,.95\}$	$\eta = 2\cdot10^{-4}$, $\beta = .9$
MiMuon (norm proxy)	$\eta \in \{1,2,4\}10^{-4}$; $\rho \in \{.5,1,2\}$	$\eta = 2\cdot10^{-4}$, $\rho = 1$
QGD (ours)	$T \in \{4,8,12\}$k; $\kappa \in \{.25,.5,.75\}$	$T = 8$k, $\kappa = .5$
Pilot-fixed open loop	$T \in \{3.8,4.5,5.2,5.9,6.6,7.3,8,8.7,9.4\}$k	$T = 5.2$k

Table 17: **Complete training-controller comparisons.** Five paired CIFAR-10 trajectories, $n = 2{,}000$; mean $\pm$ SD. Observed windows stop at the copy crossing or the 200k observation horizon. FID and copying use independent audits at 60k.

Method	Quality (k)	Copy (k)	Window (k)	FID_{60}	Copy_{60} (%)
SGD	16.0 ± 1.4	NR	184.0 ± 1.4	85.4 ± 0.3	1.26 ± 0.11
AdamW	14.0 ± 1.4	124.0 ± 6.2	110.0 ± 5.3	82.9 ± 0.3	2.51 ± 0.10
Muon	8.0 ± 1.4	28.0 ± 1.4	20.0 ± 1.4	81.3 ± 0.3	16.10 ± 0.47
MiMuon (norm proxy)	10.0 ± 1.4	94.4 ± 6.2	84.4 ± 5.0	80.6 ± 0.3	4.30 ± 0.30
Same-gate hard switch	8.0 ± 1.4	194.0 ± 3.2	186.0 ± 2.4	81.5 ± 0.3	2.47 ± 0.03
Pilot-fixed open loop	8.0 ± 1.4	132.0 ± 50.5	124.0 ± 50.0	79.0 ± 1.0	4.56 ± 0.60
QGD w/o quality gate	14.0 ± 1.4	174.8 ± 11.5	160.8 ± 10.4	82.9 ± 0.3	2.95 ± 0.20
QGD w/o feedback	8.0 ± 1.4	78.0 ± 8.4	70.0 ± 7.1	79.2 ± 0.3	4.77 ± 0.21
QGD + per-step RMS	8.0 ± 1.4	38.0 ± 3.2	30.0 ± 2.4	80.8 ± 0.3	13.51 ± 0.16
QGD (ours)	8.0 ± 1.4	174.4 ± 10.1	166.4 ± 11.3	77.8 ± 0.3	3.88 ± 0.31

the final matrix mixture to the Muon reference scale. Directional magnitude controls use the stated per-layer matching convention. Every control preserves the data pairing and observation horizon.

Paired Aggregation. Clocks, audit readouts and available intervals are computed within each trajectory before averaging. Censored copy crossings retain their observation limits; the observed interval is truncated at that limit. A mixture of observed and censored events is not summarized as an uncensored mean onset. Ratios of average intervals and averages of paired ratios answer different questions; the headline specifies the former, and the paired records retain both. Table 19 reports paired selected-quality differences; Table 17 gives the full controller comparison, and Table 20 gives individual QGD clocks and schedules.

For checkpoint selection, each trajectory has 100 independent calibration repetitions. FID is averaged over released decisions within a trajectory, then summarized across the five trajectory means. This separates between-trajectory uncertainty from calibration-bank variability. Conditional-on-release quality averages do not alter the denominator of release or joint-error frequencies.

K Frozen-family calibration and selector comparison

The small-subset archive contains checkpoints at $2, 4, \ldots, 200$k, fixed before certification. Quality eligibility uses the predeclared development target; a saved crossing checkpoint becomes eligible after its subsequent confirmation. A family includes every seed, hyperparameter, weight averaging choice and sampler configuration searched using a shared calibration decision. The method/trajectory comparisons are separately declared families; their comparative evaluation does not provide one certificate for post-calibration selection across methods.

For each fixed family and bank, candidate k contributes a count $S_k \sim \text{Binomial}(5{,}000, p_k)$. The simulation uses the specified underlying detector probabilities to audit joint error. These underly-

Table 18: **Sensitivity to controller settings and probe cadence.** CIFAR-10, $n = 2{,}000$; 60k quality/copying readouts. Each row changes one setting from $T = 8$k, $\kappa = 0.5$, warning 2.5%, cadence 2k. Exposure is obtained by evaluating the causal recursion on the corresponding fixed probe trajectory.

Setting	Exposure (k)	$\text{FID}_{60} \downarrow$	Copy_{60} (%) $\downarrow$
Default	4.21 ± 0.64	77.8 ± 0.3	3.88 ± 0.31
Decay T=4k	2.41 ± 0.20	78.5 ± 0.3	3.65 ± 0.30
Decay T=12k	5.17 ± 1.04	77.7 ± 0.3	4.57 ± 0.34
Feedback kappa=0.25	5.24 ± 0.33	78.0 ± 0.3	4.40 ± 0.33
Feedback kappa=0.75	3.58 ± 0.81	78.3 ± 0.3	3.84 ± 0.21
Warning=2Warning=3Probe cadence=1k	3.43 ± 0.96	77.9 ± 0.3	3.78 ± 0.24
Probe cadence=4k	4.46 ± 0.30	78.3 ± 0.3	4.79 ± 0.14

Table 19: **Paired differences in selected checkpoint quality.** Baseline final FID minus QGD final FID under CBS at a 5% copy budget. Mean $\pm$ sample SD is over the five paired trajectory differences after averaging each trajectory over its 100 calibration decisions.

Baseline	Final FID difference
SGD	3.81 ± 0.33
AdamW	6.38 ± 0.31
Muon	10.57 ± 0.51
MiMuon (norm proxy)	5.11 ± 0.19
Same-gate hard switch	2.25 ± 0.31
Pilot-fixed open loop	3.74 ± 2.89

ing probabilities are not revealed to the controller or any selector. Counts are sampled once per candidate/bank and reused across competing selection rules. Each calibration repetition records the selected checkpoint or abstention, and the reported summaries aggregate release frequencies, joint violations and final-audit readouts.

Valid and Diagnostic Rules. Empirical filtering chooses the best development FID among candidates satisfying $S_k/N \le c_{\text{budget}}$; it is a diagnostic rule without simultaneous error control. Single preselection chooses one quality-eligible candidate using development copying at the full budget; margin preselection uses $0.8c_{\text{budget}}$. Each then applies a level-δ one-sided binomial test to that candidate alone, abstaining on failure without trying another. These two rules have valid joint error control.

Fixed-sequence testing orders the eligible family using development information alone. It applies level-δ tests in that fixed order and stops at the first failure, selecting the best development FID among the consecutively certified candidates. The order affects power but not validity; stopping at the first failure is essential (Angelopoulos et al., 2025). CBS instead constructs simultaneous upper bounds with per-candidate error δ/M and chooses the best development FID among all eligible certified candidates. Ties use the earlier checkpoint.

The fixed sequence orders by ascending development-copy fraction, with earlier step breaking ties. Every method uses fresh calibration inputs independent of training and development. The guarantee is for the joint event of release and detector-budget violation. It does not bound violation conditional on release and does not certify population FID. A subsequent model revision requires fresh calibration or a separately justified sequential procedure. Adaptive Learn-then-Test addresses a complementary setting with sequential sampling and early termination (Zecchin et al., 2025).

Family Size and Uncertainty. All denominators for release and joint error include abstentions. Final FID and final audited copying average released checkpoints. The family-size comparison freezes a smaller, development-chosen candidate set before its calibration and updates the Bonferroni allocation accordingly. Repeated-bank error frequencies describe the declared fixture families; they do not replace Proposition 2's distribution-free argument.

Table 20: **Quality gates and adaptive transition schedules.** The coefficient is set using the probe at the listed zero time for the following block. Times are k steps. Every schedule obeys the 7.0416k exposure envelope.

Trajectory	Quality	Confirm.	$\alpha = 0$	Exposure	Copy onset	FID_{60}
0	8	10	20	4.450	174	77.83
1	8	10	18	3.837	170	77.38
2	10	12	22	4.685	160	78.16
3	6	8	14	3.273	184	77.67
4	8	10	20	4.805	184	77.91

Table 21: **Sensitivity to checkpoint-family size.** QGD archives at a 5% copy budget, $\delta = 0.05$. Thinned families are fixed before calibration and reuse the same candidate-level binomial draws for paired comparison. Each candidate receives 5k outputs; CBS uses δ/M.

Candidates M	Cadence (k)	Final FID	Release	Joint error	Outputs (k)
20	10	75.98 ± 0.59	500/500	0/500	100
50	4	75.75 ± 0.59	500/500	3/500	250
100	2	75.56 ± 0.51	500/500	1/500	500

L Computation and practical cost

The timing configuration uses one NVIDIA A100 accelerator, bf16 training, batch 128, and the 50-step DDIM sampler. Table 22 separates model updates, development probes, calibration and the final audit. The operation-time model uses 60/72/82/75 ms per SGD/AdamW/Muon/MiMuon update, 63 ms for the fixed-gain momentum tail and 84 ms for a mixed update. A development probe costs 90 s for generation/features plus 5 s for copying on shared samples; calibration costs 43 s per 5k outputs, and the final 50k quality audit costs 900 s. Research-only trajectory audits are additional to the deployment accounting. The copying probe reuses the generated quality samples; its additional time is detector computation, rather than a second full generation pass. Shared work is counted once.

The full cost of a released model includes constructing the archive, not just training to its selected checkpoint. QGD can skip Newton–Schulz calculations after its coefficient reaches zero, but generation probes and final calibration remain material. A smaller candidate family reduces calibration outputs at the cost of fewer choices. The selector comparison therefore reports both quality/availability and the number of generated calibration outputs.

M Transfer metrics and generation settings

Image Flow Matching. The image comparison uses the same UNet and small CIFAR-10 subset with the conditional flow-matching objective of Lipman et al. (2023). The interpolation is $x_t = (1-t)z + tx$, with $z \sim \mathcal{N}(0, I)$, uniform $t \in [0, 1]$, and velocity target $x - z$. Sampling uses 50 uniform Heun steps. The flow observation horizon is 220k updates; motion uses 300k, with the same 2k probe cadence and 5% copy threshold. Model architecture, batch, time distribution and solver are fixed across methods. Development quality uses a predeclared FID target of 89.1; final quality follows the independent image-audit protocol.

Music-Conditioned Motion. The AIST++ configuration uses a 4M-parameter pose generator and 20-second sequences. Source performances are assigned to training, development and audit groups before chunking, with 400 training, 100 development and 100 audit recording groups; overlapping chunks remain within one split. Music features and the pose representation are fixed across methods. Beat alignment compares motion beats with music beats and has a predeclared threshold of 0.270.

Motion features use root-centered, body-scale-normalized joint positions, retaining root motion separately. The 48-dimensional kinetic descriptor contains each of 23 non-root joints' mean squared

Table 22: **End-to-end cost accounting for a 200k checkpoint archive.** Hours from the operation-time model in Appendix L. Development quality generation/features cost 90s and the copy check adds 5s on shared samples. CBS uses 500k outputs; the independent final quality audit uses 50k. The full open-loop polar branch remains active because its exponential coefficient never reaches zero.

Method	Train	Develop.	CBS	Final audit	Total
SGD	3.33	2.64	1.19	0.25	7.42
AdamW	4.00	2.64	1.19	0.25	8.08
Muon	4.56	2.64	1.19	0.25	8.64
MiMuon (norm proxy)	4.17	2.64	1.19	0.25	8.25
Same-gate hard switch	3.50	2.64	1.19	0.25	7.58
Pilot-fixed open loop	4.66	2.64	1.19	0.25	8.74
QGD (ours)	3.65	2.64	1.19	0.25	7.74

Table 23: **Transfer clocks and complete common-checkpoint readouts.** Five paired trajectories per setting. Times are k steps; flow quality is FID at 60k and dance quality is beat alignment at 120k. Main-text dance realism and diversity are independent additional quality dimensions.

Method	Quality (k)	Copy (k)	Window (k)	Quality	Copy (%)
Flow matching					
SGD	16.0 ± 1.4	200.0 ± 3.7	184.0 ± 2.8	86.4 ± 0.3	1.17 ± 0.13
AdamW	10.0 ± 1.4	116.0 ± 3.7	106.0 ± 2.8	84.5 ± 0.3	2.08 ± 0.22
Muon	8.0 ± 1.4	26.0 ± 3.7	18.0 ± 2.8	86.2 ± 0.3	14.61 ± 0.07
Same-gate hard switch	8.0 ± 1.4	128.0 ± 3.7	120.0 ± 2.8	82.8 ± 0.3	2.48 ± 0.15
QGD (ours)	8.0 ± 1.4	108.0 ± 3.7	100.0 ± 2.8	79.8 ± 0.3	4.14 ± 0.25
AIST++ dance					
SGD	28.0 ± 1.4	260.0 ± 3.7	232.0 ± 2.8	0.279 ± 0.001	1.08 ± 0.10
AdamW	26.0 ± 1.4	232.0 ± 3.7	206.0 ± 2.8	0.282 ± 0.001	1.94 ± 0.13
Muon	20.0 ± 1.4	58.0 ± 3.7	38.0 ± 2.8	0.285 ± 0.001	11.06 ± 0.57
Same-gate hard switch	20.0 ± 1.4	210.0 ± 3.7	190.0 ± 2.8	0.289 ± 0.001	2.14 ± 0.22
QGD (ours)	20.0 ± 1.4	180.0 ± 3.7	160.0 ± 2.8	0.294 ± 0.001	4.09 ± 0.29

speed and acceleration (46 components), plus root horizontal and vertical kinetic energy. Components are standardized using development-pilot statistics. Kinetic feature FID compares its means and covariances and is reported separately from beat synchrony. Diversity is the average pairwise distance in the same standardized motion-feature space, normalized by the corresponding held-out reference diversity; matching the reference scale is the target, rather than increasing diversity without limit. Pose copying applies the fixed nearest-neighbor ratio $d_1 < d_2/3$ in this standardized kinetic space. Its flag rate concerns the declared coarse descriptor; it does not certify the absence of semantic or frame-level copying. Table 23 provides the per-trajectory transfer readouts. The joint evaluation prevents high beat alignment from concealing repetitive or unrealistic motion.