LARGE DIFFUSION MODELS NEED MORE THAN EMA

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale diffusion models are commonly trained with a constant learning rate (LR) and exponential moving average (EMA), despite the strong performance of LR decay and Schedule-Free optimization in training large language models. We study whether EMA is uniquely suited to diffusion training, or whether stronger alternatives have been underexplored. Through the lens of variance reduction, we theoretically show that averaging a constant-step trajectory has an inherent optimization limitation, whereas LR decay and Schedule-Free optimization admit vanishing stationarity guarantees under standard nonconvex assumptions. Empirically, LR decay and EMA provide complementary rather than substitutable benefits: decay alone does not outperform EMA with a constant LR, while combining decay with EMA further improves this standard baseline at most tested starting points, although the gain remains sensitive to when decay begins. Schedule-Free avoids this timing requirement and consistently outperforms EMA with a constant LR across training stages, achieving both lower fixed-evaluation objectives and better generation quality. Our analysis further shows that all three strategies substantially reduce parameter variability, yet stronger stabilization is not always better: EMA with a constant LR retains comparatively higher parameter variability, while aggressive decay combined with EMA achieves the lowest checkpoint variability but also weakens continued optimization. Schedule-Free substantially reduces parameter variability while preserving more late-stage progress, providing a better balance between parameter stability and continued optimization. Finally, the benefits of Schedule-Free extend beyond pretraining to ReFL and generator-side DMD, demonstrating its transferability across distinct diffusion optimization settings. Overall, our results suggest that large diffusion models need more than parameter averaging alone: effective optimization requires balancing variance reduction with continued optimization.

1 INTRODUCTION

Diffusion architectures have advanced rapidly from the convolutional U-Net backbones used in DDPMs (Ronneberger et al., 2015; Ho et al., 2020) to Diffusion Transformers such as DiT and SiT (Peebles & Xie, 2023; Ma et al., 2024), alongside improvements in training objectives and representation learning (Karras et al., 2022; Yu et al., 2025). In contrast, their optimization recipes have changed much less. Many large-scale diffusion models are still trained with a constant learning rate (LR) and evaluated using exponential moving average (EMA) weights, even though LR decay is standard in large-scale language-model pretraining (Hu et al., 2024; Wen et al., 2025) and Schedule-Free optimization has recently shown strong performance in the same regime (Defazio et al., 2024).

This discrepancy matters because improving the training recipe can reduce the compute needed to obtain a strong model without changing the architecture, objective, or data. Much recent work on accelerating diffusion training has instead focused on these other components. For example, REPA substantially accelerates Diffusion Transformer training by improving representation learning through alignment with pretrained visual features (Yu et al., 2025). By comparison, the optimization recipe itself has received less systematic attention, especially for modern large-scale diffusion-transformer pretraining. As a result, it remains unclear whether the widespread use of constant LR with EMA reflects a genuine optimization advantage or simply a largely inherited training convention.

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.

This motivates our central question: **does EMA have an optimization advantage specific to large-scale diffusion training, or have stronger alternatives such as LR decay and Schedule-Free optimization simply been underexplored?**

We study these strategies from a unified variance-reduction perspective. We argue that EMA, LR decay, and Schedule-Free can be understood as three different approaches to the same underlying problem: persistent variability in stochastic optimization. Their distinction is not simply whether they stabilize training, but how they reduce variability and how this interacts with continued optimization. EMA reduces variability by averaging a constant-step trajectory while leaving the underlying optimization dynamics unchanged. LR decay stabilizes the trajectory more directly by shrinking the updates, but this simultaneously reduces the magnitude of further optimization steps. Schedule-Free instead combines continued large-step optimization with averaging over an increasingly long trajectory, reducing variability without suppressing the underlying optimization dynamics as aggressively.

Our theoretical analysis places these three strategies within a common framework, clarifying how each reduces variability and exposing their different trade-offs between trajectory stability and continued optimization. In particular, averaging a constant-step trajectory has an inherent optimization limitation, whereas LR decay and Schedule-Free admit vanishing stationarity guarantees under standard nonconvex assumptions. This unified view then motivates our empirical question: which trade-off is most effective for large-scale diffusion training?

We therefore turn to the empirical consequences of this trade-off in large-scale diffusion training. Using 675M-parameter SiT-XL/2 + REPA pretraining, we compare CONSTANT LR + EMA, LR DECAY + EMA, and SCHEDULE-FREE under controlled training branches, asking how each strategy affects generation quality, optimization progress, and parameter variability.

Three findings emerge. First, LR decay and EMA are complementary rather than interchangeable: decay alone does not surpass CONSTANT LR + EMA, but DECAY + EMA can further improve the standard recipe, with gains that depend on when decay begins. Second, Schedule-Free provides more consistent improvements, outperforming CONSTANT LR + EMA across training stages without requiring a prespecified decay point. Third, the strongest stabilization is not the best optimization outcome: DECAY + EMA achieves the lowest parameter variability but weakens late-stage progress, whereas Schedule-Free remains stable at a larger LR scale and continues to make optimization progress. This allows it to strike a better balance between stability and continued optimization.

Taken together, these results suggest that EMA is highly effective but not uniquely suited to large-scale diffusion optimization. The broader lesson is that reducing trajectory variability is important but insufficient: effective optimization must also preserve continued progress. This perspective helps explain why Schedule-Free performs well in our main pretraining setting and why its benefits extend to ReFL (Xu et al., 2023) and generator-side DMD (Yin et al., 2024; Jiang et al., 2025).

2 PRELIMINARIES

2.1 WHY CONSTANT-STEP SGD HAS A NOISE FLOOR?

To see why variance reduction is necessary, consider the simplest one-dimensional quadratic $L(x) = \frac{1}{2}dx^2$ with $d > 0$, and noisy gradients $g_t = dx_t + \xi_t$, where $\xi_t \sim \mathcal{N}(0, \sigma^2)$ are independent. We use this simple quadratic example to make the mechanism explicit; Lemma 2.1 below gives the corresponding result for general smooth nonconvex objectives.

Constant-step SGD follows

$$x_{t+1} := x_t - \eta g_t = \underbrace{(1 - \eta d)x_t}_{\text{contraction toward } 0} - \underbrace{\eta \xi_t}_{\text{fresh noise}} . \tag{1}$$

The two terms have opposite effects. Without noise, $0 < \eta d < 2$ makes the iterates converge to the optimum $x^\star = 0$. With noise, however, even at the optimum we have $x_{t+1} = -\eta \xi_t \neq 0$ in general. Thus each step contracts the previous error while injecting fresh noise. **Constant-step SGD therefore does not settle at the optimum; it fluctuates around it.**

Because old perturbations are geometrically contracted while fresh noise is injected at every step,

$$x_t = (1-\eta d)^t x_0 - \eta \sum_{k=0}^{t-1} (1-\eta d)^{t-1-k} \xi_k, \qquad x_t \xrightarrow{d} \mathcal{N}\left(0, \frac{\eta\sigma^2}{d(2-\eta d)}\right), \tag{2}$$

for $0 < \eta d < 2$. Thus, the initialization vanishes, but the accumulated gradient noise does not: stable constant-step SGD converges to a distribution around the optimum rather than a deterministic point.

On the quadratic objective, this stationary variance directly produces excess loss, as

$$\mathbb{E}[L(x_\infty)] = \frac{d}{2}\,\mathrm{Var}(x_\infty) = \frac{\eta\sigma^2}{2(2-\eta d)}. \tag{3}$$

Both the parameter variance and excess loss increase with the learning rate and grow rapidly as ηd approaches the stability boundary 2. **Variance reduction therefore aims to lower this late-stage noise floor without unnecessarily suppressing further optimization progress.**

The following standard smooth nonconvex SGD bound makes the same noise-floor phenomenon explicit beyond the quadratic example (cf. Ghadimi & Lan, 2013; Bottou et al., 2018).

Lemma 2.1 (Noise floor of constant-step SGD)**.** *Under the standard smooth nonconvex assumptions in Appendix B.1.2, constant-step SGD with learning rate η satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\big[\|\nabla f(x_t)\|^2\big] \le \underbrace{\frac{2(f(x_0)-f_\star)}{\eta T}}_{\textit{optimization error}} + \underbrace{L\eta\sigma^2}_{\textit{noise floor}}. \tag{4}$$

The optimization error vanishes as T grows, but the noise floor remains under a fixed learning rate. Thus, longer training alone cannot eliminate the residual error induced by stochastic gradients.

2.2 EMA Averages a Noisy Trajectory

EMA does not change the online optimization dynamics. Instead, it forms a smoother evaluation point by averaging past iterates

$$x_t^{\mathrm{EMA}} := \beta x_{t-1}^{\mathrm{EMA}} + (1-\beta)x_t, \qquad 0 < \beta < 1. \tag{5}$$

For Eq. (1), let $b := 1-\eta d$ and define $m_t := \beta^t x_0^{\mathrm{EMA}} + (1-\beta)\sum_{j=1}^{t} \beta^{t-j} b^j x_0$ and $w_{k,t} := \sum_{j=k+1}^{t} \beta^{t-j} b^{j-1-k}$. Expanding the two recursions gives the exact EMA iterate

$$x_t^{\mathrm{EMA}} = \underbrace{m_t}_{\text{mean trajectory / optimization process}} - \underbrace{\eta(1-\beta)\sum_{k=0}^{t-1} w_{k,t}\xi_k}_{\text{stochastic fluctuation / variance}}. \tag{6}$$

Thus, the online trajectory continues to fluctuate as before, while EMA averages out part of this fluctuation. For the quadratic model in Section 2.1, the stationary variance is reduced by

$$\frac{\mathrm{Var}(x_\infty^{\mathrm{EMA}})}{\mathrm{Var}(x_\infty)} = \frac{(1-\beta)(1+\beta b)}{(1+\beta)(1-\beta b)} < 1, \qquad b = 1-\eta d, \tag{7}$$

with the derivation deferred to Appendix B.2.

Variance reduction, however, does not imply that EMA corrects the optimization bias of the underlying constant-step trajectory.

Lemma 2.2 (EMA does not in general remove constant-step optimization bias)**.** *Consider the smooth strongly convex objective*

$$f(x) = \frac{x^2}{2} + a(\sin x - x), \qquad 0 < a < 1.$$

Its unique minimizer is $x^\star = 0$. For constant-step SGD $X_{t+1} = X_t - h\{f'(X_t) + \xi_t\}$ with $\xi_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma^2)$, $\sigma^2 > 0$, and $0 < h \le 1/(1+a)$, let U follow its invariant distribution. Its mean satisfies

$$m_h := \mathbb{E}[U] = a\big(1 - \mathbb{E}[\cos U]\big) > 0 = x^\star.$$

Moreover, if $Y_t = \beta Y_{t-1} + (1-\beta)X_t$ *is the* EMA *of a stationary trajectory, then* $\mathbb{E}[Y_t] \to m_h$. *Since* f *is* $(1-a)$*-strongly convex,*

$$\mathbb{E}[f(Y_t)] - f(x^\star) \geq f(\mathbb{E}[Y_t]) - f(x^\star) \;\longrightarrow\; f(m_h) - f(x^\star) \geq \frac{1-a}{2} m_h^2 > 0.$$

Thus, EMA can reduce fluctuations around the stationary mean, but averaging alone does not in general eliminate the residual optimization error of the underlying constant-step trajectory.

The proof is given in Appendix B.2.2. Hence, EMA improves the stability of the *evaluation weights* without changing the optimization dynamics that generate them.

2.3 Learning-Rate Decay Shrinks the Noisy Trajectory

Unlike EMA, LR decay changes the online optimization dynamics themselves. With a time-varying learning rate η_t, the quadratic model in Section 2.1 becomes

$$x_{t+1} := (1 - \eta_t d)x_t - \eta_t \xi_t. \tag{8}$$

Both the deterministic update and the injected noise are now scaled by η_t. In particular, the fresh noise has variance $\eta_t^2\sigma^2$, which vanishes as $\eta_t \to 0$. Thus, LR decay reduces variance by shrinking the trajectory itself rather than averaging a noisy trajectory after the fact.

For example, consider the analytically tractable schedule $\eta_t = \eta/(2 + t\eta d)$ with $0 < \eta d < 2$. For every $t \geq 1$, the iterate has the exact form

$$x_t = \underbrace{\frac{2-\eta d}{2+(t-1)\eta d} x_0}_{\text{mean trajectory / optimization process}} - \underbrace{\frac{\eta}{2+(t-1)\eta d} \sum_{k=0}^{t-1} \xi_k}_{\text{stochastic fluctuation / variance}}. \tag{9}$$

Consequently,

$$\mathbb{E}[x_t] = \frac{2-\eta d}{2+(t-1)\eta d} x_0 = O(t^{-1}), \qquad \mathrm{Var}(x_t) = \frac{t\eta^2\sigma^2}{\big(2+(t-1)\eta d\big)^2} = O(t^{-1}). \tag{10}$$

Hence both the optimization error and parameter variance vanish asymptotically. The same mechanism also reveals the central trade-off: reducing η_t suppresses stochastic fluctuations, but simultaneously reduces the magnitude of further optimization steps.

The following standard LR-decay SGD guarantee formalizes this behavior; cf. Ghadimi & Lan (2013); Bottou et al. (2018).

Lemma 2.3 (Nonconvex convergence under LR decay)**.** *Under the standard smooth nonconvex assumptions in Appendix B.3.2, SGD with learning rates* $\{\eta_t\}$ *satisfies*

$$\frac{\sum_{t=0}^{T-1} \eta_t \,\mathbb{E}\big[\|\nabla f(x_t)\|^2\big]}{\sum_{t=0}^{T-1} \eta_t} \leq \frac{2\big(f(x_0) - f_\star\big)}{\sum_{t=0}^{T-1} \eta_t} + L\sigma^2 \frac{\sum_{t=0}^{T-1} \eta_t^2}{\sum_{t=0}^{T-1} \eta_t}. \tag{11}$$

If $\sum_t \eta_t = \infty$ *and* $\sum_t \eta_t^2 < \infty$*, both terms vanish, yielding convergence to stationarity.*

The inverse-time schedule above is used only to make the mechanism explicit. In practice, diffusion training often applies decay only near the end of training; our experiments therefore study finite late-stage cosine decays and their sensitivity to the decay starting point.

2.4 Schedule-Free Combines Trajectory Averaging with Large-Step Optimization

LR decay reduces variance by shrinking the optimization steps themselves. Schedule-Free instead keeps a constant base learning rate while averaging the optimization trajectory over an increasingly long window (Defazio et al., 2024). In its SGD form, it maintains

$$y_t := (1-\beta)z_t + \beta x_t, \tag{12}$$

$$z_{t+1} := z_t - \gamma g(y_t, \xi_t), \tag{13}$$

$$x_{t+1} := \left(1 - \frac{1}{t+1}\right) x_t + \frac{1}{t+1} z_{t+1}, \tag{14}$$

where z_t is the optimization sequence and x_t is the averaged sequence used for evaluation. The key distinction is immediate: z_t continues to take constant-scale optimization steps, while x_t averages over a growing portion of the trajectory.

For the quadratic model in Section 2.1, if $a := 1 - \gamma d(1-\beta) \in (0,1)$, then

$$\mathbb{E}[x_t] = O(t^{-1}), \qquad \mathrm{Var}(x_t) = O(t^{-1}), \qquad \mathbb{E}[L(x_t)] = O(t^{-1}), \tag{15}$$

with the derivation given in Appendix B.4. Thus, the averaged sequence becomes increasingly stable even though the optimization sequence continues to use a constant base learning rate.

The following result is restated from Chae & Kim (2026, Theorem 4.4), with minor notation changes for consistency with our presentation.

Lemma 2.4 (Nonconvex convergence of Schedule-Free SGD)**.** *Under the standard smooth nonconvex assumptions stated in Appendix B.4.2,* Schedule-Free *SGD admits a vanishing stationarity guarantee for its gradient-evaluation sequence.*

The precise bound and assumptions are given in Appendix B.4.2. Conceptually, Schedule-Free modifies this trade-off differently from LR decay: it reduces variability through trajectory averaging while retaining large-step optimization dynamics.

3 Variance Reduction in Diffusion Models

3.1 Experimental Setup: Modern Large-Scale Diffusion Pretraining

We evaluate whether the distinctions in Section 2 persist in a modern large-scale diffusion-training regime. Our main setting uses a class-conditional SiT-XL/2 model with approximately 675M parameters (Ma et al., 2024), trained on ImageNet-1K at 256×256 resolution (Deng et al., 2009) with flow matching (Lipman et al., 2023) and REPA (Yu et al., 2025). This is a strong modern diffusion-transformer recipe rather than a deliberately weak baseline. Our Constant LR + EMA baseline follows the standard REPA optimization recipe, with the same optimizer, learning rate, and EMA configuration; our main deviation is the smaller global batch size used because of computational constraints. Across all main comparisons, the model, data, objective, batch size, and evaluation protocol are fixed; only the optimization recipe is changed.

We compare three strategies:

- **Constant LR + EMA:** AdamW (Loshchilov & Hutter, 2019) with a constant learning rate of 1×10^{-4} and EMA weights for evaluation.
- **LR Decay + EMA:** independent continuations from matched late-stage checkpoints of the Constant LR + EMA trajectory, each using the same 40k-step cosine decay while retaining EMA. This isolates the effect of introducing late-stage decay and allows us to test sensitivity to its starting point.
- **Schedule-Free AdamW:** Schedule-Free AdamW (Defazio et al., 2024) with a base learning rate of 4×10^{-4}.

The learning rates are optimizer-specific rather than numerically matched, since Schedule-Free operates in a different effective LR regime from standard AdamW; the corresponding LR controls are reported in Appendix A.4.

The main trajectories are trained for up to 800k optimization steps with global batch size 128. We evaluate each method at three levels: parameter variability, optimization quality, and generation quality. Parameter variability measures how strongly the trajectory fluctuates; optimization quality is measured by the REPA objective and fixed-evaluation losses; and generation quality is measured by FID (Heusel et al., 2017), sFID (Nash et al., 2021), and Inception Score (Salimans et al., 2016). Together, these measurements distinguish *reducing parameter fluctuations* from *actually obtaining a better optimized and better generating model*. Full training and evaluation details are provided in Appendix A.1.

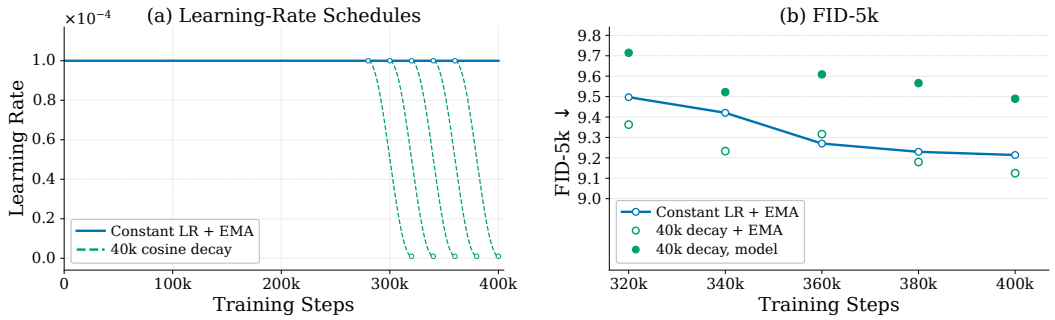

Figure 1: Late-stage cosine decay is beneficial but timing-sensitive. (a) Five independent 40k-step cosine-decay branches from the same CONSTANT LR + EMA trajectory. (b) Three-seed mean FID-5K for the matched baseline, decayed online model, and DECAY + EMA. EMA further improves all five decay endpoints, while DECAY + EMA outperforms the matched baseline at four of five starting points.

3.2 LR DECAY CAN FURTHER IMPROVE CONSTANT LR + EMA, BUT THE GAIN IS TIMING-SENSITIVE

To test whether late-stage LR decay can improve CONSTANT LR + EMA, we independently branch from several late-stage checkpoints of the same trajectory and apply the same 40k-step cosine schedule to each branch. The learning rate decays from 10^{-4} to 10^{-6}, while all other training settings remain unchanged. We evaluate each endpoint using FID-5K with three independent sampling seeds and report the mean.

The benefit of decay is timing-sensitive. Figure 1 shows that adding the same 40k-step cosine decay at different late-stage checkpoints leads to noticeably different outcomes. DECAY + EMA improves FID-5K at four of the five tested starting points, but slightly hurts at one, indicating that both the magnitude and even the sign of the gain depend on when decay begins. Full FID-5K, IS, and sFID results are reported in Appendix C.2 (Table 7).

LR decay and EMA provide complementary gains. The new comparison also shows that LR decay and EMA are not competing alternatives. Across all five decay branches, applying EMA to the decayed trajectory consistently improves FID-5K over using the corresponding online model alone. For example, at the 360k→400k branch, FID-5K improves from 9.4895 without EMA to 9.1248 with EMA. This suggests that explicit LR decay and parameter averaging provide complementary benefits and can be combined effectively.

Thus, late-stage LR decay can improve the standard recipe, but its benefit depends on when annealing begins. We next ask whether Schedule-Free can provide more consistent gains without this timing requirement.

3.3 SCHEDULE-FREE CONSISTENTLY IMPROVES OVER CONSTANT LR + EMA

Schedule-Free consistently improves generation quality. As shown in Figure 2(a), SCHEDULE-FREE ADAMW achieves lower mean FID-5K than CONSTANT LR + EMA at every evaluated checkpoint from 320k to 400k steps, improving FID-5K from 9.2139 to 9.0504 at 400k. While the baseline begins to plateau late in training, Schedule-Free continues to improve, indicating that its advantage is consistent across training stages rather than tied to a single checkpoint.

Schedule-Free also reaches a lower optimization objective. Better FID alone does not establish that Schedule-Free reaches a better optimization objective, since parameter averaging could improve sampling quality without reducing the training loss. We therefore evaluate all methods using the same fixed data, noise, class conditions, and timesteps. Under this controlled protocol, Schedule-Free achieves a lower total fixed-evaluation loss throughout the evaluated interval (Figure 2(b–d)).

Decomposing the objective shows that the denoising MSE remains close between the two methods, while Schedule-Free has a consistently lower projection loss, which accounts for most of the total-loss advantage. To check that the aggregate denoising result does not hide regressions at particular noise

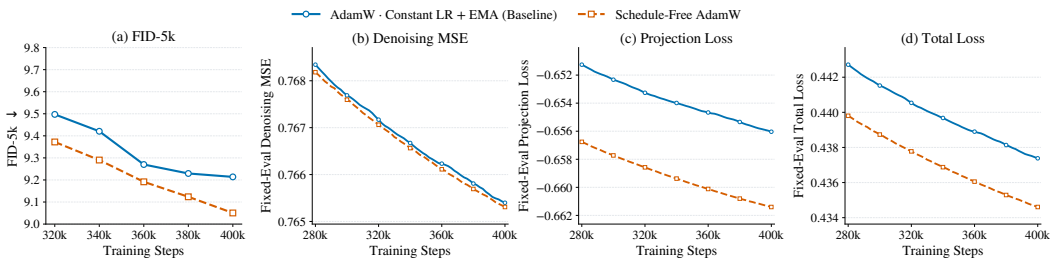

Figure 2: Generation quality and fixed-evaluation objectives for CONSTANT LR + EMA and SCHEDULE-FREE ADAMW. (a) Three-seed mean FID-5K: Schedule-Free is better at every checkpoint from 320k to 400k. (b–d) Fixed-evaluation denoising MSE, projection loss, and total loss: denoising MSE is similar, while Schedule-Free has lower projection and total losses.

levels, we further evaluate fixed timesteps in Appendix C.3. Schedule-Free achieves lower denoising MSE at all eight tested timesteps, as well as under random-timestep evaluation.

A larger learning rate alone does not explain the gain. To control for Schedule-Free's larger base learning rate, we train standard AdamW with the same 5k-step warmup and a constant learning rate of 4×10^{-4}. This run becomes unstable and produces sustained NaNs after approximately 218.7k steps, whereas Schedule-Free remains stable at the same learning-rate scale. Thus, the improvement cannot be explained by simply using a larger learning rate with standard AdamW (Appendix A.4).

The advantage persists under longer training. At 800k steps, SCHEDULE-FREE improves FID-50K from 2.0770 to 1.8735, sFID from 4.539 to 4.479, and Inception Score from 265.6 to 272.7 relative to CONSTANT LR + EMA. Selected samples are shown in Appendix E.1 (Figure 10).

3.4 SCHEDULE-FREE OFFERS A BETTER TRADE-OFF THAN CONSTANT LR + EMA AND DECAY + EMA

Reducing parameter variability helps, but minimizing it is not enough. Removing EMA from the CONSTANT LR + EMA baseline substantially degrades FID-5K in our setting. As shown in Appendix C.2 (Table 7), the degradation is approximately 1.8–4.4 points across the evaluated checkpoints. In contrast, CONSTANT LR + EMA, DECAY + EMA, and SCHEDULE-FREE ADAMW all substantially reduce parameter variability and achieve much better generation quality. Thus, the relevant question is how to reduce parameter variability without unnecessarily suppressing further optimization progress.

Lower parameter variability does not monotonically improve generation quality. Figure 3 makes this mismatch explicit. At 400k steps, DECAY + EMA has the lowest adjacent-checkpoint variability among the three main strategies, followed by SCHEDULE-FREE ADAMW and then CONSTANT LR + EMA. Yet the FID-5K ordering is different: SCHEDULE-FREE ADAMW performs best at 9.0504, compared with 9.1248 for DECAY + EMA and 9.2139 for CONSTANT LR + EMA. The same broader pattern holds when the online decay trajectory is included (Appendix A.3). Thus, stronger trajectory stabilization is beneficial up to a point, but minimizing parameter variability does not by itself produce the best generative model.

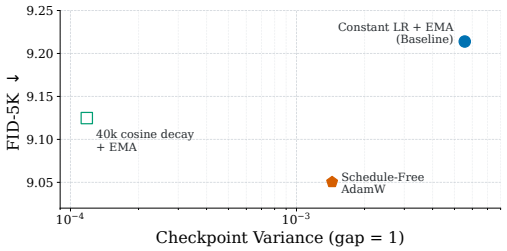

Figure 3: Late-stage checkpoint variability versus FID-5K at 400k. Variability is the trace variance between consecutive checkpoints ($g = 1$). DECAY + EMA is most stable, whereas SCHEDULE-FREE ADAMW achieves the best FID-5K.

Schedule-Free remains highly stable even at a larger learning-rate scale. Notably, the low checkpoint variability of SCHEDULE-FREE ADAMW in Figure 3 is achieved while using a $4\times$ larger base learning rate than the CONSTANT LR + EMA baseline. Despite this more aggressive learning-rate scale, SCHEDULE-FREE ADAMW still exhibits substantially lower parameter variability than the baseline. Together with the stability control in Appendix A.4, where standard AdamW becomes

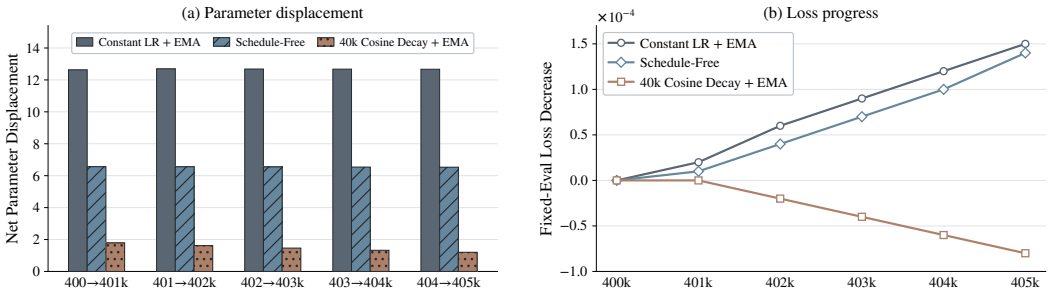

Figure 4: Late-stage parameter movement and optimization progress. (a) Net ℓ_2 displacement over five adjacent 1k-step intervals. (b) Fixed-evaluation loss decrease relative to 400k. DECAY + EMA minimizes displacement but does not improve the objective, whereas SCHEDULE-FREE ADAMW retains continued progress. The comparison uses EMA sequences for constant LR and decay and the averaged x sequence for Schedule-Free; full results are in Appendix A.3.

unstable at the same learning-rate scale, this suggests that the strong stabilization of SCHEDULE-FREE may permit a wider range of stable learning rates. One possible interpretation is that this additional stability margin gives SCHEDULE-FREE more room to operate at a larger learning-rate scale, helping preserve continued optimization progress without relying on progressively smaller updates for stability.

EMA remains beneficial across training schedules. EMA improves generation quality under both constant and decaying learning rates. Even after cosine decay stabilizes the online trajectory, EMA further reduces checkpoint variability and improves FID-5K across all tested branches (Appendix A.3; Table 7), confirming that parameter averaging and LR scheduling are complementary rather than interchangeable.

Aggressive reduction in parameter variability also suppresses optimization mobility. The mechanism in Section 2 suggests why stronger trajectory stabilization can incur a cost. LR decay scales both the stochastic noise and the deterministic optimization update by η_t; as η_t becomes small, the trajectory becomes more stable but also moves less. Figure 4(a) shows the corresponding late-stage behavior: across five consecutive 1k-step windows from 400k to 405k, the post-decay EMA trajectory has by far the smallest net parameter displacement, while SCHEDULE-FREE retains substantially more movement and CONSTANT LR + EMA moves the most. Thus, the strongest stabilization is also accompanied by the weakest late-stage parameter mobility. Because displacement alone cannot distinguish useful optimization progress from random motion, we next ask whether this movement actually lowers the objective.

Schedule-Free better preserves useful late-stage optimization. Figure 4(b) distinguishes useful optimization progress from parameter movement alone. From 400k to 405k, fixed-evaluation loss continues to decrease for both CONSTANT LR + EMA and SCHEDULE-FREE, whereas the post-decay EMA run no longer improves and instead increases slightly. Notably, CONSTANT LR + EMA also maintains a higher fixed-evaluation objective than SCHEDULE-FREE ADAMW, echoing the limitation in Section 2.2 that averaging a constant-step trajectory can reduce parameter variability without in general eliminating residual optimization error. Taken together, Figures 3 and 4 show the full trade-off: DECAY + EMA produces the strongest stabilization and the least parameter movement, but also the weakest subsequent objective progress, whereas SCHEDULE-FREE ADAMW substantially reduces variability while continuing to optimize. In our setting, Schedule-Free provides the best empirical balance between these effects.

3.5 BROADER VALIDATION ACROSS DIFFUSION TRAINING REGIMES

We next test whether the Schedule-Free gains transfer beyond SiT+REPA pretraining to two distinct diffusion training regimes.

ReFL. Following the DiffusionOPSD ReFL setup (Xu et al., 2023; Zhou et al., 2026), we replace AdamW with SCHEDULE-FREE ADAMW while keeping the remaining protocol fixed. AdamW uses the reference LR 1×10^{-5}, and SCHEDULE-FREE ADAMW is evaluated at $4\times$ and $20\times$ this baseline rate. On the directly optimized HPSv3 reward, SCHEDULE-FREE ADAMW improves the AdamW score from 6.3500 to 6.8626 at $4\times$ LR and to 9.6134 at $20\times$ LR. Cross-reward evaluation is mixed:

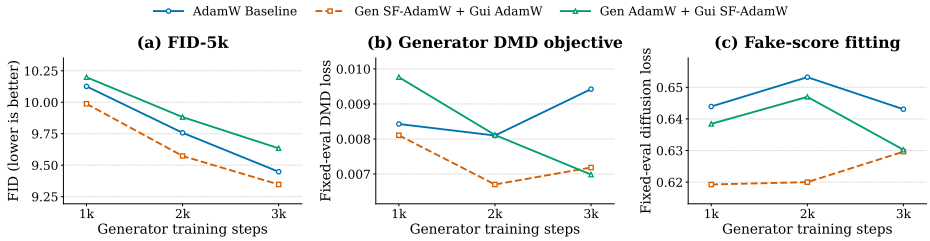

Figure 5: DMD optimizer ablation comparing all-AdamW with Schedule-Free applied only to the generator (Gen) or guidance model (Gui). (a) FID-5K: generator-side Schedule-Free improves all checkpoints, whereas guidance-side Schedule-Free does not. (b) Generator DMD objective $\mathcal{L}_{\text{DMD}}$. (c) Fake-score diffusion loss $\mathcal{L}_{\text{diff}}$; definitions are in Appendix A.6.

several orthogonal metrics improve, while others change only slightly or degrade. Complete results are reported in Appendix A.5 (Table 5).

DMD. DMD (Yin et al., 2024) jointly optimizes a generator and a guidance model, with guidance continually tracking the evolving generator distribution. Because the appropriate stopping region is unknown in advance, explicit decay must choose its starting point beforehand, while a fixed EMA timescale requires tuning to the evolving dynamics. Schedule-Free requires no prespecified decay point and continuously provides averaged evaluation parameters, making it natural for this coupled setting.

We compare the all-AdamW one-step DMDR baseline (Jiang et al., 2025) with two variants that apply SCHEDULE-FREE ADAMW only to the generator or guidance. The replaced component uses $2\times$ its baseline learning rate with a 200-step warmup; all other settings are unchanged (Appendix A.6).

Generator-side Schedule-Free improves FID-5K at every evaluated checkpoint and lowers both $\mathcal{L}_{\text{DMD}}$ and $\mathcal{L}_{\text{diff}}$. We hypothesize that Schedule-Free stabilizes the generator optimization trajectory, causing its distribution to evolve more smoothly and thereby presenting the fake-score model with an easier moving target to match. The lower generator objective and fake-score fitting loss in Figure 5(b–c) are consistent with, but do not establish, this explanation.

The guidance model plays a tracking role, adapting to the evolving fake distribution rather than directly determining the final evaluation weights. Improving guidance optimization does not necessarily translate into better generator samples. Consistent with this distinction, applying SCHEDULE-FREE only to guidance yields worse FID-5K than the all-AdamW baseline at all three checkpoints. Complete generation metrics are reported in Appendix A.6.

4 CONCLUSION AND PRACTICAL RECOMMENDATIONS

We study CONSTANT LR + EMA, LR decay, and SCHEDULE-FREE through a unified variance-reduction perspective, revealing different trade-offs between parameter stability and continued optimization. EMA stabilizes through averaging, LR decay through smaller updates, and SCHEDULE-FREE combines strong stabilization with continued optimization without requiring a prespecified decay point. These differences lead to the following practical recommendations.

When the decay point is uncertain, prefer SCHEDULE-FREE ADAMW. SCHEDULE-FREE avoids the need to choose an annealing point in advance. Its strong stability also makes it worthwhile to explore larger base learning rates than those typically used with CONSTANT LR + EMA, provided training remains stable. In our main setting, 4×10^{-4} with 5k-step warmup works well, although other regimes may require retuning.

When the training endpoint is fixed, use decay as a finishing strategy. Retaining EMA while annealing the learning rate near the end of training can further improve the standard CONSTANT LR + EMA recipe. Because the benefit depends on when annealing begins, decay is most useful when the remaining training budget is known in advance.

Choose for both stability and continued optimization. Do not select an optimization strategy solely for the lowest parameter variability. Prefer a recipe that reduces trajectory fluctuations while still preserving sufficient late-stage optimization progress.

AI USE STATEMENT

Generative AI was used for language polishing and stylistic refinement of the manuscript.

REFERENCES

Arman Bolatov, Egor Shulgin, David Li, Abduragim Shtanchaev, Sebastian U. Stich, Maxim Panov, Eric Moulines, Peter Richtárik, and Martin Takáč. Optimizers for diffusion models: A controlled benchmark. *arXiv preprint arXiv:2609.23055*, 2026. doi: 10.48550/arXiv.2609.23055. URL https://arxiv.org/abs/2609.23055.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. doi: 10.1137/16M1080173.

Shuo Cao, Jiayang Li, Xiaohui Li, Yuandong Pu, Kaiwen Zhu, Yuanting Gao, Siqi Luo, Yi Xin, Qi Qin, Yu Zhou, Xiangyu Chen, Wenlong Zhang, Bin Fu, Yu Qiao, and Yihao Liu. UniPercept: Towards unified perceptual-level image understanding across aesthetics, quality, structure, and texture. *arXiv preprint arXiv:2512.21675*, 2025. doi: 10.48550/arXiv.2512.21675. URL https://arxiv.org/abs/2512.21675.

Jiseok Chae and Donghwan Kim. Understanding schedule-free methods in nonconvex optimization: Rate guarantees and escaping saddles. *arXiv preprint arXiv:2607.09167*, 2026. doi: 10.48550/arXiv.2607.09167. URL https://arxiv.org/abs/2607.09167.

Aaron Defazio, Xingyu Alice Yang, Harsh Mehta, Konstantin Mishchenko, Ahmed Khaled, and Ashok Cutkosky. The road less scheduled. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0320.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. doi: 10.1137/120880811.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021.

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*, volume 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. MiniCPM: Unveiling the potential of small language models with scalable training strategies. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=3X2L2TFr0f.

Dengyang Jiang, Dongyang Liu, Zanyi Wang, Qilong Wu, Xin Jin, David Liu, Zhen Li, Mengmeng Wang, Peng Gao, and Harry Yang. Distribution matching distillation meets reinforcement learning. *arXiv preprint arXiv:2511.13649*, 2025. doi: 10.48550/arXiv.2511.13649. URL https://arxiv.org/abs/2511.13649.

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*, volume 35, 2022.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-Pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 36652–36663, 2023.

Hao Li, Yang Zou, Ying Wang, Orchid Majumder, Yusheng Xie, R. Manmatha, Ashwin Swaminathan, Zhuowen Tu, Stefano Ermon, and Stefano Soatto. On the scalability of diffusion-based text-to-image generation. *arXiv preprint arXiv:2404.02883*, 2024. doi: 10.48550/arXiv.2404.02883. URL https://arxiv.org/abs/2404.02883.

Zhengyang Liang, Hao He, Ceyuan Yang, and Bo Dai. Scaling laws for diffusion transformers. *arXiv preprint arXiv:2410.08184*, 2024. doi: 10.48550/arXiv.2410.08184. URL https://arxiv.org/abs/2410.08184.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Yijun Liu, Jie Huang, Zeyue Xue, Yuming Li, Ruizhe He, Haoran Li, Shijia Ge, and Siming Fu. HPSv3++: Scaling reward models across the full spectrum of diffusion model capabilities. *arXiv preprint arXiv:2606.14657*, 2026. doi: 10.48550/arXiv.2606.14657. URL https://arxiv.org/abs/2606.14657.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40, 2024. doi: 10.1007/978-3-031-72980-5_2.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. HPSv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15086–15095, 2025.

Charlie Nash, Jacob Menick, Sander Dieleman, and Peter Battaglia. Generating images with sparse representations. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7958–7968, 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

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

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Christoph Schuhmann and Romain Beaumont. LAION-Aesthetics. *LAION.AI*, 3, 2022.

Kaiyue Wen, Zhiyuan Li, Jason S. Wang, David Leo Wright Hall, Percy Liang, and Tengyu Ma. Understanding warmup-stable-decay learning rates: A river valley loss landscape view. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=m51BgoqvbP.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. doi: 10.48550/arXiv.2306.09341. URL https://arxiv.org/abs/2306.09341.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0700.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6613–6623, 2024.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025.

Z-Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, Xin Jin, Liangchen Li, Zhen Li, Zhong-Yu Li, David Liu, Dongyang Liu, Qilong Wu, Feng Yu, Zechao Zhan, Chi Zhang, Shifeng Zhang, Ruikai Zhou, and Shilin Zhou. Z-Image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025. doi: 10.48550/arXiv.2511.22699. URL https://arxiv.org/abs/2511.22699.

Wei Zhou, Xiongwei Zhu, Lingdong Kong, Bo Chen, Lei Zhang, Yongyuan Liang, Xiaoxia Hou, Ye Tian, Xian Sun, Yingshuo Wang, Linfeng Li, Shengqiong Wu, Leigang Qu, Feng Li, Wei Liu, Julian McAuley, and Tat-Seng Chua. On-policy self-distillation in diffusion models. *arXiv preprint arXiv:2608.24646*, 2026. doi: 10.48550/arXiv.2608.24646. URL https://arxiv.org/abs/2608.24646.

A Experimental Details

A.1 Main Pretraining Configuration

All main experiments use the approximately 675M-parameter SiT-XL/2 (Ma et al., 2024) for class-conditional 256×256 ImageNet-1K generation (Deng et al., 2009) with a flow-matching + REPA objective (Lipman et al., 2023; Yu et al., 2025). The main training trajectories run for up to 800k optimization steps. REPA uses DINOv2 ViT-B features (Oquab et al., 2024) and projection-loss coefficient 0.5. We use no weight decay, global batch size 128, bf16 precision, and gradient clipping with maximum norm 1.0. The dataset, architecture, objective, batch size, precision, and regularization settings are shared across the main comparisons; only the optimizer, learning rate and schedule, and EMA usage differ.

Training objective. For a clean image x, noise ϵ, diffusion timestep t, and class condition y, the interpolated input and its flow-matching velocity target are

$$x_t = \alpha_t x + \sigma_t \epsilon, \qquad v_t^* = \dot{\alpha}_t x + \dot{\sigma}_t \epsilon, \tag{16}$$

where α_t and σ_t are the signal and noise coefficients of the interpolation path, and the dots denote derivatives with respect to t. The model $v_\theta(x_t, t, y)$ predicts this target velocity. The REPA objective contains two components:

$$\mathcal{L}_{\mathrm{MSE}} = \mathrm{mean}\left[\|v_\theta(x_t, t, y) - v_t^*\|_2^2\right], \tag{17}$$

$$\mathcal{L}_{\mathrm{proj}} = -\frac{1}{K}\sum_{k=1}^{K} \mathrm{mean}\left[\left\langle \frac{z_k}{\|z_k\|_2}, \frac{\tilde{z}_k}{\|\tilde{z}_k\|_2} \right\rangle\right]. \tag{18}$$

Here, the mean in $\mathcal{L}_{\mathrm{MSE}}$ averages the squared velocity-prediction error over the evaluated samples and tensor elements. In $\mathcal{L}_{\mathrm{proj}}$, K is the number of selected frozen-encoder feature layers, z_k is the target feature from the kth selected DINOv2 layer, and $\tilde{z}_k$ is the corresponding model-projector output. The inner product is computed after ℓ_2 normalization, so this term is the negative mean cosine similarity. Consequently, lower $\mathcal{L}_{\mathrm{MSE}}$ indicates more accurate flow prediction, while lower $\mathcal{L}_{\mathrm{proj}}$ indicates better representation alignment.

With $\lambda_{\mathrm{proj}} = 0.5$, the total objective reported in our loss comparisons is

$$\mathcal{L}_{\mathrm{fixed}} = \mathcal{L}_{\mathrm{MSE}} + \lambda_{\mathrm{proj}}\mathcal{L}_{\mathrm{proj}} = \mathcal{L}_{\mathrm{MSE}} + 0.5\mathcal{L}_{\mathrm{proj}}. \tag{19}$$

For fixed-evaluation comparisons, all methods are evaluated using the same data examples, sampled noise, class conditions, and timesteps. Holding these quantities fixed reduces evaluation randomness, so differences in $\mathcal{L}_{\mathrm{MSE}}$, $\mathcal{L}_{\mathrm{proj}}$, and $\mathcal{L}_{\mathrm{fixed}}$ reflect differences between the evaluated model parameters rather than different evaluation draws.

Table 1: Main pretraining optimizer configurations.

Method	Optimizer	Base LR	Warmup	LR schedule	EMA
Constant LR + EMA	AdamW	1e-4	None	Constant	0.9999
Decay + EMA	AdamW	1e-4	None	cosine to 1e-6	0.9999
Schedule-Free	SF-AdamW	4e-4	5k linear	Schedule-Free	None

A.2 Learning-Rate Schedules

All decay branches share the same trajectory before branch step t_s. For $t_s \le t \le t_s + T$,

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_0 - \eta_{\min})\left[1 + \cos\left(\pi \frac{t - t_s}{T}\right)\right], \tag{20}$$

where $\eta_0 = 10^{-4}$, $\eta_{\min} = 10^{-6}$, and $T = 40$k. Main generation results evaluate the endpoint θ_{t_s+T} directly. Only the checkpoint-variability and late-stage progress diagnostics continue beyond the endpoint at a fixed 10^{-6}. Schedule-Free warmup is $\eta_t^{\mathrm{SF}} = 4 \times 10^{-4}(t/5000)$ for $0 \le t \le 5000$.

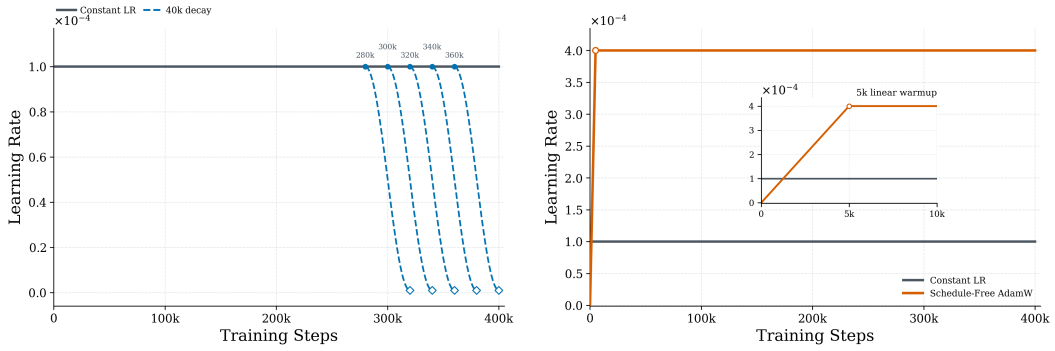

Figure 6: Learning-rate schedules used in the main experiment. (a) CONSTANT LR + EMA and independent 40k-step cosine-decay branches from different late-stage checkpoints of the same Constant-LR AdamW trajectory. Each branch decays the learning rate from 10^{-4} to 10^{-6}; different dashed curves indicate different decay starting points. (b) Constant-LR AdamW uses 10^{-4} throughout, whereas SCHEDULE-FREE ADAMW uses a base learning rate of 4×10^{-4} with a 5k-step linear warmup.

A.3 LATE-STAGE OPTIMIZATION METRICS

This section gives the complete protocols and numerical results for the late-stage diagnostics used in Section 3.4. Together, checkpoint variability, net parameter displacement, and fixed-evaluation loss progress characterize parameter variability, retained parameter movement, and continued objective improvement.

A.3.1 CHECKPOINT VARIABILITY

From each late-stage regime we sample $N = 20$ checkpoints $\theta^{(i)} \in \mathbb{R}^D$. For dimension j,

$$\bar{\theta}_j = \frac{1}{N}\sum_{i=1}^{N}\theta_j^{(i)}, \qquad \mathrm{Var}_j = \frac{1}{N-1}\sum_{i=1}^{N}(\theta_j^{(i)} - \bar{\theta}_j)^2, \tag{21}$$

and report $V_{\mathrm{total}} = \sum_{j=1}^{D}\mathrm{Var}_j$. We refer to this trace-variance statistic as *checkpoint variability*. Because it measures finite-window empirical spread, it may include both stochastic fluctuations and systematic movement and is not identified with pure theoretical stochastic variance.

The checkpoint gap g denotes the index spacing between samples in the saved-checkpoint sequence; $g = 1$ therefore selects consecutive saved checkpoints. We use 20 such checkpoints, $\theta_t, \ldots, \theta_{t+19}$. This short adjacent-checkpoint window emphasizes local parameter variability without introducing the larger systematic movement accumulated over longer gaps. Constant-LR and Schedule-Free checkpoints come directly from their late-stage trajectories. For cosine decay only, this diagnostic continues from the decay endpoint at a fixed 10^{-6} terminal rate.

Table 2: Checkpoint variability over 20 consecutive saved checkpoints (checkpoint gap $g = 1$), measured by total trace variance. Parentheses give the ratio to the Constant-LR online trajectory. Both model and EMA sequences are reported for cosine decay.

Method / sequence	Checkpoint variability (trace variance, $g = 1$)
Constant-LR online	7.1611709595e+00 (100%)
Schedule-Free averaged x	1.4373749495e-03 (0.0201%)
Constant-LR EMA	5.5516868830e-03 (0.0775%)
Cosine-decay model	7.3927e-04 (0.0103%)
Cosine-decay EMA	1.1807e-04 (0.0016%)

At checkpoint gap $g = 1$, cosine-decay EMA has the lowest local checkpoint variability, followed by the cosine-decay model, Schedule-Free averaged x, Constant-LR EMA, and the Constant-LR online trajectory.

A.3.2 Late-Stage Net Parameter Displacement

To directly measure how much parameter change each strategy retains late in training, we compute the ℓ_2 distance between checkpoints separated by $\tau = 1000$ optimization steps:

$$D_\tau(t) = \|\theta_{t+\tau} - \theta_t\|_2. \tag{22}$$

We use five consecutive 1k-step windows from 400k to 405k: $400\text{k} \to 401\text{k}$, $401\text{k} \to 402\text{k}$, $402\text{k} \to 403\text{k}$, $403\text{k} \to 404\text{k}$, and $404\text{k} \to 405\text{k}$.

Importantly, $D_\tau(t)$ measures the net change retained between the endpoints of a window, not the path length of the full optimization trajectory within that window. A larger displacement therefore does not by itself imply more effective optimization: it may reflect directed progress, stochastic motion, or both. We pair this measurement with fixed-evaluation loss progress below.

Table 3: Net parameter displacement over five adjacent 1k-step late-stage windows. Each entry is the ℓ_2 distance $\|\theta_{t+1000} - \theta_t\|_2$, not its square. For the Constant-LR baseline and cosine decay, we report both model and EMA weights; Schedule-Free uses its evaluation-time averaged x sequence.

Run	Weights	400–401k	401–402k	402–403k	403–404k	404–405k	Mean
AdamW baseline	model	69.953	68.033	68.007	68.192	68.026	68.442
AdamW baseline	EMA	12.636	12.700	12.683	12.676	12.669	12.673
Schedule-Free	x	6.5643	6.5572	6.5492	6.5408	6.5326	6.5488
40k cosine decay	model	0.909307	0.843436	0.831407	0.827545	0.820779	0.846495
40k cosine decay	EMA	1.794585	1.624687	1.471350	1.333021	1.208285	1.486386

The Constant-LR online trajectory retains the largest late-stage parameter movement, while EMA substantially reduces it. The Schedule-Free averaged x sequence moves less than Constant LR + EMA but more than both 40k cosine-decay sequences; within cosine decay, the model sequence moves less than EMA. Thus, explicit decay suppresses late-stage parameter mobility most strongly, whereas Schedule-Free retains more movement. Whether this movement corresponds to useful progress is assessed next.

A.3.3 Fixed-Evaluation Loss Progress

Net parameter displacement alone cannot determine whether parameter changes represent directed optimization progress. We therefore evaluate every checkpoint with identical data, noise, and timestep samples, reducing evaluation randomness when comparing the 400k–405k trajectories. Define the cumulative loss decrease relative to the 400k checkpoint as

$$\Delta L(t) = L_{400\text{k}} - L_t. \tag{23}$$

Here $\Delta L(t) > 0$ indicates improvement over the 400k checkpoint, and a larger value indicates more cumulative progress within the late-stage window.

Table 4: Fixed-evaluation total loss from 400k to 405k. All methods use identical fixed evaluation data, noise, and timestep sampling. Constant LR + EMA and Schedule-Free decrease throughout this interval, whereas neither the 40k Cosine Decay model nor its EMA improves in the post-decay regime.

Method	400k	401k	402k	403k	404k	405k
Constant LR + EMA	.43738	.43736	.43732	.43729	.43726	.43723
Schedule-Free	.43461	.43460	.43457	.43454	.43451	.43447
40k Cosine Decay model	.43696	.43735	.43745	.43754	.43752	.43760
40k Cosine Decay + EMA	.43691	.43691	.43693	.43695	.43697	.43699

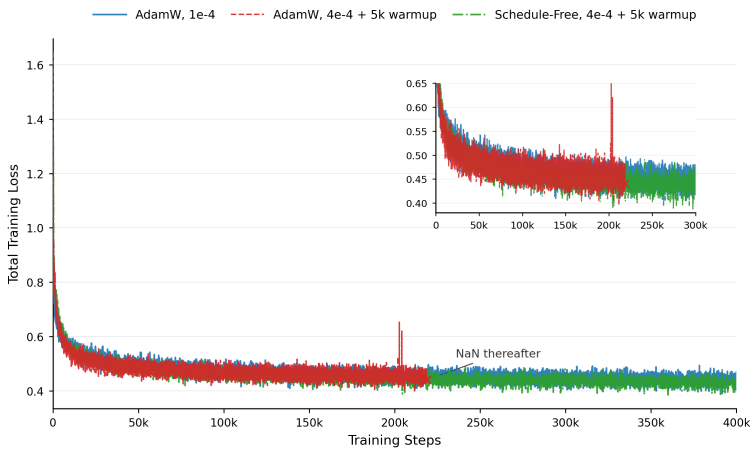

Figure 7: Stability at an aggressive learning-rate scale. Standard AdamW with learning rate 4×10^{-4} and 5k-step linear warmup develops a loss spike and sustained NaNs after approximately 218.7k steps. Schedule-Free remains stable with the same rate and warmup; standard AdamW is stable at 10^{-4}. The inset enlarges the normal-loss range.

Constant LR + EMA decreases from 0.43738 to 0.43723, and Schedule-Free decreases from 0.43461 to 0.43447. In contrast, the 40k Cosine Decay model increases from 0.43696 to 0.43760, while its EMA increases slightly from 0.43691 to 0.43699. This complements the displacement diagnostic: in the measured post-decay window, explicit decay exhibits both much smaller net movement and substantially weaker subsequent objective progress, whereas Schedule-Free retains more movement while continuing to reduce the objective.

Net displacement and fixed-evaluation progress are therefore complementary diagnostics. The former measures the parameter change retained over a fixed window; the latter tests whether continued training yields objective improvement. Together they support the empirical parameter-variability–optimization trade-off discussed in Section 3.4.

A.4 Learning-Rate Stability Ablation

We test whether Schedule-Free performs better merely because it uses a larger base learning rate. The official AdamW configuration trains stably at 10^{-4}; to match Schedule-Free's learning-rate scale, we also train AdamW at 4×10^{-4} with the same 5k-step warmup and otherwise identical settings. Figure 7 shows that this AdamW run develops a loss spike after roughly 200k steps and sustained NaNs after approximately 218.7k, whereas Schedule-Free remains stable. Thus, Schedule-Free can stably use a larger learning rate that standard AdamW cannot sustain in this setting.

A.5 ReFL Experimental Details

Following the ReFL protocol (Xu et al., 2023) in DiffusionOPSD (Zhou et al., 2026), we run ReFL on Z-Image-Turbo (Z-Image Team et al., 2025) and keep the official repository configuration fixed except for optimizer, learning rate, and warmup. HPSv3 (Ma et al., 2025) is the sole training reward. Training uses six policy GPUs and one reward GPU. Each update uses 48 prompt groups and 12 trajectories per prompt ($48 \times 12 = 576$ trajectories). Each policy GPU uses micro-batch size 1, synchronizing after accumulating 96 trajectories. Training lasts 100 updates. AdamW uses $\eta_{\text{ref}} = 10^{-5}$ without warmup. Schedule-Free uses a five-step warmup and evaluates optimizer x weights; AdamW evaluates EMA LoRA weights (Hu et al., 2022). At checkpoint 100, all runs generate from the same 1,000 fixed prompts and seeds at 1024×1024, with nine sampling steps, CFG scale 0, bfloat16, and seed 42. HPSv3 is reported as the on-reward metric; PickScore (Kirstain et al., 2023), CLIPScore (Hessel et al., 2021), HPSv2.1 (Wu et al., 2023), Aesthetic Score (Schuhmann & Beaumont, 2022), ImageReward (Xu et al., 2023), HPSv3++ (Liu et al., 2026), and the UniPercept

IAA, IQA, and ISTA scores (Cao et al., 2025) are used only as orthogonal reward models for cross-reward evaluation. Qualitative image results for ReFL are provided in Appendix E.2.

Table 5: ReFL results with HPSv3 as the training reward. HPSv3 is the directly optimized on-reward metric; all remaining columns are orthogonal reward models used only for cross-reward evaluation and provide no training signal. Fine-tuned runs are evaluated at checkpoint 100, and the Z-Image base model without reinforcement learning is included as a reference. "wu0" and "wu5" denote zero and five warmup steps, respectively. Bold indicates the best value in each column among the displayed rows.

	Training reward	Orthogonal evaluation rewards								
Run	HPSv3	Pick	CLIP	HPSv2.1	Aes	ImgR	HPSv3++	Uni IAA	Uni IQA	Uni ISTA
Z-Image base (no RL)	6.2544	22.8690	0.2754	0.2962	5.4118	0.9947	12.0641	**60.9511**	75.0656	37.7925
AdamW 1×, wu0	6.3500	22.8766	0.2754	0.2971	5.4177	0.9959	12.1271	60.7050	74.8078	38.3209
SF-AdamW 4×, wu5	6.8626	**22.8921**	0.2748	0.2995	5.4456	1.0150	**12.3485**	60.6995	**75.2258**	38.3189
SF-AdamW 20×, wu5	**9.6134**	22.8341	**0.2807**	**0.3123**	**5.7066**	**1.1197**	12.0347	51.5663	67.4061	**41.7878**

A.6 DMD Experimental Details

We study class-conditional 256×256 ImageNet generation (Deng et al., 2009) with SiT-XL/2 initialized from REPA-pretrained weights (Ma et al., 2024; Yu et al., 2025) in the one-step DMDR distribution-matching setup (Yin et al., 2024; Jiang et al., 2025). Guidance is updated once per loop and the generator once per five guidance updates. Reported steps count generator updates. All three configurations train for 3,000 generator steps and are evaluated at 1k, 2k, and 3k. No DINO or reinforcement-learning auxiliary loss is enabled.

The generator is trained with

$$\mathcal{L}_{\mathrm{DMD}} = \frac{1}{2}\,\mathrm{mean}\Big[\big(x_0 - \mathrm{sg}\left(x_0 - g_{\mathrm{DMD}}\right)\big)^2\Big], \tag{24}$$

where sg denotes stop-gradient, and the DMD direction is defined as the normalized difference

$$g_{\mathrm{DMD}} = \frac{p_{\mathrm{real}} - p_{\mathrm{fake}}}{\mathrm{mean}_{c,h,w}(|p_{\mathrm{real}}|) + \epsilon}, \qquad p_{\mathrm{real}} = x_0 - \hat{x}_{0,r}, \quad p_{\mathrm{fake}} = x_0 - \hat{x}_{0,f}. \tag{25}$$

Here, x_0 is the generator-predicted clean latent, while $\hat{x}_{0,r}$ and $\hat{x}_{0,f}$ are the clean-latent predictions obtained from the real and fake estimator modes of the same guidance model, respectively. The mean in the denominator is computed independently for each sample over the latent dimensions (c, h, w), rather than over the batch. The fake-score model is trained with the diffusion fitting objective

$$\mathcal{L}_{\mathrm{diff}} = \mathbb{E}\left[\left\|v_{\mathrm{fake}}(x_t, t, y) - (\epsilon - \tilde{x}_0)\right\|_2^2\right]. \tag{26}$$

Here $\tilde{x}_0 \sim G_\theta(z, y)$ denotes a clean latent sampled from the current generator distribution. It is conceptually the same type of clean sample as x_0 above, but is generated through the separate fake-score diffusion-fitting computation. These are also the fixed-evaluation objectives reported in Figure 5b–c.

Training uses four GPUs with per-GPU batch size 128 (global 512). In the all-AdamW baseline, both the generator and guidance model use AdamW at 2×10^{-5} without warmup. In the generator-side ablation, the generator instead uses Schedule-Free AdamW at 4×10^{-5} ($2\times$ its baseline learning rate) with a 200-step warmup and averaged x weights, while guidance retains baseline AdamW. In the guidance-side ablation, guidance instead uses Schedule-Free AdamW at 4×10^{-5} ($2\times$ its baseline learning rate) with a 200-step warmup, while the generator retains baseline AdamW. All remaining settings are held fixed. The baseline evaluates online weights. Sampling uses one step, random ImageNet class labels, CFG scale 0, and 5,000 images per checkpoint.

Table 6: DMD generation quality under the all-AdamW baseline and the two component-wise Schedule-Free ablations. Every metric is evaluated on 5,000 generated images. Within each checkpoint, bold marks the best of the three configurations. Lower FID-5K and sFID and higher IS, precision, and recall are better.

Method	Ckpt	FID-5K ↓	sFID ↓	IS ↑	Precision ↑	Recall ↑
AdamW baseline	1k	10.126	**35.310**	248.559	0.7886	0.7395
Gen SF-AdamW + Gui AdamW	1k	**9.988**	35.733	236.856	0.7850	**0.7401**
Gen AdamW + Gui SF-AdamW	1k	10.199	35.584	**256.577**	**0.7970**	0.7210
AdamW baseline	2k	9.757	**34.865**	242.844	0.7814	**0.7497**
Gen SF-AdamW + Gui AdamW	2k	**9.573**	35.261	**256.355**	**0.7950**	0.7451
Gen AdamW + Gui SF-AdamW	2k	9.882	35.529	245.570	0.7848	0.7395
AdamW baseline	3k	9.448	**34.889**	**258.754**	0.7932	0.7435
Gen SF-AdamW + Gui AdamW	3k	**9.346**	35.192	251.219	0.7890	**0.7505**
Gen AdamW + Gui SF-AdamW	3k	9.634	35.477	257.485	**0.7960**	0.7449

B Quadratic Analyses and the Classical Nonconvex SGD Bound

B.1 Constant-LR SGD

B.1.1 Stationary Distribution on Quadratic Objectives

Theorem B.1 (Stationary distribution of constant-learning-rate SGD)**.** *Consider the one-dimensional quadratic $L(x) = \frac{1}{2}dx^2$, $d > 0$, with stochastic gradient $dx_t + \xi_t$, where the $\xi_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, and deterministic x_0. The constant-step recursion is*

$$x_{t+1} = (1 - \eta d)x_t - \eta \xi_t. \tag{27}$$

If $0 < \eta d < 2$, then

$$x_t \;\Rightarrow\; x_\infty \sim \mathcal{N}\left(0, \frac{\eta \sigma^2}{d(2 - \eta d)}\right). \tag{28}$$

Proof. Writing $b = 1 - \eta d$, the first two iterates are

$$x_1 = bx_0 - \eta \xi_0,$$
$$x_2 = bx_1 - \eta \xi_1 = b^2 x_0 - \eta\big(b\xi_0 + \xi_1\big).$$

More generally, if the claimed expression holds at time t, then one further application of the recursion gives

$$\begin{aligned} x_{t+1} &= b\left(b^t x_0 - \eta \sum_{k=0}^{t-1} b^{t-1-k} \xi_k\right) - \eta \xi_t \\ &= b^{t+1} x_0 - \eta \sum_{k=0}^{t-1} b^{t-k} \xi_k - \eta \xi_t \\ &= b^{t+1} x_0 - \eta \sum_{k=0}^{t} b^{t-k} \xi_k. \end{aligned}$$

Thus induction yields

$$x_t = b^t x_0 - \eta \sum_{k=0}^{t-1} b^{t-1-k} \xi_k. \tag{29}$$

Because $|b| < 1$, $b^t x_0 \to 0$. The stochastic term is a linear combination of independent centered Gaussian noises and is therefore centered Gaussian. Independence eliminates all cross-covariances,

so its variance is

$$\operatorname{Var}\left(\eta \sum_{k=0}^{t-1} b^{t-1-k}\xi_k\right) = \eta^2 \sum_{k=0}^{t-1} b^{2(t-1-k)} \operatorname{Var}(\xi_k)$$
$$= \eta^2\sigma^2 \sum_{j=0}^{t-1} b^{2j} = \eta^2\sigma^2 \frac{1-b^{2t}}{1-b^2}.$$

Consequently,

$$\eta^2\sigma^2 \sum_{j=0}^{t-1} b^{2j} \longrightarrow \frac{\eta^2\sigma^2}{1-b^2} = \frac{\eta\sigma^2}{d(2-\eta d)}, \tag{30}$$

where the last equality uses $1-b^2 = 1-(1-\eta d)^2 = \eta d(2-\eta d)$. This proves Eq. (28). □

B.1.2 Smooth Nonconvex Stationarity

Theorem B.2 (Convergence of SGD for smooth nonconvex optimization). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and bounded below by $f_\star$. Suppose $x_{t+1} = x_t - \eta g_t$, $\mathbb{E}[g_t \mid x_t] = \nabla f(x_t)$, and*

$$\mathbb{E}[\|g_t - \nabla f(x_t)\|^2 \mid x_t] \le \sigma^2. \tag{31}$$

For $0 < \eta \le 1/L$,

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(x_t)\|^2 \le \frac{2(f(x_0)-f_\star)}{\eta T} + L\eta\sigma^2. \tag{32}$$

Choosing $\eta = \Theta(T^{-1/2})$ gives an average squared-gradient norm of $O(T^{-1/2})$, and the same order holds for the best iterate among $t < T$.

Proof. Smoothness gives

$$f(x_{t+1}) \le f(x_t) - \eta\langle \nabla f(x_t), g_t\rangle + \frac{L\eta^2}{2}\|g_t\|^2. \tag{33}$$

To evaluate the second moment of the stochastic gradient, write $\zeta_t := g_t - \nabla f(x_t)$. Conditional on x_t, the assumptions give $\mathbb{E}[\zeta_t \mid x_t] = 0$ and $\mathbb{E}[\|\zeta_t\|^2 \mid x_t] \le \sigma^2$. Hence

$$\mathbb{E}[\|g_t\|^2 \mid x_t] = \mathbb{E}[\|\nabla f(x_t) + \zeta_t\|^2 \mid x_t]$$
$$= \|\nabla f(x_t)\|^2 + 2\,\langle \nabla f(x_t), \mathbb{E}[\zeta_t \mid x_t]\rangle + \mathbb{E}[\|\zeta_t\|^2 \mid x_t],$$

and therefore

$$\mathbb{E}[\|g_t\|^2 \mid x_t] \le \|\nabla f(x_t)\|^2 + \sigma^2. \tag{34}$$

Taking conditional expectation in Eq. (33), using $\mathbb{E}[g_t \mid x_t] = \nabla f(x_t)$, and then applying the preceding second-moment bound gives

$$\mathbb{E}[f(x_{t+1}) \mid x_t] \le f(x_t) - \eta\,\langle \nabla f(x_t), \mathbb{E}[g_t \mid x_t]\rangle + \frac{L\eta^2}{2}\mathbb{E}[\|g_t\|^2 \mid x_t]$$
$$\le f(x_t) - \eta\|\nabla f(x_t)\|^2 + \frac{L\eta^2}{2}\big(\|\nabla f(x_t)\|^2 + \sigma^2\big),$$

which, after collecting the squared-gradient terms, becomes

$$\mathbb{E}[f(x_{t+1}) \mid x_t] \le f(x_t) - \left(\eta - \frac{L\eta^2}{2}\right)\|\nabla f(x_t)\|^2 + \frac{L\eta^2\sigma^2}{2}. \tag{35}$$

For $\eta \le 1/L$, $\eta - L\eta^2/2 \ge \eta/2$. Taking total expectations and rearranging therefore yields

$$\frac{\eta}{2}\mathbb{E}\|\nabla f(x_t)\|^2 \le \mathbb{E}f(x_t) - \mathbb{E}f(x_{t+1}) + \frac{L\eta^2\sigma^2}{2}.$$

Summing this inequality from $t = 0$ to $T-1$ telescopes the function values:

$$\frac{\eta}{2}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(x_t)\|^2 \le f(x_0) - \mathbb{E}f(x_T) + \frac{TL\eta^2\sigma^2}{2}$$
$$\le f(x_0) - f_\star + \frac{TL\eta^2\sigma^2}{2}.$$

Dividing by $\eta T/2$ proves Eq. (32), which is the bound stated in Eq. (4). The best-iterate statement follows from $\min_{0\le t<T} a_t \le T^{-1}\sum_{t=0}^{T-1} a_t$ applied to $a_t = \mathbb{E}\|\nabla f(x_t)\|^2$. □

B.2 CONSTANT-LR SGD + EMA

B.2.1 STATIONARY VARIANCE REDUCTION

Theorem B.3 (EMA reduces stationary variance)**.** *Under the one-dimensional quadratic setting of Appendix B.1, let $b = 1 - \eta d$, $|b| < 1$, and let $y_t = \beta y_{t-1} + (1-\beta)x_t$ with $0 < \beta < 1$. The stationary variances satisfy*

$$\frac{\mathrm{Var}(y_\infty)}{\mathrm{Var}(x_\infty)} = \frac{1-\beta}{1+\beta}\frac{1+\beta b}{1-\beta b} < 1. \tag{36}$$

Thus EMA reduces the stationary variance of the evaluation iterate.

Proof. Online SGD is $x_{t+1} = bx_t - \eta\xi_t$ and has stationary variance

$$v := \mathrm{Var}(x_t) = \frac{\eta^2\sigma^2}{1-b^2}. \tag{37}$$

We first justify that the stationary EMA process and its covariance with x_t are well defined. Take the online iterate in its stationary regime and define the causal EMA solution

$$y_t := (1-\beta)\sum_{j=0}^{\infty}\beta^j x_{t-j}.$$

Because $\mathbb{E}[x_t^2] = v < \infty$ and $0 < \beta < 1$, its partial sums are Cauchy in L^2: for $m > n$,

$$\begin{aligned}\left\|(1-\beta)\sum_{j=n+1}^{m}\beta^j x_{t-j}\right\|_{L^2} &\le (1-\beta)\sum_{j=n+1}^{m}\beta^j\|x_{t-j}\|_{L^2}\\ &\le \sqrt{v}\,\beta^{n+1} \longrightarrow 0.\end{aligned}$$

Thus the series converges in L^2, has finite variance, is jointly stationary with x_t, and satisfies $y_t = \beta y_{t-1} + (1-\beta)x_t$. It is also the unique stationary L^2 solution. Indeed, if y_t and $\widetilde{y}_t$ are two such solutions, then $D_t := y_t - \widetilde{y}_t = \beta D_{t-1}$; stationarity implies $\mathbb{E}[D_t^2] = \beta^2\mathbb{E}[D_t^2]$, and hence $D_t = 0$ almost surely. In particular, $c := \mathrm{Cov}(y_t, x_t)$ is finite by Cauchy–Schwarz.

Both stationary means are zero. Using $x_t = bx_{t-1} - \eta\xi_{t-1}$ and expanding the covariance gives

$$\begin{aligned}c &= \mathrm{Cov}(\beta y_{t-1} + (1-\beta)x_t,\, x_t)\\ &= \beta\,\mathrm{Cov}(y_{t-1}, x_t) + (1-\beta)\,\mathrm{Var}(x_t)\\ &= \beta\,\mathrm{Cov}(y_{t-1}, bx_{t-1} - \eta\xi_{t-1}) + (1-\beta)v\\ &= \beta b\,\mathrm{Cov}(y_{t-1}, x_{t-1}) - \beta\eta\,\mathrm{Cov}(y_{t-1}, \xi_{t-1}) + (1-\beta)v.\end{aligned}$$

Here y_{t-1} is measurable with respect to the history before ξ_{t-1} is sampled, so $\mathrm{Cov}(y_{t-1}, \xi_{t-1}) = 0$. Joint stationarity gives $\mathrm{Cov}(y_{t-1}, x_{t-1}) = c$, and hence

$$c = \beta bc + (1-\beta)v = \frac{1-\beta}{1-\beta b}v. \tag{38}$$

The division is valid because $|\beta b| < 1$.

Writing $w = \mathrm{Var}(y_t)$,

$$w = \beta^2 w + (1-\beta)^2 v + 2\beta(1-\beta)bc, \tag{39}$$

so Eq. (36) follows from

$$\frac{w}{v} = \frac{1-\beta}{1+\beta}\frac{1+\beta b}{1-\beta b} < 1, \tag{40}$$

because $(1+\beta)(1-\beta b) - (1-\beta)(1+\beta b) = 2\beta(1-b) > 0$. □

B.2.2 A Limitation of EMA: Persistent Optimization Bias

The preceding quadratic model is symmetric, so its invariant distribution is centered at the minimizer. This need not hold for a nonquadratic objective. Consider the one-dimensional strongly convex and smooth example

$$f(x) = \frac{x^2}{2} + a(\sin x - x), \qquad 0 < a < 1, \tag{41}$$

whose unique minimizer is $x^\star = 0$. Constant-step SGD follows

$$X_{t+1} = F_h(X_t) - h\xi_t, \qquad F_h(x) := x - hf'(x), \qquad \xi_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \tag{42}$$

with $\sigma^2 > 0$.

Theorem B.4 (EMA preserves the stationary optimization bias)**.** *For the recursion in Eq. (42), suppose $0 < h \le 1/(1+a)$. Then SGD has a unique invariant distribution π_h with finite second moment, and the law of X_t converges to π_h geometrically in the 2-Wasserstein distance. If $U \sim \pi_h$ and $m_h := \mathbb{E}[U]$, then*

$$\mathbb{E}[U^2] \le \frac{h\sigma^2}{(1-a)\{2-h(1-a)\}}, \qquad m_h = a\big(1 - \mathbb{E}[\cos U]\big) > 0 = x^\star. \tag{43}$$

If X_t is stationary and $Y_t = \beta Y_{t-1} + (1-\beta)X_t$ for $0 < \beta < 1$, then

$$\mathbb{E}[Y_t] = \beta^t \mathbb{E}[Y_0] + (1-\beta^t)m_h \longrightarrow m_h. \tag{44}$$

Thus EMA can reduce the variance around the stationary mean, but it does not move that mean to the optimizer.

Proof. Since $f''(x) = 1 - a\sin x$, we have $1 - a \le f''(x) \le 1 + a$. For $0 < h \le 1/(1+a)$,

$$0 \le F_h'(x) = 1 - hf''(x) \le q_h := 1 - h(1-a) < 1. \tag{45}$$

The mean-value theorem therefore gives $|F_h(x) - F_h(z)| \le q_h|x - z|$. Couple two SGD chains by using the same noise at every step. Their noise terms cancel, so their one-step laws satisfy

$$W_2(\nu_1 P_h, \nu_2 P_h) \le q_h W_2(\nu_1, \nu_2), \tag{46}$$

where P_h is the transition kernel of Eq. (42). The kernel preserves finite second moments. Hence it is a contraction on the complete metric space of probability measures with finite second moment, and the Banach fixed-point theorem gives a unique invariant distribution π_h and geometric convergence to it.

Because $F_h(0) = 0$, invariance, zero-mean noise, and the contraction imply

$$\mathbb{E}[U^2] = \mathbb{E}\big[(F_h(U) - h\xi)^2\big] \tag{47}$$

$$= \mathbb{E}[F_h(U)^2] + h^2\sigma^2 \le q_h^2 \mathbb{E}[U^2] + h^2\sigma^2. \tag{48}$$

Moving the $q_h^2\mathbb{E}[U^2]$ term to the left gives

$$(1 - q_h^2)\mathbb{E}[U^2] \le h^2\sigma^2.$$

From $q_h = 1 - h(1-a)$,

$$\begin{aligned} 1 - q_h^2 &= (1 - q_h)(1 + q_h) \\ &= h(1-a)\big\{2 - h(1-a)\big\}. \end{aligned}$$

This quantity is positive because $0 < q_h < 1$. Dividing by it therefore yields

$$\begin{aligned} \mathbb{E}[U^2] &\le \frac{h^2\sigma^2}{1 - q_h^2} \\ &= \frac{h\sigma^2}{(1-a)\{2-h(1-a)\}}, \end{aligned}$$

which is the second-moment bound in Eq. (43).

To derive the stationary mean, let $U \sim \pi_h$ be independent of a fresh $\xi \sim \mathcal{N}(0, \sigma^2)$. Invariance means that $U' := F_h(U) - h\xi$ also has law π_h, so $\mathbb{E}[U'] = \mathbb{E}[U] = m_h$. Taking expectation in the update gives

$$\begin{aligned} m_h &= \mathbb{E}[F_h(U) - h\xi] \\ &= \mathbb{E}[U - hf'(U)] - h\mathbb{E}[\xi] \\ &= m_h - h\mathbb{E}[f'(U)]. \end{aligned}$$

Since $h > 0$, subtracting m_h from both sides and dividing by h yields $\mathbb{E}[f'(U)] = 0$. Using $f'(x) = x + a(\cos x - 1)$, we obtain

$$\begin{aligned} 0 &= \mathbb{E}[U] + a\big(\mathbb{E}[\cos U] - 1\big) \\ &= m_h + a\big(\mathbb{E}[\cos U] - 1\big), \end{aligned}$$

and hence

$$m_h = a\big(1 - \mathbb{E}[\cos U]\big) \geq 0. \tag{49}$$

The inequality is strict: conditionally on U, the next iterate has a nondegenerate Gaussian density, so invariance implies that π_h is not supported on the countable set $2\pi\mathbb{Z}$ where $\cos U = 1$. Therefore $\mathbb{E}[\cos U] < 1$ and $m_h > 0 = x^\star$.

Finally, under stationarity of X_t, let $\mu_t := \mathbb{E}[Y_t]$ and note that $\mathbb{E}[X_t] = m_h$ for every t. Taking expectation in the EMA recursion gives

$$\mu_t = \beta\mu_{t-1} + (1-\beta)m_h.$$

Subtracting m_h from both sides and iterating yields

$$\begin{aligned} \mu_t - m_h &= \beta(\mu_{t-1} - m_h) = \cdots = \beta^t(\mu_0 - m_h), \\ \mu_t &= \beta^t\mu_0 + (1-\beta^t)m_h \longrightarrow m_h, \end{aligned}$$

because $0 < \beta < 1$. This is Eq. (44), so the EMA sequence inherits the same stationary mean bias. □

B.3 Learning-Rate Decay SGD + EMA

B.3.1 Exact Dynamics under Learning-Rate Decay and EMA

Consider the one-dimensional quadratic model from Section 2.3 under the inverse-time schedule $\eta_t = \eta/(2 + t\eta d)$. For compactness, define

$$a := \eta d, \qquad D_t := 2 + (t-1)a. \tag{50}$$

Theorem B.5 (Exact bias–variance decomposition for learning-rate decay). *Let $f(x) = \frac{1}{2}dx^2$, $d > 0$, and $g_t = dx_t + \xi_t$, where the noises are independent with $\mathbb{E}[\xi_t] = 0$ and $\mathbb{E}[\xi_t^2] = \sigma^2$. Suppose x_0 is deterministic, $0 < \eta d < 2$, and $x_{t+1} = x_t - \eta_t g_t$ with $\eta_t = \eta/(2 + t\eta d)$. Then, for every $t \geq 1$,*

$$x_t = \underbrace{\frac{2-a}{D_t}x_0}_{\textit{mean trajectory}} - \underbrace{\frac{\eta}{D_t}\sum_{k=0}^{t-1}\xi_k}_{\textit{stochastic fluctuation}}. \tag{51}$$

Consequently,

$$\mathbb{E}[x_t] = \frac{2-a}{D_t}x_0 = O(t^{-1}), \qquad \mathrm{Var}(x_t) = \frac{t\eta^2\sigma^2}{D_t^2} = O(t^{-1}), \tag{52}$$

and $\mathbb{E}[f(x_t)] = O(t^{-1})$.

Proof. Since $D_{t+1} = 2 + ta$, $a = \eta d$, and $\eta_t = \eta/(2 + ta)$, we have

$$\begin{aligned} 1 - \eta_t d = 1 - \frac{\eta d}{2+ta} = \frac{2 + ta - a}{2 + ta} = \frac{2 + (t-1)a}{2 + ta} = \frac{D_t}{D_{t+1}}, \\ \eta_t = \frac{\eta}{D_{t+1}}. \end{aligned}$$

Substituting these identities into $x_{t+1} = (1 - \eta_t d)x_t - \eta_t \xi_t$ gives

$$x_{t+1} = \frac{D_t}{D_{t+1}} x_t - \frac{\eta}{D_{t+1}} \xi_t. \tag{53}$$

Multiplying Eq. (53) by D_{t+1} and moving $D_t x_t$ to the left gives the difference identity

$$D_{t+1}x_{t+1} - D_t x_t = -\eta \xi_t.$$

Applying this identity at indices $k = 0, \ldots, t-1$ and summing yields

$$\sum_{k=0}^{t-1} \big(D_{k+1}x_{k+1} - D_k x_k\big) = -\eta \sum_{k=0}^{t-1} \xi_k,$$

$$\underbrace{(D_1 x_1 - D_0 x_0) + (D_2 x_2 - D_1 x_1) + \cdots + (D_t x_t - D_{t-1} x_{t-1})}_{D_t x_t - D_0 x_0} = -\eta \sum_{k=0}^{t-1} \xi_k.$$

All intermediate terms cancel. Since $D_0 = 2 - a$, rearranging gives

$$D_t x_t = (2-a)x_0 - \eta \sum_{k=0}^{t-1} \xi_k, \qquad x_t = \frac{2-a}{D_t} x_0 - \frac{\eta}{D_t} \sum_{k=0}^{t-1} \xi_k,$$

which is Eq. (51).

Because x_0 is deterministic and every noise variable is centered, linearity of expectation gives

$$\mathbb{E}[x_t] = \frac{2-a}{D_t} x_0 - \frac{\eta}{D_t} \sum_{k=0}^{t-1} \mathbb{E}[\xi_k] = \frac{2-a}{D_t} x_0.$$

The deterministic term contributes no variance, so

$$\mathrm{Var}(x_t) = \frac{\eta^2}{D_t^2} \mathrm{Var}\left(\sum_{k=0}^{t-1} \xi_k \right).$$

To evaluate the remaining variance, expand the centered square:

$$\begin{aligned} \mathrm{Var}\left(\sum_{k=0}^{t-1} \xi_k \right) &= \mathbb{E}\left[\left(\sum_{k=0}^{t-1} \xi_k \right)^2 \right] \\ &= \sum_{k=0}^{t-1} \mathbb{E}[\xi_k^2] + 2 \sum_{0 \le k < \ell \le t-1} \mathbb{E}[\xi_k \xi_\ell] \\ &= t\sigma^2. \end{aligned}$$

The last equality follows because independence and zero means imply $\mathbb{E}[\xi_k \xi_\ell] = \mathbb{E}[\xi_k]\mathbb{E}[\xi_\ell] = 0$ for $k \neq \ell$. Substitution proves Eq. (52). Finally,

$$\begin{aligned} \mathbb{E}[f(x_t)] &= \frac{d}{2} \mathbb{E}[x_t^2] = \frac{d}{2} \left\{ \mathbb{E}[x_t]^2 + \mathrm{Var}(x_t) \right\} \\ &= \frac{d(2-a)^2 x_0^2}{2D_t^2} + \frac{dt\eta^2\sigma^2}{2D_t^2} = O(t^{-1}), \end{aligned}$$

because the two displayed terms are respectively $O(t^{-2})$ and $O(t^{-1})$. □

Now define the EMA iterate

$$\bar{x}_t = \beta \bar{x}_{t-1} + (1-\beta) x_t, \qquad 0 < \beta < 1, \qquad \bar{x}_0 = x_0. \tag{54}$$

Theorem B.6 (Exact bias–variance decomposition for decay with EMA)**.** *Under the one-dimensional quadratic setting and inverse-time schedule of Section 2.3, suppose that x_0 is deterministic, $0 < \eta d < 2$, and $0 < \beta < 1$. Then*

$$\bar{x}_t = \underbrace{\beta^t x_0 + (1-\beta)(2-a)x_0 \sum_{j=1}^{t} \frac{\beta^{t-j}}{D_j}}_{\textit{mean trajectory / optimization process}} \underbrace{- \eta(1-\beta) \sum_{k=0}^{t-1} \left(\sum_{j=k+1}^{t} \frac{\beta^{t-j}}{D_j} \right) \xi_k}_{\textit{stochastic fluctuation}}. \tag{55}$$

Consequently,

$$\bar{m}_t := \mathbb{E}[\bar{x}_t] = \beta^t x_0 + (1-\beta)(2-a)x_0 \sum_{j=1}^{t} \frac{\beta^{t-j}}{D_j}, \tag{56}$$

$$\bar{v}_t := \mathrm{Var}(\bar{x}_t) = \eta^2(1-\beta)^2\sigma^2 \sum_{k=0}^{t-1} \left(\sum_{j=k+1}^{t} \frac{\beta^{t-j}}{D_j} \right)^2. \tag{57}$$

Therefore,

$$\mathbb{E}[f(\bar{x}_t)] = \underbrace{\frac{d}{2} \left[\beta^t x_0 + (1-\beta)(2-a)x_0 \sum_{j=1}^{t} \frac{\beta^{t-j}}{D_j} \right]^2}_{\textit{optimization term, including EMA lag}} + \underbrace{\frac{d\eta^2(1-\beta)^2\sigma^2}{2} \sum_{k=0}^{t-1} \left(\sum_{j=k+1}^{t} \frac{\beta^{t-j}}{D_j} \right)^2}_{\textit{variance contribution}}. \tag{58}$$

Proof. Unrolling the EMA recursion gives

$$\bar{x}_t = \beta^t x_0 + (1-\beta) \sum_{j=1}^{t} \beta^{t-j} x_j. \tag{59}$$

Substitution of Eq. (51) yields

$$\bar{x}_t = \beta^t x_0 + (1-\beta)(2-a)x_0 \sum_{j=1}^{t} \frac{\beta^{t-j}}{D_j} - \eta(1-\beta) \sum_{j=1}^{t} \frac{\beta^{t-j}}{D_j} \sum_{k=0}^{j-1} \xi_k. \tag{60}$$

Exchanging the order of summation in the stochastic term gives

$$\sum_{j=1}^{t} \frac{\beta^{t-j}}{D_j} \sum_{k=0}^{j-1} \xi_k = \sum_{k=0}^{t-1} \left(\sum_{j=k+1}^{t} \frac{\beta^{t-j}}{D_j} \right) \xi_k, \tag{61}$$

because the original index set $\{(j,k) : 1 \le j \le t,\ 0 \le k \le j-1\}$ is equivalently $\{(k,j) : 0 \le k \le t-1,\ k+1 \le j \le t\}$. This proves Eq. (55).

For clarity, denote the deterministic part and the coefficient of each noise variable by

$$A_t := \beta^t x_0 + (1-\beta)(2-a)x_0 \sum_{j=1}^{t} \frac{\beta^{t-j}}{D_j},$$

$$c_{t,k} := \eta(1-\beta) \sum_{j=k+1}^{t} \frac{\beta^{t-j}}{D_j}.$$

Then Eq. (55) becomes

$$\bar{x}_t = A_t - \sum_{k=0}^{t-1} c_{t,k}\xi_k.$$

Since A_t and the coefficients $c_{t,k}$ are deterministic and $\mathbb{E}[\xi_k] = 0$,

$$\mathbb{E}[\bar{x}_t] = A_t - \sum_{k=0}^{t-1} c_{t,k}\mathbb{E}[\xi_k] = A_t,$$

which is Eq. (56). For the variance, the deterministic term drops out, and expansion of the centered square gives

$$\begin{aligned}
\mathrm{Var}(\bar{x}_t) &= \mathbb{E}\left[\left(\sum_{k=0}^{t-1} c_{t,k}\xi_k\right)^2\right] \\
&= \sum_{k=0}^{t-1} c_{t,k}^2 \mathbb{E}[\xi_k^2] + 2 \sum_{0\le k<\ell\le t-1} c_{t,k}c_{t,\ell}\mathbb{E}[\xi_k\xi_\ell] \\
&= \sigma^2 \sum_{k=0}^{t-1} c_{t,k}^2 \\
&= \eta^2(1-\beta)^2\sigma^2 \sum_{k=0}^{t-1} \left(\sum_{j=k+1}^{t} \frac{\beta^{t-j}}{D_j}\right)^2.
\end{aligned}$$

The third line uses independence and centering: $\mathbb{E}[\xi_k\xi_\ell] = \mathbb{E}[\xi_k]\mathbb{E}[\xi_\ell] = 0$ for $k \neq \ell$, while $\mathbb{E}[\xi_k^2] = \sigma^2$. This proves Eq. (57).

Finally, because $f(x) = dx^2/2$,

$$\begin{aligned}
\mathbb{E}[f(\bar{x}_t)] &= \frac{d}{2}\mathbb{E}[\bar{x}_t^2] \\
&= \frac{d}{2}\left\{\mathbb{E}[\bar{x}_t]^2 + \mathrm{Var}(\bar{x}_t)\right\} \\
&= \frac{d}{2}A_t^2 + \frac{d\sigma^2}{2}\sum_{k=0}^{t-1} c_{t,k}^2.
\end{aligned}$$

Substituting the definitions of A_t and $c_{t,k}$ gives Eq. (58). □

For every fixed $0 < \beta < 1$, Eqs. (56)–(58) imply

$$\mathbb{E}[\bar{x}_t] = O(t^{-1}), \qquad \mathrm{Var}(\bar{x}_t) = O(t^{-1}), \qquad \mathbb{E}[f(\bar{x}_t)] = O(t^{-1}). \tag{62}$$

Thus EMA preserves the contraction obtained under learning-rate decay. The inverse-time schedule permits this closed-form calculation; the cosine-decay experiments in Section 3.4 show a similar empirical reduction in checkpoint variability.

B.3.2 Smooth Nonconvex Stationarity with Learning-Rate Decay

Theorem B.7 (Nonconvex stationarity of SGD with decaying learning rates). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and bounded below by $f_\star$. Consider*

$$x_{t+1} = x_t - \eta_t g_t, \qquad g_t = \nabla f(x_t) + \xi_t, \tag{63}$$

where, conditionally on the history $\mathcal{F}_t$ before sampling g_t,

$$\mathbb{E}[\xi_t \mid \mathcal{F}_t] = 0, \qquad \mathbb{E}[\|\xi_t\|^2 \mid \mathcal{F}_t] \le \sigma^2. \tag{64}$$

If $0 < \eta_t \le 1/L$, then for every $T \ge 1$,

$$\sum_{t=0}^{T-1} \eta_t \mathbb{E}\|\nabla f(x_t)\|^2 \le 2\Delta_0 + L\sigma^2 \sum_{t=0}^{T-1} \eta_t^2, \qquad \Delta_0 := f(x_0) - f_\star. \tag{65}$$

Equivalently, if $R_T \in \{0, \dots, T-1\}$ is sampled independently with $\mathbb{P}(R_T = t) = \eta_t / \sum_{k=0}^{T-1} \eta_k$, then

$$\mathbb{E}\|\nabla f(x_{R_T})\|^2 \le \frac{2\Delta_0 + L\sigma^2 \sum_{t=0}^{T-1} \eta_t^2}{\sum_{t=0}^{T-1} \eta_t}. \tag{66}$$

In particular, any schedule satisfying $\sum_{t=0}^{\infty} \eta_t = \infty$ and $\sum_{t=0}^{\infty} \eta_t^2 < \infty$ makes the right-hand side of Eq. (66) converge to zero, and hence $\mathbb{E}\|\nabla f(x_{R_T})\|^2 \to 0$. Since the minimum is no larger than this weighted average, it also follows that $\min_{0 \le t < T} \mathbb{E}\|\nabla f(x_t)\|^2 \to 0$.

Proof. By L-smoothness,

$$f(x_{t+1}) \le f(x_t) - \eta_t \langle \nabla f(x_t), g_t \rangle + \frac{L\eta_t^2}{2} \|g_t\|^2. \tag{67}$$

Because x_t is $\mathcal{F}_t$-measurable, the conditional assumptions give $\mathbb{E}[g_t \mid \mathcal{F}_t] = \nabla f(x_t)$. Expanding $\|\nabla f(x_t) + \xi_t\|^2$ and using $\mathbb{E}[\xi_t \mid \mathcal{F}_t] = 0$ gives the second-moment bound

$$\mathbb{E}[g_t \mid \mathcal{F}_t] = \nabla f(x_t), \qquad \mathbb{E}[\|g_t\|^2 \mid \mathcal{F}_t] \le \|\nabla f(x_t)\|^2 + \sigma^2. \tag{68}$$

Taking conditional expectation and using $\eta_t \le 1/L$ therefore gives

$$\mathbb{E}[f(x_{t+1}) \mid \mathcal{F}_t] \le f(x_t) - \frac{\eta_t}{2} \|\nabla f(x_t)\|^2 + \frac{L\eta_t^2 \sigma^2}{2}. \tag{69}$$

Taking total expectations and rearranging gives the one-step inequality

$$\eta_t \mathbb{E}\|\nabla f(x_t)\|^2 \le 2\big(\mathbb{E}f(x_t) - \mathbb{E}f(x_{t+1})\big) + L\eta_t^2 \sigma^2.$$

Summing it explicitly from $t = 0$ to $T - 1$ telescopes the function values:

$$\begin{aligned}
\sum_{t=0}^{T-1} \eta_t \mathbb{E}\|\nabla f(x_t)\|^2 &\le 2 \sum_{t=0}^{T-1} \big(\mathbb{E}f(x_t) - \mathbb{E}f(x_{t+1})\big) + L\sigma^2 \sum_{t=0}^{T-1} \eta_t^2 \\
&= 2\big(f(x_0) - \mathbb{E}f(x_T)\big) + L\sigma^2 \sum_{t=0}^{T-1} \eta_t^2 \\
&\le 2\big(f(x_0) - f_\star\big) + L\sigma^2 \sum_{t=0}^{T-1} \eta_t^2 \\
&= 2\Delta_0 + L\sigma^2 \sum_{t=0}^{T-1} \eta_t^2,
\end{aligned}$$

which is Eq. (65).

Let $S_T := \sum_{k=0}^{T-1} \eta_k$. Since R_T is sampled independently with $\mathbb{P}(R_T = t) = \eta_t / S_T$, the law of total expectation gives

$$\begin{aligned}\mathbb{E}\|\nabla f(x_{R_T})\|^2 &= \sum_{t=0}^{T-1} \mathbb{P}(R_T = t)\mathbb{E}\|\nabla f(x_t)\|^2 \\ &= \frac{1}{S_T}\sum_{t=0}^{T-1} \eta_t \mathbb{E}\|\nabla f(x_t)\|^2 \\ &\le \frac{2\Delta_0 + L\sigma^2 \sum_{t=0}^{T-1} \eta_t^2}{S_T},\end{aligned}$$

which proves Eq. (66). Under the stated summability conditions, the numerator is bounded while $S_T \to \infty$, so the bound converges to zero. □

B.4 Schedule-Free

B.4.1 Quadratic Variance Reduction

Theorem B.8 (Variance decay of the Schedule-Free averaged sequence)**.** *Consider the one-dimensional quadratic objective $g(x) = \frac{1}{2}dx^2$, $d > 0$, with stochastic gradient $g'(x) + \xi_t = dx + \xi_t$, where $\xi_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. Let (y_t, z_t, x_t) follow the Schedule-Free recursion in Eq. (14) from deterministic initial values. If*

$$0 \le \beta < 1, \qquad a := 1 - \gamma d(1-\beta) \in (0, 1), \tag{70}$$

then the averaged sequence $\{x_t\}$ satisfies

$$\mathbb{E}[x_t] = O(t^{-1}), \qquad \mathrm{Var}(x_t) = O(t^{-1}). \tag{71}$$

Consequently, $\mathbb{E}[x_t^2] = O(t^{-1})$, $x_t \xrightarrow{L^2} 0$, $x_t \xrightarrow{P} 0$, and $\mathcal{L}(x_t) \Rightarrow \delta_0$.

Proof. Set

$$r = \frac{\beta}{1-\beta}, \qquad a = 1 - \gamma d(1-\beta), \qquad u_t = (1-a)tx_t + az_t. \tag{72}$$

Since $\gamma d\beta = (1-a)r$,

$$z_{t+1} = az_t - (1-a)rx_t - \gamma\xi_t, \tag{73}$$

and $(t+1)x_{t+1} = tx_t + z_{t+1}$ gives

$$u_{t+1} = u_t - (1-a)rx_t - \gamma\xi_t. \tag{74}$$

Using $x_t = (u_t - az_t)/((1-a)t)$,

$$\begin{pmatrix} u_{t+1} \\ z_{t+1} \end{pmatrix} = M_t \begin{pmatrix} u_t \\ z_t \end{pmatrix} - \gamma\xi_t \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad M_t = \begin{pmatrix} 1 - r/t & ar/t \\ -r/t & a(1+r/t) \end{pmatrix}. \tag{75}$$

For $q = r/t$ and $v = (u, z)^\top$,

$$\begin{aligned}\|M_t v\|_2^2 &= ((1-q)u + aqz)^2 + (-qu + a(1+q)z)^2 &(76)\\ &= (1-2q)u^2 + a^2(1+2q)z^2 + 2q^2(u - az)^2 &(77)\\ &\le (1 - 2q + 4q^2)u^2 + a^2(1 + 2q + 4q^2)z^2. &(78)\end{aligned}$$

Because $q \to 0$ and $a^2 < 1$, there is a fixed T such that $\|M_t v\|_2 \le \|v\|_2$ for all $t \ge T$. Applying this to the mean recurrence bounds $\mathbb{E}[u_t]^2 + \mathbb{E}[z_t]^2$ and, by Cauchy–Schwarz, gives $\mathbb{E}[x_t] = O(t^{-1})$.

For centered variables $\hat{u}_t$ and $\hat{z}_t$, define $Q_t = \mathbb{E}[\hat{u}_t^2 + \hat{z}_t^2]$. Independence and zero-mean noise give

$$Q_{t+1} \le Q_t + 2\gamma^2\sigma^2, \tag{79}$$

so $Q_t = O(t)$. Again by Cauchy–Schwarz,

$$\mathrm{Var}(x_t) \le \frac{1 + a^2}{(1-a)^2 t^2} Q_t = O(t^{-1}). \tag{80}$$

Hence $\mathbb{E}[x_t^2] = O(t^{-1}) \to 0$ and $x_t \xrightarrow{L^2} 0$, $x_t \xrightarrow{P} 0$, and $\mathcal{L}(x_t) \Rightarrow \delta_0$. □

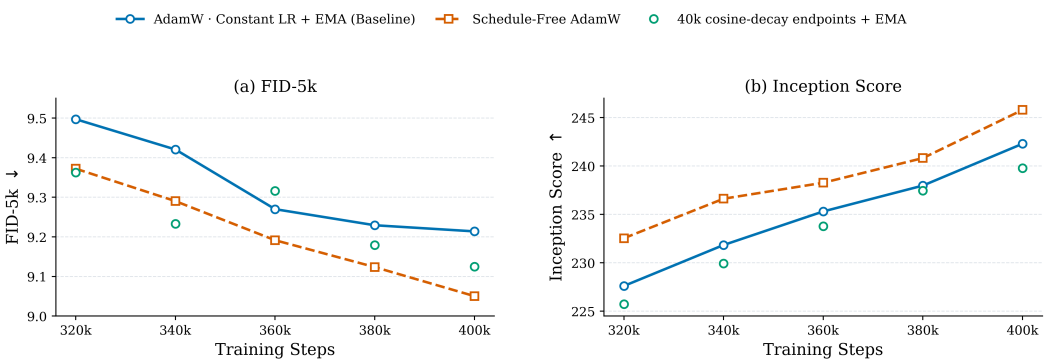

Figure 8: Late-stage generation quality for CONSTANT LR + EMA, SCHEDULE-FREE ADAMW, and 40k DECAY + EMA. (a) FID-5K, where lower is better. (b) Inception Score, where higher is better. The CONSTANT LR + EMA and SCHEDULE-FREE ADAMW curves follow continuous training trajectories. Green markers denote endpoints of independent 40k-step cosine-decay branches and should not be interpreted as a continuous trajectory. SCHEDULE-FREE ADAMW outperforms CONSTANT LR + EMA on both metrics at every checkpoint from 320k to 400k.

B.4.2 SMOOTH NONCONVEX STATIONARITY

The following theorem records the corresponding guarantee beyond the quadratic setting.

Theorem B.9 (Nonconvex stationarity of Schedule-Free SGD)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and bounded below by $f_\star$. Suppose that, conditionally on the history $\mathcal{F}_t$, the stochastic gradient evaluated at y_t satisfies*

$$\mathbb{E}[g_t \mid \mathcal{F}_t] = \nabla f(y_t), \qquad \mathbb{E}\big[\|g_t - \nabla f(y_t)\|^2 \mid \mathcal{F}_t\big] \le \sigma^2. \tag{81}$$

Starting from $x_0 = z_0$, run

$$y_t = (1-\beta)z_t + \beta x_t, \quad z_{t+1} = z_t - \gamma_t g_t, \quad x_{t+1} = (1-c_{t+1})x_t + c_{t+1} z_{t+1}, \tag{82}$$

where $0 \le \beta < 1$, $0 < \gamma \le 1/L$, and, for a warmup length $1 \le T_{\mathrm{w}} \le T$ and a constant $C > 0$,

$$\gamma_t = \gamma \min\left\{1, \frac{t+1}{T_{\mathrm{w}}}\right\}, \qquad c_{t+1} = \min\left\{\frac{(1-\beta)C\gamma_t^2}{\sum_{i=0}^{t}\gamma_i^2}, 1\right\}. \tag{83}$$

Define $p_{\mathrm{w}} := (T_{\mathrm{w}} - 1)/T$. Then

$$\begin{aligned}\min_{0\le t<T} \mathbb{E}\|\nabla f(y_t)\|^2 \le{}& \frac{2\big(f(x_0) - f_\star\big)}{\gamma(1-\beta)(1-p_{\mathrm{w}})T} + \frac{\gamma L(1-\beta)\sigma^2}{1-p_{\mathrm{w}}} \\ &+ \frac{3\beta(2-\beta)C\sigma^2(1+\log T)}{(1-p_{\mathrm{w}})T}.\end{aligned} \tag{84}$$

This result controls the gradient-evaluation points y_t and does not claim last-iterate convergence of the evaluation sequence x_T. Its proof is given by Chae & Kim (2026, Theorem 4.4).

C ADDITIONAL EXPERIMENTAL RESULTS

C.1 ADDITIONAL LATE-STAGE COMPARISONS

We jointly report FID-5K and Inception Score at late-stage checkpoints, followed by the fixed-evaluation loss components for the same optimization strategies. SCHEDULE-FREE ADAMW has lower FID-5K and higher Inception Score than CONSTANT LR + EMA at every checkpoint from 320k to 400k, while the independently branched cosine-decay endpoints vary more with decay timing.

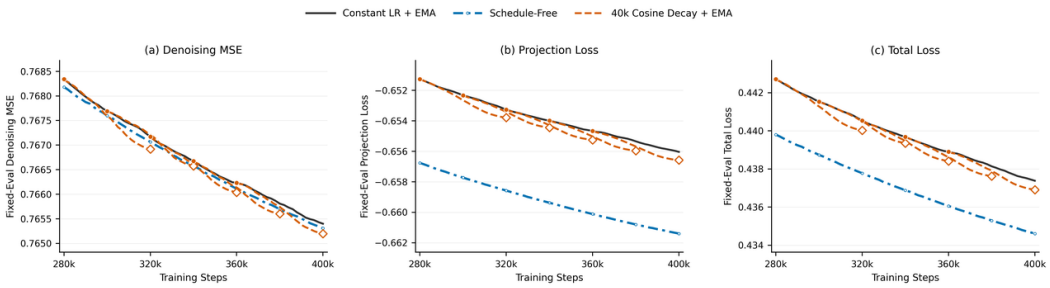

Figure 9: Fixed-evaluation objectives for CONSTANT LR + EMA, SCHEDULE-FREE ADAMW, and 40k DECAY + EMA. (a) Denoising MSE is similar across the three strategies. (b) SCHEDULE-FREE ADAMW maintains a distinctly lower projection loss. (c) Consequently, SCHEDULE-FREE ADAMW also achieves the lowest total fixed-evaluation loss throughout the evaluated interval. The cosine-decay values are endpoints of independent 40k-step branches; dashed segments connect them only as a visual guide.

C.2 MAIN FID-5K RESULTS

Table 7 reports the exact generation-quality measurements underlying the comparisons in Sections 3.2–3.4, together with the corresponding Constant LR and cosine-decay AdamW results without EMA. The first three blocks are checkpoints from continuous training trajectories. Each entry in either cosine-decay block is the endpoint of a separate 40k-step branch initialized from the checkpoint shown on the left of the arrow. The decay rows therefore compare with the Constant LR + EMA checkpoint at the same ending step, rather than forming a single trajectory.

Table 7: Complete generation-quality results for the main experiment, including Constant LR AdamW without EMA and 40k-step cosine-decay AdamW evaluated with and without EMA. The first three methods follow continuous trajectories; every cosine-decay result is the endpoint of an independent branch from a different late-stage checkpoint. FID-5K entries are means over three sampling seeds. Lower FID-5K and sFID and higher Inception Score are better.

Method	Checkpoint / branch	FID-5K	IS	sFID
SCHEDULE-FREE ADAMW	320k	9.3722	232.5443	35.1116
	340k	9.2903	236.6247	35.1530
	360k	9.1916	238.2731	35.1622
	380k	9.1237	240.8229	35.1441
	400k	9.0504	245.7993	35.1334
CONSTANT LR + EMA	320k	9.4972	227.6105	35.0881
	340k	9.4207	231.8407	35.2050
	360k	9.2700	235.3101	35.1960
	380k	9.2294	237.9821	35.2019
	400k	9.2139	242.3034	35.1140
CONSTANT LR	320k	13.7341	163.1902	38.1238
	340k	13.8222	168.1524	40.5599
	360k	11.5768	182.6766	35.3735
	380k	11.0776	190.8148	35.0794
	400k	11.1509	192.8839	36.3858
LR DECAY	280k→320k	9.7146	219.9526	35.7952
	300k→340k	9.5220	225.3971	35.6992
	320k→360k	9.6087	229.0856	35.7077
	340k→380k	9.5660	231.1641	35.8078
	360k→400k	9.4895	231.0582	36.0005
DECAY + EMA	280k→320k	9.3629	225.7234	35.5324
	300k→340k	9.2330	229.9364	35.5310
	320k→360k	9.3163	233.7810	35.4139
	340k→380k	9.1795	237.4666	35.4731
	360k→400k	9.1248	239.7905	35.6127

Schedule-Free improves FID-5K over Constant LR + EMA at all five checkpoints, with reductions ranging from 0.0784 to 0.1635, and also yields a higher Inception Score throughout. Its sFID changes are small and not uniformly favorable, so the improvement is not identical across every metric. Without EMA, cosine decay substantially improves FID-5K over the corresponding Constant LR checkpoints at all five endpoints. Cosine Decay + EMA is more timing-sensitive relative to the Constant LR + EMA baseline: four of the five branches improve endpoint FID-5K, whereas the branch starting at 320k is worse at 360k (9.3163 versus 9.2700). The complete measurements thus support the main text's distinction between the consistent gains of Schedule-Free and the start-time-dependent gains of late-stage decay.

C.3 TIMESTEP-WISE DENOISING ERROR

To test whether the aggregate denoising improvement hides a regression at any particular diffusion noise level, we evaluate the two methods under the fixed-evaluation protocol of Appendix A.1. We reuse the same evaluation data and noise for both methods and vary only the timestep, thereby removing sampling variation from the pairwise comparison. We test $t \in \{0.02, 0.10, 0.20, 0.40, 0.60, 0.80, 0.90, 0.96\}$ and additionally report the result under randomly sampled timesteps. The relative improvement in the final column is $100(\mathcal{L}_{\text{Constant+EMA}} - \mathcal{L}_{\text{SF}})/\mathcal{L}_{\text{Constant+EMA}}$, so a positive value favors Schedule-Free.

Table 8: Timestep-wise denoising MSE. Relative improvement is for Schedule-Free versus Constant + EMA.

Timestep	Constant + EMA	Schedule-Free	Relative improvement
random t	0.766211	**0.765308**	0.12%
$t = 0.02$	1.016053	**1.013539**	0.25%
$t = 0.10$	0.949475	**0.947750**	0.18%
$t = 0.20$	0.868994	**0.867809**	0.14%
$t = 0.40$	0.796613	**0.795586**	0.13%
$t = 0.60$	0.709609	**0.708319**	0.18%
$t = 0.80$	0.615276	**0.614681**	0.10%
$t = 0.90$	0.601724	**0.601448**	0.05%
$t = 0.96$	0.616707	**0.616376**	0.05%

Schedule-Free has lower denoising MSE in all eight fixed-timestep evaluations and under random timestep sampling. The relative reduction ranges from 0.05% to 0.25%: it is largest at $t = 0.02$, remains positive across the intermediate noise levels, and is still positive at $t = 0.90$ and $t = 0.96$. The randomly sampled evaluation improves by 0.12%. Thus, although the denoising differences are modest, their direction is consistent across the diffusion trajectory rather than being driven by a single timestep.

D Discussion

Scaling has long been an important axis in diffusion-model development. Diffusion Transformers improve systematically as model capacity and compute increase: increasing transformer depth, width, or the number of input tokens can lead to consistent gains in generation quality (Peebles & Xie, 2023). More recent work has further studied scaling laws for diffusion models across model size, data, compute, and pretraining performance (Liang et al., 2024; Li et al., 2024). These results motivate a related question: whether the relative effectiveness of optimization strategies may also change across model and training scales.

This question is particularly relevant when comparing our results with the recent controlled benchmark of diffusion optimizers in Bolatov et al. (2026). At the end of the benchmark budget, Schedule-Free ranks last on the primary metric for text8, QM9, and CelebA-64. On LM1B, Lion diverges on two of three seeds, while Schedule-Free finishes last among the stable runs. This aggregate result, however, hides several important differences. On QM9, Schedule-Free is last in perplexity but first in Fréchet ChemNet Distance. The experimental regimes also differ substantially from ours. Three of the four tasks in the benchmark use discrete diffusion—masked diffusion on text8, uniform diffusion on QM9, and uniform diffusion trained through Gaussian duality on LM1B—while the only Gaussian image-diffusion experiment uses a 78.7M-parameter U-Net on CelebA-64. The other models contain 16.6M, 90.2M, and 137.1M parameters for text8, QM9, and LM1B, respectively. In addition, the text8 and LM1B runs use 30k training steps rather than the 1M-step budgets used in the corresponding published studies, making the reported LM1B ordering provisional. By contrast, our main experiments train a 675M-parameter SiT-XL/2 Diffusion Transformer on ImageNet-1K at 256×256 resolution using flow matching and REPA (Ma et al., 2024; Yu et al., 2025).

The optimizer dynamics also differ between the two studies. In the benchmark, all competing non-Schedule-Free optimizers are paired with a late-stage learning-rate decay, whereas Schedule-Free retains a constant base learning rate after warmup. On text8, Schedule-Free performs strongly during intermediate training but is overtaken during the final decay phase. A similar late-stage crossover appears on QM9. In our SiT+REPA setting, the ordering is reversed: explicit decay reduces parameter variability more aggressively, whereas Schedule-Free preserves more late-stage objective progress and achieves better final generation quality. This contrast suggests that the relative ranking of optimization strategies can depend strongly on the surrounding training regime.

At the same time, the available evidence does not identify model size or dataset size as the cause of this reversal. The two studies differ not only in scale, but also in diffusion formulation, architecture,

training objective, learning-rate protocol, training budget, and data domain. The current results therefore do not support attributing the difference in optimizer ordering to any single factor.

More broadly, these observations raise the possibility that optimization strategies themselves may exhibit scale-dependent empirical behavior. EMA timescales, decay schedules, and Schedule-Free averaging may interact differently with training as model size, dataset size, and optimization budget change. A controlled study that varies model and data scale while independently retuning each optimization strategy would be needed to determine whether optimizer rankings follow a systematic scaling pattern. We leave this question for future work.

E Additional Qualitative Results

This section complements the quantitative evaluations with samples from the three training regimes studied in the paper. Section E.1 shows the main Schedule-Free pretraining result, Section E.2 compares ReFL optimizers under fixed prompts, and Section E.3 shows samples from the generator-side Schedule-Free DMD experiment.

E.1 Main Pretraining: Schedule-Free AdamW

Figure 10 provides representative class-conditional samples from the 800k-step Schedule-Free checkpoint evaluated quantitatively in Section 3.3. The grid spans animals, people, vehicles, objects, and natural or built scenes rather than selecting a single semantic category. These samples serve as a qualitative complement to the FID-50K, sFID, and Inception Score reported in the main text.

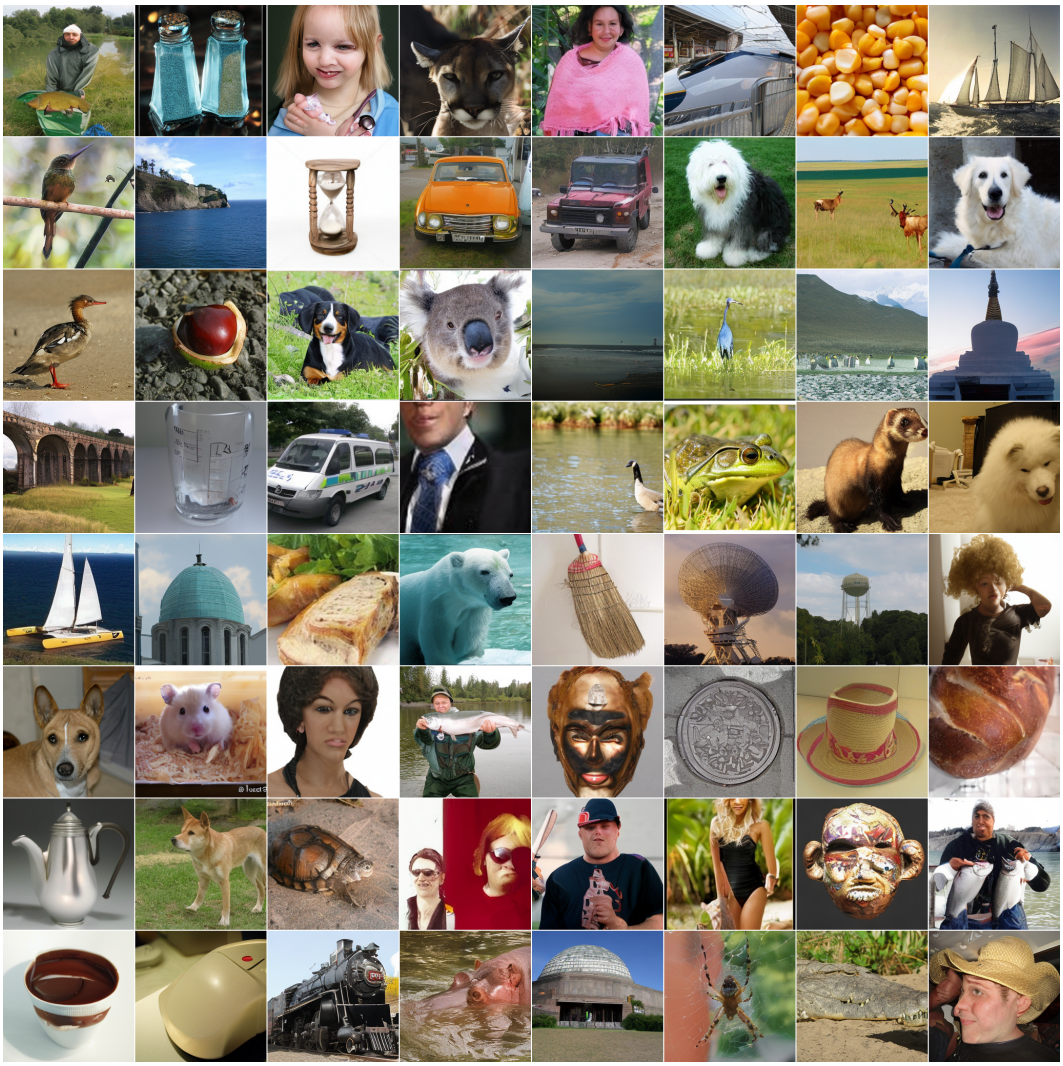

Figure 10: Selected class-conditional ImageNet-1K samples from the REPA-trained SiT-XL/2 model optimized with SCHEDULE-FREE ADAMW at $4\times$ the baseline learning rate after 800k training steps. Images are generated at 256×256 resolution.



Figure 11: Additional fixed-prompt ReFL qualitative comparison (prompt set 1). The rows compare the base model, AdamW, and Schedule-Free AdamW at two learning-rate scales; the complete prompts are printed above the columns.

E.2 ReFL Qualitative Comparisons

We provide additional fixed-prompt comparisons for ReFL. In every figure, the rows show, from top to bottom, the Z-Image base model without reinforcement learning, AdamW at $1\times$ the reference learning rate with HPSv3, Schedule-Free AdamW at $4\times$, and Schedule-Free AdamW at $20\times$. Each column uses the prompt printed above it.

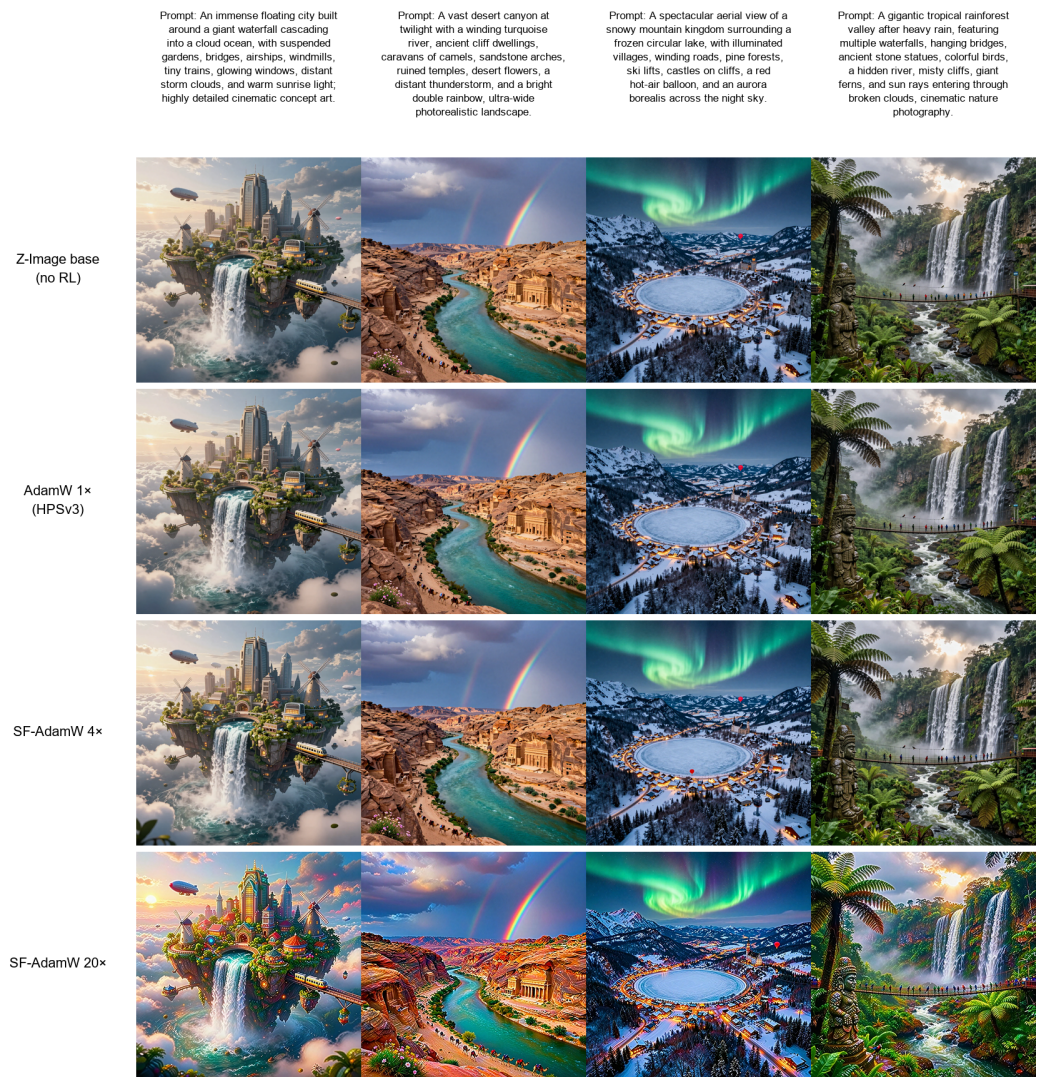

Figure 12: Additional fixed-prompt ReFL qualitative comparison (prompt set 2), using the same row ordering as Figure 11.

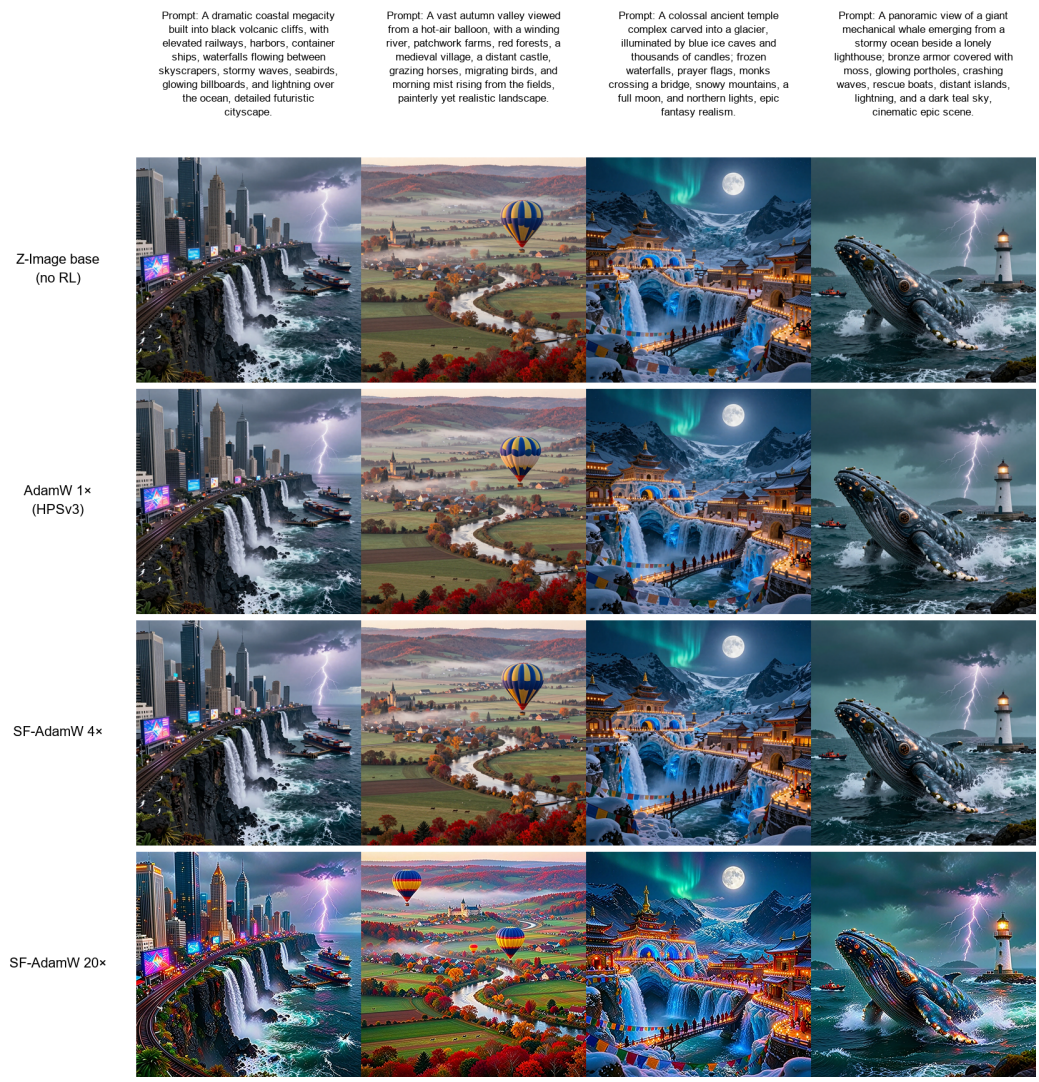

Figure 13: Additional fixed-prompt ReFL qualitative comparison (prompt set 3), using the same row ordering as Figure 11.

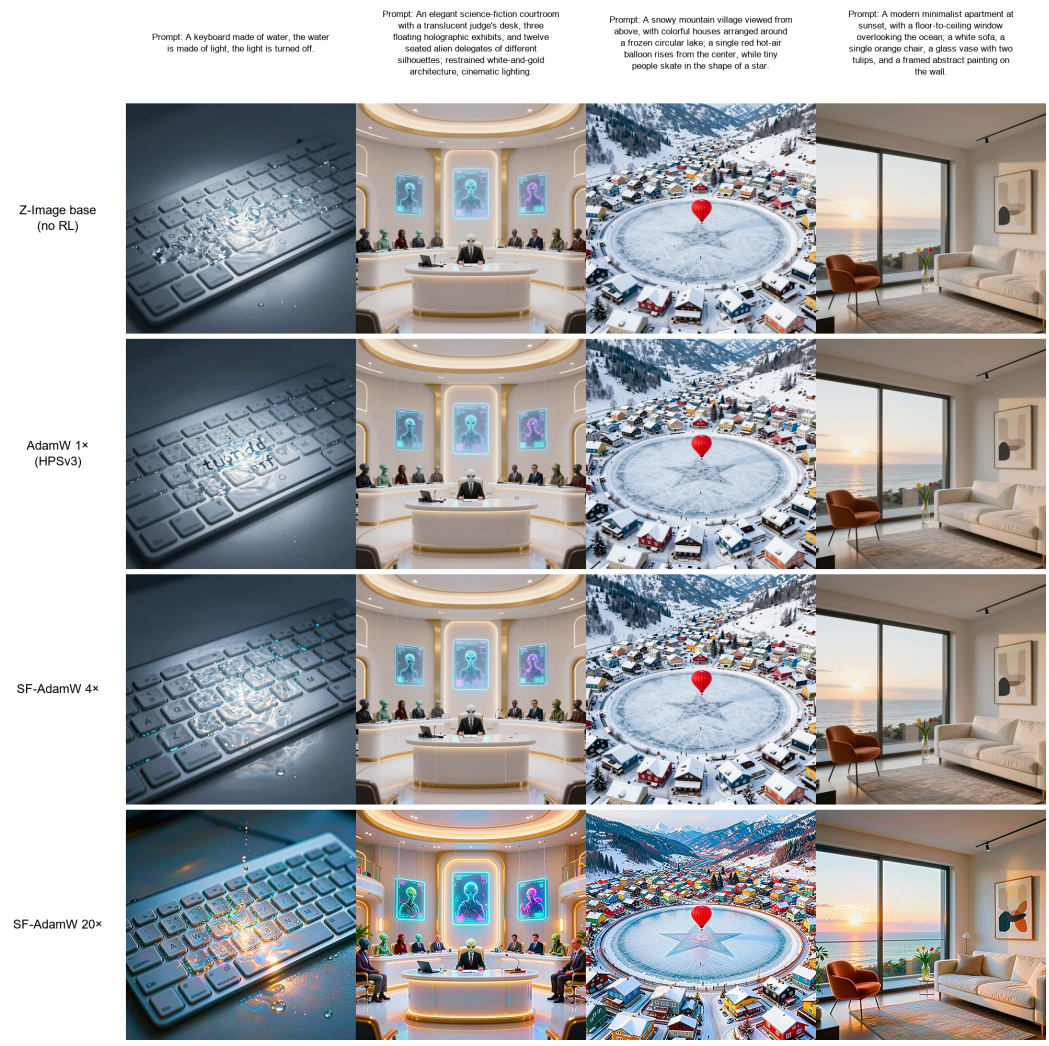

Figure 14: Additional fixed-prompt ReFL qualitative comparison (prompt set 4), using the same row ordering as Figure 11.

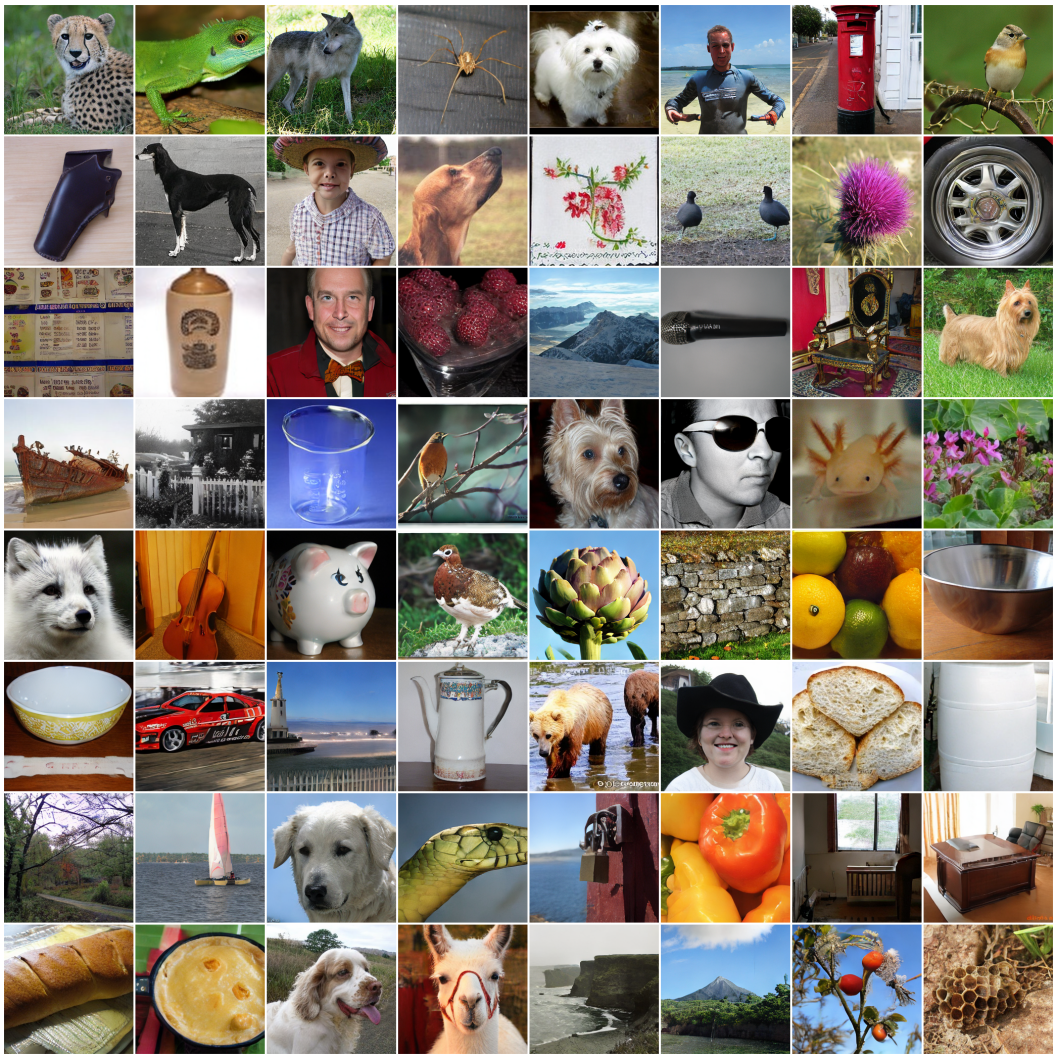

Figure 15: Selected class-conditional ImageNet samples from the best-performing DMD optimizer variant after 3k generator steps. SCHEDULE-FREE ADAMW is applied to the generator, while the guidance model retains AdamW. The figure shows an 8×8 grid of samples at 256×256 resolution.

E.3 DMD: SCHEDULE-FREE GENERATOR SAMPLES

Figure 15 visualizes the best-performing variant in the DMD optimizer ablation: Schedule-Free AdamW is applied to the generator, while the guidance model retains the baseline AdamW optimizer. At 3k generator steps, this configuration achieves the lowest FID-5K among the three variants in Figure 5 and Table 6. The samples therefore provide a qualitative view of the generator-side Schedule-Free improvement discussed in Section 3.5; the complete optimizer settings are given in Appendix A.6.