FASE: FACTOR-AWARE SPECTRAL EVOLUTION FOR GRADIENT-FREE CONTINUATION OF DIFFUSION POST-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reward-based diffusion post-training methods such as Flow-GRPO and DiffusionNFT rely on backpropagation through the generator. We propose Factor-Aware Spectral Evolution (**FASE**), a gradient-free continuation stage that starts from a gradient-trained, reward-adapted DiffusionNFT low-rank adaptation (LoRA) and performs no further backpropagation. Our design tests a constrained continuation hypothesis motivated by spectral diagnostics on Stable Diffusion 3.5 Medium (SD3.5-M): measured changes in A's row space and B's column space decrease early while ordered singular values continue to evolve, and the two factors exhibit unequal magnitudes of spectral change under the training parameterization. FASE instantiates this hypothesis by freezing the specific handoff SVD bases and optimizing their singular values using separable CMA-ES. A factor-aware adaptive controller reallocates exploration between the two factors using whitened CMA-ES mean steps while preserving total coordinate variance. Starting from task-specific DiffusionNFT checkpoints, FASE reaches GenEval 0.98, OCR 0.98, and PickScore 24.25; reference target scores from Flow-GRPO and complete DiffusionNFT runs trained from the base model provide endpoint context, not a matched from-scratch comparison. Controlled ablations isolate the adaptive factor-allocation controller by comparing it with uniform exploration and a fixed non-uniform A/B allocation from the same warm-start under matched population and generation budgets. Experiments across GenEval, OCR, and PickScore show that FASE effectively continues reward-adapted diffusion models without further backpropagation, improving all three target rewards and achieving the best observed endpoints among matched spectral-continuation variants.

1 INTRODUCTION

Reward-based post-training improves the compositional accuracy, text rendering, and preference alignment of diffusion and flow-matching (Lipman et al., 2022) generators (Black et al., 2024). Flow-GRPO (Liu et al., 2025) optimizes a stochastic reverse process with group-relative policy gradients (Shao et al., 2024), whereas DiffusionNFT (Zheng et al., 2026) learns from rewarded samples through a forward-process objective. Both update the generator by backpropagation. We ask whether reward optimization can continue effectively after a gradient-trained Low-Rank Adaptation (LoRA) (Hu et al., 2021) warm-start without further backward passes through the generator.

Evolution strategies (ES) (Salimans et al., 2017; Wierstra et al., 2014) offer a gradient-free approach: they score a population of perturbed candidates and update the search using those rewards. Recent work applies ES to reward-based language-model adaptation (Qiu et al., 2025; Korotyshova et al., 2025). For diffusion models, Calibri (Tokhchukov et al., 2026) uses CMA-ES to optimize model scaling coefficients from black-box rewards. Each diffusion candidate requires multi-step sampling, making high-dimensional full-backbone search costly. The challenge is to identify a compact, reward-adapted search space and allocate exploration effectively within it.

We propose Factor-Aware Spectral Evolution (**FASE**; Fig. 1), a gradient-free continuation stage for reward-based diffusion post-training. FASE starts from a task-specific, gradient-trained, reward-adapted DiffusionNFT LoRA; only the subsequent spectral search omits backpropagation through

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.

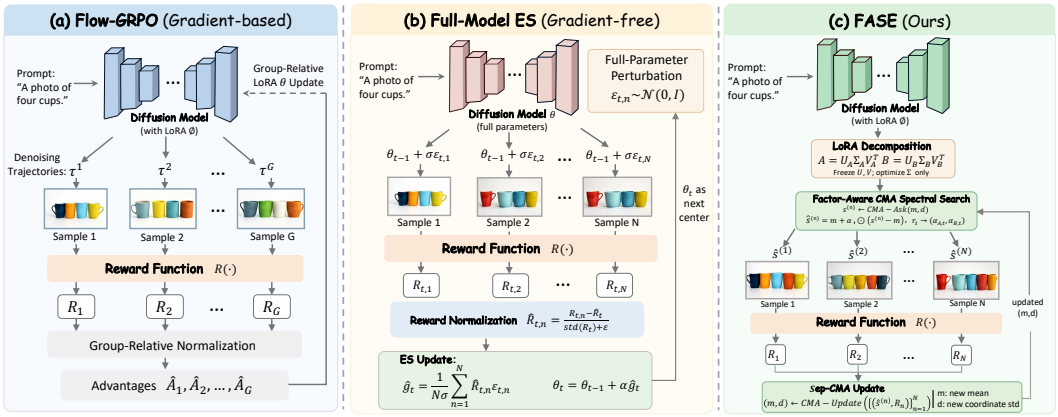


Figure 1: **Overview of reward-based diffusion post-training.** (a) Flow-GRPO updates the generator using group-relative policy gradients. (b) Full-Model ES perturbs generator parameters and updates them from population rewards (Algorithm 1). (c) As gradient-free continuation from a gradient-trained, reward-adapted DiffusionNFT LoRA handoff, FASE fixes the handoff factor SVD bases and evolves singular values with adaptive A/B exploration (Algorithm 2).

the generator. Its target-reward endpoints are reported as context against from-base baselines, not as a matched from-scratch superiority claim. FASE searches spectral coefficients within reward-adapted LoRA factor bases and allocates exploration with a variance-preserving adaptive A/B controller.

LoRA writes the update as a product of two low-rank factors, A and B. The singular value decomposition of each factor separates its left and right singular-vector bases, denoted U, V, from the singular values that scale their paired directions. A randomly initialized LoRA has not yet formed reward-adapted bases. Spectral approaches use pretrained weight components, as in PiSSA (Meng et al., 2024), or search within frozen bases from a supervised language-model LoRA, as in ESSA (Korotyshova et al., 2025). For reward-based diffusion adaptation, the questions are whether changes in the task-relevant factor subspaces diminish and whether spectral search within a selected handoff factorization remains useful.

We examine LoRA checkpoints during Flow-GRPO and DiffusionNFT training on Stable Diffusion 3.5 Medium (SD3.5-M) (Esser et al., 2024), using GenEval (Ghosh et al., 2023) (composition), OCR (Du et al., 2020) (visual text), and PickScore (Kirstain et al., 2023) (preference). Two patterns emerge. First, the measured changes in A's row space and B's column space (denoted D^A_{UV} and D^B_{UV} in Sec. 3.2) drop early, while task scores and ordered singular values continue to change. Second, under the training parameterization, A and B show unequal residual singular-value velocities. Together, these diagnostics motivate testing a constrained frozen-basis continuation hypothesis with factor-specific exploration. FASE's use of the exact handoff SVD bases is a design choice: the diagnostics neither select a uniquely preferred basis nor prove that individual singular vectors have converged.

FASE loads a DiffusionNFT LoRA, freezes its specific handoff SVD bases, and packs the singular values for optimization with separable Covariance Matrix Adaptation Evolution Strategy (CMA-ES) (Hansen & Ostermeier, 2001; Ros & Hansen, 2008). The gradient-based diagnostics motivate factor-specific exploration but do not determine the appropriate allocation during ES. An adaptive A/B controller therefore updates the exploration ratio from reward-guided, whitened CMA-ES mean steps and maps it to scales that preserve the total coordinate variance of the current singular-value distribution. Uniform spectral ES (LoRA+ES) uses the same search with the controller disabled.

On SD3.5-M, FASE reaches GenEval 0.98 (DiffusionNFT 0.98, Flow-GRPO 0.97), OCR 0.98 (Flow-GRPO 0.96, DiffusionNFT 0.97), and PickScore 24.25 (Flow-GRPO 23.62, DiffusionNFT 24.03; Table 1). FASE starts from task-specific DiffusionNFT checkpoints, whereas these gradient baselines train from the base model, so their scores are reference endpoints rather than a matched pipeline comparison. Separately, we compare Uniform (LoRA+ES), Static A/B, and FASE from the same checkpoints under matched population and generation budgets (Table 2). FASE's endpoint gains are small point estimates, and we do not claim statistical significance.

Our contributions are threefold:

- We document early decreases in measured task-relevant factor-subspace changes, continuing singular-value evolution, and parameterization-dependent A/B spectral asymmetry across two gradient post-training methods and three rewards on SD3.5-M.
- We introduce FASE, which combines spectral search in reward-adapted LoRA bases with an adaptive A/B controller that reallocates exploration while preserving total coordinate variance. After the gradient-trained DiffusionNFT handoff, the continuation requires no backpropagation through the generator.
- We report target-reward continuation on three tasks and controlled same-warm-start evaluation of the controller; endpoint differences are small point estimates and trajectories are descriptive.

2 Related Work

Reward-based diffusion post-training. DDPO optimizes image rewards through policy gradients over the denoising process (Black et al., 2024). On language models, group-relative policy optimization (GRPO) compares sampled responses within a group and updates the model by a policy gradient (Shao et al., 2024). Flow-GRPO brings that update to flow-matching generators through a stochastic sampling process (Liu et al., 2025), and DiffusionNFT learns from rewarded samples with a forward-process objective (Zheng et al., 2026). These methods update the generator with gradients. FASE instead denotes the continuation from a gradient-trained, reward-adapted DiffusionNFT LoRA: candidates are scored by forward generation, and this continuation update does not backpropagate through the generator.

Search in structured parameter spaces. ES scores a parameter distribution from population rewards (Wierstra et al., 2014; Salimans et al., 2017). On language models it can match or exceed GRPO-style RL when the search geometry is right (Qiu et al., 2025). ESSA freezes the singular vectors of each LoRA factor and evolves the singular values with full-covariance CMA-ES after a supervised warm-start on language models (Korotyshova et al., 2025). SVF and SVDiff optimize singular values in fixed pretrained bases (Sun et al., 2022; Han et al., 2023), and PiSSA fixes the subspace to the principal components of pretrained weights (Meng et al., 2024). Hyper-ES merges GRPO-style descent directions and searches combination weights (Gu et al., 2026), a different geometry from freezing LoRA singular vectors. LoRA+ and related analyses show that A and B can warrant different treatment in gradient learning (Hayou et al., 2024; Zhu et al., 2024). FASE uses that asymmetry during ES: an adaptive controller sets the A/B exploration ratio from whitened CMA-ES steps and rescales the two factors so that the total variance of the current singular-value distribution is preserved.

3 Method

3.1 Preliminaries

Evolution strategies. ES optimizes a black-box objective by evaluating a population of randomly perturbed candidate parameters (Salimans et al., 2017; Wierstra et al., 2014). Given parameters $\boldsymbol{\theta}$ and a reward $R(\boldsymbol{\theta})$, ES considers the Gaussian-smoothed objective

$$J_\sigma(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[R\left(\boldsymbol{\theta}+\sigma\boldsymbol{\epsilon}\right)\right], \tag{1}$$

where σ is the perturbation scale. Its gradient can be estimated from forward reward evaluations alone:

$$\widehat{\nabla}_{\boldsymbol{\theta}} J_\sigma(\boldsymbol{\theta}) = \frac{1}{N\sigma}\sum_{n=1}^{N}\widehat{R}_n\boldsymbol{\epsilon}_n, \qquad \boldsymbol{\epsilon}_n \sim \mathcal{N}(\mathbf{0},\mathbf{I}), \tag{2}$$

where $\widehat{R}_n$ is the normalized reward of the n-th candidate. Parameters move toward perturbations associated with higher rewards (Algorithm 1). Reward normalization produces the practical search direction used here; it need not preserve an unbiased gradient estimate of the original reward objective. Each candidate is scored using forward generation only; normalization then forms $\widehat{R}_n$. Policy-gradient methods can also use non-differentiable rewards, but the update still requires a backward pass through the training objective.

3.2 Early Decrease in Measured Factor-Subspace Change

For a frozen backbone weight W, LoRA uses the update $\Delta W = (\alpha_{\mathrm{LoRA}}/r_{\mathrm{rank}})\, BA$, with $A \in \mathbb{R}^{r_{\mathrm{rank}} \times n}$ and $B \in \mathbb{R}^{m \times r_{\mathrm{rank}}}$ (Hu et al., 2021). Here r_{rank} is the LoRA rank and α_{LoRA} is the LoRA scale. In the experiments, the rank, the scale, and the adapted layers follow DiffusionNFT (Zheng et al., 2026). To characterize how these factors evolve during gradient-based reward fine-tuning, we decompose each factor at successive checkpoints using compact SVDs.

$$A = U_A \Sigma_A V_A^\top, \qquad B = U_B \Sigma_B V_B^\top. \tag{3}$$

For $F_t \in \{A_t, B_t\}$ the checkpoint SVD is $F_t = U_{F,t} \Sigma_{F,t} V_{F,t}^\top$. Here t indexes saved checkpoints (e.g., every 5 training steps), not every optimizer update. Between adjacent checkpoints we measure two complementary quantities on each factor. We define the factor-subspace change statistic as the averaged overlap deficit

$$D^F_{\mathrm{UV}}(t) = 1 - \tfrac{1}{2}\left(\frac{\left\| U_{F,t}[:, 1{:}k]^\top U_{F,t+1}[:, 1{:}k] \right\|_F^2}{k} + \frac{\left\| V_{F,t}[:, 1{:}k]^\top V_{F,t+1}[:, 1{:}k] \right\|_F^2}{k} \right), \tag{4}$$

where $k = r_{\mathrm{rank}}$ includes all singular values. If $A \in \mathbb{R}^{r_{\mathrm{rank}} \times n}$ and $B \in \mathbb{R}^{m \times r_{\mathrm{rank}}}$ both have full rank r_{rank}, then U_A and V_B each span all of $\mathbb{R}^{r_{\mathrm{rank}}}$, making their respective overlap terms identically one. Under this condition, D^A_{UV} is half the overlap deficit of A's row space (represented by V_A), and D^B_{UV} is half the overlap deficit of B's column space (represented by U_B). Neither statistic measures convergence of individual singular vectors or an exact SVD basis. The *spectral* change of ordered singular values is

$$D^F_{\Sigma}(t) = \frac{\left\| \sigma_{F,t+1} - \sigma_{F,t} \right\|_2}{\frac{1}{2}\left(\|\sigma_{F,t+1}\|_2 + \|\sigma_{F,t}\|_2 \right) + \varepsilon}, \tag{5}$$

where $\sigma_{F,t}$ denotes those k singular values of factor F in descending order. Thus D^F_{Σ} compares ordered spectra between checkpoints rather than coefficients matched to a fixed singular-vector basis. Layer-wise values are aggregated into the curves in Fig. 2, which plots D_{UV} and D_{Σ} computed from these SVDs. The A curves are multiplied by a display factor, marked in each legend, so that they share an axis with B; that multiplier is not part of the statistic.

Observation 1. As shown in Fig. 2, D^A_{UV} and D^B_{UV} drop early and remain near zero, while the task score continues to rise and D^B_{Σ} stays active for longer. D^A_{Σ} is smaller on the same axis, which is why the plots rescale it. Because the two diagnostics measure different quantities, their magnitudes do not decompose an update or attribute reward gains causally to singular-value changes.

Conclusion 1. Across Flow-GRPO and DiffusionNFT on all three tasks, these measured row- and column-space changes decrease early while ordered singular values and task scores evolve, motivating fixed-basis continuation without implying exact-vector or exact-basis convergence.

3.3 Residual Spectral Motion Differs Between Factors

The decrease in factor-subspace change does not mean that spectral refinement has stopped. To compare the remaining motion of A and B, we use the same checkpoints as Fig. 2 and measure the per-interval spectral velocity

$$v_F(t) = \mathrm{RMS}(\Delta \Sigma_F), \qquad F \in \{A, B\}, \tag{6}$$

where $\Delta \Sigma_F$ is the vector of changes in the ordered singular values between successive checkpoints, and RMS is taken over its entries. Each factor's share of the combined velocity is

$$\mathrm{share}_F(t) = \frac{v_F(t)}{v_A(t) + v_B(t)}. \tag{7}$$

These are labeled v_A share and v_B share in Fig. 3.

Observation 2. Figure 3 plots these shares alongside the task score for DiffusionNFT and Flow-GRPO on GenEval, OCR, and PickScore. Across both methods and all three rewards, B accounts for a larger share of the spectral velocity over much of the refinement stage, while task scores continue to change.

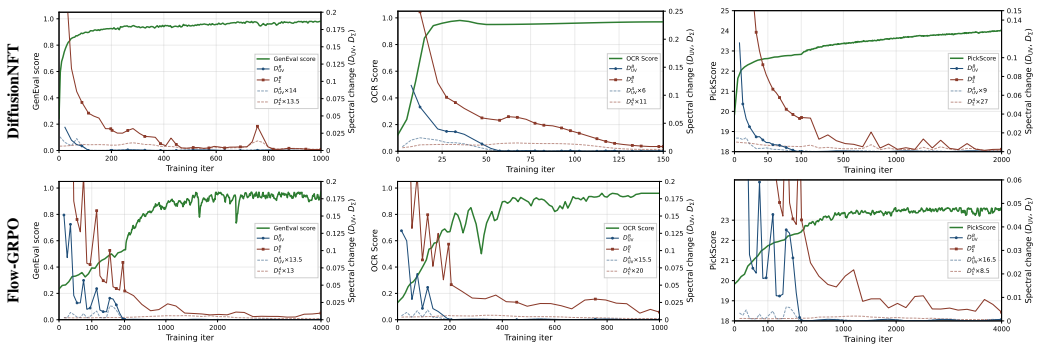


Figure 2: **Early decrease in measured factor-subspace change.** Per-factor spectral dynamics under DiffusionNFT (top) and Flow-GRPO (bottom). Solid curves are D^B_{UV} and D^B_Σ; dashed curves are D^A_{UV} and D^A_Σ, multiplied by the display factor printed in each legend. Measured changes in A's row space and B's column space decrease early, while B's ordered singular values continue to move.

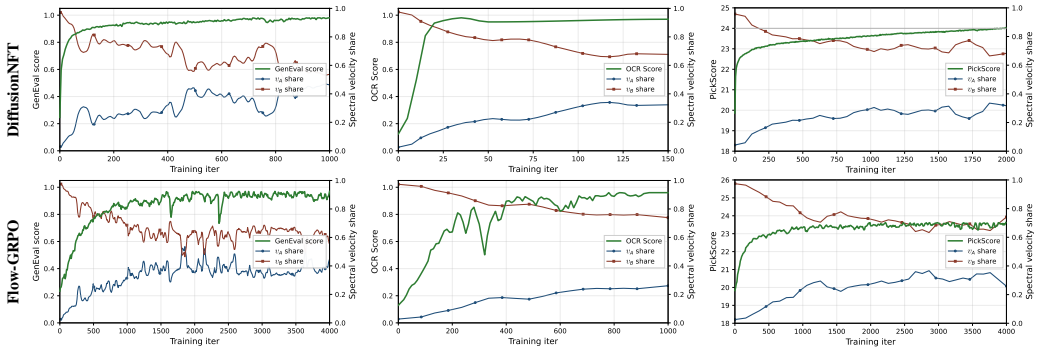


Figure 3: **Residual factor activity during gradient training.** Left axis: task score. Right axis: v_A share and v_B share, the shares share_A and share_B in Eqs. (6)–(7). Top: DiffusionNFT; bottom: Flow-GRPO; columns: GenEval, OCR, PickScore. Across both methods and all three rewards, the v_B share is larger over much of the refinement stage.

Conclusion 2. The two factors exhibit unequal residual spectral motion under the stored training parameterization. Since these velocities depend on factor scaling, they do not establish an intrinsic importance ordering or determine an optimal exploration ratio. This observation motivates testing unequal exploration of the two factors; FASE adjusts that allocation using feedback from the ES search.

3.4 Factor-Aware Spectral Evolution

FASE denotes only the continuation after a gradient-trained, reward-adapted LoRA checkpoint; our experiments use the task-specific DiffusionNFT handoffs in Sec. 4.1. The diagnostics motivate testing a constrained frozen-basis continuation hypothesis, but neither select nor establish convergence of an exact basis. FASE instantiates the hypothesis by freezing the specific handoff SVD bases and searching their spectral coefficients with separate exploration scales for A and B. An adaptive controller updates their ratio from CMA-ES mean steps while preserving the total coordinate variance within each generation (Algorithm 2).

Initialization. We compute the factor SVDs in Eq. (3), freeze their U, V matrices, and pack the singular values in $\mathbf{s}$ with separate A and B blocks. Because FASE can only rescale directions in these frozen bases, the selected warm-start defines its task-specific constrained search space; FASE then performs low-dimensional spectral refinement. Separable CMA-ES initializes $\mathbf{m} = \mathbf{s}$ and maintains per-coordinate standard deviations $\mathbf{d}$ with storage linear in the number of search coefficients (Hansen

& Ostermeier, 2001; Ros & Hansen, 2008). Adapter reconstruction retains the warm-start LoRA scaling convention.

Variance-preserving exploration. Let $r = \alpha_B/\alpha_A$ denote the exploration ratio, where α_A and α_B scale the two blocks' sampled perturbations. For each block, the CMA group energy $E_F = \sum_{i\in F} d_i^2$ is the sum of its current coordinate variances. We require

$$E_A\alpha_A^2 + E_B\alpha_B^2 = E_A + E_B, \tag{8}$$

so that changing the allocation does not change the total coordinate variance within that generation. Together with the ratio r, this gives

$$\alpha_A = \sqrt{\frac{E_A + E_B}{E_A + r^2 E_B}}, \qquad \alpha_B = r\alpha_A. \tag{9}$$

This constraint applies to the spectral search coordinates; it does not preserve the perturbation norm of the reconstructed weight update. Appendix B.1 gives the derivation.

For each of the N candidates drawn by CMA-Ask, we rescale its displacement from the current mean:

$$\hat{\mathbf{s}}^{(n)} = \mathbf{m} + \boldsymbol{\alpha} \odot \left(\mathbf{s}^{(n)} - \mathbf{m}\right), \tag{10}$$

where α_i is α_A or α_B according to the coordinate's block. Each candidate reconstructs the adapter using the frozen factor bases, generates a batch of 24 images, and receives a reward. We store the pre-update mean and coordinate scales as $\mathbf{m}^{\mathrm{pre}}$ and $\mathbf{d}^{\mathrm{pre}}$, then update CMA-ES with the stretched candidates and their rewards.

Backward-state requirements. During FASE continuation, candidate adapters are evaluated without backpropagation and require no parameter gradients, Adam moments, or denoiser activations retained for backward. Beyond ordinary inference state, streamed population evaluation needs only candidate coordinates or reproducible seeds, scalar rewards, and low-dimensional CMA-ES state. This is a structural backward-state property, not a measured reduction in peak GPU memory or an end-to-end efficiency claim; Appendix B.2 gives the complete accounting.

Adaptive allocation. After the CMA-ES update, the controller measures each block's mean displacement relative to the exploration scale applied in that generation:

$$u_F = \mathrm{RMS}\left(\frac{\Delta\mathbf{m}_F}{\mathbf{d}_F^{\mathrm{pre}}\,\alpha_F}\right), \qquad F \in \{A, B\}, \tag{11}$$

where $\Delta\mathbf{m} = \mathbf{m} - \mathbf{m}^{\mathrm{pre}}$ and division is elementwise. These whitened steps describe the reward-selected mean movement during ES, whereas the velocities in Sec. 3.3 describe changes during gradient training. Using raw mean steps would confound reward-selected motion with assigned exploration: a larger α_F or coordinate scale $\mathbf{d}_F^{\mathrm{pre}}$ mechanically produces a larger displacement and could create self-reinforcing feedback. The denominator in Eq. (11) compares movement in normalized search coordinates, while Eq. (9) prevents a ratio change from changing total coordinate variance. Thus the controller reallocates a fixed coordinate-variance budget rather than changing its total.

Every K generations, the controller averages the last K values of each u_F and exponentially smooths the log ratio of the B and A averages. A damping exponent and fixed bounds convert this state into the exploration ratio for the next generation. Algorithm 2 gives the exact update and state initialization; Appendix B.2 lists the hyperparameters. The initial ratio is shared across tasks rather than fitted to their gradient-training velocity shares.

This construction defines the three spectral-continuation settings in Sec. 4.3. *Uniform* (LoRA+ES) fixes $r \equiv 1$ (uniform A/B allocation) and disables the controller, yielding $\alpha_A = \alpha_B = 1$. *Static* is LoRA+ES with $r_0 = 5/3$ held fixed. We reserve *FASE* for the adaptive setting: Algorithm 2 starts at $r_0 = 5/3$ and updates r_t online. All three apply the variance-preserving scales and, within each task, use the same saved LoRA and ES generation budget. The returned adapter uses the optimized spectral coefficients in the frozen factor bases.

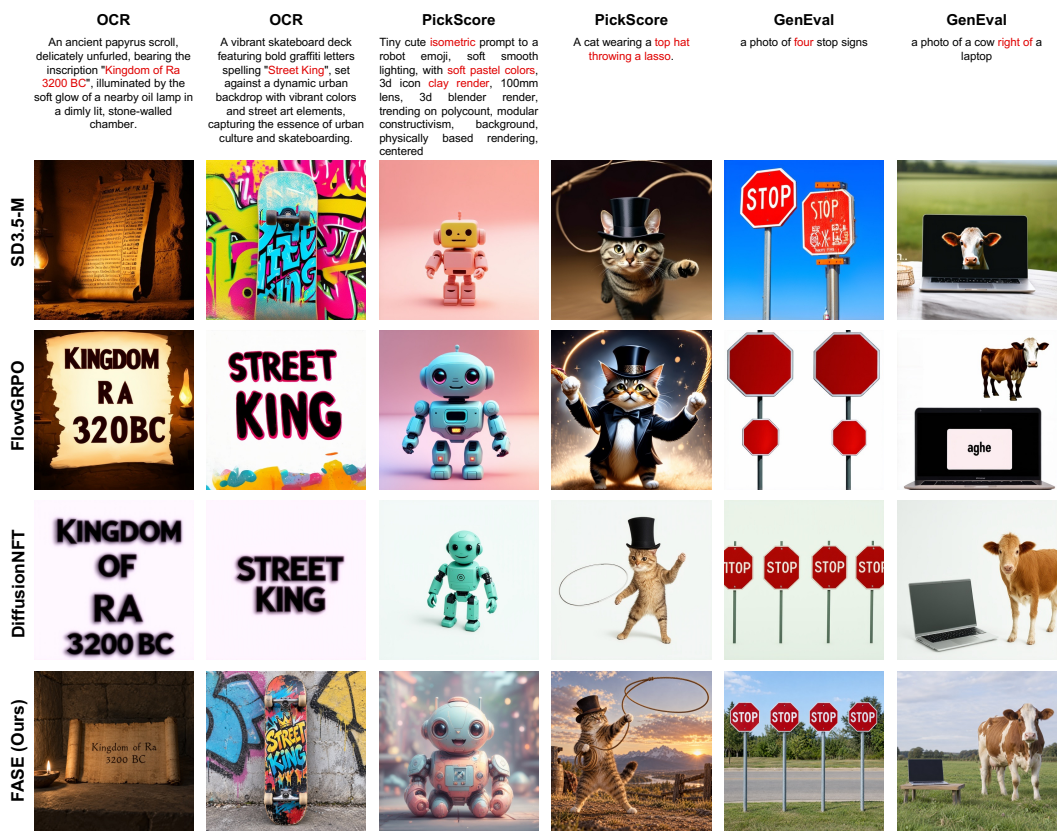


Figure 4: **Qualitative samples** on SD3.5-M. Each cell is one image for the prompt printed above it. Rows are SD3.5-M, Flow-GRPO, DiffusionNFT, and FASE. Flow-GRPO and DiffusionNFT use their from-base endpoints; FASE follows the task-specific DiffusionNFT handoff reported in Table 1. Columns are two prompts each from OCR, PickScore, and GenEval. Sampling uses guidance scale 1. Evaluation-set means, including OCR complementary scores, are in Table 1.

4 Experiments

4.1 Experimental Setup

Models and tasks. Experiments use Stable Diffusion 3.5 Medium (SD3.5-M) (Esser et al., 2024) at 512×512, following Flow-GRPO (Liu et al., 2025) and DiffusionNFT (Zheng et al., 2026). The LoRA rank, scale, and adapted layers follow DiffusionNFT. We optimize three rewards separately: GenEval (Ghosh et al., 2023) for composition, OCR with PaddleOCR (Du et al., 2020) for visual text, and PickScore (Kirstain et al., 2023) for learned preferences.

Evaluation metrics. We report GenEval, OCR, PickScore, CLIPScore (Hessel et al., 2021), HPSv2.1 (Wu et al., 2023), Aesthetic Score (Schuhmann, 2022), ImageReward (Xu et al., 2023), and UnifiedReward (Wang et al., 2025). Following DiffusionNFT (Zheng et al., 2026), model-based metrics are evaluated on DrawBench (Saharia et al., 2022), and classifier-free guidance (CFG) is re-enabled when evaluating non-OCR metrics for OCR-trained models.

Comparison protocols. Table 1 gives heterogeneous endpoint context: Flow-GRPO, DiffusionNFT, and Full-Model ES start from the base model; PiSSA+ES and Random LoRA+ES use their respective adapter initializations; FASE continues a task-specific DiffusionNFT LoRA with population $N = 12$ and 24 images per candidate. Its row is therefore not a matched end-to-end comparison. Table 2 instead controls the handoff and ES budget across Uniform (LoRA+ES), Static A/B, and FASE.

Table 1: Single-reward endpoint context on SD3.5-M (no CFG at training time). Each three-row block trains on GenEval, OCR, or PickScore (gray). CFG follows DiffusionNFT (Zheng et al., 2026) for non-OCR metrics of OCR-trained models. Flow-GRPO, DiffusionNFT, and Full-Model ES train from the base model. PiSSA+ES forms a LoRA from the pretrained weights and searches it with CMA-ES. Random LoRA+ES searches a randomly initialized LoRA with CMA-ES. FASE loads a DiffusionNFT LoRA (GenEval step 200, OCR step 50, PickScore step 100); its #Iter. counts ES generations after that load and does not include the gradient steps that produced the LoRA. Thus only its continuation stage is gradient-free. The rows are contextual endpoints, not matched from-scratch or equal-compute comparisons.

Model / Method	#Iter.	Rule-Based		Model-Based					
		GenEval	OCR	PickScore	CLIPScore	HPSv2.1	Aesthetic	ImgRwd	UniRwd
Base Model									
SD3.5-M (w/o CFG)	–	0.24	0.12	20.51	0.237	0.204	5.13	−0.58	2.02
+ CFG	–	0.63	0.59	22.34	0.285	0.279	5.36	0.85	3.03
Gradient-Based Alignment									
	4k	0.97	0.30	21.78	0.277	0.248	5.15	0.74	2.87
+ Flow-GRPO	1k	0.66	0.96	21.94	0.280	0.257	5.18	0.31	2.86
	4k	0.54	0.60	23.62	0.257	0.295	6.42	1.17	3.17
	1k	0.98	0.36	21.92	0.271	0.251	5.33	0.68	2.91
+ DiffusionNFT	150	0.54	0.97	21.63	0.281	0.246	5.19	0.37	2.81
	2k	0.53	0.64	24.03	0.270	0.315	6.17	1.29	3.40
ES-Based Alignment									
	1.5k	0.59	0.30	20.54	0.282	0.253	5.15	0.37	2.93
+ Full-Model ES	2k	0.54	0.92	17.72	0.220	0.139	3.81	−1.66	1.59
	1k	0.54	0.31	21.96	0.275	0.296	5.80	0.77	3.12
	150	0.38	0.19	19.94	0.256	0.216	5.06	−0.46	2.38
+ PiSSA + ES	450	0.46	0.70	17.69	0.197	0.116	4.18	−1.84	1.38
	300	0.45	0.16	21.83	0.263	0.263	5.52	0.31	2.74
	500	0.35	0.17	20.06	0.252	0.213	5.08	−0.49	2.30
+ Random LoRA + ES	300	0.51	0.54	18.94	0.221	0.160	4.50	−1.12	1.82
	500	0.39	0.14	21.42	0.255	0.245	5.40	0.03	2.52
	2k	0.98	0.42	21.35	0.279	0.258	5.28	0.72	3.08
+ FASE	300	0.44	0.98	17.81	0.223	0.148	3.46	0.29	1.40
	2.5k	0.58	0.61	24.25	0.272	0.320	6.20	1.30	3.45

Checkpoint selection for FASE. Figure 2 shows the same pattern for DiffusionNFT and Flow-GRPO on all three tasks. Since D^A_{UV} is already lower at the selected handoffs, we use $D^B_{\mathrm{UV}} < 0.02$ as a conservative heuristic for small measured factor-subspace change, not as proof that individual singular vectors have converged. The DiffusionNFT handoffs at steps 200, 50, and 100 (GenEval, OCR, and PickScore) are all past this point; for GenEval, step 200 retains further gradient refinement after the drop near step 100. All three settings subsequently increase each optimized target reward, supporting continuation within these fixed handoff bases (Table 2); the table does not isolate freezing or establish that the frozen bases are optimal.

Implementation and uncertainty. Appendix B.2 gives hardware, evaluation, memory-state, and hyperparameter details. The tables report the comparison run, and we do not treat small score gaps as statistically significant.

4.2 Single-Reward Diffusion Alignment

Heterogeneous endpoint context. From its task-specific DiffusionNFT handoffs, FASE reaches GenEval 0.98, OCR 0.98, and PickScore 24.25 (Table 1). The other rows use different starts and budgets, so their target scores are reference endpoints rather than evidence of from-scratch superiority; small gaps are not treated as statistically significant. Appendix A.3 instead compares continuation GPU hours from method-specific checkpoints with approximately matched rewards.

Complementary scores. For OCR, FASE raises the optimized reward from 0.950 to 0.980, while every reported complementary metric declines from the warm-start; we therefore do not claim an overall quality improvement. Appendix A.2 reports complete OCR and PickScore complementary values. Fig. 4 shows one image per method on two prompts from each benchmark. Further prompts are in Fig. 5. On the two OCR prompts, the FASE image contains the prompt's main object and target

Table 2: Controlled same-warm-start continuation ablation. DiffusionNFT reaches steps 200, 50, and 100 for GenEval, OCR, and PickScore; Uniform (LoRA+ES), Static A/B, and FASE then share that task-specific LoRA and the ES budget. Uniform uses $r \equiv 1$ with the controller disabled; Static holds $r_0 = 5/3$ fixed; FASE starts from $r_0 = 5/3$ and updates r_t online. Iter. gives the gradient step or subsequent ES generations. Gray cells indicate the optimized reward.

Model / Method	Iter.	Rule-Based		Model-Based					
		GenEval	OCR	PickScore	CLIPScore	HPSv2.1	Aesthetic	ImgRwd	UniRwd
GenEval									
DiffusionNFT	200	0.926	0.464	20.49	0.286	0.265	5.20	0.69	3.18
+ LoRA+ES (Uniform)	2k	0.970	0.40	21.18	0.275	0.252	5.24	0.70	3.04
+ LoRA+ES (Static A/B)	2k	0.973	0.41	21.27	0.277	0.255	5.26	0.71	3.06
+ FASE (Adaptive)	2k	**0.980**	0.42	21.35	0.279	0.258	5.28	0.72	3.08
OCR									
DiffusionNFT	50	0.45	0.950	17.96	0.231	0.157	3.59	0.36	1.48
+ LoRA+ES (Uniform)	300	0.42	0.970	17.87	0.227	0.152	3.47	0.24	1.41
+ LoRA+ES (Static A/B)	300	0.43	0.976	17.84	0.225	0.150	3.47	0.28	1.41
+ FASE (Adaptive)	300	0.44	**0.980**	17.81	0.223	0.148	3.46	0.29	1.40
PickScore									
DiffusionNFT	100	0.64	0.52	22.83	0.277	0.326	6.01	1.20	3.50
+ LoRA+ES (Uniform)	2.5k	0.56	0.58	24.02	0.268	0.316	6.15	1.26	3.40
+ LoRA+ES (Static A/B)	2.5k	0.57	0.60	24.18	0.270	0.318	6.18	1.28	3.43
+ FASE (Adaptive)	2.5k	0.58	0.61	**24.25**	0.272	0.320	6.20	1.30	3.45

string, while the DiffusionNFT image contains the string without that object. The reported scores are evaluation-set means.

4.3 Adaptive Factor-Aware Exploration

Table 2 compares the same task-specific handoff and ES budget across Uniform, Static, and FASE within each task; the trajectories are shown in Appendix A.4, Fig. 7. Uniform uses equal A/B allocation, Static uses a fixed non-uniform allocation, and only FASE updates the factor-allocation ratio online; the comparison therefore isolates the adaptive controller within the shared frozen-basis search.

Warm-start. All three spectral-continuation settings raise each optimized target reward. FASE moves GenEval $0.926 \rightarrow 0.980$, OCR $0.950 \rightarrow 0.980$, and PickScore $22.83 \rightarrow 24.25$; its final gains over Uniform are small point estimates and are not treated as statistically significant.

Reward over generations. FASE's reported single-run trajectory reaches Uniform's terminal OCR and GenEval values by about generations 40 and 850, respectively, and stays above Uniform on PickScore; this is a descriptive trajectory, not evidence of a statistically reliable advantage (Appendix A.4).

5 Conclusion

Direct spectral ES from random or pretrained-only LoRA bases underperforms in the heterogeneous endpoint context, indicating that effective continuation benefits from a reward-adapted handoff. Across two gradient post-training methods and three rewards, measured nontrivial factor-subspace changes diminish while ordered spectra evolve, and residual spectral motion differs between A and B under the stored parameterization. These observations motivate, rather than prove, FASE's constrained design: freezing the handoff SVD bases and combining separable CMA-ES with adaptive, variance-preserving A/B exploration, with no backward pass or backward state during continuation. Same-warm-start ablations offer descriptive single-run support for adaptive allocation, while FASE matches or numerically exceeds separately trained DiffusionNFT and Flow-GRPO reference endpoints. FASE thus provides gradient-free continuation after a gradient-trained warm-start.

REFERENCES

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024.

Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, et al. Pp-ocr: A practical ultra lightweight ocr system. *arXiv preprint arXiv:2009.09941*, 2020.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Yu Gu, Zhi Zheng, Yunpeng Ba, Xialiang Tong, Mingxuan Yuan, and Zhenkun Wang. Hyper-es: Effective evolution strategies for llm reasoning via descent direction merging. *arXiv preprint arXiv:2608.05541*, 2026.

Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7289–7300. IEEE, 2023.

Nikolaus Hansen and Andreas Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary computation*, 9(2):159–195, 2001.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. Lora+: Efficient low rank adaptation of large models. *arXiv preprint arXiv:2402.12354*, 2024.

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.

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. *arXiv preprint arXiv:2106.09685*, 2021.

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. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Daria Korotyshova, Boris Shaposhnikov, Alexey Malakhov, Alexey Khokhulin, Nikita Surnachev, Kirill Ovcharenko, George Bredis, Alexey Gorbatovski, Viacheslav Sinii, and Daniil Gavrilov. Essa: Evolutionary strategies for scalable alignment. *arXiv preprint arXiv:2507.04453*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. In *Advances in Neural Information Processing Systems*, volume 38, pp. 40783–40818, 2025.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: Principal singular values and singular vectors adaptation of large language models. *Advances in Neural Information Processing Systems*, 37: 121038–121072, 2024.

Xin Qiu, Yulu Gan, Conor F Hayes, Qiyao Liang, Yinggan Xu, Roberto Dailey, Elliot Meyerson, Babak Hodjat, and Risto Miikkulainen. Evolution strategies at scale: Llm fine-tuning beyond reinforcement learning. *arXiv preprint arXiv:2509.24372*, 2025.

Raymond Ros and Nikolaus Hansen. A simple modification in CMA-ES achieving linear time and space complexity. In *Parallel Problem Solving from Nature – PPSN X*, pp. 296–305. Springer, 2008. doi: 10.1007/978-3-540-87700-4_30.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

Christoph Schuhmann. Laion-aesthetics. `https://laion.ai/blog/laion-aesthetics/`, 2022.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Yanpeng Sun, Qiang Chen, Xiangyu He, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Jian Cheng, Zechao Li, and Jingdong Wang. Singular value fine-tuning: Few-shot segmentation requires few-parameters fine-tuning. *Advances in neural information processing systems*, 35:37484–37496, 2022.

Danil Tokhchukov, Aysel Mirzoeva, Andrey Kuznetsov, and Konstantin Sobolev. Calibri: Enhancing diffusion transformers via parameter-efficient calibration. *arXiv preprint arXiv:2603.24800*, 2026.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025.

Daan Wierstra, Tom Schaul, Tobias Glasmachers, Yi Sun, Jan Peters, and Jürgen Schmidhuber. Natural evolution strategies. *The Journal of Machine Learning Research*, 15(1):949–980, 2014.

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.

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. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations*, volume 2026, pp. 134129–134150, 2026.

Jiacheng Zhu, Kristjan Greenewald, Kimia Nadjahi, Haitz Sáez De Ocáriz Borde, Rickard Brüel Gabrielsson, Leshem Choshen, Marzyeh Ghassemi, Mikhail Yurochkin, and Justin Solomon. Asymmetry in low-rank adapters of foundation models. *arXiv preprint arXiv:2402.16842*, 2024.

A Additional Results

A.1 Additional Qualitative Samples

Fig. 5 shows further prompts from OCR, PickScore, and GenEval. Each cell is one image for the prompt printed above it.



Figure 5: **Additional qualitative samples** on SD3.5-M. Each cell is one image for the prompt printed above it. Columns are grouped by benchmark, two columns each for OCR, PickScore, and GenEval, with six prompts in each group. Sampling follows Fig. 4.

A.2 Complementary Metric Details

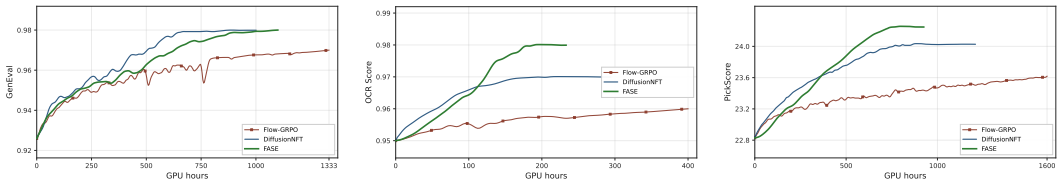
PickScore-trained FASE has higher CLIPScore, HPSv2.1, Aesthetic, ImageReward, and UnifiedReward than LoRA+ES (Uniform), although CLIPScore, HPSv2.1, and UnifiedReward remain below the warm-start (Table 2). For OCR, FASE raises the optimized reward from 0.950 to 0.980, but every reported complementary metric declines from the warm-start: GenEval $0.45 \rightarrow 0.44$, PickScore $17.96 \rightarrow 17.81$, CLIPScore $0.231 \rightarrow 0.223$, HPSv2.1 $0.157 \rightarrow 0.148$, Aesthetic $3.59 \rightarrow 3.46$, ImageReward $0.36 \rightarrow 0.29$, and UnifiedReward $1.48 \rightarrow 1.40$. These complementary endpoints also remain be-

low fully trained Flow-GRPO and DiffusionNFT OCR models (Aesthetic 5.18/5.19, UnifiedReward 2.86/2.81, PickScore 21.94/21.63).

A.3 Continuation GPU Hours

Figure 6 reports in-domain reward versus continuation GPU hours from method-specific checkpoints with approximately matched starting rewards.

Reference checkpoints and compute accounting. The starting rewards are approximately 0.926 for GenEval, 0.950 for OCR, and 22.83 for PickScore. DiffusionNFT continues from its LoRA checkpoints at optimizer updates 200, 50, and 100, respectively; FASE starts from these same DiffusionNFT checkpoints. Flow-GRPO continues from its own checkpoints chosen so that the in-domain score matches these DiffusionNFT warm-start rewards. Each curve resets elapsed GPU hours to zero at its reference checkpoint, excluding the training cost incurred before that checkpoint. The measured curves therefore report continuation cost from similar starting scores. They do not report end-to-end post-training cost or establish an end-to-end compute-efficiency claim.

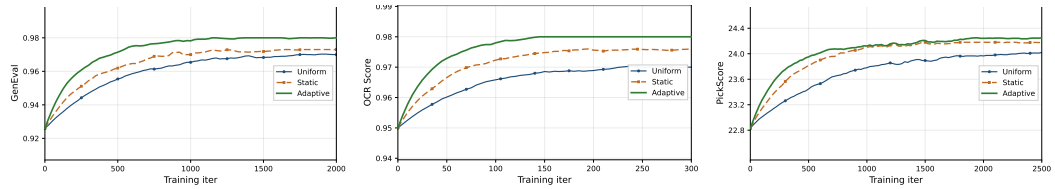


(a) GenEval (b) OCR (c) PickScore

Figure 6: **In-domain reward versus continuation GPU hours.** Curves start from method-specific checkpoints at approximately matched rewards: 0.926 (GenEval), 0.950 (OCR), and 22.83 (PickScore). DiffusionNFT and FASE share the DiffusionNFT LoRA checkpoints at updates 200, 50, and 100, respectively; Flow-GRPO uses its own checkpoints at those same starting rewards. GPU hours are measured from each reference checkpoint and exclude the preceding training cost; the curves are continuation-only, not end-to-end efficiency comparisons.

A.4 Continuation Trajectories

Figure 7 compares the matched spectral-continuation trajectories discussed in Sec. 4.3.

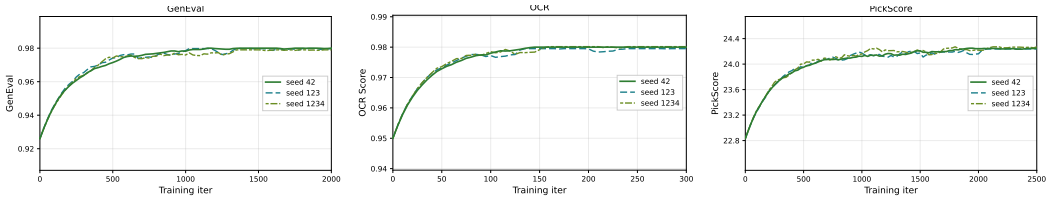


(a) GenEval (b) OCR (c) PickScore

Figure 7: **Reward during gradient-free continuation** after the saved, gradient-trained DiffusionNFT LoRA. Uniform (LoRA+ES), Static A/B, and FASE (labeled Adaptive in the plots) share the warm-start and budget and use native in-domain scores (PickScore in the low 20s). In these descriptive single-run trajectories, FASE reaches Uniform's terminal earlier on OCR and GenEval and stays above Uniform on PickScore (Table 2).

A.5 FASE Seeds

Fig. 8 shows FASE seeds 42, 123, and 1234 from the same saved LoRA. The tables report the comparison run and do not aggregate those curves.



(a) GenEval (b) OCR (c) PickScore

Figure 8: **FASE across three seeds.** Seeds 42, 123, and 1234, from the saved DiffusionNFT LoRA. Tables report the comparison run and do not aggregate these curves into a mean or an error bar.

B Method and Implementation Details

B.1 Energy-Preserving A/B Scales

Let $E_A = \sum_{i\in\mathcal{A}} d_i^2$, $E_B = \sum_{i\in\mathcal{B}} d_i^2$, and $r = \alpha_B/\alpha_A$. The stretched exploration energy is

$$E_A\alpha_A^2 + E_B\alpha_B^2 = \alpha_A^2\big(E_A + r^2E_B\big). \tag{12}$$

Imposing Eq. (8) together with $\alpha_B = r\alpha_A$ recovers Eq. (9). When $r = 1$, $\alpha_A = \alpha_B = 1$ for any positive energies, recovering LoRA+ES. For a candidate drawn from the current diagonal distribution, the sum of the coordinate variances equals its expected squared displacement from the mean. The construction preserves this sum under the A/B stretch. This is a per-generation property conditional on the current CMA scales: it does not equate the distributions of two independently evolving runs or their reconstructed weight-space perturbations.

Whitened steps u_F divide the mean displacement by the pre-update CMA scale and by α_F. This normalization removes the direct multiplicative contribution of the A/B stretch from the measured mean step. The EMA on $\log\rho$ and the exponent $\gamma < 1$ further smooth feedback from noisy RMS estimates.

The motivation statistic $v_F = \mathrm{RMS}(\Delta\Sigma_F)$ in Eq. (6) is neither E_F nor u_F, nor the raw squared magnitude $\|\Delta\Sigma_F\|_2^2$ (which overweights the higher-dimensional block). Here $|F|$ is the number of packed singular values in factor F, distinct from the CMA coordinate standard deviations d_i. The statistic diagnoses factor activity in the stored gradient-training parameterization. Its interpretation depends on the scaling convention used for those factors.

B.2 Implementation Notes

Full-Model ES and LoRA+ES. Algorithm 1 is the update used by Full-Model ES. LoRA+ES is Algorithm 2 with the controller disabled: after the factorization, set $r \equiv 1$ for every iteration and skip the ratio update, which forces $\alpha_A = \alpha_B = 1$. That is the ESSA parameterization under separable CMA-ES. FASE runs the algorithm as written, including the online update of r.

Reward evaluation. Each CMA candidate reconstructs LoRA from frozen (U, V) and the proposed $\mathbf{s}$, generates images with the same solver and timestep schedule used for the parent RL method at evaluation (40-step ODE for SD3.5-M), and scores a batch of 24 with R. No SDE conversion is required at ES time, unlike Flow-GRPO training.

Normalization. Population rewards are standardized per generation before the CMA update. Standardization preserves the ordering of rewards whenever their variance is positive, so it does not change candidate ranks for rank-based CMA selection.

Algorithm 1 Basic Evolution Strategy

Require: Initial parameters $\boldsymbol{\theta}_0$, reward function $R(\cdot)$, total iterations T, population size N, perturbation scale σ, and learning rate η.

for $t = 1$ to T **do**
  for $n = 1$ to N **do**
    Sample perturbation $\boldsymbol{\epsilon}_{t,n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    Evaluate the perturbed candidate:
    $R_{t,n} = R\left(\boldsymbol{\theta}_{t-1} + \sigma \boldsymbol{\epsilon}_{t,n}\right)$
  end for
  Normalize the population rewards to obtain $\{\widehat{R}_{t,n}\}_{n=1}^{N}$
  Estimate the ES update direction:

$$\widehat{\mathbf{g}}_t = \frac{1}{N\sigma} \sum_{n=1}^{N} \widehat{R}_{t,n} \boldsymbol{\epsilon}_{t,n}$$

  Update the model parameters:
  $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \eta \widehat{\mathbf{g}}_t$
end for
return $\boldsymbol{\theta}_T$

Algorithm 2 Adaptive A/B Spectral Exploration

Require: Warm-start LoRA (A, B), reward $R(\cdot)$, iterations T, population N, initial ratio $r_0 = \alpha_B/\alpha_A$ (raw Static prior $(\alpha_A, \alpha_B) = (0.75, 1.25)$, hence $r_0 = 5/3$), update interval K, and hyperparameters $\gamma, \beta, r_{\min}, r_{\max}$.

Factorize $A = U_A \Sigma_A V_A^\top$, $B = U_B \Sigma_B V_B^\top$; freeze all U, V.
Pack singular values $\mathbf{s} \leftarrow [\Sigma_A; \Sigma_B]$; initialize sep-CMA mean $\mathbf{m} \leftarrow \mathbf{s}$ and coordinate stds $\mathbf{d}$.
Set $r \leftarrow r_0$ and $\rho \leftarrow r_0^{1/\gamma}$.
for $t = 1$ to T **do**
  Compute group energies $E_A = \sum_{i \in \mathcal{A}} d_i^2$, $E_B = \sum_{i \in \mathcal{B}} d_i^2$.
  Map the exploration ratio to energy-preserving scales:

$$\alpha_A = \sqrt{\frac{E_A + E_B}{E_A + r^2 E_B}}, \qquad \alpha_B = r\,\alpha_A.$$

  for $n = 1$ to N **do**
    Sample a CMA candidate $\mathbf{s}^{(n)} \leftarrow \text{CMA-ASK}(\mathbf{m}, \mathbf{d})$.
    Stretch along A/B:
    $\hat{\mathbf{s}}^{(n)} = \mathbf{m} + \boldsymbol{\alpha} \odot \left(\mathbf{s}^{(n)} - \mathbf{m}\right)$, where $\alpha_i \in \{\alpha_A, \alpha_B\}$.
    Evaluate $R_n \leftarrow R\left(\hat{\mathbf{s}}^{(n)}\right)$.
  end for
  Store pre-update states $\mathbf{m}^{\text{pre}} \leftarrow \mathbf{m}$, $\mathbf{d}^{\text{pre}} \leftarrow \mathbf{d}$.
  Update CMA with stretched candidates $\{\hat{\mathbf{s}}^{(n)}, R_n\}_{n=1}^{N}$ (`check_points=False`).
  Compute the CMA mean update $\Delta\mathbf{m} \leftarrow \mathbf{m} - \mathbf{m}^{\text{pre}}$.
  Measure whitened mean steps

$$u_F = \text{RMS}\left(\frac{\Delta\mathbf{m}_F}{\mathbf{d}_F^{\text{pre}}\,\alpha_F}\right), \qquad F \in \{A, B\}.$$

  if $t \bmod K = 0$ **then**
    Average the last K values of u_F: $\bar{u}_F \leftarrow \text{Avg}(u_F)$.
    Update the log-ratio state:
    $\log \rho \leftarrow \beta \log \rho + (1 - \beta) \log(\bar{u}_B/\bar{u}_A)$.
    Set $r \leftarrow \text{clip}(\rho^\gamma, r_{\min}, r_{\max})$ (applies from iteration $t{+}1$).
  end if
end for
return LoRA reconstructed from optimized Σ_A, Σ_B and frozen U, V.

Memory. FASE continuation retains the frozen generator and reward-model parameters, the current latent, and ordinary inference workspace. After a candidate is evaluated, its coordinates or reproducible perturbation seed and scalar reward suffice; the population update also keeps the spectral mean, coordinate scales, step-size, and evolution paths maintained by CMA-ES. It requires no parameter-gradient buffers, Adam moments, or denoiser activations retained for backward, and candidates can be streamed or evaluated in waves without making population size a per-worker peak-memory multiplier. DiffusionNFT does not retain a full reverse-process trajectory, but its LoRA update still requires backward activations, parameter gradients, and optimizer state. Standard

Flow-GRPO additionally uses sampled transitions and behavior-policy log-probabilities, or their recomputation, for its policy-gradient update. These are structural backward-state distinctions, not measured peak-memory reductions or end-to-end compute-efficiency claims; the DiffusionNFT warm-start loaded by FASE was itself produced with gradient-training memory.

Recorded continuation settings. FASE loads a DiffusionNFT LoRA at optimizer updates 200 (GenEval), 50 (OCR), and 100 (PickScore). That adapter keeps the rank, scale, and adapted layers of the DiffusionNFT run. Those runs then use population $N{=}12$ with a batch of 24 images per candidate, 40-step ODE generation at 512×512, no CFG during reward optimization, and ES generation budgets 2000, 300, and 2500. The controller uses $\gamma{=}0.5$, $\beta{=}0.8$, $K{=}10$, $r \in [0.8, 2.0]$, and $r_0{=}5/3$ from raw $(\alpha_A, \alpha_B) = (0.75, 1.25)$. Hardware is 16 NVIDIA H20 GPUs. These are the FASE settings in Table 1 and the FASE (Adaptive) rows in Table 2; Uniform and Static use the corresponding LoRA+ES settings. Flow-GRPO, DiffusionNFT, Full-Model ES, and PiSSA+ES in Table 1 are trained from the base model.

AI USE STATEMENT

We used generative AI tools for code development and debugging; to assist with revisions to the title, language, and scientific-claim wording; and to provide feedback on the presentation and interpretation of experimental results.

The authors designed the experiments and methods and made all final scientific and editorial judgments. Every AI-assisted suggestion and output was reviewed and verified by the authors for technical accuracy and consistency with the experimental evidence. The authors take full responsibility for the final content of this work.