# Calibri: Enhancing Diffusion Transformers via Parameter-Efficient Calibration

Danil Tokhchukov
Lomonosov Moscow State University
firstauthor@i1.org

Aysel Mirzoeva
Lomonosov Moscow State University
secondauthor@i2.org

Andrey Kuznetsov
FusionBrain Lab
Innopolis University
secondauthor@i2.org

Konstantin Sobolev
FusionBrain Lab
Lomonosov Moscow State University
secondauthor@i2.org


Figure 1. Introducing *Calibri* – a parameter-efficient method for diffusion transformer alignment. By optimizing only $\sim 10^2$ parameters, *Calibri* significantly enhances the model's generation quality.

## Abstract

*In this paper, we uncover the hidden potential of Diffusion Transformers (DiTs) to significantly enhance generative tasks. Through an in-depth analysis of the denoising process, we demonstrate that introducing a single learned scaling parameter can significantly improve the performance of DiT blocks. Building on this insight, we propose* ***Calibri****, a parameter-efficient approach that optimally calibrates DiT components to elevate generative quality.* ***Calibri*** *frames DiT calibration as a black-box reward optimization problem, which is efficiently solved using an evolutionary algorithm and modifies just* $\sim 10^2$ *parameters. Additionally,* ***Calibri*** *introduces an innovative inference-time ensemble scaling strategy to further boost generative performance. Experimental results reveal that despite its lightweight design,* ***Calibri*** *consistently improves performance across various text-to-image models. Notably,* ***Calibri*** *also reduces the inference steps required for image generation, all while maintaining high-quality outputs.*

## 1. Introduction

In recent years, the field of visual content generation has experienced significant advancements, largely fueled by the development of diffusion models [14, 27]. Cutting-edge models like Stable Diffusion 3 [8] and FLUX [21] have redefined the landscape of modern generative frameworks. These models represent a shift from the traditional UNet architecture [28] to the more advanced Diffusion Transformer (DiT) [26], while also incorporating innovative techniques such as flow matching [23] to enhance their capabilities.

This powerful combination of a DiT backbone and flow matching has become the new de facto standard, extending far beyond text-to-image synthesis to power diverse tasks such as instruction-guided image editing [22, 37] and video generation [34].

Diffusion transformers are built from a sequence of identical blocks, each containing attention and MLP layers. Despite this uniform architecture, recent work suggests their functional contributions are highly uneven. For instance, Stable Flow [1] identified "vital layers" within the transformer, whose exclusion from generation process significantly alters the model's output. This finding implies that not all layers contribute equally to the final generation.

Building on this insight, we analyze the contribution of individual DiT blocks and uncover two surprising results. First, we find that selectively disabling certain blocks can actually improve generation quality, suggesting some may introduce detrimental artifacts. Second, we discover that a simple re-weighting of each block's output – by multiplying it with a single learned scalar – consistently enhances the model's performance over the original. These observations lead us to our central hypothesis: ***The standard DiT architecture is sub-optimally weighted, and its performance can be significantly improved through a simple post-hoc calibration of its blocks.***

Motivated by this hypothesis, we propose *Calibri*, a parameter-efficient approach designed to calibrate the contributions of DiT's architectural components and improve generation quality (Figure 1). Specifically, we frame the process of determining calibration coefficients as a black-box optimization problem with only $\sim 10^2$ parameters. The objective is to maximize the quality of model outputs, as measured by a reward model [25, 35]. To solve this optimization problem, we leverage the gradient-free evolutionary strategy CMA-ES [10, 11], which effectively identifies optimal scaling coefficients. Furthermore, we introduce a novel inference-time scaling technique, *Calibri Ensemble*, which integrates multiple calibrated models to further boost generative performance. Notably, *Calibri* also reduces the number of inference steps required for image generation, significantly improving both efficiency and quality. Extensive experiments across diverse baseline models validate the effectiveness of *Calibri* in achieving consistent performance gains without computational overhead.

In summary, our key contributions are as follows:

1. We analyze the potential of DiT architectures and reveal that their generation quality can be significantly improved through simple block scaling;
2. We propose a parameter-efficient framework, *Calibri*, which calibrates the contributions of DiT components to enhance denoising and generation quality within diffusion and flow-based models;
3. We present *Calibri Ensemble*, a novel inference-time ensemble strategy that integrates multiple calibrated *Calibri* models, leading to further improvements in generation quality;
4. Both *Calibri* and *Calibri Ensemble* are designed to seamlessly integrate with existing diffusion and flow-based models, offering a flexible and effective solution. We validate their performance by demonstrating substantial improvements in sample quality across diverse baseline methods.

## 2. Related Work

**Diffusion Models Backbones.** Early diffusion models predominantly utilized U-Net [28] backbones with residual blocks [12], pixelwise self-attention [32], and cross-attention layers for text-image conditioning [3, 15, 16, 27]. Recently, the field has shifted towards Diffusion Transformer (DiT)-based architecture [26], which received significant attention due to the scalability of transformer models [32]. One notable development is PixArt-alpha [4], which effectively applied DiT for text-conditional generation while preserving the conventional cross-attention mechanism for text-based conditioning. using a conventional cross-attention mechanism for text conditioning. A key milestone in this evolution is the introduction of the Multimodal Diffusion Transformer (MM-DiT) [8], which employs distinct transformers to process textual and visual inputs, subsequently combining their sequences through unified attention operations.

**Diffusion Model Backbone Interpretability.** Recent research has significantly advanced the understanding of diffusion model architecture, enabling novel applications. Early studies showed that cross-attention maps between text prompts and visual tokens produce high-quality saliency maps to predict spatial locations of textual concepts [31], applied in tasks like image editing [13] and layout control [5, 7]. Other works explored diffusion model components: Free-U [29] highlighted the U-Net backbone's denoising role and skip connections' contribution of high-frequency features, improving its denoising efficacy. Additionally, methods like Stable Flow [1] and FreeFlux [36] analyzed Diffusion Transformer (DiT) blocks, identifying critical layers for image formation and differentiating positional versus content-focused layers, leading to training-free image editing techniques that leverage interpretability in diffusion models.

**Visual Generative Model Alignment.** Aligning generative models, including diffusion models and rectified flows, with human feedback has significantly improved their performance. Conventional methods rely on reward models to capture human preferences [20, 25, 35, 39] and often use RLHF-inspired techniques like reward backpropagation [6, 39], Direct Preference Optimization (DPO) [33], Differentiable Diffusion Preference Optimization (DDPO) [2], and

Group Relative Policy Optimization (GRPO) [24], which typically require full model fine-tuning, making them computationally expensive.

*In contrast, our proposed Calibri method adopts a gradient-free evolutionary strategy to align diffusion and rectified flow models with human feedback. By optimizing only* $\sim 10^2$ *parameters, Calibri achieves efficient and effective alignment with human preferences.*

## 3. Method

### 3.1. Preliminaries

**Diffusion transformer architecture** The Diffusion Transformer (DiT) architecture comprises sequential DiT blocks that transform input tokens into output tokens. Two main types of DiT blocks have been introduced:

**Stantdard DiT block** [26] consists of Multi-Head Self-Attention (MHSA) layers and feed-forward layers, as illustrated in Figure 3a. Both layers apply LayerNorm to the incoming data and are modulated by a time embedding. The modulation is achieved using vectors $\alpha, \beta, \gamma$, which are generated by a distinct Multi-Layer Perceptron (MLP). Output of layers can be described by this formula:

$$\begin{aligned} x_l &= x_{l-1} + \gamma_1 MHSA(\alpha_1 LN(x_{l-1}) + \beta_1), \\ x_{l+1} &= x_l + \gamma_2 FF(\alpha_2 LN(x_l) + \beta_2), \end{aligned} \tag{1}$$

where $x_{l-1}$ denotes the input token sequence, and $x_l$ and $x_{l+1}$ represent the intermediate and final outputs.

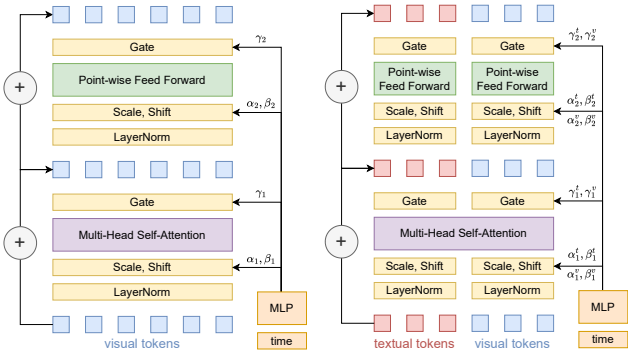
**MM-DiT block** [8] builds upon the structure of the Standard DiT block while introducing functionality for multimodal data processing. Specifically, MM-DiT combines textual and visual tokens via concatenation and processes them in parallel. Inter-modal communication is restricted to the MultiModal Attention Layer, enabling effective interaction between the two modalities. Separate modulation vectors are employed for each modality, denoted as $\alpha^v, \beta^v, \gamma^v$ for visual tokens and $\alpha^t, \beta^t, \gamma^t$ for textual tokens. Figure 3b provides a visual representation of the MM-DiT block structure, and the forward pass can be expressed as:

$$\begin{aligned} x_l^v &= x_{l-1}^v + \gamma_1^v MHSA(\alpha_1^v LN(x_{l-1}) + \beta_1^v), \\ x_l^t &= x_{l-1}^t + \gamma_1^t MHSA(\alpha_1^t LN(x_{l-1}) + \beta_1^t), \\ x_{l+1}^v &= x_l^v + \gamma_2^v FF(\alpha_2 LN(x_l^v) + \beta_2^v), \\ x_{l+1}^t &= x_l^t + \gamma_2^t FF(\alpha_2 LN(x_l^v) + \beta_2^t), \end{aligned} \tag{2}$$

where $x_l^v$ and $x_l^t$ correspond to the transformed tokens for the visual and textual modalities, respectively.

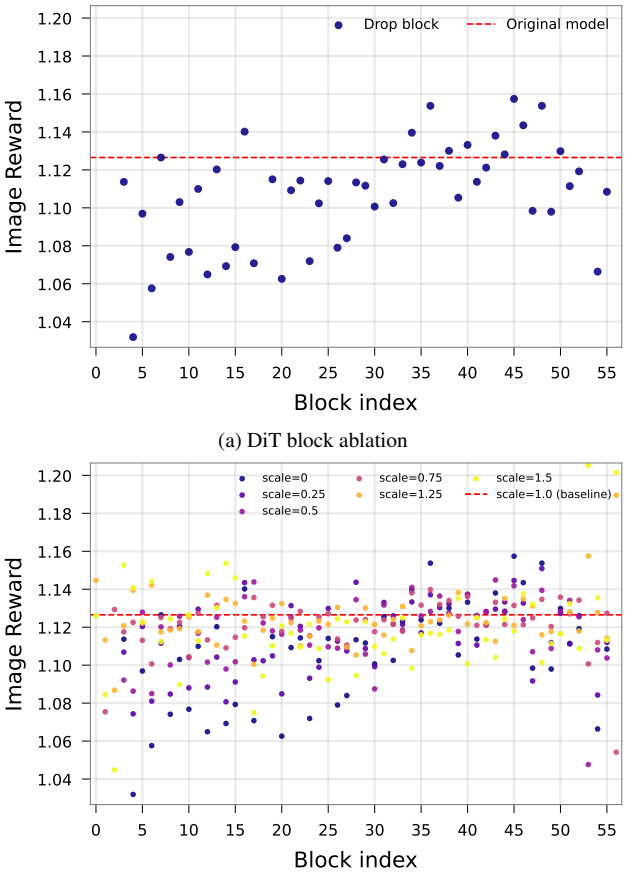
### 3.2. Motivation

Previous works [1, 36] have shown that, despite the similar architectural design across DiT blocks, their contributions to the overall model performance are uneven. Notably, Sta-




(a) DiT block scheme (b) MM-DiT block scheme

Figure 2. Illustration of DiT architectural components.




(a) DiT block ablation

(b) DiT block scaling

Figure 3. **Motivational Experiment:** Our findings reveal that the contribution of DiT blocks is not fully optimized. We demonstrate that their performance can be enhanced through a straightforward output scaling using a scalar multiplier.

ble Flow [1] identified the presence of "vital layers", whose exclusion during the inference process produces significant shifts in model outputs. Motivated by these findings, we aim to explore how the exclusion of individual DiT blocks impacts the model's overall quality in generative tasks.

To systematically evaluate the importance of individual layers within DiT, we devised a structured analysis framework. Using Qwen 3 [40] model, we first generated a set $P$ comprising $k = 64$ diverse text prompts. These prompts were used to produce a baseline set of images, $G^{base}$ utilizing FLUX [21] model. Next, for each DiT layer $l \in L$, we performed a controlled ablation by bypassing the layer output via its residual connection (i.e. in Formula 1 and 2, we multiply each $\gamma$ by 0). This process produced a collection of partially ablated image sets, $\{G^l\}$, from the same text prompts. To ensure statistical validity, we repeated each experimental configuration across 5 different random seeds.

To assess the impact of the layer, we compute Image Reward [39] scores for both the baseline image set $G^{base}$ and the ablated sets $\{G^l\}$. The results, presented in Figure 3, revealed an intriguing outcome: *removing certain layers can occasionally enhance the quality of generated images rather than degrade it.*

Building on this finding, we extended our analysis to examine the effect of scaling layer outputs. For each DiT block $l$, we generated additional sets of images, $\{G^l_s\}$, using a scaling coefficient $s$ for the block's output (i.e. in Formula 1 and 2, we multiply each $\gamma$ by $s$). We evaluated multiple scaling factors $s \in \{0, 0.25, 0.5, 0.75, 1.25, 1.5\}$, where $s = 0$ corresponds to block ablation, and $s = 1$ corresponds to the unaltered block behavior in the original model. The results, shown in Figure

Inspired by this result, we extend the experiment and generate a set of images $\{G^l_s\}$, where $s$ denotes a block output scaling coefficient, we use different $s \in \{0, 0.25, 0.5, 0.75, 1.25, 1.5\}$, $s = 0$ corresponts to block ablation, $s = 1$ corresponds to original model. The results, shown in Figure 3, presented another remarkable insight: *for each DiT block $l$, there exists an optimal scaling factor $s$ that improves the model's performance over its original configuration.*

### 3.3. Calibri

Based on the obtained insights, we present a simple yet effective method, named *Calibri*, aimed at enhancing the generative capabilities of the diffusion transformer by calibrating only a minimal subset of the model's parameters.

**Problem formulation.** The calibration process can be formulated as an optimization problem. Let $c \in \mathbb{R}^{L+1}$ represent the parameter vector of the diffusion transformer, where $L + 1$ corresponds to the total number of parameters selected for calibration. The goal is to find the optimal parameter configuration $c^*$ that maximizes the reward function:

$$c^* = \arg\max_c R(c), \tag{3}$$

where $R(*)$ is a scalar-valued function measuring the performance of the diffusion transformer on the given task.

**Search space.** The search space for calibrating the model is determined by the specific locations within the diffusion transformer where adjustments are applied. For a DiT-based diffusion or flow-based model $f_\theta(x, t, p)$, the calibration parameters are defined as $c = \omega \cup \{s_i\}_{s=1}^L$, where $\omega$ denotes output-level calibration weights, and $\{s_i\}_{s=1}^L$ represents internal-layer calibration parameters. The calibrated model output is thus expressed as $\omega f^s_\theta(x, t, p)$, where the applied calibration weights refine both external outputs and internal computations.

We introduce three levels of granularity for internal-layer calibration parameters, tailored to the structural hierarchy of diffusion transformers:

1. **Block Scaling:** As motivated in Section 3.2, block-wise scaling offers a coarse calibration technique by uniformly adjusting the outputs of Attention and MLP layers within the same architectural block using a shared scaling coefficient $\mathbf{s}$.
2. **Layer Scaling:** Extending the calibration to finer granularity, layer-wise scaling adjusts individual layers within a block using distinct coefficients. This method provides greater flexibility in refining model behavior beyond uniform block-level adjustments.
3. **Gate Scaling:** Gate-wise calibration becomes particularly important for architectures with multimodal interactions, such as MM-DiT. Here, visual and textual tokens are processed through distinct gates, each requiring specialized calibration to optimize their interaction dynamics.

**Calibration parameters search procedure.** To identify optimal calibration coefficients, we employ the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [10, 11], a powerful gradient-free optimization approach. CMA-ES optimizes an objective function $R : \mathbb{R}^d \to \mathbb{R}$ by iteratively refining a sampling distribution based on a multivariate Gaussian, $\mathcal{N}(\mu, \sigma^2 C)$, where $\mu \in \mathbb{R}^d$ represents the mean vector, $\sigma \in \mathbb{R}_{>0}$ is the step-size, and $C \in \mathbb{R}^{d\times d}$ is the covariance matrix.

The method scheme is depicted in the Figure 4. At each iteration, candidate solutions are drawn from this Gaussian distribution and evaluated using the objective function. CMA-ES updates the mean vector by moving toward higher-performing candidates while adapting the covariance matrix to reflect successful directions in the search space. This iterative refinement allows efficient exploration

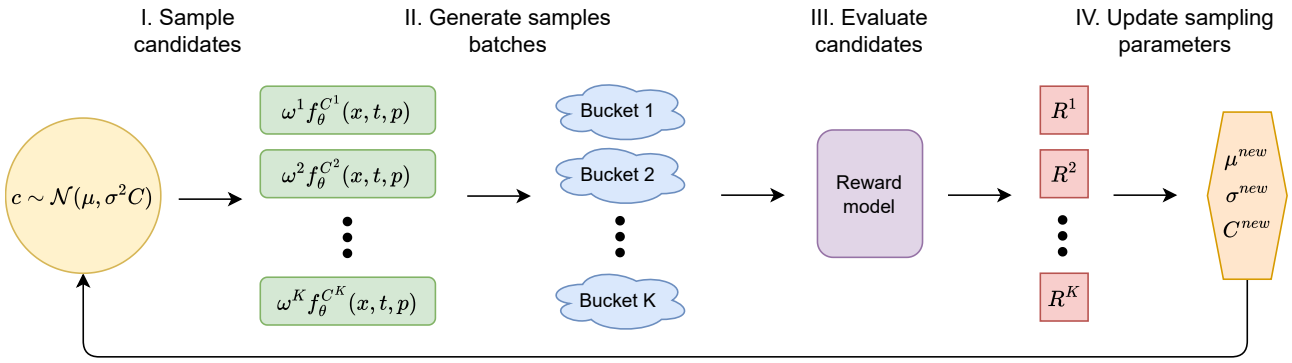



Figure 4. Illustration of calibration parameter search procedure.

and exploitation, optimizing calibration coefficients for improved model performance over successive iterations.

### 3.4. Calibri Ensemble

*Calibri* also introduces an intriguing perspective when applied in an ensemble setting. Unlike traditional inference approaches, where combining similar models might offer negligible benefits, our method enables the calibration of an ensemble of $N$ models simultaneously. Specifically, the ensemble is represented as:

$$F^{\{c_i\}_{i=1}^N}(x,t,p) = \sum_{i=1}^{N} \omega_i f_\theta^{s_i}(x,t,p|\emptyset),$$

where $\omega_i$ denotes the weight assigned to the $i$-th model, $f_\theta^{s_i}$ represents the calibrated with internal-layer calibration parameters $s_i$, and $\{x,t,p\}$ are the input signals, time step, and additional conditioning inputs, respectively.

In this case, the optimization problem 3 is reformulated as:

$$\{c_i^*\}_{i=1}^N = \arg\max_{\{c_i\}_{i=1}^N} R(\{c_i\}_{i=1}^N), \tag{4}$$

where $R(\{c_i\}_{i=1}^N)$ evaluates the overall performance of the ensemble. This ensemble calibration allows us to leverage the diversity among $N$ optimized models, resulting in enhanced generative performance and robustness.

**Relation to Classifier-Free Guidance.** *Calibri ensemble* framework seamlessly integrates into the Classifier-Free Guidance paradigm. In this specific case, the ensemble size is $N = 2$, and optimization is performed following Problem 4. By calibrating two models representing distinct guiding roles (e.g. conditional and unconditional), our method enhances generation while maintaining diversity and precision.

## 4. Experiments

**Baselines:** To evaluate the effectiveness of *Calibri*, we compare its performance across several state-of-the-art, open-source DiT-based text-to-image models. Specifically, we conduct experiments on FLUX.1-dev [21], Stable Diffusion 3.5 Medium (SD-3.5M) [8], and Qwen-Image [37], all of which represent the cutting edge in text-to-image generation. Additionally, we test *Calibri* using the SD-3.5M model checkpoint fine-tuned with Flow-GRPO [24] to analyze its performance in alignment-sensitive setups.

**Implementation Details:** For our experiments, we used train and test prompts from T2I-compbench++ [17]: train prompts were used to sample buckets for candidate evaluation in the CMA-ES algorithm and test prompts were used for intermediate reward evaluations to select the best coefficients. We used HPSv3 [25] to track image preference and Q-Align [38] to track image quality during training. For hyperparameters of CMA-ES we used commonly used parameters: the initial sigma was set to 0.25, the number of candidates was set to $4 + \lfloor 3 \ln d \rfloor \approx 20 \pm 3$ for considered models, where $d$ is the dimension of the search space and represents the number of coefficients to be tuned. We have fixed the bucket size to 16, the image resolution to 512 and the number of inference steps to 15 for training – we found it is the lowest number of steps that achieve satisfactory generation quality across several models.

**Evaluation and Metrics:** We used HPDv3 test prompts for evaluation. To measure the final metrics, we used HPSv3 [25], Q-Align [38] and ImageReward [39].

### 4.1. Calibri Design Decisions

**Search space.** To evaluate the effectiveness of scaling granularity, we consider three approaches introduced in Section 3.3: block scaling, layer scaling, and gate scaling. These scaling methods progressively increase the number of parameters available for optimization, as detailed in Table 1. All experiments are conducted using *Calibri* applied

Table 1. Comparison of various granularity levels for internal-layer calibration applied to the Flux model. All experiments are conducted using the Flux model, with evaluation performed on the HPSv3 test set.

| Scaling | N params | iters | HPSv3 | IR | Q-Align |
|---|---|---|---|---|---|
| – | – | – | 11.41 | 1.15 | 4.85 |
| Block | 57 | 200 | 13.29 | 1.17 | **4.91** |
| Layer | 76 | 410 | 13.41 | **1.24** | 4.90 |
| Gate | 114 | 960 | **13.48** | 1.18 | 4.88 |


Figure 5. Quantitative comparison of various granularity levels for internal-layer calibration applied to the Flux model.

to the Flux model, with optimization guided by the HPSv3 reward.

While gate scaling achieves the highest value of the target reward (HPSv3), it underperforms on several alternative rewards. In contrast, layer scaling yields more consistent improvements across multiple reward functions, and Figure 5 illustrates its advantage over the other scaling methods. Overall, the resulting performance across the three schemes is relatively similar, but their training speeds differ substantially, which is an important factor when choosing the appropriate scaling strategy.

**N models.** The *Calibri Ensemble* method (Section 3.4) allows us to aggregate multiple differently calibrated models into a single sampler. To validate this approach, we evaluate *Calibri Ensemble* on FLUX guided by HPSv3 reward with $N \in \{1, 2\}$ models using HPDv3 prompts. In the experiments, we use block scaling, as it empirically yields the fastest convergence. Since FLUX is a guidance-distilled

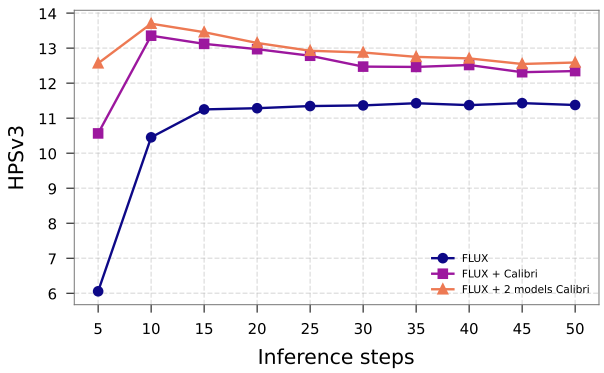



Figure 6. Comparison between *Calibri Ensemble* and original model across several inference steps.

model, we pass the same conditioning input to each model instance and then combine their contributions within the *Calibri* framework.

We also note that for $N = 2$ with block scaling, the *Calibri Ensembling* method generalizes Skip Layer Guidance (also referred to as Spatiotemporal Guidance [18]), which can be seen as a special training-free case of Autoguidance [19]. The results show that ensembling calibrated models consistently increases the HPSv3 reward across all inference steps, as illustrated in Figure 6.

**NFE.** Another notable observation in Figure 6 is that *Calibri Ensembling* shifts the optimal number of sampling steps from 30–50 in the baseline to only 10–15 steps. This substantially reduces the number of function evaluations required to achieve strong performance, making inference both faster and more computationally efficient.

### 4.2. Different Backbones

We evaluate *Calibri* on three representative T2I models: Flux, SD-3.5M, and Qwen-Image. Quantitative results, presented in Table 2, demonstrate consistent performance improvements across all baseline models when using *Calibri*. Notably, *Calibri* achieves these enhanced metrics while requiring significantly fewer inference steps – 15 steps compared to 30 for Flux, 40 for SD-3.5M, and 50 for Qwen-Image. Furthermore, qualitative comparisons in Figure 7 illustrate the superior efficiency and output quality of the proposed approach, reinforcing its effectiveness and practical advantages.

### 4.3. Combining Calibri with Alignment Methods

We assess the effectiveness of *Calibri* in combination with Alignment methods using three distinct SD-3.5M checkpoints: the pretrained base model, the Flow-GRPO [24] checkpoint aligned with the PickScore [20] reward, and the Flow-GRPO [24] checkpoint aligned with the GenEval [9]


Figure 7. Qualitative evaluation of generation quality improvements across various baseline models. Models have same NFE as in Table 2.

metric. Table 3 and Figure 8 present results for three experimental setups.

First, we examine the impact of applying *Calibri* to the base model for optimizing PickScore. Notably, this procedure not only improves the target metric (PickScore) but also enhances related metrics such as HPSv3 and Q-Align, demonstrating a broader positive impact on model performance. Additionally, the model optimized for PickScore using *Calibri* achieves comparable results to the optimization achieved by Flow-GRPO, despite *Calibri* only updating 216 parameters compared to the 18.78M parameters updated by Flow-GRPO.


Figure 8. Qualitative comparison of *Calibri* and Flow-GRPO [24] on SD-3.5M. *Calibri* achieves comparable performance with $10^5$ fewer parameters and can be combined with alignment methods to boost either the same or different target metrics.

Table 2. Quantitative evaluation of generation quality improvements across various baseline models. Notably, *Calibri* achieves superior metric scores while requiring fewer inference steps.

| Model | *Calibri* | HPSv3 | IR | Q-Align | NFE |
|---|---|---|---|---|---|
| FLUX | ✗ | 11.41 | 1.15 | 4.85 | 30 |
| | ✓ | **13.48** | **1.18** | **4.88** | 15 |
| SD-3.5M | ✗ | 11.15 | 1.10 | 4.74 | 80 |
| | ✓ | **14.10** | **1.17** | **4.91** | 30 |
| Qwen Image | ✗ | 11.26 | 1.16 | 4.55 | 100 |
| | ✓ | **12.95** | **1.18** | **4.73** | 30 |

Table 3. Comparison of *Calibri* and Flow-GRPO [24] on SD-3.5M. *Calibri* achieves comparable performance with $10^5$ fewer parameters and can be combined with alignment methods to boost either the same or different target metrics.

| Flow-GRPO | Calibri | HPSv3 | PickScore | Q-Align | NFE |
|---|---|---|---|---|---|
| ✗ | ✗ | 11.15 | 22.40 | 4.74 | 80 |
| | PickScore | **12.47** | **23.13** | **4.91** | 30 |
| PickScore | ✗ | 12.67 | 23.78 | **4.92** | 80 |
| | PickScore | **12.96** | **23.93** | 4.85 | 30 |
| GenEval | ✗ | 10.16 | 22.22 | 4.69 | 80 |
| | HPSv3 | **14.18** | 22.22 | **4.88** | 30 |

Next, we investigate the application of *Calibri* on a Flow-GRPO checkpoint that was already optimized for PickScore. Our results show that further performance improvements can be achieved by using *Calibri*, underscoring its utility in enhancing models already aligned to a specific reward.

Finally, we apply *Calibri* to a Flow-GRPO checkpoint trained to maximize the GenEval metric. As demonstrated in Table 3 and Figure 8, *Calibri* integrates efficiently with standard alignment methods and significantly boosts metric performance across the board. These findings highlight *Calibri*'s versatility and effectiveness in improving model alignment with various optimization targets.

## 5. Conclusion

In this work, we introduced *Calibri*, a novel and parameter-efficient approach to enhance the generative capabilities of Diffusion Transformers (DiTs). By uncovering the potential of a single learned scaling parameter to optimize the contributions of DiT components, we demonstrated that significant performance improvements can be achieved with minimal parameter modifications. Framing the calibration process as a black-box optimization problem solved via the CMA-ES evolutionary strategy, *Calibri* adjusts only $\sim 10^2$ parameters while delivering consistently improved generation quality. Additionally, the proposed inference-time scaling technique, *Calibri Ensemble*, effectively combines calibrated models to further enhance results.

Our extensive empirical evaluation across a range of text-to-image diffusion models confirmed the effectiveness and efficiency of *Calibri*, highlighting its ability to achieve superior generative quality with reduced computational costs. Notably, *Calibri* successfully reduces the number of inference steps required for image generation while retaining high-quality outputs, making it a practical solution for real-world applications where computational efficiency is critical.

## References


[1] Omri Avrahami, Or Patashnik, Ohad Fried, Egor Nemchinov, Kfir Aberman, Dani Lischinski, and Daniel Cohen-Or. Stable flow: Vital layers for training-free image editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7877–7888, 2025. 2, 3, 4

[2] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22563–22575, 2023. 2

[4] Junsong Chen, YU Jincheng, GE Chongjian, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-alpha : Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations*. 2

[5] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5343–5353, 2024. 2

[6] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[7] Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36:16222–16239, 2023. 2

[8] 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. 1, 2, 3, 5

[9] 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. 6

[10] Nikolaus Hansen, Sibylle D Müller, and Petros Koumoutsakos. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (cma-es). *Evolutionary computation*, 11(1):1–18, 2003. 2, 4

[11] Nikolaus Hansen, Youhei Akimoto, and Petr Baudis. CMA-ES/pycma on Github. Zenodo, DOI:10.5281/zenodo.2559634, 2019. 2, 4

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[13] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[16] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022. 2

[17] Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47:3563–3579, 2024. 5, 2

[18] Junha Hyung, Kinam Kim, Susung Hong, Min-Jung Kim, and Jaegul Choo. Spatiotemporal skip guidance for enhanced video diffusion sampling. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 6

[19] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 2024. 6

[20] 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. 2, 6

[21] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024. 1, 4, 5, 2

[22] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 2

[23] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*. 1

[24] 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. *arXiv preprint arXiv:2505.05470*, 2025. 3, 5, 6, 8, 1, 2

[25] 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*, pages 15086–15095, 2025. 2, 5, 1

[26] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 1, 2, 3

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1, 2

[29] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4733–4743, 2024. 2

[30] Dmitrii Sorokin, Maksim Nakhodnov, Andrey Kuznetsov, and Aibek Alanov. Imagerefl: Balancing quality and diversity in human-aligned diffusion models. *arXiv preprint arXiv:2505.22569*, 2025. 1

[31] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Türe. What the daam: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5644–5659, 2023. 2

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[33] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024. 2

[34] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2

[35] 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. 2

[36] Tianyi Wei, Yifan Zhou, Dongdong Chen, and Xingang Pan. Freeflux: Understanding and exploiting layer-specific roles in rope-based mmdit for versatile image editing. *arXiv preprint arXiv:2503.16153*, 2025. 2, 3

[37] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 2, 5

[38] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 5

[39] 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. 2, 4, 5

[40] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 4

# Calibri: Enhancing Diffusion Transformers via Parameter-Efficient Calibration

Supplementary Material

## Supplementary Material Structure

This supplementary document is organized as follows:

1. Section A elaborates on the limitations of the proposed methodology, providing a detailed analysis.
2. Section B analyzes the diversity of images generated by the method, both before and after incorporating the Calibri technique.
3. Section C explains the rationale behind the chosen reward model, highlighting its impact on the system's performance.
4. Section D discusses the motivation for using the CMA-ES approach as the parameter search method, justifying its effectiveness.

## A. Limitations

Our calibration coefficients selection method, *Calibri*, leverages a reward model [25] as its objective function. Reward models are trained to approximate user preferences for generated images, which enables *Calibri* to optimize the selection process effectively.

However, despite substantial advancements in reward modeling in recent years, current reward models still exhibit notable limitations. Specifically, they often demonstrate insufficient sensitivity to generation artifacts such as anatomical inconsistencies—examples include extra limbs, distorted fingers, and other visually unrealistic features, as illustrated in Figure 9.

These shortcomings in reward models can impact the performance of *Calibri*, resulting in the selection of a suboptimal set of coefficients. Addressing these limitations is crucial for further improving the robustness and overall effectiveness of the calibration process.

We anticipate that ongoing advancements in reward modeling techniques will mitigate these issues and significantly enhance their sensitivity to such artifacts, ultimately improving the performance of *Calibri* in future iterations.

## B. Generation diversity

Optimizing diffusion models using reward models often leads to a reduction in generation diversity, as highlighted by recent findings [30]. Since *Calibri* employs a reward model as its optimization objective, it is important to evaluate how this approach affects generation diversity. In Table 4, we present a comparison of generation diversity between the original model (SD-3.5M) [8] and models optimized by *Calibri* and Flow-GRPO [24].

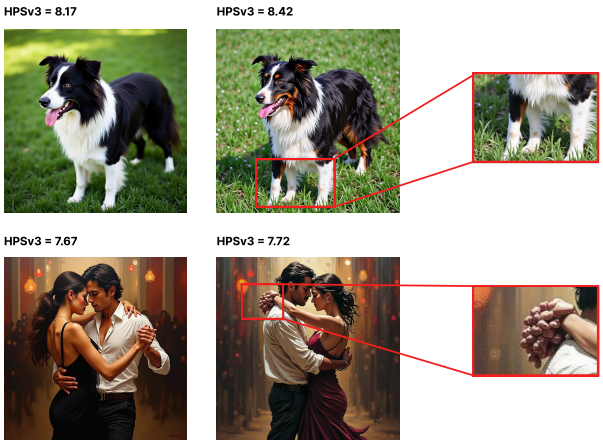



Figure 9. Limitations of modern reward models.

The results demonstrate that the generation diversity of SD-3.5M with 40 inference timesteps remains comparable to that of the model optimized by *Calibri*, which achieves comparable diversity while requiring only 15 inference timesteps. Importantly, despite its reduced inference time, the model optimized by *Calibri* exhibits significantly higher generation quality compared to the original model.

In contrast, Flow-GRPO reduces generation diversity from 0.20 to 0.15 and fails to accelerate inference time. Furthermore, when *Calibri* is applied to the model already optimized by Flow-GRPO, it does not introduce any further changes in generation diversity. This result underscores the efficiency and robustness of *Calibri* in preserving diversity while enhancing generation quality and reducing inference time.

Table 4. Comparison of Generation Diversity for SD-3.5M [8] Optimized by *Calibri* and Flow-GRPO [24].

| Flow-GRPO | Calibri | HPSv3 | PickScore | Q-Align | Dino Diversity | NFE |
|---|---|---|---|---|---|---|
| ✗ | ✗ | 11.15 | 22.40 | 4.74 | $0.20 \pm 0.06$ | 80 |
| | ✗ | 9.5 | 22.04 | 4.51 | $0.25 \pm 0.08$ | 30 |
| | PickScore | **12.47** | **23.13** | **4.91** | $0.20 \pm 0.06$ | 30 |
| PickScore | ✗ | 12.67 | 23.78 | **4.92** | $0.15 \pm 0.06$ | 80 |
| | ✗ | 12.514 | 23.76 | 4.91 | $0.15 \pm 0.05$ | 30 |
| | PickScore | **12.96** | **23.93** | 4.85 | $0.15 \pm 0.05$ | 30 |

## C. Different Rewards

To evaluate the performance of *Calibri* across different objectives, we conducted experiments using various reward


Figure 10. Illustration of *Calibri* with different rewards as objective.

Table 5. Quantitative comparison of *Calibri* layer scale on Flux [21] across different reward models.

| Calibri | HPSv3 | IR | Q-Align | PickScore | NFE |
|---|---|---|---|---|---|
| ✗ | 11.41 | 1.15 | 4.85 | 22.88 | 30 |
| HPSv3 | **13.41** | **1.24** | **4.90** | <u>23.07</u> | 15 |
| ImageReward | 11.06 | 1.17 | 4.70 | 22.47 | 15 |
| Q-Align | 11.65 | 1.0 | 4.89 | 22.36 | 15 |
| PickScore | <u>13.34</u> | <u>1.2</u> | <u>4.89</u> | **23.24** | 15 |

models as optimization objectives. The results are summarized in Table 5 and visually presented in Figure 10.

Calibration using the HPSv3 [25] reward model achieved the most significant quality improvement across all metrics, while the Pickscore [20] reward exhibited similarly strong performance. Notably, we observed that calibrating with the most effective reward model not only enhances the target metric but also yields substantial improvements across other metrics. This indicates that *Calibri* is not designed as a reward hacking method tailored to specific objectives, but rather as a general-purpose technique for improving overall generation quality.

## D. CMA-ES vs gradient-based parameter search

Gradient-based algorithms are commonly employed for alignment of diffusion models via reward maximization. However, their application to diffusion models presents

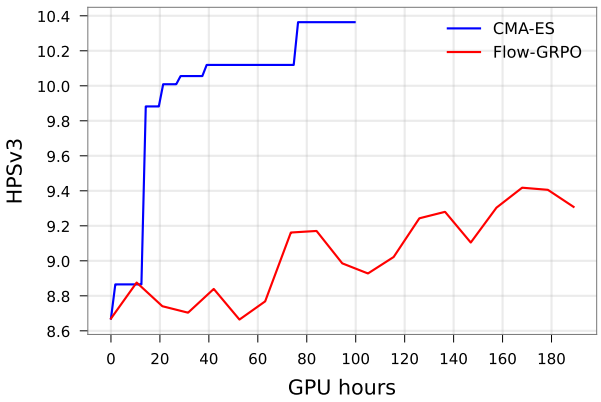
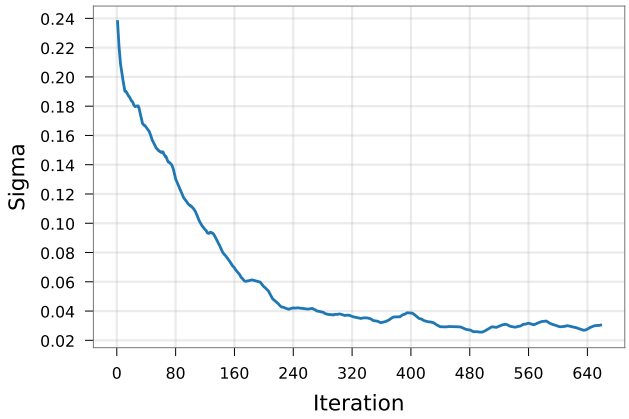



Figure 11. Comparison of CMA-ES and Flow-GRPO performance in optimizing *Calibri* coefficients.

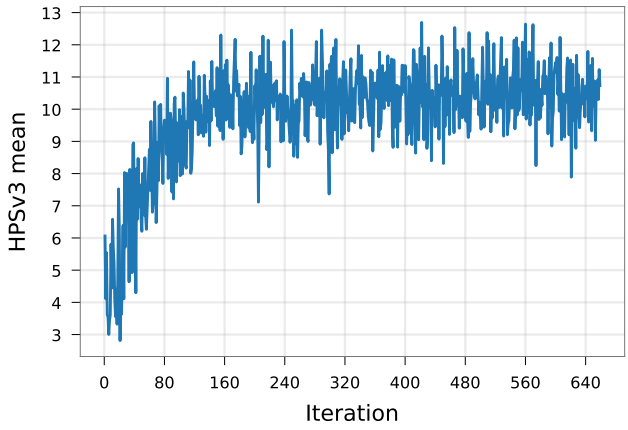
challenges due to the incompatibility of reward models with noisy latent spaces and generated images, necessitating repeated inference steps during the generation process for accurate reward computation. A recent advancement, Flow-GRPO [24], addresses this issue by reframing the diffusion process as a Markov Decision Process (MDP), enabling reward-driven training in such models. In this section, we evaluate the performance of CMA-ES and compare it to Flow-GRPO, while also providing additional analysis of CMA-ES training dynamics.

To compare CMA-ES with Flow-GRPO, we trained *Calibri* with layer scale on FLUX using these two optimizers, following the main experimental setup with evaluation during training on the T2I-Compbench++ [17] test prompts. For Flow-GRPO, we adopted the default hyperparameter configuration provided for Flux [21]. The evaluation curves obtained during training are shown in Figure 11 and demonstrate that CMA-ES is substantially more efficient than Flow-GRPO.

Additionally, we present detailed insights into CMA-ES training dynamics in Figure 12. Our analysis indicates that the *Calibri* coefficients converge effectively during training, with optimization reaching a plateau as evidenced by the stabilization of sigma and the stagnation of improvements in the training curve. These observations suggest that training with CMA-ES can be terminated once this convergence behavior is observed, optimizing computational resources without compromising performance.




(a) Sigma decrease during training and indicates when the training can be stopped.




(b) Training curve with the mean reward on buckets.

Figure 12. CMA-ES algorithm optimizes *Calibri* coefficients for layer scale FLUX.