Scaling Diffusion Transformers Efficiently via μP

Chenyu Zheng[1,2] Xinyu Zhang[2] Rongzhen Wang[1] Wei Huang[3] Zhi Tian[2] Weilin Huang[2] Jun Zhu[4] Chongxuan Li[1]

[1]Gaoling School of AI, Renmin University of China [2]ByteDance Seed [3]RIKEN AIP [4]Tsinghua University

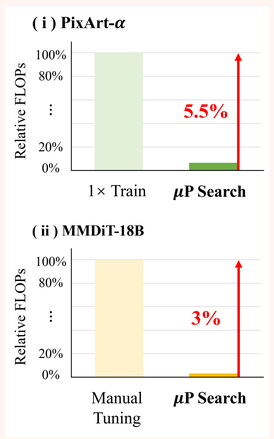
Abstract

Diffusion Transformers have emerged as the foundation for vision generative models, but their scalability is limited by the high cost of hyperparameter (HP) tuning at large scales. Recently, Maximal Update Parametrization (μP) was proposed for vanilla Transformers, which enables stable HP transfer from small to large language models, and dramatically reduces tuning costs. However, it remains unclear whether μP of vanilla Transformers extends to diffusion Transformers, which differ architecturally and objectively. In this work, we generalize standard μP to diffusion Transformers and validate its effectiveness through large-scale experiments. First, we rigorously prove that μP of mainstream diffusion Transformers, including U-ViT, DiT, PixArt-α, and MMDiT, aligns with that of the vanilla Transformer, enabling the direct application of existing μP methodologies. Leveraging this result, we systematically demonstrate that DiT-μP enjoys robust HP transferability. Notably, DiT-XL-2-μP with transferred learning rate achieves 2.9× faster convergence than the original DiT-XL-2. Finally, we validate the effectiveness of μP on text-to-image generation by scaling PixArt-α from 0.04B to 0.61B and MMDiT from 0.18B to 18B. In both cases, models under μP outperform their respective baselines while requiring small tuning cost—only 5.5% of one training run for PixArt-α and 3% of consumption by human experts for MMDiT-18B. *These results establish μP as a principled and efficient framework for scaling diffusion Transformers.*

Highlights

- We **rigorously prove** that μP of mainstream diffusion Transformers aligns with the standard Transformer, **enabling the direct application of existing μP methodologies**.
- We systematically demonstrate that **DiT under μP enjoys robust HP transferability**. Notably, DiT-XL-2-μP with transferred learning rate **achieves 2.9 times faster convergence** than the original DiT-XL-2.
- We validate that diffusion Transformers under μP outperform their respective baselines while requiring small tuning cost (e.g., 3% **FLOPs of human experts for MMDiT-18B**).

(a) Samples produced by the MMDiT-μP-18B.



(b) Efficiency of μP search.

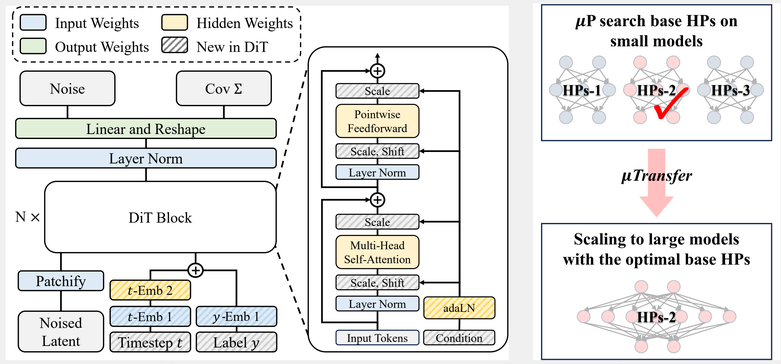
Figure 1. **Visualization results and efficiency of HP search under μP.** (a) Samples generated by the MMDiT-μP-18B model exhibit strong fidelity. (b) HP search for large diffusion Transformers is efficient under μP, **requiring only 5.5% FLOPs of a single training run for PixArt-α and just 3% FLOPs of the human experts for MMDiT-18B**.

Scaling Diffusion Transformers via μP

Table 1. **μP for Standard Transformers with Adam/AdamW optimizer.** We use purple text to highlight the differences between μP and standard parameterization (SP) in practice (e.g., Kaiming initialization), and gray text to indicate the SP settings. N is the width of the weight.

	Input weights	Hidden weights	Output weights
Multiplier	1	1	$1/N$ (1)
Init. var	σ^2	σ^2/N	σ^2 (σ^2/N)
Learning rate	η	η/N η	η

Theorem 1 (μP of diffusion Transformers) The forward passes of mainstream diffusion Transformers (U-ViT, DiT, Pixart-α, and MMDiT) can be represented within the Ne⊗orT Program. **Therefore, their μP matches the Table 1**.

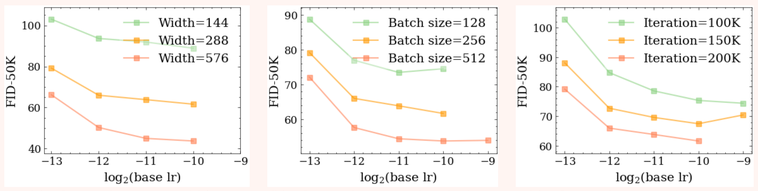


(a) Implementation of DiT-μP. (b) μTransfer.

Figure 2. **A overview of applying μP to diffusion Transformers.** (a) We illustrate the implementation of μP for DiT as an example. The *abc*-parameterization of each weight is adjusted based on its type and visualized using different colors. Modules that differ from the vanilla Transformer are also highlighted. (b) We μTransfer the optimal base HPs searched from multiple trials on small models to pretrain the target large models.

Stable Hyperparameter Transferability of μP

We empirically verify the base HP transferability of DiT under μP, across different widths, training steps, and batch sizes.

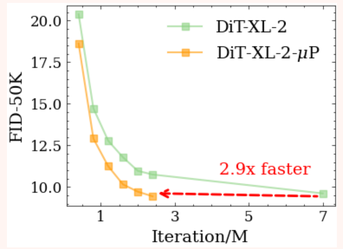


(a) HP transfer across widths. (b) HP transfer across batch sizes. (c) HP transfer across iterations.

Figure 3. **DiT-μP enjoys base HP transferability**. Unless otherwise specified, we use a model width of 288, a batch size of 256, and a training iteration of 200K. Missing data points indicate training instability, where the loss explodes. Under μP, the base learning rate can be transferred across model widths, batch sizes, and steps.

μTransfer results of DiT

Considering FID-50K, **DiT-XL-2-μP with transferred learning rate achieves 2.9× faster convergence** than the original DiT-XL-2 and a slightly better result.

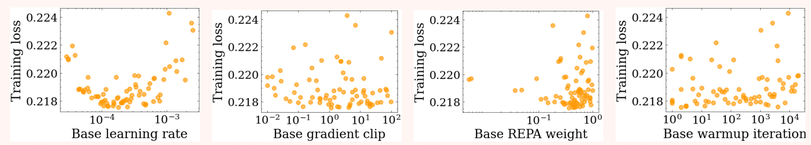


μTransfer results of PixArt-α

Table 2. **Comprehensive comparison between PixArt-α-μP and PixArt-α.** Both models are trained on the SAM dataset for 30 epochs. PixArt-α-μP (0.61B), using a base learning rate transferred from the optimal 0.04B proxy model, consistently outperforms the original baseline throughout the training process.

Epoch	Method	GenEval ↑	MJHQ FID-30K ↓	MJHQ CLIP Score ↑	MS-COCO FID-30K ↓	MS-COCO CLIP Score ↑
10	PixArt-α	0.19	38.36	25.78	34.58	28.12
	PixArt-α-μP	0.20	33.35	26.25	29.68	28.87
20	PixArt-α	0.20	35.68	26.54	30.13	28.81
	PixArt-α-μP	0.23	33.42	26.83	29.05	29.53
30	PixArt-α	0.15	42.71	26.25	37.61	28.91
	PixArt-α-μP	0.26	29.96	27.13	25.84	29.58

μTransfer results of MMDiT-18B



(a) Base learning rate. (b) Base gradient clip. (c) Base REPA loss weight. (d) Base Warm-up steps.

Figure 4. We train **0.18B MMDiT-μP proxy models** with 80 different base HPs settings. The optimal base HPs are transferred to the training of the 18B target model.

Table 3. **Evaluation results of pretrained MMDiT-18B and MMDiT-μP-18B models.** MMDiT-μP-18B achieves better benchmark results with only 3% of the manual tuning cost.

Method	GenEval Overall ↑	Human Evaluation ↑
MMDiT-18B	0.815	0.703
MMDiT-μP-18B	0.822	0.715