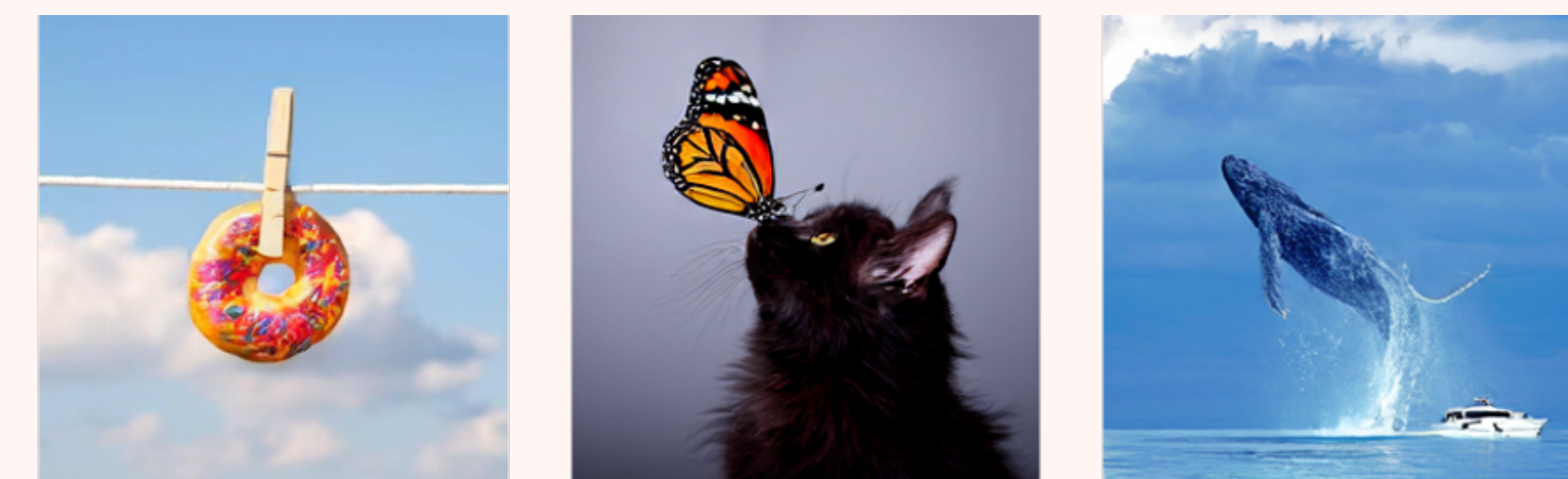


Abstract

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 (μ P) 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 μ P of vanilla Transformers extends to diffusion Transformers, which differ architecturally and objectively. In this work, we generalize standard μ P to diffusion Transformers and validate its effectiveness through large-scale experiments. First, we rigorously prove that μ P 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 μ P methodologies. Leveraging this result, we systematically demonstrate that DiT- μ P enjoys robust HP transferability. Notably, DiT-XL-2- μ P with transferred learning rate achieves 2.9 \times faster convergence than the original DiT-XL-2. Finally, we validate the effectiveness of μ P 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 μ P 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 μ P as a principled and efficient framework for scaling diffusion Transformers.

Highlights

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

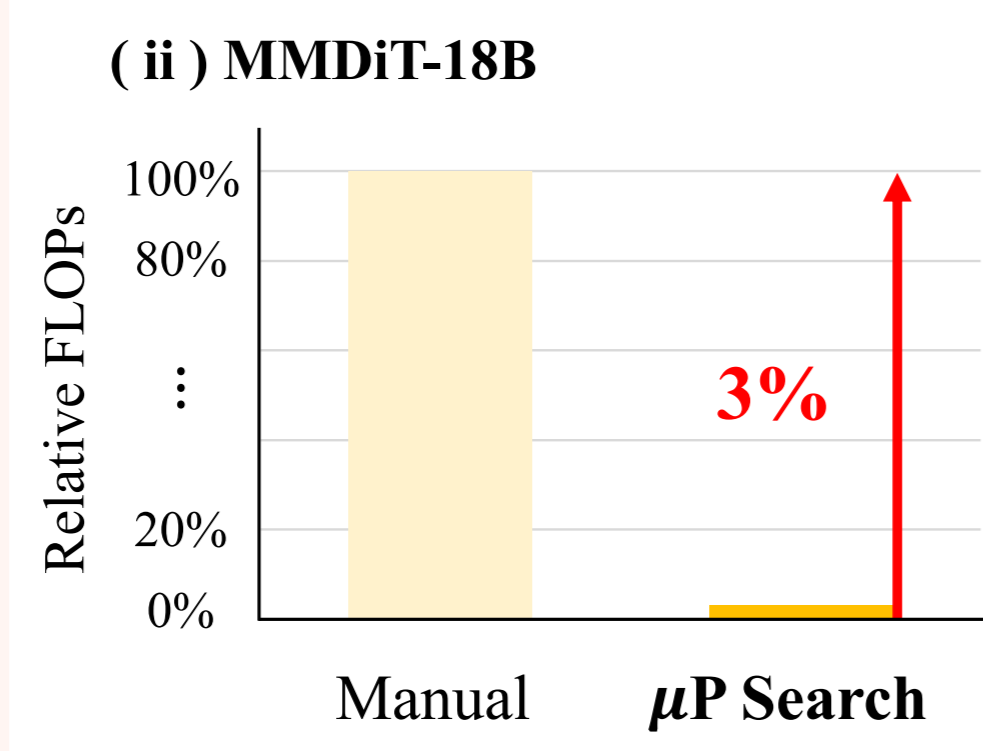
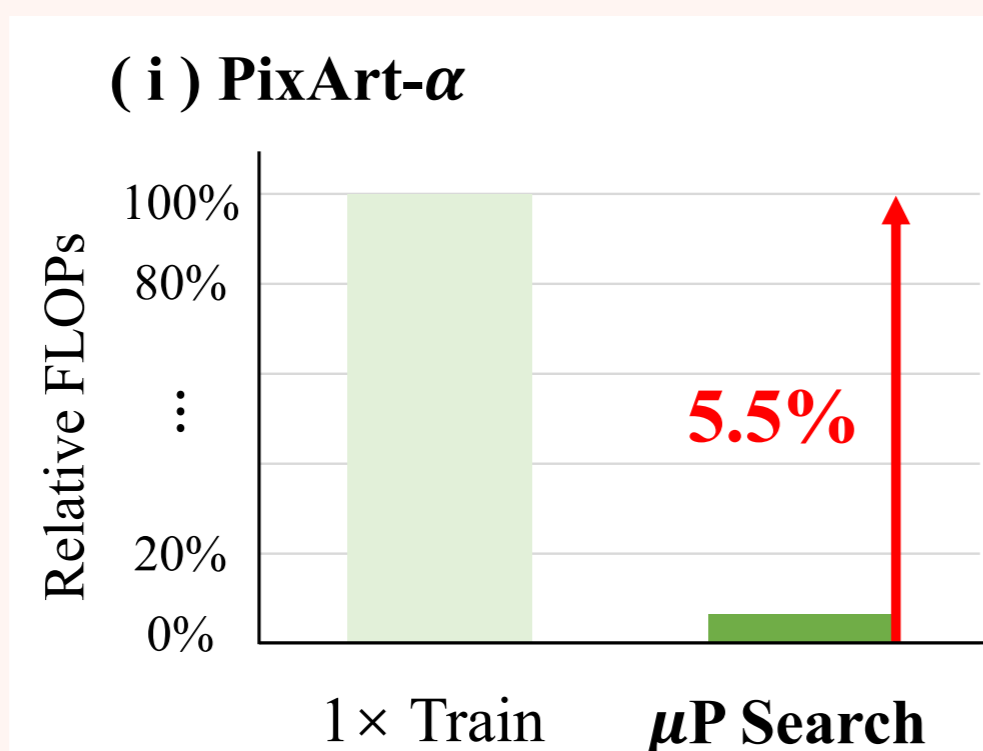


A colorful donut fixed on a white clothesline, with a bright sky and fluffy white clouds in the background
A furry little black cat looked up and saw a beautiful butterfly landing on its nose
Blue sky, calm sea, a huge blue whale leaps from the sea, and the yacht below is small



Cat in Picasso's Painting Style
A cute teddy bear holding a guitar stands in a scene with star lights and grass
Flowers grow inside the spacesuit

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



(b) Efficiency of μ P search.

Figure 1. Visualization results and efficiency of HP search under μ P. (a) Samples generated by the MMDiT- μ P-18B model exhibit strong fidelity. (b) HP search for large diffusion Transformers is efficient under μ P, 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 μ P

Table 1. μ P for Standard Transformers with Adam/AdamW optimizer. We use purple text to highlight the differences between μ P 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
Weight decay	λ	λN (λ)	λ

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

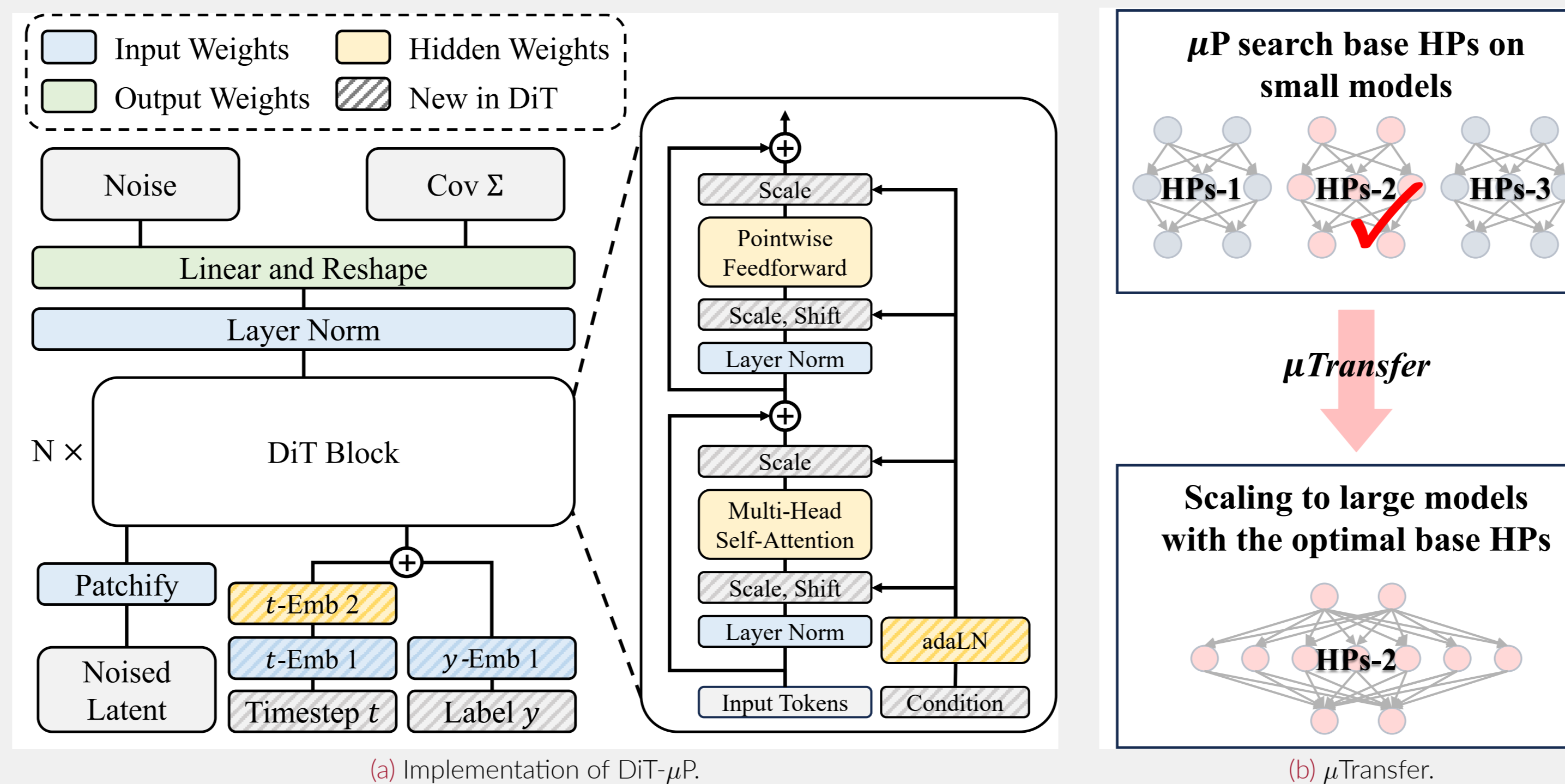


Figure 2. A overview of applying μ P to diffusion Transformers. (a) We illustrate the implementation of μ P for DiT as an example. The abc -parameterization of each weight is adjusted based on its type and visualized using different colors. (b) We μ Transfer the optimal base HPs searched from multiple trials on small models to pretrain the target large models.

Stable Hyperparameter Transferability of μ P

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

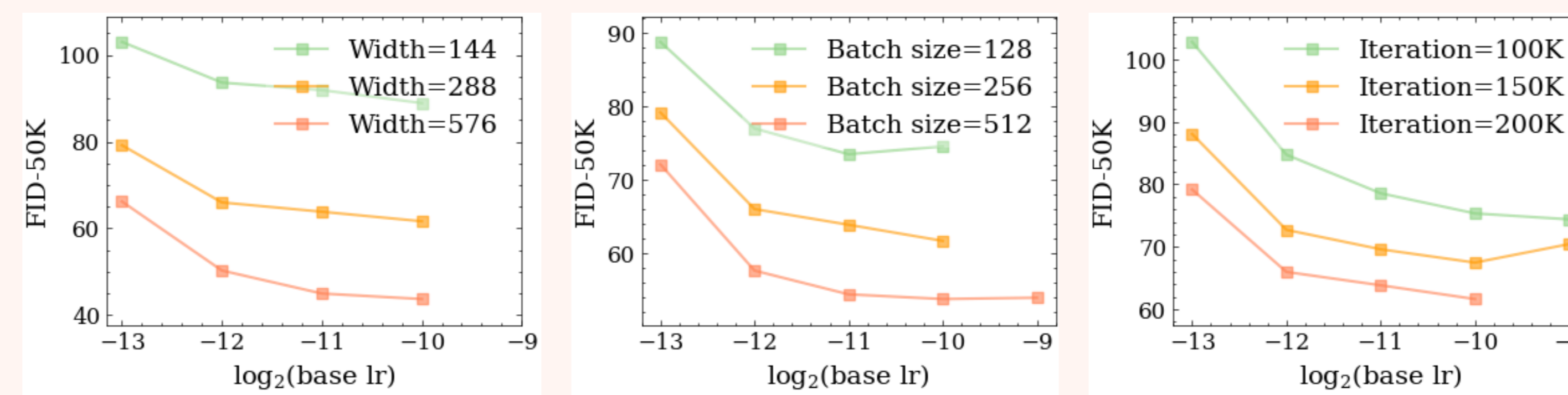
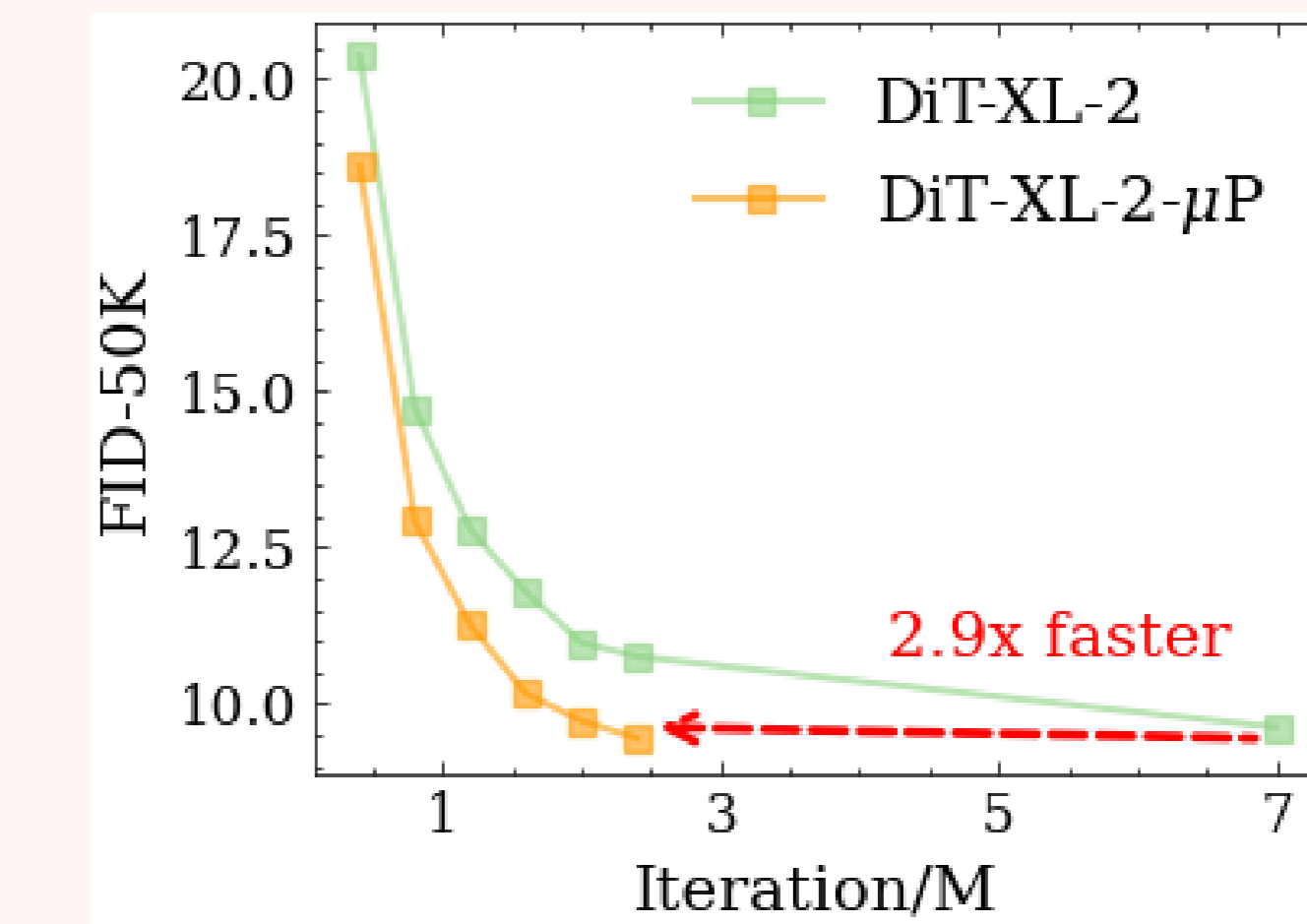


Figure 3. DiT- μ P enjoys base HP transferability. Unless otherwise specified, we use a model width of 288, a batch size of 256, and a training iteration of 200K. Missing data points indicate training instability, where the loss explodes. Under μ P, the base learning rate can be transferred across model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

μ Transfer results of DiT-XL-2 on ImageNet

Considering FID-50K, DiT-XL-2- μ P with transferred learning rate achieves 2.9 \times faster convergence than the original DiT-XL-2 and a slightly better result.



μ Transfer results of PixArt-alpha on SA-1B

Table 2. Comprehensive comparison between PixArt- α - μ P and PixArt- α . Both models are trained on the SAM dataset for 30 epochs. PixArt- α - μ P (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 \uparrow	MJHQ		MS-COCO	
			FID-30K \downarrow	CLIP Score \uparrow	FID-30K \downarrow	CLIP Score \uparrow
10	PixArt- α	0.19	38.36	25.78	34.58	28.12
	PixArt- α - μ P	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 - μ P	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 - μ P	0.26	29.96	27.13	25.84	29.58

μ Transfer results of MMDiT-18B on 800M dataset

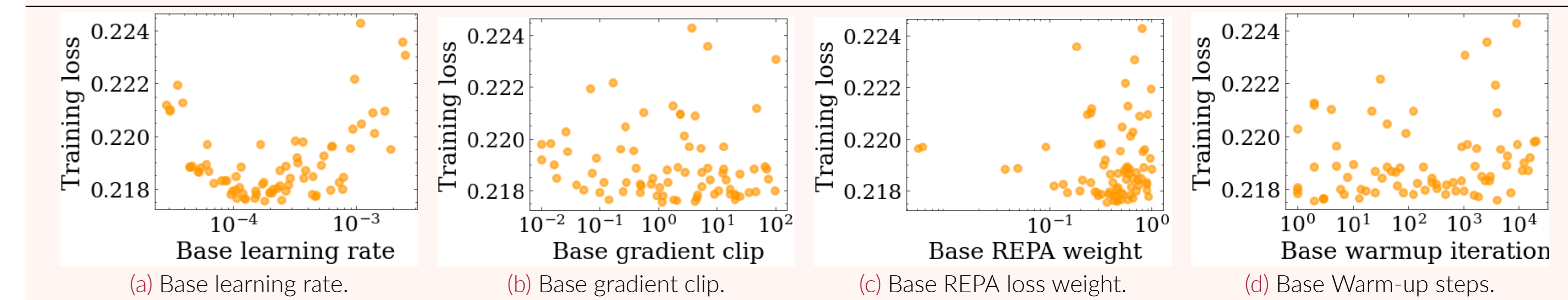


Figure 4. We train 0.18B MMDiT- μ P proxy models with 80 different base HPs settings. The optimal base HPs are transferred to the training of the 18B target model.

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

Method	GenEval Overall \uparrow	Human Evaluation \uparrow
MMDiT-18B	0.815	0.703
MMDiT- μ P-18B	0.822	0.715