An Overview of Maximal Update Parametrization (μP)

中国人民大学 高瓴人工智能学院 郑晨宇

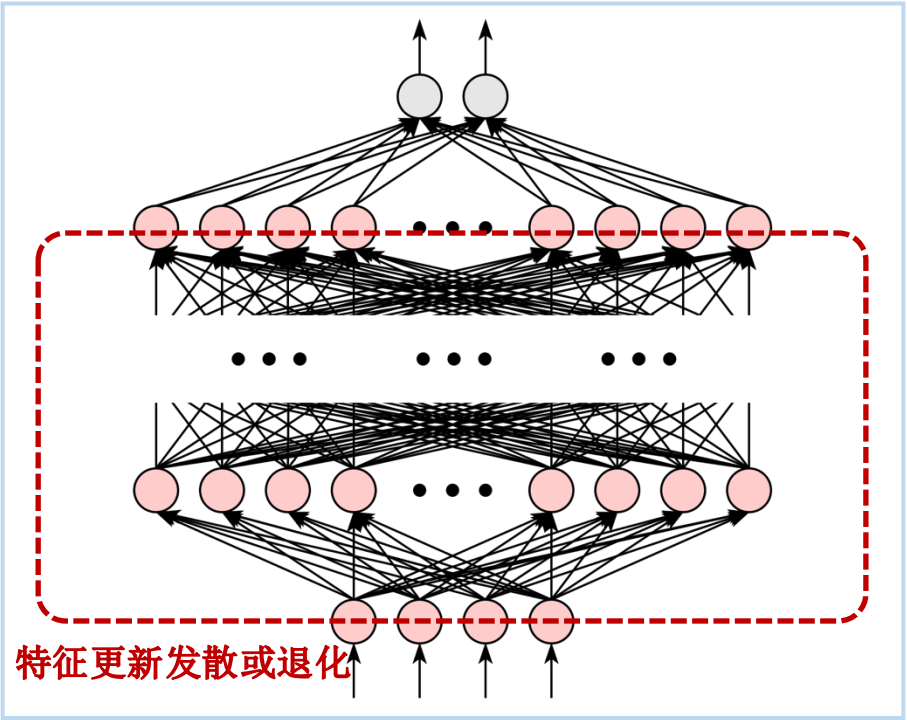
导师：李崇轩

内容纲要

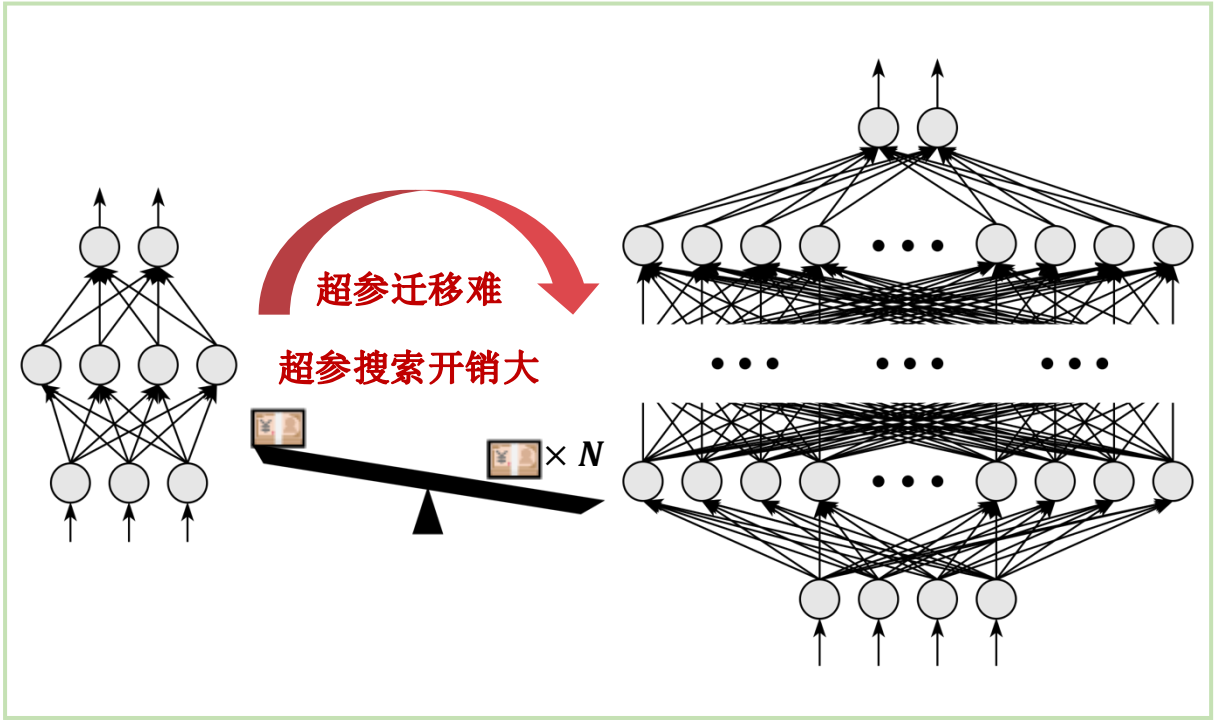
- μP 的背景与影响
- 宽度扩展的 μP 理论与实践
- 宽深联合扩展的 μP 理论与实践
- 总结与展望

1.1 μP 的实际背景——大模型训练优化难

特征学习能力在无穷宽深时失效

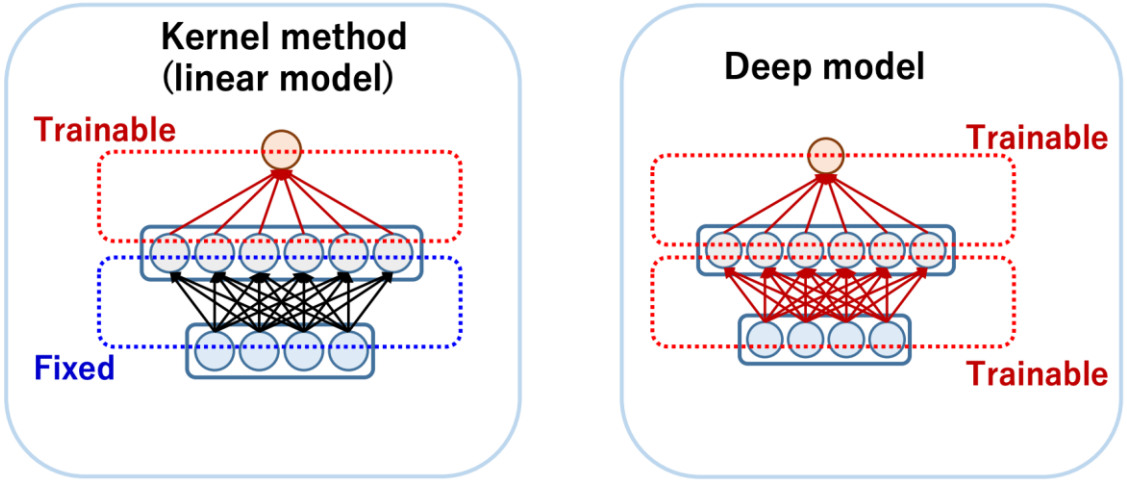


大模型训练超参数搜索开销大、迁移难



1.2 μP 的理论背景——深度学习理论发展的必经之路

- 传统深度学习理论：kernel regime（特征冻结）
  - Random feature model、Neural tangent kernel 等
- 新兴深度学习理论：feature learning regime（特征可学）
  - 无穷宽/深极限时的行为：μP/Tensor Program/DMFT/Mean-field theory/…

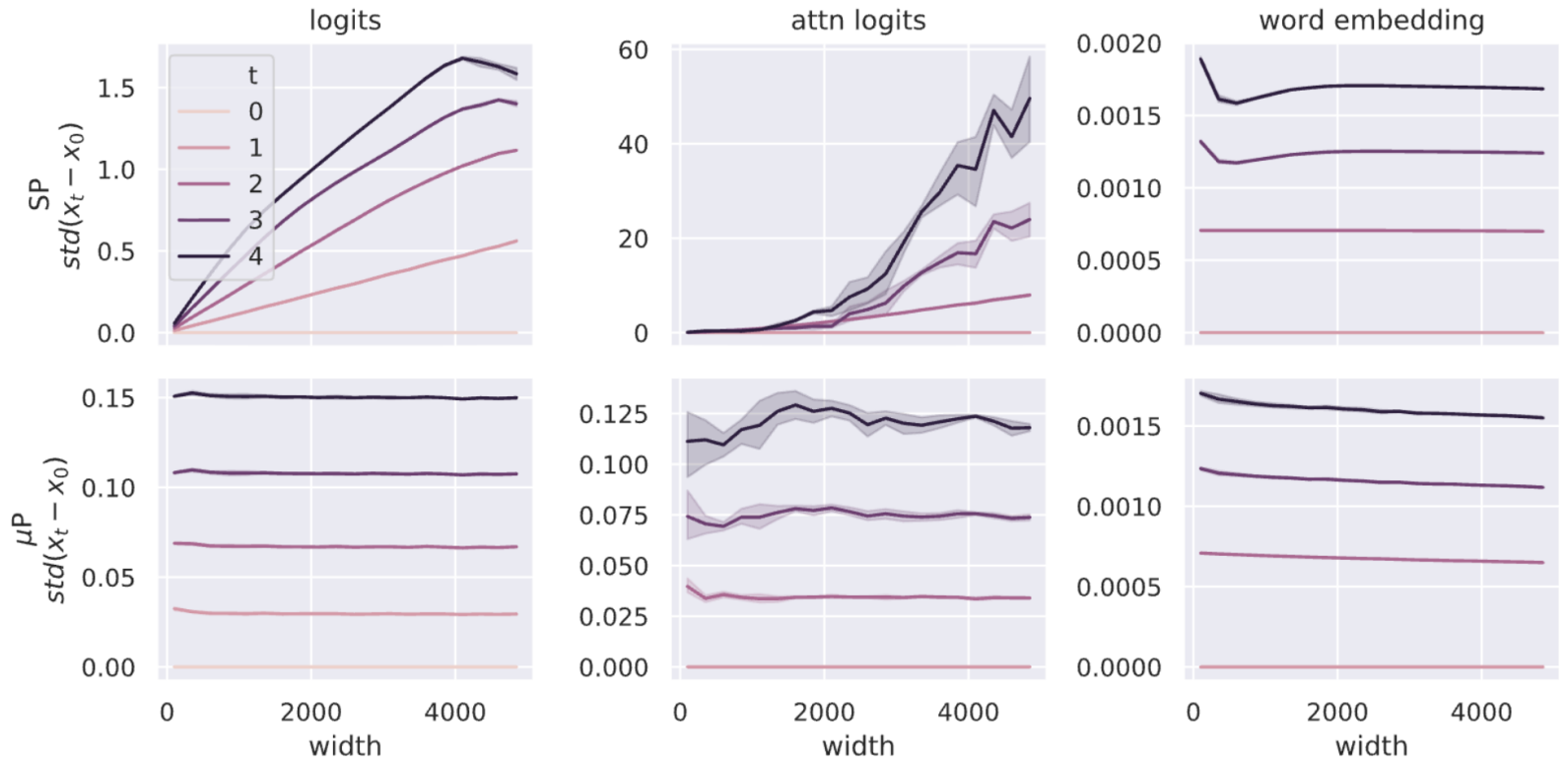


1.3 μP 的实际影响

- μP 已经作为一种准则，被应用于前沿工业界的大模型预训练
  - 发明人 Greg Yang 为 xAI co-founder，μP 被用于 Grok 全系列模型训练
  - μP 原始算法论文与 OpenAI 合作训练 GPT-3，GPT 后续系列也使用了μP
  - DeepMind 发表多篇 μP 相关的前沿探索文章，也非常重视该类技术
  - Meta 的 Llama-4 报告中也明确提到利用了类似 μP 的 “MetaP”

1.3 μP 的实际影响

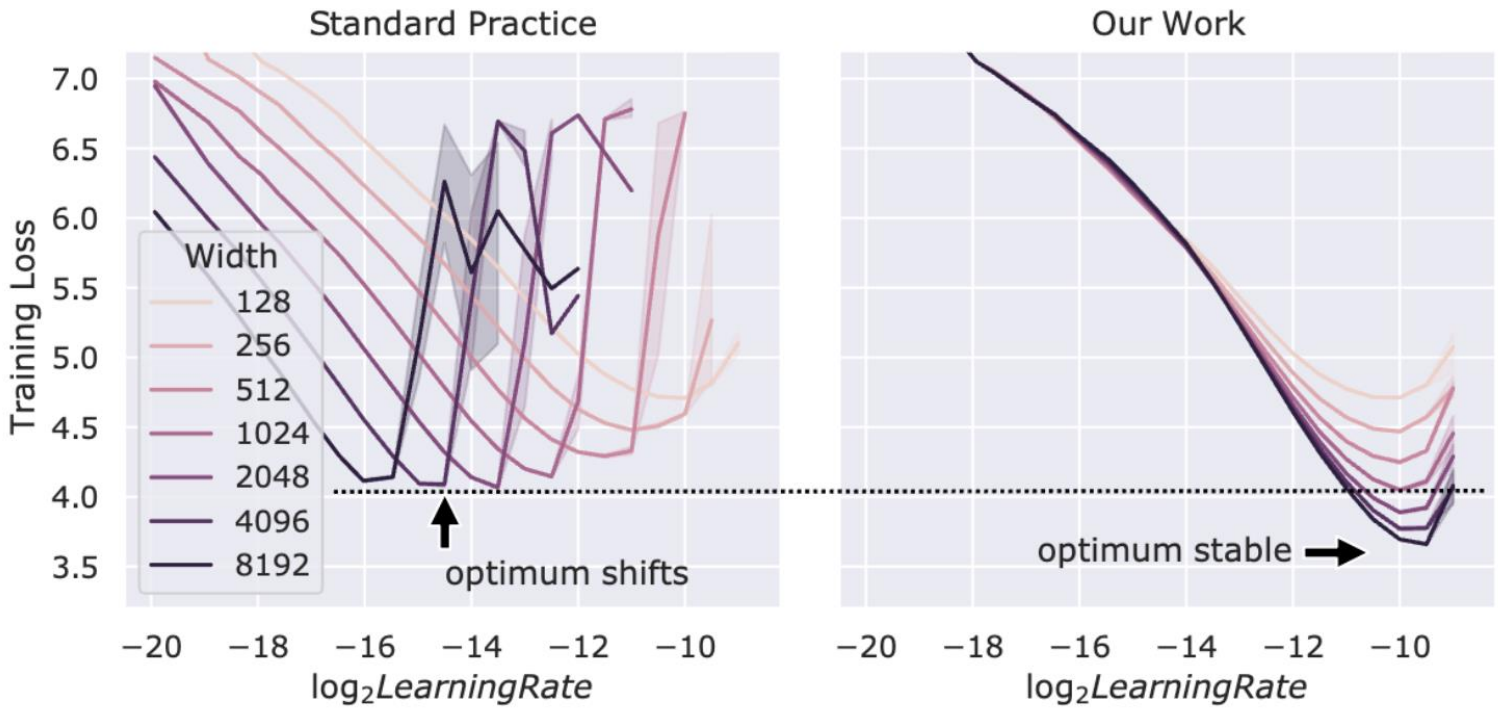
- μP 能够原则上保证规模扩展时的特征学习稳定性



Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer, NeurIPS, 2022

1.3 μP 的实际影响

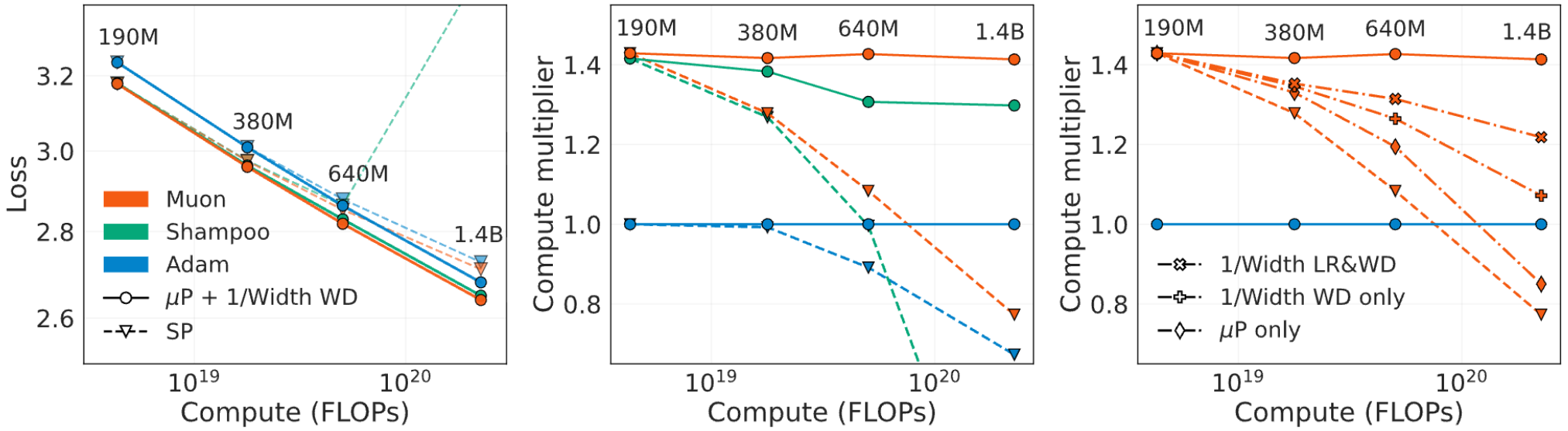
- μP 能够经验上指导超参数在规模扩展时的迁移律



Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer, NeurIPS, 2022

1.3 μP 的实际影响

- μP 能够有助于公平比较不同方法在大规模场景下的优劣



内容纲要

- μP 的背景与影响
- **宽度扩展的 μP 理论与实践**
- 宽深联合扩展的 μP 理论与实践
- 总结与展望

2.1 μP 的原则——最大稳定更新

- μP 原则上希望在保证特征变动稳定（量级和网络规模无关）的前提下尽可能更新权重（高效训练）。否则，特征随网络变大趋于爆炸或者冻结。

设 $h_l \in R^{n_l}$ 为第 l 层的特征，$\Delta h_l \in R^{n_l}$ 为其经过一步优化的更新量，μP 希望：

$$\|h_l\|_{RMS} = \Theta(1), \|\Delta h_l\|_{RMS} = \Theta(1), \qquad l \in [L]$$

$$maximize\ \Delta W_l's\ contribution\ to\ \Delta h_L, \qquad l \in [L]$$

其中 $\|v\|_{RMS} = \frac{\|v\|_2}{\sqrt{d}} = \sqrt{\frac{\sum v_i^2}{d}}$，衡量是特征向量每个维度的平均大小。

A Spectral Condition for Feature Learning, arxiv, 2023

2.2 宽度扩展 μP 的谱条件理论——原型与结果

- 最简理论原型：能反映实际情况（如 Transformer）的最简单分析对象
  - 有限层 linear MLP（$h_l = W_l h_{l-1}$），一步梯度下降，batch size 为 1
- 为了满足宽度扩展时的 μP 原则，可令权重及其变化满足以下条件

设 $W_l \in R^{n_l \times n_{l-1}}$ 为第 l 层的权重，$\Delta W_l \in R^{n_l \times n_{l-1}}$ 为其经过一步优化的变动，需要：

$$\|W_l\|_{RMS} = \Theta(1), \|\Delta W_l\|_{RMS} = \Theta(1), l \in [L]$$

其中 $\|W\|_{RMS} = \sup_{x \neq 0} \frac{\|Wx\|_{RMS}}{\|x\|_{RMS}} = \sqrt{\frac{n_{l-1}}{n_l}} \|W\|_2$。

A Spectral Condition for Feature Learning, arxiv, 2023

2.2 宽度扩展 μP 的谱条件理论——核心推导

- $\|W_l\|_{RMS} = \Theta(1)$ 的推导

$$\|h_l\|_{RMS} = \|W_l h_{l-1}\|_{RMS} \leq \|W_l\|_{RMS}\|h_{l-1}\|_{RMS} = \Theta(1)$$

$$\|h_{l-1}\|_{RMS} = \Theta(1) \Rightarrow \|W_l\|_{RMS} = \Theta(1)$$

- $\|\Delta W_l\|_{RMS} = \Theta(1)$ 的推导

$$\|\Delta h_l\|_{RMS} = \|\Delta W_l h_{l-1} + \Delta W_l \Delta h_{l-1} + W_l \Delta h_{l-1}\|_{RMS}$$

$$\leq \|\Delta W_l\|_{RMS}(\|h_{l-1}\|_{RMS} + \|\Delta h_{l-1}\|_{RMS}) + \|W_l\|_{RMS}\|\Delta h_{l-1}\|_{RMS} = \Theta(1)$$

$$\|h_{l-1}\|_{RMS}, \|\Delta h_{l-1}\|_{RMS}, \|W_l\|_{RMS} = \Theta(1) \Rightarrow \|\Delta W_l\|_{RMS} = \Theta(1)$$

注：μP 原则上想最大化的是权重更新带来的特征变化

A Spectral Condition for Feature Learning, arxiv, 2023

2.2 宽度扩展 μP 的谱条件理论——Adam 实现

- $\|W_l\|_{RMS} = \Theta(1)$ 可以由适当的高斯初始化得到，$W_l \sim N(0, \sigma_l^2)$

$$\sigma_l = \Theta\left(\frac{1}{\sqrt{n_{l-1}}}\min\left(1, \sqrt{\frac{n_l}{n_{l-1}}}\right)\right)$$

- $\|\Delta W_l\|_{RMS} = \Theta(1)$ 可以通过随宽度缩放学习率得到，如对于Adam 有

$$\|\Delta W_l\|_2 \approx \eta_l \|\mathrm{sign}(G_l)\|_2 \approx \eta_l \|\mathrm{sign}(G_l)\|_F = \Theta\left(\sqrt{\frac{n_l}{n_{l-1}}}\right)$$

$$\|\mathrm{sign}(G_l)\|_F = \sqrt{n_{l-1} n_l} \implies \eta_l = \Theta\left(\frac{1}{n_{l-1}}\right)$$

A Spectral Condition for Feature Learning, arxiv, 2023

2.2 宽度扩展 μP 的谱条件理论——启发 Muon 优化器

- $\|W_l\|_{RMS} = \Theta(1)$ 可以由适当的高斯初始化得到，$W_l \sim N(0, \sigma_l^2)$

$$\sigma_l = \Theta\left(\frac{1}{\sqrt{n_{l-1}}} \min\left(1, \sqrt{\frac{n_l}{n_{l-1}}}\right)\right)$$

- $\|\Delta W_l\|_{RMS} = \Theta(1)$ 直接启发了 Muon 的设计（不用随宽度做学习率调整）

$$\|\Delta W_l\|_2 = \eta_l \, \|UV^\top\|_2 = \eta_l = \Theta\left(\sqrt{\frac{n_l}{n_{l-1}}}\right) = \Theta(1)$$

$$\Longrightarrow \eta_l = \Theta(1) \text{ (hidden weights)}$$

A Spectral Condition for Feature Learning, arxiv, 2023

2.3 宽度扩展 μP 的实现——以 Adam 为例

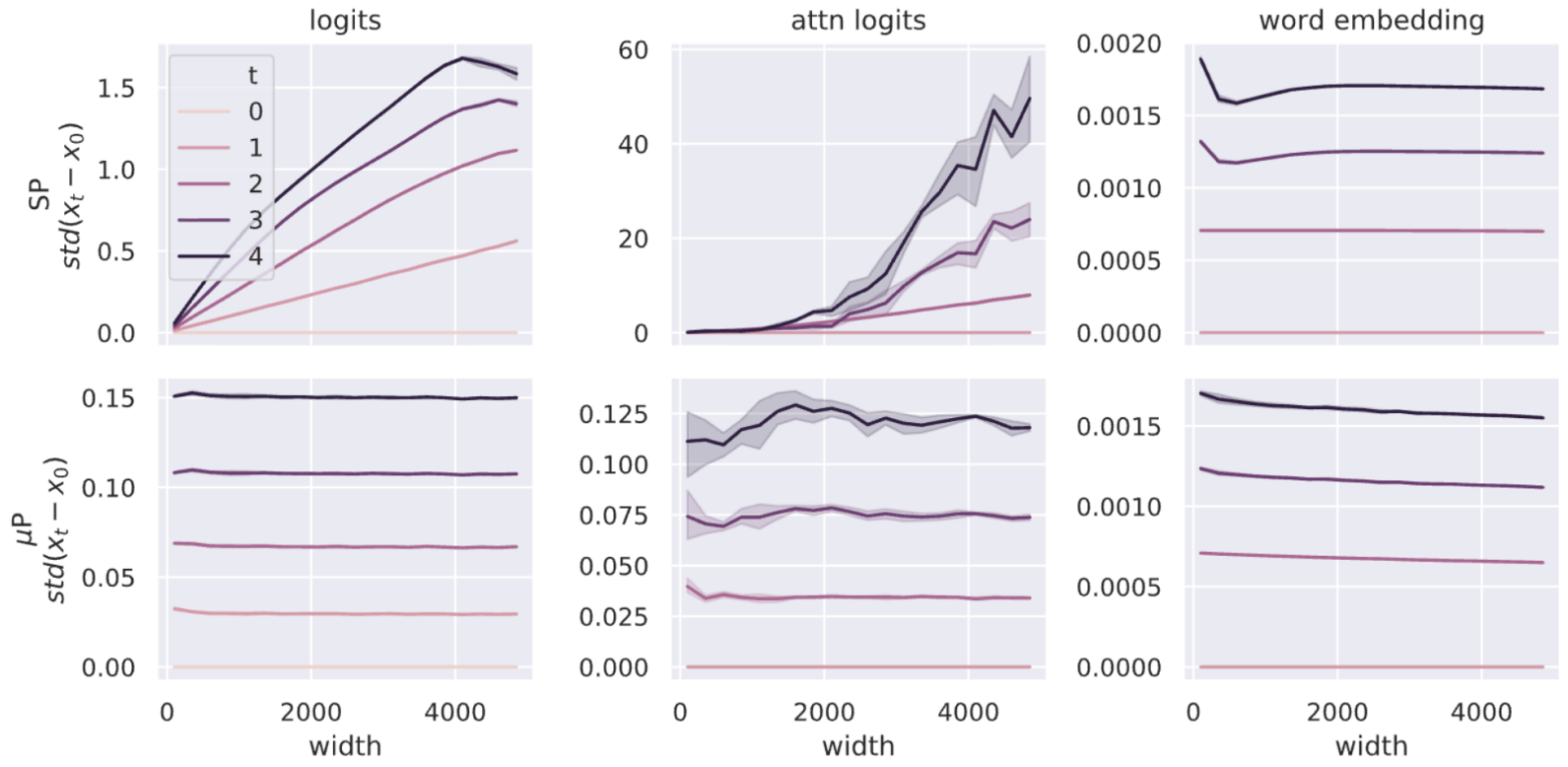
- 在 base model 上用 SP 搜索最好的超参，比如学习率 η_{base}
- 在 target model 上，使用 μP 结论缩放超参，如使用 $\eta_{base}/r_n = \Theta(1/n)$
  - r_n 为 target model 与 base model 的宽度比值，如 2048/256

	Input weights & bias	Hidden weights	Output weights
Block Multiplier	α_{base}	(α_{base})	α_{base}/r_n (α_{base})
Initial Variance	$\sigma^2_{\text{base}}/d_0$ or σ^2_{base}	$\sigma^2_{\text{base}}/r_n$	σ^2_{base} $(\sigma^2_{\text{base}}/r_n)$
Learning Rate	η_{base}	η_{base}/r_n (η_{base})	η_{base}
Weight Decay	λ_{base}	$\lambda_{\text{base}}r_n$ (λ_{base})	λ_{base}

Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer, NeurIPS, 2022

2.3 宽度扩展 μP 的实现——稳定训练

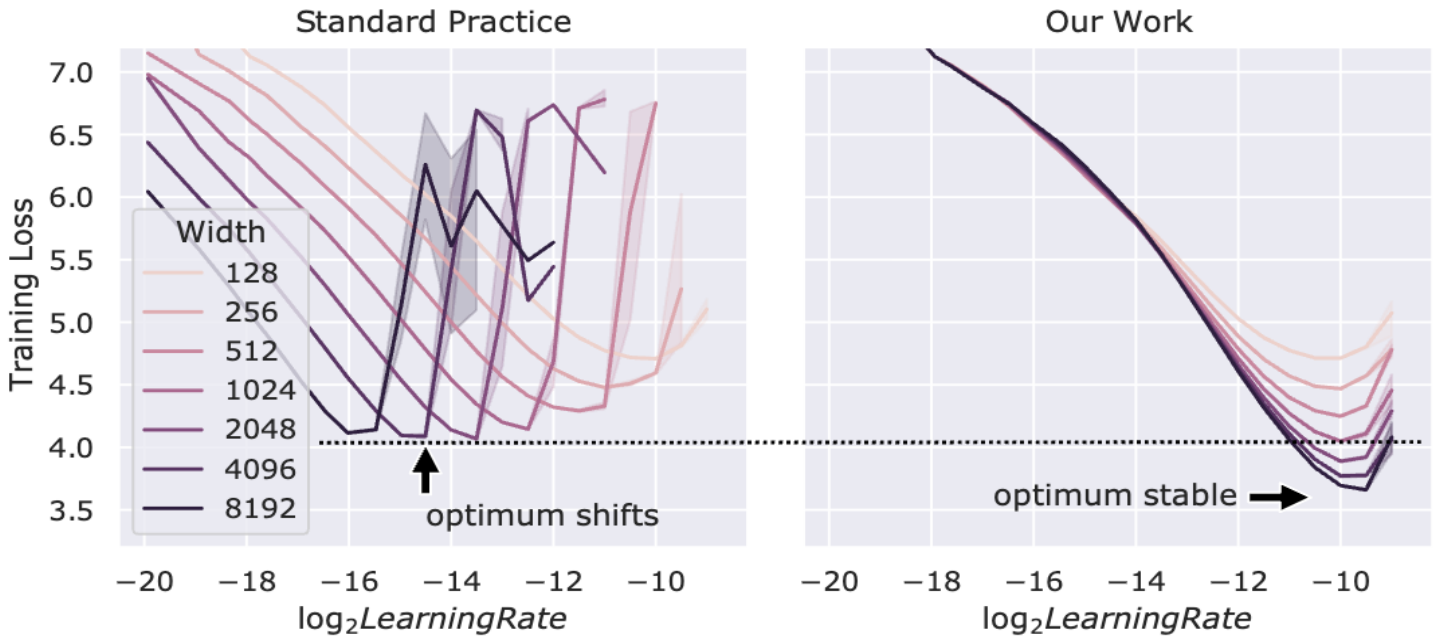
- μP 从理论上保证了不同层特征的稳定更新，从而支持大模型稳定训练



Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer, NeurIPS, 2022

2.3 宽度扩展 μP 的实现——超参迁移律/超参缩放律

- μP 参数化下的最优 base 超参数随着网络宽度增大而稳定迁移
- 直觉：如果不用这个迁移律，则特征在极限情况趋向退化/爆炸，非最优



Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer, NeurIPS, 2022

2.4 宽度扩展 μP 的 Tensor Program 理论——对象与结果

- Tensor Program 将 toy model 上谱条件得到的 μP 结论推广到一般情况
  - 一般的架构：包括 MLP、CNN、Transformer 等
  - 一般的 element-wise adaptive optimzier：包括 SGD、Adam 等
  - 训练步数为 $\Theta(1)$，且训练没收敛
- 在上述条件下，μP 结论与谱范数理论结果一致

Tensor Programs IVb: Adaptive Optimization in the Infinite-Width Limit, ICLR, 2023

2.4 宽度扩展 μP 的 Tensor Program 理论——架构范畴

- Tensor Program 的网络权重初始化要求：高斯初始化
  - 也可以拓展到一般的零均值、矩有界的分布

Setup 2.6.3 (Gaussian). *Assume*[14]

1. *Every entry of every* $W \in \mathcal{W}$ *is sampled iid from* $\mathcal{N}(0, 1/n)$.
2. *Every entry of every initial vector* $x \in \boldsymbol{x}^0$ *is sampled iid from* $\mathcal{N}(0, 1)$.
3. *The initial scalars* $\boldsymbol{c}^0$ *converge almost surely to 0.*
4. *All functions* ψ *used in* OUTERNONLIN *are pseudo-Lipschitz.*

Tensor Programs IVb: Adaptive Optimization in the Infinite-Width Limit, ICLR, 2023

2.4 宽度扩展 μP 的 Tensor Program 理论——架构范畴

- 网络前传到 final feature 的过程能由如下三个算子表示，如 Transformer：

`Avg` We can choose a existing vector $\boldsymbol{x} \in \mathbb{R}^n$ and append to the program a scalar

$$\frac{1}{n}\sum_{\alpha=1}^{n} x_\alpha \in \mathbb{R}.$$

`MatMul` We can choose a matrix $\boldsymbol{W} \in \mathbb{R}^{n\times n}$ and vector $\boldsymbol{x} \in \mathbb{R}^n$ existing in the program, and append to the program a vector

$$\boldsymbol{W}\boldsymbol{x} \in \mathbb{R}^n \quad \text{or} \quad \boldsymbol{W}^\top \boldsymbol{x} \in \mathbb{R}^n.$$

`OuterNonlin` For any integer $k, l \geq 0$, we can aggregate k existing vectors and l existing scalars to $\boldsymbol{X} \in \mathbb{R}^{n\times k}$ and $\boldsymbol{c} \in \mathbb{R}^l$, respectively. With an integer $r \geq 0$ and a pseudo-Lipschitz function $\psi : \mathbb{R}^{k(r+1)+l} \to \mathbb{R}$ (e.g., SiLU, GeLU), we append to the program a vector

$$\boldsymbol{y} \in \mathbb{R}^n, \quad y_\alpha = \frac{1}{n^r}\sum_{\beta_1,\ldots,\beta_r=1}^{n} \psi(\boldsymbol{X}_{\alpha:}; \boldsymbol{X}_{\beta_1:}; \ldots; \boldsymbol{X}_{\beta_r:}; \boldsymbol{c}^\top),$$

where the $r+1$ is called the order of ψ.

Tensor Programs IVb: Adaptive Optimization in the Infinite-Width Limit, ICLR, 2023

2.4 宽度扩展 μP 的 Tensor Program 理论——优化器范畴

- Tensor Program 理论能够考虑如下优化器集合：
  - $w_t^i = w_{t-1}^i - \eta Q_t(g_0^i, \ldots, g_t^i)$，其中 i 为参数的维度 index，g 为梯度
  - 经典的 SGD、Adam、Lion 等优化器均满足
  - 现代的矩阵预条件优化器如 Muon 等不满足

$$Q_t(g_0, \ldots, g_t) = \frac{\frac{1}{1-\beta_1^{t+1}} \sum_{s=0}^{t} (1-\beta_1)\beta_1^{t-s} g_s}{\sqrt{\frac{1}{1-\beta_2^{t+1}} \sum_{s=0}^{t} (1-\beta_2)\beta_2^{t-s} g_s^2 + \epsilon^2}}, \qquad \text{(Adam)}$$

Tensor Programs IVb: Adaptive Optimization in the Infinite-Width Limit, ICLR, 2023

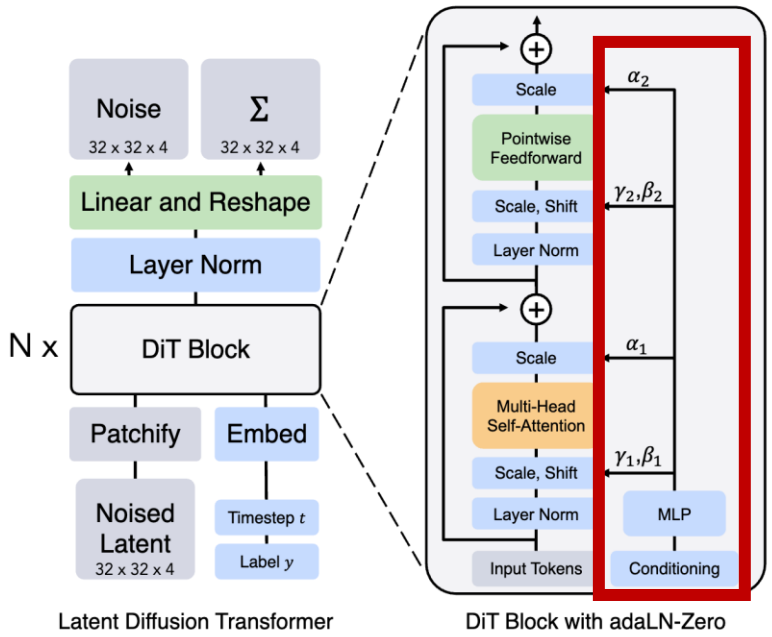
2.5 宽度扩展 μP 的 Diffusion Transformer 应用——研究问题

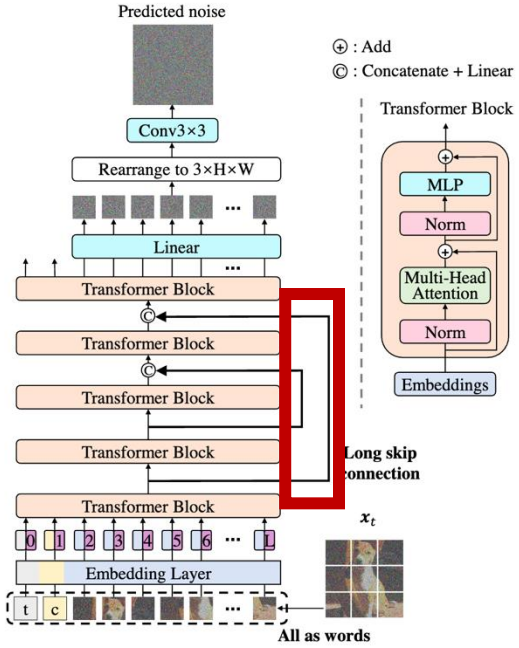
- 严谨的 μP 理论仅对 Tensor Program 可表示的架构成立，该结果对 diffusion Transformer 是否成立仍然未知
- 超参迁移性只在语言模型的场景下被验证，视觉生成任务是否成立未知
- μP 能够给 diffusion Transformers 带来多大的赋能未知

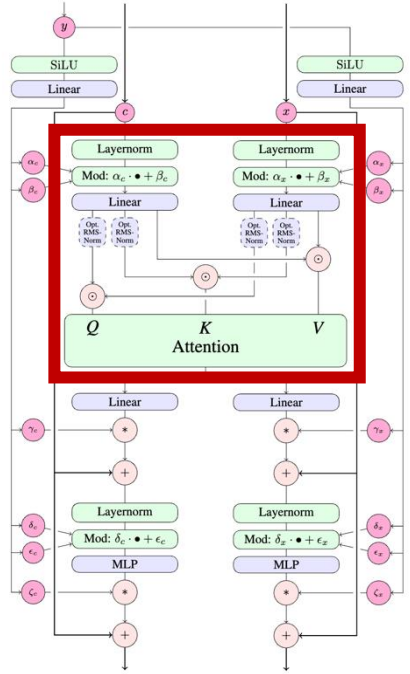
Scaling Diffusion Transformers Efficiently via μP, NeurIPS, 2025

2.5 宽度扩展 μP 的 Diffusion Transformer 应用——理论结果

- U-ViT, DiT, PixArt, MM-DiT 等主流扩散模型架构可以被 Tensor Program 表示，所以 μP 结论保持不变

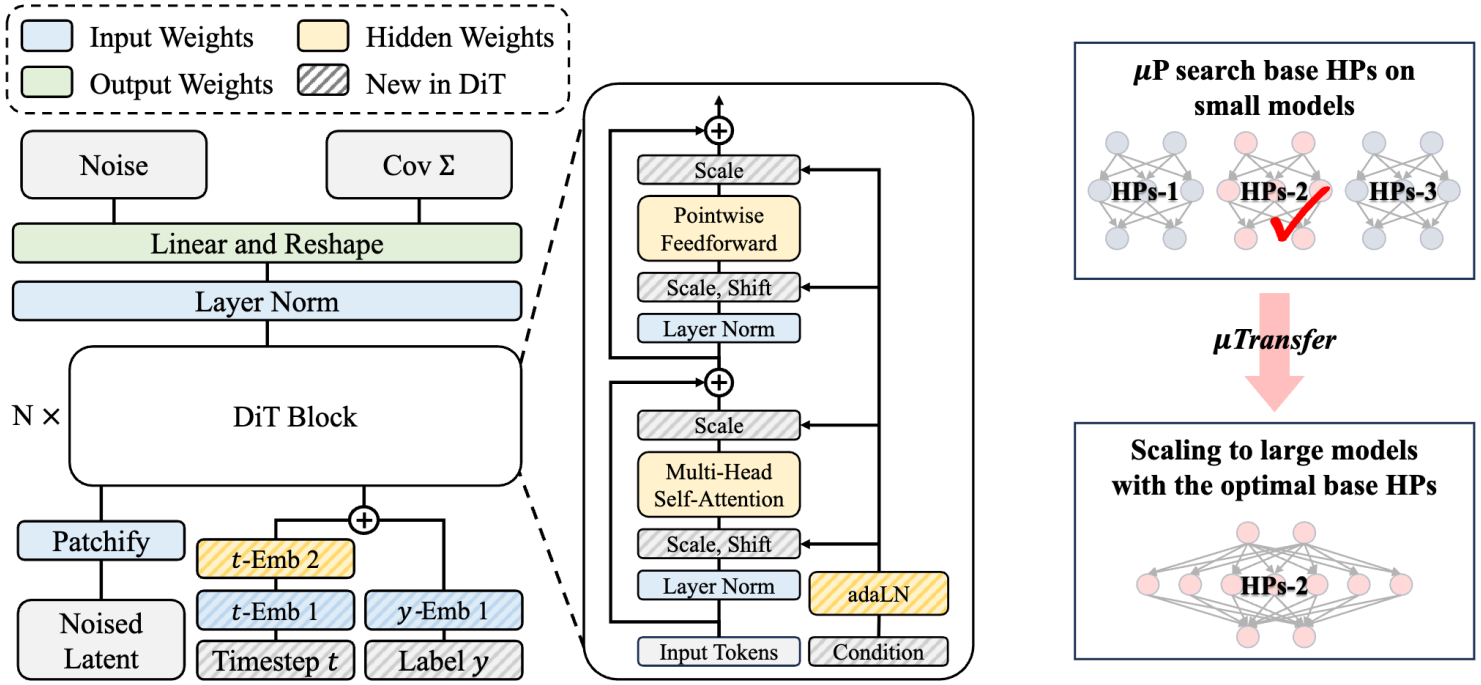






2.5 宽度扩展 μP 的 Diffusion Transformer 应用——方法

- 实现和原来一致的 μP 参数化，并进行超参迁移实验

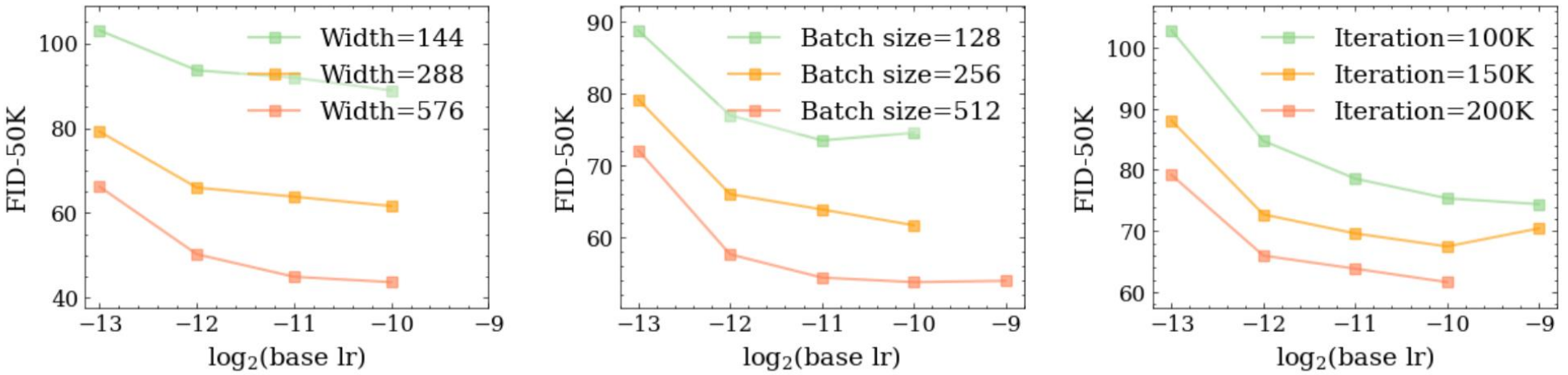


(a) Implementation of DiT-μP.

(b) μTransfer.

2.5 宽度扩展 μP 的 Diffusion Transformer 应用——DiT

- 通过 DiT 在 ImageNet 上的图像生成实验验证了超参数的迁移性

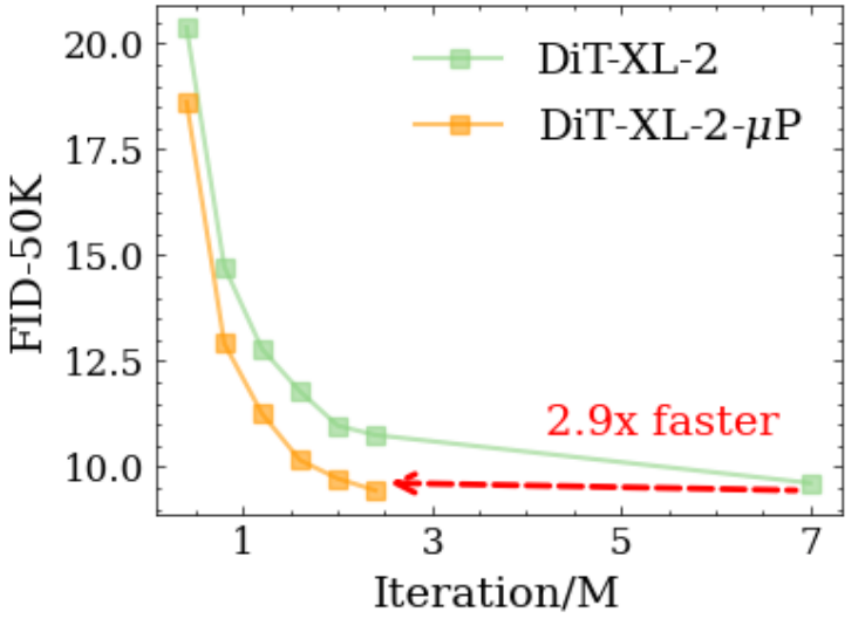


(a) HP transfer across widths. **(b)** HP transfer across batch sizes. **(c)** HP transfer across iterations.

Scaling Diffusion Transformers Efficiently via μP, NeurIPS, 2025

2.5 宽度扩展 μP 的 Diffusion Transformer 应用——DiT

- DiT 43M 搜索的学习率迁移到 675M，相较于基线取得了 **2.9** 倍的加速
  - 宽度扩大约 **4** 倍，训练步数扩大 **12** 倍实现稳定迁移



Scaling Diffusion Transformers Efficiently via μP, NeurIPS, 2025

2.5 宽度扩展 μP 的 Diffusion Transformer 应用——PixArt

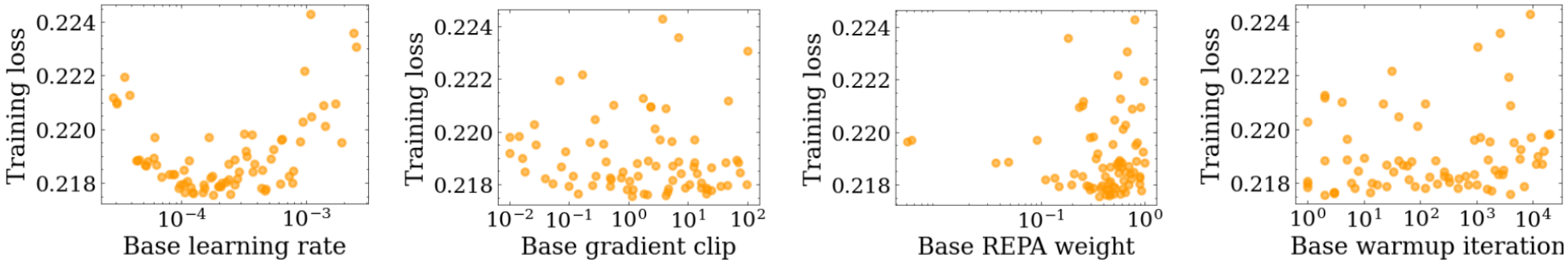
- 搜超参场景：模型 39M，batch size 176x8，训练步数 39K
- 预训练场景：模型 677M，batch size 176x32，训练步数 59K
- 5 次搜索的 FLOPs 为一次预训练的 5.5%，训练结果一致超过基线

Epoch	Method	GenEval ↑	MJHQ		MS-COCO	
			FID-30K ↓	CLIP Score ↑	FID-30K ↓	CLIP Score ↑
10	PixArt-α [8]	0.19	38.36	25.78	34.58	28.12
	PixArt-α-μP (**Ours**)	**0.20**	**33.35**	**26.25**	**29.68**	**28.87**
20	PixArt-α [8]	0.20	35.68	26.54	30.13	28.81
	PixArt-α-μP (**Ours**)	**0.23**	**33.42**	**26.83**	**29.05**	**29.53**
30	PixArt-α [8]	0.15	42.71	26.25	37.61	28.91
	PixArt-α-μP (**Ours**)	**0.26**	**29.96**	**27.13**	**25.84**	**29.58**

Scaling Diffusion Transformers Efficiently via μP, NeurIPS, 2025

2.5 宽度扩展 μP 的 Diffusion Transformer 应用——MMDiT 18B

- 搜超参场景：模型 0.18B，batch size 4096，训练步数 30K
- 预训练场景：模型 18B，batch size 4096，训练步数 200K
- 80 次搜索的 FLOPs 为一次预训练的 14.5%，大概是专家的 3%



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

Scaling Diffusion Transformers Efficiently via μP, NeurIPS, 2025

2.5 宽度扩展 μP 的 Diffusion Transformer 应用——MMDiT 18B

- 18B 的预训练实验中机评和人工评测均超过基线

Table 4 GenEval 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	Overall ↑	Single	Two	Counting	Colors	Position	Color Attribution
MMDiT-18B	0.8154	99.38	93.69	**81.88**	88.03	57.5	**68.75**
MMDiT-μP-18B	**0.8218**	**99.38**	**94.44**	79.69	**88.83**	**62.25**	68.5

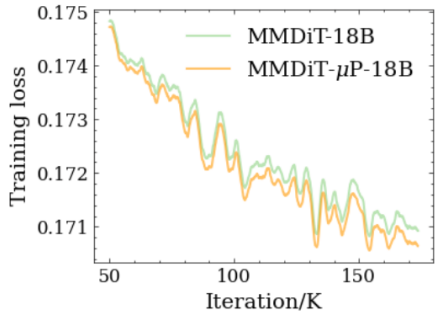


Figure 6 MMDiT-μP-18B achieves consistently lower training loss than baseline after 15K steps.

Table 5 Results of human evaluation for text-image alignment. The alignment accuracy (acc.) is computed as the average over 22,500 human alignment tests. MMDiT-μP-18B achieves superior results with only 3% of the manual tuning cost.

Method	Alignment acc. ↑
MMDiT-18B	0.703
MMDiT-μP-18B (**Ours**)	**0.715**

内容纲要

- μP 的背景与影响
- 宽度扩展的 μP 理论与实践
- 宽深联合扩展的 μP 理论与实践
- 总结与展望

2.1 宽深联合 μP 的前期研究

- 前期的宽深联合 μP 研究呈现如下特征
  - 架构不同：$h_{l+1} = h_l + \alpha_l F_l(h_l)$，考虑的残差块 $F_l(h_l)$ 不同
  - 优化器特定：特定为 SGD 或者 Adam
  - 理论推导复杂：多基于 Tensor Program、DMFT 或者梯度分析
  - 结果各不相同：不同情形的结论不同
- 不利于社区的理解与进一步推广

2.1 宽深联合 μP 的前期研究——Depth-μP

- Depth-μP 考虑单层残差块： $F_l(h_l) = W_l\phi(h_l)$，优化器为 SGD/Adam

- Depth-μP 结论：在实现宽度 μP 的前提下，$\alpha_l = \Theta\left(\frac{1}{\sqrt{L}}\right)$，学习率对应调整：

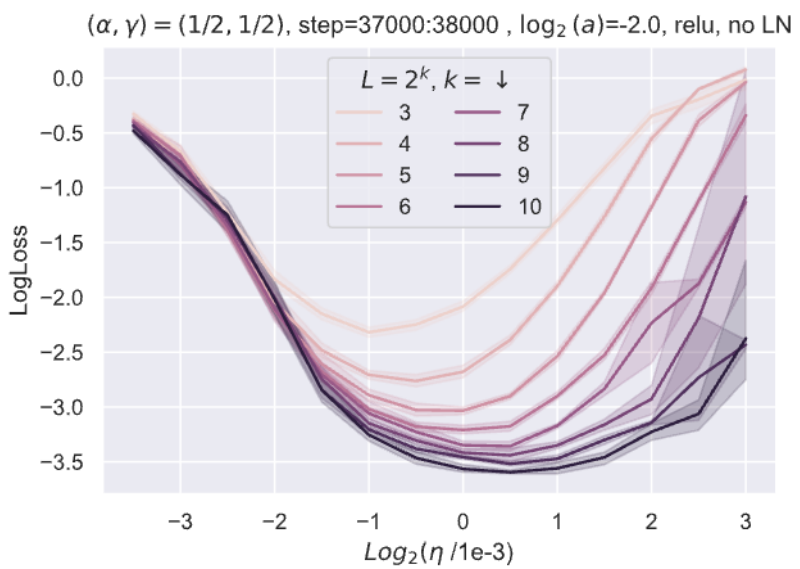
  能够实现深度 μP 原则，且最大化特征多样性

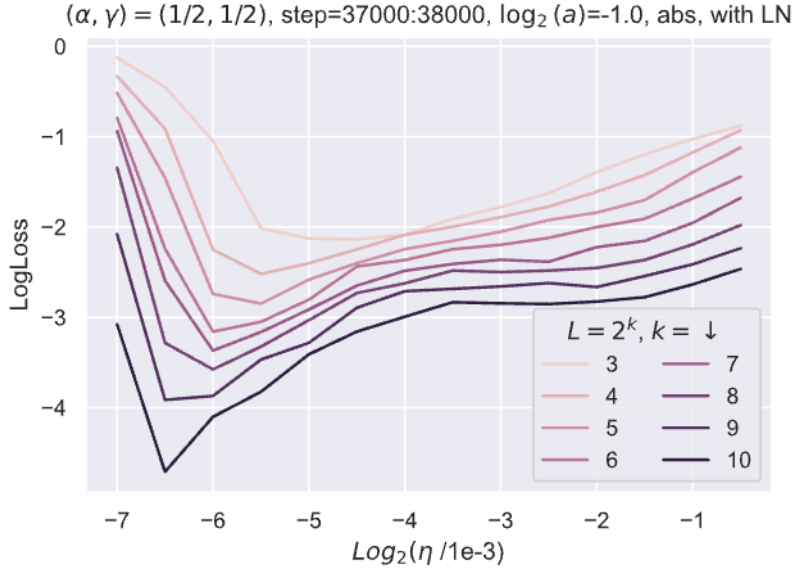
	Branch Multiplier	Learning Rate
Standard	1	? (tuned)
Depth-μP (SGD)	$a/\sqrt{\text{depth}}$	η
Depth-μP (Adam)	$a/\sqrt{\text{depth}}$	$\eta/\sqrt{\text{depth}}$

Tensor Programs VI: Feature Learning in Infinite-Depth Neural Networks, ICLR, 2024

2.1 宽深联合 μP 的前期研究——Depth-μP

- Depth-μP 在实际网络（多层残差块）训练中难以实现稳定深度超参迁移





Tensor Programs VI: Feature Learning in Infinite-Depth Neural Networks, ICLR, 2024

2.1 宽深联合 μP 的前期研究——CompleteP

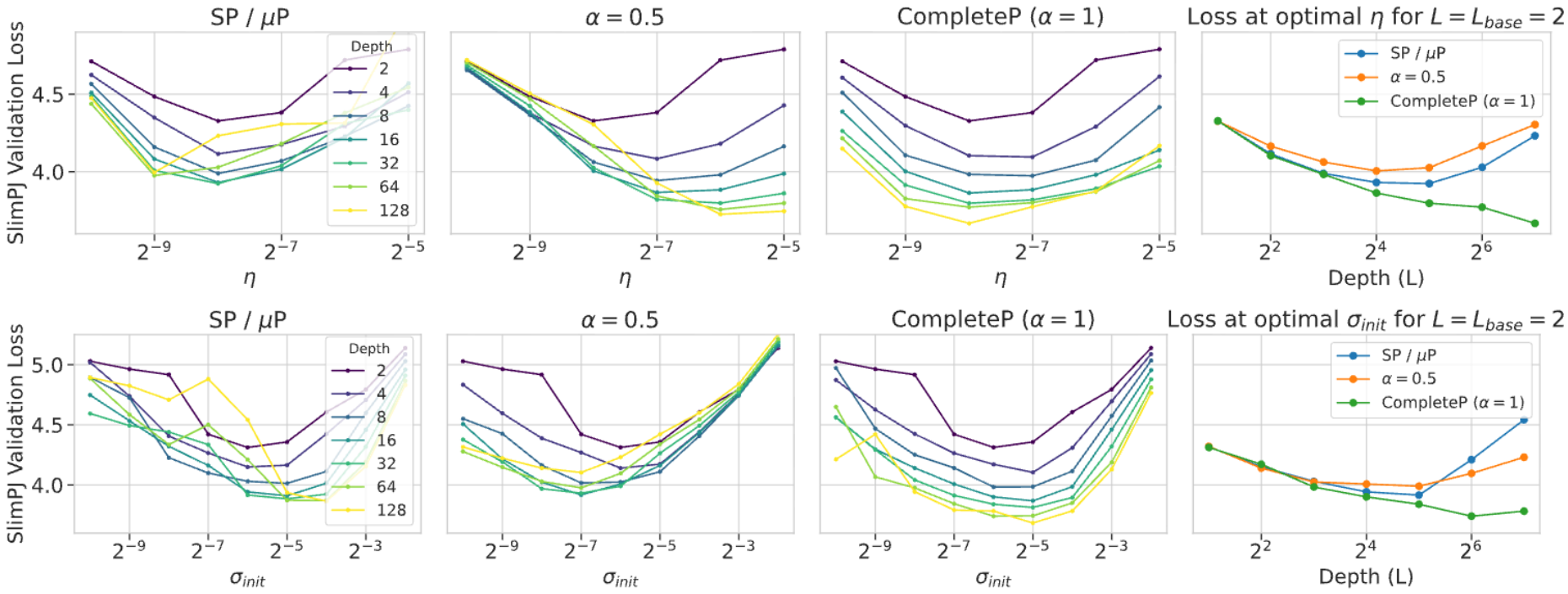
- CompleteP 考虑两层残差块： $F_l(h_l) = W_l^{(2)} W_l^{(1)} h_l$，优化器为 AdamW
- CompleteP 结论：在实现宽度 μP 的前提下，额外引入 $\alpha_l = \Theta(\frac{1}{L})$

	Input weights & biases	Hidden weights	Output weights	Hidden biases
Block Multiplier	α_{base}	α_{base}/r_L (α_{base})	α_{base}/r_n (α_{base})	α_{base}/r_L (α_{base})
Initial Variance	$\sigma^2_{\text{base}}/d_0$ or σ^2_{base}	$\sigma^2_{\text{base}}/r_n$ (σ^2_{base})	σ^2_{base}	σ^2_{base}
Learning Rate	η_{base}	η_{base}/r_n (η_{base})	η_{base}	η_{base}
Weight Decay	λ_{base}	$\lambda_{\text{base}} r_n$ (λ_{base})	λ_{base}	λ_{base}
AdamW ε	$\varepsilon_{\text{base}}/r_n$ $(\varepsilon_{\text{base}})$	$\varepsilon_{\text{base}}/(r_L r_n)$ $(\varepsilon_{\text{base}})$	$\varepsilon_{\text{base}}/r_n$ $(\varepsilon_{\text{base}})$	$\varepsilon_{\text{base}}/(r_L r_n)$ $(\varepsilon_{\text{base}})$

Don't be lazy: CompleteP enables compute-efficient deep transformers,NeurIPS, 2025

2.1 宽深联合 μP 的前期研究——CompleteP

- CompleteP 在实际 GPT-2 的训练中能够实现稳定的深度超参迁移



Don't be lazy: CompleteP enables compute-efficient deep transformers,NeurIPS, 2025

2.2 宽深联合 μP 的谱条件理论——理论原型

- 相较于宽度扩展场景的 μP 谱条件理论，理论原型有如下改进
  - 原型改进 1：残差连接为深度扩展的必备结构，因此引入残差连接
  - 原型改进 2：残差块通常有两层以上，因此考虑任意有限 k 层残差块

最简原型结构：残差块为两层的 deep linear MLP

$$h_0 = \alpha_0 W_0 x$$

$$h_l = h_{l-1} + \alpha_l \prod W_l^{(i)} h_{l-1}, \forall\, l \in [L]$$

$$h_{L+1} = \alpha_{L+1} W_{L+1} h_L$$

后续我们将表明，两层残差块（$k = 2$）是能够反映真实情况的最简模型。

2.2 宽深联合 μP 的谱条件理论—单层残差块

- 单层残差块时的理论结果：

初始化条件：
- 输入和输出层：$\alpha_0\|W_0\|_{RMS}, \alpha_{L+1}\|W_{L+1}\|_{RMS} = \Theta(1)$
- 隐藏层：$\alpha_l\|W_l\|_{RMS} = O\left(\frac{1}{\sqrt{L}}\right), \forall\, l \in [L]$

更新条件：
- 输入和输出层：$\alpha_0\|\Delta W_0\|_{RMS}, \alpha_{L+1}\|\Delta W_{L+1}\|_{RMS} = \Theta(1)$
- 隐藏层一阶条件：$\alpha_l\|\Delta W_l\|_{RMS} = \Theta\left(\frac{1}{L}\right), \forall\, l \in [L]$

- 基于宽度 μP 的实现有 $\|W_l\|_{RMS} = \Theta(1)$，所以得到 $\alpha_l = \mathrm{O}(1/\sqrt{L})$。

2.2 宽深联合 μP 的谱条件理论—单层残差块

- $\alpha_l = \mathrm{O}(1/\sqrt{L})$，调整学习率能够让一阶更新量 ϵ_1 满足 μP 条件为 $\Theta(1)$
- $\alpha_l = \Theta(1/\sqrt{L})$ 能够额外让零阶更新量 ϵ_0 也最大化为 $\Theta(1)$
  - 与原论文最大化特征多样性对应
- Depth-μP 的其它超参（如学习率）参数化都能被推导出来

$$\Delta \boldsymbol{h}_s(\boldsymbol{x}) = \Delta \boldsymbol{h}_0(\boldsymbol{x}) + \underbrace{\sum_{l=1}^{s} \alpha_l \boldsymbol{W}_l \Delta \boldsymbol{h}_{l-1}(\boldsymbol{x})}_{\boldsymbol{\epsilon}_0(s)} + \underbrace{\sum_{l=1}^{s} \alpha_l \Delta \boldsymbol{W}_l (\boldsymbol{h}_{l-1}(\boldsymbol{x}) + \Delta \boldsymbol{h}_{l-1}(\boldsymbol{x}))}_{\boldsymbol{\epsilon}_1(s)}.$$

2.2 宽深联合 μP 的谱条件理论—双层残差块

- 双层残差块时的理论结果：

初始化条件：

- 输入和输出层：$\alpha_0\|W_0\|_{RMS}, \alpha_{L+1}\|W_{L+1}\|_{RMS} = \Theta(1)$
- 隐藏层：$\alpha_l \left\|W_l^{(2)}\right\|_{RMS} \left\|W_l^{(1)}\right\|_{RMS} = \Theta\left(\frac{1}{L}\right), \forall\, l \in [L]$ (二阶更新条件造成)

更新条件：

- 输入和输出层：$\alpha_0\|\Delta W_0\|_{RMS}, \alpha_{L+1}\|\Delta W_{L+1}\|_{RMS} = \Theta(1)$
- 隐藏层一阶条件 1：$\alpha_l \left\|\Delta W_l^{(2)}\right\|_{RMS} \left\|W_l^{(1)}\right\|_{RMS} = \Theta\left(\frac{1}{L}\right), \forall\, l \in [L]$
- 隐藏层一阶条件 2：$\alpha_l \left\|W_l^{(2)}\right\|_{RMS} \left\|\Delta W_l^{(1)}\right\|_{RMS} = \Theta\left(\frac{1}{L}\right), \forall\, l \in [L]$
- 隐藏层二阶条件：$\alpha_l \left\|\Delta W_l^{(2)}\right\|_{RMS} \left\|\Delta W_l^{(1)}\right\|_{RMS} = \Theta\left(\frac{1}{L}\right), \forall\, l \in [L]$

- 基于宽度 μP 的实现有 $\|W_l\|_{RMS} = \Theta(1)$，得到 $\alpha_l = O(1/L)$，与 CompleteP 一致

Spectral Condition for μP under Width–Depth Scaling, ICLR Delta Workshop (Oral), 2026

2.2 宽深联合 μP 的谱条件理论—双层残差块

- 双层残差块相较于单层的本质区别：高阶更新量的出现
  - 令 $\epsilon_2 = \Theta(1)$ 得到二阶更新条件，并反馈得到更紧的初始化条件
  - 更深的残差块不会影响最终结果，双层残差块为最简理论原型

$$\Delta \boldsymbol{h}_s(\boldsymbol{x}) = \Delta \boldsymbol{h}_0(\boldsymbol{x}) + \underbrace{\sum_{l=1}^{s} \alpha_l \boldsymbol{W}_l^{(2)} \boldsymbol{W}_l^{(1)} \Delta \boldsymbol{h}_{l-1}(\boldsymbol{x})}_{\boldsymbol{\epsilon}_0(s)} + \underbrace{\sum_{l=1}^{s} \alpha_l \boldsymbol{W}_l^{(2)} \Delta \boldsymbol{W}_l^{(1)} (\boldsymbol{h}_{l-1}(\boldsymbol{x}) + \Delta \boldsymbol{h}_{l-1}(\boldsymbol{x}))}_{\boldsymbol{\epsilon}_1^{(1)}(s)}$$

$$+ \underbrace{\sum_{l=1}^{s} \alpha_l \Delta \boldsymbol{W}_l^{(2)} \boldsymbol{W}_l^{(1)} (\boldsymbol{h}_{l-1}(\boldsymbol{x}) + \Delta \boldsymbol{h}_{l-1}(\boldsymbol{x}))}_{\boldsymbol{\epsilon}_1^{(2)}(s)} + \underbrace{\sum_{l=1}^{s} \alpha_l \Delta \boldsymbol{W}_l^{(2)} \Delta \boldsymbol{W}_l^{(1)} (\boldsymbol{h}_{l-1}(\boldsymbol{x}) + \Delta \boldsymbol{h}_{l-1}(\boldsymbol{x}))}_{\boldsymbol{\epsilon}_2(s)}.$$

Spectral Condition for µP under Width–Depth Scaling, ICLR Delta Workshop (Oral), 2026

2.3 宽深联合 μP 的算法——实现

- 对于主流优化器（Muon-Kimi、Muon、Shampoo、SOAP、AdamW、Sophia、Lion、SSO 等），基于宽度 μP 实现，额外加入 $\alpha_l = \Theta(1/L)$ 即可

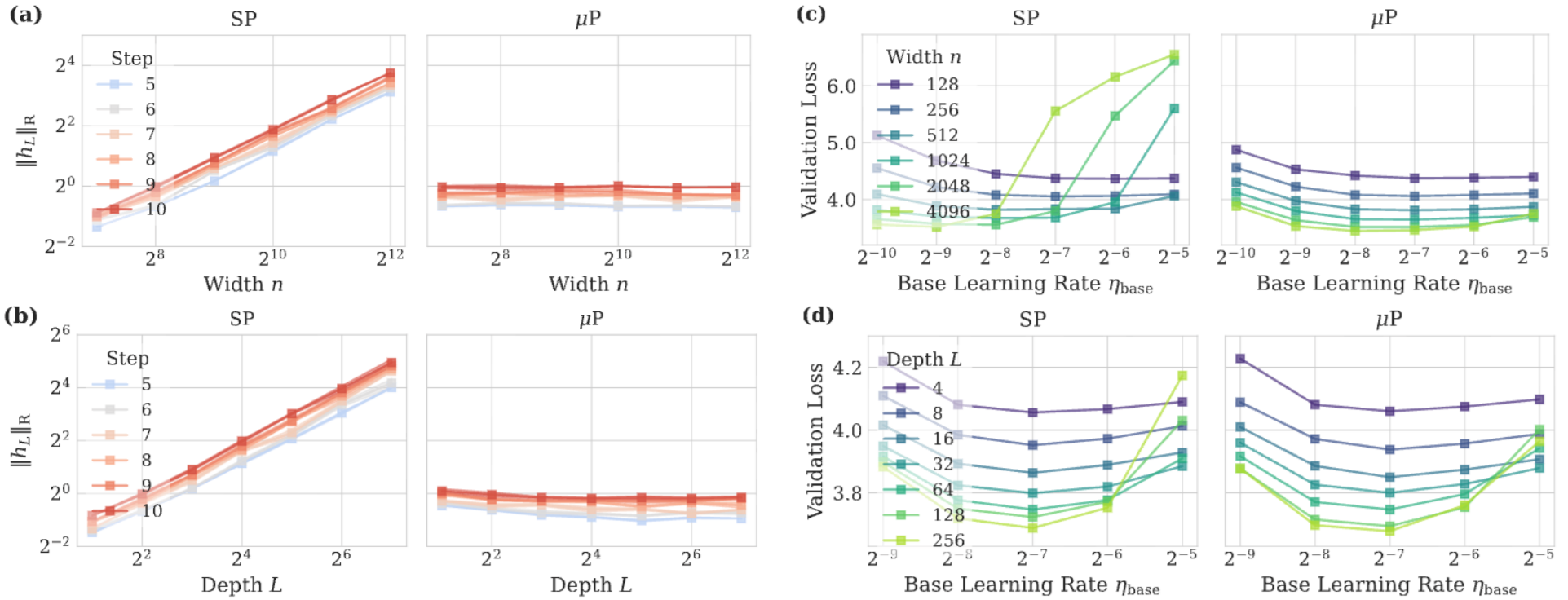
Table 2: **μP implementation for Muon-Kimi (Liu et al., 2025) with weight decay under width-depth scaling.** Entries in purple indicate differences between μP and SP, while gray shows the corresponding SP choices. Here, r_n and r_L denote the width and depth scaling ratios relative to the base model. The variance of input weights is σ^2_{base} for language and $\sigma^2_{\text{base}}/d_0$ for image.

	Input weights	Hidden weights	Output weights
Block Multiplier	α_{base}	α_{base}/r_L (α_{base})	α_{base}/r_n (α_{base})
Initial Variance	$\sigma^2_{\text{base}}/d_0$ or σ^2_{base}	$\sigma^2_{\text{base}}/r_n$ (σ^2_{base})	σ^2_{base}
Learning Rate	η_{base}	$\eta_{\text{base}}/\sqrt{r_n}$ (η_{base})	η_{base}
Weight Decay	λ_{base}	$\lambda_{\text{base}}\sqrt{r_n}$ (λ_{base})	λ_{base}

Spectral Condition for µP under Width–Depth Scaling, ICLR Delta Workshop (Oral), 2026

2.3 宽深联合 μP 的算法——Muon-Kimi 效果

- 能够实现宽深扩展时的训练稳定和超参迁移



内容纲要

- μP 的背景与影响
- 宽度扩展的 μP 理论与实践
- 宽深联合扩展的 μP 理论与实践
- 总结与展望

总结与展望

- 实践上，典型训练场景的超参宽深迁移基本被解决，大规模落地进行中
- 理论上， μP 参数化仍有很多值得探究的地方
  - μP 参数化的特有现象：超参迁移性质的充要条件仍然是个谜
  - 能否建立更精细的优化和泛化保证？