NEURAL INFORMATION PROCESSING SYSTEMS

Learning Rate Transfer for Hybrid Transformer-SSM Architectures

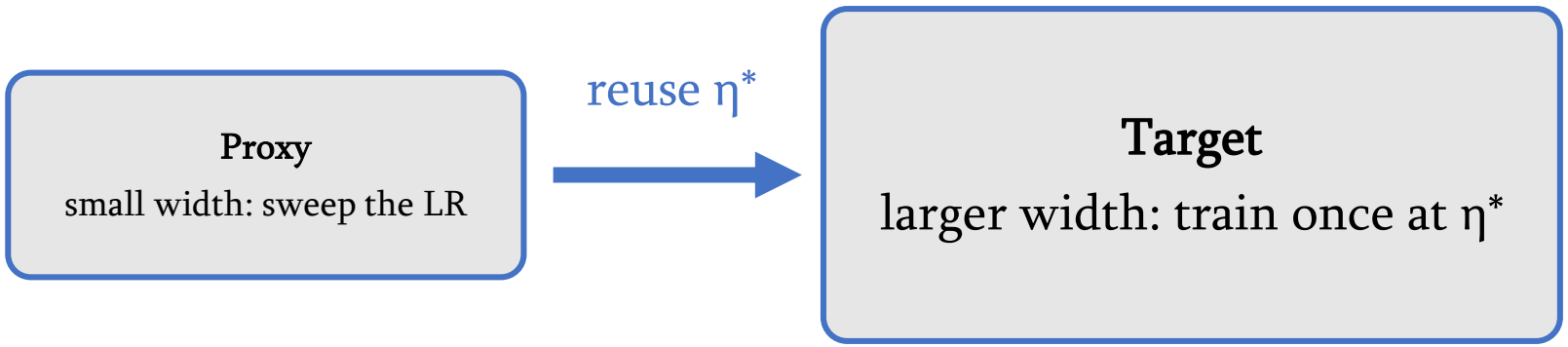
Jimin Seo[1] Gyubok Lee[2] Yeonsik Jo[3] Kiwoong Yoo[3] Yeongoon Kim[1] Minhae Oh[1] Jin Woo Koo[1] Suhwan Kim[1] Nakyung Lee[1]
Minsik Seol[1] Idris Nechnech[1] Jaehyeon Kim[4] Giho Lee[4] Jungwoo Lee[1,4]

[1]Seoul National University [2]SB Intuitions [3]LG AI Research [4]Hodoo AI

Seoul National University Cognitive Machine Learning Lab. SB Intuitions LG AI Research HodooAi

The Question

Can a learning rate tuned on a small proxy be reused on a larger hybrid?



Naïve µP with AdamW gives a near-zero LR transfer gap across widths 256–2048 (8×) and depths 4–32 (8×), up to billion-parameter scale.

Transfer gap: loss penalty of training the target at the proxy's best LR instead of its own.

Existing LR Rules

LR rules exist for Transformers and for pure SSMs, not for hybrids at practical widths.

Hybrid models pair attention with state-space model (SSM) blocks (Nemotron-H, Jamba, Falcon-H1), and their learning rate (LR) is retuned at each scale.

T	**Transformer**	**μP**: optimal LR width-independent (Yang et al., 2021)
M	**Pure SSM**	**μP-SSM**: component-specific rules (Vankadara et al., 2024)
T M	**Hybrid**	**?** To our knowledge, no prior work at practical finite widths

"No prior work" is to our knowledge.

Yang et al., 2021. Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer. NeurIPS 2021.
Vankadara et al., 2024. On Feature Learning in Structured State Space Models. NeurIPS 2024.

Comparison of Regimes

We compare 4 LR scalings (SP, naïve μP, block-wise μP and its approximated form) on 7 architectures, built from `T` Transformer, `M` Mamba and `H` hybrid blocks.

	μP derivation	Practical hybrids (this work)	μP-SSM derivation
Model	Transformer	Transformer + SSM	pure SSM
LR rule	μP	**μP** (on every block)	component-specific SSM rules
Discretization	n/a	simplified-ZOH Mamba	true zero-order hold (ZOH)
State size N_x	n/a	fixed, N_x=16	grows with width
Width	infinite-width limit	256–2048	infinite-width limit
Optimizer	SGD, Adam	AdamW	SGD

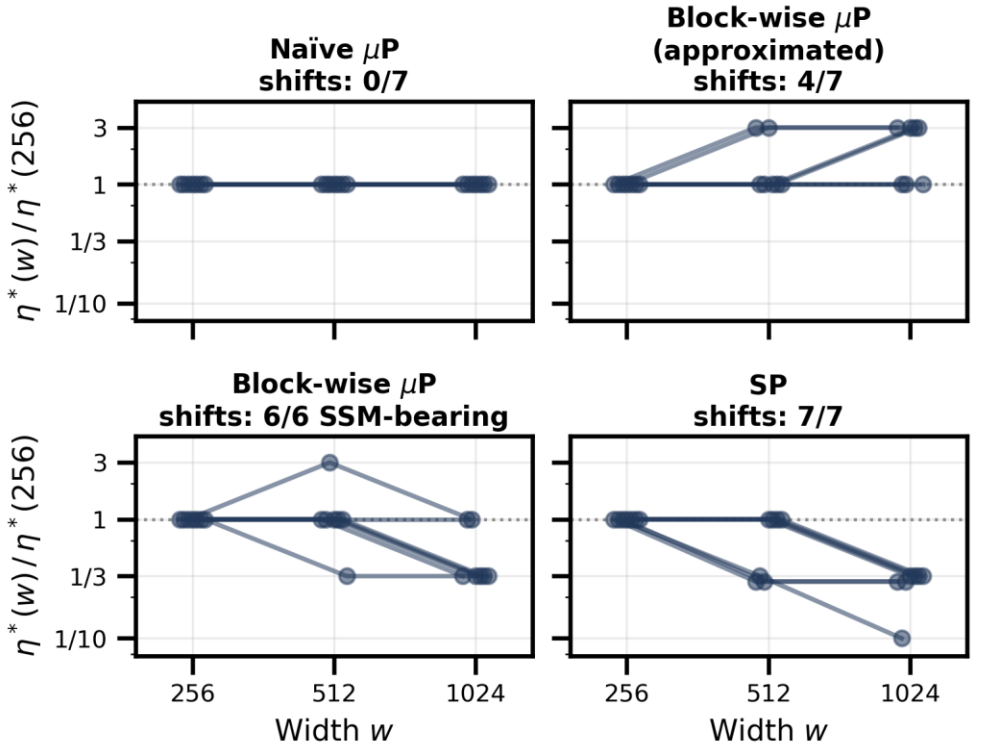
μP column: Yang et al., 2021. Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer. NeurIPS 2021.
Adam: Yang and Littwin, 2023. Tensor Programs IVb: Adaptive Optimization in the Infinite-Width Limit. arXiv:2308.01814.
μP-SSM column: Vankadara et al., 2024. On Feature Learning in Structured State Space Models. NeurIPS 2024.

Optimal LR per Scheme

Under naïve µP, the best LR stays at 1.0 at every width, for all 7 architectures

	Naïve µP			Block-wise µP (approximated)			Block-wise µP			SP		
w	256	512	1024	256	512	1024	256	512	1024	256	512	1024
MM	1.0	1.0	1.0	1.0	3.0	3.0	3.0	3.0	1.0	0.010	0.003	0.003
TT	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.003	0.001	0.0003
MTMT	1.0	1.0	1.0	1.0	1.0	1.0	1.0	3.0	1.0	0.010	0.003	0.003
MHMH	1.0	1.0	1.0	1.0	3.0	3.0	3.0	3.0	1.0	0.003	0.003	0.001
HMHM	1.0	1.0	1.0	1.0	1.0	3.0	3.0	3.0	1.0	0.003	0.003	0.001
HHHH	1.0	1.0	1.0	1.0	1.0	3.0	3.0	3.0	1.0	0.003	0.003	0.001
HTHT	1.0	1.0	1.0	1.0	1.0	1.0	3.0	1.0	1.0	0.003	0.003	0.001
Shifts		0/7			4/7			6/6 SSM-bearing			7/7	

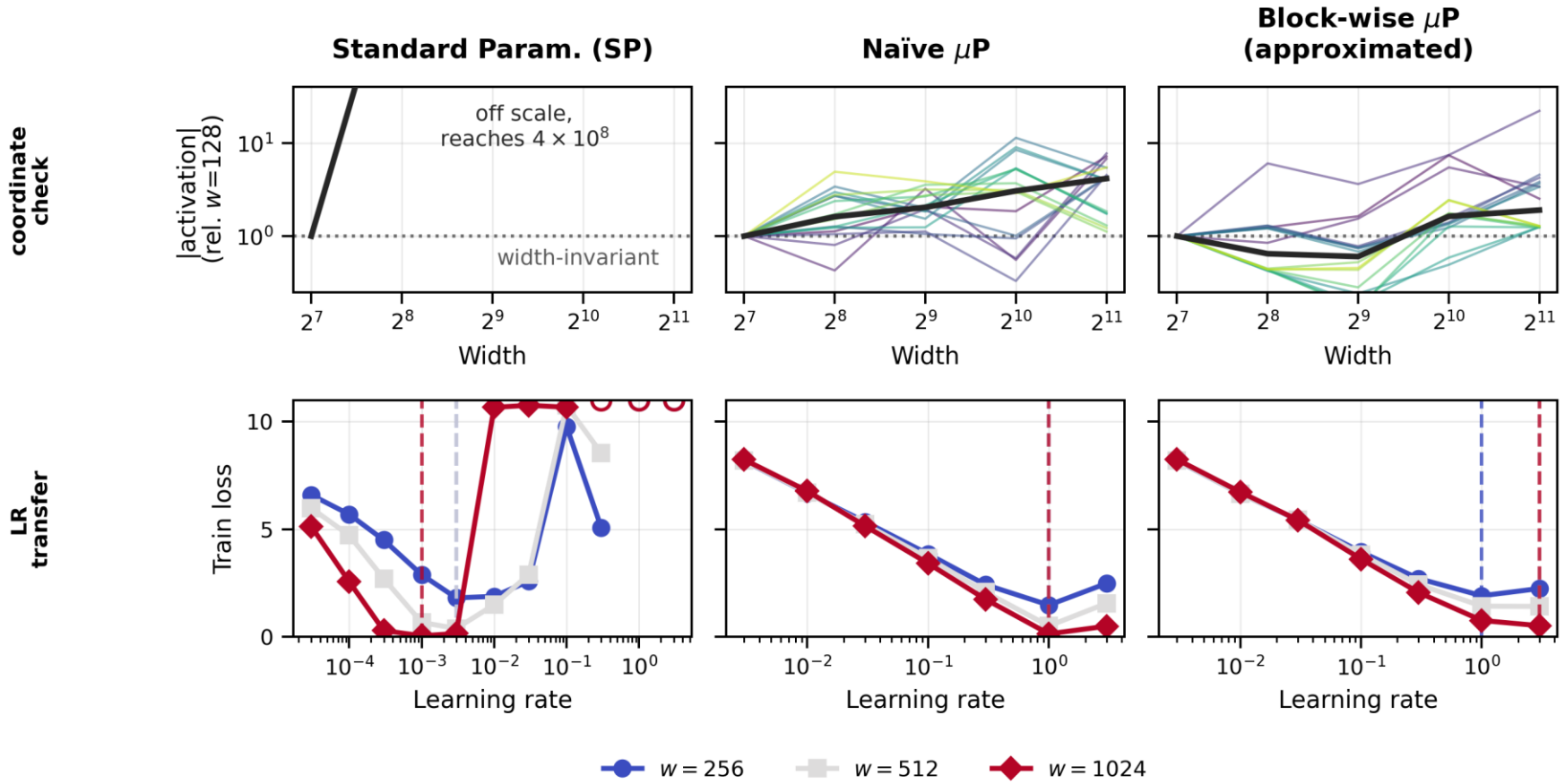
Paper Table 1.



Best LR per width (2k steps, WikiText-2);
single shading across a row means the optimum transfers.

Each architecture's best LR divided by its best LR at $w = 256$; a line flat at 1 is a transfer.

Coordinate Check vs. LR Transfer



MHMH, d = 16. Top: activation size against width (coordinate check). Bottom: training loss against LR at three widths, each width's optimum dashed. Paper Figure 1, §5.1.

The coordinate check does not predict LR transfer on hybrid models

Two Conditions for Transfer

Used as explanatory diagnostics of where transfer holds and where it breaks.

Global (Condition 1)

$$\rho(W) = \frac{\|\Delta W\|_F}{\|W\|_F} = \mathcal{O}(1) \quad \text{as } d_{\text{model}} \to \infty$$

Each step changes every weight matrix by the same fracti on of itself, at every width.

Enforced by μP's initialization and LR scaling.

Breaks when ρ grows with width, as under SP, where ρ = O(η√*d*).

Basis: μP theory (Yang et al., 2022; 2024), assumed under AdamW.

Local (Condition 2)

$$\beta_{X/Y} = \frac{\|\Delta W_X\|_F}{\|\Delta W_Y\|_F}, \quad \Delta\beta_{X/Y} \approx 0$$

The ratio of update sizes between two parts of the model, such as $A_{\log}$ and FFN, stays the same from proxy to target.

Provided, in our regime, by AdamW's per-parameter nor malization.

Breaks when β moves with width, as under SGD, where β follows the ratio of gradient sizes.

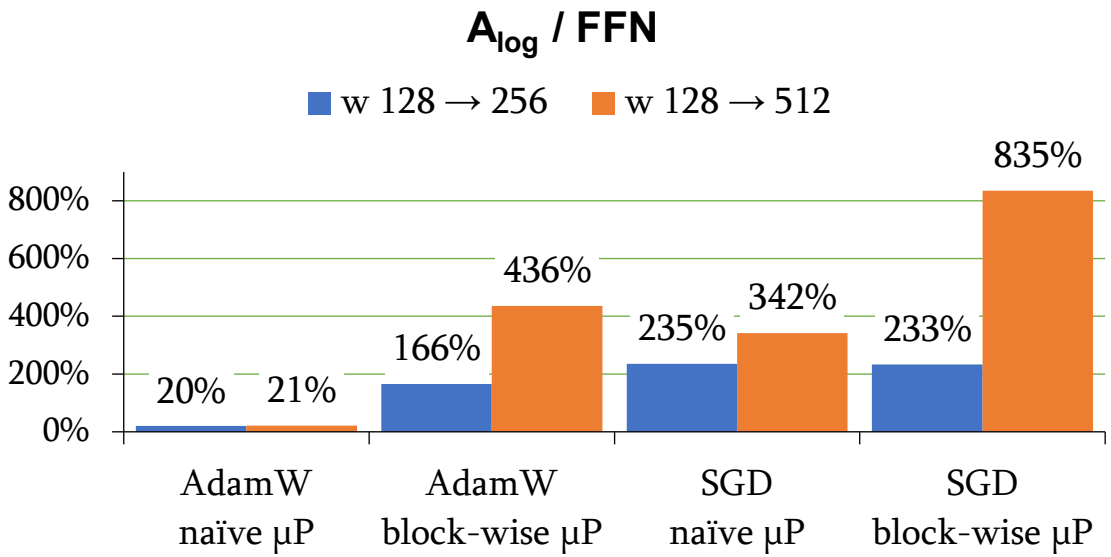
Basis: Δβ, measured between proxy and target widths on MHMH.

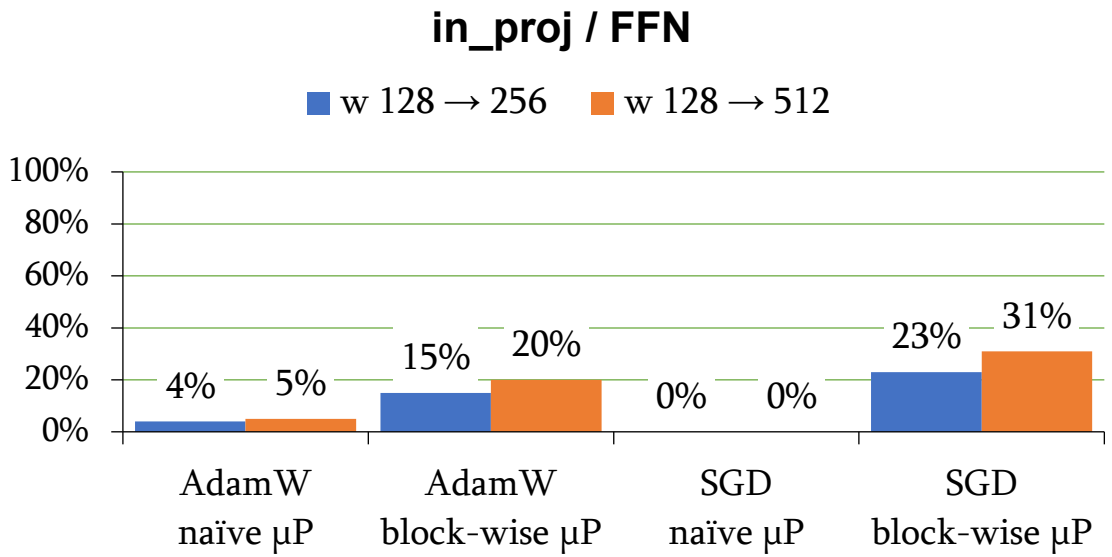
Naïve μP with AdamW meets both, consistent with its best LR staying at 1.0.

Paper §5.2–5.3.

Two Conditions for Transfer: Update Balance by Optimizer

Under AdamW with naïve μP, the update balance stays within run-to-run noise.





Δβ: shift of the update ratio $\|\Delta W_X\| / \|\Delta W_{FFN}\|$ from proxy (w 128) to target.

Same-seed reruns of the AdamW rows differ by about **20%** (the floor).

A_{log} / FFN: the pair most sensitive to width under μP-SSM 's $\eta_A \propto N_u$ rule. Only AdamW with naïve μP stays at the fl oor.

in_proj / FFN: a near-isotropic Mamba projection, for com parison. Every row stays at 0–31% (note its 0–100% axis).

Paper Table 4. MHMH, d = 16, N_x = 16; mean over steps 50–200 and 3 seeds.

How Far It Holds

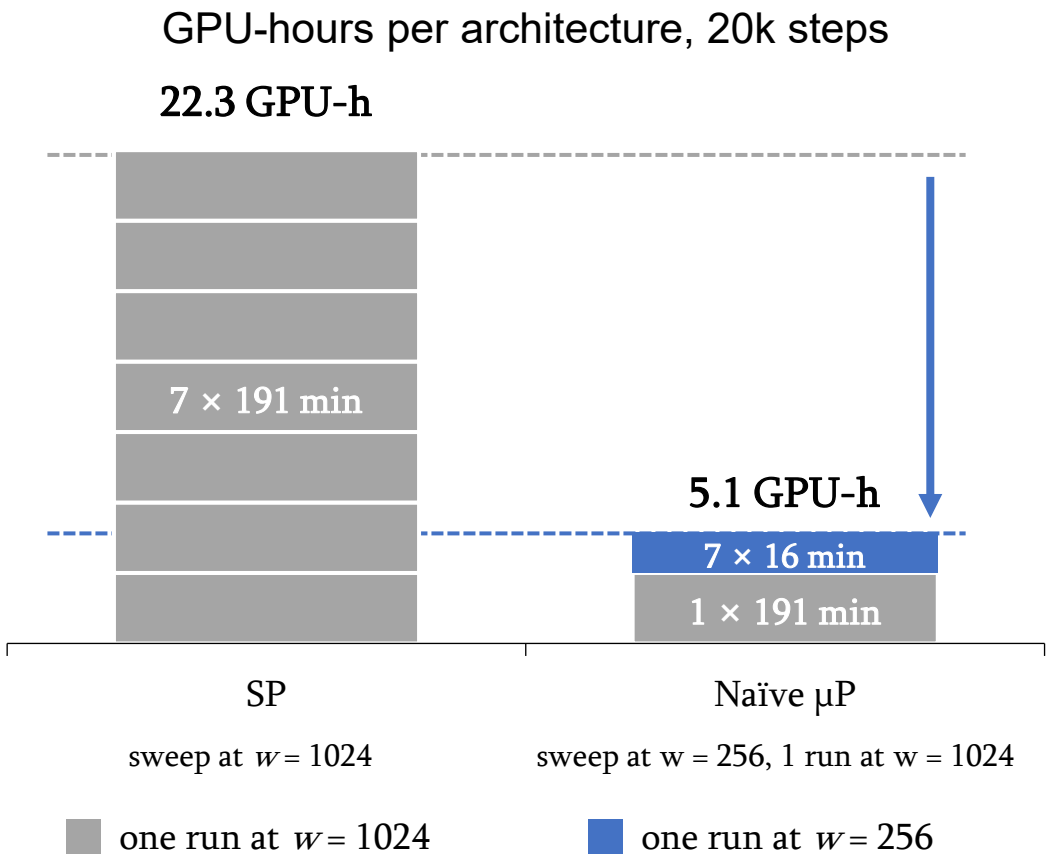
LR transfer holds along every axis we checked.

Axis	Tested	Result (naïve μP)
Width	256–2048 (8×), 7 architectures	η^*=1.0 at every width; 0.0% gap
Depth	d=8→32 with width (~58M→~624M); d=48–128 at w=1024 (up to 1.88B)	0.0% gap on MHMH and HHHH; η=1.0 within 0.043 nats of the best grid point (2k steps, one seed)
Production hybrid	Nemotron-H (Mamba-2, squared-ReLU FFN, GQA without RoPE)	η^*=1.0 at w=256–1024, 0.0% gap (SP 1.9%)
Scale, seeds, horizon	w=2048 (1.24B) 3 seeds at 20k and 60k steps; 24.9B tokens	η=1.0 is the validation-loss argmin at every seed; no divergence over 24.9B tokens (one LR, one seed)
Data	FineWeb-Edu, 3 architectures, 20k steps	0.0% gap; SP's optimum shifts 3×
Training setup	batch size, sequence length, Transformer-to-SSM ratio	the optimal LR stays width-invariant

Paper §4, Appendix C–D. Scale, horizon and seeds are each varied on their own.

Reduction in Tuning Cost

Tuning at the proxy costs about 77% less at 20k steps, in our cost model.



One run on one A6000 (HHHH, $d = 16$):

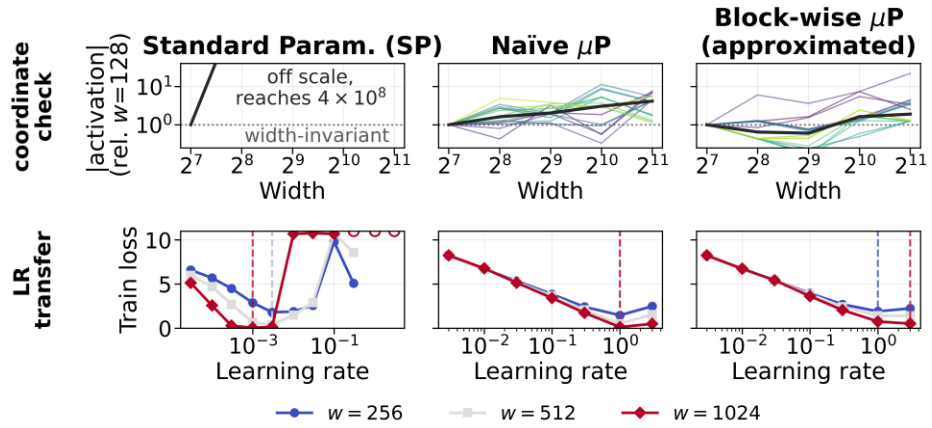
16 min at $w = 256$ and 191 min at $w = 1024$ for 20k steps; 3.1 and 17.1 min for 2k steps.

	2k steps	20k steps
SP: 7 runs at $w = 1024$	2.0 GPU-h	22.3 GPU-h
Naïve µP: 1 run at $w = 1024$, 7 runs at $w = 256$	0.65 GPU-h	5.1 GPU-h
Saving	**68%**	**77%**

Paper Appendix A: a cost model from per-configuration wall time, not an end-to-end measurement.

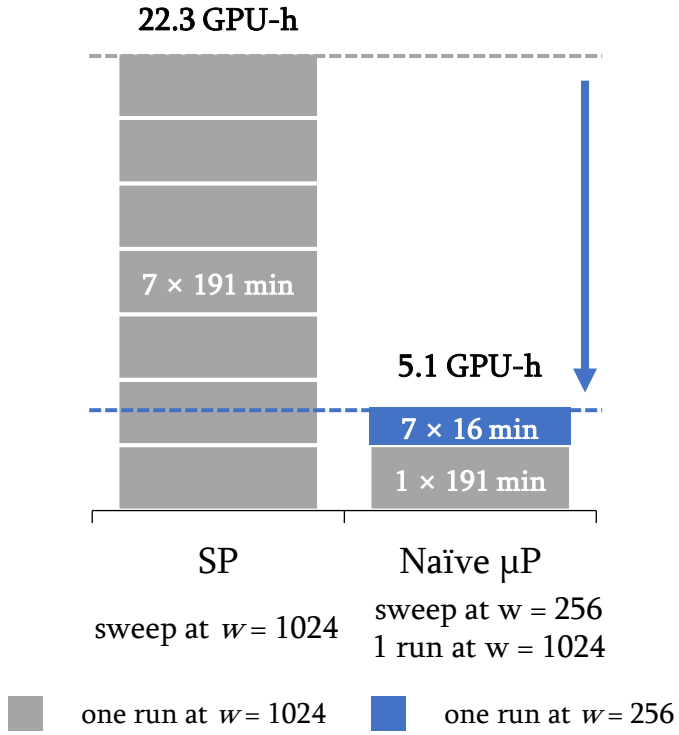
The Conclusion

Naïve µP with AdamW gives a near-zero LR transfer gap across widths 256–2048 (8×) and depths 4–32 (8×), up to billion-parameter scale.



	Naïve µP			Block-wise µP (approximated)			Block-wise µP			SP		
w	256	512	1024	256	512	1024	256	512	1024	256	512	1024
MM	**1.0**	**1.0**	**1.0**	1.0	3.0	3.0	3.0	3.0	1.0	0.010	0.003	0.003
TT	**1.0**	**1.0**	**1.0**	1.0	1.0	1.0	1.0	1.0	1.0	0.003	0.001	0.0003
MTMT	**1.0**	**1.0**	**1.0**	1.0	1.0	1.0	1.0	3.0	1.0	0.010	0.003	0.003
MHMH	**1.0**	**1.0**	**1.0**	1.0	3.0	3.0	3.0	3.0	1.0	0.003	0.003	0.001
HMHM	**1.0**	**1.0**	**1.0**	1.0	1.0	3.0	3.0	3.0	1.0	0.003	0.003	0.001
HHHH	**1.0**	**1.0**	**1.0**	1.0	1.0	3.0	3.0	3.0	1.0	0.003	0.003	0.001
HTHT	**1.0**	**1.0**	**1.0**	1.0	1.0	1.0	3.0	1.0	1.0	0.003	0.003	0.001
Shifts		**0/7**			4/7			6/6 SSM-bearing			7/7	

GPU-hours per architecture, 20k steps



More in the Paper

Also in the paper

Downstream (§4.2, Table 2):

trained at the transferred LR without retuning, MHMH-312M reaches a zero-shot averag e of 44.1; Pythia-410M and Mamba-370M retrained on the same 5B FineWeb-Edu tokens reach 42.4 and 41.9.

Nemotron-H at its released configuration (§4.3):
with the MLP ratio, SSM state, Mamba-2 group count and head dimension of Nemotron-H-8B, the optimum stays at $\eta = 1.0$ from $w = 512$ to 2048 (980M parameters).

Limits (§6)

Scale, horizon, and seeds are validated individually rather than jointly.

The infinite-width limit and μP-SSM under true ZOH with growing N_x, the regime it wa s derived for, remain open.



Learning Rate Transfer
for Hybrid Transformer-SSM Architectures

Jimin Seo[1] Gyubok Lee[2,*] Yeonsik Jo[3] Kiwoong Yoo[3] Yeongoon Kim[1]
Minhae Oh[1] Jin Woo Koo[1] Suhwan Kim[1] Nakyung Lee[1] Minsik Seol[1]
Idris Nechnech[1] Jaehyeon Kim[4] Giho Lee[4] Jungwoo Lee[3,4,†]

[1]Seoul National University [2]SB Intuitions [3]LG AI Research [4]Hodoo AI

{jimin.seo, ygoonkim, minhae.oh, jinukoo, curti05, leena, seolminsik}@cml.snu.ac.kr
{leechoch, junglee}@snu.ac.kr, gyubok.lee@sbintuitions.co.jp
{yeonsik.jo, kiwoong.yoo}@lgresearch.ai, {jaehyeon.kim, giho.lee}@hodooai.com

Abstract

We study learning rate (LR) scaling for hybrid architectures combining Transformer and State-Space Model (SSM) blocks, a class adopted by several recent production language models. In particular, we focus on the gap between the theoretical scaling rules derived for SSMs under zero-order-hold (ZOH) discretization at infinite width with growing state size, and the field-standard practical implementations using simplified-ZOH Mamba at fixed state size. Surprisingly, in this practical regime hybrid architectures achieve a **near-zero LR transfer gap** across widths 256–2048 and depths 4–32 up to billion-parameter scale using only the original μP prescription, even though SSM operations fall outside its Tensor Programs representability conditions and every parameterization we test fails the standard coordinate-check diagnostic of μP correctness. We attribute this to a two-condition decomposition of LR transfer in hybrid architectures: a *global* update-to-weight invariance, enforced by μP's initialization and LR scaling; and a *local* per-component balance, provided by AdamW's per-parameter normalization. Our observations show that the optimal LR is invariant to width up to $8\times$, that this width invariance holds across depth, sequence length, batch size, and Transformer-to-SSM ratio, and that it transfers to Nemotron-H, a production hybrid outside our custom architecture set. We hope these findings fill the gap between theoretical scaling rules and practical hybrid implementations, and stimulate further research toward bridging it.

1 Introduction

Hybrid architectures combining Transformer attention with State-Space Model (SSM) blocks have emerged as a promising direction for large language models, pairing attention's quality on long-range dependencies with SSMs' subquadratic inference cost. Controlled comparisons at 8B parameters show Mamba-2-Hybrid outperforming pure Transformers across benchmarks while running up to $8\times$ faster [Waleffe et al., 2024], and similar trends appear in recent production hybrids [NVIDIA, 2025, Jamba Team, 2025, Zuo et al., 2025]; a fundamental practical problem remains, however: *training these models requires expensive per-scale hyperparameter tuning.*

The learning rate (LR) must be retuned at each model scale because the optimum shifts with architectural choices like width and depth; at production scale this sweep can substantially exceed the

*Work done prior to joining SB Intuitions.
†Corresponding author.
‡Code is available at https://github.com/argyroneta-aquatica/hybrid-mup

Preprint.

Paper on arXiv
arxiv.org/abs/2610.01172