Parallax: Parameterized Local Linear Attention for Language Modeling

DLCT

Jul 31, 2026

Yifei Zuo[1], Dhruv Pai[2], Zhichen Zeng[3], Alec Dewulf[2], Shuming Hu[2], Zhaoran Wang[1]

[1]Northwestern University, [2]Tilde Research, [3]University of Washington

TL;DR

- We developed Local Linear Attention (**LLA**), an improvement over Softmax Attention from a test-time regression perspective.
- Parameterized Local Linear Attention (**Parallax**) is a practical variant of LLA for language modeling.

$$o_i^{\mathsf{PLX}} = \sum_{j \le i} \frac{w_{ij}(1 - t_{ij} + \bar{t}_i)}{\sum_{j' \le i} w_{ij'}} v_j = o_i^{\mathsf{SA}} - \Sigma_{KV}^{(i)} \rho_i,$$

$$\rho_i = W_R x_i, \quad t_{ij} = \rho_i^\top (k_j - q_i)$$

TL;DR

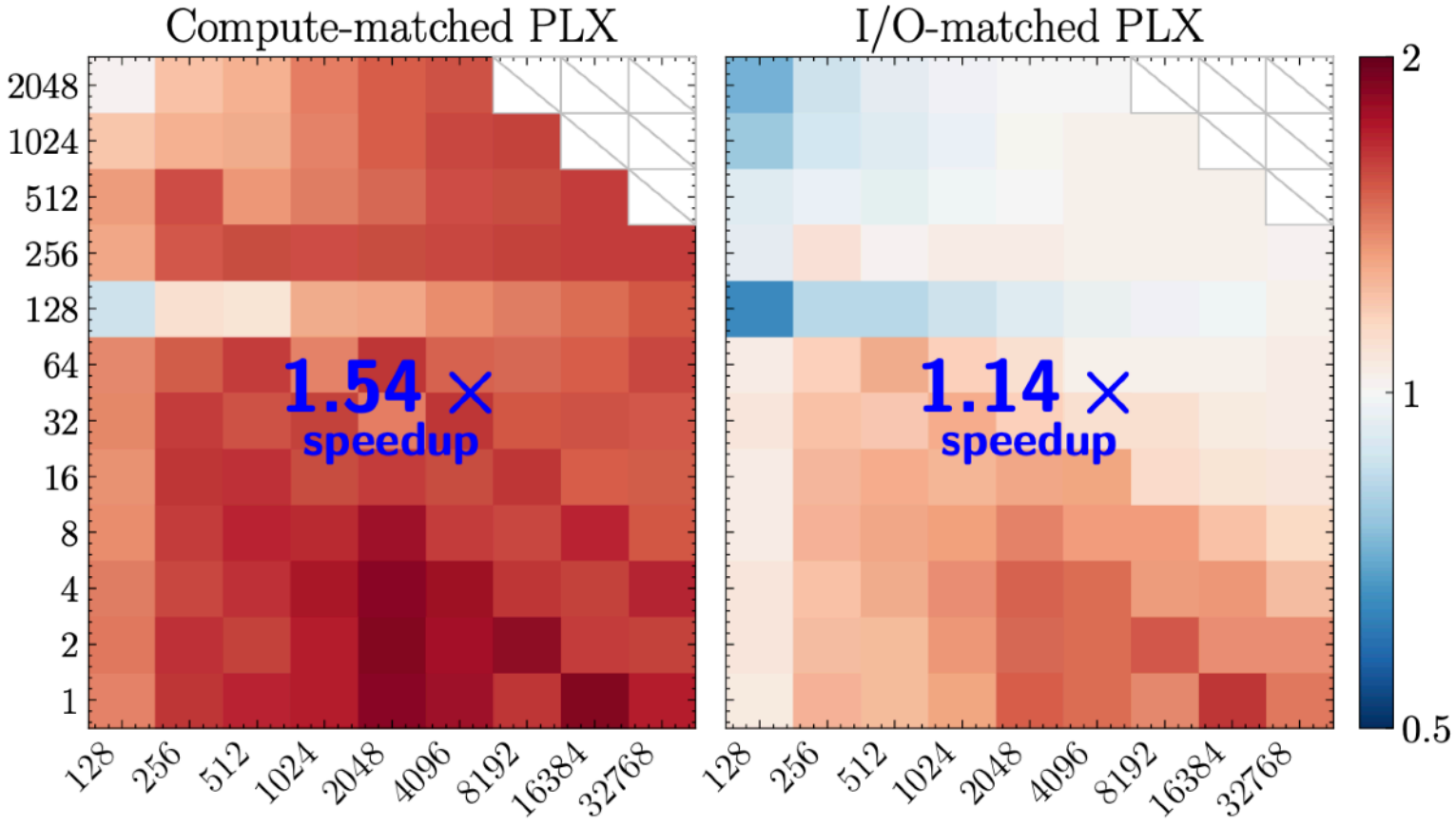
- We developed Local Linear Attention (**LLA**), an improvement over Softmax Attention from a test-time regression perspective.
- Parameterized Local Linear Attention (**Parallax**) is a practical variant of LLA for language modeling.

$$o^{\mathsf{LLA}} = o_i^{\mathsf{SA}} - (1+\eta_i)\Sigma_{KV}^{(i)}\rho_i^{\star}$$

$$o^{\mathsf{PLX}} = o_i^{\mathsf{SA}} - \Sigma_{KV}^{(i)}\rho_i$$

Efficiency

- Parallax increases arithmetic intensity and is hardware-efficient



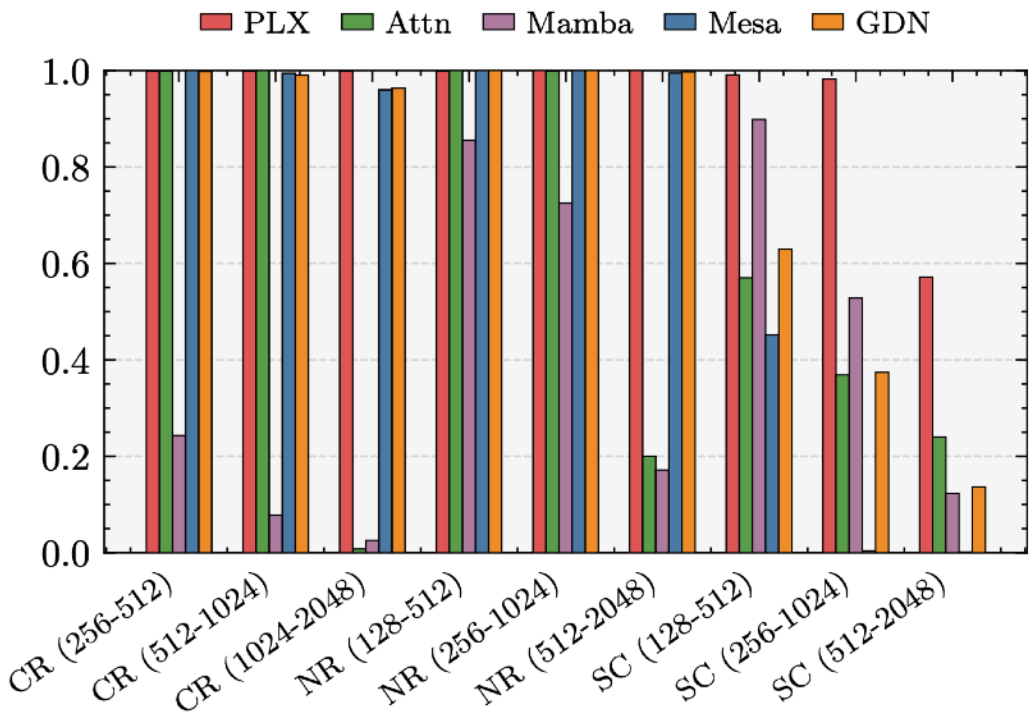
PLX-CuTeDSL vs best(FA2,FA3) in decoding.

Performance

- Parallax performs better on the MAD synthetic benchmarks.

Task	Attn	PLX	GDN	Mamba	Mesa
CMP	0.342	0.332	0.325	**0.424**	0.310
ICR	0.803	0.951	0.920	0.756	**0.998**
FCR	0.268	**0.356**	0.110	0.065	0.219
NCR	0.861	0.937	0.907	0.713	**0.999**
MEM	0.807	0.733	0.792	**0.876**	0.861
SC	0.950	**0.988**	0.939	0.950	0.568
Avg	0.672	**0.716**	0.665	0.631	0.659

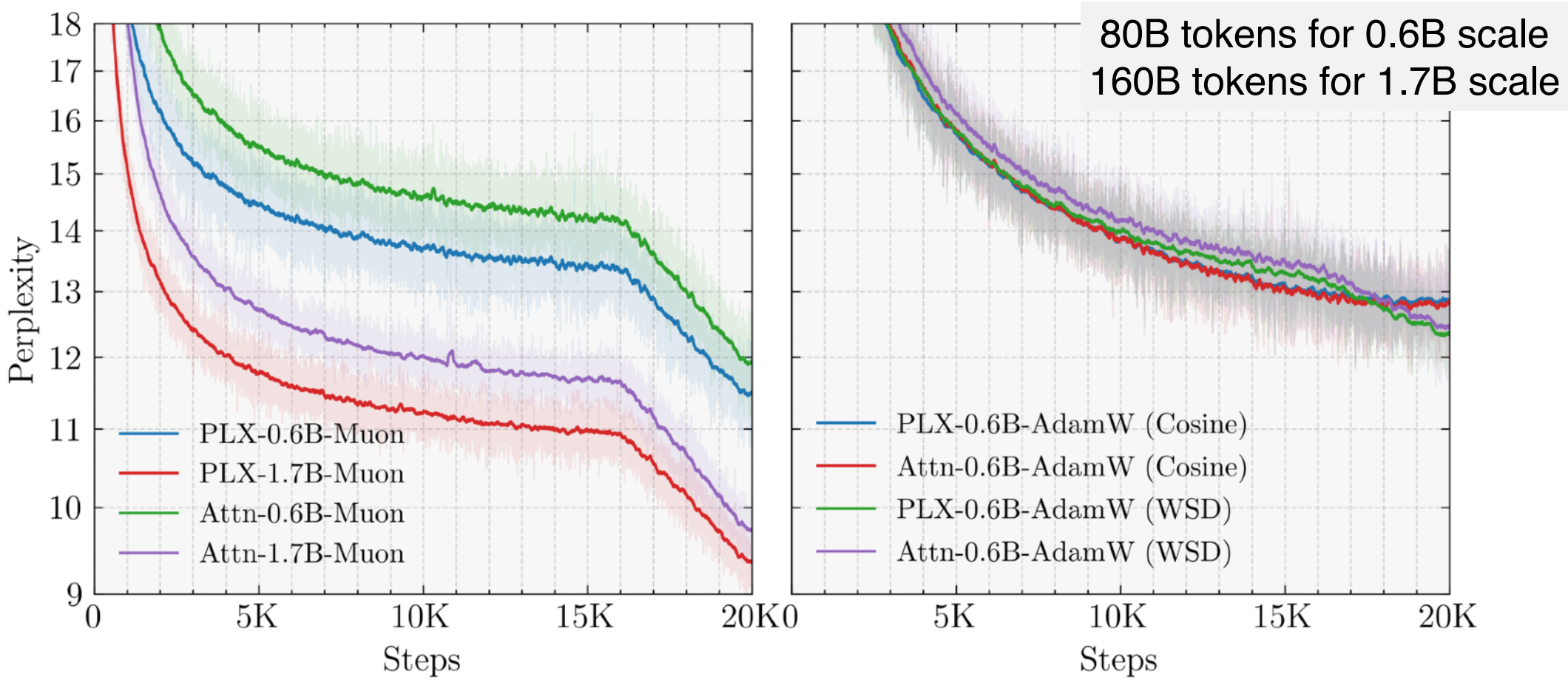
(a) MAD benchmark accuracy.



(b) MAD-challenge recall accuracy.

Performance

- Parallax outperforms SA during pretraining at the 0.6B and 1.7B scales, but its performance is optimizer-dependent.



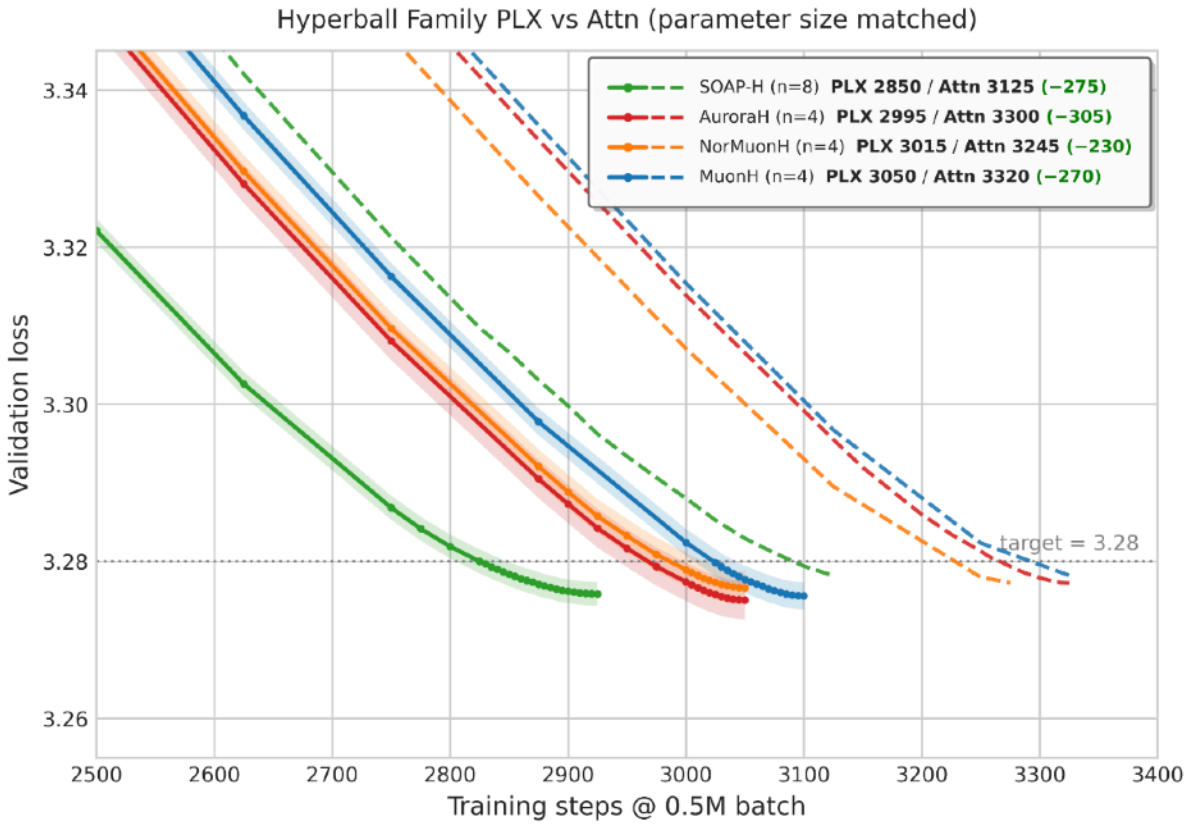
Performance

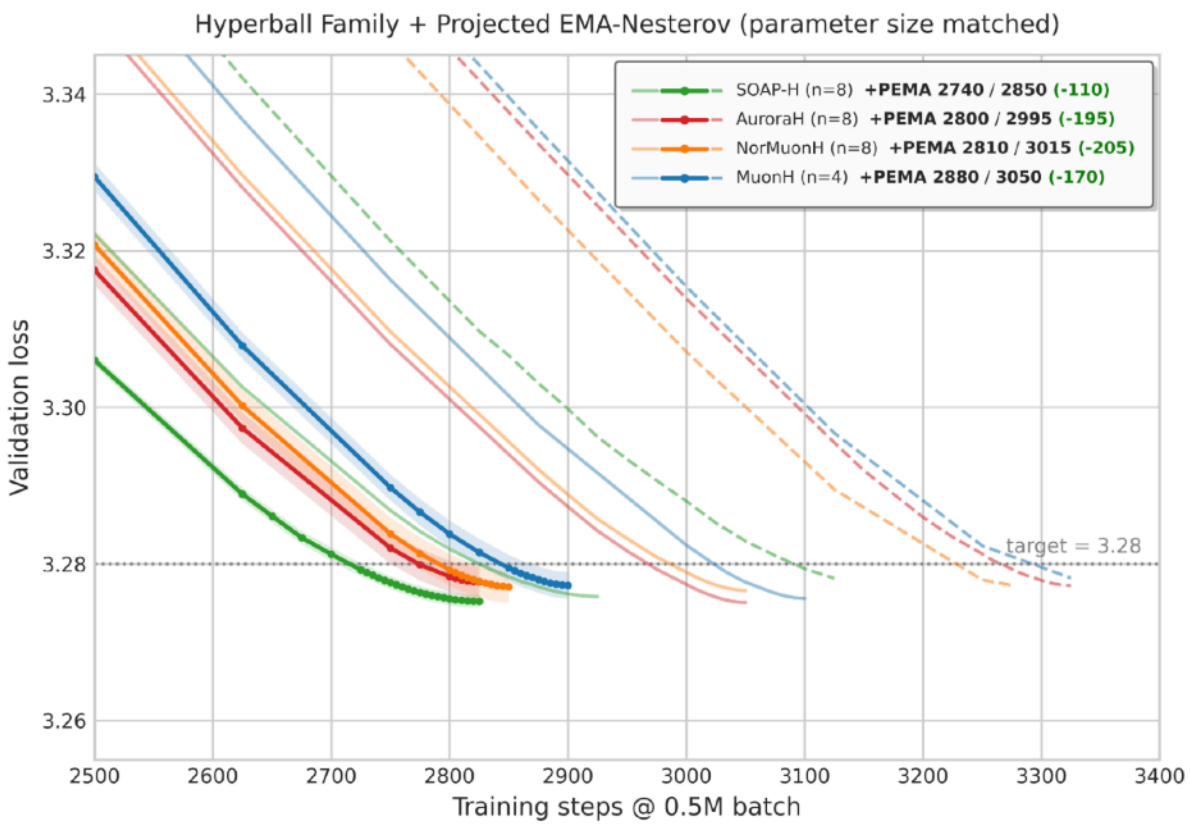
- With Muon, Parallax retains its advantage in both parameter-matched and compute-matched comparisons.

Size Optimizer	Model	RoPE ρ	LMB. ppl↓	Wiki ppl↓	LMB. acc↑	BoolQ acc↑	HSwag acc↑	PIQA acc↑	ARC-e acc↑	ARC-c acc↑	Wino. acc↑	OBQA acc↑	SciQ acc↑	Avg ↑
Cosine AdamW	Transformer	—	31.57	26.68	34.93	61.47	46.65	70.35	60.19	30.63	52.33	34.00	80.50	52.34
	Parallax	✓	29.54	26.63	36.48	58.38	46.25	69.75	58.16	30.29	52.96	36.40	74.40	51.45
WSD AdamW	Transformer	—	26.63	25.30	36.72	58.41	48.19	70.73	58.08	31.48	54.22	35.20	80.50	52.61
	Parallax	✓	26.59	25.01	37.40	57.16	48.63	71.44	60.90	32.00	52.17	35.60	78.80	52.68
0.6B Muon	Kimi DeltaAttn	—	25.16	26.81	38.29	55.29	49.30	71.11	62.12	33.19	52.57	34.20	78.50	52.73
	Gated DeltaNet	—	24.63	26.32	37.69	59.33	50.58	71.71	61.57	32.68	55.41	35.60	78.50	53.67
	Transformer	—	22.15	23.43	40.07	58.84	52.29	70.73	61.74	33.36	55.41	**37.20**	81.20	54.54
	Transformer†	—	22.35	23.36	39.45	56.17	52.35	71.92	63.01	35.49	**56.59**	36.80	82.30	54.90
	Parallax	✗	19.77	22.69	41.04	60.10	53.14	71.93	**64.27**	34.73	55.64	35.20	**83.80**	55.54
	Parallax†	✓	20.29	22.49	40.21	**61.44**	**54.48**	72.03	63.80	**36.95**	55.01	35.80	82.40	55.79
	Parallax	✓	**18.56**	**22.25**	**41.83**	60.24	53.73	**72.52**	63.51	35.49	56.20	36.80	83.60	**55.99**
1.7B Muon	Transformer	—	13.07	18.11	46.77	**64.92**	62.94	**76.39**	69.53	42.32	**61.01**	**41.00**	88.00	61.43
	Parallax	✗	10.85	17.27	49.54	62.72	64.34	75.79	70.33	**43.69**	59.91	39.80	**89.10**	61.69
	Parallax	✓	**10.80**	**17.08**	**50.26**	64.59	**64.54**	**76.39**	**73.27**	42.49	60.77	**41.00**	88.70	**62.45**

Performance

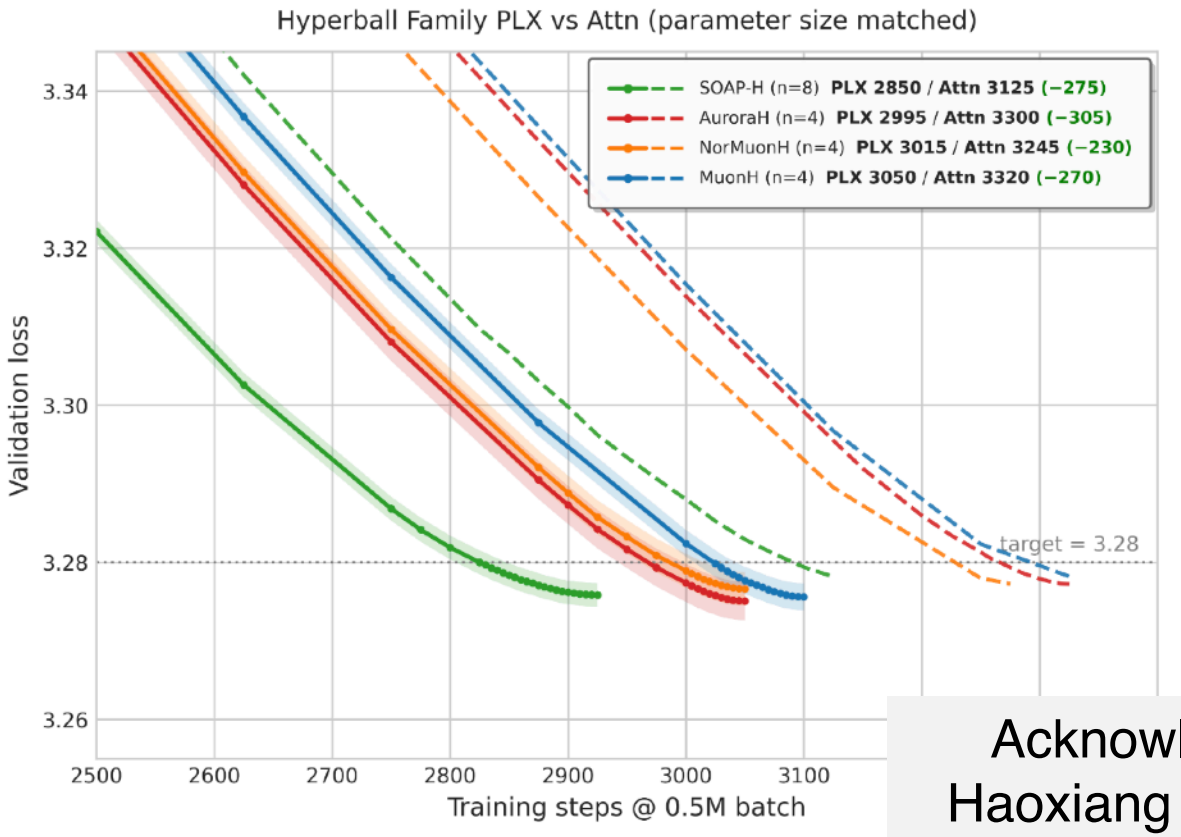
- Parallax also benefits from advanced optimizers, including Aurora, NorMuon, SOAP and their Hyperball/EMA variants.

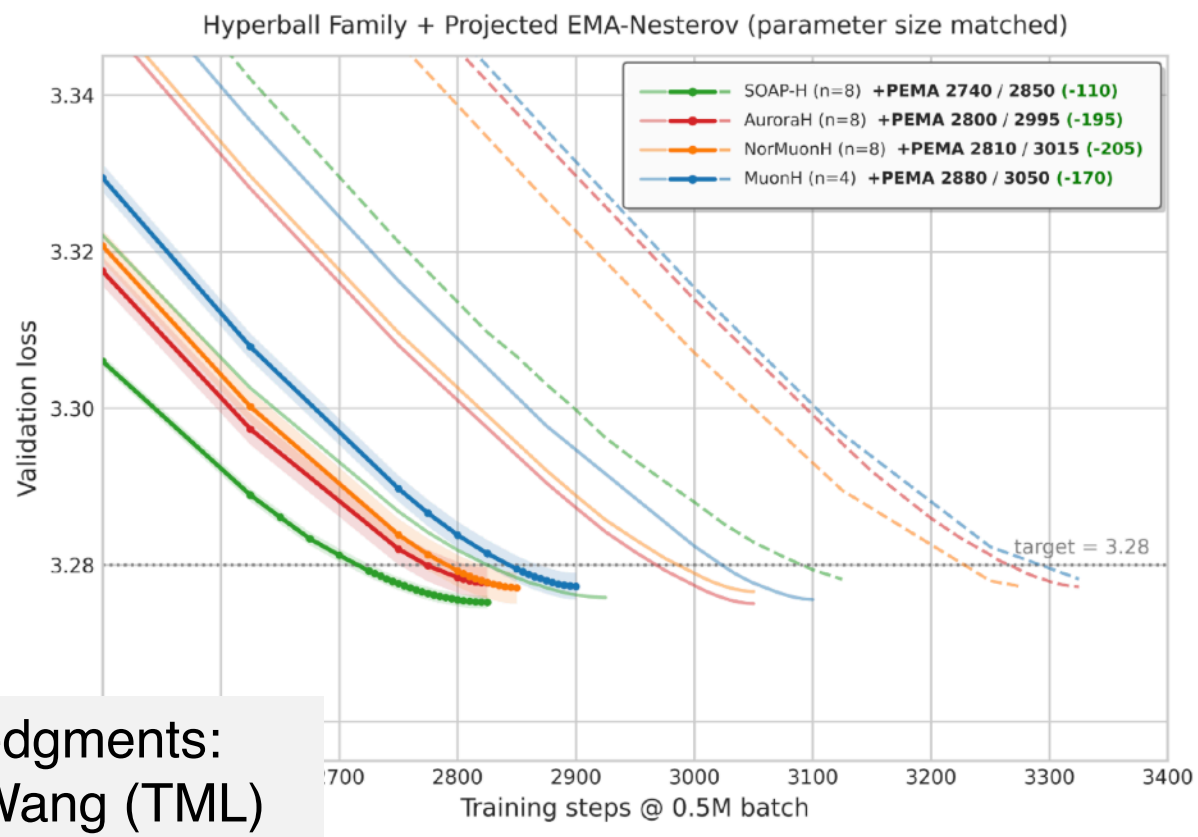




Performance

- Parallax also benefits from advanced optimizers, including Aurora, NorMuon, SOAP and their Hyperball/EMA variants.



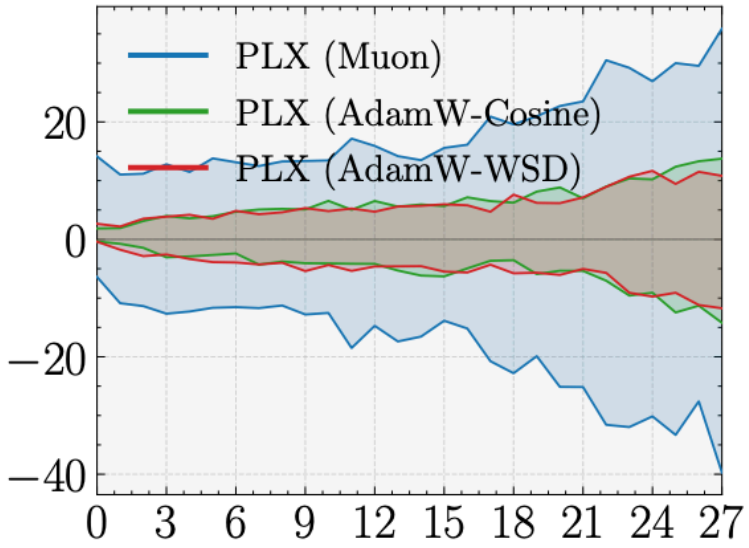


Acknowledgments:
Haoxiang Wang (TML)
Min Li (UIUC)

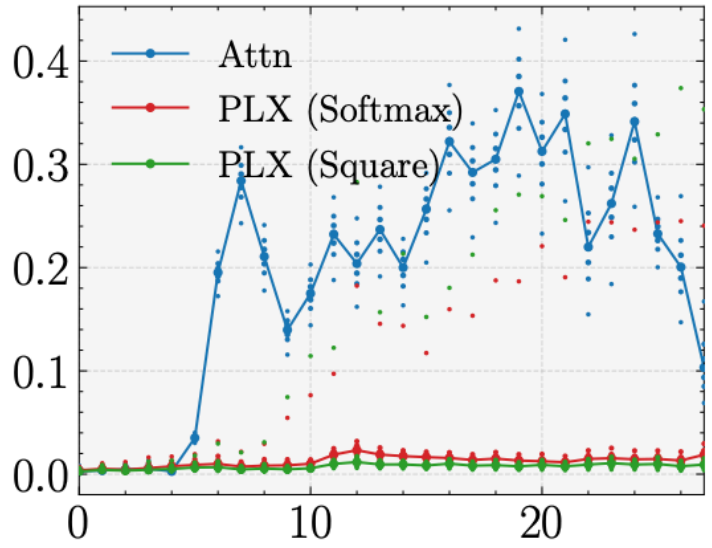
Parallax Score Patterns

- Parallax scores are signed and span a wide numerical range.

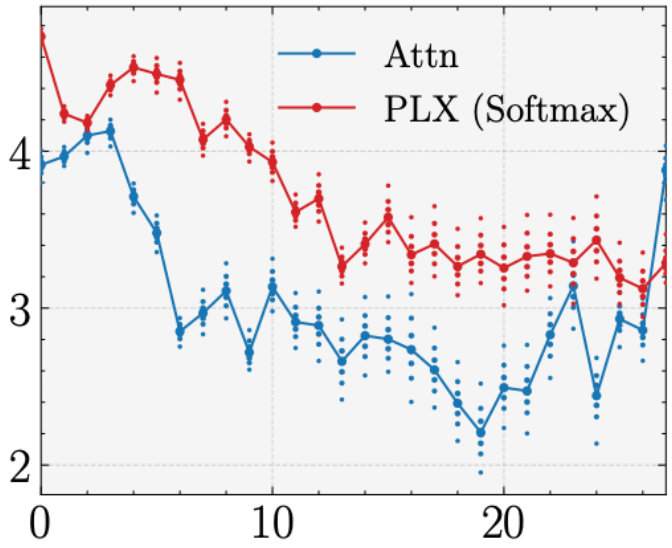
$$s_{ij} = \frac{w_{ij}(1 - t_{ij} + \bar{t}_i)}{\sum_{j' \le i} w_{ij'}} = p_{ij}(1 - t_{ij} + \bar{t}_i)$$



(a) Score ranges.



(b) Attention sink.

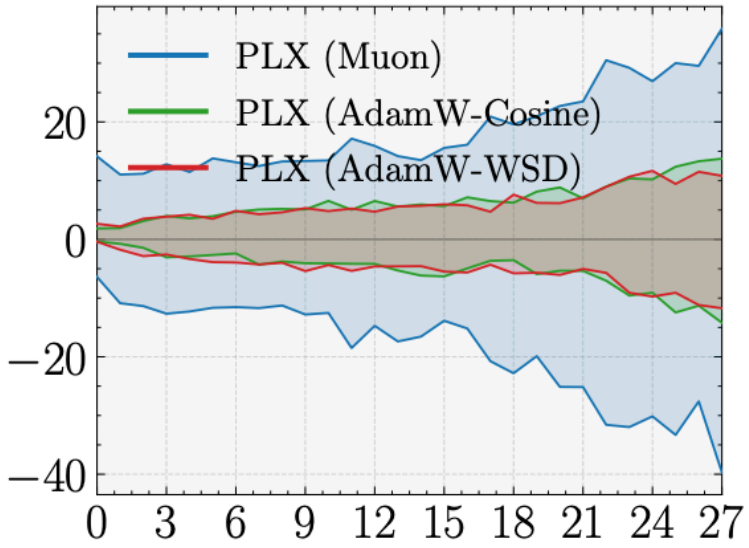


(c) Attention entropy.

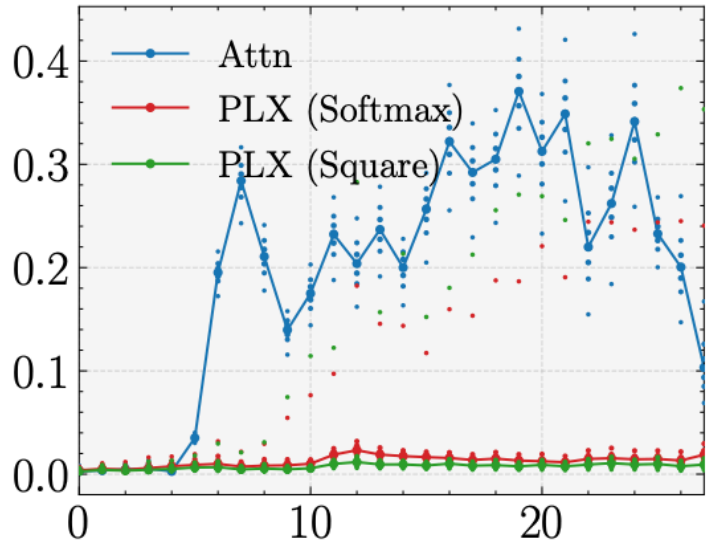
Parallax Score Patterns

- Parallax suppresses the attention sinks.

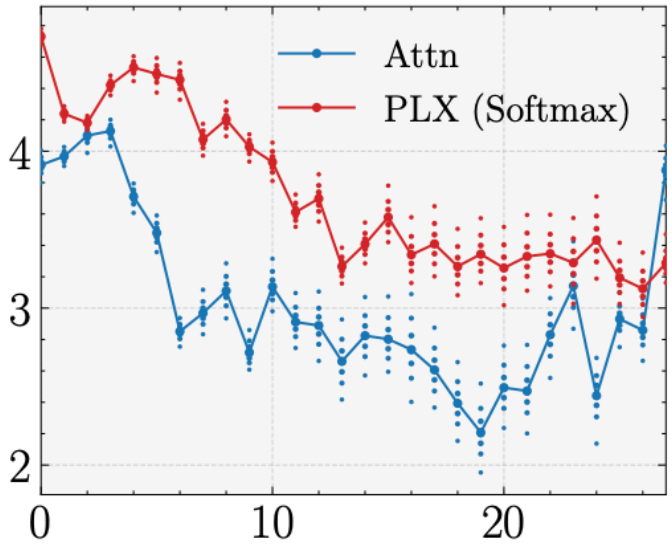
$$s_{ij} = \frac{w_{ij}(1 - t_{ij} + \bar{t}_i)}{\sum_{j' \le i} w_{ij'}} = p_{ij}(1 - t_{ij} + \bar{t}_i)$$



(a) Score ranges.



(b) Attention sink.

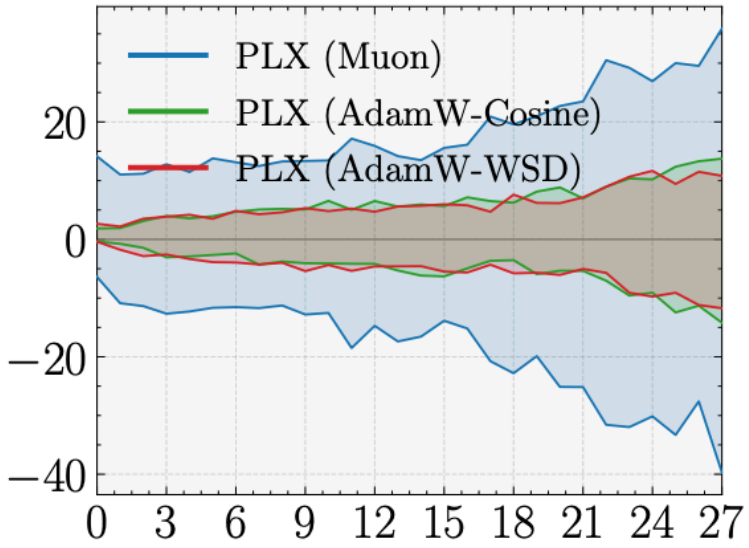


(c) Attention entropy.

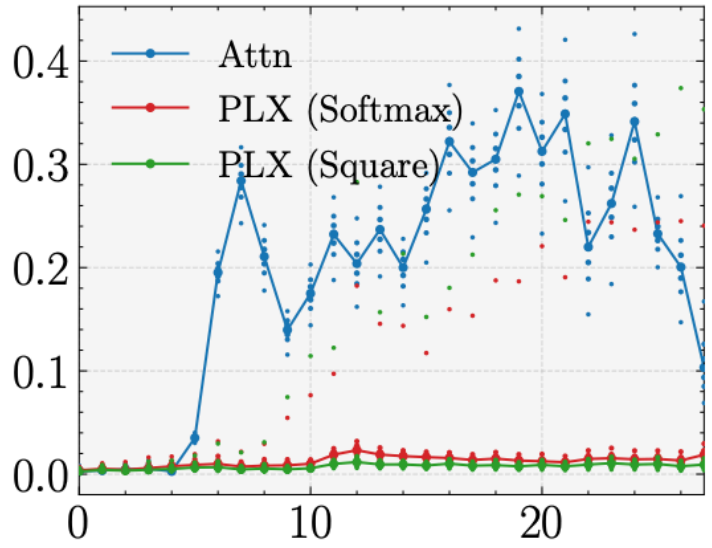
Parallax Score Patterns

- Parallax yields higher attention entropy.

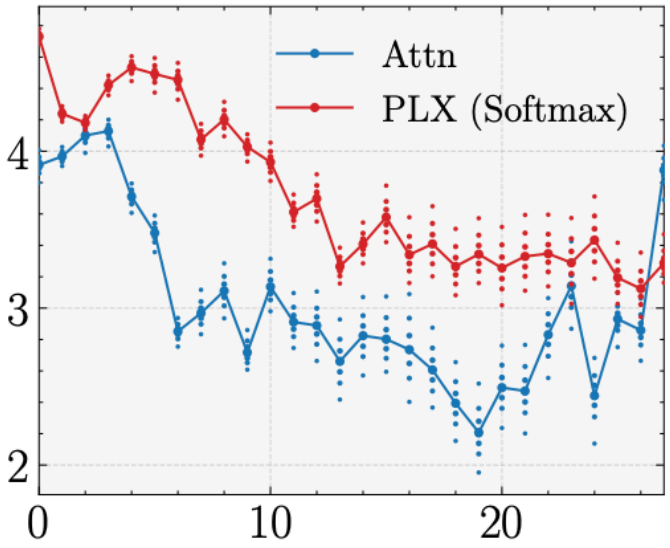
$$s_{ij} = \frac{w_{ij}(1 - t_{ij} + \bar{t}_i)}{\sum_{j' \le i} w_{ij'}} = p_{ij}(1 - t_{ij} + \bar{t}_i)$$



(a) Score ranges.



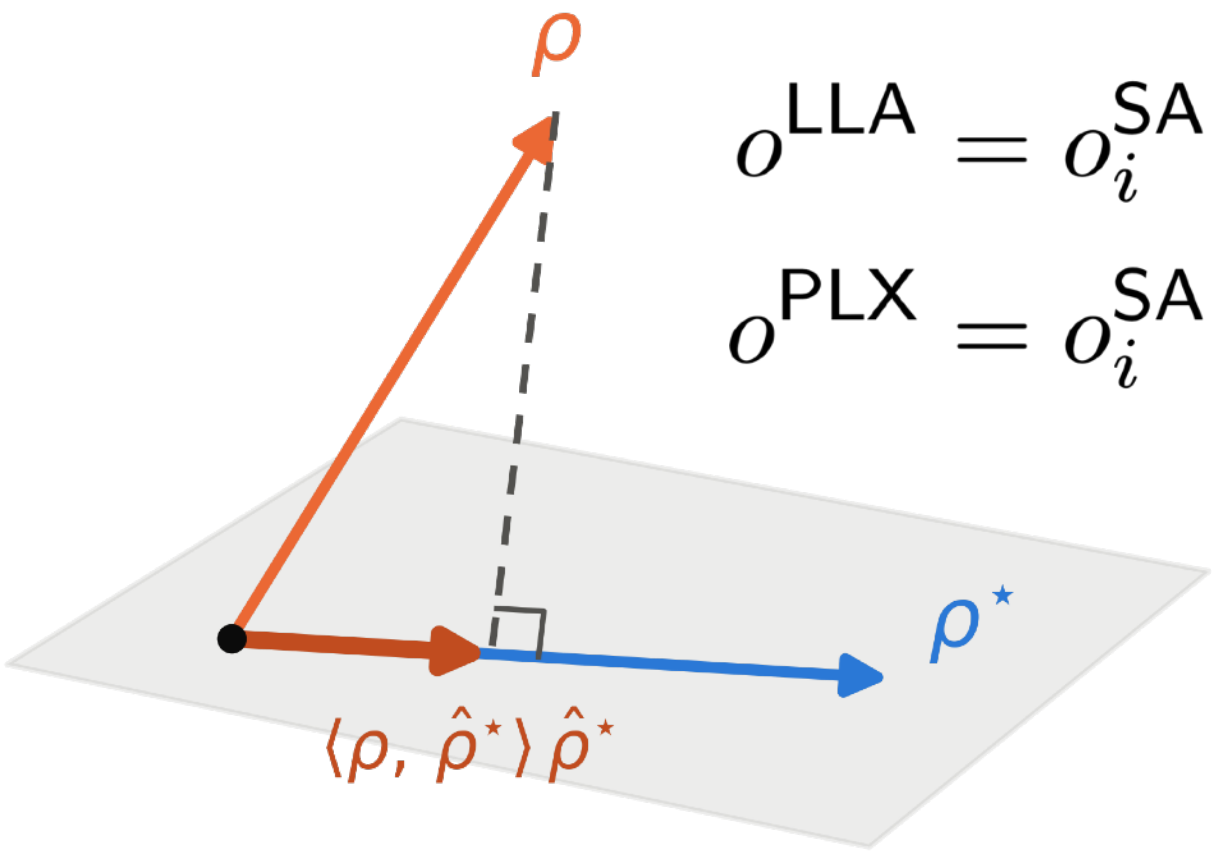
(b) Attention sink.



(c) Attention entropy.

Mechanism Analysis

- The affine structure induces magnitude requirements.

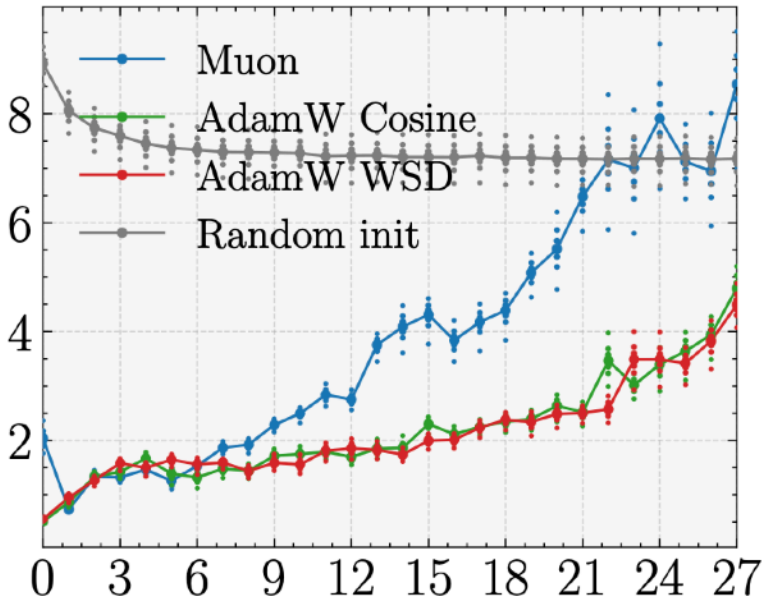


$$o^{\mathsf{LLA}} = o_i^{\mathsf{SA}} - (1+\eta_i)\Sigma_{KV}^{(i)}\rho_i^\star$$

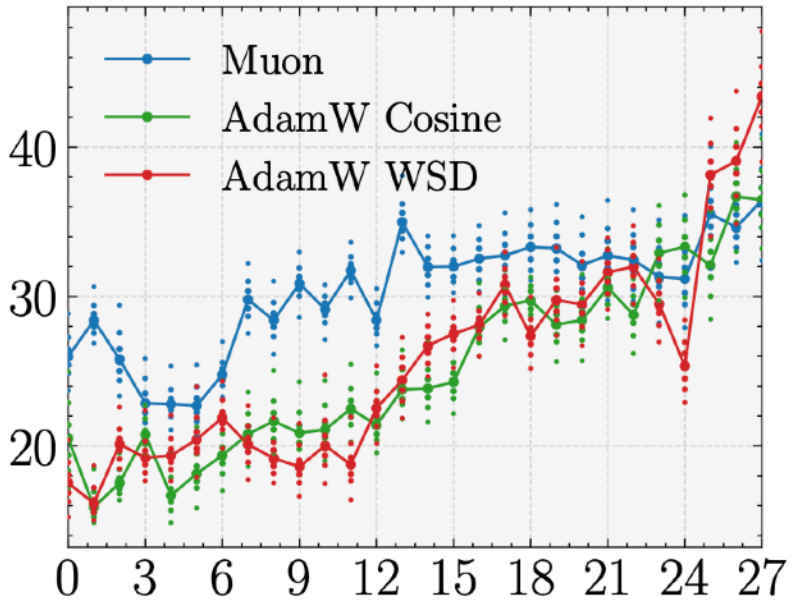
$$o^{\mathsf{PLX}} = o_i^{\mathsf{SA}} - \Sigma_{KV}^{(i)}\rho_i$$

Mechanism Analysis

- The covariance correction branch is used more effectively when trained with Muon.



(a) COR

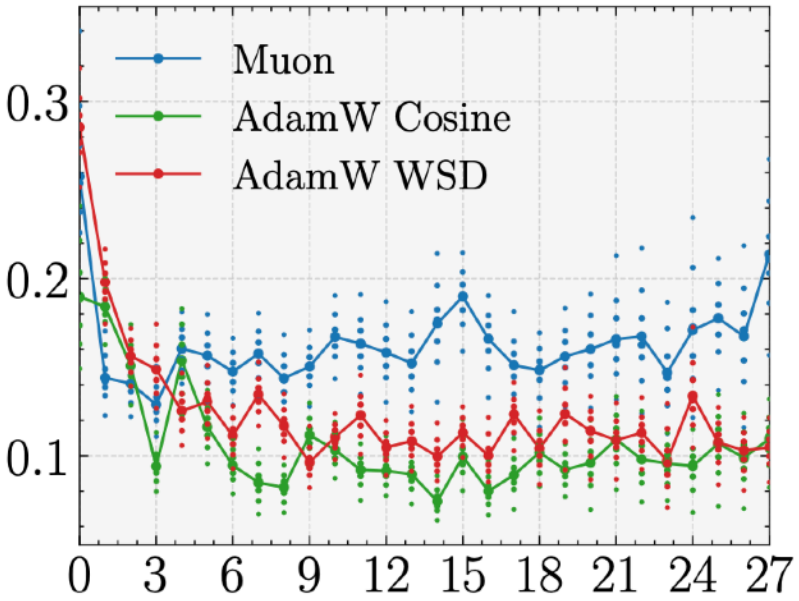


(b) $\|\mathtt{Corr}\|_F$

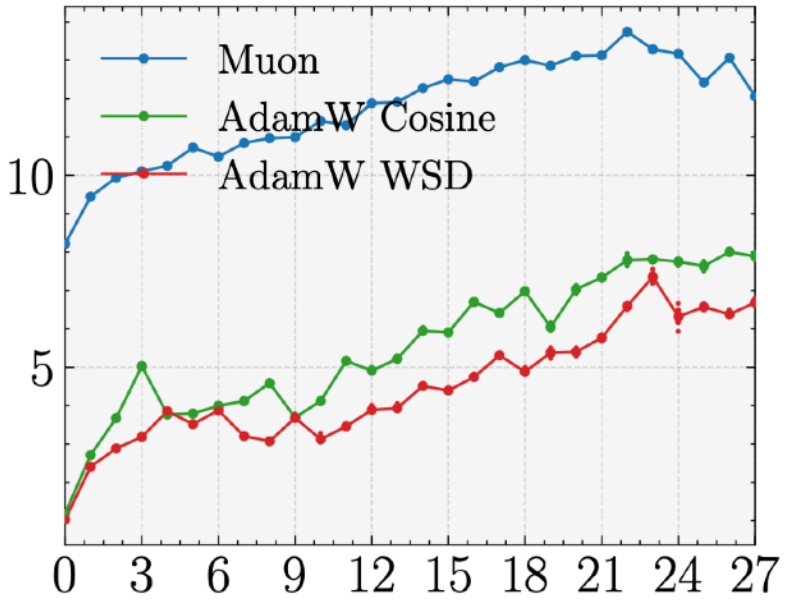
$$\mathtt{COR}_i = \|\boldsymbol{\Sigma}_{KV}^{(i)}\boldsymbol{\rho}_i\| / \|\boldsymbol{o}_i^{\mathsf{SA}}\| \qquad \mathtt{Corr}_i = \boldsymbol{\Sigma}_{VV}^{(i)-\frac{1}{2}}\boldsymbol{\Sigma}_{KV}^{(i)}\boldsymbol{\Sigma}_{KK}^{(i)-\frac{1}{2}}$$

Mechanism Analysis

- The covariance correction branch is used more effectively when trained with Muon.



(c) CPA



(d) $\|\boldsymbol{\rho}\|$

$$\mathrm{CPA}_i = \|\boldsymbol{\Sigma}_{KV}^{(i)} \boldsymbol{\rho}_i\| / \|\boldsymbol{\rho}_i\| \|\boldsymbol{\Sigma}_{KV}^{(i)}\|_2$$

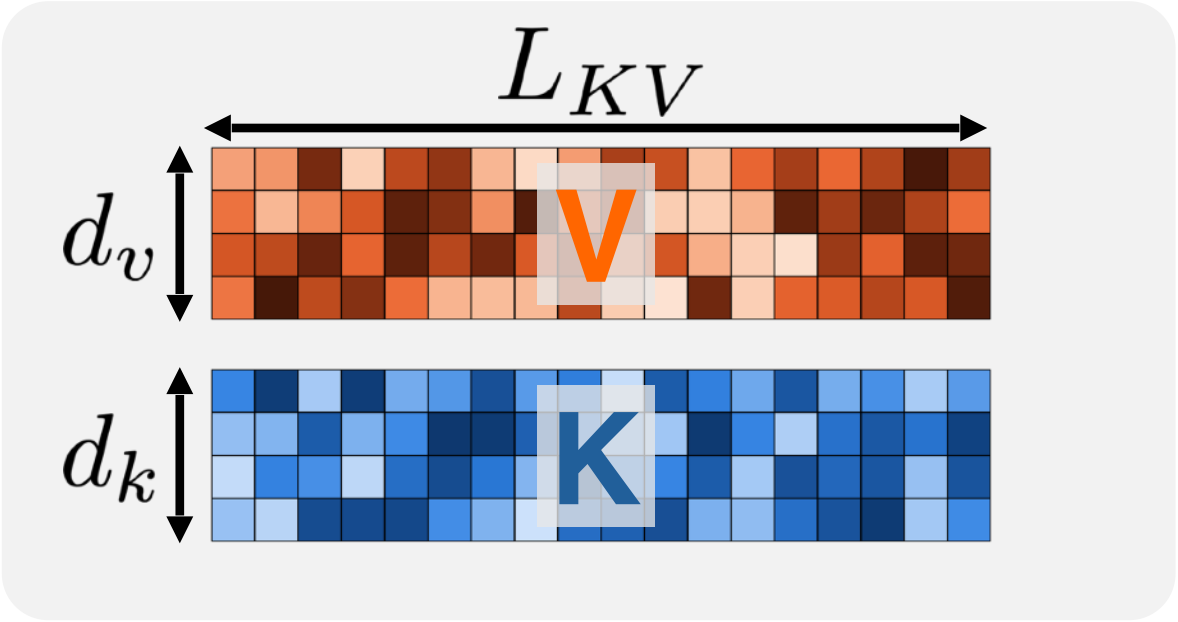
Mechanism Analysis

- Muon prevents stable rank collapse in the QK and RK circuits.
- Parallax improves the OV circuit across all optimizers.

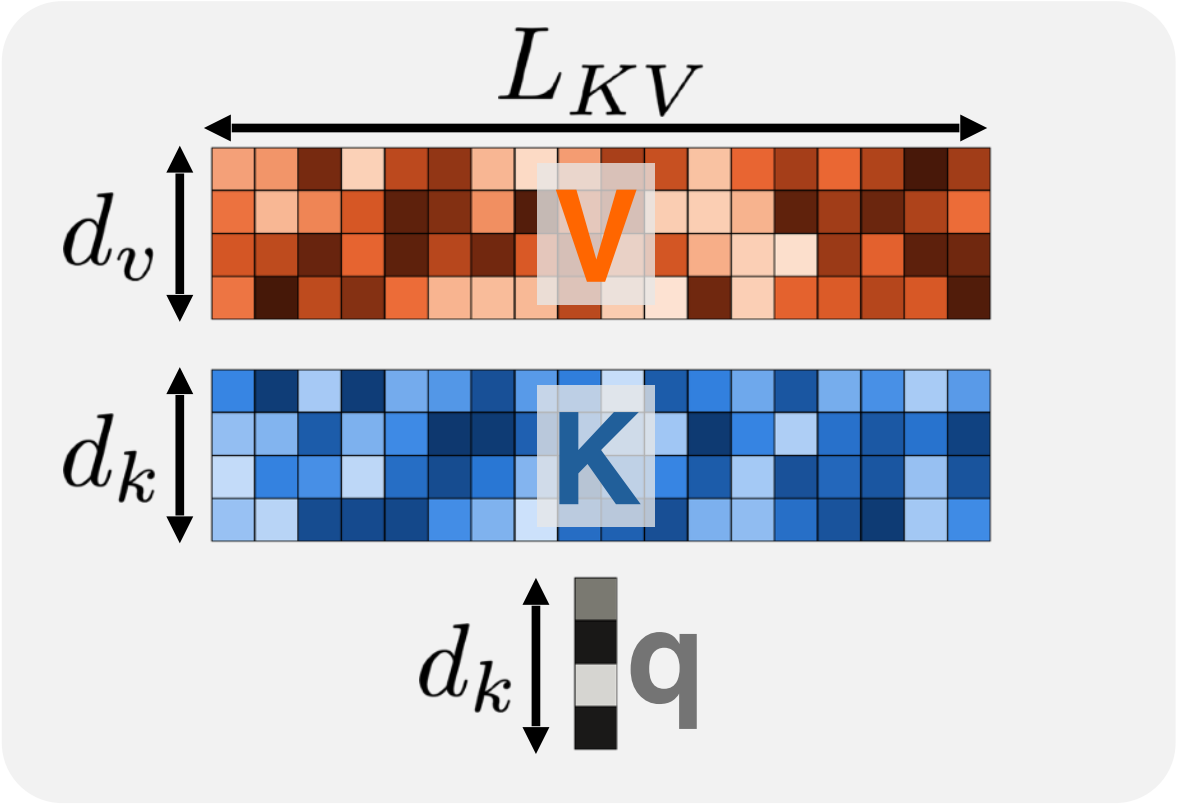
	Softmax Attention			Parallax		
	Muon	AW-Cos	AW-WSD	Muon	AW-Cos	AW-WSD
$\boldsymbol{W}_Q$	116.0	**17.4**↓	**21.7**↓	97.4	**20.9**↓	**20.7**↓
$\boldsymbol{W}_K$	105.8	**18.1**↓	**22.9**↓	106.4	**18.0**↓	**18.3**↓
$\boldsymbol{W}_V$	145.5	148.1	138.6	**199.8**↑	**177.2**↑	**184.7**↑
$\boldsymbol{W}_O$	186.6	109.8	121.6	**182.0**↑	**137.4**↑	**144.7**↑
$\boldsymbol{W}_R$		—		134.0	**9.3**↓	**11.1**↓
$\boldsymbol{W}_{QK}$	22.3	**4.6**↓	**5.9**↓	25.5	**4.9**↓	**5.1**↓
$\boldsymbol{W}_{OV}$	24.8	21.2	21.5	**34.1**↑	**30.4**↑	**30.4**↑
$\boldsymbol{W}_{RK}$		—		29.1	**9.4**↓	**9.2**↓

Table 4 Stable ranks of projection matrices and bilinear circuits, averaged across heads and layers.

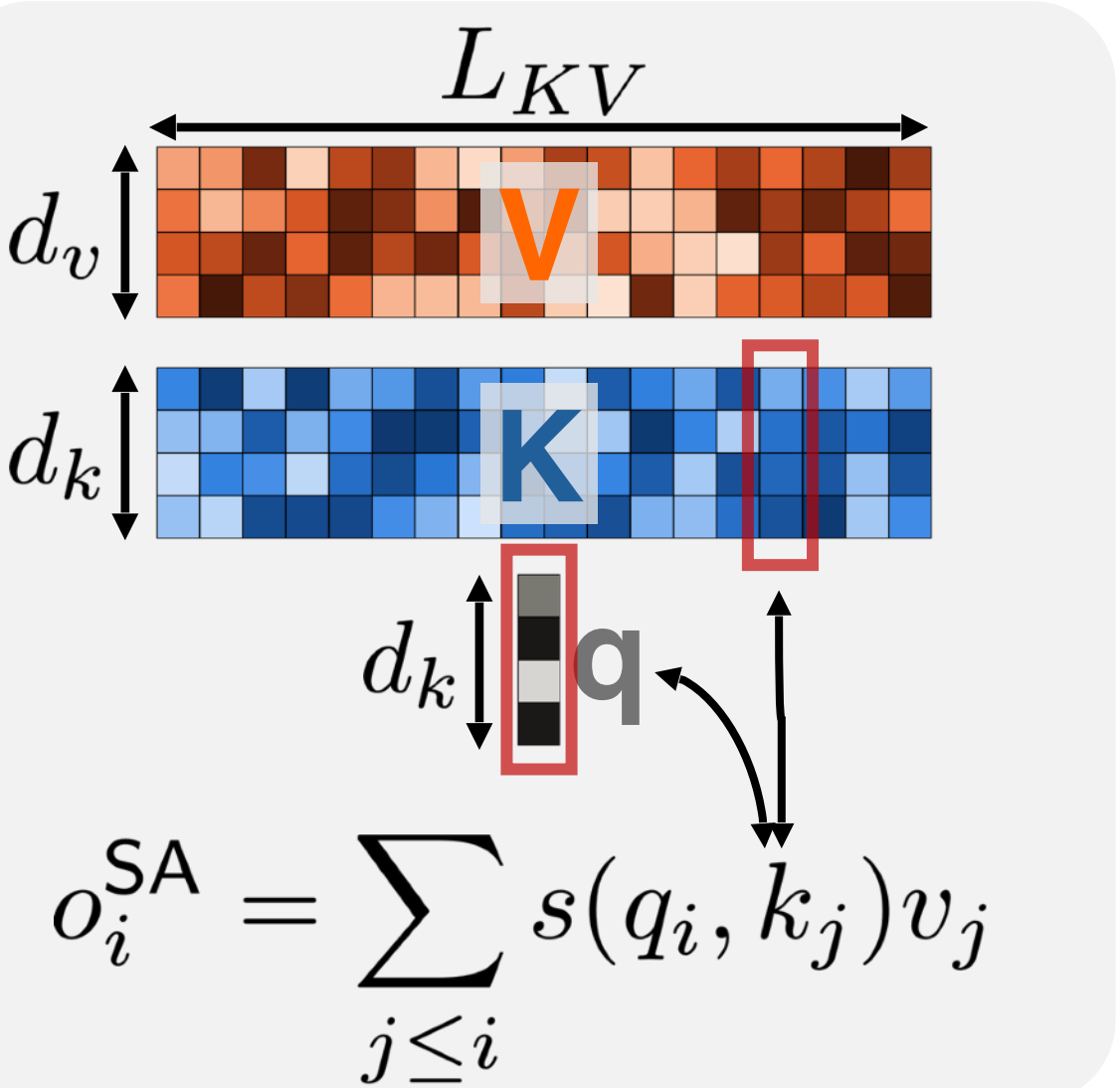
From Associative Memory to Attention



From Associative Memory to Attention

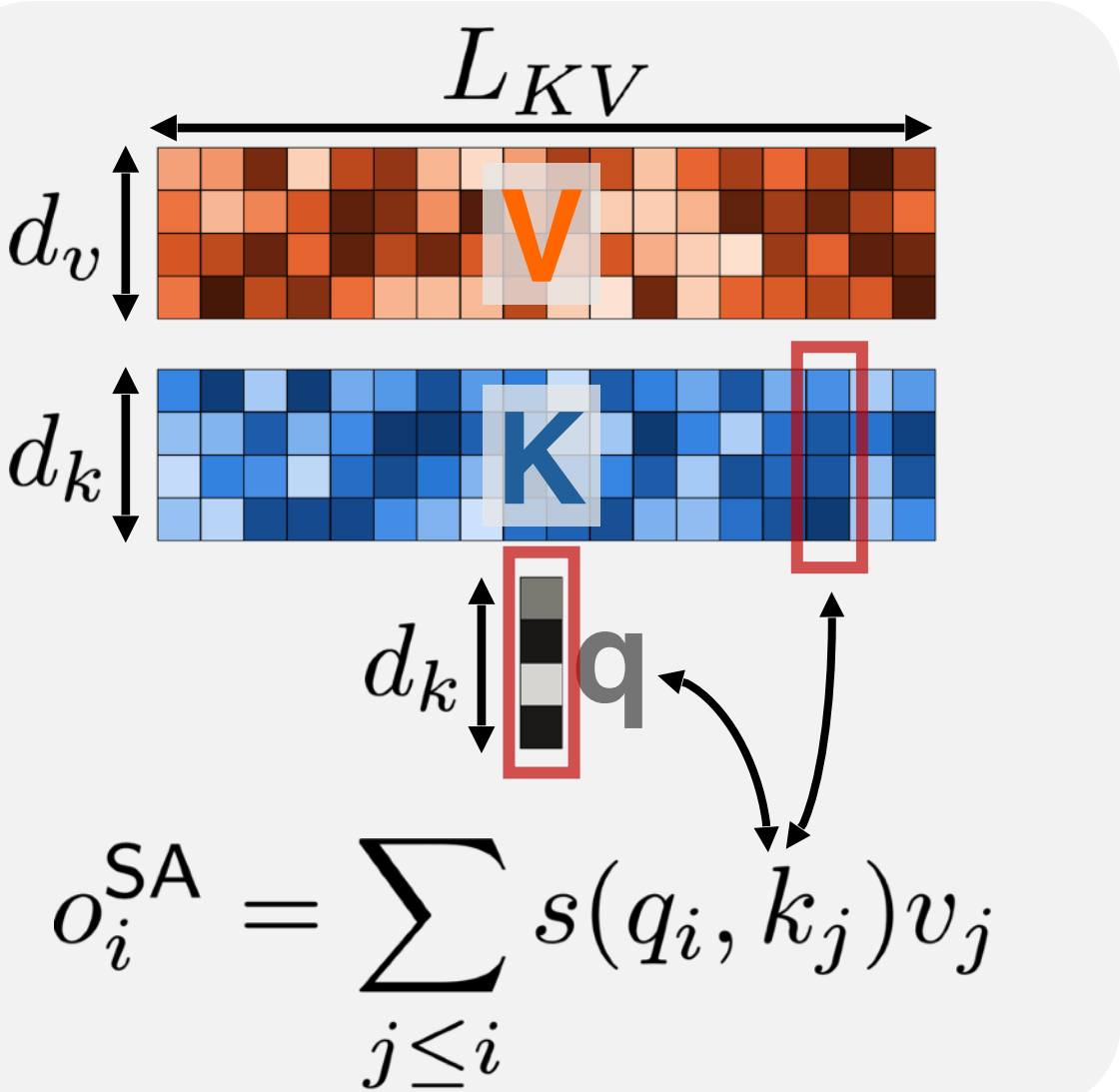


From Associative Memory to Attention



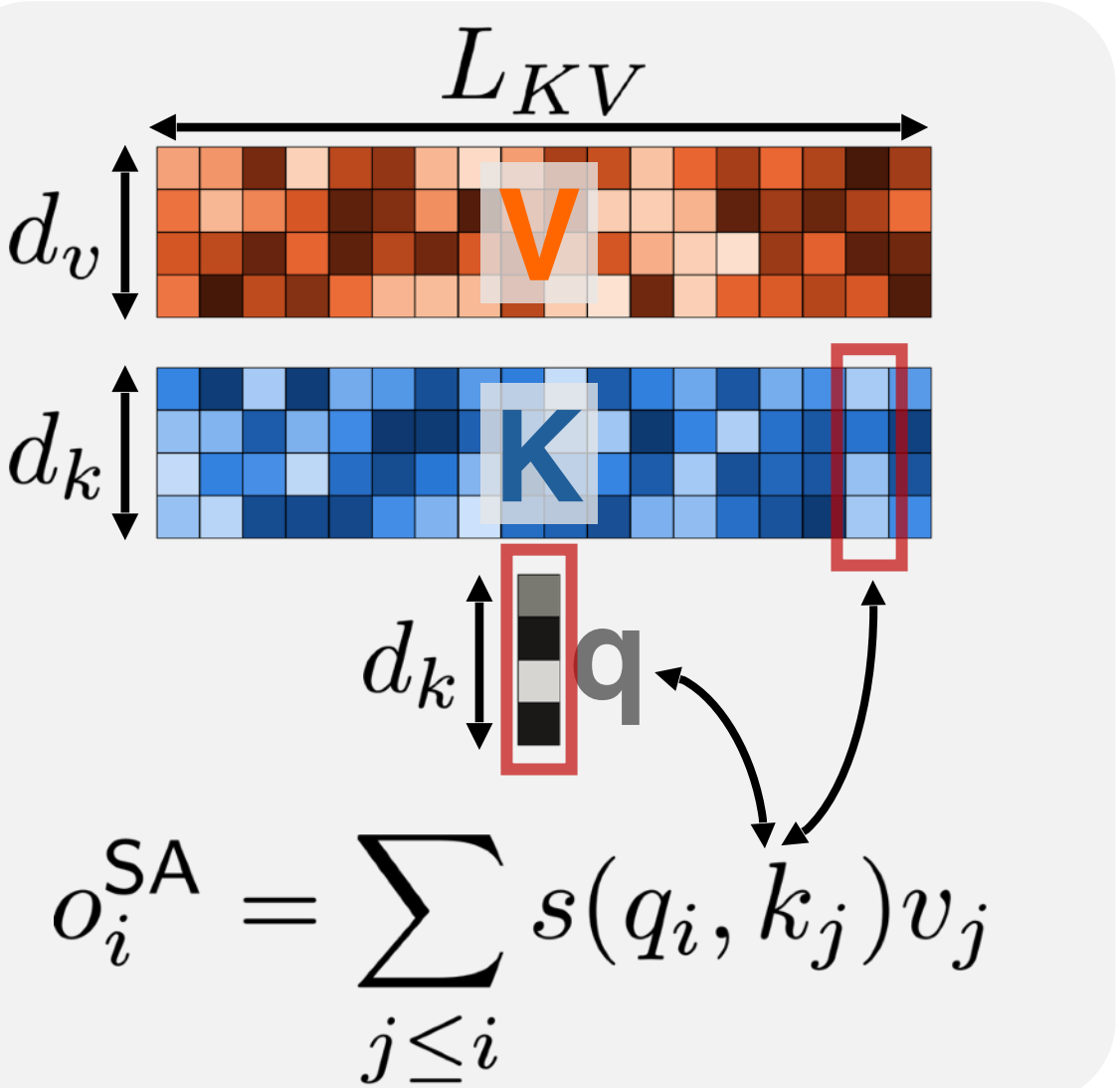
$$o_i^{\text{SA}} = \sum_{j \leq i} s(q_i, k_j) v_j$$

From Associative Memory to Attention



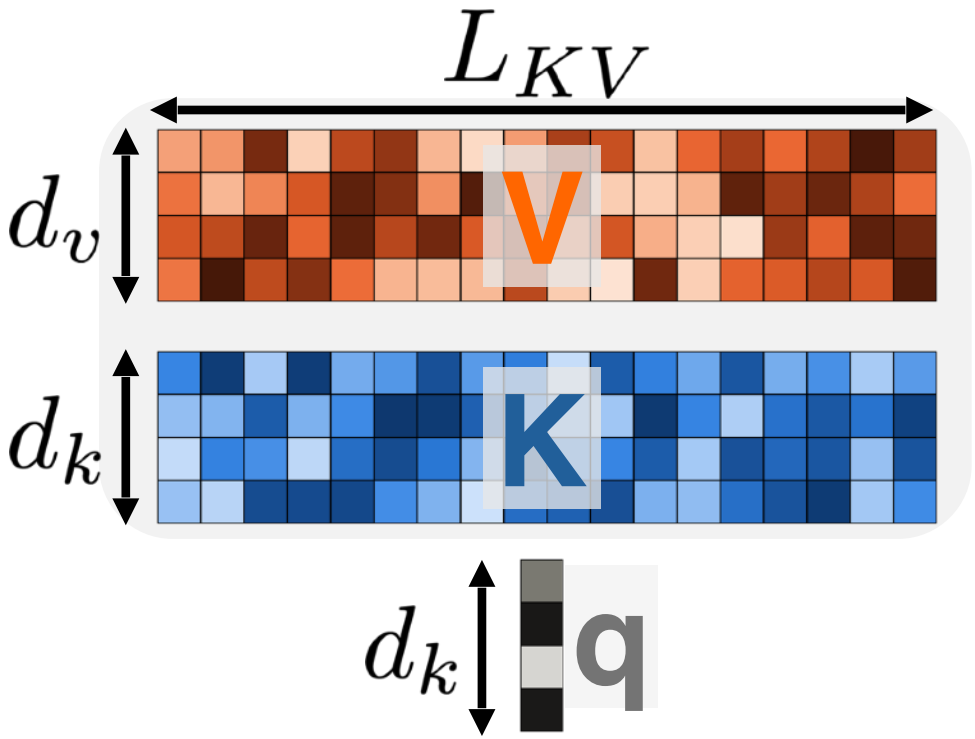
$$o_i^{\text{SA}} = \sum_{j \leq i} s(q_i, k_j) v_j$$

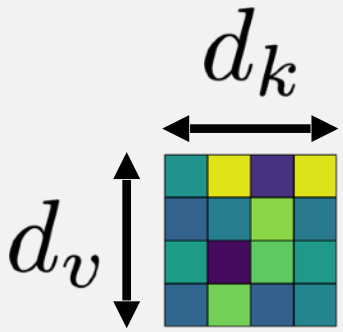
From Associative Memory to Attention



$$o_i^{\text{SA}} = \sum_{j \leq i} s(q_i, k_j) v_j$$

From Associative Memory to Attention

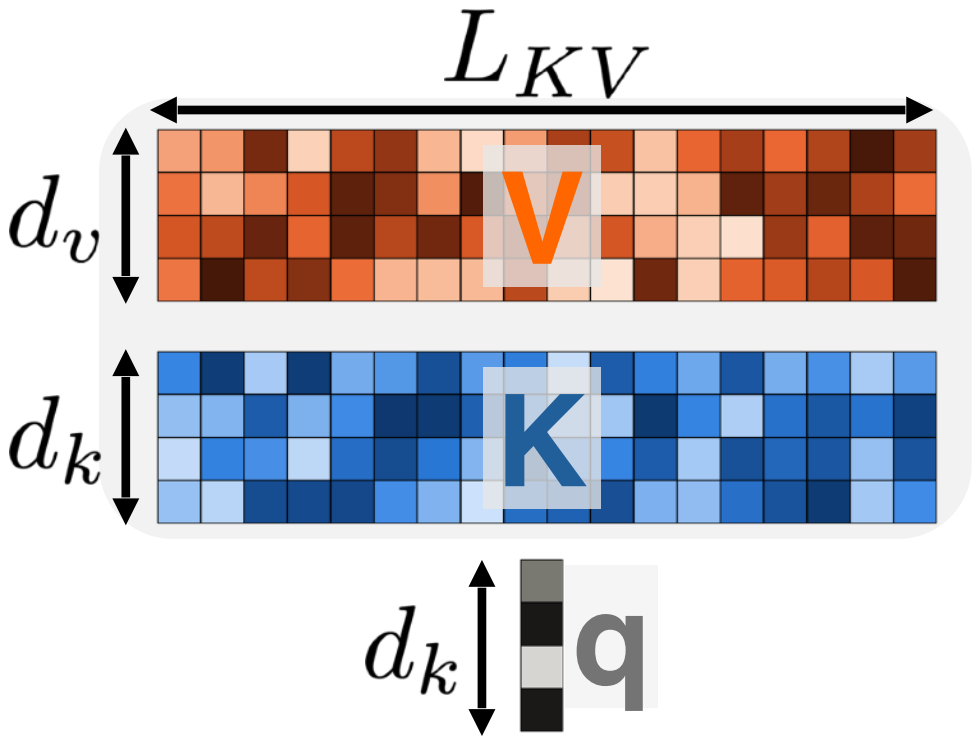


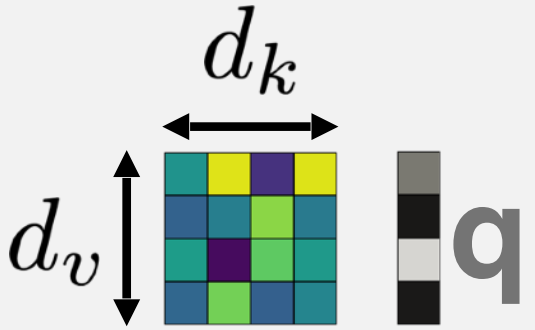


$$W_i = \sum_{j \leq i} v_j k_j^\top$$

$$o_i^{\mathsf{SA}} = \sum_{j \leq i} s(q_i, k_j) v_j$$

From Associative Memory to Attention

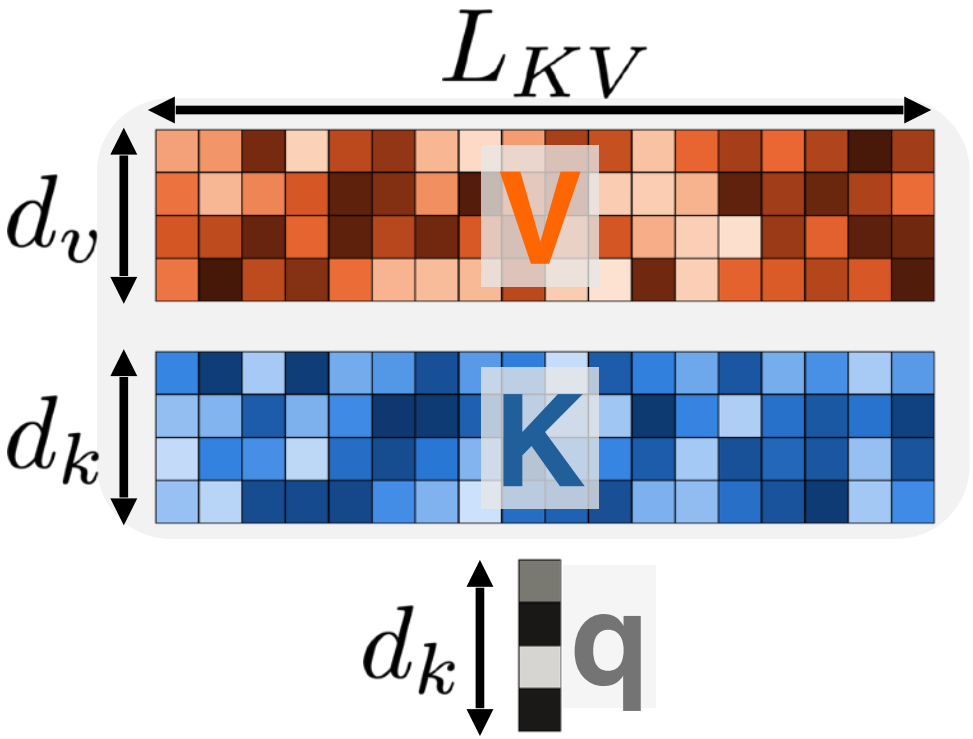


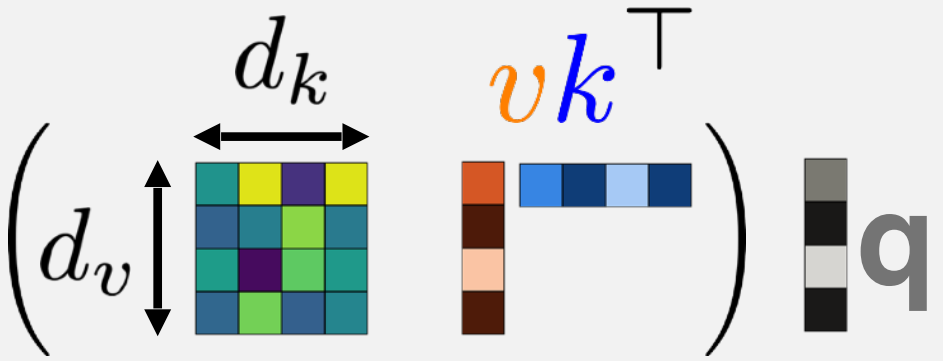


$$W_i = \sum_{j \leq i} v_j k_j^\top$$

$$o_i^{\mathsf{SA}} = \sum_{j \leq i} s(q_i, k_j) v_j$$

From Associative Memory to Attention

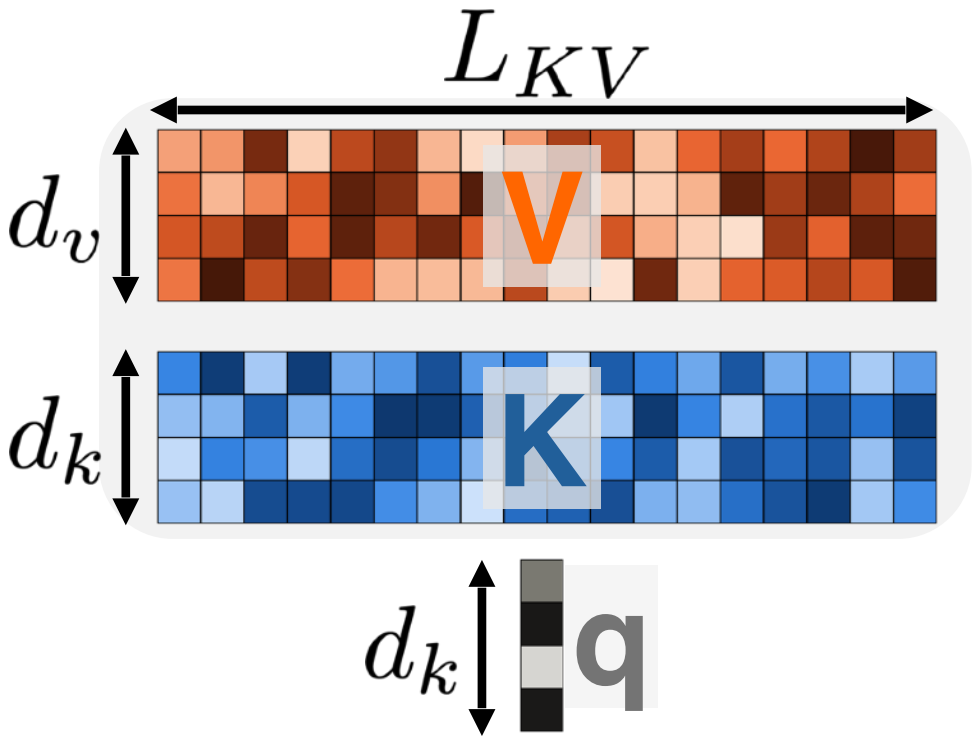


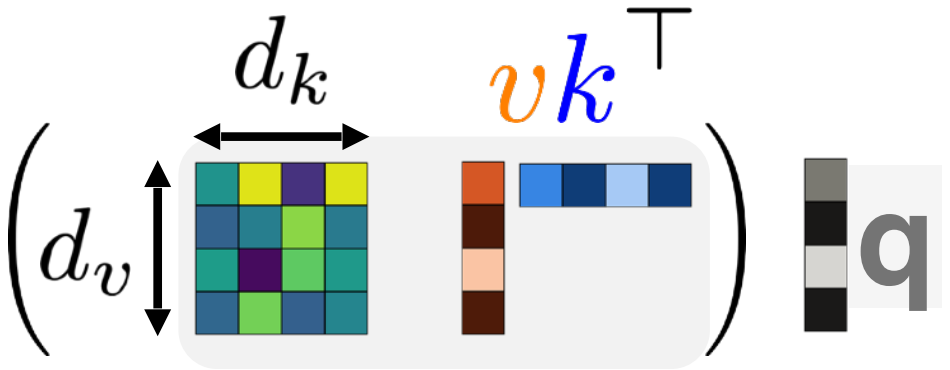


$$W_{i+1} = W_i + v_{i+1} k_{i+1}^\top$$

$$o_i^{\mathsf{SA}} = \sum_{j \leq i} s(q_i, k_j) v_j$$

From Associative Memory to Attention





$$W_{i+1} = W_i + v_{i+1} k_{i+1}^\top$$

$$o_i^{\text{SA}} = \sum_{j \le i} s(q_i, k_j) v_j$$

$$o_i^{\text{LA}} = \sum_{j \le i} v_j k_j^\top q_i = W_i q_i$$

From Associative Memory to Attention

$$\mathsf{LA}(k_l) = \sum_{j \le i} v_j k_j^\top k_l = \underbrace{v_l (k_l^\top k_l)}_{\text{recall}} + \underbrace{\sum_{j \neq l} v_j (k_j^\top k_l)}_{\text{interference}}$$

From Associative Memory to Attention

$$\mathsf{LA}(k_l) = \sum_{j \le i} v_j k_j^\top k_l = \underbrace{v_l (k_l^\top k_l)}_{\text{recall}} + \underbrace{\sum_{j \neq l} v_j (k_j^\top k_l)}_{\text{interference}}$$

- Interference vanishes only for orthogonal keys.

$$k_j^\top k_l = 0, \quad \forall j \neq l$$

- In practice, keys are correlated.
- If keys outnumber dims, interference is unavoidable.

From Associative Memory to Attention

Test Time Regression (Wang et al, 2025)

$$\hat{f} = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \mathcal{L}(f, \mathcal{D}_i) = \frac{1}{2} \sum_{j \leq i} \omega_{ij} \| f(k_j) - v_j \|^2 + \lambda \Omega(f)$$

From Associative Memory to Attention

Test Time Regression (Wang et al, 2025)

$$\hat{f} = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \mathcal{L}(f, \mathcal{D}_i) = \frac{1}{2} \sum_{j \leq i} \omega_{ij} \| f(k_j) - v_j \|^2 + \lambda \Omega(f)$$

Hypothesis Space

From Associative Memory to Attention

Test Time Regression (Wang et al, 2025)

$$\hat{f} = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \mathcal{L}(f, \mathcal{D}_i) = \frac{1}{2} \sum_{j \leq i} \omega_{ij} \| f(k_j) - v_j \|^2 + \lambda \Omega(f)$$

$$\mathcal{D}_i = \{(k_j, v_j)\}_{j=1}^{i}$$

From Associative Memory to Attention

Test Time Regression (Wang et al, 2025)

$$\hat{f} = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \mathcal{L}(f, \mathcal{D}_i) = \frac{1}{2} \sum_{j \leq i} \omega_{ij} \| f(k_j) - v_j \|^2 + \lambda \Omega(f)$$

(Weighted) MSE

From Associative Memory to Attention

Test Time Regression (Wang et al, 2025)

$$\hat{f} = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \mathcal{L}(f, \mathcal{D}_i) = \frac{1}{2} \sum_{j \leq i} \omega_{ij} \| f(k_j) - v_j \|^2 + \lambda \Omega(f)$$

Regularization

From Associative Memory to Attention

$$\hat{f} = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \mathcal{L}(f, \mathcal{D}_i) = \frac{1}{2} \sum_{j \leq i} \omega_{ij} \| f(k_j) - v_j \|^2 + \lambda \Omega(f)$$

- LA corresponds to **parametric linear regression**

$$\mathcal{H} = \{ f(x) = Wx + b \}$$

$$\hat{f} = \underset{W}{\operatorname{argmin}} \mathcal{L}(W, \mathcal{D}_i) = \frac{1}{2} \sum_{j \leq i} \omega_{ij} \| W k_j - v_j \|^2 + \lambda \| W \|_F$$

- DeltaNet (yang et al., 2024): SGD, unregularized
- MesaNet (von Oswald et al., 2026): optimal ridge regression.

From Associative Memory to Attention

$$\hat{f} = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \mathcal{L}(f, \mathcal{D}_i) = \frac{1}{2} \sum_{j \leq i} \omega_{ij} \| f(k_j) - v_j \|^2 + \lambda \Omega(f)$$

- LA corresponds to **parametric linear regression**

$$\mathsf{DeltaNet}(q_i) = \sum_{j \leq i} \beta_j v_j k_j^\top \left(\prod_{t=j+1}^{i} \left(I - \beta_t k_t k_t^\top \right) \right) q_i$$

- DeltaNet (yang et al., 2024): SGD, unregularized
- MesaNet (von Oswald et al., 2026): optimal ridge regression.

From Associative Memory to Attention

$$\hat{f} = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \mathcal{L}(f, \mathcal{D}_i) = \frac{1}{2} \sum_{j \leq i} \omega_{ij} \| f(k_j) - v_j \|^2 + \lambda \Omega(f)$$

- LA corresponds to **parametric linear regression**

$$\mathsf{MesaNet}(q_i) = \left(\sum_{j \leq i} \beta_j v_j k_j^\top \right) \left(\sum_{j \leq i} \beta_j k_j k_j^\top + \lambda I \right)^{-1} q_i$$

- DeltaNet (yang et al., 2024): SGD, unregularized
- MesaNet (von Oswald et al., 2026): optimal ridge regression.

From Associative Memory to Attention

$$\hat{f} = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \mathcal{L}(f, \mathcal{D}_i) = \frac{1}{2} \sum_{j \leq i} \omega_{ij} \| f(k_j) - v_j \|^2 + \lambda \Omega(f)$$

- SA corresponds to **nonparametric kernel regression**

$$\mathcal{H}(q_i) = \{c\}, \quad K(q_i, k_j) = \exp(q_i^\top k_j / h)$$

$$\hat{f} = \underset{f \in \mathcal{H}(q_i)}{\operatorname{argmin}} \mathcal{L}(f, \mathcal{D}_i) = \frac{1}{2} \sum_{j \leq i} K(q_i, k_j) \| c - v_j \|^2$$

- SA is a **Local Constant (NW)** estimator.
- If QKNorm is applied, the kernel becomes an RBF kernel.

From Associative Memory to Attention

$$\hat{f} = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \mathcal{L}(f, \mathcal{D}_i) = \frac{1}{2} \sum_{j \le i} \omega_{ij} \| f(k_j) - v_j \|^2 + \lambda \Omega(f)$$

- LLA corresponds to **nonparametric kernel regression**

$$\mathcal{H}(q_i) = \{ f(x) = W(x - q_i) + b \}, \quad K(q_i, k_j) = \exp(q_i^\top k_j / h)$$

$$\hat{f} = \underset{W,b}{\operatorname{argmin}} \frac{1}{2} \sum_{j \le i} K(q_i, k_j) \| W(k_j - q_i) + b - v_j \|^2 + \lambda \| W \|_F^2$$

- LLA is a **Local Linear** estimator.
- If QKNorm is applied, the kernel becomes an RBF kernel

From Associative Memory to Attention

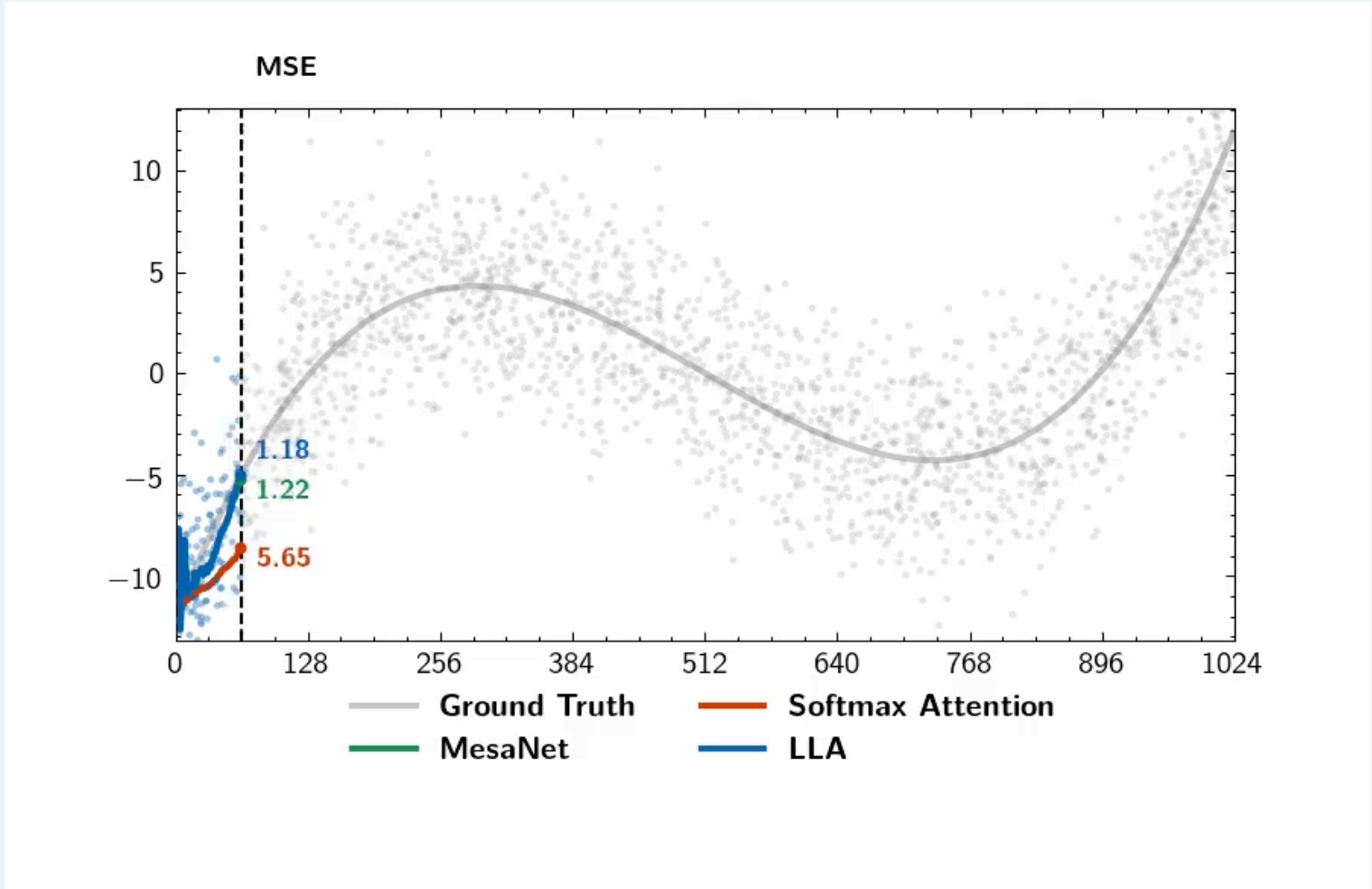
Theorem 2.1 (Bias-variance separation (Zuo et al., 2026)). *Let* $(X_i, Y_i)_{i=1}^n$ *be i.i.d. with* $X_i \in \mathbb{R}^d$ *supported on a bounded* C^2 *domain* D *and* $Y_i = f(X_i) + \varepsilon_i \in \mathbb{R}^{d_y}$, $\mathbb{E}[\varepsilon_i \mid X_i] = 0$. *Let* $\hat{f}_{\mathrm{GL}}$, $\hat{f}_{\mathrm{NW}}$, *and* $\hat{f}_{\mathrm{LL}}$ *denote the Global Linear, Nadaraya–Watson (Local Constant), and Local Linear estimators with optimal bandwidths, respectively. Under Assumptions A.1–A.4, denote* $\mathcal{R}(\hat{f}) := \mathbb{E}\int_D \|\hat{f}(x) - f(x)\|^2 \, dx$ *the integrated mean squared error, then*

$$\mathcal{R}(\hat{f}_{\mathrm{GL}}) \gg \mathcal{R}(\hat{f}_{\mathrm{NW}}) \gg \mathcal{R}(\hat{f}_{\mathrm{LL}}). \tag{2}$$

The lower bound for $\hat{f}_{\mathrm{GL}}$ *holds whenever* f *is not globally affine. The lower bound for* $\hat{f}_{\mathrm{NW}}$ *holds whenever* f *has sufficiently large normal gradient along* ∂D *(Assumption A.4).*

- LA (GL): irreducible misspecification error.
- SA (NW): unavoidable boundary bias.
- LLA (LL): minimax-optimal for C^2 target.

From Associative Memory to Attention



Local Linear Attention

For a query q_i , write $w_{ij} = \exp(q_i^\top k_j/h)$ and $z_{ij} = k_j - q_i$,

LLA computes output as

$$o_i^{\text{LLA}} = \sum_{j \le i} w_{ij} \frac{1 - \rho_i^{\star\top} z_{ij}}{\omega_i - \rho_i^{\star\top} \mu_i} v_j, \quad \rho_i^\star = \Sigma_i^{-1} \mu_i$$

$$\text{where} \quad \omega_i = \sum_{j \le i} w_{ij}, \quad \mu_i = \sum_{j \le i} w_{ij} z_{ij}, \quad \Sigma_i = \sum_{j \le i} w_{ij} z_{ij} z_{ij}^\top + \lambda I$$

Local Linear Attention

$$o_i^{\mathsf{LLA}} = \sum_{j \le i} w_{ij} \frac{1 - \rho_i^{\star\top} z_{ij}}{\omega_i - \rho_i^{\star\top} \mu_i} v_j, \quad \rho_i^{\star} = \Sigma_i^{-1} \mu_i$$

$$o_i^{\mathsf{LLA}} \xrightarrow{\lambda \to \infty} o_i^{\mathsf{SA}} = \sum_{j \le i} \mathrm{softmax}(q_i^\top k_j / h)\, v_j,$$

$$o_i^{\mathsf{LLA}} \xrightarrow[\text{drop bias}]{h \to \infty} o_i^{\mathsf{Mesa}} = \left(\sum_{j \le i} v_j k_j^\top \right) \left(\sum_{j \le i} k_j k_j^\top + \lambda I \right)^{-1} q_i.$$

Local Linear Attention

$$o_i^{\mathsf{LLA}} = \sum_{j \le i} w_{ij} \frac{1 - \rho_i^{\star\top} z_{ij}}{\omega_i - \rho_i^{\star\top} \mu_i} v_j, \qquad \rho_i^{\star} = \Sigma_i^{-1} \mu_i$$

To solve the linear system, LLA introduces a parallel conjugate gradient solver.

- High compute and I/O overhead
- A regularization–expressiveness trade-off
- Incompatibility with low-precision arithmetic

Local Linear Attention

$$o_i^{\mathsf{LLA}} = \sum_{j \le i} w_{ij} \frac{1 - \rho_i^{\star\top} z_{ij}}{\omega_i - \rho_i^{\star\top} \mu_i} v_j, \quad \rho_i^\star = \Sigma_i^{-1} \mu_i$$

$$o_i^{\mathsf{LLA}} = \sum_{j \le i} \frac{w_{ij}(1 - t_{ij})}{\sum_{j \le i} w_{ij}(1 - t_{ij})} v_j = o_i^{\mathsf{SA}} - (1 + \eta_i) \Sigma_{KV}^{(i)} \rho_i^\star,$$

$$\rho_i^\star = \Sigma_i^{-1} \mu_i, \quad t_{ij} = \rho_i^{\star\top} z_{ij}$$

Parameterized Local Linear Attention

$$o_i^{\mathsf{LLA}} = \sum_{j \le i} \frac{w_{ij}(1 - t_{ij})}{\sum_{j \le i} w_{ij}(1 - t_{ij})} v_j = o_i^{\mathsf{SA}} - (1 + \eta_i)\Sigma_{KV}^{(i)} \rho_i^{\star},$$

$$\rho_i^{\star} = \Sigma_i^{-1} \mu_i, \quad t_{ij} = \rho_i^{\star \top} z_{ij}$$

$$o_i^{\mathsf{PLX}} = \sum_{j \le i} \frac{w_{ij}\,(1 - t_{ij} + \bar{t}_i)}{\sum_{j \le i} w_{ij}\,(1 - t_{ij} + \bar{t}_i)}\, v_j = o_i^{\mathsf{SA}} - \Sigma_{KV}^{(i)} \rho_i,$$

$$\rho_i = W_R x_i, \quad t_{ij} = \rho_i^{\top} z_{ij}.$$

Parameterized Local Linear Attention

$$o_i^{\mathsf{LLA}} = \sum_{j \le i} \frac{w_{ij}(1 - t_{ij})}{\sum_{j \le i} w_{ij}(1 - t_{ij})} v_j = o_i^{\mathsf{SA}} - (1 + \eta_i)\Sigma_{KV}^{(i)} \rho_i^{\star},$$

$$\boxed{\rho_i^{\star} = \Sigma_i^{-1} \mu_i,} \quad t_{ij} = \rho_i^{\star\top} z_{ij}$$

$$o_i^{\mathsf{PLX}} = \sum_{j \le i} \frac{w_{ij}\,(1 - t_{ij} + \bar{t}_i)}{\sum_{j \le i} w_{ij}\,(1 - t_{ij} + \bar{t}_i)}\, v_j = o_i^{\mathsf{SA}} - \Sigma_{KV}^{(i)} \rho_i,$$

$$\boxed{\rho_i = W_R x_i,} \quad t_{ij} = \rho_i^{\top} z_{ij}.$$

Parameterized Local Linear Attention

$$o_i^{\mathsf{LLA}} = \sum_{j \le i} \frac{w_{ij}(1 - t_{ij})}{\sum_{j \le i} w_{ij}(1 - t_{ij})} v_j = o_i^{\mathsf{SA}} - (1 + \eta_i)\Sigma_{KV}^{(i)} \rho_i^{\star},$$

$$\rho_i^{\star} = \Sigma_i^{-1} \mu_i, \quad t_{ij} = \rho_i^{\star \top} z_{ij}$$

$$o_i^{\mathsf{PLX}} = \sum_{j \le i} \frac{w_{ij}\,(1 - t_{ij} + \bar{t}_i)}{\sum_{j \le i} w_{ij}\,(1 - t_{ij} + \bar{t}_i)}\, v_j = o_i^{\mathsf{SA}} - \Sigma_{KV}^{(i)} \rho_i,$$

$$\rho_i = W_R x_i, \quad t_{ij} = \rho_i^{\top} z_{ij}.$$

Parameterized Local Linear Attention

$$o_i^{\mathsf{LLA}} = \sum_{j \le i} \frac{w_{ij}(1 - t_{ij})}{\sum_{j \le i} w_{ij}(1 - t_{ij})} v_j = o_i^{\mathsf{SA}} - (1 + \eta_i)\Sigma_{KV}^{(i)} \rho_i^\star,$$

$$\rho_i^\star = \Sigma_i^{-1} \mu_i, \quad t_{ij} = \rho_i^{\star\top} z_{ij}$$

$$o_i^{\mathsf{PLX}} = \sum_{j \le i} \frac{w_{ij}\,(1 - t_{ij} + \bar{t}_i)}{\sum_{j \le i} w_{ij}}\, v_j = o_i^{\mathsf{SA}} - \Sigma_{KV}^{(i)} \rho_i,$$

$$\rho_i = W_R x_i, \quad t_{ij} = \rho_i^\top z_{ij}.$$

Parameterized Local Linear Attention

$$o_i^{\mathsf{PLX}} = \sum_{j \le i} \frac{w_{ij}\,(1 - t_{ij} + \bar{t}_i)}{\sum_{j \le i} w_{ij}}\, v_j = o_i^{\mathsf{SA}} - \Sigma_{KV}^{(i)} \rho_i,$$

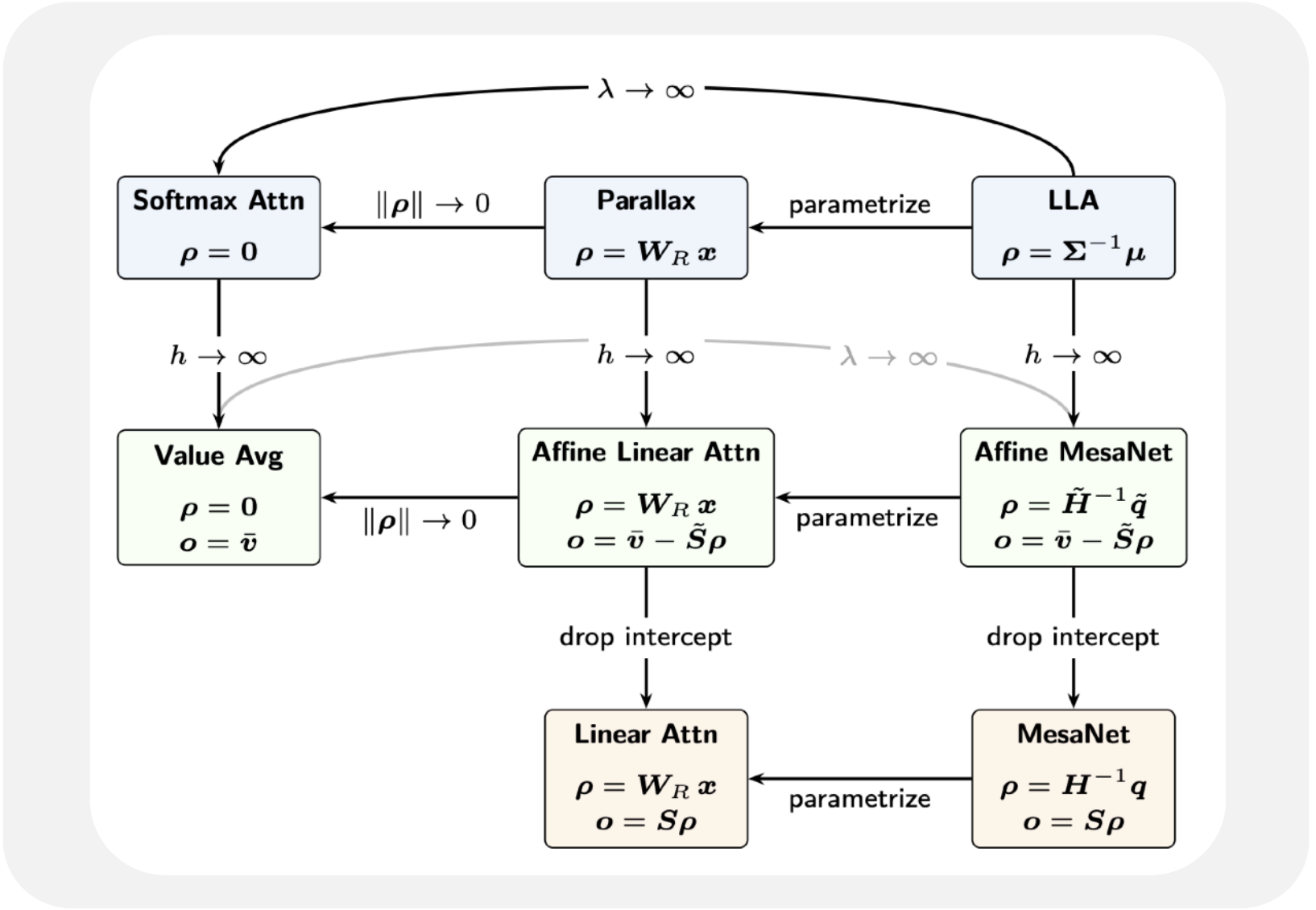
$$\rho_i = W_R x_i, \quad t_{ij} = \rho_i^\top z_{ij}.$$

$$o_i^{\mathsf{PLX}} \xrightarrow{h \to \infty} o_i^{\mathsf{Affine\ LA}} = \bar{v}_i - \tilde{S}_i \rho_i, \quad \rho_i = W_R x_i,$$

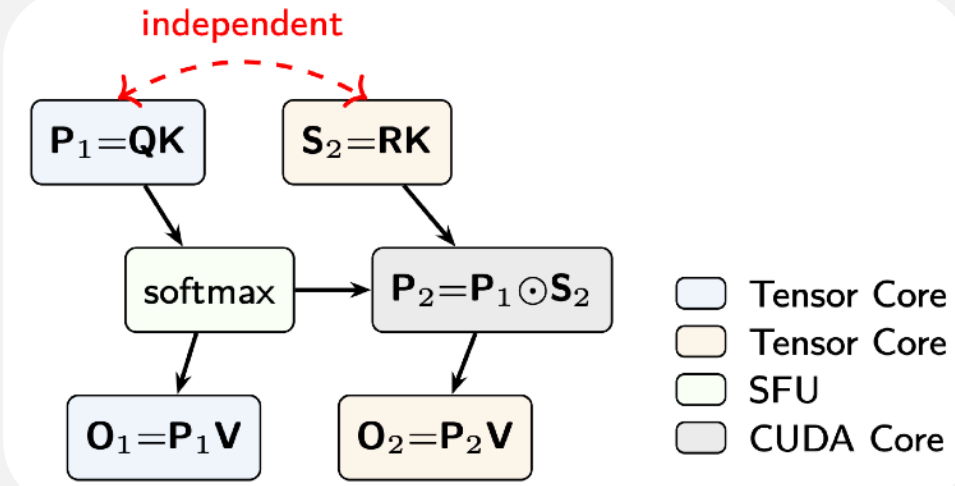
$$o_i^{\mathsf{PLX}} \xrightarrow[\text{drop bias}]{h \to \infty} o_i^{\mathsf{LA}} = S_i \rho_i, \quad \rho_i = W_R x_i$$

In LA, the query corresponds to the probe in Parallax.

Parameterized Local Linear Attention



Algorithm



Parallax increases per-head attention FLOPs without increasing I/O.

Algorithm 1 Parallax forward core computation.

Require: Input $\mathbf{Q}_r, \mathbf{R}_r, \mathbf{K}, \mathbf{V}, L, s$; Output $\mathbf{O}_r$; Running state $(\mathbf{m}_r, \mathbf{d}_{1,2}, \mathbf{O}_{1,2})$.

for $c = 1$ to $\lceil L/\mathcal{B}_c \rceil$ **do**
  $\mathbf{S}_1 \leftarrow \mathbf{Q}_r \mathbf{K}_c^\top \cdot s$ ▷ Apply Masking
  $\mathbf{m} \leftarrow \max(\mathbf{m}_r, \text{rowmax}(\mathbf{S}_1))$
  $\boldsymbol{\alpha} \leftarrow \exp2(\mathbf{m}_r - \mathbf{m})$
  $\mathbf{P}_1 \leftarrow \exp2(\mathbf{S}_1 - \mathbf{m})$
  $\mathbf{m}_r \leftarrow \mathbf{m}$
  $\mathbf{S}_2 \leftarrow \mathbf{R}_r \mathbf{K}_c^\top$ ▷ GEMM in TC
  $\mathbf{P}_2 \leftarrow \mathbf{P}_1 \odot \mathbf{S}_2$
  $\mathbf{d}_1 \leftarrow \boldsymbol{\alpha}\mathbf{d}_1 + \text{rowsum}(\mathbf{P}_1)$
  $\mathbf{d}_2 \leftarrow \boldsymbol{\alpha}\mathbf{d}_2 + \text{rowsum}(\mathbf{P}_2)$
  $\mathbf{O}_1 \leftarrow \boldsymbol{\alpha}\mathbf{O}_1 + \mathbf{P}_1 \mathbf{V}_c$
  $\mathbf{O}_2 \leftarrow \boldsymbol{\alpha}\mathbf{O}_2 + \mathbf{P}_2 \mathbf{V}_c$ ▷ GEMM in TC
end for
$\mathbf{O}_r \leftarrow \mathbf{O}_1/\mathbf{d}_1 \cdot (1 + \mathbf{d}_2/\mathbf{d}_1) - \mathbf{O}_2/\mathbf{d}_1$

Future Directions

- Gather more empirical evidence on arch–optimizer interactions, develop a theoretical understanding of AdamW degeneracy, and investigate whether architectural remedies exist.
- Post-train a Parallax model initialized from a standard attention checkpoint.
- Derive a nonparametric counterpart to DeltaNet, extending the connection between Parallax and Linear Attention.
- Scale Parallax to larger models and datasets.…

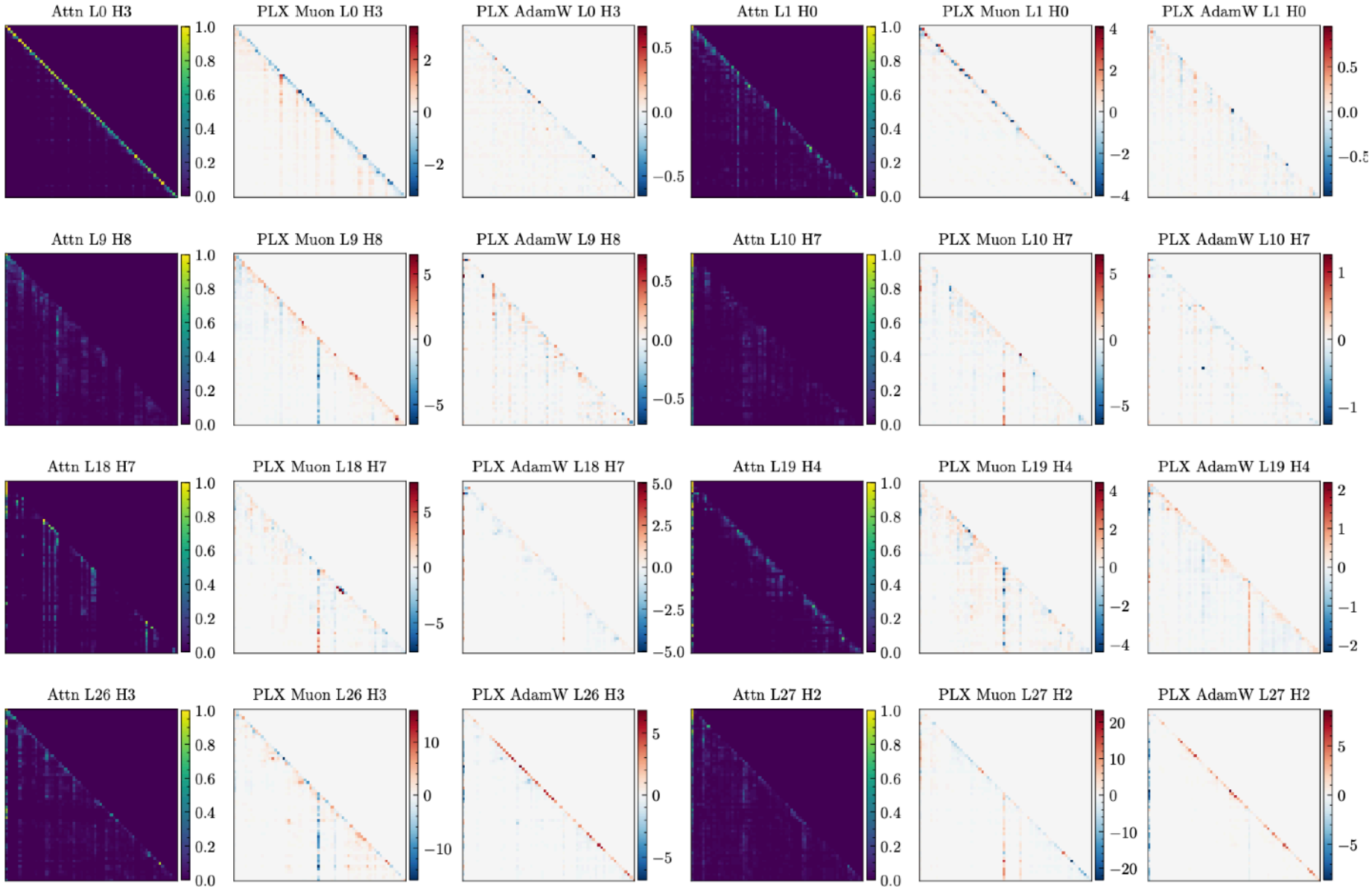
Thanks!

Boundary Amplification Factor

Proposition 3.1 (Boundary amplification is non-negative). *Denote* $\bar{z}_i = \mathbb{E}_{\boldsymbol{p}_i}[\boldsymbol{z}_{ij}]$ *and* $\boldsymbol{A}_i = \omega_i \mathrm{Var}_{\boldsymbol{p}_i}(\boldsymbol{z}_{ij}) + \lambda \boldsymbol{I} \succ 0$. *If* $\boldsymbol{\rho}_i^\star = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i$, *then* $\bar{t}_i = \mathbb{E}_{\boldsymbol{p}_i}[\boldsymbol{\rho}_i^{\star\top}\boldsymbol{z}_{ij}] \in [0,1)$, $\eta_i = \omega_i \bar{\boldsymbol{z}}_i^\top \boldsymbol{A}_i^{-1}\bar{\boldsymbol{z}}_i \geq 0$ *where* $\bar{z}_i = \mathbb{E}_{\boldsymbol{p}_i}[\boldsymbol{z}_{ij}]$. *The proof is provided in Appendix B.*

- If it is close to 0, the query is close to the weighted key center and the correction becomes pure covariance
- If its value is large, the query is close to the weighted key boundary and the correction is amplified to compensate for the boundary bias.

Visualization



Visualization