SAMformer: Unlocking the Potential of Transformers in Time Series Forecasting

Romain Ilbert[*12] **Ambroise Odonnat**[*1] **Vasilii Feofanov**[1] **Aladin Virmaux**[1] **Giuseppe Paolo**[1] **Themis Palpanas**[2] **Ievgen Redko**[1]

*Equal contribution [1]Huawei Noah's Ark Lab [2]LIPADE, Paris Descartes University

ORAL

ICML International Conference On Machine Learning

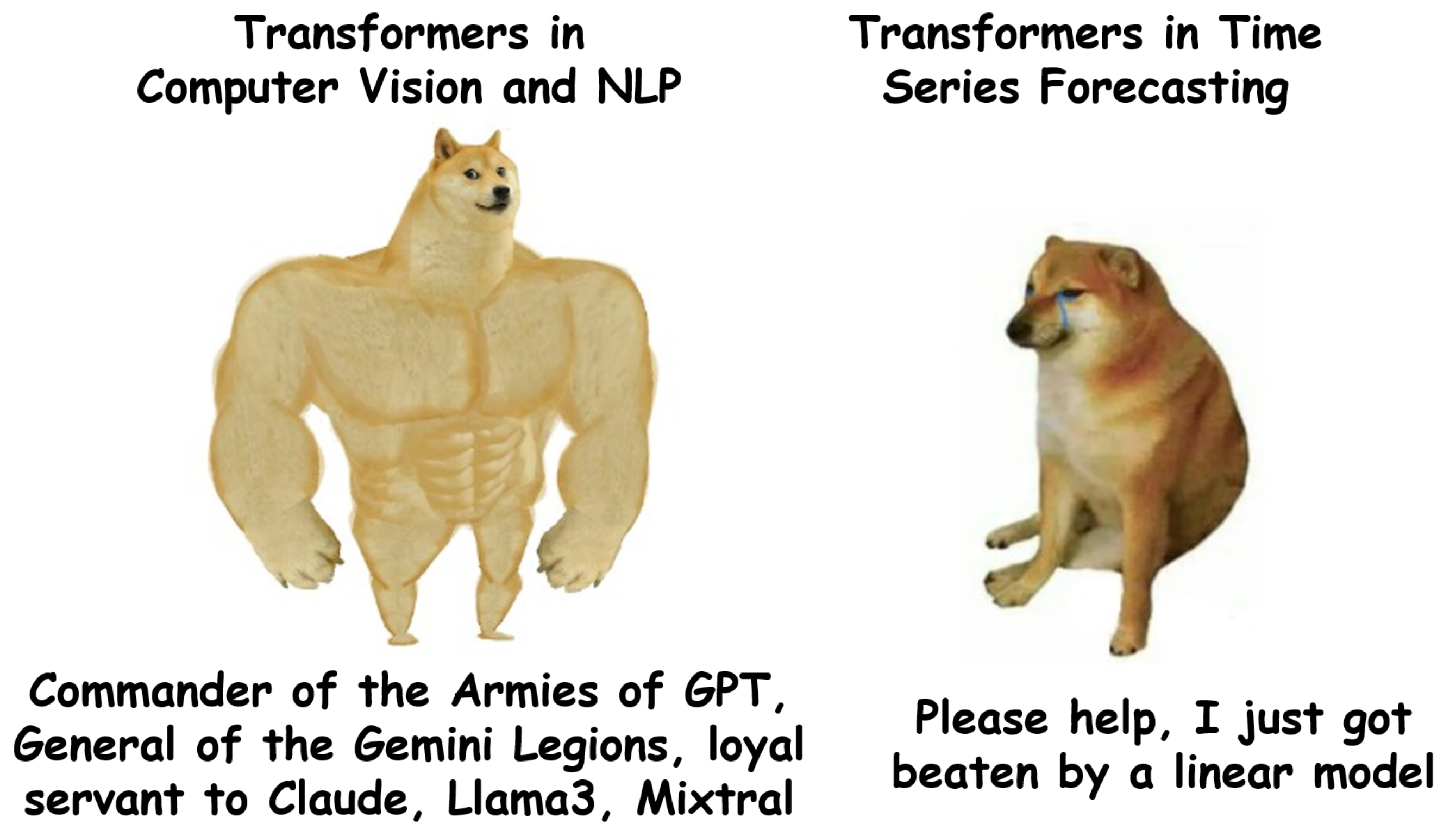
TL;DR

- Transformers show **mixed performance** in time series forecasting
- Attention is at fault for leading to a **sharp loss landscape**
- We propose SAMformer, a **shallow lightweight transformer** model
- It combines **Sharpness-Aware Minimization** (SAM) and **channel-wise attention**
- Benefits: **lightest** and **SOTA** model, **robustness**, improved **signal propagation**
- It even **surpasses MOIRAI**, the biggest open-source foundation model

Problem Setup

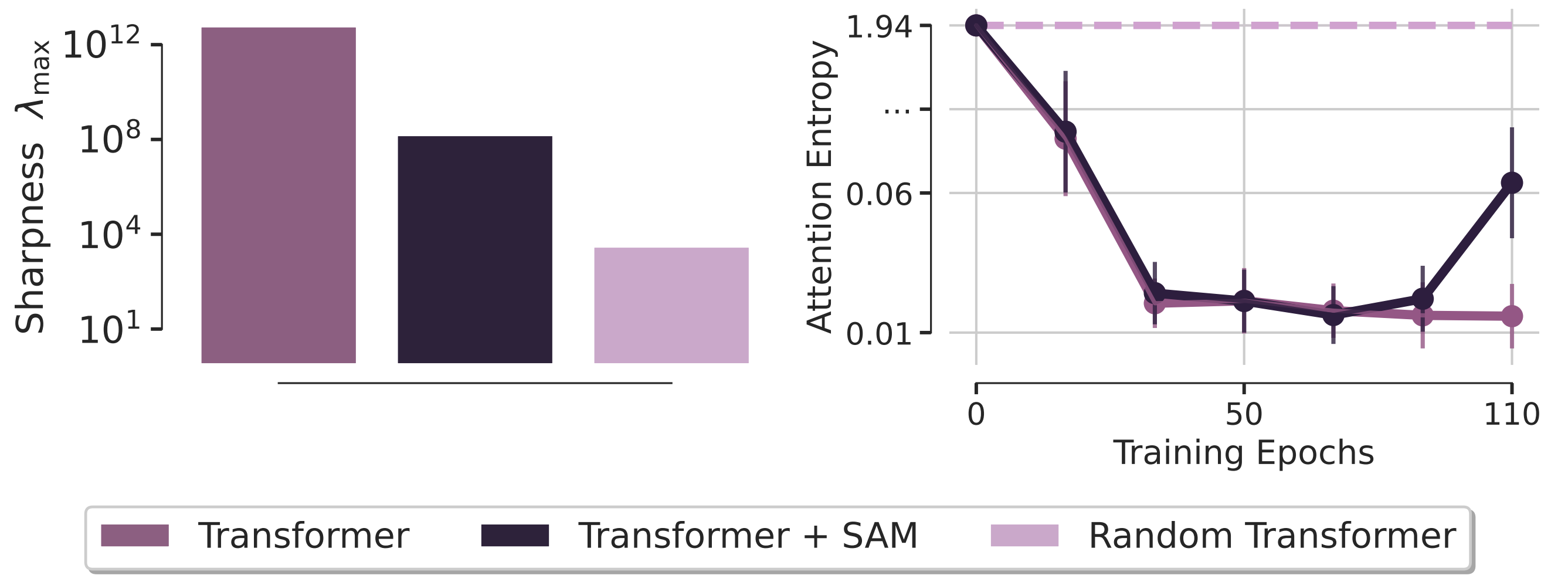
Goal: given a D-dimensional time series of length L, predict its next H values.

- Input $\mathbf{X} \in \mathbb{R}^{D\times L}$, target $\mathbf{Y} \in \mathbb{R}^{D\times H}$,
- Training set of N observations $(\{\mathbf{X}^{(i)}\}_{i=0}^{N}, \{\mathbf{Y}^{(i)}\}_{i=0}^{N})$,
- Train a predictor $f_\omega \colon \mathbb{R}^{D\times L} \to \mathbb{R}^{D\times H}$ that minimizes the MSE loss.


Transformers in Computer Vision and NLP
Transformers in Time Series Forecasting
Commander of the Armies of GPT, General of the Gemini Legions, loyal servant to Claude, Llama3, Mixtral
Please help, I just got beaten by a linear model

Trainability Issues due to the Attention

- Generate toy data according to $\mathbf{Y} = \mathbf{X}\mathbf{W}_{\text{toy}} + \boldsymbol{\varepsilon}$,
- Shallow one-layer transformer $f(\mathbf{X}) = [\mathbf{X} + \mathbf{A}(\mathbf{X})\mathbf{X}\mathbf{W}_V\mathbf{W}_O]\mathbf{W}$,
- Channel-wise attention $\mathbf{A}(\mathbf{X}) = \texttt{softmax}\left(\frac{\mathbf{X}\mathbf{W}_\mathbf{Q}\mathbf{W}_\mathbf{K}^\top\mathbf{X}^\top}{\sqrt{d_\mathrm{m}}}\right)$.

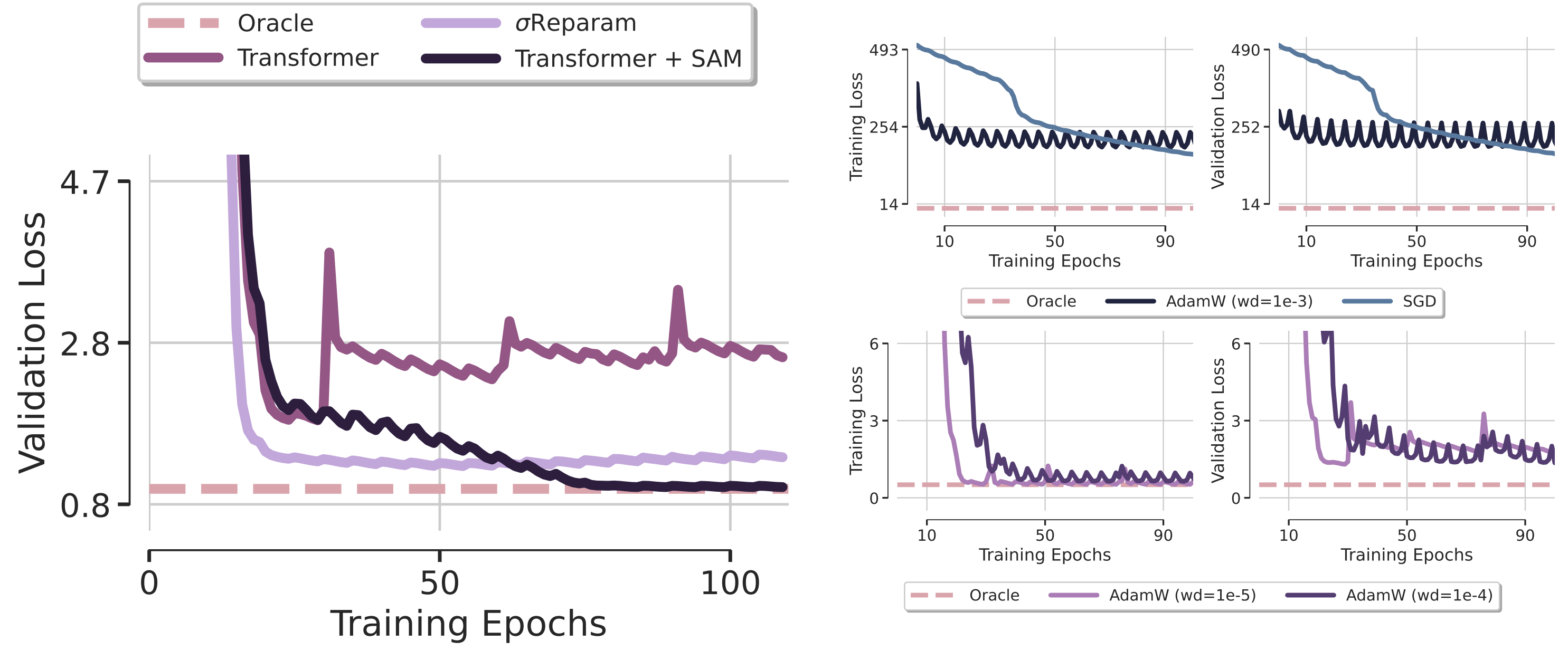


Training the attention induces an entropy collapse and a sharp loss landscape.

σReparam doesn't solve the problem, but SAM does

In theory, Transformer **can** find the optimal solution but **doesn't** in practice. Potential solutions:

- **σReparam**: replace each weight matrix $\mathbf{W}$ by $\widehat{\mathbf{W}} = \frac{\gamma}{\|\mathbf{W}\|_2}\mathbf{W}$ where $\gamma \in \mathbb{R}$ learnable,
- **Sharpness-Aware Minimization (SAM)**: replace $\mathcal{L}_{\text{train}}$ by $\mathcal{L}_{\text{train}}^{\text{SAM}}(\boldsymbol{\omega}) = \max_{\|\boldsymbol{\varepsilon}\|<\rho} \mathcal{L}_{\text{train}}(\boldsymbol{\omega} + \boldsymbol{\varepsilon})$.



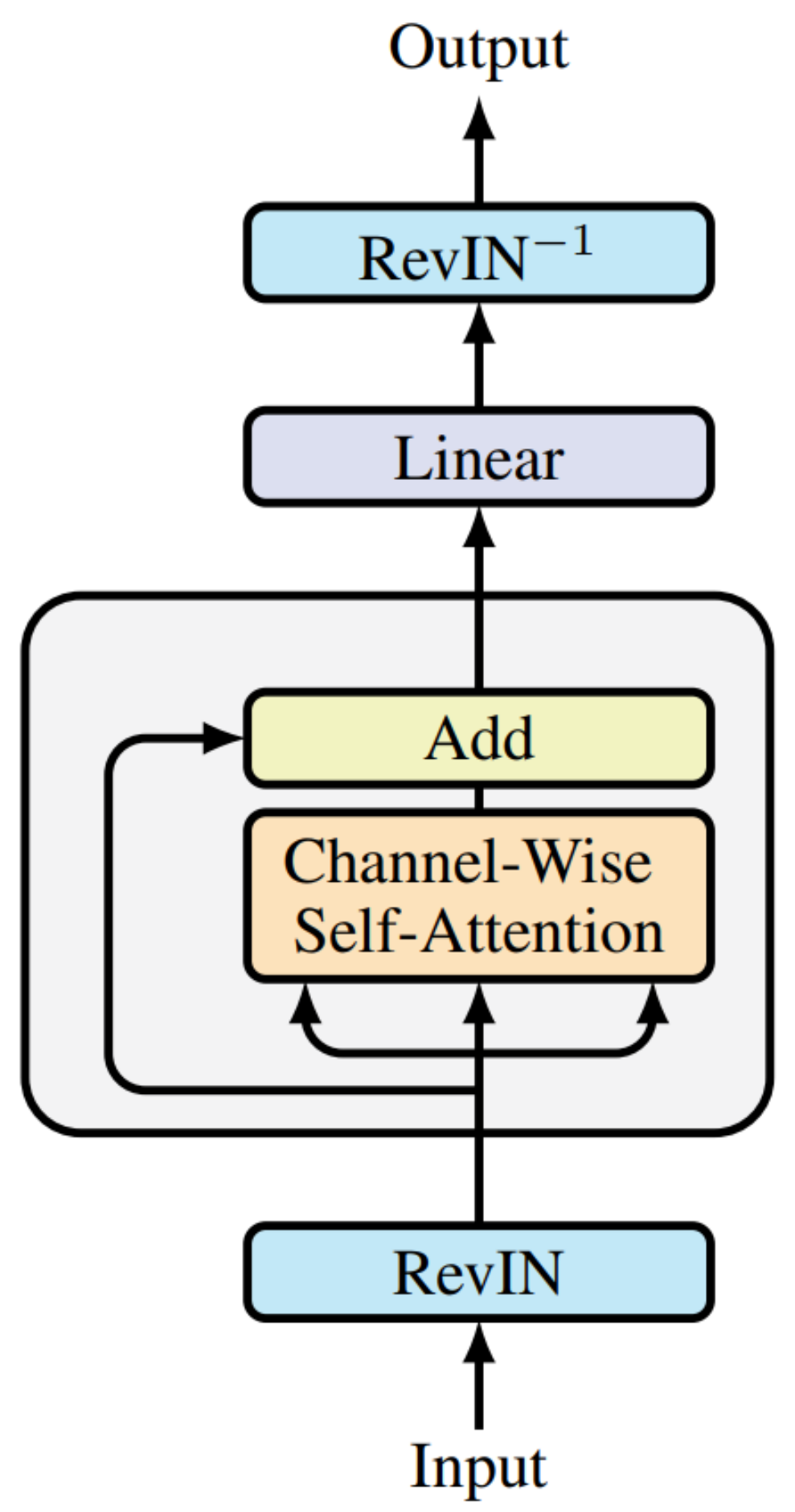
Poor generalization with SGD, Adam, and AdamW. Reducing the entropy collapse with σReparam is not sufficient but tackling the sharpness with SAM leads to the optimal solution (oracle).

SAMformer: Combining SAM and Channel-Wise Attention

- Input $\mathbf{X} \in \mathbb{R}^{D\times L}$, output $f(\mathbf{X}) \in \mathbb{R}^{D\times H}$,
- Reduce distribution shift with **RevIN**,
- **Channel-wise attention** $\mathbf{A}(\mathbf{X}) \in \mathbb{R}^{D\times D}$,
- Training with **SAM** for **smoother** loss landscape.

$$\mathbf{A}(\mathbf{X}) = \texttt{softmax}\left(\frac{\mathbf{X}\mathbf{W}_\mathbf{Q}\mathbf{W}_\mathbf{K}^\top\mathbf{X}^\top}{\sqrt{d_\mathrm{m}}}\right)$$

$$f(\mathbf{X}) = [\mathbf{X} + \mathbf{A}(\mathbf{X})\mathbf{X}\mathbf{W}_V\mathbf{W}_O]\mathbf{W}$$



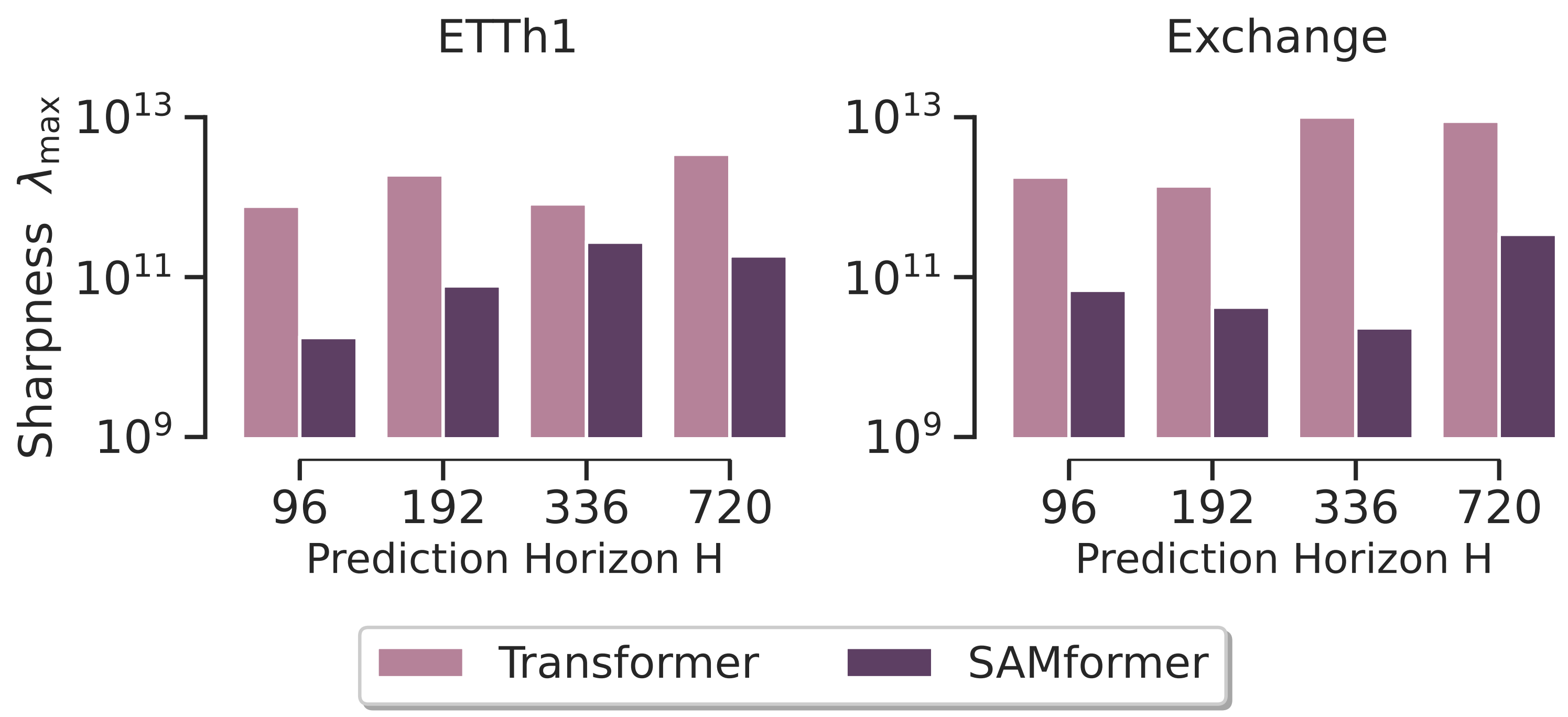
SAMformer is a shallow transformer trained with SAM.
→ One head, one encoder, SOTA performance!

Experimental Results: Easier, Better, Faster, Smoother

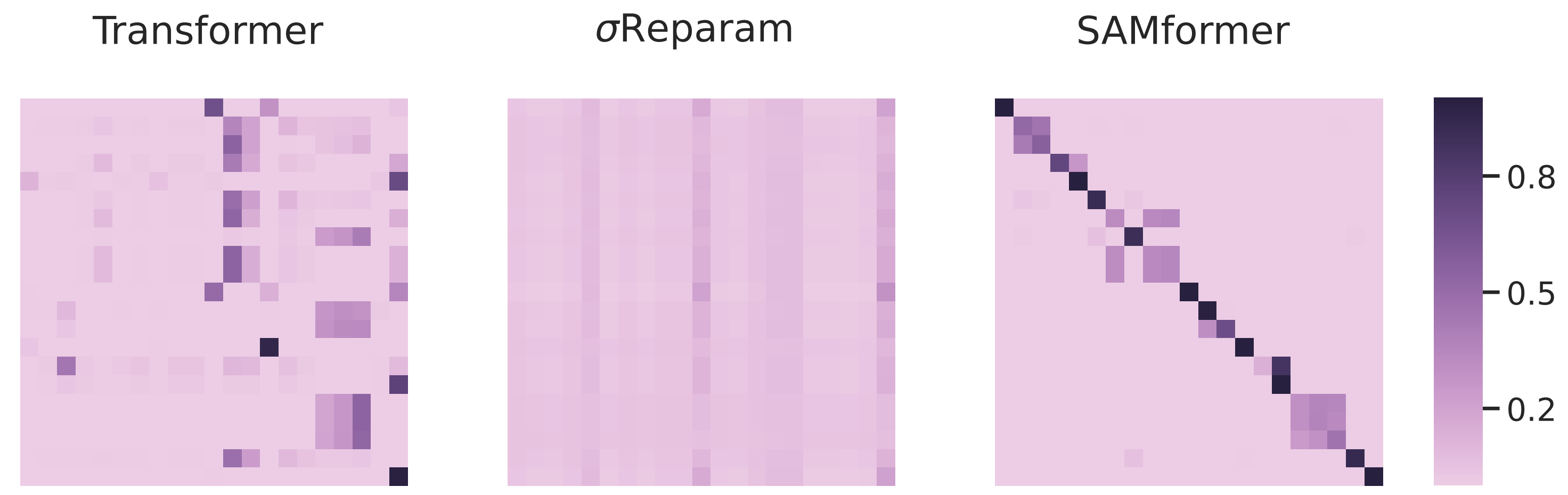
- Comparison of SAMformer to SOTA transformer-based and all-MLP models with the test MSE,
- Extensive evaluation conducted on common open-source benchmarks of various scales.

Dataset	SAMformer	`iTransformer`	`PatchTST`	`TSMixer`	`FEDformer`	`Autoformer`
	-	2024	2023	2023	2022	2021
ETTh1	**0.410**	0.454	0.469	0.437	0.440	0.496
ETTh2	**0.344**	0.383	0.387	0.357	0.437	0.450
ETTm1	**0.373**	0.407	0.387	0.385	0.448	0.588
ETTm2	**0.269**	0.288	0.281	0.289	0.305	0.327
Traffic	**0.425**	0.428	0.481	0.620	0.610	0.628
Weather	0.260	**0.258**	0.259	0.267	0.309	0.338
Overall improvement		6.58%	8.79%	13.2%	22.5%	35.9%

SAMformer outperforms all baselines while having significantly fewer parameters.



SAM leads to a smoother loss landscape and better generalization.



Channel-wise attention improves the signal propagation.

Main References

- **Chen et al.** - TMLR 2023
  TSMixer: An all-MLP architecture
- **Zhai et al.** - ICML 2023
  σReparam: Stabilizing transformer
- **Ilbert et al.** - ICML 2024 (this work)
  SAMformer: Unlocking the potential of transformers in time series forecasting

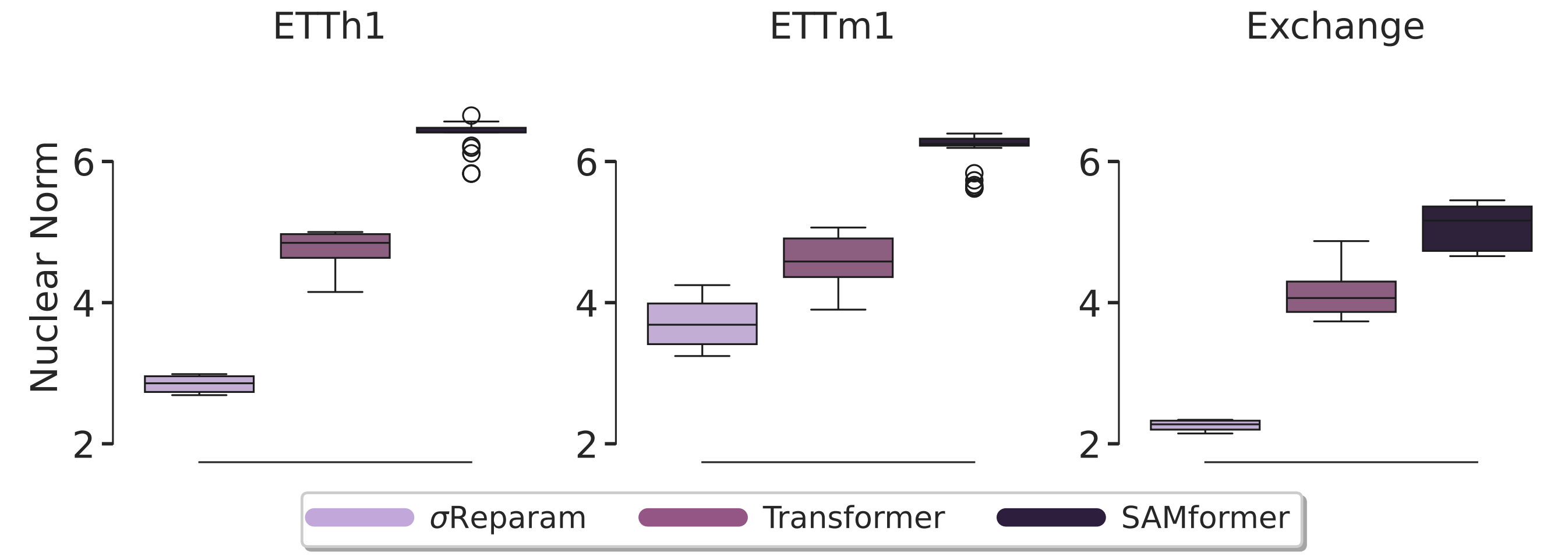
Romain Ilbert

Ambroise Odonnat

Intuition Behind the Failure of σReparam

We prove that σReparam may induce a **rank collapse** of the **attention matrix**

$$\|\mathbf{X}\mathbf{W}_Q\mathbf{W}_K^\top\mathbf{X}^\top\|_* \quad \leq \quad \underbrace{\|\mathbf{W}_Q\mathbf{W}_K^\top\|_2}_{\text{goes to } 0 \text{ with } \sigma\text{Reparam}} \quad \|\mathbf{X}\|_\mathrm{F}^2.$$



Comparison with MOIRAI

- `MOIRAI` is the biggest foundation model trained on $\mathbf{27B}$ **samples**,
- `MOIRAI` comes in three sizes: small ($\mathbf{14M}$), base ($\mathbf{91M}$) and large ($\mathbf{314M}$).

Dataset	Full-shot	Zero-shot		
	SAMformer	`MOIRAI`$_{\texttt{Small}}$	`MOIRAI`$_{\texttt{Base}}$	`MOIRAI`$_{\texttt{Large}}$
ETTh1	<u>0.410</u>	**0.400**	0.434	0.510
ETTh2	<u>0.344</u>	**0.341**	0.345	0.354
ETTm1	**0.373**	0.448	<u>0.381</u>	0.390
ETTm2	**0.269**	0.300	<u>0.272</u>	0.276
Electricity	**0.181**	0.233	<u>0.188</u>	<u>0.188</u>
Weather	0.260	<u>0.242</u>	**0.238**	0.259
Overall MSE improvement		6.9%	1.1%	7.6%

SAMformer outperforms `MOIRAI` while having significantly fewer parameters!

Take Home Message

Transformers are hard to train and perform poorly in time series forecasting.
→ Start using SAMformer for **better** performance at **lower** cost!

Want to Know More?

paper

code

ICML 2024, Vienna, Austria

Correspondence to romain.ilbert@hotmail.fr and ambroiseodonnattechnologie@gmail.com