ICML
International Conference
On Machine Learning

Robust Filter Attention: Self-attention as Precision-weighted State Estimation

Peter Racioppo
Independent Researcher
pcracioppo@gmail.com

Highlights

- Tokens are modeled as noisy observations of latent dynamics.
- Attention arises from batch state estimation.
- RFA recovers RoPE and ALiBi as limiting cases and improves long-context extrapolation.

Method	Transport	Reliability
RoPE	Rotation	Uniform
ALiBi	None	Distance bias
RFA	Dynamical transport	Propagated Precision

- We treat tokens as noisy measurements of a latent dynamical system:

$$d\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{x}(t)dt + \boldsymbol{G}d\boldsymbol{w}(t)$$

$$\boldsymbol{z}_i = \boldsymbol{C}\boldsymbol{x}(t_i) + \boldsymbol{v}_i,\ \boldsymbol{v}_i \sim N(0, \boldsymbol{R})$$

- Linear dynamics allow for closed-form state and covariance propagation.
- Each key $\boldsymbol{z}_j$ forms an estimate of the latent state at time t_i: $\hat{\boldsymbol{z}}_{ij} = e^{\boldsymbol{A}\,\Delta t_{ij}}\,\boldsymbol{z}_j$.
- The query is a reference measurement $\boldsymbol{z}_i \sim N(\boldsymbol{x}_i, \boldsymbol{R}_\Gamma)$.
- Query-key agreement is measured by the residual: $\boldsymbol{r}_{ij} = \boldsymbol{z}_i - \hat{\boldsymbol{z}}_{ij}$.

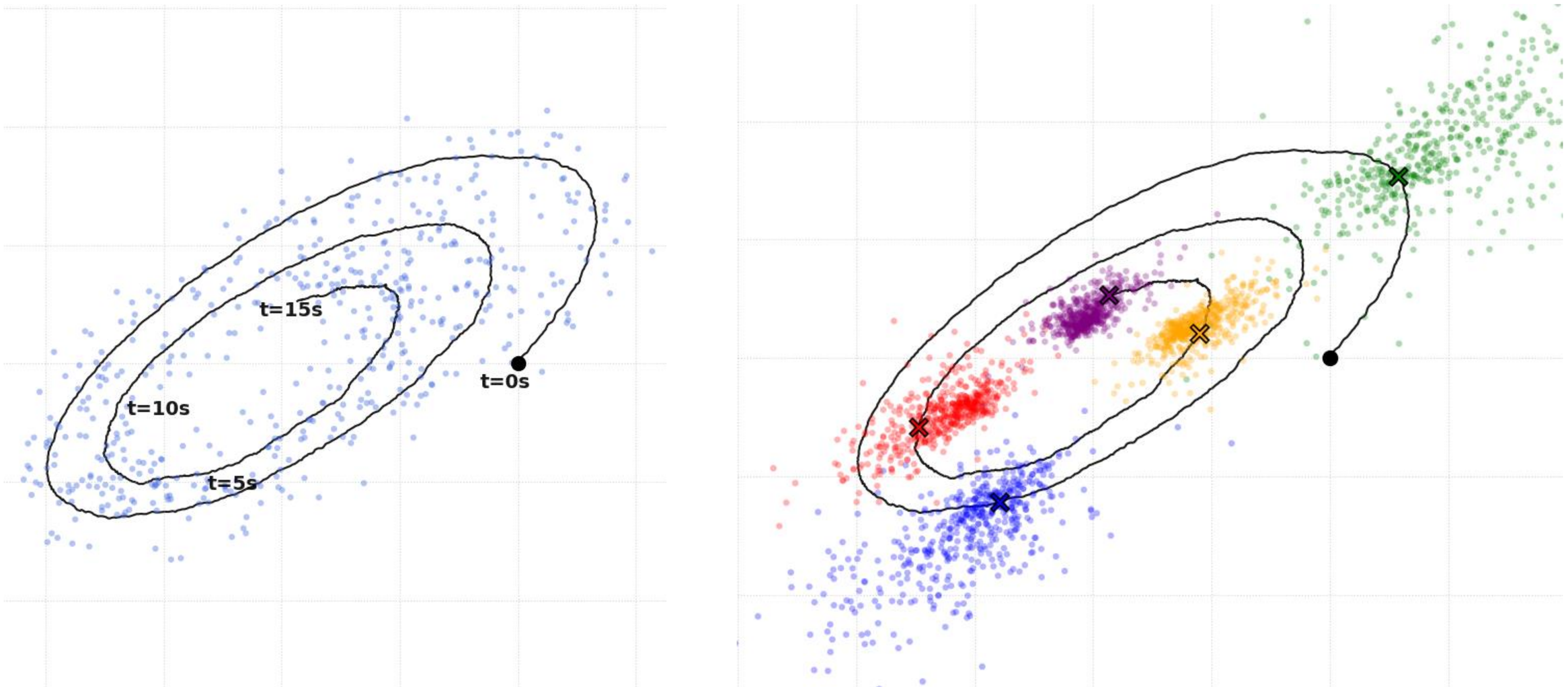

Figure 1: Tokens are modeled as noisy observations of a latent dynamical system. Each key is transported through the dynamics and combined in a precision weighted average.

Self-attention from Batch State Estimation

- $\boldsymbol{r}_{ij}$ has covariance $\boldsymbol{\Sigma}_{ij} = \underbrace{\boldsymbol{V}_{ij}}_{\text{process noise}} + \underbrace{\boldsymbol{e}^{\boldsymbol{A}\Delta t_{ij}}\boldsymbol{R}\boldsymbol{e}^{\boldsymbol{A}^{\top}\Delta t_{ij}}}_{\text{key measurement noise}} + \underbrace{\boldsymbol{R}_{\Gamma}}_{\text{query measurement noise}}$, and precision $\boldsymbol{P}_{ij} = \boldsymbol{\Sigma}_{ij}^{-1}$.
- Solving a separate estimation problem for each query gives the batch estimator:

$$\bar{\boldsymbol{z}}_i = \left(\sum_{j \le i} w_{ij}\boldsymbol{P}_{ij}\right)^{-1} \sum_{j \le i} w_{ij}\boldsymbol{P}_{ij}\hat{\boldsymbol{z}}_{ij}$$

- We reweight the precision priors by robust weights $w_{ij} = \exp(-\boldsymbol{r}_{ij}^{\top}\boldsymbol{P}_{ij}\boldsymbol{r}_{ij})$. These give rise to attention weights.
- To obtain a tractable estimator, we assume that the system matrices are *diagonalizable*, $\boldsymbol{A} = \boldsymbol{S}\,\boldsymbol{\Lambda}\,\boldsymbol{S}^{-1}$, and the decay and noise are *isotropic.*

The RFA Mechanism

- The QKV and output projections absorb the change of basis matrices $\boldsymbol{S}$ and $\boldsymbol{S}^{-1}$.
- Define forward/backward rotation kernels and a decay kernel:

$$\widetilde{\boldsymbol{\Phi}}^{-}[k,i] \coloneqq e^{-i\omega_k t_i},\ \widetilde{\boldsymbol{\Phi}}^{+}[k,i] \coloneqq e^{i\omega_k t_i},\ \boldsymbol{E}[i,j] \coloneqq e^{-\mu \Delta t_{ij}}.$$

- Define rotated queries, keys, and values:

$$\widetilde{\boldsymbol{Q}} \coloneqq \widetilde{\boldsymbol{\Phi}}^{-} \odot \boldsymbol{Q}, \quad \widetilde{\boldsymbol{K}} \coloneqq \widetilde{\boldsymbol{\Phi}}^{-} \odot \boldsymbol{K}, \quad \widetilde{\boldsymbol{V}} \coloneqq \widetilde{\boldsymbol{\Phi}}^{-} \odot \boldsymbol{V}.$$

- Residuals:

$$\left\|\boldsymbol{R}_{qk}[i,j]\right\|^2 = \|\boldsymbol{Q}_i\|^2 + \boldsymbol{E}[i,j]^2 \cdot \left\|\boldsymbol{K}_j\right\|^2 - 2\,\boldsymbol{E}[i,j] \cdot \mathrm{Re}\left(\widetilde{\boldsymbol{Q}}_i^{\dagger} \widetilde{\boldsymbol{K}}_j\right).$$

- The attention logits are: $\boldsymbol{L} = \log(\boldsymbol{P}_{\Delta t}) - \boldsymbol{P}_{\Delta t} \odot \left\|\boldsymbol{R}_{qk}\right\|^2$.
- After aggregating, we counter-rotate the values:

$$\bar{\boldsymbol{V}} = \widetilde{\boldsymbol{\Phi}}^{+} \odot \left(\widetilde{\boldsymbol{V}}\,(\boldsymbol{A} \odot \boldsymbol{E})^{\top}\right).$$

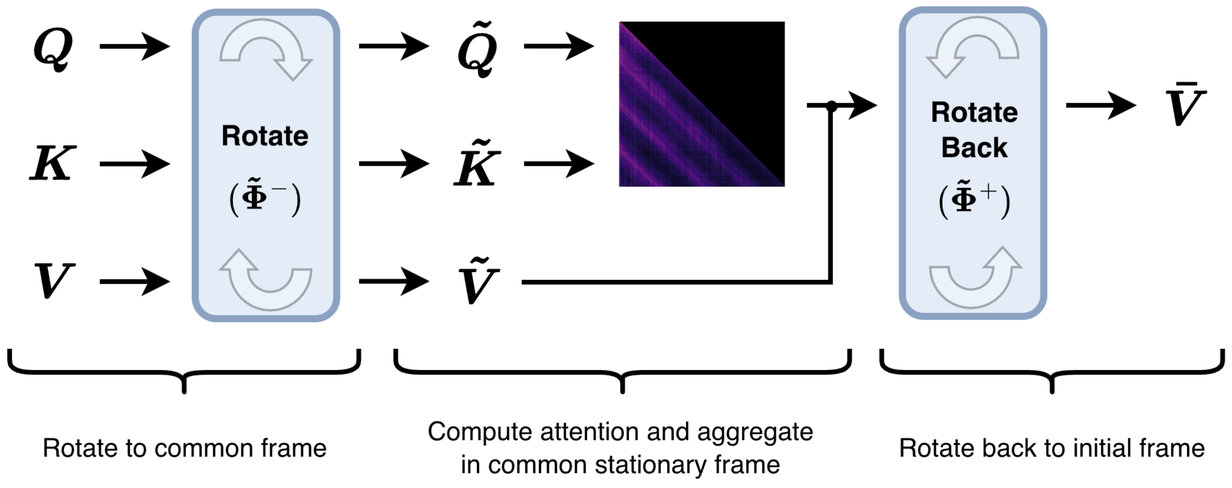

Figure 2. Queries, keys, and values are rotated into a common frame before applying attention. The attention output is then rotated back to the initial frame.

Filtering Regimes

- RFA attention heads learn distinct temporal filtering regimes through the balance between process noise and measurement noise.
- **Diffusive Regime:** Process noise dominates, causing uncertainty to grow with temporal distance, producing a recency bias similar to ALiBi.
- **Integrative Regime:** Measurement noise dominates, so transported states become more reliable after transient noise decays. Attention is initially suppressed near the diagonal and peaks at characteristic temporal lags.

Spectral Coupling

- RoPE allows high-frequency oscillations to persist indefinitely, leading to phase wrap around and poor extrapolation.
- Spectrally-Coupled RFA (SC-RFA) couples decay rate to the max rotational frequency per head $\mu_h = b \cdot \omega_{h,\max}$.
- High-frequency heads decay rapidly and specialize to local structure, while low-frequency heads persist as stable long-range integrators.

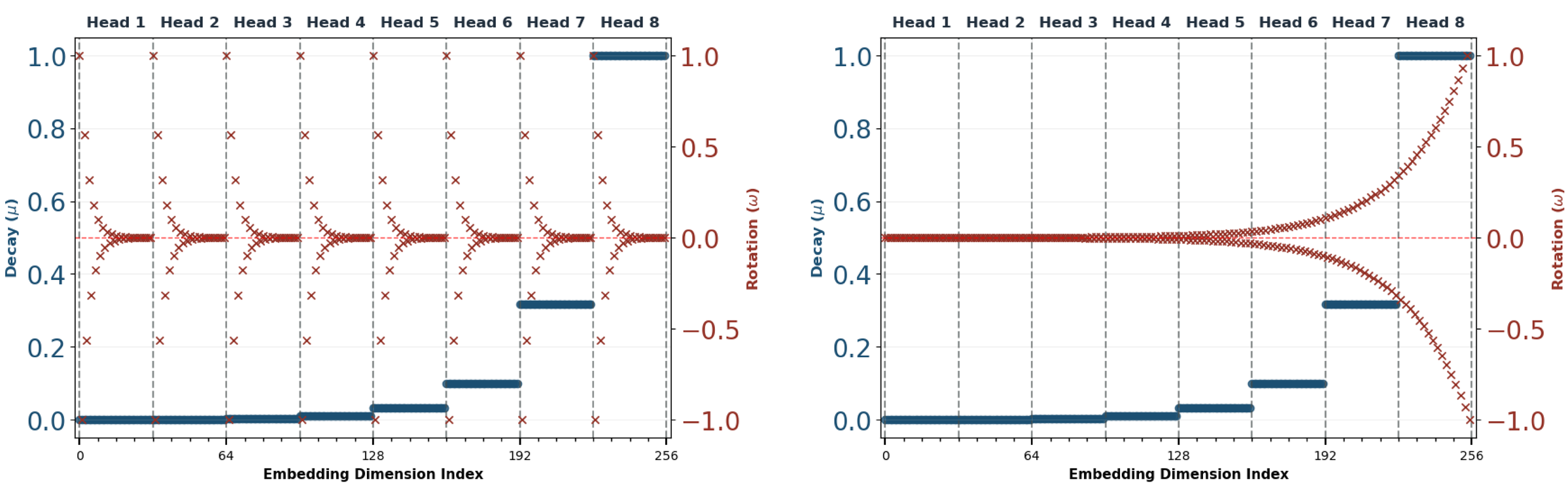

Figure 4. **Left:** RoPE assigns the same frequency spectrum to each head. **Right**: SC-RFA partitions the frequency spectrum across all heads and assigns each head a decay rate proportional to its maximum frequency.

Key Findings

- On language modeling benchmarks, RFA improves in-window perplexity relative to RoPE and stabilizes long-range behavior.
- Removing value rotation causes catastrophic long-context failure, demonstrating that aggregation must occur in a dynamically aligned temporal frame.

Model	L=512	L=1024	L=2048	L=4096
RoPE	28.48	30.94	44.21	72.69
ALiBi	28.59	27.30	26.54	**26.30**
SC-RFA (low damping)	**27.54**	26.73	29.46	37.19
SC-RFA (high damping)	27.91	**26.68**	**26.37**	28.16

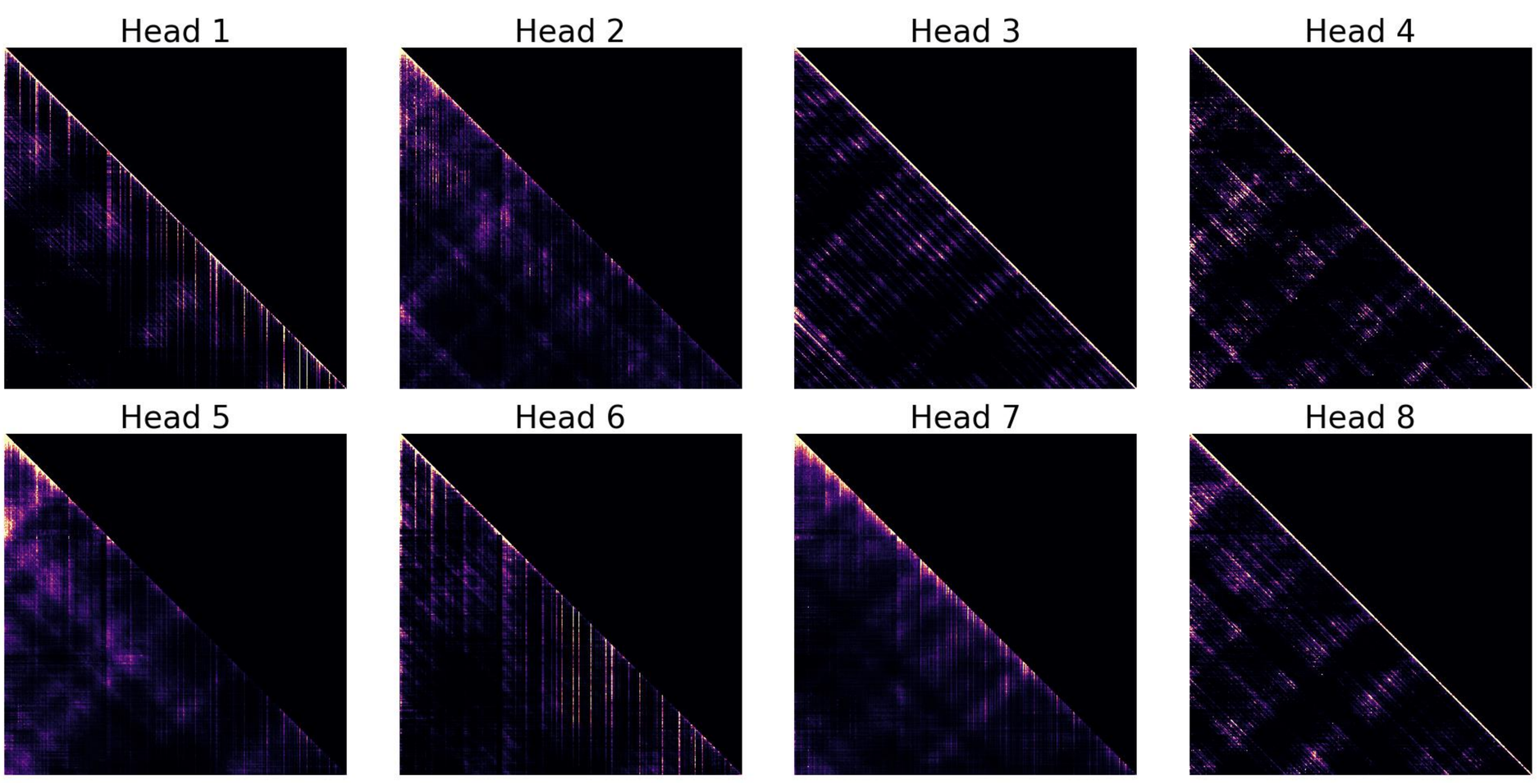

Figure 7. RoPE attention maps (at L=4096) exhibit persistent checkerboard noise.

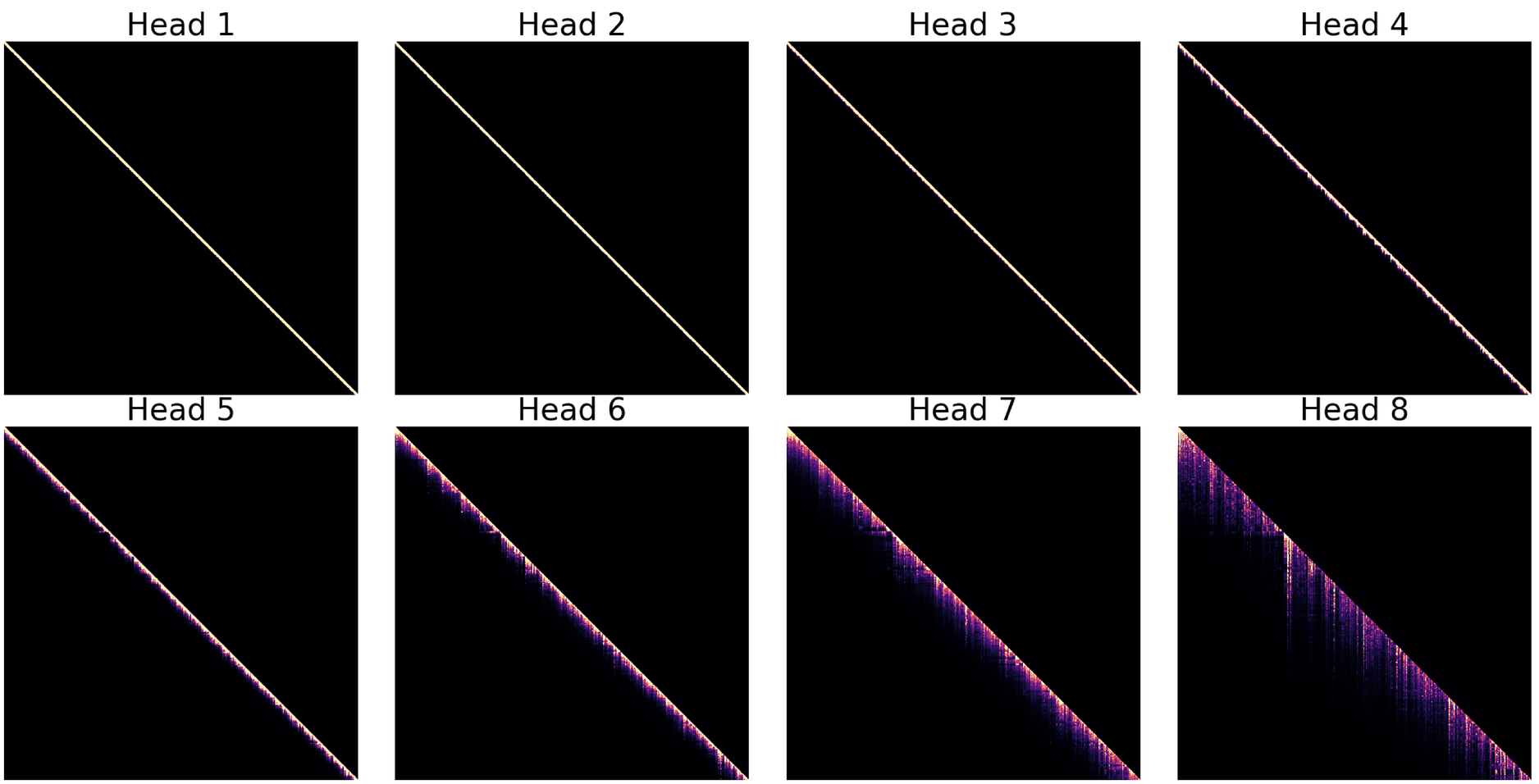

Figure 8. ALiBi attention maps remain tightly localized to the diagonal across all heads

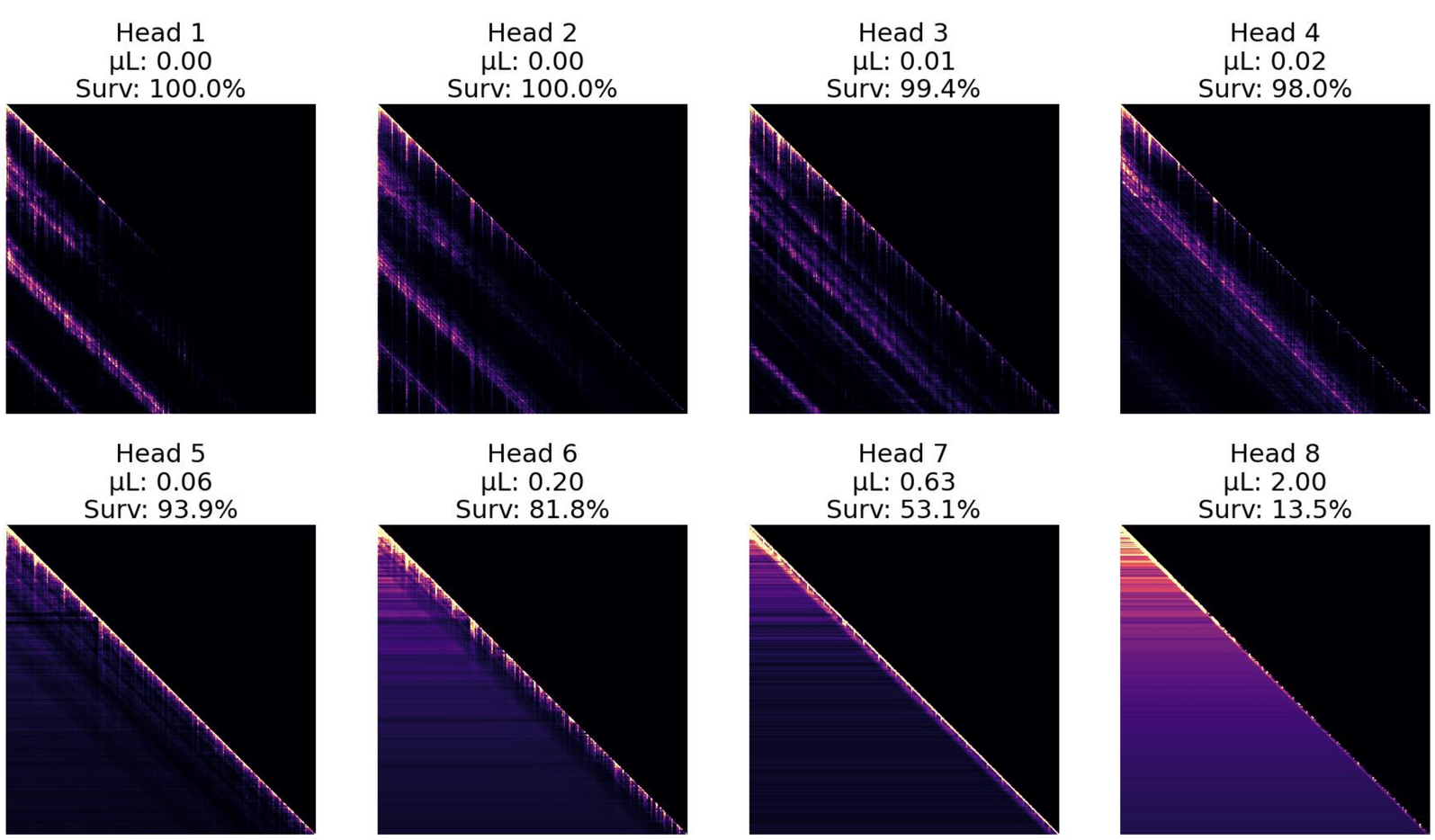

Figure 8. RFA attention maps exhibit periodic bands and integrative behavior in low-decay heads.

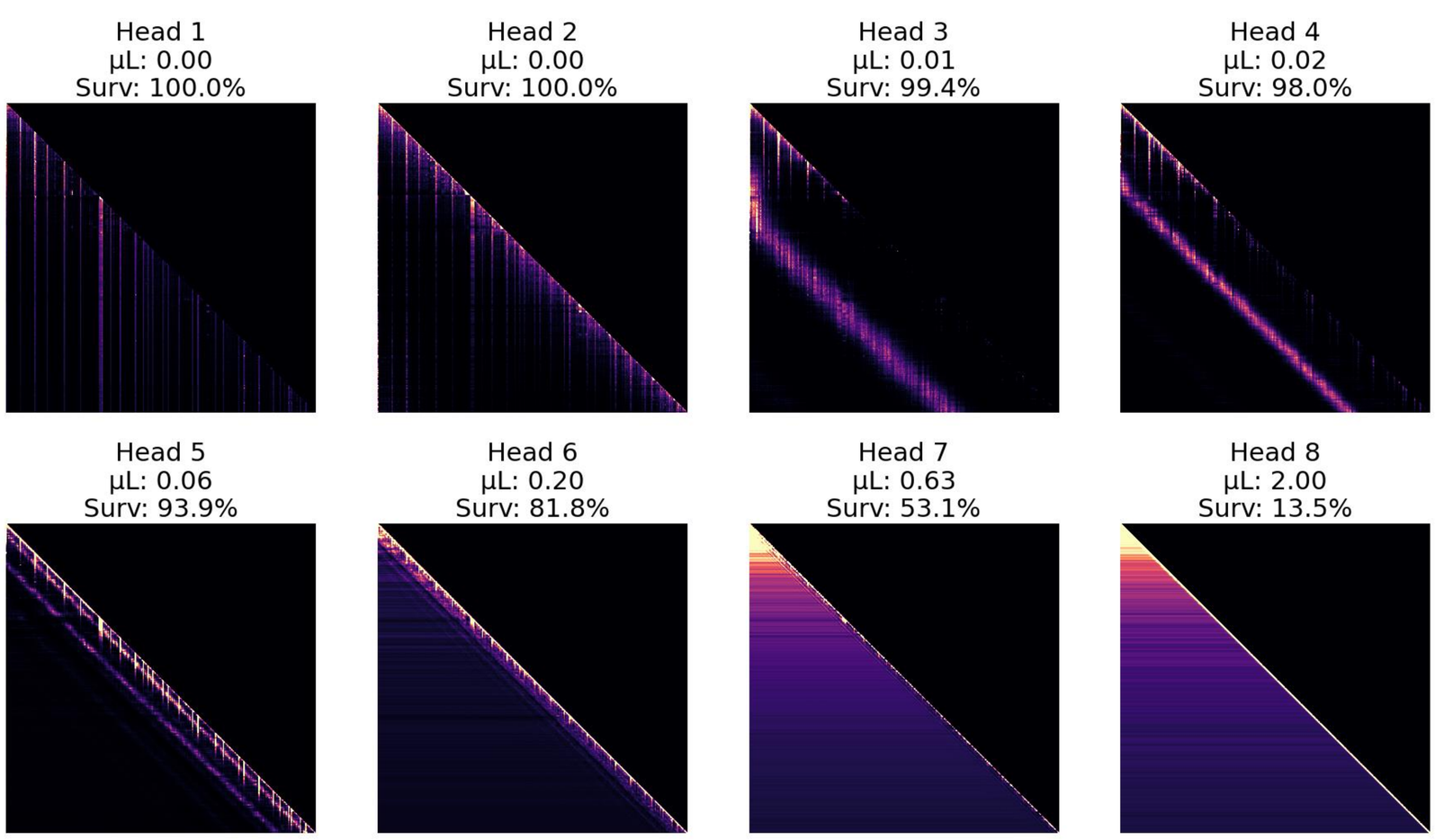

Figure 9. SC-RFA attention maps show specialization to a single lag band per head. First two heads specialize in long-range retrieval.

Conclusion

- RFA reformulates self-attention as uncertainty-aware state estimation under stochastic dynamics.
- Keys are dynamically transported and weighted according to propagated uncertainty rather than static similarity.
- RFA recovers RoPE and ALiBi as limiting cases while improving long-context stability and inducing interpretable multi-scale filtering behavior.
- Future work includes relaxing the isotropic and diagonalizable assumptions while retaining tractable parallel computation.