

It's Not RoPE that Creates Sinks: The Role of Self-Concentration and Value-Non-Mixing in Attention

Raito Kiya¹ Satoki Ohashi¹ Kosuke Sato¹ Go Kamoda^{2,3} Ryosuke Takahashi^{1,4,5}
Yuji Yamamoto^{2,3} Daiki Shiono¹ Keisuke Sakaguchi^{1,5} Goro Kobayashi^{6,1}
(¹Tohoku University ²SOKENDAI ³NINJAL ⁴MBZUAI ⁵RIKEN ⁶Preferred Networks, Inc.)



Attention Sink (AS)

- Disproportionate attention from subsequent tokens
- Sink tokens for streaming inference [Xiao+ '24]

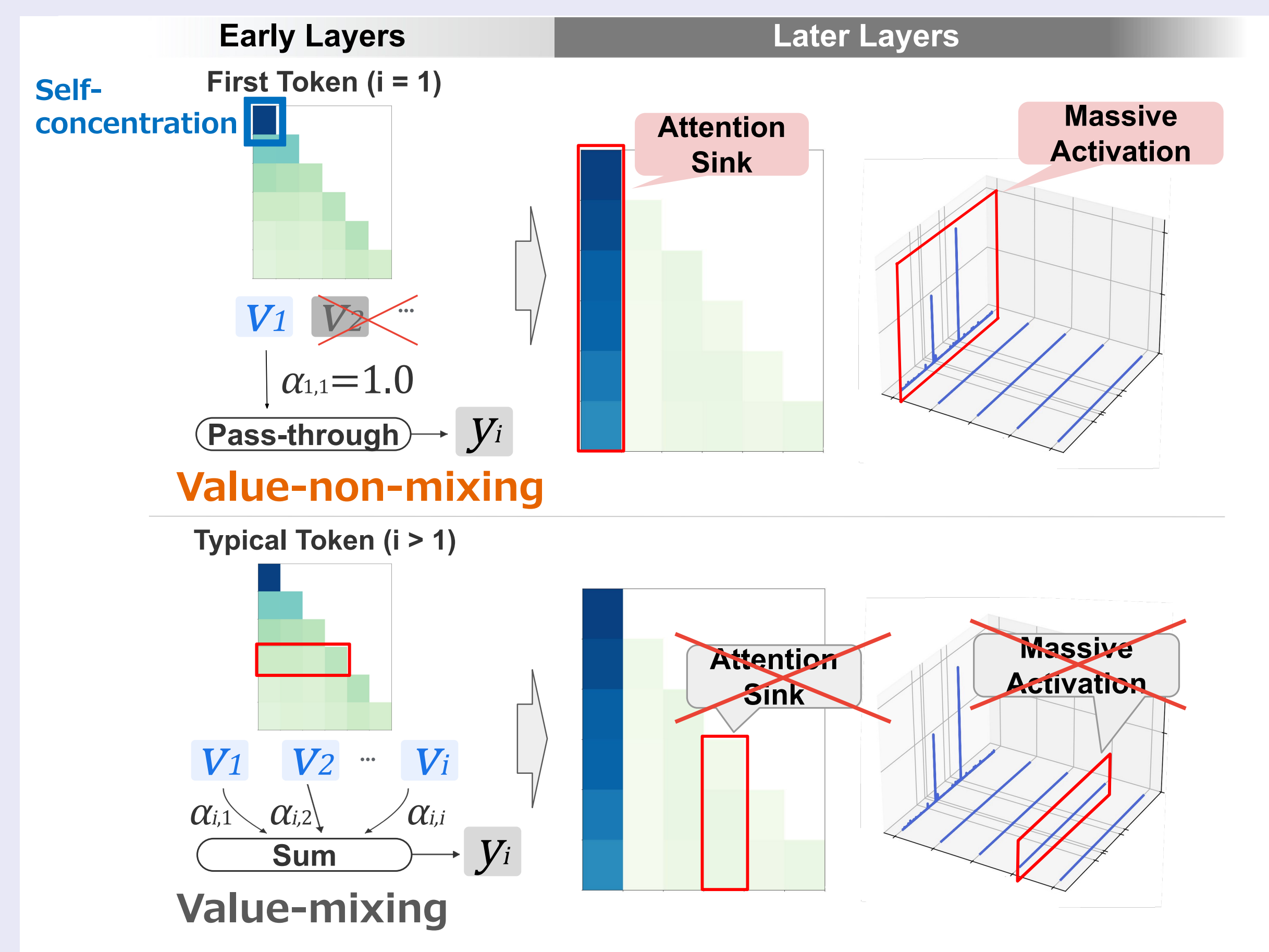
Massive Activations (MAs)

- A few hidden-state dimensions with extreme magnitudes
- Large activations as a challenge for low-bit quantization [Sun+ '24; Dettmers+ '22]

Why do AS and MAs emerge at the first position?

- We focus on first-position properties.
 - BOS
 - RoPE (index 0)
 - Self-concentration**: Pays attention only to itself
→ **Value-non-mixing**: No mixing of different Value vectors

Finding: Value-non-mixing contributes to AS and MAs

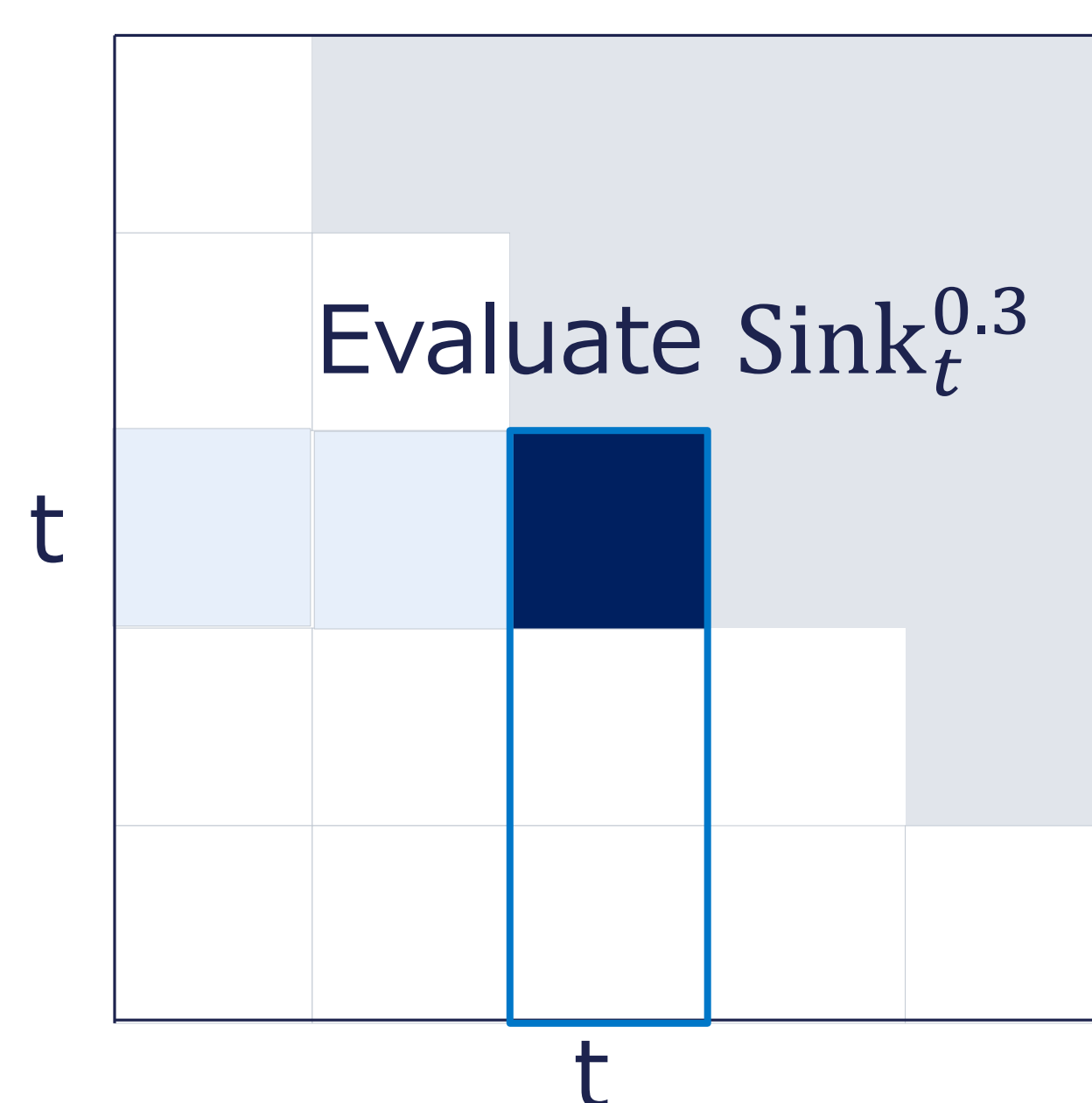


1. Self-concentration induces AS/MAs

Intervention

Forced self-concentration

$$\alpha_{t,t} = 1.0; \alpha_{t,j} = 0.0 (j \neq t) \text{ (at } t = 16)$$



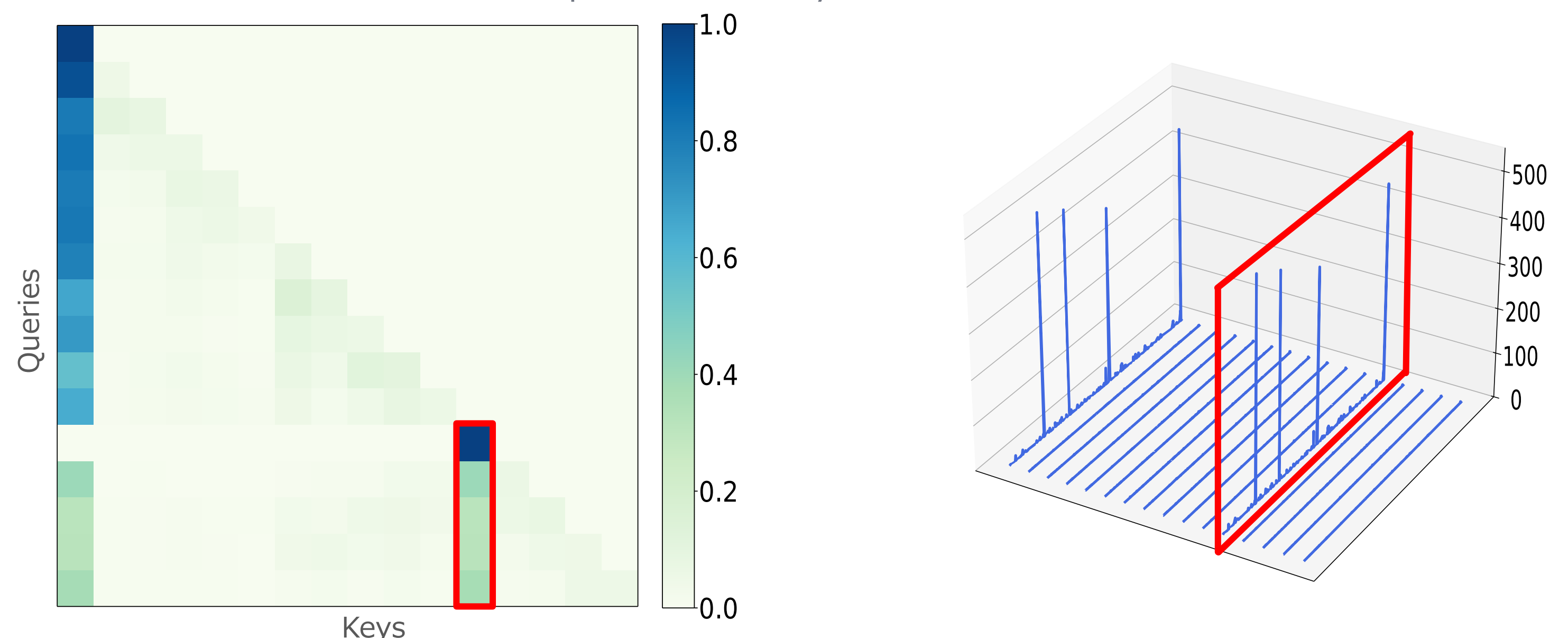
Evaluate Sink_t^{0.3}

Metric Sink_t^ε proposed by Gu+ '25

Results

Attention heatmap (left) · hidden-state activations (right)

Llama-3.2-3B · WikiText-2 · T=16 example (main: T=64)
w/o BOS · self-concentration at position 12 · layer 14



Forced self-concentration: **AS/MAs emerge**

2. Value-non-mixing contributes to AS/MAs

Self-concentration → Value-non-mixing

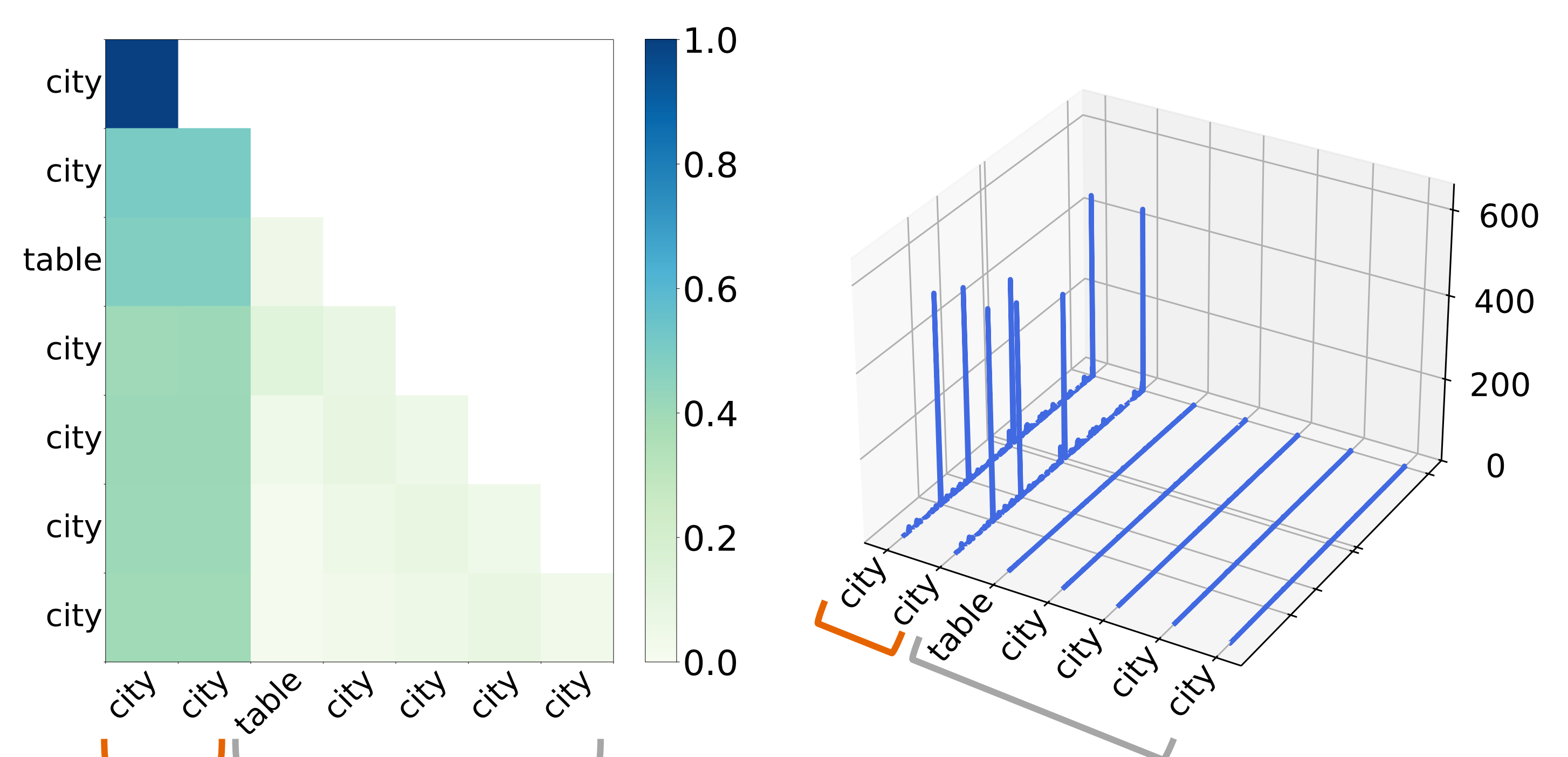
- Own Value only**
→ No mixing of **distinct Value vectors**
- RQ**
Does Value-non-mixing contribute to AS/MAs?
- Experimental design:**
Value-non-mixing without self-concentration
→ **Repeated-token inputs**

e.g. **"A A B A A A ..."**
Value-non-mixing Value-mixing

A: repeated token
B: distinct token

Results

Distinct token at position 3 · heatmap (left), activations (right)
Llama-3.2-3B · layer 14 · T = 7 example (main: T=64)



- Value-non-mixing**: **AS/MAs emerge**
- Value-mixing: **AS/MAs rarely emerge**

BOS / RoPE: First-position effects

- BOS relocation:
AS/MAs still emerge at the first position
Models with default BOS insertion · BOS moved to position 16
- First-token key-side RoPE index change:
No substantial drop in Sink₁^{0.3}
5 models · w/o BOS

Future Work

- Value-vector similarity and a continuous mixing index
- Model and input dependence
- Head-level pathways and early-layer contributions
- Naturally occurring intermediate sinks