SLIDING CRITICAL BAND IN ROPE-BASED LENGTH EXTRAPOLATION

Zifei Bai **Zhiwei Xu**
University of Michigan
Ann Arbor, MI, USA
{zifeibai,zhiweixu}@umich.edu

ABSTRACT

Context extension in RoPE-based Large Language Models has become a primary focus in the development of RoPE-based models. In this paper, we introduce Sliding Critical Band, a framework demonstrating that the dimensions requiring interpolation dynamically migrate across the spectrum under different extrapolation ratios. Building on this, we proposed Spectrum Bandwidth Exhaustion, which provides an explanation of why larger RoPE bases can enhance models' extrapolation ability. Together, these two concepts offer a more comprehensive understanding of the principles underlying context extrapolation in RoPE-based models. Evaluations on synthetic tasks and the C4 dataset validate the universality of the Sliding Critical Band across diverse scenarios. [1].

1 INTRODUCTION

Rotary Positional Encoding (Su et al., 2023) is widely adopted in transformer-based (Vaswani et al., 2017) Large Language Models (LLMs) for natural language processing (NLP) tasks. However, model performance degrades during inference when the sequence length exceeds the maximum length encountered during training. Meanwhile, in real-world scenarios, large language models frequently encounter texts of considerable length. This challenge renders the enhancement of large language models' context extrapolation capabilities an urgent problem that demands resolution.

Prior works proposed several methods to overcome this problem: Chen et al. (2023) and kaiokendev (2023) pioneered the extension of context windows by introducing Position Interpolation (PI), which rescales the relative distance between tokens to fit within the pre-trained range, followed by fine-tuning on limited long-context data. Departing from this approach, more sophisticated schemes like NTK-aware interpolation (bloc97, 2023a) and YaRN (Peng et al., 2023) have been developed.

The prevailing view holds that the success of extending the long-context window in RoPE-based transformers stems from avoiding low-frequency dimensions entering out-of-distribution (OOD) angles during long-context inference (Liu et al., 2024). However, in this paper, we demonstrate that existing explanations (Peng et al., 2023; Liu et al., 2024) for RoPE's poor extrapolation capability cannot fully predict the model's performance under mild extrapolation. Our analysis reveals a mismatch between existing OOD theories and empirical observations: During long-context inference, low-frequency rotary dimensions rotating into OOD angles are neither a sufficient nor a necessary condition for model collapse.

To bridge the gap between empirical observations and current theories, we introduce the ***Sliding Critical Band*** perspective, shifting from static periodic boundaries to a dynamic spectral window. Our contributions can be summarized as follows:

- **Re-evaluating Periodic Boundaries:** We discover that certain dimensions that complete a full rotation during pretraining harbor hidden threat: without interpolation, they can cause model collapse. Conversely, some dimensions that enter into out-of-distribution angles during inference are actually benign and performance-neutral.

[1]Code available at: https://github.com/zifei-bai/Sliding-Critical-Band-in-RoPE-based-Length-Extrapolation

- **Sliding Critical Band:** We formalize a dynamic framework where the dimensions requiring interpolation migrate across the spectrum with the proportion of extrapolation ratio.
- **Spectrum Bandwidth Exhaustion:** Our analysis reveals the mechanism underlying RoPE's base scaling: an increased RoPE base postpones the exhaustion of the available spectrum bandwidth, thereby enhancing the model's capacity for long-context extrapolation.

2 BACKGROUND

2.1 ROTARY POSITIONAL ENCODING (ROPE)

Su et al. (2023) incorporate positional information to attention mechanism by rotating $\mathbf{q} \in \mathbb{R}^d$ and $\mathbf{k} \in \mathbb{R}^d$ vectors where d is an even number. RoPE splits a d-dimensional vector into $d/2$ subvectors. For each subspace, the transformation is as follows:

$$(\mathbf{q}_i^m)' = \begin{bmatrix} \cos(m/\theta_i) & -\sin(m/\theta_i) \\ \sin(m/\theta_i) & \cos(m/\theta_i) \end{bmatrix} \mathbf{q}_i^m \tag{1}$$

where $\mathbf{q}_i^m \in \mathbb{R}^2$ is the m-th query $(0 \leq m \leq L-1)$ and the i-th subspace $(0 \leq i \leq \frac{d}{2}-1)$, $\theta_i = \theta^{\frac{-2i}{d}}$. θ is RoPE base, L is sequence length, d is embedding dimension. The same transformation is applied to $\mathbf{k}^m$ in the preceding steps. For simplicity, we denote the RoPE transformation for i-th subvector $\mathbf{x} \in \mathbb{R}^2$ as a function of

$$\mathfrak{R}_i(\mathbf{x}, m) \tag{2}$$

Then for any vector $\mathbf{v} \in \mathbb{R}^d$ at position m,

$$\mathbf{v}' = \begin{bmatrix} \mathfrak{R}_0(\mathbf{v}_{1:2}, m) \\ \mathfrak{R}_1(\mathbf{v}_{3:4}, m) \\ \vdots \\ \mathfrak{R}_{\frac{d}{2}-1}(\mathbf{v}_{d-1:d}, m) \end{bmatrix} \in \mathbb{R}^d \tag{3}$$

where $\mathbf{v}_{j:k}$ denotes the sub-vector of $\mathbf{v}$ from the j-th to the k-th components, for all $j, k \in [1, d]$.

2.2 RELATED WORKS

Context Extension in LLMs To extend the context window of pre-trained LLMs, Position Interpolation (PI) (Chen et al., 2023) linearly scales position indices, though it struggles with high-frequency information loss. Subsequent NTK-aware (bloc97, 2023a) methods, including NTK-by-parts (bloc97, 2023b) and Dynamic-NTK emozilla (2023), utilize non-uniform frequency scaling to preserve resolution in high-frequency dimensions. YaRN (Peng et al., 2023) refines this by combining non-uniform scaling with attention temperature adjustment to mitigate entropy increases. For non-linear scaling, LongRoPE (Ding et al., 2024) employs evolutionary search to identify optimal non-uniform interpolation factors, while CLEX (Chen et al., 2024) models length scaling as a continuous dynamical system using Ordinary Differential Equations (ODEs). Prior studies (Xiong et al., 2024; Rozière et al., 2024; bloc97, 2023a) found that increasing the base wavelength of RoPE alleviates the attention decay. Barbero et al. (2025) observed that lower frequencies are less affected by the relative distance and proposed p-RoPE to improve model's long-context capabilities by replacing the lowest frequencies of RoPE with No Positional Encoding (NoPE) (Haviv et al., 2022; Kazemnejad et al., 2023).

Theoretical Analysis of RoPE Theoretical analysis identifies RoPE's efficacy and limits through architectural and periodic lenses. Slash-Dominant Heads (Cheng et al., 2026) emerge from low-rank queries and cone-shaped embeddings, where the slash attention pattern is crucial in long-context inference (Xu et al., 2025; Jiang et al., 2024). From a periodic perspective, the critical dimension (Liu et al., 2024) determines a threshold for the feature dimension, beyond which extrapolation fails due to transitioning into out-of-distribution (OOD) scenarios. Men et al. (2024) concluded that RoPE's base frequency fundamentally limits the context length; falling below this lower bound results in the model exhibiting "superficial" capabilities, where model maintains low perplexity but its actual long-range information retrieval performance deteriorates.

3 PRELIMINARIES AND EXPERIMENTAL SETUP

In this section, we describe the experimental setup, including the model architecture, data generation, training, interpolation methodology, and evaluation metric.

Models. We adopt the LLaMA architecture with 4 layers, 2 attention heads, and an embedding dimension of 384 with a total of 7.1M parameters. We use Rotary Positional Encoding(Su et al., 2023) with RoPE base $\theta = 10,000$ except specified in Section 5. Character-level tokenization is used, the vocabulary size is 14, which contains 0-9, =, and special tokens [BOS], [EOS], [PAD].

Task. We primarily focus on string-copying task. The task is to exactly replicate the input sequence. The example is provided in Table 1:

Table 1: Examples of string-copying task

Input (Q: prompt, A: label)		Task difficulty
Q: 12345=	A: 12345	Length of prompt
Q: 123456789=	A: 123456789	

Data Generation. We follow the pipeline proposed in Lee et al. (2025). We generate the training data $\mathcal{D}_{\text{train}}$ of up to a fixed string length l_0 by uniformly sampling the string length $1 \leq l \leq l_0$ and generating a random sequence with length l. Denoting the input as x_i, labels as y_i,

$$\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N}, \quad \text{where Length}(x_i) \leq l_0.$$

Training and Evaluation. During pretraining, we use standard next-token-prediction loss. See more training details in Appendix A. During inference, we employ greedy decoding to generate completions. We employ both exact-match accuracy and perplexity (PPL) to assess model performance. While exact-match accuracy provides a direct measure of task success, it often fails to provide an informative signal in challenging regimes where the model has not yet mastered the task, resulting in a "zero-accuracy" plateau. However, a null accuracy does not imply identical model capabilities. To capture the fine-grained progression of model learning, we use perplexity as a continuous proxy. Perplexity remains sensitive to improvements in the model's internal representations even when the discrete accuracy remains at zero.

Interpolation Methods. To investigate the impact of different frequency components on extrapolation performance, we generalize Chen et al. (2023)'s interpolation method. Specifically, instead of uniformly scaling all dimensions, we apply interpolation only to a pre-defined subset of frequency subspaces.

Let $\mathcal{S} \subseteq \{0, 1, \ldots, \frac{d_{\text{max}}}{2} - 1\}$ be the set of indices for the subset targeted for interpolation. For a sequence length L' larger than the maximum training sequence length L_{train}, we modify the transformation function $\mathfrak{R}_i$ for each subspace as follows:

$$\mathfrak{R}_i^*(\mathbf{x}, m) = \begin{cases} \mathfrak{R}_i\left(\mathbf{x}, m/s\right) & \text{if } i \in \mathcal{S} \\ \mathfrak{R}_i(\mathbf{x}, m) & \text{if } i \notin \mathcal{S} \end{cases}, \quad s = \frac{L'}{L_{\text{train}}} \tag{4}$$

where $\mathfrak{R}_i$ is the original rotary transformation defined in equation 2, and we refer s as extrapolation ratio. Note that this selective interpolation is applied consistently to all query vectors $\mathbf{q}$ and key vectors $\mathbf{k}$ across all positions m and the smaller index corresponding to higher frequency.

By strategically choosing the index set $\mathcal{S}$ (e.g., targeting specific frequencies), we can isolate and analyze the sensitivity of the model's extrapolation capability to different spectral components of the positions encoding.

4 SLIDING CRITICAL BAND

In this section, we will re-examine the common claim regarding the failure mechanism of Rotary Positional Embedding (RoPE) extrapolation. Current literature (Peng et al., 2023; Liu et al., 2024) attributes RoPE extrapolation failure to dimensions that fail to complete a full rotation cycle during training. This boundary is formalized by the critical dimension (d_{extra}), a theoretical threshold introduced by Liu et al. (2024) to partition the spectrum based on training-time periodicity:

$$d_{\text{extra}} = \left\lceil \frac{d_{\text{max}}}{2} \log_\theta \left(\frac{L_{\text{train}}}{2\pi} \right) \right\rceil \tag{5}$$

where d_{max} is head dimension, L_{train} is the longest training sequence length, and θ is RoPE-base. [2] This threshold identifies the first rotary plane where the wavelength exceeds the training sequence length. Consequently, subspaces with indices greater than d_{extra} cover only a partial rotation during training, causing them to enter an out-of-distribution (OOD) state during long-context inference.

While the OOD hypothesis explains performance collapse under extreme extrapolation, it proves insufficient for mild extrapolation (e.g., $s = 1.1, 1.2, 1.5$). To systematically identify which frequencies require interpolation, we employ a truncated interpolation strategy. We define the set of interpolated indices $\mathcal{S}_d$ as:

$$\mathcal{S}_d = \{i \mid d \leq i \leq \frac{d_{\text{max}}}{2} - 1\}, \quad d \in [0, \frac{d_{\text{max}}}{2}] \tag{6}$$

By increasing the starting index d, we progressively remove high-frequency dimensions from the interpolation set. In our experimental setting ($L_{\text{train}} = 203, d_{\text{max}} = 192, \theta = 10000$), the theoretical threshold is $d_{\text{extra}} = 37$.

However, as the first row in Figure 1 shows, under mild extrapolation ratio, the model performance drops before d_{extra}. These results imply that even though the high-frequency dimensions complete the full rotation during training, their OOD state can still have a significant impact on model's extrapolation performance. This finding implies that the OOD state of low-frequency dimensions may not be the complete explanation for extrapolation failure.

In light of these findings, we propose a more dynamic perspective termed ***Sliding Critical Band***, which argues that the dimensions requiring interpolation are not fixed in the low-frequency region, but rather shift across the entire spectrum as the extrapolation ratio s changes.

4.1 FORMALIZING THE CRITICAL BAND

To formally define the Sliding Critical Band, we characterize it by two spectral boundaries: the Upper Bound (d_{upper}) and the Lower Bound (d_{lower}). We propose a two-step empirical search procedure: an exclusive-based search for the Upper Bound and an inclusive-based search for the Lower Bound. First, we identify the Upper Bound (d_{upper}) by progressively excluding the high-frequency dimensions from the interpolation set $\mathcal{S}$ as defined in Equation 6. As illustrated in Figure 1, we observe a "U-shaped" performance curve. Interestingly, interpolating the extreme high-frequency dimensions actually results in sub-optimal performance. Under extrapolation ratio $s = 1.2$, the model achieves its peak performance at $d = 20$, which we define as the Upper Bound d_{upper}. Second, fixing the starting index at d_{upper}, we expand the interpolation range by progressively including the lower-frequency dimensions. The set of interpolation indices $\mathcal{S}$ is a function of the threshold d:

$$\mathcal{S}_d = \{i \mid d_{\text{upper}} \leq i \leq d\}, \quad d \in [d_{\text{upper}} - 1, \frac{d_{\text{max}}}{2} - 1] \tag{7}$$

Figure 2 shows that as more dimensions are added to $\mathcal{S}$, the PPL decreases until it reaches a stable plateau. At $s = 1.2$, this convergence occurs at $d = 68$. We define this saturation point as the Lower Bound (d_{lower}), representing the limit beyond which further interpolation becomes redundant.

Based on these identified boundaries, we can now partition the entire RoPE spectrum into three functional regions relative to their impact on length extrapolation:

[2] Note that unlike the standard derivation (Liu et al., 2024), we omit the leading factor of 2 to define d_{extra} as the index of the rotary subspace (i.e., the pair of dimensions) rather than the absolute dimension index.

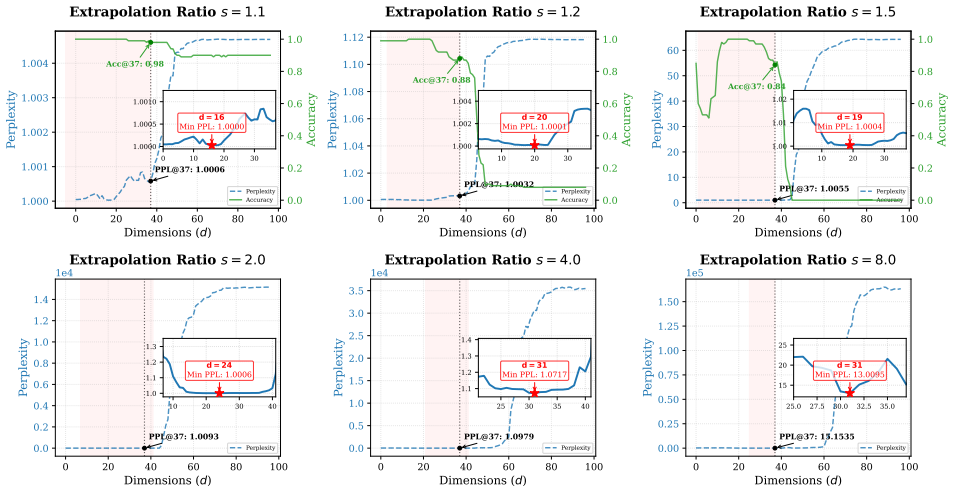

Figure 1: Exclusive-based results of PPL and Accuracy for various extrapolation ratios s for 100-digit string-copying task. Perplexity (PPL) as a function of the interpolation starting index d. The x-axis d represents the upper bound of the dimension range $[d, d_{\text{max}}/2 - 1]$ where interpolation is applied (Equation 6). Shaded regions in the main plots correspond to the magnified insets, providing a detailed view of the local minima of the perplexity for each extrapolation ratio s.

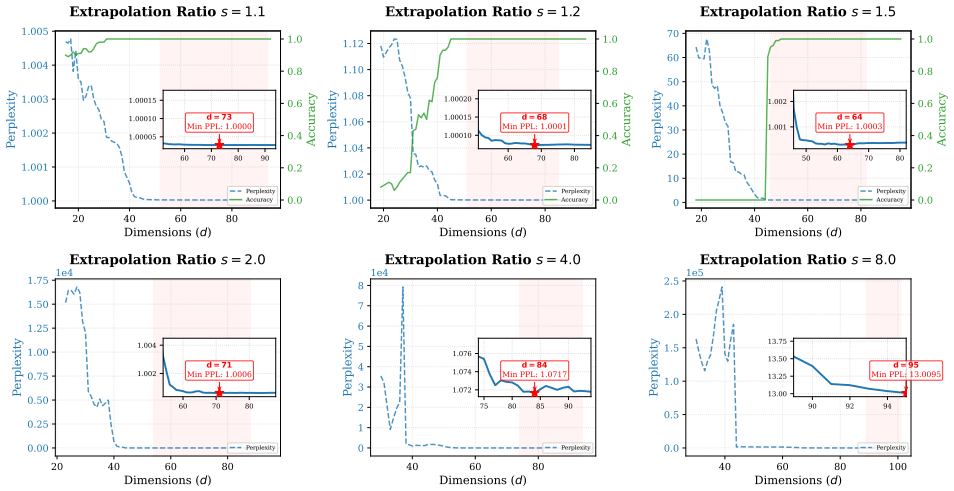

Figure 2: Inclusive-based results of PPL and Accuracy for various extrapolation ratios s for 100-digit string-copying task. Perplexity (PPL) as a function of the interpolation ending index d. The x-axis d represents the lower bound of the dimension range $[d_{\text{upper}}, d]$ where interpolation is applied (Equation 7). Shaded regions in the main plots correspond to the magnified insets, providing a detailed view of the local minima of the perplexity for each extrapolation ratio s.

- Harmful Range $[0, d_{\text{upper}})$: Dimensions that have higher frequencies than the Critical Band. This set of dimensions should not be interpolated. As these dimensions have higher frequencies, they are responsible for the local information. Therefore, interpolating these dimensions causes the model to confuse tokens that are relatively close in positional distance, leading to a decrease in model performance.
- Critical Band $[d_{\text{upper}}, d_{\text{lower}}]$: Dimensions that should be interpolated to maintain the extrapolation capability.
- Benign Range $(d_{\text{lower}}, d_{\text{max}}/2-1]$: Dimensions that have lower frequencies than the Critical Band. These dimensions are performance-neutral because their phase shifts during extrapolation are negligible; the model is largely indifferent to whether they are interpolated.

4.2 THE DYNAMICS: SLIDING ACROSS SPECTRUM

After defining the boundaries of the Critical Band, we further explore its dynamic characteristics as the extrapolation ratio s varies. This "sliding" property is key to explaining the model's extrapolation behavior.

Table 2: Critical band slides across different extrapolation ratios (s) for 50, 100, and 500-digit tasks.

Setting	$\mathbf{s = 1.1}$	$\mathbf{s = 1.2}$	$\mathbf{s = 1.5}$	$\mathbf{s = 2}$	$\mathbf{s = 4}$	$\mathbf{s = 8}$
50 digit	$[8, 48]$	$[0, 37]$	$[20, 63]$	$[25, 68]$	$[33, 87]$	$[30, 87]$
100 digit	$[16, 73]$	$[20, 68]$	$[19, 64]$	$[24, 71]$	$[31, 84]$	$[31, 95]$
500 digit	$[0, 89]$	$[23, 85]$	$[27, 94]$	$[32, 93]$	$[35, 95]$	$[35, 95]$

According to the different string-copying length settings (50, 100, 500-digit) as summarized in Table 2, the upper bound (d_{upper}) and lower bound (d_{lower}) of the Critical Band show a clear trend of shifting to the lower frequencies as the extrapolation ratio s increases. This sliding characteristic, visually demonstrated in Figure 3, reveals a hidden complexity in the relationship between frequency sensitivity and extrapolation. The functional roles of each rotary dimension are not fixed but exist in a dynamic state.

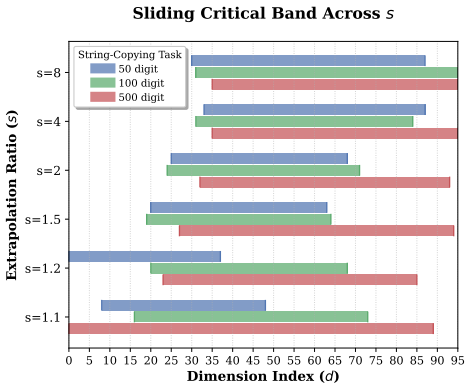

Figure 3: Sliding Critical Band across different task lengths and extrapolation ratios.

More importantly, this discovery corrects the common belief (Liu et al., 2024; Peng et al., 2023) that extrapolation failure is primarily caused by dimensions that have not completed a full rotation $d \geq d_{\text{extra}}$. Our analysis demonstrates that the maintenance of extrapolation capability does not depend on the completion of periodic rotations or the complexity of interpolation functions. Instead, it hinges entirely on whether the interpolation operation can precisely cover the Sliding Critical Band as it shifts. This insight identifies that the dynamic motion of the spectrum is a key factor affecting model performance, and it provides a viable direction for designing adaptive position encoding that can track these frequency boundaries in real time.

5 SPECTRUM BANDWIDTH EXHAUSTION

The dynamic migration of the Sliding Critical Band raises a fundamental question:

what happens when the extrapolation ratio s increases to a point where the required lower bound d_{lower} exceeds the maximum RoPE-dimension $d_{\text{max}}/2$ of the model?

In this section, we define this state as ***Spectrum Bandwidth Exhaustion***. This concept explains the ultimate collapse of extrapolation performance and also illuminates the underlying mechanism of why increasing the RoPE-base, a strategy adopted by Llama 3 (Grattafiori et al., 2024) and Code Llama (Rozière et al., 2024), is effective, and further supports Men et al. (2024)'s point of view.

As the extrapolation ratio s increases, the Critical Band slides to low-frequency dimensions. When the extrapolation ratio becomes extremely large, d_{lower} will eventually be larger than $d_{\text{max}}/2$, leading to spectrum overflow. This missing information is critical for maintaining extrapolation stability; the model will experience a sudden shift from performance degradation to model collapse. The purpose of increasing the RoPE-base is to lower the frequency floor. The frequency θ_d at the d-th dimension will then become smaller under larger RoPE-base θ compared to the original one. Therefore, the Sliding Critical Band, which would have exceeded $d_{\text{max}}/2$ when $\theta = 10,000$, is "pulled" back into

the physical dimension space when $\theta = 100,000$. Moreover, the PPL also becomes smaller, as shown in Figure 4 and Table 3.

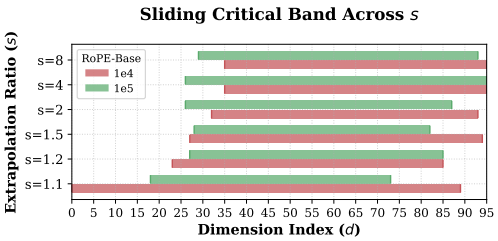

Figure 4: Sliding Critical Band for different RoPE bases with the same task length 500-digit.

Table 3: PPL results after interpolating critical band. Under large-scale extrapolation, larger RoPE-base achieves better performance.

s	$\theta = 10^4$	$\theta = 10^5$
$s = 2$	1.0013	1.0002
$s = 4$	1.0389	1.0178
$s = 8$	3.2016	1.9984

In summary, Spectrum Bandwidth Exhaustion represents the hard physical limit of RoPE-based models; the finiteness of the dimension spectrum ensures an eventual collapse. Enlarging the RoPE-base is, in essence, an act of 'buying more space' for the band to slide, thereby extending the functional horizon of the model's long-context extrapolation capabilities.

6 Recovering Slash Pattern Attention

In this section, we will discuss the necessity and effectiveness of interpolating Critical Band for enhancing the model's length extrapolation capability from the perspective of attention scores pattern. Cheng et al. (2026) proposed Slash-Dominant Heads (SDHs), a phenomenon where the attention distribution is primarily determined by RoPE and is largely unaffected by semantic content. The slash patterns play an important role in long-context inference (Xu et al., 2025; Jiang et al., 2024).

As shown in the first row of Figure 5, when the model processes length-OOD sequences, its slash pattern tendency weakens, becomes chaotic, or even disappears. This indicates that the relative distances between tokens become blurred and disordered, which we believe is the underlying reason for the model's failure during length extrapolation. Critical Band interpolation restores the structural slash attention pattern (Figure 5), thereby reinstating the model's capacity to capture relative positional dependencies between tokens.

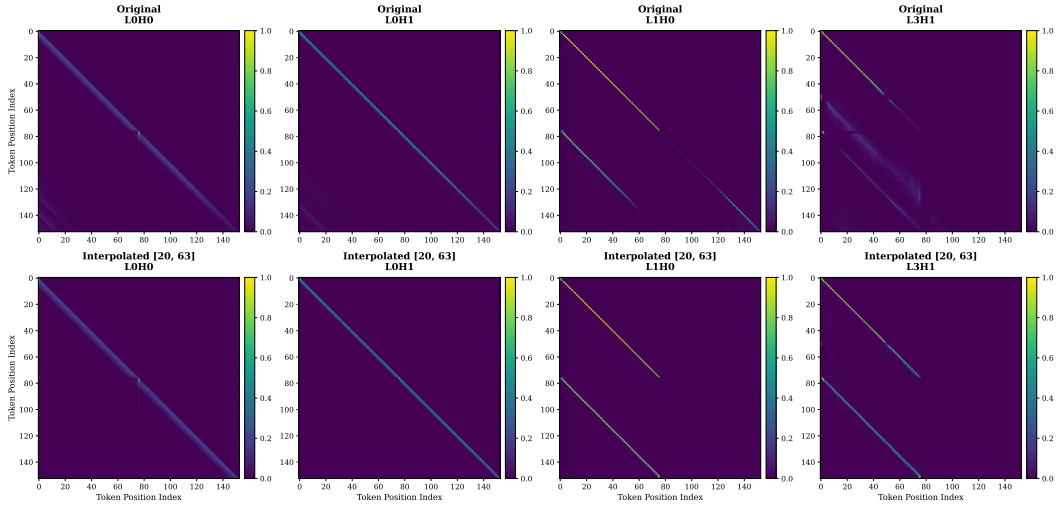

Figure 5: Comparison of attention pattern between original RoPE (first row) and Critical Band interpolation RoPE (second row) in 100-digit string-copying task with extrapolation ratio $s = 1.5$. The a-th head at b-th layer is denoted as LbHa. The slash attention pattern in the first row is weaker than that in the second row.

7 Further Validation: Real-World Scenarios

To further confirm the generalizability of the Sliding Critical Band, we conduct our experiments on LlaMA-7B (Touvron et al., 2023), specifically using the model weights adapted for the Hugging Face Transformers library (HuggyLLaMA, 2023). The dataset for evaluation is the English subset of C4 dataset (Raffel et al., 2023).

LlaMA-7B is pretrained with a $L_{\text{train}} = 2048$ context window length. We consider $s = 1.125$ ($L' = 2304$) and $s = 1.25$ ($L' = 2560$) as mild extrapolation and $s = 2.0$ ($L' = 4096$) and $s = 4.0$ ($L' = 8192$) as extreme extrapolation. The C4 dataset is truncated to this desired test length. Following the procedure in Section 4.1, we find the Critical Bands for $s = 1.125, s = 1.25, s = 2.0, s = 4.0$, as illustrated in Figure 6. Remarkably, the Critical Band exhibits an identical rightward migration as the extrapolation ratio s increases, shifting from the higher-frequency region at $s = 1.125$ toward the lower-frequency dimensions at $s = 4$. This further validation on real-world model and natural language dataset suggests that sliding dynamics are a fundamental property of RoPE-based models. We provide the detailed results and convergence plots in Figure 13, Appendix B.

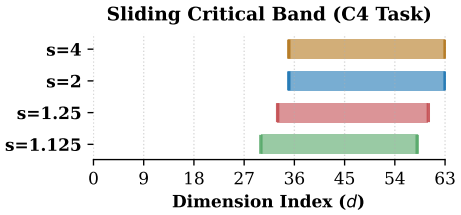

Figure 6: Sliding Critical Bands for LlaMA-7B on C4 with different extrapolation ratios.

8 Conclusion

In conclusion, our paper moves beyond the static view of RoPE extrapolation by revealing the inherent dynamic nature of spectral interpolation. We demonstrate that the requirements for interpolation are not fixed but instead evolve according to a Sliding Critical Band as the extrapolation scale increases. This framework provides a unified explanation of the effectiveness of existing methods and identifies the precise spectral dimensions responsible for stability across varying context lengths. Furthermore, we elucidate the mechanism by which increasing the RoPE base frequency enhances extrapolation, characterizing it as a strategy to delay Spectrum Bandwidth Exhaustion. The concepts of the Critical Band and Spectrum Bandwidth Exhaustion introduced in this work offer a novel perspective, which may provide insights for future research on context extension.

References

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful?, 2025. URL https://arxiv.org/abs/2410.06205.

bloc97. Ntk-aware Scaled RoPE allows LLaMA models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation. Reddit (r/LocalLLaMA), 2023a. URL https://www.reddit.com/r/LocalLLaMA/comments/14lz7j5/ntkaware_scaled_rope_allows_llama_models_to_have/.

bloc97. Add ntk-aware interpolation "by parts" correction. URL https://github.com/jquesnelle/yarn/pull/1, 2023b.

Guanzheng Chen, Xin Li, Zaiqiao Meng, Shangsong Liang, and Lidong Bing. Clex: Continuous length extrapolation for large language models, 2024. URL https://arxiv.org/abs/2310.16450.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation, 2023. URL https://arxiv.org/abs/2306.15595.

Yuan Cheng, Fengzhuo Zhang, Yunlong Hou, Cunxiao Du, Chao Du, Tianyu Pang, Aixin Sun, and Zhuoran Yang. Demystifying the slash pattern in attention: The role of rope, 2026. URL https://arxiv.org/abs/2601.08297.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens, 2024. URL https://arxiv.org/abs/2402.13753.

emozilla. Dynamically Scaled RoPE further increases context window to 8192 without fine-tuning. Reddit (r/LocalLLaMA), 2023. URL https://www.reddit.com/r/LocalLLaMA/comments/14mrgpr/dynamically_scaled_rope_further_increases/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le,

Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. Transformer language models without positional encodings still learn positional information. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1382–1390, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.99. URL https://aclanthology.org/2022.findings-emnlp.99/.

HuggyLLaMA. Llama-7b weights for transformers. https://huggingface.co/huggyllama/llama-7b, 2023.

Huiqiang Jiang, Yucheng Li, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir H. Abdi, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Minference 1.0: Accelerating pre-filling for long-context llms via dynamic sparse attention, 2024. URL https://arxiv.org/abs/2407.02490.

kaiokendev. Extending context to 8k. https://kaiokendev.github.io/til#extending-context-to-8k, 2023.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers, 2023. URL https://arxiv.org/abs/2305.19466.

Nayoung Lee, Ziyang Cai, Avi Schwarzschild, Kangwook Lee, and Dimitris Papailiopoulos. Self-improving transformers overcome easy-to-hard and length generalization challenges, 2025. URL https://arxiv.org/abs/2502.01612.

Xiaoran Liu, Hang Yan, Shuo Zhang, Chenxin An, Xipeng Qiu, and Dahua Lin. Scaling laws of rope-based extrapolation, 2024. URL https://arxiv.org/abs/2310.05209.

Xin Men, Mingyu Xu, Bingning Wang, Qingyu Zhang, Hongyu Lin, Xianpei Han, and Weipeng Chen. Base of rope bounds context length, 2024. URL https://arxiv.org/abs/2405.14591.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models, 2023. URL https://arxiv.org/abs/2309.00071.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. URL https://arxiv.org/abs/1910.10683.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. Code llama: Open foundation models for code, 2024. URL https://arxiv.org/abs/2308.12950.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL https://arxiv.org/abs/2104.09864.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. Effective long-context scaling of foundation models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4643–4663, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.260. URL https://aclanthology.org/2024.naacl-long.260/.

Ruyi Xu, Guangxuan Xiao, Haofeng Huang, Junxian Guo, and Song Han. Xattention: Block sparse attention with antidiagonal scoring, 2025. URL https://arxiv.org/abs/2503.16428.

A Experimental Details

Training and Evaluation. All the models are trained from scratch using standard next-token prediction objective and cross-entropy loss, except the LlaMA-7B in Section 7. We compute the cross-entropy loss over the entire sequence; masking is applied to the padding tokens to ensure they do not contribute to the gradient updates. Training details are as follows:

- We use AdamW optimizer with betas $\beta = (0.9, 0.99)$ and epsilon $\epsilon = 1e-12$. Weight decay is fixed to 0.1.
- We use linear warmup and cosine decay as the learning rate schedule.
- We do not use dropout.

Table 4: Training Details and Hyperparameters

Task	Total Steps	Warmup	Decay	LR	# D_{train}	Batch size	Block size
50-digit	9000	1000	2000	$5e-4$	2M	1000	128
100-digit	9000	1000	2000	$5e-4$	3M	1000	256
500-digit	9000	1000	2000	$5e-4$	3M	800	1024
500-digit, θ=1e5	9000	1000	2000	$5e-4$	3M	800	1024

B More Results

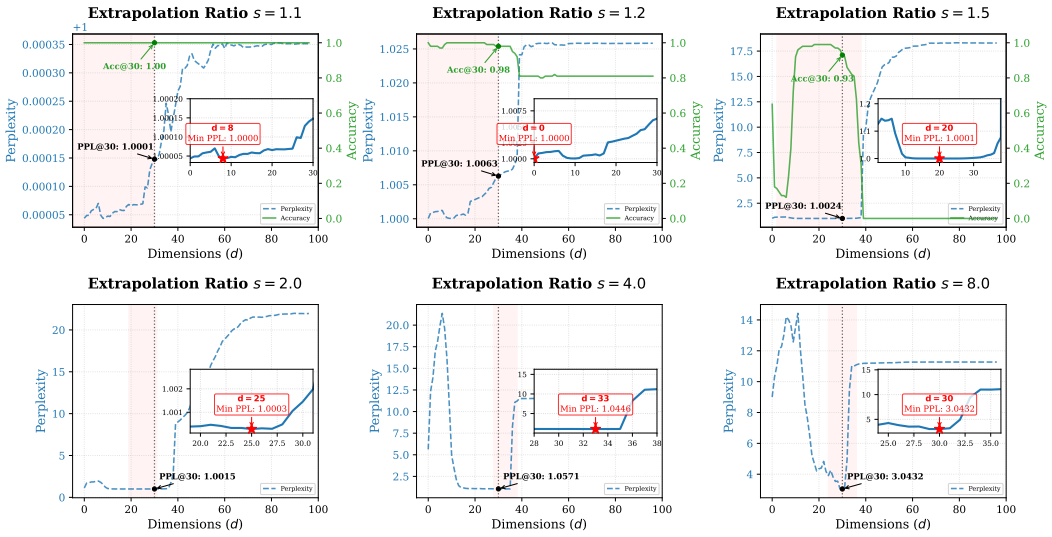

Figure 7: Exclusive-based results of PPL and Accuracy for various extrapolation ratios s for 50-digit string-copying task. Perplexity (PPL) as a function of the interpolation ending index d. The x-axis d represents the upper bound of the dimension range $[d, d_{\text{max}}/2 - 1]$ where interpolation is applied (Equation 6). Shaded regions in the main plots correspond to the magnified insets, providing a detailed view of the local minima of the perplexity for each extrapolation ratio s.

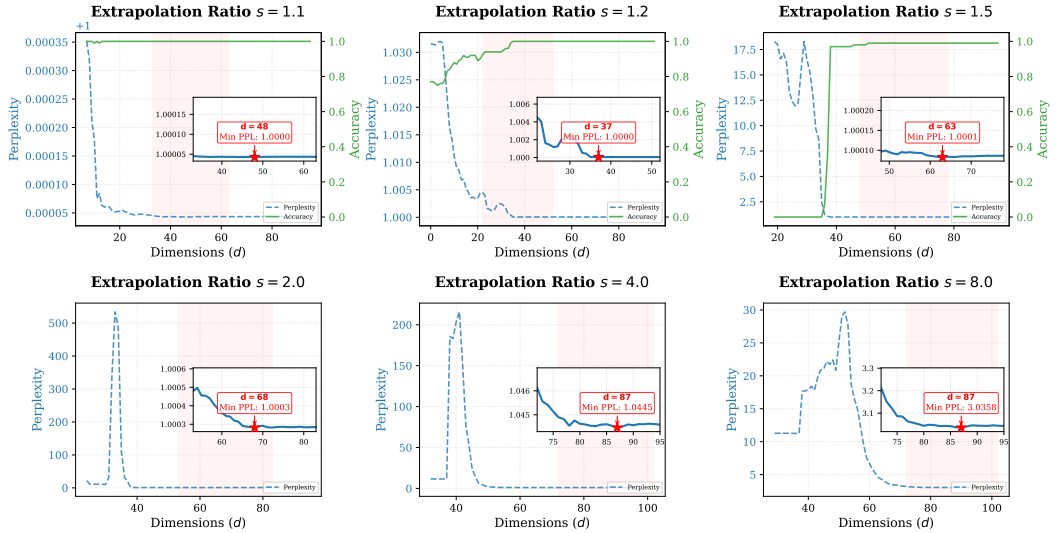

Figure 8: Inclusive-based results of PPL and Accuracy for various extrapolation ratios s for 50-digit string-copying task. Perplexity (PPL) as a function of the interpolation ending index d. The x-axis d represents the lower bound of the dimension range $[d_{\text{upper}}, d]$ where interpolation is applied (Equation 7). Shaded regions in the main plots correspond to the magnified insets, providing a detailed view of the local minima of the perplexity for each extrapolation ratio s.

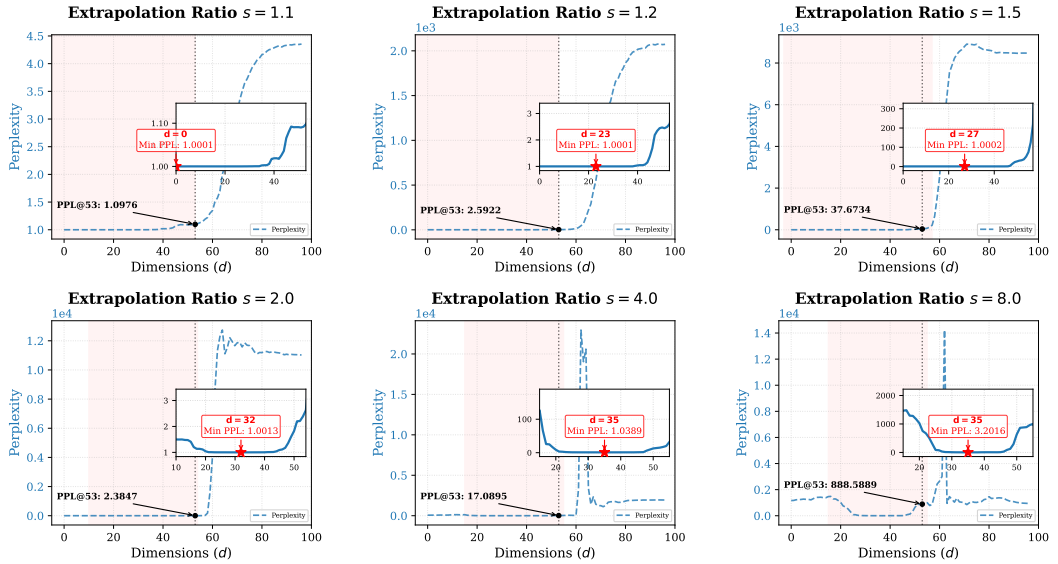

Figure 9: Exclusive-based results of PPL for various extrapolation ratios s for 500-digit string-copying task. Perplexity (PPL) as a function of the interpolation ending index d. The x-axis d represents the upper bound of the dimension range $[d, d_{\text{max}}/2 - 1]$ where interpolation is applied (Equation 6). Shaded regions in the main plots correspond to the magnified insets, providing a detailed view of the local minima of the perplexity for each extrapolation ratio s.

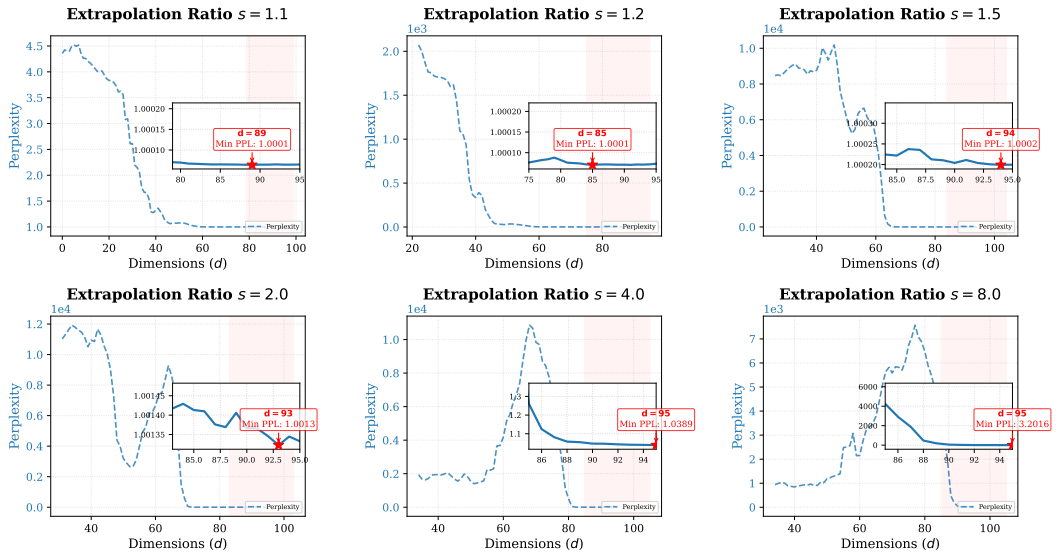

Figure 10: Inclusive-based results of PPL for various extrapolation ratios s for 500-digit string-copying task. Perplexity (PPL) as a function of the interpolation ending index d. The x-axis d represents the lower bound of the dimension range $[d_{\text{upper}}, d]$ where interpolation is applied (Equation 7). Shaded regions in the main plots correspond to the magnified insets, providing a detailed view of the local minima of the perplexity for each extrapolation ratio s.

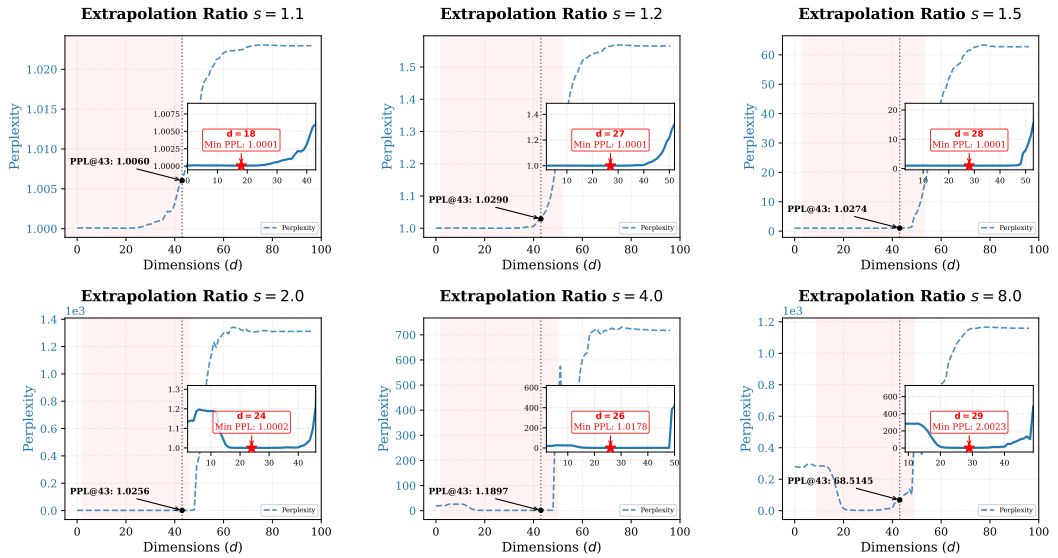

Figure 11: Exclusive-based results of PPL for various extrapolation ratios s for 500-digit string-copying task with RoPE-base $\theta = 100,000$. Perplexity (PPL) as a function of the interpolation ending index d. The x-axis d represents the upper bound of the dimension range $[d, d_{\text{max}}/2-1]$ where interpolation is applied (Equation 6). Shaded regions in the main plots correspond to the magnified insets, providing a detailed view of the local minima of the perplexity for each extrapolation ratio s.

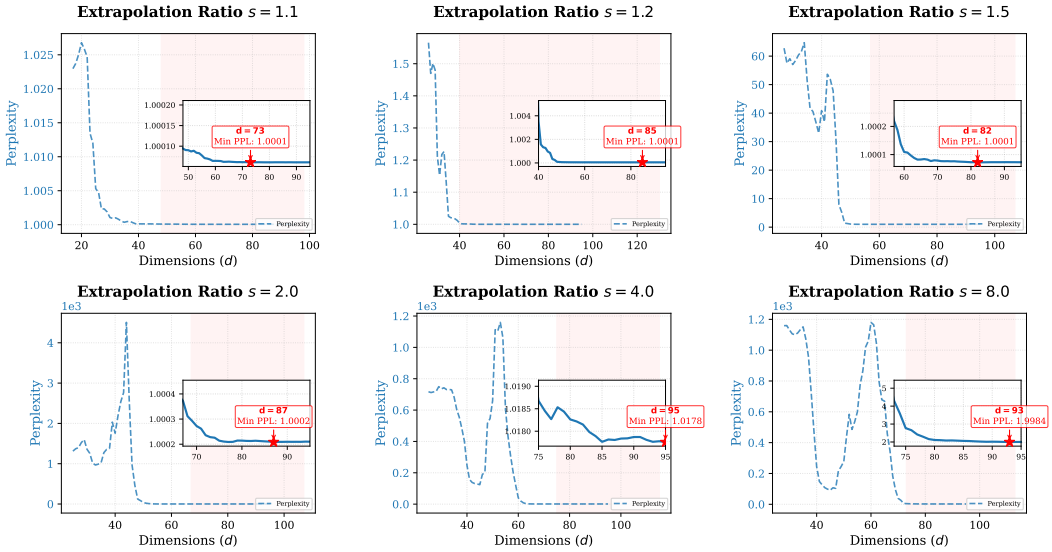

Figure 12: Inclusive-based results of PPL for various extrapolation ratios s for 500-digit string-copying task with RoPE-base $\theta = 100,000$. Perplexity (PPL) as a function of the interpolation ending index d. The x-axis d represents the lower bound of the dimension range $[d_{\text{upper}}, d]$ where interpolation is applied (Equation 7). Shaded regions in the main plots correspond to the magnified insets, providing a detailed view of the local minima of the perplexity for each extrapolation ratio s.

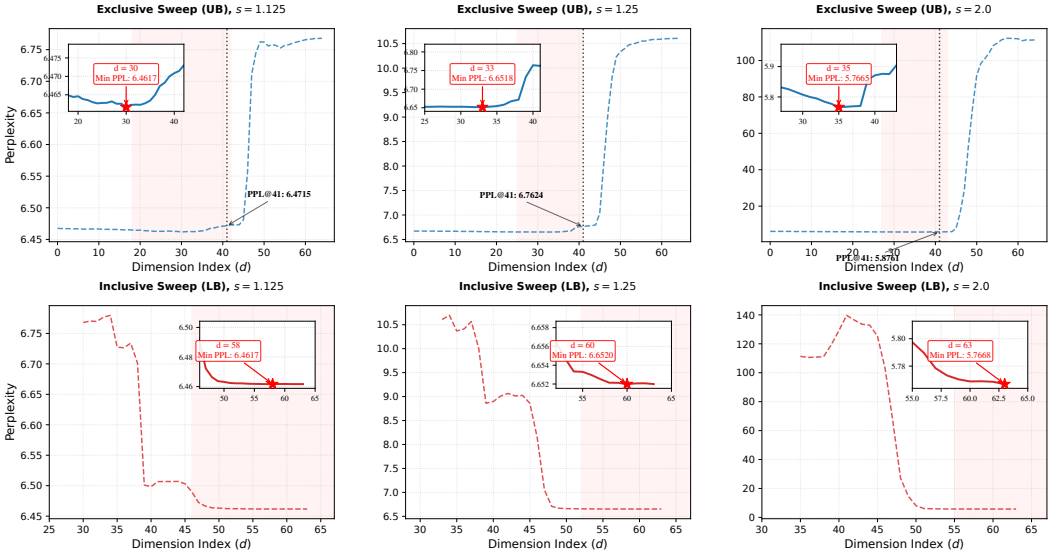

Figure 13: PPL for various extrapolation ratios s for LlaMA-7B on C4 dataset. Perplexity (PPL) as a function of the interpolation index d. For the first row, x-axis d represents the upper bound of the dimension range $[d, d_{\text{max}}/2 - 1]$ where interpolation is applied (Equation 6). For the second row, x-axis d represents the lower bound of the dimension range $[d_{\text{upper}}, d]$ where interpolation is applied (Equation 7). Shaded regions in the main plots correspond to the magnified insets, providing a detailed view of the local minima of the perplexity for each extrapolation ratio s.