[Proposal] Rethinking OCR-based Long-Context Modeling via Hybrid Visual–Textual Inputs

Yuxiang Huang
Student ID 2025310728
Dept. of CS&T
Tsinghua University

Jiahua Chen
Student ID 2025210768
Dept. of CS&T,
Tsinghua University

Fangzhou Xiong
Student ID 2025210764
Dept. of CS&T,
Tsinghua University

1 Introduction

In recent years, Large Language Models (LLMs) have demonstrated rapidly growing capabilities [5, 9, 18, 51]. Advances in model architectures [8] and training methodologies [15] have significantly expanded the boundaries of natural language understanding, forming a solid foundation for complex Artificial Intelligence (AI) applications [34], such as LLM-based multi-agent systems [21], autonomous coding agents [39], and solving complex problems through reasoning [20]. These scenarios place increasing demands on an LLM's ability to process long-context input sequences, as accurately understanding extended context is crucial for generating high-quality responses.

Pursuing longer input sequences has been a persistent objective in enhancing the core capabilities of LLMs. Early models such as BERT [12] were restricted to a maximum input length of 512 tokens, whereas state-of-the-art systems like Gemini-2.5-Pro [11] now theoretically support context windows of up to 1M tokens [10]. A widely adopted practice for enabling long-context understanding is continuous pretraining [16, 3] combined with positional embedding interpolation or extrapolation [13], which typically extends the context length from 4K or 8K tokens in the original pretraining stage to $\geq$128K tokens after long-context adaptation. However, this conventional paradigm faces challenges in both efficiency and performance. Naively feeding long-context inputs into Transformer-based models causes a quadratic increase in computation and memory access, thereby raising the latency of next-token prediction [44]. This leads to exorbitant computation costs, making both training and inference with long sequences particularly difficult. Moreover, such methods often struggle to generalize. For instance, existing video generative models are usually trained on short clips, since long sequences of visual tokens consume excessive memory, which in turn causes generated videos to lose temporal consistency over extended durations [19]. Consequently, ***limited context windows remain a bottleneck that constrains LLM performance on a wide range of complex tasks, highlighting the urgent need for an effective solution to overcome this limitation.***

One alternative solution is the family of *OCR-based* long-context processing methods [45, 7]. These approaches leverage the compression and reconstruction capabilities of visual encoders [14] by rendering long-context inputs into images, which are then encoded into visual embeddings and fed into the LLM backbone. Compared with traditional *token-based* methods, OCR-based approaches can achieve higher and more adaptive compression ratios, enabling models to handle inputs longer than those seen during training (at the cost of slightly more aggressive compression). However, these methods completely discard token-level embeddings and rely solely on visual representations, which leads to significant performance degradation on tasks requiring precise long-context retrieval and fine-grained reasoning. Thus, ***long-context processing continues to be a core bottleneck for LLMs, with the following key issues standing in the way of further progress:***

Challenge 1: The trade-off between length generalizability and fine-grained detail comprehension. Token-based methods preserve all tokens in the input sequence, thus excelling at tasks that require precise fine-grained information retrieval, such as RULER [24]. However, they struggle to generalize to

Preprint.

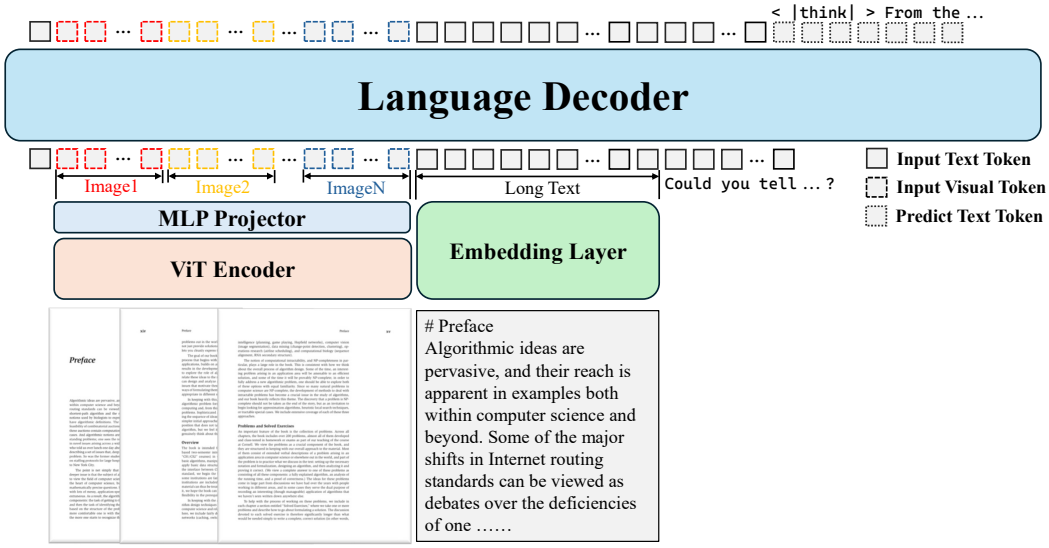

Figure 1: Our method's framework. Both visual and token embeddings are fed into the backbone.

substantially longer contexts, since unseen positional embeddings lead to out-of-distribution errors [4]. In contrast, OCR-based methods can flexibly adjust the compression ratio and thereby generalize to longer inputs with ease; yet, when the compression ratio of the visual encoder becomes high, they lose the ability to retain fine-grained details. As a result, no existing approach simultaneously achieves strong length generalizability and high-fidelity fine-grained recall.

Challenge 2: Adaptive performance–cost trade-off. In real-world applications, a scalable performance–cost trade-off is preferred for long-context processing, as one should be able to simply allocate more compute when higher precision is required. Traditional token-based methods cannot satisfy this requirement, since the computational cost for a given input length is fixed and non-adaptive. OCR-based approaches offer some flexibility by adjusting the compression ratio: lowering the compression ratio increases computation and improves performance. However, when the compression ratio becomes too high, these methods fail catastrophically on long documents, even in simple retrieval tasks. Therefore, a more reasonable performance–cost trade-off curve, i.e., maintaining acceptable performance even under higher compression ratios, is still lacking.

To this end, we propose a method that leverages the strengths of both paradigms: the fine-grained detail preservation of token-based approaches and the strong length generalizability of OCR-based methods. Specifically, we embed long documents in a hybrid manner to achieve higher performance in long-context processing: first converting them into visual embeddings, followed by token embeddings placed before the query. The resolution of the rendered images can be adjusted based on computational budgets or task requirements, while token embeddings provide precise lexical details that visual representations may fail to capture. The framework is shown in Figure 1.

2 Related Work

Here, we briefly introduce existing efficient approaches for extending the context length of LLMs. These methods can be broadly categorized into three groups: long-context modeling techniques, attention and KV-cache optimizations, and OCR-based long-context processing.

Long-Context Modeling. To avoid the data scarcity and high computational cost of pretraining on long documents, recent studies focus on extending the context length of short-context pretrained LLMs through training-free or lightweight adaptation. RoPE [41] encodes relative positional information by rotating query and key vectors in the complex plane, which makes it naturally compatible with context

length extension via interpolation or extrapolation. Building on RoPE, various methods [38, 13, 40, 56, 32] extend the context length by adjusting the base rotary frequency along different dimensions, thereby enabling models to handle input sequences over $4\times$ longer. Such extending is always conducted via a lightweight long-context continuous pretraining stage utilizing specially crafted datasets [6, 3, 16], and the extended models are evaluated on long-context benchmarks [2, 4, 24, 53, 1].

Attention and KV-cache Optimizations. To alleviate the $O(n^2)$ complexity of long-context modeling, numerous methods have been proposed to optimize the attention mechanism and KV-cache, which can be broadly categorized into sparse attention, KV-cache reduction, and token reduction. Sparse attention methods introduce sparsity patterns to reduce computation and memory access. Static sparsity [22, 48, 17] relies on manually designed patterns to minimize accuracy drop on long inputs, whereas dynamic sparse attention [28, 50, 52] selects relevant context blocks for each query in a block-sparse manner. More recently, trainable sparse attention [25, 49, 55, 36, 59] aligns sparsity between training and inference, offering near-lossless performance while significantly reducing computation and KV-cache access. KV-cache reduction seeks to reduce KV memory to enable faster inference or higher throughput: quantization methods [35, 23] store KV entries in low-bit formats; offloading-based approaches [47, 42, 26] load only activated KV blocks to the GPU, making them friendly to long-context extrapolation; and compression strategies such as selective retention [30, 58] or dimensionality reduction [33, 27] further shrink KV size. Token reduction [37, 46] reduces the input length by removing uninformative tokens before feeding into the LLM. For more comprehensive summaries, please refer to recent surveys [31, 43].

OCR-based Long-Context Processing. Among efficient long-context modeling approaches, OCR-based methods have recently gained attention for achieving strong task performance while compressing input sequences to improve efficiency. DeepSeek-OCR [45] scales the OCR model to 3B parameters and achieves state-of-the-art performance, offering a promising foundation for treating long documents as images for processing. Glymph [7] further bridges visual and textual long-context processing by adapting a multimodal LLM to handle long documents, where the input is first converted into visual embeddings. These methods leverage the bidirectional attention available in visual encoders and benefit from high compression ratios to achieve a favorable efficiency–performance trade-off; however, their performance drops considerably when higher compression is applied, particularly on documents requiring fine-grained textual understanding.

3 Early Method

3.1 Task Definition

In this section, we first introduce the task formulation of long-context processing with OCR-based inputs, followed by a detailed description of our proposed method. A summary of the notations used throughout the paper is provided in Table 1.

Task Description. We formalize the long-context instruction-following task as a triplet (I, C, R), where I denotes a concise user instruction specifying the primary goal, $C = \{c_1, \ldots, c_T\}$ represents the long textual context, and R is the desired response. The conventional objective can be written as:

$$P(R \mid I, C),$$

that is, generating an accurate response conditioned on the given instruction and the complete context. This work aims to enhance large language models' ability to understand long contexts by introducing *visual-augmented* representations that exploit structural and glyph information from the visual modality to improve comprehension and robustness in long-text scenarios.

Rendering Pipeline. To explicitly incorporate layout and glyph information into the model, we render the long text C into a sequence of visual pages or images:

$$\mathcal{V} = Render(C; \theta) = \{v_1, \ldots, v_n\},$$

where the rendering configuration is parameterized by a vector θ, which defines typography, layout, and overall visual style. Given context C and configuration θ, the rendering pipeline produces a sequence of images that serve as the long-context input to a vision-language model (VLM).

The rendered visual pages are encoded by a visual encoder into visual token embeddings:

$$V_{emb} = f_V(\mathcal{V}) \in \mathbb{R}^{N_v \times d_v}.$$

Meanwhile, the original text C is encoded by a text encoder to obtain token-level textual embeddings:

$$T_{emb} = f_T(C) \in \mathbb{R}^{N_t \times d_t},$$

preserving fine-grained semantic information at the textual level.

Cross-modal Enhancement and Concatenation Fusion. We leverage the OCR capability of the VLM to obtain text-aligned visual tokens, which are then projected into the textual space to form visually enhanced features. Formally, the projection function is defined as:

$$T_v = \phi(V_{emb}) \in \mathbb{R}^{N_v \times d_t},$$

where $\phi(\cdot)$ denotes a mapping from the visual token space to the text-aligned space.

Finally, we adopt a fusion strategy to combine textual and visual-enhanced representations:

$$\hat{T} = \mathrm{Concat}\big(T_{emb},\ T_v\big) \in \mathbb{R}^{(N_t+N_v) \times d_t},$$

which preserves the original textual representation while explicitly appending visual signals.

Generation Objective. The enhanced and projected text representation $\hat{T}$, together with the instruction I, is fed into the decoder or language model to generate the final response. The overall objective can thus be expressed as:

$$R \sim P(R \mid I, \hat{T}).$$

Table 1: The summary of notations.

Symbol	Description
I	User instruction / prompt
C	Long textual context
R	Target response
θ	Rendering configuration vector (DPI, font, layout, spacing, etc.)
$\mathcal{V} = \{v_1, \ldots, v_n\}$	Sequence of rendered visual pages
$V_{emb} \in \mathbb{R}^{N_v \times d_v}$	Visual token embeddings from the visual encoder
$T_{emb} \in \mathbb{R}^{N_t \times d_t}$	Textual token embeddings from the text encoder
$T_v \in \mathbb{R}^{N_v \times d_t}$	Visual features projected into the text-aligned space
$\hat{T} \in \mathbb{R}^{(N_t+N_v) \times d_t}$	Visually enhanced textual representation
ϕ	Projection function

3.2 Training

Since our method modifies the long-context processing pipeline of Glyph [7], additional training is required to eliminate the mismatch between training and inference. We adopt the following training pipeline.

Continuous Pretraining. We are planing to pretrain our method on ProLong [16] or LongAlign [3] to enhance the model's understanding of long-context inputs and the alignment of visually augmented representations. The main objective of the pretraining stage is to enable the visual encoder and text encoder to align semantic information across large-scale long-text inputs, while adapting to different rendering styles and layouts.

Supervised Fine-Tuning (SFT). Building upon continuous pretraining, we are plaining to perform supervised fine-tuning on the UltraChat [57] dataset to further improve the model's performance on instruction-following tasks. The SFT stage aims to teach the model to generate accurate responses by leveraging visually enhanced text representations, especially for complex or structured long texts.

3.3 Benchmarks and Expected Outcomes

Benchmarks. The model's performance will be evaluated on multiple long-context understanding and reasoning benchmarks, including LongBench [2] , RULER [24], HELMET [54], NIAH [29].

Model Baselines and Expected Outcomes. The goal of our method is to outperform Glyph [7] and GLM-4-9B-base [18] on LongBench [2] and other long-context benchmarks.

References

[1] C. An, S. Gong, M. Zhong, X. Zhao, M. Li, J. Zhang, L. Kong, and X. Qiu. L-eval: Instituting standardized evaluation for long context language models. *arXiv preprint arXiv:2307.11088*, 2023.

[2] Y. Bai, X. Lv, J. Zhang, H. Lyu, J. Tang, Z. Huang, Z. Du, X. Liu, A. Zeng, L. Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*, 2023.

[3] Y. Bai, X. Lv, J. Zhang, Y. He, J. Qi, L. Hou, J. Tang, Y. Dong, and J. Li. Longalign: A recipe for long context alignment of large language models. *arXiv preprint arXiv:2401.18058*, 2024.

[4] Y. Bai, S. Tu, J. Zhang, H. Peng, X. Wang, X. Lv, S. Cao, J. Xu, L. Hou, Y. Dong, et al. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*, 2024.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] Y. Chen, S. Qian, H. Tang, X. Lai, Z. Liu, S. Han, and J. Jia. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*, 2023.

[7] J. Cheng, Y. Liu, X. Zhang, Y. Fei, W. Hong, R. Lyu, W. Wang, Z. Su, X. Gu, X. Liu, et al. Glyph: Scaling context windows via visual-text compression. *arXiv preprint arXiv:2510.17800*, 2025.

[8] R. Child, S. Gray, A. Radford, and I. Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

[9] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[10] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[11] G. DeepMind. Gemini pro. `https://deepmind.google/models/gemini/pro/`. Accessed: 2025-11-09.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[13] Y. Ding, L. L. Zhang, C. Zhang, Y. Xu, N. Shang, J. Xu, F. Yang, and M. Yang. Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*, 2024.

[14] A. Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[15] Y. Fu, R. Panda, X. Niu, X. Yue, H. Hajishirzi, Y. Kim, and H. Peng. Data engineering for scaling language models to 128k context. *arXiv preprint arXiv:2402.10171*, 2024.

[16] T. Gao, A. Wettig, H. Yen, and D. Chen. How to train long-context language models (effectively). In *ACL*, 2025.

[17] S. Ge, Y. Zhang, L. Liu, M. Zhang, J. Han, and J. Gao. Model tells you what to discard: Adaptive kv cache compression for llms. *arXiv preprint arXiv:2310.01801*, 2023.

[18] T. GLM, A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Zhang, D. Rojas, G. Feng, H. Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

[19] Y. Gu, W. Mao, and M. Z. Shou. Long-context autoregressive video modeling with next-frame prediction. *arXiv preprint arXiv:2503.19325*, 2025.

[20] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[21] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, and X. Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.

[22] C. Han, Q. Wang, W. Xiong, Y. Chen, H. Ji, and S. Wang. Lm-infinite: Simple on-the-fly length generalization for large language models. 2023.

[23] C. Hooper, S. Kim, H. Mohammadzadeh, M. W. Mahoney, Y. S. Shao, K. Keutzer, and A. Gholami. Kvquant: Towards 10 million context length llm inference with kv cache quantization. *Advances in Neural Information Processing Systems*, 37:1270–1303, 2024.

[24] C.-P. Hsieh, S. Sun, S. Kriman, S. Acharya, D. Rekesh, F. Jia, Y. Zhang, and B. Ginsburg. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

[25] Y. Huang, B. Yuan, X. Han, C. Xiao, and Z. Liu. Locret: Enhancing eviction in long-context llm inference with trained retaining heads. 2024.

[26] Y. Huang, C. Xiao, X. Han, and Z. Liu. Nosa: Native and offloadable sparse attention. *arXiv preprint arXiv:2510.13602*, 2025.

[27] T. Ji, B. Guo, Y. Wu, Q. Guo, L. Shen, Z. Chen, X. Qiu, Q. Zhang, and T. Gui. Towards economical inference: Enabling deepseek's multi-head latent attention in any transformer-based llms, 2025. URL `https://arxiv.org/abs/2502.14837`.

[28] H. Jiang, Y. Li, C. Zhang, Q. Wu, X. Luo, S. Ahn, Z. Han, A. H. Abdi, D. Li, C.-Y. Lin, et al. Minference 1.0: Accelerating pre-filling for long-context llms via dynamic sparse attention. *Advances in Neural Information Processing Systems*, 37:52481–52515, 2024.

[29] G. Kamradt. Needle in a haystack - pressure testing llms, 2024.

[30] Y. Li, Y. Huang, B. Yang, B. Venkitesh, A. Locatelli, H. Ye, T. Cai, P. Lewis, and D. Chen. Snapkv: Llm knows what you are looking for before generation. *Advances in Neural Information Processing Systems*, 37:22947–22970, 2024.

[31] Y. Li, H. Jiang, Q. Wu, X. Luo, S. Ahn, C. Zhang, A. H. Abdi, D. Li, J. Gao, Y. Yang, et al. Scbench: A kv cache-centric analysis of long-context methods. *arXiv preprint arXiv:2412.10319*, 2024.

[32] Y. Li, T. Zhang, Z. Li, and C. Han. A training-free length extrapolation approach for llms: Greedy attention logit interpolation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8784–8804, 2025.

[33] A. Liu, B. Feng, B. Wang, B. Wang, B. Liu, C. Zhao, C. Dengr, C. Ruan, D. Dai, D. Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024.

[34] J. Liu, D. Zhu, Z. Bai, Y. He, H. Liao, H. Que, Z. Wang, C. Zhang, G. Zhang, J. Zhang, et al. A comprehensive survey on long context language modeling. *arXiv preprint arXiv:2503.17407*, 2025.

[35] Z. Liu, J. Yuan, H. Jin, S. Zhong, Z. Xu, V. Braverman, B. Chen, and X. Hu. Kivi: A tuning-free asymmetric 2bit quantization for kv cache. *arXiv preprint arXiv:2402.02750*, 2024.

[36] E. Lu, Z. Jiang, J. Liu, Y. Du, T. Jiang, C. Hong, S. Liu, W. He, E. Yuan, Y. Wang, et al. Moba: Mixture of block attention for long-context llms. *arXiv preprint arXiv:2502.13189*, 2025.

[37] J. Mu, X. Li, and N. Goodman. Learning to compress prompts with gist tokens. *Advances in Neural Information Processing Systems*, 36:19327–19352, 2023.

[38] B. Peng, J. Quesnelle, H. Fan, and E. Shippole. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023.

[39] C. Qian, W. Liu, H. Liu, N. Chen, Y. Dang, J. Li, C. Yang, W. Chen, Y. Su, X. Cong, et al. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186, 2024.

[40] N. Shang, L. L. Zhang, S. Wang, G. Zhang, G. Lopez, F. Yang, W. Chen, and M. Yang. Longrope2: Near-lossless llm context window scaling. *arXiv preprint arXiv:2502.20082*, 2025.

[41] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[42] H. Sun, L.-W. Chang, W. Bao, S. Zheng, N. Zheng, X. Liu, H. Dong, Y. Chi, and B. Chen. Shadowkv: Kv cache in shadows for high-throughput long-context llm inference. *arXiv preprint arXiv:2410.21465*, 2024.

[43] W. Sun, J. Hu, Y. Zhou, J. Du, D. Lan, K. Wang, T. Zhu, X. Qu, Y. Zhang, X. Mo, et al. Speed always wins: A survey on efficient architectures for large language models. *arXiv preprint arXiv:2508.09834*, 2025.

[44] X. Wang, M. Salmani, P. Omidi, X. Ren, M. Rezagholizadeh, and A. Eshaghi. Beyond the limits: A survey of techniques to extend the context length in large language models. *arXiv preprint arXiv:2402.02244*, 2024.

[45] H. Wei, Y. Sun, and Y. Li. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*, 2025.

[46] H. Xia, C. T. Leong, W. Wang, Y. Li, and W. Li. Tokenskip: Controllable chain-of-thought compression in llms. *arXiv preprint arXiv:2502.12067*, 2025.

[47] C. Xiao, P. Zhang, X. Han, G. Xiao, Y. Lin, Z. Zhang, Z. Liu, and M. Sun. Infllm: Training-free long-context extrapolation for llms with an efficient context memory. *Advances in Neural Information Processing Systems*, 37:119638–119661, 2024.

[48] G. Xiao, Y. Tian, B. Chen, S. Han, and M. Lewis. Efficient streaming language models with attention sinks. *arXiv*, 2023.

[49] G. Xiao, J. Tang, J. Zuo, J. Guo, S. Yang, H. Tang, Y. Fu, and S. Han. Duoattention: Efficient long-context llm inference with retrieval and streaming heads. *arXiv preprint arXiv:2410.10819*, 2024.

[50] R. Xu, G. Xiao, H. Huang, J. Guo, and S. Han. Xattention: Block sparse attention with antidiagonal scoring. *arXiv preprint arXiv:2503.16428*, 2025.

[51] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[52] S. Yang, J. Guo, H. Tang, Q. Hu, G. Xiao, J. Tang, Y. Lin, Z. Liu, Y. Lu, and S. Han. Lserve: Efficient long-sequence llm serving with unified sparse attention. *arXiv preprint arXiv:2502.14866*, 2025.

[53] H. Yen, T. Gao, M. Hou, K. Ding, D. Fleischer, P. Izsak, M. Wasserblat, and D. Chen. Helmet: How to evaluate long-context language models effectively and thoroughly. *arXiv preprint arXiv:2410.02694*, 2024.

[54] H. Yen, T. Gao, M. Hou, D. Ke, D. Fleischer, P. Izsak, M. Wasserblat, and D. Chen. Helmet: How to evaluate long-context language models effectively and thoroughly. In *International Conference on Learning Representations (ICLR) 2025*, 2025.

[55] J. Yuan, H. Gao, D. Dai, J. Luo, L. Zhao, Z. Zhang, Z. Xie, Y. Wei, L. Wang, Z. Xiao, et al. Native sparse attention: Hardware-aligned and natively trainable sparse attention. *arXiv preprint arXiv:2502.11089*, 2025.

[56] X. Zhang, S. Hu, W. Zhao, H. Wang, X. Han, C. He, G. Zeng, Z. Liu, and M. Sun. Optimal rope extension via bayesian optimization for training-free length generalization. *AI Open*, 6: 1–11, 2025.

[57] Y. Zhang, A. Ni, Z. Mao, C. H. Wu, C. Zhu, B. Deb, A. H. Awadallah, D. Radev, and Z. Rui. Enhancing chat language models by scaling high-quality instructional dialogue data. In *Proceedings of EMNLP 2023*, 2023.

[58] Z. Zhang, Y. Sheng, T. Zhou, T. Chen, L. Zheng, R. Cai, Z. Song, Y. Tian, C. Ré, C. Barrett, et al. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*, 36:34661–34710, 2023.

[59] W. Zhao, Z. Zhou, Z. Su, C. Xiao, Y. Li, Y. Li, Y. Zhang, W. Zhao, Z. Li, Y. Huang, et al. Infllm-v2: Dense-sparse switchable attention for seamless short-to-long adaptation. *arXiv preprint arXiv:2509.24663*, 2025.