Novelty Assessment Report

Paper: ProxyAttn: Guided Sparse Attention via Representative Heads
PDF URL: https://openreview.net/pdf?id=m3HXHQYmZu
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

The quadratic complexity of attention mechanisms limits the efficiency of Large Language Models (LLMs) on long-text tasks. Recently, methods that dynamically estimate block importance have enabled efficient block sparse attention, leading to significant acceleration in long-text pre-filling of LLMs. However, their block-level coarse-grained estimation inevitably leads to performance degradation at high sparsity ratios. In this work, we propose ProxyAttn, a training-free sparse attention algorithm that achieves token-level estimation by compressing the dimension of attention heads. Based on our observation of the similarity among multiple attention heads in long texts, we use the attention scores of pooled representative heads to approximate the scores for all heads. To account for the varying sparsity among heads, we also propose a block-aware dynamic budget estimation method. By combining the scores from a set of representative heads with a multi-head dynamic budget, we can achieve a more fine-grained block attention evaluation at a low computational cost. Experiments on a variety of mainstream models and extensive benchmarks confirm the underlying similarity among attention heads in long texts. Leveraging a token-level fine-grained estimation, the proposed method achieves substantial gains in performance and efficiency compared to existing methods. More precisely, ProxyAttn can achieve up to 10.3x attention acceleration and 2.4x prefilling acceleration without significant performance loss.

Core Task Landscape

This paper addresses: **efficient sparse attention for long-context language models**

A total of **50 papers** were analyzed and organized into a taxonomy with **21 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Sparse Attention Pattern Design and Selection**
- **System-Level Acceleration and Implementation**
- **Alternative Attention Mechanisms and Extensions**
- **Training and Adaptation for Long Contexts**
- **Analysis, Benchmarking, and Theoretical Foundations**
- **Domain-Specific and Application-Oriented Methods**

Complete Taxonomy Tree

- efficient sparse attention for long-context language models Survey Taxonomy
- Sparse Attention Pattern Design and Selection
  - Fixed and Structured Sparse Patterns
  - Sliding Window and Local Attention (3 papers)
    - [4] Sliding Window Attention Training for Efficient Large Language Models (Song Wen-tao, 2025) View paper
    - [10] Longlora: Efficient fine-tuning of long-context large language models (Chen, 2023) View paper
    - [31] Sinklora: Enhanced efficiency and chat capabilities for long-context large language models (Zhang Hengyu, 2024) View paper
  - Block-Sparse and Hierarchical Attention (4 papers)
    - [12] Xattention: Block sparse attention with antidiagonal scoring (Xu Ruyi, 2025) View paper
    - [24] 2-D Transformer: Extending Large Language Models to Long-Context With Few Memory (Xingyang He, 2025) View paper
    - [44] Star Attention: Efficient LLM Inference over Long Sequences (Acharya, 2024) View paper
    - [45] Every Token Counts: Generalizing 16M Ultra-Long Context in Large Language Models (Xiang Hu, 2025) View paper
  - Predefined Sparse Patterns for Specific Architectures (2 papers)
    - [25] Longcoder: A long-range pre-trained language model for code completion (Guo, 2023) View paper
    - [43] Big bird: Transformers for longer sequences (Zaheer, 2020) View paper
  - Adaptive and Dynamic Sparse Attention
  - Input-Dependent Dynamic Sparsity (5 papers)
    - [3] Flexprefill: A context-aware sparse attention mechanism for efficient long-sequence inference (Xunhao Lai, 2025) View paper
    - [11] Minference 1.0: Accelerating pre-filling for long-context llms via dynamic sparse attention (Amir Abdi, 2024) View paper
    - [22] DAM: Dynamic Attention Mask for Long-Context Large Language Model Inference Acceleration (Zhang HanZhi, 2025) View paper
    - [28] Sampleattention: Near-lossless acceleration of long context llm inference with adaptive structured sparse attention (Zhu Qian-chao, 2025) View paper
    - [49] Training-free Context-adaptive Attention for Efficient Long Context Modeling (Zeng You, 2025) View paper
  - Top-k and Scoring-Based Selection (4 papers)
    - [5] Sparser is faster and less is more: Efficient sparse attention for long-range transformers (Lou Chao, 2024) View paper

        - [27] SALE : Low-bit Estimation for Efficient Sparse Attention in Long-context LLM Prefilling (ji xiaodong, 2025) View paper
        - [34] Kascade: A Practical Sparse Attention Method for Long-Context LLM Inference (Dhruv Deshmukh, 2025) View paper
        - [46] Hashattention: Semantic sparsity for faster inference (Desai, 2024) View paper
    - Retrieval-Based and Nearest-Neighbor Sparse Attention (2 papers)
        - [15] Retrievalattention: Accelerating long-context llm inference via vector retrieval (Liu Di, 2024) View paper
        - [40] Adamas: Hadamard Sparse Attention for Efficient Long-Context Inference (Yan Siyuan, 2025) View paper
    - Learned and Training-Based Sparse Attention
    - Attention Distillation and Compression (1 papers)
        - [7] Hamming Attention Distillation: Binarizing Keys and Queries for Efficient Long-Context Transformers (Horton, 2025) View paper
    - Learnable Sparse Attention Mechanisms (2 papers)
        - [14] SparseD: Sparse Attention for Diffusion Language Models (Wang ZeQing, 2025) View paper
        - [33] Seerattention: Learning intrinsic sparse attention in your llms (Gao Yizhao, 2024) View paper
    - Hybrid and Multi-Strategy Sparse Attention
    - Mixture and Heterogeneous Sparse Attention ★ (3 papers)
        - [0] ProxyAttn: Guided Sparse Attention via Representative Heads (Anon et al., 2026) View paper
        - [1] Moa: Mixture of sparse attention for automatic large language model compression (Fu Tian-Yu, 2024) View paper
        - [16] Moba: Mixture of block attention for long-context llms (Jiang, 2025) View paper
    - Dense-Sparse Switchable Attention (2 papers)
        - [8] Infllm-v2: Dense-sparse switchable attention for seamless short-to-long adaptation (Zhao Weilin, 2025) View paper
        - [17] Lserve: Efficient long-sequence llm serving with unified sparse attention (Yang Shang, 2025) View paper
    - Multi-Stage and Hierarchical Hybrid Attention (3 papers)
        - [19] MMInference: Accelerating Pre-filling for Long-Context VLMs via Modality-Aware Permutation Sparse Attention (Li Yu-Cheng, 2025) View paper
        - [36] Retrospective Sparse Attention for Efficient Long-Context Generation (Kang, 2025) View paper
        - [39] RocketKV: Accelerating Long-Context LLM Inference via Two-Stage KV Cache Compression (Behnam, 2025) View paper
- System-Level Acceleration and Implementation
    - KV Cache Compression and Management (3 papers)
    - [21] ShadowKV: KV Cache in Shadows for High-Throughput Long-Context LLM Inference (Sun, 2024) View paper
    - [23] Exploiting Sparsity for Long Context Inference: Million Token Contexts on Commodity GPUs (Ryan Synk, 2025) View paper
    - [26] LongSight: Compute-Enabled Memory to Accelerate Large-Context LLMs via Sparse Attention (Derrick Quinn, 2025) View paper
    - Hardware-Aware and Accelerator Design (2 papers)
    - [20] Longer Attention Span: Increasing Transformer Context Length With Sparse Graph Processing Techniques (Nathaniel Tomczak, 2025) View paper
    - [50] H2EAL: Hybrid-Bonding Architecture with Hybrid Sparse Attention for Efficient Long-Context LLM Inference (Guo Xiao-tian, 2025) View paper
    - Efficient Inference Systems and Serving (1 papers)
    - [41] SparseAccelerate: Efficient Long-Context Inference for Mid-Range GPUs (Vo, 2024) View paper
- Alternative Attention Mechanisms and Extensions
    - Linear and Subquadratic Attention Approximations (3 papers)
    - [2] Efficient attention mechanisms for large language models: A survey (Sun Yutao, 2025) View paper
    - [35] LoLA: Low-Rank Linear Attention With Sparse Caching (McDermott, 2025) View paper
    - [47] Scaling Linear Attention with Sparse State Expansion (Pan Yu-qi, 2025) View paper
    - Gated and Enhanced Attention Mechanisms (1 papers)
    - [30] Gated Attention for Large Language Models: Non-linearity, Sparsity, and Attention-Sink-Free (Qiu, 2025) View paper
- Training and Adaptation for Long Contexts
    - Context Extension and Length Generalization (2 papers)
    - [37] Lag-Relative Sparse Attention In Long Context Training (Huang Wanyi, 2025) View paper
    - [38] A Little Goes a Long Way: Efficient Long Context Training and Inference with Partial Contexts (Ge, 2024) View paper
- Analysis, Benchmarking, and Theoretical Foundations
    - Empirical Analysis and Trade-off Studies (2 papers)
    - [6] The sparse frontier: Sparse attention trade-offs in transformer llms (Nawrot, 2025) View paper
    - [32] How Sparse Attention Approximates Exact Attention? Your Attention is Naturally -Sparse (Y Deng, 2024) View paper
    - Benchmarks and Evaluation Frameworks (1 papers)
    - [13] Long range arena: A benchmark for efficient transformers (Yi Tay, 2020) View paper
    - Theoretical Foundations and Mechanistic Analysis (2 papers)
    - [9] Critical attention scaling in long-context transformers (Chen Shi, 2025) View paper
    - [18] Retrieval Head Mechanistically Explains Long-Context Factuality (Wu Wen-Hao, 2024) View paper
- Domain-Specific and Application-Oriented Methods (3 papers)
    - [29] Iterative Sparse Attention for Long-sequence Recommendation (Guanyu Lin, 2025) View paper
    - [42] Long-context inference optimization for large language models: a survey (Tao Wei, 2025) View paper
    - [48] How Much Temporal Long-Term Context is Needed for Action Segmentation? (Emad Bahrami, 2023) View paper

Narrative

Core task: efficient sparse attention for long-context language models. The field has organized itself around several complementary directions. Sparse Attention Pattern Design and Selection explores how to choose which tokens to attend to, ranging from fixed patterns like sliding windows to adaptive strategies that predict importance on the fly. System-Level Acceleration and Implementation focuses on translating these patterns into fast kernels and memory-efficient serving systems, ensuring that theoretical sparsity gains materialize in practice. Alternative Attention Mechanisms and Extensions investigates fundamentally different architectures—such as linear attention or state-space models—that sidestep quadratic complexity altogether. Training and Adaptation for Long Contexts addresses how to extend pretrained models to longer sequences without prohibitive retraining costs, while Analysis, Benchmarking, and Theoretical Foundations provides the empirical testbeds and formal guarantees needed to compare methods rigorously. Finally, Domain-Specific and

Application-Oriented Methods tailor sparse attention to particular use cases like code generation or retrieval-augmented generation, where task structure can guide sparsity choices.

Within the pattern-design branch, a particularly active line of work explores hybrid and mixture-based strategies that combine multiple sparsity heuristics or dynamically route tokens to different attention modules. ProxyAttn[0] exemplifies this trend by using proxy mechanisms to blend several sparse patterns, aiming to capture both local coherence and long-range dependencies without committing to a single fixed structure. This approach contrasts with simpler uniform strategies and aligns closely with recent mixture-of-attention frameworks like Moa[1] and MoBA[16], which also leverage heterogeneous attention heads to balance efficiency and expressiveness. Meanwhile, works such as MInference[11] and SeerAttention[33] emphasize adaptive, query-driven sparsity that predicts important tokens at inference time, trading off some overhead for greater flexibility. The central tension across these methods is whether to rely on lightweight, static patterns that are easy to accelerate or to invest in more sophisticated selection mechanisms that can better preserve model quality as context lengths grow into the millions of tokens.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Moa: Mixture of sparse attention for automatic large language model compression

Authors: Fu Tian-Yu, Tianyu Fu, Huang Haofeng, Haofeng Huang, Ning, et al. (31 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Sparse attention can effectively mitigate the significant memory and throughput demands of Large Language Models (LLMs) in long contexts. Existing methods typically employ a uniform sparse attention mask, applying the same sparse pattern across different attention heads and input lengths. However, this uniform approach fails to capture the diverse attention patterns inherent in LLMs, ignoring their distinct accuracy-latency trade-offs. To address this challenge, we propose the Mixture of Attenti...

Relationship Analysis

Both papers belong to the Mixture and Heterogeneous Sparse Attention category, assigning different sparse patterns to different attention heads or layers. ProxyAttn uses representative proxy heads with pooled scores to guide sparse attention across all heads, combined with dynamic per-head budget allocation based on head sparsity characteristics. In contrast, MoA (Mixture of Attention) automatically searches and assigns distinct sparse attention configurations (patterns and their scaling rules) to different heads and layers through profiling and optimization, adapting patterns based on input sequence length rather than using proxy heads for score estimation.

---

2. Moba: Mixture of block attention for long-context llms

Authors: Jiang, Zhejun, Enzhe Lu, Liu jingyuan, Zhejun Jiang, et al. (55 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Scaling the effective context length is essential for advancing large language models (LLMs) toward artificial general intelligence (AGI). However, the quadratic increase in computational complexity inherent in traditional attention mechanisms presents a prohibitive overhead. Existing approaches either impose strongly biased structures, such as sink or window attention which are task-specific, or radically modify the attention mechanism into linear approximations, whose performance in complex re...

Relationship Analysis

Both papers belong to the Mixture and Heterogeneous Sparse Attention category, assigning different sparse patterns to different heads or layers. They overlap in addressing efficient sparse attention for long-context LLMs by leveraging head-level heterogeneity. However, ProxyAttn uses representative proxy heads with pooling to estimate block importance and applies dynamic per-head budgets, while MoBA applies Mixture of Experts principles to attention, using a gating mechanism to route different heads to different block attention patterns, allowing seamless transitions between full and sparse attention.

Contributions Analysis

Overall novelty summary. ProxyAttn proposes a training-free sparse attention algorithm that achieves token-level importance estimation by compressing attention head dimensions and using pooled representative heads as proxies. The paper sits in the 'Mixture and Heterogeneous Sparse Attention' leaf, which contains only three papers total, including ProxyAttn itself and two siblings (Moa and MoBA). This is a relatively sparse research direction within the broader taxonomy of fifty papers, suggesting the specific approach of heterogeneous head-level strategies remains less explored compared to uniform sparse patterns or adaptive selection methods found in neighboring leaves.

The taxonomy reveals that ProxyAttn's leaf is part of the 'Hybrid and Multi-Strategy Sparse Attention' branch, which sits alongside 'Fixed and Structured Sparse Patterns' and 'Adaptive and Dynamic Sparse Attention' within the broader 'Sparse Attention Pattern Design and Selection' category. Neighboring leaves include 'Dense-Sparse Switchable Attention' (two papers) and 'Multi-Stage and Hierarchical Hybrid Attention' (three papers), indicating that hybrid strategies collectively represent a moderately active area. The taxonomy's scope note clarifies that this leaf focuses on assigning different sparse patterns to different heads or layers, distinguishing it from uniform approaches in adjacent categories like 'Top-k and Scoring-Based Selection' (four papers) or 'Input-Dependent Dynamic Sparsity' (five papers).

Among fifteen candidates examined across three contributions, the block-aware dynamic budget estimation method shows the most substantial prior work overlap: ten candidates were examined, with one appearing to provide refutable prior work. The ProxyAttn algorithm itself examined only two candidates with no clear refutations, while the attention head similarity observation examined three candidates, also without refutations. These statistics reflect a limited semantic search scope rather than exhaustive coverage. The dynamic budget contribution appears to intersect with existing adaptive sparsity methods, whereas the proxy-based head pooling mechanism and the empirical similarity observation may represent more distinctive angles within the examined candidate set.

Based on the limited search of fifteen candidates, ProxyAttn appears to occupy a moderately novel position within a less-crowded research direction. The head-level heterogeneity strategy distinguishes it from more common uniform sparse patterns, though the dynamic budget component shows overlap with prior adaptive methods. The analysis does not cover the full landscape of sparse attention research, and a broader literature search might reveal additional related work, particularly in the adaptive sparsity and mixture-of-experts attention domains that neighbor this taxonomy leaf.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: ProxyAttn sparse attention algorithm

Description: The authors introduce ProxyAttn, a training-free method that estimates block importance for sparse attention by compressing along the attention head dimension rather than the sequence dimension. It uses pooled representative proxy heads to approximate attention scores for all heads, enabling fine-grained block importance evaluation at low computational cost.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Minference 1.0: Accelerating pre-filling for long-context llms via dynamic sparse attention

URL: View paper

Brief Assessment

MInference[11] focuses on identifying three fixed sparse patterns (A-shape, vertical-slash, block-sparse) and assigning them to heads offline, then building dynamic indices within those patterns. ProxyAttn compresses along the attention head dimension using pooled representative proxy heads to estimate block importance across all heads, which is a fundamentally different approach to achieving sparse attention.

---

2. A Unified Sparse Attention via Multi-Granularity Compression

URL: View paper

Brief Assessment

Multi-Granularity Compression[63] focuses on composite tokens and multi-granularity compression for sparse attention, while ProxyAttn specifically compresses along the attention head dimension using pooled representative proxy heads. The technical approaches differ fundamentally in their compression strategies.

Contribution 2: Block-aware dynamic budget estimation method

Description: The authors develop an online method to dynamically allocate different sparsity budgets to individual attention heads based on their varying sparsity characteristics. This allows diverse sparse attention patterns across heads while using unified importance scores from proxy heads.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. SparseMM: Head Sparsity Emerges from Visual Concept Responses in MLLMs

URL: View paper

Brief Assessment

SparseMM[59] focuses on identifying visual heads in MLLMs and allocating budgets based on visual relevance scores, not on dynamic sparsity budget allocation across attention heads in general transformers based on varying sparsity characteristics.

---

2. The sparse frontier: Sparse attention trade-offs in transformer llms

URL: View paper

Brief Assessment

Sparse Frontier[6] focuses on comparing existing training-free sparse attention methods across different sparsity levels and tasks, finding that 'different units of sparsification or budget adaptivity needed for different scenarios.' It does not propose a specific block-aware dynamic budget estimation method for allocating sparsity budgets to individual attention heads.

---

3. Spatten: Efficient sparse attention architecture with cascade token and head pruning

URL: View paper

Brief Assessment

Spatten[55] focuses on cascade token and head pruning for hardware acceleration in NLP attention mechanisms, not on dynamic sparsity budget allocation across attention heads in transformers for long-context LLMs.

---

4. Scene Adaptive Sparse Transformer for Event-based Object Detection

URL: View paper

Brief Assessment

Scene Adaptive Sparse[53] focuses on event-based object detection with window-token co-sparsification for visual transformers, not on dynamic budget allocation across attention heads in language models for long-context processing.

---

5. Chasing Sparsity in Vision Transformers: An End-to-End Exploration

URL: View paper

Brief Assessment

Chasing Sparsity[58] focuses on dynamic sparse subnetwork training for Vision Transformers with structured sparsity in self-attention heads, not on dynamic budget allocation across attention heads for block sparse attention in language models.

---

6. Tactic: Adaptive Sparse Attention with Clustering and Distribution Fitting for Long-Context LLMs

URL: View paper

Prior Art Analysis

Tactic[56] demonstrates prior work on dynamic budget allocation across attention heads based on varying sparsity characteristics. Both papers address the same core problem: allocating different sparsity budgets to individual attention heads based on their varying sparsity patterns. The candidate paper explicitly states that 'prior methods enforce fixed token budgets for sparse attention' and proposes a solution that 'dynamically selects tokens' to address 'variations in the importance of attention across heads, layers, and contexts.' This directly challenges the novelty claim of the original paper's block-aware dynamic budget estimation method, as Tactic[56] already implements adaptive budget allocation that accounts for head-level sparsity variations.

Evidence

Evidence 1 - **Rationale**: Both papers identify the same limitation (fixed budgets fail to account for head-level variations) and propose dynamic budget allocation as the solution. Tactic[56] explicitly addresses 'variations in the importance of attention across heads' through dynamic token selection, which is functionally equivalent to the original paper's 'block-aware budget allocation method' that provides 'different head sparsity budgets online.' - **Original**: considering the diverse sparsity among heads, we propose a block-aware budget allocation method. coupled with the unified importance score, our method can provide different head sparsity budgets online, which in turn yields diverse sparse attention masks. - **Candidate**: prior methods enforce fixed token budgets for sparse attention, assuming a set number of tokens can approximate full attention. however, these methods overlook variations in the importance of attention across heads, layers, and contexts. to address these limitations, we propose tactic, a sparsity-ad...

Evidence 2 - **Rationale**: Both methods perform budget estimation that adapts to attention sparsity patterns. The original paper uses block-level budget estimation, while Tactic[56] uses 'target fraction of total attention scores' to achieve adaptive selection, demonstrating that similar adaptive budget mechanisms existed prior to the original paper. - **Original**: given the significant discrepancy between token-level computation and the actual block budget due to attention sparsity, we employ average pooling to ensure that the budget estimation is performed at the block level. - **Candidate**: by setting a target fraction of total attention scores, tactic ensures that token selection naturally adapts to variations in attention sparsity.

Evidence 3 - **Rationale**: Both papers recognize the same fundamental observation about head-level sparsity variations. The original paper acknowledges this is 'well-established,' and Tactic[56] identifies it as a limitation of prior work that their method addresses, indicating this observation and the need for dynamic budgets was already known in the field. - **Original**: attention heads exhibit variability in sparsity. while the diversity among attention heads is a well-established observation (jiang et al., 2024), we find that it does not contradict our findings of consistency. - **Candidate**: however, these methods overlook variations in the importance of attention across heads, layers, and contexts.

---

7. Trainable dynamic mask sparse attention

URL: View paper

Brief Assessment

Trainable Dynamic Mask[54] focuses on trainable dynamic masks using value vector representations for content-aware attention, not on dynamic budget allocation across attention heads based on their sparsity characteristics.

---

8. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and â□¦

URL: View paper

Brief Assessment

Small Language Models Survey[51] provides only fragmentary mentions of attention mechanisms and sparsity without any technical detail on dynamic budget allocation methods for attention heads. The candidate does not present work that challenges the novelty of the original paper's block-aware dynamic budget estimation approach.

---

9. An integrated multi-head dual sparse self-attention network for remaining useful life prediction

URL: View paper

Brief Assessment

Dual Sparse Self-Attention[52] focuses on multi-head probsparse self-attention for remaining useful life prediction tasks, not on dynamic sparsity budget allocation across attention heads in transformers for long-context language models.

---

10. Dynamic sparse attention for scalable transformer acceleration

URL: View paper

Brief Assessment

Dynamic Sparse Attention[57] focuses on dynamic sparse patterns in attention mechanisms but does not describe a block-aware budget estimation method that allocates different sparsity budgets to individual attention heads based on their varying characteristics.

Contribution 3: Observation of attention head similarity in long contexts

Description: The authors empirically observe and establish that multiple attention heads exhibit consistency in which tokens they focus on in long-context scenarios, with the primary difference being their sparsity levels rather than token focus. This observation forms the theoretical foundation for using proxy heads.

This contribution was assessed against **3 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Attention flows: Analyzing and comparing attention mechanisms in language models

URL: View paper

Brief Assessment

Attention Flows[61] focuses on visualizing attention mechanisms during fine-tuning for classification tasks, not on analyzing attention head similarity patterns in long-context scenarios or establishing consistency in token focus across heads.

---

2. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference

URL: View paper

Brief Assessment

Smarter Better Faster[60] focuses on encoder-only models (BERT-style) for classification and retrieval tasks, not on decoder-only LLMs with long-context attention mechanisms. The architectural contexts and objectives differ fundamentally.

---

3. MuDAF: Long-Context Multi-Document Attention Focusing through Contrastive Learning on Attention Heads

URL: View paper

Brief Assessment

MuDAF[62] focuses on improving retrieval heads through contrastive learning to reduce attention distractions in multi-document QA, rather than analyzing attention head similarity patterns or token focus consistency across heads in long contexts.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] ProxyAttn: Guided Sparse Attention via Representative Heads View paper
- [1] Moa: Mixture of sparse attention for automatic large language model compression View paper
- [2] Efficient attention mechanisms for large language models: A survey View paper
- [3] Flexprefill: A context-aware sparse attention mechanism for efficient long-sequence inference View paper
- [4] Sliding Window Attention Training for Efficient Large Language Models View paper
- [5] Sparser is faster and less is more: Efficient sparse attention for long-range transformers View paper
- [6] The sparse frontier: Sparse attention trade-offs in transformer llms View paper
- [7] Hamming Attention Distillation: Binarizing Keys and Queries for Efficient Long-Context Transformers View paper
- [8] Infllm-v2: Dense-sparse switchable attention for seamless short-to-long adaptation View paper
- [9] Critical attention scaling in long-context transformers View paper
- [10] Longlora: Efficient fine-tuning of long-context large language models View paper
- [11] Minference 1.0: Accelerating pre-filling for long-context llms via dynamic sparse attention View paper
- [12] Xattention: Block sparse attention with antidiagonal scoring View paper

- [13] Long range arena: A benchmark for efficient transformers View paper
- [14] SparseD: Sparse Attention for Diffusion Language Models View paper
- [15] Retrievalattention: Accelerating long-context llm inference via vector retrieval View paper
- [16] Moba: Mixture of block attention for long-context llms View paper
- [17] Lserve: Efficient long-sequence llm serving with unified sparse attention View paper
- [18] Retrieval Head Mechanistically Explains Long-Context Factuality View paper
- [19] MMInference: Accelerating Pre-filling for Long-Context VLMs via Modality-Aware Permutation Sparse Attention View paper
- [20] Longer Attention Span: Increasing Transformer Context Length With Sparse Graph Processing Techniques View paper
- [21] ShadowKV: KV Cache in Shadows for High-Throughput Long-Context LLM Inference View paper
- [22] DAM: Dynamic Attention Mask for Long-Context Large Language Model Inference Acceleration View paper
- [23] Exploiting Sparsity for Long Context Inference: Million Token Contexts on Commodity GPUs View paper
- [24] 2-D Transformer: Extending Large Language Models to Long-Context With Few Memory View paper
- [25] Longcoder: A long-range pre-trained language model for code completion View paper
- [26] LongSight: Compute-Enabled Memory to Accelerate Large-Context LLMs via Sparse Attention View paper
- [27] SALE : Low-bit Estimation for Efficient Sparse Attention in Long-context LLM Prefilling View paper
- [28] Sampleattention: Near-lossless acceleration of long context llm inference with adaptive structured sparse attention View paper
- [29] Iterative Sparse Attention for Long-sequence Recommendation View paper
- [30] Gated Attention for Large Language Models: Non-linearity, Sparsity, and Attention-Sink-Free View paper
- [31] Sinklora: Enhanced efficiency and chat capabilities for long-context large language models View paper
- [32] How Sparse Attention Approximates Exact Attention? Your Attention is Naturally -Sparse View paper
- [33] Seerattention: Learning intrinsic sparse attention in your llms View paper
- [34] Kascade: A Practical Sparse Attention Method for Long-Context LLM Inference View paper
- [35] LoLA: Low-Rank Linear Attention With Sparse Caching View paper
- [36] Retrospective Sparse Attention for Efficient Long-Context Generation View paper
- [37] Lag-Relative Sparse Attention In Long Context Training View paper
- [38] A Little Goes a Long Way: Efficient Long Context Training and Inference with Partial Contexts View paper
- [39] RocketKV: Accelerating Long-Context LLM Inference via Two-Stage KV Cache Compression View paper
- [40] Adamas: Hadamard Sparse Attention for Efficient Long-Context Inference View paper
- [41] SparseAccelerate: Efficient Long-Context Inference for Mid-Range GPUs View paper
- [42] Long-context inference optimization for large language models: a survey View paper
- [43] Big bird: Transformers for longer sequences View paper
- [44] Star Attention: Efficient LLM Inference over Long Sequences View paper
- [45] Every Token Counts: Generalizing 16M Ultra-Long Context in Large Language Models View paper
- [46] Hashattention: Semantic sparsity for faster inference View paper
- [47] Scaling Linear Attention with Sparse State Expansion View paper
- [48] How Much Temporal Long-Term Context is Needed for Action Segmentation? View paper
- [49] Training-free Context-adaptive Attention for Efficient Long Context Modeling View paper
- [50] H2EAL: Hybrid-Bonding Architecture with Hybrid Sparse Attention for Efficient Long-Context LLM Inference View paper
- [51] A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and â View paper
- [52] An integrated multi-head dual sparse self-attention network for remaining useful life prediction View paper
- [53] Scene Adaptive Sparse Transformer for Event-based Object Detection View paper
- [54] Trainable dynamic mask sparse attention View paper
- [55] Spatten: Efficient sparse attention architecture with cascade token and head pruning View paper
- [56] Tactic: Adaptive Sparse Attention with Clustering and Distribution Fitting for Long-Context LLMs View paper
- [57] Dynamic sparse attention for scalable transformer acceleration View paper
- [58] Chasing Sparsity in Vision Transformers: An End-to-End Exploration View paper
- [59] SparseMM: Head Sparsity Emerges from Visual Concept Responses in MLLMs View paper
- [60] Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference View paper
- [61] Attention flows: Analyzing and comparing attention mechanisms in language models View paper
- [62] MuDAF: Long-Context Multi-Document Attention Focusing through Contrastive Learning on Attention Heads View paper
- [63] A Unified Sparse Attention via Multi-Granularity Compression View paper