# Novelty Assessment Report

**Paper**: FSA: An Alternative Efficient Implementation of Native Sparse Attention Kernel
**PDF URL**: https://openreview.net/pdf?id=c5mdo1hWrs
**Venue**: ICLR 2026 Conference Submission
**Year**: 2026
**Report Generated**: 2025-12-27

## Abstract

Recent advance in sparse attention mechanisms has demonstrated strong potential for reducing the computational cost of long-context training and inference in large language models (LLMs). Native Sparse Attention (NSA), one state-of-the-art approach, introduces natively trainable, hardware-aligned sparse attention that delivers substantial system-level performance boost while maintaining accuracy comparable to full attention. However, the kernel implementation of NSA forces a loop order that is only efficient with a relatively large number of query heads in each Grouped Query Attention (GQA) group, whereas existing LLMs widely adopt much smaller number of query heads in each GQA group --- such an inconsistency significantly limits the applicability of this sparse algorithmic advance. In this work, we propose **F**lash **S**parse **A**ttention (**FSA**), an alternative kernel implementation that enables efficient NSA computation across a wide range of popular LLMs with varied smaller number of query heads in each GQA group on modern GPUs. Compared to vanilla NSA kernel implementation, our empirical evaluation demonstrates that FSA achieves (i) up to 3.5x and on average 1.6x kernel-level latency reduction, (ii) up to 1.25x and 1.09x on average end-to-end training speedup on state-of-the-art LLMs, and (iii) up to 1.36x and 1.11x on average for prefill-phase speedup in LLM generative inference.



## Core Task Landscape

This paper addresses: **Efficient Sparse Attention Kernel Implementation for Long-Context Language Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **20 categories**.

### Taxonomy Overview

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

- **Sparse Attention Algorithm Design**
- **Implementation and System Optimization**
- **KV Cache Management and Compression**
- **Training and Fine-Tuning for Long Context**
- **Theoretical Analysis and Empirical Studies**
- **Alternative Efficiency Approaches**
- **Production Systems and Integrated Frameworks**

### Complete Taxonomy Tree

- Efficient Sparse Attention Kernel Implementation for Long-Context Language Models Survey Taxonomy
- Sparse Attention Algorithm Design
  - Dynamic Sparse Pattern Learning
  - Trainable Sparse Attention Mechanisms (4 papers)
    - [2] Native sparse attention: Hardware-aligned and natively trainable sparse attention (Dai, 2025) View paper
    - [23] MoGA: Mixture-of-Groups Attention for End-to-End Long Video Generation (Jia Weinan, 2025) View paper
    - [24] Seerattention: Learning intrinsic sparse attention in your llms (Gao Yizhao, 2024) View paper
    - [38] Trainable Dynamic Mask Sparse Attention (Wu, 2025) View paper
  - Heuristic-Based Dynamic Selection (6 papers)
    - [5] Xattention: Block sparse attention with antidiagonal scoring (Xu Ruyi, 2025) View paper
    - [9] Minference 1.0: Accelerating pre-filling for long-context llms via dynamic sparse attention (Amir Abdi, 2024) View paper
    - [12] Flexprefill: A context-aware sparse attention mechanism for efficient long-sequence inference (Luo Yao, 2025) View paper
    - [17] Sampleattention: Near-lossless acceleration of long context llm inference with adaptive structured sparse attention (Zhu Qian-chao, 2025) View paper
    - [29] DAM: Dynamic Attention Mask for Long-Context Large Language Model Inference Acceleration (Fan Heng, 2025) View paper
    - [32] TokenSelect: Efficient Long-Context Inference and Length Extrapolation for LLMs via Dynamic Token-Level KV Cache Selection (Wei Wu, 2024) View paper
  - Static and Hybrid Sparse Patterns
  - Block-Sparse and Structured Patterns (4 papers)
    - [4] X-former elucidator: reviving efficient attention for long context language modeling (X Miao, 2024) View paper
    - [15] SALO: an efficient spatial accelerator enabling hybrid sparse attention mechanisms for long sequences (Shen Guan, 2022) View paper
    - [20] Longer Attention Span: Increasing Transformer Context Length with Sparse Graph Processing Techniques (Nathaniel Tomczak, 2025) View paper
    - [30] ASADI: Accelerating Sparse Attention Using Diagonal-based In-Situ Computing (Hui-Ze Li, 2024) View paper

- Hierarchical and Multi-Scale Sparse Attention (2 papers)
  - [13] Progressive Sparse Attention: Algorithm and System Co-design for Efficient Attention in LLM Serving (Zhou Qihui, 2025) View paper
  - [25] Dynamic sparse attention for scalable transformer acceleration (Liu Liu, 2022) View paper
- Sparse Attention for Specialized Architectures (7 papers)
- [3] SparseD: Sparse Attention for Diffusion Language Models (Wang ZeQing, 2025) View paper
- [7] Sparse-vDiT: Unleashing the Power of Sparse Attention to Accelerate Video Diffusion Transformers (Chen Pengtao, 2025) View paper
- [35] Iterative Sparse Attention for Long-sequence Recommendation (Guanyu Lin, 2025) View paper
- [36] Audio Sparse-Transformer for Speech Classification (Hassan Salami Kavaki, 2025) View paper
- [39] Campus Abnormal Behavior Recognition With Temporal Segment Transformers (Hai Chuan Liu, 2023) View paper
- [43] SLA: Beyond Sparsity in Diffusion Transformers via Fine-Tunable Sparse-Linear Attention (Zhang Jintao, 2025) View paper
- [47] STRec: Sparse transformer for sequential recommendations (Chengxi Li, 2023) View paper

- Implementation and System Optimization
  - GPU Kernel Implementation and Optimization ★ (5 papers)
  - [0] FSA: An Alternative Efficient Implementation of Native Sparse Attention Kernel (Anon et al., 2026) View paper
  - [16] FlashInfer: Efficient and Customizable Attention Engine for LLM Inference Serving (Ye Zihao, 2025) View paper
  - [28] FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness (Dao, 2022) View paper
  - [41] DynaX: Sparse Attention Acceleration with Dynamic X:M Fine-Grained Structured Pruning (Xiao Xiong, 2025) View paper
  - [49] SparseAccelerate: Efficient Long-Context Inference for Mid-Range GPUs (Vo, 2024) View paper
  - Serving Systems and Inference Optimization (3 papers)
  - [18] Sparseserve: Unlocking parallelism for dynamic sparse attention in long-context llm serving (Zhou Qihui, 2025) View paper
  - [40] Lserve: Efficient long-sequence llm serving with unified sparse attention (Yang Shang, 2025) View paper
  - [42] Sparse Attention across Multiple-context KV Cache (Cao Ziyi, 2025) View paper
- KV Cache Management and Compression
  - KV Cache Selection and Eviction (4 papers)
  - [6] Squeezed attention: Accelerating long context length llm inference (Coleman Hooper, 2025) View paper
  - [11] Sparser is faster and less is more: Efficient sparse attention for long-range transformers (Lou Chao, 2024) View paper
  - [21] Sparq attention: Bandwidth-efficient llm inference (Ribar, 2023) View paper
  - [33] Retrievalattention: Accelerating long-context llm inference via vector retrieval (Liu Di, 2024) View paper
  - Shared and Reusable KV Cache Architectures (1 papers)
  - [8] MoSKA: Mixture of Shared KV Attention for Efficient Long-Sequence LLM Inference (Myunghyun Rhee, 2025) View paper
- Training and Fine-Tuning for Long Context
  - Context Extension via Efficient Fine-Tuning (2 papers)
  - [1] Longlora: Efficient fine-tuning of long-context large language models (Chen, 2023) View paper
  - [31] InfLLM-V2: Dense-Sparse Switchable Attention for Seamless Short-to-Long Adaptation (Zhao Weilin, 2025) View paper
  - Memory-Efficient Training Techniques (1 papers)
  - [48] Mini-Sequence Transformers: Optimizing Intermediate Memory for Long Sequences Training (Anima Anandkumar, 2024) View paper
  - End-Device and Resource-Constrained Training (1 papers)
  - [37] Minicpm4: Ultra-efficient llms on end devices (Xiao, 2025) View paper
- Theoretical Analysis and Empirical Studies
  - Approximation Theory and Convergence Analysis (1 papers)
  - [26] How Sparse Attention Approximates Exact Attention? Your Attention is Naturally -Sparse (Y Deng, 2024) View paper
  - Comparative Empirical Studies (1 papers)
  - [14] The sparse frontier: Sparse attention trade-offs in transformer llms (Nawrot, 2025) View paper
  - Survey and Taxonomy Papers (2 papers)
  - [10] Efficient Attention Mechanisms for Large Language Models: A Survey (Sun Yutao, 2025) View paper
  - [44] Sparse Attention Mechanisms in Large Language Models: Applications, Classification, Performance Analysis, and Optimization (Bai, 2024) View paper
- Alternative Efficiency Approaches
  - Linear and Low-Rank Attention (1 papers)
  - [46] Combiner: Full attention transformer with sparse computation cost (Ren, 2021) View paper
  - State-Space Models and Hybrid Architectures (1 papers)
  - [34] Overcoming Long-Context Limitations of State-Space Models via Context-Dependent Sparse Attention (Zhan, 2025) View paper
  - Context Pruning and Token Reduction (1 papers)
  - [27] Dynamic Context Pruning for Efficient and Interpretable Autoregressive Transformers (Anagnostidis, 2023) View paper
  - General Sparse Transformer Architectures (1 papers)
  - [45] Sparse is enough in scaling transformers (Jaszczur, 2021) View paper
- Production Systems and Integrated Frameworks (3 papers)
  - [19] DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models (DeepSeek-AI, 2025) View paper
  - [22] Accelerated inference with long-sequence transformers on CPUs (Y Mao, 2023) View paper
  - [50] Multivariate long-time series traffic passenger flow prediction using causal convolutional sparse self-attention MTS-Informer (Miaonan Liu, 2023) View paper

## Narrative

Core task: efficient sparse attention kernel implementation for long-context language models. The field has evolved into several interconnected branches that address complementary aspects of scaling transformers to longer sequences. Sparse Attention Algorithm Design explores structured and learned sparsity patterns that reduce computational complexity while preserving model quality, with works ranging from fixed patterns to dynamic, content-aware selection strategies. Implementation and System Optimization focuses on translating these algorithmic ideas into high-performance GPU kernels and system-level optimizations, exemplified by foundational efforts like FlashAttention[28] and specialized frameworks such as FlashInfer[16]. KV Cache Management and Compression tackles memory bottlenecks through quantization, eviction policies, and retrieval-augmented approaches. Training and Fine-Tuning for Long Context addresses the challenge of extending pretrained models to handle longer sequences efficiently, while Theoretical Analysis and

Empirical Studies provide rigorous understanding of sparsity trade-offs. Alternative Efficiency Approaches explore orthogonal methods such as linear attention or state-space models, and Production Systems integrate these techniques into deployable inference engines.

Recent work has intensified around bridging algorithmic sparsity with practical kernel efficiency. Many studies propose dynamic sparse patterns that adapt to input content, trading off selection overhead against attention savings, as seen in works like Minference[9] and SeerAttention[24]. Others emphasize co-design of sparsity and low-level optimizations, such as SparseD[3] and DynaX[41], which tailor kernel implementations to specific sparse structures. FSA[0] sits within the GPU Kernel Implementation and Optimization cluster, focusing on efficient execution of sparse attention primitives. Compared to FlashAttention[28], which pioneered IO-aware dense attention, FSA[0] extends these principles to handle irregular sparse access patterns with minimal overhead. Relative to SparseD[3], which also targets sparse kernel efficiency, FSA[0] may emphasize different sparsity formats or hardware utilization strategies. The central tension across this branch remains balancing the generality of sparse patterns against the predictability required for peak GPU performance.

## Related Works in Same Category

The following **4 sibling papers** share the same taxonomy leaf node with the original paper:

### 1. FlashInfer: Efficient and Customizable Attention Engine for LLM Inference Serving

**Authors**: Ye Zihao, Chen, Lequn, Lai, Ruihang, et al. (18 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

#### Abstract

Transformers, driven by attention mechanisms, form the foundation of large language models (LLMs). As these models scale up, efficient GPU attention kernels become essential for high-throughput and low-latency inference. Diverse LLM applications demand flexible and high-performance attention solutions. We present FlashInfer: a customizable and efficient attention engine for LLM serving. FlashInfer tackles KV-cache storage heterogeneity using block-sparse format and composable formats to optimize...

#### Relationship Analysis

Both papers belong to the GPU Kernel Implementation and Optimization category, focusing on efficient attention kernel designs for long-context scenarios. While the original paper (FSA) optimizes Native Sparse Attention (NSA) kernels by inverting loop order to handle various GQA group settings and reduce padding overhead, the candidate paper (FlashInfer) addresses KV-cache storage heterogeneity through block-sparse formats and provides a customizable attention template with JIT compilation for diverse inference scenarios. The key difference is that FSA specifically targets sparse attention kernel optimization for training and inference with NSA, whereas FlashInfer provides a general-purpose attention engine for LLM serving with focus on KV-cache management and dynamic request handling.

---

### 2. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness

**Authors**: Dao, Tri, Fu, Daniel Y., Ermon, et al. (9 authors total) | **Year/Venue**: 2022 • Neural Information Processing Systems | **URL**: View paper

#### Abstract

Transformers are slow and memory-hungry on long sequences, since the time and memory complexity of self-attention are quadratic in sequence length. Approximate attention methods have attempted to address this problem by trading off model quality to reduce the compute complexity, but often do not achieve wall-clock speedup. We argue that a missing principle is making attention algorithms IO-aware -- accounting for reads and writes between levels of GPU memory. We propose FlashAttention, an IO-awa...

#### Relationship Analysis

Both papers belong to the GPU Kernel Implementation and Optimization category, focusing on low-level kernel designs for attention mechanisms with memory access pattern optimizations. FlashAttention addresses general exact attention computation through IO-aware tiling between GPU HBM and SRAM for dense attention, while FSA specifically targets sparse attention patterns (Native Sparse Attention) by inverting loop orders to handle non-contiguous memory access when query heads per GQA group are limited. The key difference is that FlashAttention optimizes dense full attention with tiling strategies, whereas FSA extends optimization principles to sparse attention kernels with dynamic token selection.

---

### 3. DynaX: Sparse Attention Acceleration with Dynamic X:M Fine-Grained Structured Pruning

**Authors**: Xiao Xiong, Zhaorui Chen, Yue Liang, Minghao Tian, Jiaxing Shang, et al. (7 authors total) | **Year/Venue**: 2025 • International Conference on Architectural Support for Programming Languages and Operating Systems | **URL**: View paper

#### Abstract

Owning to the mechanism of self-attention, Transformers have exhibited incredible performance in a wide range of artificial intelligence tasks. With the growth of sequence length, attention computation with quadratic complexity becomes the bottleneck, and dynamic sparsity is an effective technique to alleviate this problem. However, dynamic attention sparsity for long-sequence tasks suffers from two challenges, i.e., irregular sparse patterns and heavy prediction overhead. To this end, this pape...

#### Relationship Analysis

Both papers belong to the GPU Kernel Implementation and Optimization category, focusing on low-level kernel designs for sparse attention computation. While FSA addresses efficient implementation of Native Sparse Attention (NSA) by optimizing loop ordering and memory access patterns for various GQA group settings, DynaX proposes a different approach using dynamic X:M fine-grained structured pruning with hardware co-design and block scheduling to achieve attention sparsity. The key difference is that FSA optimizes existing NSA sparse patterns through kernel implementation strategies, whereas DynaX introduces a novel dynamic pruning method with corresponding hardware acceleration.

---

### 4. SparseAccelerate: Efficient Long-Context Inference for Mid-Range GPUs

**Authors**: J Vo | **Year/Venue**: 2024 | **URL**: View paper

#### Abstract

As Large Language Models (LLMs) scale to longer context windows, the computational cost of attention mechanisms, which traditionally grows quadratically with input length, presents a critical challenge for real-time and memory-constrained deployments. Existing sparse attention techniques have sought to reduce this complexity, but they often incur significant overhead or compromise accuracy, making them less practical for large contexts on mid-range hardware. In this paper, we introduce SparseAcc...

#### Relationship Analysis

Both papers belong to the GPU Kernel Implementation and Optimization category, focusing on efficient sparse attention kernel designs for long-context language models. SparseAccelerate overlaps with FSA in addressing the computational bottleneck of attention mechanisms through sparse patterns and GPU-level optimizations, both targeting reduced latency and memory usage for long contexts. However, FSA specifically optimizes the Native Sparse Attention (NSA) kernel by inverting loop order to handle varied GQA group

settings, while SparseAccelerate introduces dynamic sparse attention patterns (Triangular, Interval-Slash, Block-Cluster) with a kernel-aware search algorithm, targeting mid-range GPUs and emphasizing Time-To-First-Token reduction starting at 16K tokens.

## Contributions Analysis

**Overall novelty summary.** The paper proposes Flash Sparse Attention (FSA), a GPU kernel implementation optimized for Native Sparse Attention (NSA) that accommodates models with small query head counts per Grouped Query Attention group. It resides in the 'GPU Kernel Implementation and Optimization' leaf, which contains five papers including the original work. This leaf sits within the broader 'Implementation and System Optimization' branch, indicating a moderately populated research direction focused on translating sparse attention algorithms into efficient low-level code. The taxonomy reveals that kernel-level optimization is a distinct but active subfield, separate from algorithmic pattern design and higher-level serving systems.

The taxonomy structure shows that FSA's leaf neighbors include 'Serving Systems and Inference Optimization', which addresses batching and memory management at the system level rather than kernel internals. Sibling papers in the same leaf likely tackle related challenges such as memory access patterns, warp utilization, or register allocation for sparse primitives. The broader 'Implementation and System Optimization' branch connects to 'Sparse Attention Algorithm Design', where methods like NSA define the sparsity patterns that FSA aims to execute efficiently. The taxonomy's scope and exclude notes clarify that FSA belongs strictly to kernel-level concerns, not algorithmic pattern selection or end-to-end serving frameworks.

Among the three contributions analyzed, the literature search examined 24 candidates total, with no refutable pairs identified. The core FSA kernel implementation examined 4 candidates with 0 refutations, while optimizations for non-contiguous memory access and empirical evaluation each examined 10 candidates with 0 refutations. This suggests that within the limited search scope—top-K semantic matches plus citation expansion—no prior work directly overlaps with FSA's specific approach to handling small query head counts in sparse attention kernels. However, the search scale is modest, and the absence of refutations reflects the examined sample rather than exhaustive coverage of all related kernel optimization literature.

Given the limited search scope of 24 candidates, the analysis indicates that FSA addresses a relatively specific gap in sparse attention kernel design. The taxonomy context reveals a moderately active research area with five papers in the same leaf, suggesting that while kernel optimization for sparse attention is an established concern, FSA's focus on small query head counts may represent a narrower technical contribution. The lack of refutable candidates among examined papers does not guarantee absolute novelty but suggests that FSA's particular optimization strategy is not prominently addressed in the top semantic matches and their citations.

---

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

### Contribution 1: Flash Sparse Attention (FSA) kernel implementation

**Description**: FSA is an alternative kernel implementation for Native Sparse Attention that inverts the loop order compared to vanilla NSA. It processes KV blocks in the outer loop and query tokens in the inner loop, eliminating padding inefficiencies when GQA groups contain few query heads, thereby enabling efficient sparse attention across diverse modern LLMs.

This contribution was assessed against **4 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

#### 1. GSAformer: Group sparse attention transformer for functional brain network analysis

**URL**: View paper

**Brief Assessment**

GSAformer[56] focuses on functional brain network analysis using group sparse attention for neuroimaging data, not on sparse attention kernel implementations for LLMs with grouped query attention.

---

#### 2. Flash Sparse Attention: An Alternative Efficient Implementation of Native Sparse Attention Kernel

**URL**: View paper

**Brief Assessment**

Flash Sparse Attention[57] is the same work as the original paper being evaluated. Both describe FSA as an alternative kernel implementation that inverts the loop order of Native Sparse Attention (NSA), processing KV blocks in the outer loop and query tokens in the inner loop to eliminate padding inefficiencies with small GQA groups. This is not a prior work that could refute novelty.

---

#### 3. Faster video diffusion with trainable sparse attention

**URL**: View paper

**Brief Assessment**

Faster Video Diffusion[55] focuses on video diffusion transformers with hierarchical coarse-fine sparse attention for video generation, not on grouped query attention efficiency in general LLMs. The technical approaches differ fundamentally in their application domains and architectural designs.

---

#### 4. Evolving Sparsity: Leveraging Token Importance Dynamics for Efficient LLM Decoding with Sparse Attention

**URL**: View paper

**Brief Assessment**

Evolving Sparsity[58] focuses on token selection mechanisms (cross-layer propagation and cross-step accumulation) for sparse attention during decoding, not on kernel implementation optimizations for grouped query attention. The candidate addresses algorithmic token importance dynamics rather than low-level kernel loop ordering and padding efficiency that FSA targets.

---

### Contribution 2: Optimizations for non-contiguous memory access and reduction

**Description**: The authors introduce specialized optimizations to handle challenges arising from FSA's inverted loop order, including index tensors for non-contiguous query token access, a separate reduction kernel to avoid atomic operations, and an online softmax kernel to ensure numerical correctness across distributed partial results.

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. AttentionRC: A Novel Approach to Improve Locality Sensitive Hashing Attention on Dual-Addressing Memory

**URL**: View paper

**Brief Assessment**

AttentionRC[67] focuses on Locality Sensitive Hashing attention with dual-addressing memory optimizations, which is a different attention mechanism than the Native Sparse Attention (NSA) framework discussed in the original paper. The technical approaches and problem domains are distinct.

---

### 2. ChunkAttention: Efficient Attention on KV Cache with Chunking Sharing and Batching

**URL**: View paper

**Brief Assessment**

ChunkAttention[66] addresses non-contiguous memory access in a different context (chunked KV cache in prefix trees for multi-tenant serving) rather than the inverted loop order challenges in FSA's sparse attention kernel. The technical approaches differ fundamentally.

---

### 3. DiffKV: Differentiated Memory Management for Large Language Models with Parallel KV Compaction

**URL**: View paper

**Brief Assessment**

DiffKV[64] addresses non-contiguous memory access in the context of differentiated KV cache compression with strided access patterns, while the original paper's FSA handles non-contiguous query token batching in sparse attention with index tensors and separate reduction kernels for online softmax.

---

### 4. TokenSelect: Efficient Long-Context Inference and Length Extrapolation for LLMs via Dynamic Token-Level KV Cache Selection

**URL**: View paper

**Brief Assessment**

TokenSelect[32] addresses non-contiguous attention sparsity at the token level for KV cache selection, but does not discuss kernel-level optimizations for non-contiguous memory access patterns, index tensors, or reduction kernels as described in the original paper's FSA implementation.

---

### 5. HARDSEA: Hybrid Analog-ReRAM Clustering and Digital-SRAM In-Memory Computing Accelerator for Dynamic Sparse Self-Attention in Transformer

**URL**: View paper

**Brief Assessment**

HARDSEA[62] focuses on hardware accelerator design for sparse attention using ReRAM and SRAM computing-in-memory architectures, not on software kernel optimizations for non-contiguous memory access patterns or reduction operations in GPU implementations.

---

### 6. Efficient deformable convnets: Rethinking dynamic and sparse operator for vision applications

**URL**: View paper

**Brief Assessment**

Efficient Deformable ConvNets[59] focuses on optimizing deformable convolution operators for vision tasks through memory access optimization and removing softmax normalization. The candidate does not address sparse attention mechanisms or the specific challenges of FSA's inverted loop order with index tensors and separate reduction kernels for attention operations.

---

### 7. Stateful large language model serving with pensieve

**URL**: View paper

**Brief Assessment**

Pensieve[61] addresses non-contiguous memory access in the context of multi-turn conversation caching across requests, not for sparse attention patterns within a single request. The technical focus differs fundamentally from FSA's sparse attention kernel optimizations.

---

### 8. Off-tanet: A lightweight neural micro-expression recognizer with optical flow features and integrated attention mechanism

**URL**: View paper

**Brief Assessment**

Off TANet[65] focuses on micro-expression recognition using optical flow features and attention mechanisms for facial analysis, not on optimizing memory access patterns or reduction operations in attention kernels for large language models.

---

### 9. Lightweight video denoising using aggregated shifted window attention

**URL**: View paper

**Brief Assessment**

Lightweight Video Denoising[63] focuses on video denoising using shifted window attention for computer vision applications, not on optimizing memory access patterns or reduction operations in attention kernels for LLMs.

---

### 10. Rethinking space-time networks with improved memory coverage for efficient video object segmentation

**URL**: View paper

**Brief Assessment**

Space Time Networks[60] focuses on video object segmentation with correspondence networks and memory reading mechanisms. The paper does not address GPU kernel optimizations, non-contiguous memory access patterns, atomic operations, or online softmax kernels for attention mechanisms - these are distinct technical domains.

---

## Contribution 3: Empirical evaluation demonstrating FSA performance gains

**Description**: The authors provide comprehensive empirical evaluation across kernel-level and end-to-end scenarios, demonstrating that FSA achieves substantial speedups over vanilla NSA and full attention in training and inference on state-of-the-art LLMs with various GQA configurations.

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. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model

**URL**: View paper

**Brief Assessment**

DeepSeek V2[52] focuses on MoE architecture with Multi-head Latent Attention (MLA) for KV cache compression, not on sparse attention kernel implementations or performance comparisons with NSA and full attention mechanisms.

---

### 2. Rectified Sparse Attention

**URL**: View paper

**Brief Assessment**

Rectified Sparse Attention[53] focuses on sparse decoding with periodic dense rectification for long-sequence generation, not on kernel-level implementation optimizations for NSA across varied GQA configurations as in the original paper.

---

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

**URL**: View paper

**Brief Assessment**

Sparse Frontier[14] focuses on comparing different sparse attention methods (e.g., vertical-slash, block-sparse, snapkv, quest) across diverse tasks and model configurations, rather than proposing a new kernel implementation like FSA. The candidate evaluates training-free sparse attention trade-offs, not kernel-level optimizations for NSA.

---

### 4. Seerattention: Learning intrinsic sparse attention in your llms

**URL**: View paper

**Brief Assessment**

SeerAttention[24] focuses on learning intrinsic sparse attention patterns through a gating mechanism for LLMs, while the original paper presents FSA as an alternative kernel implementation for Native Sparse Attention (NSA). The candidate does not demonstrate prior work on FSA's specific kernel-level optimizations or its performance comparisons with vanilla NSA implementations.

---

### 5. Native sparse attention: Hardware-aligned and natively trainable sparse attention

**URL**: View paper

**Brief Assessment**

Native Sparse Attention[2] focuses on a different sparse attention architecture (hierarchical token compression with blockwise selection) rather than the FSA kernel implementation. While both papers evaluate sparse attention performance, they address distinct technical approaches and optimization strategies.

---

### 6. Sparse sinkhorn attention

**URL**: View paper

**Brief Assessment**

Sparse Sinkhorn[54] focuses on learning sparse attention through differentiable sorting mechanisms for memory efficiency, not on optimizing grouped query attention (GQA) configurations or kernel-level implementations for native sparse attention as FSA does.

---

### 7. Trainable Dynamic Mask Sparse Attention

**URL**: View paper

**Brief Assessment**

Trainable Dynamic Mask[38] focuses on a different sparse attention mechanism (dynamic mask with content-aware and position-aware components) rather than the FSA kernel implementation optimizations for NSA that are evaluated in the original paper.

---

### 8. Sparser is faster and less is more: Efficient sparse attention for long-range transformers

**URL**: View paper

**Brief Assessment**

Sparser is Faster[11] focuses on sparse attention through learned key-value selection mechanisms (Sparse K attention), not on optimizing Native Sparse Attention (NSA) kernel implementations across different GQA configurations as FSA does.

---

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

**URL**: View paper

**Brief Assessment**

MoA[51] focuses on heterogeneous sliding-window attention patterns across different heads and input lengths, not on sparse attention kernel implementations like FSA. The candidate addresses attention span optimization rather than kernel-level performance improvements for grouped query attention configurations.

---

### 10. Longlora: Efficient fine-tuning of long-context large language models

**URL**: View paper

**Brief Assessment**

LongLoRA[1] focuses on efficient fine-tuning of long-context LLMs using shifted sparse attention during training, not on comparing sparse attention performance versus full attention in general language model training scenarios.

---

## Appendix: Text Similarity Detection

Textual similarity detection checked 28 papers and found 3 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

### 1. Flash Sparse Attention: An Alternative Efficient Implementation of Native Sparse Attention Kernel

**Detected in**: Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

## References


- [0] FSA: An Alternative Efficient Implementation of Native Sparse Attention Kernel View paper
- [1] Longlora: Efficient fine-tuning of long-context large language models View paper
- [2] Native sparse attention: Hardware-aligned and natively trainable sparse attention View paper
- [3] SparseD: Sparse Attention for Diffusion Language Models View paper
- [4] X-former elucidator: reviving efficient attention for long context language modeling View paper
- [5] Xattention: Block sparse attention with antidiagonal scoring View paper
- [6] Squeezed attention: Accelerating long context length llm inference View paper
- [7] Sparse-vDiT: Unleashing the Power of Sparse Attention to Accelerate Video Diffusion Transformers View paper
- [8] MoSKA: Mixture of Shared KV Attention for Efficient Long-Sequence LLM Inference View paper

- [9] Minference 1.0: Accelerating pre-filling for long-context llms via dynamic sparse attention View paper
- [10] Efficient Attention Mechanisms for Large Language Models: A Survey View paper
- [11] Sparser is faster and less is more: Efficient sparse attention for long-range transformers View paper
- [12] Flexprefill: A context-aware sparse attention mechanism for efficient long-sequence inference View paper
- [13] Progressive Sparse Attention: Algorithm and System Co-design for Efficient Attention in LLM Serving View paper
- [14] The sparse frontier: Sparse attention trade-offs in transformer llms View paper
- [15] SALO: an efficient spatial accelerator enabling hybrid sparse attention mechanisms for long sequences View paper
- [16] FlashInfer: Efficient and Customizable Attention Engine for LLM Inference Serving View paper
- [17] Sampleattention: Near-lossless acceleration of long context llm inference with adaptive structured sparse attention View paper
- [18] Sparseserve: Unlocking parallelism for dynamic sparse attention in long-context llm serving View paper
- [19] DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models View paper
- [20] Longer Attention Span: Increasing Transformer Context Length with Sparse Graph Processing Techniques View paper
- [21] Sparq attention: Bandwidth-efficient llm inference View paper
- [22] Accelerated inference with long-sequence transformers on CPUs View paper
- [23] MoGA: Mixture-of-Groups Attention for End-to-End Long Video Generation View paper
- [24] Seerattention: Learning intrinsic sparse attention in your llms View paper
- [25] Dynamic sparse attention for scalable transformer acceleration View paper
- [26] How Sparse Attention Approximates Exact Attention? Your Attention is Naturally -Sparse View paper
- [27] Dynamic Context Pruning for Efficient and Interpretable Autoregressive Transformers View paper
- [28] FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness View paper
- [29] DAM: Dynamic Attention Mask for Long-Context Large Language Model Inference Acceleration View paper
- [30] ASADI: Accelerating Sparse Attention Using Diagonal-based In-Situ Computing View paper
- [31] InfLLM-V2: Dense-Sparse Switchable Attention for Seamless Short-to-Long Adaptation View paper
- [32] TokenSelect: Efficient Long-Context Inference and Length Extrapolation for LLMs via Dynamic Token-Level KV Cache Selection View paper
- [33] Retrievalattention: Accelerating long-context llm inference via vector retrieval View paper
- [34] Overcoming Long-Context Limitations of State-Space Models via Context-Dependent Sparse Attention View paper
- [35] Iterative Sparse Attention for Long-sequence Recommendation View paper
- [36] Audio Sparse-Transformer for Speech Classification View paper
- [37] Minicpm4: Ultra-efficient llms on end devices View paper
- [38] Trainable Dynamic Mask Sparse Attention View paper
- [39] Campus Abnormal Behavior Recognition With Temporal Segment Transformers View paper
- [40] Lserve: Efficient long-sequence llm serving with unified sparse attention View paper
- [41] DynaX: Sparse Attention Acceleration with Dynamic X:M Fine-Grained Structured Pruning View paper
- [42] Sparse Attention across Multiple-context KV Cache View paper
- [43] SLA: Beyond Sparsity in Diffusion Transformers via Fine-Tunable Sparse-Linear Attention View paper
- [44] Sparse Attention Mechanisms in Large Language Models: Applications, Classification, Performance Analysis, and Optimization View paper
- [45] Sparse is enough in scaling transformers View paper
- [46] Combiner: Full attention transformer with sparse computation cost View paper
- [47] STRec: Sparse transformer for sequential recommendations View paper
- [48] Mini-Sequence Transformers: Optimizing Intermediate Memory for Long Sequences Training View paper
- [49] SparseAccelerate: Efficient Long-Context Inference for Mid-Range GPUs View paper
- [50] Multivariate long-time series traffic passenger flow prediction using causal convolutional sparse self-attention MTS-Informer View paper
- [51] Moa: Mixture of sparse attention for automatic large language model compression View paper
- [52] DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model View paper
- [53] Rectified Sparse Attention View paper
- [54] Sparse sinkhorn attention View paper
- [55] Faster video diffusion with trainable sparse attention View paper
- [56] GSAformer: Group sparse attention transformer for functional brain network analysis View paper
- [57] Flash Sparse Attention: An Alternative Efficient Implementation of Native Sparse Attention Kernel View paper
- [58] Evolving Sparsity: Leveraging Token Importance Dynamics for Efficient LLM Decoding with Sparse Attention View paper
- [59] Efficient deformable convnets: Rethinking dynamic and sparse operator for vision applications View paper
- [60] Rethinking space-time networks with improved memory coverage for efficient video object segmentation View paper
- [61] Stateful large language model serving with pensieve View paper
- [62] HARDSEA: Hybrid Analog-ReRAM Clustering and Digital-SRAM In-Memory Computing Accelerator for Dynamic Sparse Self-Attention in Transformer View paper
- [63] Lightweight video denoising using aggregated shifted window attention View paper
- [64] DiffKV: Differentiated Memory Management for Large Language Models with Parallel KV Compaction View paper
- [65] Off-tanet: A lightweight neural micro-expression recognizer with optical flow features and integrated attention mechanism View paper
- [66] ChunkAttention: Efficient Attention on KV Cache with Chunking Sharing and Batching View paper
- [67] AttentionRC: A Novel Approach to Improve Locality Sensitive Hashing Attention on Dual-Addressing Memory View paper