MAPSPARSE: ACCELERATING PRE-FILLING FOR LONG-CONTEXT VLMS VIA MODALITY-AWARE PERMUTATION SPARSE ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

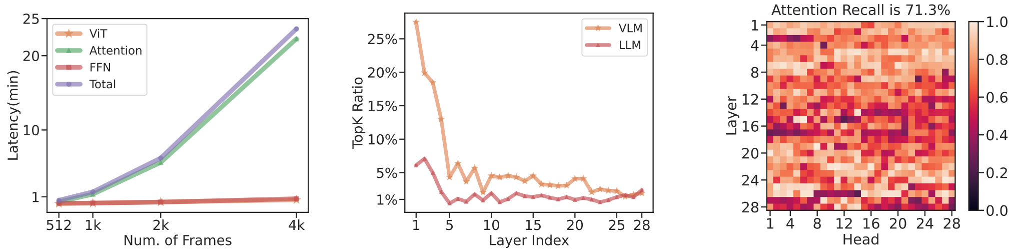
The integration of long-context capabilities with visual understanding opens up new possibilities for Vision Language Models (VLMs). However, the quadratic attention complexity during the pre-filling stage remains a major bottleneck, restricting wide deployment in real-world applications. To address this, we propose MAPSparse (Modali- ty-Aware Permutation Sparse Attention), a dynamic sparse attention method that accelerates the pre-filling stage for long-context multi-modal inputs. First, our analysis reveals that the temporal and spatial locality of video input leads to a unique sparse patterns, the *Grid pattern*. Simultaneously, VLMs exhibit markedly different sparse distributions across different modalities. We introduce a permutation-based method to leverage the unique Grid pattern and handle modality boundaries issue. By offline searching the optimal sparse patterns for each head, MAPSparse constructs the sparse distribution dynamically based on the input. We also provide optimized GPU kernels for efficient sparse computations. Notably, MAPSparse integrates seamlessly into existing VLM pipelines without any model modifications or fine-tuning. Experiments on multimodal benchmarks—including Video QA, Captioning, Vision-NIAH, and Mix Modality-NIAH—with state-of-the-art long-context VLMs (LongVila and Llava-Video) show that MAPSparse accelerates the pre-filling stage by up to 8.3× at 1M tokens while maintaining competitive performance.

1 INTRODUCTION

Scaling the context size of Vision Language Models (VLMs) allows them to handle extended temporal information from long video and text inputs, which is crucial for various applications including robotics (Black et al., 2024; Prasad et al., 2024; Cheang et al., 2024), autonomous driving (Hu et al., 2023; Wang et al., 2023; Gao et al., 2024), and healthcare (Liu et al., 2024). In addition, Zhang et al. (2024b) and Xue et al. (2024) shows that scaling the context size of VLMs can improve the resolution in the temporal dimension and leads to better performance in video understanding tasks.

However, due to the quadratic complexity of attention, the processing of the long multi-modal inputs (i.e., the pre-fill stage) can take minutes before the auto-regressive decoding. As shown in Fig. 1a, this leads to significant Time-to-First-Token latency, which hinders the wide adoption of long-context VLMs in real-world applications. Previous work (Child et al., 2019; Liu et al., 2022; Deng et al., 2024) reveals that attention matrices are typically sparse, prompting the development of sparse attention methods such as Sparse Transformer (Child et al., 2019), Swin Transformer (Liu et al., 2021), and StreamingLLM(Xiao et al., 2024). More recently, MInference (Jiang et al., 2024) propose to use dynamic sparse attention that estimate the sparse index online, and leverage optimized GPU kernels for end-to-end acceleration. However, these methods fail to exploit the unique sparse patterns in long-context VLMs, and they struggle with mixed or interleaved modalities—limiting their direct application without compromising performance.

Unlike long-text contexts, video and image inputs in VLMs exhibit spatiotemporal locality, forming grid-like attention patterns of evenly spaced vertical and horizontal lines (Fig. 2a). In mixed-modality inputs, clear modality boundaries emerge: attention between different modalities diverges



(a) VLMs' attention has heavy cost. (b) VLMs' attention is sparse. (c) Sparsity of VLMs is dynamic.

Figure 1: (a) Latency breakdown of the pre-filling stage, with 256 tokens per frame. (b) How much element in attention needs to be computed to achieve 95% recall in a 128k context. (c) Low attention recall when reusing the top-k indices from a different request. Visualizations are based on LongVILA-7B-1M (Xue et al., 2024) with a single A100.

significantly from intra-modality attention (Fig. 2b). These factors pose unique challenges for exploiting sparsity to accelerate the pre-fill stage.

In this paper, we present MAPSparse, a permutation-based dynamic sparse attention approach that significantly reduces attention FLOPs, accelerating the pre-fill stage of long-context VLMs. First, MAPSparse identifies the grid heads and leverages a *row- and column-wise permutation* to gather the sparse grid for efficient hardware computing. Next, we detect Query-boundary and 2D-boundary patterns to address inter-modality boundaries, and apply *modality-wise permutation* to gather intra-modality regions. This results in a consecutive sparse index within each modality, permitting efficient hardware implementation of sparse computing. Finally, a *Modality-Aware Sparse Attention Search Algorithm* is devised to fine-tune both inter- and intra-modality patterns offline, optimizing performance with minimal overhead.

We conduct extensive experiments using two state-of-the-art long-context VLMs, Llava-Video (Zhang et al., 2024b) and LongVila (Xue et al., 2024), across diverse video understanding tasks such as video captioning (Maaz et al., 2024), video question answering (Yu et al., 2019; Xiao et al., 2021; Mangalam et al., 2023; Fu et al., 2024), and video information retrieval (Zhang et al., 2024a). Additionally, we propose the Mixed-Modality Needle in a Hackathon task to assess multi-modal input performance. Our method effectively addresses modality boundaries, significantly accelerates the prefilling stage, and maintains high accuracy. With a 1M-length context, it achieves speedups of up to 8.3× and 1.7× over FlashAttention-2 and MInference, respectively.

2 Attention Heads in VLMs

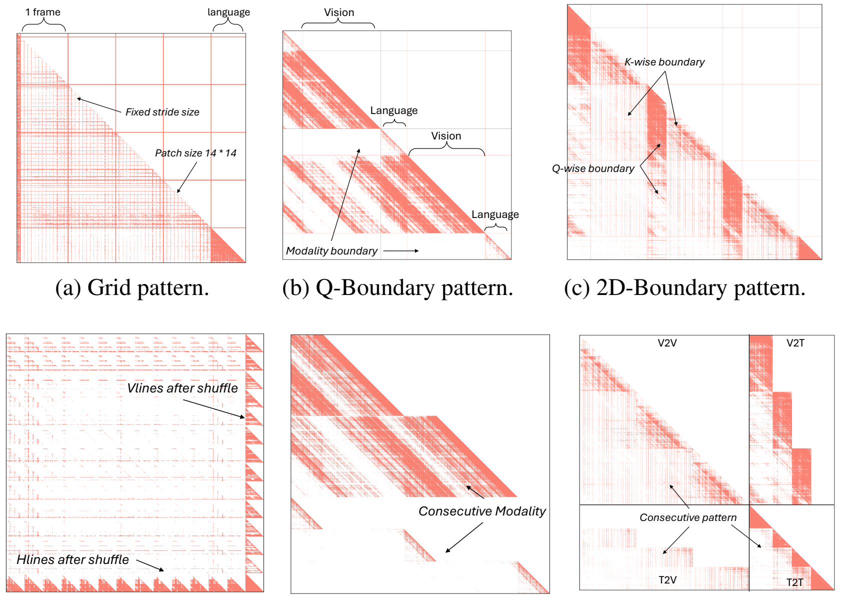
The sparsity of the attention operation in pre-trained text-only LLMs, particularly in long-context scenarios, has been extensively studied (Wu et al., 2024; Ribar et al., 2024; Jiang et al., 2024; Li et al., 2024b), with only 3% of attention weights being activated while achieving a recall rate of 96.8%. Similarly, VLMs also demonstrate notable dynamic sparsity in long-context scenarios. This section examines the shared and distinct properties of text-only LLMs and multi-modal LLMs in long-context scenarios, focusing on attention sparsity, sparse patterns, and modality boundaries.

2.1 Multi-modality Attention is Dynamically Sparse

As illustrated in Fig. 1a, for a 128k × 128k attention matrix in VLMs, retaining only the top 5.78% of attention weights on average suffices to recall 95% of total attention, indicating that each token attends to a limited subset despite long sequences. However, VLMs exhibit lower sparsity than text-only LLMs, where only 1.79% of weights achieve a 95% recall rate. Notably, the bottom layers in VLMs (e.g., the first four layers in LongVila) show reduced sparsity. Yet, due to variability across attention heads, 52.3% of heads in VLMs require less than 2% of attention to be recalled. This highlights substantial computational redundancy in VLMs, especially in long-context scenarios.

On the other hand, similar to LLMs, while the sparse nature of attention matrices remains consistent across inputs, the specific distributions of sparse attention are highly dynamic. As shown in Fig. 1c, reusing the top-k indices for 95% attention recall (derived from Fig. 1b) in a different context leads to a significant drop in performance.

2.2 The Grid Head in VLMs



(a) Grid pattern. (b) Q-Boundary pattern. (c) 2D-Boundary pattern.

(d) Permuted Grid pattern. (e) Permuted Q-Boundary pattern. (f) Permuted 2D-Boundary pattern.

Figure 2: Visualization of pre- vs. post-permutation sparsity attention patterns in VLMs.

In long-context language modeling, efficient attention mechanisms like sliding window attention (Jiang et al., 2023) and StreamingLLM (Xiao et al., 2024) exploit the locality property of text sequences. However, multi-modal inputs introduce unique geometric structures that redefine locality. As shown in Child et al. (2019), image patches exhibit locality along both vertical and horizontal directions, forming local window and slash-like patterns. Similarly, video inputs maintain locality across temporal and spatial dimensions, with frame-based sampling yielding more regular and predictable patterns.

We observe that certain VLM attention heads exhibit a **grid pattern**. While the grid's stride and starting position vary with context, the horizontal and vertical lines are evenly spaced and often symmetrical—a distinct behavior compared to text-only LLMs (Jiang et al., 2024). Fig. 2a visualizes a grid head, demonstrating how local tokens in temporal and spatial dimensions are evenly distributed within the attention map, with focus primarily on these local tokens.

2.3 Modality Boundaries in Multi-Modal Input

The input format of VLMs differs significantly from text-only LLMs. A dedicated vision encoder generates visual representations, which are processed alongside text embeddings by the LLM. Despite pretraining on large-scale datasets, the interaction and processing patterns between modalities vary considerably, leading to distinct modality boundaries in attention (Tu et al., 2024), as illustrated in Fig. 2b and 2c.

Specifically, we observe two key characteristics: 1) Intra-modality consistency: Attention within each modality follows a consistent pattern. For instance, the vision region in Fig. 2b exhibits a clear slash pattern, where critical elements are effectively clustered. 2) Modality-separated continuity: Patterns within a modality can be interrupted by boundaries from other modalities. In Fig. 2b, vision slashes are segmented by the boundary introduced by the language region.

We categorize the modality boundary patterns of VLMs into four distinct types: No-Boundary, K-Boundary, Q-Boundary, and 2D-Boundary, as illustrated in Figs. 2 and 3. 1) **No Boundary** and

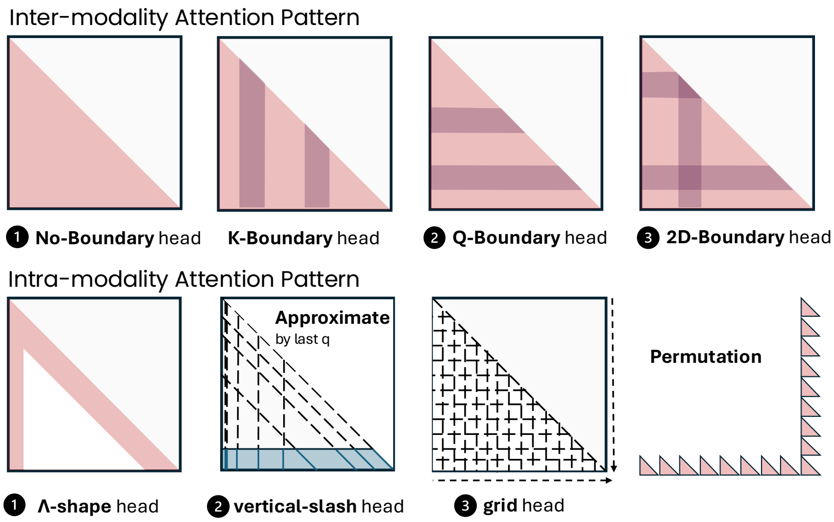


Figure 3: The framework of MAPSparse, including inter- and intra-modality sparse.

K-Boundary exhibit either no clear modality boundary or a boundary only along the key dimension, as shown in Fig. 8. Since continuity is maintained along the query dimension, these heads can be efficiently handled using intra-modality sparse patterns. 2) **Q-Boundary** refers to attention modality boundaries across the query dimension. For example, in Fig. 2b, sparse patterns like Text-to-Video and Video-to-Video appear interconnected, forming a trapezoidal structure, while a clear boundary separates Visual-to-Text and Text-to-Visual attention. 3) **2D-Boundary** occurs when modality boundaries are present in both query and key dimensions. As shown in Fig. 2c, the 2D modality boundary segments attention weights into distinct blocks. Additionally, our analysis of Audio LMs (Chu et al., 2024) and multimodal LMs (Li et al., 2025) reveals that the cross-modality boundary phenomenon persists across these architectures. These boundaries pose unique challenges and hinder the direct application of existing sparse attention approaches to multi-modal inputs.

2.4 Sparse Distributions Continuity Across Boundaries

Although sparsity patterns in VLMs are often discontinuous across modalities due to modality boundaries, we find that sparsity distributions can remain continuous across these boundaries and extrapolate to other regions of the same modality. For example, in Fig. 2b, the slash lines maintain the same relative position across different areas of the vision modality. In a more complex case, Fig. 2c shows interleaved vision and text modalities forming a mixed structure. However, by spatially aggregating regions of the same modality, we observe that sparsity distributions can extend beyond local areas and often exhibit global extrapolation potential. The upper-left region in Fig. 2c exemplifies this, where the grid pattern, initially separated by textual boundaries, becomes consecutive after spatial clustering in both row and column dimensions. To validate this observation, we conducted a quantitative attention recall experiment on mixed-modality inputs, as detailed in §4.6.

3 MAPSparse

Following the analysis in §2, we propose MAPSparse to accelerate the pre-filling stage of long-context VLMs as shown in Fig. 3. The framework consists of three modules, covering both inter- and intra-modality sparse patterns: 1) the novel Grid sparse attention, together with A-shape and Vertial-Slash pattern (Jiang et al., 2024) forms the intra-modality attention; 2) Q-Boundary and 2D-Boundary mix-modality pattern; 3) Modality-aware sparse attention search algorithm. We first do offline pattern search to identify different patterns for each attention head. Then we use online dynamic sparse approximation to build the sparse index, and finally we do dynamic sparse computation using optimized GPU kernels.

3.1 Grid Head in Multi-Modality

To better leverage the inductive bias in visual modalities (e.g., images, videos) and the vertical and horizontal structures in attention, we propose a permutation-based dynamic sparse attention for grid head, as shown in Algo. 1.

Specifically, we first perform an online search to determine the stride and phase of the grid pattern. Since only a view operation is applied to the approximate attention matrix $\hat{\boldsymbol{A}}$, the actual latency overhead remains minimal. Next, we use the identified grid stride and phase to permute the $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ tensors and compute sparse attention accordingly (see Fig. 2d). In our implementation, instead of explicitly permuting $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$, we optimize computational efficiency by dynamically loading and writing these tensors within the kernel, minimizing the overhead of tensor transposition operations. In addition to Grid sparse attention, we also employ A-shape and Vertical-Slash attention for intra-modality operations, see §C.3 for more details.

Algorithm 1 Grid Head

```
Input: $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{S \times d_h}$, stride space $s_g \in \phi_g$
# Approximate stride and phase (last_q = 64)
$\hat{\boldsymbol{A}} \leftarrow \mathrm{softmax}\left(\boldsymbol{Q}_{[-\mathrm{last\_q}:]}\boldsymbol{K}^\top/\sqrt{d} + \boldsymbol{m}_{\mathrm{casual}}\right)$
# Online search grid stide and phase
$\boldsymbol{b}_r, \leftarrow 0$
for $i \leftarrow 1$ to $|\phi_g|$ do
    if $\max(\mathrm{view}(\hat{\boldsymbol{A}}, s_{g,i})) > \boldsymbol{b}_r$ then
        $s_g \leftarrow s_{g,i}, p_g \leftarrow \mathrm{argmax}(\mathrm{view}(\hat{\boldsymbol{A}}, s_{g,i}))$
        $\boldsymbol{b}_r \leftarrow \max(\mathrm{view}(\hat{\boldsymbol{A}}, s_{g,i}))$
    end if
end for
# Permute Q, K, V tensors
$\overline{\boldsymbol{Q}}, \overline{\boldsymbol{K}}, \overline{\boldsymbol{V}} \leftarrow \mathrm{permute}\left(\boldsymbol{Q}\right), \mathrm{permute}\left(\boldsymbol{K}\right), \mathrm{permute}\left(\boldsymbol{V}\right)$
# Final dynamic sparse attention scores w/ FlashAttention (only the last and rightmost block)
$\boldsymbol{A} \leftarrow \mathrm{softmax}\left(\mathrm{sparse}(\overline{\boldsymbol{Q}\boldsymbol{K}}^\top, s_g, p_g)/\sqrt{d}\right)$
# Sparse mixed scores and values
$\boldsymbol{y} \leftarrow \mathrm{sparse}(\boldsymbol{A}\overline{\boldsymbol{V}}, s_g, p_g)$
return $\boldsymbol{y}$
```

3.2 Hybrid Modality Sparse Attention

As analyzed in §2 and illustrated in Fig. 2, modality boundaries exist in multi-modal LLMs. We classify these boundaries into four patterns: No-Boundary, K-Boundary, Q-Boundary, and 2D-Boundary. As the sparse index is continuous along the query dimension for both the No-Boundary and K-Boundary heads, we can directly apply the three intra-modality attention globally. However, for Q-Boundary and 2D-Boundary heads, MAPSparse uses a permutation-based approach to efficiently handle these modality boundaries.

Q-Boundary Head As shown in Fig.2b, Fig.2e, and §2.4, the Q-Boundary pattern shows a clear separation across modality, but the sparse distribution remains continues within each modality.

Algorithm 2 Q-Boundary Head

```
Input: $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{S \times d_h}$, modality type index $\boldsymbol{i}_m$, modality type set $m \in \phi_m$
# Permute Q tensors based on modality
$\overline{\boldsymbol{Q}} \leftarrow \mathrm{permute}\left(\boldsymbol{Q}, \boldsymbol{i}_m\right)$
# Looping over the modalities in query dimension
$\boldsymbol{y} \leftarrow \boldsymbol{0}$
for $i \leftarrow 1$ to $|\phi_m|$ do
    # Intra-modality sparse attention computation for each modality w/ FlashAttention
    $\boldsymbol{A}_{mi} \leftarrow \mathrm{softmax}\left(\mathrm{sparse}(\overline{\boldsymbol{Q}}_{mi}\boldsymbol{K}^\top, \boldsymbol{i}_{mi})/\sqrt{d}\right)$
    $\boldsymbol{y}_{mi} \leftarrow \mathrm{sparse}(\boldsymbol{A}_{mi}\boldsymbol{V})$
    # Update the modality output to the final output
    $\boldsymbol{y} \leftarrow \boldsymbol{y}_{mi} \cup \boldsymbol{y}$
end for
return $\boldsymbol{y}$
```

Algorithm 3 2D-Boundary Head

```
Input: $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{S \times d_h}$, modality type index $\boldsymbol{i}_m$, modality type set $m \in \phi_m$
# Permute Q, K, V tensors based on modality
$\overline{\boldsymbol{Q}} \leftarrow \mathrm{permute}\left(\boldsymbol{Q}, \boldsymbol{i}_m\right), \overline{\boldsymbol{K}} \leftarrow \mathrm{permute}\left(\boldsymbol{K}, \boldsymbol{i}_m\right)$
$\overline{\boldsymbol{V}} \leftarrow \mathrm{permute}\left(\boldsymbol{V}, \boldsymbol{i}_m\right)$
# Looping over the modalities in pairs
$\boldsymbol{y} \leftarrow \boldsymbol{0}$
for $i \leftarrow 1$ to $|\phi_m|$ do
    for $j \leftarrow 1$ to $|\phi_m|$ do
        # Dynamic sparse attention computation for each modality pair w/ FlashAttention
        $\boldsymbol{m}_{mi,mj} \leftarrow \mathrm{buildmask}(\boldsymbol{i}_{mi}, \boldsymbol{i}_{mj})$
        $\boldsymbol{A}_{mi,mj} \leftarrow \mathrm{softmax}(\mathrm{sparse}(\overline{\boldsymbol{Q}}_{mi}\overline{\boldsymbol{K}}_{mj}^\top, \boldsymbol{i}_{mi}, \boldsymbol{i}_{mj})/\sqrt{d} + \boldsymbol{m}_{mi,mj})$
        $\boldsymbol{y}_{mi,mj} \leftarrow \mathrm{sparse}(\boldsymbol{A}_{mi,mj}\overline{\boldsymbol{V}}_{mj})$
        # Update the modality output to the final output
        $\boldsymbol{y} \leftarrow \boldsymbol{y}_{mi,mj} \cup \boldsymbol{y}$
    end for
end for
return $\boldsymbol{y}$
```

Table 1: Performance (%) of different models and different methods on video understanding tasks evaluated at frames from 110 to 256.

Model	**FLOPs**	**VideoDC**	**ActNet-QA**	**EgoSchema**	**Next-QA**	**PerceptionTest**	**VideoMME**		**Avg.**
		test	test	test	mc	val	wo/ sub.	w/ sub.	
Llava-Video-7B # Frames: 110; Total # tokens: 20,240									
Full Attention	100%	3.66	59.6	57.0	81.2	66.1	64.7	71.0	57.6
SF-fixed	4.8%	3.26	57.3	53.3	79.8	62.9	59.9	67.1	54.8
SF-strided	41.4%	3.45	58.5	56.1	80.6	64.4	61.4	68.5	56.1
A-shape	48.2%	3.56	56.0	51.6	79.8	65.7	54.4	65.6	53.8
Tri-shape	49.0%	3.58	59.3	54.5	80.3	66.1	63.6	70.1	56.7
VisionZip	35.2%	1.35	42.1	40.5	69.5	41.4	44.9	62.1	43.1
MInference	78.8%	3.64	59.6	57.0	80.6	66.1	64.6	71.0	57.5
Ours	47.3%	3.58	**59.8**	**57.1**	80.1	**66.2**	64.5	**71.8**	**57.6**
LongVILA-7B # Frames: 256; Total # tokens: 65,800									
Full Attention	100%	2.76	59.5	61.9	80.7	58.1	60.1	65.1	55.5
SF-fixed	2.2%	1.99	51.3	59.6	76.5	55.5	57.1	63.0	52.1
SF-strided	26.6%	2.58	56.0	61.4	76.7	55.5	53.6	59.2	52.2
A-shape	29.1%	2.75	56.6	60.9	75.0	55.3	49.1	59.6	51.3
Tri-shape	29.3%	2.63	58.1	62.0	77.8	56.2	59.3	63.3	54.2
VisionZip	OOM								
MInference	47.0%	2.77	59.7	62.2	79.1	57.8	60.0	65.2	55.2
Ours	31.8%	**2.84**	**60.2**	62.2	**79.4**	57.8	60.0	**65.5**	**55.4**

Building on this insight, we propose a row-wise permutation (Algorithm 2) that groups tokens of the same modality by permuting $\boldsymbol{Q}$, and then applies offline-optimized sparse attention (A-shape, Vertical-Slash, and Grid Head) for intra-modality processing. Noted that we leverage the final segment of *each modality's* queries to dynamically approximate the sparse indices and extrapolate to the entire modality. This method enables flexibility in handling fragmented multi-modality inputs. Additionally, instead of explicitly permuting tensors, our implementation performs dynamic loading and writing inside the kernel for optimized efficiency.

2D-Boundary Head Beyond Query-Boundary, there are attention heads that exhibits modality boundaries in both query and key dimensions, as shown in Fig. 2c. Given a query token, its attention to key tokens from different modalities vary significantly, and queries from different modalities focus on keys in highly diverse patterns. To address 2D modality boundaries, we design a 2D permutation approach, that groups $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ according to their modalities. This allows us to leverage intra-modality continuity to handle each part of the the 2D boundary pattern separately and efficiently. We further illustrate this approach in Fig. 2f and detailed in Algorithm 3. Specifically, we perform permutation on both row- and column-wise for $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$, then iteratively traverse each modality pair to compute dynamic sparse attention. This 2D-Boundary will require constructing an attention mask and searching for sparse patterns in cross-modality regions. For example, in Fig. 2f, we build modality boundary indices for Vision-to-Text (bottom-left) and Text-to-Vision (upper-right) attention. This mask index construction is implemented in Triton (Tillet et al., 2019).

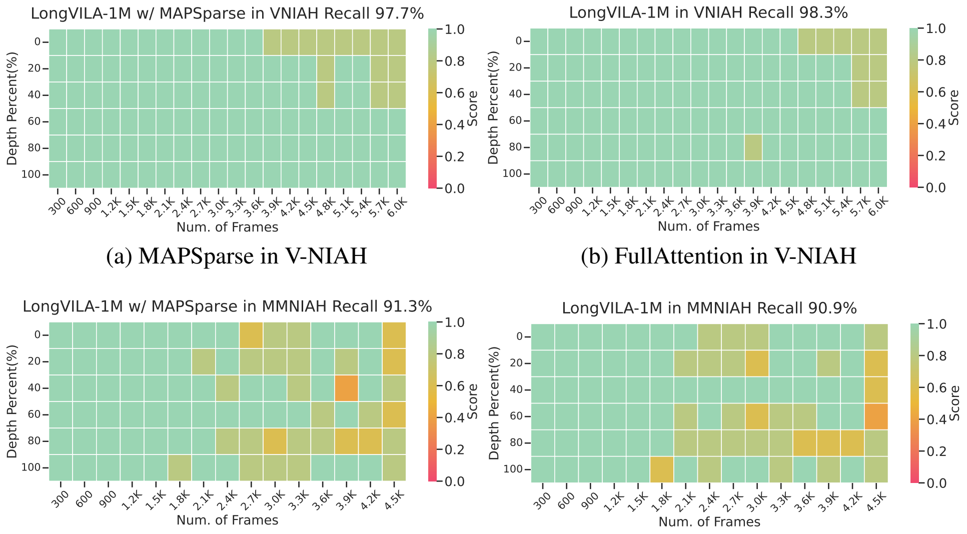
3.3 Modality-Aware Sparse Attention Search Algorithm

Due to modality boundaries in VLMs, we propose a modality-aware sparse attention pattern search algorithm (see Algorithm 4). The process unfolds in three steps: 1) intra-modality search within each modality following (Jiang et al., 2024), 2) cross-modality search across all modality pairs, and 3) inter-modality search informed by the results of the first two steps.

4 Experiments

4.1 Dataset and Baselines

Implementation Details Our experiments are conducted on two state-of-the-art long-video VLMs: Llava-Video (Zhang et al., 2024b) and LongVILA (Xue et al., 2024). We follow the MInference experimental setup, configuring the corresponding search space while adopting optimal configurations from prior work for other methods. We adjust the local window sizes of A-shape and



(a) MAPSparse in V-NIAH

(b) FullAttention in V-NIAH

(c) MAPSparse in MM-NIAH

(d) FullAttention in MM-NIAH

Figure 4: V-NIAH (Zhang et al., 2024a) and MM-NIAH results using LongVila-Qwen2-7B-1M (Xue et al., 2024).

tri-shape patterns to align FLOPs with our method. For MInference, we adopt its optimal configuration, which results in FLOPs approximately twice as high in VLMs compared to our method. Our implementation leverages Triton (Tillet et al., 2019), FlashAttention (Dao, 2024), and dynamic sparse compiler PIT (Zheng et al., 2023). For the Vertical-Slash and Grid Head patterns, we set $last_q = 64$. Latency experiments are performed on a single NVIDIA A100 using bfloat16, with greedy decoding to ensure stable results. Additional implementation details are provided in §C.

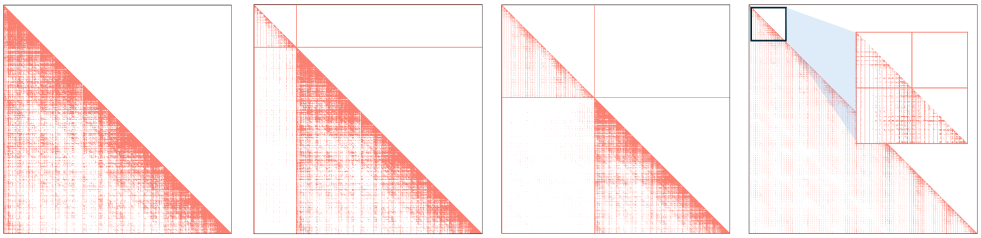
Dataset Our evaluation uses the official metrics and scripts provided by these tasks. Additionally, we introduce a Mix Modality Needle in a Haystack (MM-NIAH) task to assess VLMs' retrieval capabilities on mixed-modality inputs. Dataset details are provided in §D.

(i) Video Understanding Tasks: This includes VideoDC (Lab, 2024), ActNet-QA (Yu et al., 2019), EgoSchema (Mangalam et al., 2023), Next-QA (Xiao et al., 2021), PerceptionTest (Patraucean et al., 2024), and VideoMME (Fu et al., 2024). These benchmarks span five categories, covering tasks such as captioning and video question answering. Input lengths range from 110 frames (e.g., 20k) to 256 frames (e.g., 66k) in Llava-Video (Zhang et al., 2024b) and LongVILA (Xue et al., 2024).

(ii) Video Needle in a Haystack (V-NIAH) (Zhang et al., 2024a): A long-video retrieval task testing VLMs' performance with tokens of up to 6k frames (e.g., 1.1M tokens), where inserted images are placed at various positions.

(iii) Mix Modality Needle in a Haystack (MM-NIAH): To evaluate VLMs in mixed-modality scenarios, we construct a mix-modality version of NIAH. Specifically, 25% of the input consists of text segments inserted at the document level across different frames in long-video inputs, forming a mix-modality haystack. All other settings align with V-NIAH, including the multi-choice VQA task with randomly inserted images. This input lengths of benchmark up to 4.5k frames (e.g., 1.1M).

Baselines We include five training-free sparse attention approaches, one visual token compression method, and also incorporate FlashAttention-2 (Dao, 2024) as a baseline. 1) Spars Transformer (Fixed) (Child et al., 2019): Retains attention within each segment and allows all tokens to attend to the segment's initial tokens. 2) SparseTransformer (Strided) (Child et al., 2019): Employs local windows with dilated attention. 3) A-Shape (Xiao et al., 2024): Preserves only the sink token with local attention. 4) Tri-Shape (Li et al., 2024a; Acharya et al., 2024): Extends A-Shape by enabling full attention for all tokens to the last window's queries. 5) Vertical-Slash Pattern (Jiang et al., 2024): Focuses on specific tokens (vertical lines) and tokens at fixed intervals (slash lines). 6) VisionZip (Yang et al., 2024): A visual token compression method that reduces the number of visual



(a) All Textual Context (b) Inserted Visual (c) More Visual Context (d) All Visual Context

Figure 5: Transition of sparse patterns from textual context to visual context. (a) The vertical-slash pattern for all textual context. (b) Grid pattern appears when visual modality is appended. (c) Grid pattern dominates.

tokens per frame by evaluating tokens based on their attention scores and discarding less important ones. Full details on implementation, hyperparameters, and illustrations can be found in §C.

4.2 Long Video Understanding

Table 1 presents the performance of different methods on video understanding tasks. The results show that: 1) Our method and MInference closely approximate full attention across all tasks while requiring only half the FLOPs of MInference. 2) Static sparse patterns, such as A-shape and Tri-shape, perform reasonably well on most tasks but experience a notable performance drop in multi-choice VQA tasks like EgoSchema. Additionally, the slight increase in query full attention in Tri-shape effectively improves performance. 3) Among SF patterns, the slash pattern preserves more performance. Even when using SF-fixed with only 2%-5% of FLOPs, it still maintains strong performance on most tasks.

4.3 Video Needle In A Haystack

Fig. 4a, 4b, and 12 show the performance of different models on V-NIAH, revealing notable differences in handling long-context video retrieval as the number of processed frames increases: 1) Our method achieves results nearly identical to full attention. 2) A-shape struggles with mid-context information even at 300 frames, while Tri-shape maintains full performance until 3.9k frames (i.g. 700K tokens) before a sharp decline. 3) SF-fixed degrades at 2.1k frames (i.g. 350K tokens), while SF-strided surpasses Tri-shape, holding performance until 4.5k frames (i.g. 825K tokens). 4) MInference preserves VLM retrieval well, with only slight degradation beyond 4.8K frames.

4.4 Mix Modality Needle In A Haystack

Beyond V-NIAH, we introduce a mixed-modality NIAH test to evaluate the performance of different sparse methods on video-text inputs, in Fig. 4c, 4d, and 13. Mixed-modality inputs lead to more pronounced performance degradation across all methods. However, by incorporating inter-modality sparse patterns, our method maintains performance close to full attention, especially when compared to MInference and ours w/o inter-modality. Notably, Tri-shape and MInference show significant drops at 1.8k frames (i.g. 440K tokens) and 2.7k frames (i.g. 660K tokens).

4.5 Latency

Fig. 6 and 14 present end-to-end and kernel-level latency across different context sizes. The grid pattern significantly reduces the sparsity of the vertical-slash pattern, achieving a 2–3× speedup even at 1M tokens. Additionally, the grid pattern achieves an end-to-end speedup of 8.3× and a kernel-level speedup of 12×.

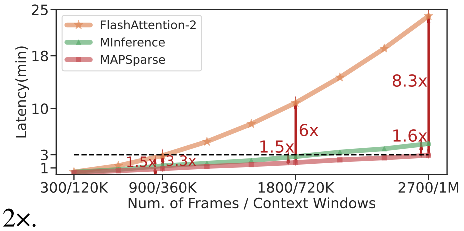


Figure 6: End-to-End Latency.

4.6 Analysis

Transition of Sparse Patterns Across Modalities Since LLMs and VLMs exhibit different sparse patterns, we examine the interplay between the Grid and Vertical-Slash pattern. As shown in Fig. 5, Llava-Video-7B primarily uses VS pattern for purely textual inputs. However, once a visual input is appended, it transitions to a Grid pattern to capture the geometric structure of the visual content. This shift occurs at the modality boundary, creating a more structured arrangement of vertical and horizontal intervals. Such behavior highlights the need for distinct sparsity strategies in visual and mixed-modality contexts, rather than simply reusing sparse patterns from LLMs for VLMs.

Sparse Index Across Modalities In Fig. 7, the sparse index achieves high recall for textual regions but fails to generalize to visual ones. To address this, we construct a sparse index from the visual modality and evaluate it on separate visual segments, each separated by modality boundaries. Remarkably, this approach extrapolates effectively across all visual segments, even when interspersed with textual boundaries. As shown in Fig. 7, the sparse index achieves high recall in the textual but fails to generalize to the visual. To address this, we construct a sparse index using the visual modality and evaluate it across distinct regions of the visual modality, separated by modality boundaries. Remarkably, this approach successfully extrapolates to all visual regions even when interrupted by text-induced boundaries.

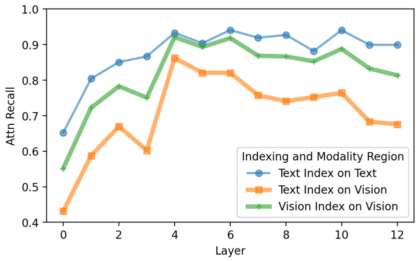


Figure 7: The sparse index does not effectively extrapolate from text to the visual modality. However, an index built within the same modality can generalize across modality boundaries.

5 Related Work

Long-Context Vision Language Models Recent VLMs have extended their context length to handle long multi-modal inputs (Zhang et al., 2024a; Xue et al., 2024; Wang et al., 2024b; Reid et al., 2024), enabling applications like long-video understanding (Fu et al., 2024; Xiao et al., 2021; Wang et al., 2024a), multi-modal information retrieval (Zhang et al., 2024a), and multi-modal chain-of-thought (Qwen, 2024). For example, Zhang et al. (2024a) leverage base LLMs' inherent long-context capacity for visual transfer learning, and Xue et al. (2024) propose multi-modal sequence parallelism to accelerate video fine-tuning. Zhang et al. (2024b) further highlight the importance of data calibration and synthetic data for improving VLM performance.

Efficiency Optimization for VLMs Despite their strong accuracy, long-context VLMs face high inference costs, limiting their deployment in long-video tasks. A common strategy is *vision token compression*—reducing video feature resolution by dropping or merging less important visual tokens (Bolya et al., 2023; Chen et al., 2024; Shen et al., 2024; He et al., 2024; Tu et al., 2024; Weng et al., 2024; Wen et al., 2024). RNN-Transformer hybrids have also been explored (Wang et al., 2024b) to balance efficiency and context length. However, these methods often assume inputs are long videos plus short text, focusing only on visual token optimization and overlooking mixed-modality inputs critical for multi-turn interactions (Huang et al., 2024).

6 Conclusion

We propose MAPSparse, a modality-aware permutation sparse attention method that accelerates long-context VLMs. It features permutation-based grid sparse attention, Q-boundary/2D-boundary patterns for mixed-modality boundaries, and a Modality-Aware Sparse Attention Search Algorithm. Our optimized GPU kernels enable end-to-end acceleration. Experiments on video understanding tasks, V-NIAH and MM-NIAH using Llava-Video and LongVila demonstrate that MAPSparse preserves full-attention performance while achieving up to 8.3× speedup at 1M tokens.

REFERENCES

Shantanu Acharya, Fei Jia, and Boris Ginsburg. Star attention: Efficient llm inference over long sequences. *CoRR*, 2024. doi: 10.48550/ARXIV.2411.17116.

Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Rich Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pre-trained image-editing diffusion models. *ICLR*, 2024.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *ICLR*, 2023.

Chilam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, Hanbo Zhang, and Minzhao Zhu. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *CoRR*, 2024. doi: 10.48550/ARXIV.2410.06158.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. *ECCV*, pp. 19–35, 2024. doi: 10.1007/978-3-031-73004-7_2.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *ArXiv preprint*, abs/1904.10509, 2019. URL https://arxiv.org/abs/1904.10509.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *ArXiv preprint*, abs/2407.10759, 2024. URL https://arxiv.org/abs/2407.10759.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *ICLR*, 2024.

Yichuan Deng, Zhao Song, and Chiwun Yang. Attention is naturally sparse with gaussian distributed input. *CoRR*, 2024. doi: 10.48550/ARXIV.2404.02690.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *CoRR*, 2024. doi: 10.48550/ARXIV.2405.21075.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *CoRR*, 2024. doi: 10.48550/ARXIV.2405.17398.

Yefei He, Feng Chen, Jing Liu, Wenqi Shao, Hong Zhou, Kaipeng Zhang, and Bohan Zhuang. Zipvl: Efficient large vision-language models with dynamic token sparsification and kv cache compression. *CoRR*, 2024. doi: 10.48550/ARXIV.2410.08584.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *CoRR*, 2023. doi: 10.48550/ARXIV.2309.17080.

Minbin Huang, Yanxin Long, Xinchi Deng, Ruihang Chu, Jiangfeng Xiong, Xiaodan Liang, Hong Cheng, Qinglin Lu, and Wei Liu. Dialoggen: Multi-modal interactive dialogue system for multi-turn text-to-image generation. *CoRR*, 2024. doi: 10.48550/ARXIV.2403.08857.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *ArXiv preprint*, abs/2310.06825, 2023. URL https://arxiv.org/abs/2310.06825.

Huiqiang Jiang, Yucheng Li, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir H. Abdi, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. MInference 1.0: Accelerating pre-filling for long-context LLMs via dynamic sparse attention. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL `https://openreview.net/forum?id=fPBACAbqSN`.

LMMs Lab. Video detail caption, 2024. URL `https://huggingface.co/datasets/lmms-lab/VideoDetailCaption`.

Yadong Li, Jun Liu, Tao Zhang, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, Chong Li, Yuanbo Fang, Dongdong Kuang, Mingrui Wang, Chenglin Zhu, Youwei Zhang, Hongyu Guo, Fengyu Zhang, Yuran Wang, Bowen Ding, Wei Song, Xu Li, Yuqi Huo, Zheng Liang, Shusen Zhang, Xin Wu, Shuai Zhao, Linchu Xiong, Yozhen Wu, Jiahui Ye, Wenhao Lu, Bowen Li, Yan Zhang, Yaqi Zhou, Xin Chen, Lei Su, Hongda Zhang, Fuzhong Chen, Xuezhen Dong, Na Nie, Zhiying Wu, Bin Xiao, Ting Li, Shunya Dang, Ping Zhang, Yijia Sun, Jincheng Wu, Jinjie Yang, Xionghai Lin, Zhi Ma, Kegeng Wu, Jia li, Aiyuan Yang, Hui Liu, Jianqiang Zhang, Xiaoxi Chen, Guangwei Ai, Wentao Zhang, Yicong Chen, Xiaoqin Huang, Kun Li, Wenjing Luo, Yifei Duan, Lingling Zhu, Ran Xiao, Zhe Su, Jiani Pu, Dian Wang, Xu Jia, Tianyu Zhang, Mengyu Ai, Mang Wang, Yujing Qiao, Lei Zhang, Yanjun Shen, Fan Yang, Miao Zhen, Yijie Zhou, Mingyang Chen, Fei Li, Chenzheng Zhu, Keer Lu, Yaqi Zhao, Hao Liang, Youquan Li, Yanzhao Qin, Linzhuang Sun, Jianhua Xu, Haoze Sun, Mingan Lin, Zenan Zhou, and Weipeng Chen. Baichuan-omni-1.5 technical report. *CoRR*, 2025. doi: 10.48550/ARXIV.2501.15368.

Yucheng Li, Huiqiang Jiang, Qianhui Wu, Xufang Luo, Surin Ahn, Chengruidong Zhang, Amir H. Abdi, Dongsheng Li, Jianfeng Gao, Yuqing Yang, and Lili Qiu. Scbench: A kv cache-centric analysis of long-context methods. *CoRR*, 2024a. doi: 10.48550/ARXIV.2412.10319.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. SnapKV: LLM knows what you are looking for before generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL `https://openreview.net/forum?id=poE54GOq2l`.

Lei Liu, Xiaoyan Yang, Junchi Lei, Xiaoyang Liu, Yue Shen, Zhiqiang Zhang, Peng Wei, Jinjie Gu, Zhixuan Chu, Zhan Qin, and Kui Ren. A survey on medical large language models: Technology, application, trustworthiness, and future directions. *CoRR*, 2024. doi: 10.48550/ARXIV.2406.03712.

Liu Liu, Zheng Qu, Zhaodong Chen, Fengbin Tu, Yufei Ding, and Yuan Xie. Dynamic sparse attention for scalable transformer acceleration. *IEEE Trans. Computers*, pp. 3165–3178, 2022. doi: 10.1109/TC.2022.3208206.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *ICCV*, pp. 9992–10002, 2021. doi: 10.1109/ICCV48922.2021.00986.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *ACL*, pp. 12585–12602, 2024. doi: 10.18653/V1/2024.ACL-LONG.679.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *NeurIPS*, 2023.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adrià Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models. *NeurIPS*, 2023.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36, 2024.

Aaditya Prasad, Kevin Lin, Jimmy Wu, Linqi Zhou, and Jeannette Bohg. Consistency policy: Accelerated visuomotor policies via consistency distillation. *CoRR*, 2024. doi: 10.48550/ARXIV.2405.07503.

Team Qwen. Qvq: To see the world with wisdom, 2024. URL `https://qwenlm.github.io/blog/qvq-72b-preview/`.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, 2024. doi: 10.48550/ARXIV.2403.05530.

Luka Ribar, Ivan Chelombiev, Luke Hudlass-Galley, Charlie Blake, Carlo Luschi, and Douglas Orr. Sparq attention: Bandwidth-efficient llm inference. *ICML*, 2024.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *CoRR*, 2024. doi: 10.48550/ARXIV.2410.17434.

Philippe Tillet, Hsiang-Tsung Kung, and David Cox. Triton: an intermediate language and compiler for tiled neural network computations. In *Proceedings of the 3rd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages*, pp. 10–19, 2019.

Dezhan Tu, Danylo Vashchilenko, Yuzhe Lu, and Panpan Xu. Vl-cache: Sparsity and modality-aware kv cache compression for vision-language model inference acceleration. *CoRR*, 2024. doi: 10.48550/ARXIV.2410.23317.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, 2024a. doi: 10.48550/ARXIV.2409.12191.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *CoRR*, 2023. doi: 10.48550/ARXIV.2309.09777.

Xidong Wang, Dingjie Song, Shunian Chen, Chen Zhang, and Benyou Wang. Longllava: Scaling multi-modal llms to 1000 images efficiently via hybrid architecture. *CoRR*, 2024b. doi: 10.48550/ARXIV.2409.02889.

Yuxin Wen, Qingqing Cao, Qichen Fu, Sachin Mehta, and Mahyar Najibi. Efficient vision-language models by summarizing visual tokens into compact registers. *CoRR*, 2024. doi: 10.48550/ARXIV.2410.14072.

Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. *ECCV*, pp. 453–470, 2024. doi: 10.1007/978-3-031-73414-4_26.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. *CoRR*, 2024. doi: 10.48550/ARXIV.2404.15574.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *ICLR*, 2024.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. *CVPR*, pp. 9777–9786, 2021. doi: 10.1109/CVPR46437.2021.00965.

Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. Longvila: Scaling long-context visual language models for long videos. *CoRR*, 2024. doi: 10.48550/ARXIV.2408.10188.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. *CoRR*, 2024. doi: 10.48550/ARXIV.2412.04467.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. *AAAI*, pp. 9127–9134, 2019. doi: 10.1609/AAAI.V33I01.33019127.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *CoRR*, 2024a. doi: 10.48550/ARXIV.2406.16852.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *CoRR*, 2024b. doi: 10.48550/ARXIV.2410.02713.

Ningxin Zheng, Huiqiang Jiang, Quanlu Zhang, Zhenhua Han, Lingxiao Ma, Yuqing Yang, Fan Yang, Chengruidong Zhang, Lili Qiu, Mao Yang, et al. Pit: Optimization of dynamic sparse deep learning models via permutation invariant transformation. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 331–347, 2023.

A Modality-Aware Sparse Attention Search Algorithm

In Algorithm 4, we show how we search for the optimal sparse attention pattern for each attention heads. To achieve the best accuracy with limited FLOPs budget, we determine which sparse pattern will be used for each attention head, and the optimal setting for the pattern in real calculation (e.g., the stride size of the grid attention, the number of vertical/slash lines in VS pattern). In Algorithm 4, we first create the search space based on a target FLOPs for each pattern, ensuring all potential candidates (i.e., different patterns with different settings) have similar computational cost. Kernel-aware here indicates the computational cost reflects the real FLOPs in GPU kernels, instead of conceptual estimations, which is crucial to achieve the optimal acceleration. Next, we go through the search space with a reference example to decide the optimal pattern and setting. Specifically, we use recall of the attention output as the objective criterion when searching for the best pattern. This approach leverages FlashAttention (Dao, 2024) to reduce GPU memory overhead and incorporates the information from the V matrix, enabling end-to-end selection of the best pattern, which further enhances performance.

Algorithm 4 Modality-aware Sparse Attention Pattern Search

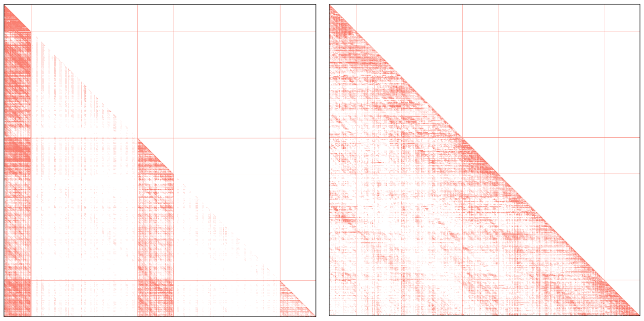
```
Input: $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{S \times d_h}$, inter-modality search space $\rho_{\text{inter}}$, intra-modality search space $\rho_{\text{intra}}$, modality type set $m \in \phi_m$, optimized sparse pattern $\mathrm{P}$

# Intra-modality sparse attention pattern search
for $i \leftarrow 1$ to $|\phi_m|$ do
    $p_{mi} \leftarrow \mathrm{KernelAwareSearch}\,(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}, m_i)$
    $\mathrm{P} \leftarrow \mathrm{P} \cup p_{mi}$
end for

# Cross-modality sparse attention pattern search
for $i \leftarrow 1$ to $|\phi_m|$ do
    for $j \leftarrow 1$ to $|\phi_m|$ do
        $p_{mi,mj} \leftarrow \mathrm{KernelAwareSearch}\,(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}, m_i, mj)$
        $\mathrm{P} \leftarrow \mathrm{P} \cup p_{mi,mj}$
    end for
end for

# Inter-modality sparse attention pattern search
for $i \leftarrow 1$ to $|\rho_{\text{inter}}|$ do
    $p_i \leftarrow \mathrm{argmin}\,(|\mathrm{sparse}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}, i) - \mathrm{attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})|$
    $\mathrm{P} \leftarrow \mathrm{P} \cup p_i$
end for
return $\mathrm{P}$
```

B Pattern Analysis



(a) K-Boundary pattern. (b) No-Boundary pattern.

Figure 8: Additional inter-modality sparse pattern.

In §2, we explain how the grid pattern emerges thanks to the unique geometric of the vision inputs. In Figure 8, we further visualize the K-Boundary pattern and No-Boundary pattern, which are additional in the mix-modality search space. Note that both K-Boundary and No-Boundary patterns takes no extra steps than pure intra-modality attention, as the sparse index in these two boundary patterns can be extracted across all rows.

C Experiment Details

C.1 Vision Language Models

We use two SOTA VLMs in our experiments: LongVILA (Xue et al., 2024) and Llava-Video (Zhang et al., 2024b). Specifically, Llava-Video use varies frames for video understanding tasks as including 32, 64, 110 frames. And as reported in their paper, Llava-Video always performs better with more frames. Therefore, we use the 110-frame variant of Llava-Video for the benchmarking. As for LongVILA, the authors released two variants, where the LongVILA-256Frame is a VLMs with 128K context length, and LongVILA-1M which is tailored for information retrieval tasks such as V-NIAH tasks. Therefore, the LongVILA-256Frame and LongVILA-1M are used respectively for the video understanding benchmarks and the V-NIAH test.

C.2 Baselines

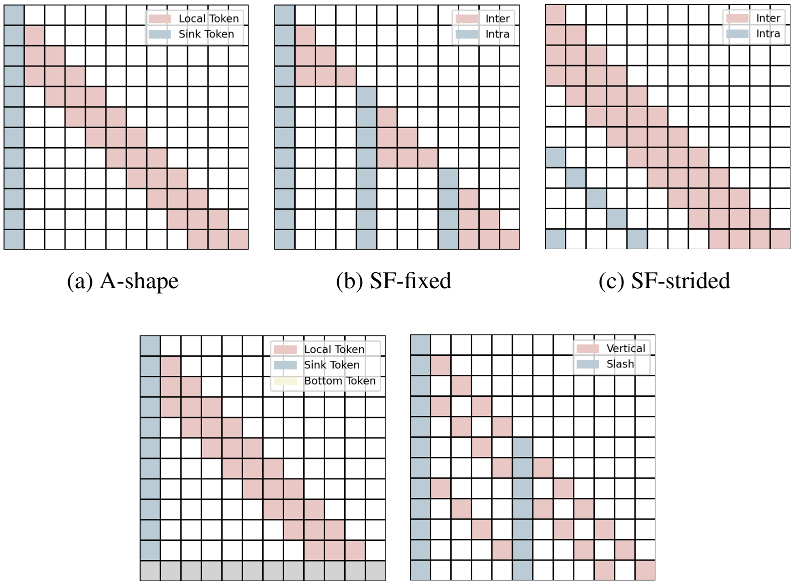
We include five sparse attention baselines in our experiments, including A-shape (Xiao et al., 2024), SF-fixed (Child et al., 2019), SF-strided (Child et al., 2019), Tri-shape (Li et al., 2024a), MInference (Jiang et al., 2024), and VisionZip (Yang et al., 2024). In Figure 9, we further visualize the patterns in these sparse attention baselines.

Note that VisionZip (Yang et al., 2024) is a visual token compression method, which directly uses the attention score in the vision tower to compress the visual tokens before feeding into the follow-up LLM. Although it is not tailored for pre-filling acceleration, it is included in our experiments as it does provide reduced FLOPs in the pre-filling stage and also provides a good comparison against token compression line of research.

C.3 A-shape and Vertical-Slash

A-shape and Vertical-Slash are used for intra-modality attention, together with the novel Grid pattern.

During the inference stage, we will perform an online estimation on the attention matrix to dynamically determine the spatial distribution our sparse indices, based on the assigned patterns and the exact input. After that, we conduct the sparse attention computations with our optimized GPU kernels. Noted that the sparse mask for vertical-slash and grid attention are dynamic, but the sparse



(a) A-shape (b) SF-fixed (c) SF-strided

(d) Tri-shape (e) Vertical-Slash (MInference)

Figure 9: The baselines of sparse attention in our experiments.

Table 2: Hyperparameters detail of baselines.

Method	Hyperparameters
A-shape	$\text{Sink} = 128, \text{Local} = 4096$
SF-fixed	$\text{Local} = \text{token\_per\_frame}, \text{vline\_stride} = \text{token\_per\_frame}$
SF-strided	$\text{Local} = \text{token\_per\_frame}, \text{vline\_stride} = \text{token\_per\_frame}$
Tri-shape	$\text{Sink} = 128, \text{Local} = 4096, \text{Bottom} = 128$
MInference	$\text{Vertical\_size} \in \{1000, 2000, 4000\}, \text{Slash\_size} \in \{1024, 2048, 4096, 6144\}$
VisionZip	$\text{dominant} = 54, \text{contextual} = 10$

mask for A-shape is static, so there is no overhead in building the dynamic masks, and only sparse calculation is required.

A-shape head. A-shape is a static sparse pattern, we simply include the first seven initial tokens plus a local window.

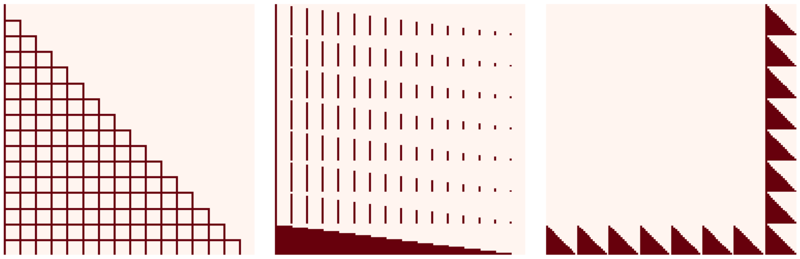
Vertical-Slash head. Due to the continuity of vertical and slash lines, we matmul the last query vector $\boldsymbol{Q}_{[-\text{last\_q}:]}$ and key vector $\boldsymbol{K}$ to produce the estimated attention matrix $\hat{\boldsymbol{A}}$, which, in turn, is used to determine the indices for the vertical $\boldsymbol{i}_v$ and slash $\boldsymbol{i}_s$ lines. After obtaining the sparse indices for the vertical and slash lines, we convert them into a sparse format $\boldsymbol{i}_{vs}$. Using these sparse indices, we perform block-sparse calculations of the attention weights and attention output.

C.4 Permutation for the Grid Pattern and Across Modality

We illustrate how the permutation is applied to the Grid pattern and the Q-boundary and 2D-boundary patterns in Figure 10 and Figure 10.

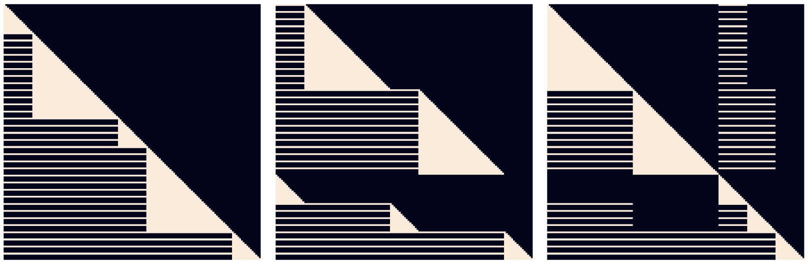
C.5 Search Space

Following (Jiang et al., 2024), we set the target FLOPs t to be the same as 1k global tokens and 4k local window tokens in the *A-shape* pattern. Additionally, we use only one sample as our calibration set from the egoschema task with no more than 25K tokens, which exhibits strong generalization and stability across different lengths and domains. The search time is approximately 15 minutes



(a) Before Permutation (b) Row-wise Permutation (c) Column-wise Permutation

Figure 10: Permutation for the Grid Pattern. (a) Before permutation. (b) Row-wise permutation. (c) Column-wise permutation.



(a) Mix-modality (b) Q-wise Permutation (c) K-wise Permutation

Figure 11: Permutation for mix-modality context. (a) Mix-modality. (b) Q-wise permutation. (c) K-wise permutation.

on a single A100. This pattern search is individually conducted for each model: Llava-Video-7B, LongVila-256Frame, and LongVila-1M. The search space is shown in Table 3.

D Benchmark Details

We evaluate our method on several video understanding benchmarks that test different aspects of video comprehension:

EgoSchema EgoSchema (Mangalam et al., 2023) is a diagnostic benchmark for very long-form video language understanding, structured as a multiple-choice question answering task. The benchmark requires models to answer questions about egocentric videos by selecting from given options (labeled A through E). The evaluation can be performed either on the full set via submission to an evaluation server, or on a released subset of 500 questions for direct scoring.

Video-MME Video-MME (Fu et al., 2024) is a comprehensive multi-modal evaluation benchmark that tests MLLMs across diverse video types and temporal dimensions. It spans 6 primary visual domains with 30 subfields and includes videos ranging from 11 seconds to 1 hour in duration. The benchmark comprises 900 videos totaling 254 hours, with 2,700 manually annotated question-answer pairs. It evaluates models' ability to process not just video frames but also integrated multi-modal inputs like subtitles and audio.

NExT-QA NExT-QA (Xiao et al., 2021) focuses on advancing video understanding from basic description to explaining temporal actions. It features both multiple-choice and open-ended QA tasks that target three key aspects: causal action reasoning, temporal action reasoning, and common scene comprehension. The benchmark is specifically designed to evaluate models' ability to reason about actions beyond superficial scene descriptions.

Attention Type	**Parameters**
Grid Attention	(frame_stride, True, False, False, 1024)
	(frame_stride, False, True, False, 1024)
	(frame_stride, False, False, True, 1024)
	(frame_stride, True, True, False, 1024)
	(frame_stride, False, True, True, 1024)
	(frame_stride, True, True, True, 1024)
	(stride, True, False, False, 1024)
	(stride, False, True, False, 1024)
	(stride, False, False, True, 1024)
	(stride, True, True, False, 1024)
	(stride, False, True, True, 1024)
	(stride, True, True, True, 1024)
A-shape	(128, 1024)
	(128, 2048)
	(128, 4096)
Vertical-Slash	(1000, 1024)
	(1000, 2048)
	(2000, 2048)
	(1000, 3096)
	(2000, 3096)
	(1000, 4096)
	(2000, 4096)
	(3500, 200)
	(1000, 2500)

Table 3: The search space for each attention pattern: 1) Grid Attention: (stride, use hline, use vline, use slash, max stride); 2) A-shape: (sink, local); 3) Vertical-Slash: (vertical size, slash size)

Perception Test The Perception Test (Patraucean et al., 2023) perce evaluates perception and reasoning skills across video, audio, and text modalities. It contains 11.6k real-world videos with an average length of 23 seconds, featuring perceptually interesting situations. The benchmark tests four key skills (Memory, Abstraction, Physics, Semantics) and various types of reasoning (descriptive, explanatory, predictive, counterfactual). Videos are densely annotated with six types of labels: multiple-choice QA, grounded video QA, object tracks, point tracks, temporal action segments, and sound segments.

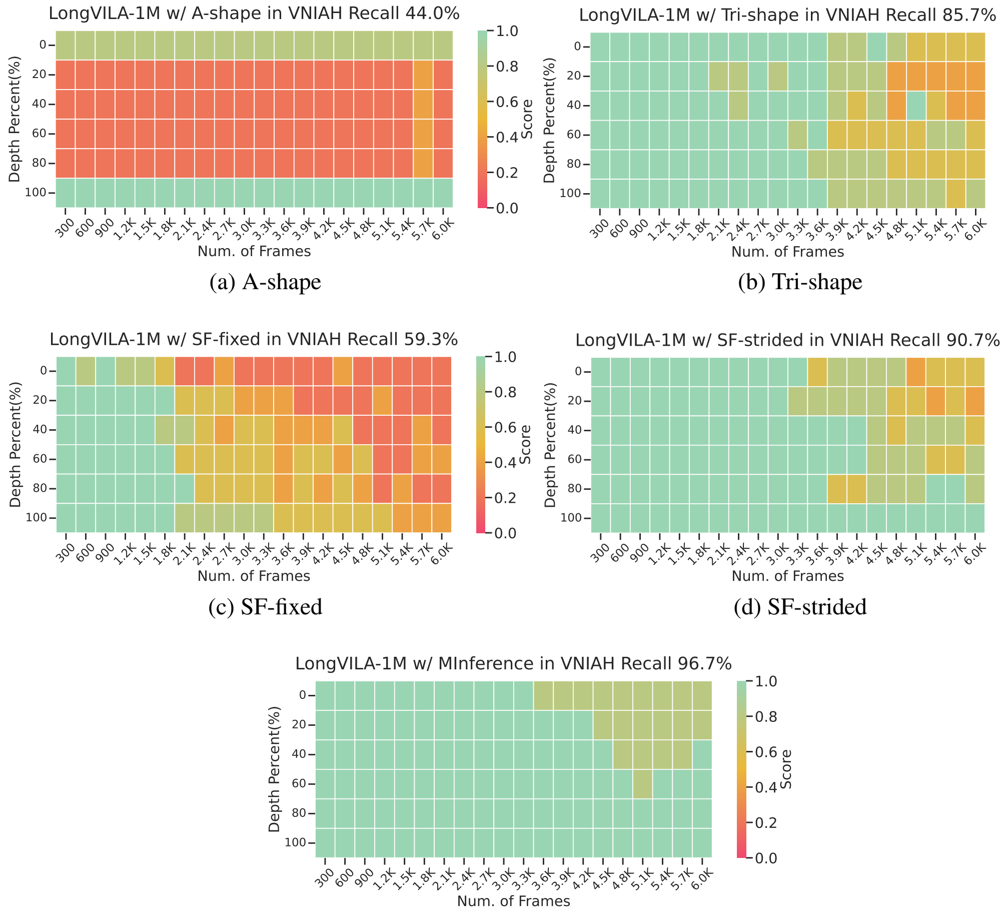
ActivityNet-QA ActivityNet-QA (Yu et al., 2019) is a large-scale VideoQA dataset consisting of 58,000 QA pairs on 5,800 complex web videos derived from the ActivityNet dataset. The benchmark is fully annotated and designed to test models' understanding of complex web videos through question answering. Unlike automatically generated datasets, ActivityNet-QA features human-annotated questions and answers, making it particularly valuable for evaluating real-world video understanding capabilities.

Video Detail Description (VideoDC) VideoDC (Lab, 2024) focuses on comprehensive video understanding through detailed descriptions. The benchmark consists of question-answer pairs generated with GPT-3.5, where questions prompt for detailed descriptions focusing on main subjects, their actions, and background scenes. The evaluation assesses the quality and completeness of video descriptions generated by models.

E Additional Experiments Results

E.1 Additional Video Needle In A Haystack Results

we further present the results of the Video Needle In A Haystack task with our baselines. The results of our method and full atttenton is shown in Figure 4.



(a) A-shape

(b) Tri-shape

(c) SF-fixed

(d) SF-strided

(e) MInference

Figure 12: Video Needle In A Haystack (Zhang et al., 2024a) results using LongVila-Qwen2-7B-1M (Xue et al., 2024).

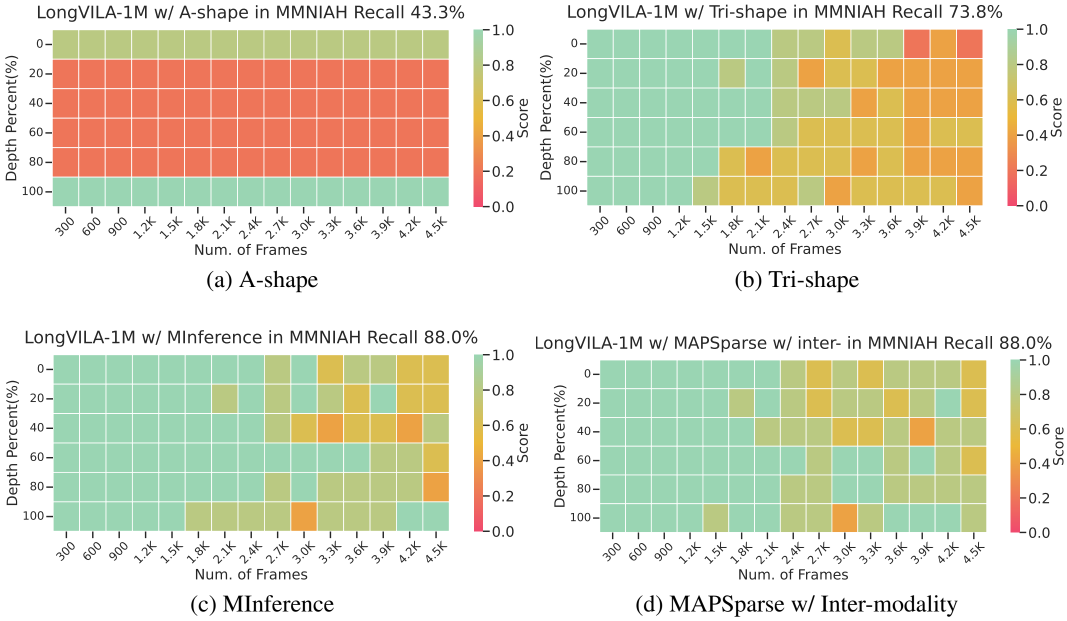
E.2 Additional Mix Modality Needle In A Haystack Results

We further present the results of the Mix Modality Needle In A Haystack task with our baselines and the inter-modality variant of our method. The results of full atttenton and MAPSparse is shown in Figure 4.

E.3 Latency Breakdown

E.4 VS Pattern vs. Grid Pattern

Both VS pattern and Grid pattern achieve strong performance on video understanding and V-NIAH tasks. However, due to the grid attention pattern observed in VLMs, the overlap between blocks covered by diagonal lines in the VS pattern is minimal, reducing sparsity within the kernel. This explains why VS pattern exhibits significantly higher latency compared to Grid pattern. Additionally, leveraging permutation-based optimization effectively reduces the number of blocks involved in kernel computation, thereby lowering latency while maintaining comparable performance.



(a) A-shape

(b) Tri-shape

(c) MInference

(d) MAPSparse w/ Inter-modality

Figure 13: Mix Modality Needle In A Haystack results using LongVila-Qwen2-7B-1M (Xue et al., 2024).

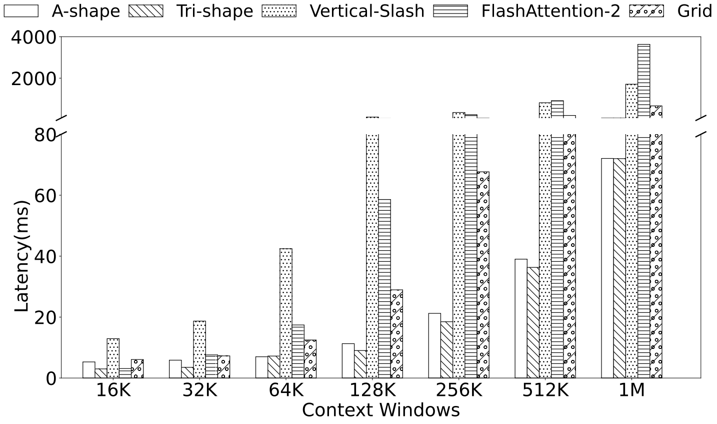


Figure 14: The latency breakdown of a single attention kernel for four sparse attention patterns and FlashAttention (Dao, 2024) across different context windows in a single A100, including the index time for dynamic sparse approximation and building dynamic sparsity. At 1M tokens, the latency for Grid is 358ms.