Fibottention: Inceptive Visual Representation Learning with Diverse Attention Across Heads

Anonymous authors
Paper under double-blind review

Abstract

Vision Transformers and their variants have achieved remarkable success in diverse visual perception tasks. Despite their effectiveness, they suffer from two significant limitations. First, the quadratic computational complexity of multi-head self-attention (MHSA), which restricts scalability to large token counts, and second, a high dependency on large-scale training data to attain competitive performance. In this paper, to address these challenges, we propose a novel sparse self-attention mechanism named *Fibottention*. Fibottention employs structured sparsity patterns derived from the Wythoff array, enabling an $\mathcal{O}(N \log N)$ computational complexity in self-attention. By design, its sparsity patterns vary across attention heads, which provably reduces redundant pairwise interactions while ensuring sufficient and diverse coverage. This leads to an *inception-like functional diversity* in the attention heads, and promotes more informative and disentangled representations. We integrate Fibottention into standard Transformer architectures and conduct extensive experiments across multiple domains, including image classification, video understanding, and robot learning. Results demonstrate that models equipped with Fibottention either significantly outperform or achieve on-par performance with their dense MHSA counterparts, while leveraging only 2% of all pairwise interactions across self-attention heads in typical settings, resulting in substantial computational savings. Moreover, when compared to existing sparse attention mechanisms, Fibottention consistently achieves superior results on a FLOP-equivalency basis. Finally, we provide an in-depth analysis of the enhanced feature diversity resulting from our attention design and discuss its implications for efficient representation learning.

1 Introduction

Transformer-based architectures, such as large foundation models, *e.g.*, GPT (Radford et al., 2018; Brown et al., 2020), BERT (Devlin et al., 2019), ALBERT (Lan et al., 2020), ViT (Dosovitskiy et al., 2020), DETR (Carion et al., 2020), D-DETR (Zhu et al., 2020), CLIP (Ke et al., 2018), have achieved dominating performance in many downstream tasks such as object detection and tracking (Dutta et al., 2024), document summarization (Radford et al., 2018), language modeling (Devlin et al., 2019), and video understanding (Reilly & Das, 2024). Compared to other deep neural networks, such as convolutional neural networks, Transformers excel when trained on large-scale datasets with extensive model parameters (Dosovitskiy et al., 2020). However, their performance typically degrades in low-data regimes (Dosovitskiy et al., 2020). The growing demand for models that can be deployed and trained effectively on diverse edge devices or within the Internet of Things (IoT) (Agarwalla et al., 2024; Reidy et al., 2023; Sun et al., 2024; Tuli & Jha, 2023; Qu et al., 2022) has driven significant research interest in developing efficient models that are both compute-efficient and data-efficient.

At the core of Transformer-based models is the multi-head self-attention (MHSA) (Kim et al., 2016; Vaswani et al., 2017; Dosovitskiy et al., 2020) mechanism. In MHSA, N input feature vectors (a.k.a. tokens) of dimension d are mapped to h query, key and value matrices $Q_i, K_i, V_i \in \mathbb{R}^{N \times d_h}$, for each head $i = 1, \ldots, h$, which are subsequently mapped to output feature vectors. Its computational bottleneck is the computation of the h attention matrices $A_i = Q_i K_i^\top / \sqrt{d_h} \in \mathbb{R}^{N \times N}$ consisting of inner products of queries and keys, which inherently limits the number of tokens due to the resulting $\mathcal{O}(N^2)$ complexity. To mitigate this limitation,

a considerable body of literature considers Transformer variants that evaluate the attention matrices only at a *sparse subset* of their entries $\Omega \subset [N] \times [N]$ of size $s = |\Omega| < N^2$, thereby lowering the incurred time complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(s)$. It has been observed that, while reducing the computational load, such *sparse attention* patterns can degrade the modeling capacity of the underlying architecture and often reduce their accuracy relative to dense attention on several benchmarks, highlighting a trade-off between efficiency and performance (Child et al., 2019; Yun et al., 2020b; Beltagy et al., 2020; Zaheer et al., 2020; Shang et al., 2022). Since MHSA is designed to facilitate interactions between all tokens, finding suitable choices for such a global support set Ω with a favorable trade-off between efficacy and efficiency is challenging due to the data, model, and instance dependence of the attention values.

Popular sparse attention strategies include local attention with fixed-size sliding windows (Wang et al., 2019; Beltagy et al., 2020; Ramachandran et al., 2019), in which only interactions between spatially proximal tokens are considered, often augmented by random attention mechanisms (Zaheer et al., 2020; Zhang & Gong, 2023). Interestingly, the majority of sparse MHSA mechanisms are designed in the context of natural language processing, with the notable exception in visual domains for high-resolution images (Esser et al., 2021; Zhang et al., 2021). Unlike natural language, images and videos exhibit strong spatial and spatiotemporal redundancy: neighboring pixels often encode similar content, and many key–query interactions in dense MHSA are unnecessary for visual representation learning. This redundancy means uniform, all-to-all attention spends substantial compute on low-utility interactions. In either domain, sparse attention has been mostly considered as a modification of MHSA that *harms* model performance while improving computational feasibility. In this work, however, we distill insights from a substantial body of literature on sparse attention mechanisms to design a versatile, deterministic sparse attention pattern that is able to improve the model performance of Transformers in a regime of limited visual data, while simultaneously improving their computational efficiency across dataset sizes.

In particular, we propose *Fibottention*, an MHSA variant that can be used as a drop-in replacement for full attention in vision Transformers (Dosovitskiy et al., 2020; Wu et al., 2021; Bertasius et al., 2021; Li et al., 2022). For each attention head A_i, Fibottention deploys a complementary sparsity pattern Ω_i based on *non-overlapping, dilated sliding windows* (Beltagy et al., 2020) with a *growing dilation* schedule to capture both local and global token interactions within each head. Concretely, we implement this using *non-overlapping generalized Fibonacci sequences* as dilation sequences $(f_n)_n$ across heads (see Table 1, Fig. 1, and Eq. (3)). Fibonacci-like schedules progressively thin out connections as distance grows, encoding dense local interactions and increasingly sparse long-range links. This structured sparsity is motivated not only by compute efficiency, but also by an inductive bias that promotes multi-scale feature aggregation and head-wise complementarity. The design choices of Fibottention are based on the following key insights: (*i*) *the principal diagonal* of an attention matrix A_i might not contain helpful information for the model (Shi et al., 2021); that (*ii*) a structured, deterministic sparsity pattern Ω, which is able to capture both local and global token interactions is desirable from a modeling (Zaheer et al., 2020; Shang et al., 2022) and efficiency perspective; (*iii*) sparsity patterns, $\Omega_1, \ldots, \Omega_h$ that *differ across* the attention heads, A_i can achieve a diversity of feature representations across heads (analogous to the intuition behind CNN-based *Inception* models (Szegedy et al., 2016)); and finally, that (*iv*) a *low overlap* between the Ω_i is desirable as it has the potential to *maximize* the diversity of the resulting feature representations while *minimizing* the total number $\sum_{i=1}^{h} |\Omega_i|$ of inner products to be calculated. Consequently, the Fibottention's structured attention increases the feature diversity across heads and exhibits a performance-enhancing inductive bias that is particularly beneficial for visual domains with *limited training data* (e.g., video understanding and robotics).

We extensively evaluate Fibottention in conjunction with diverse state-of-the-art Transformer architectures catered towards visual representation tasks, including image classification, video action recognition, and robot imitation learning (§4). On CIFAR-10/100 (Krizhevsky, 2009), it consistently surpasses Transformer baselines trained with full multi-head self-attention (MHSA) while remaining on par when trained on ImageNet-1K (Deng et al., 2009) (Table 2). As we show in Section 4.2, Fibottention's behavior is consistent across various Transformer architectures (Table 3): when incorporated into UPop (Shi et al., 2023), iFormer (Zheng, 2025), Swin-B (Liu et al., 2021), and ConViT-B (d'Ascoli et al., 2021), it consistently reduces attention compute to a small fraction of the baseline while preserving or improving accuracy depending on capacity and training regime, a feat that other sparse attention designs such as top-K attention (Gupta et al., 2021) or

Longformer (Beltagy et al., 2020) fail to achieve. In temporal domains, we see that Fibottention improves the Top-1 accuracy of TimeSformer (Bertasius et al., 2021) on Smarthome (Das et al., 2019) and NUCLA (Wang et al., 2014) (Table 5). For robot imitation learning tasks (Lift/Can/PushT), it achieves the best average task completion rates (Table 6). Overall, we observe that Fibottention can achieve large, quantifiable reductions in attention FLOPs while maintaining or enhancing predictive performance across scales, modalities, and backbones.

The remainder of this paper is organized as follows. Section 2 reviews related work on sparse, adaptive, and diverse attention for vision Transformers. Section 3 details the design of Fibottention based on Fibonacci dilation sequences defined from the Wythoff array and its head-wise masking strategy. Section 4 presents empirical evaluations of vision Transformers modified by Fibottention and other sparse attention variants on image classification tasks. Section 5 extends Fibottention to other visual domains, including video action recognition and robot imitation learning, describing training protocols and empirical findings, and Section 6 provides ablations, studies of the impact of Fibottention on head diversity and inductive bias. Section 7 concludes with limitations and avenues for future work. Finally, we provide a proof of Fibottention's $\mathcal{O}(N \log N)$ time complexity in Appendix A, implementation and hyperparameter details in Appendices B and C and further ablation studies in Appendices D and E.

2 Related Work

Vision Transformers. Derived from Transformers (Vaswani et al., 2017), which excel on long-range sequence tasks in NLP, the vision Transformer (ViT) (Dosovitskiy et al., 2020) splits images into small patches, each corresponding to a token, and has emerged as a popular architecture in visual understanding tasks. While ViTs (Dosovitskiy et al., 2020; Touvron et al., 2021) outperform CNN-based models in a variety of visual representation tasks, they require extensive training data to achieve this superior performance and exhibit a quadratic time complexity in the number of tokens N. DeiT models (Touvron et al., 2021; Wu et al., 2022) exhibit advantages over ViT in the presence of limited training data due to their knowledge distillation, but still require a quadratic time complexity with respect to N. Another line of research, which includes CvT (Wu et al., 2021) and ConViT (d'Ascoli et al., 2021) models, has achieved strong performance based on hybridization with convolutions. Mobile-oriented hybrids such as iFormer (Zheng, 2025) push this direction toward latency-constrained regimes by coupling ConvNeXt-style local processing with lightweight modulation attention, achieving favorable accuracy–latency trade-offs on smartphone-class hardware. A useful inductive bias is also provided by hierarchical models such as MViTv2 (Li et al., 2022) or Swin Transformer (Liu et al., 2021), which combine a patch merging strategy with MHSA mechanisms adapted to blocks of tokens. While all these models are compatible with the MHSA modification provided by Fibottention, we focus in our experiments on models without knowledge distillation for a more controlled experimental setup.

Sparse Attention. Despite their advancements, ViT variants share a common limitation: the MHSA mechanism inherently requires the evaluation of $\mathcal{O}(N^2)$ token interactions, posing a significant computational challenge. However, theoretical insights (Yun et al., 2020b) demonstrate that sparse attention, given an appropriate sparsity pattern, can effectively approximate any sequence-to-sequence function, mirroring the capabilities of full attention (Yun et al., 2020a). Several studies, primarily in NLP, have explored optimal sparsity patterns, emphasizing the importance of central diagonal elements in Ω_i (Clark et al., 2019; Kovaleva et al., 2019). Longformer (Beltagy et al., 2020), BigBird (Zaheer et al., 2020), and Star-Transformer (Guo et al., 2019) incorporate global and local token interactions to enhance performance (Kovaleva et al., 2019; Li et al., 2019). Regional token interactions are commonly implemented as a diagonal sliding window within Ω, which can be expanded through dilated sliding window attention to increase the receptive field without additional computational overhead (Beltagy et al., 2020; Hassani & Shi, 2022). Longformer further extends this by incorporating random token pair interactions, while sparse Transformers (Child et al., 2019) introduce specialized sparse patterns designed for efficient long-sequence generation. Collectively, these methods leverage combinations of local, global, sliding window, dilated sliding window, and random attention patterns. Later, Shi et al. (2021) observed that the main diagonal elements in Ω_i are redundant, proposing a learnable differentiable attention mask to refine the sparsity structure. Overall, compared to NLP, structured

sparse attention is less explored in vision problems, in which token interactions tend to be more redundant and contextually distinct from the language domain, and have been mostly investigated in the context of high-resolution setups (Esser et al., 2021; Zhang et al., 2021; Li et al., 2025), e.g., via the hierarchical neighborhood attention Transformer (Hassani et al., 2023); while in these works, computational efficiency is the main focus of the sparse attention design, we jointly optimize for inductive bias in Fibottention. In Section 4.2, we use a selection of the above-mentioned sparse attention modifications (adapted to the vision setting) as baselines in empirical comparisons to the adaptation of Fibottention into Vision Transformers. Beyond sparse attention, other full attention approximations have been studied, with similar potential for breaking its quadratic complexity bottleneck, such as efficient attention (Shen et al., 2021) and Linformer (Wang et al., 2020), which apply non-linear transformations of key and query matrices instead of softmax and low-rank approximations of keys and values, respectively.

Adaptive and Diverse Attention. While many sparse attention designs fix attention patterns across batches, several works explore adaptive or instance-dependent sparsity. For example, Kitaev et al. (2020); Roy et al. (2021); Wei et al. (2023) have proposed learned instance-dependent attention masks, which can be effective but impose additional model complexity dedicated to the learning of the sparsity mask. A related line of work studies variants of *top-k attention* (Zhao et al., 2019; Gupta et al., 2021; Sander et al., 2023; You et al., 2025), where each query attends dynamically only to the k most relevant keys, resulting in instance-dependent attention computations. Recently, top-k sparse attention in state-of-the-art large-scale language models (Liu et al., 2025) coupled with a lightweight indexer module has enabled efficient Transformer architectures applicable for long contexts. Fibottention, unlike learned sparse or top-k sparse attention mechanisms, is non-adaptive and does not require a conceptual and computational overhead due to its fixed sparsity patterns. A limited number of works report observations of improved empirical performance of Transformers using attention patterns that *vary across heads*; examples of such works are Longformer (Beltagy et al., 2020), which reports improved performance when combining sparse heads with and without dilation in their architectures, and Child et al. (2019), which provides evidence that differently sized sub-blocks across heads are preferable. However, we are not aware of works on fixed sparse attention patterns that systematically incorporate these observations into their mechanism design, especially in vision settings (Zhang et al., 2021), as Fibottention does with its complementary diverse sparse design. With adaptive designs, however, the benefits of diverse attention have recently been studied in language models for inference (Fu et al., 2025) or through the introduction of gating (Qiu et al., 2025).

Dynamic Token Sparsification. Several methods have explored dynamic token reduction strategies to further mitigate the quadratic complexity of ViTs. DynamicViT (Rao et al., 2021) introduces lightweight prediction modules that estimate the importance of each token at intermediate layers, progressively discarding less informative ones while maintaining the most relevant content for recognition. PS-ViT (Yue et al., 2021) adopts a progressive sampling strategy: at each iteration, the model predicts new sampling offsets to refine where tokens should be drawn, enabling it to concentrate on discriminative image regions. EViT (Liang et al., 2022) reorganizes tokens by identifying attentive and inattentive ones through class-token attention, retaining the former while merging the latter into compact representations for subsequent layers. Beyond vision-only settings, UPop (Shi et al., 2023) extends progressive pruning to vision-language Transformers. While these methods dynamically adapt the token set during inference, Fibottention sparsifies only token *interactions*, not tokens. Thus, Fibottention's design is orthogonal to dynamic token reduction approaches and can be combined with token sparsification; in combination with those ideas, Fibottention has the potential to further improve runtime efficiency and inductive biases of token-sparsified Transformer models.

3 Method

In this section, we discuss the multi-head self-attention modification dubbed Fibottention, which uses structured, diverse support sets derived from Fibonacci–Wythoff sequences in its sparse attention evaluation in each head. We start with a general sparse attention framework that allows formulating and comparing different sparsity patterns for attention heads $\{A_i\}_{i=1}^{h}$.

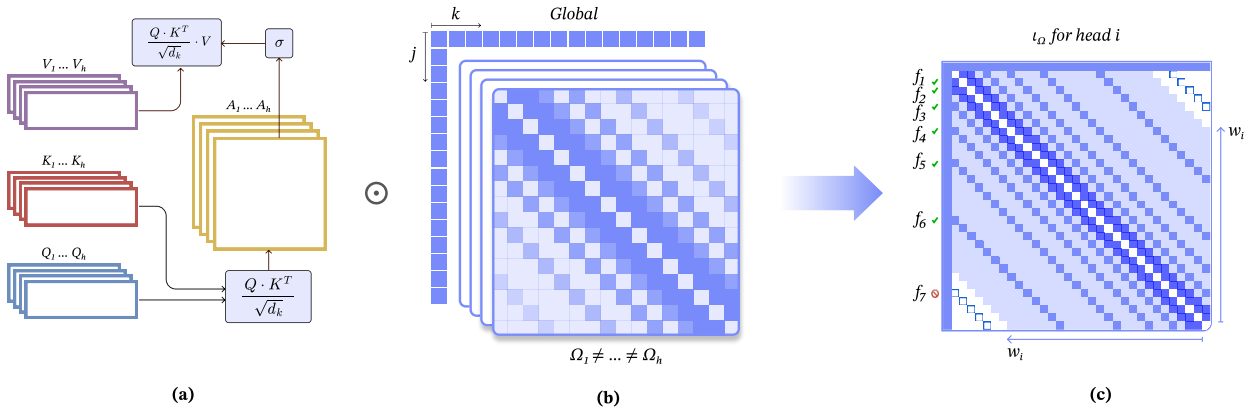


Figure 1: (a) The MHSA. (b) A general sparse attention computation strategy. A sequence of sparse support sets, $\{\Omega_i\}_{i=1}^h$, where each set selects $|\Omega_i| < N^2$ entries of the attention matrix. (c) The generalized masking strategy of Fibottention that controls sparsity of each attention matrix A_i through a dilated sequence, $(f_n)_n \subset \mathbb{N}$, and a fixed window size, w for each head. Elements on f_1 and f_2 occur exclusively in the Modified Wythoff variant.

3.1 Sparse Attention with Windowed Dilation

Instead of observing the attention matrix A_i at each of its N^2 entries, sparse attention mechanisms compute only the dot products whose indices are supported on a subset $\Omega \subset \{1, 2, \dots, N\}^2$, i.e., we can define the sparse attention matrix $A_i^\Omega \in \mathbb{R}^{N \times N}$ of the i-th head corresponding to mask Ω as

$$(A_i^\Omega)_{j,k} = \begin{cases} \frac{Q_i^{(j)\top} K_i^{(k)}}{\sqrt{d_h}}, & \text{if } (j,k) \in \Omega, \\ -\infty, & \text{if } (j,k) \notin \Omega; \end{cases} \tag{1}$$

for any $j, k \in [N]$, where $Q_i^{(j)} \in \mathbb{R}^{d_h}$ and $K_i^{(k)} \in \mathbb{R}^{d_h}$ are the j-th query vector and the k-th key vector of the i-th attention head, respectively. If $\odot$ denotes the entrywise matrix multiplication, also called Hadamard product, this can be written as $A_i^\Omega = \text{sign}(A_i) \odot (|A_i| \odot \iota_\Omega)$, where $\iota_\Omega \in \mathbb{R}^{N \times N}$ is an indicator matrix of the index set Ω that is 1 for indices $(j,k) \in \Omega$ and $-\infty$ otherwise. In this work, we study structured support sets that capture both local and global interactions while ensuring efficient inference and training through sparsity. To this end, we introduce the notion of a *dilation sequence*, $(f_n)_n \subset \mathbb{N}$, which determines the sequence of distances between indices of tokens that attend to each other and render diversity across heads. Furthermore, for a given attention head, we fix a *window size*, $1 \leq w \leq N$, which, independently of the dilation sequence, provides an upper bound for the index distance between interacting token indices in the attention matrix.

Given the sequence, $(f_n)_n$ and parameter, w, we define the *support set*, $\Omega_w^{(f_n)} \subset \{1, 2, \dots, N\}^2$ of *interacting query-key pairs dilated by* $(f_n)_n$ *of window size* w such that

$$\Omega_w^{(f_n)} = \big\{(j,k) : |j-k| \in \{f_n\}_n, |j-k| \leq w\big\}.$$

We refer to Figure 1(c) for a visualization of such support sets; $\Omega = \Omega_w^{\{f_n\}}$ represents the effective set of indices of query-key pairs for which we need to calculate the dot product in a given attention head A_i.

Several dilation sequences have been studied in both vision and language Transformer architectures. The majority of existing literature (Child et al., 2019; Beltagy et al., 2020; Zhang et al., 2021; Hassani & Shi, 2022) considers dilation sequences that are multiples of a fixed factor $c \in \mathbb{N}$, i.e., $(f_n)_n = (cn)_{n \in \mathbb{N}}$, corresponding to sliding windows with constant dilation factor c. While providing a certain level of efficiency, their attention complexity only reduces from $\mathcal{O}(Nw)$ to $\mathcal{O}(Nw/c)$, which is still of order N^2 if the window size w is chosen to be $w = \mathcal{O}(N)$. On the other hand, choosing a small window size $w = \mathcal{O}(1)$ prevents the inclusion of any global interactions. Dilation patterns based on different dilation sequences have been less explored; Li et al.

Table 1: Generalized Fibonacci sequences $\mathrm{Fib}(a_i^{\text{Wyt}}, b_i^{\text{Wyt}})$ and $\mathrm{Fib}(a_i^{\text{Wyt-m}}, b_i^{\text{Wyt-m}})$ drawn from the Wythoff array used by head i in Fibottention (default and modified variants).

i	$\boldsymbol{a_i^{\text{Wyt-m}}}$	$\boldsymbol{b_i^{\text{Wyt-m}}}$	$\boldsymbol{a_i^{\text{Wyt}}}$	$\boldsymbol{b_i^{\text{Wyt}}}$				
1	0	1	1	2	3	5	8	...
2	1	3	4	7	11	18	29	...
3	2	4	6	10	16	26	42	...
4	3	6	9	15	24	39	63	...
5	4	8	12	20	32	52	84	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...

(2019) studied exponentially dilated sequences giving rise to attention complexities of order $\mathcal{O}(N \log w) = \mathcal{O}(N \log N)$.

3.2 Diverse Sparse Attention through Fibonacci-Wythoff Dilation Sequences

While support sets derived from exponential dilation sequences lead to sparse attention matrices, it might happen that crucial query-key interactions are not captured by overly sparse patterns, deteriorating the quality of the resulting MHSA representations. At the same time, limited experimental results in (Beltagy et al., 2020; Kovaleva et al., 2019; Ding et al., 2023) indicate that varying support set patterns across attention heads can improve model performance. Furthermore, state-of-the-art sparse attention mechanisms aim for a delicate balance between covering local and global interactions (Beltagy et al., 2020; Zaheer et al., 2020), and do not necessarily include the interactions on the main diagonal of the attention matrix (Shi et al., 2021).

Motivated by these observations, we postulate that *sparse attention matrices with diverse, well-designed support patterns across attention heads* are desirable and have the potential to lead to improved learned representations.

Fibonacci Dilation Sequences. To achieve our design goals, we propose a sparse attention pattern that builds on *(generalized) Fibonacci sequences* (Sigler, 2003; Koshy, 2019). The well-known Fibonacci sequence $(f_n)_{n\in\mathbb{N}}$ is defined as the sequence of integers $(0, 1, 1, 2, 3, 5, 8, 13, \ldots)$ (OEIS Foundation Inc.) satisfying the linear recurrence relation

$$f_{n+1} = f_n + f_{n-1}, \tag{2}$$

for each $n \geq 2$, where $f_1 = 0$ and $f_2 = 1$. Binet's formula (Koshy, 2019) states that the n-th Fibonacci number satisfies $f_n = (\phi^{n-1} - \psi^{n-1})/\sqrt{5}$, where $\phi = (1 + \sqrt{5})/2 \approx 1.618$ is the golden ratio and $\psi = (1 - \sqrt{5})/2$. From this formula, it can be inferred that after initial slow growth, the sequence grows exponentially with respect to the base ϕ. Similar integer sequences can be defined from the recurrence (2) by fixing the initial elements, $f_1 = a \in \mathbb{N}$ and $f_2 = b \in \mathbb{N}$.

Given a window size w, parameters, $a, b \in \mathbb{N}$, and denoting the corresponding *generalized Fibonacci sequence*, $(f_n)_n$ by $\mathrm{Fib}(a, b)$, we can define a corresponding support set for an $N \times N$ attention matrix as $\Omega_w^{\mathrm{Fib}(a,b)} = \big\{(j, k) : |j - k| \in \mathrm{Fib}(a, b), |j - k| \leq w\big\}$. An experimental ablation study (see Section 6.6) indicates that a simple Fibonacci attention pattern can already be advantageous compared to other dilation sequences, even when deployed uniformly across heads.

Wythoff Array and its Properties. Among integer sequences based on order-2 linear recurrence relations, generalized Fibonacci sequences are attractive for creating attention support sets since by varying a and b, a variety of integer values can be covered while retaining the same long-term growth rate (see Appendix A.1, Lemma 1) as the Fibonacci numbers. Accordingly, we use h different Fibonacci-type sequences, $\mathrm{Fib}(a_i, b_i)$ with different initial values $a_1, \ldots, a_h \in \mathbb{N}$ and $b_1, \ldots, b_h \in \mathbb{N}$, giving rise to *head-specific* attention support sets. Defining also head-specific window sizes, $w_1, \ldots, w_h \leq N$, we obtain the support set Ω_i for the i-th attention head matrix A_i defined as $\Omega_{w_i}^{\mathrm{Fib}(a_i,b_i)}$ for each head index $i = 1, \ldots, h$.

Within this framework, we aim to choose the sequence parameters $(a_i, b_i)_i$ such that the following three *desiderata* are satisfied: (*i*) the overlap between different attention head support sets should be minimized, allowing for a semantic specialization of the corresponding head weights during training, (*ii*) the total size $\sum_{i=1}^{h} |\Omega_{w_i}^{\mathrm{Fib}(a_i,b_i)}|$ of the support sets should be small to retain efficiency, but within that constraint, (*iii*) as many *relevant* query-key interactions as possible should be captured by at least one attention head, i.e., the set union $\cup_{i=1}^{h} \Omega_{w_i}^{\mathrm{Fib}(a_i,b_i)}$ should be maximal.

A suitable, essentially hyperparameter-free choice can be derived from the Wythoff array (Morrison, 1980; Conway & Ryba, 2016; Chen et al., 2025), which had been originally introduced in the context of a combinatorial game (Wythoff, 1907). The Wythoff array can be considered as a collection of generalized Fibonacci sequences $\{\mathrm{Fib}(a_i, b_i)\}_{i\in\mathbb{N}}$ with specific choices a_i^{Wyt} and b_i^{Wyt} for each $i \in \mathbb{N}$ that have provably *no overlap*, but contain each integer exactly once (Morrison, 1980; Conway & Ryba, 2016). In particular, the i-th row sequence of the Wythoff array is given by the sequence, $\mathrm{Fib}(a_i^{\mathrm{Wyt}}, b_i^{\mathrm{Wyt}})$ with initial elements, $a_i^{\mathrm{Wyt}} = \lfloor \lfloor i\phi \rfloor \phi \rfloor$ and $b_i^{\mathrm{Wyt}} = \lfloor \lfloor i\phi \rfloor \phi^2 \rfloor$; see Table 1.

Fibottention. Based on the above considerations, we define a novel non-adaptive sparse attention mechanism, called *Fibottention*, that is designed as a drop-in replacement of full self-attention in multi-head self-attention blocks. In any given MHSA layer with h heads, for a given head index $i = 1 \ldots, h$, we restrict the computation of unnormalized attention weights in $A_i \in \mathbb{R}^{N\times N}$ to the support set $\Omega_i := \Omega_{w_i}^{\mathrm{Fib}(a_i^{\mathrm{Wyt}}, b_i^{\mathrm{Wyt}})}$ given as

$$\big\{(j,k) : |j-k| \in \mathrm{Fib}(a_i^{\mathrm{Wyt}}, b_i^{\mathrm{Wyt}}), |j-k| \le w_i\big\}, \tag{3}$$

where the window size w_i of the i-th head is based on two model-wide hyperparameters $w_{\min}$ and $w_{\max}$, which are chosen based on insights into the modality of the task and the data distribution. Specifically, we choose w_i based on the formula, $w_i = w_{\min} + \left\lfloor \frac{w_{\max}-w_{\min}}{h-1}(i-1) \right\rfloor$ for $i = 1, \ldots, h$, which linearly interpolates between $w_{\min} \le N$, the *minimal window size bound across heads*, and the *maximal window size bound across heads* $w_{\max}$ satisfying $w_{\min} \le w_{\max} \le N$. The resulting *spacing* of window sizes across heads is designed to further diversify the representations learned across heads as, in the case of a large disparity between $w_{\min}$ and $w_{\max}$, heads with lower indices i are biased to encode more local information, whereas heads with $w_i \approx w_{\max}$ are biased towards incorporating more global interactions. Following (1), we define Fibottention's sparse attention matrices as $A_i^{\Omega_i} = A_i \odot \iota_{\Omega_i}$ for each $i = 1, \ldots, h$ with Ω_i satisfying (3).

For Transformer architectures with several MHSA layers, we further require that the head-wise support sets are shuffled along the layer so that the i-th head uses the sets, $\{\Omega_{\pi(1)}, \ldots, \Omega_{\pi(h)}\}$ within Fibottention, where $\pi : [h] \to [h]$ is a random permutation function (fixed for each layer). We refer to Appendix B for a formal outline.

Modified Wythoff Array. While we observe excellent performance of vanilla Fibottention in image classification tasks (see Section 4), its performance degrades in tasks in other domains due to its high degree of sparsity, which might not always capture well enough important local interactions. For such cases, we propose a variant of this sparse attention mechanism that *includes two predecessor sequence elements* into each Wythoff row sequence $\mathrm{Fib}(a_i^{\mathrm{Wyt}}, b_i^{\mathrm{Wyt}})$; following the recurrence (2), we can define new initial sequence elements $b_i^{\text{Wyt-m}} = b_i^{\mathrm{Wyt}} - a_i^{\mathrm{Wyt}}$ and $a_i^{\text{Wyt-m}} = a_i^{\mathrm{Wyt}} - b_i^{\text{Wyt-m}}$ and support sets $\Omega_i = \Omega_{w_i}^{\mathrm{Fib}(a_i^{\text{Wyt-m}}, b_i^{\text{Wyt-m}})}$ for each head index i. Unlike for the original Wythoff array, it is not the case anymore that the resulting sequences contain each integer only at most once (Morrison, 1980; Conway & Ryba, 2016); on the other hand, it can be proven that this modified Fibottention shares each query-key interaction pair only across at most *three* heads (Conway & Ryba, 2016). The differences in the resulting support set patterns are visualized in Figure 5a and Table 1. We refer to Appendix B for a detailed outline that includes both the (default) Wythoff and the Modified Wythoff variants of Fibottention.

In Appendix A.3, we provide a proof that the total computational effort for inference in both Fibottention variants requires the computation of only $\mathcal{O}(N \log(w_{\max}))$ token interactions.

Table 2: Performance comparison on CIFAR-10, CIFAR-100, Tiny-ImageNet, and ImageNet datasets with all methods integrated in ViT-B, highlighting the effect of attention pruning across approaches.

Method	**Top-1 Accuracy (%) ↑**				**Pruning Ratio ↑**
	C10	C100	Tiny-IN	IN-1K	
Full Attention (Vaswani et al., 2017)	84.2	59.4	<u>75.2</u>	**75.9**	0%
Random Attention (Zaheer et al., 2020)	80.7	56.5	69.4	68.7	98.52%
Top-k Attention (Gupta et al., 2021)	81.1	57.1	72.9	73.4	98.48%
Sparse Transformer (Child et al., 2019)	81.3	58.2	70.3	68.7	98.47%
BigBird (Zaheer et al., 2020)	86.8	63.4	73.4	71.5	97.96%
Longformer (Beltagy et al., 2020)	<u>87.8</u>	<u>64.7</u>	74.3	71.6	98.47%
Linformer (Wang et al., 2020)	73.1	48.7	62.8	60.1	97.96%
Efficient Attention (Shen et al., 2021)	84.4	62.6	73.7	70.1	97.98%
Fibottention (Ours)	**91.8**	**70.7**	**79.1**	<u>75.5</u>	98.01%

4 Fibottention for Image Classification

In this section, we evaluate the performance of Fibottention across various image classification tasks in conjunction with different Transformer architectures, and compare the design to state-of-the-art sparse and efficient attention designs.

4.1 Experimental Setup

Datasets. We report Top-1 accuracy on CIFAR-10 (C10) (Krizhevsky, 2009), CIFAR-100 (C100) (Krizhevsky, 2009), Tiny-ImageNet (Deng et al., 2009), and ImageNet-1K (IN-1K) (Deng et al., 2009).

Training. For training Fibottention and other sparse attention methods with various Transformer architectures, we use the training recipe of DeiT (Touvron et al., 2021). All models are trained from *random initialization* for 100 epochs with an effective batch size of 64 using four 48GB A6000 GPUs, employing ViT-Base (ViT-B) (Dosovitskiy et al., 2020; Touvron et al., 2021) unless otherwise specified. The hyperparameters are set as $w_{\min} = 5$, and $w_{\max} = 65$ unless otherwise stated.

4.2 Experimental Results

Fibottention vs. Other Sparse Attention Mechanisms. Table 2 compares our proposed Fibottention against representative sparse-attention baselines across image classification datasets. We include a standard ViT-B (Touvron et al., 2021) with full self-attention (Vaswani et al., 2017) as a dense reference, alongside random pruning, Top-k pruning (Gupta et al., 2021), Sparse Transformer (Child et al., 2019), Efficient Attention (Shen et al., 2021), Longformer (Beltagy et al., 2020), BigBird (Zaheer et al., 2020), and Linformer (Wang et al., 2020). For a fair comparison, all sparse variants are configured to operate at a similar attention cost of approximately 0.014 GFLOPs, corresponding to pruning ratios close to 98%, whereas the dense ViT-B baseline requires 0.72 GFLOPs. Despite this extreme sparsity, Fibottention consistently delivers strong performance across datasets. It substantially improves over the ViT-B baseline using full attention on C10, C100, and Tiny-IN, while remaining competitive on the more challenging IN-1K benchmark. In particular, although a small accuracy gap remains on IN-1K, Fibottention achieves this performance while explicitly evaluating only about 2% of token-to-token interactions, highlighting a favorable accuracy–efficiency trade-off at large scale. These results indicate that a significant portion of dense self-attention in ViTs is redundant, and that carefully designed sparse patterns can preserve, or even enhance, recognition accuracy under aggressive pruning.

Table 3: Performance of Fibottention integrated in various ViT variants, achieving 94–98% attention pruning while maintaining competitive Top-1 accuracy.

ViT Variants	**Top-1 Accuracy (%) ↑**				**Pruning Ratio ↑**
	C10	C100	Tiny-IN	IN-1K	
ViT-B (Touvron et al., 2021)	84.2	59.4	75.2	**75.9**	0%
+ Fibottention	**91.8**	**70.7**	**79.1**	75.5	**98.0%**
UPop (Shi et al., 2023)	76.3	52.1	71.6	**73.9**	0%
+ Fibottention	**82.8**	**56.8**	**74.8**	72.0	**97.1%**
iFormer (Zheng, 2025)	**92.7**	**73.3**	**88.4**	**77.9**	0%
+ Fibottention	92.2	72.7	88.1	77.1	**94.4%**
Swin-B (Liu et al., 2021)	81.2	60.7	**82.8**	**79.6**	0%
+ Fibottention	**82.4**	**61.0**	81.8	78.3	**95.4%**
ConViT-B (d'Ascoli et al., 2021)	90.8	66.8	**82.9**	**82.1**	0%
+ Fibottention	**90.9**	**67.5**	82.8	80.1	**96.6%**

To better understand the impact of sparsity patterns in this high-pruning regime, we outline how different methods allocate their limited budget of token interactions: In Table 2, *BigBird* and *Longformer* denote adaptations of the sparse attention schemes introduced in (Zaheer et al., 2020; Beltagy et al., 2020), respectively, applied to ViTs under comparable pruning ratios. Since all sparse models operate at similar computational cost, observed performance differences primarily arise from how the remaining interactions are structured. Random pruning removes a fixed fraction of random attention entries without imposing any spatial or semantic structure. Although its computational cost matches that of structured methods, its accuracy is markedly lower across all datasets, including IN-1K, demonstrating that unstructured sparsity fails to preserve critical token relationships. Top-k pruning, which retains the k largest attention scores per query token, performs better than random masking, confirming that saliency-based criteria are beneficial under sparsity. However, Top-k pruning underperforms structured approaches such as BigBird, Longformer and Fibottention on C10 and C100 by 5%–10%, with a reduced performance gap on the mid-sized Tiny-IN dataset. We observe that while top-k attention exceeds the performance of most structured sparse patterns on the large-scale IN-1K dataset, Fibottention is the only mechanism that still outperforms top-k for a comparable pruning ratio and is the only efficient attention mechanism that comes close to (by 0.4% top-1 accuracy) the performance of a ViT-B trained with full attention.

We conjecture that a main reason for the strong performance of Fibottention is its head-specific, complementary masks, which enable distinct heads to capture local or global information in diverse and non-redundant feature representations. This phenomenon is further studied in Section 6.1 and Table 7.

Fibottention Integrated into Various ViT Variants. In Table 3, we investigate the effect of integrating Fibottention into a variety of ViT variants: ViT-B (Touvron et al., 2021), UPop (Shi et al., 2023), iFormer (Zheng, 2025), Swin-B (Liu et al., 2021), and ConViT-B (d'Ascoli et al., 2021) (see Appendix C.1 for further details about the setup). For ViT-B and UPop, Fibottention consistently yields higher accuracy on C10, C100, and Tiny-IN compared to their respective dense counterparts, while maintaining comparable performance on IN-1K for ViT-B and a slightly lower IN-1K accuracy for UPop. In both cases, the attention GFLOPs are reduced to only 2.0% (ViT-B) and 2.9% (UPop) of the dense baseline, demonstrating that substantial computational savings are possible even when the dense model is already well tuned. For iFormer, which already embeds a strong inductive bias via single-head attention, incorporating Fibottention results in an accuracy that remains close to the original backbone across all datasets, while reducing attention FLOPs to 5.6% of the original cost, leading to a reasonable efficiency-effectiveness trade-off. For hierarchical backbones such as Swin-B and ConViT-B, which introduce local inductive biases through token merging or convolutions, Fibottention yields accuracy that is broadly comparable to the base models. For Swin-B, the performance with Fibottention stays within a narrow band of the dense model on all four datasets, while

Table 4: Top-1 accuracy of ViT variants trained with sparse attention mechanisms, configured to compute only $\approx 2\%$ of self-attention entries.

Method	ViT-B		UPop		ConViT-B		Pruning Ratio ↑
	C10	C100	C10	C100	C10	C100	
Full Attention (Vaswani et al., 2017)	84.2	59.4	76.3	52.1	90.8	66.8	0%
Random Attention (Zaheer et al., 2020)	80.7	56.5	72.3	46.8	89.1	64.5	98.52%
Top-k Attention (Gupta et al., 2021)	81.1	57.1	71.5	36.8	89.6	64.4	98.48%
BigBird (Zaheer et al., 2020)	86.8	63.4	79.1	37.7	89.8	64.7	97.96%
Longformer (Beltagy et al., 2020)	87.8	64.7	79.6	39.4	89.2	64.3	98.47%
Fibottention (Ours)	**91.8**	**70.7**	**82.8**	**56.8**	**90.9**	**67.5**	98.01%

reducing attention GFLOPs to 4.6% of the original. For ConViT-B, Fibottention maintains similar accuracy on C10, C100, and Tiny-IN, and a slightly lower performance on IN-1K, but with attention GFLOPs reduced to 3.4% of the dense baseline. Overall, these results indicate that Fibottention is highly compatible with a range of Transformer architectures, delivering considerable computational savings together with improved performance for architectures with limited inductive bias when trained on small-sized datasets, and no or limited accuracy degradation when trained on mid-sized or large datasets compared to full attention variants, depending on the backbone and dataset.

Sparse Attention Mechanisms Across ViT Variants. Table 4 reports the performance of Fibottention and selected sparse attention methods across multiple ViT-style backbones (ViT-B, UPop, and ConViT-B) trained on C10 and C100 under a high sparsity setting where only 2% of self-attention entries are retained. We observe that among the considered sparse attention methods, only Fibottention consistently improves accuracy over the full attention baseline. This trend persists across all evaluated backbones, indicating that Fibottention's advantages robustly generalize across Transformer architectures.

5 Fibottention in Other Visual Domains

We demonstrate the versatility of Fibottention by integrating it into ViTs designed for solving visual perception tasks beyond image classification.

5.1 Video Action Classification

Datasets. We evaluate and report top-1 action classification accuracy for Fibottention using two action recognition datasets: Toyota Smarthome (Das et al., 2019) and Northwestern-UCLA Multiview Activity 3D (NUCLA) (Wang et al., 2014). The Toyota Smarthome dataset comprises ∼16K videos across 31 classes. Here, we adhere to the cross-subject (CS) and cross-view (CV2) protocols. The NUCLA dataset consists of ∼1.2K video clips with subjects performing 10 different action classes and we use the cross-subject (CS) protocol.

Table 5: Fibottention Top-1 accuracy on Smarthome and NUCLA.

Method	Smarthome		NUCLA
	CS	CV2	CS
TimeSformer (Bertasius et al., 2021)	52.2	36.6	32.9
+ BigBird (Zaheer et al., 2020)	51.4	40.1	50.9
+ Fibottention (Wythoff)	55.6	38.6	49.3
+ Fibottention (Modified Wythoff)	**57.1**	**42.3**	**59.6**

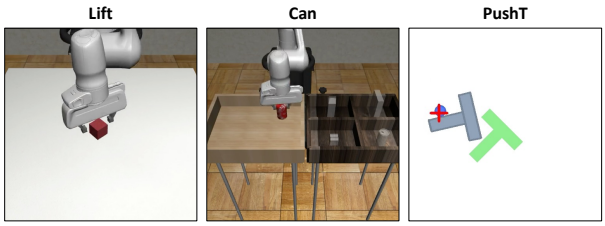


Figure 2: Sample frames from the datasets used in our robotics experiments.

Table 6: Performance of Fibottention on behavioral cloning for robotics. The average task completion accuracy is reported.

Visual Backbone	Lift	Can	PushT
ViT-B (Touvron et al., 2021)	0.980	0.960	0.678
+ BigBird (Zaheer et al., 2020)	1.000	0.880	0.690
+ Fibottention (Wythoff)	0.820	0.940	0.630
+ Fibottention (Modified Wythoff)	**1.000**	**0.960**	**0.720**

Training. For this experiment, we employ the divided-space-time attention variant of TimeSformer (Bertasius et al., 2021) for action classification. We integrate Fibottention into the spatial attention module of TimeSformer, considering that the temporal attention module already processes dense attention across the same patch in contiguous frames. For the implementation of Fibottention, we use both its Wythoff and Modified Wythoff variants. The hyperparameters for Fibottention are set to $w_{\text{min}} = 1$ and $w_{\text{max}} = 196$. Given that the TimeSformer architecture fundamentally resembles a ViT-B with additional attentional modules, it is initialized with IN-1K pre-trained weights. All video models are trained with a batch size of 32 for 15 epochs. For the Toyota Smarthome dataset, we process video clips of size $8 \times 224 \times 224$ with a sampling rate of 1/32, while for NUCLA, we use video clips of size $16 \times 224 \times 224$ with a sampling rate of 1/4.

Results. In Table 5, we compare the action classification results using TimeSformer and other attention mechanisms (BigBird, and Fibottention) integrated within TimeSformer. We observe that Fibottention with Modified Wythoff instantiation outperforms all baselines on the Smarthome and NUCLA protocols, utilizing a masking percentage of 94%. Fibottention with Modified Wythoff facilitates increased local interactions among query-key pairs compared to the original Wythoff sequences, albeit at the expense of a reduced masking ratio (by 1.5%). The Modified Wythoff proves essential in our video experiments, where capturing the temporal evolution of local patches is critical for learning discriminative spatiotemporal representations.

5.2 Robot Learning

For robotics experiments, we assess the performance of Fibottention for behavioral cloning (Florence et al., 2022) in which we aim to learn a robot policy by training a model on state-action pairs obtained from human examples.

Datasets. We evaluate three datasets: Can and Lift from Robomimic (Mandlekar et al., 2022), and PushT from Implicit Behavioral Cloning (Florence et al., 2022). In Lift, the robot must lift a cube to a specific height. In Can, the robot must move a can into a box. In PushT, the robot must align a T-shaped block with a T-shaped outline. We provide visuals of all three datasets in Figure 2.

Training. Building upon the Crossway Diffusion (Li et al., 2024) framework, we modify the architecture by substituting the ResNet visual backbone with a ViT (Dosovitskiy et al., 2020) and incorporating Fibottention into standard ViT self-attention layers. We employ a batch size of 64 and utilize ViT-B with a patch size of 8 as the visual backbone. For all other hyperparameters, including the number of epochs, we follow (Li et al., 2024).

Results. We report the average task completion accuracy in Table 6 and find that Fibottention with Modified Wythoff instantiation leads to improvements over both the baseline ViT and ViT with BigBird attention.

6 Ablation Studies and Analytical Findings

Beyond the experiments above, we perform a set of controlled ablations to analyze the effect of individual architectural and sparsity choices in Fibottention. Unless explicitly stated, all experiments in this section use ViT-B as the visual backbone.

Table 7: Head diversity statistics (Frobenius distances across last-layer features) over 10^4 CIFAR-10 images. Fibottention yields substantially higher diversity across heads.

Method	Min	Max	Median	Q1	Q3
ViT-B (Touvron et al., 2021)	27.57	61.53	43.00	37.68	48.85
+ Random Attention (Zaheer et al., 2020)	13.68	47.74	29.24	22.08	35.31
+ Sparse Transformer (Child et al., 2019)	28.84	63.71	43.27	38.27	49.65
+ Longformer (Beltagy et al., 2020)	34.49	65.27	49.26	44.52	54.29
+ BigBird (Zaheer et al., 2020)	34.15	72.08	51.86	45.67	58.55
+ Fibottention	41.63	75.95	**57.34**	51.98	63.22

6.1 Validation of Head Diversity

One of the design principles behind Fibottention is that the resulting feature representations are more *diverse across heads* than those of standard MHSA. To validate this claim, we consider the following analysis protocol: For ViT-B trained on C10 with different sparse attention mechanisms, we perform inference for single CIFAR-10 images X and compute the last-layer feature matrices $Y_i = \sigma(A_i^{\Omega_i})V_i$ for each head $i = 1, \ldots, h$, where A^{Ω_i} is as in (1). Then we measure the distances $\|Y_i - Y_j\|_F$ of feature representations across heads for each pair (Y_i, Y_j), $1 \leq i < j \leq h$ and average such relative distances in the diversity metric

$$\text{diversity}(X) = \frac{2}{h(h-1)} \sum_{i<j} \frac{\|Y_i - Y_j\|_F}{\|Y_i\|_F + \|Y_j\|_F},$$

which we report for many different input images X. Within this framework, higher values of diversity(X) indicate greater variability in information captured by different heads for a given input X. Considering 10^4 input images X, we report aggregated information (minimum, maximum, quartiles and median) of the distribution of the head diversity metric diversity(X) in Table 7 for the considered sparse attention methods. We observe that Fibottention consistently attains a significantly larger median head diversity than ViT-B and all other sparse attention mechanisms. Arguably, this is due to Fibottention's deliberate use of complementary, head-specific sparsity patterns, whereas the sparse baselines apply the same mask across heads. The increased observed head diversity indicates that Fibottention captures richer and more complementary feature representations than other sparse variants, which we believe is key to its superior performance under constrained FLOPs (see Table 2).

6.2 Validation of Inductive Bias

To validate the inductive bias in Fibottention, we examine its "inside-to-inside" attention patterns within specific regions of input images. Fibottention's structured pruning mechanism is designed to promote fo-

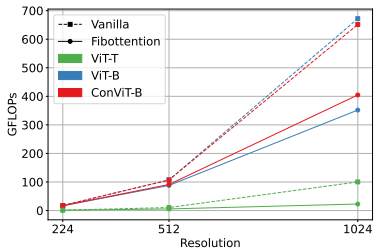


Figure 3: Inference costs of ViT-B, ViT-T, ConViT-B (Vanilla vs. Fibottention).

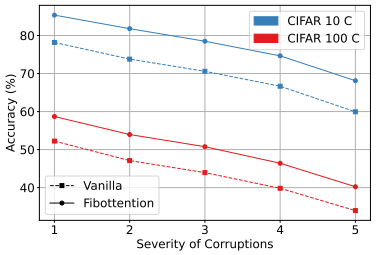


Figure 4: Test accuracy of ViT-B models for corrupted datasets CIFAR-10 C and CIFAR-100 C, trained with and without Fibottention.

Table 8: Effect of different dilation sequences $(f_n)_{n\in\mathbb{N}}$ on CIFAR-10 (Krizhevsky, 2009) and CIFAR-100 (Krizhevsky, 2009), with a fixed window size $w_i = N/3$.

Sequences	C10	C100
$(2^n)_{n\in\mathbb{N}}$	86.1	61.6
$(3^n)_{n\in\mathbb{N}}$	85.3	60.2
$(n^2)_{n\in\mathbb{N}}$	85.8	61.5
$(n^3)_{n\in\mathbb{N}}$	84.6	59.1
Fib$(1,1)$	**87.3**	**63.2**

cused attention within predefined boundaries. We hypothesize that this mechanism leads to higher average attention scores and lower variance within localized areas, reflecting a stronger inductive bias toward these regions. To test this hypothesis, we calculate the variance of attention scores within object-localized regions by aggregating and analyzing attention values across designated patches. This approach enables a direct comparison of attention distribution patterns between a Fibottention-trained ViT-B model and a corresponding model with full attention. Results averaged over 100 images show that Fibottention achieves a significantly lower variance of attention scores within localized areas (2.33×10^{-5}) compared to full attention ViT-B (8.76×10^{-5}). This reduced variance indicates that Fibottention maintains more concentrated and consistent attention within target regions.

6.3 Computational Complexity

We recall that the inference FLOPs that accrue at a Fibottention-modified self-attention layer scale with $\mathcal{O}(N \log N)$ if N is the number of tokens/patches (cf. Appendix A.3). To illustrate the implications of this in practice, we compare the projected inference cost per input of ViT-B (Dosovitskiy et al., 2020), ViT-T (Wu et al., 2022), and ConViT (d'Ascoli et al., 2021) in Figure 3, considering their dense (i.e., *vanilla*) self-attention alongside their Fibottention variants. It is well known (Dosovitskiy et al., 2020; Kaplan et al., 2020) that for smaller resolutions, a substantial part of the computation comes from non-MHSA components of the Transformer architectures, such as the feed-forward/MLP blocks (Kaplan et al., 2020), which is why the total inference FLOPs improvement of Fibottention (or any sparse attention method, for that matter) is rather limited at a 224×224 resolution. On the other hand, we see in Figure 3 that as resolution, and thus, N increases, self-attention computation becomes a much larger fraction of the total computational load so that Fibottention is able to reduce FLOP counts by up to 48%. In ConViT (d'Ascoli et al., 2021), which contains 10 GPSA and 2 MHSA modules, we apply Fibottention only to the corresponding MHSA modules.

6.4 Robustness of Fibottention

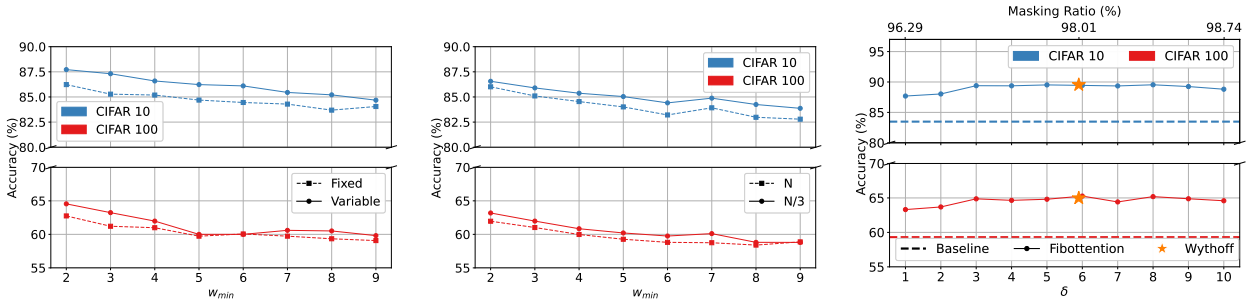
To understand Fibottention's robustness with respect to distribution shifts, we evaluate its impact on predictive performance for the C10 and C100 *corrupted* (Hendrycks & Dietterich, 2019) datasets. These datasets contain test images from the original datasets, corrupted by 19 common image corruptions with severity levels ranging between 1–5 (see Appendix E.3 for more details). For the evaluation, we use models pre-trained on default C10 and C100 data, and evaluate them on the corrupted datasets across varying severity. Figure 4 shows that Fibottention models perform consistently better than their dense MHSA counterparts across all corruption severity levels.

6.5 Choice of Dilation Sequences

Table 8 compares several choices of dilation sequences $(f_n)_n$ used to construct the diagonal offsets in the sparsity sets Ω_i, while keeping a fixed window size of $w_i = w_{\min} = w_{\max} = N/3$ for all heads. We include polynomial growth sequences such as $(n^2)_{n\in\mathbb{N}}$ and $(n^3)_{n\in\mathbb{N}}$, as well as exponential sequences $(2^n)_{n\in\mathbb{N}}$ and $(3^n)_{n\in\mathbb{N}}$ that resemble dilated patterns commonly used in prior work (Li et al., 2019). Among all options considered, the standard Fibonacci sequence $\mathrm{Fib}(1,1)$ yields the highest accuracy on both CIFAR-10 and CIFAR-100. Intuitively, Fibonacci growth is slow enough that many small offsets remain available, which reinforces local interactions, but still introduces a few longer-range diagonals. This balance appears to be better suited to image data than the more aggressive growth patterns of the alternative sequences.

6.6 Impact of Dilation Across Heads

In Figure 5(a) we examine the role of varying dilation sequences across heads. We compare two settings that both use sequences of the form $(f_n)_{n\in\mathbb{N}} = (c \cdot n)_{n\in\mathbb{N}}$: a *fixed* configuration in which all heads share the same dilation sequence, and a *variable* configuration in which each head receives a shifted version of the sequence, so that offsets differ from head to head. At comparable masking ratios, allowing the dilation patterns to vary across heads consistently improves performance. For instance, at $w_{\min} = 2$, the masking ratio increases from 72.8% for fixed sequences to 85.5% for variable sequences, while resulting in improved accuracy. This



(a) (b) (c)

Figure 5: Ablation study of (a) impact of dilation with sequences $(f_n)_{n\in\mathbb{N}} = (cn)_{n\in\mathbb{N}}$ fixed and variable across heads where $w_i = 5h_i$; (b) choice of $w_{\max}$ with sequences $(f_n)_{n\in\mathbb{N}} = \mathrm{Fib}(w_{\min}, 2w_{\min})$ where $w_{\max} = N$ and $w_{\max} = N/3$ fixed across all heads; and (c) variable dilation sequences $\mathrm{Fib}(i+\delta, i+\delta)$ for the i-th head, $i \in \{1, \ldots, 12\}$, with varying δ, vs. Wythoff.

confirms that head-wise diversity in the sparsity pattern is an effective way to maintain performance even for increased attention sparsity levels.

6.7 Choice of Window Size Limit $w_{\max}$

Figure 5(b) compares two ways of setting the maximum window size $w_{\max}$ when using Fibonacci sequences of the form $(f_n)_{n\in\mathbb{N}} = \mathrm{Fib}(w_{\min}, 2w_{\min})$: a fully global setting where $w_{\max} = N$, and a more local setting where $w_{\max} = N/3$. We observe that $w_{\max} = N/3$ consistently leads to better accuracy than $w_{\max} = N$. In image classification most of the discriminative structure tends to be spatially localized, so interactions with very distant tokens are often less informative and can even dilute object-level features. Restricting $w_{\max}$ to roughly a third of the sequence length encourages attention to focus on more relevant neighborhoods. In addition to this, we study in Appendix E.2 and Table 13 the special case of coinciding maximum and minimum window sizes, i.e., $w_{\max} = w_{\min}$, resulting in a shared fixed window size w_i across all heads. These experiments likewise suggest that moderate window sizes can be beneficial for model accuracy.

6.8 Why Wythoff?

The generalized Fibonacci sequences appearing in the Wythoff array have the special property that each positive integer appears in exactly one row, which in our context translates into minimal overlap of diagonal offsets across heads. To investigate this question, we compare in Figure 5(c) Fibottention configured with Wythoff-based generalized Fibonacci sequences against several families of head-specific Fibonacci sequences that use additional offset hyperparameters. For all models in this comparison, we fix $w_{\min} = 5$ and $w_{\max} = N/3$. We observe that the Wythoff configuration attains the highest accuracy, while having no extra hyperparameters beyond $(w_{\min}, w_{\max})$. This suggests that the highly complementary and non-overlapping head masks of Fibottention's Wythoff construction are indeed a driver of the strong model performance of Transformers using Fibottention instead of MHSA in practice.

7 Conclusion

We introduce Fibottention, an efficient, robust, $\mathcal{O}(N \log N)$ sparse attention mechanism with a fixed sparsity pattern that diversifies attention computation across heads through Fibonacci dilation sequences chosen from the Wythoff array. We implement Fibottention in conjunction with multiple state-of-the-art Transformer architectures curated for visual representation learning. Empirically, Fibottention outperforms the baselines on small-scale and mid-scale datasets and achieves comparable performance on large-scale datasets utilizing only 2–6% of token interactions in the MHSA across three diverse visual tasks. We envision that the next generation of Transformers (Gemini Team, Google, 2023) processing inputs with billions of tokens may

benefit from such optimized architectures. It remains to future work to generalize this idea to application domains that involve causal attention, such as Transformer models for natural language processing tasks or time series analysis.

References

Abhinav Agarwalla, Abhay Gupta, Alexandre Marques, Shubhra Pandit, Michael Goin, Eldar Kurtic, Kevin Leong, Tuan Nguyen, Mahmoud Salem, Dan Alistarh, et al. Enabling High-Sparsity Foundational Llama Models with Efficient Pretraining and Deployment. *arXiv:2405.03594*, 2024.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The Long-Document Transformer. *arXiv:2004.05150*, 2020.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proc. Int. Conf. Mach. Learn. (ICML)*, volume 139, pp. 813–824, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 33:1877–1901, 2020.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 213–229, 2020.

Eric Chen, Adam Ge, Andrew Kalashnikov, Ella Kim, Evin Liang, Mira Lubashev, Matthew Qian, Rohith Raghavan, Benjamin Taycher, Samuel Wang, and Tanya Khovanova. Generalizing the Wythoff array and other Fibonacci facts to Tribonacci numbers. *J. Integer Seq.*, 28(5):Article 25.5.4, 2025.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating Long Sequences With Sparse Transformers. *arXiv:1904.10509*, 2019.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What Does BERT Look at? An Analysis of BERT's Attention. In *Proc. ACL Workshop BlackboxNLP: Anal. Interpret. Neural Netw. NLP at ACL 2019*, pp. 276–286, 2019.

John Conway and Alex Ryba. The Extra Fibonacci Series and the Empire State Building. *Math. Intellig.*, 38(1):41–48, 2016.

Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota Smarthome: Real-World Activities of Daily Living. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 833–842, 2019.

Stéphane d'Ascoli, Hugo Touvron, Matthew Leavitt, Ari Morcos, Giulio Biroli, and Levent Sagun. ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases. In *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 2286–2296, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 248–255, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. NAACL-HLT*, pp. 4171–4186, 2019.

Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. LongNet: Scaling Transformers to 1,000,000,000 Tokens. *arXiv preprint arXiv:2307.02486*, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 Words: Transformers for image recognition at scale. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.

Aritra Dutta, Srijan Das, Jacob Nielsen, Rajatsubhra Chakraborty, and Mubarak Shah. Multiview Aerial Visual Recognition (MAVREC): Can Multi-view Improve Aerial Visual Perception? In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 22678–22690, 2024.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming Transformers for High-Resolution Image Synthesis. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 12873–12883, 2021.

Pete Florence, Corey Lynch, Andy Zeng, Oscar Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. *Proc. 5th Conf. Robot. Learn. (CoRL 2021)*, 164:158–168, 2022.

Tianyu Fu, Haofeng Huang, Xuefei Ning, Genghan Zhang, Boju Chen, Tianqi Wu, Hongyi Wang, Zixiao Huang, Shiyao Li, Shengen Yan, Guohao Dai, Huazhong Yang, and Yu Wang. Mixture of Attention Spans: Optimizing LLM Inference Efficiency with Heterogeneous Sliding-Window Lengths. In *Proc. 2nd Conf. Lang. Model. (COLM)*, 2025.

Gemini Team, Google. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Qipeng Guo, Xipeng Qiu, Pengfei Liu, Yunfan Shao, Xiangyang Xue, and Zheng Zhang. Star-Transformer. In *Proc. NAACL HLT*, volume 1, pp. 1315–1325, 2019.

Ankit Gupta, Guy Dar, Shaya Goodman, David Ciprut, and Jonathan Berant. Memory-efficient Transformers via Top-k Attention. In *Proc. 2nd Workshop Simple Effic. Nat. Lang. Process. (SustaiNLP) at EMNLP 2021*, pp. 39–52. Association for Computational Linguistics, 2021.

Ali Hassani and Humphrey Shi. Dilated Neighborhood Attention Transformer. *arXiv preprint arXiv:2209.15001*, 2022.

Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood Attention Transformer. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 6185–6194, June 2023.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. Learning Clip Representations for Skeleton-Based 3D Action Recognition. *IEEE Trans. Image Process.*, 27(6):2842–2855, June 2018. ISSN 1941-0042. doi: 10.1109/TIP.2018.2812099.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M Rush. Structured attention networks. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The Efficient Transformer. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.

Thomas Koshy. *Fibonacci and Lucas Numbers with Applications*, volume 2. John Wiley & Sons, Hoboken, NJ, 2019.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the Dark Secrets of BERT. In *Proc. Conf. Empirical Methods Nat. Lang. Process. 9th Int. Joint Conf. Nat. Lang. Process. (EMNLP-IJCNLP)*, pp. 4365–4374, 2019.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.

Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting. *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 32, 2019.

Xiang Li, Varun Belagali, Jinghuan Shang, and Michael S. Ryoo. Crossway Diffusion: Improving Diffusion-based Visuomotor Policy via Self-supervised Learning. In *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 16841–16849, 2024.

Xingyang Li, Muyang Li, Tianle Cai, Haocheng Xi, Shuo Yang, Yujun Lin, Lvmin Zhang, Songlin Yang, Jinbo Hu, Kelly Peng, Maneesh Agrawala, Ion Stoica, Kurt Keutzer, and Song Han. Radial Attention: $O(n \log n)$ Sparse Attention for Long Video Generation. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2025.

Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. MViTv2: Improved Multiscale Vision Transformers for Classification and Detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 4804–4814, 2022.

Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models. *arXiv preprint arXiv:2512.02556*, 2025.

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. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 10012–10022, 2021.

Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *Proc. 4th Conf. Robot. Learn. (CoRL 2021)*, volume 164, pp. 1678–1690. PMLR, 2022.

David R. Morrison. A Stolarsky array of Wythoff pairs. In V. E. Hoggatt Jr. and M. Bicknell-Johnson (eds.), *A Collection of Manuscripts Related to the Fibonacci Sequence*, volume 38, pp. 134–136. The Fibonacci Association, 1980.

OEIS Foundation Inc. The Fibonacci Numbers, 1964. URL https://oeis.org/A000045. Entry A000045: The On-Line Encyclopedia of Integer Sequences.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, et al. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2025.

Zheng Qu, Liu Liu, Fengbin Tu, Zhaodong Chen, Yufei Ding, and Yuan Xie. DOTA: detect and omit weak attentions for scalable transformer acceleration. In *Proc. ACM Archit. Support Program. Lang. Oper. Syst. (ASPLOS)*, pp. 14–26, 2022.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018.

Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 32, 2019.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 34, pp. 13937–13949, 2021.

Brendan C Reidy, Mohammadreza Mohammadi, Mohammed E Elbtity, and Ramtin Zand. Efficient deployment of Transformer models on edge TPU accelerators: A real system evaluation. In *Architecture and System Support for Transformer Models (ASSYST) Workshop, IEEE/ACM Int. Symp. Comput. Arch. (ISCA)*, 2023.

Dominick Reilly and Srijan Das. Just add π! pose induced video transformers for understanding activities of daily living. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 18340–18350, 2024.

Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. Efficient Content-Based Sparse Attention With Routing Transformers. *Trans. Assoc. Comput. Ling.*, 9:53–68, 2021.

Michael Eli Sander, Joan Puigcerver, Josip Djolonga, Gabriel Peyré, and Mathieu Blondel. Fast, differentiable and sparse top-k: a convex analysis perspective. In *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 29919–29936, 2023.

Jinghuan Shang, Kumara Kahatapitiya, Xiang Li, and Michael S. Ryoo. Starformer: Transformer with state-action-reward representations for visual reinforcement learning. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 462–479, 2022.

Zihang Shen, Mingyuan Zhu, Jiayu Cheng, Yelong Wang, Yujia Sun, Jingjing Zhang, and Jianfeng Wang. Efficient Attention: Attention with Linear Complexities. In *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, pp. 3531–3539, 2021.

Dachuan Shi, Chaofan Tao, Ying Jin, Zhendong Yang, Chun Yuan, and Jiaqi Wang. UPop: Unified and Progressive Pruning for Compressing Vision-Language Transformers. In *Proc. Int. Conf. Mach. Learn. (ICML)*, volume 202, pp. 31292–31311. PMLR, 2023.

Han Shi, Jiahui Gao, Xiaozhe Ren, Hang Xu, Xiaodan Liang, Zhenguo Li, and James Tin-Yau Kwok. SparseBERT: Rethinking the Importance Analysis in Self-attention. In *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 9547–9557, 2021.

Laurence Sigler. *Fibonacci's Liber Abaci: a translation into modern English of Leonardo Pisano's book of calculation*. Springer Science & Business Media, New York, NY, 2003.

Guangyu Sun, Matias Mendieta, Aritra Dutta, Xin Li, and Chen Chen. Towards Multi-modal Transformers in Federated Learning. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 229–246, 2024.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 2818–2826, 2016.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *Proc. Int. Conf. Mach. Learn. (ICML)*, volume 139, pp. 10347–10357, 2021.

Shikhar Tuli and Niraj K. Jha. EdgeTran: Device-Aware Co-Search Of Transformers for Efficient Inference on Mobile Edge Platforms. *IEEE Trans. Mobile Comput.*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, pp. 5998–6008, 2017.

Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-View Action Modeling, Learning, and Recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 2649–2656, June 2014. doi: 10.1109/CVPR.2014.339.

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 37, pp. 48784–48809, 2020.

Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. Multi-passage BERT: A globally normalized BERT model for open-domain question answering. In *Proc. Conf. Empirical Methods Nat. Lang. Process. 9th Int. Joint Conf. Nat. Lang. Process. (EMNLP-IJCNLP)*, pp. 5878–5882, 2019.

Cong Wei, Brendan Duke, Ruowei Jiang, Parham Aarabi, Graham W Taylor, and Florian Shkurti. Sparsifiner: Learning Sparse Instance-Dependent Attention for Efficient Vision Transformers. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 22680–22689, 2023.

Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. CvT: Introducing Convolutions to Vision Transformers. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 22–31, 2021.

Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. TinyViT: Fast Pretraining Distillation for Small Vision Transformers. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 68–85, 2022.

Willem A. Wythoff. A modification of the game of Nim. *Nieuw Arch. Wisk*, 7(2):199–202, 1907.

Chong You, Kan Wu, Zhipeng Jia, Lin Chen, Srinadh Bhojanapalli, Jiaxian Guo, Utku Evci, Jan Wassenberg, Praneeth Netrapalli, Jeremiah J Willcock, et al. Spark Transformer: Reactivating Sparsity in FFN and Attention. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2025.

Xiaoyu Yue, Shuyang Sun, Zhanghui Kuang, Meng Wei, Philip Torr, Wayne Zhang, and Dahua Lin. Vision Transformer with Progressive Sampling. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 387–396, 2021.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Are Transformers universal approximators of sequence-to-sequence functions? In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020a.

Chulhee Yun, Yin-Wen Chang, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. O(n) connections are expressive enough: Universal approximability of sparse transformers. *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 33:13783–13794, 2020b.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big Bird: Transformers for Longer Sequences. *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 33:17283–17297, 2020.

Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng Gao. Multi-Scale Vision Longformer: A New Vision Transformer for High-Resolution Image Encoding. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 2998–3008, 2021.

Zhemin Zhang and Xun Gong. Vision Big Bird: Random Sparsification for Full Attention. *arXiv:2311.05988*, 2023.

Guangxiang Zhao, Junyang Lin, Zhiyuan Zhang, Xuancheng Ren, Qi Su, and Xu Sun. Explicit Sparse Transformer: Concentrated Attention Through Explicit Selection. *arXiv preprint arXiv:1912.11637*, 2019.

Chuanyang Zheng. iFormer: Integrating ConvNet and Transformer for Mobile Application. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, pp. 22947–22961, 2025.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.

A Computational Complexity of Fibottention

In this section, we analyze the time complexity of using Fibottention within a forward pass of a Transformer architecture. To this end, we first recall an explicit formula for the n-th sequence element of a generalized Fibonacci sequence in Appendix A.1. We then use this to bound the sparsity of the evaluated attention matrix in Appendix A.2 for a given attention head. Finally, we use this result to bound the total computational complexity of Fibottention in Appendix A.3.

A.1 Generalized Binet's Formula

We state a well-known generalization of Binet's formula (Koshy, 2019) to generalized Fibonacci sequences, which states that the n-th sequence element f_n of the standard Fibonacci sequence $\mathrm{Fib}(0,1)$ (i.e., the sequence $(f_n)_n = (0,1,1,2,3,5,8,13,\cdots)$) satisfies

$$f_n = \frac{1}{\sqrt{5}}\left(\phi^{n-1} - \psi^{n-1}\right),$$

where $\phi = \frac{1+\sqrt{5}}{2}$ and $\psi = \frac{1-\sqrt{5}}{2}$ are the two solutions of the quadratic equation defining the golden ratio. This provides an explicit, non-recursive characterization of the n-th sequence element.

It turns out that this formula can be generalized to generalized Fibonacci sequences $\mathrm{Fib}(a,b)$ as defined in Section 3.2, which follow the same linear recurrence relation eq. (2), but which start at $f_1 = a$ and $f_2 = b$ for $a, b \in \mathbb{N}$, see Lemma 1.

Lemma 1 (Generalized Binet's Formula (Koshy, 2019))**.** *If $Fib(a,b) = (f_n)_{n\in\mathbb{N}}$ is the generalized Fibonacci sequence with initial values $f_1 = a$ and $f_2 = b$, then it holds that*

$$f_n = \frac{a-(b-a)\psi}{\sqrt{5}}\phi^n + \frac{(b-a)\phi - a}{\sqrt{5}}\psi^n \tag{4}$$

for each $n \geq 1$ and

$$f_n = \frac{b-a\psi}{\sqrt{5}}\phi^{n-1} + \frac{a\phi - b}{\sqrt{5}}\psi^{n-1} \tag{5}$$

for each $n \geq 2$, where $\phi = (1+\sqrt{5})/2$ and $\psi = (1-\sqrt{5})/2$.

The proof of Lemma 1 is standard and follows the proof of the conventional Binet's formula. We provide it below for completeness.

Proof of Lemma 1. We note that the defining linear recurrence relation

$$f_{n+1} = f_n + f_{n-1}.$$

of the generalized Fibonacci sequence is homogeneous and has the characteristic equation

$$x^2 - x - 1 = 0, \tag{6}$$

which has the roots ϕ and ψ as defined in Lemma 1. Due to the homogeneity of the linear recurrence relation, it follows that

$$f_n = A\phi^n + B\psi^n,$$

where A and B are constants to be determined from the initial conditions $f_1 = a$ and $f_2 = b$. In particular, from $f_1 = a$, we obtain that

$$a = A\phi + B\psi$$

and furthermore, from $f_2 = b$, we see that

$$b = A\phi^2 + B\psi^2 = A(\phi+1) + B(\psi+1),$$

where we used the characteristic equation (6) for solutions ϕ and ψ. For this, we obtain the system of equations

$$(A+B) + a = b,$$

$$A\phi + B\psi = a.$$

From the first equation, we obtain $B = (b-a) - A$. Substituting this into the second equation, this results in

$$A\phi + ((b-a) - A)\psi = a,$$

which can be rearranged to

$$A(\phi - \psi) = a - (b-a)\psi$$

and finally

$$A = \frac{a-(b-a)\psi}{\phi-\psi} = \frac{a-(b-a)\psi}{\sqrt{5}}$$

using that $\phi - \psi = \sqrt{5}$. For B, we obtain

$$B = b - a - \frac{a-(b-a)\psi}{\phi-\psi} = \frac{(b-a)\phi - a}{\phi-\psi} = \frac{(b-a)\phi - a}{\sqrt{5}}.$$

This implies equation (4). Finally, we observe that

$$\frac{a-(b-a)\psi}{\sqrt{5}}\phi = \frac{(b-a)+a\phi}{\sqrt{5}} = \frac{b+a(\phi-1)}{\sqrt{5}} = \frac{b-a\psi}{\sqrt{5}}$$

and also, that

$$\frac{(b-a)\phi - a}{\sqrt{5}}\psi = \frac{(a-b)-a\psi}{\sqrt{5}} = \frac{a(1-\psi)-b}{\sqrt{5}} = \frac{a\phi - b}{\sqrt{5}},$$

which precisely implies (5). □

A.2 Sparsity of Attention Matrix with Fibonacci Dilation Sequence

Now we are set to provide the head-wise computational overhead of the standard and modified variant of Fibottention.

Lemma 2. *Let N be the number of tokens in a multi-head self-attention block. If $(f_n)_n = Fib(a,b)$ is used as a dilation sequence for $a, b \in \mathbb{N}$ to create the attention support set $\Omega_w^{Fib(a,b)}$ as in eq. (3) for window size $w \leq N$, $a < b \leq w$, then the masked attention matrix $A^{\Omega_w^{Fib(a,b)}}$ can be computed by evaluating at most*

$$|\Omega_w^{Fib(a,b)}| \leq 2N\left(\frac{\log(\sqrt{5}w + |a\phi - b|) - \log(b - a\psi)}{\log\phi} + 1\right)$$

dot products between query and key vectors, where $\phi = (1+\sqrt{5})/2$ is the golden ratio.

Proof. Let $(f_1, \cdots, f_D)$ be the Fibonacci indices for one mask. We have $f_{k+1} = f_k + f_{k-1}$. Let $f_1 = a$ and $f_2 = b$. Let f_j be the index of the diagonal. Thus the total number of inner products to be computed for the diagonal f_j is given by $2(N - f_j)$ considering the symmetric distribution of diagonals at which the self-attention matrix is evaluated. For a total of D diagonals indicated by sequence elements $(f_1, \cdots, f_D)$ of $\mathrm{Fib}(a,b)$ such that $f_D \leq w$, but $f_{D+1} > w$, we obtain a sparsity pattern whose size is bounded by

$$|\Omega_w^{\mathrm{Fib}(a,b)}| = \sum_{j=1}^{D} 2(N - f_j).$$

Using the identity

$$\sum_{j=1}^{D} f_j + f_2 = f_{D+2}, \tag{7}$$

which holds true for generalized Fibonacci sequences $(f_n)_n = \mathrm{Fib}(a,b)$ for any a, b, we simplify this further such that

$$\begin{aligned} |\Omega_w^{\mathrm{Fib}(a,b)}| &= 2DN - 2\sum_{j=1}^{D} f_j \\ &= 2DN - 2(f_{D+2} - f_2) \\ &< 2DN - 2w + 2b \leq 2DN, \end{aligned} \tag{8}$$

where we used that $f_{D+2} \geq f_{D+1} > w$ in the first inequality and $b \leq w$ in the last inequality. Next, we find a bound on the index D of the largest sequence element f_D such that $f_D \leq w$. We solve for D such that $f_D \leq w$. Using equation (5) of Lemma 1, we observe that

$$\begin{aligned} w \geq f_D &= \frac{b - a\psi}{\sqrt{5}}\phi^{D-1} + \frac{a\phi - b}{\sqrt{5}}\psi^{D-1} \\ &\geq \frac{b - a\psi}{\sqrt{5}}\phi^{D-1} - \frac{|a\phi - b|}{\sqrt{5}}|\psi|^{D-1} \\ &\geq \frac{b - a\psi}{\sqrt{5}}\phi^{D-1} - \frac{|a\phi - b|}{\sqrt{5}} \end{aligned}$$

using that $|\psi| = \left|\frac{1-\sqrt{5}}{2}\right| \leq 1$ in the last inequality. Solving the latter inequality for D, we obtain the bound

$$D \leq \frac{\log(\sqrt{5}w + |a\phi - b|) - \log(b - a\psi)}{\log \phi} + 1,$$

using that $\phi > 1$, $\psi < 0$ and $1 \leq a \leq b$. Inserting this bound into (8), this results in the total bound

$$|\Omega_w^{\mathrm{Fib}(a,b)}| \leq 2N \left(\frac{\log(\sqrt{5}w + |a\phi - b|) - \log(b - a\psi)}{\log \phi} + 1 \right).$$

□

A.3 Time Complexity Bound for Fibottention

Lemma 2 can be used to quantify the sparsity of each head attention matrix used in the Fibottention modification of MHSA. Specifically, we recall from Section 3.2 that Fibottention uses h different sparsity patterns $\Omega_{w_1}^{\mathrm{Fib}(a_1^{\mathrm{Wyt}}, b_1^{\mathrm{Wyt}})}, \Omega_{w_2}^{\mathrm{Fib}(a_2^{\mathrm{Wyt}}, b_2^{\mathrm{Wyt}})}, \cdots, \Omega_{w_h}^{\mathrm{Fib}(a_h^{\mathrm{Wyt}}, b_h^{\mathrm{Wyt}})}$, where the initial two generalized Fibonacci sequence elements are given by $a_i^{\mathrm{Wyt}} = \lfloor \lfloor i\phi \rfloor \phi \rfloor$ and $b_i^{\mathrm{Wyt}} = \lfloor \lfloor i\phi \rfloor \phi^2 \rfloor$, where $i = 1, \ldots, h$ is a head index and $\phi = \frac{1+\sqrt{5}}{2}$ is the golden ratio (corresponding to the Wythoff array, see Table 1 for an illustration). Furthermore, the window size bounds $w_1, w_2, \cdots, w_h$ are chosen to interpolate between $w_{\min}$ and $w_{\max}$ based on the formula

$$w_i = w_{\min} + \left\lfloor \frac{w_{\max} - w_{\min}}{h - 1}(i - 1) \right\rfloor,$$

for all $i = 1, \ldots, h$.

In particular, we obtain the following result.

Theorem 3. *Assume that Fibottention is used in a Transformer block with N tokens of dimension d which contains h heads. Then the time complexity of computing all necessary query-key dot products in Fibottention can be bounded by*

$$\begin{aligned} \sum_{i=1}^{h} \frac{d}{h} |\Omega_{w_i}^{Fib(a_i^{Wyt}, b_i^{Wyt})}| &\leq 2Nd \left(2.08 \log((\sqrt{5}+1)w_{max}) - 1 \right) \\ &\leq 4.16 \cdot Nd \log(3.3 \cdot N), \end{aligned}$$

where $\log(\cdot)$ *is the natural logarithm.*

Proof. Fix a head index i and let $m_i = \lfloor i\phi \rfloor$. By definition, we have that $a_i^{\mathrm{Wyt}} = \lfloor m_i \phi \rfloor$ and $b_i^{\mathrm{Wyt}} = \lfloor m_i \phi^2 \rfloor$. Since the golden ratio ϕ satisfies $\phi^2 = \phi + 1$, we see that

$$b_i^{\mathrm{Wyt}} = \lfloor m_i \phi^2 \rfloor = \lfloor m_i \phi + m_i \rfloor = a_i^{\mathrm{Wyt}} + m_i.$$

Therefore,

$$a_i^{\mathrm{Wyt}} \phi - b_i^{\mathrm{Wyt}} = a_i^{\mathrm{Wyt}}(\phi - 1) - m_i = \frac{a_i^{\mathrm{Wyt}}}{\phi} - m_i.$$

From $a_i^{\mathrm{Wyt}} = \lfloor m_i \phi \rfloor$ we have $a_i^{\mathrm{Wyt}} \leq m_i \phi < a_i^{\mathrm{Wyt}} + 1$, hence

$$\frac{a_i^{\mathrm{Wyt}}}{\phi} \leq m_i < \frac{a_i^{\mathrm{Wyt}} + 1}{\phi} = \frac{a_i^{\mathrm{Wyt}}}{\phi} + \frac{1}{\phi},$$

which implies

$$-\frac{1}{\phi} < \frac{a_i^{\mathrm{Wyt}}}{\phi} - m_i \leq 0$$

and thus,

$$\left|a_i^{\mathrm{Wyt}}\phi - b_i^{\mathrm{Wyt}}\right| \leq \frac{1}{\phi} < 0.62$$

due to the value of the golden ratio ϕ. Moreover, since $\psi = (1-\sqrt{5})/2$ from Lemma 1 satisfies $\psi = -1/\phi < 0$ and $a_i^{\mathrm{Wyt}}, b_i^{\mathrm{Wyt}} \in \mathbb{N}$, we find the lower bound

$$\begin{aligned} b_i^{\mathrm{Wyt}} - a_i^{\mathrm{Wyt}}\psi &= b_i^{\mathrm{Wyt}} + \frac{a_i^{\mathrm{Wyt}}}{\phi} \geq \left(a_i^{\mathrm{Wyt}} + 1\right) + \frac{a_i^{\mathrm{Wyt}}}{\phi} \\ &= a_i^{\mathrm{Wyt}}\left(1 + \frac{1}{\phi}\right) + 1 = a_i^{\mathrm{Wyt}}\phi + 1 \geq \phi + 1 = \phi^2, \end{aligned}$$

using again the quadratic golden ratio equation and the fact that $a_i^{\mathrm{Wyt}} \geq 1$ for any i. Applying Lemma 2 to head i yields

$$\begin{aligned} &\left|\Omega_{w_i}^{\mathrm{Fib}(a_i^{\mathrm{Wyt}}, b_i^{\mathrm{Wyt}})}\right| \\ &\leq 2N\left(\frac{\log(\sqrt{5}\,w_i + 0.62) - \log(\phi^2)}{\log\phi} + 1\right) \\ &= 2N\left(\frac{\log(\sqrt{5}\,w_i + 0.62)}{\log\phi} - 1\right). \end{aligned}$$

Now, we observe that the dimension of the query and key vectors $(Q_i)_{j,:}, (K_i)_{j,:} \in \mathbb{R}^{d_h}$ is $d_h = d/h$ for each head, which is why we can bound the number of operations to compute Fibottention's sparse attention matrices $A_1^{\Omega_{w_1}^{\mathrm{Fib}(a_1^{\mathrm{Wyt}}, b_1^{\mathrm{Wyt}})}}, \ldots, A_h^{\Omega_{w_h}^{\mathrm{Fib}(a_h^{\mathrm{Wyt}}, b_h^{\mathrm{Wyt}})}}$ (see (1)) as

$$\frac{d}{h}\sum_{i=1}^{h}\left|\Omega_{w_i}^{\mathrm{Fib}(a_i^{\mathrm{Wyt}}, b_i^{\mathrm{Wyt}})}\right| \leq 2Nd\left(\frac{\log(\sqrt{5}\,w_{\max} + 0.62)}{\log\phi} - 1\right),$$

using that $w_i \leq w_{\max}$ for any $i = 1, \ldots, h$ and the monotonicity of $\log(\cdot)$. Since $1/\log\phi \approx 2.078086\ldots \leq 2.08$, the first bound of Theorem 3 follows.

Using then the fact that $1 + \sqrt{5} \leq 3.3$ and $w_{\max} \leq N$ yields the final bound of Theorem 3. □

To show this result, we used $b \leq w_i$ and $w_{\max} \leq N$. The statement of Theorem 3 implies that a forward pass of Fibottention (default Wythoff version) has a time complexity of $\mathcal{O}(N\log(N))$ with respect to the number of tokens. With only cosmetic adjustments, Theorem 3 and the $\mathcal{O}(N\log(N))$ attention complexity hold true for the modified Wythoff variant of Fibottention.

B Algorithmic Outline for Fibottention

In this section, we provide for completeness detailed pseudo code that facilitates the modification of a MHSA module by Fibottention. Algorithm 1 generates the generalized Fibonacci sequence elements of $\mathrm{Fib}(a, b)$ given initial elements a and b up to a window size bound w. Algorithm 2 computes the support sets $\Omega_1, \ldots, \Omega_h$ of Fibottention as described in Section 3.2 for each of the h attention heads, given L attention layers in a Transformer architecture (aggregated in tensor Ω), also incorporating dense interactions with the class token (in accordance with standard practice for ViTs). The flag *is_modified* indicates whether the Modified Wythoff variant of Fibottention is utilized or not.

Algorithm 1 Computation of Generalized Fibonacci Sequence Fib(a, b) Elements up to w.

Input: a, b, w **Output:** fib_seq
$fib\_seq \leftarrow [a, b]$
while $fib\_seq[-1] + fib\_seq[-2] \leq w$ **do**
  $next\_num \leftarrow fib\_seq[-1] + fib\_seq[-2]$
  append $next\_num$ to fib_seq
end while
return fib_seq

Algorithm 2 Generation of Sparsity Patterns of Fibottention.

Input: $L, N, h, w_{\min}, w_{\max}$, is_modified
Output: $\Omega \in \{0, 1\}^{h \times (N+1) \times (N+1)}$
$\phi \leftarrow \frac{1+\sqrt{5}}{2}$
$\Omega \leftarrow 0^{h \times (N+1) \times (N+1)}$
for each head $i \in \{1, \cdots, h\}$ **do**
  $w_i \leftarrow w_{\min}$
  if $h > 1$ **then** $+ \left\lfloor \frac{(i-1)(w_{\max} - w_{\min})}{h-1} \right\rfloor$
  end if
  $a \leftarrow \lfloor \lfloor i\phi \rfloor \phi \rfloor, \;\; b \leftarrow \lfloor \lfloor i\phi \rfloor \phi^2 \rfloor$ ▷ Wythoff pair for head i
  if is_modified **then**
    $b_{\text{Wyt-m}} \leftarrow b - a; \;\; a_{\text{Wyt-m}} \leftarrow a - b_{\text{Wyt-m}}$
    $I \leftarrow$ **getFibonacci**$(a_{\text{Wyt-m}}, b_{\text{Wyt-m}}, w_i)$
  else
    $I \leftarrow$ **getFibonacci**(a, b, w_i)
  end if
  $\Theta \leftarrow 0^{(N+1) \times (N+1)}$
  $\Theta_{0,:} \leftarrow 1; \;\; \Theta_{:,0} \leftarrow 1$ ▷ Keep class-token connections
  for each $o \in I$ **do**
    if $o \leq 0$ **then**
      continue
    end if
    for each $j \in \{0, \cdots, N - o\}$ **do**
      $(\Theta)_{j,\, j+o} \leftarrow 1; \;\; (\Theta)_{j+o,\, j} \leftarrow 1$ ▷ Allow $\pm o$ offsets
    end for
  end for
  $\Omega[i, :, :] \leftarrow \Theta$
end for
$\Omega \leftarrow$ **randomshuffle**(L, Ω) ▷ Assign masks across L layers
return Ω

Finally, Algorithm 3 demonstrates how Fibottention can be implemented in Multi-Head Self-Attention (MHSA) to compute the attention mechanism.

We note that the outlines of Algorithm 3 and Algorithm 2 focus on the correctness of the implementation without an emphasis on efficiency: In a practical sparse attention implementation, it will be sufficient to compute and store only the non-trivial (i.e., entries that are set to 1) entries of the attention mask tensor Ω, and evaluate the post-softmax attention weights A_i only corresponding entries, rather than storing h $N \times N$ matrices.

Algorithm 3 Fibottention in a Vision Transformer block.

Input: $X \in \mathbb{R}^{(N+1)\times d}$
Output: $O \in \mathbb{R}^{(N+1)\times d}$
Parameters: $\{W_i^Q, W_i^K, W_i^V\}_{i=1}^h$, W^Z, with $W_i^{Q,K,V} \in \mathbb{R}^{d\times d_h}$, $d_h = \frac{d}{h}$
Hyperparameters: $w_{\min}, w_{\max}$, *is_modified*
$\Omega \leftarrow$ **getMask**$(L, N, h, w_{\min}, w_{\max}$, *is_modified*$)$ ▷ Alg. 2
for $i = 1$ **to** h **do**
    $S_i \leftarrow (N \times N)$ matrix with $-\infty$ entries
    $Q_i \leftarrow XW_i^Q$, $K_i \leftarrow XW_i^K$, $V_i \leftarrow XW_i^V$
    $S_i[\Omega[i, \omega_1, \omega_2]] \leftarrow \frac{(Q_i)_{\omega_1}^\top (K_i)_{\omega_2}}{\sqrt{d_h}}$ for all $(\omega_1, \omega_2) \in [N]^2$ with $\Omega[i, \omega_1, \omega_2] = 1$.
    $A_i \leftarrow \text{softmax}(S_i)$
    $Z_i \leftarrow A_i V_i \in \mathbb{R}^{(N+1)\times d_h}$
end for
$Z \leftarrow \text{Concat}(Z_1, \ldots, Z_h) \in \mathbb{R}^{(N+1)\times(hd_h)}$
$O \leftarrow ZW^Z$, where $W^Z \in \mathbb{R}^{(hd_h)\times d}$
return O

C Further Implementation Details

In this section, we provide a comprehensive outline for the implementation of Fibottention within a multi-head self-attention block of a Transformer architecture.

C.1 Integration of Fibottention in Variants of ViTs

As we argued in Section 3, Fibottention can be considered as a drop-in replacement of MHSA that can be used within different vision Transformer architectures. In Section 4.2, we provided empirical results about the integration of Fibottention into various ViTs, for which we provide implementation details below: For the Swin-B (Liu et al., 2021) experiment of Table 3, we replace self-attention of Swin-B with Fibottention only in the first two stages of the model. The last two stages of the Swin-B, which are less computationally intensive due to prior patch merging modules, remain unmodified. We follow the standard training procedure of Swin-B (Liu et al., 2021). ConViT-B (d'Ascoli et al., 2021) consists of gated positional self-attention (GPSA) and MHSA blocks. We apply Fibottention only to replace the MHSA blocks and train the model following d'Ascoli et al. (2021). As iFormer (Zheng, 2025) uses single-head self-attention instead of MHSA, we replace its only attention head's attention matrix with $A_1^{\Omega_{w_{\max}}^{\text{Fib}(a_1^{\text{Wyt}}, b_1^{\text{Wyt}})}}$. To evaluate Fibottention within the UPop (Shi et al., 2023) framework, we first replace the standard MHSA blocks of the target vision-language backbone with Fibottention and subsequently apply UPop's unified and progressive pruning search to compress the modified architecture.

C.2 Experimental Configuration for Image Classification

The training settings for all image classification experiments performed in Section 4 are detailed in Table 9. We conducted all these experiments for 100 epochs using a batch size of 64 on 4 RTX A6000 GPUs.

Regarding the sparse attention baselines presented in Tables 2 and 4, BigBird (Zaheer et al., 2020) and Sparse Transformer (Child et al., 2019) denote adaptations of their respective sparse attention schemes applied to ViT-B. For BigBird, we utilize the specific hyperparameter configuration (local window size, global tokens, and random interactions) justified by our ablation study in Appendix D.2, which selects the variant that yields the best trade-off between accuracy and efficiency. For Sparse Transformer, we employ the *strided* attention variant (Child et al., 2019). Both baselines are configured to operate under pruning ratios comparable to Fibottention, ensuring that observed performance differences primarily arise from how the remaining token interactions are structured rather than disparities in computational cost.

Table 9: C10, C100, and Tiny-IN Training Settings (Touvron et al., 2021).

Input Size	224×224
Crop Ratio	0.9
Batch Size	64
Optimizer	AdamW
Optimizer Epsilon	1.0e-06
Momentum	0.9
Weight Decay	0.05
Gradient Clip	1.0
Learning Rate Schedule	Cosine
Learning Rate	1e-3
Warmup LR	1.0e-6
Min LR	1.0e-5
Epochs	100
Decay Epochs	1.0
Warmup Epochs	5
Decay Rate	0.988
Exponential Moving Average (EMA)	True
EMA Decay	0.99992
Random Resize & Crop Scale & Ratio	(0.08, 1.0), (0.67, 1.5)
Random Horizontal Flip Probability	0.5
Color Jittering	0.4
Auto-augmentation	rand-m15-n2-mstd1.0-inc1
Mixup	True
Cutmix	True
Mixup, Cutmix Probability	0.5, 0.5
Mixup Mode	Batch
Label Smoothing	0.1

D Ablations for Sparse Attention Mechanisms

In this section, we present ablations exploring the role of different hyperparameter choices within the sparse attention adaptations to ViTs on accuracy, sparsity and computational efficiency, which justify the experimental setups presented in Section 4.

Table 10: Top-1 accuracy on CIFAR-10 (C10) and CIFAR-100 (C100) for Random Attention with and without the class token, under different attention pruning ratios.

Attention Pruning Ratio (%) ↑	**w/ Class Token**		**w/o Class Token**	
	C10	C100	C10	C100
0%	**83.5**	**59.3**	**83.5**	**59.3**
20%	**83.3**	**59.1**	83.2	59.0
40%	**83.2**	58.7	82.8	**58.8**
60%	**82.7**	**58.9**	81.9	58.7
80%	**82.4**	**58.5**	81.1	58.1
90%	**81.6**	**58.1**	80.4	56.9
100%	**81.4**	**58.0**	77.5	47.9

D.1 Impact of Randomized Sparsity

In Table 10, we present the top-1 accuracy results of randomly masked (index pairs sampled uniformly at random) self-attention on C10 and C100 datasets under varying levels of sparsity. We observe that as the masking ratio increases, there is a consistent decline in accuracy, highlighting the trade-off between reducing computational cost and maintaining performance. This shows that the performance improvements obtained by the sparse attention mechanism of Fibottention reported in Table 2 are not at all observed for a random sparse sampling pattern, at any pruning ratio. This can be interpreted such that the pattern-less nature of random masking fails to identify critical token relationships underscoring the necessity of structured approaches, leading to significant information loss as sparsity increases.

D.2 Impact of Structured Sparsity

In this section, we present the role different hyperparameter choices within a BigBird (Zaheer et al., 2020) for image classification experiments. In particular, we present in Table 11 how choices of the local window size (w), global token interactions (g), and the number of random token interactions (r), impact ViT-B performance when trained on C10. We observe that $w = 2$, $g = 1$, and $r = N$ result in a model accuracy of 85.41%. However, further increasing randomness to $r = 2N$ reduced accuracy to 84.75%. This decline can be attributed to the dilution of critical token relationships in visual data, where spatially correlated information plays a vital role. Expanding the local window size to $w = 4$ resulted in a peak accuracy of 86.33%, demonstrating that larger windows capture richer token dependencies and compensate for reduced reliance on randomness. We note that a further increase of w would lower the masking ratio, increasing computational costs and diminishing efficiency. This trade-off highlights the limitations of BigBird compared to Fibottention.

Table 11: Performance comparison of masking strategies in BigBird with varying configurations. Top-1 accuracy is reported.

Configuration	**Mask Ratio**	**C10**
$w_i = 2 \mid g = 1 \mid r = N$	96.97	85.41
$w_i = 2 \mid g = 1 \mid r = 2N$	96.47	84.75
$w_i = 4 \mid g = 1 \mid r = N$	94.21	**86.33**

E Further Ablation Studies of Fibottention

In this section, we evaluate additional architectural choices in Fibottention through controlled experiments, supplementing the ablations presented in Section D.

Table 12: Top-1 accuracy of ViT-B on C10 and C100 for windowed self-attention with and without the principal diagonal. For each window size w_i we report the attention pruning ratio and accuracy when the diagonal entries are kept (left) or removed (right).

w	**w/ Main Diagonal**			**w/o Main Diagonal**		
	Pruning Ratio	C10	C100	Pruning Ratio	C10	C100
2	97.46	**86.1**	62.0	**97.97**	85.7	**62.2**
10	89.57	86.8	62.4	**90.08**	**87.0**	**63.4**
15	84.81	87.5	64.7	**85.32**	**88.0**	**64.8**
20	80.17	87.9	**64.9**	**80.69**	**88.0**	**64.9**
40	62.94	87.6	64.5	**63.45**	**87.7**	**65.0**

Table 13: Top-1 accuracy of ViT-B with windowed self-attention on CIFAR-10 (Krizhevsky, 2009) and CIFAR-100 (Krizhevsky, 2009) as a function of the shared window size $w_i = w_{\min} = w_{\max}$.

w_i	**Top-1 Accuracy (%)**		**Attention Pruning Ratio ↑**
	C10	C100	
2	85.7	62.2	97.97%
3	86.7	62.9	96.97%
4	86.8	62.9	95.97%
5	86.7	63.0	94.98%
6	86.5	62.9	93.99%
7	86.9	62.5	93.00%
8	86.3	63.1	92.02%
9	86.8	62.9	91.05%
10	86.9	63.4	90.08%
15	**88.0**	64.8	85.32%
20	**88.0**	64.9	80.69%
40	87.7	**65.0**	63.45%
80	87.0	62.9	35.24%
120	85.7	61.5	15.35%
160	83.7	60.2	3.79%
196	83.5	59.3	0%

E.1 Semantic Impact of Principal Diagonal in Self-Attention

In Table 12, we conduct an ablation study on the use of fixed local window sparsity pattern $\Omega_w = \{(j,k) \in \{1,\cdots,N\}^2 : |j-k| \leq w\}$ with window size w (identical across all heads) for attention computation, comparing configurations with (usage of Ω_w) and without the principal diagonal (usage of $\Omega_w \setminus \{(j,j) \in \{1,\cdots,N\}^2, j \in \{1,\cdots,N\}\}$). Table 12 suggests that the classification accuracy increases when removing the principal diagonal from Ω_w, consistently for different choices of w, confirming an observation of SparseBERT (Shi et al., 2021) in our setting. The pruning ratio is defined as the proportion of token interactions, $\left(\frac{N^2-|\Omega_w|}{N^2}\right) \cdot 100\%$ that are excluded during the attention computation. A larger pruning ratio corresponds to fewer token interactions being utilized, which has the potential to reduce computational costs.

E.2 Choice of Shared Window Size Limit Among Heads

Table 13 reports the top-1 accuracy of ViT-B on C10 and C100 as a function of the shared window size $w_i = w_{\min} = w_{\max}$. We observe that increasing the window size from very small neighborhoods up to a moderate range steadily improves performance, with the best results obtained for $w_i \in [15, 20]$. This indicates that incorporating local context within a limited spatial extent is beneficial for learning discriminative token representations. However, as the window size continues to grow, accuracy gradually declines alongside the attention pruning ratio. This suggests that excessively large windows dilute locality information and reduce the specialization of attention patterns. Overall, Table 13 highlights that moderate window sizes provide the best trade-off between local focus and contextual coverage.

E.3 Analysis of Corrupted Datasets

In Section 6.4, we reported results on the performance of Fibottention in the context of corrupted datasets. The top-1 accuracy numbers reported in Figure 4 contain corrupted data covering 19 corruption types, which include brightness, contrast, defocus blur, elastic transform, fog, frost, Gaussian blur, Gaussian noise, glass blur, impulse noise, JPEG compression, motion blur, pixelation, saturation, shot noise, snow, spatter, speckle

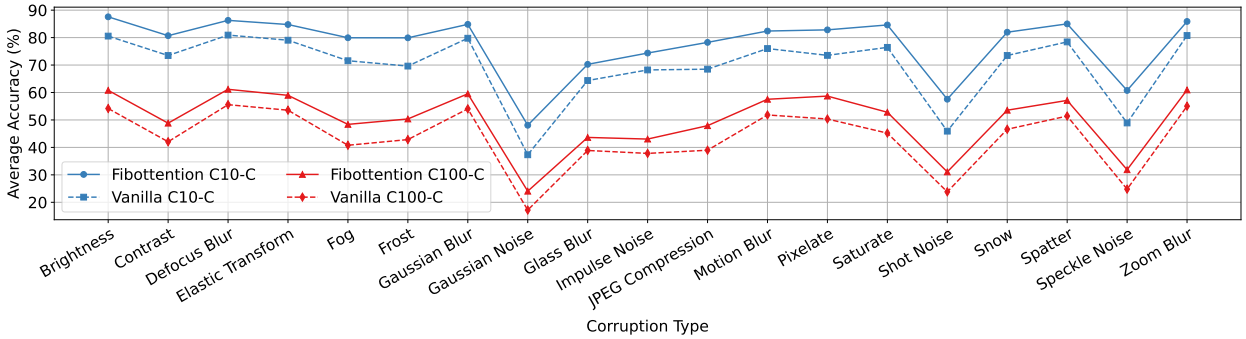


Figure 6: Performance of Fibottention compared to Vanilla ViT on C10 and C100 corrupted datasets.

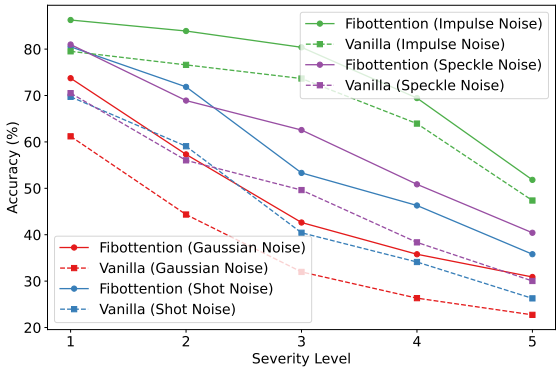


Figure 7: Performance of Fibottention ViT-B and Vanilla ViT-B on C10 corrupted dataset under four types of corruption at five levels of severity.

noise, and zoom blur. In Figure 6, we present a more detailed breakdown of the performance of Fibottention-equipped ViT-Base compared to ViT-Base using MHSA on C10 and C100 corrupted datasets (Hendrycks & Dietterich, 2019) across the different corruption types. The accuracies in Figure 6 are averaged across all five severity levels for each corruption type, providing a comprehensive view of robustness under a wide range of challenging conditions. We observe that Fibottention consistently outperforms ViT-Base on both datasets for every corruption type. Finally, we present in Figure 7 the C10 results for four representative corruptions across all five severity levels, illustrating that the improvements of Fibottention persist as corruption severity increases.