A Survey of Retentive Network

Anonymous ACL submission

Abstract

The Retentive Network (RetNet) has recently emerged as a formidable successor to the Transformer architecture. Although the self-attention mechanism excels at capturing global dependencies, its inherent quadratic complexity imposes significant memory constraints and inhibits scalability during long-sequence modeling. To overcome these challenges, RetNet introduces an innovative retention mechanism that integrates the inductive bias of recurrent neural networks with the parallelizable training advantages of attention-based models. This unified representation allows RetNet to achieve constant-time inference and linear-time training without sacrificing representational capacity. Despite the growing body of research demonstrating the efficacy of RetNet across diverse fields such as natural language processing, computer vision, and time-series analysis, a systematic synthesis of the current literature is currently unavailable. This paper presents the first comprehensive survey of Retentive Networks through a detailed examination of its architectural foundations, core innovations, and specialized variants. Furthermore, we provide a multidisciplinary analysis of its applications ranging from basic sequence tasks to complex cross-modal scenarios. Finally, we offer prospective insights and suggest strategic avenues for future inquiry to facilitate the continued evolution of RetNet in both academic research and large-scale industrial applications.

1 Introduction

The introduction of the Transformer architecture by (Vaswani et al., 2017) marked a paradigm shift in deep learning through its exclusive reliance on self-attention mechanisms. Owing to its strong capacity to model long-range dependencies and its highly parallelizable structure, the Transformer has become the dominant paradigm in natural language processing (NLP). Beyond NLP, Transformer-based models have been successfully extended to a broad range of domains, including computer vision (CV), speech processing, and scientific fields such as chemistry and bioinformatics, demonstrating their versatility in capturing complex, long-range dependencies across diverse modalities (Lin et al., 2022). Despite its strengths, the Transformer architecture faces notable limitations. During training, its quadratic time complexity makes modeling long sequences computationally costly. In the inference phase, linear memory complexity arises from storing KV cache for each token, resulting in significant memory overhead. Although various approaches have been explored to mitigate the complexity of the Transformer, achieving substantial reductions in computational overhead remains challenging (Choromanski et al., 2020; Katharopoulos et al., 2020; Wang et al., 2020).

To mitigate the computational limitations of standard Transformer architectures, a substantial body of research has been proposed. Gated linear recurrent neural networks (Qin et al., 2023; De et al.) incorporated gating mechanisms to reduce the quadratic time complexity typically associated with Transformer training. State Space Models compressed sequence data into fixed-size representations, effectively mitigating the scaling issues inherent in Transformers (Gu et al., 2021; Gu and Dao, 2023). Linear Transformers (Katharopoulos et al., 2020) further alleviated memory and computational overhead by employing linear attention mechanisms, allowing both time and memory complexity to scale linearly with sequence length. The Receptance Weighted Key Value (RWKV) leverages linear attention to reduce computational complexity and memory usage during inference (Peng et al., 2023; Li et al., 2024). Among these, RetNet (Sun et al., 2023) stands out as a compelling solution, it integrated a multi-scale retention mechanism which

*Corresponding author

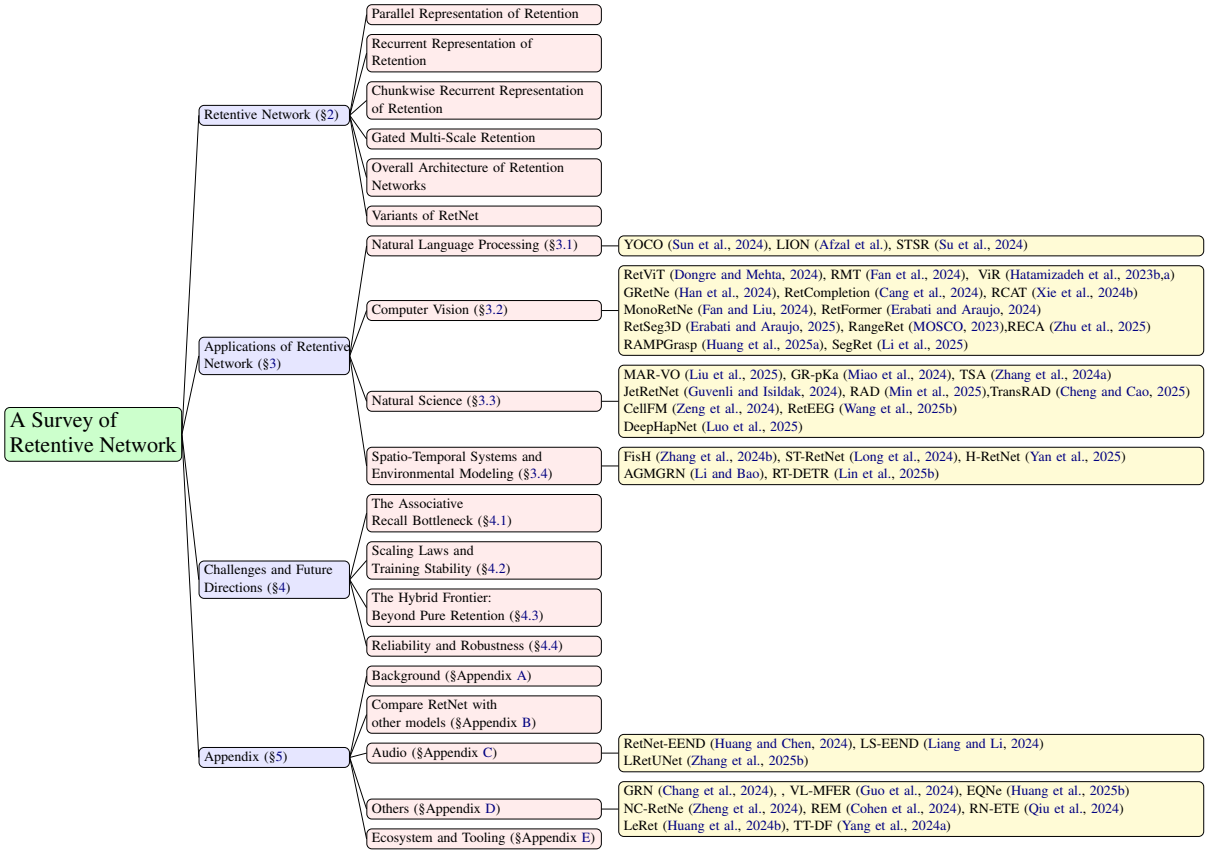


Figure 1: Structure of this paper.

employs three computational paradigms namely parallel, recurrent, and chunkwise recurrent representations. By leveraging these paradigms, RetNet achieves performance comparable to Transformers while enabling constant-time O(1) inference, reduced memory overhead, and efficient long-sequence modeling.

By employing the retention mechanism, the decay mask makes RetNet very versatile for a wide range of applications, from NLP (Cheng et al., 2024), CV (Fan et al., 2024), natural science (Luo et al., 2025) to social engineering (Yan et al., 2025). With the rapid expansion of research and applications of RetNet, this survey aims to shed light on current progress in this field.

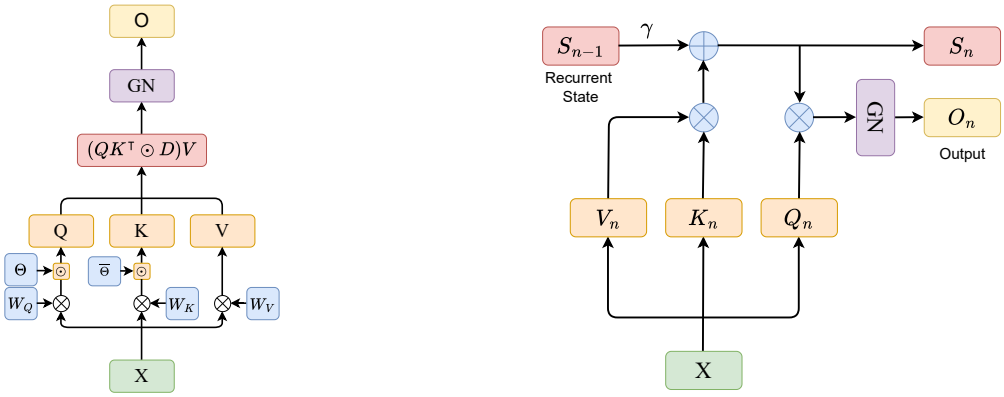
As depicted in Figure 1, the remainder of this paper is organized as follows: Section 2 delves into the principle and mechanism of RetNet, and Section 3 explores the extensive applications of RetNet in diverse domains, including NLP, CV, natural sciences, social engineering, and audio processing. Section 4 examines the primary challenges confronting RetNet and outlines prospective directions for future research. Finally, the Appendix provides foundational background and a comparison between RetNet and other models. These sections offer additional technical depth and context to the main discussion.

2 Retentive Network

Despite their effectiveness in capturing long-range dependencies, Transformer models incur high computational complexity and exhibit inefficiencies when processing long sequences (Lin et al., 2022). RetNet(Sun et al., 2023) theoretically derived the connection between recurrence and attention and proposed retention mechanism for sequence modeling. RetNet has been shown to achieve low-cost inference, efficient long-sequence modelling, Transformer-comparable performance, and parallel model training simultaneously.

2.1 Retentive Network

RetNet is constructed as a stack of L identical blocks, each comprising two core components: a Multi-Scale Retention (MSR) module and a Feed-Forward Network (FFN) module. For a given sequence of input $x = x_1 \cdots x_{|j|}$, where $|j|$ represents the length of the sequence, RetNet utilizes an autoregressive encoding method to process the sequence. The input is packed into $X^0 = [\mathbf{x}_1, \cdots, \mathbf{x}_{|j|}] \in \mathbb{R}^{|j| \times d_{\text{model}}}$, where d_{model} is the dimension of the hidden layer. The contextualized representations are then computed as follows:



(a) Parallel representation.

(b) Recurrent representation.

Figure 2: Dual form of RetNet. "GN" denotes GroupNorm.

$$X^l = \text{RetNet}_l(X^{l-1}), l \in [1, L]. \tag{1}$$

The retention mechanism, which admits both recurrent and parallel formulations, is central to the effectiveness of RetNet. Given an input sequence $X \in \mathbb{R}^{|j| \times d_{\text{model}}}$, each input vector X_n is projected into a value representation $v_n = X_n w_v$, where w_v is a learnable projection matrix. Then make the projection Q, K:

$$Q = XW^Q, \quad K = XW^K, \tag{2}$$

where $W^Q, W^K \in \mathbb{R}^{d \times d}$ are learnable matrices.

Consider a sequence modeling problem, through the state $\mathbf{s}_n \in \mathbb{R}^{d \times d}$ mapping v_n to a vector of o_n.

$$\begin{aligned} \mathbf{s}_n &= A\mathbf{s}_{n-1} + K_n^\top v_n \\ o_n &= Q_n \mathbf{s}_n = \sum_{m=1}^{n} Q_n A^{n-m} K_m^\top v_m, \end{aligned} \tag{3}$$

where K_n, Q_n is the projection of the time step n.

Further, diagonalize $A = \Lambda(\gamma e^{i\theta})\Lambda^{-1}$, where Λ is the reversible matrix, γ is the decay mask, according to Euler's formula $e^{i\theta} = [\cos\theta_1, \sin\theta_2, \cdots, \cos\theta_{d-1}, \sin\theta_d]$, then $A^{n-m} = \Lambda(\gamma e^{i\theta})^{n-m}\Lambda^{-1}$, n, m is the time step. Equation 3 becomes:

$$\begin{aligned} o_n &= \sum_{m=1}^{n} (Q_n(\gamma e^{i\theta})^n)(K_m(\gamma e^{i\theta})^{-m})^\top v_m \\ &= \sum_{m=1}^{n} \gamma^{n-m}(Q_n e^{in\theta})(K_m e^{im\theta})^\dagger v_m, \end{aligned} \tag{4}$$

where $Q_n(\gamma e^{i\theta})^n$, $K_m(\gamma e^{i\theta})^{-m}$ is the xPos (Sun et al., 2022), † is the conjugate transpose. $e^{in\theta}$ and $e^{im\theta}$ serve as rotational factors that encode positional information using complex exponential forms, where θ denotes the learnable parameters employed to model relative phase differences for the purpose of capturing sequential dependencies.

Parallel Representation of Retention As shown in Figure 2a, the retention layer is defined as:

$$\begin{aligned} &Q = (XW^Q) \odot \Theta, \quad K = (XW^K) \odot \overline{\Theta}, \\ &V = XW^V, \\ &D_{nm} = \begin{cases} \gamma^{n-m}, & n \geq m \\ 0, & n < m \end{cases}, \\ &\text{Retention}(X) = (QK^\top \odot D)V, \end{aligned} \tag{5}$$

where $\odot$ is the Hadamard product, Θ is the position-dependent modulation term, and $\overline{\Theta}$ denotes its complex conjugate, and $D \in \mathbb{R}^{|j| \times |j|}$ constitutes a unified matrix that jointly encodes causal masking and exponential decay as a function of relative positional distance.

Recurrent Representation of Retention As shown in Figure 2b, at the n-th timestep, the output is recurrently obtained as follows:

$$\begin{aligned} &S_n = \gamma S_{n-1} + K_n^\top V_n, \\ &\text{Retention}(X_n) = Q_n S_n, n = 1, \cdots, |j|. \end{aligned} \tag{6}$$

Chunkwise Recurrent Representation of Retention The input sequences are segmented into chunks. Within each chunk, the computation is carried out using the parallel representation Equation 5. In contrast, information across chunks is propagated using the recurrent representation Equation 6. Specifically, let B denote the chunk length.

The retention output of the i-th chunk is computed as follows:

$$
\begin{aligned}
Q_{[i]} &= Q_{Bi:B(i+1)}, \\
K_{[i]} &= K_{Bi:B(i+1)}, \\
V_{[i]} &= V_{Bi:B(i+1)}, \\
R_i &= K_{[i]}^{\top}(V_{[i]} \odot \zeta) + \gamma^{B} R_{i-1}, \\
\text{Retention}(X_{[i]}) &= \underbrace{(Q_{[i]} K_{[i]}^{\intercal} \odot D) V_{[i]}}_{\text{Inner-Chunk}} \\
&\quad + \underbrace{(Q_{[i]} R_{i-1}) \odot \xi}_{\text{Cross-Chunk}}, \\
\xi_{ij} &= \gamma^{i+1}, \quad \zeta_{ij} = \gamma^{B-i-1},
\end{aligned}
\tag{7}
$$

where $[i]$ indicates the i-th chunk, i.e., $x_{[i]} = [x_{(i-1)B+1}, \cdots, x_{iB}]$. ζ and ξ are exponential decay factors that modulate the influence of intra-chunk and inter-chunk information.

Gated Multi-Scale Retention In each layer, the number of retention heads is defined as $h = d_{\text{model}}/d$, where d denotes the head dimension. Each head is associated with distinct parameter matrices $W^Q, W^K, W^V \in \mathbb{R}^{d \times d}$. MSR mechanism assigns a unique decay factor γ to each head. For simplicity, identical γ values are used across different layers and kept fixed. To enhance the non-linearity of the retention layers, a swish gate (Hendrycks and Gimpel, 2016; Ramachandran et al., 2017) is introduced. Given the input X, the computation of the layer is defined as follows:

$$
\begin{aligned}
\gamma &= 1 - 2^{-5-\text{arange}(0,h)} \in \mathbb{R}^h, \\
\text{head}_i &= \text{Retention}(X, \gamma_i), \\
Y &= \text{GN}_h(\text{Concat}(\text{head}_1, \cdots, \text{head}_h)), \\
\text{MSR}(X) &= (\text{swish}(XW^G) \odot Y) W^O,
\end{aligned}
\tag{8}
$$

where $W^G, W^O \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ are learnable parameter matrices. $\text{arange}(0, h)$ denotes a vector of integers from 0 to $h-1$, used to assign distinct decay scales across h attention heads. GN denotes Group Normalization (Wu and He, 2018), applied to each head output following the SubLN strategy in (Shoeybi et al., 2019). Since each head employs a distinct γ scale, their output variances differ, which necessitates separate normalization.

Overall Architecture of Retention Networks As illustrated in Figure 3, an L-layer retention network is constructed by stacking MSR and FFN modules. The input sequence $\{x_i\}_{i=1}^{|j|}$ is first mapped to vector representations via a word embedding layer. The resulting embeddings, denoted as $X_0 = [x_1, \cdots, x_{|j|}] \in \mathbb{R}^{|j| \times d_{\text{model}}}$, serve as the initial input to the model. The final output is represented as X^L.

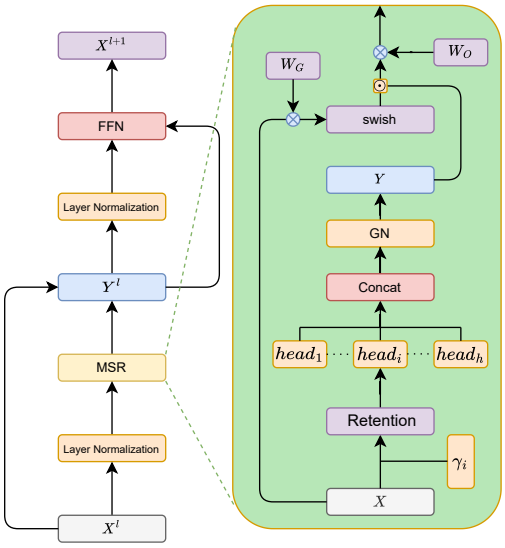


Figure 3: Overall architecture of RetNet.

$$
\begin{aligned}
Y^l &= \text{MSR}(\text{LN}(X^l)) + X^l, \\
X^{l+1} &= \text{FFN}(\text{LN}(Y^l)) + Y^l,
\end{aligned}
\tag{9}
$$

where $\text{LN}(\cdot)$ denotes the Layer Normalization function (Ba et al., 2016). The feed-forward network (FFN) is defined as

$$
\text{FFN}(X) = \text{gelu}(XW_1)W_2,
$$

where W_1 and W_2 are learnable parameter matrices, and $\text{gelu}(\cdot)$ is the Gaussian Error Linear Unit activation function.

2.2 Variants of RetNet

While RetNet provides a strong baseline for sequence modeling, its standard one-dimensional exponential decay formulation restricts its applicability to non-causal data (e.g., images) and very deep architectures. Accordingly, recent work has proposed several specialized variants that address these limitations along three complementary directions: spatially adaptive decay mechanisms, enhanced signal propagation, and improved bidirectional inference.

Spatial and Structural Adaptation. A fundamental challenge in extending RetNet to computer vision lies in bridging the mismatch between one-dimensional causal retention and two-dimensional non-causal spatial structures. **RetViT** (Dongre and Mehta, 2024) addresses this issue through a straightforward adaptation that replaces the attention blocks in ViT with one-dimensional retention, treating images as flattened sequences to achieve linear computational complexity. However, such a formulation does not explicitly encode two-dimensional spatial locality. To this end, **RMT** (Fan et al., 2024) introduces a *Manhattan Self-Attention (MaSA)* mechanism, which decomposes the decay mask into two orthogonal one-dimensional components corresponding to height and width. This design enables the retention mechanism to decay according to Manhattan distance rather than sequence index. Building on this spatial prior, **GRetNet** (Han et al., 2024) extends the approach to hyperspectral imaging by combining Manhattan-based decay with a Gaussian Multi-head Attention (GMA) mechanism, thereby effectively modeling local spectral–spatial correlations. In the three-dimensional setting, **RetFormer** (Erabati and Araujo, 2024) further generalizes the retention module to point cloud processing by incorporating decay functions based on 3D spatial coordinates.

Signal Propagation and Training Stability. As RetNet is scaled to deeper architectures, the progressive attenuation of historical information may lead to signal degradation. **DenseRetNet** (He et al., 2024) addresses this issue by incorporating the principles of Dense State Space Models (DenseSSM). Specifically, it introduces dense inter-layer connections that project hidden states from shallower layers to deeper ones, thereby improving gradient propagation and feature reuse while preserving the advantage of fully parallel training.

Bidirectional and Generative Optimization. Standard RetNet operates in a strictly causal (unidirectional) manner, which limits its applicability to tasks that require global context or bidirectional reasoning, such as image understanding and generation. To address this limitation, **LION-D** (Afzal et al., 2025) refines the RetNet architecture by reformulating the decay mechanism into a linear-attention–like structure that supports bidirectional processing while preserving RNN-style inference efficiency. For generative applications, **RetCompletion** (Cang et al., 2024) introduces a *Bi-RetNet* architecture that employs a dual-pathway design to process contextual information in both forward and backward directions. Combined with a pixel-wise inference strategy, this approach enables high-fidelity image completion and achieves substantially higher inference speed than autoregressive Transformer-based models.

A comparative summary of these variants is presented in Appendix Table 3.

3 Applications of Retentive Network

RetNet has emerged as a general-purpose sequence modeling framework whose dual formulation enables both scalable parallel training and efficient recurrent inference. Owing to these properties, it has been widely adopted across diverse application domains.

3.1 Natural Language Processing

RetNet has demonstrated significant promise in advancing NLP, particularly by addressing the scalability and memory efficiency challenges inherent in LLMs. Its innovative retention mechanism enables more efficient handling of long sequences and enhances the performance of complex reasoning tasks.

Efficient Language Architectures. One of the key challenges in NLP is the substantial memory overhead required by traditional Transformer-based models, especially with KV caches in LLMs. RetNet's retention mechanism significantly mitigates this issue by providing a more memory-efficient approach to sequence modeling. For example, YOCO (Sun et al., 2024) introduces a decoder-decoder architecture that reduces KV cache requirements while maintaining robust contextual understanding. Similarly, frameworks such as LION (Afzal et al.) and LION-D (Afzal et al., 2025) adopt fixed decay masks within RetNet, facilitating linear-time inference and maintaining the necessary parallelism for large-scale training, thus striking a balance between computational cost and performance.

Reasoning and Knowledge Representation. RetNet is well suited for complex reasoning tasks in NLP due to its ability to model structured dependencies in sequential data. This property has been effectively exploited in applications such as knowledge graph reasoning and multi-hop inference. Cheng et al. (2024) utilize RetNet as an

encoder for knowledge graph reasoning, leveraging its capacity to capture structural dependencies. Additionally, in sequence-to-sequence tasks, the STSR model (Su et al., 2024) applies RetNet's parallel retention module to speed up multi-hop reasoning. DenseRetNet (He et al., 2024) further improves feature extraction by integrating dense hidden connections, enabling better information propagation in deep reasoning processes.

3.2 Computer Vision

Extending RetNet to computer vision introduces a fundamental challenge, namely reconciling the one-dimensional causal decay inherent to the retention mechanism with the two-dimensional and generally non-causal nature of visual data.

Visual Backbones and Spatial Decay. A central line of research focuses on modifying the decay mask to encode spatial relationships. RetViT (Dongre and Mehta, 2024) demonstrates that substituting self-attention with retention blocks within Vision Transformers can improve training efficiency while preserving representational capacity. To explicitly model two-dimensional spatial structure, RMT (Fan et al., 2024) introduces Manhattan Self-Attention, which decomposes spatial interactions into orthogonal axes and applies bidirectional decay based on Manhattan distance. This spatially aware decay formulation is further extended in GRetNet (Han et al., 2024), where Gaussian-decayed retention is employed to capture local spectral and spatial correlations in hyperspectral imagery. A related strategy is adopted by RangeRet (MOSCO, 2023), which applies distance-based decay to efficiently model spatial context in LiDAR semantic segmentation.

Dense Prediction and Generation. The multi-scale nature of retention makes RetNet particularly effective for dense prediction tasks that require hierarchical feature aggregation. SegRet (Li et al., 2025) and RetSeg3D (Erabati and Araujo, 2025) incorporate multi-scale retention into two-dimensional and three-dimensional semantic segmentation frameworks, respectively, enhancing contextual modeling while maintaining computational efficiency. Extending this to industrial inspection, Multilevel RetNet (Zou et al., 2025) combines multi-scale retention with edge-enhanced attention for fine-grained infrastructure crack detection. In generative settings, RetCompletion (Cang et al., 2024) leverages a parallelized retentive decoder to enable real-time image inpainting. Similarly, regarding image restoration, RetUIE (Guan et al., 2025) introduces a retention-based channel self-attention module for underwater image enhancement to mitigate color degradation. Retention mechanisms have also been adopted in multi-modal and feature fusion scenarios. The SwiFTeR architecture (Hu et al., 2024a) and the Cross-Axis Transformer (Erickson, 2023) employ retention to aggregate visual information across spatially partitioned regions or heterogeneous inputs, supporting efficient cross-modal interaction (Wang et al., 2025c; Huang et al., 2024a).

Video and 3D Temporal Modeling. For vision tasks involving temporal dynamics, such as video understanding and point cloud processing, RetNet's recurrent formulation provides notable advantages in inference efficiency. RCAT (Xie et al., 2024b) and Maskable RetNet (Hu et al., 2024b) exploit retention to model long-range temporal dependencies for video recognition and moment retrieval. In three-dimensional perception, LION (Liu et al., 2024b) achieves linear computational complexity for point cloud sequences through groupwise retention. MonoRetNet (Fan and Liu, 2024) further explores temporal modeling by introducing a half-duplex bidirectional retention scheme for monocular depth estimation, highlighting RetNet's potential for sequential visual reasoning.

3.3 Natural Science

Scientific discovery frequently involves analyzing signals characterized by extreme sequence lengths, strong causal constraints, or complex multi-scale dynamics. RetNet is particularly well suited to such settings due to its scalable sequence modeling capability and its ability to preserve long-term dependencies with controlled computational complexity.

Physics and Signal Processing. In high-energy physics, JetRetNet (Guvenli and Isildak, 2024) models multi-scale dependencies for particle tracking. For signal classification, RetNet-based architectures have been adapted to radio-frequency modulation recognition (Han et al., 2025) and radar perception (Cheng and Cao, 2025), where both temporal continuity and spatial proximity are essential. Beyond classification, RetNet has also been employed for anomaly detection in cyber-physical systems (Min et al., 2025) and for transient stability assessment in power systems (Zhang et al.,

2024a), leveraging its capacity to capture long-term temporal irregularities.

Bio-Sequence Analysis. Biological sequences such as genomic, proteomic, and transcriptomic data pose fundamental long-context modeling challenges. RetNet has demonstrated effectiveness in this domain by enabling efficient modeling of long-range dependencies. Representative applications include haplotype assembly (Luo et al., 2025), spike protein feature analysis (Liu et al., 2024c), and large-scale transcriptomic modeling (Zeng et al., 2024). In the context of drug discovery, Peng et al. (2024) employ dual RetNet encoders to separately extract representations from drug molecules and protein targets, facilitating accurate modeling of their interactions.

Biomedical Signal and Image Processing. In electroencephalogram analysis, its recurrent formulation enables effective modeling of temporal dependencies for both decoding (Wang et al., 2025b) and denoising tasks (Wang et al., 2024a). For medical imaging, recent work has focused on improving efficiency in segmentation and classification. Models such as PFPRNet (Chu et al., 2024) and ResGDANet (Li and Huang, 2025) integrate spatial retention mechanisms to enhance feature representation, achieving favorable trade-offs between computational efficiency and diagnostic performance (ELKarazle et al., 2023; Lin et al., 2025a; Zhou et al., 2024). In particular, Ret-UNet (Guo et al., 2026) further advances segmentation by incorporating a self-retention mechanism to explicitly model global anatomical relationships, addressing the limited receptive field of traditional CNNs.

3.4 Spatio-Temporal Systems and Environmental Modeling

Spatio-temporal environmental systems involve large-scale data with strong locality and long-range temporal dependencies. RetNet effectively models these scenarios by combining scalable global context aggregation with efficient temporal dynamics.

Remote Sensing and Monitoring. In remote sensing and monitoring, RetNet facilitates the analysis of high-resolution spatial data with global contextual awareness. It has been applied to building change detection (Lin and Piao, 2024), domain-agnostic fire detection (Kim et al., 2024), and rotating object detection in aerial imagery (Liu et al., 2024a). Beyond satellite imagery, RetNet has also been employed in structural health monitoring, where it improves anomaly detection for bridge inspection (Wang et al., 2025a) and coal gangue identification (Zhang et al., 2025c). Expanding to aviation safety, recent works leverage retention mechanisms for real-time flight phase classification (Tomiło et al., 2025) and aircraft landing distance measurement (Tomiło, 2025), ensuring high precision in critical operational monitoring. These applications highlight RetNet's ability to balance modeling capacity and inference efficiency.

Spatio-Temporal Forecasting. Modeling urban dynamics and social systems requires jointly capturing spatial correlations and temporal evolution. RetNet-based architectures have demonstrated strong performance in traffic flow forecasting by effectively decoding time-dependent and spatially correlated features (Li and Bao; Zhu et al., 2024; Long et al., 2024; Yan et al., 2025). Its applicability further extends to environment and safety critical scenarios, including photovoltaic power generation forecasting under hazy conditions (Yang et al., 2024b) and real-time earthquake early warning through seismic wave representation learning (Zhang et al., 2024b).

For additional applications of RetNet beyond the domains discussed above, readers are referred to Appendix C and Appendix D.

4 Challenges and Future Directions

Despite its promising advances in balancing speed and performance, several critical challenges and uncharted territories remain that limit its broader applicability and further improvement. Below, we outline four core directions for future research that target these key limitations, aiming to refine, extend, and enhance the RetNet framework.

4.1 The Associative Recall Bottleneck

RetNet summarizes historical context through exponential decay into a fixed-size recurrent state, which inherently introduces information loss over long sequences. This phenomenon, commonly referred to as the associative recall bottleneck, differentiates retentive models from attention-based architectures that allow explicit retrieval of distant tokens.

Information Preservation. Existing retention mechanisms employ fixed or weakly adaptive decay rates, which may excessively attenuate salient

or low-redundancy information. Further research is needed to explore adaptive decay formulations or selective state update strategies that condition information retention on token importance, task requirements, or uncertainty estimates. Related ideas from structured state-space models, such as selective scanning (Gu and Dao, 2023), provide useful reference points.

In-Context Learning. The effectiveness of RetNet in few-shot and in-context learning settings has not been systematically evaluated. Understanding how state compression, decay schedules, and state dimensionality affect in-context generalization remains an open problem and is critical for assessing the suitability of RetNet in long-context reasoning tasks.

4.2 Scaling Laws and Training Stability

While RetNet demonstrates favorable efficiency properties at moderate scales, its scaling behavior with increasing model depth and parameter count remains largely unexplored.

Signal Propagation. The recurrent accumulation of hidden states may lead to signal attenuation or instability as depth increases. Prior observations in deep retentive architectures suggest the need for retention-aware normalization schemes, principled decay parameterization, and structured residual designs to ensure stable optimization (He et al., 2024). However, these design choices have not yet been systematically studied.

Generation Reliability. When relevant context has decayed, models may rely more heavily on priors, potentially increasing hallucination rates. Quantifying this effect across different decay regimes and task settings remains an important open research direction.

4.3 The Hybrid Frontier: Beyond Pure Retention

Accumulating evidence suggests that purely retentive architectures may not be optimal across all tasks and modalities, motivating the exploration of hybrid designs.

Architectural Hybridization and Integration. Purely retentive architectures are often suboptimal, driving the development of hybrids. Combining retention with sparse or retrieval-based attention mitigates recall limitations while maintaining efficiency. Furthermore, models like KARMA (Tomiło et al., 2025) integrate retention with Kolmogorov-Arnold Networks (KANs) to replace standard MLPs. This positions RetNet as a versatile backbone for compact, hardware-efficient designs across diverse paradigms.

Modality-Specific Design. Different modalities exhibit distinct structural and temporal characteristics. Applying retention mechanisms to sequential modalities, such as audio or video, while retaining attention-based modeling for spatial modalities offers a principled approach to multimodal representation learning (Zhu et al., 2025; Wang et al., 2025c).

4.4 Reliability and Robustness

The recurrent state in RetNet serves as a compressed summary of historical context and plays a critical role during inference. However, the robustness of this state under distribution shifts, noisy inputs, or adversarial perturbations has not been systematically studied. Future work should investigate how errors accumulate or propagate through the retentive state over long horizons, as well as how such effects impact downstream prediction reliability.

5 Conclusion

In summary, this paper provides a comprehensive synthesis of the existing literature on RetNet, elucidating its architectural foundations, practical applications, and the critical challenges that remain. To the best of our knowledge, this work constitutes the first dedicated review of RetNet, offering a structured perspective on its rapidly evolving ecosystem. By identifying key research gaps and future directions, we aim to provide a foundational roadmap that fosters continued innovation and exploration within this promising paradigm.

Limitations

This survey comprehensively reviews Retentive Networks and their variants across various domains. However, certain limitations remain. First, despite efforts to include all relevant literature prior to submission, some recent or niche studies might be omitted. Second, our discussion focuses on the architectural and algorithmic characteristics of RetNet models rather than application-specific implementation details. Readers are therefore encouraged to consult the original papers for detailed experimental results and practical insights.

References

Arshia Afzal, Elias Abad Rocamora, Leyla Naz Candogan, Pol Puigdemont, Francesco Tonin, Yongtao Wu, Mahsa Shoaran, and Volkan Cevher. Lion: A bidirectional framework that trains like a transformer and infers like an rnn.

Arshia Afzal, Elias Abad Rocamora, Leyla Naz Candogan, Pol Puigdemont, Francesco Tonin, Yongtao Wu, Mahsa Shoaran, and Volkan Cevher. 2025. Linear attention for efficient bidirectional sequence modeling. *arXiv preprint arXiv:2502.16249*.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.

Yueyang Cang, Pingge Hu, Xiaoteng Zhang, Xingtong Wang, Yuhang Liu, and Li Shi. 2024. Retcompletion: High-speed inference image completion with retentive network. *arXiv preprint arXiv:2410.04056*.

Qian Chang, Xia Li, and Xiufeng Cheng. 2024. Graph retention networks for dynamic graphs. *arXiv preprint arXiv:2411.11259*.

Jun Cheng, Tao Meng, Xiao Ao, and Xiaohua Wu. 2024. Pre-training retnet of simulating entities and relations as sentences for knowledge graph reasoning. In *2024 4th Asia Conference on Information Engineering (ACIE)*, pages 6–10. IEEE.

Lei Cheng and Siyang Cao. 2025. Transrad: Retentive vision transformer for enhanced radar object detection. *IEEE Transactions on Radar Systems*.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, and 1 others. 2020. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*.

Jinghui Chu, Wangtao Liu, Qi Tian, and Wei Lu. 2024. Pfprnet: A phase-wise feature pyramid with retention network for polyp segmentation. *IEEE Journal of Biomedical and Health Informatics*.

Lior Cohen, Kaixin Wang, Bingyi Kang, and Shie Mannor. 2024. Improving token-based world models with parallel observation prediction. *arXiv preprint arXiv:2402.05643*.

Soham De, Samuel L Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, Yutian Chen, Srivatsan Srinivasan, and 1 others. Griffin: Mixing gated linear recurrences with local attention for efficient language models, 2024. *URL https://arxiv.org/abs/2402.19427*, page 50.

JE Domínguez-Vidal and Alberto Sanfeliu. 2024. Force and velocity prediction in human-robot collaborative transportation tasks through video retentive networks. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9307–9313. IEEE.

Shreyas Dongre and Shrushti Mehta. 2024. Retvit: Retentive vision transformers. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–8. IEEE.

Khaled ELKarazle, Valliappan Raman, Caslon Chua, and Patrick Then. 2023. Retseg: Retention-based colorectal polyps segmentation network. *arXiv preprint arXiv:2310.05446*.

Gopi Krishna Erabati and Helder Araujo. 2024. Retformer: Embracing point cloud transformer with retentive network. *IEEE Transactions on Intelligent Vehicles*.

Gopi Krishna Erabati and Helder Araujo. 2025. Retseg3d: Retention-based 3d semantic segmentation for autonomous driving. *Computer Vision and Image Understanding*, 250:104231.

Lily Erickson. 2023. Cross-axis transformer with 3d rotary positional embeddings. *Preprint*, arXiv:2311.07184.

Dengxin Fan and Songyan Liu. 2024. Monoretnet: A self-supervised model for monocular depth estimation with bidirectional half-duplex retention. In *International Conference on Intelligent Computing*, pages 361–372. Springer.

Qihang Fan, Huaibo Huang, Mingrui Chen, Hongmin Liu, and Ran He. 2024. Rmt: Retentive networks meet vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5641–5651.

Meng Feiyu. 2024. Cfpsg: Collaborative filtering poi similarity graph enhanced retentive network for next poi recommendation. In *2024 21st International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 1–4. IEEE.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Albert Gu, Karan Goel, and Christopher Ré. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.

Meisheng Guan, Haiyong Xu, Yeyao Chen, Ting Luo, Yang Song, and Huaping Wang. 2025. Retuie: Retention-based underwater image enhancement. In *2025 International Symposium on Machine Learning and Media Computing (MLMC)*, pages 1–10. IEEE.

Chen Guo, Xinran Li, Jiaman Ma, Yimeng Li, Yuefan Liu, Haiying Qi, Li Zhang, and Yuhan Jin. 2024. Vl-mfer: A vision-language multimodal pretrained model with multiway-fuzzy-experts bidirectional retention network. *IEEE Transactions on Fuzzy Systems*.

Tianjun Guo, Weixin Zhao, and Jian Peng. 2026. Retunet: Enhancing medical image segmentation with self-retention. *Array*, 29:100653.

Ayse Asu Guvenli and Bora Isildak. 2024. B-jet tagging with retentive networks: A novel approach and comparative study. *arXiv preprint arXiv:2412.08134*.

Jia Han, Zhiyong Yu, and Jian Yang. 2025. Radio frequency-retentive network for automatic modulation classification. *Electronics Letters*, 61(1):e70203.

Zhu Han, Shuyi Xu, Lianru Gao, Zhi Li, and Bing Zhang. 2024. Gretnet: Gaussian retentive network for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*.

Ali Hatamizadeh, Michael Ranzinger, and Jan Kautz. 2023a. Vir: Vision retention networks. *arXiv. org*.

Ali Hatamizadeh, Michael Ranzinger, Shiyi Lan, Jose M Alvarez, Sanja Fidler, and Jan Kautz. 2023b. Vir: Towards efficient vision retention backbones. *arXiv preprint arXiv:2310.19731*.

Wei He, Kai Han, Yehui Tang, Chengcheng Wang, Yujie Yang, Tianyu Guo, and Yunhe Wang. 2024. Densemamba: State space models with dense hidden connection for efficient large language models. *arXiv preprint arXiv:2403.00818*.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jia Cheng Hu, Roberto Cavicchioli, and Alessandro Capotondi. 2024a. Shifted window fourier transform and retention for image captioning. *arXiv preprint arXiv:2408.13963*.

Jingjing Hu, Dan Guo, Kun Li, Zhan Si, Xun Yang, and Meng Wang. 2024b. Maskable retentive network for video moment retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1476–1485.

Jianan Huang, Xuebing Liu, Qing Zhu, Yaonan Wang, Mingtao Feng, Jiaming Zhou, Zhen Zhou, Lin Chen, and Danwei Wang. 2025a. Rampgrasp: Retentive attention-based multiscale perception grasp detection network. *IEEE Transactions on Circuits and Systems for Video Technology*.

Jingjia Huang, Jingyan Tu, Ge Meng, Yingying Wang, Yuhang Dong, Xiaotong Tu, Xinghao Ding, and Yue Huang. 2024a. Efficient perceiving local details via adaptive spatial-frequency information integration for multi-focus image fusion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9350–9359.

Kai-Wei Huang and Chia-Ping Chen. 2024. Long audio file speaker diarization with feasible end-to-end models. In *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1–6. IEEE.

Qihe Huang, Zhengyang Zhou, Kuo Yang, Gengyu Lin, Zhongchao Yi, and Yang Wang. 2024b. Leret: Language-empowered retentive network for time series forecasting. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*.

Yuyao Huang, Kai Chen, Wei Tian, and Lu Xiong. 2025b. Boost query-centric network efficiency for multi-agent motion forecasting. *IEEE Robotics and Automation Letters*.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. 2020. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR.

Sangwon Kim, In-su Jang, and Byoung Chul Ko. 2024. Domain-free fire detection using the spatial–temporal attention transform of the yolo backbone. *Pattern Analysis and Applications*, 27(2):45.

Bin Lei, Caiwen Ding, and 1 others. 2023. Flashvideo: A framework for swift inference in text-to-video generation. *arXiv preprint arXiv:2401.00869*.

Sihan Li and Juhua Huang. 2025. Resgdanet: An efficient residual group attention neural network for medical image classification. *Applied Sciences*, 15(5):2693.

Xing Li and Yuequan Bao. Adaptive gated meta graph retention network: A model for urban traffic flow prediction. *Available at SSRN 5170149*.

Zhiyuan Li, Yi Chang, and Yuan Wu. 2025. Segret: An efficient design for semantic segmentation with retentive network. *arXiv preprint arXiv:2502.14014*.

Zhiyuan Li, Tingyu Xia, Yi Chang, and Yuan Wu. 2024. A survey of rwkv. *arXiv preprint arXiv:2412.14847*.

Di Liang and Xiaofei Li. 2024. Ls-eend: Long-form streaming end-to-end neural diarization with online attractor extraction. *arXiv preprint arXiv:2410.06670*.

Ruixing Lin and Shunmei Piao. 2024. Change detection of building remote sensing images based on rmt-bit. In *2024 4th International Conference on Electronic Information Engineering and Computer Science (EIECS)*, pages 428–432. IEEE.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2022. A survey of transformers. *AI open*, 3:111–132.

Zhicheng Lin, Rongpu Cui, Limiao Ning, and Jian Peng. 2025a. Temporal features-fused vision retentive network for echocardiography image segmentation. *Sensors*, 25(6):1909.

Zhijie Lin, Zilong Zhu, Lingling Guo, Jingjing Chen, and Jiyi Wu. 2025b. Disease detection algorithm for tea health protection based on improved real-time detection transformer. *Applied Sciences (2076-3417)*, 15(4).

Zhou Linhao, Zhong Shenghua, and Xiao Zhijiao. 2024. Discovering multi-relational integration for knowledge tracing with retentive networks. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 960–968.

Jiayuan Liu, Bo Zhou, Xue Wan, Yan Pan, Zicong Li, and Yuanbin Shao. 2025. Mar-vo: A match-and-refine framework for uav's monocular visual odometry in planetary environments. *IEEE Transactions on Geoscience and Remote Sensing*.

Jing Liu, Donglin Jing, Yanyan Cao, Ying Wang, Chaoping Guo, Peijun Shi, and Haijing Zhang. 2024a. Lightweight progressive fusion calibration network for rotated object detection in remote sensing images. *Electronics*, 13(16):3172.

Zhe Liu, Jinghua Hou, Xinyu Wang, Xiaoqing Ye, Jingdong Wang, Hengshuang Zhao, and Xiang Bai. 2024b. Lion: Linear group rnn for 3d object detection in point clouds. *Advances in Neural Information Processing Systems*, 37:13601–13626.

Ziyu Liu, Yi Shen, Yunliang Jiang, Hancan Zhu, Hailong Hu, Yanlei Kang, Ming Chen, and Zhong Li. 2024c. Variation and evolution analysis of sars-cov-2 using self-game sequence optimization. *Frontiers in Microbiology*, 15:1485748.

Baichao Long, Wang Zhu, and Jianli Xiao. 2024. Stretnet: A long-term spatial-temporal traffic flow prediction method. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 3–16. Springer.

Junwei Luo, Jiaojiao Wang, Jingjing Wei, Chaokun Yan, and Huimin Luo. 2025. Deephapnet: a haplotype assembly method based on retnet and deep spectral clustering. *Briefings in Bioinformatics*, 26(1):bbae656.

Omayma Mahjoub, Sasha Abramowitz, Ruan de Kock, Wiem Khlifi, Simon du Toit, Jemma Daniel, Louay Ben Nessir, Louise Beyers, Claude Formanek, Liam Clark, and Arnu Pretorius. 2025. Sable: a performant, efficient and scalable sequence model for marl. *Preprint*, arXiv:2410.01706.

Runyu Miao, Danlin Liu, Liyun Mao, Xingyu Chen, Leihao Zhang, Zhen Yuan, Shanshan Shi, Honglin Li, and Shiliang Li. 2024. Gr-p k a: a message-passing neural network with retention mechanism for p k a prediction. *Briefings in Bioinformatics*, 25(5):bbae408.

Zhaoyi Min, Qianqian Xiao, Muhammad Abbas, and Duanjin Zhang. 2025. Retentive network-based time series anomaly detection in cyber-physical systems. *Engineering Applications of Artificial Intelligence*, 145:110215.

SIMONE MOSCO. 2023. Exploiting retentive networks in 3d lidar semantic segmentation.

Cheng Nian, Weiyi Zhang, Fasih Ud Din Farrukh, Liting Niu, Dapeng Jiang, Fei Chen, and Chun Zhang. 2024. A 77.79 gops/w retentive network fpga inference accelerator with optimized workload. In *IECON 2024-50th Annual Conference of the IEEE Industrial Electronics Society*, pages 1–7. IEEE.

Masashi Okada, Mayumi Komatsu, and Tadahiro Taniguchi. 2024. A contact model based on denoising diffusion to learn variable impedance control for contact-rich manipulation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7286–7293. IEEE.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, and 1 others. 2023. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*.

Lihong Peng, Xin Liu, Min Chen, Wen Liao, Jiale Mao, and Liqian Zhou. 2024. Mgndti: A drug-target interaction prediction framework based on multimodal representation learning and the gating mechanism. *Journal of Chemical Information and Modeling*, 64(16):6684–6698.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y. Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. 2023. Hyena hierarchy: Towards larger convolutional language models. *Preprint*, arXiv:2302.10866.

Zhen Qin, Songlin Yang, and Yiran Zhong. 2023. Hierarchically gated recurrent neural network for sequence modeling. *Advances in Neural Information Processing Systems*, 36:33202–33221.

Yufan Qiu, Yaping Liu, and Shuo Zhang. 2024. Rnete: A retentive network-based encryption traffic encoder. In *Proceedings of the 2024 3rd International Conference on Cryptography, Network Security and Communication Technology*, pages 214–219.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. 2017. Swish: a self-gated activation function. *arXiv preprint arXiv:1710.05941*, 7(1):5.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

Ruilin Su, Wanshan Zhang, and Dunhui Yu. 2024. Sequence-to-sequence multi-hop knowledge reasoning based on retentive network. In *2024 7th International Conference on Computer Information Science and Application Technology (CISAT)*, pages 360–366. IEEE.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. 2023. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*.

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. 2022. A length-extrapolatable transformer. *arXiv preprint arXiv:2212.10554*.

Yutao Sun, Li Dong, Yi Zhu, Shaohan Huang, Wenhui Wang, Shuming Ma, Quanlu Zhang, Jianyong Wang, and Furu Wei. 2024. You only cache once: Decoder-decoder architectures for language models. *Advances in Neural Information Processing Systems*, 37:7339–7361.

Paweł Tomiło. 2025. Retention mechanism based neural network model for measuring aircraft landing distance. In *2025 IEEE 12th International Workshop on Metrology for AeroSpace (MetroAeroSpace)*, pages 321–326. IEEE.

Paweł Tomiło, Jan Laskowski, and Agnieszka Laskowska. 2025. Artificial neural network model based on kolmogorov-arnold representation theorem and retention mechanism for real-time aircraft flight phases classification. *Engineering Applications of Artificial Intelligence*, 160:112004.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Baoquan Wang, Yan Zeng, and Dongming Feng. 2025a. Deep learning-based damage assessment of hinge joints for multi-girder bridges utilizing vehicle-induced bridge responses. *Engineering Structures*, 333:120148.

Bin Wang, Fei Deng, and Peifan Jiang. 2024a. Eegdir: Electroencephalogram denoising network for temporal information storage and global modeling through retentive network. *Computers in Biology and Medicine*, 177:108626.

Junliang Wang, Wenlong Hang, Shuang Liang, Qiong Wang, Badong Chen, and Jing Qin. 2025b. Convolutional retentive network for eeg decoding. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*.

Siyu Wang, Xiaocong Chen, and Lina Yao. 2024b. Retentive decision transformer with adaptive masking for reinforcement learning based recommendation systems. *arXiv preprint arXiv:2403.17634*.

Zeyu Wang, Libo Zhao, Jizheng Zhang, Rui Song, Haiyu Song, Jiana Meng, and Shidong Wang. 2025c. Multi-text guidance is important: Multi-modality image fusion via large generative vision-language model. *International Journal of Computer Vision*, pages 1–23.

Tengqing Wu. 2024. A diffusion data enhancement retentive model for sequential recommendation. In *2024 7th International Conference on Computer Information Science and Application Technology (CISAT)*, pages 114–118. IEEE.

Yuxin Wu and Kaiming He. 2018. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19.

Yunyang Xie, Kai Chen, Shenghui Li, Bingqian Li, and Ning Zhang. 2024a. Uarc: Unsupervised anomalous traffic detection with improved u-shaped autoencoder and retnet based multi-clustering. In *International Conference on Information and Communications Security*, pages 187–207. Springer.

Zexun Xie, Min Xu, Shudong Zhang, and Lijuan Zhou. 2024b. Rcat: Retentive clip adapter tuning for improved video recognition. *Electronics*, 13(5):965.

Yimo Yan, Songyi Cui, Jiahui Liu, Yaping Zhao, Bodong Zhou, and Yong-Hong Kuo. 2025. Multimodal fusion for large-scale traffic prediction with heterogeneous retentive networks. *Information Fusion*, 114:102695.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. 2025. Parallelizing linear transformers with the delta rule over sequence length. *Preprint*, arXiv:2406.06484.

Wenkui Yang, Zhida Zhang, Xiaoqiang Zhou, Junxian Duan, and Jie Cao. 2024a. Tt-df: A large-scale diffusion-based dataset and benchmark for human body forgery detection. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 429–443. Springer.

Xuan Yang, Yunxuan Dong, Lina Yang, and Thomas Wu. 2024b. Short-term photovoltaic forecasting model for qualifying uncertainty during hazy weather. *arXiv preprint arXiv:2407.19663*.

Xuan Yang, Tao Peng, Haijia Bi, and Jiayu Han. 2024c. Span-level bidirectional retention scheme for aspect sentiment triplet extraction. *Information Processing & Management*, 61(5):103823.

Yuansong Zeng, Jiancong Xie, Zhuoyi Wei, Yun Su, Ningyuan Shangguan, Shuangyu Yang, Chengyang Zhang, Wenbing Li, Jinbo Zhang, Nan Fang, and 1 others. 2024. Cellfm: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells. *bioRxiv*, pages 2024–06.

Fanlong Zhang, Huanming Chen, Quan Chen, and Jianqi Liu. 2025a. Cloud software code generation via knowledge graphs and multi-modal learning.

Lingzhe Zhang, Zewen Xiao, and Huaiyuan Wang. 2024a. Transient stability assessment of power system based on time-adapative retnet. In *2024 3rd Asia Power and Electrical Technology Conference (APET)*, pages 391–396. IEEE.

Tianning Zhang, Feng Liu, Yuming Yuan, Rui Su, Wanli Ouyang, and Lei Bai. 2024b. Fast information streaming handler (fish): A unified seismic neural network for single station real-time earthquake early warning. *arXiv preprint arXiv:2408.06629*.

Yuxuan Zhang, Zipeng Zhang, Weiwei Guo, Wei Chen, Zhaohai Liu, and Houguang Liu. 2025b. Lretunet: A u-net-based retentive network for single-channel speech enhancement. *Computer Speech & Language*, page 101798.

Zipeng Zhang, Zhencai Zhu, Bin Meng, Zheng Yang, Mingke Wu, Xinyu Cheng, Binhong Li, and Houguang Liu. 2025c. Intelligent coal gangue identification: A novel amplitude frequency sensitive neural network. *Expert Systems with Applications*, 274:126880.

Kaili Zheng, Feixiang Lu, Yihao Lv, Liangjun Zhang, Chenyi Guo, and Ji Wu. 2024. 3d human pose estimation via non-causal retentive networks. In *European Conference on Computer Vision*, pages 111–128. Springer.

Li Zhou, Dayang Wang, Yongshun Xu, Shuo Han, Bahareh Morovati, Shuyi Fan, and Hengyong Yu. 2024. Gradient guided co-retention feature pyramid network for ldct image denoising. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 153–163. Springer.

Wang Zhu, Baichao Long, and Jianli Xiao. 2024. Spatial-temporal retentive heterogeneous graph convolutional network for traffic flow prediction. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Zijie Zhu, Feng Ding, Chenglong Chu, and Fangming Zhong. 2025. Retention enhanced cross-modal attention for multi-hop vqa. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zhongliang Zou, Shuzhen Yang, Mansheng Wang, and Bo Song. 2025. Multilevel retentive networks with edge enhanced attention for infrastructure crack detection. *Engineering Structures*, 343:121191.

A Background

A.1 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are designed to model sequential and time-series data by maintaining a hidden state, which allows the network to capture temporal dependencies across time steps (Hochreiter and Schmidhuber, 1997). Formally, the RNN can be described by the following equations:

$$h_t = f_H(W_{hh} \cdot h_{t-1} + W_{hx} \cdot x_t + b_h), \quad (10)$$

$$y_t = f_O(W_{ho} \cdot h_t + b_o). \quad (11)$$

Despite their effectiveness in modeling sequential data, RNNs are prone to the vanishing gradient problem, which limits their ability to capture long-term dependencies (Bengio et al., 1994).

A.2 Transformer

The Transformer is a sequence modeling architecture that relies on self-attention to capture long-range dependencies without recurrence (Vaswani et al., 2017). Given an input sequence represented as a matrix $X \in \mathbb{R}^{n \times d}$, self-attention projects X into query, key, and value representations via learnable linear transformations:

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V, \quad (12)$$

where $W^Q \in \mathbb{R}^{d \times d_q}$, $W^K \in \mathbb{R}^{d \times d_k}$, and $W^V \in \mathbb{R}^{d \times d_v}$. The attention output is computed using scaled dot-product attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V. \quad (13)$$

To enhance representational capacity, the Transformer employs multi-head attention, which applies multiple attention operations in parallel:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V), \quad (14)$$

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^O, \quad (15)$$

where h denotes the number of heads and W^O is a learnable output projection. Multi-head attention allows the model to jointly attend to information from different representation subspaces, forming the core computational primitive of the Transformer.

B Compare RetNet with other models

B.1 Breaking the Impossible Triangle.

Table 1 highlights RetNet's distinct position on the efficiency-performance. While standard Transformers suffer from quadratic $O(N^2)$ complexity,

Architecture	Inference (per token)		Training (parallel)	
	Time	Memory	Time	Memory
Transformer (Vaswani et al., 2017)	$O(N)^{\dagger}$	$O(N)^{\dagger}$	$O(N^2)^{\ddagger}$	$O(N^2)^{\ddagger}$
Linear Transformer (Katharopoulos et al., 2020)	$O(1)$	$O(1)$	$O(N)$	$O(N)$
Hyena (Poli et al., 2023)	$O(N)$	$O(1)$	$O(N \log N)$	$O(N)$
Mamba (Gu and Dao, 2023)	$O(1)$	$O(1)$	$O(N)$	$O(N)$
DeltaNet (Yang et al., 2025)	$O(1)$	$O(1)$	$O(N)$	$O(N)$
RWKV (Peng et al., 2023)	$O(1)$	$O(1)$	$O(N)$	$O(N)$
RetNet (Sun et al., 2023)	$O(1)$	$O(1)$	$O(N)$	$O(N)$

Table 1: **Complexity Analysis of Backbone Architectures.** Comparative analysis of RetNet against Transformer, Mamba, RWKV, and other linear variants regarding time and memory complexity. N denotes the sequence length. † Standard implementation without KV cache optimization implies $O(N)$ memory growth. ‡ With optimizations like FlashAttention, memory can be reduced to linear $O(N)$, though time complexity remains quadratic.

RetNet achieves $O(1)$ inference complexity via its recurrent formulation, matching the efficiency of Linear Transformers and Mamba. Crucially, it preserves the $O(N)$ parallel training capability, effectively resolving the "impossible triangle" of training parallelism, inference efficiency, and performance. Unlike Hyena's $O(N \log N)$ convolution, RetNet's multi-scale retention offers a simpler, purely linear alternative for long-sequence modeling.

B.2 Competitive Baseline at 1.3B Scale.

Table 2 provides a controlled evaluation at the 1.3B parameter scale. In this fair setting, RetNet demonstrates strong baseline capabilities across zero-shot reasoning benchmarks (e.g., 70.07% on PIQA). Although subsequent architectures like Mamba2 achieve marginally higher accuracy (56.46% avg.) through refined state-space optimizations, RetNet (52.50% avg.) remains highly competitive against concurrent linear architectures like DeltaNet. These results validate RetNet not merely as a theoretical innovation, but as a robust foundational architecture that successfully unifies the benefits of recurrence and attention.

C Audio

In recent years, RetNet has emerged as a compelling alternative for audio-related research, offering a robust solution for modeling long-range dependencies without the quadratic overhead of standard Transformers. This shift is evident in the work of (Huang and Chen, 2024), who adapted the EEND-EDA model for long-form speaker diarization by integrating a RetNet-based encoder. Following this trajectory, Liang and Li (2024) introduced LS-EEND, where the replacement of masked self-attention with Retention facilitates linear-time inference and improved throughput. Beyond diarization, the utility of RetNet extends to the speech enhancement domain. Notably, the LRetUNet framework proposed by (Zhang et al., 2025b) employs a synergistic combination of RetNet and LSTM units to optimize time-frequency representations, specifically addressing the requirements of single-channel enhancement.

D Others

RetNet has found versatile applications in multiple domains, leveraging its capability to efficiently model long-term dependencies and handle complex sequential data.

In multimodal representation learning, VL-MFER introduces a bidirectional RetNet (Bi-RetNet) that exploits both parallel and recursive forms of multiscale retention to fuse visual and language modalities (Guo et al., 2024).

In the educational domain, a customized Retentive Module extends the original RetNet with a multiscale retention layer, capturing not only the temporal dependencies between student interactions but also their forgetting patterns, thereby enhancing the precision of learning state modeling (Linhao et al., 2024).

RetNet has also proven effective for time-series forecasting. LeRet leverages a causal retention encoder alongside multiscale retention modules to enhance nonlinear feature extraction in repaired sequences (Huang et al., 2024b). In reinforcement learning-based recommender systems, RetNet substitutes traditional masked attention with a segmented multiscale retention scheme, signifi-

Model	Wiki. ppl ↓	LMB. ppl ↓	LMB. acc ↑	PIQA acc ↑	Hella. acc ↑	Wino. acc ↑	ARC-e acc ↑	ARC-c acc ↑	Avg. acc ↑
Controlled Setting: ∼1.3B Parameters									
DeltaNet	17.71	16.88	42.46	70.72	50.93	53.35	68.47	35.66	53.60
Mamba	17.92	15.06	43.98	71.32	52.91	52.95	69.52	35.40	54.34
Mamba2	**16.56**	**12.56**	**45.66**	**71.87**	**55.67**	**55.24**	**72.47**	**37.88**	**56.46**
RetNet	19.08	17.27	40.52	70.07	49.16	54.14	67.34	33.78	52.50

Table 2: **Zero-shot Performance Comparison (1.3B Scale).** Performance comparison on language modeling and common-sense reasoning benchmarks. All models are compared at the ∼1.3B parameter scale to ensure a fair evaluation of architectural efficiency. Lower perplexity (ppl) and higher accuracy (acc) indicate better performance.

cantly improving efficiency and robustness in long-range modeling (Wang et al., 2024b). For sequential recommendation, dual RetNet modules encode both item and user history, enriching the personalized representation space (Wu, 2024). CFPSG further incorporates RetNet into a unified framework to support next-POI prediction by capturing fine-grained temporal correlations (Feiyu, 2024).

RetNet's architectural flexibility makes it well-suited for modeling motion and control dynamics. In EQNet, RetNet serves as the core of a structured state-space model, improving encoding efficiency in multi-agent motion forecasting (Huang et al., 2025b). Similarly, the NC-RetNet introduces a non-causal retention mask to access both past and future frames within blocks, improving 3D human pose estimation while maintaining low latency (Zheng et al., 2024). In TT-DF, RetNet powers the motion-guided branch, balancing long-range dependency modeling and computational cost (Yang et al., 2024a). For robotic manipulation, it is employed to process time-series inputs from learned impedance control dynamics (Okada et al., 2024).

Language and reasoning tasks also benefit from RetNet's capabilities. In ASTE, a novel bidirectional retention scheme inspired by RetNet bridges sequential and syntactic modeling gaps, boosting sentiment triplet extraction performance (Yang et al., 2024c). For world modeling, RetNet supports observation, reward, and termination prediction within REM, and is extended via the POP mechanism to generate observation sequences in parallel during imagination (Cohen et al., 2024).

System-level advancements further highlight RetNet's efficiency. A high-throughput FPGA inference accelerator incorporates RetNet to maximize hardware utility via dual-mode structure and linear computation (Nian et al., 2024). In FlashVideo, RetNet functions as the decoder, with parallel retention for training and autoregressive decoding for inference (Lei et al., 2023). Sable introduces an encoder-decoder RetNet for MARL with cross-retention and dynamic state resetting to better capture long-term dependencies in online settings (Mahjoub et al., 2025). In software engineering, RetNet is adopted as the encoder for cloud software code generation from multimodal knowledge (Zhang et al., 2025a).

In the domain of network analysis, RetNet proves invaluable in both encrypted and anomalous traffic scenarios. RN-ETE extends RetNet by incorporating a multi-resolution self-attention mechanism, enabling bidirectional retention within encrypted traffic encoding for more effective network traffic encryption and analysis (Qiu et al., 2024). On the other hand, UARC applies RetNet's retention-based reconstruction module to model long-term temporal patterns in network traffic, addressing the challenges of anomaly detection in traffic streams (Xie et al., 2024a).

In robotics, 3DMaSA (Domínguez-Vidal and Sanfeliu, 2024) adapts RetNet for predicting force and velocity signals, supporting safe and responsive human-robot interaction. In dynamic graph deep learning, Chang et al. (2024) proposed the Graph Retention Network (GRN) as a unified architecture for deep learning on dynamic graphs.

E Ecosystem and Tooling

E.1 Training and Deployment Frameworks

The current RetNet ecosystem is primarily supported by the official Microsoft *unilm* repository and community integrations in *HuggingFace*. However, there is a notable absence of RetNet in high-performance serving frameworks such as *vLLM*, *Text Generation Inference (TGI)*, or NVIDIA's

Model	Domain	Core Innovation	Key Benefit vs. Vanilla RetNet
RMT (Fan et al., 2024)	CV	**2D Manhattan Decay**: Decomposes decay into H/W axes.	Explicit modeling of 2D spatial locality in images.
RetViT (Dongre and Mehta, 2024)	CV	**Linear Backbone**: Replaces MHSA with standard MSR.	Linear complexity regarding token count; lower memory footprint.
GRetNet (Han et al., 2024)	HSI	**Gaussian Decay**: Weighted decay based on spectral similarity.	Better capture of local spectral-spatial features.
RetFormer (Erabati and Araujo, 2024)	3D	**3D Spatial Decay**: Adapted for point cloud coordinates.	Effective long-range modeling in 3D sparse data.
DenseRetNet (He et al., 2024)	NLP	**Dense Connections**: Aggregates hidden states across layers.	Improved gradient flow and training stability in deep models.
LION-D (Afzal et al., 2025)	NLP	**Bidirectional Linear**: Simplified decay for non-causal tasks.	Supports global context understanding with linear cost.
RetCompletion (Cang et al., 2024)	Gen	**Bidirectional Inference**: Dual-path context fusion.	High-speed image generation and inpainting compared to VQGAN.

Table 3: **Comparative Summary of Representative RetNet Variants.** The variants are categorized by their primary adaptation domain. "Spatial Decay" indicates whether the model modifies the standard 1D exponential decay to handle 2D/3D structures.

TensorRT-LLM.

E.2 Ecosystem Maturity Assessment

We categorize RetNet's ecosystem as "early-stage." For new practitioners, the lack of standardized, industrial-verified evaluation pipelines (like those available for Llama-based models) represents a significant barrier. The "readiness" for seamless migration from Transformer-based infrastructure to RetNet remains low due to the absence of plug-and-play kernels.

E.3 Identified Engineering Bottlenecks

Based on our profiling of existing implementations, the primary bottleneck is the **I/O-bound** nature of recurrent state updates. Current GPU memory hierarchies (HBM to SRAM) are optimized for large-scale parallel MatMuls (compute-bound). RetNet's sequential hidden state updates require frequent memory access, which currently lacks optimized "Flash-Retention" kernels to minimize latency and maximize hardware utilization on A100/H100 clusters.