Novelty Assessment Report

Paper: Block Recurrent Dynamics in Vision Transformers
PDF URL: https://openreview.net/pdf?id=gH3HhnfWLC
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

As Vision Transformers (ViTs) become standard backbones across vision, a mechanistic account of their computational phenomenology is now essential. Despite architectural cues that hint at dynamical structure, there is no settled framework that interprets Transformer depth as a well-characterized flow. In this work, we introduce the $\textbf{Block-Recurrent Hypothesis (BRH)}$, arguing that trained ViTs admit a block-recurrent depth structure such that the computation of the original L blocks can be accurately rewritten using only $k \ll L$ distinct blocks applied recurrently. Across diverse ViTs, between-layer representational similarity matrices suggest few contiguous phases. To determine whether this reflects reusable computation, we operationalize our hypothesis in the form of block recurrent surrogates of pretrained ViTs, which we call Recurrent Approximations to Phase-structured TransfORmers ($\texttt{Raptor}$). Using small-scale ViTs, we demonstrate that phase-structure metrics correlate with our ability to accurately fit $\texttt{Raptor}$ and identify the role of stochastic depth in promoting the recurrent block structure. We then provide an empirical existence proof for BRH in foundation models by showing that we can train a $\texttt{Raptor}$ model to recover 94\% of DINOv2 ImageNet-1k linear probe accuracy in only 2 blocks. To provide a mechanistic account of these observations, we leverage our hypothesis to develop a program of $\textbf{Dynamical Interpretability}$. We find $\textit{\textbf{(i)}}$ directional convergence into class-dependent angular basins with self-correcting trajectories under small perturbations $\textit{\textbf{(ii)}}$ token-specific dynamics, where $\texttt{cls}$ executes sharp late reorientations while $\texttt{patch}$ tokens exhibit strong late-stage coherence reminiscent of a mean-field effect and converge rapidly toward their mean direction and $\textit{\textbf{(iii)}}$ a collapse of the update field to low rank in late depth, consistent with convergence to low-dimensional attractors. Altogether, we find that a compact recurrent program emerges along the depth of ViTs, pointing to a low-complexity normative solution that enables these models to be studied through principled dynamical systems analysis.

Core Task Landscape

This paper addresses: **Block Recurrent Dynamics in Vision Transformers**

A total of **47 papers** were analyzed and organized into a taxonomy with **17 categories**.

Taxonomy Overview

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

- **Recurrent Mechanisms in Vision Transformers**
- **Spatial-Temporal Factorization and Multi-Scale Attention**
- **Vision Transformers for Specialized Recognition and Detection**
- **Transformer Architectural Innovations and Efficiency**
- **Domain-Specific Vision Transformer Applications**

Complete Taxonomy Tree

- Block Recurrent Dynamics in Vision Transformers Survey Taxonomy
- Recurrent Mechanisms in Vision Transformers
  - Block-Recurrent and Phase-Structured Transformers ★ (2 papers)
  - [0] Block Recurrent Dynamics in Vision Transformers (Anon et al., 2026) View paper
  - [24] Recurrent vision transformer for solving visual reasoning problems (Nicola Messina, 2022) View paper
  - Video Sequence Modeling with Recurrent Transformers (4 papers)
  - [2] TRecViT: A Recurrent Video Transformer (Patraucean, 2024) View paper
  - [6] Recurring the transformer for video action recognition (Jiewen Yang, 2022) View paper
  - [19] ViT-ReT: Vision and Recurrent Transformer Neural Networks for Human Activity Recognition in Videos (James Wensel, 2022) View paper
  - [35] Recurrent Video Masked Autoencoders (Daniel Zoran, 2025) View paper
  - Recurrent Modules for Image Restoration and Enhancement (7 papers)
  - [3] Recurrent video restoration transformer with guided deformable attention (Liang Jing-yun, 2022) View paper
  - [9] Rttlc: Video colorization with restored transformer and test-time local converter (JinJing Li, 2023) View paper
  - [26] Hyperspectral Image Denoising via Spatialâ□□Spectral Recurrent Transformer (Guanyiman Fu, 2023) View paper
  - [28] Attention-guided video super-resolution with recurrent multi-scale spatialâ□□temporal transformer (Wei Sun, 2023) View paper
  - [38] Transformer-based progressive residual network for single image dehazing (Zhe Yang, 2022) View paper
  - [40] R-Net: Recurrent and Recursive Network for Sparse-View CT Artifacts Removal (T Shen, 2019) View paper
  - [45] Transformer-based progressive (Li, 2023) View paper
- Spatial-Temporal Factorization and Multi-Scale Attention
  - Multi-Scale and Hierarchical Attention Architectures (3 papers)
  - [12] Rams-trans: Recurrent attention multi-scale transformer for fine-grained image recognition (Yunqing Hu, 2021) View paper
  - [42] Swim-Rep fusion net: A new backbone with Faster Recurrent Criss Cross Polarized Attention (Zhe Li, 2025) View paper

- [46] Adopting multiple vision transformer layers for fine-grained image representation (Fayou Sun, 2023) View paper
- Spatial-Temporal Covariance and Cross-Modal Fusion (3 papers)
- [4] Recurrence-Enhanced Vision-and-Language Transformers for Robust Multimodal Document Retrieval (Davide Caffagni, 2025) View paper
- [8] SEA-ViT: Sea Surface Currents Forecasting Using Vision Transformer and GRU-Based Spatio-Temporal Covariance Modeling (Panboonyuen, 2024) View paper
- [15] Dudocaf: Dual-domain cross-attention fusion with recurrent transformer for fast multi-contrast mr imaging (Jun Lyu, 2022) View paper
- Attention-Guided Spatial Processing and Homography Estimation (2 papers)
- [16] Recurrent Homography Estimation Using Homography-Guided Image Warping and Focus Transformer (Si-Yuan Cao, 2023) View paper
- [17] Beyond the field-of-view: Enhancing scene visibility and perception with clip-recurrent transformer (Hao Shi, 2024) View paper

- Vision Transformers for Specialized Recognition and Detection
  - Event-Based Vision and High-Speed Detection (4 papers)
  - [11] Depth AnyEvent: A Cross-Modal Distillation Paradigm for Event-Based Monocular Depth Estimation (Bartolomei Luca, 2025) View paper
  - [21] Improving multiple dense prediction performances by exploiting inter-task synergies for neuromorphic vision sensors (Tian Zhang, 2024) View paper
  - [29] PMRVT: Parallel Attention Multilayer Perceptron Recurrent Vision Transformer for Object Detection with Event Cameras (Zishi Song, 2025) View paper
  - [37] Recurrent Vision Transformers for Object Detection with Event Cameras (Mathias Gehrig, 2023) View paper
  - Medical and Multimodal Diagnostic Systems (5 papers)
  - [7] Tumor ViT-GRU-XAI: Advanced Brain Tumor Diagnosis Framework: Vision Transformer and GRU Integration for Improved MRI Analysis: A Case Study of Egypt (Mohammed Aly, 2024) View paper
  - [14] Recurrent 3-D Multi-Level Visual Transformer For Joint Classification of Heterogeneous 2-d AND 3-D Radiographic Data (Muhammad Owais, 2024) View paper
  - [22] Unified synergistic deep learning framework for multimodal 2-d and 3-d radiographic data analysis: Model development and validation (Muhammad Owais, 2024) View paper
  - [27] When CNN Meet with ViT: Towards Semi-Supervised Learning for Multi-Class Medical Image Semantic Segmentation (Wang Zi-Yang, 2022) View paper
  - [34] Explainable IRViT: Inception Recurrent Vision Transformer-Based Framework for Enhanced Breast Cancer Classification with Grad CAM Analysis (Saravanan Elumalai, 2025) View paper
  - Activity Recognition and Behavioral Analysis (3 papers)
  - [13] Vision transformer embedded video anomaly detection using attention driven recurrence (Ummay Maria Muna, 2025) View paper
  - [36] A vision transformer with recurrent neural network-based fall activity recognition system for disabled persons in smart IoT environments. (Abdulrahman Alzahrani, 2025) View paper
  - [41] SmartFallNet: A Vision Transformer and GRUâ□□Based Dynamic Model With Adaptive Kernel Attention for Precision Fall Detection (Salma Tayeb, 2025) View paper
  - Document Understanding and Multimodal Retrieval (2 papers)
  - [10] Multi-page document vqa with recurrent memory transformer (Qi Dong, 2024) View paper
  - [31] A Video Face Recognition Leveraging Temporal Information Based on Vision Transformer (Hui Zhang, 2023) View paper
- Transformer Architectural Innovations and Efficiency
  - Attention Mechanism Optimization and Replacement (2 papers)
  - [30] A recurrent vision transformer shows signatures of primate visual attention (Jonathan Morgan, 2024) View paper
  - [32] Is Attention Required for Transformer Inference? Explore Function-preserving Attention Replacement (Ren Yuxin, 2025) View paper
  - Sequential and Temporal Transformer Architectures (2 papers)
  - [39] Flexible and Efficient Spatio-Temporal Transformer for Sequential Visual Place Recognition (Lau, 2025) View paper
  - [43] Exploiting Temporal and Spatial Correlations for Efficient Visual Processing (-, 2025) View paper
  - Foundational Models and Cross-Modal Distillation (2 papers)
  - [33] OntoViT-GRU: A Conceptual Approach for Ontology-Enhanced Flood Prediction Using a Foundational Vision Transformer (Prithvi-EO-2.0) â□□ Gate Recurrent Unit (GRU) Architecture (Grujdin Ion, 2025) View paper
  - [44] VDMS: An Improved Vision Transformer-Based Model for PM2.5 Concentration Prediction (Tong Zhao, 2025) View paper
- Domain-Specific Vision Transformer Applications
  - Forecasting and Environmental Prediction (2 papers)
  - [1] On vision transformer for ultra-short-term forecasting of photovoltaic generation using sky images (Shijie Xu, 2024) View paper
  - [20] A novel integration framework for degradation-state prediction via transformer model with autonomous optimizing mechanism (Yulang Liu, 2022) View paper
  - 3D Vision and Reconstruction (2 papers)
  - [5] RRT-MVS: Recurrent Regularization Transformer for Multi-View Stereo (Jianfei Jiang, 2025) View paper
  - [47] PRAFlow_RVC: Pyramid Recurrent All-Pairs Field Transforms for Optical Flow Estimation in Robust Vision Challenge 2020 (Wan, 2020) View paper
  - Navigation and Control Systems (2 papers)
  - [18] Vision Transformers for End-to-End Vision-Based Quadrotor Obstacle Avoidance (Anish Bhattacharya, 2024) View paper
  - [25] Block-recurrent visual transformer for enhanced human detection in thermal imaging (Pham Cung Le Thien Vu, 2025) View paper
  - Biometric Authentication and Behavioral Modeling (1 papers)
  - [23] TypeFormer: Transformers for mobile keystroke biometrics (Giuseppe Stragapede, 2024) View paper

Narrative

Core task: block recurrent dynamics in vision transformers. The field has organized itself around several complementary directions. One major branch explores recurrent mechanisms in vision transformers, investigating how iterative refinement and temporal dependencies can be integrated into transformer architectures through block-recurrent structures, phase-structured designs, and hybrid recurrent-attention modules. A second branch focuses on spatial-temporal factorization and multi-scale attention, addressing how to efficiently

decompose complex visual inputs across scales and time. Additional branches examine specialized recognition and detection tasks, architectural innovations for efficiency, and domain-specific applications ranging from medical imaging to robotics. Works such as TRecViT[2] and Recurrent Video Restoration[3] illustrate how recurrent components can enhance temporal modeling, while others like RRT-MVS[5] and Recurrent Homography Estimation[16] apply these ideas to geometric vision problems.

Particularly active lines of work reveal trade-offs between computational efficiency and expressive power. Many studies adopt block-recurrent or phase-structured designs to balance the global receptive field of transformers with the parameter efficiency of recurrence, as seen in video processing tasks like Recurrent Video Restoration[3] and video anomaly detection. The original paper, Block Recurrent Dynamics[0], sits within this cluster of block-recurrent and phase-structured transformers, emphasizing structured iterative processing. Compared to nearby works such as Recurrent Visual Reasoning[24], which focuses on reasoning tasks, Block Recurrent Dynamics[0] appears to prioritize the architectural mechanism itself—how recurrent blocks can be systematically integrated into vision transformers. This contrasts with application-driven approaches like Block-recurrent Thermal Detection[25], which adapts recurrent dynamics to a specific sensing modality. Open questions remain about how best to initialize recurrent states, manage long-range dependencies, and scale these architectures to diverse visual domains.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Recurrent vision transformer for solving visual reasoning problems

Authors: Nicola Messina, G. Amato, F. Carrara, C. Gennaro, F. Falchi, et al. (9 authors total) | **Year/Venue**: 2022 | **URL**: View paper

Abstract

Although convolutional neural networks (CNNs) showed remarkable results in many vision tasks, they are still strained by simple yet challenging visual reasoning problems. Inspired by the recent success of the Transformer network in computer vision, in this paper, we introduce the Recurrent Vision Transformer (RViT) model. Thanks to the impact of recurrent connections and spatial attention in reasoning tasks, this network achieves competitive results on the same-different visual reasoning problem...

Relationship Analysis

Both papers belong to the Block-Recurrent and Phase-Structured Transformers category, exploring recurrent mechanisms in vision transformers. While the original paper investigates block-recurrent dynamics as an emergent property of trained ViTs and develops Raptor models to approximate pretrained transformers through phase-structured depth, the candidate paper introduces RViT (Recurrent Vision Transformer) as a deliberately designed recurrent architecture for visual reasoning tasks on the SVRT dataset. The key difference is that the original paper analyzes and exploits naturally emerging recurrence in foundation models, whereas the candidate paper engineers recurrence from scratch for a specific reasoning task.

Contributions Analysis

Overall novelty summary. The paper introduces the Block-Recurrent Hypothesis (BRH), proposing that trained Vision Transformers exhibit a phase-structured depth where computation across L blocks can be rewritten using k≪L distinct blocks applied recurrently. It sits within the 'Block-Recurrent and Phase-Structured Transformers' leaf, which contains only two papers total. This represents a sparse research direction within the broader taxonomy of 47 papers across 17 leaf nodes, suggesting the paper addresses a relatively unexplored aspect of vision transformer interpretability and architectural understanding.

The taxonomy reveals neighboring work in 'Video Sequence Modeling with Recurrent Transformers' (4 papers) and 'Recurrent Modules for Image Restoration' (7 papers), which apply recurrent mechanisms to specific tasks rather than analyzing inherent recurrent structure in pretrained models. The paper diverges from these application-focused directions by providing a mechanistic interpretation framework. Nearby branches in 'Spatial-Temporal Factorization' and 'Transformer Architectural Innovations' address complementary concerns about efficiency and attention mechanisms, but do not examine the dynamical flow interpretation that this work emphasizes through representational similarity analysis and phase detection.

Among 30 candidates examined across three contributions, none were identified as clearly refuting the work. The Block-Recurrent Hypothesis examined 10 candidates with 0 refutable, as did the Raptor surrogate method and the dynamical interpretability framework. This suggests limited direct prior work on block-recurrent depth structure analysis in pretrained ViTs within the search scope. The paper's focus on reusable computation phases and the role of stochastic depth in promoting recurrent structure appears distinct from existing recurrent transformer applications, though the limited search scale means potentially relevant work in mechanistic interpretability or neural network compression may exist beyond these 30 candidates.

Based on the top-30 semantic matches and taxonomy structure, the work appears to occupy a novel position at the intersection of transformer interpretability and recurrent dynamics. The sparse leaf population and absence of refuting candidates within the examined scope suggest substantive novelty, though the analysis does not cover exhaustive mechanistic interpretability literature or broader neural architecture search domains where related compression or phase-detection ideas might exist.

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

Contribution 1: Block-Recurrent Hypothesis (BRH) and empirical validation

Description: The authors formalize the Block-Recurrent Hypothesis, which states that Vision Transformers can be rewritten using a small number of parameter-tied blocks applied recurrently. They provide empirical evidence across diverse ViTs showing contiguous phase structure in layer-layer similarity matrices and demonstrate that stochastic depth promotes this recurrent block structure.

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

---

1. Three things everyone should know about vision transformers

URL: View paper

Brief Assessment

Three Things Vision[57] focuses on parallelizing sequential transformer layers for efficiency and optimization benefits, not on discovering block-recurrent depth structure with parameter-tied blocks applied recurrently as formalized in BRH.

---

2. A Manifold Representation of the Key in Vision Transformers

URL: View paper

Brief Assessment

Manifold Key Representation[68] focuses on disentangling key representations in attention mechanisms through manifold structures, not on block-recurrent depth structure or parameter-tied blocks applied recurrently across layers.

---

3. Mixture of Low-rank Experts for Transferable AI-Generated Image Detection

URL: View paper

Brief Assessment

Mixture Low-rank Experts[71] focuses on AI-generated image detection using parameter-efficient fine-tuning of CLIP-ViT models, not on block-recurrent depth structures or parameter-tied blocks in vision transformers.

---

4. RingFormer: Rethinking Recurrent Transformer with Adaptive Level Signals

URL: View paper

Brief Assessment

RingFormer[67] focuses on parameter sharing through a single transformer layer applied recurrently for computational efficiency in specific tasks (translation, image classification), not on discovering emergent block-recurrent phase structure in pretrained Vision Transformers or validating computational equivalence through representational trajectory matching.

---

5. ViT-MVT: A unified vision transformer network for multiple vision tasks

URL: View paper

Brief Assessment

ViT-MVT[65] focuses on multi-task learning with layer-adaptive sharing across different vision tasks, not on block-recurrent depth structure or parameter-tied blocks applied recurrently within a single model.

---

6. Sparse parameterization for epitomic dataset distillation

URL: View paper

Brief Assessment

Sparse Epitomic Distillation[72] focuses on dataset distillation using recurrent networks for image synthesis, not on analyzing depth structure in Vision Transformers or proposing block-recurrent computational hypotheses for ViTs.

---

7. Relaxed Recursive Transformers: Effective Parameter Sharing with Layer-wise LoRA

URL: View paper

Brief Assessment

Relaxed Recursive Transformers[70] focuses on parameter sharing through layer tying in language models with LoRA adaptation, not on the block-recurrent depth structure hypothesis in vision transformers or the empirical validation of contiguous phase structure through representational similarity analysis.

---

8. Go Wider Instead of Deeper

URL: View paper

Brief Assessment

Go Wider[73] focuses on parameter sharing across transformer blocks combined with mixture-of-experts (MoE) for width scaling, not on block-recurrent depth structure or phase-structured computation as formalized in BRH.

---

9. Minivit: Compressing vision transformers with weight multiplexing

URL: View paper

Brief Assessment

Minivit[59] focuses on weight multiplexing for model compression through parameter sharing across layers, not on formalizing block-recurrent depth structure or validating phase structure in Vision Transformers as a computational hypothesis.

---

10. Share Your Attention: Transformer Weight Sharing via Matrix-based Dictionary Learning

URL: View paper

Brief Assessment

Share Your Attention[69] focuses on weight sharing via dictionary learning for attention matrices to reduce parameters, not on block-recurrent depth structure or phase-based computational reuse across layers.

Contribution 2: Raptor: Recurrent Approximations to Phase-structured TransfORmers

Description: The authors develop Raptor, a method to train weight-tied block-recurrent approximations of pretrained ViTs that reconstruct complete internal representation trajectories. They demonstrate that a Raptor model can recover 94% of DINOv2 ImageNet-1k linear probe accuracy using only 2 recurrent blocks, providing constructive verification of functional reuse.

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

---

1. Three things everyone should know about vision transformers

URL: View paper

Brief Assessment

Three Things Vision[57] does not develop weight-tied recurrent approximations that reconstruct complete internal representation trajectories of pretrained ViTs. The parallel architecture in the candidate maintains distinct parameters across branches.

---

2. A dual-feature-based adaptive shared transformer network for image captioning

URL: View paper

Brief Assessment

Dual-feature Image Captioning[64] focuses on image captioning using dual visual features (grid and patch) with shared transformer blocks for parameter efficiency. This is fundamentally different from the original paper's work on weight-tied recurrent approximations of vision transformers for general visual tasks.

---

3. Attention mechanism for adaptive feature modelling

URL: View paper

Brief Assessment

Adaptive Feature Modelling[66] is a PhD thesis focused on attention mechanisms for feature modelling. The provided context contains only the title page and declaration, with no technical content about weight-tied recurrent approximations or vision transformers that could refute the original paper's novelty claims.

---

4. Serial Low-rank Adaptation of Vision Transformer

URL: View paper

Brief Assessment

Serial Low-rank Adaptation[63] focuses on parameter-efficient fine-tuning of vision transformers through low-rank matrix decomposition applied to attention mechanisms, not on weight-tied recurrent approximations or phase-structured depth dynamics. The candidate addresses a different technical problem (reducing fine-tuning parameters via shared low-rank matrices) compared to the original's focus on demonstrating functional reuse through recurrent block structures that reconstruct internal representation trajectories.

---

5. Cwpformer: Towards high-performance visual place recognition for robot with cross-weight attention learning

URL: View paper

Brief Assessment

Cwpformer[61] focuses on visual place recognition using vision transformers with cross-weight attention and pyramid structures for robot navigation tasks, not on weight-tied recurrent approximations of pretrained ViTs or block-recurrent depth structures.

---

6. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks

URL: View paper

Brief Assessment

VL-adapter[58] focuses on parameter-efficient transfer learning for vision-and-language tasks using adapter modules, not on weight-tied recurrent approximations of vision transformers or block-recurrent depth structures.

---

7. ViT-MVT: A unified vision transformer network for multiple vision tasks

URL: View paper

Brief Assessment

ViT-MVT[65] does not address weight-tied recurrent approximations of pretrained transformers or reconstruction of internal representation trajectories. It focuses on task-specific and task-shared parameters across multiple vision tasks.

---

8. Lightweight Recurrent Neural Network for Image Super-Resolution

URL: View paper

Brief Assessment

Lightweight Image Super-Resolution[62] focuses on recurrent neural networks for image super-resolution tasks, not on creating weight-tied block-recurrent approximations of pretrained vision transformers or analyzing their internal representation trajectories.

---

9. DTSNet: Dynamic Transformer Slimming for Efficient Vision Recognition

URL: View paper

Brief Assessment

DTSNet[60] focuses on dynamic weight sharing for efficient inference through token-adaptive compression, not on training weight-tied block-recurrent approximations that reconstruct complete internal representation trajectories of pretrained ViTs.

---

10. Minivit: Compressing vision transformers with weight multiplexing

URL: View paper

Brief Assessment

Minivit[59] develops weight-multiplexed compact models for compression purposes, not weight-tied block-recurrent approximations that reconstruct complete internal representation trajectories to verify functional reuse as described in the original paper.

Contribution 3: Dynamical Interpretability framework for Vision Transformers

Description: The authors introduce a framework for analyzing ViT depth as an iterated dynamical system. Their analysis reveals directional convergence into class-dependent angular basins, token-specific dynamics with specialized behaviors for cls and patch tokens, and collapse of update fields to low-rank subspaces consistent with convergence to low-dimensional attractors.

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

---

1. Flowing Through Layers: A Continuous Dynamical Systems Perspective on Transformers

URL: View paper

Brief Assessment

Continuous Dynamical Perspective[49] focuses on interpreting transformers as continuous ODEs via forward Euler discretization, not on analyzing ViT depth as an iterated dynamical system with phase-structured block recurrence, class-dependent angular basins, or token-specific dynamics as in the original paper.

---

2. Closed-Loop Transformers: Autoregressive Modeling as Iterative Latent Equilibrium

URL: View paper

Brief Assessment

Closed-Loop Transformers[55] focuses on autoregressive language modeling with iterative equilibrium refinement in latent space, not vision transformers or their depth dynamics. The candidate addresses a fundamentally different architecture (autoregressive transformers) and problem domain (sequence generation) compared to the original's analysis of ViT depth as dynamical systems with directional convergence.

---

3. The Mean-Field Dynamics of Transformers

URL: View paper

Brief Assessment

Mean-Field Dynamics[56] focuses on mathematical analysis of transformer attention as interacting particle systems with Wasserstein gradient flows and synchronization models. The original paper's framework analyzes ViT depth through block-recurrent structure, token-specific dynamics, and low-rank update collapse—distinct technical approaches.

---

4. Infinite limits of multi-head transformer dynamics

URL: View paper

Brief Assessment

Infinite Multi-head Limits[50] focuses on infinite-width/depth limits of transformer training dynamics using DMFT, not on analyzing ViT depth as an iterated dynamical system with convergence to class-dependent angular basins and token-specific dynamics as in the original paper.

---

5. Unveil benign overfitting for transformer in vision: Training dynamics, convergence, and generalization

URL: View paper

Brief Assessment

Benign Overfitting Vision[52] focuses on training dynamics and generalization theory for transformers through optimization analysis (three-stage convergence), not on interpreting depth as an iterated dynamical system with angular attractors and token-specific flow patterns.

---

6. A unified perspective on the dynamics of deep transformers

URL: View paper

Brief Assessment

Unified Transformer Dynamics[51] focuses on general transformer dynamics across multiple architectures (language and vision) using Vlasov equations and mean-field limits. The original paper specifically analyzes Vision Transformers through block-recurrent structure and phase-specific token dynamics, which is a distinct methodological approach.

---

7. Multi-Particle Dynamical Systems Modeling Transformers

URL: View paper

Brief Assessment

Multi-Particle Dynamics[54] focuses on modeling transformers as multi-particle dynamical systems with convergence to clusters and connections to Kuramoto oscillators, not specifically on Vision Transformers or the directional convergence into class-dependent angular basins that the original paper analyzes.

---

8. Recurrent vision transformer for solving visual reasoning problems

URL: View paper

Brief Assessment

Recurrent Visual Reasoning[24] focuses on recurrent vision transformers for visual reasoning tasks (same-different problems), not on analyzing ViT depth as a dynamical system with convergence properties. The candidate does not address directional convergence, angular basins, or low-rank attractor analysis.

---

9. Explaining transformers through dynamical systems theory

URL: View paper

Brief Assessment

Dynamical Systems Theory[48] applies Koopman operator theory to transformer decoder/encoder blocks for NLP tasks, not vision transformers. The candidate focuses on conceptualizing self-attention and feed-forward layers through Koopman operators for machine translation, whereas the original develops a framework analyzing ViT depth as iterated dynamical systems with directional convergence and token-specific dynamics for vision tasks.

---

10. The emergence of clusters in self-attention dynamics

URL: View paper

Brief Assessment

Clusters Self-Attention[53] analyzes self-attention dynamics in transformers as interacting particle systems with focus on clustering behavior and low-rank attention matrices. The original paper's framework specifically addresses ViT depth as an iterated dynamical system with directional convergence, token-specific dynamics, and low-rank update field collapse—distinct analytical perspectives not present in the candidate.

Appendix: Text Similarity Detection

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

References

- [0] Block Recurrent Dynamics in Vision Transformers View paper
- [1] On vision transformer for ultra-short-term forecasting of photovoltaic generation using sky images View paper
- [2] TRecViT: A Recurrent Video Transformer View paper
- [3] Recurrent video restoration transformer with guided deformable attention View paper
- [4] Recurrence-Enhanced Vision-and-Language Transformers for Robust Multimodal Document Retrieval View paper
- [5] RRT-MVS: Recurrent Regularization Transformer for Multi-View Stereo View paper
- [6] Recurring the transformer for video action recognition View paper
- [7] Tumor ViT-GRU-XAI: Advanced Brain Tumor Diagnosis Framework: Vision Transformer and GRU Integration for Improved MRI Analysis: A Case Study of Egypt View paper
- [8] SEA-ViT: Sea Surface Currents Forecasting Using Vision Transformer and GRU-Based Spatio-Temporal Covariance Modeling View paper
- [9] Rttlc: Video colorization with restored transformer and test-time local converter View paper
- [10] Multi-page document vqa with recurrent memory transformer View paper
- [11] Depth AnyEvent: A Cross-Modal Distillation Paradigm for Event-Based Monocular Depth Estimation View paper
- [12] Rams-trans: Recurrent attention multi-scale transformer for fine-grained image recognition View paper
- [13] Vision transformer embedded video anomaly detection using attention driven recurrence View paper
- [14] Recurrent 3-D Multi-Level Visual Transformer For Joint Classification of Heterogeneous 2-d AND 3-D Radiographic Data View paper
- [15] Dudocaf: Dual-domain cross-attention fusion with recurrent transformer for fast multi-contrast mr imaging View paper
- [16] Recurrent Homography Estimation Using Homography-Guided Image Warping and Focus Transformer View paper
- [17] Beyond the field-of-view: Enhancing scene visibility and perception with clip-recurrent transformer View paper
- [18] Vision Transformers for End-to-End Vision-Based Quadrotor Obstacle Avoidance View paper
- [19] ViT-ReT: Vision and Recurrent Transformer Neural Networks for Human Activity Recognition in Videos View paper
- [20] A novel integration framework for degradation-state prediction via transformer model with autonomous optimizing mechanism View paper
- [21] Improving multiple dense prediction performances by exploiting inter-task synergies for neuromorphic vision sensors View paper

- [22] Unified synergistic deep learning framework for multimodal 2-d and 3-d radiographic data analysis: Model development and validation View paper
- [23] TypeFormer: Transformers for mobile keystroke biometrics View paper
- [24] Recurrent vision transformer for solving visual reasoning problems View paper
- [25] Block-recurrent visual transformer for enhanced human detection in thermal imaging View paper
- [26] Hyperspectral Image Denoising via Spatialâ□□Spectral Recurrent Transformer View paper
- [27] When CNN Meet with ViT: Towards Semi-Supervised Learning for Multi-Class Medical Image Semantic Segmentation View paper
- [28] Attention-guided video super-resolution with recurrent multi-scale spatialâ□□temporal transformer View paper
- [29] PMRVT: Parallel Attention Multilayer Perceptron Recurrent Vision Transformer for Object Detection with Event Cameras View paper
- [30] A recurrent vision transformer shows signatures of primate visual attention View paper
- [31] A Video Face Recognition Leveraging Temporal Information Based on Vision Transformer View paper
- [32] Is Attention Required for Transformer Inference? Explore Function-preserving Attention Replacement View paper
- [33] OntoViT-GRU: A Conceptual Approach for Ontology-Enhanced Flood Prediction Using a Foundational Vision Transformer (Prithvi-EO-2.0) â□□ Gate Recurrent Unit (GRU) Architecture View paper
- [34] Explainable IRViT: Inception Recurrent Vision Transformer-Based Framework for Enhanced Breast Cancer Classification with Grad CAM Analysis View paper
- [35] Recurrent Video Masked Autoencoders View paper
- [36] A vision transformer with recurrent neural network-based fall activity recognition system for disabled persons in smart IoT environments. View paper
- [37] Recurrent Vision Transformers for Object Detection with Event Cameras View paper
- [38] Transformer-based progressive residual network for single image dehazing View paper
- [39] Flexible and Efficient Spatio-Temporal Transformer for Sequential Visual Place Recognition View paper
- [40] R-Net: Recurrent and Recursive Network for Sparse-View CT Artifacts Removal View paper
- [41] SmartFallNet: A Vision Transformer and GRUâ□□Based Dynamic Model With Adaptive Kernel Attention for Precision Fall Detection View paper
- [42] Swim-Rep fusion net: A new backbone with Faster Recurrent Criss Cross Polarized Attention View paper
- [43] Exploiting Temporal and Spatial Correlations for Efficient Visual Processing View paper
- [44] VDMS: An Improved Vision Transformer-Based Model for PM2.5 Concentration Prediction View paper
- [45] Transformer-based progressive View paper
- [46] Adopting multiple vision transformer layers for fine-grained image representation View paper
- [47] PRAFlow_RVC: Pyramid Recurrent All-Pairs Field Transforms for Optical Flow Estimation in Robust Vision Challenge 2020 View paper
- [48] Explaining transformers through dynamical systems theory View paper
- [49] Flowing Through Layers: A Continuous Dynamical Systems Perspective on Transformers View paper
- [50] Infinite limits of multi-head transformer dynamics View paper
- [51] A unified perspective on the dynamics of deep transformers View paper
- [52] Unveil benign overfitting for transformer in vision: Training dynamics, convergence, and generalization View paper
- [53] The emergence of clusters in self-attention dynamics View paper
- [54] Multi-Particle Dynamical Systems Modeling Transformers View paper
- [55] Closed-Loop Transformers: Autoregressive Modeling as Iterative Latent Equilibrium View paper
- [56] The Mean-Field Dynamics of Transformers View paper
- [57] Three things everyone should know about vision transformers View paper
- [58] Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks View paper
- [59] Minivit: Compressing vision transformers with weight multiplexing View paper
- [60] DTSNet: Dynamic Transformer Slimming for Efficient Vision Recognition View paper
- [61] Cwpformer: Towards high-performance visual place recognition for robot with cross-weight attention learning View paper
- [62] Lightweight Recurrent Neural Network for Image Super-Resolution View paper
- [63] Serial Low-rank Adaptation of Vision Transformer View paper
- [64] A dual-feature-based adaptive shared transformer network for image captioning View paper
- [65] ViT-MVT: A unified vision transformer network for multiple vision tasks View paper
- [66] Attention mechanism for adaptive feature modelling View paper
- [67] RingFormer: Rethinking Recurrent Transformer with Adaptive Level Signals View paper
- [68] A Manifold Representation of the Key in Vision Transformers View paper
- [69] Share Your Attention: Transformer Weight Sharing via Matrix-based Dictionary Learning View paper
- [70] Relaxed Recursive Transformers: Effective Parameter Sharing with Layer-wise LoRA View paper
- [71] Mixture of Low-rank Experts for Transferable AI-Generated Image Detection View paper
- [72] Sparse parameterization for epitomic dataset distillation View paper
- [73] Go Wider Instead of Deeper View paper