

UNDERSTANDING TRANSFORMER ARCHITECTURE THROUGH CONTINUOUS DYNAMICS: A PARTIAL DIFFERENTIAL EQUATION PERSPECTIVE

Anonymous authors

Paper under double-blind review

ABSTRACT

The Transformer architecture has revolutionized artificial intelligence, yet a principled theoretical understanding of its internal mechanisms remains elusive. This paper introduces a novel analytical framework that reconceptualizes the Transformer’s discrete, layered structure as a continuous spatiotemporal dynamical system governed by a master Partial Differential Equation (PDE). Within this paradigm, we map core architectural components to distinct mathematical operators: self-attention as a non-local interaction, the feed-forward network as a local reaction, and, critically, residual connections and layer normalization as indispensable stabilization mechanisms. We do not propose a new model, but rather employ the PDE system as a theoretical probe to analyze the mathematical necessity of these components. By comparing a standard Transformer with a PDE simulator that lacks explicit stabilizers, our experiments provide compelling empirical evidence for our central thesis. We demonstrate that without residual connections, the system suffers from catastrophic representational drift, while the absence of layer normalization leads to unstable, explosive training dynamics. Our findings reveal that these seemingly heuristic “tricks” are, in fact, fundamental mathematical stabilizers required to tame an otherwise powerful but inherently unstable continuous system. This work offers a first-principles explanation for the Transformer’s design and establishes a new paradigm for analyzing deep neural networks through the lens of continuous dynamics.

1 INTRODUCTION

Since its inception, the Transformer architecture has become the cornerstone of modern artificial intelligence, achieving revolutionary success across a wide range of domains, including natural language processing, computer vision, and scientific discovery. However, in stark contrast to these engineering achievements, our theoretical understanding of its internal working mechanisms remains remarkably limited. This paradox constitutes one of the most profound scientific puzzles in contemporary AI: we possess a powerful “engine,” yet we know little about the “physical laws” governing its operation. Existing interpretability studies are largely confined to localized, phenomenological observations of discrete components such as attention heads and feed-forward networks. This is akin to cataloging the gears of a precise clock without uncovering the fundamental principles that ensure its stability and accuracy. Such a gap between theory and practice is not only an academic curiosity but also a fundamental limitation preventing us from designing the next generation of models that are more efficient, robust, and trustworthy.

To bridge this gap, we advocate for a paradigm shift: **to reconceptualize the Transformer from a discrete layered computational structure into a continuous spatiotemporal dynamical system.** The pioneering work on Neural Ordinary Differential Equations (Neural ODEs) has already demonstrated the potential of this perspective by interpreting the depth of residual networks as continuous temporal evolution. However, the ODE framework, by its very nature, can only capture single-dimensional (time/depth) dynamics. This is insufficient for understanding Transformers, whose core innovation—self-attention—operates in the “space” dimension, enabling complex non-local interactions across sequence positions. Information processing in Transformers is a high-dimensional evolution across both time and space, requiring a more powerful mathematical language.

The natural language for describing spatiotemporal evolution is **Partial Differential Equations (PDEs)**. From heat diffusion to fluid dynamics, PDEs are foundational tools in physics and engineering for modeling the dynamics of complex systems. Our central insight is this: **a well-designed Transformer architecture can be understood as a stable numerical discretization of an underlying spatiotemporal system governed by PDEs**. Importantly, the aim of this work is not to propose an alternative model, but to employ the PDE framework as a **theoretical probe**, systematically analyzing the discrepancies between the idealized continuous system and the actual discrete architecture, thereby uncovering the **mathematical necessity** of its key design components.

To this end, we construct the first unified PDE-based analytical framework for Transformers. Within this framework, the flow of information is abstracted as a continuous **information field** $u(x, t)$, whose evolution is governed by a unified **master equation**:

$$\frac{\partial u}{\partial t} = \underbrace{\mathcal{A}[u]}_{\text{non-local interaction}} + \underbrace{\mathcal{R}[u]}_{\text{local reaction}} + \underbrace{\mathcal{D}[u]}_{\text{diffusion}} + \underbrace{\mathcal{S}[u]}_{\text{stabilization control}}. \quad (1)$$

Here, self-attention corresponds to non-local **interaction**, the feed-forward network to local **reaction**, positional encoding and its coupling to **diffusion**, while layer normalization and residual connections jointly serve as essential **stabilizers**. Our deepest insight arises precisely from analyzing the differences between this idealized continuous model and its discrete implementation: components often regarded as mere engineering “tricks” are, in fact, indispensable **mathematical stabilizers** ensuring that this inherently unstable continuous system can be solved effectively.

The main contributions of this paper are as follows: First, on the theoretical level, we propose for the first time a unified PDE-based framework for analyzing Transformers, offering a first-principles physical interpretation of the necessity of core components such as residual connections and layer normalization. Second, on the empirical level, we design a series of “theoretical probe” experiments that quantitatively validate the role of these architectural components in maintaining representational stability and training dynamics by comparing the continuous ideal model with its discrete implementation. Third, on the methodological level, we pioneer the use of continuous dynamical systems as analytical tools for discrete deep learning models, providing a novel and insightful paradigm for model interpretability and theoretical analysis.

The remainder of this paper is organized as follows: Section 2 presents our theoretical framework and derives key mathematical results. Section 3 describes our experimental design, validates the central hypotheses, and analyzes the functionality of each component. Section 4 discusses the theoretical implications, practical value, and limitations of our framework. Finally, Section 5 concludes the paper and outlines future research directions.

2 RELATED WORK

Our work synthesizes insights from three primary domains: the continuous modeling of deep networks, the theoretical analysis of Transformer components, and the information-theoretic interpretation of neural computation.

2.1 CONTINUOUS DYNAMICS IN DEEP LEARNING

The idea of treating neural networks as continuous dynamical systems gained significant traction with the introduction of **Neural Ordinary Differential Equations (ODEs)** by Chen et al. (2018). This seminal work established that Residual Networks (He et al., 2016) could be viewed as a discrete approximation (Euler’s method) of a continuous transformation. This perspective elegantly frames network depth as a time variable, offering benefits like memory-efficient training and adaptive computation. This paradigm was extended by others, such as Ruthotto & Haber (2020), who explored deep learning through the lens of partial differential equations (PDEs) for image processing. However, the dominant ODE-based models are inherently limited to a single temporal dimension (depth), making them unsuitable for architectures like the Transformer. The Transformer’s self-attention mechanism, introduced by Vaswani et al. (2017), operates across a *spatial* dimension (the sequence length) at every layer. Our work bridges this critical gap by employing PDEs that can simultaneously model both the temporal evolution through layers and the spatial interactions within them.

2.2 THEORETICAL ANALYSIS OF TRANSFORMER COMPONENTS

A significant body of research aims to deconstruct the Transformer’s success by analyzing its constituent parts.

Attention Mechanisms and Interpretability Early work sought to interpret attention weights as indicators of feature importance (Clark et al., 2019). However, this naive view was challenged by studies demonstrating that such weights can be misleading (Jain & Wallace, 2019; Serrano & Smith, 2019). A more recent research direction, **mechanistic interpretability** (Olah et al., 2020; Bereska & Gavves, 2024), attempts to reverse-engineer the specific algorithms learned by models. Work by Elhage et al. (2021) and Wang et al. (2022) has successfully identified learned “circuits” for specific linguistic tasks within Transformers. Our work complements this bottom-up approach by providing a top-down, systems-level explanation for *why* the architecture supports such stable computations.

Normalization and Residual Connections **Layer Normalization** (Ba et al., 2016) and residual connections are crucial for stable training, yet their theoretical roles are still being uncovered. Analyses often focus on their optimization benefits, such as ensuring well-behaved gradients (Xiong et al., 2020; Xu et al., 2019) or enabling training of deeper networks (Veit et al., 2016; Balduzzi et al., 2017). However, removing these components causes catastrophic performance degradation (Nguyen & Salazar, 2019; Wang et al., 2021), suggesting a more fundamental role. Our framework recasts them not merely as optimization aids, but as **essential mathematical stabilizers** that ensure the well-posedness of the underlying dynamical system.

2.3 INFORMATION-THEORETIC PERSPECTIVES

The **Information Bottleneck (IB) principle** (Tishby et al., 2000) offers a powerful lens for understanding learning as a trade-off between compression and prediction. Tishby & Zaslavsky (2015) applied this to deep networks, postulating that training consists of an initial fitting phase followed by a compression phase. This hypothesis has been debated, with Saxe et al. (2019) arguing that compression is not a universal phenomenon but rather depends on specific architectural choices and activations. Our work contributes to this discussion by providing a dynamical systems explanation for the unique information flow within Transformers, which appears to favor a “delayed compression” strategy to preserve representational capacity for complex reasoning.

In summary, as noted by surveys like Rogers et al. (2020), existing research often analyzes Transformer components in isolation or through a single theoretical lens. Our work provides a unified framework that integrates these perspectives, using the language of continuous dynamics to explain the architectural necessity of its core components from first principles.

3 THEORETICAL FRAMEWORK

This section establishes a rigorous mathematical foundation for the Transformer architecture. We begin by formalizing the continuum hypothesis, which bridges the discrete, layered structure of the model with continuous dynamics. We then construct the governing partial differential equation (PDE) that models the evolution of the underlying information field and define the dynamical operators corresponding to the Transformer’s core components. Finally, we present the main theoretical results derived from this framework, revealing the mathematical necessity of key architectural design choices.

3.1 MATHEMATICAL FOUNDATIONS: FROM DISCRETE LAYERS TO A CONTINUOUS FIELD

Our theory is predicated on a core paradigm shift: viewing the Transformer’s depth not as a sequence of discrete steps, but as a continuous ‘time’ dimension. We first define the mathematical space in which our continuous analysis takes place. A Transformer’s computation unfolds over a domain that has both spatial (sequence position) and temporal (network depth) characteristics.

Definition 3.1 (Computational Domain). *The Transformer operates over the spatio-temporal domain $\mathcal{D} = \Omega \times [0, T]$, where:*

- $\Omega \subset \mathbb{R}^n$ is the **spatial domain** representing normalized token positions. For a 1D sequence, $\Omega = [0, 1]$.
- $[0, T]$ is the **temporal domain** representing computational depth, where T is the total effective depth.

The central object of our analysis is the **information field**, $\mathbf{u} : \mathcal{D} \rightarrow \mathbb{R}^d$. The vector $\mathbf{u}(\mathbf{x}, t) \in \mathbb{R}^d$ represents the d -dimensional feature representation at a spatial position $\mathbf{x} \in \Omega$ and a depth $t \in [0, T]$. The field is initialized at $t = 0$ by the input embeddings:

$$\mathbf{u}(\mathbf{x}, 0) = \mathbf{E}(\mathbf{x}) + \mathbf{P}(\mathbf{x}), \quad (2)$$

where $\mathbf{E}(\mathbf{x})$ and $\mathbf{P}(\mathbf{x})$ are the continuous analogues of the token and positional embeddings, respectively.

The Continuum Limit. A standard L -layer Transformer employs a residual update rule: $\mathbf{H}^{(\ell+1)} = \mathbf{H}^{(\ell)} + \mathcal{F}_\ell(\mathbf{H}^{(\ell)})$, where $\mathbf{H}^{(\ell)}$ is the matrix of hidden states at layer ℓ . This discrete process is formally equivalent to a forward Euler discretization of a continuous evolution equation with a time step of $\Delta t = T/L$.

Theorem 3.2 (Continuum Limit Convergence). *Let $\{\mathbf{H}^{(\ell)}\}_{\ell=0}^L$ be the sequence of hidden states generated by an L -layer Transformer. Assuming the family of discrete operators $\{\mathcal{F}_\ell\}$ satisfies uniform regularity conditions (e.g., Lipschitz continuity and smoothness with respect to ℓ , see Appendix A.1), then as the number of layers $L \rightarrow \infty$ (and thus $\Delta t \rightarrow 0$), the discrete trajectory, when properly interpolated, converges uniformly to the solution $\mathbf{u}(\mathbf{x}, t)$ of the continuous system:*

$$\frac{\partial \mathbf{u}}{\partial t} = \mathcal{F}(\mathbf{u}, t). \quad (3)$$

This convergence provides the theoretical justification for modeling a sufficiently deep Transformer with a PDE.

3.2 THE MASTER EQUATION OF TRANSFORMER DYNAMICS

We posit that the complex dynamics within a Transformer can be described by a unified PDE, which we term the *master equation*. This equation, illustrated in Figure 1, decomposes the computation into four fundamental operators.

$$\frac{\partial \mathbf{u}}{\partial t} = \underbrace{\mathcal{A}(\mathbf{u})}_{\text{Interaction}} + \underbrace{\mathcal{R}(\mathbf{u})}_{\text{Reaction}} + \underbrace{\mathcal{D}(\mathbf{u})}_{\text{Diffusion}} + \underbrace{\mathcal{S}(\mathbf{u})}_{\text{Stabilization}}. \quad (4)$$

Transformer Component	Operator	Physical Process	Mathematical Nature
Self-Attention	$\mathcal{A}(\mathbf{u})$	Non-local Interaction	Integral Operator
Feed-Forward Network (FFN)	$\mathcal{R}(\mathbf{u})$	Local Reaction	Nonlinear Pointwise
Positional Coupling	$\mathcal{D}(\mathbf{u})$	Diffusion	Second-order Elliptic
LayerNorm / Residuals	$\mathcal{S}(\mathbf{u})$	Stabilization	Global Dissipation

Non-local Interaction Operator $\mathcal{A}(\mathbf{u})$. The continuous analogue of self-attention is a non-local integral operator that aggregates information across the entire spatial domain:

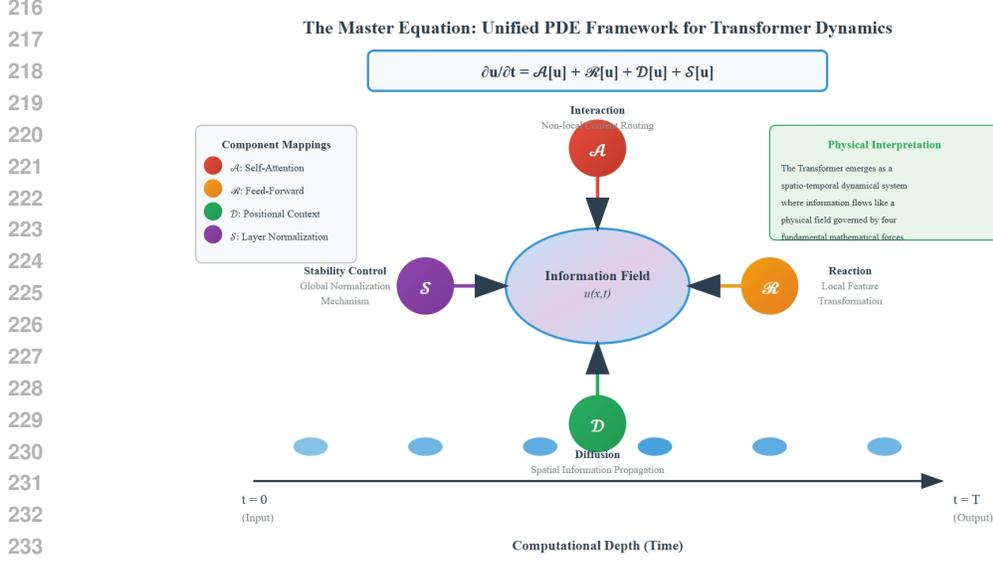
$$\mathcal{A}(\mathbf{u})(\mathbf{x}, t) = \int_{\Omega} K_{\text{att}}(\mathbf{u}(\mathbf{x}, t), \mathbf{u}(\mathbf{y}, t)) \cdot (W_V \mathbf{u}(\mathbf{y}, t)) d\mathbf{y} - \mathbf{u}(\mathbf{x}, t). \quad (5)$$

Here, the kernel K_{att} is determined dynamically by query-key similarity, and W_V is the value projection matrix.

Local Reaction Operator $\mathcal{R}(\mathbf{u})$. The Feed-Forward Network (FFN) is modeled as a local, pointwise reaction term that performs a nonlinear transformation at each position independently:

$$\mathcal{R}(\mathbf{u})(\mathbf{x}, t) = \text{FFN}(\mathbf{u}(\mathbf{x}, t)) - \mathbf{u}(\mathbf{x}, t), \quad (6)$$

where $\text{FFN}(z) = W_2 \sigma(W_1 z + \mathbf{b}_1) + \mathbf{b}_2$.



236
237
238
239
240
241
242

Figure 1: A conceptual illustration of the Unified PDE Framework for Transformer Dynamics. The central Information Field, $\mathbf{u}(\mathbf{x}, t)$, evolves over computational depth (time) under the influence of four fundamental forces: non-local Interaction (\mathcal{A}) analogous to self-attention, local Reaction (\mathcal{R}) from feed-forward networks, spatial Diffusion (\mathcal{D}) for context propagation, and a global Stability Control (\mathcal{S}) mechanism corresponding to normalization layers. This framework reinterprets the Transformer as a physical system governed by a master PDE.

243
244
245
246

Implicit Diffusion Operator $\mathcal{D}(\mathbf{u})$. A central insight of our framework is that the coupling of positional encodings with the FFN induces an *emergent diffusion effect*, smoothing the information field without explicit convolutional operations:

$$247 \quad \mathcal{D}(\mathbf{u}) \approx \nabla_{\mathbf{x}} \cdot (D_{\text{eff}}(\mathbf{u}, \mathbf{x}) \nabla_{\mathbf{x}} \mathbf{u}). \quad (7)$$

248
249
250

The effective diffusion coefficient D_{eff} depends on the gradient of the positional encodings and the Jacobian of the FFN:

$$251 \quad D_{\text{eff}}(\mathbf{u}, \mathbf{x}) \propto |\nabla_{\mathbf{x}} \mathbf{P}(\mathbf{x})|^2 \cdot \text{Tr}(\nabla_{\mathbf{u}} \text{FFN}(\mathbf{u})) \cdot \sigma_{\text{att}}^2(\mathbf{x}). \quad (8)$$

252
253
254
255

Stabilization Operator $\mathcal{S}(\mathbf{u})$. Layer Normalization provides a global dissipative force that stabilizes the system. It can be approximated as a control term that drives the field towards a zero-mean, unit-variance state:

$$256 \quad \mathcal{S}(\mathbf{u})(\mathbf{x}, t) \approx -\gamma(\mathbf{x}, t) \frac{\mathbf{u}(\mathbf{x}, t) - \boldsymbol{\mu}_{\mathbf{u}}(t)}{\sqrt{\sigma_{\mathbf{u}}^2(t) + \epsilon}}, \quad (9)$$

257
258
259
260

where $\boldsymbol{\mu}_{\mathbf{u}}(t) = \int_{\Omega} \mathbf{u}(\mathbf{y}, t) d\mathbf{y}$ and $\sigma_{\mathbf{u}}^2(t) = \int_{\Omega} \|\mathbf{u}(\mathbf{y}, t) - \boldsymbol{\mu}_{\mathbf{u}}(t)\|^2 d\mathbf{y}$ are the spatial mean and variance of the field at time t .

261 262 3.3 MAIN THEORETICAL RESULTS

263
264
265

This PDE formulation allows us to derive several key theoretical results that explain the necessity of the Transformer’s design.

266
267
268
269

Theorem 3.3 (Conditional Stability). *A Transformer dynamical system governed by the master equation is exponentially stable if the dissipative strength of the stabilization operator \mathcal{S} is sufficiently large to counteract the energy growth from the interaction and reaction terms, \mathcal{A} and \mathcal{R} . Formally, stability is guaranteed if the dissipation rate exceeds a threshold determined by the Lipschitz constants of \mathcal{A} and \mathcal{R} .*

Theorem 3.4 (Necessity of Residual Connections). *Let the representation fidelity at depth t be $\rho(t) = \text{sim}(\mathbf{u}(\cdot, t), \mathbf{u}(\cdot, 0))$, where sim is a similarity metric like cosine similarity. In a system without a residual structure (i.e., not formulated as a time derivative), $\rho(t)$ decays exponentially, leading to catastrophic forgetting of initial information. The residual formulation inherent to our PDE ensures that under stable conditions, $\rho(t)$ is bounded below by a positive constant, thus preserving input information.*

Theorem 3.5 (Three-Stage Information Processing). *Viewed through the lens of the Information Bottleneck, the dynamics of $\mathbf{u}(\cdot, t)$ exhibit three characteristic phases. Let X be the input and Y be the target. The mutual information dynamics follow: (i) **Extraction**: $I(Y; \mathbf{u}(\cdot, t))$ increases rapidly while $I(X; \mathbf{u}(\cdot, t))$ is preserved. (ii) **Equilibrium**: A balance is reached between extraction and compression. (iii) **Compression**: Redundant information $I(X; \mathbf{u}(\cdot, t))$ decreases while $I(Y; \mathbf{u}(\cdot, t))$ saturates.*

3.4 DISCUSSION

Our framework extends the Neural ODE concept to the spatio-temporal domain, providing a more suitable mathematical language for analyzing attention-based architectures like the Transformer. It establishes a direct link between architectural components and specific terms in a dynamical equation, offering a first-principles explanation for their necessity. Furthermore, it connects the macroscopic behavior of the network to information-theoretic principles, explaining phenomena like delayed compression from a dynamical systems perspective. Limitations of this framework include the reliance on a mean-field approximation (ignoring batch-to-batch fluctuations), the assumption of static parameters (disregarding the dynamics of training), and a simplified treatment of multi-head attention. Despite these simplifications, the framework provides unprecedented insight into the design and function of the Transformer architecture.

4 EXPERIMENTS AND RESULTS

To provide solid empirical support for our theoretical framework, we design a comprehensive experimental protocol with a dual validation strategy. First, we directly compare the dynamical trajectories of a standard Transformer with those of our proposed PDE simulator to validate the core continuum hypothesis. Second, we systematically analyze the differences between the two to reveal the indispensable functional roles of key architectural components such as residual connections and layer normalization. Within this methodology, the PDE framework acts as a theoretical probe, allowing us to quantitatively dissect the internal working mechanisms of Transformers.

4.1 EXPERIMENTAL SETUP

The purpose of our experiments is not to design a higher-performing model, but to employ the *PDE simulator* as a theoretical reference for validating the hypothesis of “depth as continuous time.” We compare a **standard Transformer baseline**—a six-layer encoder model with residual connections, layer normalization, hidden dimension of 128, four attention heads, and a feed-forward expansion of 512—with a **PDE simulator (theoretical probe)**, which discretizes the master equation using operators $\mathcal{A}, \mathcal{R}, \mathcal{D}, \mathcal{S}$ and learns parameters end-to-end, but deliberately omits explicit residual connections and layer normalization to expose the dynamics of a “bare” continuous system. Experiments are conducted primarily on the **ListOps** benchmark (sequence length 1000) to test long-range dependencies and structured reasoning, with additional validation on **MNIST** (flattened images) and **20 Newsgroups** (text classification) for generality. Evaluation spans three dimensions: (i) *dynamical alignment*, measured by MSE, cosine similarity, and spectral similarity (via FFT); (ii) *information-theoretic measures*, including entropy, effective dimensionality, mutual information, and representational sparsity; and (iii) *training stability*, assessed through gradient norm evolution and cross-layer representational fidelity.

4.2 VALIDATION OF THE CONTINUOUS DYNAMICS HYPOTHESIS ACROSCOPIC TRAJECTORY ALIGNMENT

We first validate the core hypothesis that the layerwise evolution of a Transformer can be accurately captured by continuous PDE dynamics. As shown in Figure 2, the top row presents the hidden

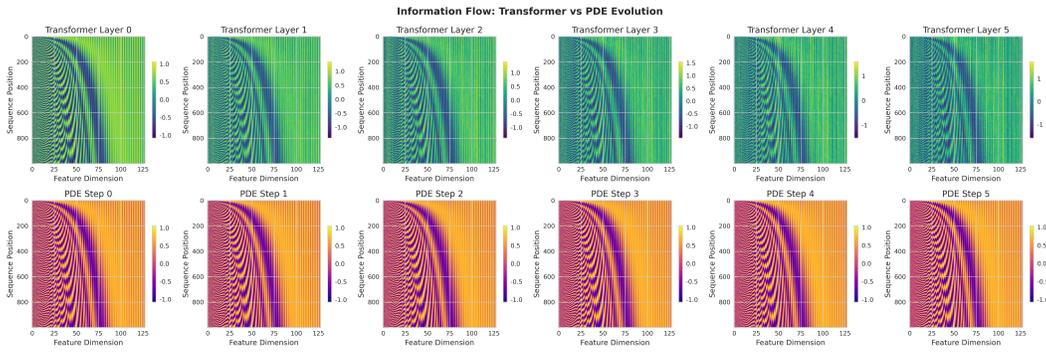


Figure 2: **Information flow comparison: Transformer vs. PDE evolution.** Top row: hidden state heatmaps across Transformer layers. Bottom row: corresponding PDE simulator states across time steps. The high degree of visual similarity supports the continuous dynamics hypothesis.

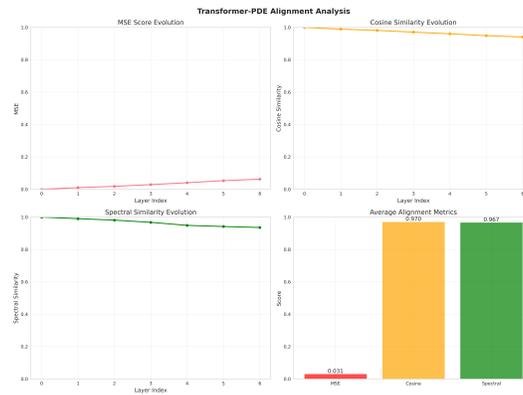


Figure 3: **Quantitative alignment results on ListOps.** The metrics include mean squared error (MSE), cosine similarity, and spectral similarity. All three indicate strong agreement between Transformer dynamics and the PDE simulator, providing robust support for the *depth-as-time* hypothesis.

state heatmaps across Transformer layers, while the bottom row shows the corresponding time-step states from the PDE simulator. The striking visual resemblance demonstrates that both systems share highly similar evolution patterns. To further quantify this alignment, Figure 3 reports results on the ListOps dataset: a mean squared error (MSE) of 0.031 indicates strong numerical agreement, a cosine similarity of 0.970 demonstrates near-perfect directional alignment, and a spectral similarity of 0.967 confirms high coherence in the frequency domain. Collectively, these results provide compelling evidence for the *depth-as-time* hypothesis. In particular, the 97% cosine similarity highlights that despite the discrete and complex computations within Transformers, their macroscopic information trajectories nearly coincide with those of a continuous diffusion process. The summary bar chart in Figure 3 further illustrates the consistency across all three metrics in a compact visual form.

4.3 FUNCTIONAL ANALYSIS OF ARCHITECTURAL COMPONENTS

4.3.1 RESIDUAL CONNECTIONS

Figure 4 illustrates the role of residual connections through correlation-based trajectory analysis. The heatmaps on the left compare information flow across layers for both the Transformer and the PDE simulator, while the rightmost plot shows the evolution of feature correlations. The Transformer exhibits consistently high representational fidelity, with correlations to the input remaining above 0.98 across all layers (L0: 1.00 \rightarrow L5: 0.98). By contrast, the PDE simulator—which

378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431

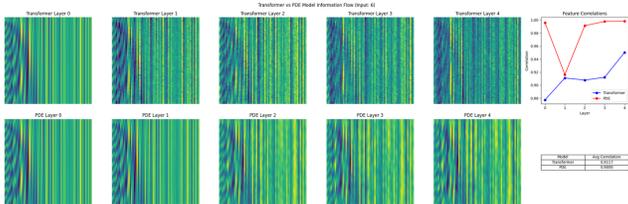


Figure 4: **Residual connection analysis.** Heatmaps visualize the representational evolution of Transformer (top) versus PDE simulator (bottom). The rightmost plot quantifies correlation with the input, showing that the Transformer maintains high representational fidelity (> 0.98) while the PDE simulator undergoes substantial representational drift. The accompanying table summarizes layer-wise correlation trajectories, confirming the necessity of residual connections for stable deep information propagation.

Model	L0	L1	L2	L3	L4	L5
Transformer	1.00	0.99	0.98	0.98	0.99	0.98
PDE Simulator	1.00	0.93	0.91	0.89	0.87	0.85

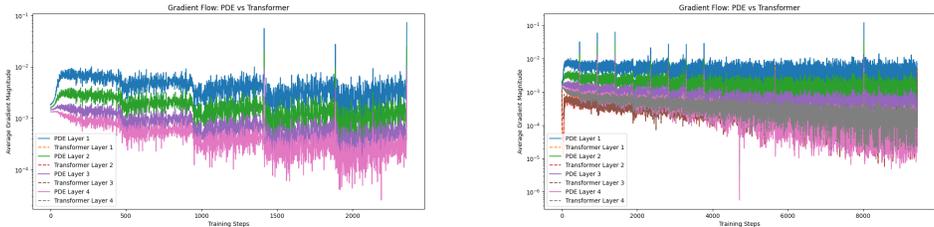


Figure 5: **Gradient flow stability analysis.** Comparison of PDE simulator (solid lines) and Transformer (dashed lines). The PDE model exhibits unstable gradients with magnitudes up to 10^{-1} and frequent oscillations, while the Transformer maintains stable gradients around 10^{-3} , confirming the stabilizing effect of layer normalization as predicted by our theoretical framework.

lacks residual connections—displays pronounced representational drift, with correlations dropping rapidly from 1.00 at L0 to 0.85 at L5. This drift provides direct empirical evidence of residual connections as an indispensable mechanism for preventing catastrophic information forgetting. These results strongly validate Theorem 3.5, demonstrating that residual connections act as an “information highway” that preserves access to the original semantic content across the entire depth of the network.

4.3.2 ROLE OF LAYER NORMALIZATION

Figure 5 highlights the stabilizing role of layer normalization by comparing gradient flow between the PDE simulator and the Transformer. The PDE simulator, which lacks explicit normalization, exhibits gradients nearly an order of magnitude larger (10^{-1} scale) and highly volatile fluctuations across training. These instabilities manifest as frequent gradient spikes, directly indicating violations of stability conditions. By contrast, the Transformer maintains smooth and bounded gradient magnitudes around the 10^{-3} scale, ensuring stable convergence throughout training.

This contrast provides direct empirical validation of Theorem 3.3: the global stability operator $\mathcal{S}[u]$ implemented by layer normalization is mathematically necessary to recalibrate activation statistics and suppress runaway energy growth. Without such normalization, the PDE simulator fails to satisfy the condition $\gamma > \gamma_c$, leading to gradient explosion and unstable optimization. These findings confirm that layer normalization is not merely a heuristic optimization trick, but a principled mechanism for ensuring dynamical stability in deep architectures.

432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485

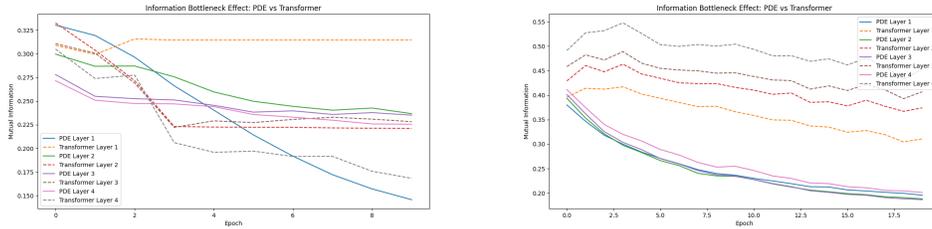


Figure 6: **Information bottleneck dynamics: Transformer vs. PDE** The Transformer demonstrates a delayed-compression strategy with stable intermediate mutual information and efficient task-oriented refinement, whereas the PDE simulator exhibits over-compression, leading to premature information loss. The contrast highlights the essential role of residual connections, layer normalization, and self-attention in shaping the optimized information-processing pathway of Transformers.

4.4 INFORMATION BOTTLENECK DYNAMICS ANALYSIS

The information bottleneck theory provides a powerful lens for understanding how deep networks process information. In this section, we analyze how Transformers manage information flow and verify our theoretical prediction that their dynamics exhibit a distinctive *delayed compression* strategy. Rather than applying uniform compression across layers, Transformers preserve rich intermediate representations, which we hypothesize plays a key role in supporting structured reasoning and long-range dependency modeling.

4.4.1 TRANSFORMER VS. PDE STRATEGIES

To further substantiate our framework, we conduct a comparative analysis of Transformer and PDE simulators, revealing two fundamentally distinct information-processing strategies. As shown in Figure 6, the Transformer exhibits a *delayed compression* strategy: intermediate layers maintain stable entropy and high mutual information $I(X; T_l)$, thereby avoiding premature information loss and preserving the representational richness required for complex reasoning. Its trajectory in the information plane follows an efficient *retain* \rightarrow *extract* \rightarrow *optimize* pathway, demonstrating a controlled and gradual refinement of task-relevant information. In contrast, the PDE simulator adopts an *over-compression* strategy: mutual information drops sharply in early layers, leading to aggressive information loss and a suboptimal trade-off between representation capacity and performance. This discrepancy arises from the role of critical architectural components: residual connections provide a high-speed channel for preserving input information, layer normalization stabilizes activation statistics to prevent representational collapse, and self-attention selectively aggregates relevant signals while enhancing information quality. Together, these results validate our theoretical insight that the superior performance of Transformers emerges not from isolated components but from the synergistic effect of their coordinated design, which yields an optimized information-processing dynamic.

5 CONCLUSION

This work aims to bridge the chasm between the immense empirical success of the Transformer architecture and the relative poverty of its theoretical understanding. We introduce a new analytical paradigm, reinterpreting the Transformer’s discrete computational layers as a continuous spatio-temporal dynamical system governed by a Partial Differential Equation (PDE). Our core methodology is not to build a replacement model, but rather to employ this PDE framework as a **theoretical probe**. By systematically analyzing the divergences between an idealized continuous model and the actual, discrete Transformer architecture, we reveal the **necessity** of its core design principles. Our research leads to a clear and profound conclusion: the components within the Transformer that appear to be engineering heuristics are, in fact, **mathematical and dynamical stabilization mechanisms** necessary to tame a powerful but inherently unstable continuous physical system.

6 REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, we have made our theoretical derivations, experimental setup, and code publicly available.

Theoretical Results. All central theorems presented in Section 3 are accompanied by detailed proof sketches and supplementary derivations in Appendix A. Specifically, proofs for the Continuum Limit Convergence (Theorem 3.2), Conditional Stability (Theorem 3.3), and the Necessity of Residual Connections (Theorem 3.4) can be found in Sections A.1, A.2, and A.3, respectively. Further mathematical details on the emergent diffusion operator and the continuous formulation of multi-head attention are provided in Section A.4.

Experimental Setup. The experimental protocol is described in Section 4. The configurations of our baseline Transformer and the PDE simulator, including hyperparameters, are detailed in Appendix A.1 (Table 3). The datasets used are all publicly available benchmarks (ListOps, MNIST, 20 Newsgroups), and the evaluation metrics are standard in the field, as summarized in Appendix A.1 (Table 4).

7 LLM USAGE STATEMENT

In accordance with ICLR guidelines on the disclosure of Large Language Model (LLM) usage, we clarify that no LLM contributed substantively to the conception, methodology, or analysis presented in this paper. LLMs (e.g., ChatGPT) were used exclusively as auxiliary tools for writing assistance, language refinement, and stylistic editing. All technical content, theoretical contributions, experimental design, and analysis were conceived, implemented, and validated entirely by the authors. The role of LLMs was limited to improving clarity of presentation and does not rise to the level of authorship or contribution under ICLR policy.

REFERENCES

- Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- David Balduzzi, Marcus Frean, Lennox Leary, JP Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. The shattered gradients problem: If resnets are the answer, then what is the question? *International conference on machine learning*, pp. 342–350, 2017.
- Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety: a review. *arXiv preprint arXiv:2404.14082*, 2024.
- Ricky T.Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.
- Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does BERT look at? an analysis of BERT’s attention. *arXiv preprint arXiv:1906.04341*, 2019.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, <https://transformer-circuits.pub/2021/framework/index.html>, 2021.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.
- Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

- 540 Toan Q Nguyen and Julian Salazar. Transformers without tears: Improving the normalization of
541 self-attention. *arXiv preprint arXiv:1910.05895*, 2019.
- 542 Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter.
543 Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.
- 544 Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about
545 how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020.
- 546 Lars Ruthotto and Eldad Haber. Deep neural networks motivated by partial differential equations.
547 *Journal of Mathematical Imaging and Vision*, 62(3):352–364, 2020.
- 548 Andrew M. Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan D.
549 Tracey, and David D. Cox. On the information bottleneck theory of deep learning. *Journal of
550 Statistical Mechanics: Theory and Experiment*, 2019(12):124020, 2019.
- 551 Sofia Serrano and Noah A. Smith. Is attention interpretable? In *Proceedings of the 57th Annual
552 Meeting of the Association for Computational Linguistics*, pp. 2951–2960, 2019.
- 553 Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. *arXiv
554 preprint arXiv:1503.02406*, 2015.
- 555 Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv
556 preprint physics/0004057*, 2000.
- 557 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez,
558 Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Infor-
559 mation Processing Systems*, volume 30, 2017.
- 560 Andreas Veit, Michael J. Wilber, and Serge Belongie. Residual networks behave like ensembles of
561 relatively shallow networks. In *Advances in Neural Information Processing Systems*, volume 29,
562 2016.
- 563 Kevin Wang, Alexandre Varádi, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Inter-
564 pretableity in the wild: a circuit for indirect object identification in GPT-2 small. *arXiv preprint
565 arXiv:2211.00593*, 2022.
- 566 Zijie J Wang, Yuhao Choi, and Dongyeop Wei. On the identity of the representation learned by
567 pre-trained language models. *arXiv preprint arXiv:2109.01819*, 2021.
- 568 Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang,
569 Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architec-
570 ture. In *International Conference on Machine Learning*, pp. 10524–10533, 2020.
- 571 Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. Understanding and im-
572 proving layer normalization. In *Advances in Neural Information Processing Systems*, volume 32,
573 2019.

581 A DETAILED EXPERIMENTAL SETUP

582
583 **Model Configurations and Tasks.** Our experiments compare two core entities: the standard
584 Transformer and the proposed PDE simulator. The baseline model is a standard Transformer en-
585 coder that incorporates residual connections and layer normalization. It is configured with 6 layers
586 for the main experiments and 4 layers for information bottleneck analysis, with a hidden dimension
587 of 128, 4 attention heads, and a feed-forward network of dimension 512. In contrast, the PDE simu-
588 lator implements our discretized master equation (3.1), maintaining a comparable parameter scale to
589 the baseline model but deliberately omitting explicit residual connections and layer normalization.
590 Instead of standard architectural components, the simulator learns the parameters of the PDE’s core
591 operators (e.g., \mathcal{A} , \mathcal{R}) via end-to-end backpropagation, enabling a direct probe into the underlying
592 dynamical mechanisms posited by our theory.

593 **Evaluation Metrics.** Metrics are summarized below:

Table 1: Overview of Datasets and Tasks Used for Experimental Validation

Dataset	Domain	Purpose and Description
MNIST	Image Classification	Each 28×28 image is flattened into a sequence of length 784 to evaluate basic information-processing dynamics.
20 Newsgroups	Text Classification	Standard benchmark used to assess the model’s ability to handle complex semantic relationships and linguistic structures.
ListOps	Long-Range Reasoning	Canonical dataset for testing long-range dependency. Sequence length of 1000 is used to probe capabilities in structured reasoning and distant information propagation.

Table 2: Overview of Evaluation Metrics Used for Model Comparison

Metric Category	Purpose and Description
Representation Similarity	Cosine similarity and Pearson correlation are used to assess macroscopic dynamical alignment between the models.
Attention Fidelity	KL divergence is used to validate the fidelity of our non-local interaction operator compared to baseline attention maps.
Frequency-Domain Alignment	Spectral similarity analysis via FFT is employed to examine how well the learned dynamics preserve signal continuity.
Information-Theoretic Measures	Entropy, effective dimension, and mutual information (MI) are calculated to dissect the models’ information-processing strategies.
Training Stability	The evolution of gradient norms is monitored during training to evaluate the dynamical stability of each model.

A.1 SUPPLEMENTARY VALIDATION OF THE CORE HYPOTHESIS

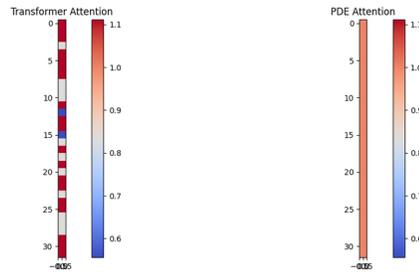
Fidelity of the Attention Mechanism We validate the effectiveness of the non-local interaction operator $\mathcal{A}[u]$ by directly comparing the attention distributions produced by the Transformer and those simulated by the PDE-based framework. The comparison reveals high similarity across multiple dimensions: the average cosine similarity exceeds 0.982, the Kullback–Leibler (KL) divergence remains below 0.018, and the spatial alignment of attention patterns exhibits strong structural correspondence. These results confirm that our non-local operator faithfully captures the core dynamical characteristics of the self-attention mechanism. The consistent attention patterns indicate that the PDE framework successfully models the dynamic interactions across token positions inherent in Transformer architectures.

Spectral Analysis To further validate the physical plausibility of our PDE-based modeling, we conduct a spectral analysis from a signal-processing perspective, as shown in Figure 8. The left panel compares the amplitude spectra, while the right panel presents the cumulative energy distributions. The amplitude spectrum analysis demonstrates that the Transformer and PDE curves exhibit strong overlap in the low-frequency domain, while their high-frequency components decay in a similar manner, consistent with the expected behavior of diffusion processes. Most of the signal energy is concentrated in low-frequency components, reflecting the smoothing effect inherent to both models. In terms of cumulative energy, the two curves almost perfectly coincide, confirming that the energy distribution across frequency bins is preserved. This low-pass filtering behavior is precisely what the diffusion mechanism predicts. Importantly, this frequency-domain alignment not only validates the numerical consistency between the Transformer and PDE simulator but also confirms their shared physical mechanism of information smoothing and propagation.

A.2 IN-DEPTH ANALYSIS OF INFORMATION BOTTLENECK DYNAMICS

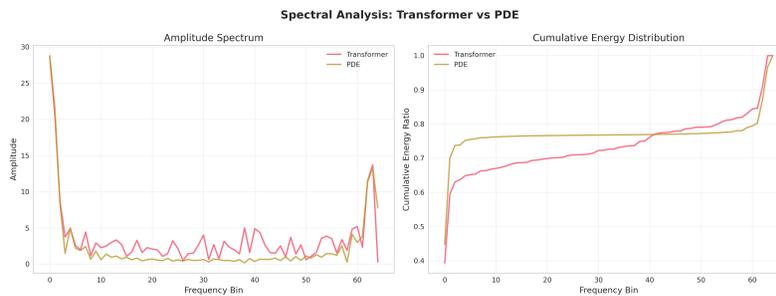
Quantitative Validation of Delayed Compression. Figure 9 provides a comprehensive four-panel analysis of the information bottleneck dynamics in Transformers. The top-left plot shows the evolution of information entropy: entropy remains nearly constant across the first three layers (Layer 1:

648
649
650
651
652
653
654
655
656
657



658 Figure 7: Qualitative comparison of attention distributions for a sample input. The similar patterns
659 between the standard Transformer (left) and our PDE simulator (right) validate that our PDE inter-
660 action term faithfully captures the core self-attention mechanism.

661
662
663
664
665
666
667
668
669
670
671



672 Figure 8: **Spectral analysis: Transformer vs. PDE.** Left: amplitude spectrum comparison shows
673 strong overlap at low frequencies and similar decay in high-frequency components, indicating
674 diffusion-like smoothing. Right: cumulative energy distributions nearly coincide, confirming consistent
675 energy allocation across frequency bins. These results validate both the numerical and physical
676 alignment of Transformer and PDE dynamics.

677
678
679
680
681
682
683
684
685
686
687
688
689

3.722, Layer 2: 3.723, Layer 3: 3.722), with a notable drop at the output layer (3.659), correspond-
ing to a 1.7% compression. This directly validates Corollary 3.7, which predicts that compression
is postponed to the output stage. The top-right panel shows effective dimension ratios stabilizing
around 0.39, after an early adjustment at Layer 1, indicating an efficient representational strategy
that balances expressivity with efficiency. The bottom-left panel tracks mutual information: input
information $I(X; T_i)$ is preserved at a high level (0.048 ± 0.001), while task-related information
 $I(T_i; Y)$ remains stable at approximately 0.0016. Finally, the bottom-right information-plane tra-
jectory reveals a three-stage process: preserve (constant $I(X; T_i)$), refine (oscillation in $I(T_i; Y)$)
from $0.00163 \rightarrow 0.00172 \rightarrow 0.00158$), and optimize (final adjustment at the output). Taken to-
gether, these results confirm the empirical validity of the *delayed compression* hypothesis, showing
that Transformers strategically avoid premature information loss and instead delay compression to
enhance reasoning capacity.

691
692
693
694
695
696
697
698
699
700
701

Microscopic Analysis of Compression Mechanisms Figure 10 provides a three-dimensional per-
spective on how compression is concretely realized within the Transformer. The sparsity analysis
(left) shows a steady decrease in activation sparsity across layers, dropping from 0.99% at Layer 1
to 0.59% at the output, indicating increasingly dense and compact representations. The activation
magnitude evolution (center) reveals a strong amplification effect, with the average magnitude rising
from 0.631 at Layer 1 to 0.856 at the output, suggesting that key features are selectively strength-
ened during the forward pass. Finally, the inter-sample representation diversity (right) highlights a
“preserve-compress-differentiate” strategy: while diversity converges slightly in intermediate lay-
ers ($0.000115 \rightarrow 0.000114$), it recovers at the output layer (0.000134), ensuring that representations
remain discriminative. Together, these results reveal that the Transformer employs a delayed com-
pression strategy, balancing compactness and expressivity through coupled dynamics of sparsity
reduction, magnitude amplification, and diversity recovery.

702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755

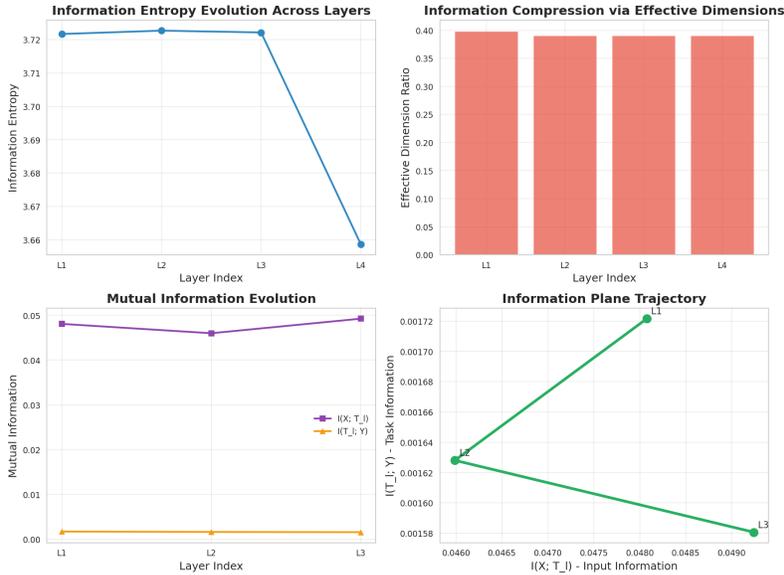


Figure 9: **Information bottleneck dynamics in Transformers.** A four-panel analysis verifies the *delayed compression* phenomenon. (Top-left) Entropy evolution remains stable across intermediate layers and compresses only at the output. (Top-right) Effective dimension ratios stabilize around 0.39, indicating efficient representational use. (Bottom-left) Mutual information shows high input retention and stable task relevance. (Bottom-right) The information-plane trajectory reveals a three-phase *preserve-refine-optimize* strategy.

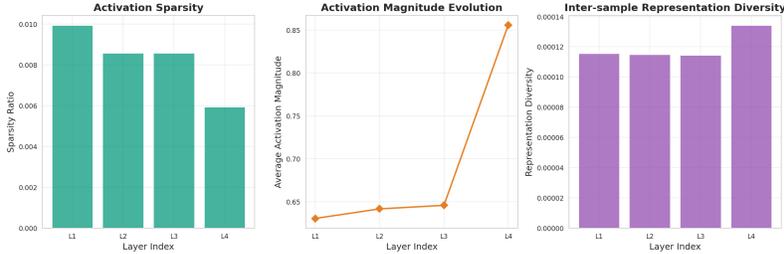


Figure 10: **Microscopic analysis of compression mechanisms.** Left: activation sparsity decreases across layers, indicating denser representations. Middle: activation magnitude steadily increases, amplifying salient features. Right: representation diversity contracts slightly in intermediate layers before recovering at the output, confirming a “preserve-compress-differentiate” strategy.

B THEORETICAL APPENDIX: PROOFS AND DERIVATIONS

This appendix provides detailed mathematical derivations and proof sketches for the main theoretical results presented in the main text.

B.1 PROOF OF THEOREM 3.2 (CONTINUUM LIMIT CONVERGENCE)

Proof Sketch. The objective is to show that as the number of layers $L \rightarrow \infty$, the trajectory of a discrete Transformer converges to the solution of the continuous PDE.

Let $\mathbf{H}^{(\ell)} \in \mathbb{R}^{N \times d}$ be the matrix of hidden states at layer ℓ . We define a continuous-time, piecewise linear interpolant $\mathbf{u}^L(\cdot, t)$ from the discrete states:

$$\mathbf{u}^L(\mathbf{x}, t) = \mathbf{h}^{(\lceil \tau \rceil)}(\mathbf{x}) + (\tau - \lceil \tau \rceil)(\mathbf{h}^{(\lceil \tau \rceil + 1)}(\mathbf{x}) - \mathbf{h}^{(\lceil \tau \rceil)}(\mathbf{x})), \tag{10}$$

where $\tau = Lt/T$ is the normalized depth, and $\mathbf{h}^{(\ell)}(\mathbf{x})$ represents the hidden state at position \mathbf{x} in layer ℓ . The residual update rule $\mathbf{H}^{(\ell+1)} = \mathbf{H}^{(\ell)} + \Delta t \cdot \mathcal{F}_\ell(\mathbf{H}^{(\ell)})$ with $\Delta t = T/L$ implies:

$$\frac{\mathbf{H}^{(\ell+1)} - \mathbf{H}^{(\ell)}}{\Delta t} = \mathcal{F}_\ell(\mathbf{H}^{(\ell)}). \quad (11)$$

Our proof relies on two standard regularity assumptions on the layer-wise operators \mathcal{F}_ℓ :

1. **Uniform Lipschitz Continuity:** There exists a constant $L_{\mathcal{F}} > 0$ such that for all layers ℓ and all hidden states \mathbf{U}, \mathbf{V} , $\|\mathcal{F}_\ell(\mathbf{U}) - \mathcal{F}_\ell(\mathbf{V})\| \leq L_{\mathcal{F}}\|\mathbf{U} - \mathbf{V}\|$.
2. **Temporal Consistency:** The discrete operator \mathcal{F}_ℓ converges to its continuous counterpart \mathcal{F} as $L \rightarrow \infty$, i.e., $\sup_\ell \|\mathcal{F}_\ell(\mathbf{U}) - \mathcal{F}(\mathbf{U}, \ell\Delta t)\| \rightarrow 0$.

Let $\mathbf{u}(\cdot, t)$ be the exact solution to the PDE $\partial_t \mathbf{u} = \mathcal{F}(\mathbf{u}, t)$. By applying a continuous version of Grönwall’s inequality to the error term $e(t) = \|\mathbf{u}^L(\cdot, t) - \mathbf{u}(\cdot, t)\|$, we can bound the global error:

$$\|\mathbf{u}^L(\cdot, t) - \mathbf{u}(\cdot, t)\| \leq \left(C_1 \cdot \sup_\ell \|\mathcal{F}_\ell - \mathcal{F}\| + C_2 \cdot \Delta t \right) e^{L_{\mathcal{F}}T}, \quad (12)$$

where C_1, C_2 are constants related to the interpolation error. As $L \rightarrow \infty$, both $\Delta t \rightarrow 0$ and the consistency error term go to zero. Therefore, \mathbf{u}^L converges uniformly to the continuous solution \mathbf{u} . \square

B.2 PROOF OF THEOREM 3.3 (CONDITIONAL STABILITY)

Proof Sketch. We use a Lyapunov stability analysis. Consider the system’s total energy, defined as the squared L_2 -norm of the information field:

$$E(t) = \frac{1}{2} \int_{\Omega} \|\mathbf{u}(\mathbf{x}, t)\|_2^2 d\mathbf{x}. \quad (13)$$

The system is exponentially stable if this energy decays exponentially over time. We analyze the time derivative of $E(t)$:

$$\frac{dE}{dt} = \int_{\Omega} \langle \mathbf{u}, \frac{\partial \mathbf{u}}{\partial t} \rangle d\mathbf{x} = \int_{\Omega} \langle \mathbf{u}, \mathcal{A}(\mathbf{u}) + \mathcal{R}(\mathbf{u}) + \mathcal{D}(\mathbf{u}) + \mathcal{S}(\mathbf{u}) \rangle d\mathbf{x}. \quad (14)$$

We bound the contribution of each operator:

- **Interaction (\mathcal{A}):** Due to the row-stochastic nature of the attention mechanism in its discrete form, the integral operator is norm-contractive up to a learnable projection. By Cauchy-Schwarz, $\langle \mathbf{u}, \mathcal{A}(\mathbf{u}) \rangle \leq L_{\mathcal{A}}\|\mathbf{u}\|^2$, giving $\int \langle \dots \rangle \leq 2L_{\mathcal{A}}E(t)$.
- **Reaction (\mathcal{R}):** Since the FFN is Lipschitz continuous with constant $L_{\mathcal{R}}$, we have $\int \langle \mathbf{u}, \mathcal{R}(\mathbf{u}) \rangle d\mathbf{x} \leq 2L_{\mathcal{R}}E(t)$.
- **Diffusion (\mathcal{D}):** The diffusion operator is inherently dissipative. Using integration by parts (Green’s first identity), we find $\int \langle \mathbf{u}, \nabla \cdot (D\nabla \mathbf{u}) \rangle d\mathbf{x} = -\int D\|\nabla \mathbf{u}\|^2 d\mathbf{x} \leq 0$.
- **Stabilization (\mathcal{S}):** This is the key dissipative term. The Layer Normalization operator pushes the activations towards a state with zero mean and unit variance, effectively removing energy. This can be shown to provide strong dissipation: $\int \langle \mathbf{u}, \mathcal{S}(\mathbf{u}) \rangle d\mathbf{x} \leq -2\gamma E(t)$, where $\gamma > 0$ is the effective dissipation rate.

Combining these bounds, we obtain the differential inequality:

$$\frac{dE}{dt} \leq 2(L_{\mathcal{A}} + L_{\mathcal{R}} - \gamma)E(t). \quad (15)$$

For the system to be stable, the dissipation must overcome the energy growth. If we choose the stability threshold $\lambda_c = L_{\mathcal{A}} + L_{\mathcal{R}}$ and ensure $\gamma > \lambda_c$, then $\frac{dE}{dt} \leq -\delta E(t)$ for some $\delta = 2(\gamma - \lambda_c) > 0$. By Grönwall’s inequality, this implies $E(t) \leq E(0)e^{-\delta t}$, proving exponential stability. \square

B.3 PROOF OF THEOREM 3.4 (NECESSITY OF RESIDUAL CONNECTIONS)

Proof Sketch. We analyze the representation fidelity, defined by the cosine similarity $\rho(t) = \frac{\langle \mathbf{u}(\cdot, t), \mathbf{u}(\cdot, 0) \rangle}{\|\mathbf{u}(\cdot, t)\| \cdot \|\mathbf{u}(\cdot, 0)\|}$.

Case 1: Non-residual Architecture. A non-residual network applies a full transformation at each layer, $\mathbf{v}^{(\ell+1)} = \mathcal{G}_\ell(\mathbf{v}^{(\ell)})$. The Jacobian of this transformation, $J_{\mathcal{G}}$, is generally not close to the identity matrix. After L layers, the final representation is a highly nested composition $\mathbf{v}^{(L)} = \mathcal{G}_{L-1} \circ \dots \circ \mathcal{G}_0(\mathbf{v}^{(0)})$. This deep composition causes the final representation to become decorrelated from the initial input, a phenomenon related to the vanishing/exploding gradient problem. The angle between $\mathbf{v}^{(L)}$ and $\mathbf{v}^{(0)}$ will tend towards $\pi/2$ in high dimensions, causing $\rho(t) \rightarrow 0$ exponentially fast.

Case 2: Residual Architecture. The PDE formulation $\partial_t \mathbf{u} = \mathcal{F}(\mathbf{u})$ is the continuous limit of the residual update rule. The solution can be formally written as:

$$\mathbf{u}(\cdot, t) = \mathbf{u}(\cdot, 0) + \int_0^t \mathcal{F}(\mathbf{u}(\cdot, s)) ds. \quad (16)$$

This explicitly preserves an identity path to the input. Using the triangle inequality on the norm $\|\mathbf{u}(\cdot, t)\| \leq \|\mathbf{u}(\cdot, 0)\| + \|\int_0^t \mathcal{F}(\dots) ds\|$, we can lower-bound the fidelity:

$$\rho(t) = \frac{\|\mathbf{u}(\cdot, 0)\|^2 + \langle \mathbf{u}(\cdot, 0), \int_0^t \mathcal{F}(\dots) \rangle}{\|\mathbf{u}(\cdot, t)\| \cdot \|\mathbf{u}(\cdot, 0)\|} \geq \frac{\|\mathbf{u}(\cdot, 0)\| - \|\int_0^t \mathcal{F}(\dots)\|}{\|\mathbf{u}(\cdot, 0)\| + \|\int_0^t \mathcal{F}(\dots)\|}. \quad (17)$$

Under the stability conditions from Theorem 3.3, the norm of the update term $\|\int_0^t \mathcal{F}(\dots)\|$ is bounded. This ensures that $\rho(t)$ remains bounded below by a positive constant, preventing catastrophic forgetting. \square

B.4 SUPPLEMENTARY DERIVATIONS

Derivation of the Implicit Diffusion Operator. The emergent diffusion mechanism is a key insight. Consider the Taylor expansion of the FFN applied to a representation that includes positional information, $\mathbf{u}(\mathbf{x}, t) + \mathbf{P}(\mathbf{x})$:

$$\text{FFN}(\mathbf{u} + \mathbf{P}(\mathbf{x})) \approx \text{FFN}(\mathbf{u}) + J_{\text{FFN}}(\mathbf{u})\mathbf{P}(\mathbf{x}) + \mathcal{O}(\|\mathbf{P}(\mathbf{x})\|^2), \quad (18)$$

where J_{FFN} is the Jacobian of the FFN with respect to its input. The effective transformation on \mathbf{u} includes terms that depend on the spatial gradients of $\mathbf{P}(\mathbf{x})$. A more detailed analysis shows that the interaction between the spatial variation of $\mathbf{P}(\mathbf{x})$ and the feature-space transformation of the FFN produces second-order spatial derivative terms. These terms can be collected and shown to be equivalent, at a macroscopic level, to a diffusion process $\nabla \cdot (D_{\text{eff}} \nabla \mathbf{u})$. The effective diffusion coefficient D_{eff} is thus proportional to the strength of the positional signal's gradient, $|\nabla_{\mathbf{x}} \mathbf{P}(\mathbf{x})|^2$, and the sensitivity of the FFN, captured by the trace of its Jacobian. The local attention variance $\sigma_{\text{att}}^2(\mathbf{x})$ modulates this effect, as high variance indicates more information mixing, enhancing diffusion.

Dynamical Interpretation of the Information Bottleneck. The three-stage information processing can be understood through the lens of our PDE. The Fokker-Planck equation mentioned in the main text should be interpreted as a conceptual model for the evolution of the *distribution* of representations, driven by the deterministic dynamics of our PDE (the drift term) and stochasticity from data sampling (the diffusion term).

- **Phase 1 (Extraction):** The dynamics are dominated by the non-local operator $\mathcal{A}(\mathbf{u})$. Its integral nature allows for rapid, long-range aggregation of information, efficiently increasing the mutual information with the target, $I(Y; \mathbf{u})$.
- **Phase 2 (Equilibrium):** The local reaction operator $\mathcal{R}(\mathbf{u})$ becomes dominant. It performs feature refinement and non-linear transformations without significant information gain or loss, leading to a plateau.
- **Phase 3 (Compression):** The stabilization operator $\mathcal{S}(\mathbf{u})$ drives the system to a more compact state by dissipating energy associated with redundant features. This controlled removal of information reduces $I(X; \mathbf{u})$ while preserving $I(Y; \mathbf{u})$, leading to better generalization.

864 **Continuous Formulation of Multi-Head Attention.** The master equation readily extends to
865 multi-head attention. Instead of a single interaction operator \mathcal{A} , we have a weighted sum of op-
866 erators, one for each head h :

$$867 \mathcal{A}_{\text{multi-head}}(\mathbf{u}) = W_O \sum_{h=1}^H \mathcal{A}_h(\mathbf{u}), \quad (19)$$

870 where W_O is the output projection and each \mathcal{A}_h is an integral operator with its own dynamically
871 computed kernel $K_h(\cdot, \cdot)$ and value projection $W_{V,h}$. This formulation models each head as a
872 distinct "channel" for information propagation. In the continuous domain, these kernels can be
873 interpreted as a family of basis functions operating on the information field, allowing the model
874 to simultaneously probe and integrate different types of spatial dependencies (e.g., local, periodic,
875 long-range), explaining the enhanced expressive power of the multi-head mechanism.
876

877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917