Future Research Directions for the AMTF and GATF Frameworks

Soumadeep Ghosh

Kolkata, India

Abstract

The Adaptive Measure-Theoretic Filter (AMTF) and Geometry-Aware Transformer Filter (GATF) establish foundational frameworks for state estimation in chaotic dynamical systems by integrating Oseledets decomposition with modern filtering and deep learning architectures. This paper identifies eight principal directions for extending these frameworks: (1) particle Transformer filters combining sequential Monte Carlo with geometry-aware attention; (2) sparse attention mechanisms for high-dimensional state spaces; (3) continuous-time neural stochastic differential equation formulations; (4) rigorous convergence rate analysis; (5) uncertainty quantification respecting geometric structure; (6) joint state-parameter estimation in partially observed systems; (7) multi-scale temporal architectures exploiting Lyapunov timescale separation; and (8) comprehensive experimental validation across diverse application domains. For each direction, we present formal problem statements, mathematical frameworks, architectural specifications, and connections to relevant literature. Collectively, these directions chart a path toward transforming GATF into a mature methodology with solid theoretical foundations, efficient implementations, and demonstrated practical impact.

The paper ends with “The End”

Contents

1 Introduction **3**

2 Particle Transformer Filters **3**
2.1 Motivation and Problem Statement . . . 3
2.2 Mathematical Framework . . . 4
2.3 Theoretical Properties . . . 5
2.4 Algorithmic Specification . . . 6

3 Sparse Attention Mechanisms for High-Dimensional State Spaces **6**
3.1 Motivation and Problem Statement . . . 6
3.2 Mathematical Framework . . . 7
3.3 Complexity Analysis . . . 8
3.4 Theoretical Properties . . . 8

4 Continuous-Time Neural Stochastic Differential Equation Formulations **9**
4.1 Motivation and Problem Statement . . . 9
4.2 Mathematical Framework . . . 9
4.3 Training Objective . . . 10
4.4 Numerical Challenges and Solutions . . . 10

5 Theoretical Convergence Rate Analysis 10
5.1 Motivation and Problem Statement 10
5.2 Mathematical Framework 11
5.3 Main Theoretical Results 11

6 Uncertainty Quantification 12
6.1 Motivation and Problem Statement 12
6.2 Mathematical Framework 12

7 Partially Observed Systems with Unknown Parameters 13
7.1 Motivation and Problem Statement 13
7.2 Mathematical Framework 14

8 Multi-Scale Temporal Modeling 14
8.1 Motivation and Problem Statement 14
8.2 Mathematical Framework 15

9 Experimental Validation 16
9.1 Benchmark Suite 16
9.2 Baseline Methods 16
9.3 Evaluation Metrics 16

10 Additional Research Directions 17
10.1 Equivariant GATF Architectures 17
10.2 Interpretability and Visualization 17
10.3 Online Learning and Adaptation 17

11 Conclusion 18

List of Figures

1 Research agenda structure. 3
2 Particle GATF architecture. 5
3 Sparse attention patterns for GATF. 8
4 Continuous-time GATF architecture. 10
5 Theoretical convergence rate comparison. 12
6 Uncertainty quantification in GATF. 13
7 Hierarchical GATF for joint state-parameter estimation. 14
8 Multi-scale GATF architecture. 15
9 Expected comparative performance on Lorenz-63. 17

List of Tables

1 Comparison of sparse attention patterns 8
2 Proposed benchmark suite for GATF validation 16

1 Introduction

Classical filtering theory, originating with the Kalman filter [1] and extended to non-linear settings via the Kushner–Stratonovich equation [2], assumes a fixed probabilistic structure. However, many real-world systems exhibit chaotic dynamics where local stability properties vary dramatically across the state space. The Adaptive Measure-Theoretic Filter (AMTF) [6] addresses this limitation by dynamically adjusting its filtration structure based on geometric information from the underlying attractor, incorporating Oseledets decomposition [4] to enable geometry-aware adaptive gain mechanisms.

Building upon AMTF, the Geometry-Aware Transformer Filter (GATF) [7] synthesizes measure-theoretic foundations with the Transformer architecture [5], aligning multi-head attention mechanisms with Oseledets subspaces and incorporating Lyapunov-weighted attention scores. This creates a filtering framework that inherits rigorous probabilistic foundations while leveraging the representational power and parallelism of modern attention-based architectures.

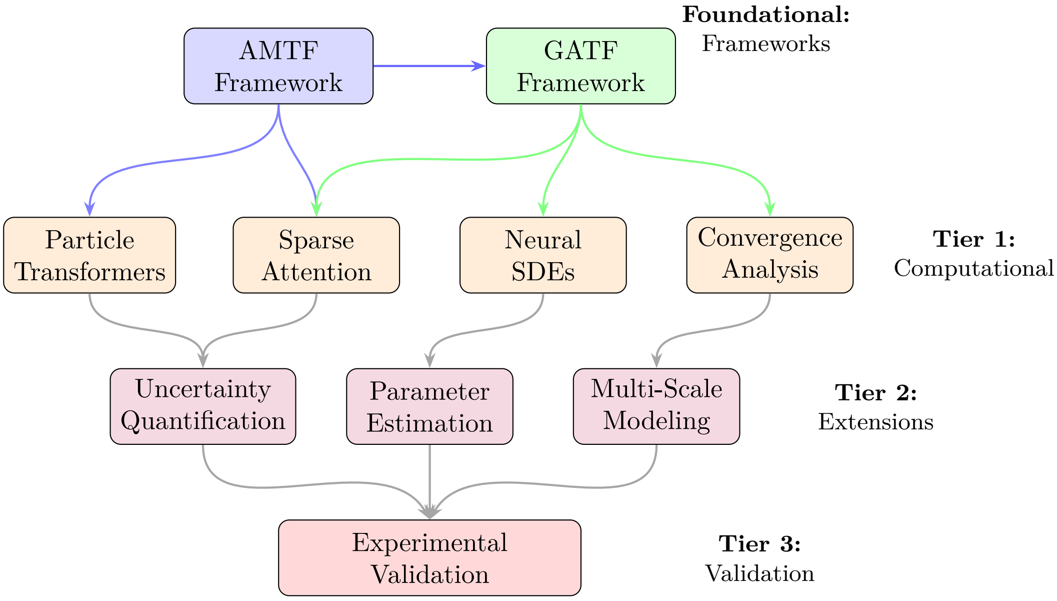


Figure 1: Research agenda structure.

Shows the progression from foundational AMTF and GATF frameworks through computational extensions (Tier 1), theoretical and architectural extensions (Tier 2), to experimental validation (Tier 3).

This paper presents a comprehensive research agenda that addresses both theoretical gaps and practical extensions of the AMTF/GATF framework. The structure follows a tiered approach illustrated in Figure 1, progressing from foundational computational innovations through theoretical deepening to empirical validation.

2 Particle Transformer Filters

2.1 Motivation and Problem Statement

The current GATF framework operates on continuous density representations through the Zakai equation [3], which becomes computationally intractable in high-dimensional state spaces. The Zakai SPDE requires discretization over the full state space, with computational complexity scaling exponentially in dimension—the curse of dimensionality renders this approach impractical when $n > 10$.

Problem 2.1 (Particle GATF Design)**.** Design a particle-based implementation of the GATF architecture that:

1. Represents the filtering distribution $p_t(x|Z_{0:t})$ through N weighted particles $\{(x_t^{(i)}, w_t^{(i)})\}_{i=1}^N$
2. Leverages Oseledets-aligned attention mechanisms for importance sampling proposals
3. Maintains the theoretical guarantees of GATF stability in the particle approximation
4. Achieves computational complexity $O(N \cdot n \cdot h)$ where h is the number of attention heads

2.2 Mathematical Framework

Definition 2.2 (Particle GATF Filtering Distribution)**.** The particle approximation to the GATF filtering density is:

$$\hat{p}_t^N(x|Z_{0:t}) = \sum_{i=1}^{N} w_t^{(i)} \delta_{x_t^{(i)}}(x) \tag{1}$$

where δ_x denotes the Dirac measure centered at x, and weights satisfy $\sum_i w_t^{(i)} = 1$.

Definition 2.3 (Attention-Based Importance Proposal)**.** The geometry-aware importance sampling proposal distribution takes the form:

$$q(x_{t+1}|x_t^{(i)}, Z_{0:t+1}) = \sum_{j=1}^{k} \alpha_j^{(t)}(x_t^{(i)}) \cdot \mathcal{N}\left(x_{t+1}; \mu_j^{(t)}(x_t^{(i)}), \Sigma_j^{(t)}(x_t^{(i)})\right) \tag{2}$$

where k is the number of Oseledets subspaces and the mixture weights are derived from Lyapunov-weighted attention scores:

$$\alpha_j^{(t)}(x) = \frac{\exp(\beta \cdot \lambda_j(t))}{\sum_{\ell=1}^{k} \exp(\beta \cdot \lambda_\ell(t))} \tag{3}$$

Each component (μ_j, Σ_j) is aligned with Oseledets subspace $E_j(x)$:

$$\mu_j^{(t)}(x) = x + \Delta t \cdot \Pi_t^{(j)} f(x, \theta_t) \tag{4}$$

$$\Sigma_j^{(t)}(x) = \Delta t \cdot \Pi_t^{(j)} \sigma(x)\sigma(x)^\top (\Pi_t^{(j)})^\top + \epsilon I \tag{5}$$

Definition 2.4 (Oseledets-Guided Resampling)**.** Stratified resampling preserves particle diversity along unstable manifolds by computing effective sample size within Voronoi cells:

$$N_j^{\text{eff}} = \left(\sum_{i: x_t^{(i)} \in \mathcal{V}_j} (w_t^{(i)})^2 \right)^{-1} \tag{6}$$

where $\mathcal{V}_j$ is a Voronoi cell in the Oseledets-adapted metric. Resampling is triggered independently in each cell when $N_j^{\text{eff}} < N_{\text{thresh}}$.

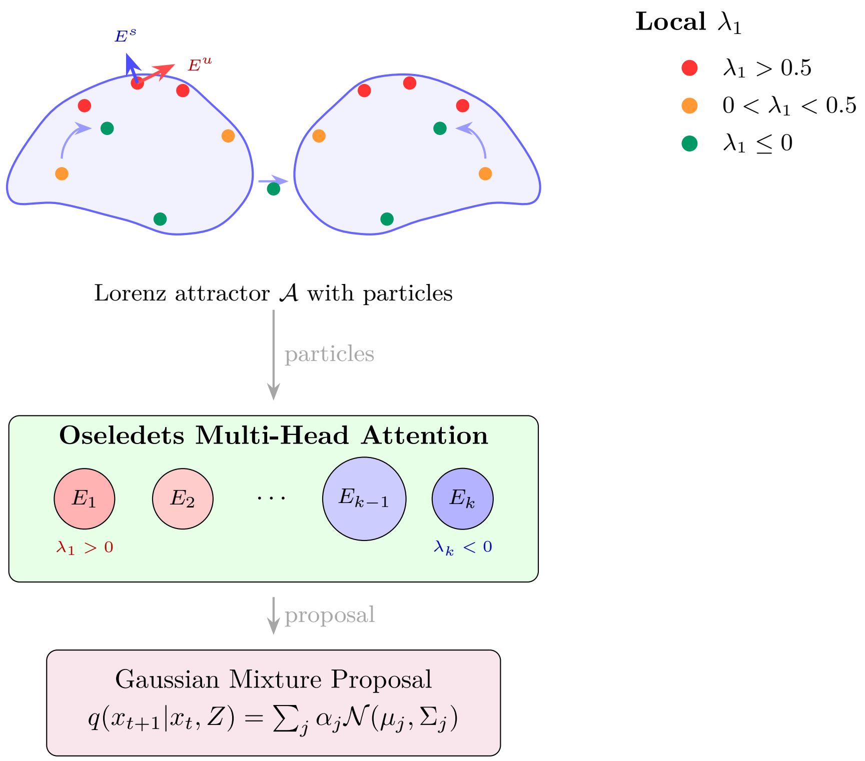


Figure 2: Particle GATF architecture.

Particles distributed on the Lorenz strange attractor are colored by local maximal Lyapunov exponent. The Oseledets Multi-Head Attention processes particles through subspace-aligned heads (E_1 unstable to E_k stable), generating mixture proposals for importance sampling.

2.3 Theoretical Properties

Proposition 2.5 (Unbiasedness)**.** *The particle GATF estimator is asymptotically unbiased:*

$$\lim_{N\to\infty} \mathbb{E}\left[\hat{p}_t^N(x|Z_{0:t})\right] = p_t(x|Z_{0:t}) \tag{7}$$

almost surely with respect to the observation measure.

Proposition 2.6 (Variance Reduction)**.** *The Oseledets-aligned importance proposal achieves variance reduction over standard bootstrap proposals:*

$$\operatorname{Var}\left[\hat{p}_t^{\mathrm{GATF}}\right] \leq \operatorname{Var}\left[\hat{p}_t^{\mathrm{bootstrap}}\right] \cdot \prod_{j:\lambda_j>0} e^{-c\lambda_j \Delta t} \tag{8}$$

where the product is over unstable Lyapunov exponents and $c > 0$ is a constant depending on the proposal alignment quality.

Conjecture 2.7 (Central Limit Theorem)**.** *Under suitable regularity conditions, the particle GATF satisfies:*

$$\sqrt{N}\left(\hat{p}_t^N - p_t\right) \xrightarrow{d} \mathcal{N}(0, \Sigma_t^{\mathrm{GATF}}) \tag{9}$$

where the asymptotic covariance Σ_t^{GATF} is smaller than the standard particle filter covariance along unstable Oseledets directions.

2.4 Algorithmic Specification

Algorithm 1 Particle Geometry-Aware Transformer Filter (P-GATF)

```
Require: Initial particles $\{x_0^{(i)}\}_{i=1}^N$, observations $\{Z_t\}$, Transformer weights $\Theta$
Ensure: Particle approximation $\{(x_t^{(i)}, w_t^{(i)})\}_{i=1}^N$
Initialize weights $w_0^{(i)} = 1/N$ for all $i$
Initialize Oseledets basis via QR decomposition at each particle
for $t = 1, 2, \ldots, T$ do
    // Lyapunov Estimation (parallel over particles)
    for each particle $i = 1, \ldots, N$ do
        Update local Lyapunov estimates $\hat{\lambda}_j(t, x^{(i)})$ via recursive QR
        Compute local Oseledets projectors $\Pi_t^{(j)}(x^{(i)})$
    end for
    // Attention-Based Importance Sampling
    Compute attention weights $\alpha_j^{(t)}$ via Lyapunov-weighted softmax
    for each particle $i = 1, \ldots, N$ do
        Sample component $j \sim \text{Categorical}(\alpha_1^{(t)}, \ldots, \alpha_k^{(t)})$
        Sample $\tilde{x}_{t+1}^{(i)} \sim \mathcal{N}(\mu_j^{(t)}(x_t^{(i)}), \Sigma_j^{(t)}(x_t^{(i)}))$
    end for
    // Weight Update
    for each particle $i = 1, \ldots, N$ do
        Compute likelihood: $\ell^{(i)} = p(Z_{t+1}|\tilde{x}_{t+1}^{(i)})$
        Compute transition: $\tau^{(i)} = p(\tilde{x}_{t+1}^{(i)}|x_t^{(i)})$
        Compute proposal: $q^{(i)} = q(\tilde{x}_{t+1}^{(i)}|x_t^{(i)}, Z_{0:t+1})$
        Update weight: $\tilde{w}_{t+1}^{(i)} = w_t^{(i)} \cdot \ell^{(i)} \cdot \tau^{(i)}/q^{(i)}$
    end for
    Normalize: $w_{t+1}^{(i)} = \tilde{w}_{t+1}^{(i)} / \sum_j \tilde{w}_{t+1}^{(j)}$
    // Oseledets-Guided Resampling
    Partition particles into Voronoi cells $\{\mathcal{V}_j\}$ based on Oseledets metric
    for each cell $j$ do
        Compute effective sample size $N_j^{\text{eff}}$
        if $N_j^{\text{eff}} < N_{\text{thresh}}$ then
            Resample particles within cell $j$
        end if
    end for
    // Transformer Refinement (optional)
    $x_{t+1}^{(i)} \leftarrow \tilde{x}_{t+1}^{(i)} + \text{MHA}_{\text{Osel}}(\{\tilde{x}_{t+1}^{(j)}\}_{j=1}^N)$
end for
return $\{(x_t^{(i)}, w_t^{(i)})\}$
```

3 Sparse Attention Mechanisms for High-Dimensional State Spaces

3.1 Motivation and Problem Statement

Standard self-attention in GATF has computational complexity $O(n^2)$ in the spatial discretization dimension, making it prohibitive for applications such as climate modeling (state dimension $\sim 10^6$) or large-scale robotics. The quadratic scaling arises from the dense attention matrix in the standard formulation [5].

Problem 3.1 (Sparse GATF Design)**.** Design sparse attention patterns for GATF that:

1. Reduce computational complexity from $O(n^2)$ to $O(n \cdot s(n))$ where $s(n) = o(n)$
2. Preserve the geometric structure captured by Oseledets decomposition
3. Maintain stability guarantees under sparsification
4. Adaptively allocate attention bandwidth based on local Lyapunov exponents

3.2 Mathematical Framework

Definition 3.2 (Sparse Attention Mask)**.** A sparse attention pattern is defined by a binary mask $M \in \{0,1\}^{n\times n}$:

$$\text{SparseAttention}(Q, K, V; M) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} \odot M + (1 - M) \cdot (-\infty)\right) V \tag{10}$$

where $\odot$ denotes elementwise multiplication and the $-\infty$ terms zero out masked positions after softmax.

We propose three complementary sparsity patterns:

Definition 3.3 (Oseledets-Block Sparse Attention)**.** Attention is restricted to queries and keys within the same Oseledets subspace:

$$M_{ij}^{\text{Osel}} = \mathbf{1}\left[\exists \ell : x_i \in E_\ell \text{ and } x_j \in E_\ell\right] \tag{11}$$

This yields complexity $O(n \cdot \bar{d})$ where $\bar{d}$ is the average subspace dimension.

Definition 3.4 (Flow-Neighborhood Sparse Attention)**.** Attention is restricted to points within ϵ-neighborhoods under the flow-induced metric:

$$M_{ij}^{\text{flow}} = \mathbf{1}\left[d_{\text{flow}}(x_i, x_j) < \epsilon\right] \tag{12}$$

where the flow metric is:

$$d_{\text{flow}}(x, y) = \inf_{T>0} \left\{ \int_0^T \|\phi_s(x) - \phi_s(y)\| \, ds \right\}^{1/2} \tag{13}$$

with ϕ_s the flow map.

Definition 3.5 (Lyapunov-Thresholded Adaptive Sparsity)**.** Full attention in chaotic regions; local attention elsewhere:

$$M_{ij}^{\text{Lyap}}(t) = \begin{cases} 1 & \text{if } \lambda_1(x_i, t) > \tau \text{ or } \lambda_1(x_j, t) > \tau \\ M_{ij}^{\text{local}} & \text{otherwise} \end{cases} \tag{14}$$

where $\tau > 0$ is a threshold and M^{local} is a fixed local attention pattern.

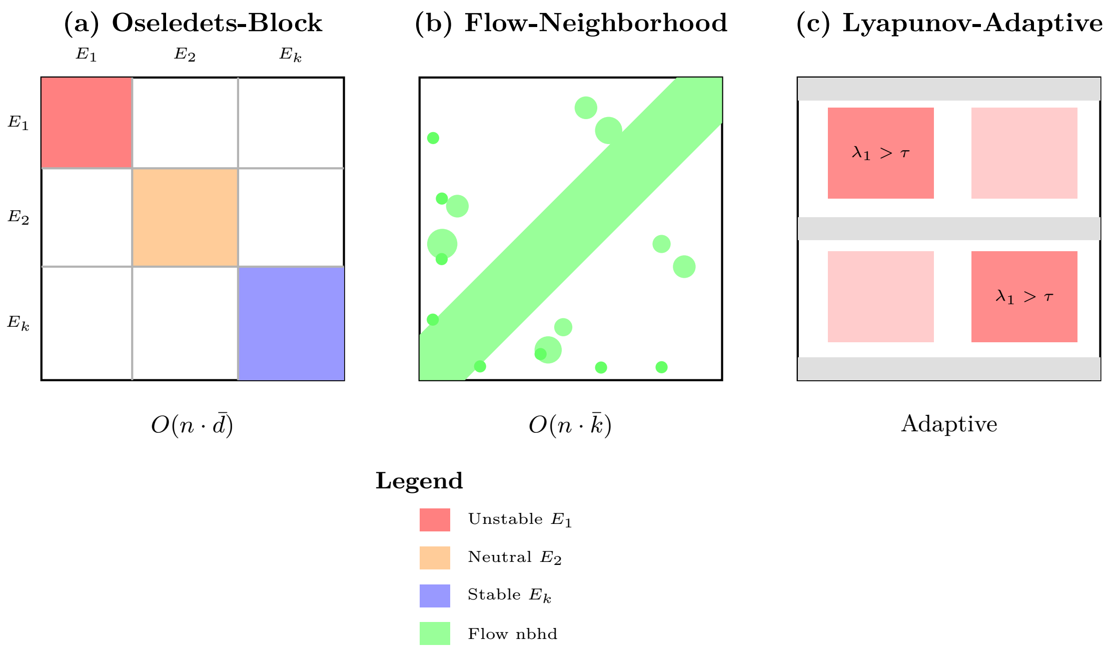


Figure 3: Sparse attention patterns for GATF.

(a) Oseledets-block sparsity restricts attention within subspaces E_j. (b) Flow-neighborhood sparsity connects points along the attractor flow topology. (c) Lyapunov-adaptive sparsity uses dense attention only in chaotic regions ($\lambda_1 > \tau$) with cross-attention between them.

3.3 Complexity Analysis

Table 1: Comparison of sparse attention patterns

Sparsity Pattern	Mask Density	Complexity	Preserves Geometry
Dense (baseline)	1	$O(n^2)$	Yes
Oseledets-block	$\sum_\ell d_\ell^2/n^2$	$O(n \cdot \bar{d})$	Yes (by construction)
Flow-neighborhood	$\bar{k}/n$	$O(n \cdot \bar{k})$	Approximately
Lyapunov-threshold	Adaptive	$O(n \cdot n_{\text{chaotic}} + n \cdot s_{\text{local}})$	In chaotic regions
Combined	Product	$O(n \cdot \min(\bar{d}, \bar{k}))$	Yes

3.4 Theoretical Properties

Theorem 3.6 (Sparse GATF Stability). *Let $M^{(t)}$ be a sequence of sparse attention masks satisfying:*

1. ***Connectivity**: The graph induced by $M^{(t)}$ is connected for all t*
2. ***Oseledets coverage**: For each unstable subspace E_j with $\lambda_j > 0$, the restriction $M^{(t)}|_{E_j \times E_j}$ has density at least $\rho_{\min} > 0$*

Then the sparse GATF update is mean-square stable:

$$\mathbb{E}\left[\|\rho_t^{\text{sparse-GATF}} - \rho_t^*\|_{L^2}^2\right] \leq C \cdot e^{-\gamma' t} \tag{15}$$

where $\gamma' = \gamma \cdot \rho_{\min}$ and γ is the dense GATF convergence rate.

Proposition 3.7 (Approximation Error Bound)**.** *The sparse attention approximation error is bounded by:*

$$\|\mathrm{SparseAttn}(Q,K,V;M)-\mathrm{Attention}(Q,K,V)\|_F \leq \|V\|_F \cdot \sqrt{1-\|M\|_0/n^2} \cdot \max_{(i,j):M_{ij}=0} A_{ij} \quad (16)$$

where $A = \mathrm{softmax}(QK^\top/\sqrt{d_k})$ *and* $\|M\|_0$ *is the number of nonzeros.*

4 Continuous-Time Neural Stochastic Differential Equation Formulations

4.1 Motivation and Problem Statement

The discrete-time GATF update introduces discretization artifacts and handles irregular observation times awkwardly. A continuous-time formulation would unify discrete Transformer updates with the continuous-time Zakai framework [3] while enabling natural treatment of asynchronous observations.

Problem 4.1 (Continuous-Time GATF)**.** Develop a continuous-time GATF formulation that:

1. Parameterizes the drift and diffusion of the filtering SPDE using attention-based neural networks
2. Evolves the Oseledets decomposition continuously via the Lyapunov differential equation
3. Modulates filtering gain as a continuous function of time via attention mechanisms
4. Handles irregular observation times without discretization artifacts
5. Preserves the Girsanov transformation structure from AMTF

4.2 Mathematical Framework

Definition 4.2 (Continuous-Time Neural Filtering SPDE)**.** The filtering density $\rho_t(x)$ evolves according to:

$$d\rho_t(x) = \mathcal{L}_\theta^*[\rho_t](x)\,dt + \mathcal{A}_\phi[\rho_t](x)\,dt + \rho_t(x)h(x)^\top R^{-1}dZ_t \quad (17)$$

where:

- $\mathcal{L}_\theta^*$ is a neural network-parameterized adjoint generator:

$$\mathcal{L}_\theta^*[\rho](x) = -\nabla\cdot(f_\theta(x,\rho)\cdot\rho) + \frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial x_i\partial x_j}\left([\sigma_\theta\sigma_\theta^\top]_{ij}\rho\right) \quad (18)$$

- $\mathcal{A}_\phi[\rho_t]$ is the continuous-time attention correction:

$$\mathcal{A}_\phi[\rho_t](x) = \int_{\mathcal{A}} K_\phi(x,y;\Lambda_t)\nabla_y\rho_t(y)\cdot\Pi_t^{(u)}(y)\,dy \quad (19)$$

Definition 4.3 (Continuous Oseledets Evolution)**.** The Oseledets projectors evolve according to the Lyapunov differential equation:

$$\frac{d}{dt}\Pi_t^{(j)} = \left[D_xf(\hat{x}_t),\Pi_t^{(j)}\right] + \text{Gram-Schmidt correction} \quad (20)$$

where $[\cdot,\cdot]$ denotes the commutator and $\hat{x}_t$ is the current state estimate.

Definition 4.4 (Continuous Attention Kernel)**.** The attention kernel takes the form:

$$K_\phi(x,y;\Lambda_t) = \frac{\exp\left(\langle q_\phi(x), k_\phi(y)\rangle/\sqrt{d} + \beta \sum_j \lambda_j(t) \langle \Pi_t^{(j)} x, \Pi_t^{(j)} y\rangle\right)}{\int_{\mathcal{A}} \exp\left(\langle q_\phi(x), k_\phi(z)\rangle/\sqrt{d} + \beta \sum_j \lambda_j(t) \langle \Pi_t^{(j)} x, \Pi_t^{(j)} z\rangle\right) dz} \tag{21}$$

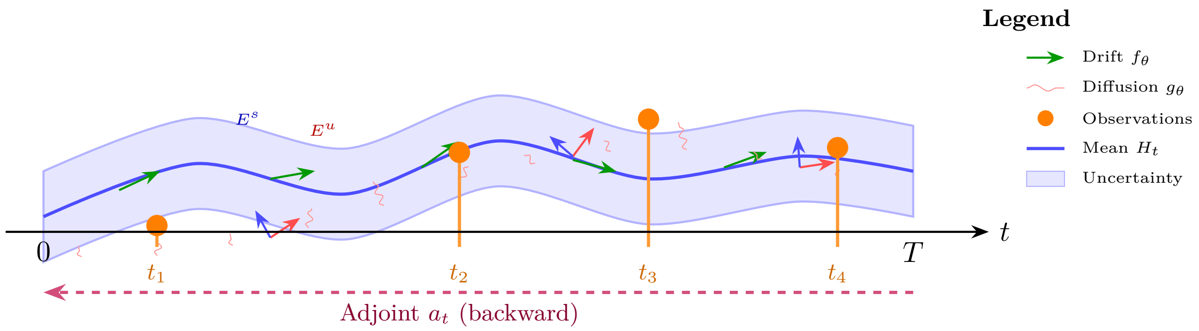


Figure 4: Continuous-time GATF architecture.

The filtering density evolves as a continuous "tube" (blue shaded region) around the mean trajectory (dark blue), driven by drift f_θ (green arrows) and diffusion g_θ (red noise). Observations arrive at irregular times t_1, t_2, t_3, t_4 (orange impulses). Oseledets frames (red/blue arrows) rotate continuously. The adjoint a_t (dashed purple) flows backward for gradient computation.

4.3 Training Objective

Definition 4.5 (Continuous-Time GATF Loss)**.** The training objective combines filtering accuracy with physics-informed regularization:

$$\mathcal{L}(\theta,\phi) = \underbrace{\mathbb{E}\left[\int_0^T \|H_t - H_t^*\|^2 dt\right]}_{\text{filtering loss}} + \lambda_Z \underbrace{\mathbb{E}\left[\int_0^T \|\mathcal{R}_{\text{Zakai}}[\rho_t]\|^2 dt\right]}_{\text{Zakai residual}} + \lambda_O \underbrace{\mathbb{E}\left[\int_0^T \|\Pi_t^{(j)}(\Pi_t^{(j)})^\top - \Pi_t^{(j)}\|_F^2 dt\right]}_{\text{Oseledets orthogonality}} \tag{22}$$

where $\mathcal{R}_{\text{Zakai}}$ is the Zakai equation residual and H_t^* is the target hidden state.

4.4 Numerical Challenges and Solutions

1. **Stiffness**: The Lyapunov spectrum may span several orders of magnitude. *Solution*: Employ implicit-explicit (IMEX) splitting schemes.

2. **Boundary Conditions**: The compact attractor $\mathcal{A}$ imposes constraints on the density support. *Solution*: Employ reflecting boundary conditions or penalize density mass outside $\mathcal{A}$.

3. **Adjoint Computation**: Backpropagation through the neural SDE requires computing gradients through stochastic dynamics. *Solution*: Use the adjoint sensitivity method [14]:

$$da_t = -a_t^\top \frac{\partial f_\theta}{\partial H}\bigg|_{H_t} dt - a_t^\top \frac{\partial g_\theta}{\partial H}\bigg|_{H_t} dW_t \tag{23}$$

5 Theoretical Convergence Rate Analysis

5.1 Motivation and Problem Statement

While both AMTF [6] and GATF [7] establish mean-square stability, the specific convergence rates remain unquantified. This gap limits principled comparison with classical methods and hyperparameter selection.

Problem 5.1 (Convergence Rate Characterization)**.** Establish rigorous convergence rates for GATF that:

1. Quantify dependence on architectural parameters (number of heads h, key dimension d_k)
2. Characterize the role of Lyapunov spectrum in convergence speed
3. Provide non-asymptotic (finite-time) bounds suitable for practical applications
4. Enable principled comparison with classical filtering methods

5.2 Mathematical Framework

Consider the GATF filtering density ρ_t^{GATF} approximating the true conditional density ρ_t^*. We decompose the total error:

$$\|\rho_t^{\text{GATF}} - \rho_t^*\|_{L^2} \leq \underbrace{\|\rho_t^{\text{GATF}} - \rho_t^{\text{disc}}\|_{L^2}}_{\text{approximation error}} + \underbrace{\|\rho_t^{\text{disc}} - \rho_t^*\|_{L^2}}_{\text{discretization error}} \tag{24}$$

Definition 5.2 (Approximation Class)**.** Let $\mathcal{F}_{h,d_k,L}$ denote the class of GATF architectures with h attention heads, key dimension d_k, and L Transformer layers. Define the approximation error:

$$\epsilon_{\text{approx}}(h, d_k, L) = \inf_{f \in \mathcal{F}_{h,d_k,L}} \sup_{\rho^* \in \mathcal{P}} \|f(\cdot) - \rho^*\|_{L^2} \tag{25}$$

where $\mathcal{P}$ is the class of filtering densities arising from systems satisfying standard regularity assumptions.

5.3 Main Theoretical Results

Theorem 5.3 (GATF Convergence Rate)**.** *Under assumptions:*

1. *The attractor $\mathcal{A}$ is compact with diameter D and box-counting dimension $d_{\mathcal{A}}$*
2. *The observation function h satisfies uniform observability with constant $\alpha > 0$*
3. *The Lyapunov spectrum satisfies $\lambda_1 > \lambda_2 > \cdots > \lambda_k$ with spectral gap $\Delta = \lambda_1 - \lambda_2$*
4. *Attention weights are bounded: $\|W^Q\|, \|W^K\|, \|W^V\| \leq M$*

the GATF approximation error satisfies:

$$\mathbb{E}\left[\|\rho_t^{\text{GATF}} - \rho_t^*\|_{L^2}^2\right] \leq \underbrace{C_1 \cdot \frac{d_{\mathcal{A}}}{h}}_{\textit{head complexity}} + \underbrace{C_2 \cdot \frac{1}{d_k}}_{\textit{key dimension}} + \underbrace{C_3 \cdot e^{-\gamma(\Delta,\alpha)t}}_{\textit{exponential decay}} + \underbrace{C_4 \cdot \frac{\log N}{N}}_{\textit{sample complexity}} \tag{26}$$

where $\gamma(\Delta, \alpha) = \min(\Delta, \alpha)$ is the effective convergence rate and N is the number of training samples.

Corollary 5.4 (Optimal Head Allocation)**.** *Optimal head allocation satisfies $h_j^* \propto d_j \cdot |\lambda_j|$ where $d_j = \dim(E_j)$ is the Oseledets subspace dimension and $|\lambda_j|$ the corresponding Lyapunov exponent magnitude.*

Corollary 5.5 (Comparison with EKF)**.** *For systems with positive maximal Lyapunov exponent $\lambda_1 > 0$, GATF achieves faster convergence than EKF when:*

$$h \cdot d_k > C \cdot \lambda_1/\alpha \tag{27}$$

That is, the architecture must be sufficiently expressive to capture unstable dynamics.

Conjecture 5.6 (Adaptive Optimality)**.** *The Lyapunov-weighted attention mechanism achieves near-optimal adaptation:*

$$\mathbb{E}\left[\|\rho_t^{\mathrm{GATF}} - \rho_t^*\|_{L^2}^2 \,\middle|\, \textit{regime change at } t_0\right] \leq C \cdot (t - t_0)^{-1} + \textit{lower order} \tag{28}$$

for $t > t_0 + O(\lambda_1^{-1})$, matching information-theoretic limits.

Convergence Rate Comparison

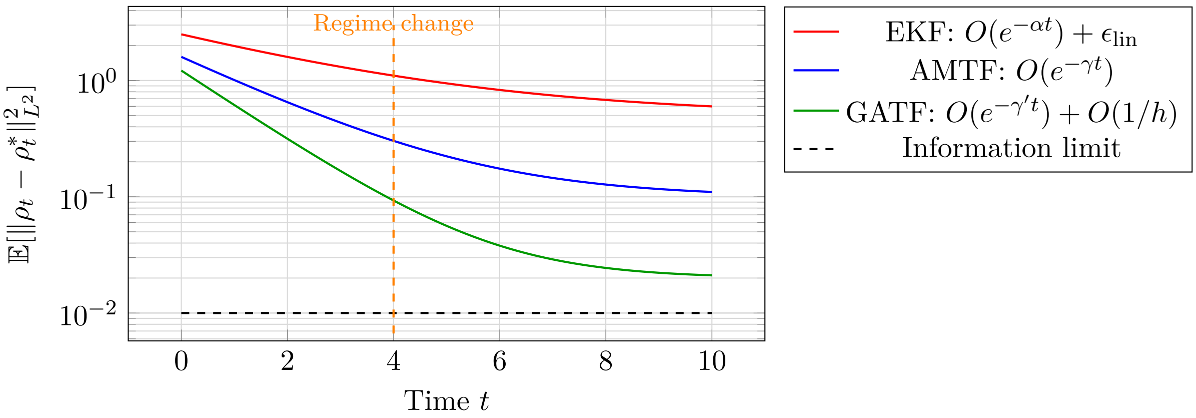


Figure 5: Theoretical convergence rate comparison.

GATF (green) achieves faster exponential decay than AMTF (blue) and EKF (red), with lower asymptotic error floor. The regime change at $t = 4$ causes temporary error increase, with GATF recovering fastest.

6 Uncertainty Quantification

6.1 Motivation and Problem Statement

The current GATF framework produces point estimates and filtering densities but does not explicitly account for uncertainty arising from: (1) finite training data, (2) approximate Lyapunov exponent estimation, (3) model misspecification, or (4) numerical discretization. For safety-critical applications, calibrated uncertainty estimates are essential.

Problem 6.1 (Geometry-Aware Uncertainty Quantification)**.** Develop uncertainty quantification methods for GATF that:

1. Provide calibrated predictive distributions respecting geometric structure
2. Decompose uncertainty into aleatoric (inherent) and epistemic (reducible) components
3. Propagate uncertainty through the Oseledets decomposition and attention mechanisms
4. Enable principled decision-making under uncertainty

6.2 Mathematical Framework

Definition 6.2 (GATF Uncertainty Decomposition)**.** The total predictive uncertainty is:

$$\mathrm{Var}[\rho_t^{\mathrm{GATF}}(x)] = \underbrace{\mathbb{E}_\theta\left[\mathrm{Var}[\rho_t|\theta]\right]}_{\text{aleatoric}} + \underbrace{\mathrm{Var}_\theta\left[\mathbb{E}[\rho_t|\theta]\right]}_{\text{epistemic}} \tag{29}$$

where the expectation and variance over θ capture uncertainty in GATF parameters.

Definition 6.3 (Oseledets-Structured Uncertainty)**.** We decompose epistemic uncertainty along Oseledets subspaces:

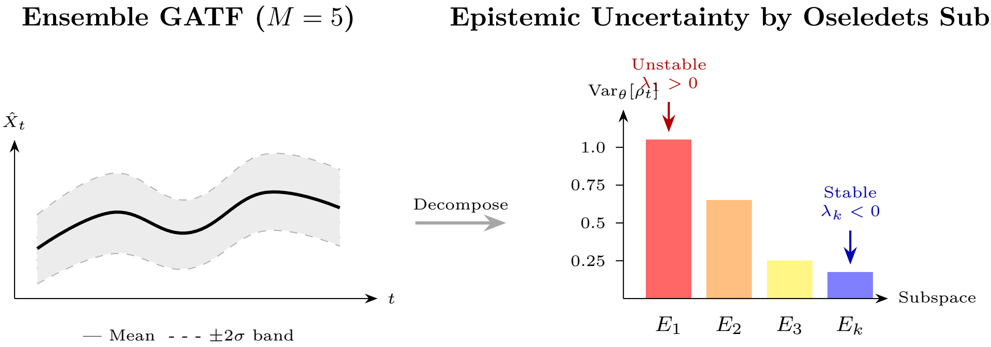
$$\mathrm{Var}_\theta[\rho_t] = \sum_{j=1}^{k} \Pi_t^{(j)} \mathrm{Var}_\theta[\rho_t] (\Pi_t^{(j)})^\top \tag{30}$$

This reveals which dynamical directions contribute most to uncertainty—typically unstable directions with $\lambda_j > 0$.

Definition 6.4 (Geometry-Aware Calibration Error)**.** Standard calibration metrics are adapted to respect Oseledets structure:

$$\mathrm{GACE} = \sum_{j=1}^{k} w_j \cdot \left| \mathbb{P}[\Pi_t^{(j)}(X_t - \hat{X}_t) \in C_\alpha^{(j)}] - (1-\alpha) \right| \tag{31}$$

where $C_\alpha^{(j)}$ is the α-credible region in subspace E_j and $w_j \propto |\lambda_j|$ weights by dynamical importance.



Ensemble GATF $(M = 5)$

Epistemic Uncertainty by Oseledets Subspace

Figure 6: Uncertainty quantification in GATF.

Left: Ensemble of $M = 5$ GATF models (colored curves) produce the ensemble mean (black) and $\pm 2\sigma$ uncertainty bands (gray dashed/shaded). Right: Epistemic uncertainty decomposed by Oseledets subspace shows larger uncertainty along unstable directions (E_1, $\lambda_1 > 0$) than stable directions (E_k, $\lambda_k < 0$).

7 Partially Observed Systems with Unknown Parameters

7.1 Motivation and Problem Statement

Many real-world chaotic systems have parameters that drift slowly, switch between discrete regimes, or are simply unknown. The current GATF assumes known parameters in the dynamics. Joint state-parameter estimation requires extending the framework to handle coupled fast (state) and slow (parameter) dynamics.

Problem 7.1 (Joint State-Parameter Estimation)**.** Extend GATF to jointly estimate:

1. The hidden state X_t evolving on the attractor $\mathcal{A}$
2. Unknown parameters $\theta_t \in \Theta$ that may drift or switch
3. The parameter-dependent Oseledets decomposition

while maintaining stability guarantees and regime detection capabilities.

7.2 Mathematical Framework

Definition 7.2 (Augmented State Space)**.** Define the augmented state:

$$\tilde{X}_t = \begin{pmatrix} X_t \\ \theta_t \end{pmatrix} \in \mathbb{R}^{n+p} \tag{32}$$

evolving according to:

$$d\tilde{X}_t = \begin{pmatrix} f(X_t, \theta_t) \\ g(\theta_t) \end{pmatrix} dt + \begin{pmatrix} \sigma(X_t) & 0 \\ 0 & \sigma_\theta \end{pmatrix} d \begin{pmatrix} W_t \\ W_t^\theta \end{pmatrix} \tag{33}$$

where $g(\theta_t)$ models slow parameter drift and σ_θ is the parameter diffusion.

Definition 7.3 (Parameter-Dependent Oseledets Decomposition)**.** The Oseledets splitting depends on the parameter:

$$T_x\mathcal{A}(\theta) = E_1(x;\theta) \oplus E_2(x;\theta) \oplus \cdots \oplus E_k(x;\theta) \tag{34}$$

with Lyapunov exponents $\lambda_j(\theta)$ varying smoothly in θ.

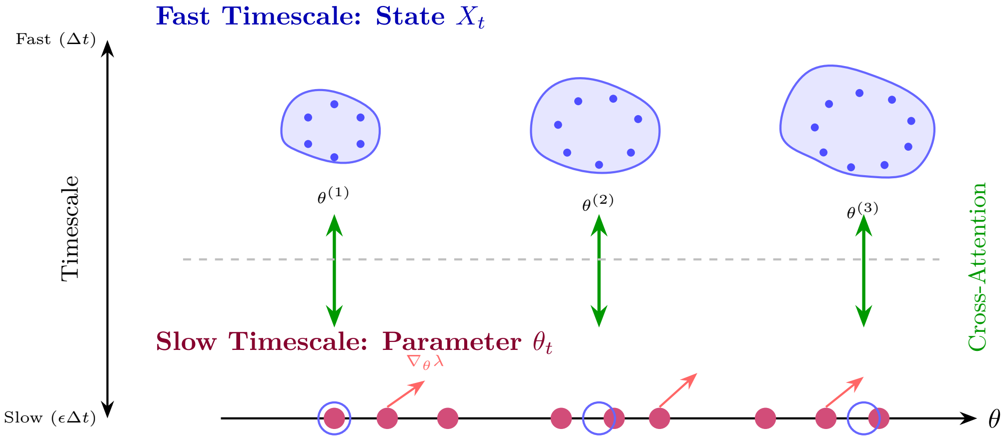


Figure 7: Hierarchical GATF for joint state-parameter estimation.

Fast timescale (top): State particles evolve on parameter-dependent attractors with different shapes for $\theta^{(1)}, \theta^{(2)}, \theta^{(3)}$. Slow timescale (bottom): Parameter particles (purple) updated less frequently along the parameter axis, with attention guided by Lyapunov gradients (red arrows). Bidirectional cross-attention (green) couples the two levels.

Theorem 7.4 (Hierarchical GATF Stability)**.** *Under timescale separation (parameter dynamics slower than state dynamics by factor $\epsilon \ll 1$), the hierarchical GATF satisfies:*

$$\mathbb{E}\left[\|\rho_t(X,\theta) - \rho_t^*(X,\theta)\|_{L^2}^2\right] \le C_X e^{-\gamma_X t} + C_\theta \epsilon^{-1} e^{-\gamma_\theta \epsilon t} \tag{35}$$

where γ_X is the state convergence rate and γ_θ depends on parameter identifiability.

8 Multi-Scale Temporal Modeling

8.1 Motivation and Problem Statement

Many chaotic systems exhibit separation of timescales—climate features fast atmospheric dynamics ($\sim$ days), slower oceanic circulation ($\sim$ months), and very slow ice sheet evolution ($\sim$ millennia). The Lyapunov spectrum naturally encodes these timescales through the exponential rates λ_j.

Problem 8.1 (Multi-Scale GATF Design)**.** Design a multi-scale GATF architecture that:

1. Operates separate attention mechanisms at different temporal resolutions
2. Uses the Lyapunov spectrum to guide timescale decomposition
3. Mediates information transfer between scales via cross-scale attention
4. Achieves computational efficiency by allocating resources according to timescale

8.2 Mathematical Framework

Definition 8.2 (Lyapunov Timescale)**.** Each Oseledets subspace E_j has characteristic timescale:

$$\tau_j = |\lambda_j|^{-1} \tag{36}$$

We partition subspaces into S scale groups:

$$\mathcal{S}_s = \{j : \tau_s^{\text{min}} \le \tau_j < \tau_s^{\text{max}}\}, \quad s = 1, \dots, S \tag{37}$$

Definition 8.3 (Scale-Specific GATF Update)**.** At scale s, operating with time step $\Delta t_s \approx \tau_s$:

$$\rho_{t+\Delta t_s}^{(s)}(x) = \rho_t^{(s)}(x) + \text{MHA}_s\left(\Pi^{(\mathcal{S}_s)}\rho_t\right) + \kappa_t^{(s)}\left(\nabla\rho_t^{(s)}\right)^\top \Pi^{(u,s)} \tag{38}$$

where $\Pi^{(\mathcal{S}_s)} = \sum_{j\in\mathcal{S}_s} \Pi^{(j)}$ projects onto the subspaces at scale s.

Definition 8.4 (Cross-Scale Attention)**.** Information flows between scales via:

$$\text{CrossScaleAttn}(Q^{(s)}, K^{(s')}, V^{(s')}) = \text{softmax}\left(\frac{Q^{(s)}(K^{(s')})^\top}{\sqrt{d_k}} + B_{ss'}\right)V^{(s')} \tag{39}$$

where $B_{ss'}$ is a learned bias encoding scale relationships.

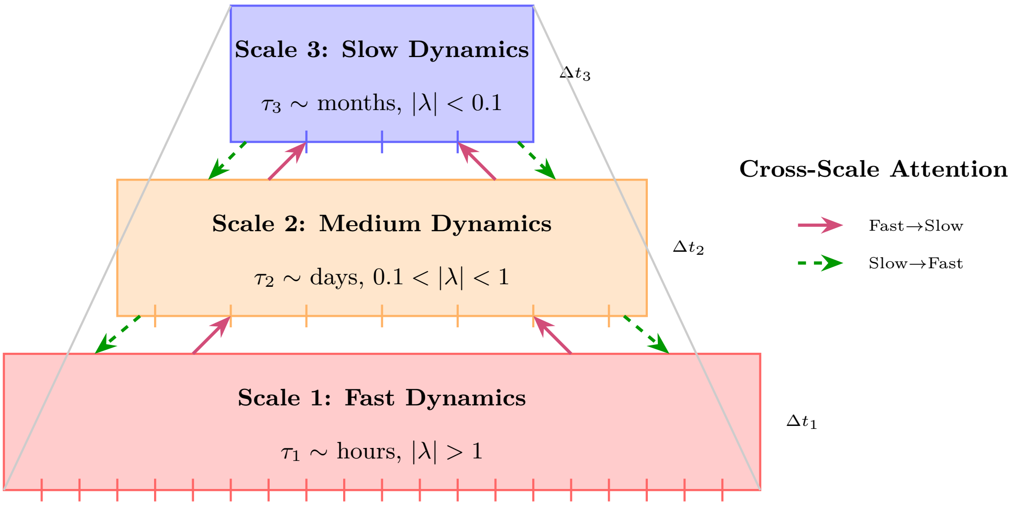


Figure 8: Multi-scale GATF architecture.

Three temporal scales form a pyramid, each operating at different update frequencies: fast (Δt_1, many ticks), medium (Δt_2), and slow (Δt_3, few ticks). Cross-scale attention mediates information flow: upward (purple solid) aggregates fast dynamics to slower scales; downward (green dashed) provides slow context to fast scales. Scale assignment is guided by Lyapunov exponent magnitudes $|\lambda|$.

Proposition 8.5 (Multi-Scale Speedup)**.** *Let $R = \tau_{\max}/\tau_{\min}$ be the timescale ratio. The multi-scale GATF achieves computational speedup:*

$$\frac{\textit{Cost(single-scale)}}{\textit{Cost(multi-scale)}} \approx \frac{R}{\sum_{s=1}^{S} |\mathcal{S}_s| \cdot R/2^{s-1}} \tag{40}$$

For typical chaotic systems with $S \approx \log_2 R$ scales, this yields speedup $O(\log R)$.

9 Experimental Validation

9.1 Benchmark Suite

Table 2: Proposed benchmark suite for GATF validation

Benchmark	**Dim**	**Type**	**Challenge**	**Key Metric**
Lorenz-63	3	ODE	Baseline chaotic	RMSE during regime transition
Lorenz-96	40+	Spatiotemporal	High-dimensional	Scalability, pattern correlation
Kuramoto-Sivashinsky	∞	PDE	Continuous spectrum	Spectral prediction skill
Double pendulum	4	Hamiltonian	Mixed regular/chaotic	Regime classification accuracy
Rössler system	3	ODE	Different topology	Topological feature preservation
ERA5 reanalysis	$\sim 10^6$	Real data	Operational	Forecast skill (ACC, RMSE)
S&P 500 volatility	~ 500	Financial	Regime switching	VaR exceedance, Sharpe ratio

9.2 Baseline Methods

1. **Classical Filtering:** Extended Kalman Filter (EKF), Unscented Kalman Filter (UKF), Ensemble Kalman Filter (EnKF) [13], Particle Filter (PF) [12].

2. **Deep Learning:** LSTM/GRU, Temporal Convolutional Networks (TCN), Standard Transformer [5], Neural ODE/SDE [14].

3. **Ablated GATF Variants:** GATF w/o Oseledets MHA, GATF w/o Lyapunov weighting, GATF w/o geometric PE, GATF w/o adaptive gain.

9.3 Evaluation Metrics

Filtering Accuracy:

- Root Mean Square Error: RMSE $= \sqrt{\frac{1}{T}\sum_t \|X_t - \hat{X}_t\|^2}$
- Continuous Ranked Probability Score (CRPS)
- Log-likelihood: $\sum_t \log p(X_t|Z_{0:t})$

Regime Detection:

- Detection latency (compare with $O(\lambda_1^{-1})$ bound)
- False positive rate
- F1 score for regime classification

Calibration:

- Expected Calibration Error (ECE)
- Geometry-Aware Calibration Error (GACE, Definition 6.4)

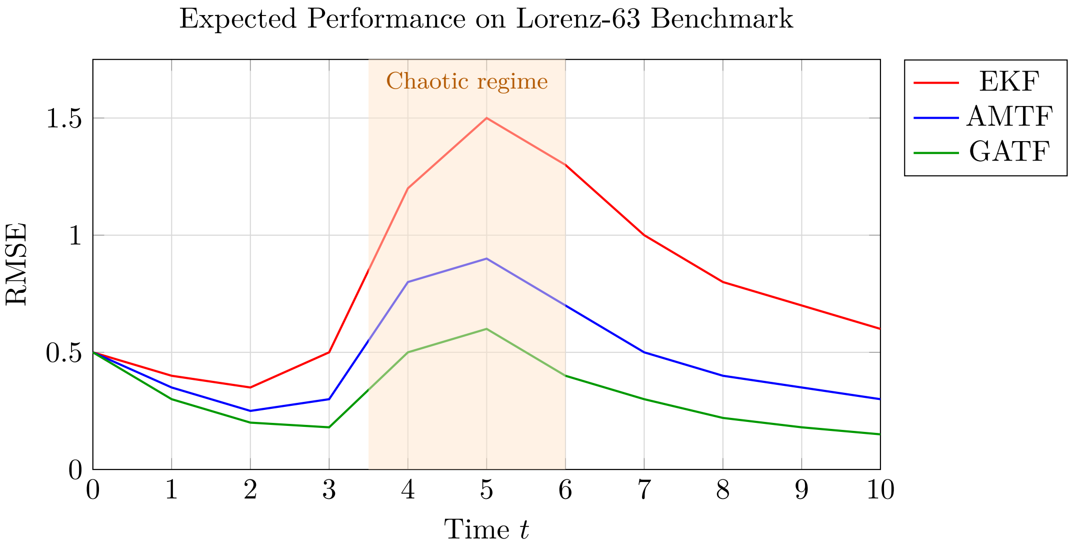


Figure 9: Expected comparative performance on Lorenz-63.

GATF (green) maintains lowest RMSE, particularly during the chaotic regime transition (orange region), due to geometry-aware attention and Lyapunov-weighted corrections.

10 Additional Research Directions

10.1 Equivariant GATF Architectures

When dynamical systems possess symmetries (rotational invariance in fluid flows, permutation invariance in particle systems), the GATF should respect these symmetries.

Definition 10.1 (Oseledets-Equivariant Attention)**.** For a symmetry group G acting on state space:

$$\text{MHA}_{\text{equiv}}(g \cdot Q, g \cdot K, g \cdot V) = g \cdot \text{MHA}_{\text{equiv}}(Q, K, V) \quad \forall g \in G \tag{41}$$

This requires the Oseledets projectors to transform appropriately:

$$\Pi^{(j)}(g \cdot x) = g \cdot \Pi^{(j)}(x) \cdot g^{-1} \tag{42}$$

10.2 Interpretability and Visualization

Definition 10.2 (Attention Attribution)**.** Attribute filtering decisions to specific Oseledets subspaces:

$$\text{Attribution}_j(X_t) = \sum_{\text{head } i \in E_j} \|\text{head}_i\|_F \cdot \lambda_j(t) \tag{43}$$

This enables visualization of which dynamical directions drive filtering corrections.

10.3 Online Learning and Adaptation

Definition 10.3 (Continual GATF)**.** Adapt GATF parameters online as the system evolves:

$$\theta_{t+1} = \theta_t - \eta_t \nabla_\theta \mathcal{L}(\rho_t^{\text{GATF}}, Z_t) \tag{44}$$

with learning rate η_t modulated by estimated Lyapunov exponents to ensure stability.

11 Conclusion

This research agenda presents a comprehensive path forward for the AMTF and GATF frameworks, spanning computational innovations (particle methods, sparse attention, continuous-time formulations), theoretical foundations (convergence rates, uncertainty quantification), architectural extensions (hierarchical estimation, multi-scale modeling), and empirical validation across diverse application domains.

The eight primary directions are interconnected—progress in one area will inform and accelerate others. Collectively, these efforts aim to transform the initial GATF framework into a mature methodology with:

- **Solid theoretical foundations**: Rigorous convergence guarantees and uncertainty quantification
- **Efficient implementations**: Scalable algorithms for high-dimensional chaotic systems
- **Demonstrated practical impact**: Validated performance across weather prediction, finance, robotics, and neuroscience applications

The synthesis of measure-theoretic filtering with modern deep learning architectures, guided by the geometric structure of chaotic dynamics through Oseledets decomposition, represents a promising paradigm for adaptive state estimation in complex real-world systems.

References

[1] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[2] H. J. Kushner, "Dynamical equations for optimal nonlinear filtering," *Journal of Differential Equations*, vol. 3, no. 2, pp. 179–190, 1967.

[3] M. Zakai, "On the optimal filtering of diffusion processes," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 11, no. 3, pp. 230–243, 1969.

[4] V. I. Oseledets, "A multiplicative ergodic theorem. Lyapunov characteristic numbers for dynamical systems," *Trudy Moskovskogo Matematicheskogo Obshchestva*, vol. 19, pp. 179–210, 1968.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[6] S. Ghosh, "A Novel Adaptive Filtration Architecture," 2026. Kolkata, India.

[7] S. Ghosh, "The Geometry-Aware Transformer Filter," 2026. Kolkata, India.

[8] P. E. Protter, *Stochastic Integration and Differential Equations*, 2nd ed. Springer, 2005.

[9] G. Benettin, L. Galgani, A. Giorgilli, and J.-M. Strelcyn, "Lyapunov characteristic exponents for smooth dynamical systems and for Hamiltonian systems," *Meccanica*, vol. 15, no. 1, pp. 9–30, 1980.

[10] L. Arnold, *Random Dynamical Systems*, Springer Monographs in Mathematics. Springer, 1998.

[11] A. Bain and D. Crisan, *Fundamentals of Stochastic Filtering*. Springer, 2009.

[12] P. Del Moral, *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications.* Springer, 2004.

[13] G. Evensen, “The ensemble Kalman filter: Theoretical formulation and practical implementation,” *Ocean Dynamics*, vol. 53, no. 4, pp. 343–367, 2003.

[14] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, “Neural ordinary differential equations,” in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[15] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations.* Springer, 1992.

Glossary

Adaptive Filtration

A time-varying sigma-algebra $\mathcal{G}_t$ that incorporates both observational and geometric information from the underlying dynamical system; in GATF, realized through attention over Oseledets subspaces.

Attention Mechanism

A function computing weighted sums of values based on query-key compatibility scores; enables selective focus on relevant information without fixed locality constraints.

Attractor

A compact invariant set $\mathcal{A} \subset \mathbb{R}^n$ to which trajectories converge; may exhibit chaotic (strange) dynamics with positive Lyapunov exponents.

Doléans-Dade Exponential

The stochastic exponential $\mathcal{E}_t(M)$ of a local martingale M, satisfying $d\mathcal{E}_t = \mathcal{E}_{t-}dM_t$; used for measure changes via Girsanov theorem.

Filtration

An increasing family of sigma-algebras $\{\mathcal{F}_t\}_{t\geq 0}$ representing information available up to time t; fundamental to conditional expectation and martingale theory.

GATF

Geometry-Aware Transformer Filter; the proposed unified architecture combining AMTF measure-theoretic foundations with Transformer attention mechanisms.

Girsanov Theorem

Fundamental result relating probability measures under which a process is Brownian motion with different drifts; enables adaptive measure construction in filtering.

Lyapunov Exponent

A quantity λ measuring the average exponential rate of divergence ($\lambda > 0$) or convergence ($\lambda < 0$) of nearby trajectories; characterizes chaos.

Multi-Head Attention

Parallel attention computations with different learned projections, enabling joint attention to information from different representation subspaces.

Oseledets Decomposition

The splitting $T_x\mathcal{A} = \bigoplus_i E_i(x)$ of tangent space into subspaces with distinct Lyapunov exponents, guaranteed by the Multiplicative Ergodic Theorem.

Particle Filter
A sequential Monte Carlo method representing the filtering distribution through weighted samples; enables non-parametric density estimation.

Positional Encoding
Injection of sequence position information into Transformer representations; in GATF, replaced with geometric encodings based on Oseledets structure.

Scaled Dot-Product Attention
The attention function $\mathrm{softmax}(QK^\top/\sqrt{d_k})V$ using dot products scaled by $\sqrt{d_k}$ for numerical stability.

Semimartingale
A stochastic process decomposable into a local martingale and finite variation process; the natural domain for stochastic integration.

Sparse Attention
Attention mechanisms with structured sparsity patterns reducing computational complexity from $O(n^2)$ to sub-quadratic.

Transformer
A neural architecture based entirely on attention mechanisms without recurrence; enables parallel sequence processing with global receptive fields.

Zakai Equation
A stochastic PDE governing the unnormalized conditional density in nonlinear filtering; the linear counterpart to the Kushner–Stratonovich equation.

The End