From Visual Vocabulary to Grammar: An Interpretable Paradigm for Scientific Discovery in Spatiotemporal Data

Zhen Yuan Yeo [1] **Timothée Levilly** [1] **Ying Chen Lim** [2] **Ngoc Thi Nguyen** [2 3] **Eun Ho Song** [4] **Kyeongmee Lee** [4] **Alexandre Thiéry** [1] **Jwa-Min Nam** [4] **N. Duane Loh** [2 3]

[1]*Department of Statistics and Data Science, National University of Singapore, Singapore 117546, Singapore* [2]*Department of Physics, National University of Singapore, Singapore 117551, Singapore* [3]*Centre for Bio-imaging Sciences, National University of Singapore, Singapore 117557, Singapore* [4]*Department of Chemistry, Seoul National University, Seoul 08826, South Korea*. Correspondence to: yeozy@nus.edu.sg.

1. Abstract

Scientific spatiotemporal datasets often contain rich and recurring structure, yet lack clear labels, observables, or governing models. While modern machine learning methods can extract patterns from such data, they frequently produce representations that are difficult to interpret and relate to scientific states or dynamical regimes.

We present an interpretable, data-driven paradigm for spatiotemporal motif discovery that represents complex dynamics using discrete spatial states and their temporal organization. The approach is based on the observation that many systems repeatedly visit a finite set of characteristic spatial configurations, which can be treated as a visual vocabulary (i.e. tokens), while system behavior is captured by transitions and sequences between these configurations (i.e. grammar).

The paradigm separates symbolic spatial description from temporal organization through four conceptual stages: **segmentation**, **featurization**, **tokenization**, and **motif learning**. This decomposition yields compact, interpretable summaries of spatiotemporal structure without assuming trajectories, labels, or mechanistic rules. By representing high-dimensional image sequences in symbolic form, the approach enables characterization, comparison, and classification of dynamical regimes in unfamiliar datasets, supporting discovery and hypothesis generation in noisy and heterogeneous scientific systems.

2. Introduction

Modern scientific experiments increasingly produce spatiotemporal datasets whose structure is rich but often unknown [1, 2, 3]. In many settings, there are no labels, no clear observables, and no established models to guide analysis [4, 5]. Commonly used workflows rely on domain-specific heuristics or black-box deep learning methods that obscure interpretation [6, 7]. As a result, scientists may recover statistically separable clusters or predictive signals, yet still lack a meaningful description of system states, their transitions, and underlying dynamical regimes [1, 4]. This motivates discovery-first approaches that prioritize interpretability [1, 8].

The challenge of spatiotemporal interpretation: Interpreting spatiotemporal data remains difficult despite advances in machine learning [1, 4, 5]. Mechanistic models require strong assumptions and predefined states [9, 10]. Trajectory-based analyses fail when the interacting objects cannot be definitely identified and tracked, or systems where objects interact as a collective (multi-particle interactions) [11]. Data-driven models capture patterns, but entangle space and time in opaque representations. The central bottleneck is not model capacity, but the ability to interpret their outputs in scientifically meaningful terms.

A language-inspired view of scientific dynamics: Many complex systems repeatedly visit a limited set of spatial configurations [12]. These configurations form a visual vocabulary of states. Temporal evolution can be represented as sequences of vocabulary elements. The statistics and ordering of these sequences define a grammar that characterizes system dynamics. This view reframes spatiotemporal analysis as learning vocabulary and grammar directly from data.

Spatiotemporal Motif paradigm: This work introduces an interpretable, data-driven paradigm for discovering structure in spatiotemporal datasets. We explicitly separate spatial description from temporal organization, shown in Figure 1. The proposed paradigm does not assume labels, trajectories, or mechanistic rules. Instead, it learns symbolic representations of recurrent spatial patterns and how they relate to one another over time, summarizing system behavior in human-interpretable terms.

3. Conceptual stages of the proposed framework

Segmentation identifies local regions where dynamics between interacting objects occur and restricts analysis to signal-bearing areas. This removes irrelevant background and defines the basic units of observation. The goal is not perfect object identification, but consistent localization of meaningful activity.

Featurization encodes local spatial structure into representations that are comparable across time and conditions. The use of symmetry-invariant features preserve meaningful geometry while suppressing nuisance variation. This stage defines the visual vocabulary from which states are learned.

Tokenization groups similar spatial representations into a finite set of discrete tokens. Each token corresponds to a recurring spatial state observed in the data. Discretization converts high-dimensional observations into interpretable symbols.

Motif learning analyzes how tokens evolve over

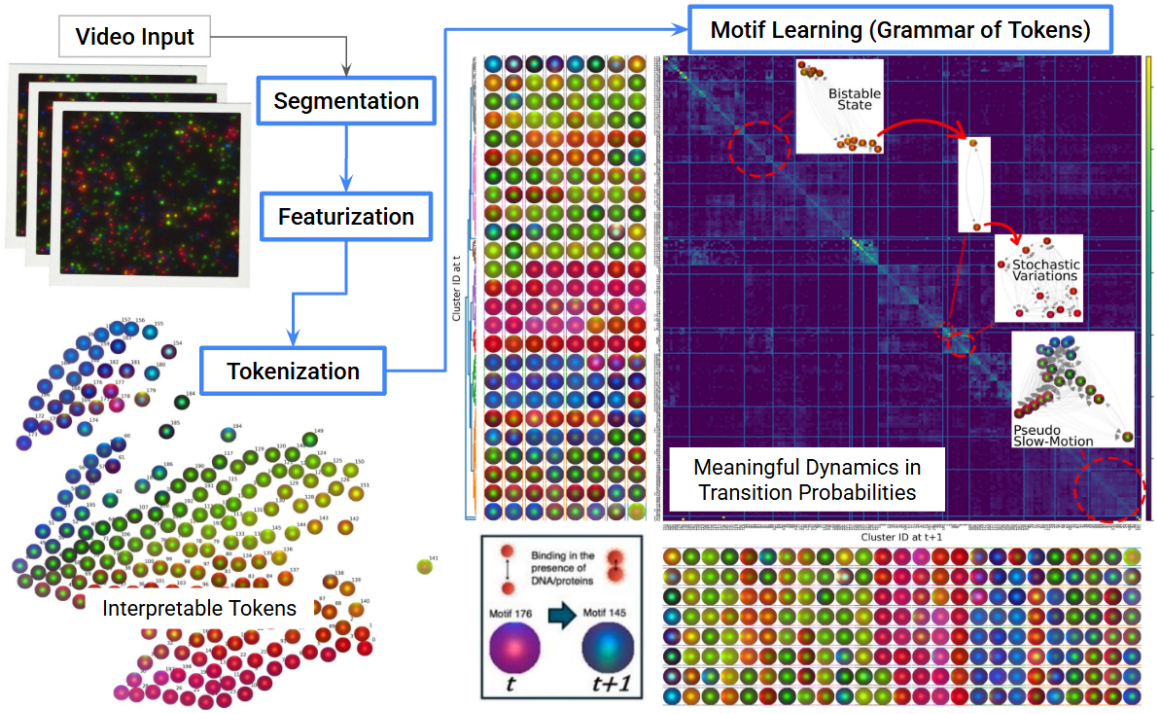

Fig. 1: Overview of the unsupervised machine learning discovery pipeline. Spatiotemporal video data [13] are transformed into interpretable spatial tokens via segmentation, featurization, and tokenization, followed by motif learning that captures transition statistics and recurrent sequences. The resulting token vocabulary and grammar summarize meaningful system dynamics across time.

time. Transition statistics between the tokens reveal persistent sequences, cycles, and dynamical regimes. These patterns form a grammar that compactly summarizes system dynamics.

4. Results and Discussion

Interpretable insights: The approach produces a vocabulary of spatial states and a grammar describing their temporal organization. Differences between experimental conditions appear as changes in token usage or probability of transition between tokens. The outputs support classification, comparison, and hypothesis generation without relying on opaque latent spaces.

Applicability and scope: The paradigm is suited for time-resolved datasets with unknown or novel dynamics. It is particularly useful in noisy, crowded, or heterogeneous systems where trajectories and identities of the interacting objects are unreliable. For example, the exemplary dataset (shown in Figure 1) was obtained from [13] and segmented with Nanopics [14]. The framework is designed for exploratory analysis and early-stage discovery.

Contrast with hand-crafted pipelines: The framework does not strongly rely on domain knowledge of a system, unlike traditional hand-crafted analysis where states and their transitions are defined a priori. It formalizes this process algorithmically by learning discrete states and their transition structure directly from data.

Limitations and assumptions: This approach assumes that recurring spatial structure exists in the data. It does not infer causality or mechanistic laws directly. Interpretability of the tokens and their grammar depends on meaningful segmentation and feature choices. The paradigm complements, rather than replaces, domain-specific modeling.

5. Conclusion

This work provides a language-based paradigm for interpreting scientific spatiotemporal data. By learning visual vocabulary and grammar directly from observations, it bridges data-driven analysis and scientific understanding. This approach enables systematic discovery of scientific insights in unfamiliar datasets and offers a foundation for interpretable models of complex dynamics.

Acknowledgments

The authors gratefully acknowledge the National University of Singapore (NUS), Institute for Digital Molecular Analytics and Science (IDMxS) and the Bio&Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. RS-2023-00222838).

References

[1] Peter Y Lu, Samuel Kim, and Marin Soljačić. Extracting interpretable physical parameters from spatiotemporal systems using unsupervised learning. *Phys. Rev. X.*, 10(3):031056,

September 2020.

[2] Ting-Ting Gao, Baruch Barzel, and Gang Yan. Learning interpretable dynamics of stochastic complex systems from experimental data. *Nat. Commun.*, 15(1):6029, July 2024.

[3] Karl Lapo, Sara M Ichinaga, and J Nathan Kutz. A method for unsupervised learning of coherent spatiotemporal patterns in multiscale data. *Proc. Natl. Acad. Sci. U. S. A.*, 122(7):e2415786122, February 2025.

[4] Xing Yan, Zhou Zang, Yize Jiang, Wenzhong Shi, Yushan Guo, Dan Li, Chuanfeng Zhao, and Letu Husi. A spatial-temporal interpretable deep learning model for improving interpretability and predictive accuracy of satellite-based PM2.5. *Environ. Pollut.*, 273(116459):116459, March 2021.

[5] Christian M Mulomba, Vogel M Kiketa, David M Kutangila, Pescie H K Mampuya, Junior N Mukenze, Landry M Kasunzi, Kyandoghere Kyamakya, Tasho Tashev, and Selain K Kasereka. Applying causal machine learning to spatiotemporal data analysis: An investigation of opportunities and challenges. *IEEE Access*, 13:141832–141857, 2025.

[6] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable AI: A review of machine learning interpretability methods. *Entropy (Basel)*, 23(1):18, December 2020.

[7] Chenxi Liao, Masataka Sawayama, and Bei Xiao. Unsupervised learning reveals interpretable latent representations for translucency perception. *PLoS Comput. Biol.*, 19(2):e1010878, February 2023.

[8] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, 1(5):206–215, May 2019.

[9] M C Cross and P C Hohenberg. Pattern formation outside of equilibrium. *Rev. Mod. Phys.*, 65(3):851, July 1993.

[10] Rebecca Hoyle. *Pattern Formation: An Introduction to Methods*. Cambridge University Press, Cambridge, England, March 2006.

[11] Oleksandr Chepizhko, Eduardo G Altmann, and Fernando Peruani. Optimal noise maximizes collective motion in heterogeneous media. *Phys. Rev. Lett.*, 110(23):238101, June 2013.

[12] Vasileios Basios and Dónal Mac Kernan. Symbolic dynamics, coarse graining and the monitoring of complex systems. *Int. J. Bifurcat. Chaos*, 21(12):3465–3475, December 2011.

[13] Sungi Kim, Jeong-Eun Park, Woosung Hwang, Jinyoung Seo, Young-Kwang Lee, Jae-Ho Hwang, and Jwa-Min Nam. Optokinetically encoded nanoprobe-based multiplexing strategy for MicroRNA profiling. *J. Am. Chem. Soc.*, 139(9):3558–3566, March 2017.

[14] Zhen Yuan Yeo, Eun Ho Song, Kyeongmee Lee, Jwa-Min Nam, and N Duane Loh. Biosensing using CNNs to detect noisy but persistent nanoparticle binding events on supported lipid bilayer systems. In *AI4X 2025 International Conference*, 2025.