# Explainable Long Video Understanding through Dynamic Motion Tokens and Temporal Causal Discovery

Henri M. Rose
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
rose403@buffalo.edu

**Abstract**

Long video understanding remains a central challenge in artificial intelligence due to the complexity of temporal dependencies, the volume of redundant visual data, and the opacity of deep learning models. This paper proposes a framework that integrates dynamic motion tokens and temporal causal discovery to produce interpretable analyses of extended video sequences. Dynamic motion tokens are learned representations that condense local motion patterns into discrete, semantically meaningful units while preserving temporal ordering. Temporal causal discovery then identifies directed causal relationships among these tokens across time, yielding a graph-based explanation of event progression. The system is designed to support explainability by design rather than post-hoc interpretation. The paper examines the structural trade-offs involved in token granularity, causal graph sparsity, and computational efficiency. It also discusses architectural choices for large-scale deployment, including distributed processing pipelines and memory-bounded inference. Robustness considerations are addressed, particularly concerning distribution shift and adversarial perturbations. Fairness and policy implications are explored in the context of video surveillance and content moderation applications. The framework is contrasted with existing methods such as SlowFast, VideoMAE, and transformer-based architectures, highlighting the benefits of causal explainability in high-stakes domains. The work concludes with a forward-looking discussion of governance and sustainability, arguing that transparent causal models are essential for accountable video analytics in societal infrastructure. This research contributes to the growing intersection of explainable artificial intelligence and temporal reasoning, offering a principled pathway toward trustworthy long video understanding.

**Keywords**

long video understanding, explainability, motion tokens, temporal causal discovery, dynamic representation, video analytics, causal inference, system architecture.

## 1. Introduction

The ability to interpret long video sequences is increasingly critical for applications ranging from autonomous driving and medical imaging to surveillance and content moderation. While modern deep learning models have achieved impressive accuracy on short clips, extending these successes to minute-long or hour-long videos introduces profound challenges in memory, computation, and interpretability. Traditional video architectures often rely on spatiotemporal convolutions or attention mechanisms that aggregate information over long horizons, yet the resulting representations remain opaque. Explainability becomes a pressing requirement when these systems are deployed in contexts where decisions must be justified to regulators, users, or affected communities. This paper proposes a framework that addresses

both the technical demands of long video understanding and the need for transparent, causal explanations.

The approach centers on two complementary innovations: dynamic motion tokens and temporal causal discovery. Dynamic motion tokens transform raw optical flow or frame differences into a compact sequence of discrete symbols, each representing a distinct motion pattern such as a person walking, a vehicle turning, or an object rotating. These tokens are learned in an unsupervised or weakly supervised manner and maintain temporal coherence, enabling the system to reason about events at multiple time scales. Temporal causal discovery then applies algorithms from causal inference to infer directed influences among tokens over time, producing a graph whose edges represent causal relationships. This graph serves as an explicit explanation of how past motion leads to future motion, allowing human analysts to verify, critique, and debug the model's reasoning.

The paper is organized as follows. Section 2 reviews related work in video understanding and explainable AI. Section 3 details the construction and properties of dynamic motion tokens. Section 4 describes temporal causal discovery techniques adapted for video. Section 5 discusses system architecture and deployment trade-offs. Section 6 addresses robustness, fairness, and policy implications. Section 7 concludes with reflections on governance and future research directions.

## 2. Related Work and Background

Long video understanding has been tackled through several families of models. Spatiotemporal convolutional networks such as I3D and SlowFast process clips by stacking 3D convolutions or by sampling frames at different rates [1, 2]. These models achieve strong performance on benchmarks like Kinetics and Something-Something but typically limit temporal context to a few seconds due to memory constraints. Transformer-based architectures, including ViViT and VideoMAE, extend the receptive field through self-attention mechanisms, yet they still struggle with hour-long videos because quadratic attention complexity and fixed-length token sequences do not scale gracefully [3, 4]. Recent efforts such as HY-Himmel introduce hierarchical interleaved motion encoding to handle longer sequences [7], but the resulting representations remain largely uninterpretable.

Explainability in video models has often relied on post-hoc saliency maps or attention visualization [5]. Methods like Grad-CAM highlight regions that influence a classification decision, but these heatmaps provide only spatial attribution without temporal structure [6]. Causal explanations are increasingly recognized as more faithful to human reasoning. Temporal causal discovery, originally developed for time-series analysis in fields like economics and neuroscience, has been adapted to video by treating token sequences as multivariate time series [8, 9, 10]. However, the application of such methods to long video has been limited by the need to discretize continuous motion into meaningful tokens without losing temporal resolution. This gap motivates the present work.

## 3. Dynamic Motion Tokens for Explainable Representation

Dynamic motion tokens are the foundational layer of the proposed framework. Instead of operating on raw pixels or dense flow fields, the system first extracts motion features from overlapping temporal windows. These features are then quantized into a finite set of tokens using a learned vocabulary. Each token corresponds to a prototypical motion pattern that can be visually interpreted by human annotators. For example, a token might represent a person raising their hand, a car accelerating, or a flag waving. The tokenization process is dynamic in

the sense that the vocabulary adapts to the domain of interest through unsupervised clustering or through a vector quantization autoencoder trained on unlabeled video data.

The choice of token granularity introduces a fundamental trade-off between expressiveness and interpretability. A small vocabulary yields coarse tokens that are easy to label but may miss subtle distinctions needed for accurate causal inference. A large vocabulary produces fine-grained tokens that capture nuanced motion but become harder for humans to map to semantic concepts. The optimal granularity depends on the application. In surveillance scenarios, coarse tokens such as “person walking” and “person standing” may suffice, whereas in surgical video analysis, finer distinctions like “tool rotating clockwise” are necessary. The system allows the vocabulary size to be set as a hyperparameter, and the resulting tokens can be audited for semantic consistency.

Temporal coherence is preserved by encoding the order and duration of tokens. A long video is thus represented as a sequence of token indices with associated time stamps. This representation is inherently interpretable: an analyst can inspect the token sequence to understand what motions occurred and when. Moreover, the token sequence is much shorter than the original video, enabling efficient processing of hour-long recordings. The compression ratio is determined by the temporal window length and the stride. Shorter windows yield more tokens and higher temporal resolution but increase computational cost. This trade-off must be managed carefully for real-time applications.

## 4. Temporal Causal Discovery in Video Sequences

Once a video is tokenized, the next step is to infer causal relationships among the tokens. Temporal causal discovery treats the token sequence as a multivariate time series where each variable corresponds to a token type and each time step indicates the presence or absence of that token. The goal is to learn a directed acyclic graph (DAG) or a directed graph with possible cycles representing causal influences over time. Algorithms such as PC, FCI, or Greedy Equivalence Search can be applied, but they are designed for continuous data and may require adaptation for discrete token sequences [8, 11]. A more scalable approach uses variants of Granger causality or structural equation models regularized with sparsity constraints [12].

The resulting causal graph provides an explicit explanation of event dynamics. For instance, the graph might reveal that a token representing “person opening door” is consistently followed by “person entering room,” and that this relationship holds across multiple videos. Such patterns can be validated by human experts. The sparsity of the graph is a critical design parameter. If too many edges are retained, the graph becomes cluttered and hard to read; if too few, important causal links may be missed. Regularization techniques, such as lasso or thresholding based on confidence intervals, help balance this trade-off.

Causal discovery in long videos also faces challenges from non-stationarity and unknown confounders. Lighting changes, camera cuts, or domain shifts can introduce spurious correlations that the discovery algorithm might mistake for causal links. Robustness can be improved by incorporating domain knowledge as prior constraints or by aggregating graphs across multiple videos to filter out chance dependencies [10]. Moreover, the discovered graph can be used to generate counterfactual explanations: given an observed token sequence, what would have happened if a particular token had been absent? This capability is especially valuable for high-stakes decision support.

## 5. System Architecture and Deployment Considerations

Integrating dynamic motion tokens and temporal causal discovery into a deployable system requires careful architectural design. The typical pipeline consists of three stages: tokenization, causal discovery, and explanation generation. Tokenization is the most computationally intensive stage because it involves extracting motion features and applying quantization. For long videos, this stage can be parallelized by dividing the video into overlapping segments and processing them on separate compute nodes. The segments are then reassembled into a continuous token sequence. This map-reduce approach scales linearly with video length, assuming sufficient hardware.

Causal discovery operates on the token sequence and can be executed on a single machine if the token vocabulary is not too large. Vocabulary sizes in the range of 100 to 1000 tokens are manageable, but scaling to 10,000 tokens may require distributed graph inference algorithms. Memory constraints are another concern: the causal graph is of order O(V^2) where V is the number of token types, so for large vocabularies the graph becomes dense and impractical to store. A practical solution is to learn a sparse graph incrementally using online or streaming causal discovery methods [13]. Alternatively, the graph can be approximated with a fixed topology, such as a Markov chain of order k, which reduces complexity.

Deployment in real-world settings also involves latency and throughput requirements. For real-time surveillance, tokenization must occur at frame rate, which necessitates optimized inference engines on edge devices. Causal discovery can be performed periodically offline to update the graph, while online inference uses the precomputed graph to explain new tokens quickly. This hybrid approach balances responsiveness with accuracy. The system must also handle multiple video streams simultaneously, which imposes additional load balancing and data synchronization challenges.

Governance of such a system includes logging and audit trails. Every explanation is accompanied by the causal graph and the token sequence that produced it, enabling post-hoc accountability. The system can be configured to flag high-uncertainty inferences for human review, reducing the risk of false causal claims. Sustainability considerations arise from the energy consumption of large-scale video processing. Efficient tokenization algorithms, perhaps using lightweight neural networks or optical flow approximations, can mitigate the carbon footprint.

## 6. Robustness, Fairness, and Policy Implications

Any deployed video understanding system must contend with robustness to distribution shift. The causal discovery approach offers a natural advantage: causal relationships, once learned, are often more invariant under changing environmental conditions than mere correlations. However, if the training videos are biased toward certain lighting, camera angles, or demographic groups, the discovered graphs may encode spurious causal links. For example, a token representing “person walking” might be incorrectly linked to a token representing “car approaching” if the two events are correlated in the training data due to a particular intersection geometry. Domain adaptation techniques, such as adversarial training or invariant risk minimization, can help the model learn causal structures that generalize across domains [14, 15].

Fairness concerns are especially acute in video analytics used for law enforcement or hiring. Dynamic motion tokens might be sensitive to body shape, gait, or clothing, leading to biased causal graphs that treat certain demographic groups differently. The explainability of the framework provides a mechanism for fairness auditing: regulators can inspect the token

dictionary and the causal graph to identify whether any tokens are disproportionately associated with protected attributes. If such associations are found, the vocabulary can be re-clustered or the causal discovery algorithm can be constrained to ignore attributes that are not causally relevant. Additionally, the system can be required to produce counterfactual explanations for each decision, which can be tested for group fairness.

Policy implications extend to privacy. Token sequences are abstractions that discard pixel-level details, yet they may still encode sensitive information about behaviors. Differential privacy mechanisms can be applied during tokenization to prevent re-identification. Furthermore, causal graphs can be designed to reveal only high-level event patterns rather than individual activities, aligning with privacy-by-design principles. Regulatory frameworks such as the European Union's Artificial Intelligence Act may mandate explainability for high-risk AI systems, and the proposed framework directly satisfies such requirements by providing transparent, causal explanations.

The deployment of such systems in public spaces also raises questions of power and accountability. Who decides which motion patterns are encoded into tokens? How are causal links validated? A participatory governance model that includes domain experts, community representatives, and ethicists can help ensure that the system serves the public interest. The technical flexibility of the framework—allowing adjustments to vocabulary size, sparsity, and sensitivity—supports iterative refinement through stakeholder feedback.

## 7. Conclusion

This paper has presented a framework for explainable long video understanding that combines dynamic motion tokens with temporal causal discovery. The approach addresses the twin challenges of scalability and transparency by transforming raw video into discrete, interpretable symbols and then inferring causal structures among them. Through detailed discussion of architectural trade-offs, robustness, fairness, and policy, the paper demonstrates that causal explainability is not merely an added feature but a foundational design principle for high-stakes video analytics.

Future work should explore end-to-end learning of motion tokens jointly with causal discovery, perhaps using variational autoencoders or neural causal models. The integration of human feedback loops, where analysts can correct erroneous causal edges, could further improve trust. Sustainability remains an open challenge; efficient tokenization with low energy overhead is essential for widespread adoption. Finally, cross-modal extensions that incorporate audio or text descriptions could enrich the causal graph and provide multisensory explanations. As video permeates every facet of societal infrastructure, the need for systems that can be both understood and held accountable will only grow. The framework proposed here offers a principled step toward that goal.

## References

1. Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the Kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6299–6308).

2. Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). SlowFast networks for video recognition. In Proceedings of the IEEE International Conference on Computer Vision (pp. 6202–6211).

3. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). ViViT: A video vision transformer. In Proceedings of the IEEE International Conference on Computer Vision (pp. 6836–6846).

4. Tong, Z., Song, Y., Wang, J., & Wang, L. (2022). VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In Advances in Neural Information Processing Systems (Vol. 35, pp. 10078–10093).

5. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (pp. 618–626).

6. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In International Conference on Machine Learning (pp. 2668–2677).

7. Jin, H., Yi, H., Zhao, W., Luo, J., Ye, S., Guan, Z., ... & Yu, T. (2026). HY-Himmel Technical Report: Hierarchical Interleaved Multi-stream Motion Encoding for Long Video Understanding. arXiv preprint arXiv:2605.08158.

8. Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., & Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. Science Advances, 5(11), eaau4996.

9. Tank, A., Covert, I., Foti, N., Shojaie, A., & Fox, E. (2022). Neural Granger causality for nonlinear time series. Journal of Machine Learning Research, 23(1), 4560–4620.

10. Peters, J., Janzing, D., & Schölkopf, B. (2017). Elements of causal inference: Foundations and learning algorithms. MIT Press.

11. Spirtes, P., Glymour, C., & Scheines, R. (2000). Causation, prediction, and search (2nd ed.). MIT Press.

12. Shojaie, A., & Michailidis, G. (2010). Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. Biometrika, 97(3), 519–538.

13. Zhu, P., Zhao, S., Deng, H., & Han, F. (2025). Attentive radiate graph for pedestrian trajectory prediction in disconnected manifolds. IEEE Transactions on Intelligent Transportation Systems.

14. Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant risk minimization. arXiv preprint arXiv:1907.02893.

15. Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. Proceedings of the IEEE, 109(5), 612–634.

16. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.

17. Lakkaraju, H., Arsov, N., & Leskovec, J. (2020). Interpretable machine learning: A path to more trustworthy AI. Nature Machine Intelligence, 2(7), 361–362.

18. Pearl, J. (2009). Causality: Models, reasoning, and inference (2nd ed.). Cambridge University Press.

19. Goyal, R., Kahou, S. E., Michalski, V., Pal, C., & Bengio, Y. (2019). The “something something” video database for learning and evaluating visual common sense. In Proceedings of the IEEE International Conference on Computer Vision (pp. 5842–5850).

20. Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7794–7803).