Measurement-First World Models, Takens-Based Transformers, and Trajectory Computation

Kevin R. Haylett

June 2026
Pensée

Abstract

These notes examine a measurement-first approach to world modelling, contrasting it with current machine-learning framings based on latent manifolds, image patches, diffusion landscapes, and static representations. The central claim is that topology should not be treated as the primitive object. Rather, topology emerges from ordered measurement, memory, and trajectory reconstruction. In this framing, an attention-based model is not limited to language. It may be understood more generally as a finite trajectory engine, capable of operating over symbolic, visual, sensory, or experimental time-series data, provided the architecture preserves the temporal structure of measurement. A Takens-Based Transformer, suitably modified, may therefore provide a route toward video, CCD, signal, and multimodal world modelling in which each frame or measurement field is treated as a state within a dynamical trajectory rather than as a static image patch.

1 Starting Point: The Limitation of Static-Vector Thinking

A common criticism of large language models is that they operate over words or word-like vectors and therefore cannot possess a true world model. This criticism is partly correct but often too shallow.

The weakness is not that language models are merely "word machines." The deeper point is that any serious model of intelligence must be understood as a dynamical system. A language model does not simply manipulate static word vectors. Its internal representations evolve across a sequence. Each token is recontextualised by the prior trajectory. The attention mechanism does not merely retrieve fixed meanings; it constructs relations across an unfolding symbolic path.

This means an LLM may be better understood as a finite symbolic trajectory engine. Its native domain is language, but the deeper computational principle is trajectory continuation under constraint.

Yann LeCun's JEPA/world-model programme is relevant here because it explicitly shifts attention toward video, physical prediction, and abstract latent representation. Meta describes V-JEPA as predicting missing or masked parts of video in an abstract representation space, rather than directly predicting pixels. V-JEPA 2 is presented as a self-supervised foundation world model trained on video, aimed at visual understanding, prediction, and robot-control capabilities.

However, the concern is that this still often remains inside a stochastic machine-learning frame. It seeks better latent representations and world models, but may not fully articulate the prior dynamical question: how does measured temporal experience construct the topology on which such a world model depends?

2 Measurement Before Topology

The central inversion is this:

> **Measurement comes first.**
> Time series comes second.
> Memory binds the sequence.
> **Topology is reconstructed** from the measured trajectory.

In much current AI language, one often finds an implicit order such as:

data → representation → latent manifold → prediction

The measurement-first order is different:

measurement → ordered time series → memory → phase-space reconstruction → emergent topology → continuation / prediction / action

This is not just a philosophical distinction. It changes the architecture.

If the topology is assumed first, the model searches for a useful latent space. If measurement is primary, the model must preserve the structure of observation over time. The topology is then an achieved reconstruction, not a pre-existing container.

This is where Takens-style thinking becomes important. Delay-coordinate embedding is used in nonlinear dynamics to reconstruct aspects of a system's phase space from observations over time. Standard accounts of Takens' theorem describe it as giving conditions under which the dynamics of a system can be reconstructed from time-delayed observations.

The key point for this research trajectory is that the time series is not incidental. It is the carrier of the system.

3 Why "Patches" Matters

The word "patches" reveals an important architectural inheritance from computer vision. In the Vision Transformer, an image is split into fixed-size patches, each patch is linearly embedded, positional information is added, and the resulting sequence is passed into a transformer.

This is effective, but it carries a particular assumption: the input is treated as a spatial field divided into local regions.

A patch is therefore like a patch of grass: a local region cut from a larger surface.

This becomes problematic when applied to time-series reconstruction. If a delay embedding or Hankel construction is first converted into a two-dimensional image-like object and then divided into patches, the model risks treating reconstructed dynamics as a spatial image rather than as a temporal state reconstruction.

The distinction is:

time series → delay embedding → image patches → standard self-attention

versus

time series → delay-coordinate state reconstruction → attention over dynamical states

The first approach uses Takens-like structure as a preprocessing device. The second makes the reconstructed temporal state the object of attention.

The sharper criticism is:

They have not made attention Takens-based. They have made Takens image-based, then applied patch attention.

This distinction matters because the Takens embedding already carries time. Sequentially patching or rasterising the embedding may lose, blur, or spatialise some of the temporal information that the embedding was meant to preserve.

4 Video Frames as Measured Field Tokens

A possible extension of the Takens-Based Transformer is to treat video not as a sequence of images to be patchified, but as a sequence of measured fields.

A CCD or video frame is not merely a picture. It is a finite measured field produced by an instrument. A video sequence is therefore a measured trajectory:

$\text{frame}_1 \to \text{frame}_2 \to \text{frame}_3 \to \text{frame}_4 \to \ldots$

Each frame may be treated as an image-token or field-token, but not in the same sense as a word token or image patch. It is a full measured state, or a compressed representation of such a state, situated within a temporal sequence.

The model input could therefore be:

video frame + support channels → trajectory embedding → attention over reconstructed dynamical states

The support channels may include:

time index, exposure, sensor geometry, camera calibration, optical flow, depth estimate, uncertainty, provenance, motion estimate, object hypothesis, task instruction, language prompt, and memory state.

The important point is that these channels should not simply be flattened into anonymous feature space. They have different functional roles. Some carry measurement. Some carry instrument conditions. Some carry uncertainty. Some carry symbolic instruction. Some carry memory or history.

In this view, text remains essential, but not because text is the whole model. Text acts as an entry channel, task channel, naming channel, indexing channel, and communicative bridge. The video or CCD sequence supplies the measured trajectory.

5 Diffusion Models as Potential-Landscape Traversal

Diffusion models have become central to modern image and video generation. Denoising Diffusion Probabilistic Models describe generation as a learned denoising process connected to nonequilibrium thermodynamics and denoising score matching. Latent diffusion models move this process into a learned latent space, using sequential denoising autoencoders and allowing conditioning through mechanisms such as cross-attention. Recent video-generation systems also use patch-like latent representations; OpenAI's Sora technical report describes extracting spacetime patches from compressed video, with those patches acting as transformer tokens.

From the present perspective, diffusion models may be interpreted as traversing a learned landscape of potential. They begin from noise or partial structure and iteratively unfold toward a coherent image or video state.

This is dynamical in a broad sense:

noise → partial structure → stronger structure → coherent generated field

For video:

noise field → latent motion possibilities → stabilised spatiotemporal sequence

This is powerful, but it is not the same as measurement-first reconstruction. Diffusion models learn landscapes of plausible generation. A Takens/TBT-style system would aim to reconstruct landscapes of measured continuation.

The distinction can be stated as:

Diffusion builds a landscape of potential. TBT reconstructs a landscape of measured continuation.

A combined architecture might be especially interesting: diffusion for generative possibility, TBT for measurement-constrained trajectory reconstruction, and language for symbolic task control.

6 Core Conceptual Distinctions

The emerging distinction is not LLMs versus world models, or diffusion versus transformers. It is a deeper distinction between different kinds of trajectory.

- LLMs continue symbolic trajectories.
- Diffusion models traverse generative potential landscapes.
- Video models learn spatiotemporal continuations.
- Takens/TBT models reconstruct dynamical state from measurement history.

A genuine world model may need all of these, but ordered correctly.
The proposed order is:

1. Measurement
2. Time series
3. Memory
4. Delay/state reconstruction
5. Emergent topology
6. Trajectory continuation
7. Generative possibility
8. Symbolic control and exposition
9. Action or intervention

This places measurement and memory prior to topology.

7 Possible Research Paths

7.1 Video-TBT Prototype

A first research path would be to build a modified Takens-Based Transformer for simple video sequences.

The aim would not be high-quality video generation. The aim would be dynamical reconstruction.

Suitable test cases might include: a dot moving left to right, a ball bouncing under gravity, an object disappearing behind an occluder, a car approaching or receding, a rotating object, or a simple two-body interaction.

The model should be tested on whether it preserves temporal structure, predicts continuation, handles occlusion, and distinguishes plausible continuation from merely image-like continuation.

The input should be full-frame or compressed-frame tokens, with support channels attached.

7.2 Support-Channel Formalism

A second research path is to formalise support channels.

Rather than treating all inputs as ordinary features, the architecture would distinguish between:

primary measured field, timing channel, provenance channel, uncertainty channel, instrument channel, task channel, language channel, memory channel, and action channel.

This may become a defining feature of the architecture. The support channels do not merely add information. They constrain admissible trajectory evolution.

7.3 Patch-Based Transformer versus Trajectory-Based Transformer

A third research path is comparative.

One could compare:

Vision Transformer patch-token processing, spacetime-patch video diffusion, ordinary time-series transformer, and Takens-Based Transformer.

The question would be: Which architecture best preserves dynamical information?

The tests should not be ordinary classification accuracy alone. They should include trajectory continuation, occlusion recovery, phase consistency, time reversal sensitivity, sensor perturbation, and physical plausibility.

7.4 Hybrid Diffusion-TBT Architecture

A fourth research path is a hybrid model.

The TBT component reconstructs the measured dynamical trajectory. The diffusion component generates possible futures. The support-channel system constrains which futures are admissible.

This would allow the model to distinguish between: what is possible, what is plausible, what is measured, what is admissible, and what is merely generatively attractive.

That distinction is central to a measurement-first world model.

7.5 CCD / Instrument-Based World Modelling

A fifth research path is to begin not with internet video, but with controlled CCD sequences.

This would align more closely with the measurement-first framing. The CCD is a parallel measurement instrument. Each frame is a finite measured field. The sequence of frames forms a directly controlled experimental trajectory.

Possible experiments: moving light source, two-source interference-like patterns, near-field LED/CCD interactions, rotating object under fixed illumination, shadow motion, reflection, refraction, or simple mechanical oscillation.

The advantage is that provenance, instrument geometry, exposure, and uncertainty can be explicitly recorded as support channels.

7.6 Language as Entry Channel Rather Than Whole Model

A sixth research path is to integrate language carefully.

The text prompt should not dominate the measured field. Instead, text should function as an instruction and indexing channel.

Examples: “track the moving object,” “predict the next frame,” “estimate whether the object is approaching,” “identify whether the trajectory changes,” “describe the measured continuation.”

In this arrangement, language does not replace the world model. It interrogates and conditions the trajectory model.

7.7 Mathematical Development: From Embedding to Functional Symbolic Trajectory

A final research path is conceptual and mathematical.

The Takens embedding may be treated as one important mechanism within a wider theory of Functional Symbolic Trajectories.

The deeper object is not the embedding itself. The deeper object is the finite trajectory through measured, symbolic, and remembered states.

This suggests a hierarchy:

finite measurement → symbolic registration → ordered trajectory → memory → reconstruction → topology → continuation → exposition

This may provide a bridge between the Takens-Based Transformer, Geofinitism, measured numbers, support channels, uncertainty, provenance, and admissibility.

8 Provisional Research Claim

A world model should not begin from a latent topology. It should begin from finite measurement.

The topology of the world, for the model, is reconstructed from ordered measurement, memory, and trajectory traversal.

Attention-based architectures need not be restricted to language. They may operate over any suitably represented measured trajectory. But when moving beyond language, architecture matters. A video frame, CCD image, or sensor field should not automatically be reduced to a patch-token image grammar. It may instead be treated as a measured state within a temporal trajectory.

Diffusion models show that modern AI can traverse landscapes of potential. Takens-Based Transformers suggest a different but complementary possibility: reconstructing landscapes of measured continuation.

The research opportunity is therefore to build architectures in which symbolic trajectory, measured trajectory, support channels, memory, uncertainty, and generative possibility are brought into one nonlinear dynamical framework.

9 Open Questions and Directions for the Research Program

The following questions and suggestions are offered to help guide concrete implementation, prototyping, and theoretical refinement of the ideas presented above.

1. **Implementation of delay embedding in Video-TBT** How should delay-coordinate embedding be realised when the observables are high-dimensional video frames or CCD fields? Should the embedding operate directly on sequences of (compressed) frame tokens, on learned per-frame embeddings, or via a hierarchical approach (coarse temporal embedding + local spatial refinement)? What embedding dimension and delay parameters are appropriate for image-like data?

2. **Integration of heterogeneous support channels** How can support channels (timing, provenance, uncertainty, instrument geometry, task instructions, etc.) be incorporated without disrupting the dynamical structure of the reconstructed trajectory? Should they themselves be delay-embedded, or treated as conditioning signals that constrain the evolution operator?

3. **Evaluation metrics for dynamical fidelity** Beyond standard accuracy or generation quality, what quantitative diagnostics best assess whether a model preserves dynamical structure? Suggested candidates include: trajectory continuation error, occlusion recovery, phase-space consistency (e.g., Lyapunov exponents or recurrence plots), time-reversal sensitivity, robustness to sensor perturbation, and physical plausibility under known dynamics (e.g., conservation laws or Newtonian motion).

4. **Scaling to real video and CCD data** Delay embedding works elegantly on low-dimensional attractors. What strategies (learned embeddings, multi-scale or hierarchical delay structures, or hybrid latent + explicit delay approaches) are needed to make the framework tractable for realistic video or high-resolution CCD sequences while retaining the measurement-first character?

5. **Hybrid Diffusion–TBT architectures** What is a concrete design for combining the two paradigms? One possibility: the TBT maintains the reconstructed measured trajectory and admissible region defined by support channels, while a diffusion component proposes futures *conditioned on* that admissible region. How should training objectives and inference procedures be structured so that the model can distinguish measured continuation from generatively plausible futures?

6. **Relationship to existing dynamical-systems approaches in machine learning** How does the proposed TBT framework relate to (and potentially extend) recent work that interprets transformers and state-space models through the lens of delay embeddings, Koopman operators, or Neural ODEs? Are there opportunities for hybridisation with reservoir computing or physics-informed architectures?

7. **Controlled experimental testbeds** Beyond the simple synthetic cases already proposed (bouncing ball, rotating objects, occlusion), what minimal real-world CCD or sensor experiments would provide the strongest early validation of the measurement-first claim while keeping provenance and calibration fully known?

These questions are intended to support the development of working prototypes and to sharpen the distinctions between measurement-first trajectory reconstruction and existing generative or latent approaches.