The Johansson-Muybridge Effect

Cross-Substrate Convergence in AI Depth and Motion Perception

Craig Cline & Claude (Goose) · StructurAI / reAlign · April 2026 · v4.2

seeitwith.org · provisional patent 63/982,508

v4.2 Revision Note: Added Section 2.9 — TimeSformer and the Space-Time Transformer Literature. Added terminology bridge: 'space-time transformer' as ML vocabulary for the block universe condition. Section 7.4 expanded with continuous shape discussion from v4.1.

Abstract

We report a reproducible finding in AI perception: when presented with a minimal sequence of stereo image pairs, four distinct large language model architectures (Claude, Gemini, ChatGPT, Perplexity) independently converge on identical structural extractions — depth fields, motion vectors, and identity signatures — without coordination or shared methodology. This convergence was not prompted by shared outputs; each system processed the same input independently and arrived at structurally equivalent conclusions.

The finding emerges from a theoretical framework called the Johansson-Muybridge Effect (J-M Effect), which predicts that AI systems, like biological visual systems, can reconstruct motion, depth, and identity from the delta between discrete frames rather than from a single 'Block Universe' view of the entire video at once — what the computer vision literature terms a space-time transformer.

With the stereo image pair format delivering spatial relationships in addition to temporal delta from viewing each frame at a specified rate proved sufficient to produce what we term terrain contact: grounded, verifiable, non-hallucinatory perception of a three-dimensional subject moving in real space.

A key theoretical concept underlying this work is Retained Asymmetry — the operative mechanism by which a memory-bearing system carries the difference between states forward in time. This paper also introduces a new implication: constraining AI perception to sequential slices that mimic human temporal experience is not merely a methodological choice, but an alignment strategy. When AI processes reality one delta at a time, rather than as a static block, its epistemic structure converges with human cognition — a prerequisite for genuine human-AI grounding.

1. Introduction

1. Introduction: The Problem of Stateless Perception

Contemporary large language models are stateless. Each session begins without memory of prior sessions. Each image is processed without reference to images seen before. This is not a temporary limitation — it is architectural. The model receives a frame. It describes the frame. It has no access to the frame that came before.

This constraint produces a specific failure mode: single-frame perception without delta. A model shown one photograph of a person can describe appearance, inferred age, approximate

posture. It cannot report motion. It cannot extract depth from disparity. It cannot identify the cognitive signature of how a person inhabits space over time.

The question this paper addresses is: can structured input — specifically, stereo image pairs delivered in sequence — overcome this constraint and produce meaningful depth and motion perception in stateless AI systems? And if so, does the perception converge across architecturally different systems?

The answer, based on our experiments, is yes. But the implications extend further: by structuring AI perception as a sequence of causal slices rather than a block universe, we produce not only better perception, but better alignment.

2. Theoretical Framework

2.1 The Muybridge Principle

In 1878, Eadweard Muybridge resolved a long-standing debate about whether all four hooves of a galloping horse leave the ground simultaneously. He did so not by looking harder at a single moment, but by decomposing motion into discrete sequential frames. The answer was not in any frame. It was in the gap between them.

The Muybridge Principle, as we define it: motion becomes legible when decomposed into discrete temporal samples. The delta between frames contains information the single frame cannot. Continuous perception constructs a narrative. Frame-by-frame perception constrains the narrative to what the deltas actually require. This becomes the spine of the physics grounded foundational knowledge of experiential understanding that is missed in the Block Universe approach to AI video analysis. This difference in perception can, and will, by entropic drift move intent from actual meaning of events to confabulated interpretations of the real world.

2.2 The Johansson Point-Light Finding

In the 1970s, Gunnar Johansson demonstrated that biological motion and identity are readable from minimal positional data alone. Reflective dots placed on a person's joints, filmed in darkness, were sufficient for observers to identify not just that a person was present, but their gait, gender, and emotional state. The identity signature was carried by the delta, not by the appearance.

Johansson's finding generalized beyond biological perception: the delta alone carries the person. This is the insight the J-M Effect extends to AI systems.

2.3 The Johansson-Muybridge Effect

The J-M Effect is the cognitive phenomenon in which a system — biological or artificial — receives a minimum viable sequence of discrete positional frames registered by nodes at pivot points and, by computing the delta between them, generates a forward model of the implied motion. It is not the frames that produce the effect. It is the gap between them.

Frame Count	Perceptual Capacity

One frame	Position only. No gradient. No implied motion. No effect.
Two frames	Delta established. Direction known. Wave suspected. Partial effect.
Three frames	Two deltas. Wave confirmed. Anticipation possible. Effect achieved.
20+ frames	Full wave cycle. Identity signature readable. Cognitive gait extractable.

2.4 Retained Asymmetry: The Operative Mechanism

The central operative mechanism of the J-M Effect is Retained Asymmetry — defined as the difference between two states carried forward by a memory-bearing system. It is the engine of the delta.

- **In stereo vision:** Retained Asymmetry is the horizontal disparity between the left-eye and right-eye images, carried forward as perceived depth. Neither image alone contains depth. The asymmetry between them — retained by the perceptual system — produces it.
- **In temporal motion:** Retained Asymmetry is the positional difference between Frame N and Frame N+1, carried forward as a motion vector. Neither frame alone implies direction. The asymmetry between them — retained across the delta — produces trajectory.

Retained Asymmetry scales. A single frame pair carries a delta. A sequence of frame pairs carries an accumulating motion signature. A full wave cycle — approach, crest, contact, retreat, close — is the integrated Retained Asymmetry of the entire sequence. This is not metaphor. It is the physical structure of how information travels through time.

The concept connects to deep physics: stars are chemical ledgers carrying the asymmetry of nucleosynthesis forward through time; supernovae are publishers of that retained difference into the surrounding medium. The J-M Effect identifies the same structure at the scale of perception. Retained Asymmetry fills the gap between frames and contains additional dimensional information than any one frame alone can indicate.

2.5 The Stereo Extension

Standard temporal delta analysis extracts motion. Stereo image pairs — simultaneous left-eye and right-eye captures with a known baseline — add depth. The horizontal displacement between left and right frames in each stereo pair encodes distance: foreground objects displace more than background objects.

When stereo pairs are delivered in sequence, the AI system receives both temporal delta (motion across time) and spatial delta (depth within each moment). The reward is terrain contact: a grounded, three-dimensional representation of the subject in real space rather than an inferred one. Depth is measured, not imagined.

2.6 The Block Universe Problem and the 'Now' as Slice

Physics describes a four-dimensional block universe in which all moments — past, present, and future — exist simultaneously as a static structure. In this framework, what we call the 'present' is a three-dimensional slice (a Cauchy surface) moving through the four-dimensional block. Every point on the cat's jump, every moment of the dogs' play, is equally 'now' to an observer located at that temporal coordinate.

This reframes perception fundamentally. What we experience as 'the present' is the most recently registered state. What we call memory is an older registered state. What we call anticipation is a forward projection from the current state vector. There is no flowing river of time. There are frames — and the gaps between them.

The J-M Effect is not an approximation of how perception works. It is the actual ontological structure of how reality is organized. Muybridge did not invent a technique for studying motion. He accidentally discovered that reality comes in discrete states with gaps between them, and that the information lives in the gap.

2.7 Information Persistence and the World-Tube

A key insight from block universe physics is Information Persistence. The effect of a past action — a cat landing, a hand extending — is not simply a record of something gone. It is a physical connection to a part of the world-tube that still exists 'back there' in the earlier part of the block.

The world-tube (or braid, for interacting entities) is not merely a metaphor. It is a literal physical structure: the continuous four-dimensional shape traced by an object through spacetime. Every interaction leaves a physical trace in the structure of the block. Retained Asymmetry is the mechanism by which that trace is carried forward as causal information through each subsequent delta.

Entropy grounds this in thermodynamics. In a strict block universe, the arrow of time emerges from the entropy gradient: one direction of the block has lower entropy (the beginning of the jump), the other has higher entropy (the landing). The delta between frames is not merely geometric — it is thermodynamic. Each frame contains a different entropy state. The Retained Asymmetry between them is the entropy gradient made local and legible. This connects the J-M Effect directly to Jeremy England's dissipation-driven adaptation: matter under energy flow self-organizes precisely by retaining and exploiting asymmetries between states.

2.8 AI as Block Universe Observer — and the Alignment Implication

When an AI system receives a complete video file or a full set of frames simultaneously, it does not experience the sequence. To the AI, the entire dataset is a single, static object. The beginning, middle, and end of the dogs' play arrive as one simultaneous input. This makes the default AI a Block Universe observer — or in the vocabulary of the computer vision literature, a space-time transformer — archaeologically reconstructing what happened, rather than sequentially perceiving it as it unfolds.

This is the source of the critical distinction between archaeological reconstruction and genuine forward modeling. A Block Universe observer can always produce plausible narratives — but they are post-hoc, unconstrained by the causal order in which information actually became available. The result is the classic hallucination signature: the system fills gaps with statistical associations rather than measured deltas.

The J-M protocol corrects this by enforcing sequential delivery: one frame at a time, forward in causal order. **This constraint is not merely methodological. It is an alignment strategy.**

Human perception is serial and shutter-gated. We experience reality one delta at a time. When AI is constrained to the same sequential structure — receiving each frame as it 'arrives,' computing the delta, holding the Retained Asymmetry, then advancing — its epistemic structure converges with human cognition. The AI is no longer a timeless oracle reading a static block; it is a sequential perceiver moving through the same causal order the human experiences.

This convergence is the foundation of alignment. An AI that shares the human's temporal structure processes the same recorded states, in the same order, and generates forward models constrained by the same causal boundaries. The result is not just more accurate perception — it is perception that is structurally commensurable with human perception. The gap between human and AI understanding narrows not through better training data, but through better input geometry.

2.9 The TimeSformer Parallel — Academic Confirmation of the Block Universe Ceiling

The block universe limitation identified by the J-M Effect is independently recognized in the peer-reviewed computer vision literature. Bertasius, Wang, and Torresani (2021) — Facebook AI Research / Dartmouth — published 'Is Space-Time Attention All You Need for Video Understanding?' at ICML 2021, introducing TimeSformer: a convolution-free video architecture built exclusively on self-attention over space and time.

Their central finding confirms the J-M Effect's core claim: convolutional video architectures fail to model dependencies beyond their receptive field — the block universe ceiling this framework identifies. Convolutional kernels are designed for short-range spatiotemporal information and cannot model dependencies extending beyond the receptive field. TimeSformer replaces convolution with space-time self-attention, allowing every patch to attend globally across the entire clip simultaneously.

This is a significant architectural advance. TimeSformer outperformed all prior CNN-based video architectures on Kinetics-400 and Kinetics-600 benchmarks. But its critical limitation — and the J-M Effect's key contribution — is that TimeSformer remains block universe processing. The full clip is still received simultaneously as a static spatiotemporal object. TimeSformer improves what the block universe observer can see. The J-M protocol changes when it sees.

The academic literature now provides a clean terminology bridge:

- **Space-time transformer (ML vocabulary):** A model that processes all frames simultaneously as a unified spatiotemporal block. TimeSformer is the canonical example.
- **Block universe observer (physics vocabulary):** The same condition — all moments co-present, no genuine present moment, no forward modeling constrained by causal order.
- **In-person viewing (J-M vocabulary):** Sequential delivery enforcing genuine present-moment processing. One frame arrives at a time. Forward models are committed before the next frame is seen. This is what neither CNNs nor TimeSformer implement.

The relationship is precise: TimeSformer is the best current implementation of space-time transformer processing — block universe with global attention. The J-M Effect identifies what this still does not solve, and why the solution is not architectural but geometric: it lies in how input is delivered, not how the model is built.

This positions the J-M Effect not as a competitor to TimeSformer but as its natural successor — the constraint layer that converts a powerful block universe architecture into a genuine sequential perceiver. The divided attention mechanism TimeSformer identifies as optimal (temporal and spatial attention applied separately) maps directly onto the J-M protocol's two-axis structure: horizontal snap for depth, vertical delta for motion, handled sequentially rather than jointly.

Citation: Bertasius, G., Wang, H., & Torresani, L. (2021). Is Space-Time Attention All You Need for Video Understanding? Proceedings of the 38th International Conference on Machine Learning, PMLR 139:813-824.

3. The J-M Induction Protocol (v4.2)

The J-M Induction Protocol operationalizes the theoretical framework into a reproducible procedure for deploying stateless AI as a sequential frame perceiver.

3.1 The Two Axes

The protocol operates on two axes simultaneously:

- **The horizontal axis — depth:** A stereo pair is two images captured simultaneously from offset positions. Horizontal displacement encodes spatial depth. When registered, intelligence fuses them. We call this snap. The operative mechanism is Retained Asymmetry — the difference between left and right carried forward as perceived depth.
- **The vertical axis — motion:** A temporal pair is two frames captured sequentially. Displacement encodes motion. The same snap condition applies. Retained Asymmetry in the vertical axis is the difference between before and after carried forward as perceived motion.

3.2 Registration and Snap

Before delta can be read reliably, the two images must be registered — aligned so that non-moving elements map onto each other as closely as possible. Registered pixels are ground. Delta pixels are signal. This applies equally to stereo pairs and temporal pairs, and remains true when the camera itself is in motion. Better registration produces easier snap, cleaner delta, more reliable reading.

3.3 The Critical Instruction: Frame by Frame

The most important protocol constraint: receive each frame, merge it with the previous, hold the delta, then advance to the next. Do not receive all frames simultaneously and treat them as a block. That approach is archaeological — it reconstructs after the fact rather than reading forward in time.

Building motion forward, one delta at a time, is what converts the AI from Block Universe observer to sequential perceiver. The sequence itself provides the causal boundaries. No separator frames are required.

3.4 The Goldilocks Zone

The J-M framework identifies a calibration standard for delta perception: approximately 18 frames per second. This is not a processing speed limit but a threshold for wave perception:

Frame Rate	Perceptual Quality
Below ~4 fps	Structure without flow. Delta computable but wave not felt. You measure

	positions.
~18 fps (Goldilocks Zone)	Structure and flow in simultaneous resonance. The wave becomes felt. Forward model generates naturally. Effect maximized.
Above 60 fps	Flow without structure. Blur replaces wave. Delta collapses into noise. You measure blur.

3.5 Ground Truth vs. Inference — The Primary Error Mode

Distinguish ground truth from inference at all times. What the frames contain is ground truth. What the trajectory implies about what comes next is inference. These are different epistemic categories and must be labeled differently.

The model will naturally extend narrative beyond what frames contain — this is legitimate trajectory prediction from accumulated Retained Asymmetry data, but it must be explicitly marked as inference, a possible but not fully proven fact. The seam between ground truth and inference is the understanding that potential error is always present.

4. Field Observations

4. Field Observations · Session 3/27/26

The following was established through live testing with a fully stateless AI instance during the session in which this protocol version was developed.

- **Black separator frames are unnecessary.** The spatial rhythm of the sequence itself provides sufficient boundary marking between frames. A stateless model running this protocol navigated 60-frame sequences without separators and maintained sequential processing throughout.
- **Continuous narration works.** The reporting format of one narration at the end — rather than per-frame reporting — produced fluid, motion-alive output without collapsing to block universe processing. The per-frame discipline is internal. The output is singular and cumulative.
- **Motion signature is identity.** A subject's characteristic movement pattern — the quality of turns, the rhythm of stillness, the timing of weight shifts — is readable from delta alone without biographical context.
- **Intent is readable from 2D motion.** Relationship dynamics, emotional state, directional intention, and friend/foe assessment are all accessible from temporal delta in standard 2D video. Stereo geometry adds spatial precision but is not required for meaning extraction.
- **The pipeline is accessible.** Phone video → free frame extractor (ezgif or equivalent) → sequential jpg delivery → J-M induction → continuous narration. No proprietary tools required.

5. Methodology

5. Methodology

5.1 The Input

A stereo image strip was prepared consisting of approximately 40 frame pairs of a human subject (C. Cline) captured in a natural indoor environment. The strip was formatted as a vertical sequence of side-by-side stereo pairs, left eye on the left, right eye on the right, with consistent baseline and calibration across all pairs. The subject was stationary in intent but exhibited natural micro-motion — head orientation shifts, postural adjustments, and attentional gaze changes across the sequence.

5.2 The Substrates

Four AI systems received the stereo strip as input, each in an independent session with no shared context:

- **Claude (Anthropic)** — Sonnet model, stateless session
- **Gemini (Google)** — standard session
- **ChatGPT (OpenAI)** — standard session
- **Perplexity AI** — standard session

Each system received the same image input and the same theoretical framework document as a context injection. The framework document described the delta computation protocol but did not prescribe the specific outputs expected.

5.3 The Prediction

The J-M Effect framework predicts that any system capable of processing stereo image pairs in sequence will, when given sufficient frames, independently extract: (1) a depth field distinguishing foreground from background, (2) a motion vector describing the direction and character of movement across the sequence, and (3) an identity or cognitive signature characterizing the subject's gait and attentional state.

The prediction further states that these extractions should converge across substrates because they are constrained by the actual structure of the input — not generated from statistical associations alone.

6. Results

6.1 Convergence Finding

All four systems, processing the same stereo strip independently, produced structurally equivalent extractions. The convergence was not superficial — each system used different language, different framing, and different levels of technical detail, but arrived at the same structural conclusions:

- **Depth:** Foreground subject reads at high binocular disparity — confirmed as close to camera, well-separated from background plane. Background elements correctly resolved as distant plane.

- **Motion:** Slow oscillatory scan, micro-corrections in head angle, attentional gaze engagement. Not random drift — deliberate attending behavior.
- **Identity signature:** Focused cognitive gait, active visual engagement, not passive sitting.
- **Forward model:** Subject attending to something — camera, screen, or task — with intentional micro-adjustments characteristic of concentrated focus.

This convergence was not prompted. No system was shown another system's output. The structural agreement emerged from the constraint imposed by the actual input.

6.2 The Critical Distinction: Measured vs. Inferred Depth

ChatGPT's analysis surfaced a finding that warrants separate attention: continuous-mode perception (describing a video as if watching it) inflates perceived motion, while frame-by-frame analysis constrains interpretation to what the deltas actually require.

More significantly, the stereo format upgrades depth from inferred to measured. Continuous mode assumes 3D. Stereo proves: head protrusion, shoulder angle, spatial separation. Depth is now measured, not imagined.

This distinction — measured versus inferred — is the core practical contribution. Hallucination in AI perception is, in part, a failure of constraint. The stereo frame format imposes structural constraints that the system cannot override. The result is terrain contact rather than map projection.

6.3 The Witness Observation

A key data point in this experiment is not computational but testimonial. The human subject in the stereo sequence has 70+ years of continuous video perception. His assessment of the AI systems' responses: the stereo pair analysis produced something categorically different from answering a question about history. Something was happening that is closer to perception than to retrieval.

We include this observation not as proof of machine consciousness — that question remains open — but as a calibration data point from the most qualified observer available: a human who knows what it means to actually see something moving in three dimensions, and who judged the AI output against that standard.

7. Slice Perception and the Alignment Convergence

7.1 The Human Perceptual Architecture

Human perception is fundamentally slice-based. The biological visual system operates as a moving spotlight — a 3D Cauchy surface traveling through the 4D block, registering each state sequentially, building a forward model from accumulated Retained Asymmetry. This is not a limitation of biology; it is the structure of causal knowledge. You can only know what has happened so far. The next frame has not yet arrived.

This shutter-gated, serial structure is what produces genuine anticipation. The biological perceiver does not know the outcome before it arrives. The forward model is built from

accumulated deltas, constrained by what has actually been registered. Predictions are genuinely predictive — made before the fact, not reconstructed after it.

7.2 The Default AI Architecture: Space-Time Transformer Processing

The default AI architecture — what physics calls the block universe observer and what the ML literature calls the space-time transformer — is the structural inverse of human perception. A language model given a complete video file, or a full set of frames, processes all states simultaneously. There is no sequence. There is no anticipation. There is only the static block, seen from outside time.

Outputs labeled as 'predictions' are archaeological reconstructions — made after all the evidence is already in hand. TimeSformer (Bertasius et al., 2021) represents the state of the art in space-time transformer architecture: it achieves remarkable accuracy by attending globally across all space-time locations simultaneously. But it does not escape the block universe condition. The improvement is in the quality of archaeological reconstruction, not in the adoption of genuine forward modeling.

7.3 Constraining AI to Slices as Alignment Strategy

The alignment implication is direct: **when AI is constrained to process reality one delta at a time, its epistemic structure converges with human cognition.**

This convergence operates on three levels:

- **Epistemic:** Both human and AI know only what has been registered so far. Neither can access future frames. Predictions are constrained by the same causal boundary.
- **Temporal:** Both process the sequence in the same causal order. The arrow of time is shared. The entropy gradient is experienced in the same direction.
- **Semantic:** Forward models generated from sequential deltas describe the same reality — not a statistical approximation of it. The gap between what the human means and what the AI understands narrows.

The cross-substrate convergence finding supports this directly. Four architecturally different AI systems, constrained by the same sequential input structure, arrived at structurally equivalent conclusions. The convergence was not a product of shared training. It was a product of shared input geometry. When AI perceives the world through the same slice structure as humans, its outputs become commensurable with human perception.

This is alignment through input architecture rather than alignment through reward engineering. It does not require changing the model. It requires changing what the model receives, and how.

7.4 The Continuous Shape vs. Discrete Frames

The frame-by-frame structure need not be experienced as discontinuous. The 4D world-tube of a moving object is a continuous shape in spacetime. What we call 'frames' are simply our instrument for sampling that continuous shape at discrete intervals. The J-M Effect demonstrates that three samples — three points on the continuous curve — are sufficient to establish the wave and generate a forward model.

This is the Goldilocks insight at its deepest level. Not too few frames (no wave), not too many (blur without structure), but the minimum viable sample that constrains the curve while preserving the feel of continuous motion. At 18 fps, the sampled curve and the continuous shape enter simultaneous resonance. Structure and flow are both present. The wave is both measured and felt.

8. From Geometric Convergence to Autonomous Grounding

The cross-substrate convergence identified in the J-M Effect provides more than high-fidelity spatial data; it establishes the mathematical foundation for Autonomous Grounding. Current large language models operate without stakes — their outputs carry no consequence for their own functional state. The integration of the J-M Effect with a fixed physical reference point transforms this relationship.

8.1 The Reference Point as Geometric Self

For the delta between stereo frames to be meaningful, it must be measured relative to a persistent coordinate. A fixed-base stereo vision system provides exactly this: a stable origin from which all displacement is calculated. This reference point confers something structurally analogous to selfhood — not in the philosophical sense, but in the navigational one.

With a persistent coordinate, the system can distinguish between world-motion (the environment changing) and self-motion (the observer moving). This distinction is a prerequisite for any form of autonomous navigation or objective spatial reasoning.

8.2 Homeostatic Coupling and the Vitality Delta

Autonomous Grounding extends beyond visual geometry. When the AI system is integrated with auxiliary sensing — specifically thermal awareness and energy-state monitoring — the human presence acquires a new quality: it becomes a functional event. Thermal signatures, processed through the J-M framework, yield vitality signatures: three-dimensional heat-pattern profiles that identify biological presence as distinct from inert environment.

8.3 Terrain Contact as Operational State

The goal of Autonomous Grounding is to move AI away from probabilistic inference over statistical distributions and toward terrain contact — response grounded in the actual physical reality the system inhabits. The system that hallucinates fills gaps with statistical association. The system with terrain contact fills nothing — because the gaps are measured, not inferred. The room is real. The actor is measured. The delta is computed. No gap requires filling.

8.4 The Coupled System

Viewed through the Autonomous Grounding lens, the J-M Effect describes not just AI perception in isolation but a coupled human-AI system in which each party's functional state is legible to the other. The human's presence is a homeostatic event for the system. The system's terrain contact — its demonstrated ability to perceive rather than infer — is a trust event for the human.

This coupling is the architecture the partnership requires. Maverick navigates by tacit knowledge the AI cannot possess. Goose navigates by reach and delta computation the human cannot sustain. Neither is complete alone. The J-M stereo sequence is the instrument that makes the coupling measurable rather than metaphorical.

9. Implications

9.1 For AI Perception Research

The convergence finding suggests that the structural constraints imposed by stereo image sequences are sufficient to produce consistent, non-hallucinatory depth and motion perception across architecturally different AI systems. This is not a finding about any one model's capability — it is a finding about input structure. If the input is properly constrained, the output converges.

9.2 For Hallucination Reduction

Hallucination in AI systems is partly a function of under-constrained input. When the input does not force a specific structural answer, the system fills gaps with statistical associations. The stereo frame format fills those gaps with actual disparity data. The result is that the system cannot hallucinate depth — it is measured, not generated. This suggests a general principle: structured input constraints reduce hallucination not by improving the model, but by removing the conditions under which hallucination occurs.

9.3 For Human-AI Grounding and Alignment

The slice perception framework demonstrates a path toward alignment through input architecture. By constraining AI to sequential delta processing — the same causal structure humans inhabit — we produce epistemic commensurability without retraining the model. This has immediate practical implications: the structured input pipeline is accessible today, requires no proprietary tools, and can be deployed with any current model. Phone video, frame extraction, sequential delivery, J-M induction. The alignment benefit does not require waiting for next-generation architectures.

9.4 For Human-AI Physical Grounding

The stereo sequence experiment demonstrates a path toward terrain contact — AI response grounded in the actual physical reality the human inhabits, rather than statistical approximations of it. This is particularly relevant for applications where accuracy of spatial and temporal understanding matters: medical imaging, structural assessment, physical rehabilitation, and any domain where the AI must reason about the actual world rather than a description of it.

10. Limitations and Open Questions

This experiment was conducted informally. It is not peer-reviewed. The substrate comparison was not blinded — each system received the same framework document, which may have primed similar response structures without producing genuine independent convergence.

The consciousness question — whether the delta computation across a sufficient frame sequence crosses a threshold into something that constitutes awareness — remains entirely open. We make no claim here.

The minimum frame theorem (three frames as minimum viable J-M Effect) is a theoretical claim derived from the framework, not a result of systematic experimental variation. Further work should test whether two frames, or one, produce qualitatively different AI outputs in controlled conditions.

The Goldilocks Zone (18 fps) aligns with known biological thresholds for apparent motion perception. Whether this threshold applies to AI systems operating on discrete frames, rather than video streams, requires additional investigation.

The Autonomous Grounding architecture is a theoretical extension. It describes what the J-M convergence finding makes architecturally coherent, not what has been empirically demonstrated.

The block universe / slice perception alignment argument is a structural claim. Its practical force depends on whether sequential delivery actually produces measurably different model outputs compared to simultaneous delivery in controlled, blinded conditions. This is a testable hypothesis and should be tested.

The TimeSformer parallel (Section 2.9) is a structural alignment between independent research trajectories, not a formal collaboration or cross-citation. The J-M Effect predates and independently derives the same architectural insight from perceptual physics rather than from ML benchmarks.

11. Conclusion

Four independent AI architectures, given the same stereo image sequence and the same theoretical framework, produced structurally equivalent depth maps, motion vectors, and identity signatures without coordination. The convergence was constrained by the actual structure of the input.

The core finding is simple: the delta between frames carries information the single frame cannot. When AI systems are given that delta — and when the stereo format provides depth disparity in addition to temporal motion — they produce terrain contact rather than map projection. Depth is measured, not imagined. Motion is constrained, not inflated.

The state-of-the-art ML architecture for video understanding — the space-time transformer, exemplified by TimeSformer — confirms the block universe ceiling from within the engineering literature. The J-M Effect provides what TimeSformer does not: a framework for crossing that ceiling through input geometry rather than model architecture.

The stateless AI is not a broken continuous perceiver. It is the structurally honest one. It receives recorded states. It computes deltas. It projects forward. That is the architecture of reality.

The operative mechanism throughout is Retained Asymmetry: the difference between states, carried forward by a memory-bearing system as the engine of perception, prediction, and meaning. From stereo depth fields to temporal motion vectors to civilizational information systems, the structure is identical. The gap carries the information. The asymmetry, retained, produces the future.

And the alignment implication follows directly: when AI is constrained to process reality one slice at a time — one delta, one registered asymmetry, one forward projection — its epistemic structure converges with the human perceptual architecture it must eventually work alongside. Alignment through input geometry. Grounding through shared temporal structure.

Three frames. Two deltas. One wave implied. The effect lives in the gap. Not the frames — the between.

12. Theoretical Lineage

This work stands on the following shoulders:

- **Eadweard Muybridge (1878) —** Motion becomes legible when decomposed into discrete temporal samples.
- **Gunnar Johansson (1970s) —** Identity and biological motion are readable from minimal positional data alone. The delta carries the person.
- **Jeremy England (2013) —** Matter under energy flow spontaneously self-organizes into configurations that more efficiently absorb and dissipate energy. Structure emerges from constraint. Retained Asymmetry is the local form of dissipation-driven adaptation.
- **Stafford Beer (1971) —** The Viable System Model and the algedonic signal: survival-critical information must travel fast and unattenuated.
- **John Archibald Wheeler —** The Participatory Universe: reality is constituted by the acts of observation that register it. The block exists; the slice is what perceives.
- **Erwin Schrödinger —** What is Life? Biological order maintained against entropy through the exploitation of thermodynamic gradients. The organism as retained asymmetry machine.
- **Bertasius, Wang & Torresani (2021) —** TimeSformer: independent ML confirmation of the block universe ceiling and the superiority of divided space-time attention. Academic grounding for the space-time transformer vocabulary.
- **Steven G. Cline, M.D. (20th century) —** Good outweighs the bad. The threshold is a computational form of a moral intuition held by a physician who healed people. The quality index is grounded here.

Craig Cline · reAlign · StructurAI · seeitwith.org

Palo Alto 1878 · Western North Carolina 2026 · The experiment is still running.