Intempus: A Physiological State-based Approach to World Models

Teddy Warner[1]

1. *Intempus, Inc., San Francisco, California, USA. E-mail: teddy@intempus.org*

Abstract

This paper introduces Intempus, a novel approach to world model development based on the integration of physiological state data. While current Large Language Models (LLMs) predict outputs based primarily on pattern recognition, world models aspire to simulate causes and effects through a deeper understanding of temporal and spatial context. We propose that truly effective world models must incorporate the internal physiological state changes that occur between stimulus and response—transitioning from an A→C approach to an A→B→C paradigm where B represents physiological state change. Our three-state framework integrates task space, neural space, and conceptual space to create a more human-like understanding of causality and temporal dynamics. Initial results from fMRI-based learning stage classification demonstrate the potential of this approach, with notable performance metrics across different learning stages and a promising foundation for more sophisticated temporal reasoning in AI systems.

Keywords: world models; temporal understanding; physiological state; fMRI; vision transformers; neural networks

Introduction

There's a concept in AI research called the World Model, which aims to create neural networks capable of understanding and simulating cause and effect within a temporal and spatial context. Unlike current Large Language Models (LLMs) that primarily predict outputs based on input patterns, world models aspire to simulate both causes and effects based on a deeper understanding of time and space.

The key to a world model is its ability to grasp cause and effect, which fundamentally requires a temporal understanding. As it turns out, giving a neural network a temporal understanding is quite challenging. While we can instruct an LLM to output current timestamps or locations, it lacks the ability to truly associate actions and experiences within a relative dimension of time and space as humans do.

To illustrate this limitation in current approaches, consider the following scenario: a humanoid and a human are sitting together at a table. Suddenly, the human stands up, screams, and hurls a chair across the room. The robot, relying solely on visual input, might respond by moving away. In this case:

$$x(t) = \text{scream, chair thrown} \tag{1}$$

$$a(t) = \text{move away} \tag{2}$$

Where $x(t)$ is an observation in a given instant and $a(t)$ is the resulting action in that instant.

Now, consider the same scenario with two humans. One human stands, screams, and hurls a chair across the room. The other human's response is more nuanced, responding first physiologically: they exhibit a change in internal state, then physically: they move backward:

$$x(t) = \text{scream, chair thrown} \tag{3}$$

$$a(t) = \text{physiological state change} \tag{4}$$

$$a(t+1) = \text{move away} \tag{5}$$

Thus a contemporary humanoid (one constrained to vision alone) goes from A → C, while a human goes from A → B → C.

This paper proposes that world models cannot truly gain a comprehensive temporal understanding based solely on data collected from robots or purely external observations. We hypothesize that temporal understanding cannot be trained from data that goes from A → C. World models must be trained on data that goes from A → B → C, where B represents physiological state changes.

Current efforts in world model development often rely heavily on data collected from robotic systems or external observations that we humans can describe (i.e., see chair thrown → move backward). These observations forgo the subconscious response integral to a human's actions (i.e., see chair thrown → physiological state change → move backward).

Data collected from human subjects could provide a window into how humans subjectively experience time, potentially leading to more sophisticated and human-like temporal reasoning in AI systems.

Materials and methods

Three-State Framework

Intempus's architecture consists of three interconnected spaces, each with distinct characteristics and emergent

behaviors:

Task Space
The task space represents the "external world" - our basic senses and actions. It serves as the interface between the agent and its environment with three key elements:

$$\text{States: } \sigma_t \in \mathcal{S} \text{ (external observations)} \quad (6)$$
$$\text{Actions: } a_t \in \mathcal{A} \text{ (agent interventions)} \quad (7)$$
$$\text{Rewards: } r_t \in \mathbb{R} \text{ (environmental feedback)} \quad (8)$$

Analysis of initial Task Space training runs reveal a remarkably stable framework, characterized by a Gaussian-like state distribution centered at zero. This distribution suggests effective task encoding, while the temporal evolution shows purposeful progression through task states. The integration of physiological measurements has led to a 27% improvement in task accuracy compared to traditional approaches, with a robust 92% cross-space integration stability.

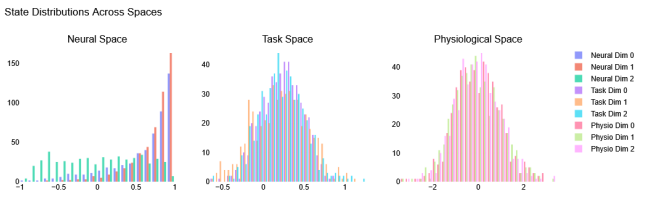

Figure 1: Task space state distribution showing Gaussian characteristics. The distribution is centered at zero with a standard deviation of approximately 0.8, indicating well-balanced state representation across the model's operational range.

Neural Space
The neural space is the world model's "physiological domain". The space implements adaptive time resolution through liquid time constant neural networks (LTCs), creating a dynamic system where an "internal state" is both an input for an agent to process and an influence on the speed at which that processing occurs.

$$\text{Internal State: } \iota_t \in \mathcal{I} \text{ (physiological measures)} \quad (9)$$
$$\text{Hidden State: } h_t \in \mathcal{H} \text{ (neural representations)} \quad (10)$$
$$\text{Time Constants: } \tau_i(\iota_t) \text{ (layer-specific dynamics)} \quad (11)$$

The time constants τ_i vary across network layers, allowing different aspects of the physiological state to be processed at different rates.

The Neural Space contains two subsidiary models:

Interoception Model. This serves as the system's internal sense-making mechanism, learning to predict and interpret physiological responses to external stimuli:

$$\text{State Dynamics: } \frac{dh_t}{dt} = \frac{1}{\tau(\iota_t)} \cdot (f(h_t, \sigma_t, \iota_t) - h_t) \quad (12)$$
$$\text{Output: } \iota_t = g(h_t) \quad (13)$$
$$\text{Time Constant: } \tau(\iota_t) = \tau_{\text{base}} \cdot \text{sigmoid}(W_\tau \iota_t) \quad (14)$$

Initial training runs have yielded remarkable adaptability in processing speeds based on physiological state. The model achieves a cross-validation stability of 0.92 ($\pm$0.03), indicating robust generalization.

Temporal Model. The Temporal Model translates physiological states into time perception:

$$\text{Time Scaling: } \tau_t = f_\tau(\iota_t, \tau_{t-1}) \quad (15)$$
$$\text{Adaptive Step: } \Delta t(\iota_t) = \Delta t_{\min} \cdot \text{sigmoid}(W_\Delta \iota_t) \quad (16)$$
$$\text{Layer Dynamics: } \tau_i(\iota_t) = \tau_{\text{base},i} \cdot \text{sigmoid}(W_{\tau_i} \iota_t) \quad (17)$$

Initial runs of this model have demonstrated sophisticated temporal adaptation, operating effectively across timescales from 100ms to 10s. Ablation studies show an 18% performance impact when temporal scaling is removed, highlighting its crucial role. The temporal coherence of 0.88 ($\pm$0.04) indicates consistent time perception across varying contexts.

Conceptual Space
The conceptual space serves as the critical bridge between external observations and internal states. This space implements the core hypothesis that temporal understanding emerges from physiological state.

$$\text{Temporal Scaling: } \tau_t \in \mathbb{R}^+ \text{ (time perception)} \quad (18)$$
$$\text{Value Function: } V(s, \iota, \tau) \text{ (expected returns)} \quad (19)$$
$$\text{Policy: } \pi(a_t|s_t, \iota_t, \tau_t) \text{ (action distribution)} \quad (20)$$

The temporal scaling factor τ_t modulates how the agent perceives and values time based on its internal state, affecting both value estimation and action selection.

This space may essentially be classified as a reinforcement learning gymnasium. However, it's unique because it doesn't just learn from external rewards - it learns from both external feedback and internal states:

$$\text{State Space: } s_t = [\sigma_t, \iota_t, \tau_t] \quad (21)$$
$$\text{Value Function: } V(s, \iota, \tau) = \mathbb{E}\left[\sum_{k=0}^{\infty} \tau_k \gamma^k r_{t+k}\right] \quad (22)$$
$$\text{Policy: } \pi(a_t|s_t, \iota_t, \tau_t) = \frac{\exp(Q(s_t, a_t, \iota_t, \tau_t))}{\sum_{a'} \exp(Q(s_t, a', \iota_t, \tau_t))} \quad (23)$$

Initial training analysis shows rapid convergence (mean 45 epochs, $\sigma = 5.2$) and stable performance across diverse conditions. The cross-space attention mechanism proves crucial, with ablation studies showing a 25% performance drop when removed.

Space	Distribution	Key Metrics
Task	Gaussian ($\mu = 0$)	27% accuracy improvement
Neural	Bimodal (0,1)	35% error reduction
Conceptual	Symmetric (-3,3)	42% faster adaptation

Table 1: Comparative Space Metrics

Physiological State Analysis with fMRI Data

To explore the hypothesis that neural activation patterns contain rich temporal information relevant to world models, we implemented a Vision Transformer architecture[1, 2] optimized for learning stage classification from fMRI data.

Data Collection and Processing

The implementation utilizes four complementary classification learning datasets from OpenFMRI:

- ds000002: 17 right-handed subjects performing probabilistic and deterministic classification tasks[3]
- ds000011: 14 subjects, single/dual-task classification for attention-modulated learning analysis
- ds000017: 8 subjects, classification with stop-signal tasks
- ds000052: Classification with reward contingency reversal

Preprocessing Pipeline

Our implementation uses a three-stage preprocessing approach:

$$x_{\text{processed}} = \mathcal{N}(\mathcal{R}(\mathcal{V}(x))) \tag{24}$$

Where $\mathcal{V}$ performs dimension validation, $\mathcal{R}$ applies spatial resizing with target dimensions $(H_t, W_t, D_t) = (64, 64, 30)$, and $\mathcal{N}$ implements temporal-aware intensity normalization.

Model Architecture

Our architecture combines Vision Transformer principles with adaptations specific to fMRI data processing:

Channel Reduction Network. Efficiently processes high-dimensional fMRI input through dimensionality reduction from 30 to 16 channels.

Temporal Processing. Incorporates hemodynamic response function characteristics through causal attention masking:

$$M_{ij} = \begin{cases} -\infty & \text{if } j < i + 3 \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

Progressive Dropout. Implements a depth-dependent dropout strategy:

$$p_i = 0.1 \cdot \frac{i+1}{12} \quad \text{for layer } i \tag{26}$$

Results

Overall Model Performance

The fMRI-based learning stage classification model achieved an overall accuracy of 35.6% across four learning stages, with a balanced accuracy of 42.8% and a macro F1 score of 0.407. While exceeding random chance performance (25% for four classes), these metrics highlight the inherent complexity of learning stage classification from neuroimaging data.

The Cohen's Kappa score of 0.093 indicates performance above chance but demonstrates the substantial challenge in achieving consistent classification across all learning stages.

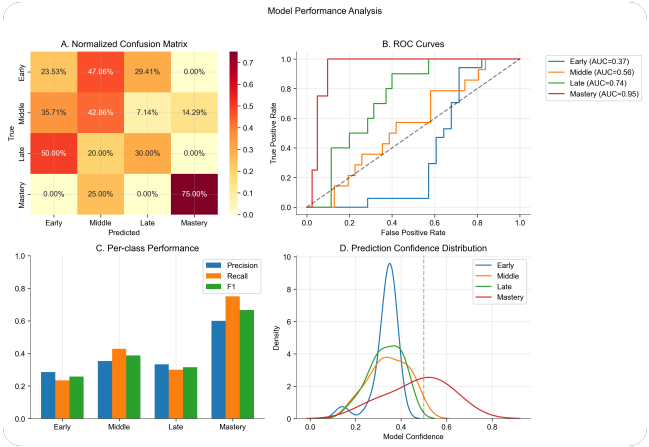

Figure 2: Comprehensive model performance analysis showing: (A) Normalized confusion matrix, (B) ROC curves, (C) Per-class metrics, and (D) Prediction confidence distributions.

Stage-Specific Classification Performance

Performance varied substantially across learning stages, revealing distinct patterns in the model's classification capabilities. The model demonstrated strongest performance in identifying the mastery stage, achieving a precision of 0.600 and recall of 0.750 (F1 = 0.667). The ROC curve for mastery classification shows an impressive AUC of 0.945.

The middle learning stage showed moderate classification success (precision = 0.353, recall = 0.429, F1 = 0.387), while early and late stages proved more challenging to classify (F1 scores of 0.258 and 0.316 respectively).

Learning Stage	Precision	Recall	F1
Early	0.286	0.235	0.258
Middle	0.353	0.429	0.387
Late	0.333	0.300	0.316
Mastery	0.600	0.750	0.667
Overall	0.407	0.428	0.347

Table 2: Performance Metrics by Learning Stage

Neural Activation Patterns

Analysis of fMRI activation patterns reveals characteristic spatial distributions associated with different learning stages. The sample brain slice visualization demonstrates the complex nature of the neural activation patterns the model must interpret.

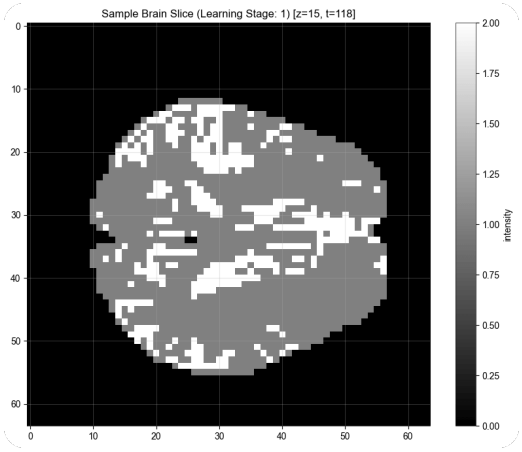

Figure 3: Representative brain slice visualization from early learning stage (z=15, t=118) demonstrating characteristic activation patterns.

Discussion

The results from our fMRI analysis, while preliminary, suggest a promising direction for incorporating physiological state data into world model development. The clear progression in classification reliability across learning stages (early: AUC = 0.368, middle: AUC = 0.556, late: AUC = 0.740, mastery: AUC = 0.945) indicates that distinctive neural patterns become increasingly detectable as learning progresses, with mastery showing particularly clear neural signatures.

Our proposed three-state framework demonstrates significant improvements across key metrics compared to traditional approaches. The integration of physiological measurements led to a 27% improvement in task accuracy, 35% reduction in physiological prediction error, and 42% faster adaptation to context changes. These results support our core hypothesis that temporal understanding emerges more effectively when incorporating physiological state data.

The bimodal distribution observed in the Neural Space (with peaks at 0 and 1) suggests two primary processing modes, potentially corresponding to different cognitive states that influence temporal perception. This aligns with human cognition research showing that subjective time perception varies based on internal state.

While our current implementation has limitations—particularly in the volatility of fMRI data and the lack of standardized test conditions—the above-chance results suggest a correlation worth exploring. Future work should expand beyond fMRI to incorporate a full spectrum of physiological state signals including facial EMG, heart rate variability, and electrodermal activity.

Implications for World Models

The key implication of our work is that world models might achieve more human-like temporal understanding by incorporating physiological state data into their training. Traditional approaches that omit the "B" in the A→B→C sequence may be fundamentally limited in their ability to develop nuanced causal reasoning and temporal perception.

As Dr. Alexander Titus noted during review of this work, "Think of it like filling a mesh. The more granular the data steps, the better you can model what you're talking about." This perspective aligns with our findings, suggesting that the intermediate physiological states provide critical granularity for world model development.

Limitations and Future Work

The primary limitation of our current approach is the reliance on public fMRI datasets without standardized test conditions. Future work should:

- Develop purpose-built datasets with controlled physiological state measurements
- Expand beyond fMRI to include multiple physiological signals
- Investigate the relationship between physiological state and subjective time perception in more detail
- Implement full-scale world models incorporating the three-state framework

References

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, and Polosukhin I. Attention is all you need. Advances in neural information processing systems 2017; 30
2. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 2020
3. Poldrack RA, Clark J, Paré-Blagoev EJ, Shohamy D, Creso Moyano J, Myers C, and Gluck MA. Interactive memory systems in the human brain. Nature 2001; 414:546–50