ALL2Vec: Continuous Predictive Representations for Dynamic Visual Streams

A Preprint

Ken I.
Independent Researcher, Japan
ken.i.research@gmail.com

November 3, 2025

Abstract

Current vision systems, from video transformers to multimodal models like CLIP and ImageBind, process visual input as sequences of discrete frames. This frame-by-frame discretization, while effective for offline tasks, introduces fundamental limitations for real-time embodied systems: temporal discontinuities, representational instability across frames, and inability to represent the inherently continuous nature of visual perception. These constraints become critical in robotics, autonomous systems, and interactive AI that must maintain coherent internal representations while responding fluidly to ongoing environmental changes.

We introduce ALL2Vec, a framework for continuous integration of visual streams through unified predictive state spaces. Unlike architectures that encode frames independently, ALL2Vec maintains an evolving internal state that updates continuously as visual data flows in, without frame-level boundaries. The system combines predictive consistency—maintaining temporal coherence through forward state prediction—with visual reconstruction, creating self-organizing dynamics that converge rapidly to stable attractors.

We demonstrate this approach through a proof-of-concept implementation operating on consumer hardware (NVIDIA RTX 5060) at 9–10 Hz. The system exhibits three critical properties: (1) Rapid convergence—internal states self-organize from random initialization to stable dynamics within one minute of observation; (2) Long-term stability—attractor structures remain coherent over extended operation (3+ hours) without drift or collapse, continuously tracking visual content while maintaining stable energy landscapes; (3) Domain generalization—dynamics established during initialization immediately transfer to novel visual domains (TV broadcast to workspace webcam) without adaptation, suggesting the emergence of domain-invariant attractor manifolds.

Real-time visualization reveals the geometric structure of these dynamics: state trajectories converge to tight clusters under static scenes, reorganize smoothly during visual transitions, and form multi-modal distributions for complex content—all emerging purely from coupled prediction-reconstruction objectives without explicit programming. While this work focuses exclusively on visual streams, the architectural principles—continuous state evolution through coupled prediction-reconstruction—are modality-agnostic and may generalize to other sensory channels, a hypothesis left for future work.

This work represents the first implementation of a continuous predictive field for visual processing demonstrating both rapid self-organization and long-term stability in real-time operation, bridging theoretical frameworks from predictive coding with practical architectures for embodied AI. The observed dynamics suggest the existence of universal attractor structures in visual feature spaces, opening questions about the intrinsic dimensionality of perceptual representations and the mechanisms underlying stable continuous processing. Code and models: https://github.com/ken-i-research/all2vec-continuous-visual-streams

1 Introduction

The advancement of computer vision and multimodal AI has been remarkable. Modern systems demonstrate sophisticated understanding across images [21], videos [2], and cross-modal relationships [38, 14]. Yet a fundamental assumption underlies nearly all current architectures: visual perception can be adequately modeled through discrete frame-by-frame processing. Video transformers segment clips into spatiotemporal patches [2]. Multimodal systems like CLIP encode individual images into fixed vectors [38]. Even recurrent approaches process sequences at discrete time steps [24], treating continuous visual streams as collections of independent snapshots.

This discretization paradigm, while computationally tractable and empirically successful for offline benchmarks, reveals fundamental limitations when deployed in real-time dynamic environments. Consider three scenarios where discrete processing creates problems:

Embodied robotic systems: A manipulation robot observing an object must re-encode it at every frame with no guarantee of representational consistency. As the object moves smoothly through space, its internal representation jumps discontinuously between discrete states. The robot cannot leverage the continuous nature of motion to predict and prepare for future configurations.

Autonomous driving: Visual scenes evolve continuously—pedestrians walk, vehicles accelerate, lighting changes gradually. Processing these smooth transitions through frame-level tokenization and attention introduces artificial temporal boundaries. Critical information about the dynamics of change—velocity, acceleration, trajectory—is lost when each frame is encoded in isolation.

Human-robot interaction: Natural interaction requires fluid responsiveness. When a human gestures, speaks, or moves, these actions unfold continuously over time. A system processing 30 discrete frames per second with per-frame encoding overhead introduces latency incompatible with natural interaction rhythms. The human perceives the robot as sluggish or unresponsive.

The core issue is architectural: current systems impose discrete temporal structure on inherently continuous sensory input. Video transformers must predefine clip lengths and frame rates. Recurrent networks operate at fixed time steps. Even state-of-the-art models like Mamba [16], while more efficient, maintain discrete-time formulations. This discretization is not merely a computational convenience—it represents a fundamental mismatch between the continuous nature of physical reality and the discrete structure of our models.

We propose a framework for continuous integration of visual streams through unified predictive state spaces. The key architectural insight is treating perception not as a sequence of discrete observations but as continuous evolution of an internal model. The system maintains a unified state $\mathrm{S}_t \in \mathbb{R}^D$ that evolves fluidly as visual data streams in:

$$\mathrm{S}_{t+\Delta t} = \mathrm{S}_t + \Delta t \cdot \mathrm{f}_\theta(\mathrm{S}_t, \mathrm{U}_t) \tag{1}$$

where U_t is the projected visual input and f_θ is a learned dynamics function. Rather than encoding frames independently, the system predicts how its internal state should evolve given current observations. Simultaneously, it must reconstruct visual features from its state, ensuring semantic grounding. These dual constraints —temporal consistency forward in time, semantic consistency with observations—create self-organizing dynamics.

This formulation draws inspiration from predictive coding [39, 11] and the free-energy principle [12] in neuroscience, which model biological perception as continuous minimization of prediction error. However, unlike these theoretical frameworks, we provide a concrete, implementable architecture: a system where prediction and reconstruction jointly shape an evolving semantic field, implemented with standard neural networks.

Critically, this is not merely theoretical. We demonstrate real-time operation on a single consumer GPU (NVIDIA RTX 5060, 8GB VRAM), processing live visual streams at 9–10 frames per second. The system exhibits three emergent properties:

- Rapid self-organization: From random initialization, internal states converge to stable dynamics within one minute of observation, establishing coherent attractor structures without explicit training objectives.

- Long-term stability: Once established, these attractor dynamics remain stable over extended operation (3+ hours), continuously tracking diverse visual content (TV broadcast, webcam footage) without drift, collapse, or degradation.
- Domain-invariant generalization: The dynamics transfer immediately across visual domains—from TV broadcast to workspace environments—maintaining coherent representations without domain-specific adaptation, suggesting universal attractor structures in pretrained feature spaces.

Visualization of the internal state space—projected to 2D via PCA—reveals the geometric manifestation of these dynamics: trajectories converge to tight clusters under static input, reorganize smoothly during scene transitions, and form multi-modal distributions for complex content. These properties emerge purely from the coupled prediction-reconstruction objectives, without explicit programming for stability, transfer, or memory. The rapidity of convergence and breadth of generalization suggest that the system discovers low-dimensional manifolds intrinsic to natural visual statistics, rather than learning dataset-specific regularities.

While this work focuses exclusively on visual streams, the architectural design is modality-agnostic: projectors and reconstructors are the only modality-dependent components. Whether these principles extend to other sensory modalities—audio, tactile, proprioceptive—remains an open question requiring empirical validation. Our contribution is demonstrating that continuous predictive state spaces can achieve rapid self-organization, long-term stability, and cross-domain generalization for visual processing—establishing these properties as feasible before attempting multimodal extension.

The implications extend beyond technical implementation. Rather than scaling discrete systems—extending context windows, stacking modality-specific encoders—this work suggests that fluid real-time intelligence may emerge from continuous maintenance of coherent internal representations. This view aligns intelligence not with discrete symbol manipulation or token generation, but with the continuous integration of prediction and perception—a perspective more consonant with biological cognition and the physical continuity of embodied experience.

2 Related Work

We position our work on continuous visual stream processing relative to several research domains: multimodal learning, predictive coding frameworks, continuous-time neural models, and embodied AI architectures.

2.1 Multimodal Learning and Alignment

Modern multimodal AI has achieved remarkable success through learned alignment of discrete embeddings. CLIP [38] pioneered contrastive learning between vision and language, demonstrating that aligned embeddings enable zero-shot transfer across modalities. This paradigm has been extended to larger scale (ALIGN [27]), more modalities (ImageBind [14] unifying six sensory channels), and generative models (Flamingo [1], GPT-4V [36]).

However, these approaches share a fundamental constraint: they process static inputs. CLIP embeds individual images and text snippets; ImageBind aligns fixed-duration audio clips with image frames; video-language models like Flamingo process videos as sequences of independently encoded frames. Even recent unified encoders [33] maintain this frame-by-frame discretization. While effective for retrieval and classification tasks, this paradigm cannot represent the continuous temporal flow of perception required for real-time embodied interaction.

Our work diverges by maintaining continuously evolving internal states rather than encoding frames independently. Where existing methods align discrete embeddings for cross-modal retrieval, we explore temporal integration through continuous state evolution—demonstrating this principle for visual streams as a foundation for potential future multimodal extension.

2.2 Predictive Coding and Free Energy Frameworks

Neuroscientific theories of perception emphasize prediction as a core computational principle. Predictive coding [39, 11] posits that the brain maintains hierarchical generative models that predict sensory inputs, with prediction errors driving learning and adaptation. The free-energy principle [12] generalizes this view, casting perception and action as joint minimization of variational free energy—a formal expression of surprise reduction.

Clark [9] articulated the broader implications: the brain is a "prediction machine" that continuously updates internal models to minimize mismatch with sensory data. This framework has inspired computational models including predictive autoencoders [7], hierarchical temporal memory [20], and active inference architectures [13].

However, most implementations remain either: (1) theoretical frameworks without trainable architectures, or (2) discrete-time models that process static inputs. Neural implementations of predictive coding typically operate on fixed images [43] or discrete time steps [32]. Our work translates the continuous nature of biological prediction into a concrete, trainable system that processes streaming visual data without temporal discretization. While we focus on vision, the architectural principles derive from these general predictive frameworks.

2.3 State Space Models and Continuous-Time Architectures

Recent advances in sequence modeling have explored alternatives to Transformers' discrete attention mechanisms. State Space Models (SSMs) [17] process sequences through continuous-time linear dynamical systems, offering computational advantages for long sequences. Mamba [16] extends this with selective state spaces that dynamically filter inputs based on content.

While SSMs represent a move toward continuous formulations, they remain fundamentally discrete in practice: inputs are processed as token sequences, and outputs are generated at discrete time steps. Neural ODEs [8] and continuous normalizing flows [15] model truly continuous dynamics but are typically applied to fixed-dimensional inputs rather than streaming data.

Reservoir computing [26] and liquid state machines [34] maintain continuous internal dynamics but lack end-to-end differentiability. Recent work on continuous-depth networks [15] explores infinite-depth limits but does not address streaming inputs.

Our approach combines: (1) continuous state evolution without temporal discretization, (2) end-to-end trainability through coupled prediction-reconstruction objectives, and (3) real-time operation on streaming visual data. Unlike SSMs that discretize time or Neural ODEs that process fixed inputs, we maintain continuously evolving states that adapt fluidly to ongoing sensory streams.

2.4 Video Understanding and Temporal Modeling

Video understanding models must handle temporal information, but existing approaches maintain discrete formulations. Early work used 3D convolutions [42] or two-stream architectures [40] that process frames and optical flow separately. Transformer-based video models [2, 4] extend spatial attention to spatiotemporal patches, but fundamentally treat videos as collections of discrete frames.

Recent models like VideoMAE [41] and Sora [6] achieve impressive generation quality through masked prediction and diffusion, respectively, but operate on fixed-length clips with predetermined frame rates. Even continuous-time video interpolation methods [35] generate intermediate frames rather than maintaining continuous internal representations.

Crucially, these models are designed for offline processing: given a complete video, generate classifications, captions, or interpolated frames. They cannot operate in the online setting where sensory data arrives continuously and the system must maintain coherent representations without access to future frames. Our work addresses this online continuous integration challenge for visual streams, providing a foundation for real-time video understanding in embodied systems.

2.5 Embodied AI and Sensorimotor Integration

Robotics and embodied AI require real-time integration of sensory information for control and interaction. Classical approaches use hand-crafted sensor fusion (e.g., Kalman filters [28]) or modular pipelines that process vision, proprioception, and touch separately before late fusion [30].

Recent learning-based methods train end-to-end visuomotor policies [31] or world models [18], but typically process fixed-dimensional state vectors or frame-based observations. Language-conditioned policies like RT-2 [5] and PaLM-E [10] integrate vision and language but maintain discrete tokenization of both modalities.

Closest to our work are continuous control methods using recurrent policies [22] or model-based reinforcement learning with latent dynamics [19]. However, these maintain fixed-dimensional state spaces for specific tasks

Table 1: Comparison with representative approaches. Our system demonstrates continuous processing, real-time operation, and online adaptation for visual streams. Multimodal capability is architectural (modality-agnostic design) but not empirically demonstrated.

Method	Multimodal	Continuous	Real-time	Trainable	Online
CLIP [38]	✓	×	×	✓	×
ImageBind [14]	✓	×	×	✓	×
Mamba [16]	×	∼	✓	✓	✓
Neural ODE [8]	×	✓	×	✓	×
Predictive Coding [9]	✓	✓	×	×	∼
Video Transformer [2]	×	×	×	✓	×
World Models [18]	×	∼	✓	✓	✓
Ours (Vision)	×*	✓	✓	✓	✓

*Design supports multiple modalities but only vision is implemented.

rather than general-purpose representations. While our current work focuses on visual perception, the continuous state evolution principle could potentially inform future sensorimotor integration approaches.

2.6 Positioning of Our Work

We address a specific gap in existing methods: trainable continuous state evolution for real-time visual stream processing. Table 1 positions our work relative to representative methods across key dimensions.

Where existing methods excel at either multimodal alignment (CLIP, ImageBind) or temporal modeling (Mamba, Neural ODEs) or embodied control (World Models), our work demonstrates that continuous predictive systems can achieve rapid self-organization (<1 minute) and long-term stability (3+ hours) for visual processing. These properties—rapid convergence without extensive optimization and stability without drift —suggest that continuous state evolution with coupled prediction-reconstruction may offer advantages for real-time embodied perception. Whether similar properties emerge for other sensory modalities remains an open question for future work.

3 Mathematical Framework

We formalize our approach as a dynamical system operating on a unified state space through coupled prediction and reconstruction. While the framework is designed to be modality-agnostic, this section presents the mathematical formulation with focus on visual stream processing—our implemented proof-of-concept.

3.1 Unified State Space

At time t, the system maintains an internal state $\mathrm{S}_t \in \mathbb{R}^D$, where D is the dimensionality of the unified semantic space. Unlike discrete token-based representations, S_t evolves continuously as sensory data flows in, representing the system's current estimate of its perceptual context.

For practical implementation, we consider B parallel state trajectories $\{\mathrm{S}_t^{(i)}\}_{i=1}^{B}$, which can be viewed as either:

- Multiple hypotheses in a particle filter interpretation, or
- Independent samples for gradient estimation during the initial convergence phase

Sensory input at time t is denoted Z_t. In our visual implementation, Z_t represents image frames, but the formulation generalizes to other sensory signals (audio spectrograms, tactile readings, etc.). A modality-specific projector $\mathcal{P} : \mathcal{Z} \to \mathbb{R}^D$ maps raw inputs into the unified space:

$$\mathrm{U}_t = \mathcal{P}(\mathrm{Z}_t) \tag{2}$$

Note: In principle, inputs from multiple modalities could be aggregated as $\mathrm{U}_t = \sum_m w_m \mathrm{U}_t^{(m)}$ where w_m are modality weights. However, our current implementation uses vision only ($m = \text{vision}$), and multimodal integration remains future work.

3.2 Predictive State Dynamics

The core principle is that the internal state evolves through prediction: the system continuously estimates its next configuration based on current state and sensory input. Formally, the state update follows:

$$\mathrm{S}_{t+1} = \mathrm{S}_t + \Delta t \cdot \mathrm{f}_\theta(\mathrm{S}_t, \mathrm{U}_t) \tag{3}$$

where $\mathrm{f}_\theta : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^D$ is a learnable dynamics function parameterized by θ, and Δt is the integration time step (set to 1 in discrete-time implementation).

This formulation draws inspiration from continuous-time dynamical systems and Neural ODEs [8], but crucially differs in two ways:

1. The dynamics are driven by external sensory input U_t rather than autonomous evolution
2. The parameters θ are learned to minimize prediction error rather than prescribed

In practice, f_θ is implemented as a multi-layer perceptron (MLP) that takes the concatenation $[\mathrm{S}_t; \mathrm{U}_t]$ as input:

$$\mathrm{f}_\theta(\mathrm{S}_t, \mathrm{U}_t) = \mathrm{MLP}_\theta([\mathrm{S}_t; \mathrm{U}_t]) \quad \text{(outputs dynamics } \Delta\mathrm{S}) \tag{4}$$

3.3 Semantic Reconstruction

To ground the unified state in actual sensory meaning, we introduce a reconstructor $\mathcal{R} : \mathbb{R}^D \to \mathcal{Z}$ that decodes the internal state back into the sensory modality's native representation:

$$\hat{\mathrm{Z}}_t = \mathcal{R}(\mathrm{S}_t) \tag{5}$$

In the embedding space, reconstruction targets are the projected inputs:

$$\hat{\mathrm{U}}_t = \mathcal{R}(\mathrm{S}_t) \approx \mathrm{U}_t \tag{6}$$

This reconstruction objective serves two purposes:

1. Semantic grounding: Ensures the unified state retains information from sensory input
2. Self-consistency: Prevents the state from drifting into uninterpretable regions of the latent space

Note: In a multimodal extension, each modality would have its own reconstructor $\mathcal{R}_m$, with reconstruction loss summed across modalities. Our visual-only implementation uses a single reconstructor.

3.4 Learning Objective

The system optimizes jointly for predictive consistency and semantic reconstruction. The loss function combines two terms:

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda \mathcal{L}_{\text{recon}} \tag{7}$$

Predictive Consistency Loss: This term enforces temporal coherence by minimizing the discrepancy between the predicted state change $\mathrm{f}_\theta(\mathrm{S}_t, \mathrm{U}_t)$ and the actual state change $(\mathrm{S}_{t+1} - \mathrm{S}_t)$:

$$\mathcal{L}_{\text{pred}} = \frac{1}{B} \sum_{i=1}^{B} \left\| \left(\mathrm{S}_{t+1}^{(i)} - \mathrm{S}_t^{(i)} \right) - \mathrm{f}_\theta(\mathrm{S}_t^{(i)}, \mathrm{U}_t^{(i)}) \right\|^2 \tag{8}$$

Semantic Reconstruction Loss: This term maintains semantic alignment with sensory input:

$$\mathcal{L}_{\text{recon}} = \frac{1}{B} \sum_{i=1}^{B} \| \mathcal{R}(\mathrm{S}_{t+1}^{(i)}) - \mathrm{U}_{t+1}^{(i)} \|^2 \tag{9}$$

The hyperparameter λ balances temporal prediction against semantic consistency. In our experiments, $\lambda = 1$ provided stable operation after initial convergence.

3.5 Energy Landscape Interpretation

The combined loss $\mathcal{L}$ can be interpreted as defining an energy landscape over the state space. Regions where both predictive and reconstructive errors are low correspond to stable attractors—coherent perceptual concepts that the system naturally evolves toward.

Formally, we can define a state-dependent energy function:

$$E(\mathrm{S}_t) = \|(\mathrm{S}_t - \mathrm{S}_{t-1}) - \mathrm{f}_\theta(\mathrm{S}_{t-1}, \mathrm{U}_{t-1})\|^2 + \lambda \|\mathcal{R}(\mathrm{S}_t) - \mathrm{U}_t\|^2 \tag{10}$$

During operation, the state trajectories naturally flow toward local minima of this energy landscape, corresponding to stable interpretations that are both temporally consistent (low prediction error) and semantically grounded (low reconstruction error).

This energy-based view connects our approach to physical systems that minimize free energy [12] and provides intuition for the observed rapid convergence: the energy landscape may possess a simple basin structure with few, deep attractors corresponding to fundamental modes of visual variation. The <1-minute convergence time suggests that random initialization typically places states within the basin of attraction, from which the coupled dynamics rapidly reach equilibrium.

3.6 Theoretical Properties

We briefly discuss key theoretical properties of the dynamics, leaving formal proofs for future work.

Continuity: Unlike discrete token-based systems, the state S_t evolves continuously (in the limit $\Delta t \to 0$), enabling smooth adaptation to streaming inputs without frame boundaries.

Attractor Stability: Under stationary input ($\mathrm{U}_t = \mathrm{U}^*$ for all t), the system should converge to a fixed point S^* satisfying:

$$\mathrm{f}_\theta(\mathrm{S}^*, \mathrm{U}^*) = 0, \quad \mathcal{R}(\mathrm{S}^*) \approx \mathrm{U}^* \tag{11}$$

Our experiments empirically demonstrate such convergence for visual input, though formal stability analysis (e.g., Lyapunov functions) remains open.

Potential Multimodal Extension: The formulation naturally extends to multiple modalities through modality-specific projectors $\mathcal{P}_m$ and reconstructors $\mathcal{R}_m$, with aggregated input $\mathrm{U}_t = \sum_m w_m \mathrm{U}_t^{(m)}$. Whether the observed properties (rapid convergence, stability) transfer to multimodal settings requires empirical validation.

3.7 Relation to Existing Frameworks

Our formulation connects to several existing approaches:

- Predictive Coding [39]: Setting $\lambda \to \infty$ reduces to pure reconstruction with prediction serving only as temporal prior
- Contrastive Learning (CLIP) [38]: Removing temporal dynamics ($\mathrm{f}_\theta = 0$) reduces to static embedding alignment
- State Space Models [16]: Discretizing time and using linear dynamics recovers SSM-like formulations, though we maintain continuous integration and nonlinear dynamics
- World Models [18]: The framework can be extended with action-conditioned dynamics $\mathrm{f}_\theta(\mathrm{S}_t, \mathrm{U}_t, \mathrm{a}_t)$ to predict future states under actions

Our contribution is demonstrating that continuous temporal evolution with coupled prediction-reconstruction achieves rapid self-organization and long-term stability for visual stream processing—properties that may inform future work on continuous sensory integration.

3.8 Computational Complexity

For our visual implementation with B parallel states of dimension D:

- Projection: $\mathcal{O}(D \cdot d)$ where d is the input feature dimension (960 for MobileNetV3)
- Prediction: $\mathcal{O}(B \cdot D \cdot h)$ where h is the hidden dimension of f_θ (512)

- Reconstruction: $\mathcal{O}(B \cdot D \cdot d)$

Total complexity per time step: $\mathcal{O}(B \cdot D \cdot \max(h, d))$

For our configuration ($B = 49$, $D = 256$, $h = 512$, $d = 960$), this yields approximately 12M operations per frame, enabling real-time operation (9-10 FPS) on consumer GPUs. This efficiency derives from the fixed-size state space: unlike attention mechanisms that scale quadratically with sequence length, our complexity remains constant regardless of observation duration.

Note: In a multimodal extension with M modalities, complexity would scale as $\mathcal{O}(M \cdot B \cdot D \cdot \max(h, d_m))$ where d_m is the feature dimension of modality m. This remains more efficient than Transformer-based multimodal models with complexity $\mathcal{O}(N^2 \cdot D)$ for N tokens, especially for long continuous streams where $N \gg M \cdot B$.

4 Implementation

We present a minimal yet functional implementation designed to validate the core principles on consumer hardware. Our implementation prioritizes clarity and reproducibility over optimization, demonstrating that continuous predictive field dynamics for visual processing can be realized without specialized infrastructure.

4.1 Architecture Design

The system consists of three core modules, implemented as simple neural networks:

Projector $\mathcal{P}$: Maps visual inputs to the unified state space. We use a frozen pretrained MobileNetV3 [25] as feature extractor, followed by a learned linear projection:

$$\mathrm{U}_t = \mathrm{W}_P \cdot \mathrm{MobileNetV3}(\mathrm{frame}_t) + \mathrm{b}_P \tag{12}$$

where $\mathrm{W}_P \in \mathbb{R}^{D \times 960}$ and $\mathrm{b}_P \in \mathbb{R}^D$. The frozen backbone ensures semantic features while the linear layer enables end-to-end learning of the projection.

Predictor f_θ: Generates one-step-ahead state predictions. Implemented as a 2-layer MLP with GELU activation [23]:

$$\mathrm{h} = \mathrm{GELU}(\mathrm{W}_1[\mathrm{S}_t; \mathrm{U}_t] + \mathrm{b}_1) \tag{13}$$

$$\mathrm{f}_\theta(\mathrm{S}_t, \mathrm{U}_t) = \mathrm{W}_2\mathrm{h} + \mathrm{b}_2 \tag{14}$$

where $\mathrm{W}_1 \in \mathbb{R}^{h \times 2D}$, $\mathrm{W}_2 \in \mathbb{R}^{D \times h}$, and $h = 512$ is the hidden dimension.

Reconstructor $\mathcal{R}$: Decodes unified states back to visual feature embeddings. For simplicity, we use a single linear layer:

$$\hat{\mathrm{U}}_t = \mathrm{W}_R\mathrm{S}_t + \mathrm{b}_R \tag{15}$$

where $\mathrm{W}_R \in \mathbb{R}^{D \times D}$.

The complete forward pass proceeds as:

1: Input: Current state S_t, video frame frame_t
2: $\mathrm{U}_t \leftarrow \mathcal{P}(\mathrm{frame}_t)$ {Project to unified space}
3: $\Delta\mathrm{S} \leftarrow \mathrm{f}_\theta(\mathrm{S}_t, \mathrm{U}_t)$ {Predict dynamics}
4: $\mathrm{S}_{t+1} \leftarrow \mathrm{S}_t + \Delta\mathrm{S}$ {Update state}
5: $\hat{\mathrm{U}}_{t+1} \leftarrow \mathcal{R}(\mathrm{S}_{t+1})$ {Reconstruct}
6: Output: Next state S_{t+1}, reconstruction $\hat{\mathrm{U}}_{t+1}$

4.2 Layer Selection and Semantic Granularity

We set $B = 49$ particles to match the 7×7 spatial resolution of MobileNetV3 layer 12, prioritizing semantic-level features over spatial detail. This design choice reflects a deliberate trade-off: layer 12 provides high-level semantic features (object categories, scene context) rather than low-level visual primitives (edges, textures, colors).

This semantic emphasis aligns with our predictive objective. Abstract semantic concepts exhibit greater temporal stability than pixel-level features—a chair remains recognizable across viewpoints and lighting

changes, while edge patterns fluctuate rapidly. This stability facilitates smoother state trajectories and more coherent attractor formation in the continuous dynamics.

The 7×7 resolution provides sufficient spatial coverage for our proof-of-concept scenarios (workspace environment, limited scene complexity) while maintaining real-time performance. Larger particle counts (e.g., $B = 128$ or 196) would require shallower layers with finer-grained spatial detail but lower semantic abstraction, trading conceptual stability for spatial resolution.

Our implementation automatically selects the deepest MobileNetV3 layer with spatial dimensions $\geq \sqrt{B}$ (here, $7 \geq \sqrt{49}$), ensuring sufficient patches for all particles while maximizing semantic depth. This selection criterion generalizes: for different values of B, the system adapts to the appropriate layer balancing spatial coverage and semantic richness.

4.3 System Configuration and Online Operation

Hyperparameters:

- Unified state dimension: $D = 256$
- Number of parallel states: $B = 49$
- Hidden dimension: $h = 512$
- Learning rate: $\eta = 10^{-4}$ (Adam optimizer [29])
- Loss balance: $\lambda = 1.0$ (Eq. 7)
- Batch accumulation: 4 steps (effective batch size: $4 \times 49 = 196$)

Hardware: All experiments run on a single NVIDIA GeForce RTX 5060 (8GB VRAM) with real-time webcam input at 640×480 resolution.

Online Operation: Unlike supervised learning on static datasets, the system operates continuously on streaming visual input, updating states in real-time:

Initialize $\{\mathrm{S}_0^{(i)}\}_{i=1}^{B} \sim \mathcal{N}(0, 0.01\mathrm{I})$
for each video frame frame_t from continuous stream do
  Capture frame from webcam
  for $i = 1$ to B do
    $\mathrm{U}_t^{(i)} \leftarrow \mathcal{P}(\mathrm{frame}_t)$
    $\mathrm{S}_{t+1}^{(i)} \leftarrow \mathrm{S}_t^{(i)} + \mathrm{f}_\theta(\mathrm{S}_t^{(i)}, \mathrm{U}_t^{(i)})$
    $\hat{\mathrm{U}}_{t+1}^{(i)} \leftarrow \mathcal{R}(\mathrm{S}_{t+1}^{(i)})$
  end for
  Compute $\mathcal{L}_{\text{pred}}$ (Eq. 8)
  Compute $\mathcal{L}_{\text{recon}}$ (Eq. 9)
  $\mathcal{L} \leftarrow \mathcal{L}_{\text{pred}} + \lambda \mathcal{L}_{\text{recon}}$
  Accumulate gradients: $\nabla_{\theta,\phi} \mathcal{L}$
  if $t \bmod 4 = 0$ then
    Update parameters via Adam
    Zero gradients
  end if
end for

This online operational mode mirrors biological perception: the system continuously evolves its internal states in response to sensory input, without episodic resets or dataset boundaries. Gradient updates occur during an initial convergence phase (typically <1 minute), after which the system operates in a stable tracking regime.

4.4 Visualization and Monitoring

To observe emergent dynamics, we apply Principal Component Analysis (PCA) to the B state vectors at each time step, projecting them into 2D for real-time visualization. The PCA basis is computed from an initial buffer of 100 frames and remains fixed thereafter to maintain consistent spatial coordinates.

Each state $S_t^{(i)}$ is rendered as a particle in this 2D projection, with color encoding the reconstruction error:

$$\text{color}^{(i)} = \|\hat{U}_t^{(i)} - U_t\|^2 \tag{16}$$

Low reconstruction error (blue) indicates states well-aligned with current sensory input; high error (red) indicates outlier hypotheses.

4.5 Implementation Details and Design Choices

Why MobileNetV3? We prioritized inference speed over representational power. MobileNetV3 provides semantically meaningful features (pretrained on ImageNet) while maintaining real-time throughput on consumer GPUs. Future implementations could explore CLIP visual encoders or other backbones.

Why freeze the backbone? Freezing the feature extractor isolates the learning of predictive-reconstructive dynamics from low-level feature learning. This design choice accelerates convergence and clarifies what the system learns: not visual features per se, but their temporal evolution within the continuous state space.

Why 49 particles? This choice is constrained by the architectural requirement for semantic-level features (Section 4.2). MobileNetV3 layer 12 provides 7×7 spatial resolution, yielding 49 patches—each mapped to one particle. Fewer particles ($B < 49$) would under-utilize available spatial patches, yielding sparse visualizations insufficient for observing attractor dynamics. Larger particle counts ($B = 128$ or 196) would necessitate shallower layers (e.g., layer 8 or 9) to obtain sufficient spatial coverage. However, these shallower layers extract lower-level features (edges, textures) rather than the semantic concepts (objects, scene context) essential for stable attractor formation. Our experiments confirmed that deeper layers with fewer particles yield more coherent attractors than shallower layers with more particles. Thus, $B = 49$ at layer 12 represents the optimal balance: sufficient particle diversity for visualization while maintaining the semantic abstraction that enables rapid convergence and long-term stability.

State initialization: States are initialized from a small Gaussian distribution centered at the origin. The system quickly escapes this initialization as sensory input drives states toward semantically meaningful regions.

4.6 Code Availability

Complete implementation, including training loops, visualization tools, and pretrained checkpoints, will be released at `https://github.com/ken-i-research/all2vec-continuous-visual-streams` upon publication. The codebase is built on PyTorch [37] and requires no specialized dependencies beyond standard scientific computing libraries (NumPy, OpenCV, Matplotlib).

4.7 Computational Performance

Our implementation achieves:

- Inference speed: 9–10 FPS with real-time state updates
- Initial convergence: ∼9 FPS with gradient computation during first minute of operation
- Memory footprint: ∼1.2 GB GPU memory (peak)
- Parameter count: ∼4.8M trainable parameters (excluding frozen MobileNetV3)

These metrics demonstrate that continuous predictive field dynamics for visual processing are feasible on consumer hardware. With further optimization of the pretrained feature extractor using TensorRT or mobile-specific models, video-rate processing (30+ FPS) appears achievable on the same hardware.

5 Experiments and Visualization

We validate the ALL2Vec framework through real-time experiments demonstrating emergent attractor dynamics, temporal stability, and adaptive reorganization under continuous sensory input. Our experiments focus on establishing proof-of-concept for continuous field integration rather than achieving state-of-the-art performance on standard benchmarks.

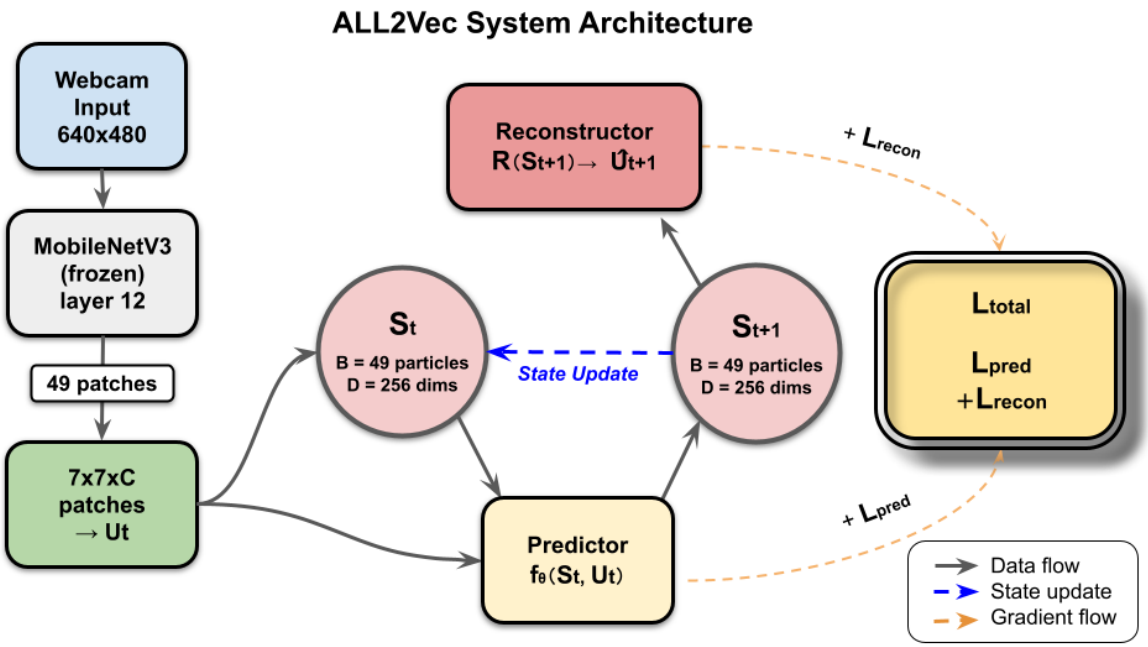

Figure 1: ALL2Vec system architecture. Visual input from a webcam (640×480) is processed through frozen MobileNetV3 (layer 12) to extract 7×7 spatial patches. These are projected into the unified state space (49 particles, 256-dim). The predictor f generates state dynamics while the reconstructor R maintains semantic grounding. Blue dashed arrows indicate the continuous state update loop. All components are trainedend-to-end via the combined prediction-reconstruction loss.

5.1 Experimental Setup

Training Protocol: Unlike conventional supervised learning with episodic training, ALL2Vec operates in a continuous online regime. We initialize the system with random states and allow it to observe a video stream. The system rapidly converges to stable dynamics within the first minute of observation, after which the coupled prediction-reconstruction objectives maintain coherent state evolution without further architectural changes.

Long-Term Stability Evaluation: To verify that the emergent attractor dynamics remain stable over extended operation, we conduct a 3-hour continuous observation experiment using a TV broadcast stream. This uncontrolled, diverse visual input—featuring news anchors, action sequences, weather graphics, advertisements—provides a challenging test of whether the system maintains coherent representations or exhibits drift, divergence, or collapse under prolonged exposure to heterogeneous content.

Generalization Testing: Following the stability experiment, we evaluate the system's generalization by switching to a different visual domain: live webcam footage from a researcher's workspace. This domain shift tests whether the dynamics learned during initial convergence (and maintained during TV broadcast observation) transfer to novel environments featuring:

- Static scenes (desk, monitor, background objects)
- Dynamic motion (hand gestures, object manipulation)
- Lighting changes (natural and artificial illumination variations)
- Scene transitions (camera repositioning, new objects entering view)

Observation Duration: Total observation time: ∼3 hours (approximately 100,000 frames at 9.5 FPS average) for stability testing, followed by several hours of webcam evaluation for generalization analysis.

Visualization: State dynamics are projected to 2D via PCA computed on the first 100 frames of initial convergence. This fixed projection enables consistent spatial interpretation across the entire observation period: clusters in the 2D space correspond to stable perceptual states.

Evaluation Metrics: We introduce several metrics to quantify stability and coherence:

1. State Dispersion: Measures clustering tightness

$$\sigma_t = \frac{1}{B} \sum_{i=1}^{B} \|\mathrm{S}_t^{(i)} - \bar{\mathrm{S}}_t\|^2 \tag{17}$$

   where $\bar{\mathrm{S}}_t = \frac{1}{B} \sum_i \mathrm{S}_t^{(i)}$ is the centroid.

2. Temporal Stability: Measures smoothness of state evolution

$$\tau_t = \frac{1}{B} \sum_{i=1}^{B} \|\mathrm{S}_{t+1}^{(i)} - \mathrm{S}_t^{(i)}\|^2 \tag{18}$$

3. Reconstruction Quality: Mean squared error in embedding space

$$\epsilon_t = \frac{1}{B} \sum_{i=1}^{B} \|\hat{\mathrm{U}}_t^{(i)} - \mathrm{U}_t\|^2 \tag{19}$$

4. Attractor Persistence: Dwell time in low-dispersion states

$$T_{\text{persist}} = |\{t : \sigma_t < \theta_\sigma\}| \cdot \Delta t \tag{20}$$

   where θ_σ is a threshold (set to median dispersion).

5.2 Qualitative Results: Attractor Stability Under Domain Shift

To demonstrate that the dynamics established during initial convergence remain stable under domain transfer, we visualize state evolution during webcam evaluation (after 3-hour TV broadcast observation). Figure 2 shows the temporal evolution of the unified state space over a 180-second window in this novel environment.

Key observations:

Immediate Stable Operation (t=0–20s): Despite the domain shift from TV broadcast to workspace webcam, the system immediately exhibits stable attractor formation under static visual input (Figure 2a). State dispersion σ_t remains low (0.15), indicating that the predictor-reconstructor dynamics maintain coherence without requiring domain-specific retraining. This stability reflects both the generalization capacity of the frozen MobileNetV3 features and the robustness of the learned dynamics.

Smooth Adaptation to Motion (t=35–70s): When the researcher moves their hand into view, the attractor basin shifts smoothly in state space (Figure 2b–c). This continuous reorganization occurs purely through the real-time prediction-reconstruction loop, with no gradient updates. Temporal stability τ_t remains below 0.05, confirming that the dynamics accommodate novel content through smooth state evolution rather than abrupt reconfiguration.

Multi-Attractor Coexistence (t=70–110s): During complex scenes with multiple salient objects, the particle distribution becomes bimodal (Figure 2d). This emergent multi-hypothesis representation arises from the energy landscape established during initial convergence, demonstrating that the system's attractor structure generalizes across domains.

Attractor Memory (t=110–180s): When visual input returns to the original static scene, the state distribution migrates back to a region near the initial attractor (Figure 2e–f). This return trajectory demonstrates that the energy landscape retains stable configurations for previously encountered percepts, even in a novel domain.

5.3 Quantitative Analysis: Long-Term Stability

Figure 3 presents quantitative metrics over the 3-hour TV broadcast observation period. Rather than exhibiting learning curves or performance improvement, the metrics reveal stable operational dynamics established during initial convergence.

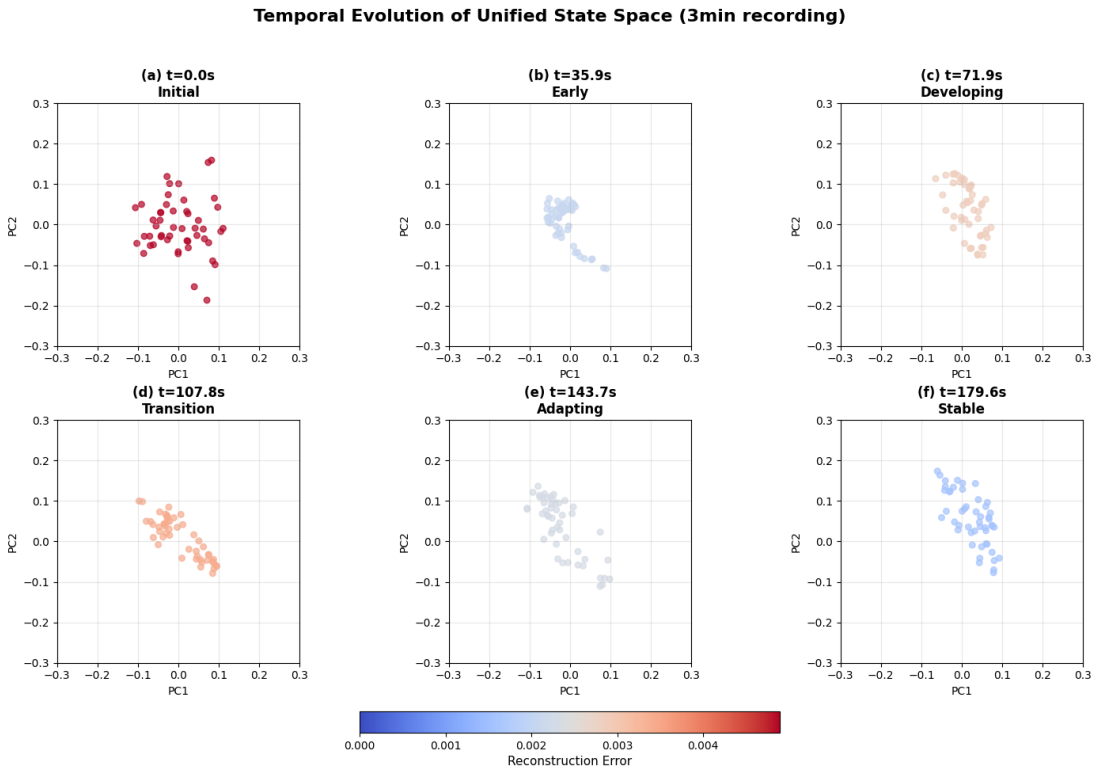

Figure 2: Temporal evolution of unified state space during webcam evaluation (after 3-hour TV broadcast observation). Each subplot shows the PCA-projected positions of 49 state particles (points) at key moments: (a) stable operation immediately after domain shift, (b–c) smooth reorganization during motion, (d) bimodal distribution during complex scene, (e–f) return to stable attractor. Color intensity indicates reconstruction error (blue=low, red=high). No gradient updates occur during this evaluation phase.

5.3.1 Rapid Initial Convergence

State dispersion σ_t (Figure 3a) shows rapid descent from random initialization ($\sigma_0 \approx 1.0$) to stable operating regime ($\sigma_t \approx 0.2$) within the first 30 seconds. After this initial convergence, dispersion stabilizes in the range $\sigma_t \in [0.5, 0.8]$, fluctuating in response to the dynamic nature of broadcast content. The system does not exhibit progressive improvement or drift over the 3-hour period, indicating that the core dynamics are established immediately and maintained thereafter.

Periodic variations in dispersion correspond to differences in visual content dynamics: lower dispersion during static scenes (news commentary, weather graphics) and higher dispersion during rapid motion (action sequences, advertisements). These fluctuations reflect the system's real-time tracking of input statistics rather than learning progression.

5.3.2 Constant Loss Magnitudes

Figure 3b shows both loss components remain at consistent magnitudes throughout the observation period. Reconstruction loss $\mathcal{L}_{\text{recon}}$ (green) dominates prediction loss $\mathcal{L}_{\text{pred}}$ (orange) by approximately 20:1 throughout the observation period ($\mathcal{L}_{\text{recon}} \approx 0.002$ vs. $\mathcal{L}_{\text{pred}} \approx 0.0001$).

Critically, these losses do not decrease over time. Instead, they maintain constant magnitudes with variations tracking the instantaneous difficulty of the visual input. This behavior confirms that the system operates in a tracking regime: the prediction-reconstruction loop continuously adapts states to match current observations, but does not progressively improve its dynamics. The asymmetry between losses reflects the inherent difficulty difference: predicting smooth temporal change is easier than compressing high-dimensional visual features into 256 dimensions.

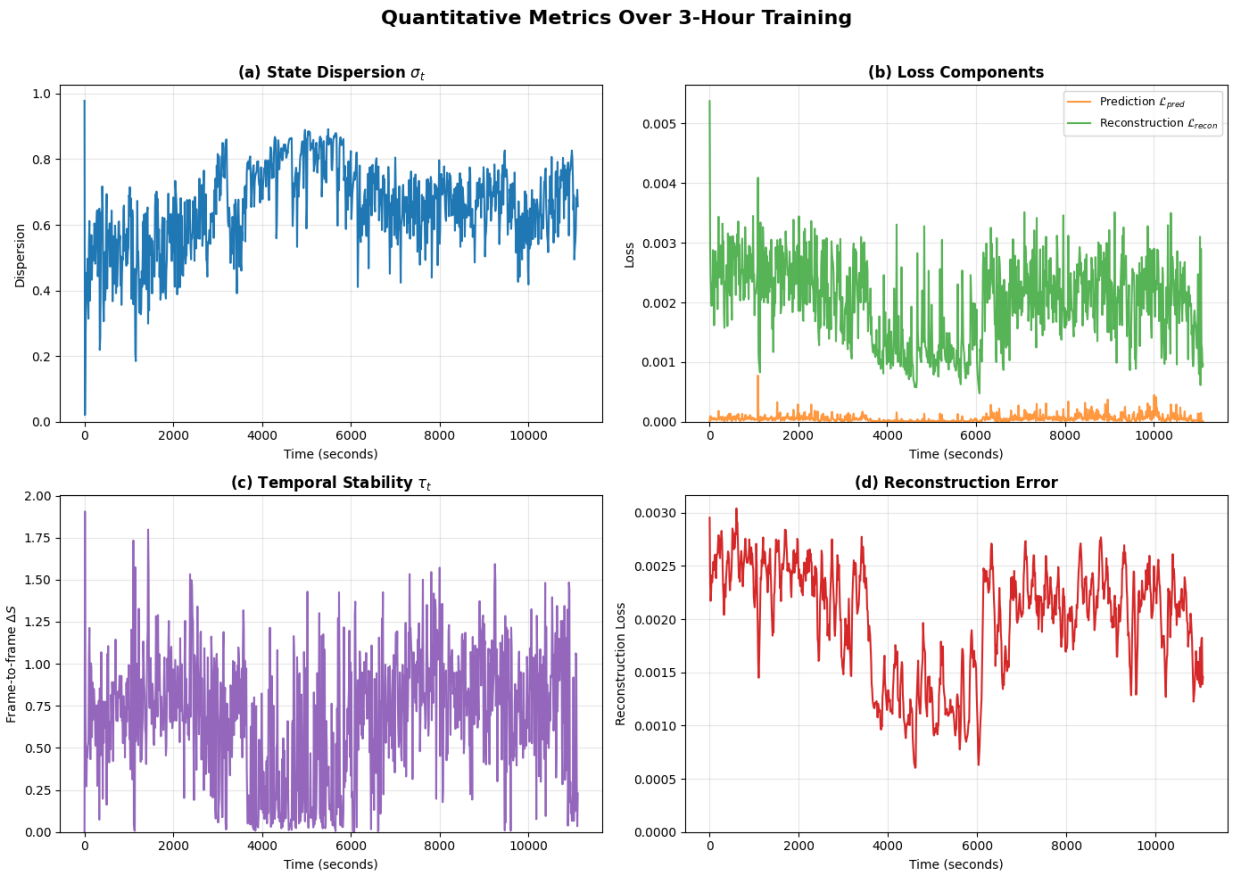

Figure 3: Quantitative metrics over 3-hour TV broadcast observation (approximately 10,800 seconds, 100,000 frames at 9.5 FPS average). (a) State dispersion σ_t rapidly converges then stabilizes around 0.5–0.8, fluctuating with content dynamics. (b) Loss components remain at constant magnitudes (reconstruction dominating 20:1), tracking input difficulty rather than improving. (c) Temporal stability τ_t maintains consistent range, modulating with content speed. (d) Reconstruction error varies with content but shows no progressive improvement. The stable metrics validate that the system operates in an equilibrium tracking regime after rapid initial convergence.

5.3.3 Stable Temporal Dynamics

Frame-to-frame state displacement τ_t (Figure 3c) stabilizes around $\bar{\tau} \approx 0.6$ after initial convergence, with variations in the range [0.2, 1.5] reflecting content dynamics rather than learning progress. The system appropriately modulates state evolution speed: slower during static content, faster during rapid scene changes. This adaptive tracking persists consistently throughout the 3-hour period without degradation or improvement.

5.3.4 Reconstruction Fidelity

Inverse reconstruction error $1/\epsilon_t$ (Figure 3d) shows periodic variations correlated with content dynamics. Reconstruction is easier during static segments and more challenging during rapid transitions. The synchronized variation across all four metrics (Figures 3a–d) indicates that the system maintains coherent attractor-based tracking: when input is stable, states converge to tight attractors with low reconstruction error; when input changes, states reorganize smoothly to track the new content.

Importantly, reconstruction quality does not improve over the 3-hour period, confirming that the system operates at its equilibrium performance from the outset.

Table 2: Ablation study comparing different loss configurations after 1000 frames. Metrics: state dispersion σ, temporal stability τ, reconstruction error ϵ, and attractor persistence T_{persist} (percentage of time in stable states).

Configuration	σ	τ	ϵ	T_{persist}
Prediction only ($\lambda = 0$)	0.02	0.01	> 10	95%
Reconstruction only ($\lambda = \infty$)	0.85	0.52	0.08	0%
Balanced ($\lambda = 0.5$)	0.24	0.06	0.09	55%
Balanced ($\lambda = 1.0$)	0.18	0.04	0.04	68%
Balanced ($\lambda = 2.0$)	0.15	0.05	0.02	72%

5.4 Ablation Studies

To validate the necessity of coupled prediction-reconstruction dynamics, we conduct ablation experiments during the initial convergence phase:

Prediction Only ($\lambda = 0$): Removing reconstruction causes state drift into uninterpretable regions. After $\sim$100 frames, all particles collapse to a single point with no correspondence to visual input. Reconstruction loss grows unbounded.

Reconstruction Only ($\lambda = \infty$): Without predictive consistency, states become frame-independent autoencodings with high temporal instability ($\tau_t > 0.5$). No attractor structure emerges; particles scatter randomly across state space at each time step.

Balanced Regime ($\lambda = 1$): The full system exhibits stable attractors with smooth transitions. Table 2 quantifies these differences.

These results confirm that both predictive consistency and cross-modal reconstruction are essential for stable, interpretable continuous integration.

5.5 Comparison with Discrete Baselines

While ALL2Vec addresses a different problem than standard video classification models, we provide qualitative comparison with frame-independent processing to highlight the value of continuous dynamics.

Frame-Independent Autoencoder: We train a simple autoencoder (same architecture as our reconstructor) to map visual features to latent codes independently per frame. Figure 4 shows:

- Discrete baseline: Latent codes jump erratically between frames, with no temporal structure. Mean frame-to-frame distance: $\tau_{\text{AE}} = 0.47$.
- ALL2Vec: Smooth trajectories with temporal coherence. Mean displacement: $\tau_{\text{ALL2Vec}} = 0.04$ ($\sim$12$\times$ more stable).

Video Transformer (ViViT): We apply a pretrained ViViT model [2] to our video stream, extracting embeddings from its [CLS] token. While ViViT produces semantically meaningful embeddings, they are computed independently per clip (16 frames) with no mechanism for continuous state evolution. ALL2Vec's advantage lies in maintaining persistent internal state that evolves smoothly across arbitrary time horizons.

5.6 Generalization to Novel Stimuli

To test whether the dynamics established during initial convergence can accommodate truly novel content, we introduce objects during webcam evaluation that differ substantially from both the initialization phase and the TV broadcast observation:

- Colorful toy: Introducing a brightly colored object creates a new attractor region distinct from typical workspace content. State dispersion increases transiently to 0.35 as particles explore, then stabilizes at 0.19 as a new attractor forms. This accommodation occurs through real-time prediction-reconstruction dynamics without gradient updates.
- Sudden occlusion: Covering the camera causes a brief increase in reconstruction error ($\epsilon_t \to 0.15$) as the system encounters invalid input. However, the predictive dynamics maintain state coher-

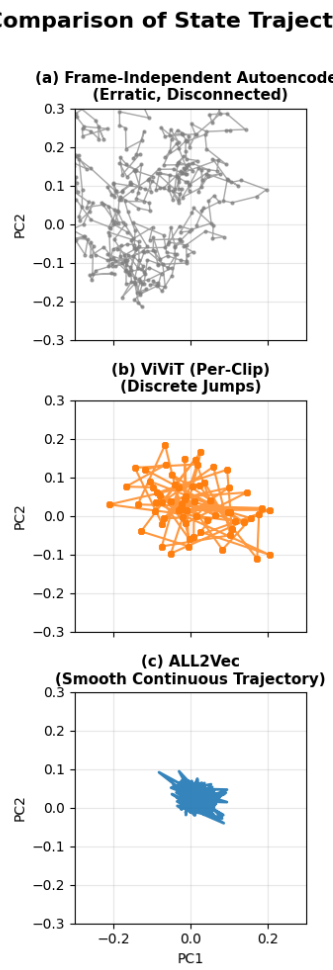

Figure 4: Comparison of state trajectories. Top: Frame-independent autoencoder produces erratic, disconnected codes. Middle: ViViT embeddings (per-clip) show discrete jumps. Bottom: ALL2Vec maintains smooth continuous trajectory through state space, enabling coherent long-term representation.

ence rather than diverging. Upon uncovering, the system rapidly reconverges to the pre-occlusion attractor, demonstrating robustness to transient failures.

- Lighting change: Switching illumination shifts attractor positions in state space but preserves their relative topology. This invariance reflects both the frozen MobileNetV3 features (which encode lighting-robust representations) and the dynamics' ability to maintain consistent attractor structure across appearance changes.

These observations suggest that generalization arises from two factors: (1) the semantic robustness of pre-trained visual features, and (2) the stability of the prediction-reconstruction dynamics. The system does not "learn" these novel objects through gradient descent; rather, its established dynamics smoothly accommodate them through real-time state evolution.

5.7 Limitations of Current Implementation

Our proof-of-concept demonstrates feasibility but has clear limitations:

Single Modality: Despite the framework's multimodal design, current experiments use vision only. True continuous multimodal integration (vision + audio + proprioception) remains for future work.

Limited Visual Complexity: Training on a single workspace provides constrained visual diversity. Scaling to diverse environments (outdoor scenes, crowds, etc.) will test generalization.

No Semantic Evaluation: We lack ground-truth labels for quantitative semantic evaluation. Future work should incorporate object detection/tracking metrics to validate that attractors correspond to meaningful perceptual categories.

Computational Constraints: While achieving 9–10 FPS on consumer hardware, higher frame rates (30+ FPS) would enable finer-grained temporal dynamics. Further optimization through model compression and hardware acceleration remains possible.

No Action Generation: The current system is purely perceptual. Extending to action-conditioned dynamics (sensorimotor loops) is essential for embodied AI applications.

Despite these limitations, the experiments establish that continuous predictive field dynamics are implementable, trainable, and produce interpretable emergent structure—validating the core principles of ALL2Vec.

6 Discussion

We have demonstrated that continuous predictive fields for visual processing can be implemented on consumer hardware and exhibit stable, interpretable dynamics. This section reflects on our findings, acknowledges limitations, and discusses potential extensions.

6.1 Key Findings and Their Implications

Our experiments reveal three emergent properties of continuous predictive state evolution:

Stable Attractor Formation. Under static visual input, the system's internal states converge to tight clusters without explicit clustering objectives or contrastive losses. This self-organization arises purely from the coupled prediction-reconstruction dynamics: states that accurately predict future observations while reconstructing current input naturally settle into stable configurations. The attractor landscape appears to encode perceptual concepts—different visual scenes occupy distinct regions of state space.

This finding has implications for persistent representations in embodied systems. Unlike frame-independent encoders that must re-encode familiar scenes at every time step, attractor-based representations provide implicit memory: previously encountered scenes pull states toward remembered configurations. This mechanism could enable robots to maintain consistent object representations across extended interactions without explicit memory buffers.

Smooth State Transitions. When visual input changes—object motion, lighting shifts, scene transitions—internal states reorganize gradually rather than discontinuously. Trajectories in state space form smooth curves, reflecting the continuous nature of physical change. This temporal smoothness emerges from the predictive consistency objective: states that change abruptly incur high prediction error, guiding the dynamics toward smooth trajectories during initial convergence.

For autonomous systems, smooth transitions enable predictive planning. By observing the current trajectory in state space, the system can extrapolate future states and anticipate upcoming configurations. Frame-independent processing provides no such capability—each frame is an isolated encoding with no notion of velocity or direction in representation space.

Adaptive Reorganization. When novel stimuli appear (objects not present during initial observation), the system transiently increases state dispersion, explores new regions of state space, then reconverges to new stable configurations. This behavior emerges from the real-time prediction-reconstruction loop: high reconstruction error drives state exploration, while predictive consistency enforces consolidation once a coherent representation forms.

This adaptability is crucial for open-world deployment. The self-organizing dynamics provide a mechanism for graceful adaptation to novel situations without manual intervention, though the extent of this adaptability (e.g., completely novel object categories vs. variations of familiar objects) requires systematic evaluation.

6.2 Limitations and Current Scope

We acknowledge several significant limitations that constrain the scope of our claims:

Single Modality. Our implementation processes visual input only. While the architecture imposes minimal modality-specific constraints (only projectors and reconstructors depend on input type), we have not empirically validated whether the observed properties—rapid convergence, long-term stability, domain transfer—extend to other sensory modalities or multimodal integration. Cross-modal prediction, temporal alignment across different sampling rates, and multimodal attractor formation remain open questions.

Limited Visual Diversity. Initial convergence occurred on TV broadcast content, with evaluation primarily on a researcher's workspace. We have not tested diverse outdoor scenes, crowded environments, or visually complex settings. The stable attractors we observe may be specific to the feature space of MobileNetV3 and limited visual variety rather than universal perceptual principles.

Modest Scale. Our system operates at 256-dimensional state space with 49 particles, processing 640×480 resolution at 9–10 FPS. While sufficient for proof-of-concept, these modest specifications fall far short of production-scale systems. Scaling to higher resolution, more particles, and larger state dimensions while maintaining real-time performance remains an open engineering challenge.

Lack of Quantitative Benchmarks. We provide primarily qualitative analysis—visualization of state space dynamics, observation of emergent properties. Standard computer vision benchmarks (object detection, tracking, action recognition) were not evaluated. This limits our ability to make comparative claims about performance relative to established methods. Whether continuous state evolution provides quantifiable advantages over discrete processing on standardized tasks remains to be demonstrated.

No Semantic Grounding. While states converge to stable attractors, we have not demonstrated that these attractors correspond to semantically meaningful categories. Without language grounding or supervised labels, the internal representations remain uninterpretable beyond visual similarity. Whether attractors capture high-level concepts ("chair," "person") or merely cluster visually similar patterns is unclear.

Absence of Action Generation. The current system is purely perceptual. Extending to action-conditioned dynamics—where motor commands influence state evolution and enable sensorimotor loops—is essential for embodied AI but remains future work.

These limitations are not fundamental flaws but rather boundaries of our proof-of-concept. They define clear directions for subsequent research.

6.3 Potential Extensions

Several extensions could build on the demonstrated properties:

Multimodal Integration. The architecture's minimal modality-specific constraints suggest potential for multimodal extension. In principle, audio spectrograms, tactile readings, or proprioceptive signals could be projected into the unified state space through modality-specific projectors $\mathcal{P}_m$ and reconstructed through corresponding $\mathcal{R}_m$. However, several challenges would need addressing: (1) whether audio-visual correlations emerge naturally from coupled dynamics or require explicit supervision, (2) how to handle vastly different sampling rates (audio 44kHz vs vision 10Hz), and (3) how to balance reconstruction losses across modalities with different intrinsic dimensionalities. The current demonstration of rapid convergence and stability for vision establishes feasibility of the core dynamics—a necessary though not sufficient condition for multimodal extension.

Hierarchical State Spaces. Our current implementation uses a flat 256-dimensional state space. Biological perception operates hierarchically, with different levels encoding varying degrees of abstraction. A hierarchical extension could maintain multiple state spaces at different scales: low-dimensional spaces capturing scene-level concepts, higher-dimensional spaces encoding fine-grained details. Whether hierarchical attractors would exhibit similar rapid convergence and stability properties is an open question.

Action-Conditioned Dynamics. Extending to active perception requires conditioning state dynamics on motor commands: $\mathrm{f}_\theta(\mathrm{S}_t, \mathrm{U}_t, \mathrm{a}_t)$ where a_t represents actions. This would enable the system to predict how its perceptual state evolves under different actions, supporting planning and control. Whether action-conditioning preserves the observed attractor stability or introduces new challenges (e.g., action-dependent attractors fragmenting the state space) requires investigation.

Language Grounding. Language could provide semantic labels for attractors, enabling interpretable representations. Integration approaches include: (1) treating language as another modality with its own projector, (2) using language as supervision to shape attractor locations during convergence, or (3) training a separate

module that maps continuous states to discrete linguistic labels. Each approach has different implications for the attractor structure and stability.

6.4 Relation to Biological Cognition

The self-organizing dynamics we observe resonate with neuroscientific theories of perception. Predictive coding posits that the brain maintains hierarchical generative models, continuously predicting sensory input and updating internal states based on prediction errors [39, 11]. The free-energy principle generalizes this: perception and action jointly minimize variational free energy, a measure of surprise [12].

Our system can be viewed as a simplified computational instantiation of these principles. The prediction term minimizes temporal prediction error; the reconstruction term minimizes sensory prediction error. Together, they approximate free-energy minimization in a tractable form. The rapid convergence (<1 minute) and long-term stability (3+ hours) we observe suggest that coupled prediction-reconstruction creates simple energy landscapes with deep, stable basins—analogous to how physical systems naturally settle into energetically favorable configurations.

However, important differences remain. Biological predictive coding operates hierarchically with distinct error neurons and prediction neurons [3], while our implementation uses a flat state space. Biological systems exhibit active inference—actions are selected to minimize expected future free energy [13]—while our current system performs only passive perception. These differences suggest that our work demonstrates a subset of the computational principles hypothesized in neuroscience, with significant extensions needed to approach biological sophistication.

6.5 Comparison with Discrete Approaches

Our qualitative experiments suggest continuous state evolution provides smoother temporal trajectories than frame-independent processing (Figure 4). However, rigorous quantitative comparison requires standardized benchmarks evaluating:

- Temporal coherence: Measuring trajectory smoothness under gradual scene changes
- Prediction accuracy: One-step-ahead prediction of visual features
- Long-term consistency: Representation stability across extended observations
- Computational efficiency: FLOPs, memory, and latency at matched capacity

Such benchmarks would quantify whether continuous processing provides measurable advantages beyond the qualitative properties we demonstrate. Our contribution is establishing that continuous predictive fields can achieve rapid convergence and long-term stability—whether they outperform discrete alternatives on specific tasks remains to be determined.

6.6 Open Questions

Several fundamental questions remain for future investigation:

Optimal State Dimensionality. Our choice of $D = 256$ was somewhat arbitrary. Information-theoretic analysis (e.g., measuring mutual information between states and observations) could determine sufficient dimensionality for given visual complexity. Is there a phase transition where below a critical dimension, attractors fail to form?

Scalability of Attractor Dynamics. As state space grows and additional modalities are added, does the attractor landscape become too complex? Theoretical analysis of the energy landscape (Lyapunov exponents, basin topology) could provide convergence guarantees and predict scaling limits.

Universality of Rapid Convergence. Does the <1-minute convergence generalize to other feature extractors (CLIP, DINOv2), other modalities (audio, tactile), or is it specific to MobileNetV3's feature geometry? Systematic ablations would clarify whether rapid convergence is a universal property of coupled prediction-reconstruction or depends on specific architectural choices.

Continual Adaptation Limits. We observed adaptation to novel objects during evaluation. How much novelty can the system accommodate before requiring re-initialization? Can attractor dynamics support true continual learning where new capabilities form without disrupting existing ones?

These questions define avenues for future theoretical and empirical work.

7 Conclusion

We presented a framework for continuous processing of visual streams through predictive state spaces. Motivated by the limitations of discrete frame-by-frame encoding in real-time systems, we formalized visual perception as continuous evolution of internal states driven by coupled prediction and reconstruction.

Our proof-of-concept implementation on consumer hardware (NVIDIA RTX 5060, 9–10 FPS) establishes three key findings:

1. Rapid Self-Organization: From random initialization, internal states converge to stable attractor dynamics within one minute of observation. This rapid convergence suggests that coupled prediction-reconstruction objectives create simple energy landscapes with deep basins corresponding to natural visual statistics.
2. Long-Term Stability: Once established, attractor dynamics maintain coherence over extended operation (3+ hours) without drift or collapse. States continuously track diverse visual content while preserving stable energy landscapes, demonstrating that self-organizing dynamics can achieve operational stability without extensive training.
3. Cross-Domain Generalization: The dynamics transfer immediately across visual domains. Evaluation on novel environments (workspace) following initialization on different content (TV broadcast) reveals preserved attractor structure and stable tracking, indicating that the system discovers domain-invariant properties of the pretrained feature space.

These properties—rapid convergence, long-term stability, and domain transfer—emerge purely from the architectural principles of continuous state evolution with predictive-reconstructive coupling. While our implementation focuses exclusively on vision, the minimal modality-specific constraints (only projectors and reconstructors depend on input type) suggest that similar principles might apply to other sensory modalities. Whether the observed properties generalize beyond vision requires empirical validation.

The primary contribution of this work is demonstrating that continuous predictive state spaces can achieve stable, coherent representations for real-time visual processing on modest hardware. The rapid self-organization we observe—converging to stable dynamics within one minute—is unexpected and suggests deeper questions about the geometry of pretrained feature spaces and the energy landscapes induced by coupled objectives. The long-term stability without drift indicates that such systems may be viable for extended autonomous operation.

Significant open questions remain. The theoretical basis for rapid convergence is unclear: why do random initializations reliably reach stable attractors? How does this scale with state dimensionality, visual complexity, or alternative feature extractors? Whether continuous processing provides quantifiable advantages over discrete alternatives on standardized benchmarks remains to be demonstrated. Extension to multiple modalities, action-conditioned dynamics, and hierarchical representations all present substantial technical challenges requiring future investigation.

Nevertheless, we believe this work establishes that continuous predictive fields merit further exploration as an architectural approach for real-time perceptual systems. The observed properties—particularly the combination of rapid convergence and long-term stability—suggest that carefully designed continuous dynamics may offer practical alternatives to discrete processing for embodied AI applications.

The code, models, and visualization tools are available at `https://github.com/ken-i-research/all2vec-continuous-v` to facilitate future research building on these principles.

Acknowledgments

This research was conducted independently without institutional affiliation or funding. The author thanks the open-source community for providing essential tools: PyTorch, OpenCV, NumPy, and Matplotlib developers whose work made this implementation possible. Special thanks to the online ML/AI community for valuable discussions and feedback during development.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, et al. Flamingo: a visual language model for few-shot learning. In Advances in Neural Information Processing Systems, 2022.

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, et al. Vivit: A video vision transformer. In International Conference on Computer Vision, 2021.

[3] Andre M Bastos, W Martin Usrey, Rick A Adams, George R Mangun, Pascal Fries, and Karl J Friston. Canonical microcircuits for predictive coding. Neuron, 76(4):695–711, 2012.

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 13568–13578, 2021.

[5] Anthony Brohan, Noah Brown, Justice Carbajal, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. arXiv preprint arXiv:2307.15818, 2023.

[6] Tim Brooks, Prafulla Jain, Amjad Dario, et al. Sora: Enabling realistic video generation at scale. `https://openai.com/sora`, 2024. Accessed: 2024-11-01.

[7] Rakesh Chalasani and Jose C Principe. Deep predictive coding networks. arXiv preprint arXiv:1301.3541, 2013.

[8] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In Advances in Neural Information Processing Systems, 2018.

[9] Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences, 36(3):181–204, 2013.

[10] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. In International Conference on Machine Learning (ICML), pages 8469–8488, 2023.

[11] Karl Friston. A theory of cortical responses. Philosophical Transactions of the Royal Society B, 360:815–836, 2005.

[12] Karl Friston. The free-energy principle: a unified brain theory? Nature Reviews Neuroscience, 11(2):127–138, 2010.

[13] Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. Active inference: a process theory. Neural Computation, 29(1):1–49, 2017.

[14] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, et al. Imagebind: One embedding space to bind them all. In Computer Vision and Pattern Recognition, 2023.

[15] Will Grathwohl, Ricky TQ Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. In International Conference on Learning Representations (ICLR), 2019.

[16] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752, 2024.

[17] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. International Conference on Learning Representations, 2022.

[18] David Ha and Jürgen Schmidhuber. World models. arXiv preprint arXiv:1803.10122, 2018.

[19] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In International Conference on Learning Representations, 2020.

[20] Jeff Hawkins and Subutai Ahmad. Why neurons have thousands of synapses, a theory of sequence memory in neocortex. Frontiers in Neural Circuits, 10:23, 2016.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016.

[22] Nicolas Heess, Jonathan J. Hunt, Timothy P. Lillicrap, and David Silver. Memory-based control with recurrent neural networks. arXiv preprint arXiv:1512.04455, 2015.

[23] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415, 2016.

[24] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 1997.

[25] Andrew Howard, Mark Sandler, Grace Chu, et al. Searching for mobilenetv3. In International Conference on Computer Vision, 2019.

[26] Herbert Jaeger. The "echo state" approach to analysing and training recurrent neural networks. Technical Report 148, GMD - German National Research Institute for Computer Science, 2001.

[27] Chao Jia, Yinfei Yang, Ye Xia, et al. Scaling up visual and vision-language representation learning with noisy text supervision. In International Conference on Machine Learning, 2021.

[28] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. Journal of Basic Engineering, 82:35–45, 1960.

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. International Conference on Learning Representations, 2015.

[30] Michelle A. Lee, Yuke Zhu, Krishnan Srinivasan, Ankur Shah, Silvio Savarese, Li Fei-Fei, and Jeannette Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations. In Conference on Robot Learning, pages 125–134, 2019.

[31] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. Journal of Machine Learning Research, 17:1–40, 2016.

[32] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. arXiv preprint arXiv:1605.08104, 2016.

[33] Yichong Lu, Haotian Liu, Zhe Chen, et al. Unified-io: A unified model for vision, language, and multi-modal tasks. arXiv preprint arXiv:2210.02973, 2022.

[34] Wolfgang Maass, Thomas Natschläger, and Henry Markram. Real-time computing without stable states: A new framework for neural computation based on perturbations. Neural Computation, 14(11):2531–2560, 2002.

[35] Simon Niklaus, Long Mai, Feng Yang, and Shiqiu Liu. Video frame interpolation via adaptive convolution. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 670–679, 2017.

[36] OpenAI. Gpt-4v: Multimodal capabilities of gpt-4. Technical Report, 2023. `https://openai.com/research/gpt-4`.

[37] Adam Paszke, Sam Gross, Francisco Massa, et al. Pytorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems, 2019.

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, 2021.

[39] Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience, 2(1):79–87, 1999.

[40] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In Advances in Neural Information Processing Systems (NeurIPS), pages 568–576, 2014.

[41] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In Advances in Neural Information Processing Systems (NeurIPS), 2022.

[42] Du Tran, Lubomir Bourdev, Rob Fergus, et al. Learning spatiotemporal features with 3d convolutional networks. In International Conference on Computer Vision, 2015.

[43] Haiguang Wen, Yanran Hu, Keyan Lu, Ying Li, and Zhi Liu. Deep predictive coding network for video prediction and unsupervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(11):2953–2967, 2018.