

UNITY Framework: Continuous Neural Field Architecture for Universal Multimodal Learning Without Attention or Recurrence

Ian Patel*
Spyder Sync

August 13, 2025

Abstract

The UNITY Framework is a neural architecture that models diverse input modalities as continuous, evolving fields updated through localized interaction rules. It eliminates traditional tokenization, attention mechanisms, and recurrence, instead relying on field diffusion equations with learned flow updates and a content-aware streaming memory system. The framework unifies text, image, audio, and video processing into a single architecture, achieving transformer-level accuracy with sublinear memory growth and near-constant per-step computational complexity, enabling deployment on low-resource hardware.

1 Technical Field

This invention relates to machine learning and artificial intelligence, specifically to architectures for efficient multimodal neural processing with continuous field representations.

2 Background

Existing architectures such as Transformers rely heavily on tokenized representations and global attention operations, which incur:

- $O(N)$ memory scaling with sequence length.
- High compute cost due to large matrix multiplications.
- Separate architectural designs for different modalities.

State Space Models (SSMs) reduce memory complexity but remain bound to sequential processing constraints and modality-specific designs.

*Inventor, Founder & CEO of Spyder Sync. Email: ianpatel15c@gmail.com

3 Summary of the Invention

The UNITY Framework addresses these limitations by introducing:

1. **Continuous Field Representation** – All modalities are mapped into a shared spatiotemporal field space without tokenization.
2. **Local Field Diffusion + Learned Flow** – Updates follow discretized partial differential equations with trainable terms, eliminating global attention.
3. **Content-Aware Stream Memory** – Unbounded, non-decaying memory slots dynamically update based on semantic similarity.
4. **Universal Encoders/Decoders** – Single encoder-decoder family supports all modalities without specialized designs.

4 Detailed Description

4.1 Universal Field Encoders

Any input modality X is transformed into a spatiotemporal field representation $H(\mathbf{x}, t) \in \mathbb{R}^d$.

- **Text**: Byte-level continuous embeddings with normalization flows.
- **Image**: Local Laplacian embedding of pixel patches.
- **Audio**: Spectral flow transforms with temporal derivatives.
- **Video**: Optical-flow-enhanced motion fields plus residual encoding.

4.2 Flow Layers (Core Innovation)

Updates follow a generalized diffusion equation:

$$\frac{\partial H}{\partial t} = \alpha \nabla^2 H + F(H, \nabla H) \quad (1)$$

where:

- $\nabla^2 H$ = Laplacian operator for local neighborhood smoothing.
- $F(H, \nabla H)$ = trainable neural flow function for content-specific updates.

The discrete form is:

$$H(\mathbf{x}, t+1) = H(\mathbf{x}, t) + \alpha \sum_j (H(\mathbf{x}_j, t) - H(\mathbf{x}, t)) + F(H(\mathbf{x}, t), \nabla H(\mathbf{x}, t)) \quad (2)$$

4.3 Stream Memory

Dynamic memory slots M_k persist across processing steps without decay:

$$M_k(t+1) = M_k(t) + G(H(\mathbf{x}, t), M_k(t)) \quad (3)$$

where G is a similarity-based gating function.

4.4 Output Decoding

- **Text:** Gradient-guided field collapse into symbol sequences.
- **Image:** Inverse diffusion for spatial reconstruction.
- **Structured Data:** Field curl operations for graph/tree outputs.

5 Advantages

- **Token-free:** Removes discretization bottlenecks.
- **$O(1)$ Memory per Step:** Suitable for extremely long contexts (1M+ steps).
- **Multimodal Native:** Processes all modalities with same core mechanism.
- **Hardware-friendly:** Runs on low-end GPUs and CPUs with near-constant step cost.

6 Claims

1. A machine learning architecture comprising:
 - a universal encoder configured to map multiple input modalities into a continuous spatiotemporal field;
 - a plurality of flow layers configured to update the field using localized diffusion and a trainable flow function;
 - a streaming memory module configured to retain and update content representations based on semantic similarity without fixed decay; and
 - a decoder configured to transform the updated field into an output corresponding to the original modality or a target modality.
2. The architecture of claim 1, wherein the localized diffusion is implemented via a discrete Laplacian operator over neighborhood field elements.
3. The architecture of claim 1, wherein the trainable flow function applies a learned update rule based on local gradients of the field.
4. The architecture of claim 1, wherein the universal encoder supports text, images, audio, and video without modality-specific structural changes.
5. The architecture of claim 1, wherein the streaming memory module employs similarity-based gating to update long-term memory slots with new field information.
6. A method of processing multimodal input data, comprising:
 - mapping the input to a continuous spatiotemporal field;
 - iteratively applying localized diffusion and trainable flow updates to the field;
 - updating a content-aware streaming memory module with field state information;
 - decoding the final field state into one or more output modalities.

7 Potential Applications

- General-purpose AI models
- Low-resource multimodal assistants
- Edge device machine vision
- Real-time transcription and translation
- Scientific simulation acceleration