ChatGPT

Robust Toroidal Attention Representations: Insights from NeRFs, Transformers, and Holographic Memory

Introduction

Modern AI research is converging on innovative paradigms for **robust and efficient internal representations**. Key approaches include **neural implicit scene representations** (exemplified by Neural Radiance Fields, or NeRFs), **Transformer-based world models** leveraging attention, and **brain-inspired holographic/hyperdimensional memory systems**. Each offers unique strengths – *NeRFs* provide continuous, coordinate-based encodings of complex spaces, *Transformers* excel at flexible attention over sequences or features, and *hyperdimensional (holographic) memory* distributes information across high-dimensional vectors for noise resilience [1] [2]. Researchers are even exploring *topological representations* (e.g. using a **torus** for inherently periodic data) to eliminate edge effects in modeling cyclic patterns [2] [3]. In this survey, we synthesize recent (2022–2025) advances across these areas. Our goal is to inspire **ML-friendly strategies to represent the "Donut's" toroidal attention state** – i.e. a latent attention map with toroidal (doughnut-like) topology – in a **continuous, robust, and manipulable** form.

Neural Scene Representations (NeRFs and Implicit Fields)

Neural Radiance Fields (NeRFs) have revolutionized 3D scene representation by using *implicit neural functions* to model space. Unlike traditional 3D data structures (voxels, point clouds, meshes) which suffer from discretization or integration issues, implicit neural representations offer **continuous functions parameterized by neural networks** [4]. A NeRF maps input coordinates (a 3D position along with a viewing direction) to an output color and density, effectively encoding the entire scene as a continuous volumetric field [5]. Through differentiable volume rendering, NeRF learns photorealistic scenes from 2D images alone [5], demonstrating that a neural network can **compactly store a complex 3D structure** and interpolate smoothly between observations. This implicit approach – often enhanced by sinusoidal position encodings to capture high-frequency details – naturally handles *periodic continuity* (e.g. representing angles or rotations), a property we can harness for toroidal domains.

Research since the original 2020 NeRF has greatly extended the idea. NeRF variants now handle **dynamic scenes, larger environments, and semantic content** [6]. Notably, implicit scene representations are being integrated into **world models for reinforcement learning**. For example, Shim *et al.* (2023) introduce SNeRL, which uses a NeRF-based latent encoder to provide an agent with a **3D-aware state representation** from images [7]. This NeRF-derived representation, especially when augmented with semantic features, led to improved data efficiency and performance on complex control tasks compared to conventional 2D CNN encoders [7] [8]. These advances suggest that NeRF-like networks can serve as general **coordinate-based memory**: given a coordinate (which could encode spatial location, time, or other indices on a manifold), the network outputs the associated content. **Implementing a toroidal attention map** could likewise use a NeRF-style implicit function defined over a 2D torus (topologically, the product of two circles). In practice, this means using periodic coordinate encodings (e.g. sine/cosine for angles) and training a small MLP to represent the attention value at any continuous coordinate on the torus. Such a model would store the attention state as a smooth field on a donut-shaped domain, inherently eliminating boundary discontinuities (since the input space is circular in each dimension).

Transformer World Models and Attention Mechanisms

Transformers have become central to sequence modeling and world models, thanks to the self-attention mechanism that flexibly connects information across time and space. In world-model architectures (e.g. for model-based RL or video prediction), Transformers offer the ability to attend over long histories and multiple modalities. A recent example is the **Transformer-based World Model (TWM)**, which uses a Transformer-XL backbone to learn an environment's dynamics for an RL agent [9]. The transformer operates over sequences of latent states, actions, and even rewards – in fact, TWM feeds *predicted rewards back into the sequence* – allowing the model to attend to **compact state representations and their outcomes** across time [10]. This design lets the agent **imagine trajectories** in latent space and evaluate them, yielding strong sample-efficiency gains over prior RNN-based approaches [9] [11]. The takeaway is that attention enables flexible, on-the-fly retrieval of relevant parts of an agent's experience (or imagined experience) without rigid memory slots.

When considering a **toroidal attention state**, Transformers can be adapted to respect the torus topology. In conventional sequence models, positional encodings are often linear; however, if the sequence or memory has a natural wrap-around (e.g. angles 0° and 360° being equivalent), we can use **periodic positional encodings** or coordinate embeddings on a circle. This ensures the attention mechanism "knows" that the end of the sequence is adjacent to the beginning in a cyclic sense. Indeed, research on *toroidal architectures* shows that using a torus can eliminate boundary artifacts – e.g. image models on a toroidal grid treat the left and right edges as neighbors [12] [3]. For temporal data, one could map time-of-day or phases of a cycle onto a circular manifold so that a transformer model seamlessly attends across the 24h boundary or the end of a game level, etc. More concretely, we might arrange the attention state as a 2D array (or 1D ring) and *tile it* or use modulo indexing so that attention weights between "wrap-around" positions are as strong as local neighbors. This way, a **continuous attractor** can be formed within the transformer's latent memory: the attention focus could move around the torus indefinitely. The idea is analogous to how **grid cells** in the brain maintain an activity "bump" on a toroidal manifold of possible positions [13] [14]. The transformer's job would be to update and query this bump (the attention distribution) based on inputs, without ever encountering a hard boundary – effectively treating the latent state space itself as doughnut-shaped.

Another role for Transformers is to interface with more explicit memory systems. Transformers can perform **content-based addressing** (via attention queries and keys), which aligns with how one might query a continuous or distributed memory of the toroidal state. We will discuss next how a memory could be structured to support such queries robustly. For now, it's clear that Transformers provide the "glue" that can connect high-level decisions to low-level state representations: they can **read from and write to a specialized memory** (be it a coordinate-based NeRF or a hyperdimensional vector) by learning appropriate attention operations. In summary, Transformers endow the system with **sequence prediction, integration of modalities (state, action, reward)**, and the ability to **attend to relevant parts of a large representational space** – all of which are needed to manipulate a complex toroidal attention state over time.

Holographic Memory and Hyperdimensional Embeddings

While NeRFs and Transformers address *continuous representation* and *attention/sequence modeling* respectively, **Holographic or Hyperdimensional memory** addresses *robustness and associative retrieval*. **Hyperdimensional Computing (HDC)**, also known as **Vector Symbolic Architecture (VSA)**, represents information as very high-dimensional vectors (often thousands of dimensions) [15]. These *hypervectors* use **distributed encoding**: each piece of information is smeared across all dimensions, and conversely each dimension participates in representing many pieces. The power of HDC comes from a set of algebraic operations on hypervectors: - **Bundling (superposition)**: combining multiple vectors into one (e.g. by addition) to represent a set or collection. - **Binding**: associating two pieces of information (e.g. via component-wise multiplication or circular convolution), which produces a new vector encoding their relationship [15]. - **Permutation**: shuffling or rotating a vector's components to encode sequences or roles (e.g. to distinguish "item A in position 1" vs "item A in position 2") [15].

Crucially, these operations produce results that remain in the *same vector space*, allowing iterative composition and easy comparison via similarity (dot product, Hamming distance, etc.). The **strengths of hyperdimensional representations** include remarkable **robustness to noise and failure** – because information is redundant across thousands of dimensions, random noise or deletion of some components does not wipe out the encoded content [16]. In effect, high-dimensional distributed codes have error-correcting properties: even if a stored memory is slightly corrupted, it can often be decoded or recognized since the **statistical signal of the pattern** still stands out above noise. They are also computationally efficient on modern hardware (vector operations that can be parallelized) and naturally suited for **associative memory tasks** like pattern completion and similarity search [16].

To illustrate, consider a **holographic memory** analog: in cognitive science, holographic models of memory (e.g. Holographic Reduced Representations by Plate) allow multiple memories to be superposed in one vector, and a cue can retrieve a specific memory by **pattern interference** rather than by a direct index lookup [17]. For example, if we encode each memory item as a high-dimensional wave pattern, storing many items means summing all their vectors into one "hologram." When a query pattern (cue) is presented, it will **resonate** with the matching memory and produce a maximal similarity, effectively reading out that item without ever storing addresses [17] [18]. This is analogous to how an optical hologram stores images: the image is distributed across the entire holographic plate, and shining the correct reference beam on it causes the desired image to emerge. Even fragments of the hologram can reproduce the whole image (with some loss of clarity), demonstrating *graceful degradation*. Similarly, hyperdimensional memories degrade gracefully – remove or flip a few percent of the bits of a hypervector, and it will still be closer to the true pattern than to a random vector, so the system can still recognize/reconstruct it.

Recent research validates the benefits of hyperdimensional/holographic approaches. For instance, **SpikeHD** (Zou *et al.*, 2022) combined spiking neural networks with hyperdimensional computing to create an efficient online learning system. The result was a model that **significantly improved learning performance, achieved substantial robustness to noise and hardware bit failures, and required fewer parameters** than a purely spiking neural net [19] [20]. The HDC component in SpikeHD maps SNN outputs into high-dimensional space and performs classification there, effectively leveraging the error tolerance of distributed vectors to compensate for noisy spiking outputs [21] [22]. This example shows that hyperdimensional representations can be plugged into neural architectures to boost reliability and reduce complexity. Another example on the cognitive side is the development of a **Holographic Declarative Memory module** for the

ACT-R cognitive architecture (Kelly et al., 2025), which replaces symbolic discrete memory with distributed vectors to better model human-like recall and generalization [23].

For our toroidal attention scenario, hyperdimensional embeddings suggest a way to store and manipulate the entire attention state as a single **holistic vector**. Instead of having a large grid of attention values, one could encode the whole pattern into a hypervector. For example, one might assign a base hypervector to each location on the torus (each "position" in the attention map) and then combine them (via bundling) weighted by the current attention intensities to produce a single memory vector. This single vector would thus **embed the full attention distribution** across the torus. Querying or focusing on a particular region of attention could be done by constructing a probe vector (e.g. binding the query location's base vector with an "attention read" key) and measuring its similarity with the memory vector – effectively retrieving how much attention is at that location (through the dot product). This is a form of content-addressable access, much like transformer attention but in a single-shot associative recall: the probe *interferes* with the superposed memory and picks out the component corresponding to the queried location [17]. Because the memory is distributed, such a system would be robust to faults: if some dimensions of the vector are corrupted (or if some neurons storing it die out), the information can still be recovered from the remaining dimensions with only a slight loss of fidelity.

Holographic Codes and Robust Representations

The term **holographic codes** originates from quantum error correction and theoretical physics, but the core ideas translate well as analogies for robust representation in ML. Holographic **error-correcting codes** (such as the HaPPY code by Pastawski *et al.*, 2015 [24]) are inspired by the holographic principle in quantum gravity, which suggests that information about a volume (the "bulk") can be fully encoded on a lower-dimensional boundary. In practical terms, these codes distribute logical information across many physical qubits in such a way that **any sufficiently large subset of those qubits can reconstruct the original information**. There is no single point of failure: the encoding is highly **redundant and spatially distributed** (like a hologram). Recent work in this vein has shown that by using *geometric and topological structures* for encoding data, one can achieve **inherent error resilience** without active error-correction algorithms. For example, Nye (2025) proposes a "geometric holographic memory" that stores data in topological invariants of a geometrical construct, providing **natural error protection through topology** [25]. The system encodes information in continuous geometric patterns rather than discrete bits, and changes in these patterns (unless they break a topological invariant) do not easily corrupt the stored data [25] [26]. In fact, the data can be preserved *without any external error correction*, because the **structure of the code itself (e.g. a knotted or toroidal pattern)** imposes constraints that prevent small perturbations from altering the represented value [26]. This approach was shown to yield **robust memory storage**, with error resistance surpassing traditional bitwise storage in theoretical analysis [26].

Translating this to a neural context, we can think of designing the representation of our attention state such that *the content is protected by the form*. A simple example of a topological invariant in a toroidal attention map might be the total sum of attention or a particular winding number if we treat the attention distribution as a loop – ensuring such invariants could make the representation resilient to noise. More directly, using ideas from holographic codes means **spreading information broadly** so that it is over-determined and *any partial observation can still infer the whole*. In practice, the hyperdimensional approach already moves toward this by smearing data across many dimensions. One could further borrow techniques from error-correcting codes (like adding parity checks or using redundant basis functions) to ensure certain consistency in the attention vector. The bottom line is that **robust representations intentionally** sacrifice some efficiency or sparsity to gain resilience**: they encode data with overlaps and constraints reminiscent of how a hologram or a topological code works. This way, even if part of the network's representation is damaged (e.g. some units knocked out or some weights disturbed), the underlying** "attention state" can still be reconstructed** from the remaining pieces, much as a hologram fragment still contains a blurry whole image.

Toward a Toroidal Attention State Representation

Bringing these threads together, we can envision an architecture for the **"Donut's" toroidal attention state** that leverages *implicit continuous representation*, *attention-based read/write*, and *distributed robust encoding*. Below are some key ideas and components for such a design:

- **Toroidal Coordinate Mapping**: We treat the attention state as a continuous function defined on a torus (e.g. $f(\theta, \phi)$ on a 2D torus, or $g(\alpha)$ on a 1D circle if only one angle is needed). This can be implemented by a small neural field network that takes in angular coordinates and outputs the "attention intensity" at that point. The NeRF paradigm offers a template for this – using an MLP with periodic (sinusoidal) position encodings to naturally handle the $0/2\pi$ wrap-around [2] [27]. By training this network on whatever target attention distribution is needed (possibly via gradient descent or online updates from an agent's cognition), we get a **smooth, interpolable representation** of attention over the torus. Crucially, because the domain is a torus, the network learns that, for example, $(\theta = 359^\circ)$ is next to $(\theta = 0^\circ)$ with no discontinuity. Techniques like the **Toroidal Autoencoder** have even enforced such latent torus structures via specialized loss functions (e.g. a circular spring loss to ensure a uniform coverage of the torus) [28], showing that neural networks *can learn toroidal manifolds* for latent variables. This coordinate-based representation means our attention state isn't stored as a big array of parameters, but rather **implicitly** in the weights of the MLP – much like NeRF compresses a scene. It's memory-efficient and can generalize (fill in) between learned points. Changes to the attention (shifting focus, spreading out, etc.) could be effected by **modifying the network's inputs or latent codes** rather than brute-force moving a bump in an array.

- **Continuous Attractor Dynamics**: To maintain and update the attention state in a stable way, we can incorporate principles from *continuous attractor networks*. In neuroscience, as noted, head direction cells form a ring attractor and grid cells a torus attractor such that the neural activity can drift smoothly but remains constrained to a ring/torus manifold [14]. We could implement a similar idea in an ML-friendly manner: for example, a recurrent layer (or even the implicit network itself with a feedback loop) that ensures the attention pattern tends to form a localized "bump" that can move but not vanish. One could initialize the attention MLP's output as a peaked Gaussian on the torus and then have a recurrence (through the transformer or a separate update rule) that shifts this peak based on inputs (e.g. if an agent's focus moves). Because the toroidal representation has no edges, the **bump can roam freely** without distortion [13]. The attractor aspect means if no new input perturbs it, the attention will stay where it is – providing a form of short-term memory. Implementing this might draw on techniques from *neural ODEs or continuous dynamics*, or even energy-based models that have a continuous manifold of states as minima.

- **Hyperdimensional Memory Encoding**: In parallel, or as an alternative, the entire toroidal attention map at any given time could be encoded into a single high-dimensional vector (a hypervector). Each point on the torus can be assigned a random hypervector $H(\theta, \phi)$ from a known set; the current attention distribution $A(\theta, \phi)$ can then be encoded as $M = \sum_{\theta,\phi} A(\theta,\phi) \cdot H(\theta,\phi)$ (an example of **weighted bundling** of all location vectors, weighted by attention value). This *memory vector* M now **contains the attention state in a distributed form**. A nice property is that storing a new attention pattern could be as simple as replacing or adding to this vector; comparing two attention states is just a vector similarity; and partial patterns can be queried. For instance, to read the attention at a specific location (θ_0,ϕ_0), one can take the hypervector $H(\theta_0,\phi_0)$ for that location and compute its dot product with M. If M contains a significant contribution from that location's vector (i.e. if attention was high there), the dot product will be high; if not, it will be near random. This is essentially a **content-addressable query** in one step, reminiscent of how a cue retrieves a superposed memory by interference [17]. Because of the holographic overlap, this memory is robust: even if M is noisy or partially corrupted, the correlation with the true $H(\theta_0,\phi_0)$ will on average still be higher than with other random vectors, so the system can tolerate damage. Such a hyperdimensional memory could serve as a **backup or auxiliary representation** for the attention state, enabling quick similarity checks (e.g. "has my current attention distribution been seen before?" by comparing M to stored prototypes) or serving as a bridge between symbolic and subsymbolic processing (as in neuro-symbolic systems [29]).

- **Transformer Integration (Query/Update)**: A Transformer (or any differentiable attention module) can be the controller that **reads from and writes to these representations**. For example, the Transformer's queries and keys could be designed to interact with the hyperdimensional memory: the transformer's **query vector could serve as a cue** to retrieve information from the distributed memory (via a similarity score) [17], effectively acting like an associative read. Conversely, the transformer's output could be used to **update the implicit torus representation**. One approach is to treat the implicit NeRF-like network as an external function and use the transformer's output to adjust its parameters or to modulate its latent inputs (if the implicit model has a latent code). Another approach is to have the transformer produce a *delta attention pattern* (like an image delta on the torus grid), which is then added to the current implicit function output (and perhaps re-fit or encoded back into the hypervector memory). The high bandwidth of attention mechanisms means the Transformer can potentially attend to the entire toroidal state (through either sampling many points or via the compressed memory vector) and decide where to increase or decrease attention. In essence, the Transformer provides a **learned rule for attention shifts** given the agent's goals or observations, while the underlying representation (implicit + hyperdimensional) ensures those shifts are applied in a smooth and robust way (on the torus and with redundant encoding).

- **Robustness via Redundancy and Topology**: By combining the above, our toroidal attention state gains multiple layers of robustness. The **topological continuity** of the torus means no artificial edge discontinuities – the representation is rotation/shift-friendly and can reuse learned patterns at 0° when at 360° etc., which is a form of generalization. The **distributed hypervector encoding** means any single neuron or component is not critical, as the information is smeared across many parts (this is analogous to the **topological invariant protection** mentioned earlier, where the data is in the global structure) [26]. Even if the implicit network suffers minor weight perturbations or the hypervector loses some bits, the attention pattern can be recovered or maintained from the remaining redundancy. Moreover, we could introduce explicit regularizers: for instance, encourage the implicit network to produce a certain total attention mass (invariant) or use an autoencoder that reconstructs the attention from random partial samples, forcing a holographic property that any portion of the representation still contains clues to the whole. These ideas mirror the goals of holographic codes – **any subset of the representation is informative about the full state**, up to some degradation. In practice, such redundancy can be traded off with precision, but given modern large models often have excess capacity, using some of that capacity for resilience is worthwhile.

To conclude, representing a **"Donut's toroidal attention state"** in an ML-friendly way is an ambitious but plausible endeavor. By drawing inspiration from **NeRFs** (to encode continuous fields on irregular domains), **Transformers** (to flexibly control and query those fields through attention), **hyperdimensional memory** (to store and retrieve states with high noise tolerance), and **holographic codes** (to distribute information for built-in error correction), we can design a system where an agent's attention is not just a static array of numbers, but a **living, robust manifold**. Such a system could maintain focus over long durations, smoothly transition as the context shifts, and recover from perturbations – all desirable qualities for advanced world models and cognitive architectures. The research frontier (2022–2025) is already exploring pieces of this puzzle: for example, using tori for representation of cyclic data [2], building transformer world models that internalize environment state [9], and adopting high-dimensional vector representations for memory and reasoning [16]. The synthesis outlined above is one possible path to a **toroidal attention module** that is both **geometrically structured** and **robustly distributed** – a promising direction for the next generation of intelligent systems.

Sources:

1. Meléndez, J. *et al.* (2025). **Toroidal Topology Computation using Neuro-Symbolic AI and Hyperdimensional Computing** – discusses leveraging torus structures for cyclic data and the strengths of hyperdimensional (vector symbolic) representations [2] [15].

2. Xiao, W. *et al.* (2025). **Neural Radiance Fields for the Real World: A Survey** – a comprehensive survey of NeRFs, implicit 3D representations and their advantages over traditional discretized representations [4] [5].

3. Shim, H. *et al.* (2023). **SNeRL: Semantic-aware Neural Radiance Fields for Reinforcement Learning** – demonstrates using a NeRF-based model as a world-model component, yielding improved sample efficiency in RL tasks [7] [8].

4. Robine, J. *et al.* (2023). **Transformer-based World Models Are Happy With 100k Interactions** (ICLR 2023) – introduces a Transformer-XL world model that attends over latent states, actions, and rewards, enabling imagination and efficient learning [9] [10].

5. Zou, Z. *et al.* (2022). **SpikeHD: Memory-inspired spiking hyperdimensional network for robust online learning** – combines spiking neural nets with hyperdimensional computing to achieve noise-robust learning with smaller models [19] [20].

6. Rutledge-Taylor, M.F. *et al.* (2014). **Dynamically Structured Holographic Memory** – cognitive model where memories are high-dimensional vectors, supporting superposition and cue-based retrieval (the basis for later Holographic Declarative Memory in ACT-R) [17] [18].

7. Gardner, R.J. *et al.* (2022). **Toroidal topology of population activity in grid cells** (*Nature*, 602:123–128) – provides evidence that grid cells in the entorhinal cortex form a toroidal activity pattern for spatial position, a biological inspiration for toroidal representation in neural systems [13] [14].

8. Nye, L. (2025). **Geometric Holographic Memory: Efficient and Error-resilient Data Storage** – describes encoding information in geometric/topological forms (inspired by holographic codes) to achieve inherent error correction and robust memory without active error-checking [25] [26].

---

[1] [2] [3] [12] [15] [16] [27] [28] [29] Toroidal Topology Computation using Neuro-Symbolic AI and Hyperdimensional Computing | by John Meléndez (馬強安 - 在臺灣) | Medium
https://medium.com/@John_Melendez/toroidal-topology-computation-using-neuro-symbolic-ai-and-hyperdimensional-computing-c4e3062e8f21

[4] [5] [6] Neural Radiance Fields for the Real World: A Survey
https://arxiv.org/html/2501.13104v1

[7] [8] proceedings.mlr.press
https://proceedings.mlr.press/v202/shim23a/shim23a.pdf

[9] [10] [11] [Quick Review] Transformer-based World Models Are Happy With 100k Interactions
https://liner.com/review/transformerbased-world-models-are-happy-with-100k-interactions

[13] [14] Toroidal topology of population activity in grid cells | Nature
https://www.nature.com/articles/s41586-021-04268-7?error=cookies_not_supported&code=3c4b2880-2510-4f50-9771-f360eea87387

[17] [18] arxiv.org
https://arxiv.org/pdf/2511.11848

[19] [20] [21] [22] Memory-inspired spiking hyperdimensional network for robust online learning | Scientific Reports
https://www.nature.com/articles/s41598-022-11073-3?error=cookies_not_supported&code=9a72251a-1ab4-4f4e-89cb-0e2d6792bba6

[23] Robert WEST | Professor (Associate) | Doctor of Psychology | Carleton University, Ottawa | Institute of Cognitive Science | Research profile
https://www.researchgate.net/profile/Robert-West-23

[24] Holographic code | Error Correction Zoo
https://errorcorrectionzoo.org/c/holographic

[25] [26] Geometric Holographic Memory: Efficient and Error-resilient Data Storage
https://jhap.du.ac.ir/article_459.html