The Cortical Canvas: Spatiotemporal Substrate Allocation as a Unifying Principle for Multimodal Embodied Intelligence

Abstract

We present the Recursive Omnimodal Video Action Model (ROAM), whose central architectural primitive — the spatiotemporal canvas — exhibits deep structural parallels with mammalian cortical organization. The canvas is a 3D grid of embedding vectors onto which heterogeneous sensory and motor streams are placed at designated positions, then processed by a shared transformer backbone with looped (recurrent) computation. We argue that this design implicitly recapitulates several principles of cortical computation: (1) bandwidth-proportional area allocation mirrors cortical magnification; (2) local dense attention combined with learned sparse global connections reproduces the small-world wiring of cortical columns and long-range white matter tracts; (3) variable-depth recurrence per canvas region parallels the heterogeneous temporal frequencies and metabolic allocation rates observed across cortical areas; and (4) progressive sharpening of attention across loop iterations mirrors the hierarchical refinement cascade from primary sensory cortex to prefrontal association areas. We formalize these parallels, identify where they hold and where they break, and propose a series of experiments designed to test the functional predictions this cortical framing generates. The goal is not to claim biological fidelity but to establish a bidirectional bridge: neuroscience principles that inform architecture design, and architectural behaviors that generate testable neuroscience hypotheses.

1. Introduction

The dominant paradigm in multimodal machine learning is the modality-specific encoder followed by a shared fusion layer [Radford et al., 2021; Alayrac et al., 2022]. Vision gets a ViT, language gets a causal transformer, actions get an MLP, and some late-fusion mechanism stitches them together. This works. But it is fundamentally at odds with how biological intelligence organizes multimodal processing.

The mammalian neocortex does not contain separate processors for vision, touch, audition, and motor control joined by a fusion bottleneck. It contains a single, approximately uniform computational substrate — the cortical sheet — that is *allocated* to different functions through a combination of genetic priors, developmental self-organization, and experience-dependent plasticity [Mountcastle, 1978; Kaas, 1987]. Visual cortex is not fundamentally different

tissue from motor cortex; it is the same six-layer columnar architecture receiving different inputs and wired to different outputs. The computational principles are shared. The *layout* — which regions get how much cortical real estate, and how they connect to each other — encodes the inductive bias.

ROAM takes this principle seriously. Its core data structure is a spatiotemporal canvas: a 3D grid of shape (T, H, W) where each position holds a d_{model}-dimensional embedding vector. Different *regions* of this grid are designated for different modalities — visual patches occupy the largest contiguous block (analogous to V1-V4 occupying approximately 50% of primate cortex), proprioceptive state occupies a smaller strip (analogous to somatosensory cortex), action outputs occupy another strip (motor cortex), and reward signals occupy a small corner (orbitofrontal cortex). The transformer backbone operates over the flattened canvas via self-attention, seeing no explicit grid structure — only positional encodings and attention masks reveal the spatial layout.

This design has a remarkable consequence: the same model weights process all modalities. The only thing that changes between a 7-DOF robot arm and a quadruped, between a computer control agent and a multi-robot assembly system, is the canvas layout — which regions are allocated to which streams and how large they are. The weights encode *physics* and *computation*. The layout encodes *embodiment* and *task structure*. This separation mirrors the cortical distinction between columnar microcircuitry (shared computation) and cortical area allocation (specialized function).

The goal of this paper is to develop this analogy rigorously. We do not claim that ROAM *is* a cortex, or that attention *is* synaptic transmission. Biological cortex is vastly more complex in its biochemistry, plasticity mechanisms, neuromodulation, and developmental processes than any transformer will be for decades. What we claim is weaker but, we believe, more useful: that the spatiotemporal canvas architecture and cortical organization share a set of *computational-level principles* [Marr, 1982] — principles about how to organize heterogeneous information processing on a uniform substrate — and that these shared principles generate non-trivial predictions in both directions. From neuroscience to architecture: the cortical analogy suggests specific design choices (bandwidth-proportional allocation, local-dense plus global-sparse connectivity, variable processing depth) that we implement and evaluate. From architecture to neuroscience: the model's emergent behaviors (functional specialization from uniform weights, spontaneous hub formation in sparse connectivity, variable iteration depth correlating with computational difficulty) suggest hypotheses about cortical function that could be tested experimentally.

The paper proceeds as follows. Section 2 surveys the relevant neuroscience and machine learning background. Section 3 describes ROAM's architecture with explicit reference to the implemented code. Section 4 develops the cortical parallels in detail — the meat of the paper. Section 5 proposes experiments designed to test the framework's predictions. Section 6 discusses limitations and open questions.

1.1 Why This Matters

The practical motivation is straightforward: if the cortical canvas framework is correct, it tells us how to scale multimodal embodied models. Instead of designing bespoke fusion architectures for each new modality combination, we allocate canvas real estate according to information bandwidth, wire local regions densely and global regions sparsely, and let the shared backbone discover the computational structure. This is how evolution scaled the cortex: by expanding and subdividing cortical real estate while preserving columnar microcircuitry [Krubitzer, 2007].

The theoretical motivation is deeper. There is a growing body of work suggesting that the structure of intelligence is constrained not by the specific substrate (biological neurons vs. artificial ones) but by the *computational problem* of building world models from heterogeneous sensorimotor streams under resource constraints [Friston, 2010; Hawkins & Ahmad, 2016]. If this is true, then independently evolved solutions to the same problem should share structural features. The cortical canvas hypothesis is a specific, testable version of this claim: the structural features shared between ROAM and cortex are not coincidental but reflect convergent solutions to the same organizational problem.

1.2 Scope and Caveats

We restrict our analysis to the neocortex, specifically the principles governing area allocation, intra-area vs. inter-area connectivity, and temporal processing hierarchy. We do not address subcortical structures (thalamus, basal ganglia, cerebellum), neuromodulatory systems (dopamine, serotonin, acetylcholine), or developmental processes in any detail. Each of these has potential architectural analogues in ROAM (the thalamus as an attention routing mechanism, dopamine as reward signal, the cerebellum as a forward model), but developing these would require a separate treatment.

We also acknowledge that the analogy is strongest at Marr's computational level (what problem is being solved) and algorithmic level (what representations and transformations are used), and weakest at the implementation level (how it is physically realized). Attention is not synaptic transmission. Backpropagation is not Hebbian learning. Positional encoding is not retinotopic mapping in the strict developmental sense. We are explicit about where the analogy breaks.

---

2. Background and Related Work

2.1 Cortical Organization Principles

The neocortex is a 2-4mm thick sheet of neural tissue, approximately 2500 cm^2 in humans when unfolded, organized into six layers with a repeating columnar microstructure [Mountcastle, 1997]. Despite this uniformity at the microscale,

the cortex is functionally heterogeneous at the macroscale: different regions process different modalities and perform different computations. The tension between microscale uniformity and macroscale specialization is one of the central puzzles of neuroscience.

Cortical magnification. The amount of cortical area allocated to a sensory or motor function is roughly proportional to the information bandwidth of that function [Daniel & Whitteridge, 1961]. In the primate visual system, the fovea (which subtends roughly 2 degrees of visual angle but provides the highest acuity) has a cortical magnification factor approximately 30 times that of the periphery. More generally, V1 through V4 together occupy approximately 50% of macaque cortex [Van Essen, 2003], reflecting the dominance of vision in primate information processing. Somatosensory cortex allocates disproportionate area to the hands and lips (the cortical homunculus of Penfield and Boldrey [1937]), reflecting their high receptor density and behavioral importance. Motor cortex follows a similar distorted mapping. The principle is: more bandwidth or more behavioral relevance equals more cortical real estate.

Connectivity: local dense, global sparse. Within a cortical column (approximately 300 μm diameter), connectivity is dense — neurons make thousands of synapses primarily with their neighbors [Douglas & Martin, 2004]. Between distant cortical areas, connectivity is sparse and specific, mediated by long-range axonal projections through the white matter [Markov et al., 2014]. This architecture forms a small-world network: high local clustering combined with a small number of long-range shortcuts [Bassett & Bullmore, 2006]. Small-world networks are efficient for information routing (low average path length) while maintaining high local processing capacity (high clustering coefficient). They also operate near a critical regime that maximizes the dynamic range of neural responses [Shew & Plenz, 2013].

Temporal hierarchy. Different cortical areas operate at different characteristic temporal frequencies. Primary sensory cortex (V1, A1, S1) responds to rapid fluctuations on the order of 10-100 ms. Higher sensory areas (V4, IT) integrate over hundreds of milliseconds. Association areas (prefrontal, parietal) maintain representations over seconds to minutes [Murray et al., 2014; Hasson et al., 2015]. This temporal hierarchy is not an accident but reflects the computational role of each area: early areas must track fast-changing sensory input, while higher areas must maintain stable representations for planning and decision-making. The hierarchy is supported by differences in intrinsic neural time constants (mediated by NMDA receptor density and recurrent connectivity strength) and by the laminar pattern of feedforward vs. feedback connections [Felleman & Van Essen, 1991].

Predictive coding. A prominent theoretical framework posits that the cortex implements hierarchical Bayesian inference, with each cortical area maintaining a generative model that predicts the activity of the area below it [Rao & Ballard, 1999; Friston, 2005]. Prediction errors propagate upward (feedforward); predictions propagate downward (feedback). Processing continues at each level

until prediction error is minimized. This framework provides a natural account of recurrent processing: the depth of recurrence at each level reflects the difficulty of explaining the current sensory input given the model's priors. Easy (predicted) stimuli require few iterations; surprising stimuli require many.

2.2 Relevant Machine Learning Architectures

Diffusion Transformers (DiTs). The Diffusion Transformer [Peebles & Xie, 2023] replaces the U-Net backbone of standard diffusion models with a transformer, using adaptive layer normalization (AdaLN) to condition on the diffusion timestep. CogVideoX [Yang et al., 2024] extends this to video generation using 3D spatiotemporal attention. ROAM builds on this by replacing standard DiT blocks with looped blocks and extending the token space to include non-visual modalities.

Looped transformers and adaptive compute. The idea of sharing weights across transformer layers (looping) dates to Universal Transformers [Dehghani et al., 2019], which also introduced per-token adaptive halting via a pondering mechanism. More recently, the Oro paper [2024] demonstrated that looped language models achieve competitive performance with fewer unique parameters, with the key innovation being an exit gate mechanism that learns to allocate compute across loops. ROAM extends this idea to the multimodal diffusion setting.

Mixture of Experts (MoE) and conditional computation. The exit gate can be viewed as a temporal form of conditional computation [Bengio et al., 2016]: instead of routing tokens to different expert modules (spatial routing), tokens are routed to different numbers of loop iterations (temporal routing). This connection to MoE is important because it suggests that the benefits of conditional computation — specialization, efficiency, and capacity scaling — apply in the temporal domain as well as the spatial domain.

Multimodal fusion architectures. Flamingo [Alayrac et al., 2022], Gato [Reed et al., 2022], and RT-2 [Brohan et al., 2023] represent different approaches to multimodal fusion: cross-attention gating, sequence interleaving, and action-token appending, respectively. ROAM's canvas approach is closest to Gato's sequence interleaving but with spatial structure: instead of flattening all modalities into a single sequence, ROAM assigns each modality a spatial region on the canvas, preserving geometric relationships.

Sparse attention mechanisms. Longformer [Beltagy et al., 2020], BigBird [Zaheer et al., 2020], and more recently the block-sparse attention patterns of Mixtral [Jiang et al., 2024] combine local window attention with global tokens. ROAM's sparse connectivity module differs in that the global connections are *learned* rather than hand-designed: a set of continuous logits are trained via straight-through estimation to discover which long-range connections the model needs.

2.3 Neuroscience-Inspired AI

Thousand Brains Theory. Hawkins and Ahmad [2016] propose that each cortical column independently learns a model of its input, and that cortical columns reach consensus through lateral connections. The canvas can be viewed through this lens: each canvas position is an independent "column" that maintains a d_{model}-dimensional model of its local input, and positions coordinate through attention. The key difference is that ROAM's positions interact through a global attention mechanism rather than local lateral inhibition, allowing arbitrary long-range coordination.

Hyperdimensional computing. Kanerva [2009] proposes that the brain represents and manipulates information using high-dimensional vectors (on the order of 10,000 dimensions) where the key operations are element-wise addition (superposition), element-wise multiplication (binding), and permutation (sequencing). ROAM's canvas embeddings are in a d_{model}-dimensional space (typically 256-1024) and are combined through attention, which can be viewed as a soft form of the binding and superposition operations.

Active inference and the free energy principle. Friston [2010] proposes that biological agents minimize variational free energy, a quantity that bounds surprise (negative log evidence). Under this framework, perception is inference (explaining sensory input), action is inference (bringing sensory input in line with predictions), and the allocation of processing resources across cortical areas reflects the expected reduction in free energy from additional processing. ROAM's variable-depth looping, where the exit gate decides when a canvas region has been sufficiently processed, can be viewed as a discrete approximation to the free energy principle: each loop iteration reduces the "surprise" (loss) of the model's predictions, and the gate halts when the marginal reduction falls below threshold.

---

3. Architecture

ROAM's architecture has three components: the spatiotemporal canvas, the processing backbone, and the adaptive compute mechanism. We describe each with reference to the implementation.

3.1 The Spatiotemporal Canvas

The canvas is a 3D tensor of shape (T, H, W) where each position (t, h, w) holds a d_{model}-dimensional embedding vector. The `SpatiotemporalCanvas` class (implemented in `roam/canvas.py`) manages the canvas lifecycle: creation, modality placement, extraction, and output masking.

Positional encoding. Each canvas position receives a 3D sinusoidal positional encoding, computed by `SinusoidalPositionalEncoding3D`. The d_{model}-

dimensional encoding vector is split into three components (d_t, d_h, d_w) corresponding to temporal, vertical, and horizontal position, each generated by standard sinusoidal encoding at different frequencies. This creates a smooth manifold over the canvas: nearby positions have similar encodings, and distances in canvas space are approximately preserved in encoding space.

Modality embeddings. In addition to positional encodings, each canvas position receives a learned modality-type embedding: a single d_{model}-dimensional vector per modality (visual, text, action, proprioceptive, reward), added to every position assigned to that modality. This tells the model *what kind* of information occupies each position, independent of *where* it is.

Region assignment. The `CanvasLayout` dataclass specifies which 3D bounding box on the canvas belongs to which modality. Region boundaries are pure configuration — they do not appear in the model weights. Changing the layout (e.g., adding more visual positions, removing the text region, adding a second action column for a bimanual robot) requires no retraining; it only changes which positions are populated and masked.

Output masking. The `get_output_mask()` method returns a boolean tensor indicating which positions participate in the diffusion process (i.e., have noise added and loss computed). Input positions (past visual frames, current proprioceptive state) are conditioning context; output positions (future visual frames, future actions, reward prediction) are the targets.

3.2 The Processing Backbone

The backbone is a stack of transformer blocks, each of which can be looped multiple times with shared weights. Two implementations exist.

v1: LoopedDiffusionBlock (`roam/blocks.py`). Standard transformer block (AdaLN-conditioned self-attention + SwiGLU FFN) augmented with a loop embedding per iteration and a probabilistic exit gate. The exit gate (`roam/exit_gate.py`) uses a two-layer MLP on the mean-pooled canvas representation to produce an instantaneous exit probability via sigmoid. Survival probabilities accumulate across loops to form a proper distribution over exit times, and KL divergence from uniform serves as entropy regularization.

v2: ROAMv2Block (`roam/v2/block.py`). Replaces dense self-attention with position-aware attention and sparse connectivity, adds progressive sharpening, and uses a more efficient halting mechanism. The four v2 components are:

1. *Geometric position-aware attention* (`geometric_attention.py`): Position encodings are concatenated onto the query and key vectors rather than added to the input. This decomposes the attention score into $\text{score}(i,j) = \text{content}(i,j) + \text{position}(i,j)$, allowing the model to learn *structural interaction patterns* (e.g., "action positions always attend to proprioceptive positions") as an explicit geometric prior, separate from content-based attention.

2. *Local window + learned sparse global connectivity* (`sparse_connectivity.py`): `LocalWindowMask` precomputes a boolean mask allowing each position to attend to its spatiotemporal neighbors within a window of $(\pm w_t, \pm w_h, \pm w_w)$. `LearnedSparseConnectivity` maintains a continuous logit matrix that is discretized via straight-through top-k estimation during training, selecting exactly k_{global} long-range connections per token. `CombinedAttentionMask` takes the union of local and global masks.

3. *Progressive sharpening* (`progressive_sharpening.py`): The attention logits at loop ℓ are multiplied by a sharpening factor $\beta(\ell)$ that increases monotonically across loops. At loop 0, $\beta = \beta_{\min}$ (soft, exploratory attention); at loop $K-1$, $\beta = \beta_{\max}$ (sharp, precise attention). The schedule can be linear, exponential, or learned. A monotonicity regularizer penalizes $\beta(\ell) > \beta(\ell+1)$.

4. *Efficient halting* (`efficient_halting.py`): A single learned d_{model}-dimensional direction vector $\mathbf{h}$. Each token's halting score at each loop is $\text{softplus}(\mathbf{x} \cdot \mathbf{h} + b)$. Scores accumulate across loops; when the cumulative score exceeds a threshold, the token halts. Cost: d_{model} parameters total, one dot product per token per loop — approximately 128 times cheaper than the v1 per-loop MLP gate.

3.3 Adaptive Compute and Halting

The halting mechanism creates a direct analogue to variable metabolic allocation. Each token (canvas position) independently decides when it has been processed enough. Tokens representing visually complex scenes or novel proprioceptive states may loop many times; tokens representing static backgrounds or well-predicted joint angles may halt early. The total computation invested in each canvas region emerges from the interaction between the halting scores and the input data, not from any hardcoded allocation.

During training, all loops are executed (with halted tokens' states held constant), and the halting entropy loss encourages a spread of halting times. During inference, tokens that have halted are excluded from attention computation, providing a direct wall-clock speedup proportional to the fraction of tokens that halt early.

---

4. Cortical Parallels

This section develops the six core parallels between the canvas architecture and cortical organization.

4.1 Bandwidth-Proportional Allocation as Cortical Magnification

The cortical magnification factor (CMF) describes the number of millimeters of cortical surface per degree of visual angle, per millimeter of skin surface, or per semitone of auditory frequency. It is highest for high-acuity, high-bandwidth sensory channels and lowest for low-bandwidth, low-precision channels. The foveal representation in V1 has a CMF roughly 30 times that of the far periphery [Daniel & Whitteridge, 1961]. The hand occupies disproportionate somatosensory cortex relative to the back [Penfield & Boldrey, 1937]. And critically, this allocation is not fixed: cortical maps reorganize in response to altered input, with deafferented regions being colonized by neighboring representations [Merzenich et al., 1984].

The canvas implements this principle directly. In the production layouts described in the CORTICAL_CANVAS design document, vision occupies 50-70% of canvas positions, tactile 10-15%, proprioception 3-5%, language 3-5%, and reward 1-3%. These proportions are not arbitrary; they reflect the approximate information bandwidth of each stream. Vision at camera resolution delivers orders of magnitude more bits per second than proprioception (a few joint angles) or reward (a single scalar). Allocating canvas positions proportional to bandwidth gives each modality roughly equal *representation density*: the number of embedding dimensions per bit of input information.

The deeper question is whether this allocation should be hand-designed (as currently implemented) or learned. In the cortex, initial allocation is genetically specified and then refined by experience. In ROAM, the layout is currently a hyperparameter. We propose (Section 5) experiments to make allocation learnable, with the prediction that the learned allocation will converge to something close to the bandwidth-proportional hand-designed layout, much as the cortex's experience-dependent reorganization preserves the gross allocation pattern while refining the details.

Where the analogy holds: Proportional allocation, the principle that more bandwidth justifies more compute substrate, and the functional consequence that finer-grained representation of high-bandwidth modalities improves overall task performance.

Where the analogy breaks: Cortical magnification is continuous and smooth (the CMF varies gradually across the cortical sheet). The canvas layout has hard boundaries between regions. Cortical reorganization is gradual, activity-dependent, and mediated by Hebbian plasticity. Canvas layout changes require explicit reconfiguration. Future work on soft allocation via learned modality-affinity vectors could close this gap.

4.2 Local Dense + Global Sparse Connectivity as Cortical Wiring

Within a cortical column, neurons are densely interconnected. A typical pyramidal neuron in layer 2/3 makes approximately 7,000 synapses, primarily with other

neurons within a 500 μm radius [Braitenberg & Schuz, 1998]. Between distant cortical areas, connections are sparse, specific, and mediated by long-range axons bundled into white matter tracts. The corpus callosum connects homotopic areas of the two hemispheres; the arcuate fasciculus connects Broca's and Wernicke's areas for language; the inferior longitudinal fasciculus connects occipital and temporal cortex for visual object recognition [Catani & Thiebaut de Schotten, 2008].

This wiring pattern forms a small-world network: the average path length between any two cortical areas is short (approximately 2-3 synaptic relays), while local clustering is high [Bassett & Bullmore, 2006]. Small-world networks are near-optimal for several computational objectives: they maximize the ratio of information integration to wiring cost [Bullmore & Sporns, 2012], they support both segregated (modular) and integrated (global) processing, and they operate near a critical point that maximizes dynamic range and information transmission [Shew & Plenz, 2013].

ROAM's `CombinedAttentionMask` implements a structurally isomorphic wiring pattern. The `LocalWindowMask` creates dense connectivity within a spatiotemporal neighborhood of $(\pm w_t, \pm w_h, \pm w_w)$, analogous to intra-columnar connectivity. The `LearnedSparseConnectivity` module discovers exactly k_{global} long-range connections per token, analogous to inter-area white matter tracts. The combined mask is the union: each token attends to all its local neighbors and to its k_{global} most important distant partners.

The key innovation is that the global connections are *learned*, not hand-designed. During training, continuous logits are maintained for all possible $N \times N$ token pairs. Top-k selection (with straight-through gradient estimation) picks the k highest-scoring connections. The sparsity loss (L1 on sigmoid of logits) encourages the model to use as few global connections as possible. After training, the connectivity pattern is fixed.

This learning process has a direct cortical analogue: the formation of long-range cortico-cortical connections during development. Initial connectivity is diffuse and exuberant; activity-dependent pruning eliminates connections that are not functionally useful, leaving a sparse but precisely targeted pattern [Innocenti & Price, 2005]. The ROAM training process — starting from uniform logits and pruning via gradient-driven sparsification — mirrors this developmental trajectory.

What the model can learn. The local-dense + global-sparse pattern enables a specific computational profile: high-bandwidth local processing (each modality region processes its own stream at full resolution) combined with low-bandwidth cross-modal binding (selected tokens from each region exchange information globally). This is precisely the pattern needed for embodied intelligence: process visual, proprioceptive, and action streams at their native temporal and spatial resolution within their respective regions, but coordinate across modalities through a small number of "hub" tokens that carry compressed summaries.

Connection to criticality. The small-world topology that emerges from local-dense + global-sparse wiring places the system near a critical regime in the information-theoretic sense. At criticality, correlation lengths diverge, meaning that local perturbations can propagate globally — a desirable property for a system that needs to integrate information across modalities. The number of global connections k_{global} is a control parameter: too few and the system fragments into disconnected modules; too many and it becomes fully connected (losing the computational efficiency of modularity). We predict (Section 5) that the optimal k_{global} will correspond to a phase transition in task performance, analogous to the critical connectivity at which cortical networks maximize information transmission.

4.3 Variable Recurrence as Variable Processing Frequency

Cortical areas differ not only in what they process but in *how fast* they process it. Neurons in primary visual cortex (V1) respond to stimuli with latencies of 40-60 ms and can follow temporal modulations up to approximately 100 Hz [Hawken et al., 1996]. Neurons in inferotemporal cortex (IT) have longer latencies (80-120 ms) and integrate over longer timescales. Prefrontal cortex neurons maintain stable representations over seconds, supporting working memory and planning [Funahashi et al., 1989; Murray et al., 2014].

This temporal hierarchy is supported by differences in intrinsic neural dynamics. Higher cortical areas have longer intrinsic timescales, measured by the autocorrelation decay time of neural activity [Murray et al., 2014]. The gradient of intrinsic timescales along the cortical hierarchy is remarkably consistent across species and sensory modalities, suggesting that it is a fundamental organizing principle [Hasson et al., 2015].

ROAM's halting mechanism creates an analogous temporal hierarchy. Each canvas position independently decides how many loop iterations to execute via the `EfficientHalting` module. Positions processing fast-changing, high-bandwidth streams (visual frames at the current timestep) will require many iterations to fully process the complex input. Positions representing slowly changing state (a reward signal, a strategic plan) may halt after one or two iterations because their input is simple and well-predicted.

But there is a subtlety: in the cortex, the temporal hierarchy is *intrinsic* (slow areas have slow dynamics regardless of input), while in ROAM, the halting depth is *adaptive* (the same position may loop many times on a difficult input and few times on an easy one). This is actually a feature, not a bug. The cortical temporal hierarchy is also modulated by input complexity — V1 neurons fire longer bursts in response to novel or unexpected stimuli [Vinken et al., 2017]. The halting mechanism captures both the intrinsic component (through the learned bias b in the halting direction) and the adaptive component (through the input-dependent halting score $\mathbf{x} \cdot \mathbf{h}$).

Connection to metabolic allocation. Cortical processing has a metabolic

cost: neural activity consumes glucose and oxygen, delivered by blood flow. The brain allocates metabolic resources to active cortical areas via neurovascular coupling — the mechanism that drives fMRI signals [Logothetis, 2008]. Areas engaged in demanding computation receive increased blood flow. The halting mechanism is a computational analogue: loop iterations are the "metabolic resource," and the halting score is the "demand signal." The total number of loop iterations invested in a canvas region is the architectural equivalent of the total blood flow delivered to a cortical area during a cognitive task.

Connection to the free energy principle. Friston's free energy framework [2010] proposes that the brain minimizes variational free energy, which bounds the surprise of sensory observations. Under this framework, the amount of processing allocated to each cortical area reflects the expected reduction in free energy from additional computation. If the current generative model already predicts the sensory input well (low surprise), little processing is needed. If the prediction is poor (high surprise), more processing is allocated to update the model. ROAM's halting mechanism implements this directly: the halting score measures the "readiness" of each token's representation. When the representation is well-formed (low residual error), the halting score increases and the token exits. When the representation is still uncertain, the token continues looping. The entropy regularization on halting scores ensures that the system does not collapse to a trivial solution (all tokens halting immediately or all looping to maximum), just as the free energy bound prevents the brain's inference from degenerating.

4.4 Spatiotemporal Biases and Topographic Maps

The cortex is organized topographically: nearby cortical neurons represent nearby features of the sensory or motor space. In V1, neighboring neurons respond to neighboring locations in the visual field (retinotopy). In A1, neighboring neurons respond to neighboring sound frequencies (tonotopy). In S1, neighboring neurons respond to neighboring body parts (somatotopy). These maps are approximately continuous — nearby cortical positions map to nearby sensory positions — but they are also distorted, with high-acuity regions magnified relative to low-acuity regions [Wandell et al., 2007].

ROAM's 3D sinusoidal positional encoding creates an analogous topographic structure. The encoding assigns each canvas position (t, h, w) a vector in $\mathbb{R}^{d_{\text{model}}}$ such that nearby positions have similar encodings (via the smoothness of sinusoidal functions) and distant positions have dissimilar encodings. This means that the dot product between two positions' encodings is a monotonically decreasing function of their canvas distance, creating a *soft locality bias* in the attention mechanism: nearby positions are a priori more likely to attend to each other than distant ones.

The `GeometricPositionEncoding` in v2 goes further. Instead of adding positional encoding to the input (as in standard transformers), it maintains separate

learned position vectors that are concatenated onto the query and key projections. The attention score decomposes additively:

$$\text{score}(i,j) = \underbrace{\mathbf{q}_i^{\text{content}} \cdot \mathbf{k}_j^{\text{content}}}_{\text{content similarity}} + \underbrace{\mathbf{q}_i^{\text{pos}} \cdot \mathbf{k}_j^{\text{pos}}}_{\text{interaction geometry}}$$

The content term captures "do these two tokens have semantically related content?" The position term captures "given the spatial relationship between positions i and j, should they interact?" This decomposition allows the model to learn *structural interaction patterns* — for example, "action positions at time t should always attend to proprioceptive positions at time $t-1$" — as a geometric prior, separate from whatever content happens to occupy those positions.

Euclidean canvas vs. non-Euclidean kinematics. The canvas embedding space is approximately Euclidean: the sinusoidal encoding is a smooth map from the 3D grid $\mathbb{Z}^3$ into $\mathbb{R}^{d_{\text{model}}}$, and distances in encoding space approximate L2 distances on the grid. But the kinematic manifold of a robot is not Euclidean. The configuration space of a 7-DOF arm is a 7-torus (each revolute joint has S^1 topology), and distances in configuration space are measured by geodesics on this torus, not by L2 distance in Euclidean space.

This mismatch is addressed by two mechanisms. First, the learned sparse global connections can override the local geometry: if two canvas positions are far apart on the grid but functionally coupled (e.g., the gripper and the visual object it is approaching), the learned connectivity will create a direct connection between them. Second, the content-based component of the attention score is not biased by position — it depends only on the semantic content of the tokens. So even when the positional bias says "these positions are far apart and should not interact," the content-based score can override this if the content strongly demands interaction. The learned global connections provide the architectural capacity, and the content-based attention provides the dynamic routing.

This mirrors a feature of cortical topographic maps: they are the *default* wiring pattern, but they can be overridden by top-down attention. A monkey fixating on a stimulus in its peripheral visual field can attend to that stimulus despite its peripheral position on the retinotopic map, via feedback from prefrontal cortex that enhances the gain of the relevant V1 neurons [Reynolds & Heeger, 2009].

4.5 Progressive Sharpening as Cortical Processing Cascade

When a visual stimulus appears, information cascades through the cortical hierarchy in a characteristic sequence: V1 responds first (40-60 ms), then V2 (50-70 ms), then V4 (80-100 ms), then IT (100-130 ms) [Lamme & Roelfsema, 2000]. But this feedforward sweep is followed by recurrent processing within and between areas, and the character of the processing changes over time. Early feedforward activity carries coarse, categorical information; later recurrent activity refines

this into fine-grained, context-dependent representations [Vanrullen & Thorpe, 2002].

Grossberg's Adaptive Resonance Theory (ART) [Grossberg, 1976] formalizes this as a two-phase process: an initial bottom-up matching phase (fast, approximate) followed by top-down expectation matching (slow, precise). When bottom-up and top-down signals match ("resonate"), the representation stabilizes. When they mismatch, a reset signal triggers exploration of alternative interpretations. Rao and Ballard's predictive coding framework [1999] gives a complementary account: each cortical level generates a prediction of the level below, and processing iterates between levels until prediction errors are minimized.

ROAM's progressive sharpening schedule implements a computational analogue. At loop 0, the sharpening factor $\beta = \beta_{\min}$ (typically 1.0), and attention is soft — every token attends to a broad distribution over other tokens, much like the coarse initial feedforward sweep. At loop $K-1$, $\beta = \beta_{\max}$ (typically 4.0-8.0), and attention is sharp — each token focuses narrowly on a few key partners, much like the refined recurrent processing that follows the feedforward sweep.

The gradient-flow benefit is direct. Sharp attention (high β) is computationally powerful — it approximates hard selection, enabling precise read-write operations on the canvas. But sharp attention has vanishing gradients: if a token's attention is concentrated on one key, the gradient signal to all other keys is near zero, making it impossible to discover new useful interactions via SGD. Soft attention (low β) has well-distributed gradients but limited computational power (everything is averaged together). Progressive sharpening resolves this tension: gradients flow through the soft early loops, discovering which interactions are useful, while the sharp late loops exploit these interactions for precise computation.

This is literally a learned cortical processing cascade. The early loops correspond to V1-like processing: broad, feature-detecting, gradient-rich. The late loops correspond to PFC-like processing: focused, decision-making, computationally precise. The β schedule is the architectural analogue of the feedforward-to-recurrent processing transition, and it is *learned* rather than hand-designed (when using the `learned` schedule type), meaning the model discovers its own processing hierarchy from data.

4.6 The Canvas as a Thousand Brains

Hawkins and Ahmad [2016] propose the Thousand Brains Theory of intelligence: each cortical column independently learns a complete model of its sensory input, including its location in the object's reference frame. Columns vote on the identity and pose of the perceived object through lateral connections. The key insight is that columns are not feature detectors that each capture a fragment of the object; they are *complete models* that each observe the object from a different viewpoint (or touch point, or auditory position).

The canvas architecture has a compelling parallel. Each canvas position is a

d_{model}-dimensional embedding vector that is updated through self-attention over the full canvas. After processing, each position's embedding reflects not just its local input but the global context of the entire canvas, as filtered through attention. In this sense, each position is an independent model of the world, conditioned on its own local input plus its attended context.

The attention mechanism plays the role of Hawkins' lateral voting. Positions with consistent representations reinforce each other (high attention weights flow between tokens with similar content). Positions with inconsistent representations suppress each other (divergent tokens receive low attention weights). Over loop iterations, the canvas converges to a globally consistent state — a consensus among thousands of local models, mediated by attention rather than lateral inhibition.

This perspective makes a specific prediction: after training, each canvas position's embedding should contain enough information to reconstruct (decode) not just its own modality but a coarse version of the *entire* canvas. A visual position should encode not just its local pixel patch but also a compressed representation of the robot's proprioceptive state and the current action. This is testable via linear probing: train a linear decoder on individual canvas positions and measure how much information about distant modalities is recoverable. If the Thousand Brains analogy holds, this cross-modal information should be substantial.

Connection to hyperdimensional computing. Kanerva's framework [2009] describes how high-dimensional random vectors can serve as a basis for associative memory: two randomly chosen vectors in $\mathbb{R}^d$ are approximately orthogonal when d is large, so superposition (addition) can store multiple items in a single vector with minimal interference. ROAM's d_{model}-dimensional embeddings operate in exactly this regime: each canvas position's embedding is a superposition of its local input, its positional encoding, its modality embedding, and the information gathered from attention over other positions. The high dimensionality of the embedding space (typically 256-1024) ensures that these components can coexist with minimal interference.

4.7 Emergence of Functional Specialization from Uniform Architecture

Perhaps the deepest parallel between the canvas and the cortex is this: functional specialization emerges from uniform architecture plus structured input.

The cortex is histologically uniform at the microscale: the same six-layer columnar structure appears throughout [Mountcastle, 1978]. Yet the functional diversity of cortical areas is enormous: V1 computes orientation selectivity, MT computes motion, IT recognizes objects, Broca's area produces speech. How does uniform hardware produce specialized function? The answer is structured input: V1 receives retinotopic input from the lateral geniculate nucleus; A1 receives tonotopic input from the medial geniculate; and S1 receives somatotopic input from the ventral posterior nucleus. The structured input, combined with the cortex's learning rules, drives the emergence of specialized computation in each

area [O'Leary et al., 2007].

The canvas architecture has the same property. All canvas positions are processed by the same transformer weights. There are no modality-specific layers, no special processing for vision vs. proprioception vs. action. The only differences are: (a) which positions receive which input (via the layout), (b) the positional encoding of each position, and (c) the modality-type embedding. These are *exactly* the cortical analogues: which input pathway projects to which area, the topographic map, and the area identity signal (which in the cortex is partly genetic, via area-specific transcription factors like EMX2 and PAX6 [O'Leary et al., 2007]).

We predict that after training, the hidden states at different canvas positions will become functionally specialized despite the shared weights: visual positions will develop representations tuned to spatial features, action positions will develop representations tuned to motor plans, and the boundary between regions will sharpen through the learned sparse connectivity. This is directly testable via representation similarity analysis (RSA) [Kriegeskorte et al., 2008]: compare the representational geometry at visual vs. action vs. proprioceptive canvas positions and measure the degree of functional differentiation.

4.8 Where the Analogy Breaks

Honesty demands that we enumerate the breaks.

Fixed grid vs. irregular topology. The cortex is not a regular grid. It is a folded sheet with gyri and sulci, and its functional areas have irregular shapes and variable sizes. The canvas is a rectilinear 3D grid with hard region boundaries. This means the canvas cannot naturally represent the cortex's smooth transitions between areas (e.g., the gradual shift from V1 to V2 at the V1/V2 border).

Sinusoidal encoding vs. developmental self-organization. The canvas's positional encoding is a fixed mathematical function. Cortical topographic maps emerge through activity-dependent developmental processes (Hebbian learning, spike-timing-dependent plasticity, competition for neurotrophic factors). The fixed encoding constrains the canvas to a specific manifold structure that may not match the optimal geometry for a given task.

Attention vs. synaptic transmission. Self-attention computes a weighted sum over all attended tokens, with weights determined by content similarity. Synaptic transmission is binary (fire or not), modulated by synaptic strength, and subject to complex dynamics (short-term facilitation/depression, long-term potentiation/depression, neuromodulation). The information-processing implications are different: attention is a soft, continuous routing mechanism; synaptic transmission is a hard, plastic, state-dependent one.

No inhibition. Cortical computation relies critically on inhibitory interneurons, which implement gain control, competition (winner-take-all), and oscillatory dynamics [Douglas & Martin, 2004]. The transformer has no explicit inhibitory mechanism. The softmax in attention provides implicit competition (attention

weights sum to one), but this is a weaker form of inhibition than the cortex's diverse inhibitory cell types provide.

No neuromodulation. The cortex's processing is profoundly shaped by neuromodulatory systems (dopamine, norepinephrine, serotonin, acetylcholine) that globally alter the gain, plasticity, and dynamics of cortical circuits [Dayan, 2012]. ROAM has no analogue. The diffusion timestep conditioning (via AdaLN) is the closest mechanism, globally modulating the processing at each step, but it is a scalar rather than the multi-dimensional neuromodulatory "landscape" of the biological brain.

No developmental trajectory. Cortical organization is shaped by a protracted developmental process involving neuronal migration, axon guidance, synaptogenesis, and pruning [Stiles & Jernigan, 2010]. ROAM's architecture is fixed at initialization and modified only through gradient-based training. There is no analogue of the critical periods, experience-dependent refinement, or maturational changes that shape cortical function.

These breaks define the limits of the analogy. They also define opportunities: each break suggests a specific architectural extension (irregular mesh topology, learned positional encoding, explicit inhibitory units, neuromodulatory conditioning) that could close the gap and potentially improve performance.

---

5. Proposed Experiments

We propose seven experiments designed to test specific predictions of the cortical canvas framework.

5.1 Cortical Magnification Scaling Law

Hypothesis: Task performance scales as a power law with the number of canvas positions allocated to the task-relevant modality, analogous to the cortical magnification scaling law relating acuity to cortical area.

Method: Train a series of ROAM models on the same visuomotor task (e.g., Bridge V2 pick-and-place) with varying proportions of canvas allocated to vision: 20%, 35%, 50%, 65%, 80% of total positions. Hold all other hyperparameters constant. Measure action prediction MSE, video prediction FVD, and downstream task success rate.

Prediction: Performance improves as a power law $P \propto A^{\alpha}$ where A is the number of visual positions and $\alpha \in (0.3, 0.7)$, with diminishing returns above approximately 60% allocation. This would mirror the psychophysical finding that visual acuity scales as a power law with cortical magnification factor.

Control: Repeat with uniform random allocation (same total positions, but distributed uniformly across modalities). This tests whether *structured* allocation matters or only total canvas size.

5.2 Sparse Connectivity Phase Transition

Hypothesis: There exists a critical value of k_{global} (global connections per token) at which task performance transitions sharply from failure to success, analogous to the percolation threshold in small-world networks.

Method: Train ROAM v2 on a multi-modal task (video prediction + action prediction) with $k_{\text{global}} \in \{0, 2, 4, 8, 16, 32, 64, N\}$ (where N = full connectivity). Measure task loss and also graph-theoretic properties of the learned connectivity: average path length, clustering coefficient, modularity.

Prediction: Performance is poor for $k_{\text{global}} < k^*$ and good for $k_{\text{global}} > k^*$, with k^* corresponding to the percolation threshold at which the graph becomes connected. At k^*, the graph-theoretic properties will match those of small-world networks (high clustering, low path length). This value may be surprisingly small (on the order of $\log N$, as predicted by random graph theory).

Caveat from existing results: Experiment 19 (v2 experiments) showed dense attention outperforming sparse at $N = 120$ by 3.47 times, with the hypothesis that sparse connectivity only helps at $N > 1000$. This experiment extends the investigation to production-scale canvases.

5.3 Emergent Functional Specialization via RSA

Hypothesis: After training on multimodal data, the representational geometry of canvas positions will show functional specialization: visual positions will cluster separately from proprioceptive positions, which will cluster separately from action positions, despite all positions being processed by identical weights.

Method: Train ROAM on Bridge V2 (video + proprioception + action). After training, extract hidden states from all canvas positions on a held-out set. Compute representational dissimilarity matrices (RDMs) for visual, proprioceptive, and action positions separately. Compare RDMs using second-order RSA [Kriegeskorte et al., 2008].

Prediction: Visual position RDMs will be most similar to RDMs from pretrained vision models (e.g., DINOv2). Proprioceptive position RDMs will be most similar to RDMs computed from joint-angle similarity. Action position RDMs will be most similar to RDMs from task success labels. The cross-modal similarity (e.g., visual RDM vs. proprioceptive RDM) will be lower than within-modal similarity, indicating functional differentiation.

5.4 Halting Depth Correlates with Computational Difficulty

Hypothesis: The average halting depth (number of loops) at each canvas position correlates with the computational difficulty of that position's prediction task, analogous to the finding that cortical metabolic activity (fMRI BOLD signal) increases with task difficulty [Carpenter et al., 1999].

Method: Train ROAM v2 with halting enabled on a task with graded difficulty (e.g., video prediction where some frames contain predictable backgrounds and others contain novel object interactions). Record the per-position halting depth and the per-position prediction loss.

Prediction: Strong positive correlation ($r > 0.5$) between halting depth and prediction loss, with the highest halting depths occurring at visual positions depicting contact events, object deformations, or occlusion boundaries — the "hard" parts of visual prediction. Proprioceptive positions during smooth motion will halt early; positions during contact transitions will loop longer.

5.5 Cross-Modal Probing (Thousand Brains Test)

Hypothesis: After training, each canvas position's embedding contains decodable information about distant modalities, not just its own modality — consistent with the Thousand Brains prediction that each "column" maintains a model of the whole world.

Method: After training ROAM on video + proprioception + action, train linear probes on individual canvas positions: (a) probe visual positions for proprioceptive state, (b) probe proprioceptive positions for visual features, (c) probe action positions for both visual features and proprioceptive state. Measure probe accuracy.

Prediction: Probe accuracy will be above chance for all cross-modal combinations, with the strongest cross-modal signal in positions near the boundary between regions (where local attention spans both modalities) and in positions identified as "hubs" by the sparse connectivity.

5.6 Progressive Sharpening Ablation and Cortical Cascade Analogy

Hypothesis: The progressive sharpening schedule is necessary for the model to solve tasks requiring both coarse recognition and fine-grained control, and the optimal schedule matches the temporal profile of the cortical feedforward-to-recurrent processing cascade.

Method: Train ROAM v2 with four conditions: (a) constant $\beta = 1.0$ (always soft), (b) constant $\beta = 8.0$ (always sharp), (c) progressive sharpening from 1.0 to 8.0 (the cortical cascade condition), and (d) *reverse* sharpening from 8.0 to 1.0 (anti-cortical). Evaluate on a task requiring both object recognition (needs broad, coarse attention) and precise grasping (needs focused, sharp attention).

Prediction: Progressive sharpening outperforms all baselines. Constant-soft fails at precise control (cannot sharply select action targets). Constant-sharp fails at recognition (gradients cannot discover relevant visual features). Reverse sharpening performs worst (sharp early loops have no useful gradient signal for the soft late loops that would need to exploit them).

Caveat from existing results: Experiment 18 showed mild sharpening ($\beta_{\max} =$

2) helping contact detection by 1.30 times F1 while aggressive sharpening ($\beta_{\max} = 8$) hurt (0.66 times). This may reflect the small canvas size ($N = 120$); we predict that the benefit of aggressive sharpening will increase with canvas size, as larger canvases have more tokens to route between.

5.7 Zero-Shot Morphology Transfer (Dynamic Canvas Test)

Hypothesis: A model trained on multiple body topologies can control an unseen body topology at inference time, given only a canvas layout description, because the model weights encode topology-independent physics while the layout encodes topology-dependent structure.

Method: Train ROAM on three robot morphologies (6-DOF arm, 7-DOF arm, quadruped) using different canvas layouts per morphology. Evaluate zero-shot on three held-out morphologies (5-DOF arm, bimanual, hexapod).

Prediction: Zero-shot performance on unseen morphologies exceeds random policy by at least 5 times and reaches at least 0.3 times the performance of a morphology-specific specialist trained from scratch. Performance will be highest for the 5-DOF arm (closest to training distribution) and lowest for the hexapod (most different from training bodies). The model will exhibit qualitatively correct behaviors (reaching toward targets, maintaining balance) even when quantitative accuracy is poor.

---

6. Discussion

6.1 What Does This Architecture Know About Physical Intelligence?

The cortical canvas hypothesis, if the experimental evidence supports it, implies something about the *structure of physical intelligence itself*. The argument is as follows:

1. Biological cortex evolved to solve the problem of multimodal sensorimotor control under resource constraints.
2. ROAM's canvas was designed (independently of cortical neuroscience) to solve the same problem in silicon.
3. The two architectures share non-trivial structural features: bandwidth-proportional allocation, local-dense + global-sparse connectivity, variable processing depth, progressive refinement from coarse to fine.
4. These shared features are unlikely to be coincidental. They reflect constraints imposed by the problem itself — specifically, the need to process heterogeneous streams at different bandwidths and timescales on a finite compute substrate.

If this argument is correct, then the cortical canvas architecture is not arbitrary. It is a *convergent design*: the design that any sufficiently capable system would arrive at when solving the multimodal embodied intelligence problem under

realistic resource constraints. This is a strong claim, and we do not insist on it. But it is a productive hypothesis because it generates specific predictions (Section 5) and suggests specific architectural improvements (add neuromodulation, add developmental plasticity, add inhibition).

6.2 Practical Implications

Beyond the theoretical interest, the cortical canvas framework has practical design implications:

Scaling recipe. To scale ROAM to larger, more complex tasks, allocate canvas positions proportional to information bandwidth, wire local regions densely and global regions sparsely with $k_{\text{global}} \approx \log N$, and set max loops proportional to the depth of the required computational cascade. This recipe is directly derived from the cortical analogy and can be applied without expensive hyperparameter search.

Morphology-agnostic control. The dynamic canvas layout (Section 3 of the DYNAMIC_MORPHOLOGY document) enables a single model to control arbitrary robot bodies. The cortical analogy predicts that this will work because physics is topology-independent, and the canvas layout is the only thing that encodes topology. This prediction is testable (Experiment 5.7).

Interpretability. The cortical parallel provides a vocabulary for interpreting the model's internal representations. Instead of opaque attention maps, we can analyze the model's behavior in terms of cortical magnification (which canvas regions are most active), functional specialization (which regions develop modality-specific representations), and metabolic allocation (which regions use the most loop iterations). This structured interpretability framework may be more useful than generic attention visualization.

6.3 Limitations

The most significant limitation is that the parallels developed here are at the computational and algorithmic levels, not the implementation level. We cannot conclude from architectural similarity that the model and the cortex are doing the same thing. They may be converging on the same organizational structure for different reasons, or the parallels may be superficial. The experiments proposed in Section 5 are designed to test whether the parallels have functional consequences — whether the cortical-inspired design choices actually improve performance over non-cortical alternatives. If they do, the framework is useful regardless of whether it is biologically "correct."

A second limitation is scale. Our experiments to date (v1 and v2) have been conducted at toy scale ($N = 120$, $d_{\text{model}} = 128 - 256$). Several of the cortical parallels (sparse connectivity, cortical magnification scaling) may only manifest at production scale ($N > 1000$). The CogVideoX graft provides a path to production scale, but the full experimental program requires significant compute.

A third limitation is the static nature of the current layout. Cortical organization is dynamic: cortical maps reorganize in response to altered input, injury, and learning. The current canvas layout is a fixed hyperparameter. Making it learnable (via the soft allocation mechanism outlined in the CORTICAL_CANVAS document) is a natural next step but introduces optimization challenges (the layout loss landscape may be non-convex).

6.4 Future Directions

Thalamic attention routing. The thalamus serves as a gateway between cortical areas, routing information based on behavioral state. Adding a "thalamic" routing mechanism that dynamically adjusts which canvas regions can communicate would enable task-dependent connectivity.

Cerebellar forward model. The cerebellum is widely believed to implement a forward model for motor control [Wolpert et al., 1998]. Adding a separate lightweight network that predicts the next canvas state and provides prediction error as a conditioning signal could improve motor control performance.

Neuromodulatory conditioning. Replacing the scalar diffusion-timestep conditioning with a multi-dimensional "neuromodulatory" vector that encodes task state, novelty, reward history, and urgency could enable richer context-dependent processing.

Developmental curriculum. Instead of training on the full multimodal canvas from the start, begin with a "neonatal" canvas (coarse visual input, no language, simple motor output) and progressively expand, mirroring the developmental expansion of cortical function. This may improve training stability and final performance.

References

Alayrac, J.-B., et al. (2022). Flamingo: a Visual Language Model for Few-Shot Learning. NeurIPS.

Bassett, D. S. & Bullmore, E. T. (2006). Small-world brain networks. The Neuroscientist, 12(6), 512-523.

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The Long-Document Transformer. arXiv:2004.05150.

Bengio, E., et al. (2016). Conditional computation in neural networks for faster models. arXiv:1511.06297.

Braitenberg, V. & Schuz, A. (1998). Cortex: Statistics and Geometry of Neuronal Connectivity. Springer.

Brohan, A., et al. (2023). RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. arXiv:2307.15818.

Bullmore, E. T. & Sporns, O. (2012). The economy of brain network organization. Nature Reviews Neuroscience, 13(5), 336-349.

Carpenter, P. A., et al. (1999). Graded functional activation in the visuospatial system with the amount of task demand. Journal of Cognitive Neuroscience, 11(1), 9-24.

Catani, M. & Thiebaut de Schotten, M. (2008). A diffusion tensor imaging tractography atlas for virtual in vivo dissections. Cortex, 44(8), 1105-1132.

Daniel, P. M. & Whitteridge, D. (1961). The representation of the visual field on the cerebral cortex in monkeys. Journal of Physiology, 159(2), 203-221.

Dayan, P. (2012). Twenty-five lessons from computational neuromodulation. Neuron, 76(1), 240-256.

Dehghani, M., et al. (2019). Universal Transformers. ICLR.

Douglas, R. J. & Martin, K. A. C. (2004). Neuronal circuits of the neocortex. Annual Review of Neuroscience, 27, 419-451.

Felleman, D. J. & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. Cerebral Cortex, 1(1), 1-47.

Friston, K. (2005). A theory of cortical responses. Philosophical Transactions of the Royal Society B, 360(1456), 815-836.

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

Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. Journal of Neurophysiology, 61(2), 331-349.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. Biological Cybernetics, 23(3), 121-134.

Hasson, U., et al. (2015). Hierarchical process memory: memory as an integral component of information processing. Trends in Cognitive Sciences, 19(6), 304-313.

Hawken, M. J., Shapley, R. M., & Grosof, D. H. (1996). Temporal-frequency selectivity in monkey visual cortex. Visual Neuroscience, 13(3), 477-492.

Hawkins, J. & Ahmad, S. (2016). Why neurons have thousands of synapses, a theory of sequence memory in neocortex. Frontiers in Neural Circuits, 10, 23.

Innocenti, G. M. & Price, D. J. (2005). Exuberance in the development of cortical networks. Nature Reviews Neuroscience, 6(12), 955-965.

Jiang, A. Q., et al. (2024). Mixtral of Experts. arXiv:2401.04088.

Kaas, J. H. (1987). The organization of neocortex in mammals: implications for theories of brain function. Annual Review of Psychology, 38(1), 129-151.

Kanerva, P. (2009). Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. Cognitive Computation, 1(2), 139-159.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis. Frontiers in Systems Neuroscience, 2, 4.

Krubitzer, L. (2007). The magnificent compromise: cortical field evolution in mammals. Neuron, 56(2), 201-208.

Lamme, V. A. F. & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. Trends in Neurosciences, 23(11), 571-579.

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. Nature, 453(7197), 869-878.

Markov, N. T., et al. (2014). A weighted and directed interareal connectivity matrix for macaque cerebral cortex. Cerebral Cortex, 24(1), 17-36.

Marr, D. (1982). Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. MIT Press.

Merzenich, M. M., et al. (1984). Somatosensory cortical map changes following digit amputation in adult monkeys. Journal of Comparative Neurology, 224(4), 591-605.

Mountcastle, V. B. (1978). An organizing principle for cerebral function: the unit model and the distributed system. In The Mindful Brain (pp. 7-50). MIT Press.

Mountcastle, V. B. (1997). The columnar organization of the neocortex. Brain, 120(4), 701-722.

Murray, J. D., et al. (2014). A hierarchy of intrinsic timescales across primate cortex. Nature Neuroscience, 17(12), 1661-1663.

O'Leary, D. D. M., Chou, S.-J., & Sahara, S. (2007). Area patterning of the mammalian cortex. Neuron, 56(2), 252-269.

Peebles, W. & Xie, S. (2023). Scalable Diffusion Models with Transformers. ICCV.

Penfield, W. & Boldrey, E. (1937). Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. Brain, 60(4), 389-443.

Radford, A., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. ICML.

Rao, R. P. N. & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience, 2(1), 79-87.

Reed, S., et al. (2022). A Generalist Agent. arXiv:2205.06175.

Reynolds, J. H. & Heeger, D. J. (2009). The normalization model of attention. Neuron, 61(2), 168-185.

Shew, W. L. & Plenz, D. (2013). The functional benefits of criticality in the cortex. The Neuroscientist, 19(1), 88-100.

Stiles, J. & Jernigan, T. L. (2010). The basics of brain development. Neuropsychology Review, 20(4), 327-348.

Van Essen, D. C. (2003). Organization of visual areas in macaque and human cerebral cortex. In The Visual Neurosciences (pp. 507-521). MIT Press.

Vanrullen, R. & Thorpe, S. J. (2002). Surfing a spike wave down the ventral stream. Vision Research, 42(23), 2593-2615.

Vinken, K., et al. (2017). Adaptation can explain evidence for encoding of probabilistic information in macaque inferior temporal cortex. Current Biology, 27(22), 3519-3525.

Wandell, B. A., Dumoulin, S. O., & Brewer, A. A. (2007). Visual field maps in human cortex. Neuron, 56(2), 366-383.

Wolpert, D. M., Miall, R. C., & Kawato, M. (1998). Internal models in the cerebellum. Trends in Cognitive Sciences, 2(9), 338-347.

Yang, Z., et al. (2024). CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. arXiv:2408.06072.

Zaheer, M., et al. (2020). Big Bird: Transformers for Longer Sequences. NeurIPS.