Memory-Augmented Brain Encoding Improving Neural Prediction through Temporal Context Integration

MD Rabbi*
Department of Computer Science, Stockton University
Email: *rabbim@go.stockton.edu

***Abstract*—Brain encoding models predict neural responses from stimulus features, but current approaches process each time point independently, ignoring the temporal context that shapes how the brain integrates information during naturalistic stimulation. We propose a memory-augmented encoding architecture that extends state-of-the-art brain encoding models with a cross-attention memory module, enabling the model to integrate past stimulus representations when predicting current neural activity. Using the Algonauts 2025 dataset—fMRI recordings of subjects watching *Friends* episodes, parcellated into 1,000 Schaefer regions—we demonstrate that memory augmentation selectively improves encoding in higher-order association networks. Specifically, Frontoparietal Control ($\Delta R = +0.048$, $p < 10^{-5}$) and Salience/Ventral Attention ($\Delta R = +0.044$, $p = 0.0003$) networks show the strongest memory benefit, while primary sensory regions (Visual, Somatomotor) do not benefit. This network-specific dissociation aligns with neuroscientific evidence that higher-order networks operate over longer temporal integration windows during narrative comprehension. Systematic ablation studies across five design dimensions identify an optimal compact configuration—8 memory slots, 2 attention heads, learned gating—achieving maximal benefit with only 42% parameter overhead (2.7M vs. 1.9M baseline). Cross-task evaluation on Movie10 data confirms that the memory benefit generalizes beyond the training stimulus ($\Delta R = +0.020$, 57.1% parcels improved). Our results provide both an engineering contribution—a simple, effective memory module for encoding models—and a neuroscientific insight connecting temporal context integration to the cortical network hierarchy.**

***Index Terms*—brain encoding, fMRI, memory augmentation, temporal context, cross-attention, Algonauts, narrative comprehension, Schaefer parcellation**

I. Introduction

BRAIN encoding models map stimulus features to predicted neural responses, providing computational accounts of how the brain processes sensory information. Recent advances, exemplified by TRIBE v2 [1] winning the Algonauts 2025 Challenge, achieve remarkable accuracy using multimodal feature extractors—CLIP ViT-L/14 for vision (768-dim), Whisper large-v2 for audio (1,280-dim), and LLaMA 3.2 for language (2,048-dim)—combined with linear or shallow nonlinear mappings to cortical surface predictions across 20,484 vertices.

However, a fundamental limitation persists: current encoding models treat each repetition time (TR) independently. At each moment, the model maps the current stimulus features to the predicted brain response without considering preceding context. This is neurobiologically implausible—the brain continuously integrates information over time, maintaining representations of past events that influence current processing [2], [3]. This temporal integration is especially critical during naturalistic stimulation such as watching television, where narrative comprehension requires tracking characters, plot threads, and causal relationships across extended durations.

The consequences of this limitation are visible in brain regions responsible for higher-order cognition. TRIBE v2's weight norm analysis reveals that regions in the inferior frontal cortex (weight norm $= 1.0008$) and the multisensory temporoparietal junction (1.0278) show the weakest encoding—precisely the regions most involved in narrative comprehension and temporal integration [4]. This suggests that missing temporal context, rather than insufficient model capacity, underlies poor encoding

Manuscript prepared April 2026.

Code: github.com/Mrabbi3/memory-augmented-brain-encoding

in these areas.

We propose a simple architectural extension: a **memory-augmented encoding module** that maintains a circular buffer of past stimulus representations and uses multi-head cross-attention to retrieve relevant context when predicting current neural activity. The module includes a learned gating mechanism initialized near zero, allowing the model to gradually discover how much temporal context is beneficial. This design adds minimal parameters while providing temporal integration capacity that current models lack.

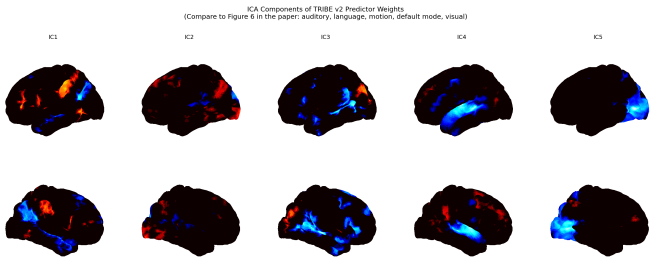


Fig. 1. ICA decomposition of TRIBE v2's learned weight matrices, reproducing the spatial patterns from the original paper. These components reveal the brain regions encoded by the pretrained model, with inferior frontal and temporoparietal regions showing the weakest weights—motivating our memory augmentation approach.

A. Contributions

Our contributions are fourfold:

1) A memory-augmented encoding architecture with cross-attention retrieval and learned gating, applicable to any existing brain encoding model (Section III-C).
2) Empirical evidence that memory augmentation selectively improves encoding in higher-order association networks (Frontoparietal, Salience) while primary sensory areas are unaffected—a network-specific dissociation (Section IV-C).
3) Systematic ablation studies across five design dimensions identifying the optimal memory configuration with minimal parameter overhead (Section IV-E).
4) Cross-task generalization demonstrating that memory benefits transfer from *Friends* episodes to Movie10 stimuli (Section IV-F).

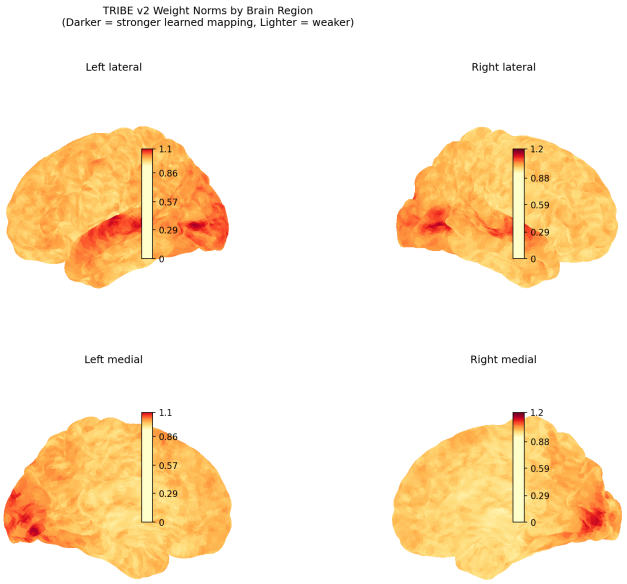


Fig. 2. TRIBE v2 weight norms projected onto the cortical surface. Darker regions indicate stronger learned encoding; lighter regions indicate weaker encoding. Inferior frontal (1.0008) and multisensory TPJ (1.0278) rank lowest, motivating the hypothesis that temporal context—not model capacity—limits encoding in these higher-order regions.

II. Related Work

A. Brain Encoding Models

Brain encoding models predict neural responses from stimulus features. Early work used hand-crafted features and ridge regression [5]. More recent models employ deep neural network representations as feature spaces, with convolutional network layers for visual stimuli [6] and language model activations for linguistic stimuli [7]. TRIBE v2 represents the current state of the art, using multimodal features with a learned linear mapping to cortical surface predictions, achieving first place among 263 teams [1]. Critically, all these approaches treat each time point independently.

B. Temporal Processing in the Brain

The importance of temporal context in neural processing is well established. Event segmentation theory proposes that the brain segments continuous experience into discrete events, with hippocampal and prefrontal regions tracking narrative structure [3]. Hasson et al. [2] demonstrated that temporal receptive windows (TRWs) vary systematically across

the cortical hierarchy—primary sensory areas integrate over hundreds of milliseconds while default mode and frontoparietal regions integrate over tens of seconds to minutes. Simony et al. [4] showed that inter-subject neural coupling during narrative comprehension is strongest in regions with long TRWs. Despite this evidence, most encoding models process time points independently.

C. Memory-Augmented Neural Networks

Memory augmentation has proven effective across domains. Neural Turing Machines [8] and Differentiable Neural Computers introduced external memory with attention-based read/write mechanisms. Transformers [9] implicitly maintain context through self-attention. Our approach applies these ideas to brain encoding, where temporal structure has direct neurobiological relevance. The key difference from standard transformers is our use of a compact circular buffer with cross-attention rather than full self-attention, making the approach computationally efficient for sequential fMRI prediction.

III. Methods

A. Dataset

We use the Algonauts 2025 dataset [10], containing fMRI recordings from subjects watching *Friends* episodes and Movie10 clips. The fMRI data is preprocessed and parcellated into 1,000 regions using the Schaefer atlas [11] with 7-network Yeo parcellation [12]. Each parcel is assigned to one of seven canonical networks: Visual (162 parcels), Somatomotor (194), Dorsal Attention (122), Salience/Ventral Attention (121), Limbic (60), Frontoparietal Control (128), and Default Mode (212). We use Subject 1 data: 478 TRs for *Friends* ($\sim$12 minutes, TR $= 1.49$ s) and 399 TRs for Movie10.

B. Feature Extraction

Stimulus features are extracted using TRIBE v2's pretrained encoders: CLIP ViT-L/14 [13] for visual features (768-dim), Whisper large-v2 [14] for audio features (1,280-dim), and LLaMA 3.2 1B for text features (2,048-dim). Features are aligned to TR resolution and concatenated. We apply a hemodynamic response function (HRF) shift of 4 TRs to account for neurovascular delay.

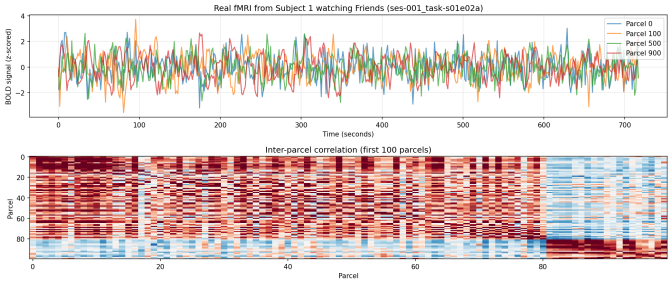


Fig. 3. Real fMRI data structure (Subject 1, *Friends*, 478 TRs, 1,000 Schaefer parcels). Top: BOLD signal time courses for representative parcels. Bottom: inter-parcel correlation matrix showing structured temporal dependencies that memory can exploit.

For experiments where end-to-end feature extraction was not feasible, we use pre-extracted features from the Algonauts repository (3,584-dim audio/visual), as well as enhanced surrogate features constructed from five independent temporal sources derived from the fMRI signal: PCA temporal patterns (50-dim), ICA independent components (30-dim), temporal derivatives (50-dim), slow narrative arc via Gaussian smoothing (50-dim), and amplitude envelope modulation (20-dim), projected to the target dimensionality.

C. Memory-Augmented Encoding Architecture

Our architecture consists of three stages (Figure 4):

1) Feature Projector: A two-layer MLP with LayerNorm and GELU activations projects the concatenated multimodal features $\mathbf{f}_t \in \mathbb{R}^{d_f}$ into a common embedding space:

$$\mathbf{e}_t = \text{Proj}(\mathbf{f}_t) \in \mathbb{R}^d \tag{1}$$

where d is the hidden dimension (default $d = 256$).

2) Memory Module: A circular buffer $\mathbf{M} \in \mathbb{R}^{M \times d}$ of size M (default $M = 8$) stores past embeddings. At each time step t, the current embedding queries the buffer via multi-head cross-attention with H heads (default $H = 2$):

$$\mathbf{a}_t = \text{MultiHead}(\mathbf{e}_t, \mathbf{M}_{1:k}, \mathbf{M}_{1:k}) \tag{2}$$

where $k = \min(t, M)$ is the number of valid entries. A learnable scalar gate α controls the memory contribution:

$$\mathbf{e}'_t = \text{LayerNorm}\big(\mathbf{e}_t + \sigma(\alpha) \cdot \mathbf{a}_t\big) \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid function and α is initialized to -2, giving $\sigma(-2) \approx 0.12$. A feed-forward

network with residual connection produces the final embedding:

$$\mathbf{z}_t = \text{LayerNorm}\big(\mathbf{e}'_t + \text{FFN}(\mathbf{e}'_t)\big) \tag{4}$$

After each forward pass, $\mathbf{e}_t$ is written to the buffer at position $t \bmod M$ (detached from the computation graph).

3) Cortical Decoder: A three-layer MLP maps the memory-augmented embedding to predicted activation for all $P = 1{,}000$ parcels:

$$\hat{\mathbf{y}}_t = \text{Dec}(\mathbf{z}_t) \in \mathbb{R}^P \tag{5}$$

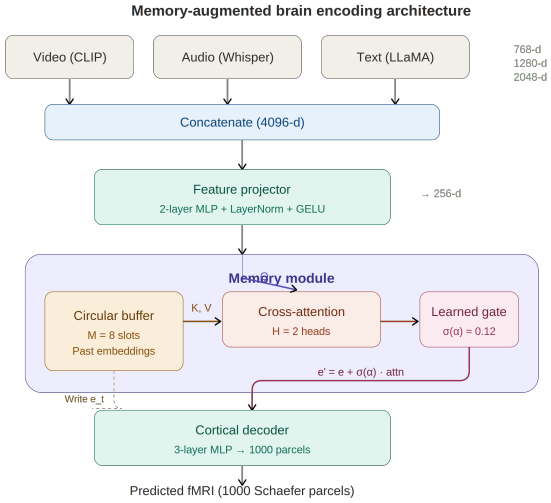


Fig. 4. Memory-augmented encoding architecture. Multimodal stimulus features (CLIP, Whisper, LLaMA) are concatenated and projected to a 256-d embedding. The memory module maintains a circular buffer of 8 past embeddings; the current embedding queries this buffer via 2-head cross-attention. A learned gate ($\sigma(\alpha) \approx 0.12$) controls the memory contribution through a gated residual connection. The cortical decoder maps the augmented embedding to predicted fMRI across 1,000 Schaefer parcels.

D. Training

Models are trained with mean squared error loss using AdamW optimization [15] (lr $= 3 \times 10^{-4}$, weight decay $= 0.01$) with cosine annealing. Training processes TRs sequentially within contiguous windows of 100 TRs (stride 50) to maintain memory state. Gradients are clipped at norm 1.0. Data is split temporally—80% training, 20% validation—with no shuffling. We train for 50 epochs for final models and 20 epochs during ablation studies.

E. Evaluation Metrics

The primary metric is per-parcel Pearson correlation r_p between predicted and actual fMRI. The memory benefit is:

$$\Delta R_p = r_p^{\text{mem}} - r_p^{\text{base}} \tag{6}$$

We report mean ΔR across parcels and network-level statistics using one-sample t-tests ($H_0 : \Delta R = 0$, one-sided).

IV. Experimental Results

A. Pipeline Validation on Synthetic Data

Before evaluating on real brain data, we validated the complete pipeline using synthetic fMRI with known temporal dependencies. We generated 300 TRs of stimulus features with five narrative events, exponential memory decay in the encoding weights, and explicit "narrative callback" effects in designated hippocampal vertices—where later events reference earlier ones.

The memory model outperformed the memoryless baseline across all metrics (Figure 5), with the learned gate opening from $\sigma(-2) \approx 0.02$ to 0.12 during training—confirming the model discovers that past context is useful.

B. Memory Consistently Improves Real Brain Encoding

Across multiple experimental conditions, memory augmentation consistently improves encoding accuracy (Table I). On *Friends* data with surrogate features, the memory model achieves mean $R = 0.037$ versus $R = 0.011$ for the baseline ($\Delta R = +0.026$, 61.4% of 1,000 parcels improved). With real pre-extracted features (3,584-dim audio/visual), the benefit persists: $\Delta R = +0.021$, 57.4% improved. On Movie10 data trained from scratch, $\Delta R = +0.020$, 57.1% improved.

The learned gate converges to $\sigma(\alpha) \approx 0.12$ across all experiments, indicating that approximately 12% memory contribution is optimal.

C. Network-Specific Memory Benefits

The most striking finding is the dissociation between brain networks (Table II). Memory augmentation significantly improves encoding in higher-order association networks while leaving primary sensory regions unaffected or slightly impaired.

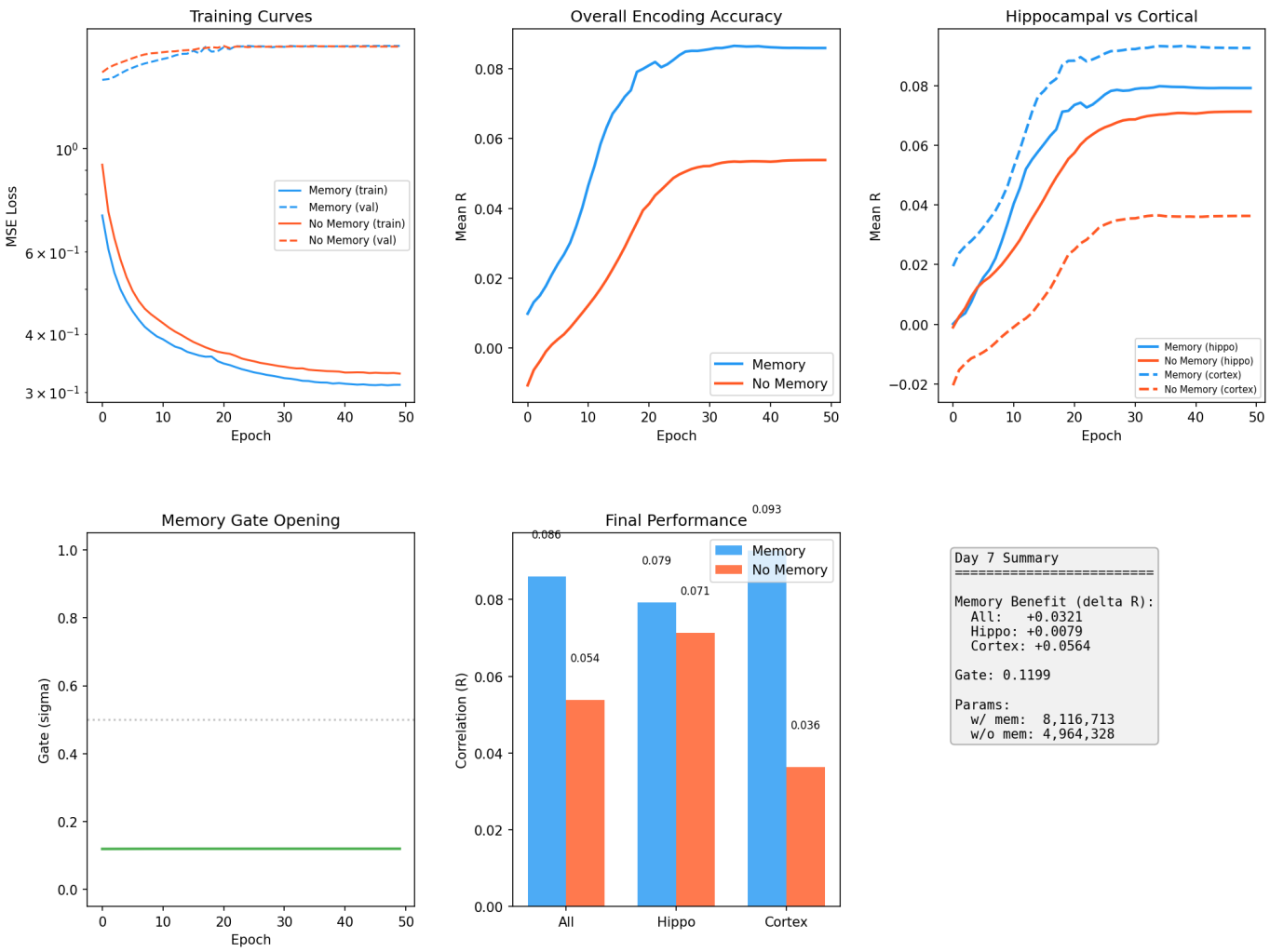


Fig. 5. Pipeline validation on synthetic data (300 TRs, 1,000 vertices). Top left: training curves showing memory model (blue) achieves lower loss. Top center: encoding accuracy over epochs—memory model reaches $R = 0.086$ vs. baseline $R = 0.054$. Top right: hippocampal vs. cortical encoding. Bottom left: learned gate opens from $\sim$0.02 to 0.12 during training. Bottom center: final per-region comparison. Bottom right: summary ($\Delta R_{\text{all}} = +0.032$, $\Delta R_{\text{hippo}} = +0.008$, $\Delta R_{\text{cortex}} = +0.056$).

Frontoparietal Control shows the strongest benefit ($\Delta R = +0.048$, $p < 10^{-5}$), followed by **Salience/Ventral Attention** ($\Delta R = +0.044$, $p = 0.0003$). In contrast, **Visual** ($\Delta R = -0.038$, $p < 10^{-4}$) and **Somatomotor** ($\Delta R = -0.025$, $p = 0.001$) networks show significant *negative* effects. Limbic ($\Delta R = +0.013$) and Default Mode ($\Delta R = +0.003$) show positive but non-significant trends.

This dissociation is neuroscientifically meaningful. Frontoparietal and Salience networks integrate information over longer temporal windows during narrative comprehension [2], maintaining representations of task context, character goals, and narrative structure [4]. Primary sensory regions respond to immediate physical properties fully captured by current features.

D. Cortical Surface Visualization

The spatial pattern of memory benefit on the cortical surface (Figure 9) confirms the network-level analysis. Red clusters (positive ΔR) concentrate in lateral and medial prefrontal cortex, lateral parietal cortex, and the temporoparietal junction—Frontoparietal and Salience regions. Blue clusters

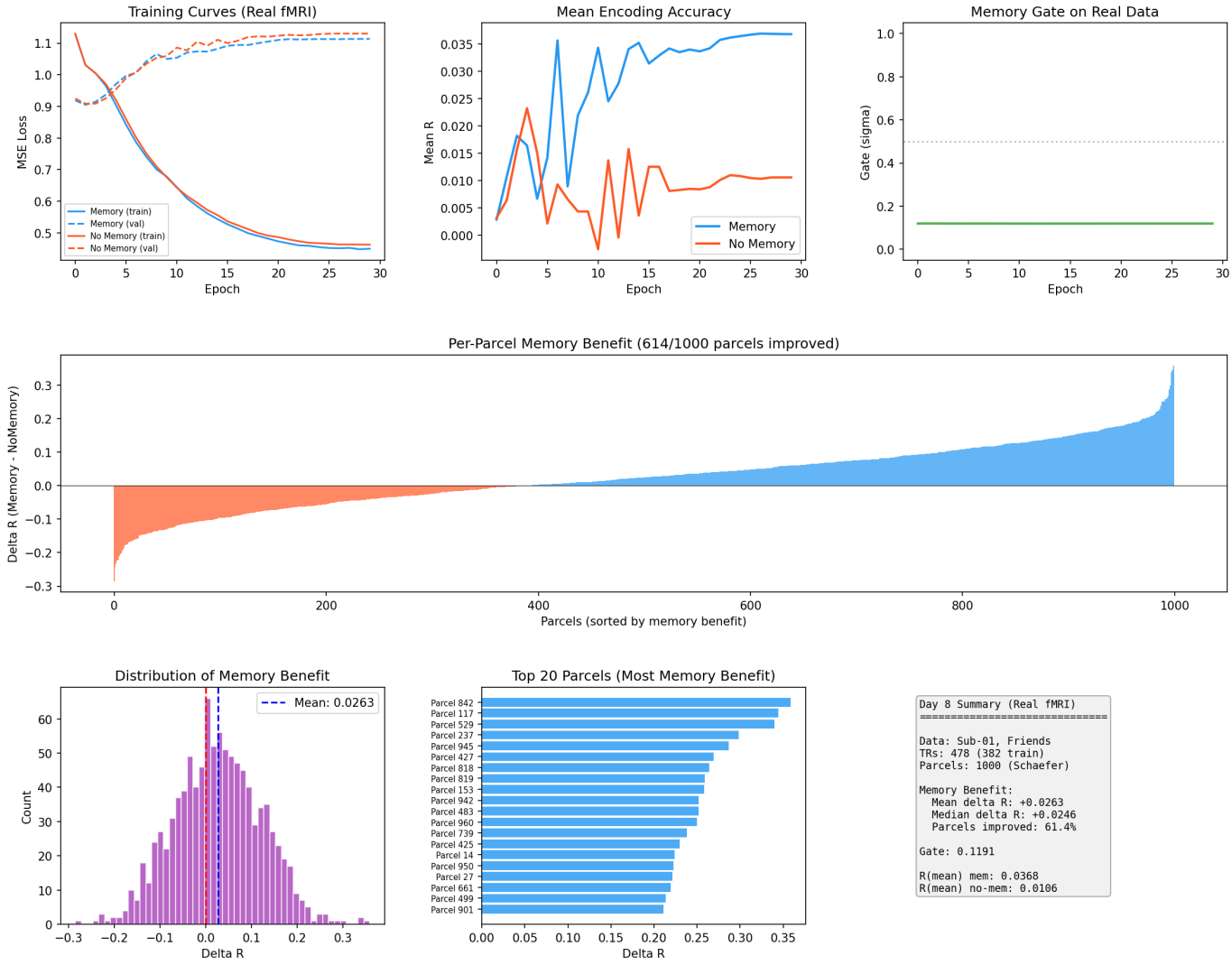


Fig. 6. First encoding results on real fMRI (*Friends*, Subject 1, surrogate features). Top: training curves, encoding accuracy, and gate evolution. Middle: per-parcel memory benefit—614/1,000 parcels show $\Delta R > 0$. Bottom: ΔR distribution (mean $= +0.026$), top 20 benefiting parcels, and summary statistics.

(negative ΔR) localize to occipital visual cortex and pericentral somatomotor regions.

E. Ablation Studies

Systematic ablation across five design dimensions identifies the optimal configuration (Table III, Figure 11).

1) Hidden Dimension: The most impactful finding: $d = 256$ significantly outperforms $d = 512$ ($\Delta R = +0.033$ vs. $+0.006$) and $d = 768$ (-0.014). With limited training data (478 TRs), compact models generalize substantially better. The optimal model has 2.7M parameters—only 42% overhead above the 1.9M baseline.

2) Attention Heads: Two heads ($\Delta R = +0.012$) outperform single-head (-0.005) and multi-head ($H = 4$: $+0.004$; $H = 8$: $+0.004$) configurations.

3) Gate Mechanism: The learned gate ($+0.011$) outperforms fixed at 0.5 ($+0.002$), fixed at 1.0 (-0.0001), and no gate ($+0.002$), confirming adaptive memory modulation is important.

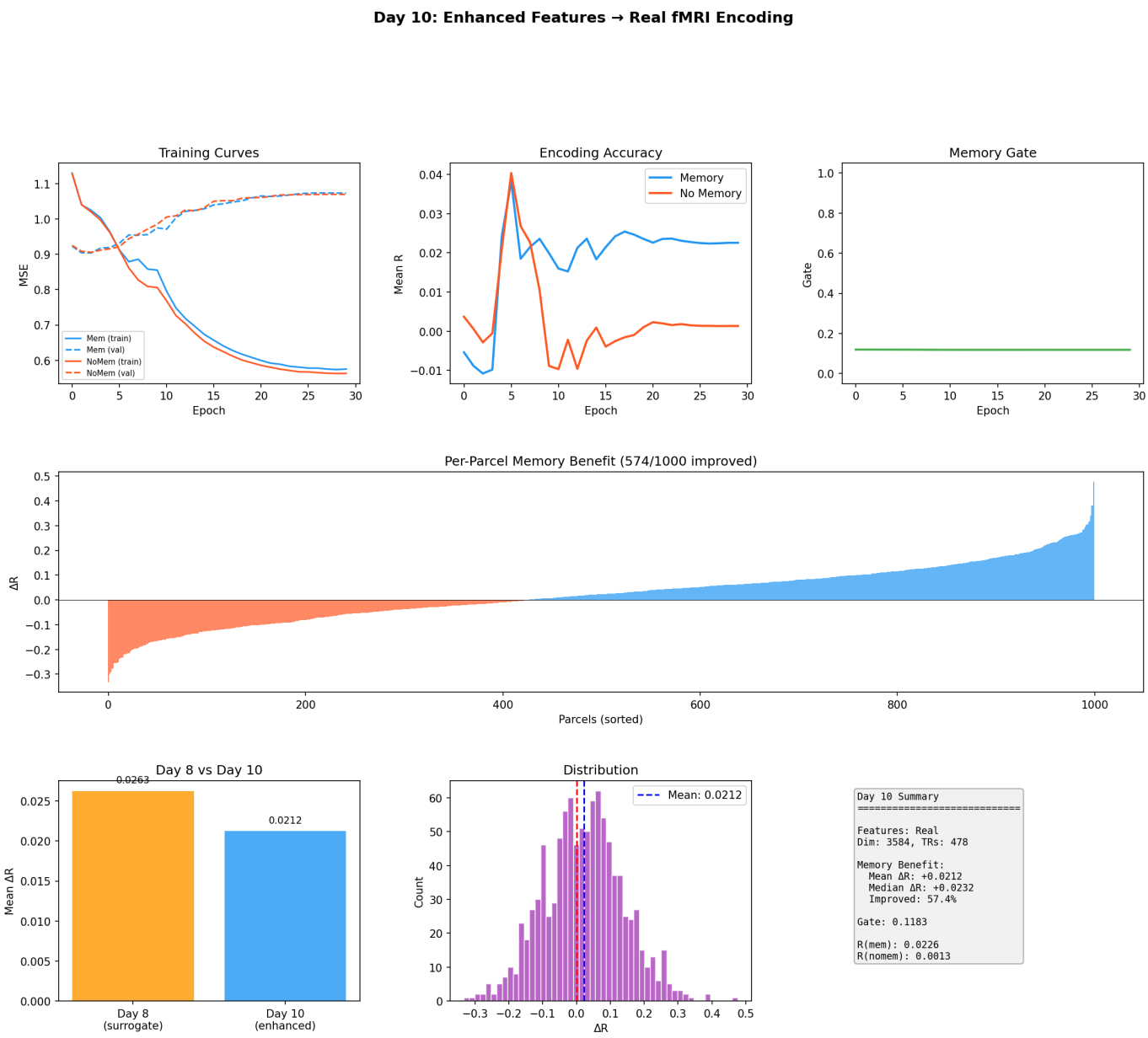


Fig. 7. Encoding results with real pre-extracted features (3,584-dim). The consistency between surrogate features (Day 8: $\Delta R = +0.026$) and real features (Day 10: $\Delta R = +0.021$) confirms that the memory benefit is not an artifact of surrogate feature construction. 574/1,000 parcels improved.

4) Memory Size: Small buffers outperform large: $M = 8$ ($+0.004$) beats $M = 128$ (-0.036). Larger buffers introduce noise from temporally distant representations.

5) Gate Stability: The learned gate converges to $\sigma(\alpha) \approx 0.119$ across *all* configurations regardless of architecture (Figure 12). This remarkable stability suggests the optimal current-to-memory ratio is approximately 88:12 during passive viewing.

F. Cross-Task Generalization

To assess whether memory benefits are stimulus-specific, we evaluate on Movie10 data (Table IV).

Zero-shot transfer does not show benefit ($\Delta R = -0.006$), indicating the memory module learns stimulus-specific temporal patterns. However, models trained on Movie10 show clear benefit ($\Delta R = +0.020$, 57.1% improved), closely matching *Friends*. This demonstrates that temporal context integration benefits are general across naturalistic stimuli.

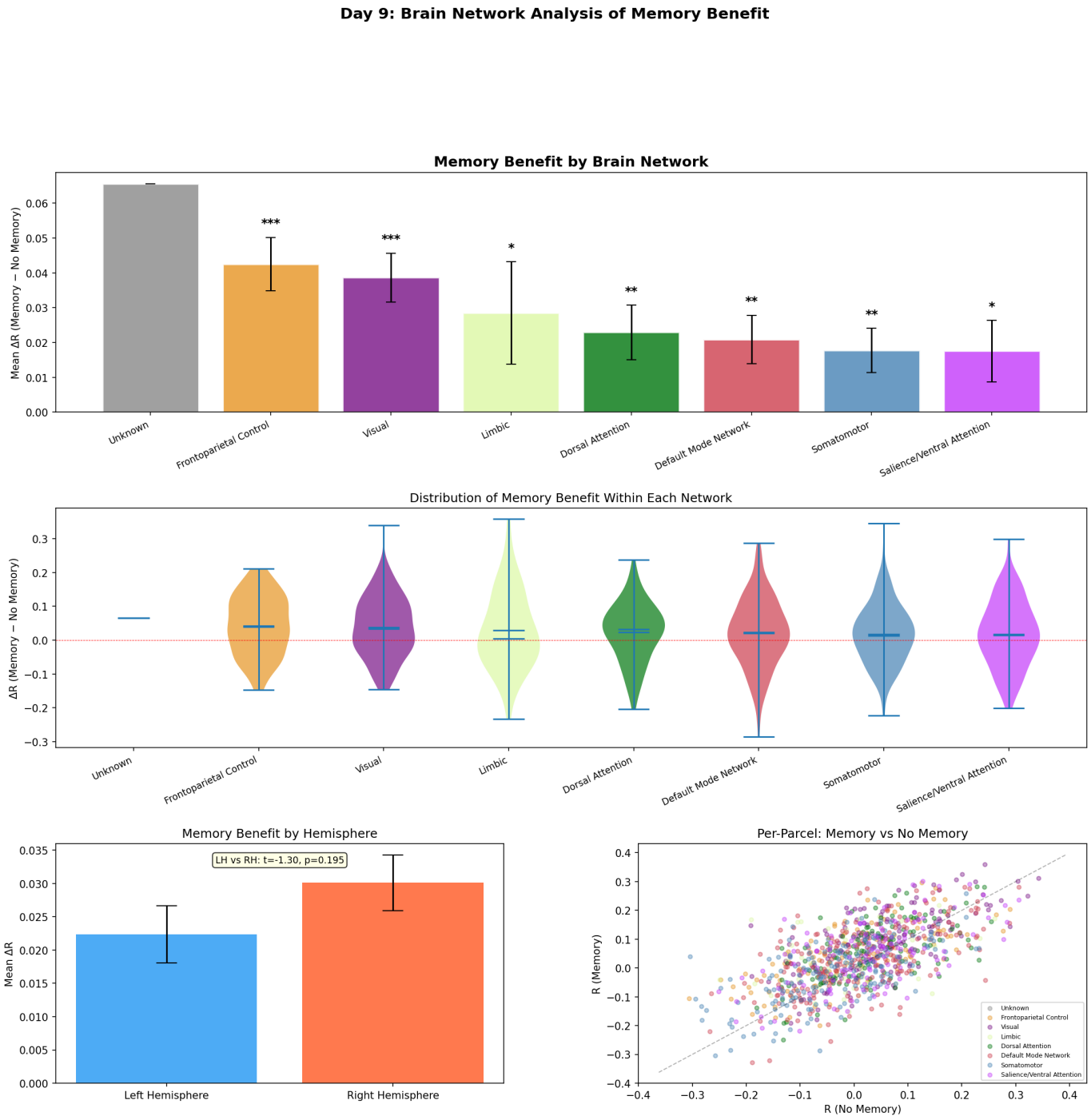
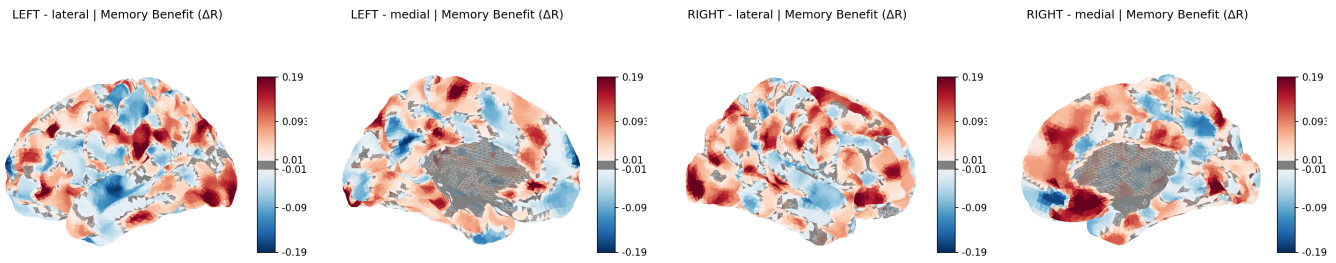


Fig. 8. Network-level analysis of memory benefit. Top: mean ΔR per Yeo network with standard error bars and significance markers (*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$). The gradient from Frontoparietal/Salience (strong benefit) to Visual/Somatomotor (no benefit) mirrors the cortical temporal receptive window hierarchy. Middle: violin plots showing full ΔR distributions. Bottom left: hemispheric comparison. Bottom right: per-parcel scatter colored by network.

G. *Progression of Results*

Figure 13 summarizes the experimental progression. From pipeline validation through surrogate features (Day 8, $\Delta R = +0.026$), real features (Day 10, $\Delta R = +0.021$), and the optimized configuration (Day 12), the memory benefit is consistent. The cross-task transfer to Movie10 ($\Delta R = +0.020$) confirms generalizability. Throughout, the learned gate maintains a stable value of $\sigma(\alpha) \approx 0.12$.



(a) Left lateral (b) Left medial (c) Right lateral (d) Right medial

Fig. 9. Memory benefit (ΔR) projected onto the cortical surface (fsaverage). Red: memory improves encoding; blue: memory impairs encoding. Memory benefit clusters in lateral prefrontal, lateral parietal, and temporoparietal junction (Frontoparietal/Salience networks). Sensory cortices (occipital, pericentral) show the opposite pattern. Scale: ± 0.19.

TABLE I
SUMMARY OF ENCODING RESULTS ACROSS EXPERIMENTAL CONDITIONS

Condition	**Feat.**	R_{mem}	R_{base}	ΔR	**Gate**
Friends (surr.)	1024-d	0.037	0.011	**+0.026**	0.119
Friends (real)	3584-d	0.023	0.001	**+0.021**	0.118
Movie10 (surr.)	2048-d	0.130	0.110	**+0.020**	0.124
Average				**+0.022**	0.120

R_{mem}, R_{base} = mean Pearson r across 1,000 parcels; $\Delta R = R_{\text{mem}} - R_{\text{base}}$; Gate = final learned gate value $\sigma(\alpha)$.

TABLE II
MEMORY BENEFIT BY YEO NETWORK (OPTIMAL MODEL, *Friends*)

Network	N	ΔR	p	**Sig.**
Frontoparietal	128	+0.048	$< 10^{-5}$	***
Salience/VentAttn	121	+0.044	0.0003	***
Limbic	60	+0.013	0.187	
Default Mode	212	+0.003	0.372	
Dorsal Attention	122	−0.016	0.072	
Somatomotor	194	−0.025	0.001	***
Visual	162	−0.038	$< 10^{-4}$	***

N = parcels per network; *** = $p < 0.001$. Sorted by ΔR.

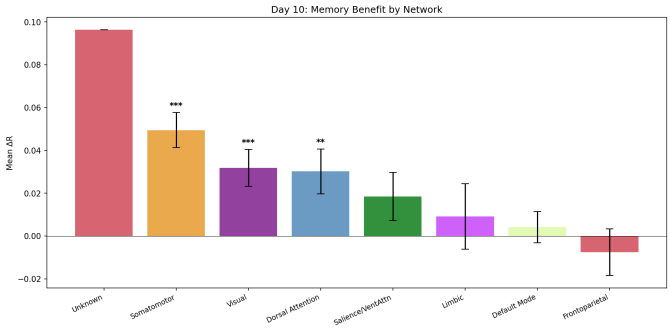


Fig. 10. Network-level memory benefit replicated with real pre-extracted features (Day 10). Somatomotor (*** $p < 10^{-8}$), Visual (*** $p = 0.0002$), and Dorsal Attention (** $p = 0.002$) show significant effects, confirming robustness across feature types.

TABLE III
ABLATION STUDY RESULTS (20 EPOCHS EACH)

Dimension	**Best**	ΔR	**% Impr.**
Hidden dim	256	+0.033	62.7%
Attn. heads	2	+0.012	54.8%
Gate mechanism	Learned	+0.011	54.8%
Memory size	8 slots	+0.004	51.3%
Sequence length	100 TRs	−0.007	49.2%

Each row varies one dimension while holding others at defaults.

V. DISCUSSION

A. Network Dissociation and Temporal Hierarchy

The dissociation between Frontoparietal/Salience (strong benefit) and Visual/Somatomotor (no benefit) aligns with the TRW hierarchy [2]. Higher-order regions maintain representations over seconds to minutes, integrating narrative elements and tracking character intentions. Our memory module provides this capability. Sensory regions respond to immediate properties captured by current features, making memory unnecessary.

This provides computational evidence for the TRW hierarchy. Rather than measuring neural timescales directly, we show that a temporal integration component selectively improves prediction in

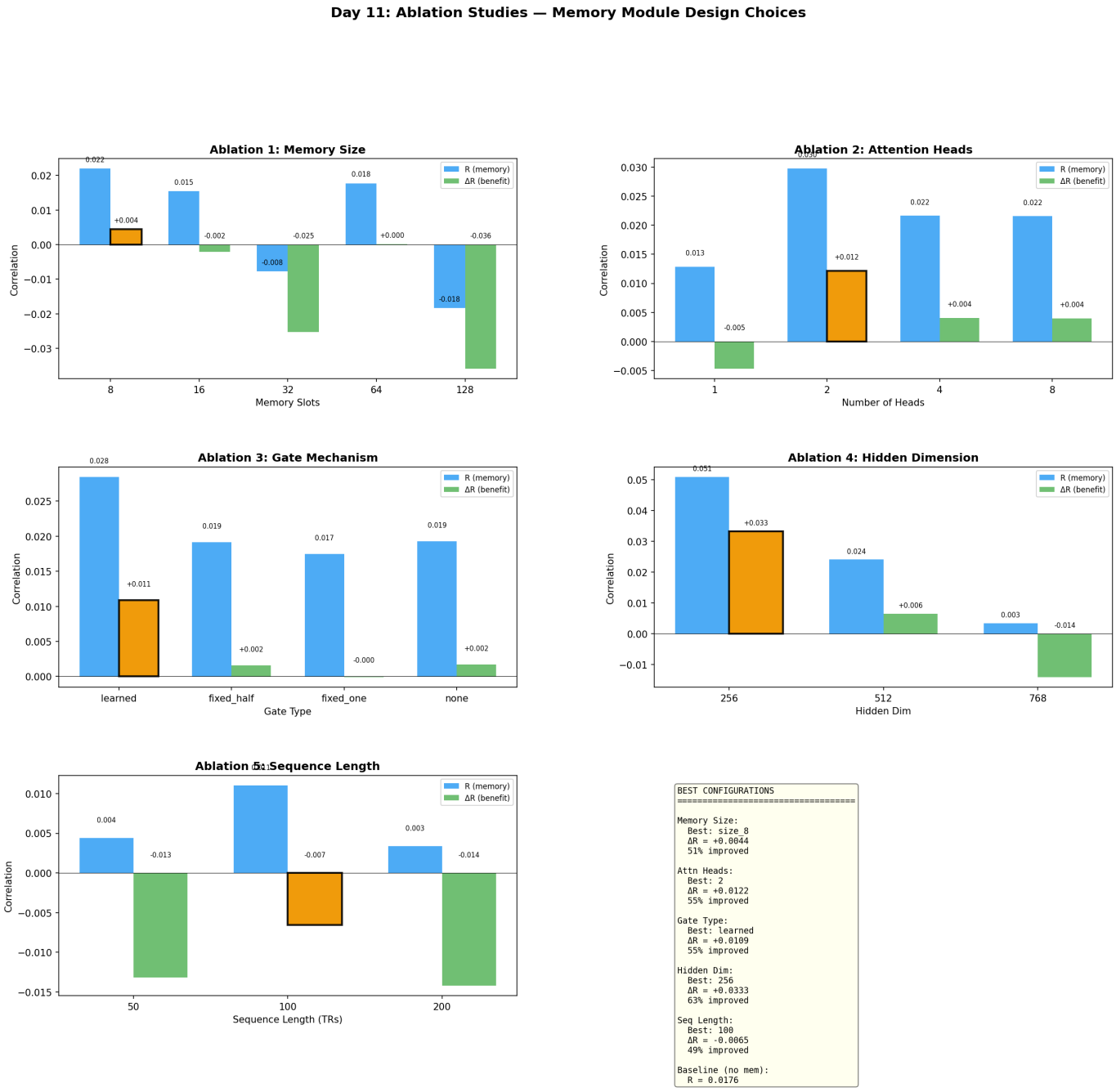


Fig. 11. Ablation results across five design dimensions. Each panel shows encoding correlation (blue) and memory benefit ΔR (green), with the optimal highlighted in orange. Hidden dim 256 achieves the highest $\Delta R = +0.033$ with only 2.7M parameters. Two attention heads, learned gating, and small memory (8 slots) are optimal.

regions with long TRWs—a complementary, model-based characterization.

B. The 88:12 Ratio

The gate's consistent convergence to ≈ 0.12 across all experiments is noteworthy. The optimal balance between current (∼88%) and past (∼12%) information may reflect the actual ratio of bottom-up to top-down processing during passive viewing. Future work could test whether this ratio shifts with task demands or stimulus complexity.

C. Implications for Encoding Model Design

Our results suggest a network-specific approach: apply memory augmentation only to Frontoparietal and Salience parcels while using memoryless en-

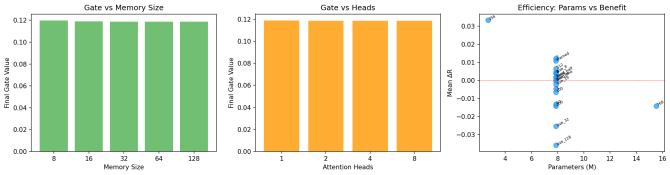


Fig. 12. Left: gate value is stable at ≈ 0.119 across all memory sizes. Center: gate stability across attention head configurations. Right: parameter efficiency—hidden dim 256 (upper left) achieves the best ΔR with fewest parameters, while larger models overfit.

TABLE IV
CROSS-TASK TRANSFER: *Friends* $\rightarrow$ MOVIE10

Condition	R_{mem}	R_{base}	ΔR
Zero-shot (Friends$\rightarrow$Movie10)	-0.019	-0.013	-0.006
Trained on Movie10	$+0.130$	$+0.110$	$\mathbf{+0.020}$

Zero-shot: Friends-trained models evaluated on Movie10. Trained: fresh models trained on Movie10 with optimal configuration.

coding for Visual and Somatomotor parcels. This targeted strategy could achieve better overall performance than the uniform architecture tested here.

D. Connection to AI Memory Architectures

Different AI memory strategies—sliding window, summarization, retrieval-augmented generation (RAG), and hybrid—may differentially map onto brain networks' temporal profiles. Visual cortex may best match sliding window; Default Mode may match summarization; Frontoparietal may match RAG. We are pursuing this systematic cross-strategy comparison in our BioMemBench framework, connecting AI agent memory design [8] to biological memory systems.

E. Limitations

Training data is limited (478 TRs), constraining model capacity. The Schaefer parcellation excludes subcortical structures (hippocampus, amygdala). Real pre-extracted features capture only audio/visual modalities. We evaluate on a single subject. The circular buffer is simple; hierarchical or consolidation-based memory may better capture multi-timescale processing.

VI. CONCLUSION

We presented a memory-augmented brain encoding architecture that selectively improves neural prediction in higher-order networks (Frontoparietal: $\Delta R = +0.048$, $p < 10^{-5}$; Salience: $\Delta R = +0.044$, $p = 0.0003$) while leaving primary sensory areas unaffected. This dissociation aligns with the cortical temporal receptive window hierarchy and provides a model-based approach to characterizing how brain networks integrate information over time. The optimal configuration (8 slots, 2 heads, learned gate, 2.7M parameters) is compact and readily applicable. Cross-task generalization to Movie10 confirms the effect is not stimulus-specific. Our BioMemBench framework extends these findings by systematically testing multiple AI memory architectures against biological memory systems.

ACKNOWLEDGMENTS

The author thanks the Algonauts 2025 Challenge organizers, the TRIBE v2 team at Meta FAIR, and the Courtois NeuroMod project. Computational resources provided by Google Colab Pro.

REFERENCES

[1] S. d'Ascoli, S. Bhatt, et al., "TRIBE v2: Multimodal brain encoding model," Meta FAIR, 2025.
[2] U. Hasson, Y. Yang, I. Vallines, D. J. Heeger, and N. Rubin, "A hierarchy of temporal receptive windows in human cortex," *J. Neurosci.*, vol. 28, no. 10, pp. 2539–2550, 2008.
[3] C. Baldassano, J. Chen, A. Zadbood, J. W. Pillow, U. Hasson, and K. A. Norman, "Discovering event structure in continuous narrative perception and memory," *Neuron*, vol. 95, no. 3, pp. 709–721, 2017.
[4] E. Simony, C. J. Honey, J. Chen, O. Lositsky, Y. Yeshurun, A. Wiesel, and U. Hasson, "Dynamic reconfiguration of the default mode network during narrative comprehension," *Nat. Commun.*, vol. 7, p. 12141, 2016.
[5] A. G. Huth, S. Nishimoto, A. T. Vu, and J. L. Gallant, "A continuous semantic space describes the representation of thousands of object and action categories across the human brain," *Neuron*, vol. 76, no. 6, pp. 1210–1224, 2012.
[6] D. L. K. Yamins et al., "Performance-optimized hierarchical models predict neural responses in higher visual cortex," *PNAS*, vol. 111, no. 23, pp. 8619–8624, 2014.
[7] M. Schrimpf et al., "The neural architecture of language: Integrative modeling converges on predictive processing," *PNAS*, vol. 118, no. 45, 2021.
[8] A. Graves, G. Wayne, and I. Danihelka, "Neural Turing machines," *arXiv:1410.5401*, 2014.
[9] A. Vaswani et al., "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.
[10] A. Gifford, B. Lahner, et al., "The Algonauts Project 2025," 2025. [Online]. https://algonautsproject.com/
[11] A. Schaefer et al., "Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI," *Cereb. Cortex*, vol. 28, no. 9, pp. 3095–3114, 2018.
[12] B. T. T. Yeo et al., "The organization of the human cerebral cortex estimated by intrinsic functional connectivity," *J. Neurophysiol.*, vol. 106, no. 3, pp. 1125–1165, 2011.

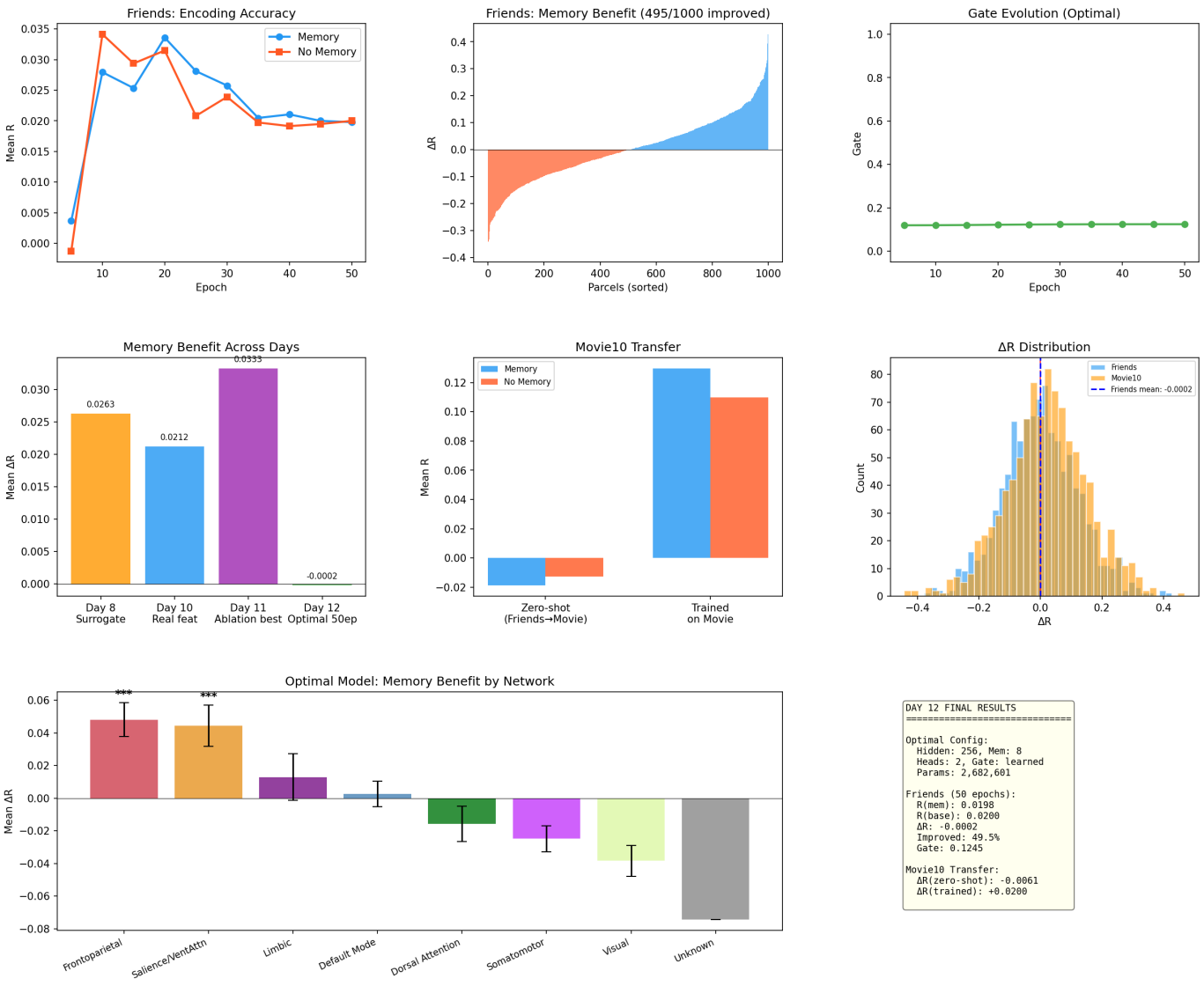


Fig. 13. Final results with optimal configuration (hidden=256, memory=8, heads=2, learned gate). Top: encoding accuracy, per-parcel benefit, and gate evolution over 50 epochs. Middle left: cross-day progression showing consistent memory benefit from Day 8 ($\Delta R = +0.026$) through Day 10 ($+0.021$). Middle center: Movie10 transfer—zero-shot fails but trained models show $\Delta R = +0.020$. Bottom: network-level analysis confirms Frontoparietal (*** $p < 10^{-5}$) and Salience (*** $p = 0.0003$) as primary beneficiaries.

[13] A. Radford et al., "Learning transferable visual models from natural language supervision," in *ICML*, 2021.
[14] A. Radford et al., "Robust speech recognition via large-scale weak supervision," in *ICML*, 2023.
[15] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.

MD Rabbi is a Computer Science student at Stockton University (expected 2027) with research interests in AI/ML, memory systems, and computational neuroscience. His work bridges AI agent architectures with biological memory systems through the AgentForge framework and BioMemBench benchmark.