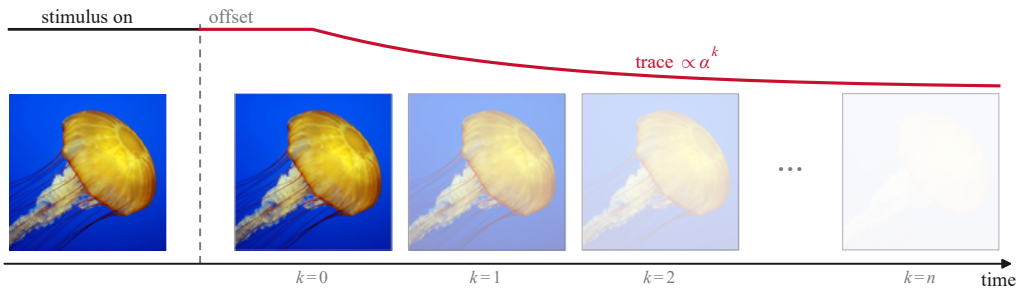
NEURAL NETWORKS SPONTANEOUSLY DISSOCIATE VISUAL FROM INFORMATIONAL PERSISTENCE USING ICONIC MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Tied classically to decaying persistence in early visual cortex, *iconic memory* is a high-capacity sensory trace that briefly outlasts its stimulus. Yet, fading within the cortex is not confined to any single region, nor does any reverberant activity that counteracts such fading cease uniformly. The ventral stream sustains retrograde insignia intractably, and the recurrence that re-amplifies this decaying activity exists at nearly every level of the visual hierarchy, rendering *iconic* behavior hard to isolate: *If a decaying, re-entrant visual trace is available everywhere, what distinguishes the stage that yields an* ***icon*** *from the rest?* Appended to neural networks where such stage-wise differentiation already exists, would a fading percept let us determine its functional use? Confer benefits beyond what a mere expansion in parameter space warrants? To find out, a proper test should heed all stages equally, letting the network determine its own specialized regions, *if any*. Consequently, we place the same lightweight ICONICBlock following each stage of quaternate-depth convolutional and attention-based classifiers—governed by a learnable retention factor (α) to index temporal persistence, re-entry gain (γ) to index feedback strength, succeeded by canonical divisive normalization, all initialized identically—and train on image classification tasks. Serendipitously, ResNet, DenseNet, ConvNeXt, ventral stream model CORnet-S, transformer model Swin-Tiny, despite their varied architectures, converge on the same final structure: All concentrate *iconic* persistence at the sensory input and the categorical readout—the two ends we argue double as network's artificial counterparts to dissociating early *visual persistence* and late *informational persistence* intra-cortex, the two telltale components of *iconic memory*. Dependent on natural image structure, this particular emergence finds no expression when non-ecological or phase-scrambled images are used during learning. Furthermore, the putatively V1-correspondent early gain is subsumed when fixed Gabor filters are prepended and, while the late stage jump is preserved, its divergence is amplified with increasing task complexity. Across architectures, ICONICBlocks yield relative clean gains exceeding 8%, mean corruption gains 18% and resilience in varied adversarial attack budgets—a rarely seen simultaneous improvement that through its robustness profile hints at improved alignment with the visual cortex, a finding we further corroborate with representational-similarity tests.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Deep neural networks model the primate ventral stream as a feedforward hierarchy of increasingly abstract representations (Yamins et al., 2014; Güçlü & van Gerven, 2015), with successive stages implying a corresponding temporal progression (Cichy et al., 2016; DiCarlo et al., 2012), an approach that has proven useful for both object recognition and neural prediction alike (Cadieu et al., 2014; Yamins & DiCarlo, 2016; Schrimpf et al., 2018). The biological visual stream, however, is not solely feedforward (Felleman & Van Essen, 1991; Lamme et al., 1998; Lamme & Roelfsema, 2000) nor is there any monotonic temporal course that rigidly orders its dynamics (Schmolesky et al., 1998; Bullier, 2001). Lateral and feedback connections, recirculating activity within and across cortical stages, are abundant (Felleman & Van Essen, 1991), allowing the representations to continue being shaped beyond the domain of any initial feedforward sweep (Kietzmann et al., 2019). As a result, certain sensory activity persists long after the stimulus that evoked it has disappeared—a phenomenon closely related to landmark vision literature on *iconic sensory memory* and *fading memory* (Sperling, 1960; Coltheart, 1980; Duysens et al., 1985; Benucci et al., 2009; Nikolić et al., 2009; Teeuwen et al., 2021) (Fig. 1). Yet sensory persistence within the cortex, coursing through a densely interconnected circuitry, is thus not easy to confine to a single cortical region: Lingering post-stimulus traces have already been observed in primary visual cortex (V1) (Nikolić et al., 2009; Teeuwen et al., 2021), higher ventral stream (V4) (Zamarashkina et al., 2020) and inferotemporal cortex (IT) (Rolls & Tovee, 1994; Kovács et al., 1995), where their usable duration appears to vary with the visual task that they are asked to sustain (Rensink, 2014). Iconic memory, on this emerging view, is less an inert and monotonic sensory buffer than a rapidly fading, but richly distributed, continuation of the representations through which vision ordinarily proceeds.

To map the elusive functional organization of cortical processing or, in our case, to pinpoint the exact locus of *persistence*, researchers have increasingly turned to computational models of vision for assistance. Most prominent among these models remain deep neural networks (Lindsay, 2021; Mehrer et al., 2021; Storrs et al., 2021; Mathis et al., 2024), whose successive differentiation of stages characterizes a biologically corresponding hierarchy of increasingly abstract representations similar to that of cortex (Yamins et al., 2014; Güçlü & van Gerven, 2015; Cichy et al., 2016). One model that makes the correspondence explicit is CORnet-S, whose fourfold stage distribution corresponds to cortical layers V1, V2, V4 and IT (Kubilius et al., 2019)—a network we use as a backbone and a biological mapping we adopt as an interpretive framework throughout.

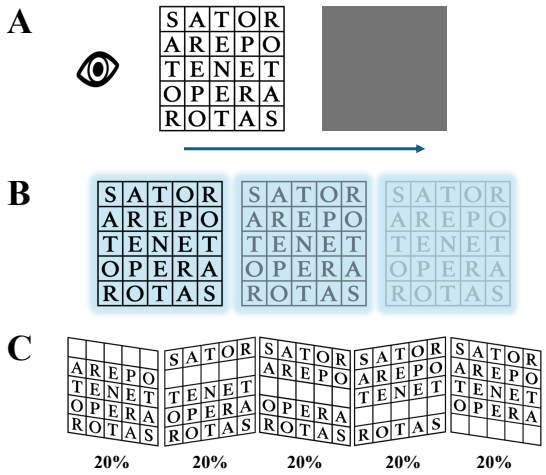


Figure 1: **Partial report of iconic persistence. (A)** A briefly viewed stimulus disappears from the visual field (offset). **(B)** Its representation inside cortex, though rapidly fading, persists longer. **(C)** Any of the five post-cued rows can be recalled with comparable accuracy, indicating spatially uniform retention of the fading percept.

Furthermore, research centered on the recurrent neural dynamics that sensory persistence is intimately tied to has also seen considerable progress, with broad computational parallels appearing in state-of-the-art architectures, such as Vision Mamba, whose state-retention dynamics echo the persistence and decay characteristics of cortical processing (Murray et al., 2014; Chaudhuri et al., 2015; Gu & Dao, 2024; Zhu et al., 2024). Nevertheless, for all the alleged successes in mirroring the ventral stream sequentially and the substantial progress made in incorporating these fuller recurrent dynamics, computational models of the ventral stream have seldom treated the *distribution of recurrence*, driven by deep computational demand, as an object of inference in itself, nor have they denoted what the temporal attrition of the representations conveys more than as a task-enforced caveat to be overcome. Indeed, while prior work has equipped convolutional stages with task optimized time constants, it has done so mainly as a means to comparing simple decay models against feedforward or recurrent processing (Nayebi et al., 2018). Other recurrent ar-

chitectures explicitly prescribe where the recurring should take place or how extensively each stage be reiterated, inspired mainly by primate vision (Kubilius et al., 2019). Yet, as a whole, these implementations still are insufficient to distinguish a hierarchy genuinely shaped by a network-wide persistence demand, from one whose amounts were merely stipulated beforehand or learned as auxiliary controls, which naturally leaves the following unresolved: *Given a hierarchical visual system in which each stage is offered a chance to retain its own fleeing activity, which stages actually do, by how much and what does an eventual distribution so uneven reveal about what each stage is for?*

We introduce **ICONICNet**, a family of augmented image classifiers designed to make these very questions testable. Each of the successive stages of a standard neural network receives the identically initialized, lightweight, persistence-capable IconicBlock: an intra-recurrent fading memory reservoir governed by learnable persistence parameter α and a re-entry gain parameter γ. Since no one stage is differentiated beforehand, any structural divergence must arise from the interaction between representational depth and task objectives of the network rather than from unequal architectural provision. Consequently, the final inter-stage profile reveals at which depth persistence should most strongly be retained. Across transformer-based, convolutional and recurrent ventral stream architectures (Kubilius et al., 2019; Liu et al., 2021; 2022), the learned α and γ that define the fleeting percept follow a monotonic increase and a non-monotonic U-shape, respectively.

Left free to allocate persistence, the networks thus recover a distinction long drawn in psychophysics between visible and informational persistence (Coltheart, 1980; Loftus & Irwin, 1998). They place the former at their seemingly V1-correspondent stage, where fading persistence has been most directly tied to iconic memory (Nikolić et al., 2009; Teeuwen et al., 2021; Yang et al., 2024) and the latter at the final, inferotemporal-correspondent stage, where classification pressure sustains category-selective structure beyond any raw sensory trace, much like temporal-cortex neurons continue, for hundreds of milliseconds after a brief flash, to process the stimulus as though it were still in view (Rolls & Tovee, 1994; Keysers et al., 2005).

The final hierarchy, while proves robust across varied computational backbones, is nevertheless not immutable. Ecological content and complexity of the handed visual task have varying impress on where inside the network persistence should most be sustained, consistent with accounts of a biomimetic network emergence.

2 BIOLOGICAL BACKGROUND

MEMORY Any notion of what constitutes knowing or utilizing information in organisms endowed with a nervous system is shaped, in one form or another, *by memory* (Tulving, 1985; Eichenbaum, 2000). Long-term memory consolidation, and therefore knowledge accumulation, depends on a cascade of short-lived memoranda, selection, stabilization and reorganization of which ultimately yield enduring cortical representations (Dudai, 2004; Frankland & Bontempi, 2005). Yet in a sensory world marked by inherent transience, even immediate memory must be able to retain some viable trace, beyond the stimulus from which it arose (Gegenfurtner & Sperling, 1993).

VISUAL PERSISTENCE At the very beginning of this remarked retention sits *iconic memory*—a fading, high-capacity sensory buffer that outlasts its stimulus by a few hundred milliseconds (Sperling, 1960; Neisser, 1967) and whose rapidly decaying afterimage can be easily effaced when a mask (new stimulus) is presented (Averbach & Coriell, 1961). Yet must *what survives* merely *remain visible*? When Coltheart (1980) decomposed the sustained trace into two distinct components, aside from the well-attested, visible persistence part that arised at an earlier sensory stage—the continued phenomenal experience of seeing the stimulus after offset—there appeared an additional, informational persistence part, the continued availability of information about that stimulus, which he termed iconic memory proper as emerges only at a considerably later, downstream processing stage. Such a distinction naturally raises more questions: *If the stimulus is gone and its visible mark already receding, in what form and in which reaches of the cortex can its information remain?*

INFORMATIONAL PERSISTENCE A more parsimonious account is that no fixed or fragmentary version of the percept need ever have existed. Along the visual hierarchy, image-bound structure yields to perceptual shape, through the same Gestalt grouping that temporal-integration studies require of the iconic trace for pattern completion (Eriksen & Collins, 1967; Di Lollo, 1980; Wagemans et al., 2012), and, ultimately, to representations increasingly organized by object category (Bracci &

Op de Beeck, 2016). Therefore, although able to coexist with its precursor, a trace of the vanished stimulus propagating across these levels does not need to remain in the form with which it began (Tomić & Bays, 2024; Matic et al., 2025; Akyürek, 2025). Regarding this late-comer persistence, multilayer studies of iconic memory draw an even more intriguing picture, demonstrating that iconic memory's usable duration varies with task demands, and positing, if such a variation does exist, it should be due to distinct levels of the visual hierarchy utilizing the percept in alternating forms (Rensink, 2014). Aside from the iconic nomenclature, it has been known that the deepest stages of the cortex employ a richer recurrent partake. Indeed, due to the cortical framework we adopt, it would not be surprising to observe that a module embedded in inferotemporal cortex (IT) will demonstrate dynamics more closely related to the informational part than to the visual one (Keysers et al., 2005; Kovács et al., 1995; Akrami et al., 2009; Kiani et al., 2007), with telltale properties, such as increased recurrence and processing demand during more challenging tasks (Spitzer & Richmond, 1991; Kar et al., 2019; Kar & DiCarlo, 2021).

AN ICON AS AN ARCHIVE What we can safely assume of informational persistence is, therefore, a consequential role in formation of object category. To decode a scene, the cortex works by integrating edges into shapes and shapes into Gestalt-like features that which deep neural networks, on separate occasions, have been demonstrated can mimic the same progression: with edges binding across early to intermediate layers into progressively integrated wholes that, at finalized depth, give way to an altogether categorical order (Zeiler & Fergus, 2014; Zeman et al., 2020). Given that the spatial hierarchy is already shared, can we expect, with ICONICBlocks granting each stage the *temporal freedom* it never had, the network to rediscover this same immanent hierarchy and *dissociate visual from informational persistence?*

3 METHODOLOGY

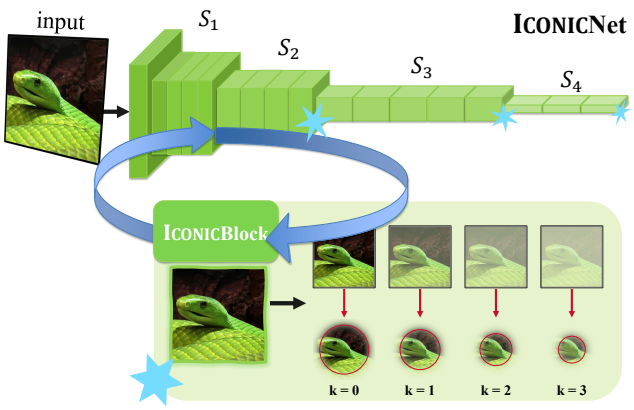


Figure 2: **ICONICNet architecture.** Each stage S_1–S_4 of a standard backbone receives an identically initialized ICONICBlock, placed broadly at the locations marked by stars that denote stage endpoints. Within each block, the stage activation is written once into a trace that fades geometrically across $K = 4$ internal steps, a gate reads from what remains at each step, and the accumulated readout reenters the stage through a learned gain.

NEURAL NETWORKS AND THE CORTICAL STAGES WITH WHICH THEY CORRESPOND While no exact neural scaffold exists that can precisely bridge cortical areas and the stages of neural networks, prior work has supported a coarser alignment along an axis of continuously changing representations (Yamins et al., 2014; Güçlü & van Gerven, 2015; Cichy et al., 2016). Following this notation, extending the canonical CORnet-S stage arrangements, where an architecture allows a natural four-fold decomposition, we broadly approximate these successive depths as a V1–V2–V4–IT progression. Where no such four-fold grouping exists, such as in EfficientNet (Tan & Le, 2019), we more broadly ask whether an increasing or non-monotonic hierarchy is at all present, either in the exact form or in a more elongated one.

INTRODUCING TIME INTO A SPATIAL VISUAL HIERARCHY Standard feedforward networks operating on two-dimensional images lack an intrinsic temporal axis—depth specifies a sequence of representational transformations, yet says nothing about how long any one of those transformations should remain available. Therefore, to instantiate the fading dynamics warranted of iconic persistence, we approximate time within each processing stage as a sequence of retrieval opportunities from an exponentially decaying trace. The offset—the moment when the iconic trace kicks in—required of iconic memory's instantiation is imposed by the same

limitation, which we manufacture: The activation the stage has already formed stands in for the stimulus, the trace written from it stands for the icon that outlasts it after offset and the internal steps for the successive delays at which partial report has repeatedly probed for an icon and found less and less of it reachable (Sperling, 1960; Averbach & Coriell, 1961). Within a single forward pass, then, a stage can read from a trace that is already fading.

The resultant ICONICBlock writes the stage activation into a trace once, at the manufactured moment of offset, lets that trace regulate access to a fixed content representation as it fades, and returns the accumulated readout to the stage through a learned gain (Fig. 2).

WRITING AND RETAINING A TRACE The stage activation x_ℓ is inscribed once, through a normalized projection, into a trace m_ℓ that is thereafter left only to fade,

$$m_{\ell,k} = \alpha_\ell^k \, m_\ell, \qquad k = 0, \dots, K-1, \tag{1}$$

over a small, fixed number K of internal steps, with $\alpha_\ell \in (0, 1)$ a learned retention factor for every stage. The exponential form follows the icon as it has been measured in macaque primary visual cortex, where post-stimulus activity predicts report along the same decaying course of iconic report (Teeuwen et al., 2021; Lu et al., 2005; Baek et al., 2016; Smith & Busch, 2026). The block is specified in full in Appendix A.

The trace that thus forms intra-stage receives no further input and undergoes no further transformation of its content across these steps—a deliberate departure from the recurrent models of the ventral stream that let each pass through a stage refine its representation (Kubilius et al., 2019; Nayebi et al., 2018; Kietzmann et al., 2019; Kar et al., 2019). Recurrent processing in V1 arrives late, carries figure-ground segregation and perceptual grouping and is abolished by a backward mask that leaves the stimulus-driven residue intact (Lamme & Roelfsema, 2000; Lamme et al., 2002; Fahrenfort et al., 2007). That residue is what predicts report after offset (Teeuwen et al., 2021), and the quantitative accounts of iconic availability accordingly model the buffer as a trace that relaxes passively toward quiescence (Zylberberg et al., 2009; Tomić & Bays, 2024). Restricting the trace to scalar decay follows the same norm (Appendix B).

READING WHAT REMAINS A trace that fades but is never consulted serves no purpose; the question is how the stage should draw on what remains at each step. What partial report loses with delay are reportable items, the precision of those still retained changing comparatively little (Pratte, 2018), and the models built on that paradigm have accordingly kept the information stored apart from the time-dependent process that governs access to it (Gegenfurtner & Sperling, 1993). We keep the same two apart by projecting a content representation q_ℓ once from the written trace and holding it fixed across steps, while the fading trace regulates a gate that determines how much of q_ℓ each step is allowed to pass through,

$$z_{\ell,k} = \sigma\left(W_\ell^x x_\ell + W_\ell^m m_{\ell,k}\right), \qquad u_\ell = \frac{1}{K} \sum_{k=0}^{K-1} z_{\ell,k} \odot q_\ell, \tag{2}$$

so that early steps, where the trace is still strong, admit nearly all of the stored content, while later steps, where the trace has decayed, admit progressively less—mirroring the declining availability that partial report measures with increasing cue delay (Sperling, 1960).

RE-ENTRY THROUGH GAIN CONTROL Once the intra-stage readout is complete and memory has completely faded, the trace must feed back into the stage as the experience of that memory. Because each block sits on activations of different scales and carries projections of its own, the readout passes through canonical divisive normalization in recurrent-amplification form (Carandini & Heeger, 2012; Heeger & Zemlianova, 2020) and admitted back through a learned gain γ_ℓ,

$$\hat{x}_\ell = x_\ell + \gamma_\ell \, \mathcal{N}(u_\ell; x_\ell), \qquad \gamma_\ell \in (0, 1), \tag{3}$$

where normalizing to the activation of the host stage, without any learnable parameter the network could exploit, isolates the strength of re-entry in γ_ℓ and makes it the same quantity at every depth (Appendix C). Retention and re-entry are thereby kept answerable to different questions—α_ℓ setting how long the trace persists, γ_ℓ how strongly what it yields is admitted back into the stage—so that a persistent trace may exert weak influence or a brief one, strong, and the two remain identifiable as long as no decay-dependent multiplier reaches the readout beside γ_ℓ, which is why α_ℓ enters through the gate alone, why the steps are averaged with fixed weights and why neither the write nor

Table 1: Clean top-1, mean corruption accuracy (mCA), mean PGD accuracy and learned memory parameters on Tiny ImageNet for iconic and baseline networks. PGD accuracy is averaged over $\epsilon \in \{0.25, 0.5, 1\}/255$. Accuracies and memory parameters are means over $n = 3$.

					α				γ			
Model	Variant	Clean Top-1	mCA	PGD_ϵ	S1	S2	S3	S4	S1	S2	S3	S4
ResNet-18	Iconic	$\mathbf{60.45 \pm 0.38}$	$\mathbf{25.46 \pm 0.65}$	$\mathbf{25.68 \pm 0.93}$	0.076	0.109	0.220	0.964	0.642	0.383	0.356	0.974
	Baseline	57.71 ± 0.40	21.33 ± 0.45	21.78 ± 0.37	–	–	–	–	–	–	–	–
ConvNeXt-Tiny	Iconic	$\mathbf{60.01 \pm 0.46}$	$\mathbf{33.42 \pm 0.45}$	$\mathbf{33.82 \pm 0.63}$	0.202	0.449	0.517	0.556	0.522	0.285	0.242	0.903
	Baseline	55.80 ± 0.66	30.35 ± 0.53	31.23 ± 0.70	–	–	–	–	–	–	–	–
CORnet-S	Iconic	$\mathbf{46.25 \pm 0.25}$	$\mathbf{26.89 \pm 0.23}$	$\mathbf{29.80 \pm 0.77}$	0.256	0.297	0.288	0.328	0.508	0.471	0.163	0.798
	Baseline	45.89 ± 0.73	25.03 ± 0.71	27.83 ± 0.57	–	–	–	–	–	–	–	–
DenseNet-121	Iconic	$\mathbf{58.00 \pm 0.04}$	$\mathbf{32.45 \pm 0.71}$	$\mathbf{24.17 \pm 0.34}$	0.218	0.368	0.456	0.552	0.752	0.336	0.270	0.630
	Baseline	57.46 ± 0.11	31.04 ± 0.23	19.91 ± 0.21	–	–	–	–	–	–	–	–
Swin-T	Iconic	$\mathbf{40.42 \pm 0.33}$	$\mathbf{23.22 \pm 0.16}$	$\mathbf{20.72 \pm 0.33}$	0.110	0.113	0.149	0.187	0.223	0.167	0.156	0.247
	Baseline	38.38 ± 0.37	21.67 ± 0.52	20.19 ± 0.49	–	–	–	–	–	–	–	–

the trace-driven term carries a learnable scale (Appendix B). Every stage receives the same block under the same initialization, and nothing regularizes or constrains how retention and re-entry order themselves across depth, so that whatever profile they come to trace is the network's own preference under the classification objective (Appendix A).

4 Results

Persistence Strength (α) Intra-stage: ***How Long Is a Fading Percept Useful for?*** Trained on natural images, the retention factor that governs how long the fading trace is useful for (Fig.) climbs through the hierarchy in every backbone we tested (Table 1). In ICONICResNet-18 the climb is steep, in ICONICConvNeXt, more gradual. In ICONICCORnet-S it is ever faint, and retention stays within a narrow band from V1 to IT (Fig. 6). Deeper stages hold their fading traces longer before those traces cease to influence access.

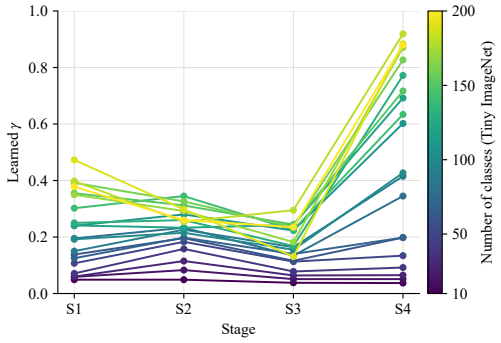


Figure 3: **Increasing task complexity drives informational persistence.**

The cortex keeps the same ordering. Intrinsic timescales, measured from resting single-neuron autocorrelation, lengthen from sensory toward association cortex in the primate (Ogawa & Komatsu, 2010; Murray et al., 2014; Cavanagh et al., 2016)—a gradient that holds under several timescale definitions at once and is conserved across mice, rats, macaques, and humans (Spitmaan et al., 2020; Song et al., 2024; Zeisler et al., 2025; Gao et al., 2020; Manea et al., 2022). Within the visual system, stimulus representations grow more temporally stable from V1 into the deeper areas of the rodent ventral stream (Piasini et al., 2021), and response decay timescales rise with anatomical hierarchy rank across mouse visual cortex (Siegle et al., 2021); in humans, the same gradient surfaces as a hierarchy of temporal receptive windows—sensory cortex tracking the moment, higher-order regions accumulating over sentences, paragraphs, and whole narrative events (Hasson et al., 2008; Lerner et al., 2011; Honey et al., 2012; Baldassano et al., 2017). The progression is as much about forgetting as about integration: when context switches, sensory cortex realigns within moments, while regions at the hierarchy's summit retain the superseded context for many seconds (Chien & Honey, 2020).

Circuit models of primate and human cortex derive this gradient from a single ingredient, a gradient of local recurrent excitatory strength across architecturally similar areas (Kiebel et al., 2008; Chaudhuri et al., 2015; Demirtaş et al., 2019; Li & Wang, 2022). What α acquires is an analogous gradient, arising across stages that start identically initialized but shaped by nothing beyond classi-

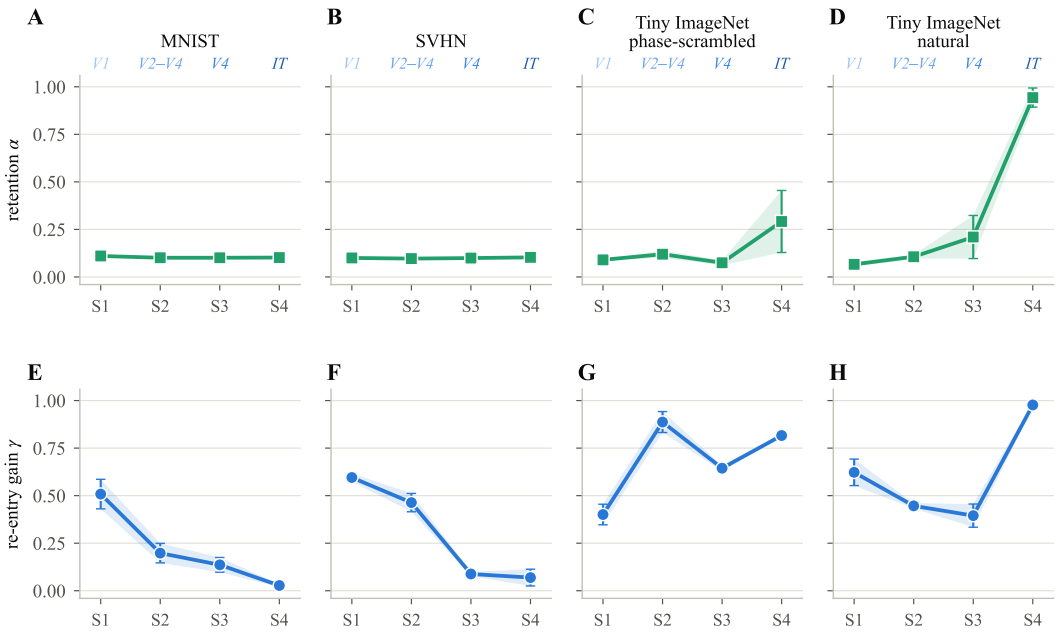


Figure 4: **Persistence is learned only from natural images.** Retention α (A–D) does not budge on impoverished datasets like MNIST or SVHN, varies at random on scrambled images and rises through the hierarchy on natural images. Re-entry gain γ (E–H) recovers its middle depression and terminal rebound only in the natural condition (ICONICResNet-18; $n = 5$).

fication loss—a convergence that computational accounts of neural timescales have begun to look for in task-performing networks (Zeraati et al., 2026) and that broader theories of the hierarchy predict of any system that is forced to integrate at depth what it could have merely registered at entry (Himberger et al., 2018; Wolff et al., 2022).

RE-ENTRY STRENGTH (γ) INTER-STAGE: ***Where Is a Fading Percept Most Needed?*** Across all tested quaternate models, γ magnitude that denotes total re-entry gain of ICONICBlocks follows $S_1 > S_2 > S_3 < S_4$ (Table 1). For 2D CNNs, it is not easy to clearly differentiate persistence without a magnitude axis, in contrast to a temporal one. Thus, we, with some putative interpretation, take this inter-stage magnitude strength as the reconfirmation of persistence, determined by how strongly the whole ICONICBlock is preferred there. The profile parallels the representational crossover that both neural networks and the human ventral stream exhibit between shape-dominant early processing and category-dominant late processing (Zeman et al., 2020)—and it does so in every architecture we test, from pure convolutional to transformer to an explicit ventral-stream model, despite none of them sharing a single temporal parameter at initialization. V1, on the shape-dominant side, already hosts strong recurrent contextual modulation—figure-ground enhancement driven by feedback from higher areas, appearing roughly 80–100 ms after onset and abolished by masking that interrupts re-entrant signals (Lamme, 1995; Zipser et al., 1996; Roelfsema et al., 2002). The intermediate stages sit at the crossing, where neither mode of processing places heavy demand on re-entrant amplification. At the summit, an altogether different recurrent regime takes over. Because behaviorally sufficient object-identity solutions in IT require additional time for challenging images, contingent on recurrent input from ventrolateral prefrontal cortex (Kar et al., 2019; Kar & DiCarlo, 2021), and PIT neurons increasingly encode object-level instead of textural structure (Wang & Ponce, 2026). The same non-monotonic pattern appears across all architectures, and its late rebound is affected by task complexity directly.

TASK DEMANDS DRIVE THE FINAL JUMP If IT's recurrent demand intensifies with task difficulty in primates (Kar et al., 2019; Spitzer & Richmond, 1991), and if we purportedly claim to have achieved the same demand structure, should we expect the S_3-to-S_4 jump to follow a similar trajectory? Yes—and across an ever-ranging axis of complexity.

With few training images per class, the U-shape is absent; as data grows, S_4 is the first stage to diverge (Fig. 8). With few classes, the rebound is modest; as classes accumulate, it widens on a graded dose-response curve (Fig. 7). Most telling of all, collapsing 200 fine-grained Tiny ImageNet classes into 20 coarse superclasses—same images, same architecture—nearly abolishes both hierarchies at once: α flattens rather than climbing, and the γ rebound at S_4 shrinks to a fraction of itself (Fig. 5). A network that needs only to tell animals from vehicles can afford to let its fading traces dissipate roughly uniformly; one that must tell a golden retriever from a Labrador cannot, and so it builds, from nothing beyond classification pressure, a depth-dependent persistence structure in which the last stage clings to its trace the way inferotemporal cortex clings to its recurrent input when the object is hard to resolve (Kar et al., 2019; Kar & DiCarlo, 2021).

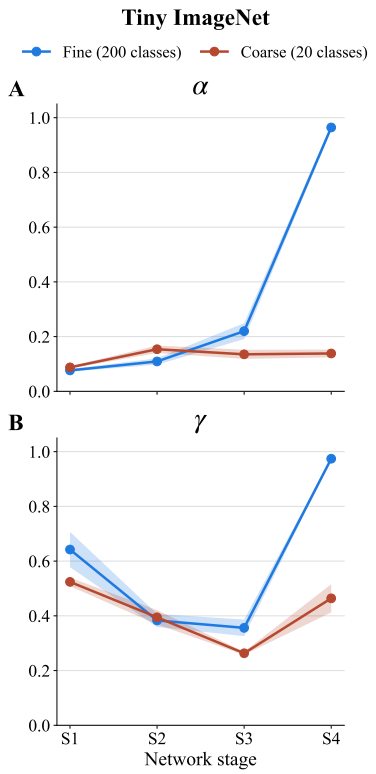


Figure 5: **Task complexity drives the terminal jump.** Learned α (A) and γ (B) at each stage of ICONICResNet-18 trained on the same Tiny ImageNet images labeled with either the 200 fine-grained classes or 20 coarse superclasses. Under coarse labels, α flattens and the γ rebound at S_4 shrinks.

Is there then a continuous thread of persistence fed from the initial sensory reservoir all the way up to the final stage, or is the S_3-to-S_4 jump insignificant—a mere computational artifact of capacity bottlenecked so far and finally given room? To test structural coherency, we phase-scrambled Tiny ImageNet, destroying edge structure, contour continuity, and Gestalt grouping cues while preserving the global spectral statistics, and trained the same architectures on the result (Fig. 4). The early elevation collapses. S_1, which on natural images had demanded the highest re-entry gain of the first three stages, loses its privileged position; the ordering across S_1–S_3 dissolves into inconsistency. The S_4 rebound survives, and it should—be it an edge or a noisy nuisance, whatever information has arrived at S_4 is all the network has, and it must still classify; the categorical imperative is unchanged, only the raw material differs. Together with the class-count manipulation, this constitutes a double dissociation: task complexity modulates the terminal jump while leaving the early stages untouched; phase scrambling modulates the early stages while leaving the terminal jump intact. The two ends of the U are driven by orthogonal pressures—one by the structure of the input, the other by the structure of the required output—and whatever S_4 holds on to is not a downstream echo of the sensory trace that S_1 was clinging to. What Coltheart (1980) split by behavioral dissociation into visible and informational persistence, the network splits by gradient pressure into two regimes that can be independently created and destroyed, using the same ether.

NON-ECOLOGICAL OR NON-COMPLEX DATASETS On MNIST and SVHN the temporal hierarchy is entirely absent (Fig. 4A–B, E–F). Retention stays at initialization, and re-entry gain declines monotonically toward zero at S_4—the very stage that on natural images had demanded the strongest trace now wants the least. Stimuli that lack the compositional richness of natural scenes—the part-whole relationships, the occlusion, the within-category variability that the ventral stream evolved to resolve—offer no gradient pressure for the kind of hierarchical persistence that ecological vision demands.

VONENET FRONT-END SUBSUMES S_1 DEMAND The fixed Gabor-filter frontend of Dapello et al. (2020), parameterized to match empirical distributions of primate V1 simple and complex cells, provides a direct test of what drives the early elevation (Fig. 6, purple). With orientation-selective filtering already in place—the very computation V1 neurons are classically understood to perform (Hubel & Wiesel, 1962; Jones & Palmer, 1987)—S_1 is relieved of forming these filters on its own, and its re-entry gain drops accordingly, leaving S_2, itself unchanged, as the early peak. The rest of the hierarchy is unaffected — the intermediate trough holds and the terminal rebound at S_4

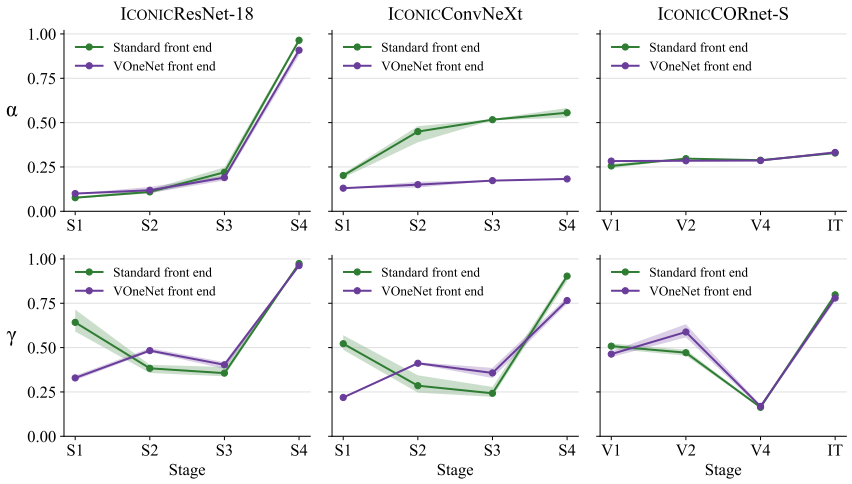


Figure 6: Learned values of the iconic memory parameters α (top) and γ (bottom) at each processing stage of three ICONICNet variants, trained with either a standard front end or a VOneNet one, where lines give the mean across seeds and shaded bands their spread. VOne front-end subsumes S_1.

stays where it was, consistent with the evidence that IT's recurrent demand is about object identity (Kar et al., 2019), not edge extraction, and cannot be satisfied by any amount of V1-like filtering.

WHAT IS WORTH PERSISTING CHANGES Depth pushes neural representations from local sensory features toward increasingly object and category-organized information, and the Gestalt-like effects that networks do acquire appear disproportionately late. The late resurgence of persistence coincides with the stage at which initially local contour and shape correspondences have been integrated into a more global object representation, paralleling the ventral stream progression from local edge coding in early visual cortex, through contour component coding in V4, toward increasingly whole object representations in IT.

5 DISCUSSION

Yet, sensory persistence does not end with vision. One of the most surprising results was ICONICNet preserving the γ hierarchy across audio event classification on FSD50K (Fonseca et al., 2022) (Appendix G). Due to the scope and nature of which, we denote ECHOICNet and leave for future work to dwell on. In line with these results, that the models converge on a representation favors fading sensory memory strongly at its earliest, putatively V1-correspondent stage—the stage where such transitory persistence, while spanning with no lesser reach the rest of the cortical hierarchy, has been most directly tied to iconic memory (Nikolić et al., 2009; Teeuwen et al., 2021; Yang et al., 2024). That they independently elevate persistence at the final stage, where classification pressure steers the network towards a regime of category-selective sustenance in which informational structure has accrued enough prominence as to become pivotal in shaping the network—separating its persistence from the visible iconic counterpart and into the formally distinct informational component (Coltheart, 1980; Loftus & Irwin, 1998). That the intra-stage persistence lets the network temporally recover the Murray et al. (2014)'s cortical time-scale hierarchy where deeper layers dwell longer. That none of these were a priori imposed, yet all emerged, we argue demonstrates how biological and artificial systems can become more aligned not only through shared architectural enrichment but also through the *universality* of what they are denied. A universality that may be traced to the increasingly prominent pursuit of understanding how disparate systems arrive at similar internal models of the world (Huh et al., 2024; Sucholutsky et al., 2025). With one closely corroborative recent framework (Bosch et al., 2026), suggesting on one hand of the spectrum, that such convergences may as much be shaped by *shared limitations* as they are by *shared capacities*—with the common limitation assigned here being *the impermanence*.

AI USE STATEMENT

In this work, we used generative AI tools for code debugging and checking the veracity of the claims made to see if they carry any tangible value against the broader literature. We have not used generative AI tools for writing/composition, decision-making or any paper-altering addition and the remaining required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work performs secondary computational analyses of previously collected human neuroimaging data from Cichy et al. (2016). No new human participants were recruited and no new human data were collected for this study. The original study reports that all participants provided written informed consent, that data collection was approved by the Institutional Review Board of the Massachusetts Institute of Technology and that the study was conducted in accordance with the Declaration of Helsinki.

REPRODUCIBILITY STATEMENT

The ICONICBlock is specified in full in Appendix A (Eqs. 4–8), with the initialization and regularization of every block parameter listed in Table 2, the divisive gain and its bounds given in Appendix C, the backbones, datasets, evaluation protocol and VOneNet front ends in Appendix D and the representational similarity analysis described in Appendix F.

REFERENCES

Athena Akrami, Yan Liu, Alessandro Treves, and Bharathi Jagadeesh. Converging neuronal activity in inferior temporal cortex during the classification of morphed stimuli. *Cerebral Cortex*, 19(4): 760–776, 2009. doi: 10.1093/cercor/bhn125.

Elkan G. Akyürek. Temporal integration as an adaptive process in visual perception, attention, and working memory. *Neuroscience & Biobehavioral Reviews*, 170:106041, 2025. doi: 10.1016/j.neubiorev.2025.106041.

E. Averbach and A. S. Coriell. Short-term memory in vision. *Bell System Technical Journal*, 40(1): 309–328, 1961. doi: 10.1002/j.1538-7305.1961.tb03987.x.

Thomas Bachlechner, Bodhisattwa Prasad Majumder, Henry Mao, Gary Cottrell, and Julian McAuley. ReZero is all you need: Fast convergence at large depth. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 1352–1361, 2021.

Jongsoo Baek, Luis Andres Lesmes, and Zhong-Lin Lu. qPR: An adaptive partial-report procedure based on bayesian inference. *Journal of Vision*, 16(10):25, 2016. doi: 10.1167/16.10.25.

Christopher Baldassano, Janice Chen, Asieh Zadbood, Jonathan W. Pillow, Uri Hasson, and Kenneth A. Norman. Discovering event structure in continuous narrative perception and memory. *Neuron*, 95(3):709–721.e5, 2017. doi: 10.1016/j.neuron.2017.06.041.

Andrea Benucci, Dario L. Ringach, and Matteo Carandini. Coding of stimulus sequences by population responses in visual cortex. *Nature Neuroscience*, 12(10):1317–1324, 2009. doi: 10.1038/nn.2398.

Victoria Bosch, Rowan P. Sommers, Adrien Doerig, and Tim C. Kietzmann. The umwelt representation hypothesis: Rethinking universality. *Trends in Cognitive Sciences*, 2026. doi: 10.1016/j.tics.2026.07.004. in press.

Stefania Bracci and Hans Op de Beeck. Dissociations and associations between shape and category representations in the two visual pathways. *The Journal of Neuroscience*, 36(2):432–444, 2016. doi: 10.1523/JNEUROSCI.2314-15.2016.

Jean Bullier. Integrated model of visual processing. *Brain Research Reviews*, 36(2–3):96–107, 2001. doi: 10.1016/S0165-0173(01)00085-6.

Charles F. Cadieu, Ha Hong, Daniel L. K. Yamins, Nicolas Pinto, Diego Ardila, Ethan A. Solomon, Najib J. Majaj, and James J. DiCarlo. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLOS Computational Biology*, 10(12):e1003963, 2014. doi: 10.1371/journal.pcbi.1003963.

Matteo Carandini and David J. Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62, 2012.

Sean E. Cavanagh, Joni D. Wallis, Steven W. Kennerley, and Laurence T. Hunt. Autocorrelation structure at rest predicts value correlates of single neurons during reward-guided choice. *eLife*, 5: e18937, 2016. doi: 10.7554/eLife.18937.

Rishidev Chaudhuri, Kenneth Knoblauch, Marie-Alice Gariel, Henry Kennedy, and Xiao-Jing Wang. A large-scale circuit mechanism for hierarchical dynamical processing in the primate cortex. *Neuron*, 88(2):419–431, 2015. doi: 10.1016/j.neuron.2015.09.008.

Hsiang-Yun Sherry Chien and Christopher J. Honey. Constructing and forgetting temporal context in the human cerebral cortex. *Neuron*, 106(4):675–686.e11, 2020. doi: 10.1016/j.neuron.2020.02.013.

Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6:27755, 2016. doi: 10.1038/srep27755.

Radoslaw Martin Cichy, Gemma Roig, Alex Andonian, Kshitij Dwivedi, Benjamin Lahner, Alex Lascelles, Yalda Mohsenzadeh, Kandan Ramakrishnan, and Aude Oliva. The algonauts project: A platform for communication between the sciences of biological and artificial intelligence. In *Conference on Cognitive Computational Neuroscience*, 2019. URL https://arxiv.org/abs/1905.05675.

Max Coltheart. Iconic memory and visible persistence. *Perception & Psychophysics*, 27(3):183–228, 1980.

Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David Cox, and James J. DiCarlo. Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13073–13087, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/98b17f068d5d9b7668e19fb8ae470841-Abstract.html.

Murat Demirtaş, Joshua B. Burt, Markus Helmer, Jie Lisa Ji, Brendan D. Adkinson, Matthew F. Glasser, David C. Van Essen, Stamatios N. Sotiropoulos, Alan Anticevic, and John D. Murray. Hierarchical heterogeneity across human cortex shapes large-scale neural dynamics. *Neuron*, 101(6):1181–1194.e13, 2019. doi: 10.1016/j.neuron.2019.01.017.

Vincent Di Lollo. Temporal integration in visual memory. *Journal of Experimental Psychology: General*, 109:75–97, 1980.

James J. DiCarlo, Davide Zoccolan, and Nicole C. Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012. doi: 10.1016/j.neuron.2012.01.010.

Yadin Dudai. The neurobiology of consolidations, or, how stable is the engram? *Annual Review of Psychology*, 55:51–86, 2004. doi: 10.1146/annurev.psych.55.090902.142050.

Jacques Duysens, Guy A. Orban, Jacques Crémieux, and Hugo Maes. Visual cortical correlates of visible persistence. *Vision Research*, 25(2):171–178, 1985. doi: 10.1016/0042-6989(85)90110-5.

Howard Eichenbaum. A cortical–hippocampal system for declarative memory. *Nature Reviews Neuroscience*, 1(1):41–50, 2000.

Charles W. Eriksen and James F. Collins. Some temporal characteristics of visual pattern perception. *Journal of Experimental Psychology*, 74(3):476–484, 1967. doi: 10.1037/h0024765.

Johannes J. Fahrenfort, H. Steven Scholte, and Victor A. F. Lamme. Masking disrupts reentrant processing in human visual cortex. *Journal of Cognitive Neuroscience*, 19(9):1488–1497, 2007. doi: 10.1162/jocn.2007.19.9.1488.

Daniel J. Felleman and David C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1):1–47, 1991. doi: 10.1093/cercor/1.1.1.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. FSD50K: An open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2022.

Paul W. Frankland and Bruno Bontempi. The organization of recent and remote memories. *Nature Reviews Neuroscience*, 6(2):119–130, 2005. doi: 10.1038/nrn1607.

Richard Gao, Ruud L. van den Brink, Thomas Pfeffer, and Bradley Voytek. Neuronal timescales are functionally dynamic and shaped by cortical microarchitecture. *eLife*, 9:e61277, 2020. doi: 10.7554/eLife.61277.

Karl R. Gegenfurtner and George Sperling. Information transfer in iconic memory experiments. *Journal of Experimental Psychology: Human Perception and Performance*, 19(4):845–866, 1993. doi: 10.1037/0096-1523.19.4.845.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

Umut Güçlü and Marcel A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27): 10005–10014, 2015.

Uri Hasson, Eunice Yang, Ignacio Vallines, David J. Heeger, and Nava Rubin. A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience*, 28(10):2539–2550, 2008.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

David J. Heeger and Klavdia O. Zemlianova. A recurrent circuit implements normalization, simulating the dynamics of V1 activity. *Proceedings of the National Academy of Sciences*, 117(36): 22494–22505, 2020. doi: 10.1073/pnas.2005417117.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Kevin D. Himberger, Hsiang-Yun Chien, and Christopher J. Honey. Principles of temporal processing across the cortical hierarchy. *Neuroscience*, 389:161–174, 2018. doi: 10.1016/j.neuroscience.2018.04.030.

Christopher J. Honey, Thomas Thesen, Tobias H. Donner, Lauren J. Silbert, Chad E. Carlson, Orrin Devinsky, Werner K. Doyle, Nava Rubin, David J. Heeger, and Uri Hasson. Slow cortical dynamics and the accumulation of information over long timescales. *Neuron*, 76(2):423–434, 2012. doi: 10.1016/j.neuron.2012.08.011.

David H. Hubel and Torsten N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160(1):106–154, 1962. doi: 10.1113/jphysiol.1962.sp006837.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Judson P. Jones and Larry A. Palmer. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1233–1258, 1987. doi: 10.1152/jn.1987.58.6.1233.

Kohitij Kar and James J. DiCarlo. Fast recurrent processing via ventrolateral prefrontal cortex is needed by the primate ventral stream for robust core visual object recognition. *Neuron*, 109(1): 164–176.e5, 2021. doi: 10.1016/j.neuron.2020.09.035.

Kohitij Kar, Jonas Kubilius, Kailyn Schmidt, Elias B. Issa, and James J. DiCarlo. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, 22(6):974–983, 2019. doi: 10.1038/s41593-019-0392-5.

Christian Keysers, Da-Ke Xiao, Peter Földiák, and David I. Perrett. Out of sight but not out of mind: The neurophysiology of iconic memory in the superior temporal sulcus. *Cognitive Neuropsychology*, 22(3–4):316–332, 2005. doi: 10.1080/02643290442000103.

Roozbeh Kiani, Hossein Esteky, Koorosh Mirpour, and Keiji Tanaka. Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology*, 97(6):4296–4309, 2007. doi: 10.1152/jn.00024.2007.

Stefan J. Kiebel, Jean Daunizeau, and Karl J. Friston. A hierarchy of time-scales and the brain. *PLoS Computational Biology*, 4(11):e1000209, 2008.

Tim C. Kietzmann, Courtney J. Spoerer, Lynn K. A. Sörensen, Radoslaw M. Cichy, Olaf Hauk, and Nikolaus Kriegeskorte. Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, 116(43):21854–21863, 2019. doi: 10.1073/pnas.1905544116.

Gyula Kovács, Rufin Vogels, and Guy A. Orban. Cortical correlate of pattern backward masking. *Proceedings of the National Academy of Sciences of the United States of America*, 92(12):5587–5591, 1995. doi: 10.1073/pnas.92.12.5587.

Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib J. Majaj, Elias B. Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Aran Nayebi, Daniel Bear, Daniel L. K. Yamins, and James J. DiCarlo. Brain-like object recognition with high-performing shallow recurrent ANNs. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Victor A. F. Lamme. The neurophysiology of figure–ground segregation in primary visual cortex. *Journal of Neuroscience*, 15(2):1605–1615, 1995. doi: 10.1523/JNEUROSCI.15-02-01605.1995.

Victor A. F. Lamme and Pieter R. Roelfsema. The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, 23(11):571–579, 2000. doi: 10.1016/S0166-2236(00)01657-X.

Victor A. F. Lamme, Hans Supèr, and Henk Spekreijse. Feedforward, horizontal, and feedback processing in the visual cortex. *Current Opinion in Neurobiology*, 8(4):529–535, 1998. doi: 10.1016/S0959-4388(98)80042-1.

Victor A. F. Lamme, Karl Zipser, and Henk Spekreijse. Masking interrupts figure–ground signals in V1. *Journal of Cognitive Neuroscience*, 14(7):1044–1053, 2002. doi: 10.1162/089892902320474490.

Ya Le and Xuan Yang. Tiny ImageNet visual recognition challenge. Technical report, Stanford University, 2015.

Yulia Lerner, Christopher J. Honey, Lauren J. Silbert, and Uri Hasson. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, 31(8): 2906–2915, 2011.

Songting Li and Xiao-Jing Wang. Hierarchical timescales in the neocortex: Mathematical mechanism and biological insights. *Proceedings of the National Academy of Sciences*, 119(6): e2110274119, 2022. doi: 10.1073/pnas.2110274119.

Grace W. Lindsay. Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, 33(10):2017–2031, 2021. doi: 10.1162/jocn_a_01544.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021. doi: 10.1109/ICCV48922.2021.00986.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022. doi: 10.1109/CVPR52688.2022.01167.

Geoffrey R. Loftus and David E. Irwin. On the relations among different measures of visible and informational persistence. *Cognitive Psychology*, 35:135–199, 1998.

Zhong-Lin Lu, James Neuse, Stephen Madigan, and Barbara Anne Dosher. Fast decay of iconic memory in observers with mild cognitive impairments. *Proceedings of the National Academy of Sciences*, 102(5):1797–1802, 2005. doi: 10.1073/pnas.0408402102.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Ana M. G. Manea, Anna Zilverstand, Kamil Ugurbil, Sarah R. Heilbronner, and Jan Zimmermann. Intrinsic timescales as an organizational principle of neural processing across the whole rhesus macaque brain. *eLife*, 11:e75540, 2022. doi: 10.7554/eLife.75540.

Mackenzie Weygandt Mathis, Adriana Perez Rotondo, Edward F. Chang, Andreas S. Tolias, and Alexander Mathis. Decoding the brain: From neural representations to mechanistic models. *Cell*, 187(21):5814–5832, 2024. doi: 10.1016/j.cell.2024.08.051.

Karla Matic, Issam Tafech, Peter König, and John-Dylan Haynes. Temporal dynamics and readout latency in perception and iconic memory. *Journal of Vision*, 25(13):3, 2025. doi: 10.1167/jov.25.13.3.

Johannes Mehrer, Courtney J. Spoerer, Emer C. Jones, Nikolaus Kriegeskorte, and Tim C. Kietzmann. An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, 118(8):e2011417118, 2021. doi: 10.1073/pnas.2011417118.

John D. Murray, Alberto Bernacchia, David J. Freedman, Ranulfo Romo, Jonathan D. Wallis, Xinying Cai, Camillo Padoa-Schioppa, Tatiana Pasternak, Hyojung Seo, Daeyeol Lee, and Xiao-Jing Wang. A hierarchy of intrinsic timescales across primate cortex. *Nature Neuroscience*, 17(12):1661–1663, 2014. doi: 10.1038/nn.3862.

Aran Nayebi, Daniel Bear, Jonas Kubilius, Kohitij Kar, Surya Ganguli, David Sussillo, James J. DiCarlo, and Daniel L. K. Yamins. Task-driven convolutional recurrent models of the visual system. In *Advances in Neural Information Processing Systems*, volume 31, pp. 5290–5301, 2018.

Ulric Neisser. *Cognitive Psychology*. Appleton-Century-Crofts, New York, 1967.

Danko Nikolić, Stefan Häusler, Wolf Singer, and Wolfgang Maass. Distributed fading memory for stimulus properties in the primary visual cortex. *PLOS Biology*, 7(12):e1000260, 2009. doi: 10.1371/journal.pbio.1000260.

Tadashi Ogawa and Hidehiko Komatsu. Differential temporal storage capacity in the baseline activity of neurons in macaque frontal eye field and area V4. *Journal of Neurophysiology*, 103(5):2433–2445, 2010. doi: 10.1152/jn.01066.2009.

Eugenio Piasini, Liviu Soltuzu, Paolo Muratore, Riccardo Caramellino, Kasper Vinken, Hans Op de Beeck, Vijay Balasubramanian, and Davide Zoccolan. Temporal stability of stimulus representation increases along rodent visual cortical hierarchies. *Nature Communications*, 12:4448, 2021. doi: 10.1038/s41467-021-24456-3.

Michael S. Pratte. Iconic memories die a sudden death. *Psychological Science*, 29(6):877–887, 2018. doi: 10.1177/0956797617747118.

Ronald A. Rensink. Limits to the usability of iconic memory. *Frontiers in Psychology*, 5:971, 2014. doi: 10.3389/fpsyg.2014.00971.

Pieter R. Roelfsema, Victor A. F. Lamme, Henk Spekreijse, and Holger Bosch. Figure–ground segregation in a recurrent network architecture. *Journal of Cognitive Neuroscience*, 14(4):525–537, 2002. doi: 10.1162/08989290260045756.

Edmund T. Rolls and Martin J. Tovee. Processing speed in the cerebral cortex and the neurophysiology of visual masking. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 257(1348):9–15, 1994. doi: 10.1098/rspb.1994.0087.

Tim Salimans and Diederik P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in Neural Information Processing Systems*, volume 29, pp. 901–909, 2016.

Matthew T. Schmolesky, Youngchang Wang, Doug P. Hanes, Kirk G. Thompson, Stefan Leutgeb, Jeffrey D. Schall, and Audie G. Leventhal. Signal timing across the macaque visual system. *Journal of Neurophysiology*, 79(6):3272–3278, 1998. doi: 10.1152/jn.1998.79.6.3272.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? Technical report, bioRxiv, 2018.

Joshua H. Siegle, Xiaoxuan Jia, Séverine Durand, Sam Gale, Corbett Bennett, et al. Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature*, 592(7852):86–92, 2021. doi: 10.1038/s41586-020-03171-x.

Paul Justin Connor Smith and Niko A. Busch. Pre-stimulus pupil-linked arousal enhances initial stimulus availability and accelerates decay in iconic memory. *Attention, Perception, & Psychophysics*, 88:182, 2026. doi: 10.3758/s13414-026-03326-4.

Min Song, Eun Ju Shin, Hyojung Seo, Alireza Soltani, Nicholas A. Steinmetz, Daeyeol Lee, Min Whan Jung, and Se-Bum Paik. Hierarchical gradients of multiple timescales in the mammalian forebrain. *Proceedings of the National Academy of Sciences*, 121(51):e2415695121, 2024. doi: 10.1073/pnas.2415695121.

George Sperling. The information available in brief visual presentations. *Psychological Monographs: General and Applied*, 74(11):1–29, 1960.

Mehran Spitmaan, Hyojung Seo, Daeyeol Lee, and Alireza Soltani. Multiple timescales of neural dynamics and integration of task-relevant signals across cortex. *Proceedings of the National Academy of Sciences*, 117(36):22522–22531, 2020. doi: 10.1073/pnas.2005993117.

Hedva Spitzer and Barry J. Richmond. Task difficulty: Ignoring, attending to, and discriminating a visual stimulus yield progressively more activity in inferior temporal neurons. *Experimental Brain Research*, 83(2):340–348, 1991. doi: 10.1007/BF00231157.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Training very deep networks. In *Advances in Neural Information Processing Systems*, volume 28, pp. 2377–2385, 2015.

Katherine R. Storrs, Tim C. Kietzmann, Alexander Walther, Johannes Mehrer, and Nikolaus Kriegeskorte. Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of Cognitive Neuroscience*, 33(10):2044–2064, 2021. doi: 10.1162/jocn_a_01755.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Christopher J. Cueva, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nathan Cloos, Nikolaus Kriegeskorte, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert

Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew K. Lampinen, Klaus-Robert Müller, Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment. *Transactions on Machine Learning Research*, October 2025. URL https://openreview.net/forum?id=Hiq7lUh4Yn.

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114. PMLR, 2019.

Rob R. M. Teeuwen, Catherine Wacongne, Ulf H. Schnabel, Matthew W. Self, and Pieter R. Roelfsema. A neuronal basis of iconic memory in macaque primary visual cortex. *Current Biology*, 31 (24):5401–5414.e4, 2021. doi: 10.1016/j.cub.2021.09.052.

Ivan Tomić and Paul M. Bays. A dynamic neural resource model bridges sensory and working memory. *eLife*, 12:RP91034, 2024. doi: 10.7554/eLife.91034.3.

Endel Tulving. Memory and consciousness. *Canadian Psychology*, 26(1):1–12, 1985.

Twan van Laarhoven. L2 regularization versus batch and weight normalization. *arXiv preprint arXiv:1706.05350*, 2017.

Johan Wagemans, James H. Elder, Michael Kubovy, Stephen E. Palmer, Mary A. Peterson, Manish Singh, and Rüdiger von der Heydt. A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychological Bulletin*, 138(6):1172–1217, 2012. doi: 10.1037/a0029333.

Binxu Wang and Carlos R. Ponce. Neuronal tuning aligns dynamically with object and texture manifolds across the visual hierarchy. *Nature Neuroscience*, 29:864–875, 2026. doi: 10.1038/s41593-026-02207-1.

Annemarie Wolff, Nareg Berberian, Mehrshad Golesorkhi, Javier Gomez-Pilar, Federico Zilio, and Georg Northoff. Intrinsic neural timescales: Temporal integration and segregation. *Trends in Cognitive Sciences*, 26(2):159–173, 2022. doi: 10.1016/j.tics.2021.11.007.

Daniel L. K. Yamins and James J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, 2016.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models explain the neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

Yiling Yang, Johanna Klon-Lipok, Katharine Shapcott, Andreea Lazar, and Wolf Singer. Dynamic fading memory and expectancy effects in the monkey primary visual cortex. *Proceedings of the National Academy of Sciences*, 121(8):e2314855121, 2024. doi: 10.1073/pnas.2314855121.

Polina Zamarashkina, Dina V. Popovkina, and Anitha Pasupathy. Timing of response onset and offset in macaque V4: stimulus and task dependence. *Journal of Neurophysiology*, 123(6):2311–2325, 2020. doi: 10.1152/jn.00586.2019.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, volume 8689 of *Lecture Notes in Computer Science*, pp. 818–833. Springer, 2014. doi: 10.1007/978-3-319-10590-1_53.

Zachary R. Zeisler, Marques Love, Ueli Rutishauser, Frederic M. Stoll, and Peter H. Rudebeck. Consistent hierarchies of single-neuron timescales in mice, macaques, and humans. *Journal of Neuroscience*, 45(19):e2155242025, 2025. doi: 10.1523/JNEUROSCI.2155-24.2025.

Astrid A. Zeman, J. Brendan Ritchie, Stefania Bracci, and Hans Op de Beeck. Orthogonal representations of object shape and category in deep convolutional neural networks and human visual cortex. *Scientific Reports*, 10:2453, 2020. doi: 10.1038/s41598-020-59175-0.

Roxana Zeraati, Anna Levina, Jakob H. Macke, and Richard Gao. Neural timescales from a computational perspective. *Nature Neuroscience*, 29(7):1534–1547, 2026. doi: 10.1038/s41593-026-02343-8.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision Mamba: Efficient visual representation learning with bidirectional state space model. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62429–62442. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhu24f.html.

Karl Zipser, Victor A. F. Lamme, and Peter H. Schiller. Contextual modulation in primary visual cortex. *Journal of Neuroscience*, 16(22):7376–7389, 1996. doi: 10.1523/JNEUROSCI.16-22-07376.1996.

Ariel Zylberberg, Stanislas Dehaene, Gabriel B. Mindlin, and Mariano Sigman. Neurophysiological bases of exponential sensory decay and top-down memory retrieval: A model. *Frontiers in Computational Neuroscience*, 3:4, 2009. doi: 10.3389/neuro.10.004.2009.

A THE ICONICBLOCK IN FULL

PLACEMENT One block follows each of the four stages of the backbone, its output taking the place of the stage output that the next stage receives: in ResNet-18, after each of its four residual stages (64, 128, 256 and 512 channels), and in ConvNeXt-Tiny, after each of its four stages (96, 192, 384 and 768 channels).

COMPUTATION Written out in full, the block computes

$$m_\ell = \phi(\mathrm{GN}^\circ(W_\ell^w * x_\ell)), \qquad m_{\ell,k} = \alpha_\ell^k \, m_\ell, \tag{4}$$

$$q_\ell = \phi(\mathrm{GN}(W_\ell^p * m_\ell)), \tag{5}$$

$$z_{\ell,k} = \sigma(W_\ell^x * x_\ell + b_\ell + W_\ell^m * m_{\ell,k}), \qquad W_{\ell,\mathrm{c}}^m = D_{\ell,\mathrm{c}}/\|D_{\ell,\mathrm{c}}\|_2, \tag{6}$$

$$u_\ell = \frac{1}{K}\sum_{k=0}^{K-1} z_{\ell,k} \odot q_\ell, \tag{7}$$

$$\hat{x}_\ell = x_\ell + \gamma_\ell \, \rho_\ell \, u_\ell, \qquad \alpha_\ell = \sigma(a_\ell), \qquad \gamma_\ell = \gamma_{\max}\, \sigma(g_\ell), \tag{8}$$

where W_ℓ^w, W_ℓ^p and W_ℓ^x are 1×1 convolutions, of which only W_ℓ^x carries a bias b_ℓ; W_ℓ^m is a bias-free 1×1 convolution whose output rows are held at unit norm through a learned direction D_ℓ; ϕ is the GELU nonlinearity; GN° is a group normalization without learnable scale or shift and GN one with them, both using 32 groups at every stage of either backbone; and ρ_ℓ is the divisive gain of Appendix C. We use $K = 4$ and $\gamma_{\max} = 1$ throughout. Eqs. 4–8 reduce to Eqs. 1–3 of the main text with $\mathcal{N}(u_\ell; x_\ell) = \rho_\ell u_\ell$.

WHAT KEEPS α READABLE Three choices ensure that α_ℓ alone sets the temporal course of the trace's hold upon access. The write normalization carries no affine parameters, so that the trace is written at a fixed scale and no learnable rescaling of m_ℓ can counterfeit a slower or swifter decay. The rows of W_ℓ^m follow the weight normalization of Salimans & Kingma (2016) with its length frozen at unity: a reparameterization that ordinarily decouples the length of a weight vector from its direction here removes that length altogether, leaving α_ℓ^k as the only learnable magnitude upon the trace-driven term of the gate. The steps, finally, are averaged with fixed weights $1/K$, and no representation of k is supplied, so that no temporal preference enters the readout besides the one α_ℓ expresses. The allocation and legibility terms combine additively beneath the sigmoid, a soft conjunction standing in for the product by which Gegenfurtner & Sperling (1993) composed their transfer rate. Because the content normalization is affine and follows W_ℓ^p directly, it removes the scale of W_ℓ^p, whose initialization accordingly shapes the pattern of q_ℓ and leaves its magnitude to the normalization.

INITIALIZATION AND REGULARIZATION Table 2 lists the initialization of every block parameter, identical at every stage, together with whether it is subject to the backbone's weight decay.

Table 2: Initialization and regularization of ICONICBlock parameters, identical at every stage. The logits are initialized so that $\alpha_\ell^{(0)} = 0.1$ and $\gamma_\ell^{(0)} = 0.1$.

Parameter	Role	Initialization	Weight decay
W_ℓ^w	write	Kaiming normal (fan-out)	yes
W_ℓ^p	content projection	$\mathcal{N}(0,\, 0.01^2)$	yes
GN scale, shift	content normalization	$1,\, 0$	yes
W_ℓ^x	allocation	0	yes
b_ℓ	gate bias	-2	no
D_ℓ	trace direction	$\mathcal{N}(0, 1)$	no
a_ℓ	retention logit	$\mathrm{logit}(0.1)$	no
g_ℓ	re-entry logit	$\mathrm{logit}(0.1/\gamma_{\max})$	no

Weight decay is withheld from a_ℓ, g_ℓ, b_ℓ and D_ℓ. On the first three it would draw every stage toward $\alpha_\ell = \frac{1}{2}$, $\gamma_\ell = \frac{1}{2}\gamma_{\max}$ and half-open gates, a prior indistinguishable in its effect from a learned preference, while on the directions D_ℓ, whose effective weight is invariant to their scale, it would exert no regularizing effect at all, altering only the effective learning rate (van Laarhoven, 2017). The allocation weights begin at zero and the gate bias at -2, following the practice of

biasing gates toward closure so that a gated block begins near the identity and opens only where training finds reason to (Srivastava et al., 2015); access therefore begins near $\sigma(-2) \approx 0.12$ and varies, at first, with the trace alone. The initial re-entry gain $\gamma_\ell^{(0)} = 0.1$ serves the same end, in the manner of residual branches initialized at or near zero (Bachlechner et al., 2021; Liu et al., 2022): by Appendix C, the initial injection remains below $0.1\,\|x_\ell\|_2$ wherever $r_\ell < 100$, so that the block enters a backbone it at first scarcely disturbs. With $\alpha_\ell^{(0)} = 0.1$ at every stage, the trace reaches the later steps only faintly at the outset, at a tenth of its written magnitude by the second step and a hundredth by the third, and whatever stage-wise persistence the network comes to exhibit is acquired in training.

B Why Retention Is Confined to the Gate

Consider the more literal rendering of a decaying trace, in which each step of the readout is additionally weighted by its surviving magnitude, all else being as in Eqs. 4–8:

$$\tilde{u}_\ell = \frac{1}{K} \sum_{k=0}^{K-1} \alpha_\ell^k \, z_{\ell,k} \odot q_\ell. \tag{9}$$

Geometric Reduction Wherever access varies little from step to step, $z_{\ell,k} \approx z_\ell$—and exactly so for the plain decaying readout, in which no gate intervenes and $z_{\ell,k} \equiv 1$—the weights form a geometric series and

$$\tilde{u}_\ell \approx c(\alpha_\ell)\, z_\ell \odot q_\ell, \qquad c(\alpha) = \frac{1}{K} \sum_{k=0}^{K-1} \alpha^k = \frac{1-\alpha^K}{K(1-\alpha)}, \tag{10}$$

a static rescaling of the gated content, increasing monotonically on $(0,1)$ from $1/K$ as $\alpha \to 0$ to 1 as $\alpha \to 1$. The step-wise structure of the readout then conveys nothing that a single rescaling could not, and the rescaling is of precisely the kind that γ_ℓ exercises.

Gradient Coupling Let L denote the training objective, $\delta_\ell = \partial L / \partial \hat{x}_\ell$ and $v_\ell = z_\ell \odot q_\ell$. Because ρ_ℓ is computed under a stop-gradient, it enters the backward pass as a constant, and with $G_\ell = \sum_n \rho_{\ell,n} \langle \delta_{\ell,n}, v_{\ell,n} \rangle$ summed over the samples n of a batch, Eqs. 8 and 10 give

$$\frac{\partial L}{\partial \gamma_\ell} = c(\alpha_\ell)\, G_\ell, \qquad \frac{\partial L}{\partial \alpha_\ell} = \gamma_\ell\, c'(\alpha_\ell)\, G_\ell, \qquad \text{whence} \qquad \frac{\partial L}{\partial \alpha_\ell} = \frac{\gamma_\ell\, c'(\alpha_\ell)}{c(\alpha_\ell)} \frac{\partial L}{\partial \gamma_\ell}. \tag{11}$$

The factor $\gamma_\ell c'(\alpha_\ell)/c(\alpha_\ell)$ is positive on $(0,1)$, as are the Jacobians $\alpha_\ell(1-\alpha_\ell)$ and $\gamma_\ell(1-\gamma_\ell/\gamma_{\max})$ of the sigmoid parameterizations, so the logits a_ℓ and g_ℓ receive gradients of the same sign at every step: with the gain held at its detached value, the objective depends upon the two only through the product $\gamma_\ell\, c(\alpha_\ell)$, is flat along its level sets, and cannot tell one parameter from the other. Under SGD with momentum or Adam, whose updates follow a running average of the gradient, the two logits consequently move in the same direction for as long as their positive multipliers vary slowly relative to the averaging horizon, and from identical initial values at every stage their profiles across depth are bound to share an ordering. This is the lockstep we observed under the coupled readout, in which α and γ traced one and the same U-shaped profile across depth.

Why a Gate Passive decay leaves the trace at every step a scaled copy of the one written, so that the only property of the trace the retention factor can alter is its magnitude. Let the trace reach the readout through any map f positively homogeneous of degree $p \geq 0$ in it, $f(\lambda m) = \lambda^p f(m)$ for $\lambda > 0$, a class that contains bias-free convolutions, ReLUs and their compositions ($p = 1$) as well as normalizations ($p = 0$). Any readout $\sum_k w_k f(m_{\ell,k})$ whose weights w_k do not depend on α_ℓ then reduces exactly to $C_p(\alpha_\ell)\, f(m_\ell)$ with $C_p(\alpha) = \sum_k w_k \alpha^{kp}$, of which Eq. 10 is the case $p = 1$, $w_k = 1/K$. Where the magnitude $C_p(\alpha_\ell)$ is left in place, the retention factor becomes a second gain beside γ_ℓ, coupled to it as in Eq. 11; where it is removed, by a degree-zero map or by dividing the readout by $C_p(\alpha_\ell)$, α_ℓ vanishes from the forward pass and receives no gradient at all. The retention factor can therefore bear upon re-entry other than as a magnitude only if the fading trace passes through a map that is not homogeneous in it, and such a map must weigh the trace against some scale or quantity that the decay leaves untouched, be it a bias, a threshold, the intrinsic scale of a

nonlinearity or the present feedforward state. Placed on the path of the content, as a nonlinearity or biased threshold applied to $m_{\ell,k}$ before readout, that comparison would let the content change as the trace fades, the transformation that Eq. 1 excludes and the retained precision of surviving items (Pratte, 2018) argues against. Placed on the path of access, weighing $\alpha_\ell^k\, W_\ell^m * m_\ell$ against $W_\ell^x * x_\ell + b_\ell$ and letting the outcome decide how much of an unchanged q_ℓ is admitted, it is Eq. 6, of which the sigmoid is the simplest smooth form; the comparison must, moreover, act element-wise, since one returning a single scalar per sample would merely rescale the readout once more. Within the gate, finally, a learnable scale upon the trace drive would let the strength of the trace's hold upon access be set by that scale as readily as by α_ℓ, reopening on a smaller stage the confound the gate removes, and it is this route that the affine-free write and the unit-norm rows of W_ℓ^m foreclose.

THE GATED READOUT Under Eq. 7, α_ℓ reaches the readout through the gate alone,

$$\frac{\partial u_\ell}{\partial \alpha_\ell} = \frac{1}{K} \sum_{k=1}^{K-1} k\, \alpha_\ell^{k-1}\, z_{\ell,k}\, (1 - z_{\ell,k}) \odot (W_\ell^m * m_\ell) \odot q_\ell, \tag{12}$$

a direction that vanishes at $k = 0$ and is in general no multiple of u_ℓ, so that $\partial L/\partial \alpha_\ell$ is no longer tied to $\partial L/\partial \gamma_\ell$. Linearizing the gate about its allocation term $h_\ell = W_\ell^x * x_\ell + b_\ell$,

$$u_\ell \approx \sigma(h_\ell) \odot q_\ell + c(\alpha_\ell)\, \sigma'(h_\ell) \odot (W_\ell^m * m_\ell) \odot q_\ell, \tag{13}$$

shows what α_ℓ governs in its place: the weight of the trace-driven component of access relative to the feedforward one, a change in the pattern of the readout that the scalar gain of Appendix C preserves, while changes in its overall magnitude are largely absorbed by that gain, leaving the magnitude of re-entry chiefly to γ_ℓ.

C THE DIVISIVE GAIN

The normalization of Eq. 3 is $\mathcal{N}(u_\ell; x_\ell) = \rho_\ell u_\ell$, with

$$\rho_\ell = \mathrm{sg}\left[\mathrm{clip}\left(\frac{r_\ell^{-1}}{1 + r_\ell},\, \rho_{\min},\, \rho_{\max}\right)\right], \qquad r_\ell = \frac{\|u_\ell\|_2}{\|x_\ell\|_2}, \tag{14}$$

norms taken per sample over the entire $C_\ell \times H_\ell \times W_\ell$ activation and floored at $\epsilon = 10^{-6}$, $\mathrm{sg}[\cdot]$ denoting the stop-gradient and $(\rho_{\min}, \rho_{\max}) = (0.01, 100)$. The factor r_ℓ^{-1} brings the candidate to the scale of its stage, amplifying a weak candidate more than a strong one in the manner of normalization by recurrent amplification (Heeger & Zemlianova, 2020), and the divisive term, whose semi-saturation constant is fixed at unity, compresses it again in proportion to its original relative strength (Carandini & Heeger, 2012). Within the bounds, that is for $0.0099 \lesssim r_\ell \lesssim 9.51$,

$$\|\rho_\ell u_\ell\|_2 = \frac{\|x_\ell\|_2}{1 + r_\ell}, \tag{15}$$

so that the injected update $\gamma_\ell \|x_\ell\|_2/(1 + r_\ell)$ takes its scale from the stage's own activation and from γ_ℓ, and is only mildly compressed as the raw candidate grows; it is this property that allows γ_ℓ to be read as a strength of re-entry comparable across stages whose activations differ in scale. Where r_ℓ falls below the lower bound, the gain saturates at $\rho_{\max}$ and the normalized candidate, $100\, r_\ell \|x_\ell\|_2$, falls short of $\|x_\ell\|_2/(1 + r_\ell)$; where it rises above the upper, the gain saturates at $\rho_{\min}$, and the normalized candidate grows as $0.01\, r_\ell \|x_\ell\|_2$, keeping the update below $\gamma_\ell \|x_\ell\|_2$ for $r_\ell < 100$. The statistic is computed under a stop-gradient, so that the gain acts as a forward-only controller: gradients reach u_ℓ as though ρ_ℓ were constant, and neither α_ℓ nor any network activation or weight can seek advantage by steering the normalization itself.

D EXPERIMENTAL DETAILS

BACKBONES AND BASELINES The block is attached to ResNet-18 and ResNet-50 (He et al., 2016), ConvNeXt-Tiny (Liu et al., 2022), DenseNet-121, Swin-T (Liu et al., 2021) and CORnet-S (Kubilius et al., 2019), in every case after each of the four stages the backbone already possesses and with the placement of Appendix A, and every ICONIC variant is compared against the same backbone trained without the blocks.

DATASETS The stage-wise profiles of Table 1 and of Figs. 6 and 5 are obtained on Tiny ImageNet (Le & Yang, 2015), whose 200 classes are collapsed into 20 superclasses for the coarse-label condition, and whose images, for the scrambled condition of Fig. 4, are phase-scrambled by randomizing the Fourier phase of every image while its amplitude spectrum is kept, which removes edges, contours and grouping cues and leaves the global spectral statistics in place. MNIST and SVHN serve as the non-ecological condition of the same figure. The class-count sweep of Fig. 7 trains on subsets of CIFAR-100 growing from 5 to 100 classes in steps of 5, and the data-size sweep of Fig. 8 on 1, 2, 4, 8, 16, 32, 64, 128, 256 and 500 training images per class. The ResNet-50 and ResNet-18 comparisons of Tables 3, 4 and 5 use ImageNet-100, a 100-class subset of ImageNet, and its corrupted counterpart ImageNet-100-C, the fifteen corruption types of Hendrycks & Dietterich (2019) at severities one to five.

EVALUATION Clean accuracy is top-1 accuracy on the held-out images of each dataset. Mean corruption accuracy (mCA) is top-1 accuracy averaged over the fifteen corruption types and five severities of Hendrycks & Dietterich (2019), which Tables 3 and 4 also report per corruption type. Adversarial accuracy is measured under projected gradient descent (Madry et al., 2018) at $\epsilon \in \{0.25, 0.5, 1\}/255$ and reported as the mean over the three budgets. Accuracies are averaged over seeds, with standard deviations where given, and the learned α_ℓ and γ_ℓ of every figure and table are read from the trained models and averaged over the same seeds.

VONENET FRONT ENDS The VOneNet variants of Fig. 6 follow the strategy of Dapello et al. (2020): the stem of each backbone, its first convolutional block, is replaced by their VOneBlock, a fixed Gabor filter bank parameterized to match empirical distributions of primate V1 simple and complex cells and left untrained, followed by the trainable bottleneck convolution of the same work, which maps the output of the filter bank to the channel count the first stage expects. The four stages and the ICONICBlocks that follow them are left exactly as in the standard variant, so that the two curves of the figure differ only in what precedes S_1.

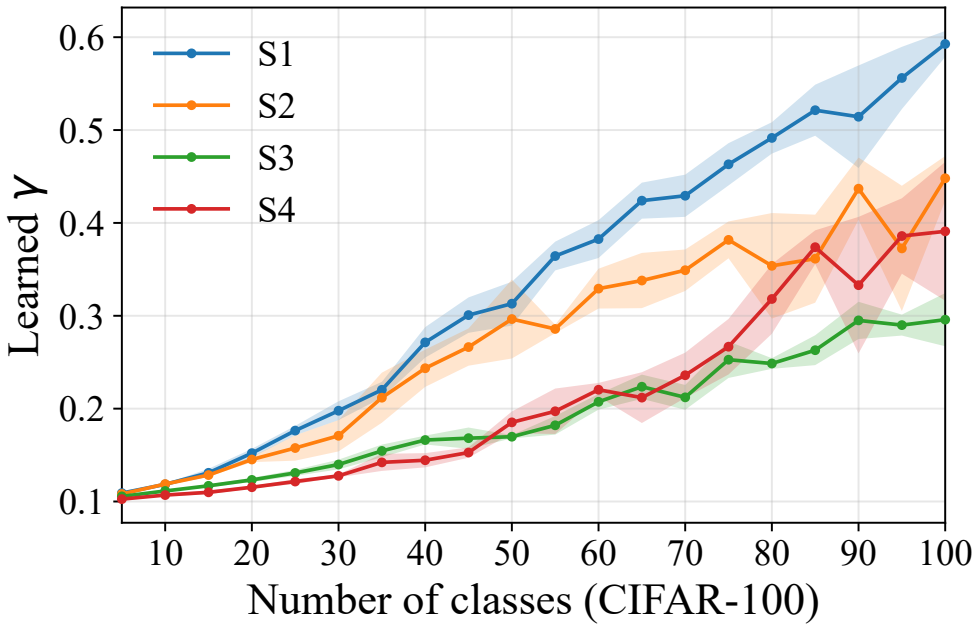


Figure 7: Learned γ at each of the four stages as the number of CIFAR-100 classes seen during training grows from 5 to the full 100, with the solid line tracing the mean and the shaded band the spread across seeds. All four stages begin from nearly the same small value when only a handful of classes are present and climb steadily as more classes are added, separating as they go so that S1 ends highest, S2 and S4 settle in between, and S3 stays the lowest across most of the range, with S4 becoming noticeably less stable once the class count passes roughly 75.

E Additional Results

This appendix collects the class-count, data-size and stage-ablation sweeps (Figs. 7, 8 and 9) and the ImageNet-100-C robustness comparisons (Tables 3 and 4).



Figure 8: Learned γ at each of the four stages as the number of training images per class grows from 1 to 500, with the settings placed at equal intervals along the horizontal axis and error bars marking the spread across seeds. Every stage holds γ close to zero through the low-data regime and only begins to depart from it once the per-class count reaches the low hundreds, after which S4 rises far more steeply than the rest, S1 follows at a clear distance, and S2 and S3 grow the least and end almost level with one another.

Table 3: ImageNet-100-C robustness comparison (5-seed averages). Per-category mCA (%) averaged over severities 1–5. Best per row in **bold**.

Corruption	ResNet-50	IconicResNet-50
Noise		
Gaussian Noise	19.60%	**26.24%**
Shot Noise	19.88%	**26.42%**
Impulse Noise	14.00%	**19.84%**
Blur		
Defocus Blur	34.34%	**36.96%**
Glass Blur	42.42%	**46.56%**
Motion Blur	46.78%	**51.60%**
Zoom Blur	46.78%	**50.74%**
Weather		
Snow	32.76%	**40.80%**
Frost	34.50%	**41.52%**
Fog	32.94%	**37.38%**
Brightness	61.06%	**64.96%**
Digital		
Contrast	23.42%	**29.28%**
Elastic Transform	64.00%	**68.00%**
Pixelate	66.54%	**70.44%**
JPEG Compression	56.52%	**59.48%**
Clean Top-1	74.14% $\pm$0.43	**77.83%** $\pm$0.26
mCA	39.74% $\pm$0.52	**44.68%** $\pm$0.61

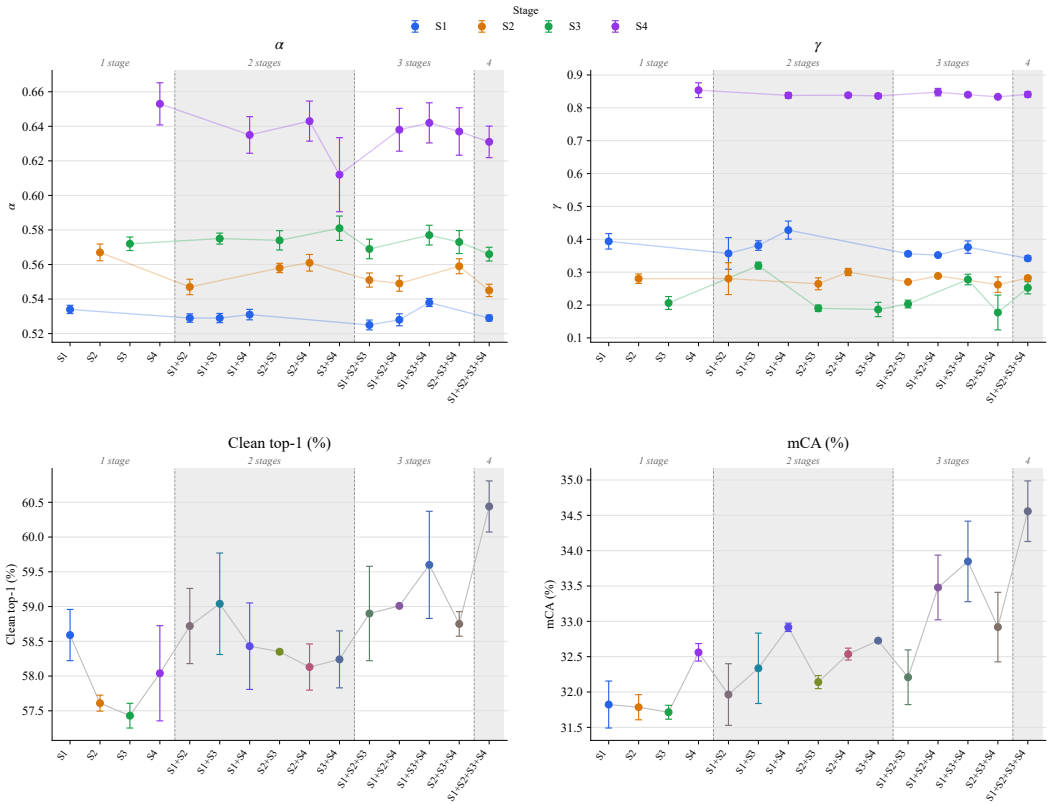


Figure 9: Learned iconic parameters and accuracy across all 15 stage configurations of the ablation, with configurations grouped along the horizontal axis by the number of stages carrying an IconicBlock (one, two, three and all four, alternately shaded). The top row shows the learned α (left) and γ (right) of each active stage, colored by stage and joined across configurations so that the value a given stage settles on can be followed as other stages are switched on around it, while the bottom row reports clean top-1 accuracy (left) and mean corruption accuracy, mCA (right), where each point takes the blended color of the stages active in that configuration. Error bars denote ± 1 standard deviation across seeds, and the full S1+S2+S3+S4 configuration reaches the highest clean top-1 and mCA of all fifteen.

Table 4: ImageNet-100-C robustness comparison. Per-category mCA (%) averaged over severities 1–5. Best per row in **bold**.

Corruption	ResNet-18	IconicResNet-18
Noise		
Gaussian Noise	21.86%	**25.80%**
Shot Noise	21.44%	**26.30%**
Impulse Noise	15.89%	**17.80%**
Blur		
Defocus Blur	33.56%	**38.40%**
Glass Blur	39.89%	**45.00%**
Motion Blur	42.27%	**51.60%**
Zoom Blur	44.29%	**48.90%**
Weather		
Snow	32.41%	**41.80%**
Frost	36.22%	**44.60%**
Fog	38.36%	**46.10%**
Brightness	61.74%	**68.30%**
Digital		
Contrast	27.09%	**35.60%**
Elastic Transform	56.66%	**65.50%**
Pixelate	61.46%	**70.50%**
JPEG Compression	53.68%	**61.20%**
Clean Top-1	73.00%	**79.16%**
mCA	39.12%	**45.80%**

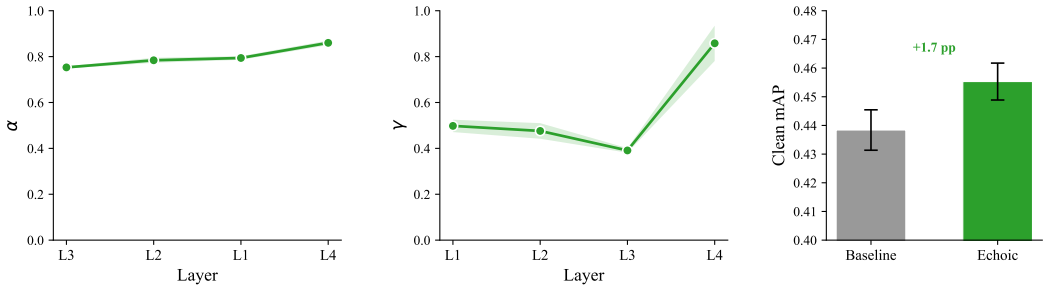
F Representational Similarity Analysis

To assess whether iconic memory blocks shift internal representations toward those of human visual cortex, we performed a representational similarity analysis (RSA) on the 92-image stimulus set of the Algonauts 2019 benchmark (Cichy et al., 2016; 2019), which provides group-averaged brain representational dissimilarity matrices (RDMs) from fMRI (early visual cortex and inferior temporal cortex) and from MEG (early and late post-stimulus windows). Each trained model variant was presented with all 92 images, and activations were read out at the input of each residual stage and before the final pooling, then spatially averaged to yield one feature vector per image. Model RDMs were computed as $1 - \rho_{\text{Spearman}}$ between these vectors, and brain–model alignment was measured as the Spearman correlation between the upper-triangular entries of the model and brain RDMs (Table 5). Because no permutation tests, bootstrap intervals or noise-ceiling estimates were computed, and because the models were trained on Tiny ImageNet rather than on the stimulus distribution, we treat these alignment scores as supplementary and descriptive evidence.

Table 5: ImageNet-100 top-1 accuracy and brain-model alignment on the Algonauts 2019 92-image set (Cichy et al., 2016; 2019). Alignment is the Spearman correlation between the upper-triangular entries of the model and brain RDMs.

		fMRI		MEG	
Model	IN-100 Acc.	EVC	IT	Early	Late
Iconic ResNet50	**77.7%**	**+0.079**	**+0.157**	**+0.196**	**+0.191**
Vanilla ResNet50	74.3%	+0.037	+0.130	+0.061	+0.166
Vanilla CORnet-S	76.0%	+0.018	+0.076	+0.010	+0.157
Vanilla ConvNeXt	75.7%	−0.012	+0.041	−0.035	+0.089

G EchoicNet



Sensory persistence extends to audio. To test whether the learned memory dynamics transfer from vision to audio, we trained an echoic ConvNeXt-Tiny on FSD50K, a large-scale audio event classification dataset containing over 51,000 variable-length clips across 200 sound categories. Each clip was resampled to 32 kHz, truncated or padded to 4 seconds, and converted to a 128-bin log-mel spectrogram (1024-point FFT, 800-sample window, hop 320, 20–14,000 Hz), yielding a single-channel time–frequency image that the ConvNeXt backbone processes with its standard stride-2 stem and stochastic depth 0.1. The echoic arm places an independent iconic memory block after each of the four ConvNeXt stages, while the matched baseline uses the identical backbone and initialization with no memory blocks. Both conditions were trained for 60 epochs with AdamW (learning rate 2.5×10^{-4} linearly scaled from 5×10^{-4} at batch 128, weight decay 0.05), a five-epoch linear warm-up followed by cosine annealing, gradient clipping at 1.0, and mixed-precision training, over three seeds (3, 11, 18). Validation used three-crop averaging and final evaluation used eight crops.