PERSISTENT MEMORY FOR VISUAL RECOGNITION UNDER CORRUPTION AND MISSING EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Can persistent memory support visual recognition when the input is corrupted or incomplete, particularly when labeled training data are limited? We investigate this question through the Vision Hopfield Memory Network (V-HMN), a model that integrates associative retrieval with iterative feature refinement across layers. Separate local and global memory banks store training-derived representations of patch neighborhoods and image-level context. During inference, each branch retrieves a weighted combination of stored patterns and uses the discrepancy between the retrieved and current representations to guide feature updates. This design makes information from training examples available through both learned parameters and persistent representations, while keeping the memory fixed at evaluation. The retrieved entries and their weights also make the memory component of inference inspectable. We study this mechanism through common image corruptions, controlled occlusion and masking, and training with limited labels, restricting memory construction to the declared training budget. Comparisons with the same architectural scaffold trained without retrieval, alongside interventions that disable reads or alter memory contents, distinguish the effects of learning with memory from those of accessing it during inference. Our analyses examine whether memory contributes beyond clean recognition performance, how its contribution changes with label availability, and when retrieval corrects predictions or introduces errors. By connecting these outcomes to local and global retrieval, memory coverage, and the entries used during computation, we investigate both the potential and the limitations of persistent memory for recognition under degraded evidence.

1 INTRODUCTION

When an image loses part of its evidence, a classifier must use the information that remains. Parameter fitting makes training information available through learned transformations. Persistent memory also supplies stored representations to the current computation. This creates a second route for supporting recognition, but its usefulness depends on what the query retrieves.

The Vision Hopfield Memory Network (V-HMN) stores training-derived local neighborhoods and global context throughout a visual recognition model. Similarity-dependent weights combine stored values, and learned residual updates refine intermediate features before branch fusion. Training labels organize retention; inference uses neither those labels nor new writes. Tracing these operations exposes the information actually retrieved.

We evaluate V-HMN through recognition under degraded inputs, interventions on its memory reads, learning from limited training subsets, and traces of the stored information used during inference. Four findings connect the model to these questions:

- On ImageNet-C, V-HMN has 48.10% mean corruption error versus 60.57% for ResNet-50 V1, with 76.50% versus 76.15% clean accuracy under different training recipes. Smaller-benchmark comparisons retain both external models and separately trained controls (Section 5).
- Within fitted models, selective reads improve dataset-level corruption accuracy over bank means. Local/global interventions distinguish the source of this gain from dependence on the nonselective pathway (Section 5.2).

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.

- V-HMN has the highest reported mean at 10% and 30% training fractions on CIFAR-10, CIFAR-100, and Fashion-MNIST. A separate budget-controlled study identifies its lowest-budget gain over the scaffold and its schedule sensitivity (Section 5.3).
- Actual read traces, content interventions, and signed score accounting expose stored information and its influence on a defined computation (Section 5.4).

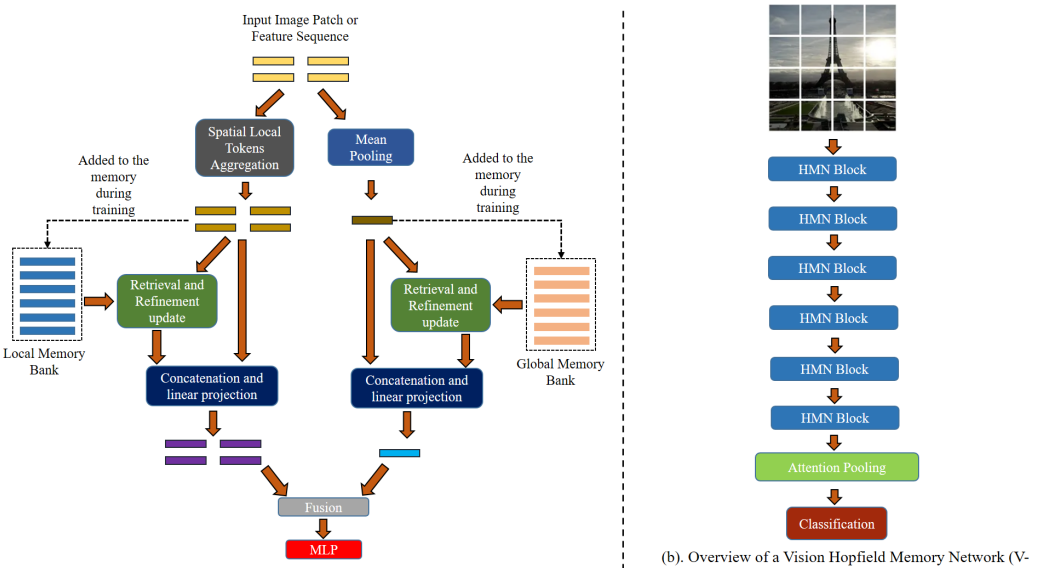


Figure 1: V-HMN architecture. Left: local-neighborhood and pooled global queries retrieve representations from separate banks, then combine initial and updated features through projection and fusion. Right: stacked blocks lead to attention pooling and classification. Dashed arrows denote training writes; evaluation freezes the banks. This six-block illustration differs from the staged ImageNet configuration (Appendix A).

2 Related Work

Associative retrieval and visual memory. Modern Hopfield networks connect associative retrieval to attention (Ramsauer et al., 2021); their energy guarantees do not automatically transfer to a learned vision block. Associative Transformer updates low-rank memory with bottleneck-attended batch tokens, then projects it for Hopfield token reconstruction; evaluation freezes its memory (Sun et al., 2025). V-HMN retains individual sample-derived local/global representations in per-class queues and corrects intermediate queries using their weighted values. The cache model's output-level label votes motivate our exemplar control (Orhan, 2018). Pretrained-neighbor classification, image–text retrieval, and replaceable multimodal memory provide other uses of stored information (Geirhos et al., 2025; Iscen et al., 2023; Kang et al., 2025). Our focus is the effect of intermediate reads relative to the trained and output-level controls evaluated here.

Prototypes, labels, and explanations. Prototypical Networks address episodic novel-class recognition (Snell et al., 2017), unlike our fixed-class label budgets. ProtoPNet links part comparisons directly to classification (Chen et al., 2019); V-HMN transforms weighted mixtures further. Attention weights alone need not explain predictions (Jain & Wallace, 2019); traces identify reads, and interventions measure their influence.

Robustness and missing evidence. Common-corruption accuracy differs from perturbation stability (Hendrycks & Dietterich, 2019); occlusion motivates separate masks and compositional modeling (Kortylewski et al., 2020). AugMix's augmentation/consistency training and MoireMix's procedural augmentation provide alternative robustness mechanisms (Hendrycks et al., 2020; Matsuo et al.,

2026), motivating our augmentation-by-memory comparison. Synthetic-corruption gains need not extend to natural shifts (Taori et al., 2020); our claims concern the measured protocols.

3 VISION HOPFIELD MEMORY NETWORK

3.1 LOCAL AND GLOBAL BRANCHES

Figure 1 follows the forward path through local/global reads, fusion and classification; experiment-specific depths and widths are given in Appendix A. For one image, let $H \in \mathbb{R}^{N \times d_e}$ denote its N tokens after block-input normalization. The local branch projects an overlapping $w \times w$ token neighborhood to $q_i^L \in \mathbb{R}^d$; the global branch projects mean-pooled tokens to $q^G \in \mathbb{R}^d$. Each branch has its own bank $M^b \in \mathbb{R}^{K_b \times d}$, $b \in \{L, G\}$, with capacity K_b. Banks are separate across blocks and store depth-specific representations.

Let V^b be the set of occupied slot indices. Suppressing branch and block indices, retrieval uses cosine-softmax weights:

$$a_j(q; M) = \frac{\exp(\sqrt{d}\, \widehat{q}^\top \widehat{M}_j)}{\sum_{k \in V} \exp(\sqrt{d}\, \widehat{q}^\top \widehat{M}_k)}, \qquad m(q; M) = \sum_{j \in V} a_j(q; M) M_j. \tag{1}$$

Hats denote vectors normalized with a small ϵ guarding zero norms. Similarity uses normalized queries and keys with fixed scale $\sqrt{d}$; the weighted values M_j retain their original magnitudes.

The native refinement is $q_{t+1} = q_t + \beta[m(q_t) - q_t]$, with an unconstrained learned scalar β per branch and block, shared across iterations. Primary smaller-benchmark runs use one step, denoted q^+. A learned projection maps the concatenated initial and refined queries $[q; q^+]$ to $\mathbb{R}^{d_e}$. The global output is broadcast across tokens and added to the local output before residual and feedforward transformations. At the final layer, learned token scores s_i give pooling weights $p_i = \mathrm{softmax}_i(s)$ and $v = \sum_i p_i H_i$ for the linear classifier.

The predictive-coding-inspired residual has an analytical fixed-target interpretation, but is not generally gradient descent on $\frac{1}{2}\|q - m(q)\|^2$, whose full gradient includes the retrieval Jacobian. Training retains gradients through retrieval; the coefficient is unconstrained and each iteration requeries memory, without an imposed energy-descent guarantee (Appendix F).

A read bypass sets its correction to zero ($\lambda = 0$ in Figure 4); native reads use $\lambda = 1$. Projections, fusion, residual paths and pooling remain. An empty bank uses $m(q) = q$, making the update an identity. Depth and strength interventions are specified in Appendix F.

3.2 WRITES, PERSISTENCE, AND PROVENANCE

Each branch partitions its slots into equal per-class queues, replacing the oldest entry within a class. Local writes retain up to five projected neighborhoods with largest feature norm per image; global writes retain the projected pooled feature. Stored values are detached buffers, while classification loss trains query projections and refinement coefficients. Multiple slots or augmented views from one source do not add independent examples.

Primary smaller-benchmark training reads the existing banks, then commits detached loss-forward queries and source/view identities after backward and the optimizer step. The queries use the same crop/flip view and pre-optimizer encoder, without an extra forward or epoch-end refresh. The ImageNet and CIFAR-10 models use their experiment-specific population protocols (Appendices H and I).

Evaluation freezes learned parameters and banks and never writes images, labels or representations. Figure 4 details the controlled training lifecycle. Checkpoints preserve values, occupancy, queue positions and provenance. Source IDs, view transforms, token coordinates, write steps and slot generations link actual reads to training views. Models without source metadata support latent-slot inspection only.

3.3 SELECTIVE RETRIEVAL AND THE NONSELECTIVE PATHWAY

For a bank with K valid entries, let $\mu = K^{-1} \sum_j M_j$. A single native update decomposes exactly as

$$q^+ = (1 - \beta)q + \beta\mu + \beta[m(q) - \mu]. \tag{2}$$

Uniform weights retain the first two terms: a learned rescaling of the query and a bank-dependent offset. The last term isolates the deviation supplied by selective weighting. Thus replacing a read by its bank mean preserves a nonselective transformation that bypass removes. The unconstrained β need not make that transformation a contraction. For the concatenation projection $P[q; q^+] = Aq + Bq^+ + b$, this gives

$$P[q; q^+] = [A + (1 - \beta)B]q + \beta B\mu + b + \beta B[m(q) - \mu]. \tag{3}$$

With frozen banks, this single-step uniform-read map can be absorbed into the existing projection weights and bias, giving an equivalent read-bypassed computation. Bypassing the original weights instead returns $[A + B]q + b$. The uniform-trained control learns with evolving sample-derived means, so this inference equivalence does not establish equivalence to scaffold training. Only the final term depends on selective weighting.

3.4 MEASUREMENTS AND INSPECTABLE READS

We pair query and projected-feature discrepancies with accuracy, margin, loss and rescue/harm. Clean pre-read and native post-read states are model-produced references, not ground-truth semantic states. The exact correction identity and fixed-target interpretation in Appendix F formulate diagnostic predictions without guaranteeing recognition improvement.

A trace records the branch, block, query/token, occupied slot indices, weights, source identities and returned mass. Top-k displays retain omitted mass $1 - \sum_{j \in T_k} a_j$. Removal selects entries from that input's unmodified read, renormalizes the remaining weights and reruns the ordinary suffix without changing the bank. Cumulative-mass and random-removal controls record actual entry counts and removed mass. A named-block intervention isolates its read operators; our exploratory probe uses block three of six.

4 EXPERIMENTAL COMPARISONS

Questions and protocols. Three-repeat smaller-benchmark studies compare trained models and inference interventions; ImageNet evaluates one V-HMN model and ResNet-50 V1 under their respective recipes. The fractional-data study reports multi-run mean accuracies and SDs at fixed fitting fractions. Its training fractions differ from the selection-inclusive budgets in the controlled study (Appendix I).

Trained models and inference controls. The separately trained scaffold retains all non-memory computation; full-minus-scaffold measures the total method difference under a common recipe. The uniform-trained model learns its own representations and banks under mean reads. Bypass and mean-read interventions instead change a fixed V-HMN checkpoint. External comparisons use ResNet-18 (He et al., 2016), a post-hoc exemplar head on scaffold features, and AiT-S, a frozen-memory adaptation of the author release. Architecture-specific tuning and selected settings are given in Appendix A. The fractional-data models are ViT (Dosovitskiy et al., 2021), Swin-ViT (Liu et al., 2021), MLP-Mixer (Tolstikhin et al., 2021), MetaFormer instantiated as PoolFormer (Yu et al., 2022), Vim (Zhu et al., 2024), and a local AiT wrapper distinct from the AiT-S adaptation.

Data and selection. Tiny ImageNet uses a stratified 90%/10% fitting/selection split of official training data. The 200-epoch schedule was fixed from runtime measurements before final comparisons, retaining three independent repetitions. Clean selection maximizes accuracy, then minimizes NLL. Primary training uses crop/flip without corruption augmentation or mixed labels. CIFAR-100 budgets include selection labels: 50, 150, and 500 examples per class split as 45/5, 135/15, and 450/50. Separately nested fitting and selection pools remain disjoint and are shared across models. Only fitting examples populate memory in these repeated training comparisons. Full recipes and the 35,200-update lowest-budget sensitivity are specified in Appendix A.

The released 15 corruption types and five severities are never used for fitting or selection. Corrected Tiny ImageNet-C pairs with the author's separate bounding-box crops (Hendrycks, 2021); canonical validation supplies primary clean and mask scores. The cohort mapping and preprocessing are specified in Appendix A. Masks remove nested cells of an 8×8 grid or one randomly placed square, with RGB fitting-mean fill before normalization. Fractions $\{0, .10, .25, .50, .75\}$ and three realizations are paired across models. Realized areas and non-nested square placement are specified in Appendix E.

Read-operation and augmentation factorials. Local/global read factorials use bypass or uniform means as distinct references, retaining all non-memory computation. Each policy applies across all blocks with weights, banks and inputs fixed; earlier interventions can change downstream queries. The separate AugMix factorial applies the published consistency objective to V-HMN and scaffold with identical views and steps; only the unmixed view supplies writes. Appendix F.1 defines branch allocations, and Appendix D.1 specifies AugMix.

Endpoints and uncertainty. Primary endpoints are raw mean corruption error $\bar{E} = \frac{1}{75} \sum_{c,s} E_{c,s}$ and separate grid/square accuracy AUCs, trapezoidally integrated over nominal removal on $[0, .75]$ and divided by .75. Clean accuracy supplies the zero-removal endpoint. Errors are stored as fractions and displayed as percentages. CIFAR-100-C and Tiny ImageNet-C use no AlexNet normalization; ImageNet-C additionally reports benchmark-normalized mCE. Rescue minus harm equals the fixed-checkpoint accuracy difference. For the repeated smaller-benchmark comparisons, sample SD describes three training repetitions; separate 2,000-replicate source bootstraps retain every image's associated conditions and fixed checkpoints together. These intervals do not estimate training-population uncertainty. The uniform-trained comparisons and associative comparator are targeted follow-ups motivated by inspected benchmark results. Settings are fixed before their final evaluations, with checkpoint selection restricted to clean development data; uniform-trained settings are inherited from the paired V-HMN runs.

5 Results

5.1 Recognition with corrupted and missing evidence

How does the complete model recognize degraded inputs? Table 1 compares separately trained models at full label budgets and separately labels inference-time bypass in V-HMN.

Comparison with an associative model. V-HMN has lower mean corruption error than our frozen-memory AiT-S adaptation by 6.70 ± 0.73 points on Tiny ImageNet and 12.81 ± 0.44 points on CIFAR-100 (paired mean $\pm$ sample SD, three seeds). The masking differences depend on dataset and geometry (Table 1). Architectures and clean-selected recipes differ in this external-model comparison.

The paired clean-to-corruption accuracy drop is 4.42 points smaller for V-HMN than AiT-S on CIFAR-100, and 5.83 points larger on Tiny ImageNet's matching corruption-source cohort (paired SDs 0.29 and 1.04 points); this is descriptive, not a causal clean-accuracy adjustment.

Mask geometry changes the ranking. V-HMN exceeds ResNet-18 in CIFAR grid accuracy AUC under fitting-mean fill; ResNet-18 leads on the other CIFAR endpoints and all Tiny endpoints. On CIFAR-100, V-HMN's higher raw grid NLL AUC relative to ResNet-18 becomes lower after matched clean-development temperature scaling (Table 9). The trained scaffold has lower full-budget corruption error and higher grid/square AUC than V-HMN on both datasets, with modest further gains from its exemplar head. Figure 6 retains every severity and masking level.

Learning with uniform reads. Does selective weighting help when the nonselective pathway learns its own representations? Relative to uniform-trained V-HMN, the original model changes corruption error by $+0.51 \pm 0.59$ points on CIFAR-100 and -0.09 ± 0.21 on Tiny. On Tiny it has lower clean accuracy and lower grid/square AUC by 1.08 ± 0.57/0.83 ± 0.36 points. This comparison does not establish a general selective-training advantage. Clean-development temperature scaling changes the Tiny corruption-NLL ordering but leaves the positive masking-NLL gaps and all Top-1 decisions unchanged (Table 9).

Table 1: Clean-trained recognition with degraded evidence. Mean ± sample SD of three training repetitions; all values are percentages. C error averages all 15 corruptions and five severities; grid/square accuracy AUC is normalized over $[0, .75]$. Clean and masks use canonical images; Tiny corruption-paired crops differ (Appendix A). Bypass reuses the V-HMN checkpoint; scaffold and uniform-read controls are separately trained. The uniform-trained control learns its own representations and banks. CIFAR uses 500 labels per class, including selection. AiT-S is our frozen-memory adaptation of the author release. The upper rows compare V-HMN and its controls; italic rows identify external models with clean-selected recipes. Ranking compares trained models within each dataset; bypass is unranked. Bold marks numerical best values (ties retained), not significance.

Model	Clean ↑	C error ↓	Grid AUC ↑	Square AUC ↑
Tiny ImageNet				
V-HMN	50.68 ± 0.09	82.87 ± 0.48	17.78 ± 0.56	19.48 ± 0.35
V-HMN, bypass at inference	32.49 ± 0.68	89.56 ± 0.18	13.03 ± 0.46	12.26 ± 0.30
Trained scaffold	51.48 ± 0.35	82.60 ± 0.15	18.04 ± 0.11	19.93 ± 0.48
Uniform-trained V-HMN	51.20 ± 0.12	82.96 ± 0.42	18.87 ± 0.08	20.31 ± 0.20
Scaffold + exemplar	51.64 ± 0.34	82.50 ± 0.20	18.17 ± 0.21	20.08 ± 0.59
ResNet-18	$\mathbf{63.96} \pm 0.04$	$\mathbf{76.88} \pm 0.23$	$\mathbf{21.36} \pm 0.15$	$\mathbf{23.75} \pm 0.18$
AiT-S (adapted)	33.63 ± 0.38	89.57 ± 0.29	12.68 ± 0.83	10.13 ± 0.61
CIFAR-100				
V-HMN	64.07 ± 0.23	51.96 ± 0.17	29.85 ± 0.62	20.93 ± 0.28
V-HMN, bypass at inference	51.11 ± 1.90	62.17 ± 1.25	23.40 ± 1.68	16.84 ± 1.18
Trained scaffold	64.29 ± 0.14	51.60 ± 0.46	29.87 ± 0.21	21.12 ± 0.41
Uniform-trained V-HMN	63.91 ± 0.80	51.46 ± 0.57	30.04 ± 0.35	20.87 ± 0.59
Scaffold + exemplar	64.46 ± 0.20	$\mathbf{51.31} \pm 0.36$	$\mathbf{30.20} \pm 0.27$	21.26 ± 0.47
ResNet-18	$\mathbf{76.90} \pm 0.21$	51.44 ± 0.20	26.18 ± 0.53	$\mathbf{27.15} \pm 0.80$
AiT-S (adapted)	55.68 ± 0.50	64.78 ± 0.32	29.55 ± 2.65	21.88 ± 1.13

Recognition at ImageNet scale. How does V-HMN compare with ResNet-50 on ImageNet-C? V-HMN has 12.47 points lower corruption error than torchvision ResNet-50 `IMAGENET1K_V1`, with similar clean accuracy (Table 2). The models use different training recipes and their specified clean preprocessing (Appendix H). The comparison therefore assesses the evaluated systems rather than isolating an architecture effect. The result establishes an advantage over this evaluated checkpoint under their respective protocols.

Table 2: ImageNet recognition and read interventions. Clean accuracy (↑) and raw corruption error (↓) are percentages; mCE (↓) uses the official ImageNet-C normalization over all 75 conditions. The three V-HMN rows share one trained model. Learned parameters, including classifiers: 26.00M versus 25.56M; sample-written memory is additional state. Different training recipes and clean preprocessing are specified in Appendix H; these are checkpoint point estimates, not repeated-training means. Ranking compares full V-HMN with ResNet only; interventions are unranked. Bold marks numerical best values (ties retained), not significance.

Model / inference reads	Clean (%) ↑	C error (%) ↓	mCE ↓
V-HMN: full	**76.50**	**48.10**	**61.17**
V-HMN: bypass	70.56	56.87	72.39
V-HMN: uniform	75.56	49.61	63.12
Public ResNet-50 V1	76.15	60.57	76.43

5.2 Source of the memory contribution

The model comparisons assess complete recognition systems. We now hold V-HMN's learned parameters and stored representations fixed to identify the contribution of its read operation. Full reads improve corruption accuracy over bypass by 6.69 ± 0.30 points on Tiny and 10.20 ± 1.41 on CIFAR-100; gains over frozen bank means are 1.85 ± 0.23 and 3.43 ± 0.18 (Figure 2). Both the nonselective pathway and selective weighting contribute within these models. The read gains

are larger on matching clean inputs, so they include dependence needed for clean recognition (Appendix C). Within the evaluated ImageNet V-HMN, full reads add 8.77/1.52 points of corruption accuracy over bypass/uniform reads, alongside 5.94/0.94 points on clean images. Selective weighting contributes in this checkpoint, including to clean recognition; the image-paired intervals in the appendix condition on these fitted weights.

Where do the corruption differences arise? Sources recognized by both ImageNet models when clean comprise 70.37% of the cohort and contribute 11.29 of the 12.47-point corruption gap. Within the shared-clean-correct groups, full-minus-uniform corruption accuracy is 1.75/3.15/2.35 points on ImageNet/CIFAR/Tiny. Appendix C.2 retains all four groups, their sizes and unconditional contributions. This outcome-conditioned partition describes the full-cohort result; it is not a causal clean-accuracy adjustment.

Local and global retrieval contributions. Which selective pathway supplies the gain over bank means? Averaging the two orders of enabling selective reads, local/global corruption-accuracy contributions are 0.75 ± 0.10/2.68 ± 0.21 points on CIFAR and 0.27 ± 0.29/1.59 ± 0.32 on Tiny (Figure 2). The additional selective contribution is chiefly global in both fitted settings; the Tiny local effect varies across repetitions. CIFAR's broader bypass-referenced allocations are instead 9.35 ± 0.87/0.85 ± 0.54 points. The reference therefore distinguishes strong local read dependence from the larger global contribution of selective weighting. The allocations sum to the full-minus-reference effect; their interaction must not be added again. Appendix F.1 gives all four cells, paired-clean results and covariance. The component needed by a fitted inference path need not be the component supplying its additional query-dependent benefit. These factorials localize an inference effect; they do not establish the cause of the fractional-data advantage.

Geometry and training recipe. Utility also changes with what remains visible: CIFAR reads improve square accuracy AUC but reduce accuracy at 50% removal by 0.90 ± 0.51 points. Black fill changes that effect to $+2.28 \pm 0.56$ and reverses the tested grid-accuracy ordering against ResNet (Appendix E).

Does selective retrieval still contribute with corruption-oriented training? Under AugMix, full reads exceed bank means by 0.78 ± 0.27 corruption-accuracy points and 2.11 ± 0.07/2.44 ± 0.07 grid/square AUC points. This inference contribution persists, while the total full-minus-scaffold corruption-accuracy difference is -0.23 ± 0.63 points.

5.3 Learning from limited training data

How does V-HMN compare with other models when fewer training examples are available?

Table 3: Clean Top-1 accuracy (%) from the three-seed fractional-data study, with reported means and SDs. Fractions count fitting examples; the final row gives V-HMN-only 50% observations. MetaFormer uses PoolFormer; AiT is the local wrapper, distinct from the AiT-S adaptation. Appendix I.1 gives the protocol and complete observations. Ranking uses only the seven-model 10%/30% comparisons. Bold marks numerical best values (ties retained), not significance.

Model	CIFAR-10 ↑		CIFAR-100 ↑		Fashion-MNIST ↑	
	10%	30%	10%	30%	10%	30%
ViT	72.73 ± 0.42	83.94 ± 0.33	40.48 ± 0.80	57.40 ± 0.79	87.17 ± 0.40	89.71 ± 0.45
Swin-ViT	70.37 ± 0.94	80.89 ± 0.24	35.25 ± 0.59	51.77 ± 0.60	88.42 ± 0.48	90.34 ± 0.15
MLP-Mixer	76.14 ± 0.16	85.53 ± 0.40	41.94 ± 0.98	56.22 ± 0.75	87.16 ± 0.63	89.41 ± 0.13
MetaFormer	49.92 ± 4.08	60.86 ± 6.05	19.39 ± 1.26	37.47 ± 3.19	85.02 ± 1.22	88.78 ± 1.03
Vim	69.02 ± 2.13	79.58 ± 0.37	36.61 ± 1.25	49.67 ± 0.79	86.09 ± 1.03	88.18 ± 2.43
AiT	67.89 ± 0.55	80.69 ± 1.36	35.84 ± 0.67	54.59 ± 0.87	84.33 ± 0.38	87.62 ± 0.32
V-HMN	$\mathbf{80.22} \pm 0.29$	$\mathbf{88.67} \pm 0.21$	$\mathbf{43.21} \pm 1.07$	$\mathbf{62.42} \pm 0.29$	$\mathbf{89.18} \pm 0.16$	$\mathbf{91.04} \pm 0.22$
V-HMN, 50%	91.19 ± 0.38		68.93 ± 0.35		91.53 ± 0.12	

The fractional-data study compares clean recognition across seven models. V-HMN has the highest reported mean in all six dataset/fraction comparisons (Table 3). At 10% on CIFAR-10, its reported

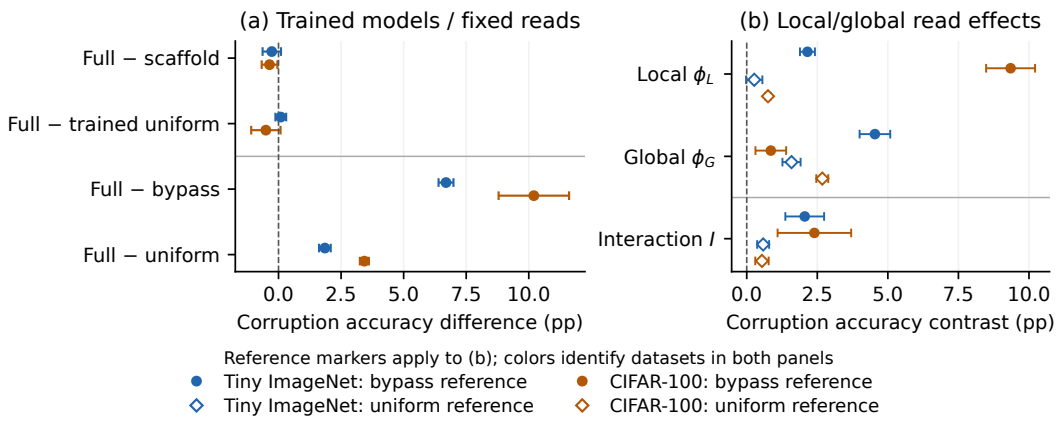


Figure 2: Trained-model comparisons and allocation of fixed-checkpoint read effects (mean ± sample SD of three paired training-run differences). (a) Above the divider: full V-HMN versus separately trained scaffold and uniform-read models. Below: full versus bypassed or uniform reads in the same checkpoint. (b) Each branch is selective or uses the indicated bypass (filled) or uniform (open) reference. Local and global symmetric allocations average a branch's effect over the two states of the other branch. They sum to the full-minus-reference effect; the interaction is a separate nonadditivity contrast and is not added again. Corruption accuracy averages all 75 conditions. Clean and masking allocations are reported in the appendix. Colors identify datasets throughout.

mean exceeds MLP-Mixer by 4.08 percentage points. At 30% on CIFAR-100, its reported mean exceeds ViT by 5.02 percentage points. The additional 50% observations extend V-HMN's reported learning curve; comparator measurements at that fraction were not reported. The fractions count fitting examples: 5,000/15,000/25,000 for CIFAR and 6,000/18,000/30,000 for Fashion-MNIST. They are not total label budgets including selection. Appendix I gives this protocol and separately reported diagnostic trajectories.

The budget-controlled CIFAR-100 study asks whether V-HMN improves on its separately trained scaffold with explicit fitting/selection accounting. At 50 labels per class (45 fitting, five selection), V-HMN gains 0.84 ± 0.29 corruption-accuracy points over its scaffold and 0.22 ± 0.20 over the exemplar head. The scaffold ordering reverses at 150 and 500; the normalized log-label-area difference is -0.13 ± 0.23 points. ResNet-18 has lower corruption error at all three budgets. Extending the lowest-budget trajectory from 3,600 to 35,200 updates changes the scaffold gap to 0.42 ± 0.57 points (Figure 10). The fractional-data comparison establishes a broad reported advantage at its fitting fractions; the controlled study identifies a budget-dependent gain and its schedule sensitivity.

5.4 Retrieval transparency

V-HMN exposes the representations used by its memory reads. Figure 3 illustrates local/global nearest-prototype correspondences; Appendix G adds recorded forward reads, source identities, weights and output interventions.

Inspection and functional influence. Local panels show source regions; global panels show whole-image neighbors. Figure 14 separately records one local query within the measured two-branch intervention at block three. Across three Tiny models on the same 256 corruption-paired clean images, replacing read values from both block-three banks with norm-matched random directions, while retaining their keys, changes accuracy by -5.47 points (conditional 95% source-bootstrap interval $[-8.46, -2.73]$); value permutation and uniform means each give -0.39 with intervals spanning zero. Figure 16 retains all recorded conditions and the label-informed value controls.

Accounting for a specified score. At block three, signed path contributions reconstruct the uniform-to-selective difference in $h = z_y - z_{c^*}$, with c^* fixed at the native endpoint. The groups satisfy the

accounting identity within recorded numerical tolerance (Appendix G). This fixed-competitor score uses a different reference from the bypass-defined prediction categories; it need not equal a change in the max-over-other margin.

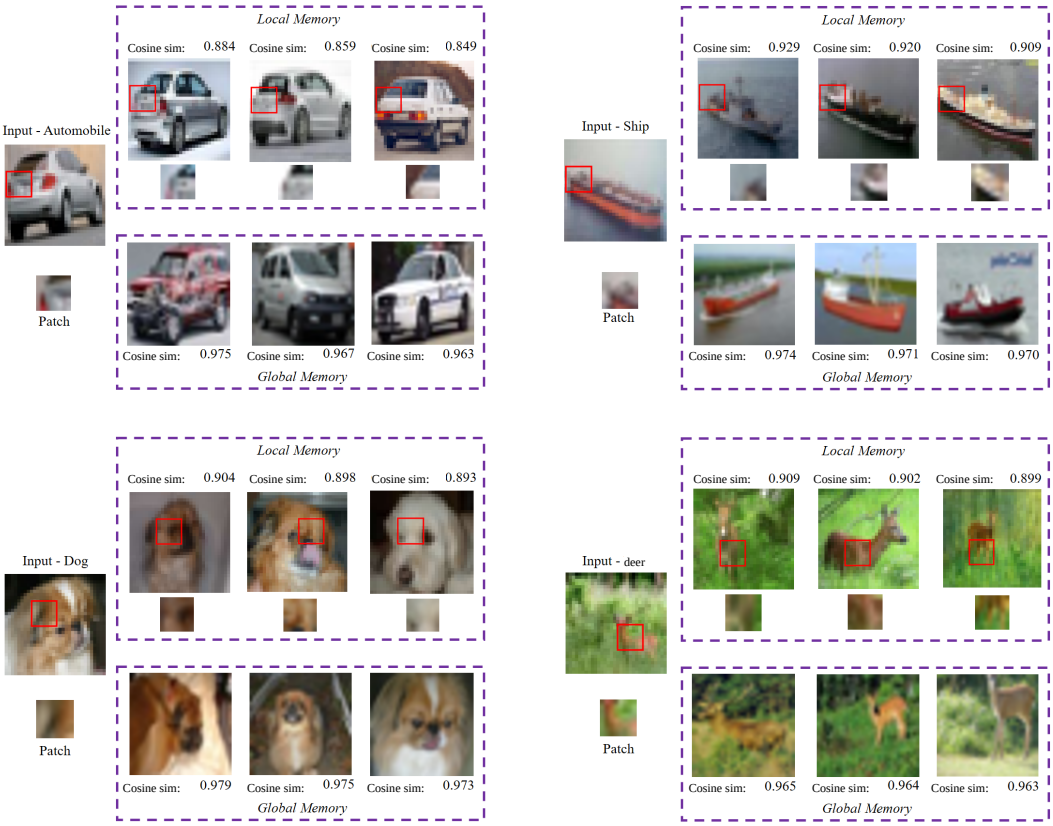


Figure 3: Local and global memory examples on CIFAR-10. Attention-pooling weights select the red query region; the visualization displays 8×8 patches. Local/global neighbors come from a separate training-query index, with cosine-similarity scores. Automobile, ship, dog and deer panels retain all original examples. Appendix I.4 specifies this illustration; Appendix G records actual forward reads and intervention effects.

6 Discussion and Conclusion

V-HMN integrates local and global training representations with feature refinement. It improves corruption recognition over the evaluated AiT-S adaptation and, on ImageNet-C, ResNet-50 V1. The smaller-benchmark trained controls delimit its full-budget advantage. Within fitted models, selective weighting contributes beyond bank means; its branch allocation differs from bypass dependence. The fractional-data study reports the highest mean among the listed models at both tested comparative fractions on three datasets. Budget-controlled scaffold comparisons locate budget and schedule boundaries. Finally, actual reads and functional interventions expose stored information and account for a defined output-score difference.

These results concern specified corruptions and masks, not arbitrary natural shifts or adversarial attacks. Fixed-class fitting fractions differ from selection-inclusive budgets. Three repetitions limit training-population inference; ImageNet compares one fitted V-HMN with a differently trained reference. Class-guided retention uses labels, and memory can reinforce misleading queries. Coverage, drift and residual geometry alone do not establish semantic recovery.

AI USE STATEMENT

The authors developed V-HMN and led preparation of this submission. Generative AI assisted with research refinement, literature work, implementation, mathematical and result analysis, figures and tables, and language editing, including sentence structure, proofreading, bulk editing, and formatting. The authors take responsibility for the submission's contents.

REPRODUCIBILITY STATEMENT

Appendix A specifies training, data, memory writes, evaluation, and model selection; Appendix F defines the correction measurements. Appendices H and I document the ImageNet and additional small-image protocols. Tables distinguish reported aggregates from statistics computed from paired predictions. The accompanying manuscript source contains the figures, tables, bibliography, and style files needed to compile the paper. Experimental code, per-example predictions, and model checkpoints are retained separately; the manuscript source package does not itself provide those experimental artifacts.

ETHICS STATEMENT

This work uses public image-recognition datasets. Source-linked memory traces may expose training examples, so sharing them requires respecting the datasets' permissions and any identifiable content. The benchmark results do not establish safety-critical deployment reliability.

REFERENCES

Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems*, 2019. URL https://arxiv.org/abs/1806.10574.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.11929.

Robert Geirhos, Priyank Jaini, Austin Stone, Sourabh Medapati, Xi Yi, George Toderici, Abhijit Ogale, and Jonathon Shlens. Towards flexible perception with visual memory. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 19056–19081, 2025. URL https://proceedings.mlr.press/v267/geirhos25a.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

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. URL https://arxiv.org/abs/1512.03385.

Dan Hendrycks. Clarification of Tiny ImageNet-C source images and clean counterparts. Author comment on the official robustness repository, issue 39, January 30, 2021. URL https://github.com/hendrycks/robustness/issues/39#issuecomment-770255986. Includes the author-provided clean bounding-box crop archive; accessed September 19, 2026.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1903.12261.

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1912.02781.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Hu_Squeeze-and-Excitation_Networks_CVPR_2018_paper.html.

Ahmet Iscen, Alireza Fathi, and Cordelia Schmid. Improving image recognition by retrieving from web-scale image-text data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19295–19304, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Iscen_Improving_Image_Recognition_by_Retrieving_From_Web-Scale_Image-Text_Data_CVPR_2023_paper.html.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 2019. URL https://arxiv.org/abs/1902.10186.

Dahyun Kang, Ahmet Iscen, Eunchan Jo, Sua Choi, Minsu Cho, and Cordelia Schmid. Memory-modular classification: Learning to generalize with memory replacement. *Transactions on Machine Learning Research*, 2025. URL https://arxiv.org/abs/2504.06021.

Adam Kortylewski, Ju He, Qing Liu, and Alan Yuille. Compositional convolutional neural networks: A deep architecture with innate robustness to partial occlusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. URL https://arxiv.org/abs/2003.04490.

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. URL https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper.html.

Yuto Matsuo, Yoshihiro Fukuhara, Yuki M. Asano, Rintaro Yanagi, Hirokatsu Kataoka, and Akio Nakamura. MoireMix: A formula-based data augmentation for improving image classification robustness. *arXiv preprint arXiv:2603.25109*, 2026. URL https://arxiv.org/abs/2603.25109.

Emin Orhan. A simple cache model for image recognition. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://papers.nips.cc/paper/8214-a-simple-cache-model-for-image-recognition.pdf.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2008.02217.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1703.05175.

Yuwei Sun, Hideya Ochiai, Zhirong Wu, Stephen Lin, and Ryota Kanai. Associative transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2309.12862v4.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 3319–3328. PMLR, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2007.00644.

Ilya O. Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. MLP-Mixer: An all-MLP architecture for vision. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://papers.nips.cc/paper/2021/hash/cba0a4ee5ccd02fda0fe3f9a3e7b89fe-Abstract.html.

Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. MetaFormer is actually what you need for vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2111.11418.

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, 2024. URL https://proceedings.mlr.press/v235/zhu24f.html.

A TRAINING, DATA, AND EVALUATION PROTOCOLS

Table 4: Experiment-specific recipe overview. The detailed split, optimizer choices for external models, and evaluation settings follow below. A reported training schedule does not identify the selected epoch. ImageNet's serialized epoch is 302; its original selection rule is unavailable.

Experiment	Optimization	Training views	Memory and selection
Budget-controlled CIFAR-100 / Tiny	200 epochs; Adam, LR .001, WD 5×10^{-5}; batch 256; 5-epoch warmup, cosine	Random crop and horizontal flip	Detached writes after backward/step; held-out training labels select clean Top-1, then NLL
Fractional CIFAR / Fashion-MNIST	400 epochs; Adam, LR .001, WD 5×10^{-5}; batch 256; 5-epoch warmup, cosine	Crop/flip, AutoAugment, MixUp, CutMix	Aggregate-specific bank state and selection unavailable; separate retained trajectories detailed in Appendix I
CIFAR-10-C models	400 epochs; Adam, LR .001, WD 5×10^{-5}; batch 256; 5-epoch warmup, cosine	Random crop and horizontal flip	Forward-pass writes; clean-test selection at epochs 391/392/384
ImageNet V-HMN	310-epoch documented recipe; AdamW, LR .001, WD .03; batch 1,000; warmup and cosine; clip .1	RandAugment, MixUp, CutMix, random erasing	Forward-pass write implementation; supplied populated banks frozen for evaluation; original selection unavailable
ImageNet ResNet-50 V1	90 epochs; SGD, LR .1, momentum .9, WD 10^{-4}; batch 256; LR divided by 10 every 30 epochs	Bilinear random resized crop, horizontal flip	No sample bank; evaluated public V1 weights

Architecture and splits. The small-image V-HMN has six blocks, token/retrieval width 256, 4×4 stride-4 RGB patches, learned positional embeddings, and one training-time refinement. A normalized 3×3 token neighborhood projects from $9 \cdot 256$ to 256; the global branch projects the mean of normalized tokens. Each branch concatenates initial and updated queries ($512 \to 256$). Their sum passes through GELU and dropout 0.1 into the residual path; the second pre-normalized residual MLP has ratio two. Final attention pooling and a linear classifier produce logits. The initial unconstrained branch coefficient is $\beta = .2$.

Dataset	Labels/class	Fitting	Selection	Tokens	Slots local/global
Tiny ImageNet, 64^2	500	90,000	10,000	256	2,400/1,000
CIFAR-100, 32^2	50	4,500	500	64	2,500/1,000
CIFAR-100, 32^2	150	13,500	1,500	64	2,500/1,000
CIFAR-100, 32^2	500	45,000	5,000	64	2,500/1,000

Slots are per block. Tiny allocates 12 local and five global slots per class; CIFAR allocates 25 and ten. Class-stratified source rankings split official training images into disjoint 90/10 fitting/selection pools. CIFAR budgets use nested classwise prefixes within each pool; fitting sources never become selection sources at another budget. Counts include all labels used for model selection. No external pretraining, distillation, unlabeled remainder, or evaluation sources enter fitting or memory.

Training and checkpoint selection. Seeds $s \in \{2026, 2027, 2028\}$ determine splits and initialization. Training stochasticity, data ordering, and crop/flip streams use $s + 10000$, $s + 20000$, and $s + 30000$; view draws additionally depend on epoch and source ID. Controlled models share these streams and matching initial tensors. Primary training uses cross-entropy, Adam at 10^{-3}, weight decay $5 \cdot 10^{-5}$, batch 256 without dropping the last batch, five warm-up epochs, and cosine decay to

$5 \cdot 10^{-6}$ over 200 epochs. This schedule was fixed from development runtime measurements before the primary comparisons. Selection maximizes clean held-out Top-1, then minimizes NLL, then retains the earliest epoch.

Zero padding is eight pixels for Tiny and four for CIFAR, followed by a uniform native-size crop and horizontal flip with probability $1/2$. There is no primary corruption augmentation, erasing, MixUp, or CutMix. Tiny RGB normalization uses mean $(.485, .456, .406)$ and SD $(.229, .224, .225)$; CIFAR uses $(.5071, .4867, .4408)$ and $(.2675, .2565, .2761)$. The 50-label sensitivity completes 35,200 updates, matching full-budget training, with an epoch cap of 1,956. Selected checkpoints can precede that endpoint (Table 17).

Comparator recipes. The scaffold retains query projections, concatenation/fusion, normalization, residual paths, pooling, and classifier. Its two updates are identities throughout training and inference. The uniform-trained control retains the same bank sizes, write schedule, learned coefficients, and branch computations, but replaces each read by its own bank's valid-entry mean from initialization. It inherits the paired V-HMN recipe and clean-selection rule without a new hyperparameter search.

ResNet-18 uses a 3×3 stride-1 stem without max-pooling. Two prespecified 20-epoch trials per dataset/budget/seed compare SGD (momentum .9, learning rate .1, weight decay $5 \cdot 10^{-4}$) and Adam (learning rate 10^{-3}, weight decay $5 \cdot 10^{-5}$). Clean selection chooses the optimizer, breaking ties by NLL and then candidate order. The selected recipe is trained anew for 200 epochs.

The exemplar comparator stores normalized pooled scaffold features and labels from up to 5,000 distinct class-balanced fitting sources: 25/class on Tiny, 50/class on CIFAR, or all 45/class at the lowest budget. With normalized query z and stored (z_i, y_i),

$$p_{\mathrm{ex}}(c \mid z) = \sum_i \mathrm{softmax}_i(z^\top z_i/\tau)\, \mathbf{1}[y_i = c], \qquad p = (1-\eta)p_{\mathrm{scaffold}} + \eta p_{\mathrm{ex}}.$$

Clean selection chooses $\tau \in \{.05, .1, .2\}$ and $\eta \in \{0, .25, .5, .75, 1\}$ by Top-1, then NLL, then smaller mixture/temperature. Source ranking uses seed 2026. The store is frozen before evaluation, with no additional model training or labels.

Associative Transformer comparator. We use AiT-S dimensions from Sun et al. (2025): two blocks, width 768, MLP width 3,072, twelve self-attention heads, and eight bottleneck heads. Each block has 32 priors of dimension 32 and selects 512 batch-pooled tokens per prior and bottleneck head. Patch size four is retained at both image resolutions, making the Tiny configuration a declared spatial adaptation. We label this comparator "AiT-S (adapted)": it is a frozen-memory adaptation of author revision `107ae934`, rather than a reproduction of the paper's numerical results.

The implementation retains that release's learned prior-query projection, residual before proposal layer normalization, interleaved head coordinates, rowwise ℓ_2 normalization across the feature coordinates of the moving-average priors, and flattened patch classifier without active positional or class-token embeddings. The paper indexes its update denominator over priors and does not display the release's learned query projection or pre-normalization prior residual; we retain the release's operations. A single dot-product Hopfield update at scale one and both source residual paths reconstruct the tokens. Training combines cross-entropy with coefficient-.01 importance/load balancing and uses a .1 new-state moving-average weight. Classification reads the proposed detached state; commitment occurs after backward, with a cloned old state preventing aliasing of the auxiliary-loss graph. Classification gradients reach token processing, memory-to-token projection, and the head; auxiliary gradients train bottleneck queries and keys. Detached writes do not add a classification gradient through the bottleneck value/output projections. Evaluation reads frozen priors and skips the bottleneck update. This resolves persistent test-state mutation in the release, but does not reproduce its unspecified alternative test-time bottleneck-ablation pathway.

Two 20-epoch development trials per dataset compare AdamW learning rates 10^{-4} and 10^{-5}, with weight decay .01, five warm-up epochs, and cosine floor 10^{-6}. Clean held-out accuracy selects the configuration, breaking ties by NLL and then trial order. Final training uses the same 200 epochs, physical batch 256, crop/flip views, normalization, source splits, and seed conventions as the campaign. The paper's CIFAR settings instead use 100 epochs and batch 512. Batch pooling makes the physical writing batch part of the method; gradient accumulation is not used to claim equivalence to batch 512. This external comparison measures the declared models and recipes, not a controlled

change to V-HMN's architecture. For seeds 2026, 2027, and 2028 respectively, the evaluated AiT-S checkpoints selected by clean Top-1 (ties: lower NLL, then earlier epoch) have epoch/update counts Tiny ImageNet: 30/10,560, 179/63,008, 33/11,616 (completed 200 epochs/70,400 updates); CIFAR-100: 199/35,024, 190/33,440, 177/31,152 (completed 200 epochs/35,200 updates).

The selected optimizer and exemplar parameters, including run-specific choices, are given below; candidate grids alone do not specify the evaluated models.

Table 5: Actual development-selected comparator settings, in seed order 2026/2027/2028. The selected SGD recipe uses peak learning rate 0.1, weight decay 5×10^{-4}, and momentum 0.9; Adam uses peak learning rate 10^{-3} and weight decay 5×10^{-5}. AiT-S is the frozen-memory adaptation of the author release; its AdamW recipe uses peak learning rate 10^{-4} and weight decay 10^{-2}. Each 200-epoch trajectory uses physical batch 256, five-epoch linear warm-up to the peak rate, then cosine decay. The final learning-rate floor is 5×10^{-6} for ResNet-18 and 10^{-6} for AiT-S. Selected epochs identify the evaluated checkpoints, not the completed trajectory endpoints.

Dataset	Model	Labels/class	Selected optimizer	Selected epoch
CIFAR-100	AiT-S adaptation	500	AdamW/AdamW/AdamW	199/190/177
CIFAR-100	ResNet-18	50	Adam/Adam/Adam	168/155/79
CIFAR-100	ResNet-18	150	SGD/Adam/Adam	178/194/177
CIFAR-100	ResNet-18	500	SGD/SGD/SGD	181/172/183
Tiny ImageNet	AiT-S adaptation	500	AdamW/AdamW/AdamW	30/179/33
Tiny ImageNet	ResNet-18	500	SGD/SGD/SGD	189/189/199

Table 6: Selected exemplar temperatures τ and mixture weights η, in seed order 2026/2027/2028. Predictions are $(1-\eta)p_{\text{scaffold}} + \eta p_{\text{exemplar},\tau}$. The source budget counts distinct permitted fitting images, not memory views or selection images. These retrieval temperatures are separate from post-hoc classifier-output temperature scaling.

Dataset	Labels/class	Sources	Seed 2026 (τ, η)	Seed 2027	Seed 2028
CIFAR-100	50	4,500	$(0.1, 0.5)$	$(0.1, 0.75)$	$(0.1, 0.75)$
CIFAR-100	150	5,000	$(0.1, 0.75)$	$(0.1, 0.5)$	$(0.05, 0.5)$
CIFAR-100	500	5,000	$(0.05, 0.5)$	$(0.1, 0.75)$	$(0.05, 0.25)$
Tiny ImageNet	500	5,000	$(0.05, 0.5)$	$(0.1, 0.75)$	$(0.1, 0.25)$

Memory lifecycle and resources. Loss forwards read the pre-update bank and capture detached initial queries from the same view. After backward and the optimizer step, each image writes up to five largest-norm local queries and one global query per block to class-balanced rings, in batch order. No extra forward or clean refresh occurs. Labels choose write queues, never test reads. Checkpoints preserve weights, values, validity, pointers, source/view/token IDs, write metadata, optimizer and RNG state, configuration, and ordered class mappings. Each run uses one H100 80 GB with deterministic FP32 operations and TF32 disabled; independent runs may execute concurrently. Evaluation freezes all weights, banks, and normalization state.

Corruption cohorts. Released CIFAR-100-C and corrected Tiny ImageNet-C images supply the 15 primary corruptions at all five severities, equally weighted. The corrected Tiny release (Zenodo 8206060) includes frost and the verified 200-class mapping. No extra crop is applied to released corruptions. CIFAR pairs with its 10,000 official test images; Tiny's 10,000 canonical validation images supply its primary clean and mask results. Tiny corruption-paired clean measurements instead use the benchmark author's separate 10,000 bounding-box crops (Hendrycks, 2021). Their 256×256 RGB images are resized to 64×64 by PIL bilinear interpolation before normalization, without another JPEG encoding. Same-name standard Tiny test images are not these counterparts. All paired mechanism measurements use these supported source correspondences.

Masks and uncertainty. Masks operate on deterministic raw RGB before normalization, with fitting-image channel means as primary fill. An 8×8 image grid removes $\lfloor 64\rho + 1/2 \rfloor$ cells from a source-keyed ranking, nested across fractions. A square minimizes integer-side area error, ties

toward the smaller side, and samples an unclipped valid top-left. Source-keyed placement draws are shared across fractions and models; squares are not assumed nested. Seed 2026 and realizations 0–2 define the streams. Fractions are $0, .1, .25, .5, .75$, with actual pixel areas recorded. Three realizations are averaged per source, and the clean prediction supplies zero removal. Separate black-fill and centered-square sensitivities at .25/.5 do not enter the primary AUC. Source bootstrap intervals use 2,000 replicates and retain every derivative of each resampled image. Training uncertainty is the sample SD of three independent repetitions. One percentage point is the provisional meaningful-effect reference, not a criterion for successful research.

AugMix. The separate CIFAR factorial uses width three, random chain depth one–three, severity three, and Dirichlet/Beta parameter one. Operations are autocontrast, equalize, posterize, rotate, solarize, shear-x/y, and translate-x/y, with reference operator ranges. Both models receive base-view cross-entropy plus 12 times Jensen–Shannon divergence across that view and two independently keyed AugMix views. Only the unmixed base forward supplies writes. This is a distinct augmentation/consistency recipe, never pooled with the primary condition.

A.1 Memory lifecycle and frozen evaluation

Figure 4 complements the architecture in Figure 1 by making the read/write timing explicit. In the primary smaller-benchmark implementation, the forward pass reads the existing bank, while detached source queries are committed after backward and the optimizer step. Evaluation performs reads only. The read strength in this diagram is an intervention on the native correction, rather than a change to the learned update coefficient.

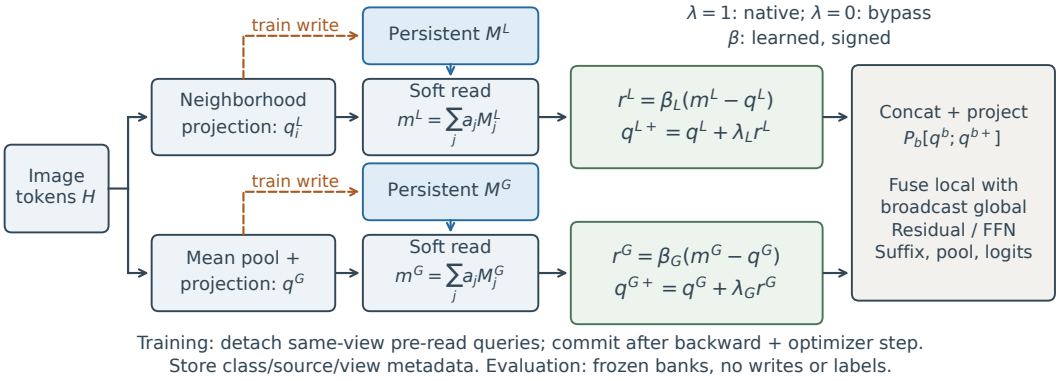
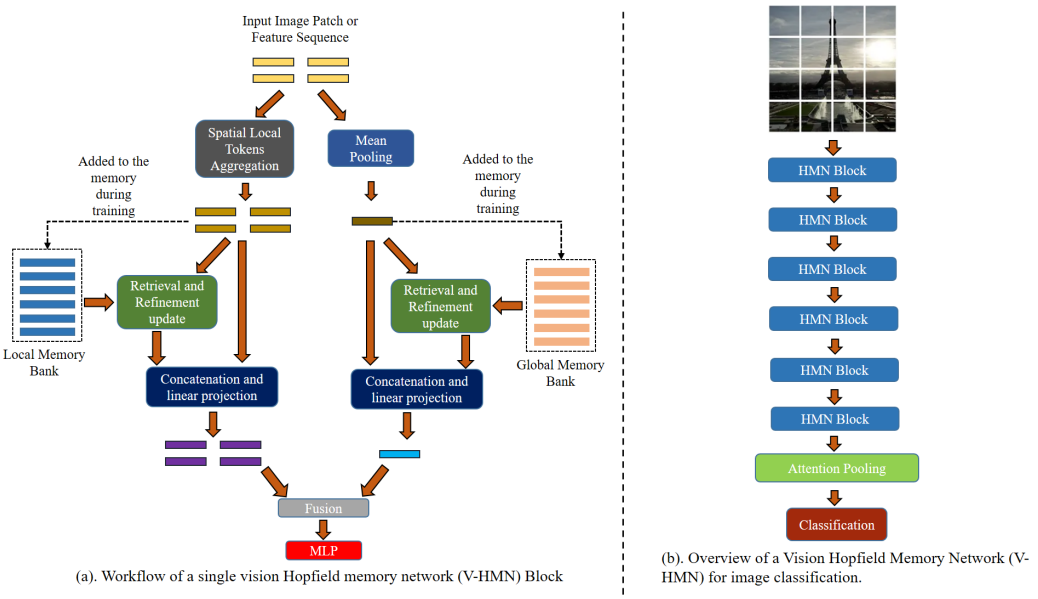


Figure 4: Local and global queries read training-derived representations before branch fusion and classification. Primary smaller-benchmark training commits detached queries and source metadata after backward and the optimizer step; evaluation freezes parameters and banks. Read weights and corrections expose the memory contribution within the larger computation.

The resulting evaluation interventions alter the read computation while keeping learned parameters, bank values, validity and source metadata fixed. Historical forward-pass writing is documented separately for the corresponding experiments.

A.2 Complete architecture overview

Figure 5 retains both architecture panels at appendix scale. The left panel connects spatial neighborhood queries and pooled global queries to separate memories and the feature update; the right panel shows the six-block configuration. This is the smaller-image layout, not a claim that the staged ImageNet model has six blocks. The exact normalization, residual computation, bank dimensions and writing regimes are specified in Sections 3 and A.



(a). Workflow of a single vision Hopfield memory network (V-HMN) Block

(b). Overview of a Vision Hopfield Memory Network (V-HMN) for image classification.

Figure 5: Complete V-HMN overview: local/global memory retrieval and refinement, branch concatenation/projection and fusion, the MLP, and the stacked-block classifier. The schematic keeps the training write arrows distinct from inference reads; normalization and residual paths are given in the method. All content of the original two-panel diagram is retained.

B Complete benchmark summaries

Table 1 gives the full-budget comparisons, including the fixed-checkpoint read bypass. This section adds reduced label budgets, probability-quality endpoints, the separate AugMix recipe, and the extended lowest-budget schedule. Values are means and sample standard deviations across three training repetitions. Individual predictions, repetitions, and corruption–severity scores remain in the machine-readable scientific artifacts; conditions are not treated as independent training runs.

Table 7: Reduced-budget primary endpoints, mean ± sample SD over three independent training repetitions. All metrics are percentages. Corruption error averages all 15 types and five severities; mask AUC integrates over nominal removal and divides by .75. Canonical and corruption-paired CIFAR-100 clean endpoints are identical in every repetition; the common clean value is shown once. Primary clean-trained full-budget comparisons appear in Table 1. Labels include fitting and selection. Lower error and higher accuracy/AUC are better. Ranking compares trained models within each label budget; bypass is unranked. Bold marks numerical best values (ties retained), not significance.

Model	Clean ↑	C error ↓	Grid AUC ↑	Square AUC ↑
CIFAR-100; 5,000 total labels; crop/flip, 200 epochs				
V-HMN	23.97 ± 0.45	79.95 ± 0.29	9.14 ± 0.90	8.63 ± 0.11
V-HMN, reads bypassed	22.41 ± 0.51	81.10 ± 0.30	8.38 ± 1.02	8.06 ± 0.32
Trained scaffold	23.01 ± 0.49	80.79 ± 0.43	8.61 ± 0.67	8.37 ± 0.45
ResNet-18	**33.58** ± 0.65	**76.45** ± 0.76	**15.50** ± 0.48	**13.76** ± 0.62
Scaffold + exemplar	23.53 ± 0.21	80.17 ± 0.10	8.94 ± 0.70	8.62 ± 0.55
CIFAR-100; 15,000 total labels; crop/flip, 200 epochs				
V-HMN	42.67 ± 0.10	65.58 ± 0.28	18.97 ± 0.30	15.19 ± 0.10
V-HMN, reads bypassed	36.36 ± 0.93	70.77 ± 0.70	14.61 ± 0.17	12.35 ± 0.05
Trained scaffold	43.29 ± 0.73	65.10 ± 0.57	19.28 ± 0.60	14.98 ± 0.38
ResNet-18	**58.09** ± 2.06	**60.57** ± 1.46	**22.97** ± 0.34	**21.49** ± 0.69
Scaffold + exemplar	43.56 ± 0.71	64.75 ± 0.43	19.59 ± 0.65	15.16 ± 0.36

Table 8: Complete augmentation and schedule endpoints, mean ± sample SD over three independent training repetitions. All metrics are percentages. Corruption error averages all 15 types and five severities; mask AUC integrates over nominal removal and divides by .75. Canonical and corruption-paired CIFAR-100 clean endpoints are identical in every repetition; the common clean value is shown once. Primary clean-trained full-budget comparisons appear in Table 1. Labels include fitting and selection. Lower error and higher accuracy/AUC are better. Ranking compares trained models within each recipe/schedule; bypass is unranked. Bold marks numerical best values (ties retained), not significance.

Model	Clean ↑	C error ↓	Grid AUC ↑	Square AUC ↑
CIFAR-100; 50,000 total labels; AugMix				
V-HMN	67.75 ± 1.14	42.52 ± 0.87	$\mathbf{32.26} \pm 0.30$	$\mathbf{24.71} \pm 0.50$
V-HMN, reads bypassed	62.44 ± 1.04	48.63 ± 1.07	30.25 ± 0.47	21.71 ± 0.33
Trained scaffold	$\mathbf{68.10} \pm 0.30$	$\mathbf{42.30} \pm 0.24$	32.14 ± 0.41	24.26 ± 0.26
CIFAR-100; 5,000 total labels; 35,200 updates				
V-HMN	$\mathbf{25.36} \pm 0.44$	$\mathbf{78.71} \pm 0.37$	$\mathbf{11.33} \pm 0.79$	$\mathbf{8.97} \pm 0.22$
V-HMN, reads bypassed	12.22 ± 3.51	89.81 ± 2.79	4.86 ± 1.35	4.63 ± 0.89
Trained scaffold	24.74 ± 0.45	79.13 ± 0.40	11.19 ± 0.38	8.72 ± 0.14

Model	Probabilities	Clean ↓	Corruption ↓	Grid AUC ↓	Square AUC ↓
Tiny ImageNet					
V-HMN	Raw	3.67 ± 0.014	9.64 ± 0.040	9.97 ± 0.176	9.22 ± 0.274
	T-scaled	2.14 ± 0.004	4.61 ± 0.026	4.77 ± 0.050	4.52 ± 0.086
Trained scaffold	Raw	3.51 ± 0.025	9.56 ± 0.039	10.31 ± 0.204	9.10 ± 0.314
	T-scaled	2.12 ± 0.011	4.66 ± 0.030	4.92 ± 0.037	4.54 ± 0.075
Uniform-trained	Raw	3.47 ± 0.035	9.20 ± 0.250	8.96 ± 0.033	8.33 ± 0.133
	T-scaled	2.13 ± 0.004	4.63 ± 0.073	$\mathbf{4.58} \pm 0.030$	4.38 ± 0.026
ResNet-18	Raw	$\mathbf{1.65} \pm 0.002$	$\mathbf{4.12} \pm 0.013$	4.70 ± 0.063	4.39 ± 0.109
	T-scaled	$\mathbf{1.65} \pm 0.002$	$\mathbf{4.13} \pm 0.015$	4.71 ± 0.071	4.40 ± 0.109
Scaffold + exemplar	Raw	2.39 ± 0.187	4.49 ± 0.260	$\mathbf{4.62} \pm 0.257$	$\mathbf{4.33} \pm 0.199$
	T-scaled	2.28 ± 0.059	4.51 ± 0.068	4.66 ± 0.017	$\mathbf{4.36} \pm 0.121$
AiT-S (adapted)	Raw	6.38 ± 4.686	14.56 ± 10.628	10.95 ± 5.674	12.88 ± 7.054
	T-scaled	3.04 ± 0.067	5.42 ± 0.082	4.93 ± 0.689	5.39 ± 0.719
CIFAR-100					
V-HMN	Raw	2.26 ± 0.025	4.04 ± 0.018	6.69 ± 0.134	10.76 ± 0.216
	T-scaled	1.36 ± 0.011	2.21 ± 0.009	$\mathbf{3.43} \pm 0.054$	4.85 ± 0.074
Trained scaffold	Raw	2.30 ± 0.029	4.02 ± 0.088	6.86 ± 0.093	10.41 ± 0.367
	T-scaled	1.38 ± 0.014	2.20 ± 0.034	3.49 ± 0.029	4.75 ± 0.097
Uniform-trained	Raw	2.31 ± 0.057	3.97 ± 0.070	6.66 ± 0.096	10.65 ± 0.816
	T-scaled	1.39 ± 0.025	$\mathbf{2.18} \pm 0.019$	$\mathbf{3.43} \pm 0.010$	4.83 ± 0.212
ResNet-18	Raw	$\mathbf{0.97} \pm 0.012$	2.39 ± 0.024	3.68 ± 0.014	$\mathbf{3.88} \pm 0.185$
	T-scaled	$\mathbf{0.96} \pm 0.013$	2.41 ± 0.030	3.74 ± 0.007	$\mathbf{3.98} \pm 0.197$
Scaffold + exemplar	Raw	1.58 ± 0.213	$\mathbf{2.33} \pm 0.075$	$\mathbf{3.50} \pm 0.274$	4.73 ± 0.824
	T-scaled	1.44 ± 0.033	2.22 ± 0.019	$\mathbf{3.43} \pm 0.012$	4.70 ± 0.166
AiT-S (adapted)	Raw	3.92 ± 0.065	7.99 ± 0.163	9.06 ± 1.098	13.46 ± 1.233
	T-scaled	1.73 ± 0.017	2.98 ± 0.036	3.44 ± 0.268	4.49 ± 0.279

Table 9: Probability quality before and after matched clean-selection temperature scaling: mean ± sample SD over three training repetitions. All entries are negative log-likelihood in nats (lower is better); corruption averages all 15 types and five severities equally. Masking entries integrate NLL over removed-area fractions $[0, 0.75]$, including the canonical clean endpoint, and divide by 0.75. Tiny ImageNet corruption uses its separate released-image cohort. Each positive temperature is fitted only to that run's original clean model-selection examples, whose labels were already used for checkpoint or exemplar-head selection; the same temperature applies to every final condition. Top-1 predictions are unchanged. The exemplar kernel and mixture are fixed. AiT-S denotes the frozen-memory adaptation of the author release; external-model recipes differ. The raw/scaled rows share a trained checkpoint and are not independent repetitions. Ranking is separate within each dataset and raw/scaled probability setting. Bold marks numerical best values (ties retained), not significance.

Effect of matched probability scaling. Mean corruption-NLL differences (V-HMN minus adapted AiT) contract from -4.920 ± 10.656 to -0.803 ± 0.079 nats on Tiny ImageNet and from -3.956 ± 0.167 to -0.771 ± 0.041 on CIFAR-100 after matched clean-selection scaling. Values are mean $\pm$ sample SD of three paired seed differences. The CIFAR square-masking difference reverses from -2.701 ± 1.448 to $+0.355 \pm 0.341$ nats, favoring adapted AiT after scaling. Against the uniform-trained control on Tiny ImageNet, the scaled corruption difference is -0.020 ± 0.049 nats; grid and square NLL-area differences are $+0.193 \pm 0.020$ and $+0.149 \pm 0.072$ nats, respectively, so the uniform-trained masking-loss advantage persists. Masking NLL areas are normalized over removal fractions $[0, 0.75]$, and lower loss is better. These changes concern output probabilities; Top-1 accuracy is unchanged.

Comparator	Clean gap g_Q	Corruption gap g_C	Gap change $g_C - g_Q$
CIFAR-100-C			
Trained scaffold	-0.22 ± 0.16	-0.36 ± 0.31	-0.14 ± 0.17
Uniform-trained	$+0.16 \pm 0.70$	-0.51 ± 0.59	-0.67 ± 0.18
Scaffold + exemplar	-0.39 ± 0.12	-0.65 ± 0.19	-0.26 ± 0.17
ResNet-18	-12.83 ± 0.23	-0.52 ± 0.36	$+12.31 \pm 0.51$
AiT-S (adapted)	$+8.40 \pm 0.33$	$+12.81 \pm 0.44$	$+4.42 \pm 0.29$
Tiny ImageNet-C			
Trained scaffold	-0.17 ± 0.66	-0.27 ± 0.37	-0.10 ± 0.36
Uniform-trained	$+0.61 \pm 0.48$	$+0.09 \pm 0.21$	-0.52 ± 0.27
Scaffold + exemplar	-0.30 ± 0.70	-0.37 ± 0.41	-0.07 ± 0.33
ResNet-18	-13.05 ± 0.60	-5.98 ± 0.27	$+7.07 \pm 0.47$
AiT-S (adapted)	$+12.53 \pm 1.76$	$+6.70 \pm 0.73$	-5.83 ± 1.04

Table 10: Clean recognition and corruption degradation across all comparators. Each gap is V-HMN accuracy minus comparator accuracy, in percentage points; entries are mean $\pm$ sample SD of three paired seed differences. Q is the matching corruption-source clean cohort, and C averages all 15 corruptions and five severities equally. Tiny ImageNet uses the separate corruption-paired clean images here, rather than the canonical clean images used for clean classification and masking. A positive $g_C - g_Q$ means a smaller absolute accuracy drop for V-HMN; it does not imply higher corruption accuracy. This is descriptive, not a causal clean-accuracy adjustment. SDs are computed after subtraction within each seed, retaining covariance. ResNet-18 and AiT-S use their declared external-model recipes; AiT-S is the frozen-memory adaptation of the author release.

Table 11: Tiny ImageNet clean-reference cohorts for the six full-budget prediction systems. Canonical clean images anchor masking AUC; corruption-paired clean images anchor clean-to-corruption contrasts. Mean $\pm$ sample SD across three fitted repetitions, in percent; the final column gives within-repetition differences in percentage points before aggregation. The paired-clean preparation follows the released corruption source geometry, rather than treating these two clean scores as interchangeable. Ranking is separate for each clean cohort; cohort differences are descriptive. Bold marks numerical best values (ties retained), not significance.

Model	Canonical (%) ↑	Paired (%) ↑	Difference (pp)
V-HMN	50.68 ± 0.09	31.25 ± 0.85	19.43 ± 0.90
Scaffold	51.48 ± 0.35	31.42 ± 0.62	20.06 ± 0.65
Uniform-trained	51.20 ± 0.12	30.64 ± 0.76	20.57 ± 0.83
ResNet-18	$\mathbf{63.96} \pm 0.04$	$\mathbf{44.30} \pm 0.73$	19.66 ± 0.73
Scaffold + exemplar	51.64 ± 0.34	31.55 ± 0.68	20.09 ± 0.55
AiT-S, frozen-memory adaptation	33.63 ± 0.38	18.72 ± 1.10	14.91 ± 0.72

B.1 Corruption heterogeneity

The heatmaps retain all 15 primary corruption types and five severities. Each cell averages the three selected checkpoints of the named model. Rounding in the display does not enter the equally weighted primary aggregate. CIFAR-100 uses the full label budget here; its lower-budget aggregates and learning curves use the same suite.

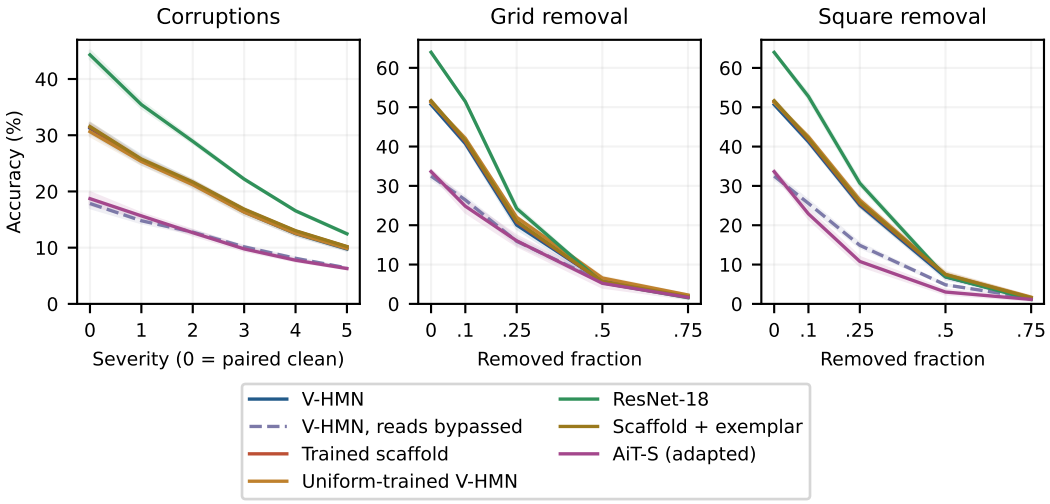


Figure 6: Complete Tiny ImageNet curves; full-budget CIFAR-100 continues below. Lines and shading show means and sample SDs across three paired runs. Dashed curves bypass reads in the same V-HMN checkpoints; uniform-trained uses nonselective training reads. Severity zero uses corruption-paired clean images; masks use canonical clean images.

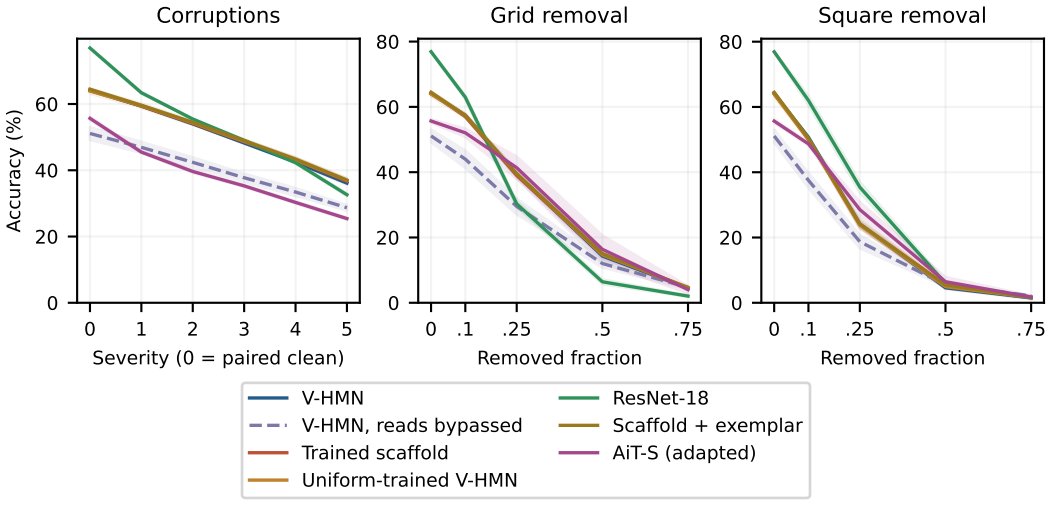


Figure 6 (continued): Full-budget CIFAR-100, with the same read conditions and uncertainty convention.

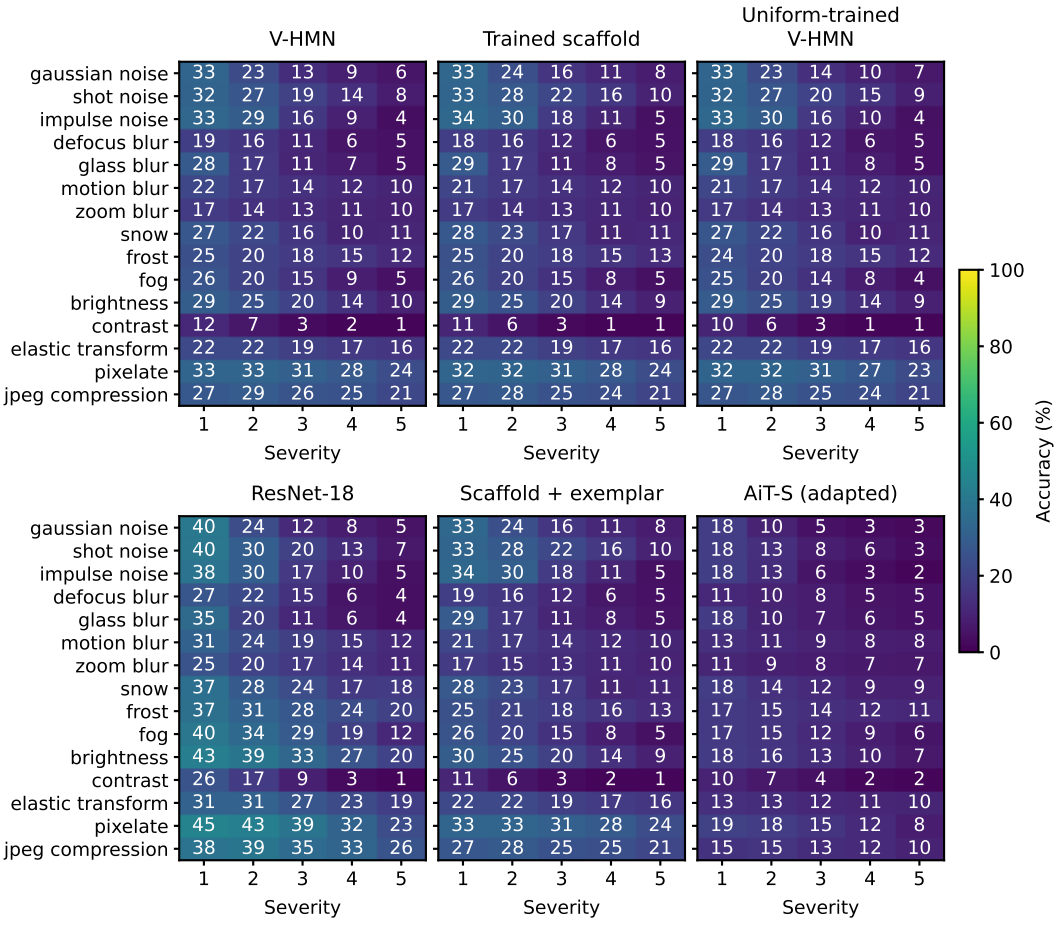


Figure 7: Tiny ImageNet-C accuracy (%) for every primary corruption and severity. The corrected release includes frost.

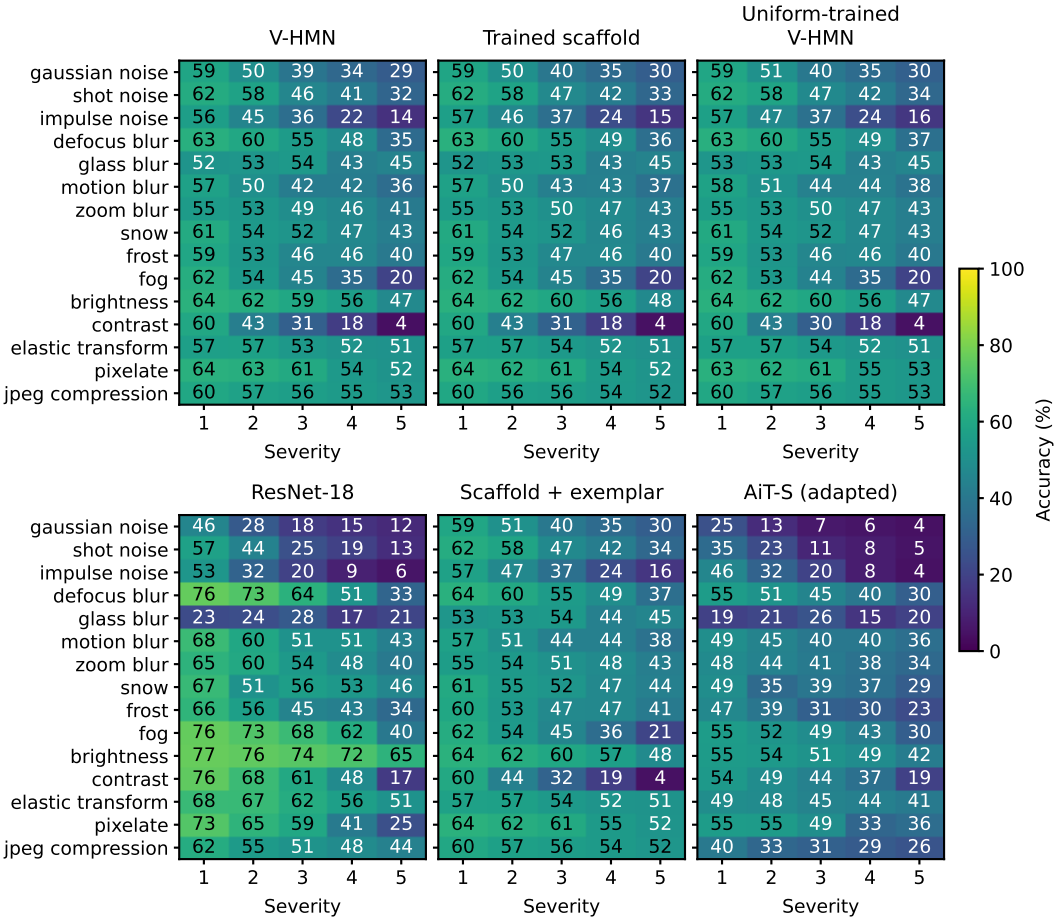


Figure 8: CIFAR-100-C accuracy (%) at 500 labels per class.

C MEMORY CONTROLS AND STATISTICAL INTERPRETATION

Let F, B, M, and S denote full V-HMN, its read bypass, its uniform-mean read intervention, and the independently trained scaffold. For each seed and endpoint, $F - S = (F - B) + (B - S)$ and $F - B = (F - M) + (M - B)$. We compute every contrast before its sample SD. In particular, $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$; adding component standard deviations would discard their dependence. The $B - S$ contrast compares different training histories under bypass, rather than isolating a pure regularization effect.

The uniform-mean intervention retains the parameters and banks learned with selective retrieval. Equations 2–3 explain why it preserves more of the fitted computation than bypass. The uniform-trained control instead learns all its parameters and bank entries under nonselective reads. These are different comparisons.

Table 12: V-HMN minus the uniform-trained control, paired within training seed before computing mean and sample SD. Differences are percentage points in the named endpoint. Negative corruption-error differences favor selective V-HMN; positive accuracy/AUC differences favor it.

Dataset	Clean	C error	Grid AUC	Square AUC
Tiny ImageNet	-0.53 ± 0.06	-0.09 ± 0.21	-1.08 ± 0.57	-0.83 ± 0.36
CIFAR-100	$+0.16 \pm 0.70$	$+0.51 \pm 0.59$	-0.19 ± 0.38	$+0.05 \pm 0.81$

Table 13: Tiny ImageNet components of the fixed-checkpoint read effect: full retrieval F, uniform valid-entry mean M, and bypass B. Entries are paired mean $\pm$ sample SD over three training repetitions, in percentage points of the named metric. Negative corruption-error differences and positive accuracy/AUC differences favor the first computation. These interventions keep the selectively trained weights and banks fixed; they are distinct from the uniform-trained model. CIFAR-100 crop/flip and AugMix components appear in Table 19, which uses corruption accuracy rather than error.

Contrast	Clean	Paired clean	C error	Grid AUC	Square AUC
Tiny ImageNet; crop/flip					
Full − mean	4.18 ± 0.32	4.35 ± 0.40	-1.85 ± 0.23	2.15 ± 0.50	2.09 ± 0.47
Mean − bypass	14.01 ± 0.68	9.04 ± 0.43	-4.84 ± 0.50	2.61 ± 0.62	5.13 ± 0.59

The full-minus-frozen-mean accuracy gap also changes relative to matching clean inputs. Its corruption-minus-paired-clean difference is -2.50 ± 0.16 points on Tiny and -0.14 ± 0.43 on full-budget CIFAR-100. Each difference is formed within a training repetition before computing its sample SD; Tiny uses the corruption author's matching clean crops. Positive read effects under corruption therefore coexist with larger mean effects on clean inputs. This descriptive comparison does not causally adjust for clean recognition, and the small CIFAR difference does not establish equivalence.

Centered-retrieval development pilot. Centered retrieval uses $q^+ = q + \beta[m(q) - \mu]$, with μ the valid-entry bank mean, in place of the original update. Its correction vanishes under uniform weights; empty banks preserve the query. For fixed nonempty banks and one enabled read, setting $A_c = A - \beta B$ and $b_c = b + \beta B\mu$ in the concatenation projection preserves the original output. This correspondence does not preserve training with changing banks or the fixed-parameter bypass intervention. We compared original and centered retrieval with one paired CIFAR-100 seed (912026) over 50 epochs, using 45,000 fitting and 5,000 held-out training images. Original/centered clean accuracy was $63.36/63.02\%$, and mean development-corruption accuracy was $47.96/48.31\%$. The centered-minus-original differences were -0.34 and $+0.35$ percentage points, respectively. The corruption mean equally weights speckle noise, Gaussian blur, spatter, and saturation at severities 1, 3, and 5. Clean selection chose epochs 48/48 (8,448/8,448 updates); the evaluated checkpoints need not be the trajectory endpoints. The variant did not meet the declared development criteria. The fixed thresholds were at least one point of development-corruption improvement and at most 0.3 percentage points of clean-accuracy decrease. This single-seed pilot supplies no training-run

uncertainty estimate or final benchmark evidence; official corruption-test images were not used for this decision.

C.1 Retrieval selectivity and a bounded attention-scale follow-up

We distinguish $K_{\text{eff}} = 1/\sum_j \alpha_j^2$ from entropy perplexity $\exp(H)$, where $H = -\sum_j \alpha_j \log \alpha_j$ and K counts valid slots. The native cosine-logit inverse temperature is $\sqrt{256} = 16$; this is separate from the learned, unconstrained β. Keys are normalized and values are not. The clean block-three bank dispersion $[K^{-1}\sum_j \|v_j - \mu\|^2]^{1/2}$ is 25.52/3.46 for CIFAR-100 and 19.51/2.08 for Tiny ImageNet (L/G; averages over runs). Nonuniform attention and a nonzero $\beta B(m - \mu)$ establish selective computation, not beneficial decisions. Complete-cohort entropy and top-five mass are paired with rescue/harm, margins, and NLL; full perturbed-cohort query vectors were not retained, so their projected decomposition is restricted to the saved cases.

Dataset	Cohort/site	K_{eff}/K L (%)	G (%)	$R_{S/A}$ L	G
CIFAR-100	Clean/block 3	6.68 ± 0.98	35.67 ± 2.29	0.483 ± 0.032	0.408 ± 0.029
CIFAR-100	Cases/block 1	1.98 ± 0.25	20.88 ± 7.78	0.781 ± 0.002	0.504 ± 0.128
CIFAR-100	Cases/block 3	7.70 ± 0.91	35.30 ± 3.10	0.485 ± 0.031	0.412 ± 0.058
CIFAR-100	Cases/block 6	67.36 ± 5.43	40.14 ± 7.24	0.109 ± 0.019	0.381 ± 0.062
Tiny	Clean/block 3	18.14 ± 1.37	26.76 ± 3.08	0.474 ± 0.033	0.303 ± 0.024
Tiny	Cases/block 1	3.35 ± 0.57	20.45 ± 0.37	0.845 ± 0.012	0.417 ± 0.021
Tiny	Cases/block 3	17.16 ± 1.55	31.03 ± 1.68	0.486 ± 0.026	0.293 ± 0.036
Tiny	Cases/block 6	42.69 ± 15.59	82.31 ± 4.37	0.195 ± 0.082	0.088 ± 0.018

Table 14: Retrieval selectivity, mean and sample SD across three fitted runs. L/G denote local/global reads. Clean rows use all 256 recorded diagnostic images; case rows use every retained first-ID outcome case from the nine pixel-input conditions, with the same cases at each site. These outcome-selected cases are descriptive, not representative population estimates. $R_{S/A} = \mathbb{E}\|\beta B[m(q) - \mu]\|/\mathbb{E}\|\beta B[\mu - q]\|$ compares selective and affine correction norms, not the full affine branch output. The vectors can cancel, so this is not an explained-accuracy fraction.

As a targeted development follow-up, $r_\alpha \in \{0.5, 1, 2\}$ multiplies the existing cosine logits at every read; r_α is distinct from the correction vector r. Three frozen full-budget CIFAR-100 checkpoints use their 5,000 original training-derived model-selection images, four non-primary corruption types, and severities 1, 3, and 5. Promotion requires mean corruption-accuracy gain ≥ 1 percentage point, a positive gain in every repetition, and mean clean-accuracy decrease ≤ 0.3 points. Relative to $r_\alpha = 1$, the corruption-accuracy changes are -0.16 ± 0.21 and -1.42 ± 0.19 points for $r_\alpha = 0.5$ and 2; mean clean changes are -0.98 and -1.25 points. Neither candidate qualifies; the native scale is retained. These development thresholds are selection criteria, not significance tests.

C.2 Clean-correctness partitions

For each model pair, matching clean predictions divide sources into both correct, only V-HMN/full correct, only comparator/uniform correct, and neither correct. For each group g, its fraction w_g and conditional corruption-accuracy difference Δ_g obey

$$A_{\text{corrupt}}(F) - A_{\text{corrupt}}(C) = \sum_g w_g \Delta_g.$$

Every source carries all 75 released corruption conditions. Tiny uses its corruption-paired clean crops, not canonical validation. Products are formed within each repetition before aggregation; an empty group has undefined conditional accuracy and zero contribution. Table 15 reports all four groups and the complete-cohort difference. These outcome-conditioned strata describe where differences occur; they do not adjust away training or architectural differences. Image-bootstrap intervals resample each complete source bundle and condition on the fitted checkpoints.

Clean-correctness group	Share (%)	Corruption gap (pp)	Contribution (pp)
ImageNet: public ResNet-50 V1			
All sources	100	12.47 [12.34, 12.61]	12.47 [12.34, 12.61]
Both correct	70.37	16.04 [15.89, 16.19]	11.29 [11.17, 11.41]
V-HMN only	6.13	20.46 [19.85, 21.08]	1.25 [1.20, 1.31]
Comparator only	5.78	−5.03 [−5.58, −4.46]	−0.29 [−0.32, −0.26]
Both wrong	17.72	1.24 [1.06, 1.45]	0.22 [0.19, 0.26]
ImageNet: uniform-read intervention			
All sources	100	1.52 [1.47, 1.57]	1.52 [1.47, 1.57]
Both correct	73.82	1.75 [1.69, 1.82]	1.29 [1.25, 1.34]
V-HMN only	2.68	7.46 [7.01, 7.87]	0.20 [0.18, 0.22]
Comparator only	1.74	−4.70 [−5.25, −4.17]	−0.08 [−0.09, −0.07]
Both wrong	21.76	0.49 [0.40, 0.59]	0.11 [0.09, 0.13]
CIFAR-100: uniform-read intervention			
All sources	100	3.43 ± 0.18	3.43 ± 0.18
Both correct	57.17 ± 0.28	3.15 ± 0.33	1.80 ± 0.19
V-HMN only	6.91 ± 0.49	24.62 ± 0.43	1.70 ± 0.15
Comparator only	3.34 ± 0.16	-17.30 ± 0.38	-0.58 ± 0.03
Both wrong	32.59 ± 0.28	1.55 ± 0.01	0.51 ± 0.01
Tiny ImageNet: uniform-read intervention			
All sources	100	1.85 ± 0.23	1.85 ± 0.23
Both correct	23.69 ± 0.95	2.35 ± 0.93	0.55 ± 0.20
V-HMN only	7.56 ± 0.16	14.11 ± 0.73	1.07 ± 0.05
Comparator only	3.21 ± 0.24	-10.24 ± 0.65	-0.33 ± 0.04
Both wrong	65.54 ± 1.05	0.86 ± 0.05	0.57 ± 0.03

Table 15: Exhaustive clean-correctness decomposition. Full V-HMN is compared with the named computation or checkpoint. Groups use the two predictions on the matching clean image; "V-HMN only" and "Comparator only" mean only that computation is clean-correct. The corruption gap is full-minus-comparator accuracy averaged equally over all 75 conditions within the group. Multiplying by its source share gives the unconditional contribution; four contributions sum to the all-source gap within each repetition. CIFAR/Tiny entries are means ± sample SD over three paired training repetitions, with products formed before averaging. ImageNet entries are fixed-checkpoint point estimates; brackets give 95% base-image bootstrap intervals (1,000 replicates), retaining every corruption and both computations together. All-source ImageNet intervals reuse the designated original checkpoint-comparison intervals printed elsewhere; group-specific intervals use the separate clean-correctness bootstrap. No interval is selected by its width. ImageNet has no training-run SD. Tiny uses the separate corruption-paired clean crops. These post-hoc strata retain the entire cohort; they do not causally adjust for clean recognition or replace the primary comparison.

C.3 UNCERTAINTY AND PAIRED OUTCOMES

Training-run uncertainty uses three independent repetitions and paired differences. The separate base-image bootstrap resamples source IDs, retaining their corruptions, severities, mask realizations, and supplied checkpoints together. These intervals are conditional on the fitted models. Small differences across three repetitions do not establish equivalence.

Rescue is $P(F \text{ correct}, B \text{ wrong})$ and harm is $P(F \text{ wrong}, B \text{ correct})$; their difference equals the accuracy gain. Both use the full cohort denominator. Conditional rescue and harm divide the corresponding counts by bypass errors and successes, respectively. NLL and true-class margin remain separate outcomes.

Matched temperature scaling. Following Guo et al. (2017), we fit one positive scalar T per model and training repetition by minimizing NLL on the same permitted clean model-selection examples used for checkpoint or exemplar-head selection. These labels are reused rather than newly held out for calibration. For classifier logits z, $p_T(y \mid x) = \mathrm{softmax}(z(x)/T)_y$. This rescales outputs, not memory-retrieval logits; all weights, banks, and exemplar-head choices remain frozen. We solve the convex problem in inverse temperature $b = 1/T$ by derivative bisection with prospective numerical bounds $T \in [0.05, 100]$. All 36 optima lie inside these bounds. The exemplar head uses $p_T(y \mid x) = p(y \mid x)^{1/T} / \sum_c p(c \mid x)^{1/T}$ with its original kernel and mixing weight. The same fitted temperature is applied to clean, all corruption, and both masking evaluations. Positive T preserves class ordering; every saved Top-1 decision is verified unchanged, retaining the original tie convention. NLL differences are paired within seed before their sample SD is computed. Unless marked temperature-scaled, intervention NLL values use the original unscaled outputs. NLL measures probability quality through both confidence and correctness; it is not a calibration-only statistic. Clean-derived output scaling does not guarantee calibration under distribution shift.

	Tiny ImageNet			CIFAR-100		
Model	Seed 2026	Seed 2027	Seed 2028	Seed 2026	Seed 2027	Seed 2028
V-HMN	2.761	2.775	2.793	2.619	2.634	2.652
Trained scaffold	2.655	2.730	2.683	2.622	2.634	2.641
Uniform-trained	2.656	2.626	2.598	2.608	2.630	2.661
ResNet-18	0.990	0.997	0.992	0.950	0.937	0.939
Scaffold + exemplar	1.034	0.694	1.054	1.014	0.642	1.212
AiT-S (adapted)	1.862	6.997	1.960	3.820	3.842	3.835

Table 16: Fitted positive classifier-output temperatures T, dimensionless. No fitted temperature reaches either numerical bound. Each value uses that run's clean selection cohort; no corruption-evaluation labels enter the fit.

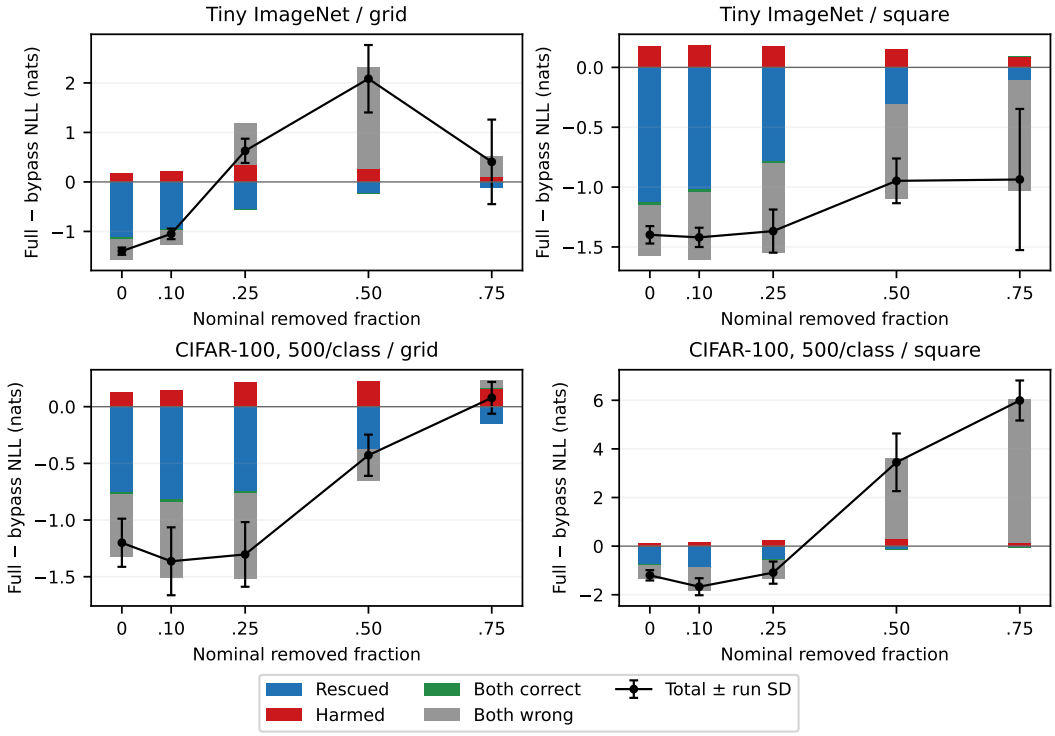


Figure 9: Outcome accounting of the fixed-checkpoint loss change on the complete primary masking cohorts. Each colored contribution averages $(\ell_F - \ell_B)\mathbf{1}\{\text{outcome}\}$ over all image–realization outcomes before averaging the three training runs. The four contributions sum to the black total; positive values mean reads worsen NLL. Bars stack positive and negative contributions separately. Error bars show the sample SD of the paired total across training runs, not a source interval. Clean is the canonical cohort; each nonzero fraction uses three mask realizations. Both-wrong predictions may assign different classes. Outcomes depend on both predictions, so the partition is descriptive rather than causal.

D LABEL AVAILABILITY, AUGMENTATION, AND OPTIMIZATION

The CIFAR-100 fitting/selection allocations are 45/5, 135/15, and 450/50 examples per class; memory draws only from the fitting pool. Section 5.3 presents the same-budget gains, crossing learning curves, and normalized log-label-area comparison. These do not support a general label-saving factor. Figure 10 retains the complete absolute curves and paired contrasts.

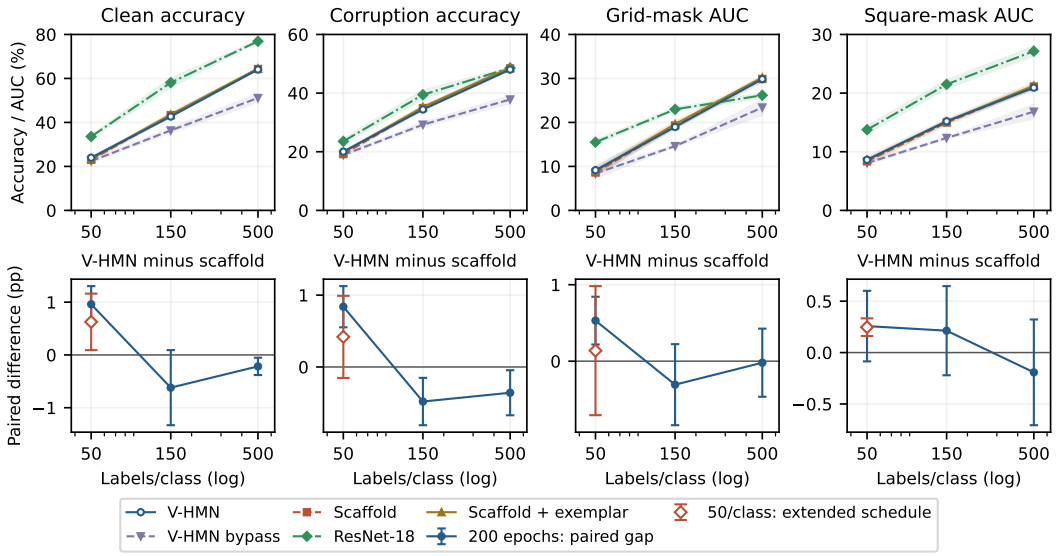


Figure 10: Label availability and schedule sensitivity on CIFAR-100. Top: absolute clean, corruption, and normalized masking accuracy AUC under crop/flip training for 200 epochs. Bottom: V-HMN minus its separately trained scaffold, computed within each paired seed before the mean and sample SD (bars). Hollow diamonds show the 50-label comparison after 35,200-update trajectories; clean-selected checkpoints can precede those endpoints. Top shading is sample SD over three training repetitions, not a confidence interval. Labels/class include fitting/selection allocations of 45/5, 135/15, and 450/50. Corruption accuracy weights all 75 conditions equally. All absolute axes start at zero; contrast axes retain zero and may differ in range. Lines connect observed budgets without extrapolating label savings.

The extended lowest-budget trajectories complete 35,200 updates, but clean selection often chooses an earlier checkpoint. Table 17 retains each selected epoch and update. Table 18 decomposes the change in fitted read dependence across schedules; its components are not causal fractions of learning.

Table 17: Training exposure and checkpoint selection at 50 labels per class. Each cell gives selected epoch / optimizer updates for the evaluated clean-selected checkpoint, not the completed trajectory. The 200-epoch runs completed 3600 updates with 200 clean-selection opportunities; extended runs completed 35200 updates with 1956 opportunities. Exact repeated-example exposure counts are retained with the per-run scientific outputs.

	200 epochs		Extended	
Seed	V-HMN	Scaffold	V-HMN	Scaffold
2026	144 / 2592	88 / 1584	1215 / 21870	942 / 16956
2027	105 / 1890	66 / 1188	1788 / 32184	1470 / 26460
2028	104 / 1872	50 / 900	1515 / 27270	599 / 10782

Table 18: Uniform-read accounting across lowest-budget training schedules. Entries are paired mean ± sample SD in accuracy/AUC percentage points across three training repetitions. F is full retrieval, M uniform bank means, and B bypass. The same clean-selected checkpoints and source cohorts supply all interventions. Differences telescope, $F - B = (F - M) + (M - B)$, within each schedule and for the paired schedule change; covariance is retained. This exploratory control was specified after inspecting the primary schedule results. Completed schedule lengths do not imply that evaluated checkpoints have identical update counts. This is not causal mediation or an allocation of semantic causes.

Endpoint	Schedule/contrast	$F - M$	$M - B$	$F - B$
Clean	200 epochs	$+0.55 \pm 0.49$	$+1.02 \pm 0.45$	$+1.56 \pm 0.21$
Clean	35,200 updates	$+2.05 \pm 0.42$	$+11.09 \pm 3.07$	$+13.14 \pm 3.07$
Clean	Extended − equal	$+1.50 \pm 0.46$	$+10.08 \pm 3.51$	$+11.58 \pm 3.09$
Corruption accuracy	200 epochs	$+0.26 \pm 0.32$	$+0.89 \pm 0.32$	$+1.14 \pm 0.01$
Corruption accuracy	35,200 updates	$+1.80 \pm 0.56$	$+9.30 \pm 2.38$	$+11.10 \pm 2.42$
Corruption accuracy	Extended − equal	$+1.54 \pm 0.37$	$+8.41 \pm 2.59$	$+9.95 \pm 2.41$
Grid AUC	200 epochs	-0.20 ± 0.04	$+0.96 \pm 0.14$	$+0.76 \pm 0.13$
Grid AUC	35,200 updates	$+0.02 \pm 0.34$	$+6.44 \pm 1.00$	$+6.46 \pm 0.84$
Grid AUC	Extended − equal	$+0.22 \pm 0.32$	$+5.48 \pm 0.93$	$+5.70 \pm 0.78$
Square AUC	200 epochs	$+0.05 \pm 0.10$	$+0.51 \pm 0.29$	$+0.56 \pm 0.21$
Square AUC	35,200 updates	$+0.27 \pm 0.22$	$+4.07 \pm 1.14$	$+4.33 \pm 0.97$
Square AUC	Extended − equal	$+0.22 \pm 0.13$	$+3.56 \pm 0.98$	$+3.77 \pm 0.93$

D.1 AugMix as a distinct training recipe

AugMix changes data processing and the objective for both V-HMN and its scaffold, with identical views and optimizer steps. Only the unmixed forward populates V-HMN banks. Table 19 gives the complete recipe interaction and read decomposition; the main paper discusses its selective-retrieval gains on masking alongside the total trained comparison.

Table 19: Uniform-read accounting across training recipes, in accuracy/AUC percentage points, mean ± sample SD of three paired training repetitions. F denotes full reads, M uniform bank means, and B bypass. The same seed pairing is retained between recipes. The components telescope, $F - B = (F - M) + (M - B)$, including for the AugMix-minus-crop/flip interaction. Their SDs must not be added; the covariance is retained. This is an exploratory sensitivity analysis, not causal mediation or an allocation of explained accuracy. No source-image interval is reconstructed from aggregate values.

Endpoint	Recipe/contrast	$F - M$	$M - B$	$F - B$
Clean	Crop/flip	$+3.57 \pm 0.47$	$+9.40 \pm 2.13$	$+12.96 \pm 1.72$
Clean	AugMix	$+1.79 \pm 0.44$	$+3.52 \pm 0.36$	$+5.31 \pm 0.13$
Clean	AugMix − crop/flip	-1.77 ± 0.70	-5.88 ± 2.28	-7.65 ± 1.59
Corruption accuracy	Crop/flip	$+3.43 \pm 0.18$	$+6.77 \pm 1.57$	$+10.20 \pm 1.41$
Corruption accuracy	AugMix	$+0.78 \pm 0.27$	$+5.33 \pm 0.40$	$+6.11 \pm 0.22$
Corruption accuracy	AugMix − crop/flip	-2.65 ± 0.41	-1.44 ± 1.88	-4.10 ± 1.62
Grid AUC	Crop/flip	$+0.68 \pm 0.32$	$+5.77 \pm 1.20$	$+6.45 \pm 1.11$
Grid AUC	AugMix	$+2.11 \pm 0.07$	-0.10 ± 0.23	$+2.01 \pm 0.25$
Grid AUC	AugMix − crop/flip	$+1.43 \pm 0.25$	-5.86 ± 1.10	-4.44 ± 1.01
Square AUC	Crop/flip	$+1.55 \pm 0.39$	$+2.54 \pm 1.01$	$+4.09 \pm 0.90$
Square AUC	AugMix	$+2.44 \pm 0.07$	$+0.56 \pm 0.73$	$+3.00 \pm 0.73$
Square AUC	AugMix − crop/flip	$+0.89 \pm 0.38$	-1.98 ± 1.61	-1.09 ± 1.31

E Mask geometry and fill sensitivity

The primary grid removes $\lfloor 64\rho + 1/2 \rfloor$ cells; its nonzero realized fractions are $9.375, 25, 50, 75\%$ at both resolutions. Square sides minimize absolute area error, with smaller-side tie breaking. At 64×64, realized fractions are $9.7656, 25, 49.4385, 73.8525\%$; at 32×32, they are

$9.7656, 25, 51.6602, 76.5625\%$. AUC integrates the declared nominal coordinates, so we do not silently substitute realized areas. Square placements share source-keyed random draws but are not nested.

The same removed area can produce different read effects under different geometries. At exactly 25% removal, the square-minus-grid difference in full-minus-bypass accuracy is $+6.10 \pm 0.86$ points on Tiny and -4.47 ± 0.47 on CIFAR-100. These contrasts concern the fitted models; corresponding total method differences are smaller and uncertain.

Fill sensitivity holds masks and removed source pixels fixed while changing their replacement statistics. On CIFAR-100 at 25% grid removal, V-HMN minus ResNet changes from $+9.07 \pm 1.91$ points with fitting-mean fill to -14.22 ± 0.32 with raw black fill. At 50% square removal, V-HMN's full-minus-bypass effect changes from -0.90 ± 0.51 to $+2.28 \pm 0.56$. Centering the mean-filled square leaves a negative read effect, -1.15 ± 0.76. These sensitivity conditions are descriptive and do not replace the primary endpoints. Fill comparisons use matching mean-fill controls under identical execution settings.

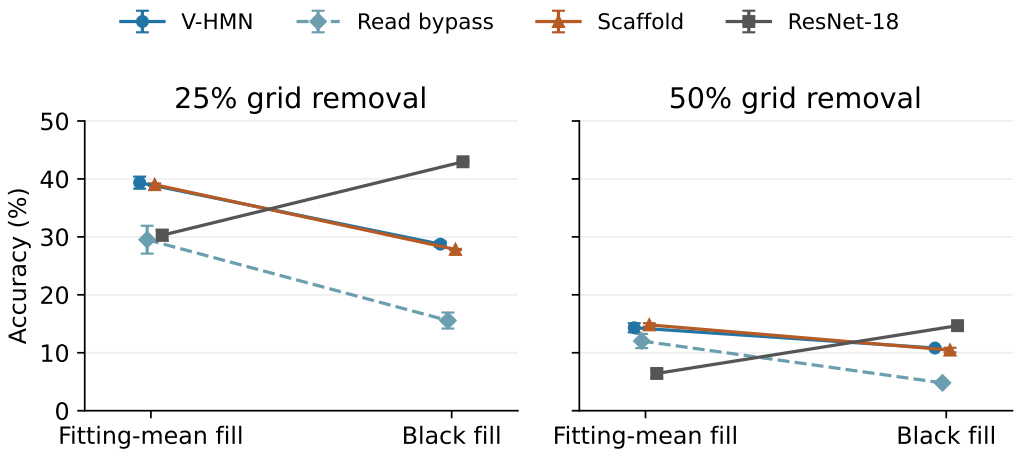


Figure 11: The CIFAR-100 grid-mask comparison depends on fill. Identical removed cells, source images, and three mask realizations are evaluated with raw fitting-mean RGB or raw black RGB before normalization. Points and bars are means and sample SDs across three training repetitions. Both prescribed fractions are shown. Lines connect the two categorical fill conditions; small horizontal offsets separate methods. A read advantage within V-HMN can increase while its ranking against ResNet-18 reverses.

Table 20: Paired model contrasts for Tiny ImageNet mask sensitivity. The shift is the sensitivity model contrast minus the matched primary-condition model contrast, in percentage points. Positive shifts indicate that the first-named method gains relative to the second under the sensitivity condition. All realizations remain within source-image clusters. Entries are training-seed means ± sample SD.

Condition	Contrast	n	Primary effect	Sensitivity effect	Shift
Black grid 25%	full − bypass	3	4.13 ± 0.87	7.60 ± 0.95	3.46 ± 1.42
Black grid 25%	full − scaffold	3	-0.76 ± 0.95	-0.65 ± 1.95	0.11 ± 2.88
Black grid 25%	bypass − scaffold	3	-4.89 ± 1.10	-8.24 ± 1.04	-3.35 ± 2.11
Black grid 50%	full − bypass	3	0.11 ± 0.44	1.03 ± 0.33	0.92 ± 0.77
Black grid 50%	full − scaffold	3	0.15 ± 0.50	-0.69 ± 0.82	-0.84 ± 1.28
Black grid 50%	bypass − scaffold	3	0.04 ± 0.26	-1.72 ± 0.64	-1.76 ± 0.89
Black square 25%	full − bypass	3	10.23 ± 0.04	14.47 ± 2.02	4.24 ± 1.98
Black square 25%	full − scaffold	3	-0.50 ± 1.26	-0.59 ± 2.38	-0.10 ± 2.64
Black square 25%	bypass − scaffold	3	-10.73 ± 1.29	-15.07 ± 0.80	-4.34 ± 0.78
Black square 50%	full − bypass	3	2.01 ± 0.50	3.78 ± 1.33	1.77 ± 1.32
Black square 50%	full − scaffold	3	-0.48 ± 1.10	-0.87 ± 1.98	-0.38 ± 1.88
Black square 50%	bypass − scaffold	3	-2.49 ± 0.61	-4.65 ± 0.77	-2.15 ± 0.57
Centered square 25%	full − bypass	3	10.36 ± 0.11	7.74 ± 0.57	-2.62 ± 0.46
Centered square 25%	full − scaffold	3	-0.27 ± 1.74	0.09 ± 1.97	0.36 ± 0.55
Centered square 25%	bypass − scaffold	3	-10.63 ± 1.64	-7.66 ± 1.40	2.97 ± 0.52
Centered square 50%	full − bypass	3	1.91 ± 0.68	1.25 ± 0.74	-0.65 ± 0.34
Centered square 50%	full − scaffold	3	-0.44 ± 1.13	-0.36 ± 1.32	0.08 ± 0.34
Centered square 50%	bypass − scaffold	3	-2.35 ± 0.46	-1.61 ± 0.59	0.74 ± 0.14

This sensitivity comparison does not include ResNet-18.

F CORRECTION, PATHWAY, AND MEMORY-CONTENT DIAGNOSTICS

Measurement sites. The selected-site probes hold the prefix input of block three fixed and rerun the ordinary suffix. They use the first 256 numeric source IDs in each paired final cohort, giving 90/100 CIFAR classes and 140/200 Tiny classes, across all three checkpoints. Prefix channel means and RMS statistics come from 256 class-stratified fitting sources. Pixel masks use one realization in these bounded probes. Feature noise at multipliers $0, .1, .25, .5$ and replacement of 25%/50% of channels by their fitting means perturb representations, not released benchmark images. These exploratory cohorts are not population-representative evaluations.

Fixed-target interpretation. For $\mathcal{F}(q) = \frac{1}{2}\|q - m(q)\|^2$, the chain rule gives

$$\nabla_q \mathcal{F}(q) = [I - J_m(q)]^\top [q - m(q)],$$

where J_m is the readout Jacobian. The native residual is therefore not generally the negative gradient of $\mathcal{F}$. Holding $m_t = m(q_t)$ fixed instead defines $E_t(q) = \frac{1}{2}\|q - m_t\|^2$, with

$$E_t\big(q_t + \beta[m_t - q_t]\big) = (1 - \beta)^2 E_t(q_t).$$

For a nonzero residual, this surrogate decreases exactly when $0 < \beta < 2$, with equality at $\beta = 0, 2$; it is nonincreasing for $0 \le \beta \le 2$. The learned scalar is unconstrained and each iteration recomputes the readout, so this identity provides no common-energy descent guarantee. Holding the target fixed here is an analytical construction: training retains gradients through the query-dependent read, while sample-written bank values remain detached.

Geometry and decisions. For a degraded query q, specified clean reference c in the same coordinate system, and fixed realized residual $r = \beta[m(q) - q]$, scaling the correction gives

$$\|q + \lambda r - c\|^2 - \|q - c\|^2 = 2\lambda\langle q - c, r\rangle + \lambda^2 \|r\|^2 . \tag{4}$$

For $\lambda > 0$ and $r \neq 0$, discrepancy decreases only when $\lambda < -2\langle q - c, r\rangle / \|r\|^2$. The learned signed β is included in r; negative alignment can be outweighed by an excessive step. For nonzero r, the diagnostic minimizer on $[0, 1]$ is $\mathrm{clip}(-\langle q - c, r\rangle / \|r\|^2, 0, 1)$. It uses the clean counterpart and is not an inference rule. We distinguish clean PRE from the fixed native clean POST reference,

and projected displacement Br from query displacement r. For one iteration with q, $m(q)$ and b fixed, changes preserving $A + B$ and βB preserve the projected branch while changing query-space geometry. Projected measurements therefore provide a useful additional reference, without guaranteeing decision improvement.

The projected curves below pair feature discrepancy with classification and branch-specific loss changes. Under CIFAR prefix noise, reads harm predictions despite moving the measured queries closer to clean references; feature replacement and image masking also produce different effects. Both clean references are retained.

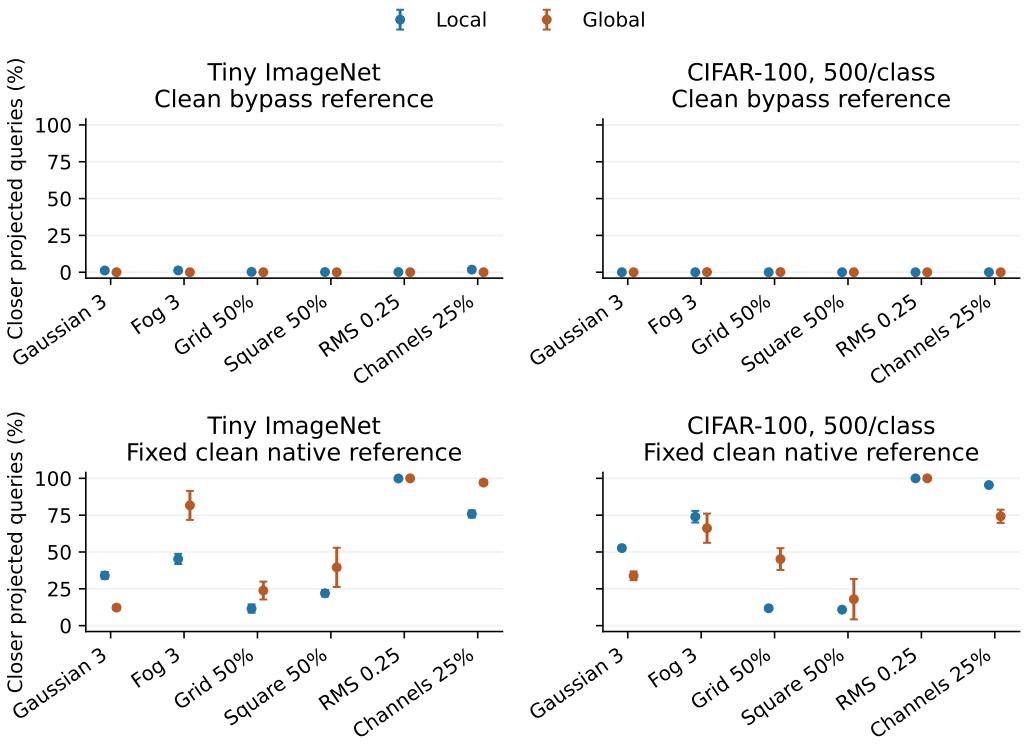


Figure 12: Projected branch discrepancy on the fixed mechanism cohorts. Points and bars are means and sample SDs across three training repetitions. Each point averages the fraction of closer queries within each source, then across sources. The six displayed conditions represent the two recorded released corruptions, the two primary mask geometries at nominal 50%, and the two representation perturbation types; the saved scientific artifacts retain all 14 conditions and all label budgets. Clean references are model outputs, not semantic ground truth.

F.1 Local and global read contributions

For accuracy-valued endpoint A_{lg}, the first factorial uses bypass for a disabled read, retaining all non-memory computation. The second uses the valid-bank mean, retaining the native update, and enables selective weighting instead. Each binary policy applies to its branch across every block. Learned parameters, stored state, input images, projections and refinement depth remain fixed; internal queries need not, because earlier interventions change downstream inputs. Within each repetition, we calculate

$$I = A_{11} - A_{10} - A_{01} + A_{00}, \qquad \phi_L = \tfrac{1}{2}[(A_{10} - A_{00}) + (A_{11} - A_{01})],$$

$$\phi_G = \tfrac{1}{2}[(A_{01} - A_{00}) + (A_{11} - A_{10})], \qquad \phi_L + \phi_G = A_{11} - A_{00}.$$

These allocations average the two orders of enabling reads. Adding the two effects measured with the other read enabled would count an extra I. All contrasts and their sample SDs use paired seed

values; corruption error is converted to accuracy before interpretation. Interactions describe fitted read operations, not biological synergy or the result of retraining a branch. Clean, corruption and masking endpoints remain separate.

Table 21: Algebraic attribution of the fixed-checkpoint local/global read contrast. All entries are paired mean ± sample SD across three training repetitions, in percentage points; corruption uses accuracy. I is the interaction. The symmetric allocations ϕ_L and ϕ_G sum to full minus both bypassed before rounding; they average two branch-addition orders and are not unique causal attributions.

Endpoint	ϕ_L	ϕ_G	I	Full − bypass
CIFAR-100				
Canonical clean	11.00 ± 1.12	1.96 ± 0.64	1.73 ± 1.58	12.96 ± 1.72
Corruption-paired clean	11.00 ± 1.12	1.96 ± 0.64	1.73 ± 1.58	12.96 ± 1.72
Corruption accuracy	9.35 ± 0.87	0.85 ± 0.54	2.40 ± 1.30	10.20 ± 1.41
Grid AUC	5.25 ± 0.77	1.20 ± 0.40	0.56 ± 0.87	6.45 ± 1.11
Square AUC	2.24 ± 0.66	1.86 ± 0.25	0.03 ± 0.83	4.09 ± 0.90
Tiny ImageNet				
Canonical clean	5.23 ± 0.68	12.95 ± 0.52	4.31 ± 1.24	18.18 ± 0.66
Corruption-paired clean	4.12 ± 0.73	9.27 ± 0.80	2.97 ± 0.70	13.39 ± 0.08
Corruption accuracy	2.15 ± 0.26	4.54 ± 0.54	2.05 ± 0.69	6.69 ± 0.30
Grid AUC	0.85 ± 0.12	3.91 ± 0.36	0.82 ± 0.74	4.76 ± 0.45
Square AUC	1.75 ± 0.09	5.46 ± 0.24	1.15 ± 0.54	7.22 ± 0.15

Table 22: Absolute selective–uniform factorial endpoints. S_{lg} enables selective weighting in the named local/global branch when its bit is one; zero uses the valid-bank mean with the native update. All policies apply across every block. Accuracy and normalized accuracy AUC are percentages, with mean ± sample SD across three fitted repetitions. S_{11} and S_{00} reuse the original full and uniform predictions. Ranking compares the four policies separately within each endpoint row. Bold marks numerical best values (ties retained), not significance.

Endpoint (% ↑)	S_{00}	S_{10}	S_{01}	S_{11}
CIFAR-100				
Canonical clean	60.51 ± 0.25	61.86 ± 0.22	62.60 ± 0.20	$\mathbf{64.07} \pm 0.23$
Corruption-paired clean	60.51 ± 0.25	61.86 ± 0.22	62.60 ± 0.20	$\mathbf{64.07} \pm 0.23$
Corruption accuracy	44.61 ± 0.31	45.09 ± 0.49	47.01 ± 0.22	$\mathbf{48.04} \pm 0.17$
Grid AUC	29.17 ± 0.49	29.18 ± 0.34	$\mathbf{30.06} \pm 0.79$	29.85 ± 0.62
Square AUC	19.37 ± 0.43	20.61 ± 0.18	19.85 ± 0.38	$\mathbf{20.93} \pm 0.28$
Tiny ImageNet				
Canonical clean	46.50 ± 0.24	46.71 ± 0.51	49.81 ± 0.36	$\mathbf{50.68} \pm 0.09$
Corruption-paired clean	26.90 ± 1.16	28.12 ± 1.57	29.14 ± 0.85	$\mathbf{31.25} \pm 0.85$
Corruption accuracy	15.28 ± 0.67	15.26 ± 0.87	16.58 ± 0.47	$\mathbf{17.13} \pm 0.48$
Grid AUC	15.64 ± 0.27	16.28 ± 0.13	16.81 ± 0.49	$\mathbf{17.78} \pm 0.56$
Square AUC	17.40 ± 0.56	18.21 ± 0.62	18.33 ± 0.26	$\mathbf{19.48} \pm 0.35$

Refinement depth. Evaluation requeries memory at each iteration. The models were trained at depth one; two or three iterations sharply reduce accuracy in the bounded Tiny probe. External strength scales the final realized displacement after all iterations rather than replacing each learned coefficient.

Table 23: Query-dependent local/global contributions relative to uniform reads in both branches. Each cell is a paired mean ± sample SD across three fitted checkpoints, in percentage points; corruption uses accuracy and masking uses normalized accuracy AUC. The symmetric allocations average each branch effect with the other branch selective and uniform. They sum to full-selective minus all-uniform before display rounding; the interaction is already split between them and must not be added again. This is an algebraic allocation of the declared whole-model intervention, not a unique semantic attribution.

Endpoint	Local allocation	Global allocation	Interaction	Full − uniform
CIFAR-100				
Canonical clean	1.42 ± 0.11	2.15 ± 0.40	0.12 ± 0.18	3.57 ± 0.47
Corruption-paired clean	1.42 ± 0.11	2.15 ± 0.40	0.12 ± 0.18	3.57 ± 0.47
Corruption accuracy	0.75 ± 0.10	2.68 ± 0.21	0.54 ± 0.24	3.43 ± 0.18
Grid AUC	-0.10 ± 0.23	0.79 ± 0.33	-0.22 ± 0.30	0.68 ± 0.32
Square AUC	1.16 ± 0.18	0.40 ± 0.29	-0.16 ± 0.14	1.55 ± 0.39
Tiny ImageNet				
Canonical clean	0.54 ± 0.21	3.63 ± 0.51	0.66 ± 0.37	4.18 ± 0.32
Corruption-paired clean	1.67 ± 0.44	2.68 ± 0.53	0.88 ± 0.42	4.35 ± 0.40
Corruption accuracy	0.27 ± 0.29	1.59 ± 0.32	0.58 ± 0.21	1.85 ± 0.23
Grid AUC	0.81 ± 0.26	1.34 ± 0.37	0.34 ± 0.12	2.15 ± 0.50
Square AUC	0.98 ± 0.21	1.10 ± 0.37	0.33 ± 0.11	2.09 ± 0.47

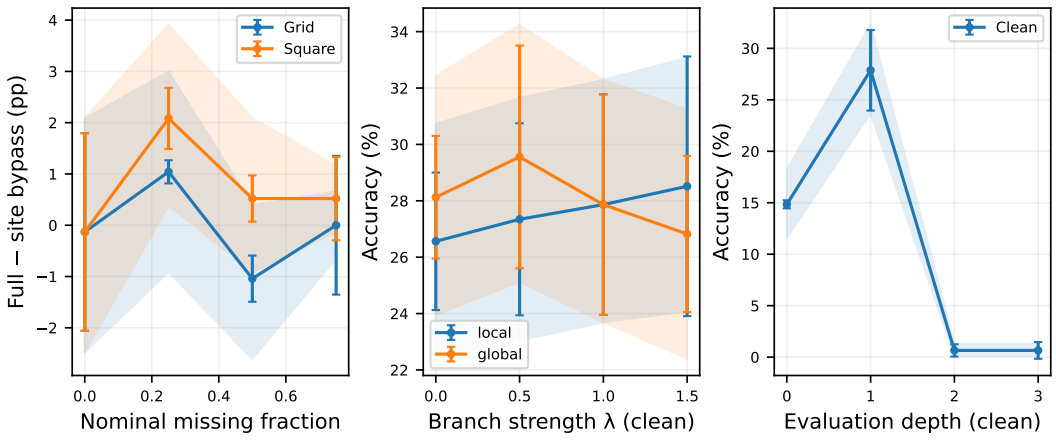


Figure 13: Tiny selected-site strength and whole-model depth diagnostics on the fixed paired cohort. Points show means and error bars sample SD across three training repetitions; shaded bands are 95% source-bootstrap intervals conditional on those fitted checkpoints. These exploratory curves do not select a replacement inference depth from final benchmark data.

Permuting slot labels or source metadata alone leaves Equation 1 unchanged. Likewise, a shared permutation of key–value rows is an invariance, not a harmful-content control. Content interventions must alter values, query–value associations, or the occupied set; their labels must describe the actual operator.

F.2 Capacity and distinct sources

At each CIFAR budget, the selected banks contain 21,000 occupied slots and a union of 1,000 distinct fitting sources. Thus increasing the fitting set reduces fractional stored-source coverage (22.22%, 7.41%, 2.22%) without changing that source count. Slots, views, and source images are distinct quantities. This observation alone does not explain the learning curve.

Frozen-bank probes use the first 1,000 canonical final records, three checkpoints, clean images, and one realization of each 50% mask. They reduce occupied capacity classwise to approximately one quarter, or retain slot allocation while copying fewer distinct sources within each class. A norm-matched version rescales copied values to destination norms. The intervention is applied at block three, with banks restored afterward. Tables report paired accuracy changes; differences in view multiplicity and stored norms remain recorded in the underlying artifacts. Destructive replacement establishes sensitivity to contents, rather than semantic appropriateness of ordinary retrieval. CIFAR-100 has 2,500/1,000 local/global entries from 500/1,000 distinct sources per bank. Capacity reduction leaves 600/200 entries from 396/200 sources. The fixed-slot coverage intervention retains 100/200 sources. Tiny ImageNet has 2,400/1,000 local/global entries from 600/1,000 distinct sources per bank. Capacity reduction leaves 600/200 entries from 430–438/200 sources. The fixed-slot coverage intervention retains 200/200 sources. Ranges describe variation across the three checkpoints; source counts are not summed across banks.

Table 24: Frozen-bank controls at block three on the first 1,000 canonical final images and one realization of each 50% mask. Full minus controlled computation, paired mean ± sample SD over three checkpoints. Positive accuracy and negative NLL favor the full bank. Rescue and harm are percentages of all images. These fixed-cohort diagnostics are distinct from corruption-suite and mask-AUC endpoints. Joint key–value permutations produce no prediction changes in this cohort.

Input	Bank control	Δ acc. (pp)	Rescue (%)	Harm (%)	Δ NLL (nats)
CIFAR-100					
Clean	Reduced capacity	-0.27 ± 0.42	0.53 ± 0.25	0.80 ± 0.17	0.000 ± 0.004
Clean	Reduced sources	-0.13 ± 0.23	0.77 ± 0.15	0.90 ± 0.10	-0.000 ± 0.014
Clean	Norm-matched sources	0.03 ± 0.15	0.87 ± 0.06	0.83 ± 0.15	-0.001 ± 0.015
Grid 50%	Reduced capacity	-0.20 ± 0.17	0.43 ± 0.12	0.63 ± 0.15	0.020 ± 0.051
Grid 50%	Reduced sources	-0.40 ± 0.17	0.53 ± 0.47	0.93 ± 0.31	-0.004 ± 0.017
Grid 50%	Norm-matched sources	-0.37 ± 0.29	0.60 ± 0.52	0.97 ± 0.23	0.002 ± 0.020
Square 50%	Reduced capacity	-0.03 ± 0.25	0.10 ± 0.10	0.13 ± 0.15	-0.019 ± 0.122
Square 50%	Reduced sources	-0.00 ± 0.17	0.20 ± 0.10	0.20 ± 0.10	0.119 ± 0.130
Square 50%	Norm-matched sources	-0.00 ± 0.17	0.17 ± 0.12	0.17 ± 0.15	0.132 ± 0.125
Tiny ImageNet					
Clean	Reduced capacity	-0.00 ± 0.20	0.77 ± 0.21	0.77 ± 0.06	-0.009 ± 0.008
Clean	Reduced sources	-0.07 ± 0.21	0.80 ± 0.10	0.87 ± 0.15	-0.012 ± 0.003
Clean	Norm-matched sources	-0.03 ± 0.29	0.93 ± 0.06	0.97 ± 0.23	-0.017 ± 0.009
Grid 50%	Reduced capacity	-0.00 ± 0.10	0.17 ± 0.15	0.17 ± 0.06	0.013 ± 0.017
Grid 50%	Reduced sources	-0.13 ± 0.12	0.23 ± 0.25	0.37 ± 0.29	-0.052 ± 0.012
Grid 50%	Norm-matched sources	-0.17 ± 0.12	0.23 ± 0.25	0.40 ± 0.17	-0.042 ± 0.028
Square 50%	Reduced capacity	-0.00 ± 0.17	0.10 ± 0.10	0.10 ± 0.10	0.002 ± 0.032
Square 50%	Reduced sources	-0.10 ± 0.17	0.17 ± 0.15	0.27 ± 0.12	-0.005 ± 0.044
Square 50%	Norm-matched sources	-0.10 ± 0.35	0.27 ± 0.12	0.37 ± 0.23	-0.013 ± 0.042

G Retrieval transparency and functional influence

Source-linked traces record the actual branch, layer, iteration, query position, slot IDs, weights, values, corrections, bank state, source IDs, view transforms, and write metadata. For local entries

the displayed region is the source query neighborhood, not the full receptive field of a deep feature. Displayed weights retain their original normalization and report omitted mass.

Examples are the first source identifier within each outcome stratum, using the fixed condition priority grid 50%, square 50%, grid 25%, square 25%, grid 75%, then square 75%: successful correction, harmful correction, unchanged success, and unresolved error. This is reproducible selection from a specified cohort, rather than a selection for visual appeal. Selected-entry removal targets cumulative attention mass and reruns the suffix with the remaining weights renormalized. Random removal matches the entry count and chooses the closest mass among 16 draws; actual mass gaps can be large, so it is not an exact mass-matched control. Value replacements preserve keys and weights, with norm matching and explicitly label-informed same/other-class diagnostics kept separate from ordinary inference. The two examples juxtapose opposite prediction effects under the same block-level read intervention. Their displayed central local query identifies part of the accessed information; it does not account for the entire two-branch effect. Table 26 separately accounts for a uniform-to-selective fixed-competitor score difference.

Table 25: Original-batch decision changes for the displayed source images. Classes are full / selected-site bypass predictions. Both local and global reads are bypassed at block 3 only; all other blocks remain enabled. Differences are full minus bypass. Positive true-class margin change favors full retrieval, whereas negative NLL change favors it. These are individual examples selected by the fixed outcome rule, not an estimate of population performance.

Outcome	Source	Mask	Classes	Δ margin (logits)	Δ NLL (nats)
rescued	110023	Grid 50%	163 / 138	+14.856	−12.749
harmed	110045	Grid 50%	141 / 137	−7.285	+4.668
unchanged correct	110067	Grid 50%	86 / 86	+2.593	−0.098
unchanged wrong	110000	Grid 50%	113 / 113	+1.256	−1.249

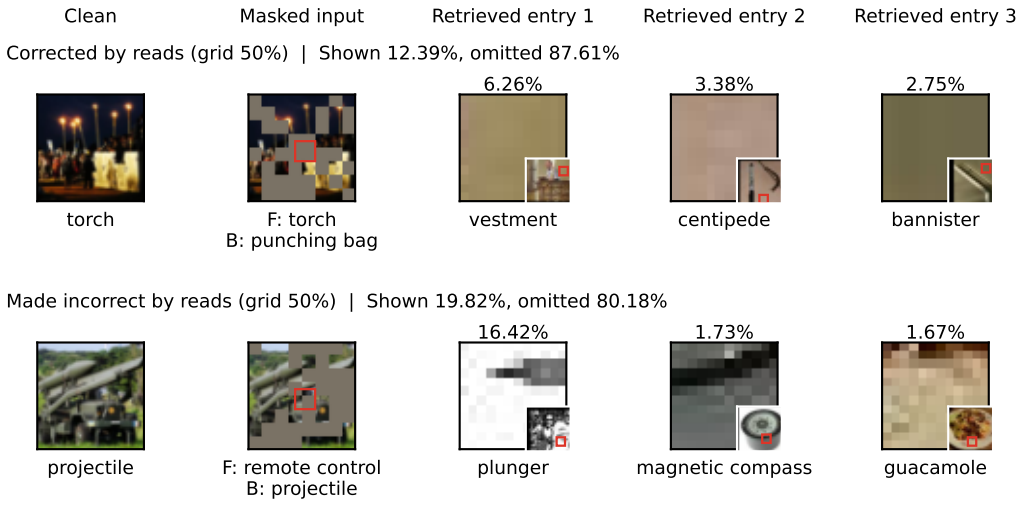


Figure 14: Actual stored information used by one Tiny ImageNet model at block 3. The same fixed-rule corrected and newly incorrect examples compare full reads (F) with both local and global reads bypassed at this block (B); other blocks stay native. The three entries come from one central local query, not the whole block-wide intervention. Each crop shows its recorded source-query neighborhood; the inset retains the full training view and red neighborhood box. Original weights and shown/omitted mass are retained. Names are stored-image labels, not class probabilities for the input. The red input box marks the local query neighborhood, not its complete receptive field. The appendix gives checkpoint details, all four outcomes, and functional measurements, including separate uniform-to-selective fixed-competitor score accounting (Table 26).

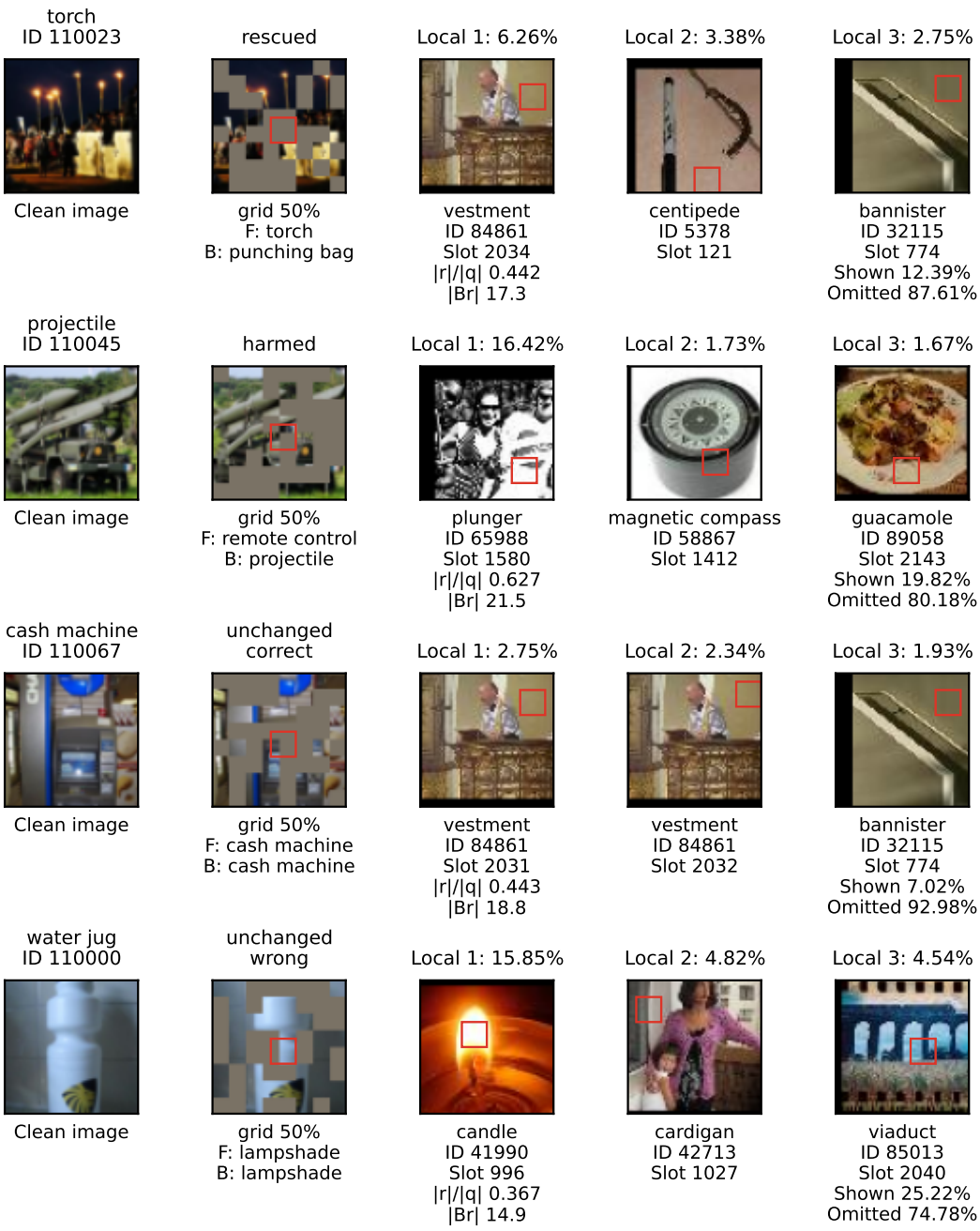


Figure 15: All four fixed-rule Tiny ImageNet examples at block three, local branch, seed 2026 and clean-selected epoch 189: successful correction, harmful correction, unchanged success and unresolved error. Outcome categories compare both reads enabled with both bypassed at this block. The displayed central local query and its three highest-weight entries are only part of that whole-site intervention. Source labels identify stored examples, weights retain their original normalization, and displayed mass leaves the remaining entries explicit. Stored neighborhoods do not exhaust deep-feature receptive fields.

A specified score difference. We additionally use path integration (Sundararajan et al., 2017) for these same four examples. At block three, the prefix is fixed. The baseline Z_0 is the fused pre-nonlinearity output with both reads uniform, and Z_1 uses both selective reads. For one query and occupied slot j, $d_j = \beta(a_j - 1/K)BM_j$. Lifting these contributions to local token positions and broadcasting global contributions gives $Z_1 - Z_0 = \sum_j \Delta Z_j$. We define $h(Z) = z_y(Z) - z_{c^*}(Z)$ with $c^* = \arg\max_{c \neq y} z_c(Z_1)$ fixed along the path. At Z_0, this score need not equal $z_y(Z_0) - \max_{c \neq y} z_c(Z_0)$; consequently its endpoint difference is not generally a change in the max-over-other margin. With

$$G = \int_0^1 \nabla_Z h(Z_0 + t(Z_1 - Z_0))\, dt, \qquad C_j = \langle G, \Delta Z_j \rangle_F,$$

the fundamental theorem of calculus gives $\sum_j C_j = h(Z_1) - h(Z_0)$. We use frozen parameters and banks, differentiating only through the ordinary suffix with respect to Z. Gauss–Legendre quadrature is bounded at 16, 32 and 64 nodes, requiring residual at most $10^{-3} + .01|h(Z_1) - h(Z_0)|$ and stability of displayed groups between successive rules. Grouping by source occurs after signed slot contributions are calculated. These scores depend on the chosen reference, site, score and path; they are not deletion effects or a complete causal account. The rescue/harm strata above use bypass, a different baseline.

Table 26: Uniform-to-selective fixed-competitor score accounting on the same four original trace examples. The first contribution sums only the three displayed local entries at the central query; other local covers all remaining local entries and query positions. Global includes broadcast contributions. These three groups plus the numerical residual equal $\Delta h = h(Z_1) - h(Z_0)$, in raw logit units, where $h(Z) = z_y(Z) - z_{c^*}(Z)$ and c^* is the highest-scoring nontrue class at the selective endpoint Z_1. That class stays fixed: $h(Z_1)$ equals the true-class margin, but $h(Z_0)$ need not. Thus Δh is not generally a change in the re-maximized true-class margin. Outcome labels retain their separate full-versus-site-bypass definition. Each row is one example from one checkpoint, without a population uncertainty interval.

Outcome	Source	c^*	Shown	Other local	Global	Residual	Δh	Nodes
rescued	110023	161	-0.015	$+0.302$	-2.098	$-3.81e-06$	-1.811	32
harmed	110045	141	$+0.002$	$+0.379$	$+0.441$	$-9.63e-06$	$+0.822$	32
unchanged correct	110067	132	-0.003	-0.018	$+1.668$	$+1.26e-07$	$+1.647$	32
unchanged wrong	110000	113	$+0.015$	$+0.900$	-2.998	$-7.73e-07$	-2.082	32

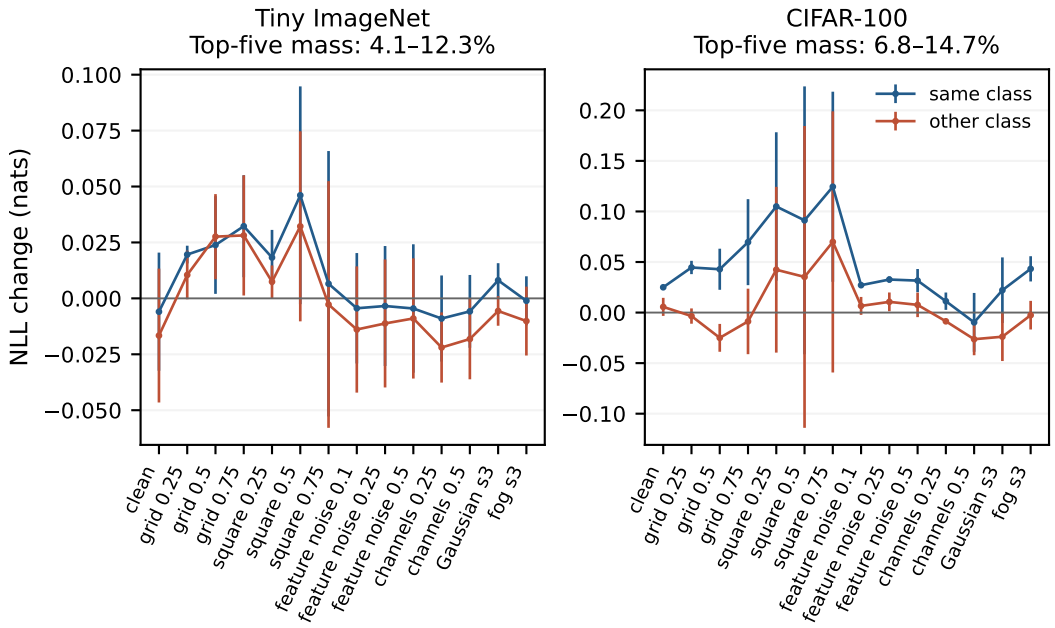


Figure 16: Functional influence of selected local values at block three, across every recorded condition of the fixed 256-image cohorts. Label-informed diagnostics replace the five highest-weight entries per query with norm-matched same-class or other-class source values, preserving keys and weights. Titles give the range of condition-mean selected mass; five entries account for only part of each read. The vertical axis is intervention-minus-full NLL: positive means worse loss. Points show means and bars sample SD across three fitted checkpoints. Ground-truth labels are used only in this diagnostic. This tests the influence of selected contents, not a complete explanation of the classifier or an evaluation-time inference rule.

H IMAGENET EVALUATION

ImageNet evaluation protocol. We evaluate one V-HMN model with fixed weights and populated banks (Table 27). The model contains 25,997,825 learned parameters, including the classifier, and 26,880,000 sample-written memory values counted separately. Its staged configuration has widths and latent dimensions 128/224/320/512, depths 3/3/4/2, 14×14 input patches, 256 spatial tokens, and 5,000/3,000 local/global slots per block. All 96,000 saved entries are valid; source-image provenance is unavailable. Frozen inference uses 1 refinement step and float32 arithmetic. The documented ImageNet recipe uses 310 epochs, AdamW, batch size 1,000, initial learning rate 0.001, cosine decay with initial linear warmup, weight decay 0.03, gradient clipping at norm 0.1, and RandAugment, MixUp, CutMix, and random erasing. The evaluated checkpoint records epoch 302; its original selection cohort and rule are unavailable. This epoch field does not establish a final-epoch selection. The original resolved training command and write history are unavailable. We evaluate all 50,000 validation images with the sorted-synset class mapping: RGB, bicubic antialiased short-edge resize to 256, center crop to 224, and normalization by $(0.485, 0.456, 0.406)/(0.229, 0.224, 0.225)$. We make no new checkpoint selection, write no evaluation representations, and do not repopulate memory. ImageNet-C uses all 15 corruption types and five severities, with released RGB images and no additional resize or crop. Raw mean error weights the 75 conditions equally; mCE uses the matching AlexNet reference errors. Full, bypassed, and uniform reads share one checkpoint. Intervals use 1,000 bootstrap replicates and resample base images together with all their perturbations and read modes; they do not estimate training-run uncertainty.

Public ImageNet reference. We use torchvision's ResNet-50 `IMAGENET1K_V1` checkpoint.[1] Its documented recipe trains on clean ImageNet-1K for 90 epochs with SGD, learning rate 0.1, momentum 0.9, weight decay 10^{-4} and global batch size 256. The learning rate decreases tenfold every 30 epochs; augmentation uses bilinear random resized crops to 224×224 and horizontal flips. This fixed model supplies an external comparison under its own recipe, rather than a matched training repetition.

Table 27: ImageNet-C performance and read interventions for two fixed checkpoints with different training recipes. Clean Top-1 is accuracy (%); C error is raw mean error (%) over all 75 ImageNet-C conditions; mCE uses the official AlexNet reference. NLL is raw and unscaled. Higher accuracy and lower error, mCE, and NLL are better. V-HMN full, bypass, and uniform rows are inference settings of the same checkpoint. Clean resizing is bicubic for V-HMN and the official bilinear preset for ResNet-50; both preserve released corruption pixels without resizing or cropping. Learned parameters: 26.00M for V-HMN and 25.56M for ResNet-50; V-HMN's sample-written memory is additional stored state. Ranking compares full V-HMN with ResNet only; interventions are unranked. Bold marks numerical best values (ties retained), not significance.

Model	Clean ↑	C error ↓	mCE ↓	Clean NLL ↓	C NLL ↓
V-HMN: full	**76.50**	**48.10**	**61.17**	1.058	**2.537**
V-HMN: bypass	70.56	56.87	72.39	1.389	3.561
V-HMN: uniform	75.56	49.61	63.12	1.253	2.794
Public ResNet-50 V1	76.15	60.57	76.43	**0.962**	3.274

Full retrieval changes mean corruption accuracy relative to bypass by +8.77 percentage points (conditional 95% interval [+8.64, +8.91]); rescue/harm probabilities are 13.65%/4.88%. Full retrieval changes mean corruption accuracy relative to uniform by +1.52 percentage points (conditional 95% interval [+1.47, +1.57]); rescue/harm probabilities are 4.52%/3.01%. Rescue minus harm equals the accuracy change. Subtracting the corresponding clean read gaps gives $+2.83$/$+0.58$ points against bypass/uniform reads, respectively; these are descriptive contrasts, not causal adjustments. Full-minus-bypass accuracy is positive for 14/15 corruption-type averages but negative in 3/75 individual conditions. Negative type averages occur for fog (-1.95 points). Full-minus-uniform accuracy is positive for 15/15 corruption-type averages but negative in 8/75 individual conditions.

[1] Released `resnet50-0676ba61.pth`; pinned torchvision recipe.

V-HMN minus the public ResNet-50 reference gives clean and mean corruption accuracy gaps of $+0.35$ $[+0.05, +0.66]$ and $+12.47$ $[+12.34, +12.61]$ percentage points, respectively. Subtracting the clean gap from the corruption gap gives $+12.12$ $[+11.82, +12.42]$ points. Positive values indicate a smaller clean-to-corruption accuracy decrease for V-HMN, not necessarily higher absolute corruption accuracy. Brackets are 95% paired image-bootstrap intervals conditional on these checkpoints; each resampled source retains its clean image and all 75 corruptions, preserving covariance. This descriptive difference does not isolate a memory effect or estimate training-run variability.

V-HMN has lower aggregate corruption error and NLL under these checkpoint-specific protocols, while the public reference has lower clean NLL. Accuracy gains occur across all 75 conditions; fog has higher type-average NLL for V-HMN despite higher accuracy. The within-checkpoint controls separately measure the effect of using this V-HMN model's persistent memory.

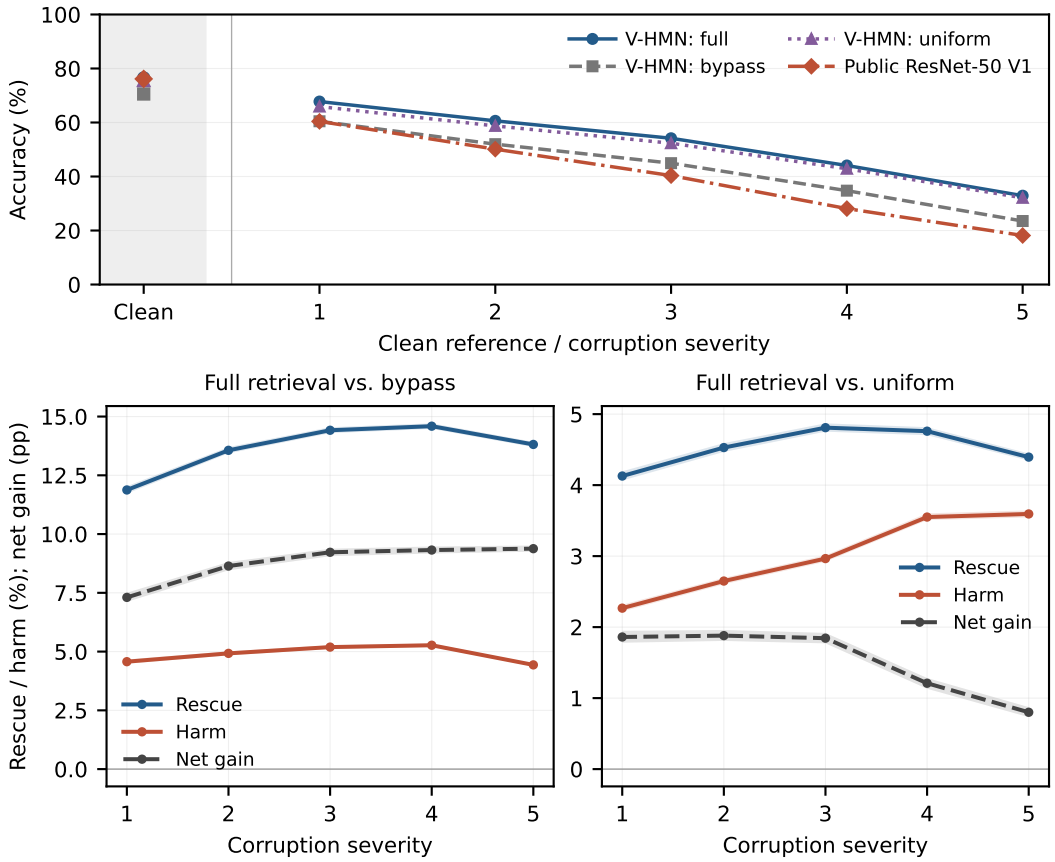


Figure 17: ImageNet scale context across severity, averaging all 15 corruption types. Top: accuracy point estimates for one V-HMN checkpoint with full, bypass or uniform reads, and a public ResNet-50 V1 checkpoint with a different recipe. Isolated clean markers use the respective declared preprocessing; released corruptions receive no further resizing or cropping. Bottom: read effects. Rescue means full retrieval is correct and the indicated intervention is wrong; harm is the reverse, both over the full cohort. Rescue minus harm is the accuracy gain. Bands are 95% paired image-bootstrap intervals preserving all perturbations and read conditions for each resampled source. They condition on the V-HMN checkpoint; these displays do not estimate training-run variability.

H.1 Additional clean ImageNet classification context

Table 28 preserves the complete clean-classification comparison, including input sizes, parameter counts and the reported ViT reimplementation marker. ResNet (He et al., 2016), SENet (Hu et al., 2018), ViT (Dosovitskiy et al., 2021) and MLP-Mixer (Tolstikhin et al., 2021) provide scale context

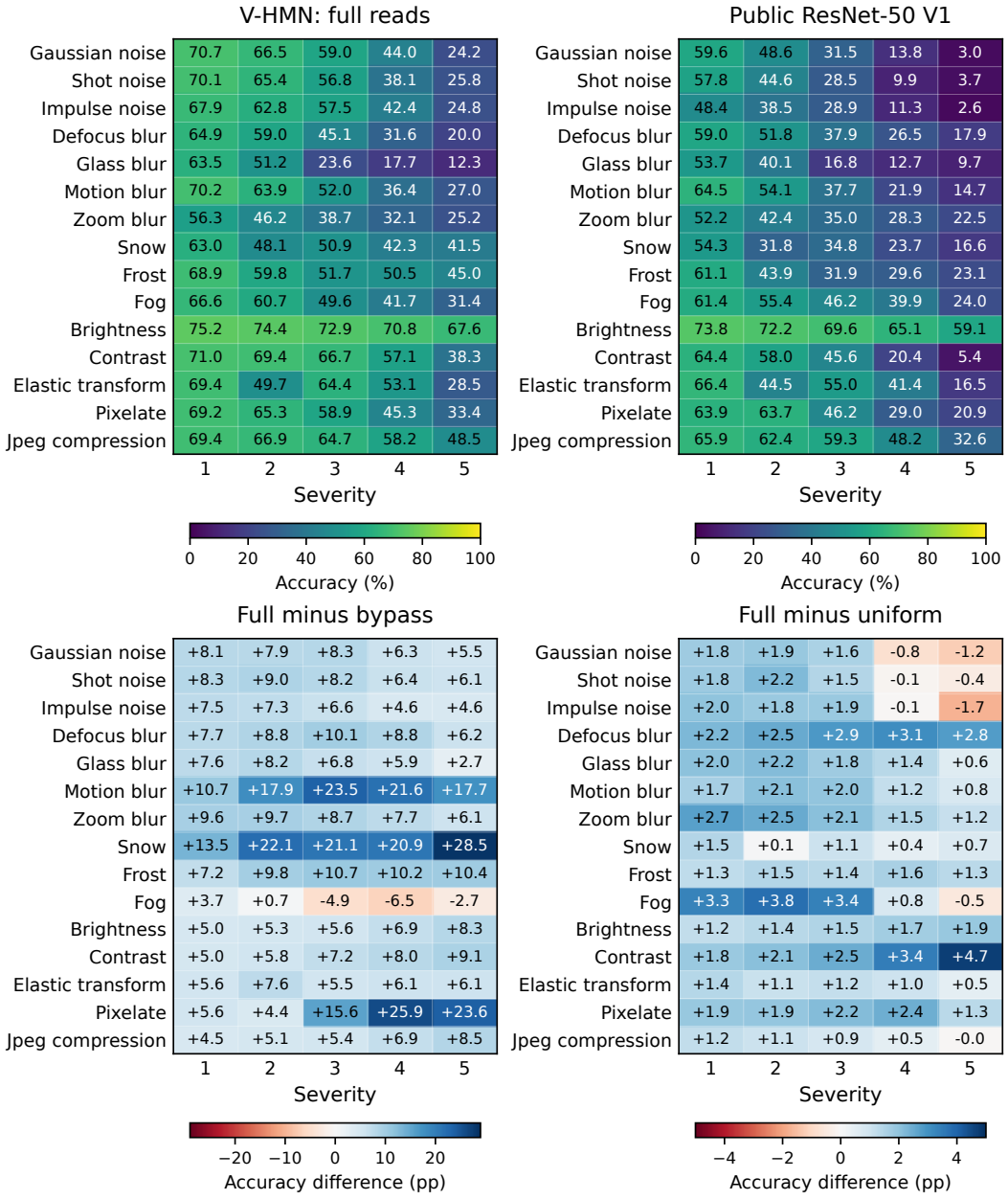


Figure 18: All 75 released ImageNet-C conditions for the evaluated V-HMN model and public ResNet-50 V1. Top: absolute Top-1 accuracy, using the same 0–100% color scale. Bottom: full-minus-bypass and full-minus-uniform accuracy, in percentage points; positive values favor full retrieval. Signed panels have separate symmetric scales centered at zero. Every cell uses the same 50,000 source images, with values rounded only for annotation. Corruption types and severities receive equal weight in the aggregate. V-HMN read modes share weights and banks; ResNet-50 V1 has its own training recipe and clean preprocessing. Cells are conditional point estimates for the fixed checkpoints, not independent training repetitions or significance tests.

Table 28: Clean ImageNet-1K classification context with original input sizes, parameter counts, and Top-1 accuracies. These rows use different recipes and resolutions and are not ranked. The dagger marks the reported ViT reimplementation at 224 pixels. The reported V-HMN value 76.7% and the evaluated model at 76.50% in Table 2 are distinct observations, not rounding of one result.

Model	Input (px)	Params	Top-1 (%) ↑
ResNet-50 (He et al., 2016)	224	26M	76.1
SENet-50 (Hu et al., 2018)	224	28M	76.7
ViT-B/16 (Dosovitskiy et al., 2021)	384	86M	77.9
ViT-B/16† (Dosovitskiy et al., 2021)	224	86M	76.2
ViT-L/16 (Dosovitskiy et al., 2021)	384	307M	76.5
MLP-Mixer-B/16 (Tolstikhin et al., 2021)	224	59M	76.4
V-HMN	224	26M	76.7

under their stated resolutions and recipes. The table places the reported V-HMN classification result among models of different sizes and input resolutions. It is not a common-recipe architecture comparison and has no uncertainty estimate. The 76.7% reported observation remains separate from the 76.50% clean evaluation paired with ImageNet-C in Table 2; the latter supplies the within-model full, bypass and uniform comparisons used in the present corruption analysis.

I ADDITIONAL SMALL-IMAGE EXPERIMENTS

The CIFAR-10-C evaluations below use three fixed trained models. The fractional-data and architecture tables report multi-run aggregates, with their original uncertainty convention. Separate diagnostic trajectories are identified explicitly; they are not substituted for those aggregate observations.

Released CIFAR-10-C follow-up. These evaluations use three V-HMN models trained with crop/flip augmentation and one public ResNet-56, each trained with 50,000 CIFAR-10 examples and selected on clean-test accuracy under the different recipes below. Frozen evaluation covers all 15 released corruption types at five severities. Full, bypass, and uniform reads share each V-HMN checkpoint. The available native T02b implementation writes detached queries after each branch read, before backward; the records do not specify a separate refresh pass.

The V-HMN runs use 400 epochs of Adam (learning rate 10^{-3}, weight decay 5×10^{-5}), five warm-up epochs and cosine decay; selected epochs are 391, 392 and 384.

Each V-HMN model has six width-256 blocks with 4×4 input patches, a 3×3 local window and 2,500 local/1,000 global slots per block. The public ResNet-56 checkpoint[2] uses 200 epochs of Nesterov SGD (learning rate 0.1, momentum 0.9, weight decay 5×10^{-4}) with cosine decay; its author log reports selection at epoch 198. Both recipes use batch size 256, random crop and horizontal flip, without corruption augmentation, MixUp or CutMix. Evaluation uses all 10,000 clean test images and their released corrupted counterparts at 32×32, without resizing or cropping, normalized by channel means $(.4914, .4822, .4465)$ and standard deviations $(.2023, .1994, .2010)$.

Paired read effects. Full minus bypass changes clean and corruption accuracy by $+66.17 \pm 6.55$ and $+55.81 \pm 4.27$ percentage points, respectively; the corresponding NLL changes are -4.289 ± 1.462 and -3.856 ± 1.287 nats. Full minus uniform changes clean and corruption accuracy by $+2.86 \pm 0.57$ and $+4.59 \pm 1.09$ percentage points, respectively; the corresponding NLL changes are -0.138 ± 0.044 and -0.448 ± 0.211 nats. Contrasts are computed within seed before their three-run mean $\pm$ sample SD. Positive accuracy and negative NLL differences favor full reads. Uniform is an inference intervention, not a separately trained control. Selective retrieval improves corruption accuracy over a bank-mean read in each repetition. The full-versus-bypass gap is larger on clean inputs, so the corruption gap also reflects a broader dependence on the trained inference computation. Compared with the public reference, full V-HMN has lower mean corruption error despite lower clean accuracy, while the reference has lower clean and corruption NLL.

[2] `https://github.com/chenyaofo/pytorch-cifar-models`; released `cifar10_resnet56-187c023a.pt`.

Model / inference	Clean acc. ↑	C error ↓	Clean NLL ↓	C NLL ↓
V-HMN: full reads	91.11 ± 0.07	$\mathbf{24.18} \pm 0.83$	0.502 ± 0.014	1.733 ± 0.076
Same checkpoints: bypass	24.94 ± 6.51	79.99 ± 3.68	4.791 ± 1.465	5.588 ± 1.214
Same checkpoints: uniform	88.26 ± 0.52	28.76 ± 1.23	0.640 ± 0.047	2.181 ± 0.202
Public ResNet-56	**94.37**	27.14	**0.252**	**1.565**

Table 29: Released CIFAR-10-C. Accuracy and raw mean corruption error are percentages; NLL is in nats. Corruption means weight all 75 conditions equally. V-HMN values are mean ± sample SD across three historical runs; bypass and uniform reuse those checkpoints. ResNet-56 is one fixed checkpoint, so no training-run SD is estimated. The text specifies the different recipes and historical clean-test selection. Ranking compares full V-HMN with ResNet only; interventions are unranked. Bold marks numerical best values (ties retained), not significance.

Custom perturbation protocols. The Gaussian-noise and block-occlusion tests used custom inputs, not released CIFAR-C or ImageNet-C; the refinement comparison changes training depth. Its retrieval-stability statistic measures perturbed-top-one membership in the clean top-five set, and its cosine compares top-one prototype vectors. Neither is a weighted-read recovery or prediction-rescue measure, and these illustrative statistics alone do not establish a controlled memory benefit. The retained tables preserve their original population and uncertainty conventions.

Fractional-data protocol and retained trajectories. Table 3 reports the fractional-data comparison. At 10%, 30%, and 50%, CIFAR uses 5,000, 15,000, and 25,000 fitting examples; Fashion-MNIST uses 6,000, 18,000, and 30,000. The experiment description specifies 400 epochs of Adam, learning rate 10^{-3}, batch size 256, weight decay 5×10^{-5}, five warm-up epochs, cosine decay, and crop/flip/AutoAugment/MixUp/CutMix. These recipe settings do not identify the exact selection epoch or bank state for each aggregate. The AiT row identifies the local `assoc_tf` wrapper, not the current AiT-S frozen-memory adaptation or an exact CVPR reproduction.

The best/final records in Table 32 describe separate retained trajectories. Their logs use official-test accuracy for selection, and all 27 retained best V-HMN checkpoints have empty sample banks. Their classification scores remain valid observations under that protocol, but do not demonstrate retrieval from stored training samples. Neither a single retained checkpoint nor a later experiment replaces the published multi-run aggregate. In particular, 50 fitting examples per CIFAR-100 class differs from the selection-inclusive 50-label budget of 45 fitting and five selection examples. The budget-controlled scaffold study provides a distinct controlled comparison; it is not pooled with the fractional-data table.

I.1 Complete fractional-data observations

The fractional-data study asks how the complete models recognize clean test images from restricted fitting sets. Tables 30 and 31 retain all reported observations, including the V-HMN-only 50% setting. The fitting counts and 400-epoch recipe are defined above. Selection-label access is not included in those fitting fractions. V-HMN has the highest reported mean among these models in each of

Table 30: V-HMN accuracy (%) at three fitting fractions. Entries retain the reported mean ± standard deviation; the comparison reports three seeds and does not specify the SD divisor. The 50% observations have no reported comparator at that fraction; rows have different fitting budgets and are not ranked.

Fitting fraction	CIFAR-10 ↑	CIFAR-100 ↑	Fashion-MNIST ↑
10%	80.22 ± 0.29	43.21 ± 1.07	89.18 ± 0.16
30%	88.67 ± 0.21	62.42 ± 0.29	91.04 ± 0.22
50%	91.19 ± 0.38	68.93 ± 0.35	91.53 ± 0.12

the six 10%/30% comparisons. The 50% observations extend its measured curve without supplying a cross-model comparison. The uncertainty remains the original mean and SD; aggregate cells do not supply paired-run intervals or identify a causal memory mechanism. The separately retained trajectories below have their own selected and final observations and are not pooled with these tables.

Table 31: Recognition from fractional training sets: Top-1 accuracy (%). C10/C100 denote CIFAR-10/100 and FM denotes Fashion-MNIST. Entries retain the reported mean ± standard deviation; the comparison reports three seeds and does not specify the SD divisor. Baseline identities follow the experiment: ViT, Swin-ViT, MLP-Mixer, PoolFormer (MetaFormer), Vim, and the local AiT wrapper; this AiT row is distinct from the frozen-memory AiT-S adaptation. Ranking is separate within each dataset/fraction column; higher accuracy is better. Bold marks numerical best values (ties retained), not significance.

Model	C10 10%	C10 30%	C100 10%	C100 30%	FM 10%	FM 30%
ViT	72.73 ± 0.42	83.94 ± 0.33	40.48 ± 0.80	57.40 ± 0.79	87.17 ± 0.40	89.71 ± 0.45
Swin-ViT	70.37 ± 0.94	80.89 ± 0.24	35.25 ± 0.59	51.77 ± 0.60	88.42 ± 0.48	90.34 ± 0.15
MLP-Mixer	76.14 ± 0.16	85.53 ± 0.40	41.94 ± 0.98	56.22 ± 0.75	87.16 ± 0.63	89.41 ± 0.13
MetaFormer	49.92 ± 4.08	60.86 ± 6.05	19.39 ± 1.26	37.47 ± 3.19	85.02 ± 1.22	88.78 ± 1.03
Vim	69.02 ± 2.13	79.58 ± 0.37	36.61 ± 1.25	49.67 ± 0.79	86.09 ± 1.03	88.18 ± 2.43
AiT	67.89 ± 0.55	80.69 ± 1.36	35.84 ± 0.67	54.59 ± 0.87	84.33 ± 0.38	87.62 ± 0.32
V-HMN	$\mathbf{80.22} \pm 0.29$	$\mathbf{88.67} \pm 0.21$	$\mathbf{43.21} \pm 1.07$	$\mathbf{62.42} \pm 0.29$	$\mathbf{89.18} \pm 0.16$	$\mathbf{91.04} \pm 0.22$

Table 32: Fractional-data summaries and separately retained trajectories. The reported column preserves the original aggregate means and SD verbatim. Retained columns use only the separately saved run directories with seed labels 1, 2, and 3, with sample SD ($n-1$ denominator). In those retained records, best selects the maximum official-test accuracy across 400 epochs, final uses epoch 400, and all 27 best checkpoints have zero occupied entries in their 12 local/global banks. These selection and bank observations describe the retained columns only; they are not assigned to the published aggregates without matching run lineage. Fit-label counts exclude selection labels in the retained protocol, so those runs do not implement the current fitting/selection annotation budgets.

Dataset	Train %	Fit labels	Reported acc. (%)	Best acc. (%)	Final acc. (%)
CIFAR-10	10	5,000	80.22 ± 0.29	80.39 ± 0.16	80.29 ± 0.15
CIFAR-10	30	15,000	88.67 ± 0.21	88.85 ± 0.22	88.72 ± 0.27
CIFAR-10	50	25,000	91.19 ± 0.38	91.52 ± 0.33	91.42 ± 0.31
CIFAR-100	10	5,000	43.21 ± 1.07	43.59 ± 0.91	43.34 ± 0.94
CIFAR-100	30	15,000	62.42 ± 0.29	62.51 ± 0.15	62.37 ± 0.25
CIFAR-100	50	25,000	68.93 ± 0.35	69.14 ± 0.38	69.08 ± 0.39
Fashion-MNIST	10	6,000	89.18 ± 0.16	89.29 ± 0.19	88.96 ± 0.15
Fashion-MNIST	30	18,000	91.04 ± 0.22	91.09 ± 0.06	90.91 ± 0.16
Fashion-MNIST	50	30,000	91.53 ± 0.12	91.63 ± 0.18	91.28 ± 0.25

I.2 Training with different refinement depths

Does the number of refinement steps affect recognition when it also changes training? Table 33 compares depths zero through three on the three named datasets. The source describes separate models trained at each depth under the small-image recipe; this comparison differs from changing the inference depth of one fitted model. The highest reported mean is at depth two on each dataset. The depth-three means are lower, so the table does not support monotonic improvement with refinement. These are reported means and standard deviations rather than reconstructed paired contrasts.

Figure 19 retains the complete corresponding CIFAR-10 custom-corruption plot. Its eleven stated settings comprise Gaussian-noise standard deviations $\{.05, .10, .20, .30\}$, square-occlusion area fractions $\{.05, .10, .20\}$, and contrast factors $\{.5, .75, 1.25, 1.5\}$. The figure reports a mean across these custom corruptions; exact weighting, perturbation repetitions, clipping, fill and seed identifiers are not specified. It is separate from the released CIFAR-C suite and the controlled masking curves. The custom plot shows how performance changes with the trained refinement configuration. It supplies additional perturbation context, while the fixed-checkpoint analyses in Appendix C separately test the effect of altering inference reads.

Table 33: Clean Top-1 accuracy (%) of models trained with the indicated refinement depth; FM denotes Fashion-MNIST. Entries retain the reported mean ± standard deviation; the comparison reports three seeds and does not specify the SD divisor. The last two columns retain reported custom CIFAR-10 accuracies (%), with dispersion and repetitions unspecified; dashes denote unreported results. Training and inference depths change together, unlike the fixed-checkpoint depth probe. Ranking compares training depths separately within each dataset. Bold marks numerical best values (ties retained), not significance.

Depth	CIFAR-10 ↑	CIFAR-100 ↑	FM ↑	Custom ↑	Occlusion ↑
0	93.56 ± 0.10	75.84 ± 0.14	92.05 ± 0.08	71.65	87.24
1	93.94 ± 0.11	76.58 ± 0.09	92.27 ± 0.06	73.79	89.08
2	$\mathbf{94.28} \pm 0.13$	$\mathbf{76.59} \pm 0.16$	$\mathbf{92.48} \pm 0.06$	**74.35**	**89.79**
3	93.99 ± 0.07	76.41 ± 0.09	92.40 ± 0.05	—	—

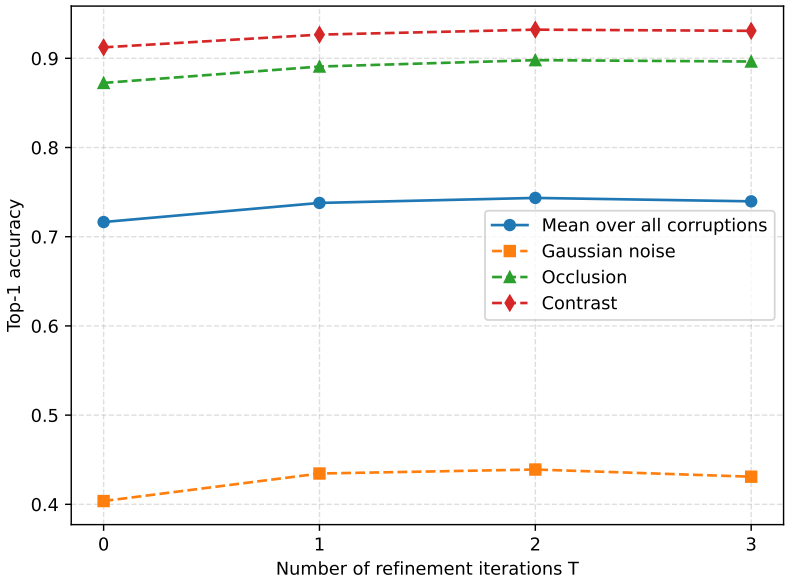


Figure 19: CIFAR-10 recognition under custom Gaussian noise, square occlusion and contrast changes, for models trained with different refinement depths. The complete original plot is retained, including depth three. Dispersion and the number of training repetitions for this plot are unspecified; the clean table's standard deviations do not supply uncertainty for these curves.

Table 34: Historical memory-construction controls (Top-1 accuracy, %): mean ± sample SD over three unique seeds. Best denotes test-selected checkpoints; final is the recorded final epoch. Online denotes clean-label online writes; no writes denotes strong augmentation without sample writes; refresh denotes strong augmentation with refresh. Ranking is separate for each write/augmentation regime and each best/final column. Bold marks numerical best values (ties retained), not significance.

		CIFAR-10		CIFAR-100	
Regime	Memory construction	Best ↑	Final ↑	Best ↑	Final
Online	Class-balanced ring	91.11 ± 0.08	90.88 ± 0.15	65.45 ± 0.31	65.17 ± 0.14
	Unbalanced ring	91.11 ± 0.12	91.00 ± 0.16	$\mathbf{65.54} \pm 0.10$	$\mathbf{65.26} \pm 0.18$
	Learned prototypes	91.21 ± 0.13	91.03 ± 0.17	64.47 ± 0.18	64.10 ± 0.15
	EMA class means	$\mathbf{91.46} \pm 0.26$	$\mathbf{91.29} \pm 0.32$	64.96 ± 0.12	64.72 ± 0.16
No writes	Class-balanced ring	93.67 ± 0.02	93.58 ± 0.05	75.83 ± 0.13	75.68 ± 0.16
	Unbalanced ring	93.71 ± 0.05	93.68 ± 0.02	75.96 ± 0.61	75.84 ± 0.58
	Learned prototypes	$\mathbf{94.32} \pm 0.16$	$\mathbf{94.22} \pm 0.12$	$\mathbf{76.46} \pm 0.26$	$\mathbf{76.37} \pm 0.20$
	EMA class means	93.65 ± 0.12	93.58 ± 0.11	75.66 ± 0.35	75.59 ± 0.41
Refresh	Class-balanced ring	93.56 ± 0.30	93.33 ± 0.34	75.86 ± 0.12	75.63 ± 0.05
	Unbalanced ring	93.66 ± 0.21	93.50 ± 0.25	76.07 ± 0.25	75.77 ± 0.35
	Learned prototypes	$\mathbf{94.19} \pm 0.21$	$\mathbf{94.15} \pm 0.24$	$\mathbf{76.60} \pm 0.21$	$\mathbf{76.41} \pm 0.19$
	EMA class means	94.06 ± 0.23	93.93 ± 0.24	76.39 ± 0.21	76.26 ± 0.23

Learned prototypes add trainable parameters; EMA stores one vector per class. These variants do not match both slot capacity and trainable capacity. In the no-write regime, sample rings and EMA receive no sample writes, whereas learned prototypes remain trainable.

Table 35: Clean-test window and nominal bank-capacity results. Cells are the reported mean ± SD over three stated seeds, in percent. Seed-level observations and the SD denominator are unavailable for these tables; the reported SD is retained without conversion, and no paired interval is inferred. These are reported training configurations, not new fixed-checkpoint bank interventions. Ranking is separate for window size, local capacity, and global capacity in each dataset. Bold marks numerical best values (ties retained), not significance.

Quantity	Setting (accuracy ↑)	CIFAR-10	CIFAR-100	Fashion-MNIST
Local window	3	$\mathbf{93.94} \pm 0.11$	$\mathbf{76.58} \pm 0.09$	$\mathbf{92.27} \pm 0.06$
Local window	5	93.22 ± 0.07	74.80 ± 0.07	91.94 ± 0.14
Local window	7	92.60 ± 0.67	70.67 ± 0.08	91.23 ± 0.35
Local slots	1500	93.93 ± 0.09	76.47 ± 0.29	92.32 ± 0.07
Local slots	2500	93.94 ± 0.11	76.58 ± 0.09	92.27 ± 0.06
Local slots	3500	$\mathbf{94.14} \pm 0.01$	76.42 ± 0.14	92.36 ± 0.14
Local slots	4500	94.04 ± 0.05	$\mathbf{76.59} \pm 0.13$	$\mathbf{92.44} \pm 0.03$
Global slots	500	93.91 ± 0.10	75.93 ± 0.12	92.23 ± 0.10
Global slots	1000	93.94 ± 0.11	$\mathbf{76.58} \pm 0.09$	$\mathbf{92.27} \pm 0.06$
Global slots	1500	$\mathbf{94.15} \pm 0.17$	76.41 ± 0.03	92.12 ± 0.04
Global slots	2000	93.94 ± 0.14	76.19 ± 0.04	92.09 ± 0.17

I.3 Clean small-image classification

Table 36 compares clean recognition across CIFAR-10, CIFAR-100, SVHN and Fashion-MNIST. It retains all seven model rows and the originally reported parameter figures. The baseline labels identify ViT (Dosovitskiy et al., 2021), Swin-ViT (Liu et al., 2021), MLP-Mixer (Tolstikhin et al., 2021), PoolFormer within MetaFormer (Yu et al., 2022), Vim (Zhu et al., 2024), and the local associative wrapper associated with AiT (Sun et al., 2025). This last row is distinct from the frozen-memory AiT-S adaptation evaluated under the controlled protocol. V-HMN has the highest reported mean on each dataset in this comparison. The default recipe specifies 400 epochs with the augmentation described in Appendix I; the table does not identify each aggregate's selected checkpoint or bank state. These results establish complete-model classification performance under the reported protocol. The separately trained scaffold and fixed-model interventions answer the different question of what

Table 36: Clean classification under the small-image comparison protocol. Accuracy (%) retains the reported mean ± SD over three seeds. Parameter counts (millions) are the original descriptive architecture figures, not a performance ranking or a measurement of sample-memory storage. Ranking is separate within each accuracy column; parameter counts are not ranked. Bold marks numerical best values (ties retained), not significance.

Model	CIFAR-10 ↑	CIFAR-100 ↑	SVHN ↑	Fashion-MNIST ↑	Params (M)
ViT	91.66 ± 0.08	72.56 ± 0.01	96.11 ± 0.29	91.83 ± 0.15	7.16
Swin-ViT	87.94 ± 0.10	66.58 ± 0.37	95.89 ± 0.17	91.79 ± 0.17	6.92
MLP-Mixer	92.65 ± 0.26	73.35 ± 0.39	96.91 ± 0.06	91.46 ± 0.32	8.71
MetaFormer	75.40 ± 1.56	48.76 ± 0.75	86.54 ± 2.71	91.42 ± 0.21	6.99
Vim	82.83 ± 3.87	63.29 ± 2.83	86.63 ± 1.51	88.63 ± 1.49	7.12
AiT	92.97 ± 0.30	72.91 ± 0.17	95.98 ± 0.06	91.51 ± 0.04	7.15
V-HMN	$\mathbf{93.94} \pm 0.11$	$\mathbf{76.58} \pm 0.09$	$\mathbf{97.16} \pm 0.04$	$\mathbf{92.27} \pm 0.06$	7.12

memory contributes. Parameter figures describe learned architectures; persistent stored values are additional state.

I.4 LOCAL AND GLOBAL PROTOTYPE ILLUSTRATION

The visualization procedure forms a separate index of training-image queries at the final block. It selects the local query location using attention-pooling weights and compares local and pooled global queries to this index. The source launcher specifies CIFAR-10 and an 8×8 visualization patch size. Its cosine function normalizes both operands and returns their dot product; neighbors are ordered from largest to smallest score. Thus the figure's scores are cosine similarities, not distances. Figure 20 preserves all four examples and their numerical scores, with the score label and the deer spelling corrected.

The local panels show spatially aligned patches while the global panels show whole source images. These panels make the local/global distinction visible but do not supply the native read weights, omitted attention mass, or a prediction intervention. They are therefore distinct from the recorded forward traces and functional tests in Appendix G. The illustrated checkpoint and exact index instance are not linked to the evaluated performance models.

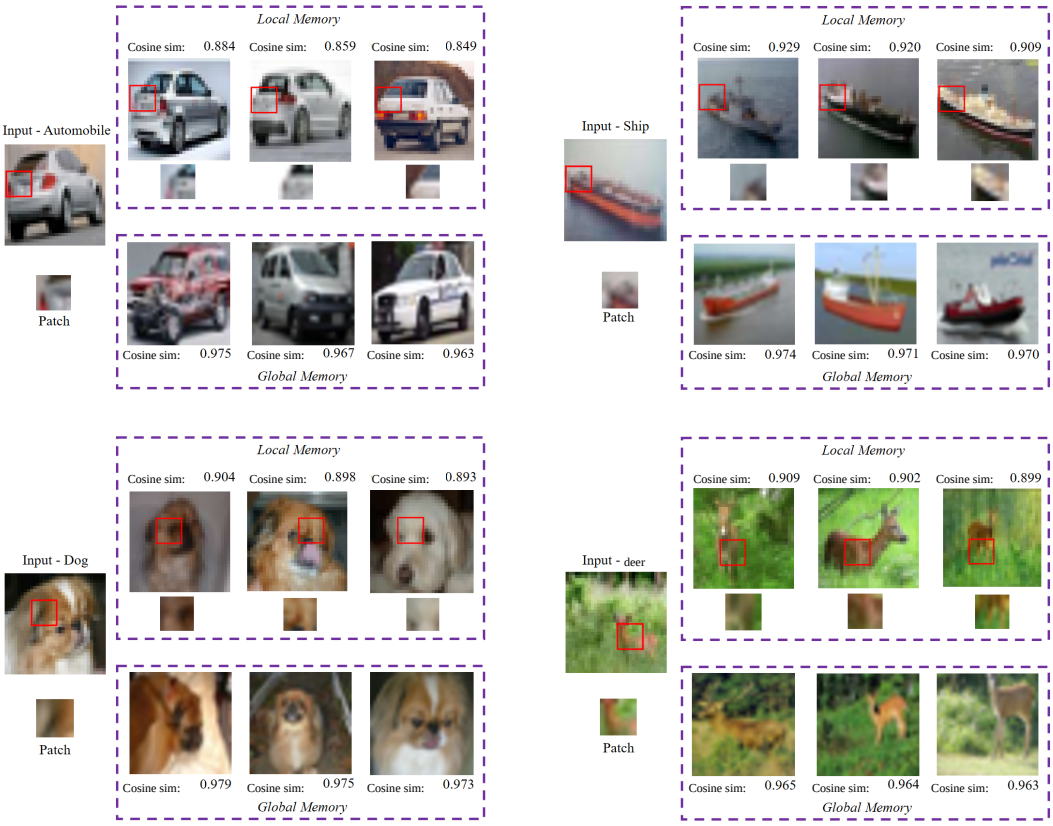


Figure 20: Complete local/global nearest-prototype illustration on CIFAR-10, using 8×8 patches specifically for visualization. Automobile, ship, dog and deer panels retain the input, selected query region, local source regions and global source images. Red boxes locate the attention-pooling-selected query and corresponding source patches. Labels report cosine similarity to the separate training-query index, not attention probabilities. The example panels do not establish that these neighbors are the persistent slots used by the classifier or that visual resemblance caused a correct prediction.