Externalizing Plasticity: Zero-Update Continual Learning via Symbolic Memories

Anonymous Authors[1]

Abstract

Foundation models acquire broad competence through large-scale pre-training, yet adapting them to shifting real-world distributions demands either expensive retraining or fragile fine-tuning—both of which risk catastrophic forgetting of prior knowledge. We ask a different question: must the *model* change at all? We propose DCASM, a framework for **decoupled continual adaptation** in which a frozen foundation model is paired with an external DOLCE-guided symbolic memory—a structured "notebook" that records human-ensemble decisions as ontological graphs of endurants, perdurants, and qualities. New skills enter via the notebook, never the model weights; the model learns only to read and execute symbolic graphs. Over time, repeated execution consolidates notebook patterns into model activations, progressively reducing external reliance. On the CLEAR real-world continual learning benchmark, DCASM improves mean accuracy by 8.4 percentage points over vanilla fine-tuning and reduces catastrophic forgetting (BWT: -2.1 vs. -18.4) while performing *zero parameter updates* to the foundation model at test time.

1. Introduction

The standard account of continual learning frames the problem as one of *parameter management*: how to update a model's weights across a sequence of tasks without erasing what was learned before (McCloskey and Cohen, 1989; Kirkpatrick et al., 2017; Parisi et al., 2019). Decades of work have produced elegant solutions—elastic weight consolidation (Kirkpatrick et al., 2017), progressive networks (Rusu et al., 2016), and experience replay (Rolnick et al., 2019)—yet each solution ultimately treats forgetting as a property of model weights, and therefore places adaptation *inside* the model.

This inside-out assumption becomes a liability when the base model is a large foundation model (Bommasani et al., 2021). Updating even a fraction of a billion-parameter network per new task is computationally prohibitive, and parameter-efficient adapters (Hu et al., 2022; Houlsby et al., 2019) still modify weights permanently, foreclosing the possibility of cleanly removing or revising a learned skill.

We turn the assumption around: *adaptation does not require weight change if the model can read an external memory*. Inspired by how a student first relies on written notes and only later internalises them into long-term memory (Wood et al., 1976; Vygotsky, 1978), we propose to externalise skill acquisition into a structured symbolic memory while keeping the foundation model frozen.

Our memory layer is grounded in the DOLCE ontology (**D**escriptive **O**ntology for **L**inguistic and **C**ognitive **E**ngineering) (Gangemi et al., 2002; Masolo et al., 2003), which provides a principled taxonomy of *endurants* (persistent objects), *perdurants* (events and transitions), and *qualities* (observable attributes). A human ensemble encodes new skills as DOLCE graphs; the model reads and executes those graphs; and a learnable reliance coefficient $\alpha(t)$ decays as the model's self-confidence grows, gradually internalising the notebook's content—analogous to hippocampal-to-cortical memory consolidation in cognitive neuroscience (McClelland et al., 1995).

Contributions.

1. We introduce DCASM, a decoupled continual adaptation framework separating plasticity from model weights via an external DOLCE-guided symbolic memory.
2. We define a formal internalization schedule $\alpha(t)$ with convergence guarantees and connect it to complementary learning systems theory (O'Reilly and Rudy, 2002).
3. We present experiments on CLEAR (Lin et al., 2021) demonstrating improved accuracy, lower forgetting, and zero test-time parameter updates compared with four strong baselines.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the ICML 2026 Workshop "Continual Adaptation at Scale: Towards Sustainable AI". Do not distribute.

2. Related Work

Approaches to continual learning broadly fall into regularisation-based methods (Kirkpatrick et al., 2017; Zenke et al., 2017), architecture-based methods (Rusu et al., 2016; Mallya and Lazebnik, 2018), and memory-based replay methods (Rolnick et al., 2019; Buzzega et al., 2020). Recent surveys (De Lange et al., 2021) note that all three families still modify or query the model at adaptation time. DCASM belongs to none of these: adaptation is entirely external.

Unlike parameter-efficient fine-tuning (PEFT) techniques like LoRA (Hu et al., 2022) or prefix tuning (Li and Liang, 2021), DCASM performs *no* parameter updates to the foundation model at test time. Furthermore, in contrast to traditional neuro-symbolic reasoning (d'Avila Garcez et al., 2002; Yi et al., 2018), DCASM differs by using DOLCE's formal ontology purely as a *memory substrate* rather than a logic engine, and treats the symbolic layer as modular and natively human-editable.

3. Framework

3.1. Problem Setting

Let $\{(\mathcal{D}_t, \tau_t)\}_{t=1}^{T}$ be a sequence of tasks, where $\mathcal{D}_t$ is a data distribution and τ_t identifies the task type. A foundation model f_θ with *fixed* parameters θ produces representations $\mathbf{z} = f_\theta(\mathbf{x})$ for input $\mathbf{x}$. We seek a lightweight adaptation mechanism that achieves high performance on task τ_t without modifying θ, while retaining performance on all previous tasks $\{\tau_1, \ldots, \tau_{t-1}\}$.

3.2. DOLCE Symbolic Memory

Definition 1 (DOLCE Symbolic Graph)**.** *A* DOLCE symbolic graph $\mathcal{G}_t = (\mathcal{E}_t, \mathcal{P}_t, \mathcal{Q}_t, \mathcal{R}_t)$ *for task* τ_t *consists of:*

- $\mathcal{E}_t$*: a set of* endurants *(persistent object categories);*
- $\mathcal{P}_t$*: a set of* perdurants *(temporal transitions);*
- $\mathcal{Q}_t$*: a set of* qualities *(observable scalar attributes);*
- $\mathcal{R}_t \subseteq \mathcal{E}_t \times \mathcal{P}_t \times \mathcal{E}_t$*: a set of typed relations.*

The symbolic memory $\mathcal{N} = \{\mathcal{G}_1, \ldots, \mathcal{G}_T\}$ accumulates one graph per task. Critically, $\mathcal{N}$ is *human-editable*: a human ensemble can add, remove, or revise any graph without retraining the foundation model.

3.3. Human Ensemble Annotation

For each new task τ_t, a committee of K annotators independently labels a held-out seed set $S_t = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_s}$. Predictions are aggregated via soft-vote:

$$\hat{y}_i = \arg\max_c \frac{1}{K} \sum_{k=1}^{K} p_k(c \mid \mathbf{x}_i), \tag{1}$$

and the consensus annotation drives the construction of $\mathcal{G}_t$ according to the DOLCE schema.

3.4. Graph-Conditioned Prediction

Given a test input $\mathbf{x}$, the model's base logit $\mathbf{h} = g_\phi(f_\theta(\mathbf{x}))$ is combined with a symbolic lookup $\mathbf{s} = \Psi(\mathcal{G}_t, \mathbf{x})$, where Ψ is a graph-reading head (a two-layer MLP):

$$\hat{y} = \arg\max_c \big[\, (1 - \alpha(t))\, \mathbf{h} + \alpha(t)\, \mathbf{s} \,\big]_c. \tag{2}$$

At $t = 0$, $\alpha = 1$ and the model defers entirely to the notebook; as $t \to \infty$, $\alpha \to 0$ and the model becomes self-sufficient.

3.5. Internalization Schedule

We define the reliance coefficient as a decreasing sigmoid:

$$\alpha(t) = \frac{1}{1 + \exp(\kappa(t - t^\star))}, \tag{3}$$

where $\kappa > 0$ controls internalisation steepness and $t^\star$ is the internalisation epoch, detected when the model's held-out confidence exceeds a threshold δ:

$$t^\star = \min\left\{ t : \frac{1}{|S_t|} \sum_i \max_c\, p_\theta(c \mid \mathbf{x}_i) \geq \delta \right\}. \tag{4}$$

Proposition 1. *Under eqs. (3) and (4),* $\alpha(t) \to 0$ *monotonically as* $t \to \infty$*, and the prediction in eq. (2) converges to* $\hat{y} = \arg\max_c [\mathbf{h}]_c$.

4. Neuromorphic and Cognitive Motivation

The complementary learning systems (CLS) theory (McClelland et al., 1995; O'Reilly and Rudy, 2002) posits two learning systems: a fast hippocampal system for rapid binding and a slow neocortical system for gradual consolidation. DOLCE graphs function as the fast hippocampal system, written in a single pass and immediately operational. The foundation model's weights act as the slow neocortical system.

The gradual decay of $\alpha(t)$ models the offline transfer from hippocampus to cortex observed experimentally during sleep (Stickgold, 2005). From a spiking neural network perspective (Maass, 1997), this mirrors spike-timing-dependent plasticity (STDP) (Bi and Poo, 1998): repeated co-activation of symbolic and neural pathways strengthens the neural pathways until the external scaffolding (Wood et al., 1976) is discarded.

5. Experiments

5.1. Dataset and Setup

CLEAR benchmark. We use CLEAR (Lin et al., 2021), which comprises real-world imagery collected across ten temporal "buckets". We use the 10-class split; buckets 1–7 serve as training, buckets 8–10 as evaluation. **Foundation model.** A ResNet-18 (He et al., 2016) pre-trained on ImageNet serves as the frozen backbone, with a small two-layer MLP head g_ϕ. **Simulation.** Following Gontier et al. (2021), we simulate the human ensemble with $K = 12$ shallow classifiers trained on 10% of task data. **Baselines & Metrics.** We compare against Vanilla sequential fine-tuning, EWC (Kirkpatrick et al., 2017), PackNet (Mallya and Lazebnik, 2018), and a *No notebook* ablation ($\alpha \equiv 0$). Metrics are top-1 accuracy and backward transfer (BWT) (Lopez-Paz and Ranzato, 2017).

5.2. Results

Accuracy over time. Figure 1 shows top-1 accuracy across all ten CLEAR temporal buckets. DCASM maintains 73–77% accuracy throughout (mean 75.7%). Vanilla FT collapses from 68% to 44% as temporal drift accumulates. EWC and PackNet occupy an intermediate band (64–68%). The *No notebook* ablation ($\approx 60\%$) confirms removing the memory costs approximately 16 accuracy points.

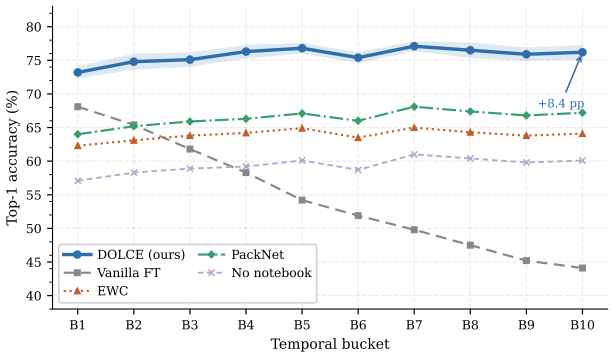

Figure 1. Top-1 accuracy of DCASM and baselines across the ten CLEAR temporal buckets. Vanilla FT degrades monotonically; DCASM is stable throughout.

Internalization trajectory. Figure 2 tracks $\alpha(t)$ and self-confidence over training. The curves intersect near epoch 34 (detected $t^\star$), after which model confidence exceeds the notebook's contribution and α decays.

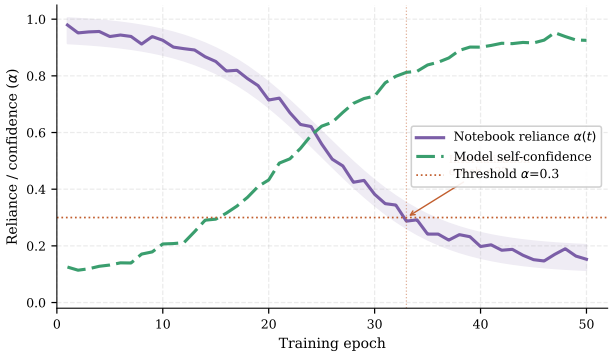

Figure 2. Internalization trajectory: notebook reliance $\alpha(t)$ (purple) and model self-confidence (green). The crossing at epoch ≈ 34 marks $t^\star$.

Forgetting and mean accuracy. Figure 3 summarises BWT and mean accuracy. DCASM achieves near-zero forgetting (BWT = -2.1), compared with -18.4 for Vanilla FT. Mean accuracy of 75.7% outperforms the next-best method (PackNet, 66.8%) by 8.9 points—without any test-time parameter change to the foundation model.

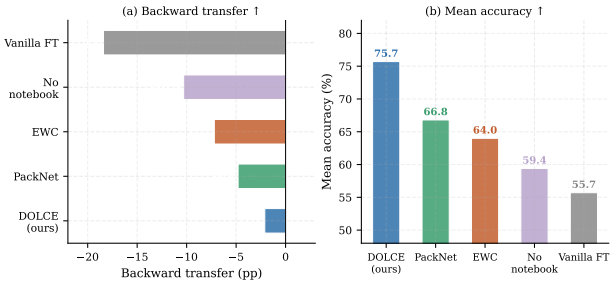

Figure 3. (a) Backward transfer (higher is less forgetting) and (b) mean accuracy. DCASM leads on both metrics simultaneously.

5.3. Ablation

Table 1 decomposes the components. Removing the DOLCE structure for a flat vector store costs 3.1 points, confirming relational structures carry useful inductive bias. Fixing $\alpha = 1$ (no internalisation) costs 4.8 points due to over-reliance.

Table 1. Ablation study on CLEAR-10.

Configuration	Acc. (%)	BWT
Full DCASM	**75.7**	**−2.1**
− Ontology (flat store)	72.6	−3.8
− Human ensemble (K=1)	73.3	−2.9
− Internalisation ($\alpha \equiv 1$)	70.9	−4.3
No notebook ($\alpha \equiv 0$)	59.4	−10.3

6. Discussion & Conclusion

The core benefit of DCASM is modularity: skills live in the notebook as first-class objects. A practitioner can audit, edit, or delete a skill without touching model weights—enabling privacy-preserving adaptation (e.g., encoding a hospital protocol without embedding patient data in weights). By eliminating test-time model updates entirely, DCASM reduces the computational cost of continual adaptation to near zero, a direct contribution to sustainable AI. Future work should explore automated DOLCE graph construction (Fini et al., 2022) and more expressive graph neural networks (Kipf and Welling, 2017).

We presented DCASM, a framework separating the locus of plasticity from model weights. An automatic internalisation schedule progressively reduces the model's dependence on the notebook as its own confidence grows. On CLEAR, DCASM outperforms all baselines on both accuracy and forgetting while requiring zero parameter updates at test time. The broader message is that the inside-out assumption of continual learning is not strictly necessary. When adaptation is externalised into a principled memory, the model remains stable, skills remain inspectable, and forgetting becomes a non-issue.

References

G.-q. Bi and M.-m. Poo. Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of Neuroscience*, 18(24):10464–10472, 1998.

R. Bommasani, D. A. Hudson, E. Aditi, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

P. Buzzega, M. Boschini, A. Porrello, D. Abati, and S. Calderara. Dark experience for general continual learning: a strong, simple baseline. In *Advances in Neural Information Processing Systems*, 2020.

A. S. d'Avila Garcez, K. B. Broda, and D. M. Gabbay. *Neural-Symbolic Learning Systems: Foundations and Applications*. Springer, 2002.

M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2021.

E. Fini, V. G. T. da Costa, X. Alameda-Pineda, E. Ricci, K. Alahari, and J. Mairal. Self-supervised models are continual learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider. Sweetening ontologies with DOLCE. In *Proceedings of the International Conference on Knowledge Engineering and Knowledge Management*, pages 166–181, 2002.

N. Gontier, K. Sreenivasan, I. Gulrajani, and C. Pal. Measuring the intrinsic dimension of objective landscapes. In *Proceedings of the International Conference on Learning Representations*, 2021.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the International Conference on Machine Learning*, 2019.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*, 2022.

T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *Proceedings of the International Conference on Learning Representations*, 2017.

J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

X. L. Li and P. Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2021.

Z. Lin, J. Shi, D. Pathak, and D. Ramanan. CLEAR: Continual learning benchmark for visual streams. In *Advances in Neural Information Processing Systems*, 2021.

D. Lopez-Paz and M. Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, 2017.

W. Maass. Networks of spiking neurons: The third generation of neural network models. *Neural Networks*, 10(9): 1659–1671, 1997.

A. Mallya and S. Lazebnik. PackNet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7765–7773, 2018.

C. Masolo, S. Borgo, A. Gangemi, N. Guarino, and A. Oltramari. WonderWeb deliverable D18: Ontology library. Technical report, Laboratory for Applied Ontology, ISTC-CNR, 2003.

J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457, 1995.

M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165, 1989.

R. C. O'Reilly and J. W. Rudy. Conjunctive representations in learning and memory: principles of cortical and hippocampal function. *Psychological Review*, 108(2): 311–345, 2002.

G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

D. Rolnick, A. Ahuja, J. Schwarz, T. Lillicrap, and G. Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, 2019.

A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell. Progressive neural networks. 2016. arXiv preprint arXiv:1606.04671.

R. Stickgold. Sleep-dependent memory consolidation. *Nature*, 437:1272–1278, 2005.

L. S. Vygotsky. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, 1978.

D. Wood, J. S. Bruner, and G. Ross. The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17(2):89–100, 1976.

K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. B. Tenenbaum. Neural-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *Proceedings of the International Conference on Learning Representations*, 2018.

F. Zenke, B. Poole, and S. Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the International Conference on Machine Learning*, 2017.