Lost or Hidden? A Concept-Level Forgetting in Supervised Continual Learning

Katarzyna Filus
Institute of Theoretical and Applied Informatics
Polish Academy of Sciences, Gliwice, Poland
kfilus@iitis.pl

Kamil Faber
AGH University of Krakow
Krakow, Poland
kfaber@agh.edu.pl

Roberto Corizzo
American University
Washington DC, USA
rcorizzo@american.edu

Christopher Kanan
University of Rochester
Rochester, New York, USA
ckanan@cs.rochester.edu

Abstract

Continual learning studies how models can adapt to new tasks while retaining previously acquired knowledge. Although a broad spectrum of methods has been proposed to mitigate catastrophic forgetting, the field remains predominantly performance-driven, with limited insight into what forgetting actually corresponds to within the vision model's representation space. Prior work has primarily analyzed forgetting through task-level performance or coarse measures of representational drift, without disentangling output-level accessibility from changes in finer-grained internal structure. To this end, we propose a diagnostic framework that leverages Sparse Autoencoders (SAEs) to define a task-anchored latent feature space, enabling analysis of how task-specific information evolves at a finer granularity, where individual SAE latents are treated as concept proxies for recurring and relatively disentangled visual patterns in the model's internal computations. Within this framework, we decompose forgetting into apparent concept deletion, recoverability, and decodability. We show that a large portion of seemingly lost concept-level information can often be recovered under linearity assumption, with concept decodability degrading as more tasks are introduced. Overall, our findings suggest that a significant part of concept-level forgetting can be attributed to changes in the representational accessibility rather than complete information erasure.

1 Introduction

The degradation of previously learned knowledge during sequential training, known as catastrophic forgetting, remains a central challenge in continual learning [33]. Various forgetting mitigation strategies have been proposed, including rehearsal-based methods [30], regularization and distillation [24], as well as hybrid methods [5]. Despite this progress, evaluation remains largely performance-driven, equating forgetting to decrease in task accuracy after learning new tasks, while leaving unresolved the fundamental questions of how forgetting manifests within models [26]. Recent works notice this gap and attempt to shift from the output-level evaluation to representations [18, 26, 6, 15]. Previous studies measure representational drift using mostly similarity measures [22, 18, 26]. Other works [6, 15] show that much apparent performance loss can be recovered via retrained classifiers or representation constraints. These works establish that task-relevant information can persist despite output-level forgetting and motivate our study. However, analyses of dense, entangled representations provide limited insight into the persistence and accessibility of individual task-specific features, motivating our analysis at the level of concept proxies

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

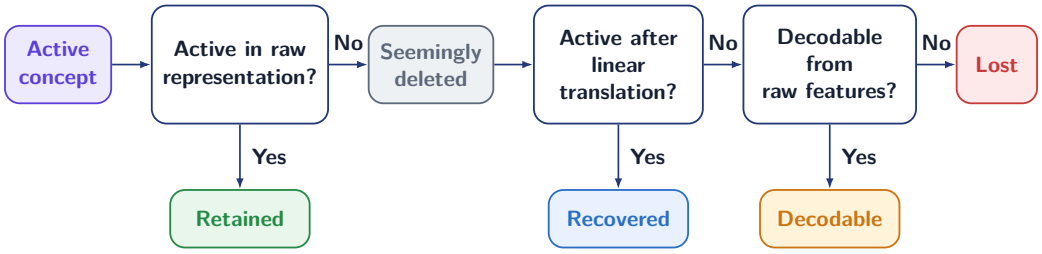


Figure 1: The taxonomy of concept transitions for task t data after task $t+1$ training.

Similarly to how humans recognize objects by combining recurring concepts, recent progress in mechanistic interpretability enables the study of neural representations in terms of more disentangled and interpretable features used to perform learned tasks [2]. It is particularly important because deep networks are considered to rely on superposition, where multiple concepts – understood as recurring patterns within the model computation – are encoded in overlapping neural directions, making feature-level analysis difficult [2]. Methods such as Sparse Autoencoders (SAEs) [19] and transcoders [9] map representations into higher-dimensional sparse latent spaces that can isolate more atomic and interpretable features. Although these latent dimensions do not correspond to the ground-truth semantic concepts, they often capture recurring and relatively disentangled patterns in model computations, serving as useful *concept proxies* that often align with abstract motifs such as furry animals or red objects [2]. In the following, we refer to them as *concepts* and leverage them as a structured basis for analyzing how fine-grained information evolves during continual learning.

Studying forgetting at a fine-grained level should be as important as in terms of accuracy degradation, indicating whether the model retains knowledge and supports stable interpretations rather than shifts representations. Mechanistic interpretability offers a promising basis for this analysis. The closest work to ours [26] uses transcoders to study geometric transformations of features in a shared latent space. However, it focuses on *how* features transform, rather than analyzing task-specific information accessibility. Our approach complements this perspective by anchoring the analysis to individual tasks and examining the subsequent accessibility of their concept proxies through activation dynamics, linear recovery, and concept-level decodability.

In this work, we propose a concept-level diagnostic framework for analyzing forgetting through the accessibility of task-specific information. Using Sparse Autoencoders (SAEs), we define a fixed, task-anchored latent space in which individual features act as concept proxies, enabling fine-grained tracking of how information evolves during continual training. Within this framework, we decompose forgetting into concept deletion, recoverability, and decodability, distinguishing between information that is removed, misaligned, or still present but not directly accessible. Our results show that much of the seemingly lost concept-level information remains linearly recoverable, indicating that forgetting often reflects reduced accessibility under the evaluated readouts rather than complete erasure. At the same time, some concepts become less linearly decodable over time, making loss of accessibility functionally equivalent to forgetting under standard linear readouts. Moreover, continual strategies differ in how well they preserve the fine-grained information. Our findings provide a unified view of forgetting that links representational drift, readout limitations, and concept-level information dynamics. Our contributions can be summarized as follows:

- We introduce a concept-level diagnostic framework for analyzing forgetting in continual learning, leveraging SAEs to define a task-anchored latent space whose features act as concept proxies. It enables studying task-specific knowledge preservation and linear recoverability through concept activation dynamics and task- and concept-level decodability.
- We define a taxonomy of concept transitions after learning subsequent tasks in continual learning: retained, seemingly deleted, recovered, decodable, and lost concepts (see Fig. 1).
- We show that a substantial portion of task-specific information remains linearly recoverable at the level of individual concepts, and that continual learning strategies differ in how well they preserve the fine-grained knowledge and its recoverability.

The code implementation of our framework is available at: `https://github.com/lifelonglab/lost-or-hidden-neurips-2026`.

2 Related work

Most work in continual learning quantifies forgetting through degradation in the model's classifier performance [32, 31, 20, 21]. Most works focused on developing mitigation strategies, including replay-based methods, regularization and distillation, architectural, and hybrid variants [33, 29, 12], with forgetting typically evaluated through accuracy-based metrics. Although this line of work has established effective methods for reducing catastrophic forgetting, it offers limited insight into how forgetting manifests within internal representations.

More recent studies relate forgetting to representational drift and changes in feature geometry [6, 15, 26, 18]. The authors of [6] shows that classifier-level forgetting can overestimate degradation of the underlying representations. These findings suggest that task-relevant information may persist despite output-level forgetting, providing an important motivation for forgetting analysis beyond task-level performance. However, existing works have focused on global properties of dense, entangled representations, leaving the persistence of finer-grained information less explored. In contrast, we move beyond task-level probing on raw features by using an SAE-defined, task-anchored concept space to analyze whether more fine-grained task-relevant information remains preserved and linearly decodable at the concept level. The authors of [15] enforces linear separability across tasks through backward feature projection. In contrast, we use linear mappings not as a training constraint, but as a diagnostic tool to measure how much past information remains linearly recoverable under standard continual learning strategies, and where this recoverability breaks down.

In terms of understanding forgetting, mechanistic interpretability offers a complementary perspective by decomposing neural representations into more atomic and interpretable components, e.g., by using Sparse Autoencoders (SAEs) and transcoders [19, 9, 2]. Such methods construct overcomplete, sparse latent spaces that expose disentangled patterns, offering a stronger basis for feature-level analysis. Recent work in continual learning uses SAEs to mitigate forgetting through sparse feature distillation in language models [34], activation regularization [27], and SAE-induced geometric constraints on representation drift [11]. These works use SAEs to constrain continual updates. Our contribution is a complementary, post-hoc diagnostic framework. We use task-anchored SAEs to examine whether old task-specific information remains directly accessible, recoverable through linear translation, or residually decodable. Such an analysis can complement accuracy-based testing by revealing how different mitigation strategies affect the fine-grained task information.

The closest work to ours is [26], in which transcoders are used to map representations from different tasks into a shared space and describes forgetting in terms of geometric transformations such as rotations and scaling. While this work provides a more feature-focused perspective on forgetting, our framework differs in its explicit task anchoring and focus on the accessibility of previously learned, task-specific information. In contrast, our goal is not to characterize how features transform but to study and quantify whether task-specific information remains accessible at the concept level. To this end, we use task-anchored SAE latent spaces and analyze concept activation dynamics, linear recoverability, and concept-level decodability. This allows us to distinguish between information that is preserved but misaligned, implicitly encoded, or no longer accessible under linear readouts. This enables us to directly examine different outcomes of concept-level forgetting, including retention, apparent deletion, recovery through linear translation, and residual decodability, rather than relying on global representational properties.

3 Forgetting as Loss of Accessibility: A Concept-Level Framework

In this section, we formulate our concept-level analysis of forgetting as a loss of accessibility to fine-grained task-specific information rather than complete information loss. The central question is whether information that no longer supports the original task readout after continual training has been erased, or it remains present in the representation but is misaligned, less active, or harder to decode. Our framework is diagnostic, analyzing how past task information evolves within the model.

Problem setup. Let $\mathcal{D}^t = \{(x_i^t, y_i^t)\}_{i=1}^{N_t}$ denote the dataset of task t. Let f_{θ_t} denote the continual model's feature extractor after training on task t, and $f_{\theta_{t+s}}$ the same extractor after further training on tasks $t+1, \ldots, t+s$. We study representations of task t data under these two models:

$$h_t(x) = f_{\theta_t}(x), \quad h_{t+s}(x) = f_{\theta_{t+s}}(x), \quad x \in \mathcal{D}^t. \tag{1}$$

The goal is to analyze how the representation of task t evolves after subsequent training and to what extent task-relevant information remains accessible.

Linear translation of representations. Continual training may introduce geometric shifts in the representation space without necessarily removing the information. To study this behavior, we introduce a linear translation mapping:

$$T_{t+s\to t}(h) = W_{t+s}h + b_{t+s}, \tag{2}$$

trained to align $h_{t+s}(x)$ with $h_t(x)$ using task t training data:

$$\min_{W,b} \mathbb{E}_{x\sim\mathcal{D}^t_{\text{train}}} \left\| W h_{t+s}(x) + b - h_t(x) \right\|_2^2 . \tag{3}$$

The learned mapping $T_{t+s\to t}$ is then applied to t test data. We use linear mappings to match standard readouts in continual learning, enabling recovery analysis under typical decision heads. This choice is supported by prior work showing that forgetting can be mitigated by retraining linear heads [6], that task representations may remain linearly related [15], and aligns with the linear representation hypothesis [10, 2, 26]. If task-t information can be recovered via a linear mapping, forgetting may reflect representational change rather than complete erasure.

Concept proxies via sparse autoencoders. To analyze forgetting with finer granularity than task accuracy or global representation similarity, for each task t, we train an SAE on the train split of $h_t(x)$, and use the trained SAE's encoder to obtain different test representations:

$$z_t(x) = \text{SAE}_t(h_t(x)), \quad z_{t+s}(x) = \text{SAE}_t(h_{t+s}(x)), \quad z_T(x) = \text{SAE}_t(T_{t+s\to t}(h_{t+s}(x))), \tag{4}$$

where $z(x) \in \mathbb{R}^K$ denotes latent activations. The SAE trained on task t defines a fixed coordinate system – an anchor for concept-level analysis to track how specific representational components evolve across training. Individual task-anchored latent features serve as **concept proxies**. This interpretation is motivated by evidence that SAEs can extract features that are more interpretable and monosemantic than individual neurons or alternative representation bases [19]. However, sparsity alone does not guarantee monosemanticity, which requires empirical assessment [28, 16, 13]. In our setting, quantitative quality assessment in App. L and qualitative examples showing shared visual motifs in App. J support the use of SAE latents as proxies for coherent, fine-grained patterns. Strong task-level concept-based probe performance further supports their task relevance. We therefore refer to these features as **concepts**, without assuming that each latent corresponds to a ground-truth semantic entity (see App. A for a detailed definition).

Identifying active concepts through binarization of activations. Following recent work showing that binarization of SAE features can provide an effective alternative representation [14, 1], and to focus on salient and reusable features, we introduce a frequency-based binarization rule to assess whether the concept is **active**. For each latent dimension k, we compute its mean activation over task t training data μ_k and use it to define a mean-based binarization rule [1]:

$$a_k(x) = \mathbb{I}[z^k(x) > \mu_k]. \tag{5}$$

Then, a concept k is considered **active** for task t and a frequency threshold τ if

$$\frac{1}{|\mathcal{D}^t_{\text{test}}|} \sum_x a_k(x) \geq \tau, \tag{6}$$

as consistently and strongly activated, corresponding to a likely informative and reusable feature. Therefore, we consider: raw features $h(x)$, continuous $z(x)$ and binarized $a(x)$ latent activations.

Analysis modules.
(1) Binary concept activation analysis. We analyze the structure of the binary activation space by measuring: i) the number of *active concepts*, ii) the *deletion ratio* (fraction of concepts active at the model at time t but inactive at $t+s$; by subtracting this value from 1, we obtain the ratio of retained concepts), iii) *regained concept count* after applying $T_{t+s\to t}$ and *regained activation mass*, defined as the fraction of lost activation magnitude on seemingly deleted concepts that is restored by the linear translation T:

$$R_{\text{mass}} = \frac{\sum_{k\in\mathcal{D}} \max(0, M^k_{T(t+s)} - M^k_{t+s})}{\sum_{k\in\mathcal{D}} \max(0, M^k_t - M^k_{t+s})}, \tag{7}$$

where $\mathcal{D}$ denotes the set of seemingly deleted concepts, and M^k_t, M^k_{t+s}, and $M^k_{T(t+s)}$ denote the mean activation magnitudes of concept k at task t, after subsequent training, and after linear translation, respectively. These metrics quantify how concept-level information is preserved, potentially lost, or recoverable.

(2) Concept prediction from raw features. We assess concept-level decodability by predicting binary concept activations from raw features. For each concept k, we define a binary label:

$$y_k(x) = \mathbb{I}[z^k_t(x) > 0], \tag{8}$$

and train a classifier to predict $y_k(x)$ from $h_{t+s}(x)$. We use a less strict threshold than in the activity analysis, since the selected latent features have already been identified as salient concept proxies at the dataset level and to reduce sparse space class imbalance. We evaluate performance using balanced accuracy and F1 score on the $t + s$ test splits. High values indicate that concept-level information remains decodable from the representation. Mean scores reflect overall concept preservation, while their distribution reflects that concepts can have different decodability levels.

(3) Task prediction via linear probes. We train linear classifiers on task t data after task t training to predict task labels using: (i) raw features $h_t(x)$, (ii) concept activations $z_t(x)$. We evaluate these trained probes on test t data at t, $t+s$, and $T_{t+s\to t}$ using accuracy. This measures how well task-level information remains linearly decodable when using raw features and concept-level features, and whether translation can fix some of the linear readout issues.

The concept taxonomy. We distinguish five possible outcomes for a task-t concept proxy after training on task $t + s$: ***Retained***, when the concept is active at t and remains active in the raw $t + s$ representation; ***Seemingly deleted***, when it was active at t but is absent in raw $t + s$ under the binarization rule; ***Recovered***, when it is absent in raw $t + s$ but reappears after translation; ***Decodable***, when it does not reappear as an active concept, but its presence can be predicted from the representation ($\text{F1} > 0$); and ***Lost***, when it is neither recovered by translation nor decodable ($\text{F1} = 0$). ***Decodable*** concepts do not reactivate but remain predictable (non-zero performance), whereas ***Lost*** concepts are neither recoverable nor decodable. Among decodable concepts, balanced accuracy and F1 quantify recoverability, with higher values indicating greater accessibility (see Fig. 1).

4 Experimental setup

Our experimental design aims to evaluate the accessibility-based formulation of forgetting introduced in Section 3. Our **continual learning setting** includes evaluation of our framework on 3 task-incremental benchmarks: i) **2seq-CIFAR10**, splitting CIFAR-10 [23] into 2 tasks, 5 classes each; ii) **2seq-tiny-ImageNet**, using the first 40 classes of tiny-ImageNet [7], divided into 2 tasks with 20 classes each, enabling a coarse-grained comparison with 2seq-CIFAR10; iii) **10seq-tiny-ImageNet**, using a split of tiny-ImageNet into 10 tasks with 20 classes each. We leverage four representative continual learning strategies: i) **SGD**, corresponding to naive fine-tuning without knowledge retention, serves as a baseline for natural continual-learning behavior; ii) Learning without Forgetting (**LWF**) [24], a distillation-based strategy; iii) Elastic Weight Consolidation (**EWC**), a regularization-based strategy [21], and iv) Dark Experience Replay ++ (**DER++**) [5], a hybrid strategy combining experience replay with knowledge distillation and regularization. As the backbone, we use Resnet18 [17], one of the most popular models in continual image classification [3, 25]. To ensure reproducibility, we use implementations, data splits and hyperparameters from *Mammoth*[1] [4]. Appendix B reports Task-Incremental accuracies: all strategies achieve relatively strong performance on new tasks, while exhibiting different forgetting levels, making them suitable for our analysis.

In our **concept-level analysis setup**, for each task, we train an SAE with `BatchTopKSAE` regime via `overcomplete` [2] on the same Mammoth's train/test splits. We use Mean Squared Error (MSE) with an `overcomplete`'s dead-neuron reactivation loss weighted by 10^{-2}. Across experiments, manual inspection showed reconstruction $R^2 > 0.6$ and close to 0% dead-neuron rate, indicating

[1] `https://github.com/aimagelab/mammoth`

[2] `https://github.com/KempnerInstitute/overcomplete`

meaningful and non-degenerate SAEs. We also examined top 9 activating images for selected SAE latents and found consistent shared visual motifs, indicating coherent patterns rather than arbitrary feature mixtures (App. J). App. L shows that active SAE latents can achieve substantially higher monosemanticity scores (MS) [28] than randomized baselines, quantitatively supporting their quality and coherence as concept proxies. We use a single linear layer for translation, and Logistic Regression for task-level and concept-level probes. Though nonlinear mappings are less constrained and relevant under standard readouts, we additionally evaluate a simple nonlinear translator in App. I, which yields no significant gains. We verified the analysis robustness across 10 runs (App. D), 6 frequency-based binarization thresholds (App. E), 5 SAE batch sizes (App. H), and 4 SAE K values (App. G). The results show consistent behavior of key metrics and support our conclusions' robustness. Additional experiments with a Vision Transformer (ViT) [8] backbone (App. N) further support the robustness of our findings across architectures. We validate that the selected active neurons retain most task-relevant information through SAE-based deletion experiments on non-active neurons under different parameters and binarization thresholds (App. E, F, G, H). We used NVIDIA GH200 GPUs for continual training and Titan RTX GPU for SAE training and evaluation. Average execution time for an example task and 10 runs was 204.199 ± 42.473 s on the Titan GPU. We report hyperparameter values in App. K. Our code is available at `https://github.com/lifelonglab/lost-or-hidden-neurips-2026`.

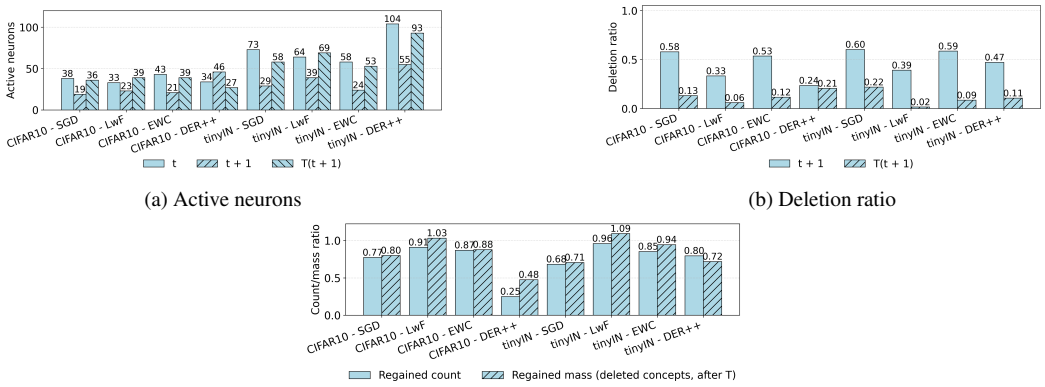
5 Experimental analysis

Our experiments follow the *analysis modules* defined in Section 3: (1) Binary concept activation analysis, (2) Concept prediction from raw features and (3) Task prediction via linear probes.

5.1 Concept retention, deletion and recoverability: concept activation analysis

Fig. 2a shows the number of active neurons for 2seq-CIFAR10 and 2seq-tiny-ImageNet. Under the same binarization rule, substantially more latent neurons are active for 2seq-tiny-ImageNet, likely reflecting its higher complexity. In all cases, only a small fraction of the SAE latent space is active, corresponding to salient, reusable features. The number of active neurons decreases when task-t data are encoded after training on task $t+1$ for most setups, but a large fraction is recovered after linear translation. An exception is 2seq-CIFAR10 with DER++, where the number of active concepts increases at $t+1$, possibly due to replay-driven expansion of the active feature set. Fig. 2b reveals clear differences between continual strategies: deletion ratios are the highest for SGD and EWC, lower for DER++ and LwF. Linear translation noticeably improves deletion ratios for 2seq-tiny-ImageNet, especially for LwF, where deletion becomes almost zero. Fig. 2c shows that the regained count ratio and the regained mass are the highest for LwF, followed by EWC. It suggests that these strategies may preserve a more linearly recoverable fine-grained space. Overall, these results indicate that many apparently forgotten concepts are not irreversibly lost, but become less readable in the original task-t coordinate system due to approximately linear representational drift. As a qualitative complement, Appendix J compares the most activating images for example **retained** concepts at t and $t+1$ (Figs. J.1, J.2, J.3). Shared visual motifs indicate overlapping information across the two representations. We also show **recovered** concepts (Figs. J.4, J.5, J.6), whose common motifs further support that concept-level information can be linearly restored.

Figure 3 shows deletion ratios for each task t after all subsequent tasks $t+s$ on 10seq-tiny-ImageNet. For raw representations, deletion ratios increase with s under all strategies, with consistently higher values for SGD and EWC. Under LwF and DER++, later tasks start from lower deletion levels (at $t+1$), suggesting improved resistance to degradation. For DER++, deletion increases progressively with s, although with a decelerating trend over time (log-like). Across all tasks, linear alignment substantially reduces deletion ratios, indicating that much of the drift is partially reversible. The strength of this reduction is weaker for DER++ than for LwF. The regained count ratios in Fig. 4 follow similar patterns as in the bar plots: they are the highest for LwF and EWC, comparable for SGD and DER++. Increases in deletion rates at higher s and slight decreases in regained counts (SGD, DER++, EWC), suggest gradually reduced linear recoverability of concept activations.

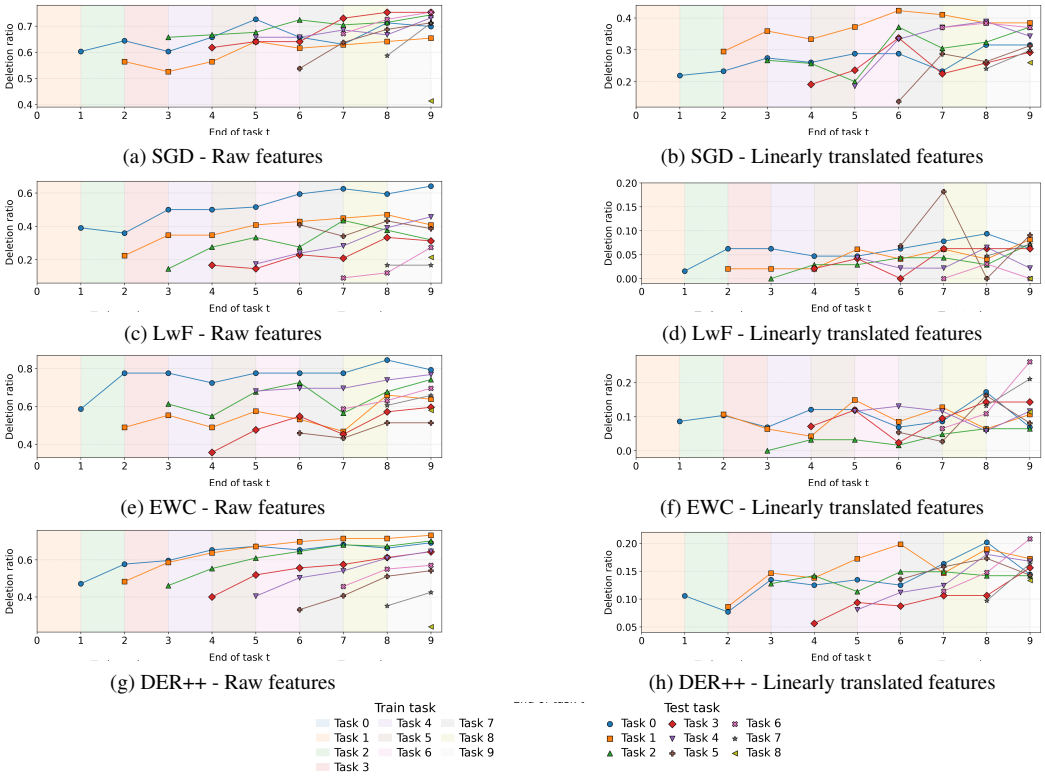


(a) Active neurons

(b) Deletion ratio

(c) Regained concept statistics

Figure 2: Concept activation analysis: **Active concept count**, (seeming) **deletion ratio** and **regained concept statistics** for 2seq-CIFAR10 and 2seq-tiny-ImageNet.



(a) SGD - Raw features

(b) SGD - Linearly translated features

(c) LwF - Raw features

(d) LwF - Linearly translated features

(e) EWC - Raw features

(f) EWC - Linearly translated features

(g) DER++ - Raw features

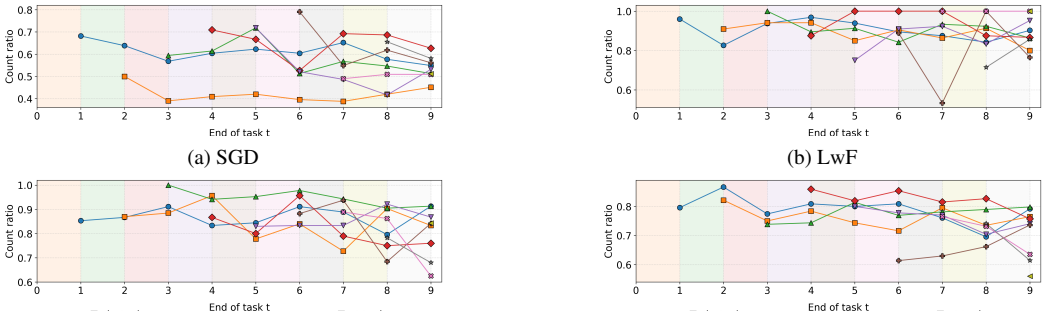
(h) DER++ - Linearly translated features

(i) Common legend

Figure 3: **Deletion ratio** for all tasks of 10seq-tiny-ImageNet throughout the continual training.

5.2 Concept decodability: concept prediction from raw features

Fig. 5a reports the mean balanced accuracy and F1 of linear classifiers trained to predict concepts that become inactive in raw task-t representations after learning task $t+1$, with the corresponding score distributions shown in Figs. 5c (F1) and 5b (balanced accuracy). The results show that LwF and EWC preserve substantially higher concept decodability at $t+1$ than SGD and DER++, suggesting that, beyond coarse-grained class information [6], finer-grained concept-level information can also remain accessible, especially under forgetting mitigation with strategies differing in decodability

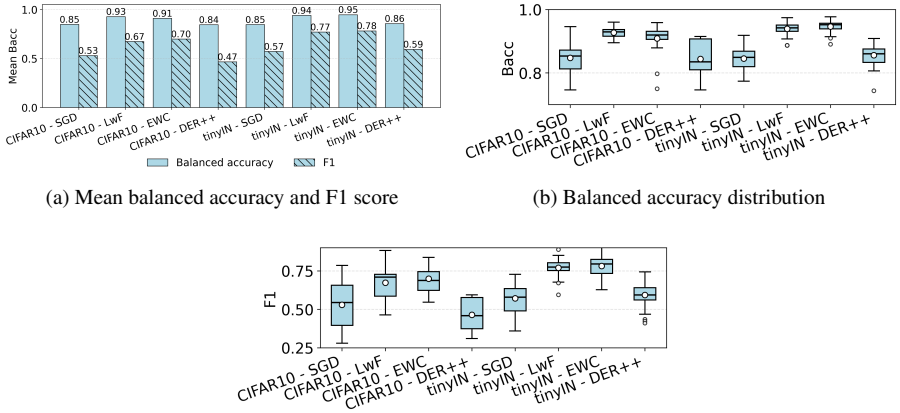


(a) SGD (b) LwF

(c) EWC (d) DER++

Figure 4: **Regained concept count ratio** for all tasks of 10seq-tiny-ImageNet throughout the continual training after the linear translation. We use the same legend as in Fig. 3.

preservation (LwF, EWC being superior to DER++). At the same time, the spread of balanced accuracy and F1 across concepts indicates that the recoverability is non-uniform: some concepts remain easily decodable, whereas others degrade strongly. For LwF and EWC, the distributions are shifted toward higher values and are more compact, further indicating their broader and more consistent preservation of information at the level of concept proxies. Overall, these results suggest that fine-grained forgetting is often selective and partial rather than uniform or complete.

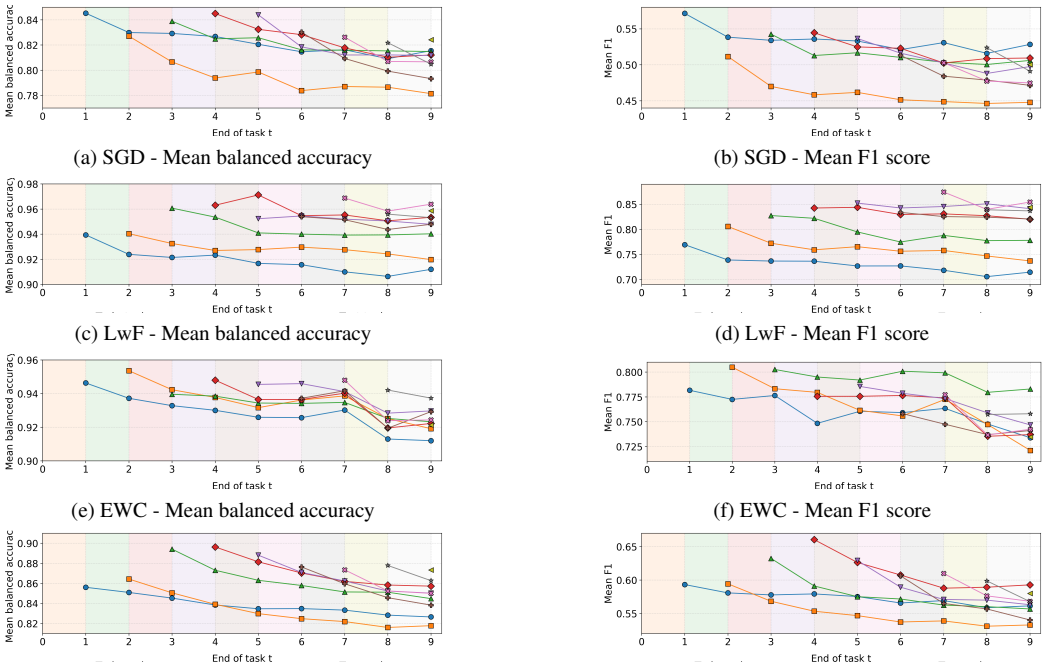


(a) Mean balanced accuracy and F1 score (b) Balanced accuracy distribution

(c) F1 score distribution

Figure 5: Mean **balanced accuracy** and **F1 score** along with their distributions for the concept prediction probe on 2seq-CIFAR10 and 2seq-tiny-ImageNet.

Fig. 6 shows the evolution of mean balanced accuracy and F1 of concept predictions throughout all future continual tasks for all tasks. These plots also show significant differences: for SGD and DER++ both balanced accuracy and F1 drop with subsequent training tasks. For LwF it is almost constant (only some small drops appear) showing that this forgetting-aware strategy preserves task information also at the more fine-grained level of concepts. As the ground truth labels are constant (task t data at t determines them), this is true decodability degradation caused by possible information loss or increasing readout issues at the level of individual concepts. Our analysis did not identify fully **lost** neurons at the strictest possible criterion, i.e., $\mathrm{F1} = 0$, suggesting that concept-level information was at most partially lost rather than entirely erased in our setup. As a qualitative complement, in Appendix J we compare the top activating images of selected **decodable** concepts (see Figs. J.7, J.8). Concepts with varying decodability (F1, balanced accuracy) still share motifs across t and $t+1$, supporting that at least some part of concept-level information can remain decodable.



(a) SGD - Mean balanced accuracy

(b) SGD - Mean F1 score

(c) LwF - Mean balanced accuracy

(d) LwF - Mean F1 score

(e) EWC - Mean balanced accuracy

(f) EWC - Mean F1 score

(g) DER++ - Mean balanced accuracy

(h) DER++ - Mean F1 score

Figure 6: **Mean balanced accuracy** and **mean F1 score** for the **concept prediction probe** on 10seq-tiny-ImageNet for all tasks throughout continual training. We use the same legend as in Fig. 3.

5.3 Task prediction via linear probes

Fig. 7 reports task 0 accuracy for 2seq-CIFAR10 and 2seq-tiny-ImageNet, using probes trained on continual model features and SAE latents. As expected, the largest drops at $t + 1$ are observed for SGD and EWC. Linear translation restores most performance. However, incomplete recovery for all strategies except LwF indicates that degradation cannot be explained solely by linear misalignment, pointing to residual readability limitations and partial loss of task information. Importantly, the recovery of accuracy for latent-based probes suggests that task-relevant information remains largely compatible with the original task-anchored concept basis after alignment, and thus provides functional evidence that the translation restores access to fine-grained information at the level of the analyzed concept proxies. This further indicates that a large portion of fine-grained information is not fully deleted due to forgetting, but becomes misaligned and less directly accessible. The remaining performance gap after alignment implies, however, partial degradation of concept-level structure, consistent with the observed variability in F1 scores for concept prediction.

To further understand the temporal dynamics of probe performance, we provide an extended analysis across all tasks t after $t + s$ of 10seq-tiny-ImageNet in Appendix C (Fig. C.1 – continual model features-based probes, Fig. C.2 – concept-based probes). Accuracy at $t + s$ without translation declines as s increases for all strategies, with the strongest effect on SGD and EWC. Operating on SAE latents seems to stabilize the probe accuracies on non-translated data for LwF and DER++. Translation recovers much of the performance at all s, thus information recoverability persists over time. However, it slightly decreases as s grows for some strategies (SGD, DER++), further suggesting that stronger forgetting may involve more advanced drifts or partial information erasure.

6 Conclusion

In this paper, we proposed an SAE-based framework that defines a fixed, more disentangled concept space to analyze forgetting in continual learning through task-specific knowledge preservation and recoverability under a linear readout assumption. Our results show that forgetting mitigation strategies affect not only the magnitude but also the nature of forgetting. Methods such as LwF keep representational drift largely linear, making class- and concept-level information more decodable



(a) Representation-level

(b) Concept-level

Figure 7: **Accuracy of probes** predicting 2seq-CIFAR10 and 2seq-tiny-ImageNet task 0 classes from the continual model features (representation-) and SAE latent activations (concept-level).

and reversible. In contrast, other strategies induce less recoverable distortions. Overall, in the studied settings, much apparent forgetting reflects approximately linear representational changes, with seemingly deleted information often recoverable through linear remapping. Importantly, our results suggest that knowledge preservation extends beyond coarse-grained class information to a more fine-grained level. At the same time, our results indicate that some part of the concept-level information becomes no longer fully accessible in a linear setting. Under the common assumption of a linear readout as a classifier, this leads to a practical loss of decodability, which can be functionally equivalent to forgetting. Thanks to the observation of the distributions of concept-prediction scores we find that concept-level forgetting is not uniform. At the same time, our results suggest that continual learning strategies preserve past knowledge at different levels: some retain their functionality more at the level of outputs (DER++), whereas others better preserve it at the level of the internal organization of concept proxies (LwF, EWC). Overall, our findings suggest that, in supervised continual learning, past fine-grained structure of model knowledge and computations are often not lost but hidden: they may remain linearly recoverable, although their preservation depends on the continual learning strategy. This emphasizes the importance of preserving such structures for interpretable continual learning. As improving transparency and trustworthiness of continual models is a positive societal impact, future work will extend our framework to broader settings and problems, as well as design strategies that better preserve fine-grained task-relevant information.

Limitations. One may view the use of SAE latent features in a concept-level analysis as a limitation, since they are concept proxies rather than definitive human-interpretable concepts or ground-truth semantic entities. However, they provide an automatic, label-free, and computationally feasible approximation of more disentangled, fine-grained structure in the continual model's representation space, supported by strong task-level concept-based probe performance, quantitative quality assessment in App. L, and qualitative examples in App. J showing clear shared motifs. Nevertheless, these results do not establish a one-to-one correspondence between SAE features and semantic concepts. Our taxonomy relies on an operational definition of concept loss. We use $\mathrm{F1} = 0$ to define *Lost* concepts, indicating zero measured decodability under the evaluated linear readout. This does not establish that the underlying information has been erased, nor does a low but nonzero $\mathrm{F1}$ imply useful preservation. Residual decodability might be too weak to support reliable readout and can be functionally equivalent to partial concept loss. Fitting linear translators requires access to old-task data. This requirement is compatible with our framework's purpose as a post-hoc diagnostic tool, but limits its direct use as a replay-free continual learning strategy. Future methods informed by our findings would need to remove this dependence. Smaller data subsets may offer a practical trade-off between diagnostic cost and quality when computational resources or analysis time are limited. App. M shows that smaller old-task subsets can support concept-recovery diagnostics, with slightly reduced reconstruction quality and recovery performance. Finally, a potential negative societal impact is that concept-level analyses may increase confidence in deploying continual models in high-stakes settings, even though the identified concepts are proxies and may miss sensitive or safety-critical information.

Acknowledgments

We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2025/018871. This research was supported by the START Scholarship of the Foundation for

Polish Science (FNP) for outstanding young scholars, agreement No. START 017.2025 and the Polish Minister of Science and Higher Education scholarship for outstanding young researchers, agreement No. SMN/20/1546/2024

References

[1] D. Aswal and C. Hudelot. ConceptGuard: Neuro-Symbolic Safety Guardrails via Sparse Interpretable Jailbreak Concepts, 2025.

[2] L. Bereska and E. Gavves. Mechanistic Interpretability for AI Safety – A Review. *arXiv preprint arXiv:2404.14082*, 2024.

[3] A. Bian, W. Li, H. Yuan, C. Yu, M. Wang, Z. Zhao, A. Lu, P. Ji, and T. Feng. Make Continual Learning Stronger via C-Flat. *Advances in Neural Information Processing Systems*, 37:7608–7630, 2024.

[4] M. Boschini, L. Bonicelli, P. Buzzega, A. Porrello, and S. Calderara. Class-Incremental Continual Learning into the eXtended DER-verse. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[5] P. Buzzega, M. Boschini, A. Porrello, D. Abati, and S. Calderara. Dark Experience for General Continual Learning: a Strong, Simple Baseline. *Advances in Neural Information Processing Systems*, 33:15920–15930, 2020.

[6] M. Davari, N. Asadi, S. Mudur, R. Aljundi, and E. Belilovsky. Probing Representation Forgetting in Supervised and Unsupervised Continual Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16712–16721, 2022.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[9] J. Dunefsky, P. Chlenski, and N. Nanda. Transcoders Find Interpretable LLM Feature Circuits. *Advances in Neural Information Processing Systems*, 37:24375–24410, 2024.

[10] N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

[11] K. Faber, K. Filus, and Ł. Pawela. Sparse autoencoders can induce local geometry for mitigating catastrophic forgetting. *arXiv preprint arXiv:2610.22439*, 2026.

[12] K. Faber, C. Kanan, V. Lomonaco, and R. Corizzo. Continual anomaly detection: A comprehensive survey and research roadmap. *Preprints*, 2026.

[13] K. Filus and S. Pokuciński. Measuring monosemanticity in sparse autoencoders via latent activation coherence. *arXiv preprint arXiv:2607.17770*, 2026.

[14] J. Gallifant, S. Chen, K. Sasse, H. Aerts, T. Hartvigsen, and D. Bitterman. Sparse Autoencoder Features for Classifications and Transferability. In *Conference on Empirical Methods in Natural Language Processing*, pages 29927–29951, 2025.

[15] Q. Gu, D. Shim, and F. Shkurti. Preserving linear separability in continual learning by backward feature projection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24286–24295, 2023.

[16] R. Härle, F. Friedrich, M. Brack, B. Deiseroth, S. Waeldchen, P. Schramowski, and K. Kersting. Measuring and guiding monosemanticity. *Advances in Neural Information Processing Systems*, 38:167468–167508, 2026.

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[18] J. Hu, Z. Lian, Z. Wen, C. Li, G. Chen, X. Wen, B. Xiao, and M. Tan. Continual Knowledge Adaptation for Reinforcement Learning. *arXiv preprint arXiv:2510.19314*, 2025.

[19] R. Huben, H. Cunningham, L. R. Smith, A. Ewart, and L. Sharkey. Sparse Autoencoders Find Highly Interpretable Features in Language Models. In *International Conference on Learning Representations*, 2024.

[20] R. Kemker, M. McClure, A. Abitino, T. Hayes, and C. Kanan. Measuring catastrophic forgetting in neural networks. In *AAAI Conference on Artificial Intelligence*, volume 32, 2018.

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

[22] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton. Similarity of Neural Network Representations Revisited. In *International Conference on Machine Learning*, 2019.

[23] A. Krizhevsky, G. Hinton, et al. Learning Multiple Layers of Features from Tiny Images. 2009.

[24] Z. Li and D. Hoiem. Learning without Forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017.

[25] M. Masana, X. Liu, B. Twardowski, M. Menta, A. D. Bagdanov, and J. Van De Weijer. Class-incremental learning: survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5513–5533, 2022.

[26] S. Masip, G. M. van de Ven, J. Ferrando, and T. Tuytelaars. Putting a Face to Forgetting: Continual Learning meets Mechanistic Interpretability. *arXiv preprint arXiv:2601.22012*, 2026.

[27] E. Ning, W. Xue, D. Lou, and Y. Guo. From weights to features: Sae-guided activation regularization for llm continual learning. *arXiv preprint arXiv:2606.26629*, 2026.

[28] M. Pach, S. Karthik, Q. Bouniot, S. Belongie, and Z. Akata. Sparse Autoencoders Learn Monosemantic Features in Vision-Language Models. *Advances in Neural Information Processing Systems*, 2025.

[29] 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.

[30] D. Rolnick, A. Ahuja, J. Schwarz, T. Lillicrap, and G. Wayne. Experience Replay for Continual Learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[31] A. Soutif-Cormerais, A. Carta, A. Cossu, J. Hurtado, V. Lomonaco, J. Van de Weijer, and H. Hemati. A Comprehensive Empirical Evaluation on Online Continual Learning. In *IEEE/CVF International Conference on Computer Vision*, pages 3518–3528, 2023.

[32] E. Verwimp, R. Aljundi, S. Ben-David, M. Bethge, A. Cossu, A. Gepperth, et al. Continual Learning: Applications and the Road Forward. *Transactions on Machine Learning Research*, 2024.

[33] L. Wang, X. Zhang, H. Su, and J. Zhu. A Comprehensive Survey of Continual Learning: Theory, Method and Application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5362–5383, 2024.

[34] M. Zhang, Y. Li, L. Li, D. Shen, H. Xiong, and Y. Sun. Sae-fd: Sparse autoencoder feature distillation for continual learning of large language models. *arXiv preprint arXiv:2605.25525*, 2026.

A Concept definition

In this work, we understood a **concept** as a computational representation of a shared, recurring pattern that helps a model structure information. Methods from mechanistic interpretability such as Sparse Autoencoders (SAEs) can extract such representations via self-supervised learning, often providing more coherent patterns than standard continual models. Analogous to human perception, in which recognizing an object (e.g. a dog) relies on features like ears, a nose, or paws, model concepts capture recurring patterns in data. While they do not fully align with human semantic concepts and remain approximate (see Fig. A.1), potentially less coherent representations, they provide a practical and well-established way to model abstract, fine-grained information, serving as useful **concept proxies**, to which we refer to as **concepts**. Studying whether these concepts persist in continual learning is equally as important as accuracy, as it reveals whether the model truly retains knowledge and preserves a consistent way of organizing information and computations, rather than merely sustaining performance through shifting internal representations. These all can be also connected to whether the model maintains stable and trustworthy interpretations, which is crucial for explainability and trustworthiness in artificial intelligence.



Figure A.1: SAE latents provide cleaner, more concept-aligned representations than raw model features, and are therefore used here as proxies for concepts. i) Feature from a CL model corresponds to a single neuron in the representation layer in CL model; ii) SAE allows us to build a concept proxy, disentangling features; iii) ground truth semantic concept in real-world

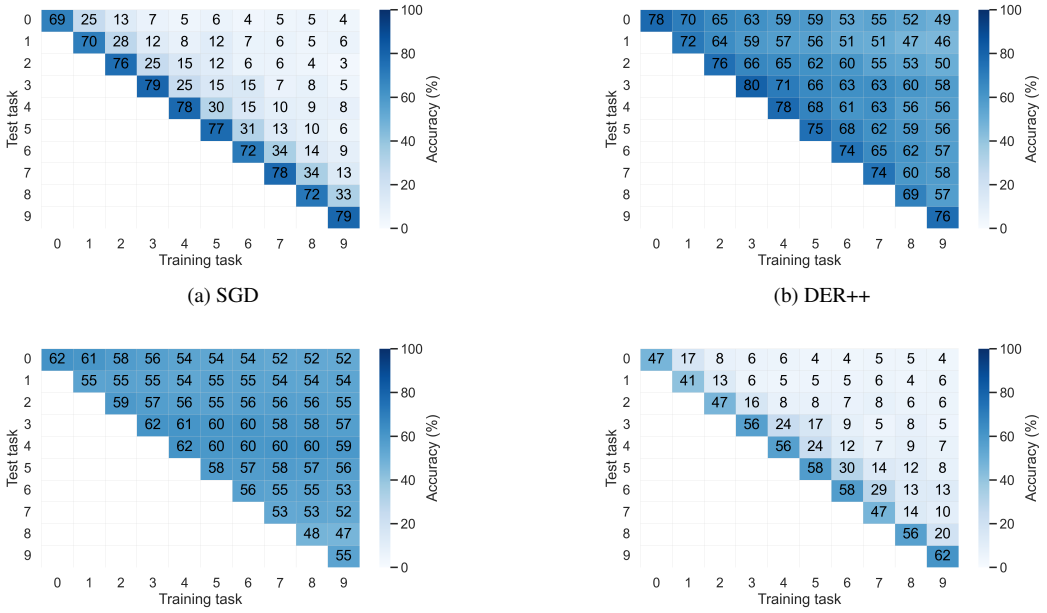
B Accuracy of the original continual models' classification heads

Table B.1 presents the accuracy results for both tasks for the 2seq-CIFAR10 case. After learning the second task, performance on the first task drops significantly for SGD, and notably less for DER++ and LwF, while accuracy on the new task remains similarly high for SGD, LwF and DER++ (96.4 for SGD, 96.5 for DER++, 95.3 for LwF), and slightly lower for EWC (89). This suggests that on CIFAR10, all methods learn the new task well, but forgetting mitigation mainly affects retention of the old one, with LwF preserving past-task accuracy best, DER++ offering an intermediate preservation, and SGD, EWC forgetting the most.

Table B.1: Task-IL accuracy (%) on 2seq-CIFAR10. For task 0, we report the accuracy after learning task 0. For task 1, we report the accuracies on task 0 and task 1 after learning task 1.

Method	**Task 0 after** t=0	**Task 0 after** t=1	**Task 1 after** t=1
SGD	93.30	66.94	96.36
DER++	93.42	85.58	96.48
LwF	91.36	90.82	95.32
EWC	72.34	46.08	88.96

We present the accuracy matrices for all tasks under the Task-IL setting for 10seq-tiny-ImageNet in Fig. B.1. The four examined strategies resulted in expected behaviors. SGD achieves reasonable accuracy on the currently learned task, but older-task performance drops sharply as training proceeds – as an indication of strong forgetting. DER++ provides the best overall Task-IL accuracy, with noticeably better retention of past tasks than SGD while still maintaining strong performance on newer tasks. In contrast, LwF is more uniform across tasks, but at a lower accuracy level, suggesting better stability at the cost of plasticity and overall per-task performance.



(a) SGD

(b) DER++

(c) LwF

(d) EWC

Figure B.1: Accuracy of the original continual models' classification heads in the task incremental learning setup on 10seq-tiny-ImageNet.

C Evolution of task-level linear probes performance when trained on raw features and latent features

Figures C.1 and C.2 show the temporal dynamics of feature- and concept-based probe accuracy for task t, measured on task-t data after subsequent tasks $t + s$. In all cases, accuracy on raw task-t representations declines as s increases, but with different dynamics: under SGD and EWC it drops quickly and stabilizes only slightly above 0, whereas under DER++ and LwF it remains much higher. DER++ preserves task performance in non-translated data better than LwF, likely due to combining replay and distillation. For LwF, accuracy after linear translation stays nearly constant over time, indicating that the apparent forgetting is largely reversible by a simple linear remapping. For SGD, EWC and DER++, translation recovers much of the lost performance, but less so as s grows, further suggesting that stronger forgetting may involve more advanced drifts or partial information erasure. Similar trends both at the concept level and the level of continual model entangled features, show that these effects persists even in sparse, more disentangled representations. Additionally, operating at the concept level appears to slow down the decline for DER++ and LwF relative to the raw-feature probe. Importantly, the recovery of accuracy for probes operating on SAE latents suggests that task-relevant information remains largely compatible with the original task-anchored concept basis after alignment. This indicates that a significant portion of fine-grained, concept-level information is not fully deleted during continual learning, but instead becomes misaligned and less directly accessible. Moreover, this recovery provides functional evidence that the linear translation of the continual model features restores access to fine-grained concept-based information.

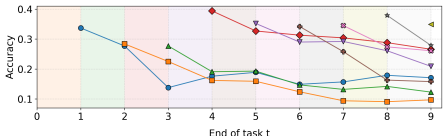
D Stability of concept-based forgetting analysis across runs

Box plots of key metric values across 10 runs for an example configuration (2seq-tiny-ImageNet, LwF) presented in Fig. D.1 show that the results are largely stable across runs. The probe-based measures, especially the representation- and concept-level probe accuracies and the mean balanced accuracy and F1 for concept prediction, show the most compact distributions, indicating high consistency of the main conclusions. Larger variability is observed for the regained count and mass ratio. This suggests that the exact number of regained concepts may vary between runs, but the overall conclusion that linear alignment restores a substantial fraction of the concept information remains robust.

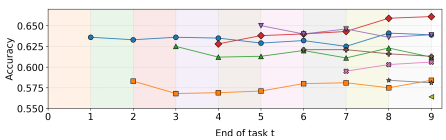


(a) SGD - Raw features

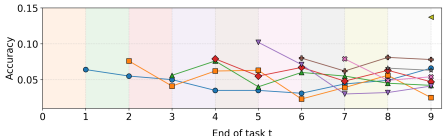
(b) SGD - Linear translation



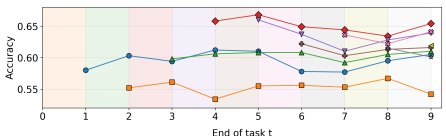
(c) LwF - Raw features



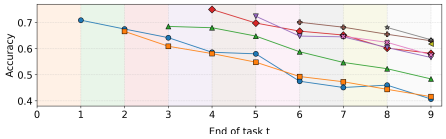
(d) LwF - Linear translation



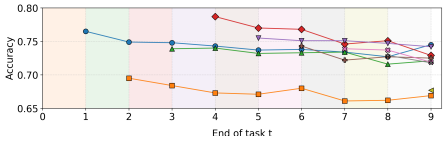
(e) EWC - Raw features



(f) EWC - Linear translation



(g) DER++ - Raw features



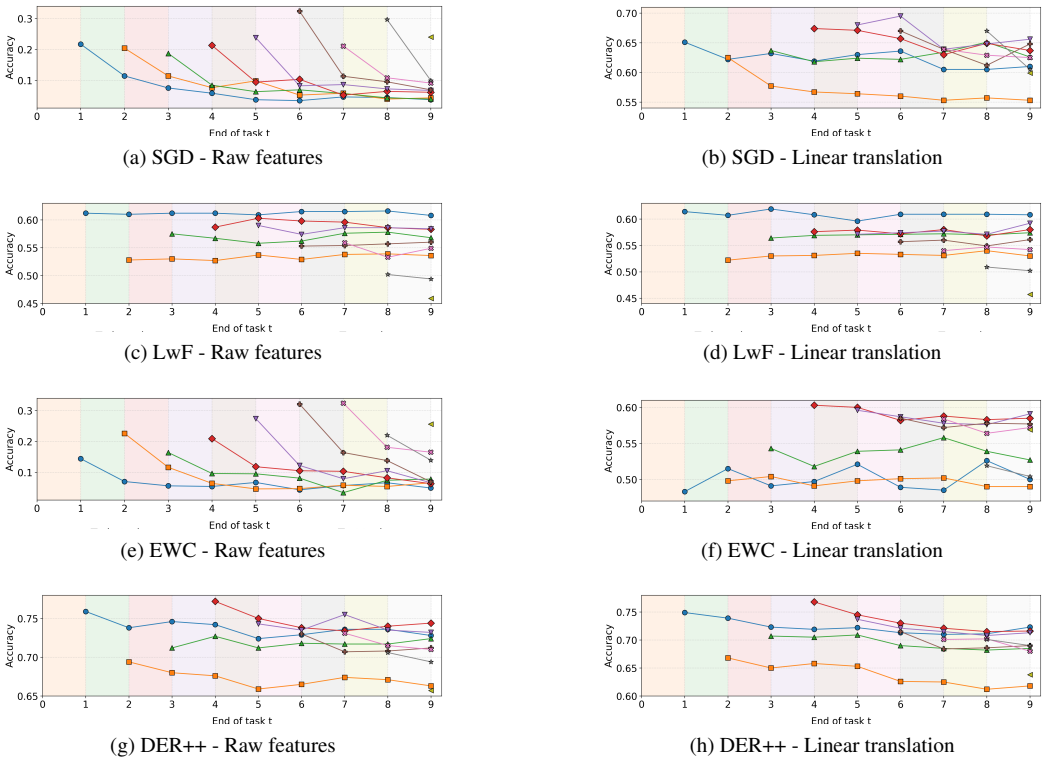
(h) DER++ - Linear translation

Figure C.1: **Accuracy of probes** for 10seq-tiny-ImageNet classes prediction throughout continual training based on the **raw representations** from the continual model (representation-level). Accuracy is higher in all cases after applying the linear translation to the raw features. We use the same legend as in Fig. 3.

E Impact of frequency-based binarization threshold on concept-based forgetting analysis

With the frequency-based binarization criterion, a latent is considered active if it exceeds its task-t train-set mean on at least a specified fraction of task-t samples. As shown in Fig. E.1 for the example LwF 2seq-tiny-ImageNet case, increasing this threshold reduces the number of detected active neurons, as could be predicted, but the overall forgetting picture remains largely unchanged. Across all thresholds, raw $t+1$ representations of t exhibit fewer active concepts and non-zero deletion, while linear alignment restores most of the lost activity and keeps both feature-level and concept-level probe accuracy close to the task-t reference. This indicates that our concept-forgetting observations are not an artifact of a particular binarization threshold, but are qualitatively robust across a range of frequency-based thresholds.

To verify that the set of binary active neurons is functionally relevant, we additionally measure task-0 classification accuracy after zeroing out, through the SAE decoder, all latent neurons that are not marked as active under the given frequency-based rule. Figure E.2b shows that this filtering leads only to a limited accuracy drop across the examined thresholds, confirming that the selected active neurons retain most of the task-level information. For 2seq-tiny-ImageNet under LwF, the performance remains relatively stable across thresholds, although slightly larger drops appear for thresholds above 0.05. This suggests that very permissive thresholds include less functionally relevant activations, while stricter thresholds can entail some loss of the task-relevant information. Therefore, thresholds around 0.05 provide a reasonable trade-off between retaining task information and enforcing a meaningful notion of concept activity. We therefore use 0.05 in all setups, including CIFAR10, to keep the analysis unified across datasets.



(a) SGD - Raw features

(b) SGD - Linear translation

(c) LwF - Raw features

(d) LwF - Linear translation

(e) EWC - Raw features

(f) EWC - Linear translation

(g) DER++ - Raw features

(h) DER++ - Linear translation

Figure C.2: **Accuracy of probes** for 10seq-tiny-ImageNet classes prediction throughout continual training based on the **SAE latent neuron activations** (concept-level). Accuracy is higher in all cases after applying the linear translation to the raw features. We use the same legend as in Fig. 3.

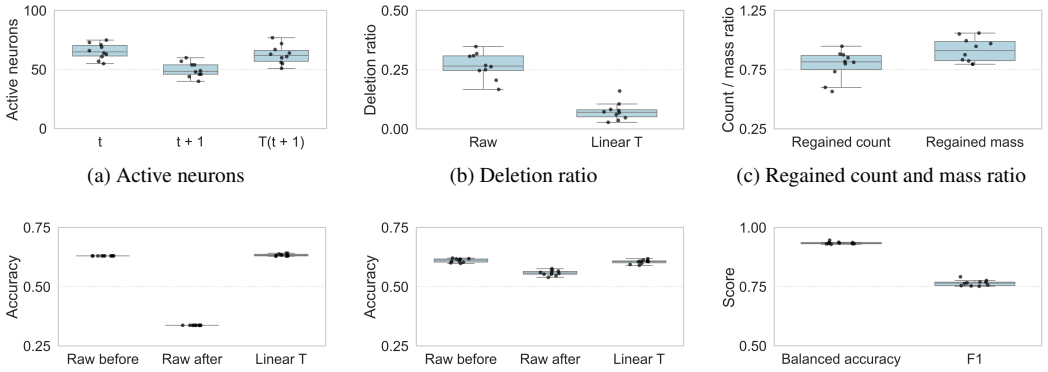
F Impact of 0.05 frequency-based binarization on probe-based task accuracy

As an additional check, to examine whether the active set of neurons under the frequency-based binarization rule used in our main experiments retains the functionally relevant task information for both datasets, we perform an experiment, in which we zero-out the neurons that are overall not active under the binarization rule (i.e. due to values lower than mean train activations or due to the frequency). We use the SAE's decoder for this intervention by first zeroing out the activations of the selected neurons, decoding the modified latent vector, and comparing it with the decoder output obtained from the raw latent activations. We then use this difference to update the original input representation, rather than replacing it directly with the decoded vector, in order to avoid propagating the SAE reconstruction error.

Fig. F.1 shows that using the frequency-based threshold of 0.05 preserves most of the probe accuracy across all configurations, both at the representation and concept level. This indicates that the set of neurons selected as active retains the majority of task-relevant information despite being only a subset of the full latent space. Overall, these results support the reliability of the adopted binarization rule as the basis for our concept-level analysis across all examined settings.

G Impact of SAE's K on concept-based forgetting analysis

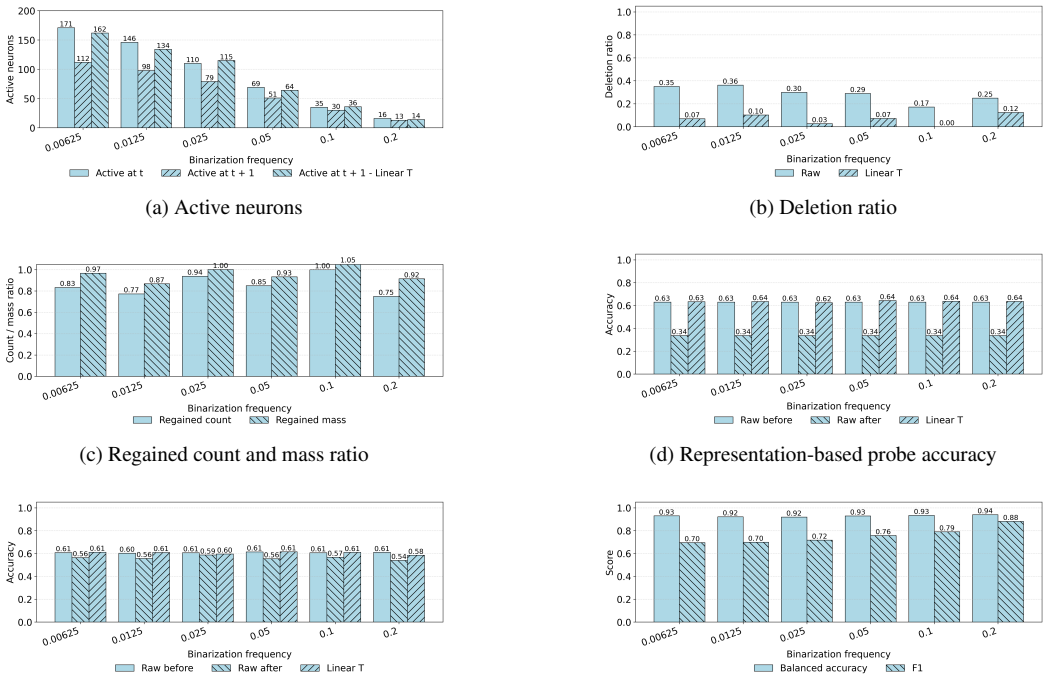
Fig. G.1 shows the impact of the sparsity-control parameter K used for the BatchTopK SAE on the key metrics for the example 2seq-tiny-ImageNet + LwF setup. Overall, the key results are qualitatively stable across the tested K values, indicating that our conclusions are not sensitive to this hyperparameter. The main difference lies in the number of active neurons, which is expected since parameter K directly controls the sparsity of the activations. Although some other small quantitative differences in other metrics can be observed, the overall patterns remain unchanged: raw task-$t+1$ representations show concept deletion, linear alignment recovers a large fraction of the



(a) Active neurons (b) Deletion ratio (c) Regained count and mass ratio

(d) Representation-based probe accuracy (e) Concept-based probe accuracy (f) Mean balanced accuracy and F1

Figure D.1: **Stability analysis**: box plots of the distribution of values of our key measures for the 2seq-tiny-ImageNet on the example LwF continual learning strategy case for 10 individual runs. For all measures bounded to the $[0, 1]$ range, the y-axes were rescaled to dataset-specific intervals of width 0.5 in order to facilitate visual comparison of the spread across plots



(a) Active neurons (b) Deletion ratio

(c) Regained count and mass ratio (d) Representation-based probe accuracy

(e) Concept-based probe accuracy (f) Mean balanced accuracy and F1

Figure E.1: The **impact of frequency-based binarization threshold** on metrics for 2seq-tiny-ImageNet task 0 under LwF.

lost information, and both task- and concept-level probe performance remain close to the reference values. Importantly, most values fall within the spread previously observed across repeated runs of the same method, suggesting that the effect of K is comparable to the run-to-run variability.

To verify that the set of binary active neurons is functionally relevant under different K values, we additionally measure task-0 classification accuracy after zeroing out, through the SAE decoder, all latent neurons that are not marked as active under the given frequency-based rule. Figure G.2 shows that the approach is robust to the chosen K for BatchTopK SAE regime.



(a) Representation-level (b) Concept-level

Figure E.2: The **impact of frequency-based binarization threshold** on accuracy of probes built for prediction of 2seq-tiny-ImageNet task 0 classes under LwF based on the raw representations from the continual model (representation-level) and based on the SAE latent neuron activations (concept-level). All accuracy values are computed for the task 0 data after training on task 0, but for the "Raw before filtered" case, we use our SAE decoder to zero out the neurons that are not active binary under our binarization rule. We examine frequency-based threshold in $[0.00625, 0.0125, 0.025, 0.05, 0.1, 0.2]$.



(a) Representation-level (b) Concept-level

Figure F.1: Accuracy of probes built for prediction of 2seq-CIFAR10 and 2seq-tiny-ImageNet task 0 classes based on the raw representations from the continual model (representation-level) and based on the SAE latent neuron activations (concept-level). Both accuracy values are computed for the task 0 data after training on task 0, but for the $F(t)$ case, we use our SAE decoder to zero out the neurons that are not active binary under our binarization rule (e.g. the frequency of their values that are higher than the activation threshold is lower than the frequency threshold).

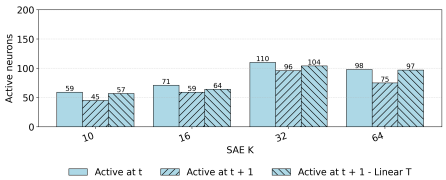
H Impact of SAE batch size on concept-based forgetting analysis

Fig. H.1 shows the impact of the SAE training batch size on the key metrics for the example 2seq-tiny-ImageNet + LwF setup. Overall, the results are qualitatively stable across the tested batch sizes, indicating that our conclusions are not sensitive to this hyperparameter. Although small quantitative differences appear, the overall patterns remain unchanged: raw task-$t+1$ representations show concept deletion, linear alignment recovers a large fraction of the lost information, and both task- and concept-level probe performance stay close to the reference values. Similarly to the results for K, most values remain within the spread observed across repeated runs, suggesting that the effect of batch size is comparable to the run-to-run variability.

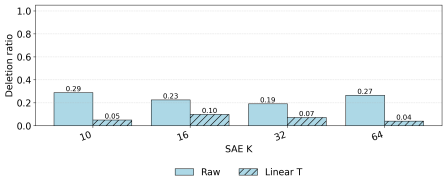
To verify that the set of binary active neurons is functionally relevant under different SAE batch sizes, we additionally measure task-0 classification accuracy after zeroing out, through the SAE decoder, all latent neurons that are not marked as active under the given frequency-based rule. Figure H.2 shows that the approach is robust to the chosen batch size for SAE training.

I Impact of using a non-linear translation model on concept decodability and accessibility

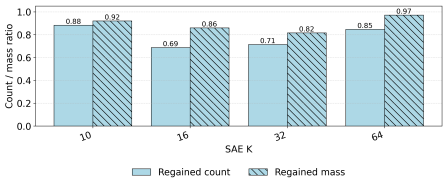
Nonlinear recoverability setup. To test whether the information could be better restored by a low-capacity nonlinear mapping instead of a linear mapping, we additionally considered a nonlinear translation model. It was implemented as a `torch` model with three fully connected layers and GELU nonlinearities after the first two layers. The hidden dimensionality was set equal to the input dimensionality. The model was trained, analogously to the linear translator, to predict the task-t representations from the task-$(t+s)$ representations by minimizing mean squared error. We trained



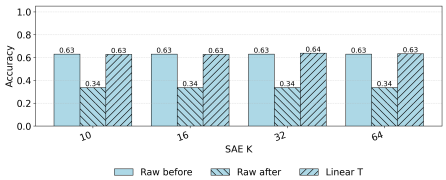
(a) Active neurons



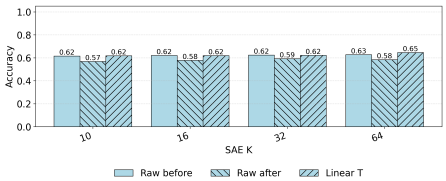
(b) Deletion ratio



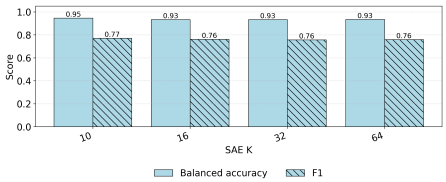
(c) Regained count and mass ratio



(d) Representation-based probe accuracy

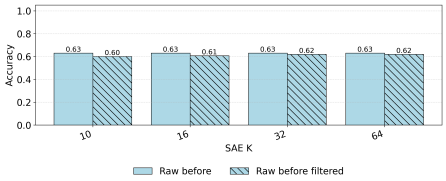


(e) Concept-based probe accuracy

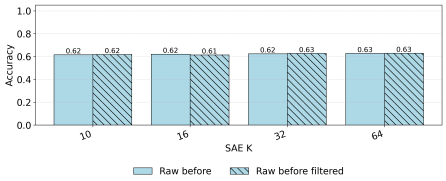


(f) Mean balanced accuracy and F1

Figure G.1: The **impact of SAE's K** on metrics for 2seq-tiny-ImageNet task 0 under LwF.



(a) Representation-level

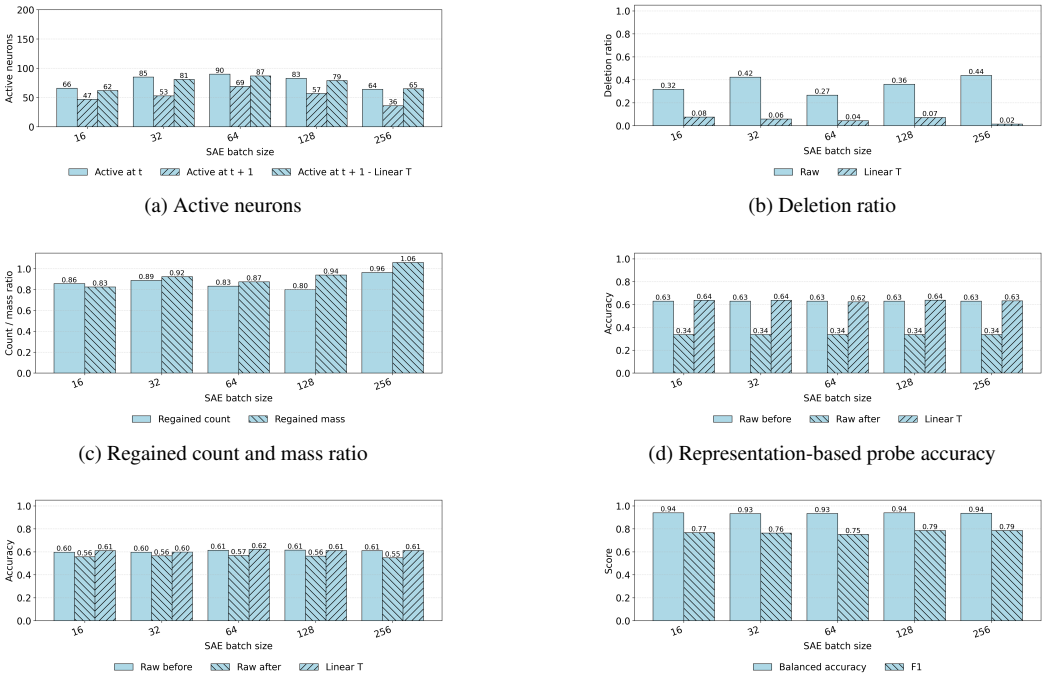


(b) Concept-level

Figure G.2: The **impact of SAE's K** on accuracy of probes built for prediction of 2seq-tiny-ImageNet task 0 classes under LwF based on the raw representations from the continual model (representation-level) and based on the SAE latent neuron activations (concept-level). All accuracy values are computed for the task 0 data after training on task 0, but for the "Raw before filtered" case, we use our SAE decoder to zero out the neurons that are not active binary under our binarization rule.

it for 100 epochs using AdamW, with learning rate 10^{-3}, weight decay 10^{-4}, batch size 128, and a validation split of 0.2.

Fig. I.1 shows that the nonlinear translation model does not provide substantial improvements over the linear translation model. In most cases, the nonlinear model yields slightly worse performance across all considered metrics. Only two exceptions can be observed for deletion/regained count metrics, namely for 2seq-CIFAR10 with DER++ and 2seq-tiny-ImageNet with SGD, where the nonlinear model leads to minor improvements. However, these gains are marginal and correspond to restoring only a very small number of additional concepts (1-3 concepts), without meaningful impact on the overall metrics. Overall, both models achieve highly comparable performance, with the linear translation often performing slightly better. The overall results, therefore, indicate that the recoverable component of the representational drift is largely linear, both at the task level and at the level of individual concepts, and that simple nonlinear mappings do not provide additional explanatory power in this setting. While it is possible that more expressive nonlinear models could recover additional information, identifying such models and their configurations would require exploring a largely unconstrained architectural and hyperparameter space, which is beyond the scope of this work.



(a) Active neurons

(b) Deletion ratio

(c) Regained count and mass ratio

(d) Representation-based probe accuracy

(e) Concept-based probe accuracy

(f) Mean balanced accuracy and F1

Figure H.1: The **impact of SAE batch size** on metrics for 2seq-tiny-ImageNet task 0 under LwF.



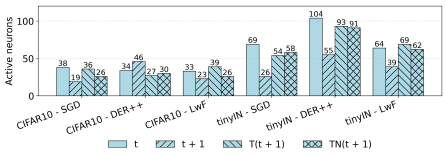
(a) Representation-level

(b) Concept-level

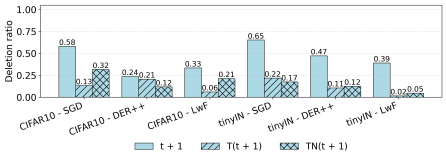
Figure H.2: The **impact of SAE batch size** on accuracy of probes built for prediction of 2seq-tiny-ImageNet task 0 classes under LwF based on the raw representations from the continual model (representation-level) and based on the SAE latent neuron activations (concept-level). All accuracy values are computed for the task 0 data after training on task 0, but for the "Raw before filtered" case, we use our SAE decoder to zero out the neurons that are not active binary under our binarization rule.

J Qualitative analysis of concepts

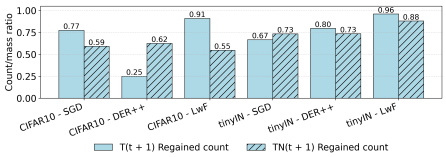
In this section, we provide a qualitative examination of SAE neurons across datasets and continual learning strategies. The goal of this analysis is to visually assess whether SAE latents after tasks t and $t+1$ correspond to reasonably coherent and interpretable patterns. For data from task t, the visual coherence of the top-activating examples suggests that SAE latent neurons capture recurring patterns in the representation space and suggests that it may therefore serve as a useful proxy for concept detectors. Comparing task t examples after training on task $t+1$ with those obtained immediately after task t provides a qualitative indication of concept preservation. Visual similarity suggests that the concept remains at least partially encoded. Substantial differences, visible even in the most prototypical and central part of the neuron activation space, point either to information loss or to representational changes that prevent faithful recovery of the original concept. These results further support analysis of cases for which concept prediction $F1 > 0$ and can be used as a visual sanity check. In particular, we study the visual quality of neurons identified for task t after t and $t+1$ training, by inspecting their top-9 activating images, ordered by activation rank from 1 to 9. To make



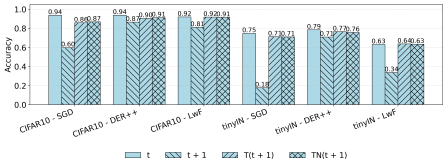
(a) Active neurons



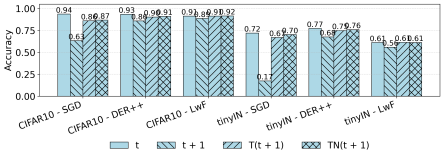
(b) Deletion ratio



(c) Regained count ratio



(d) Representation-based probe accuracy



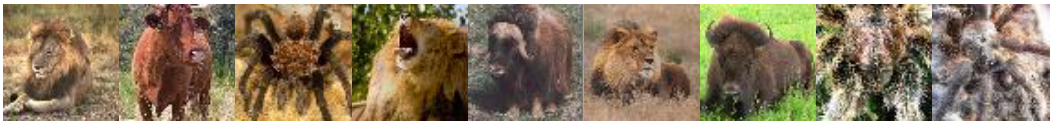
(e) Concept-based probe accuracy

Figure I.1: The **impact of using a non-linear translation model** on different metrics compared to using linear translation. It is visible that the utilized model makes the recoverability and decodability worse in most cases, while in two improved cases, the improvement is marginal.

this comparison meaningful in the presence of representation drift, the maximally activating examples at $t + 1$ are determined using the linearly translated feature vectors. This allows us to examine how the center of the concept changes after subsequent training, and whether the same visual motif can still be recovered despite the apparent shift in representation space.



(a) After t training



(b) After $t + 1$ training, linearly translated

Figure J.1: The example **retained** concept, which stayed active in raw vector after $t + 1$. Top-ranked examples for latent neuron 274 are presented at t (top) and at $t + 1$ (bottom). The setup was 2seq-tiny-ImageNet with DER++. The pictures show a **clear motif**: brown, furry objects, and the motif is shared between representations at t and $t + 1$.

In the qualitative analysis, we present representative examples of concepts drawn from the proposed taxonomy, illustrating different post-task outcomes such as **retained** (Figs. J.1, J.2, J.3), **deleted** (Figs. J.4, J.6, J.7, J.8), **recovered** (Figs. J.4, J.6, J.5), and **decodable** (Figs. J.7, J.8) concepts. These examples complement the quantitative results by showing how the central visual structure of individual concepts expressed with top-9 activating images changes after subsequent task learning. Notably, across the cases examined in our analysis, we did not observe any concepts that would fall into the **Lost** category at the strictest possible F1 threshold, i.e., concepts that were neither recoverable through translation nor decodable from the representation in our setup.



(a) After t training



(b) After $t + 1$ training, linearly translated

Figure J.2: The example **retained** concept, which stayed active in raw vector after $t + 1$. Top-ranked examples for latent neuron 370 are presented at t (top) and at $t + 1$ (bottom). The setup was 2seq-tiny-ImageNet with DER++. The pictures show a **clear motif**: penguins, and the motif is shared between representations at t and $t + 1$.



(a) After t training

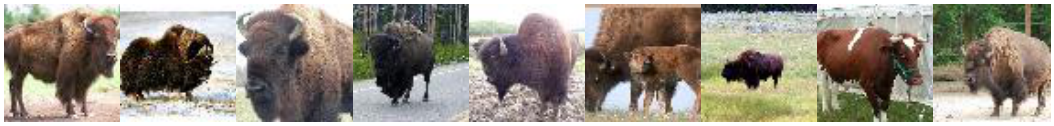


(b) After $t + 1$ training, linearly translated

Figure J.3: The example **retained** concept, which stayed active in raw vector after $t + 1$. Top-ranked examples for latent neuron 425 are presented at t (top) and at $t + 1$ (bottom). The setup was 2seq-tiny-ImageNet with DER++. The pictures show a **clear motif**: transport vehicles, and the motif is shared between representations at t and $t + 1$.

In many of presented cases images representing the concepts at t show a clear common motif, e.g. brown, furry objects (Fig. J.1), penguins (Fig. J.2), bisons (Fig. J.4) or birds (Fig. J.7), regardless of the $t+1$ outcome of the concept transition, which was **retained** for the first two examples, **recovered** for the third example and **decodable** for the fourth example. Some lower-quality neurons can be showed as well, e.g. see Figs. J.6 and J.8. Nevertheless, even in these cases, some recurring objects or motifs can still be identified. For example, in Fig. J.6, several highly activating examples contain backpacks or towers, while in Fig. J.8, recurring elements such as cats and cars can be observed. This behavior is consistent with standard SAEs, which typically extract features of varying quality and degrees of disentanglement. Overall, the presented examples still suggest that the SAE and our mean-based and frequency-based thresholding criteria capture at least partially disentangled features associated with recurring visual patterns, supporting our use of these latent units as proxies for concepts.

Across all presented examples, the top-9 activating images at task t and after learning task $t + 1$ remain highly aligned, indicating that the centers of the corresponding concepts' (latent features) activation spaces are largely preserved. This is reflected in shared visual motifs and, in many cases, even repeated images among the strongest activations. Importantly, this observation is valid even for the seemingly deleted, recovered and decodable categories. Fig. J.5 shows that some contamination of the task t concept center can occur at $t + 1$. While at task t the dominant motifs include green plants, water, and bisons on grass, at task $t + 1$ an out-of-pattern image of a girl in a sombrero appears among the top activations. These qualitative observations provide graphical support for the quantitative results in the paper, which indicate that a large portion of concept-level knowledge is preserved during updates on subsequent tasks, although some information may become harder to read out or partially lost.



(a) After t training

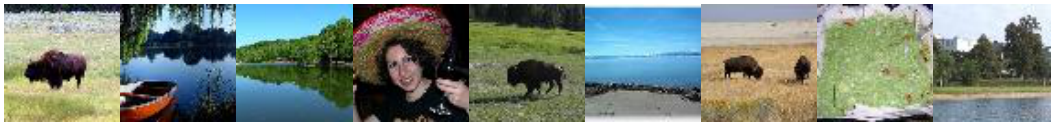


(b) After $t + 1$ training, linearly translated

Figure J.4: The example of a **lower quality concept**. The concept can be described as **seemingly deleted**, as it was absent in raw $t + 1$ and **recovered**, because it reappeared in translated $t + 1$. Top-ranked examples for latent neuron 687 are presented at t (top) and at $t + 1$ (bottom). The setup was 2seq-tiny-ImageNet with LwF. The concept obtained a relatively **high decodability level**: mean bacc was 0.97 and F1 was 0.8 for concept decodability. It had 76 images with > 0 activations at t and 77 at translated $t + 1$.



(a) After t training



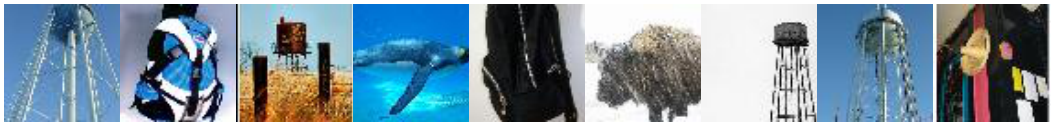
(b) After $t + 1$ training, linearly translated

Figure J.5: The example of a **lower quality concept**. The concept can be described as **seemingly deleted**, as it was absent in raw $t + 1$ and **recovered**, because it reappeared in translated $t + 1$. Top-ranked examples for latent neuron 358 are presented at t (top) and at $t + 1$ (bottom). The setup was 2seq-tiny-ImageNet with DER++. The concept obtained a **medium decodability level**: F1 (0.6) at mean bacc of 0.97 for concept decodability.

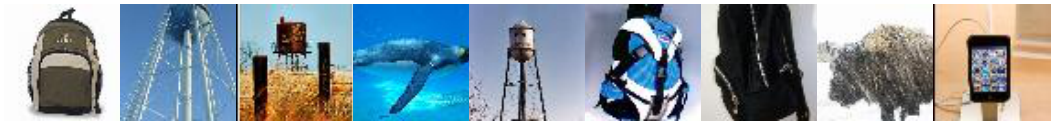
K Hyperparameters

In our **concept-level analysis setup**, for each task, we train an SAE with `BatchTopKSAE` regime via the `overcomplete` [3] on the same Mammoth's train/test splits. We use the unified setup: batch size $B = 16$, top-$K = 10$, 10 epochs, learning rate $5 \cdot 10^{-3}$, expansion rate 2, Mean Squared Error (MSE) loss augmented with an `overcomplete`'s dead-neuron reactivation loss weighted by 10^{-2}. Across experiments, manual inspection showed reconstruction $R^2 > 0.6$ and close to 0% dead-neuron rate, indicating meaningful and non-degenerate SAEs. As a qualitative complement, we examine top 9 activating images for selected SAE latents and find consistent shared visual motifs, indicating that they capture coherent patterns rather than arbitrary feature mixtures (Appendix J). We use frequency-based binarization threshold $\tau = 0.05$. We use a single linear layer as for translation, trained for 100 epochs with `AdamW` using MSE, learning rate 10^{-3}, weight decay 10^{-4}, batch size 128, and a validation split 0.2. For task-level probes, we use Logistic Regression with `lbfgs` solver and 1500 maximum iterations, while for concept-level prediction we use Logistic Regression with `liblinear` solver, 1000 iterations, and `balanced` class weight. We verified the analysis robustness across 10 runs (App. D), frequency-based binarization thresholds (App. E, {0.00625, 0.0125, 0.025, 0.05, 0.1, 0.2}), SAE batch size (App. H, $B \in \{16, 32, 64, 128, 256\}$) and SAE K (App. G, $K \in \{10, 16, 32, 64\}$). The results show consistent behavior of key metrics, with variability largely within run-to-run fluctuations, and support robustness of our conclusions. We further validate that the selected active neurons retain most task-relevant information with a SAE-based experiments

[3] `https://github.com/KempnerInstitute/overcomplete`



(a) After t training



(b) After $t+1$ training, linearly translated

Figure J.6: The example of a **lower quality concept**. The concept can be described as **seemingly deleted**, as it was absent in raw $t+1$ and **recovered**, because it reappeared in translated $t+1$. Top-ranked examples for latent neuron 146 are presented at t (top) and at $t+1$ (bottom). The setup was 2seq-tiny-ImageNet with LwF. The concept obtained a relatively **high decodability level**: F1 of 0.71 and mean bacc of 0.92 for concept decodability.



(a) After t training



(b) After $t+1$ training, linearly translated

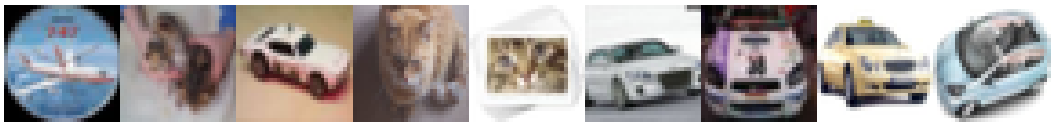
Figure J.7: The example of a **good quality concept**. The concept can be described as **seemingly deleted**, as it was absent in raw $t+1$ and **decodable**, because it didn't reappear in translated $t+1$ but obtained concept decodability $F1 > 0$. Top-ranked examples for latent neuron 14 are presented at t (top) and at $t+1$ (bottom). The setup was 2seq-CIFAR10 with SGD. The concept obtained a relatively **low concept decodability level**: F1 of 0.44 and mean bacc of 0.88.

on deletion of non-active neurons under different values of parameters and under our binarization rule for all methods and datasets, further supporting the use of threshold 0.05. We used multiple NVIDIA GH200 GPUs for continual learning experiments and 1 NVIDIA Titan RTX GPU for SAE training and evaluation. The average execution time of the proposed concept-level analysis pipeline, excluding continual model training but including the saving of all required and additional data for analysis, was 204.199 ± 42.473 s per task on the Titan RTX GPU.

L Using SAE latents as concept proxies

As we use SAE latents as concept proxies in our framework, in this section we assess their feature coherence through two complementary analyses. First, we measure whether active SAE latents respond to coherent visual patterns, comparing their Monosemanticity Scores (MS) [28] with scores obtained from a permutation baseline. Second, we compare SAE latents with features from a reconstruction-trained dense autoencoder, PCA, and Gaussian random projections. Together, these analyses help us assess whether SAE latents capture coherent patterns and whether they provide a more suitable feature space for concept-level analysis than these alternatives.

MS [28] measures how coherent are the images that strongly activate a given SAE latent neuron. Pairwise cosine similarities are computed between embeddings from a pretrained encoder. For a neuron k, representing a given concept proxy, latent activations are used to weight similarities between samples. The MS is then defined as the activation-weighted average of pairwise embedding



(a) After t training



(b) After $t + 1$ training, linearly translated

Figure J.8: The example of a **lower quality concept**. The concept can be described as **seemingly deleted**, as it was absent in raw $t + 1$ and **decodable**, because it didn't reappear in translated $t + 1$ but obtained concept decodability $F1 > 0$. Top-ranked examples for latent neuron 677 are presented at t (top) and at $t + 1$ (bottom). The setup was 2seq-CIFAR10 with SGD. The concept obtained a relatively **low decodability level**: F1 of 0.28 and mean bacc of 0.75.

similarities, assigning higher scores to neurons whose most activating examples correspond to coherent visual patterns, and lower scores to neurons activating for diverse unrelated images [28].

We first analyze an example configuration: 2seq-Tiny-ImageNet under the LwF continual learning strategy. For each active SAE latent identified using our frequency-based binarization criterion, we compute MS using embeddings extracted from the continual model's backbone. As a baseline, we also compute the score after randomly permuting the encoder embeddings. This permutation disrupts the correspondence between samples and their embeddings while preserving the embedding-space similarity structure.

Figure L.1 shows that active SAE latents achieve higher MS values than the permutation baseline. Scores vary across latents, reflecting differences in feature coherence, but the overall separation indicates that the latents capture recurring visual patterns rather than arbitrary associations.

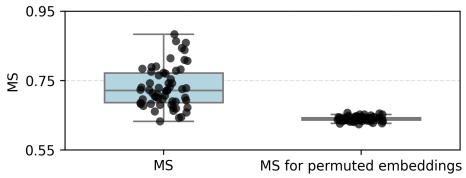


Figure L.1: Monosemanticity Score (MS) for active SAE latents in the 2seq-Tiny-ImageNet + LwF configuration. The box plots compare scores for active SAE latents with scores obtained after permuting encoder embeddings. Higher MS indicates that a latent's strongly activating examples form a more coherent visual pattern.

SAEs are designed to decompose dense representations into sparse latent features. Although these features are not guaranteed to be perfectly disentangled or human-interpretable, the sparsity objective is intended to produce more semantically coherent features than representations learned without using that objective [28]. This property supports fine-grained analysis and qualitative inspection of top-activating examples, as shown in Appendix J.

In principle, concept proxies could be extracted using other feature spaces, such as dense autoencoders, Principal Component Analysis (PCA) or random projection. By comparison, a standard autoencoder is trained to only reconstruct its input (without sparsity constraints), PCA identifies directions of maximal variance, and random projections produce untrained linear combinations of the original features. Although our framework does not require SAE latents specifically, our rationale to use SAEs was that their sparsity objective encourages individual features to respond to a narrower range of patterns, making them more suitable for feature-level analysis and qualitative inspection. The comparisons below test whether this motivation is reflected in the measured coherence of the resulting features.

To compare these alternative representations quantitatively, we extend the analysis above to include a reconstruction-trained dense autoencoder (DenseAE) with the same latent dimensionality as the SAE, PCA, and Gaussian random projections. We report the mean and standard deviation of MS across features, alongside the corresponding scores computed using permuted embeddings. The comparison uses the 2seq-Tiny-ImageNet, LwF, Task 0 configuration.

Table L.1: Monosemanticity Score (MS) for SAE latents and alternative feature spaces in the 2seq-Tiny-ImageNet, LwF, Task 0 configuration. Permuted MS is computed after randomly permuting the encoder embeddings. Values are mean $\pm$ standard deviation.

Method	MS (mean $\pm$ std)	Permuted MS (mean $\pm$ std)
SAE	$\mathbf{0.7277 \pm 0.0405}$	0.6389 ± 0.0017
DenseAE	0.6455 ± 0.0199	0.6389 ± 0.0008
PCA	0.6390 ± 0.0016	0.6389 ± 0.0002
Random projection	0.6414 ± 0.0048	0.6389 ± 0.0003

The MS values for DenseAE, PCA, and random projection are close to their respective permutation-based references. SAE latents show a larger separation from the corresponding reference, indicating greater feature coherence under this measure. These results support the use of SAE latents as concept proxies in our framework. SAEs provide a latent feature space whose individual dimensions exhibit more coherent visual patterns than the alternatives considered. Although latents still remain proxies rather than guaranteed human-interpretable concepts, their top-activating examples can be inspected manually, as in Appendix J, or analyzed with image-captioning models. While similar analyses can be conducted in other feature spaces, the lower MS values observed for those alternatives suggest that interpreting individual dimensions may be less reliable.

M Effect of Translator Training Data

The linear translator provides a mechanism for recovering representations of an earlier task, which is an important part of our diagnostic framework. Training this translator requires access to data from the previous task. In this section, we investigate how much old-task data is needed for the translator to remain useful, and wether we can only use a fraction of the training dataset to make the solution more efficient. In the performed experiment, we varied the fraction of old-task training data used to fit the translator and measured both its reconstruction quality and the downstream effect on concept recovery. This analysis also characterizes whether translator training can operate with a limited data budget, much like replay-based approaches, for future applications beyond diagnostics.

Table M.1: Translator reconstruction quality as a function of the fraction of old-task training data. Values are mean $\pm$ standard deviation over five runs.

Training fraction	MSE	Cosine similarity	R^2
1%	0.156 ± 0.003	0.952 ± 0.001	0.740 ± 0.005
5%	0.081 ± 0.001	0.974 ± 0.0004	0.860 ± 0.002
10%	0.065 ± 0.0002	0.979 ± 0.0001	0.890 ± 0.0004
20%	0.055 ± 0.0001	0.983 ± 0.00008	0.910 ± 0.0003
50%	0.047 ± 0.0002	0.985 ± 0.00006	0.920 ± 0.0003
100%	0.044 ± 0.0003	0.986 ± 0.00009	0.930 ± 0.0005

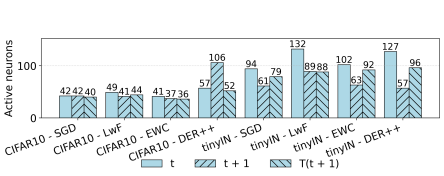
We conducted the experiment for the example 2seq-tiny-ImageNet setting, using Task 0 data as the anchor under the LwF strategy. For each data fraction, we repeated translator training five times using 1% (100 samples), 5%, 10%, 20%, 50%, or 100% of the original data. Table M.1 reports the mean and standard deviation across runs.

Reconstruction quality decreases as the training subset becomes smaller. However, translators trained on 10–20% of the data still obtain high test cosine similarity (approximately 0.98) and achieve R^2 values around 0.9, which in many cases can be acceptable. We further evaluated concept recovery with a translator trained on 10% of the data. The fraction of regained latents decreased from approximately 100% with the full-data translator to approximately 80%, consistent with the higher translation error. Despite this reduction, concept recovery remained clearly observable. These results suggest that the

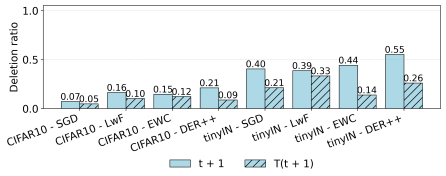
translator's data requirements can be reduced when access to old-task data is limited, although doing so entails some loss in reconstruction quality and recovery performance.

N Vision Transformer results

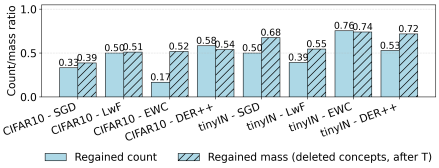
To analyzed the robustness of our framework to other neural network architectures, in this appendix we additionaly performed a smaller experiment with the Vision Transformer (ViT) model. Figures N.1–N.3 present these additional experiments on 2seq-CIFAR10 and 2seq-tiny-ImageNet. The results are qualitatively consistent with our main findings. Although training on subsequent tasks leads to apparent concept deletion, linear translation restores a large fraction of the lost concept in terms of regained concept count and mass (Fig. N.1). Concept prediction results further show that fine-grained information remains decodable to varying degrees (Fig. N.2) with substantial part being possible to decode. Moreover, also in this case, the recovery of representation- and concept-level probe accuracy after alignment provides additional evidence of preserved task-relevant information (Fig. N.3). Although the relative rankings of forgetting mitigation strategies in some cases differs from those observed with ResNet18, the central finding remains consistent across backbones: much of the seemingly forgotten fine-grained information remains linearly recoverable, suggesting representational misalignment rather than complete information erasure.



(a) Active neurons

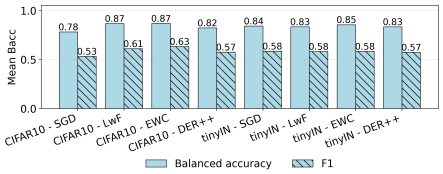


(b) Deletion ratio

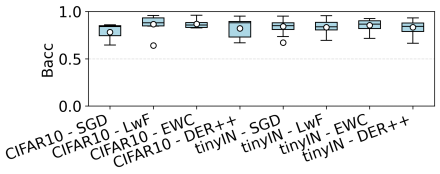


(c) Regained concept statistics

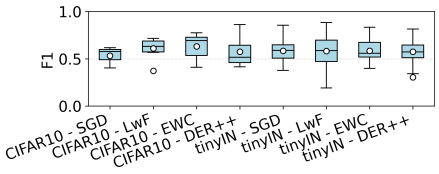
Figure N.1: ViT analysis: Concept activation analysis: **Active concept count**, (seeming) **deletion ratio** and **regained concept statistics** for 2seq-CIFAR10 and 2seq-tiny-ImageNet.



(a) Mean balanced accuracy and F1 score

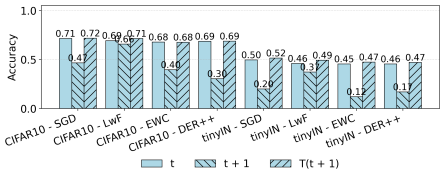


(b) Balanced accuracy distribution

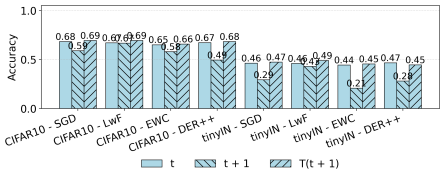


(c) F1 score distribution

Figure N.2: ViT analysis: Mean **balanced accuracy** and **F1 score** along with their distributions for the concept prediction probe on 2seq-CIFAR10 and 2seq-tiny-ImageNet.



(a) Representation-level



(b) Concept-level

Figure N.3: **ViT analysis: Accuracy of probes** predicting 2seq-CIFAR10 and 2seq-tiny-ImageNet task 0 classes from the continual model features (representation-) and SAE latent activations (concept-level).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope. Our main contributions are a concept-level framework, defining a taxonomy of concept transitions after learning subsequent tasks in continual learning and observations regarding the transitions of concepts in continual learning. We formulate our framework and taxonomy in Section 3. We then use our framework with its three main Analysis modules (Binary concept activation analysis, Concept prediction from raw features and Task prediction via linear probes) to examine forgetting through the lens of changes in fine-grained patterns in the model computations obtained via SAEs (which we treat as concept proxies, and refer to as concepts in the paper). The results are presented in Section 5. The section is divided into three subsections, each corresponding to a single analysis module, in the same order as they are introduced in Section 3.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We created a separate "Limitations" paragraph in Section 6.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: In our paper, we use freely available datasets and libraries (Mammoth, overcomplete – links to their repositories provided in the main paper) to perform continual training and Sparse Autoencoders training. For continual training, we mostly use the default setting for common continual strategies and data splits of the Mammoth library. We also provide the values for all the crucial parameters in the main paper, Appendix K, and our code for reproducibility at `https://github.com/lifelonglab/lost-or-hidden-neurips-2026`.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We use freely available datasets in our experiments provided via the Mammoth library with their splits. We use open-source library overcomplete to train our SAEs. When it comes to our own code implementing the framework, we provide it at `https://github.com/lifelonglab/lost-or-hidden-neurips-2026`.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We describe our experimental setup in Section 4 and Appendix K, and the full details are provided with the code. In Section 4, we describe the libraries (Mammoth, overcomplete), the methods, and the datasets used (provided via Mammoth with its splits). We describe the parameters used in the study and also refer to different Appendices, in which we test the robustness of different parameters on an example setup (e.g. robustness across different runs – Appendix D, frequency threshold values – Appendix E, SAE K – Appendix G, SAE batch size – Appendix H). We also present in Appendix F a sanity check of the chosen frequency threshold, in which we examine whether the active set of neurons under the frequency-based binarization rule used in our main experiments retains the functionally relevant task information for both datasets. This experiment uses SAE's decoder to delete concepts that are inactive our binarization rule to see whether they sufficiently support the performance on a given task.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: While the main paper does not include error bars for all the experiments (it would be too computationally expensive, especially from the continual training side), we do perform additional robustness analysis in the Appendices for example configurations (e.g. across 10 runs – Appendix D, frequency threshold values – Appendix E, SAE K – Appendix G, SAE batch size – Appendix H). We believe that these additional stability analyses support the robustness of the main findings and conclusions.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: As our framework does not require dedicated hardware, we report the main computational resources used, including GPUs, in Section 4. We also provide the measured runtime for a single task of the concept-based analysis, reported as mean $\pm$ standard deviation over 10 runs. This time excludes continual training and saving its outputs, as these are not part of the framework itself (which takes significant time for computations and results generation). In practice, the reported runtime is overestimated, since our implementation additionally saves large amounts of metadata used for the paper's analysis. The full research project required additional compute beyond the experiments reported in the paper, as substantial prototyping, testing, and preliminary experimentation were needed before obtaining the final version of the framework and the results presented in the main paper. All the data saved for the purpose of the experiments, including the results of continual training and the results of the framework required a few hundred gigabytes of disk space.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

Justification: When it comes to the data-related concerns, we use freely available datasets, which eliminates the majority of the described concerns. We cite the datasets properly in the main paper. We do not use any deprecated datasets, but standard computer vision benchmarks. We mention one potential negative societal impact of our study that we see in Limitations, however we find more positive ones, as our framework aims to improve the explainability of continual learning and make one of its biggest concerns – forgetting – more understandable. We also did our best to properly communicate the details of our experimental setup (datasets, libraries used, parameters used, etc.) to make reproducibility of the paper high. For this purpose, we also share our code at `https://github.com/lifelonglab/lost-or-hidden-neurips-2026`.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This paper emphasizes the importance of preserving fine-grained representational structure, measured through SAE-based concept proxies, as a step toward more interpretable and trustworthy continual learning, offering a positive societal impact. We mention this in Section 6. In out discussion of Limitations, we also mentioned that concept-level analyses of forgetting may increase confidence in deploying continual models in high-stakes settings, even when the identified concepts remain proxies and may fail to capture socially sensitive, domain-specific, or safety-critical information, which creates a potential negative societal impact.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly reference (links to the repositories, paper citations) and respect all the resources used in the paper (libraries, datasets).

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide an implementation of our framework compatible with standard Python libraries. We share our code at `https://github.com/lifelonglab/lost-or-hidden-neurips-2026` under the MIT License.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]