Semantic Concept Steering Breaks the Explanation Drift Loop in Continual Learning

Yehonatan Elisha
Tel Aviv University

Oren Barkan
The Open University

Noam Koenigstein
Tel Aviv University

Abstract

Catastrophic forgetting in continual learning (CL) manifests not only as accuracy degradation but also as *explanation drift*: models maintain predictive accuracy while silently shifting attention from diagnostic features to spurious correlations. This work identifies and resolves a previously overlooked structural flaw in explanation-aware CL: pixel-level saliency regularization derives supervision targets from the model's own frozen saliency maps, creating a *self-referential drift loop* that actively reinforces spurious attention across tasks. We introduce **C^{3}L** (Concept-Consistent Continual Learning), which breaks this loop by anchoring regularization to class-consistent semantic concepts (e.g., "curved beak", "striped pattern") discovered offline, independent of model state, providing (i) target independence from model drift, and (ii) cross-instance semantic consistency across every instance of a class. We identify that our concept-based objective, combined with approximate LRP relevance conservation, induces a *crowding-out effect*: since concept and background regions compete for the relevance of the same prediction, enforcing high relevance within concept regions implicitly suppresses spurious background correlations without explicit penalization. C^{3}L does not require human-annotated concept sets, as it employs a fully automated concept discovery and spatial grounding. Extensive evaluation across five benchmarks and over 20 baselines demonstrates state-of-the-art performance in accuracy, forgetting, explanation quality, spurious correlation mitigation, and out-of-distribution robustness. Our code is provided at: `https://github.com/yonisGit/concept_consistent_continual_learning`

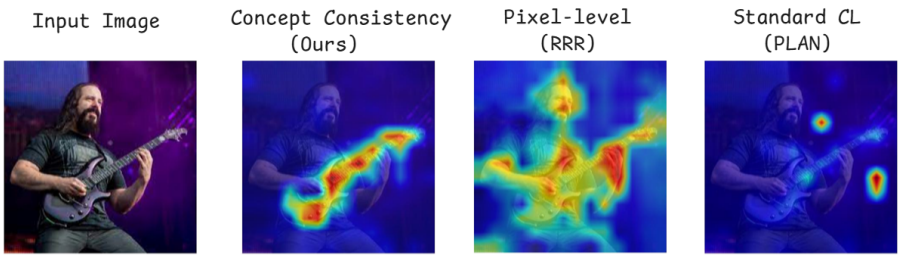


Figure 1: **Mitigating Explanation Drift via Semantic Anchoring.** Saliency maps for "Electric Guitar" sample after continual learning. Standard CL (PLAN) suffers from explanation drift, shifting attention to spurious background artifacts. While pixel-level methods (RRR) attempt to mitigate this, they result in noisy, non-specific attributions. Our C^{3}L anchors model representations to semantically meaningful concepts, maintaining focus on diagnostic features like the fretboard and guitar body.

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

1 Introduction

Deep neural networks trained on static datasets catastrophically forget previously learned knowledge when adapted to sequential tasks [1, 2]. This phenomenon poses fundamental challenges for deploying adaptive systems in dynamic environments, from medical diagnosis systems learning new diseases to autonomous vehicles adapting to novel scenarios [3]. While the continual learning literature has developed sophisticated mechanisms to preserve predictive accuracy through replay [4, 5], regularization [6, 7], and architectural adaptations [8, 9], these approaches treat models as black boxes operating in parameter or feature space - preserving *what* is predicted, not *how*.

The explanation drift problem. Ebrahimi et al. [10] showed that catastrophic forgetting manifests not only as accuracy degradation but also as *explanation drift*: models maintain accuracy while silently shifting attention from diagnostic features (e.g., bird beak shapes) to spurious correlations (e.g., background foliage). This instability in reasoning is not captured by standard metrics, yet has critical implications - when such spurious correlations fail under distribution shift, performance can collapse [11, 12]. Existing explanation-aware CL methods address this drift by regularizing pixel-level saliency maps [10, 13, 14], but they share a fundamental and previously overlooked structural flaw: Pixel-level methods derive supervision targets from the model's own frozen saliency at buffer insertion. If the model has already drifted toward spurious features at that point, these patterns become the regularization target, effectively reinforcing the drift across future tasks. This *self-referential drift loop* is not merely an attribution noise issue - it is a structural property of any scheme that derives targets from model-state-dependent saliency.

Breaking the loop with semantic concept anchoring. We present **C^{3}L** (Concept-Consistent Continual Learning) (Figure 2), which breaks this loop by grounding regularization in *class-consistent semantic concepts* - human-interpretable visual attributes (e.g., "curved beak") discovered *offline*, prior to and independent of any model state. This yields two structural properties no previous explanation-aware method can provide: *(i) target independence* - concept targets cannot encode model drift as they are produced independently of the learner's state; and *(ii) cross-instance semantic consistency* - every instance of a class is regularized toward the same diagnostic attributes, inducing a class-level invariant beyond instance-specific saliency supervision.

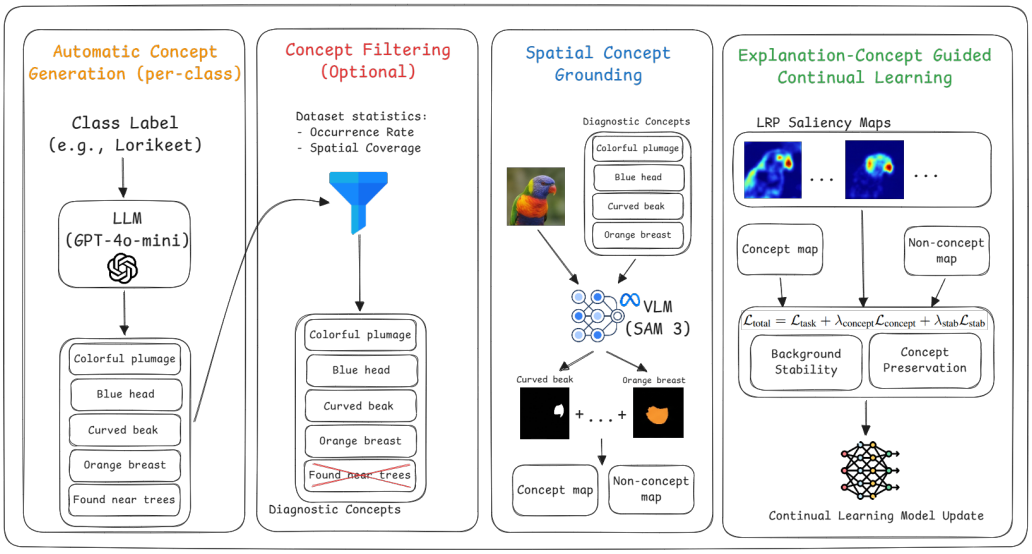


Figure 2: **C^{3}L Overview.** Discriminative visual concepts are discovered offline via an LLM and spatially grounded using SAM 3, producing model-state-independent supervision targets that break the self-referential drift loop of prior pixel-level methods. During continual training, $\mathcal{L}_{\text{concept}}$ anchors relevance to semantic regions via a *crowding-out effect*, while $\mathcal{L}_{\text{stab}}$ preserves background spatial structure against task-specific drift.

Furthermore, we identify a *crowding-out effect*: under approximate LRP relevance conservation, concept and background regions compete for the relevance of the same prediction. Enforcing high

relevance within semantically grounded concept regions therefore implicitly suppresses spurious background attributions, without any explicit penalization, turning semantic anchoring into an implicit suppression mechanism, which we validate empirically. A complementary background stability term preserves the *spatial structure* of background relevance while allowing its magnitude to shrink, preventing freed relevance mass from drifting toward new task-specific artifacts rather than toward diagnostic concepts. Finally, a fully automated framework, leveraging LLMs for concept discovery [15] and open-vocabulary segmentation for spatial grounding [16], eliminates the manual annotation burden that has historically prevented concept-based approaches [17] from scaling to the many-class, sequential-task regime.

Contributions:

1. **Problem identification:** We formalize the self-referential drift loop as a structural obstacle in pixel-level explanation-aware CL (Section 3).
2. **Crowding-out mechanism:** We identify the crowding-out effect induced by approximate LRP conservation, which turns semantic anchoring into an implicit background-suppression mechanism, and validate it quantitatively.
3. **$\mathbf{C^3L}$ framework:** A novel concept-guided continual learning framework. It operates without manual annotation and leverages semantic anchors to eliminate self-referential drift and drive the generalization of transferable concepts in the model (§4.5).
4. **State-of-the-art results:** C^3L outperforms nearly twenty baselines across five benchmarks on accuracy, forgetting, semantic localization, spurious correlation mitigation, and OOD robustness, on both CNN and ViT architectures.

2 Related Work

Continual learning strategies. CL methods can be roughly divided into several types. Replay-based methods [4, 5, 18, 19] store exemplars and interleave them during new task training. Wan et al. [20] show that replay is essential for maintaining feature learning dynamics across tasks. Regularization-based methods [6, 7, 21] add parameter regularization terms or function regularization terms to transfer knowledge from the old model to the new one. Architecture-based methods often add new trainable modules to expand the feature space for new tasks. Prompt-based [9, 22, 23] methods keep the backbone frozen and learn task-specific input prompts. Adapter-based methods [24–26] insert learnable modules or selectively fine-tune components while freezing pretrained weights. All these operate in parameter, feature, or sample space without preserving *semantic reasoning* - which visual features the model attends to. This leaves models vulnerable to explanation drift [10].

Explainable AI and saliency maps. Explainability methods aim to reveal which input features drive a model's prediction. Model-agnostic approaches such as LIME [27] and SHAP [28], as well as learned explainers [29], apply across domains, and dedicated attribution techniques have been developed for recommender systems [30–34] and for language models [35–38]. In computer vision, explanations are typically given as saliency maps, produced by gradient-based methods such as Grad-CAM [39] and its extensions to similarity models [40] and transformers [41]; by relevance-propagation methods that approximately conserve relevance, such as LRP [42, 43] and AttnLRP [44]; and by path-integration methods originating from Integrated Gradients [45], later extended to integration over intermediate representations [46, 47], stochastic and probabilistic paths and baselines [48, 49], and metric-adapted baselines [50]. Related work revisits the completeness property underlying conservation-based attribution [51] and the human-aligned evaluation of saliency maps [52]. Concept-based explanations link attributions to human-interpretable concepts [53–55], and concept-guided relevance has recently been used as a *training* signal to steer models away from spurious correlations in a single fine-tuning stage, by explicitly suppressing background relevance [12]. $\mathbf{C^3L}$ differs in both setting and mechanism: it targets continual learning, where explanations drift as the model is updated on a sequence of tasks; it identifies the self-referential drift loop of pixel-level explanation-aware CL methods and replaces their model-derived targets with learner-independent concept targets; and rather than suppressing the background, it stabilizes its spatial structure with a scale-invariant term, while background magnitude decreases through the crowding-out effect.

Explanation-aware continual learning. Recognizing that catastrophic forgetting manifests as explanation drift alongside accuracy loss, Ebrahimi et al. [10] introduced RRR, which stores Grad-CAM [39] saliency maps alongside replay exemplars and penalizes attention shift, demonstrating that models frequently preserve accuracy while fundamentally altering their decision process. Subsequent

work extends this paradigm: Saha and Roy [13] propose saliency-guided exemplar packing to maximize buffer diversity, and Liu et al. [14] introduce task-adaptive saliency masks that adjust dynamically to forgetting patterns. While these methods show that explanation consistency provides valuable supervision beyond classification loss, they lead to a *self-referential drift loop*. This drift loop is not merely an attribution noise issue - it is a structural property of any scheme that derives targets from model-state-dependent saliency. The problem is compounded by well-documented brittleness of pixel-level attributions: they are notoriously noisy and prone to highlighting spurious correlations [56, 57], can be sensitive to imperceptible input perturbations [58] and may emphasize high-frequency texture artifacts rather than semantically meaningful object parts [11].

$\mathbf{C}^3\mathbf{L}$ breaks this loop, replacing model-state-dependent pixel targets with class-consistent, model-state-independent concept-level supervision generated entirely offline before training.

Concept discovery and visual grounding. Early concept-based models required extensive manual annotation [53, 17]. Label-free CBMs [15] use LLMs to generate concepts without supervision. SAM 3 [16] enables precise spatial localization of textual concepts. In this work, we do not aim to construct a bottleneck model for inherent interpretability. Rather, we leverage these automated discovery and grounding pipelines as an offline pre-processing stage to synthesize rich, dense supervision signals. This allows us to anchor the model's attention to semantic regions during continual learning, effectively replacing expensive manual concept annotation with scalable, cross-instance semantic consistent priors.

Positioning. Prior region-level CL relies on model-state-dependent targets (e.g., CAM), creating a *self-referential drift loop* we identify. $\mathbf{C}^3\mathbf{L}$ eliminates this loop via three axes no prior method addresses simultaneously: (1) *target independence*: offline generation prevents drift reinforcement; (2) *semantic specificity*: supervision targets are fine-grained attributes rather than coarse boundaries; and (3) *cross-instance consistency*: class-shared targets induce robust invariants. Finally, Table 11 isolates semantic specificity, confirming a significant advantage over object-level masks using the identical SAM 3 model.

3 Method

3.1 Problem Formulation

We consider the class-incremental continual learning setting [4], where a model encounters a sequence of T tasks $\{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_T\}$. Each task $\mathcal{T}_t$ provides a training set $\mathcal{D}_t = \{(x_i, y_i)\}_{i=1}^{N_t}$ containing N_t labeled examples from K_t new classes $\mathcal{C}_t = \{c_1^{(t)}, \ldots, c_{K_t}^{(t)}\}$ disjoint from previous tasks classes. During training on task $\mathcal{T}_t$, the model must learn to classify all classes encountered so far, $\mathcal{C}_{1:t} = \bigcup_{j=1}^{t} \mathcal{C}_j$, while only having access to $\mathcal{D}_t$ and a small replay buffer $\mathcal{M}$ containing stored samples from $\mathcal{D}_{1:t-1}$. Let $f_\theta : \mathcal{X} \to \mathcal{Y}$ denote our classifier and $E_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^{H \times W}$ an explanation function mapping input-label pairs to saliency maps. Beyond maximizing accuracy, we aim for *semantic stability*: maintaining attention on semantically meaningful concept regions that define each class.

Concept-grounded reasoning. For each class $c \in \mathcal{C}_{1:t}$, let $\Psi_c = \{\psi_1, \psi_2, \ldots, \psi_K\}$ denote a set of visual concepts discovered for that class (§3.2). For sample x with label y, let $M_{\text{concept}}^{(x)} \in \{0,1\}^{H \times W}$ denote regions where concepts from Ψ_y are present (§3.2), and $M_{\text{bg}}^{(x)} = \mathbf{1} - M_{\text{concept}}^{(x)}$ its complement. Concept regions $M_{\text{concept}}^{(x)}$ identify meaningful features (e.g., wings, wheels) that should remain attended across tasks, while explanations in the $M_{\text{bg}}^{(x)}$ regions should not develop spurious correlations. By enforcing both positive supervision (attend to concepts) and negative supervision (stabilize backgrounds), we guide models toward focusing on robust, transferable features.

3.2 Automated Concept Discovery and Grounding

Automatic concept generation. Manual concept annotation [17] is impractical at the scale of continual learning. We adopt Oikarinen et al. [15]'s label-free approach, querying a large language model to generate discriminative textual attributes for each class $c \in \mathcal{C}_t$, yielding initial concepts $\Psi_c^{\text{raw}} = \{\psi_1, \ldots, \psi_K\}$ that are class-discriminative and linguistically interpretable. Large language

models encode substantial knowledge about the visual appearance of object categories, enabling semantically meaningful attributes aligned with human perception.

Concept filtering (Optional). We optionally validate concepts on held-out data $\mathcal{C}_c^{\text{VAL}}$ using: (1) **Occurrence rate** $\nu_{\text{occ}}(\psi) \geq \tau_{\text{occ}}$ - fraction of instances where concept ψ can be spatially localized; (2) **Coverage** ν_{cov} - IoU between the union of all concept masks and the class-object segment, ensuring concepts are neither too sparse nor dense. Failing concepts are pruned, yielding $\Psi_c^* \subseteq \Psi_c^{\text{raw}}$. Importantly, Appendix E.6 shows this step is optional: unvalidated concept sets also achieve strong performance, indicating our framework tolerates some noise. Further details on the filtering mechanism can be found in the Appendix.

Spatial concept grounding. For each replay sample $(x, y) \in \mathcal{M}$, we use SAM 3 [16] to localize concepts from Ψ_y^*. Given image x and a batch of concept prompts (one noun phrase per concept), SAM 3 produces binary masks $\{B_i\}_{i=1}^M$ with confidence scores $\{s_i\}_{i=1}^M$ for each concept $\psi \in \Psi_y^*$. We aggregate high-confidence regions ($s_i \geq \tau_{\text{conf}}$):

$$M_{\text{concept}}^{(x)} = \bigcup_{\psi \in \Psi_y^*} \bigcup_{i : s_i(\psi) \geq \tau_{\text{conf}}} B_i^{(\psi)}. \tag{1}$$

SAM 3's batched queries enable efficient single-pass grounding of all class concepts, eliminating the need for task-specific segmentation models or pixel-level annotations. The concept sets are shared across all instances of a class, and the masks are produced independently of the learner's state, ensuring cross-instance semantic consistency absent in pixel-level methods. Appendix E.5 demonstrates that C^{3}L remains robust even under weak concept grounding conditions, as demonstrated by a targeted noise robustness evaluation.

3.3 Concept-Consistent Continual Learning

Explanation extraction. We obtain pixel-level relevance $S_x \in \mathbb{R}^{H \times W}$ using architecture-specific methods satisfying relevance conservation [43]: Layer-wise Relevance Propagation (LRP) [42] with ϵ-rule ($\epsilon = 10^{-6}$) for CNNs, and AttnLRP [44] for ViTs. LRP-based methods capture comprehensive layer-wise attribution. See Appendix E.2 for empirical comparison with different explanation methods.

Training objective. When training on task $\mathcal{T}_t$, we optimize:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_{\text{concept}} \mathcal{L}_{\text{concept}} + \lambda_{\text{stab}} \mathcal{L}_{\text{stab}} \tag{2}$$

where $\mathcal{L}_{\text{task}}$ is cross-entropy on $\mathcal{D}_t$ and replay buffer $\mathcal{M}$.

The two regularization terms described below implement a *correct-not-preserve* strategy, in contrast to conventional saliency alignment ($\|S - S^{\text{ref}}\|_1$) which mostly anchor the model to reference explanations that may themselves encode spurious correlations.

Concept Overlap Preservation ($\mathcal{L}_{\text{concept}}$). To anchor attention to class-defining features, we enforce high relevance within concept regions:

$$\mathcal{L}_{\text{concept}} = \frac{1}{|\mathcal{M}|} \sum_{(x,y) \in \mathcal{M}} \text{BCE}(S_x \odot M_{\text{concept}}^{(x)}, M_{\text{concept}}^{(x)}) \tag{3}$$

where $\odot$ denotes element-wise multiplication, and BCE denotes the binary cross entropy function. Because concept targets are model-state-independent, they cannot encode drift, breaking the self-referential loop that afflicts pixel-level methods. Notably, $\mathcal{L}_{\text{concept}}$ can also be applied to all training samples, as discussed in Sec. 4.2.

The Crowding-Out Effect. Minimizing $\mathcal{L}_{\text{concept}}$ also has an implicit effect on background regions. Let S_x satisfy the approximate LRP conservation property [43]:

$$\sum_{i \in \mathcal{C} \cup \mathcal{B}} S_x(i) \approx f_\theta(x)_y, \tag{4}$$

where $\mathcal{C}$ and $\mathcal{B}$ index concept and background pixels respectively, and $f_\theta(x)_y$ is the logit of the target class. Denoting by $\rho_{\mathcal{C}} = \sum_{i \in \mathcal{C}} S_x(i) / \sum_i S_x(i)$ the share of relevance assigned to concept regions, Eq. 4 gives $\sum_{i \in \mathcal{B}} S_x(i) \approx f_\theta(x)_y \, (1 - \rho_{\mathcal{C}})$. Concept and background regions therefore compete for

the relevance of the same prediction: increasing the concept share via $\mathcal{L}_{\text{concept}}$ reduces background relevance unless the target logit grows proportionally. We term this coupling the *crowding-out effect*; it suppresses background relevance *without explicit background penalization*. Since conservation is only approximate in networks with non-linearities, relevance is signed, and $\mathcal{L}_{\text{concept}}$ operates on min-max normalized maps, we treat crowding-out as a mechanism hypothesis rather than a guarantee, and validate it empirically in Appendix D.

Background Stability ($\mathcal{L}_{\text{stab}}$). The crowding-out effect reduces background relevance *magnitude*, but does not control its *spatial pattern*. Without additional regularization, freed relevance mass could drift toward new task-specific artifacts rather than concentrating on diagnostic concepts. To prevent this, we store a reference saliency S_x^{ref} when sample (x, y) first enters $\mathcal{M}$ after task $\mathcal{T}_j$, and penalize spatial drift in background regions during subsequent tasks ($t > j$):

$$\mathcal{L}_{\text{stab}} = \frac{1}{|\mathcal{M}|} \sum_{(x,y)\in\mathcal{M}} \left(1 - \frac{\langle S_x \odot M_{\text{bg}}^{(x)},\ S_x^{\text{ref}} \odot M_{\text{bg}}^{(x)} \rangle}{\|S_x \odot M_{\text{bg}}^{(x)}\|_2 \cdot \|S_x^{\text{ref}} \odot M_{\text{bg}}^{(x)}\|_2 + \epsilon} \right). \tag{5}$$

Using cosine similarity is a deliberate design choice: it penalizes *spatial and structural* drift in the background while leaving the magnitude free to decrease under crowding-out. Although S_x^{ref} is model-derived, this term does not reintroduce the self-referential drift loop: it is confined to background regions and is scale-invariant, constraining *where* background relevance lies but not *how much* of it there is. The corrective signal comes from the model-independent concept term, which remains free to reduce background relevance, so spurious patterns present at buffer insertion cannot be reinforced in magnitude. Crucially, we stabilize rather than suppress backgrounds, since the pretrained backbone encodes robust contextual priors (e.g., geometry, scale) that support generalization and should not be erased. Table 6 confirms this design outperforms removing background regularization.

A detailed discussion on objective design choices is provided in Appendix C.

Isolating the contribution of semantic guidance. We rigorously verify that the observed gains stem from C^{3}L, rather than the capabilities of VLMs (Appendix H). Specifically: (1) removing foundation models entirely by training on PartImageNet with ground-truth annotations - C^{3}L still outperforms all baselines; and (2) fine-grained concept masks outperform coarse object-level masks generated by the same SAM 3 model. Together, these results demonstrate that the semantic dual-regularization objective is essential to mitigating explanation drift.

4 Experiments

We present a comprehensive empirical evaluation of C^{3}L across diverse continual learning scenarios, model architectures, and evaluation protocols. Our experiments address four key research questions:
(1) **RQ1** (§4.2): Does concept-level regularization reduce catastrophic forgetting more effectively than state-of-the-art methods?
(2) **RQ2** (§4.3): Does C^{3}L mitigate spurious correlations and explanation drift?

(3) **RQ3** (§4.4): What are the contributions of C^{3}L's key components and design choices - dual regularization, buffer size, spatial grounding, explanation methods, task granularity, and concept discovery mechanism?
(4) **RQ4** (§4.5): Can C^{3}L's training mechanism improve model robustness to out-of-distribution shifts (ImageNet-A and ObjectNet)? **Success here suggests the model has internalized and generalized transferable semantic concepts rather than dataset-specific artifacts.**

4.1 Experimental Setup

Datasets. We evaluate on five benchmarks spanning fine-grained recognition and out-of-distribution generalization: CUB-200-2011 (**CUB**) [59] (200 bird species), **DomainNet** [60] (345 classes across 6 domains), **ImageNet-R** [61] (200 classes with artistic renditions), **ImageNet-A** [62] (200 classes with natural adversarial examples), **ObjectNet** [63] (113 classes with pose/background variations). These datasets represent both typical class-incremental benchmarks (CUB, DomainNet) and challenging distribution shifts (ImageNet-R/A, ObjectNet) that test robustness beyond standard continual learning settings. Also, they exhibit a significant domain gap with ImageNet (i.e., the pre-trained dataset).

Following standard protocols [5, 22], we split datasets into 10 sequential tasks (5 for DomainNet) with equal classes per task under class-incremental learning: task boundaries known during training but not at test, with classification over all encountered classes. For ObjectNet, 3 randomly sampled classes are excluded so that the remaining 110 classes divide evenly into tasks (Appendix B).

Evaluation metrics. We report three standard continual learning metrics [4, 64]: (i) **Average Accuracy (ACC)** $\uparrow$: the average accuracy on all the tasks after the last task has been learned, (ii) **Average Forgetting (F)** $\downarrow$: mean accuracy drop on previous tasks, $F = \frac{1}{T-1}\sum_{i=1}^{T-1} \max_{j\in\{i,\dots,T\}} A_{i,j} - A_{i,T}$, where $A_{i,j}$ is accuracy on task i after training on task j, (iii) **Average Incremental Accuracy (AIA)** $\uparrow$: mean accuracy after learning each task, capturing the model's performance throughout the entire learning trajectory. For explanation evaluation (§4.3), we measure: (i) **Spurious Correlation Mitigation (Debiasing)**: following Lee et al. [65] and Bayasi et al. [66] we evaluate ACC over both Fitzpatrick17K (**FITZ**) [66] and Split ImageNet-100C (**I-100C**) [65] to evaluate the model's reliance on spurious correlations. (ii) **Semantic Localization**: following Chefer et al. [41] we measure the mIoU, pixel accuracy (pixAcc), and mAP of the explanations on CUB. More details provided in Appendix G.

Baselines. We compare against nearly 20 state-of-the-art methods spanning multiple paradigms. For CNNs: *Regularization-based*: EWC [6], LwF [67], PEC [21]; *Replay-based*: ER [68], iCaRL [4], DER++ [5], ER-ACE [18], CSReL-CL [19]; *Explanation-aware*: RRR [10], EPR [13] TASS [14]. For ViTs: *Prompt-based*: L2P [9], DualPrompt [22], CODA-Prompt [23]; *Adapter-based*: MOS [24], MIN [25], PLAN [26]. We also provide a ViT-based implementation for RRR named RRR-ViT which uses Grad-CAM-ViT [41]. Joint Training is standard supervised fine-tuning on the i.i.d. union of all tasks' data, without concept supervision; it is commonly regarded as an upper bound for continual learning methods that use the same supervision. In this work, we focus on continual learning for widely used architectures such as CNNs and ViTs. While C^3L could potentially be extended to concept bottleneck models (CBMs), systematically adapting CBMs to continual learning constitutes a research problem in its own and is therefore left for future work.

Controlled comparisons. To isolate the benefit of the choice of explanation method from C^3L, we implement two controlled variants of RRR, called RRR-LRP and RRR-ViT-AttnLRP (for ResNet-18 and ViT-B respectively), that use our explanation methods instead of Grad-CAM. To isolate replay benefits, we add replay-augmented variants (-R) following Wang et al. [9].

All replay methods use identical buffer size ($|\mathcal{M}| = 1000$) for fair comparison. Section 4.4 shows that C^3L achieves competitive results against all baselines (with $|\mathcal{M}| = 1000$) using a small buffer of size $|\mathcal{M}| = 200$. We follow original code-bases for implementation and hyperparameter settings to ensure the best possible performance.

Implementation Details. We use ResNet-18 [69] (ImageNet-1K pretrained) and ViT-B/16 [70] (ImageNet-21K pretrained). C^3L employs LRP [42] for CNNs and AttnLRP [44] for ViTs; all baselines use Grad-CAM [39] as a unified evaluation protocol. Concept discovery follows Oikarinen et al. [15] with spatial grounding via SAM 3 [16]. Full configuration details are provided in Appendix B.

Training Details. Regularization is applied exclusively to replay samples ($|\mathcal{M}| = 1000$). Results are averaged over 5 runs (mean $\pm$ std), with statistical significance assessed via paired t-tests ($p < 0.05$). Full hyperparameter settings and training protocols are provided in Appendix B.

Finally, comprehensive implementation details are provided in the Appendix. Across all tables, the best results are in **bold**, and the second-best are underlined.

4.2 RQ1: Standard Continual Learning Performance

Table 1 presents comprehensive results across ResNet-18 and ViT-B. Due to space constraints, the complete set of results, including all evaluated methods, is reported on Table 17 in the Appendix. C^3L achieves consistent improvements in accuracy and forgetting across all datasets, outperforming state-of-the-art and explanation-based methods. Notably, $\mathcal{L}_{\text{concept}}$ could also be applied to all training samples. However, we restrict it to replay examples to ensure a fair comparison with prior replay methods. Results for this variant are reported in the full results table in the Appendix.

A critical question is whether gains stem from using superior explanation methods (LRP vs. Grad-CAM) rather than concept-level regularization. Our controlled comparisons provide clear evidence for the latter. RRR-ViT-AttnLRP (using AttnLRP instead of Grad-CAM) improves over RRR-ViT.

Table 1: Quantitative comparison of Continual Learning performance across five benchmarks using ResNet-18 and ViT-B architectures. We report Average Accuracy (ACC), Average Forgetting (F), and Average Incremental Accuracy (AIA). We report results over 5 trials.

Method	CUB			DomainNet			ImageNet-R			ImageNet-A			ObjectNet		
	ACC ↑	F ↓	AIA ↑	ACC ↑	F ↓	AIA ↑	ACC ↑	F ↓	AIA ↑	ACC ↑	F ↓	AIA ↑	ACC ↑	F ↓	AIA ↑
ResNet-18															
Joint Training	83.04±0.51	-	-	74.99±0.22	-	-	76.31±0.90	-	-	71.22±0.42	-	-	73.42±0.91	-	-
Replay-based															
ER	64.34±0.92	14.25±0.58	68.42±0.84	61.92±0.81	8.92±0.55	61.99±0.76	63.38±0.68	11.85±0.49	66.15±0.72	55.01±0.45	19.43±0.66	59.80±0.38	57.84±0.52	17.24±0.58	62.01±0.55
DER++	71.45±0.72	9.62±0.45	75.14±0.65	68.02±0.58	5.74±0.32	67.70±0.55	69.81±0.48	7.82±0.35	72.15±0.52	61.43±0.38	12.53±0.44	66.41±0.32	64.44±0.42	11.34±0.48	68.80±0.45
ER-ACE	69.73±0.78	10.85±0.49	73.32±0.72	66.44±0.65	6.42±0.38	66.22±0.62	68.21±0.55	8.93±0.42	70.75±0.58	59.83±0.42	14.82±0.52	64.71±0.36	62.74±0.46	13.14±0.52	67.11±0.48
CSReL-CL	71.92±0.82	9.25±0.52	75.54±0.76	67.44±0.69	6.12±0.42	67.23±0.65	69.42±0.59	8.24±0.46	71.83±0.62	60.74±0.45	13.42±0.56	65.51±0.39	63.83±0.49	12.05±0.56	67.94±0.53
Explanation-aware															
RRR	72.21±0.78	9.10±0.44	75.81±0.73	68.25±0.33	5.66±1.06	67.95±0.54	70.13±0.65	7.66±0.62	72.43±0.50	61.73±0.43	12.06±0.21	66.68±0.31	64.82±0.32	10.87±0.35	69.16±0.64
RRR-LRP	73.25±0.65	8.42±0.38	76.74±0.58	69.30±0.52	5.13±1.28	68.92±0.49	71.15±0.45	6.92±0.92	73.32±0.46	63.01±0.35	10.24±0.69	68.12±0.28	66.20±0.36	9.23±0.42	70.74±0.41
EPR	72.86±0.72	8.75±0.41	76.32±0.65	68.82±0.42	5.36±0.82	68.41±0.51	70.68±0.55	7.21±0.84	72.88±0.48	62.45±0.39	11.15±0.52	67.42±0.30	65.54±1.04	10.02±0.38	70.05±0.52
TASS	70.24±0.85	10.35±0.56	74.12±0.78	66.91±0.72	6.74±0.45	66.73±0.68	68.92±0.62	8.63±0.49	71.42±0.65	60.13±0.48	14.25±0.62	65.01±0.42	63.01±0.52	12.83±0.58	67.44±0.55
C^{3}L	**79.73±0.31**	**5.32±0.63**	**82.24±0.35**	**73.01±0.23**	**3.14±0.55**	**72.51±0.88**	**74.35±0.46**	**4.23±0.29**	**75.84±0.65**	**67.84±0.72**	**6.52±0.46**	**70.83±0.41**	**71.12±0.53**	**5.82±0.25**	**72.23±0.92**
ViT-B															
Joint Training	86.12±0.23	-	-	81.31±0.29	-	-	84.59±0.31	-	-	75.31±0.19	-	-	78.96±0.24	-	-
Prompt-based															
L2P	70.82±1.24	7.94±0.65	74.24±0.91	71.85±0.78	4.23±0.42	72.40±0.83	71.49±0.69	4.85±0.56	71.91±0.74	54.94±0.82	6.83±0.44	56.11±0.57	64.53±0.70	6.24±0.55	66.12±0.69
CODA-Prompt	74.74±0.79	5.42±0.42	78.13±0.67	74.95±0.82	3.12±0.38	75.50±0.78	72.35±0.65	3.82±0.41	72.69±0.64	54.24±0.68	6.92±0.40	55.49±0.53	67.82±0.66	4.83±0.42	68.24±0.55
Adapter-based															
MOS	75.94±0.68	4.82±0.41	79.53±0.55	76.14±0.70	2.12±0.39	76.81±0.66	76.88±0.24	4.15±0.61	78.35±0.53	61.34±0.68	3.52±0.29	62.26±0.35	71.53±0.67	3.54±0.31	71.82±0.56
PLAN	76.12±0.70	4.83±0.42	79.74±0.56	76.03±0.83	2.24±0.28	76.76±0.67	76.62±0.55	4.02±0.31	78.93±0.84	60.25±0.69	4.52±0.32	61.14±0.55	70.82±0.68	3.42±0.30	71.23±1.03
Replay-based															
RRR-ViT	76.94±1.17	4.82±0.61	78.53±1.14	78.32±1.21	2.01±0.49	78.19±1.16	78.83±0.94	4.12±0.50	80.84±0.93	60.52±1.18	5.43±0.23	63.24±0.64	70.19±1.17	3.82±0.51	70.94±1.16
RRR-ViT-AttnLRP	78.09±0.85	4.17±0.58	80.34±0.88	78.47±1.08	1.92±0.37	78.56±0.90	79.48±0.76	3.35±1.01	80.87±0.66	63.02±0.56	4.24±0.26	63.51±0.59	71.61±0.87	3.33±0.42	71.88±0.89
L2P-R	69.62±1.25	6.34±0.56	77.25±0.92	74.23±0.29	3.42±0.53	74.82±1.04	76.82±0.90	3.91±0.57	77.34±0.89	58.33±0.23	6.14±0.21	59.71±0.68	67.24±1.21	5.32±0.56	68.15±0.61
CODA-R	77.53±0.99	4.23±0.42	80.24±0.77	78.02±0.82	3.02±0.28	78.21±0.78	78.54±0.66	2.82±0.31	79.52±0.64	61.24±0.89	4.82±0.42	61.43±0.55	71.53±0.78	3.24±0.30	71.62±0.67
MOS-R	78.82±0.65	3.84±0.29	82.52±0.53	79.03±0.78	1.52±0.42	79.42±0.54	80.92±0.42	2.43±0.28	81.55±0.31	67.84±0.96	2.24±0.27	66.25±0.42	73.62±0.35	2.52±0.29	73.89±0.93
PLAN-R	79.24±0.56	3.52±0.28	82.14±0.54	78.62±0.59	1.83±0.27	78.92±0.65	80.13±0.51	2.58±1.29	80.89±0.42	65.48±0.57	3.05±0.31	65.83±0.51	72.85±0.56	2.84±0.28	73.06±0.55
C^{3}L	**83.65±0.28**	**2.92±0.24**	**84.64±0.37**	**80.43±0.55**	**1.23±0.62**	**80.82±0.26**	**82.74±0.76**	**1.89±0.33**	**83.92±0.45**	**71.14±0.59**	**1.53±0.41**	**70.93±0.38**	**76.83±0.56**	**1.92±0.83**	**77.92±0.78**

However, C^{3}L with AttnLRP achieves far greater gains. Since both use identical AttnLRP, this provides evidence for the benefit of concept-level over pixel-level regularization: semantically grounded regions provide substantially stronger supervision than raw saliency. The same pattern holds for ResNet-18.

Replay-based methods mostly outperform other methods in both CNNs and ViTs. This aligns with recent findings [20] showing that replay-based methods benefit from preserving feature learning dynamics. We further observe that for ResNet-18, explanation-aware methods achieve very good performance, and are mostly better than regular replay-based methods. However, pixel-level methods may not distinguish between spurious correlations and semantic features. In contrast, C^{3}L focuses on validated semantic concepts, providing cleaner supervision, which is also evident by the performance. This gap is further emphasized on Sec. 4.3. ViTs exhibit substantially lower forgetting than CNNs across all methods, likely reflecting the more stable representations of large-scale pretrained transformers. Notably, RRR-ViT sometimes outperforms prompt-based methods with replay (L2P-R, CODA-R), demonstrating that explanation consistency provides valuable regularization even for transformers. Nevertheless, C^{3}L maintains consistent advantages across both architectures, demonstrating its effectiveness across distinct architectural families.

4.3 RQ2: Explanation Drift and Spurious Correlations

Standard continual learning protocols do not assess whether methods mitigate spurious correlations or maintain focus on semantically meaningful features. To address this limitation, we adopt two complementary evaluation protocols: (1) **Spurious Correlation Mitigation**, measured by accuracy on FITZ and I100C, evaluating model reliance on spurious features (skin tone in FITZ and injected noise in I-100C); (2) **Semantic Localization**, evaluated using mIoU, pixAcc, and mAP of explanations on CUB annotated segments[1], to quantify explanation drift. For (1), we follow prior work [66, 65] and fine-tune a CNN model (ResNet-18 pretrained on ImageNet-1K). For (2), we evaluate the same CUB-finetuned model from Section 4.2. We compare C^{3}L against representative baselines from Table 1 and Table 17. In addition, for protocol (1), we include two recent class-incremental learning methods for debiasing: BiasPruner (**BP**) [66] and **BGS** [65]. Further experimental and implementation details provided in Appendix G. As shown in Table 2, C^{3}L consistently achieves the best performance across both accuracy and semantic localization metrics. The improvement in accuracy is consistent with our hypothesis that C^{3}L anchors attention to semantically relevant features, thereby preventing the emergence of spurious correlations. Notably, C^{3}L outperforms specialized debiasing methods on accuracy across both datasets. In contrast, pixel-level explanation-aware baselines (RRR, EPR) and other continual learning methods (L2P, MOS, MIN) show lower semantic localization scores and spurious correlation mitigation performance. These results show that concept-grounded regularization systematically redirects model attention from irrelevant cues to semantically meaningful features,

[1] `https://data.caltech.edu/records/w9d68-gec53`

thereby mitigating explanation drift, improving the quality of model reasoning, and limiting the accumulation of spurious correlations.

Table 2: Evaluation of explanation quality and stability using segmentation metrics (mIoU, pixAcc, mAP) on the CUB dataset and ACC on the FITZ and the I-100C datasets. Higher is better for all metrics.

	ResNet-18				
Metric	RRR	EPR	BGS	BP	**C^{3}L**
ACC (FITZ)	38.69	40.58	56.85	58.13	**60.47**
ACC (I-100C)	31.49	32.97	51.39	48.52	**64.19**
mIoU	52.59	53.83	49.65	50.24	**58.77**
pixAcc	66.92	67.40	63.22	62.47	**79.96**
mAP	69.71	68.32	65.89	61.51	**83.92**
	ViT-B				
Metric	RRR-ViT	L2P	MOS	MIN	**C^{3}L**
mIoU	54.55	42.02	54.17	51.39	**59.86**
pixAcc	72.05	56.15	71.45	62.23	**80.51**
mAP	68.28	43.59	67.86	55.43	**85.83**

Table 3: Ablation study analyzing the impact of removing concept regularization ($\mathcal{L}_{concept}$) and background stability ($\mathcal{L}_{stab}$) terms.

Configuration	**ImageNet-R**			**CUB**		
	ACC ↑	F ↓	AIA ↑	ACC ↑	F ↓	AIA ↑
C^{3}L	82.74	1.89	83.92	83.65	2.92	84.64
w/o $\mathcal{L}_{\text{concept}}$	77.12	3.82	78.13	73.74	5.19	72.95
w/o $\mathcal{L}_{\text{stab}}$	78.94	3.01	78.32	78.42	3.86	78.98
w/o both	72.27	9.13	71.56	70.92	8.48	71.40

4.4 RQ3: Ablation Studies

We dissect C^{3}L's components to understand their individual contributions. All ablations use ImageNet-R with ViT-B unless otherwise noted. We address: (1) the benefit of each regularization term, (2) impact of replay buffer size, (3) concept-level vs. object-level segmentation (§E.4) - **emphasizes the benefit from using concept guidance**, (4) alternative explanation methods (§E.2), (5) task granularity (§E.3), (6) effect of replay vs. no-replay training using concepts (§E.1), and (7) SAM 3 vs. Grounded SAM (§E.5). Space constraints limit the main paper to questions (1) and (2). The remaining ablations appear in the Appendix E.

Regularization terms. Table 3 shows results for component ablations. On ImageNet-R, both terms cause similar performance drops, indicating both are important. On CUB, removing $\mathcal{L}_{\text{concept}}$ causes larger drops, suggesting this term is especially impactful for fine-grained datasets by strengthening concept attention. Yet, removing $\mathcal{L}_{\text{stab}}$ also hurts performance, confirming its role preventing drift to spurious backgrounds and textures that correlate with specific tasks. The asymmetric importance confirms our design intuition: concept overlap provides positive supervision, while stability provides negative supervision.

Impact of Replay Buffer Size. We evaluate C^{3}L on ImageNet-R using the ViT-B backbone with varying replay buffer sizes $|\mathcal{M}| \in \{200, 500, 1000\}$. Standard replay methods degrade severely with smaller buffers due to overfitting and reduced representative variance. As presented in Table 4, C^{3}L demonstrates robustness to reduced replay capacity. With the standard buffer size of $|\mathcal{M}| = 1000$, C^{3}L consistently outperforms all baselines. Notably, even when the buffer size is reduced to $|\mathcal{M}| = 500$, our method maintains superior performance, surpassing all compared methods, even those using the full 1,000-sample buffer. In the highly constrained regime of $|\mathcal{M}| = 200$, C^{3}L remains highly competitive. While the performance gap naturally narrows, our method frequently matches and even exceeds the performance of strong baselines trained with larger buffers. This data efficiency is consistent with our hypothesis: concept-grounded regularization anchors the model to semantic features which can be more robust and transferable than the pixel-level patterns relied upon by standard replay and saliency methods, thus requiring fewer examples to maintain stability.

4.5 RQ4: Robustness to Out-of-Distribution (OOD) Samples

We evaluate robustness to OOD samples using a setup inspired by [71], which studies feature transferability by training on a subset of ImageNet classes. Specifically, we construct a compact fine-tuning dataset for C^{3}L by randomly selecting 5 images per class from half of the ImageNet-1K [72] categories, resulting in a total of 2,500 training images. This sparse sampling is motivated by computational considerations and serves to assess data efficiency under limited supervision, while random class selection ensures broad semantic coverage. Experiments are conducted using ResNet-18 and ViT-B models pretrained on ImageNet-1K. Robustness is evaluated specifically on ImageNet-A and ObjectNet (restricted to classes overlapping ImageNet-1K) as these benchmarks

Table 4: Performance analysis on ImageNet-R (ViT-B) with varying replay buffer sizes ($|\mathcal{M}|$) to assess data efficiency.

Method	$\|\mathcal{M}\|$ =200			$\|\mathcal{M}\|$ =500			$\|\mathcal{M}\|$ =1,000		
	ACC ↑	F ↓	AIA ↑	ACC ↑	F ↓	AIA ↑	ACC ↑	F ↓	AIA ↑
RRR-ViT	71.48	7.25	70.20	73.83	6.11	73.55	78.83	4.12	80.84
L2P-R	72.83	4.44	72.98	73.60	4.18	76.32	76.82	3.91	77.34
CODA-R	73.80	3.43	73.41	75.13	3.16	77.25	78.54	2.82	79.52
MOS-R	77.28	3.88	78.95	78.57	3.21	78.92	80.92	2.43	81.55
PLAN-R	77.78	3.45	79.65	78.23	3.18	80.11	80.13	2.58	80.89
C^{3}L	**80.40**	**2.65**	**81.89**	**81.26**	**2.11**	**82.99**	**82.74**	**1.89**	**83.92**

Table 5: Out-of-distribution robustness evaluation. We report accuracy before and after fine-tuning. For ResNet-18 we compared: RRR / C^{3}L, and for ViT-B: RRR-ViT / C^{3}L.

Model	Dataset	Before	After
ResNet-18	ImageNet-A	8.59	7.92 / **20.15**
	ObjectNet	21.88	21.56 / **36.93**
ViT-B	ImageNet-A	13.48	13.18 / **30.82**
	ObjectNet	33.92	34.49 / **58.04**

contain challenging samples that commonly cause failures in ImageNet-pretrained models. Further details are provided in the Appendix. Table 5 details the classification accuracy before and after fine-tuning with C^{3}L in comparison with RRR (RRR-ViT). We observe a significant increase in accuracy when employing C^{3}L's semantic-based regularization. In contrast, RRR yields only negligible gains (and sometimes even decrease the accuracy), highlighting its lower robustness capabilities in this setting. This isolates the impact of semantic features, demonstrating their substantial capacity to modulate model behavior. Full experiments, including EPR and TASS, along segmentation evaluation are provided in the Appendix.

5 Conclusion

We presented C^{3}L, a framework that breaks the self-referential drift loop and mitigates explanation drift by anchoring representations to semantically grounded concepts. By regularizing attention toward semantic attributes rather than unstable pixel-level saliency, C^{3}L promotes learning of transferable features that better resist spurious correlations and explanation drifts. Our automated framework eliminates the need for manual annotation, making concept-based supervision scalable for continual learning. Extensive evaluation across five benchmarks demonstrates that C^{3}L establishes state-of-the-art performance in continual learning, particularly under challenging out-of-distribution shifts. Limitations and future work are detailed in the Appendix.

Acknowledgments and Disclosure of Funding

This work was supported by the Ministry of Innovation, Science & Technology, Israel, and by the Israeli Science Foundation (ISF grant 977/26).

References

[1] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165, 1989.

[2] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4): 128–135, 1999.

[3] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

[4] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2001–2010, 2017.

[5] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 15920–15930, 2020.

[6] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

[7] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning (ICML)*, pages 3987–3995, 2017.

[8] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

[9] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 139–149, 2022.

[10] Sayna Ebrahimi, Suzanne Petryk, Akshay Gokul, William Gan, Joseph E Gonzalez, Marcus Rohrbach, and Trevor Darrell. Remembering for the right reasons: Explanations reduce catastrophic forgetting. In *International Conference on Learning Representations (ICLR)*, 2021.

[11] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019.

[12] Yehonatan Elisha, Oren Barkan, and Noam Koenigstein. Concept-guided fine-tuning: Steering vits away from spurious correlations to improve robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17035–17045, June 2026.

[13] Gobinda Saha and Kaushik Roy. Saliency guided experience packing for replay in continual learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5273–5283, 2023.

[14] Xialei Liu, Jiang-Tian Zhai, Andrew D Bagdanov, Ke Li, and Ming-Ming Cheng. Task-adaptive saliency guidance for exemplar-free class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23954–23963, 2024.

[15] Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *International Conference on Learning Representations (ICLR)*, 2023.

[16] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations (ICLR)*, 2026.

[17] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning (ICML)*, pages 5338–5348, 2020.

[18] Lucas Caccia, Rahaf Aljundi, Nader Asadi, Tinne Tuytelaars, Joelle Pineau, and Eugene Belilovsky. New insights on reducing abrupt representation change in online continual learning. In *International Conference on Learning Representations (ICLR)*, 2022.

[19] Ruilin Tong, Yuhang Liu, Javen Qinfeng Shi, and Dong Gong. Coreset selection via reducible loss in continual learning. In *International Conference on Learning Representations*, volume 2025, pages 13964–13999, 2025.

[20] Hongyi Wan, Shiyuan Ren, Wei Huang, Miao Zhang, Xiang Deng, Yixin Bao, and Liqiang Nie. Understanding the forgetting of (replay-based) continual learning via feature learning: Angle matters. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=6UIer20oYA`.

[21] Michał Zając, Tinne Tuytelaars, and Gido van de Ven. Prediction error-based classification for class-incremental learning. In *International Conference on Learning Representations*, volume 2024, pages 8239–8267, 2024.

[22] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*, pages 631–648, 2022.

[23] James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11909–11919, 2023.

[24] Hai-Long Sun, Da-Wei Zhou, Hanbin Zhao, Le Gan, De-Chuan Zhan, and Han-Jia Ye. MOS: Model surgery for pre-trained model-based class-incremental learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pages 20699–20707, 2025.

[25] Kai Jiang, Zhengyan Shi, Dell Zhang, Hongyuan Zhang, and Xuelong Li. Mixture of noise for pre-trained model-based class-incremental learning. *Advances in neural information processing systems*, 38: 44776–44802, 2025.

[26] Xiequn Wang, Zhan Zhuang, and Yu Zhang. Plan: Proactive low-rank allocation for continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2909–2918, 2025.

[27] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1135–1144, 2016.

[28] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

[29] Oren Barkan, Yuval Asher, Amit Eshel, Yehonatan Elisha, and Noam Koenigstein. Learning to explain: A model-agnostic framework for explaining black box models. In *IEEE International Conference on Data Mining (ICDM)*, pages 944–949, 2023.

[30] Oren Barkan, Veronika Bogina, Liya Gurevitch, Yuval Asher, and Noam Koenigstein. A counterfactual framework for learning and evaluating explanations for recommender systems. In *Proceedings of the ACM Web Conference (WWW)*, pages 3723–3733, 2024.

[31] Liya Gurevitch, Veronika Bogina, Oren Barkan, Yahlly Schein, Yehonatan Elisha, and Noam Koenigstein. LXR: Learning to explain recommendations. *ACM Transactions on Recommender Systems*, 4(2), 2025.

[32] Mikhail Baklanov, Veronika Bogina, Yehonatan Elisha, Yahlly Schein, Liron Allerhand, Oren Barkan, and Noam Koenigstein. Refining fidelity metrics for explainable recommendations. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2025.

[33] Oren Barkan, Yahlly Schein, Yehonatan Elisha, Veronika Bogina, Mikhail Baklanov, and Noam Koenigstein. Fidelity-aware recommendation explanations via stochastic path integration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

[34] Dor Arviv, Yehonatan Elisha, Oren Barkan, and Noam Koenigstein. Extracting interaction-aware monosemantic concepts in recommender systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

[35] Oren Barkan, Edan Hauon, Avi Caciularu, Ori Katz, Itzik Malkiel, Omri Armstrong, and Noam Koenigstein. Grad-SAM: Explaining transformers via gradient self-attention maps. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 2882–2887, 2021.

[36] Oren Barkan, Yonatan Toib, Yehonatan Elisha, and Noam Koenigstein. A learning-based approach for explaining language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM)*, 2024.

[37] Oren Barkan, Yonatan Toib, Yehonatan Elisha, Jonathan Weill, and Noam Koenigstein. LLM explainability via attributive masking learning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

[38] Oren Barkan, Yehonatan Elisha, Yonatan Toib, Jonathan Weill, and Noam Koenigstein. Improving LLM attributions with randomized path-integration. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

[39] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[40] Oren Barkan, Omri Armstrong, Amir Hertz, Avi Caciularu, Ori Katz, Itzik Malkiel, and Noam Koenigstein. GAM: Explainable visual similarity and classification via gradient activation maps. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM)*, 2021.

[41] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021.

[42] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[43] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 193–209, 2019.

[44] Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. AttnLRP: Attention-aware layer-wise relevance propagation for transformers. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 135–168. PMLR, 2024.

[45] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328, 2017.

[46] Oren Barkan, Yehonatan Elisha, Yuval Asher, Amit Eshel, and Noam Koenigstein. Visual explanations via iterated integrated attributions. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023.

[47] Oren Barkan, Yehonatan Elisha, Jonathan Weill, Yuval Asher, Amit Eshel, and Noam Koenigstein. Deep integrated explanations. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM)*, 2023.

[48] Oren Barkan, Yehonatan Elisha, Jonathan Weill, Yuval Asher, Amit Eshel, and Noam Koenigstein. Stochastic integrated explanations for vision models. In *IEEE International Conference on Data Mining (ICDM)*, 2023.

[49] Yehonatan Elisha, Oren Barkan, and Noam Koenigstein. Probabilistic path integration with mixture of baseline distributions. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM)*, 2024.

[50] Oren Barkan, Yehonatan Elisha, Jonathan Weill, and Noam Koenigstein. BEE: Metric-adapted explanations via baseline exploration-exploitation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[51] Ziv Weiss Haddad, Oren Barkan, Yehonatan Elisha, and Noam Koenigstein. Soft local completeness: Rethinking completeness in XAI. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[52] Yehonatan Elisha, Seffi Cohen, Oren Barkan, and Noam Koenigstein. Rethinking saliency maps: A cognitive human aligned taxonomy and evaluation framework for explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

[53] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning (ICML)*, pages 2668–2677, 2018.

[54] Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence*, 5:1006–1019, 2023.

[55] Yehonatan Elisha, Oren Barkan, Ziv Weiss Haddad, and Noam Koenigstein. ConEx: Human-interpretable saliency maps via concept-aware attribution. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026.

[56] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9505–9515, 2018.

[57] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un)reliability of saliency methods. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 267–280, 2019.

[58] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3681–3688, 2019.

[59] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[60] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1406–1415, 2019.

[61] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021.

[62] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.

[63] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[64] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 532–547, 2018.

[65] Donggyu Lee, Sangwon Jung, and Taesup Moon. Continual learning in the presence of spurious correlations: Analyses and a simple baseline. In *12th International Conference on Learning Representations, ICLR 2024*, 2024.

[66] Nourhan Bayasi, Jamil Fayyad, Alceu Bissoto, Ghassan Hamarneh, and Rafeef Garbi. Biaspruner: Debiased continual learning for medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 90–101. Springer, 2024.

[67] Zhizhong Li and Derek Hoiem. Learning without forgetting. In *European Conference on Computer Vision (ECCV)*, pages 614–629, 2016.

[68] Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *ICLR (Poster)*, 2019.

[69] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[70] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[71] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

[72] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[73] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[74] Sabina B van Rooij and Gertjan J Burghouts. Guided sam: Label-efficient part segmentation. In *International Conference on Pattern Recognition*, pages 291–306. Springer, 2024.

[75] Zichao Zeng and Jan Boehm. Exploration of an open vocabulary model on semantic segmentation for street scene imagery. *ISPRS International Journal of Geo-Information*, 13(5), 2024. ISSN 2220-9964. doi: 10.3390/ijgi13050153. URL `https://www.mdpi.com/2220-9964/13/5/153`.

[76] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Hao Zhang, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

[77] Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1820–1828, 2021.

[78] Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, Qihang Yu, and Alan Yuille. Partimagenet: A large, high-quality dataset of parts. In *European Conference on Computer Vision*, pages 128–145. Springer, 2022.

[79] Katharina Prasse, Patrick Knab, Sascha Marton, Christian Bartelt, and Margret Keuper. Dcbm: Data-efficient visual concept bottleneck models. *arXiv preprint arXiv:2412.11576*, 2024.

[80] Sukrut Rao, Sweta Mahajan, Moritz Böhle, and Bernt Schiele. Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In *European Conference on Computer Vision*, pages 444–461. Springer, 2024.

[81] Itay Benou and Tammy Riklin Raviv. Show and tell: Visually explainable deep neural nets via spatially-aware concept bottleneck models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30063–30072, 2025.

A Appendix Overview

This appendix provides comprehensive supplementary materials to support the main paper. We organize the content as follows:

- **Appendix B**: Detailed implementation specifications, hyperparameters, and training protocols.
- **Appendix C**: Theoretical and empirical justification for method design choices.
- **Appendix D**: Discussion and analysis about the crowding-out phenomenon.
- **Appendix E**: Extended ablation studies including replay contribution, explanation methods, task granularity, and segmentation approaches.
- **Appendix F**: Additional experiments on non-pretrained networks and results for Table 1.
- **Appendix G**: Detailed explanation quality and robustness (RQ4) experimental details with full results.
- **Appendix H**: Discussion on foundation model usage and fairness and PartImageNet experiment.
- **Appendix I**: Qualitative visualizations and case studies.
- **Appendix J**: Failure cases analysis.
- **Appendix K**: Discussion on limitations of our method and future work.

B Implementation Details

Full Implementation Details. We use **ResNet-18** [69] (ImageNet-1K pretrained) for CNN experiments and ViT-B/16 **(ViT-B)** [70] (ImageNet-21K pretrained) for transformer experiments. For explanation generation, we employ LRP [42] with ϵ-rule ($\epsilon = 10^{-6}$) for CNNs and AttnLRP [44] for ViTs. For saliency map evaluation (e.g., Tab. 2), explanations for all baselines are generated post-training using Grad-CAM [39] for CNNs and Grad-CAM-ViT [41] for ViTs. While C^3L yields marginally higher scores when evaluated on its native LRP maps, we report Grad-CAM results to maintain a unified evaluation protocol. C^3L's saliency maps are min-max normalized to $[0, 1]$ prior to BCE computation. Concept discovery follows the approach from Oikarinen et al. [15] which resulting in a concept set per class, filtered by occurrence rate ($\nu_{\text{occ}} \geq 0.15$) and coverage ($\nu_{\text{cov}} \geq 0.2$). Spatial grounding uses SAM 3 [16] with confidence threshold $\tau_{\text{conf}} = 0.3$. We validate these configuration choices in Appendix E.

Full Training Details. We apply regularization only to replay samples, as current task samples lack reference explanations. The replay buffer $\mathcal{M}$ (size $|\mathcal{M}| = 1000$) is populated via uniform random sampling from previous tasks. Hyperparameters: $\lambda_{\text{concept}} = 1.0$, $\lambda_{\text{stab}} = 1.5$ for CNNs, and $\lambda_{\text{concept}} = 0.8$, $\lambda_{\text{stab}} = 1.2$ for ViTs. These values were selected based on validation set performance (validated in the Appendix). Results are averaged over 5 runs with different random seeds and task orderings. We report mean $\pm$ standard deviation and assess statistical significance ($p < 0.05$) using paired t-tests comparing C^3L against the best-performing baseline in each experimental setting.

Training Protocols. For our method, we employ stochastic gradient descent (SGD) for both model architectures, using a batch size of 128 and a learning rate of 0.001. Training is conducted for 20 epochs on ImageNet-R, ImageNet-A, ObjectNet, and DomainNet, and for 50 epochs on the CUB dataset. All baselines were used according to the recommended setup reported in their papers and code.

Task Splits. All datasets are split into tasks with an equal number of classes per task: CUB (200 classes), ImageNet-R (200), and ImageNet-A (200) into 10 tasks of 20 classes, and DomainNet (345 classes) into 5 tasks of 69 classes. ObjectNet contains 113 classes that overlap with ImageNet-1K, which cannot be divided evenly into 10 tasks; we therefore randomly exclude 3 classes and split the remaining 110 classes into 10 tasks of 11 classes each. All methods use the same class subset and splits.

Table 6: Evaluation on different background design for C^3L using ViT-B on ImageNet-R and CUB.

Configuration	ImageNet-R			CUB		
	ACC ↑	F ↓	AIA ↑	ACC ↑	F ↓	AIA ↑
C^3L	82.74	1.89	83.92	83.65	2.92	84.64
C^3L-NoReplay	79.82	3.84	78.99	80.94	3.27	81.78
w/o $\mathcal{L}_{\text{stab}}$	78.94	3.01	78.32	78.42	3.86	78.98

Concept Discovery. We used the implementation from [15] with GPT-4o-mini [73] and used the prompts from the original github repository [2] without the "around" prompts.

Spatial Grounding Parameters. For SAM 3 segmentation, we use confidence threshold $\tau_{\text{conf}} = 0.3$ and used the Ultralytics implementation [3]. Lower thresholds increase false positives (irrelevant regions), while higher thresholds miss valid concept instances. We validate this choice in Appendix E.7.

Replay Buffer. The replay buffer size is set to $|\mathcal{M}| = 1000$ for fair comparison with baselines. We use uniform random sampling across previous tasks to populate the buffer. Each task contributes equally to maintain balanced representation.

Computational Environment. All experiments are conducted on NVIDIA A100 GPUs using PyTorch. For both ResNet-18 and ViT-B experiments, we use single-GPU training.

C Method Design Justification

C.1 Background Regularization Strategy

A central design choice in C^3L is how to regularize non-concept (background) regions. Unlike pixel-level alignment methods that indiscriminately anchor explanations to a reference saliency map, C^3L introduces a semantic prior that explicitly reallocates relevance toward intrinsic class attributes while constraining background behavior.

Rationale. Our design is motivated by three considerations.

First, *mitigating spurious correlations*: by enforcing full relevance within discovered concept regions, we can leverage the finite nature of the relevance budget. While relevance conservation is approximate in modern deep networks [43], empirically, maximizing relevance on foreground concepts encourages a redistribution of mass away from the background. This "crowding-out" effect helps suppress spurious cues without requiring explicit penalization. Second, *incomplete concept discovery*: automated concept generation may omit valid diagnostic features (e.g., fine-grained parts). Explicit background suppression risks eliminating such implicitly learned cues present in the pretrained model. Third, *context preservation*: background regions encode stable contextual information (e.g., pose, geometry, scale) learned during pretraining, which can support generalization and should not be erased.

To balance these factors, $\mathcal{L}_{\text{stab}}$ enforces *stability*, preserving the spatial structure of background relevance, rather than suppression. While the relative magnitude of background relevance naturally decreases due to relevance reallocation, anchoring its pattern prevents drift toward new, task-specific artifacts.

Empirical Support. We compare three strategies: (i) the full C^3L, (ii) no background regularization (w/o $\mathcal{L}_{\text{stab}}$ on Section 4.4), and (iii) explicit background suppression without replay (C^3L-NoReplay on Section E.1). Experiments with ViT-B on ImageNet-R and CUB (Table 6) show that stabilization outperforms both alternatives; since (iii) also removes replay, the comparison with (ii) is the one that isolates the background term, confirming that preserving pretrained contextual priors while preventing spurious correlations yields the best trade-off.

[2] https://github.com/Trustworthy-ML-Lab/Label-free-CBM

[3] https://docs.ultralytics.com/models/sam-3/

D Quantitative Evidence for the Crowding-Out Effect

In Section 3.3, we introduced the concept of *relevance reallocation* as a mechanism by which enforcing high attribution within concept regions indirectly suppresses spurious background correlations. This phenomenon, which we term the *crowding-out effect*, operates through the approximate conservation of relevance mass in neural networks [43]. Specifically, when the model is encouraged to concentrate relevance within semantically meaningful concept regions via $\mathcal{L}_{\text{concept}}$, the finite relevance budget naturally reallocates away from non-diagnostic background regions, even without explicit suppression.

While Section C.1 provided empirical evidence via ablation studies, a critical question remains: does relevance mass actually move from background to concept regions during training, or do both regions simply change independently? To directly test the crowding-out hypothesis, we track the spatial distribution of relevance across different image regions throughout continual learning.

D.1 Relevance Mass Flow Analysis

Experimental Protocol. We train C^3L on CUB using the standard 10-task split (20 classes per task) with a ViT-B backbone. For each sample x in the test set of task $\mathcal{T}_1$, we partition pixels into three regions:

(1) **Concept regions (C):** Pixels within $M^{(x)}_{\text{concept}}$, corresponding to discovered semantic attributes.

(2) **Background regions (B):** Pixels within $M^{(x)}_{\text{bg}}$, representing non-concept areas.

(3) **Total (T):** All pixels in the image.

At the end of each task $\mathcal{T} \in \{1, 3, 5, 8, 10\}$, we freeze the model and compute relevance maps S_x using AttnLRP [44]. For each region $R \in \{C, B, T\}$, we measure:

$$R_{\text{total}}(R) = \sum_{i \in R} S_x(i), \quad R_{\text{rel}}(R) = \frac{R_{\text{total}}(R)}{f_\theta(x)_y} \tag{6}$$

where $R_{\text{total}}(R)$ is the (signed) relevance assigned to region R and $R_{\text{rel}}(R)$ is the fraction of the target logit $f_\theta(x)_y$ attributed to it. Under approximate conservation (Eq. 4), $R_{\text{rel}}(T) \approx 1$ and $R_{\text{rel}}(C) \approx \rho_{\mathcal{C}}$, the concept share defined in Section 3.3. We also compute the **Concentration Ratio (CR)**:

$$\text{CR} = \frac{R_{\text{rel}}(C)}{R_{\text{rel}}(B)} \tag{7}$$

which directly captures the shift of relevance from background to concept regions. Under the crowding-out hypothesis, we expect: (1) $R_{\text{rel}}(C)$ to increase monotonically with training, (2) $R_{\text{rel}}(B)$ to decrease monotonically, (3) CR to increase, reflecting concentration toward concepts, (4) $R_{\text{rel}}(C) + R_{\text{rel}}(B) = R_{\text{rel}}(T)$ to remain close to 1 (approximate relevance conservation).

We average these metrics over all test samples from $\mathcal{T}_1$ and report mean values.

D.2 Results and Analysis

Table 7 presents the relevance mass flow analysis. The results provide strong quantitative evidence for the crowding-out effect:

As training progresses from $\mathcal{T}_1$ to $\mathcal{T}_{10}$, the fraction of the logit attributed to concept regions increases, while background relevance decreases. This shift occurs despite the model learning 180 new classes across subsequent tasks, which typically induces explanation drift toward spurious features [10]. The concentration ratio CR increases from 1.13 after $\mathcal{T}_1$ to 2.57 after $\mathcal{T}_{10}$. This demonstrates that relevance becomes progressively more concentrated in semantically meaningful concept regions relative to backgrounds. The relevance maps account for 98–100% of the target logit at every stage ($R_{\text{rel}}(T) \in [0.98, 1.00]$), so approximate conservation holds throughout training. The 0.20 gain in concept share is therefore matched by a 0.18 reduction in background share: relevance is reallocated from background to concept regions rather than added on top of it, consistent with the crowding-out effect described in Section 3.3.

Table 7: Relevance mass flow analysis across training tasks on CUB $\mathcal{T}_1$ test set. Results demonstrate reallocation of relevance from background to concept regions, validating the crowding-out mechanism.

Training Stage	$R_{\text{rel}}(C) \uparrow$	$R_{\text{rel}}(B) \downarrow$	CR $\uparrow$	$R_{\text{rel}}(T)$
After $\mathcal{T}_1$	0.52	0.46	1.13	0.98
After $\mathcal{T}_3$	0.61	0.37	1.65	0.98
After $\mathcal{T}_5$	0.65	0.33	1.97	0.98
After $\mathcal{T}_8$	0.71	0.28	2.54	0.99
After $\mathcal{T}_{10}$	0.72	0.28	2.57	1.00
Absolute Change	+0.20	-0.18	+1.44	+0.02
Relative Change	+38.5%	-39.1%	+127.4%	–

These results validate the theoretical mechanism proposed in Section 3.3: by enforcing high relevance within concept regions through $\mathcal{L}_{\text{concept}}$, the model satisfies this constraint by reallocating relevance mass away from background regions. The background stabilization term $\mathcal{L}_{\text{stab}}$ prevents drift to new spurious artifacts while allowing this magnitude reduction. Together, the dual regularization creates a *crowding-out effect* where semantically grounded concepts compete for and capture an increasing share of the relevance budget, naturally suppressing background correlations without explicit penalization.

This analysis provides the first direct quantitative evidence for the crowding-out effect in concept-based continual learning. Unlike previous work that measures only final explanation quality [10, 13], tracking relevance flow throughout training reveals the mechanism by which concept regularization mitigates catastrophic forgetting. The reallocation of relevance from backgrounds to concepts demonstrates that semantic anchoring actively reshapes the model's decision-making process, not merely preserving existing attention patterns.

Importantly, this redistribution occurs without explicit background suppression. The crowding-out effect emerges naturally from the interaction between concept overlap enforcement ($\mathcal{L}_{\text{concept}}$), background stability ($\mathcal{L}_{\text{stab}}$), and approximate relevance conservation. This validates our design choice in favor of stability over suppression (Appendix C.1): while background relevance *magnitude* decreases due to crowding-out, its *spatial pattern* remains anchored to pretrained priors, preserving useful contextual information while preventing spurious correlations.

E Extended Ablation Studies

E.1 Impact of Replay in Semantic-Centered Training

To disentangle the respective contributions of experience replay and semantic regularization, we evaluate C^{3}L under four distinct configurations:

1. **C^{3}L**: our full method.
2. **C^{3}L-NoReplay**: a replay-free variant that performs class-incremental training using only semantic regularization on the current task data. In this setting, no replay buffer is maintained and training is conducted solely on samples from the newly introduced task, with saliency maps generated on-the-fly during optimization. Concept discovery is performed once prior to training each task, while concept grounding is applied at the sample level during training. The objective is to explicitly align the model's explanations with concept-specific regions, using the following regularization terms in addition to the task loss $\mathcal{L}_{\text{task}}$:

$$\mathcal{L}_{\text{concept}} = \text{BCE}(S_x \odot M^{(x)}_{\text{concept}}, M^{(x)}_{\text{concept}}) \tag{8}$$

$$\mathcal{L}_{\text{stab}} = \left\| M^{(x)}_{\text{bg}} \odot S_x \right\|_2^2 \tag{9}$$

3. **C^{3}L-FullReplay**: a variant that applies experience replay to both semantic regularization terms. This configuration follows class-incremental training with concept regularization

Table 8: Ablation study disentangling the contribution of experience replay versus semantic regularization on ImageNet-R and CUB (ViT-B).

	ImageNet-R			CUB		
Method	ACC ↑	F ↓	AIA ↑	ACC ↑	F ↓	AIA ↑
ER	72.27±0.82	9.13±0.54	71.56±0.71	70.92±0.94	8.48±0.62	71.40±0.88
C^3L-NoReplay	79.82±0.65	3.84±0.42	78.99±0.58	80.94±0.72	3.27±0.38	81.78±0.66
C^3L-FullReplay	80.32±0.58	2.96±0.39	80.04±0.52	80.13±0.64	3.42±0.51	81.52±0.59
C^3L	**82.74±0.76**	**1.89±0.33**	**83.92±0.45**	**83.65±0.28**	**2.92±0.24**	**84.64±0.37**

on both current and replayed samples, using the same replay mechanism as C^3L and augmenting $\mathcal{L}_{\text{task}}$ with the following losses:

$$\mathcal{L}_{\text{concept}} = \frac{1}{|\mathcal{M}|} \sum_{(x,y)\in\mathcal{M}} \text{BCE}(M^{(x)}_{\text{concept}} \odot S_x, M^{(x)}_{\text{concept}} \odot S^{\text{ref}}_x) \tag{10}$$

$$\mathcal{L}_{\text{stab}} = \frac{1}{|\mathcal{M}|} \sum_{(x,y)\in\mathcal{M}} \left\| M^{(x)}_{\text{bg}} \odot \left(S_x - S^{\text{ref}}_x\right) \right\|_2^2 \tag{11}$$

4. **ER**: standard experience replay without any semantic or explanation-based regularization.

Table 8 reports results on ImageNet-R and CUB using a ViT-B backbone. The full C^3L consistently achieves the best overall performance, indicating that combining positive supervision on concept regions with stabilizing constraints on non-concept areas yields the most effective trade-off between plasticity and stability. Essentially, C^3L-FullReplay is very similar to RRR-ViT-AttnLRP, as it preserves explanation consistency while applying a slightly different distance function.

Notably, C^3L-NoReplay outperforms standard ER despite not storing any past samples. Moreover, it surpasses C^3L-FullReplay on the CUB dataset and achieves comparable performance on ImageNet-R. These findings suggest that semantic regularization alone can serve as an alternative to experience replay, particularly in scenarios where storing training data is infeasible due to privacy, memory, or deployment constraints. Intuitively, anchoring learning to semantically meaningful concepts identified during pretraining encourages the model to preserve its reasoning patterns across tasks, even in the absence of explicit replayed examples.

E.2 Explanation method

Table 9: Performance comparison using different explanation methods (AttnLRP, Grad-CAM, IIA, T-Attr) within the C^3L framework (ViT-B, ImageNet-R).

Explanation Method	**ACC ↑**	**F ↓**	**AIA ↑**
C^3L-AttnLRP (Ours)	82.74 ± 0.76	1.89 ± 0.33	83.92 ± 0.45
C^3L-GradCAM	80.92 ± 0.84	2.43 ± 0.41	82.15 ± 0.52
C^3L-IIA	81.85 ± 0.77	2.08 ± 0.35	83.12 ± 0.46
C^3L-T-Attr	81.47 ± 0.79	2.21 ± 0.37	82.68 ± 0.48

We compare AttnLRP against three alternatives: Grad-CAM [39], Integrated Iterated Attribution (IIA) [46], and Transformer Attribution (T-Attr) [41]. To conduct this analysis, we implemented three additional C^3L variants, each replacing the LRP-based explanation creation with the corresponding baseline method. Table 9 presents the results (method names reflect their adjusted C^3L variants). The findings show a clear advantage for LRP in this process, likely because LRP propagates relevance through multiple layers, offering a more comprehensive view of the model's internal representations. However, the relatively small gap indicates that C^3L is somewhat robust to explanation method choice, as long as the method produces reasonably faithful attributions.

E.3 Impact of Task Granularity

To evaluate the robustness of C^3L under varying task granularities, we conduct an ablation study on ImageNet-R by varying the number of splits. Following the protocol in [23], we divide the 200 classes of ImageNet-R into $T \in \{5, 10, 20\}$ tasks, corresponding to 40, 20, and 10 classes per task, respectively. We compare our method against the top-performing replay-augmented baselines from Table 1 using the ViT-B backbone.

As presented in Table 10, performance naturally degrades for all methods as the number of tasks increases, due to the higher frequency of model updates and the resulting accumulation of catastrophic forgetting steps. However, C^3L consistently maintains the highest accuracy performance (ACC and AIA) and lowest forgetting (F) across all settings. Notably, in the challenging 20-task setting, where forgetting is typically most severe, C^{3}L outperforms the nearest competitor (PLAN-R) by a significant margin. This demonstrates that anchoring representations to stable semantic concepts provides a more durable protection against forgetting than architectural or prompt-based regularization alone, particularly in long-sequence scenarios.

Table 10: Robustness analysis under varying task granularities (5, 10, and 20 split tasks) on ImageNet-R using the ViT-B backbone.

Method	5 Tasks			10 Tasks			20 Tasks		
	ACC ↑	F ↓	AIA ↑	ACC ↑	F ↓	AIA ↑	ACC ↑	F ↓	AIA ↑
L2P-R	80.52	3.15	82.13	76.82	3.91	77.34	74.12	7.85	75.20
CODA-R	81.45	2.68	82.90	78.54	2.82	79.52	75.30	6.12	76.85
MOS-R	82.37	1.95	83.86	80.92	2.43	81.55	76.45	4.88	78.10
PLAN-R	82.18	2.09	83.49	80.13	2.58	80.89	76.90	4.50	78.55
C^{3}L	**84.63**	**1.10**	**86.15**	**82.74**	**1.89**	**83.92**	**81.43**	**2.21**	**82.38**

E.4 Concept-Level vs. Object-Level Segmentation

Table 11 presents results for ViT-B on the ImageNet-R and CUB datasets. To assess whether fine-grained concept supervision provides advantages over coarse object-level segmentation, we replace the discovered concept masks with object segmentation masks. Specifically, we prompt SAM 3 using class names (e.g., "siberian husky") instead of concept lists (e.g., "blue eyes, fluffy ears, pointed snout"). This substitution leads to a consistent degradation in performance, indicating that concept-level supervision captures sub-object parts and attributes (e.g., "curved beak" or "striped wing") that offer more precise and informative guidance than binary object masks.

While object segmentation conveys category-level information (i.e., identifying *what* object is present), it does not specify *which* visual features are diagnostically relevant for classification. In contrast, concept-level masks emphasize discriminative attributes, resulting in denser and more localized supervision that more effectively constrains model attention. Additionally, object-level segmentations often encompass redundant or weakly informative regions, for example, assigning high importance to the entire body of a "Brown Bear", whereas concept-level supervision can focus on a minimal yet sufficient subset of salient parts.

This experiment isolates the contribution of concept-level guidance. Although "C^{3}L (Object segments)" relies on supervision from the same foundation model as C^{3}L (SAM 3), its performance is substantially inferior, indicating that the gains stem from semantic concept guidance rather than from the foundation model itself.

Table 11: Performance comparison between using fine-grained concept masks versus coarse object-level segmentation masks (ViT-B).

	ImageNet-R			CUB		
Mask Type	ACC ↑	F ↓	AIA ↑	ACC ↑	F ↓	AIA ↑
C^{3}L	82.74 ± 0.76	1.89 ± 0.33	83.92 ± 0.45	83.65 ± 0.28	2.92 ± 0.24	84.64 ± 0.37
C^{3}L (Object segments)	79.93 ± 0.91	2.84 ± 0.41	81.06 ± 0.58	78.47 ± 1.24	4.01 ± 0.66	81.64 ± 0.38

E.5 SAM 3 vs. Grounded SAM

Table 12: Impact of the spatial grounding foundation model: comparison between SAM 3 and Grounded SAM on ImageNet-R (ViT-B).

Grounding Method	ACC ↑	F ↓	AIA ↑
C^3L + Grounded SAM	81.98 ± 0.82	2.12 ± 0.38	83.25 ± 0.73
C^3L	$\mathbf{82.74 \pm 0.76}$	$\mathbf{1.89 \pm 0.33}$	$\mathbf{83.92 \pm 0.45}$

Prompt Complexity Robustness Analysis. Prior work suggests that GroundedSAM underperforms on complex or verbose prompts [74, 75]. To further justify the use of SAM 3 over GroundedSAM (Table 12), we conduct a targeted evaluation of robustness to prompt complexity. We randomly sample 50 classes from the CUB dataset and filter their automatically discovered concept sets to include concepts more likely to be verbose, fine-grained descriptions (i.e., concepts strictly longer than two words), resulting in a set of 286 textual concepts. We evaluate these concepts across 500 images (10 randomly selected images per class), prompting both SAM 3 and GroundedSAM with the exact same text prompts. We measure the detection hit-rate, defined as the percentage of prompts yielding a segmentation mask with a confidence score ≥ 0.3, alongside the average confidence score across all predicted masks. SAM 3 maintains a high hit-rate of 86% with an average confidence of 0.67, whereas GroundedSAM's hit-rate drops significantly to 59% with an average confidence of 0.45. These findings suggest that GroundedSAM's ability to localize fine-grained, verbose semantic attributes is inferior to that of SAM 3. Notably, this establishes the C^3L + GroundedSAM vs. C^3L ablation (Table 12) not merely as a foundation model comparison, but as a noise robustness evaluation for our framework. Despite the degradation in mask quality and the omission rate introduced by GroundedSAM, C^3L retains competitive continual learning performance, underscoring the resilience of our dual-regularization objective to noisy or incomplete semantic grounding.

Results. As we can see on Table 12 replacing SAM 3 with GroundedSAM [76] for concept localization slightly decreases performance. SAM 3's advantage lies in its ability to handle multiple concepts in a single batched query (one noun-phrase prompt per concept), returning confidence-scored masks that enable more accurate aggregation. GroundedSAM processes concepts sequentially, requiring multiple forward passes. Nevertheless, the modest performance gap indicates that C^3L is relatively robust to grounding method choice - as long as concepts are reasonably well-localized, the framework provides substantial benefits. Finally, GroundedSAM is more suitable for object segmentation rather than concept segmentation [16].

E.6 Concept Filtering: Required vs. Optional

We evaluate whether the concept filtering step (§3.2) is necessary by comparing performance with and without filtering.

Table 13: Ablation study on the necessity of the concept filtering and validation step across ImageNet-R and CUB.

	ImageNet-R			CUB		
Configuration	ACC ↑	F ↓	AIA ↑	ACC ↑	F ↓	AIA ↑
C^3L (no validation)	81.93 ± 0.84	2.17 ± 0.38	83.15 ± 0.52	82.47 ± 0.89	3.24 ± 0.41	83.52 ± 0.76
C^3L (with validation)	$\mathbf{82.74 \pm 0.76}$	$\mathbf{1.89 \pm 0.33}$	$\mathbf{83.92 \pm 0.45}$	$\mathbf{83.65 \pm 0.28}$	$\mathbf{2.92 \pm 0.24}$	$\mathbf{84.64 \pm 0.37}$
MOS-R (Best baseline)	80.92±0.42	2.43±0.28	81.55±0.31	78.82±0.65	3.84±0.29	82.52±0.53

Filtering provides modest improvements, but even unvalidated concepts substantially outperform the best baseline. This demonstrates robustness: GPT-4o-mini generates predominantly valid concepts, and the framework tolerates some noisy concepts through the stability term's robustness.

E.7 SAM 3 Confidence Threshold

We evaluate sensitivity to the SAM 3 confidence threshold $\tau_{\text{conf}} \in \{0.2, 0.3, 0.4, 0.5\}$. Table 14 present these results.

Table 14: Sensitivity analysis of the SAM 3 confidence threshold (τ_{conf}) for concept grounding on ImageNet-R (ViT-B).

τ_{conf}	ACC ↑	F ↓	AIA ↑
0.2	82.15 ± 0.85	2.08 ± 0.61	83.24 ± 0.62
0.3 (our choice)	$\mathbf{82.74 \pm 0.76}$	$\mathbf{1.89 \pm 0.33}$	$\mathbf{83.92 \pm 0.45}$
0.4	82.43 ± 0.79	1.95 ± 0.46	83.61 ± 0.48
0.5	81.87 ± 0.83	2.14 ± 0.79	82.98 ± 0.95

We can observe that $\tau_{\text{conf}} = 0.3$ yields the best overall performance. Lower thresholds introduce noisy false-positive regions, whereas higher thresholds exclude valid concept instances. Notably, performance varies smoothly across a wide range of threshold values, indicating that the method is not overly sensitive to the specific choice of τ_{conf}.

E.8 Hyperparameter Sensitivity Analysis

Table 15 shows performance across different values of λ_{concept} and λ_{stab} on ImageNet-R (ViT-B).

Table 15: Sensitivity analysis for regularization hyperparameters $\lambda_{concept}$ and λ_{stab} on ImageNet-R (ViT-B).

λ_{concept}	λ_{stab}	ACC ↑	F ↓	AIA ↑
0.5	0.6	80.83 ± 0.89	2.34 ± 0.41	81.34 ± 0.64
0.8	**1.2**	82.74 ± 0.76	1.89 ± 0.33	83.92 ± 0.45
1.0	1.5	82.51 ± 0.81	1.95 ± 0.36	82.42 ± 0.81
1.5	2.0	81.92 ± 0.87	2.11 ± 0.38	81.85 ± 0.75
2.0	2.5	81.24 ± 0.93	2.28 ± 0.42	81.37 ± 0.53

Performance only slightly varies across the tested range, indicating reasonable robustness to hyperparameter choices. We select $\lambda_{\text{concept}} = 0.8$, $\lambda_{\text{stab}} = 1.2$ based on validation set performance.

F Additional Experiments

F.1 Performance on Non-Pretrained Networks

To assess whether C^{3}L's benefits extend beyond pretrained networks, we evaluate on ResNet-18 trained from random initialization on CUB.

Table 16: Continual learning performance on CUB using a non-pretrained ResNet-18 backbone.

Method	ACC ↑	F ↓	AIA ↑
ER	42.17 ± 1.35	18.94 ± 0.92	45.83 ± 1.24
DER++	48.23 ± 1.18	14.62 ± 0.78	51.47 ± 1.09
EPR	51.02 ± 0.98	12.84 ± 0.26	53.57 ± 0.75
RRR	49.84 ± 1.24	13.95 ± 0.82	52.93 ± 1.15
C^{3}L	$\mathbf{59.21 \pm 1.08}$	$\mathbf{10.74 \pm 0.71}$	$\mathbf{60.18 \pm 0.97}$
Joint Training	68.45 ± 0.89	–	–

C^{3}L outperforms baselines by large margin also without pretraining, though absolute performance is lower than with ImageNet pretraining.

F.2 Full Continual Learning Results

Table 17 presents full results for Table 1 from the main paper. C^{3}L's with $\mathcal{L}_{\text{concept}}$ applied to all training samples, is denoted as C^{3}L++. Unlike C^{3}L, however, C^{3}L++ does not follow a continual

Table 17: Quantitative comparison of Continual Learning performance across five benchmarks using ResNet-18 and ViT-B architectures. We report Average Accuracy (ACC), Average Forgetting (F), and Average Incremental Accuracy (AIA). We report results over 5 trials.

Method	CUB			DomainNet			ImageNet-R			ImageNet-A			ObjectNet		
	ACC ↑	F ↓	AIA ↑	ACC ↑	F ↓	AIA ↑	ACC ↑	F ↓	AIA ↑	ACC ↑	F ↓	AIA ↑	ACC ↑	F ↓	AIA ↑
ResNet-18															
Joint Training	83.04±0.51	-	-	74.99±0.22	-	-	76.31±0.90	-	-	71.22±0.42	-	-	73.42±0.91	-	-
Regularization-based															
EWC	55.15±0.95	22.34±0.72	60.82±1.05	54.40±0.88	14.25±0.64	54.81±0.92	55.24±0.77	18.42±0.58	58.36±0.83	48.04±0.61	28.52±1.15	52.75±0.55	48.42±0.68	26.24±0.79	52.04±0.73
LwF	57.62±0.83	20.15±0.65	62.34±0.92	56.27±0.75	12.82±0.52	56.53±0.84	57.81±0.69	16.85±0.48	60.92±0.76	48.83±0.54	26.83±0.98	54.11±0.49	49.94±0.62	24.12±0.71	53.73±0.66
PEC	62.23±0.76	15.94±0.62	66.71±0.85	60.34±0.68	9.82±0.45	60.42±0.72	62.12±0.58	12.93±0.52	64.84±0.65	52.74±0.48	22.14±0.82	58.01±0.44	55.02±0.55	19.82±0.64	59.20±0.59
Replay-based															
ER	64.34±0.92	14.25±0.58	68.42±0.84	61.92±0.81	8.92±0.55	61.99±0.76	63.38±0.68	11.85±0.49	66.15±0.72	55.01±0.45	19.43±0.66	59.80±0.38	57.84±0.52	17.24±0.58	62.01±0.55
iCaRL	66.82±0.88	12.74±0.54	70.65±0.79	64.00±0.75	7.85±0.48	63.92±0.69	65.72±0.64	10.43±0.45	68.25±0.65	57.22±0.41	17.15±0.59	62.13±0.35	60.11±0.48	15.42±0.54	64.32±0.51
DER++	71.45±0.72	9.62±0.45	75.14±0.65	68.02±0.58	5.74±0.32	67.70±0.55	69.81±0.48	7.82±0.35	72.15±0.52	61.43±0.38	12.53±0.44	66.41±0.32	64.44±0.42	11.34±0.48	68.80±0.45
ER-ACE	69.73±0.78	10.85±0.49	73.32±0.72	66.44±0.65	6.42±0.38	66.22±0.62	68.21±0.55	8.93±0.42	70.75±0.58	59.83±0.42	14.82±0.52	64.71±0.36	62.74±0.46	13.14±0.52	67.11±0.48
CSReL-CL	71.92±0.82	9.25±0.52	75.54±0.76	67.44±0.69	6.12±0.42	67.23±0.65	69.42±0.59	8.24±0.46	71.83±0.62	60.74±0.45	13.42±0.56	65.51±0.39	63.83±0.49	12.05±0.56	67.94±0.53
Explanation-aware															
RRR	72.21±0.78	9.10±0.44	75.81±0.73	68.25±0.33	5.66±1.06	67.95±0.54	70.13±0.65	7.66±0.62	72.43±0.50	61.73±0.43	12.06±0.21	66.68±0.31	64.82±0.32	10.87±0.35	69.16±0.64
RRR-LRP	73.25±0.65	8.42±0.38	76.74±0.58	69.30±0.52	5.13±1.28	68.92±0.49	71.15±0.45	6.92±0.92	73.32±0.46	63.01±0.35	10.24±0.69	68.12±0.28	66.20±0.36	9.23±0.42	70.74±0.41
EPR	72.86±0.72	8.75±0.41	76.32±0.65	68.82±0.42	5.36±0.82	68.41±0.51	70.68±0.55	7.21±0.84	72.88±0.48	62.45±0.39	11.15±0.52	67.42±0.30	65.54±1.04	10.02±0.38	70.05±0.52
TASS	70.24±0.85	10.35±0.56	74.12±0.78	66.91±0.72	6.74±0.45	66.73±0.68	68.92±0.62	8.63±0.49	71.42±0.65	60.13±0.48	14.25±0.62	65.01±0.42	63.01±0.52	12.83±0.58	67.44±0.55
C^3L	**79.73±0.31**	**5.32±0.63**	**82.24±0.35**	**73.01±0.23**	**3.14±0.55**	**72.51±0.88**	**74.35±0.46**	**4.23±0.29**	**75.84±0.65**	**67.84±0.72**	**6.52±0.46**	**70.83±0.41**	**71.12±0.53**	**5.82±0.25**	**72.23±0.92**
C^3L++	83.91±0.28	4.81±0.52	83.18±0.31	74.72±0.19	2.84±0.45	74.83±0.72	75.26±0.38	3.82±0.24	75.48±0.58	69.82±0.61	5.89±0.39	69.47±0.35	72.11±0.44	5.26±0.21	72.55±0.81
ViT-B															
Joint Training	86.12±0.23	-	-	81.31±0.29	-	-	84.59±0.31	-	-	75.31±0.19	-	-	78.96±0.24	-	-
Prompt-based															
L2P	70.82±1.24	7.94±0.65	74.24±0.91	71.85±0.78	4.23±0.42	72.40±0.83	71.49±0.69	4.85±0.56	71.91±0.74	54.94±0.82	6.83±0.44	56.11±0.57	64.53±0.70	6.24±0.55	66.12±0.69
DualPrompt	72.43±0.81	6.82±0.54	75.94±0.69	73.22±0.75	3.74±0.39	73.83±0.71	72.12±0.67	4.33±0.43	72.46±0.86	54.05±0.70	7.24±0.42	55.04±0.45	66.12±0.58	5.52±0.44	66.35±0.57
CODA-Prompt	74.74±0.79	5.42±0.42	78.13±0.67	74.95±0.82	3.12±0.38	75.50±0.78	72.35±0.65	3.82±0.41	72.69±0.64	54.24±0.68	6.92±0.40	55.49±0.53	67.82±0.66	4.83±0.42	68.24±0.55
Adapter-based															
MOS	75.94±0.68	4.82±0.41	79.53±0.55	76.14±0.70	2.12±0.39	76.81±0.66	76.88±0.24	4.15±0.61	78.35±0.53	61.34±0.68	3.52±0.29	62.26±0.35	71.53±0.67	3.54±0.31	71.82±0.56
MIN	75.45±0.83	5.23±0.54	79.12±0.69	75.75±0.85	2.74±0.41	76.30±0.81	76.15±0.68	4.31±0.36	78.54±0.66	61.92±0.81	3.14±0.28	62.42±0.57	72.15±0.55	3.62±0.42	72.61±0.69
PLAN	76.12±0.70	4.83±0.42	79.74±0.56	76.03±0.83	2.24±0.28	76.76±0.67	76.62±0.55	4.02±0.31	78.93±0.84	60.25±0.69	4.52±0.32	61.14±0.55	70.82±0.68	3.42±0.30	71.23±1.03
Replay-based															
RRR-ViT	76.94±1.17	4.82±0.61	78.53±1.14	78.32±1.21	2.01±0.49	78.19±1.16	78.83±0.94	4.12±0.50	80.84±0.93	60.52±1.18	5.43±0.23	63.24±0.64	70.19±1.17	3.82±0.51	70.94±1.16
RRR-ViT-AttnLRP	78.09±0.85	4.17±0.58	80.34±0.88	78.47±1.08	1.92±0.37	78.56±0.90	79.48±0.76	3.35±1.01	80.87±0.66	63.02±0.56	4.24±0.26	63.51±0.59	71.61±0.87	3.33±0.42	71.88±0.89
L2P-R	69.62±1.25	6.34±0.56	77.25±0.92	74.23±0.29	3.42±0.53	74.82±1.04	76.82±0.90	3.91±0.57	77.34±0.89	58.33±0.23	6.14±0.21	59.71±0.68	67.24±1.21	5.32±0.56	68.15±0.61
CODA-R	77.53±0.99	4.23±0.42	80.24±0.77	78.02±0.82	3.02±0.28	78.21±0.78	78.54±0.66	2.82±0.31	79.52±0.64	61.24±0.89	4.82±0.42	61.43±0.55	71.53±0.78	3.24±0.30	71.62±0.67
MOS-R	78.82±0.65	3.84±0.29	82.52±0.53	79.03±0.78	1.52±0.42	79.42±0.54	80.92±0.42	2.43±0.28	81.55±0.31	67.84±0.96	2.24±0.27	66.25±0.42	73.62±0.35	2.52±0.29	73.89±0.93
PLAN-R	79.24±0.56	3.52±0.28	82.14±0.54	78.62±0.59	1.83±0.27	78.92±0.65	80.13±0.51	2.58±1.29	80.89±0.42	65.48±0.57	3.05±0.31	65.83±0.51	72.85±0.56	2.84±0.28	73.06±0.55
C^3L	**83.65±0.28**	**2.92±0.24**	**84.64±0.37**	**80.43±0.55**	**1.23±0.62**	**80.82±0.26**	**82.74±0.76**	**1.89±0.33**	**83.92±0.45**	**71.14±0.59**	**1.53±0.41**	**70.03±0.38**	**76.83±0.56**	**1.92±0.83**	**77.92±0.78**
C^3L++	87.21±0.22	2.62±0.20	86.79±0.31	81.02±0.48	1.10±0.55	80.89±0.21	83.15±0.67	1.72±0.29	84.21±0.38	73.58±0.51	1.37±0.36	73.81±0.32	77.12±0.49	1.72±0.74	78.48±0.69

learning replay setting and is therefore not considered our primary method. Since Joint Training uses no concept supervision, C^3L++ can exceed it (e.g., on CUB).

G Additional Experimental Details

G.1 Explanation Quality

Motivation and Protocol Overview. To assess whether C^3L grounds attributions in semantically meaningful features rather than spurious artifacts, we adopt two complementary evaluation protocols. (1) **Spurious Correlation Mitigation:** We measure the model's ability to ignore spurious biases using final accuracy on FITZ [66] and Split ImageNet-100C [65], following standard evaluation practices in prior debiasing work [66, 65]. (2) **Semantic Localization:** We evaluate whether explanations remain focused on semantic object regions and avoid explanation drift by computing segmentation-based metrics on saliency maps, following Chefer et al. [41], on the CUB dataset, which provides ground-truth segmentation annotations[4].

Datasets. Split ImageNet-100C (I-100C) [65] is derived from Split ImageNet-100, which consists of 10 sequential tasks, each containing 10 disjoint classes randomly sampled from ImageNet-1K. To introduce controlled spurious correlations, we inject Gaussian noise disproportionately: $(\alpha \times 100)\%$ of the samples from one randomly selected class are corrupted, while only 1% of samples from the remaining classes are affected, for a skew ratio $\alpha \geq 0.01$. We consider seven bias levels by varying the skew ratio.

Fitzpatrick17k (FITZ) [77] is a large-scale clinical dermatology dataset comprising 16,577 images across 114 skin conditions, each annotated with Fitzpatrick Skin Type (FST) on a six-point scale. Due to its pronounced skin-tone imbalance-approximately 47% of images correspond to FST I–II and only about 13% to FST V–VI-the dataset is widely used to evaluate robustness and bias mitigation under real-world demographic skew.

Training Configuration. We employ the ResNet-18 and ViT-B backbones pre-trained on ImageNet-1K and ImageNet-21k respectively. For spurious correlation mitigation, we finetune ResNet-18 on both FITZ and I-100C using the exact same setup reported in the original codebase of every baseline. Specifically, for BGS, we used its best performing configuration on the accuracy metric reported in the paper, which is DER+BGS. For semantic localization, both models are fine-tuned on the CUB using the same protocol from the main paper.

[4] `https://data.caltech.edu/records/w9d68-gec53`

Semantic Localization Evaluation. Here, our goal is to asses the quality of semantic localization and explanation map consistency. We evaluate explanation quality using the CUB dataset, which provides ground-truth object segmentation masks. Following Chefer et al. [41], we threshold saliency maps at the 50th percentile to produce binary masks and compute: (1) **mIoU**: Mean Intersection-over-Union between explanation masks and ground-truth objects, (2) **pixAcc**: Pixel-wise accuracy (correctly classified pixels / total pixels), and (3) **mAP**: Mean Average Precision across IoU thresholds $[0.5, 0.55, \ldots, 0.95]$. Higher values indicate that explanations better localize semantic object regions rather than spurious backgrounds. Since C^3L is trained to align relevance with concept regions on the object, improved localization is partly expected by design; the spurious-correlation results on FITZ and I-100C provide complementary evidence that does not rely on segmentation.

G.2 Robustness Tests Further Details

OOD Evaluation Protocol. Our setup is inspired by [71], which studies feature transferability by training on a subset of ImageNet classes; here, we fine-tune on a class subset and evaluate on OOD benchmarks.

Dataset Construction. We curate a lightweight dataset by sampling 5 images from 500 randomly selected ImageNet-1K [72] classes resulting with 2,500 samples. This ensures broad semantic coverage while rigorously testing data efficiency under limited supervision.

Models and Benchmarks. We utilize ImageNet-pretrained ResNet-18 and ViT-B models. Testing is performed on ImageNet-A and ObjectNet (restricted to overlapping classes), which represent challenging OOD shifts that ImageNet-pretrained models mostly fail with. Performance is measured using classification accuracy (%) and mean Intersection over Union (mIoU). For mIoU, we compare the explanation maps, computed as in Tab. 2, against object segmentations generated by prompting SAM 3 with the ground-truth class labels.

Training. We execute the complete C^3L framework-comprising concept discovery, validation, and grounding - on the 2,500-sample subset. First, we perform automatic concept generation [15] for the 500 selected classes. We then validate these candidate concepts using 40 images per class, yielding a filtered concept set for each category. Subsequently, we ground these concepts for every sample in the training subset by prompting SAM 3 with the filtered concepts (one prompt per concept), thereby generating concept maps. Additionally, we employ LRP and AttnLRP to produce explanations for each image. Finally, the models are fine-tuned exclusively on this subset using the C^3L approach, leveraging the derived concept maps and explanations as supervision. For both RRR and RRR-ViT we used the same setup reported by the authors [10] for the fine-tuning phase.

Full results. We additionally evaluate EPR and TASS, including their ViT-adapted variants (EPR-ViT, TASS-ViT), to enable a comprehensive comparison with explanation-aware baselines. The ViT adaptations follow standard practice, using Grad-CAM for ViTs [41] and generating explanations from the `[CLS]` token, as commonly adopted in prior work.

The results mirror the trends observed in the main paper: C^3L consistently outperforms all explanation-aware methods in robustness to out-of-distribution data.

H Foundation Models Usage and Fairness Clarification

H.1 Performance without Foundation Models Usage

To disentangle the semantic guidance and foundation models factors, we remove foundation models from the training pipeline and instead conduct experiments on the **PartImageNet** dataset [78], which provides ground-truth part-level annotations. Although part annotations are not equivalent to semantic concepts, this setting offers a controlled evaluation that isolates the effect of semantic guidance. The results therefore illustrate the benefits of leveraging semantic features for continual learning independently of foundation models.

Table 18: Out-of-distribution robustness evaluation. We report accuracy (%) and mIoU on ImageNet-A and ObjectNet before and after fine-tuning with C^3L using a sparse subset. We compare EPR, TASS, RRR, and C^3L (for ViT-B, the ViT variants EPR-ViT, TASS-ViT, and RRR-ViT).

Model	Dataset	Metric	Before	After			
				EPR	TASS	RRR	**C^3L**
ResNet-18	ImageNet-A	Acc	8.59	8.49	6.34	7.92	**20.15**
		mIoU	21.32	21.78	20.83	20.39	**43.27**
	ObjectNet	Acc	21.88	23.08	21.62	21.56	**36.93**
		mIoU	35.41	35.27	34.68	34.16	**52.91**
ViT-B	ImageNet-A	Acc	13.48	15.34	12.91	13.18	**30.82**
		mIoU	26.38	26.53	25.82	25.43	**45.32**
	ObjectNet	Acc	33.92	34.26	32.43	34.49	**58.04**
		mIoU	38.12	37.65	37.19	37.02	**54.83**

While we aim to ensure a fair comparison with existing baselines, we note that equipping those methods with the same foundation-model guidance used in our approach would be conceptually similar to applying our method or to the experimental setting presented in Table 11.

Experimental Setup. We evaluate a ResNet-18 trained from scratch on a class-incremental split of PartImageNet. We used 150 classes (out of 158, removing 8 randomly sampled classes) partitioned into 10 tasks of 15 classes each.
Baselines: Standard ER [68], DER++ [5], RRR [10], EPR [13], and our original C^3L.
C^3L-PG (Part-Guided): Instead of using GPT-4o-mini or SAM 3, we used the dataset's ground-truth part masks as the semantic concepts for our dual regularization objective.

Results and Discussion. Table 19 demonstrates that even when trained from scratch without foundation model priors, concept-level guidance outperforms all baselines.

Table 19: Continual Learning on PartImageNet. Comparison of concept-guided learning against state-of-the-art baselines.

Method	ACC ↑	F ↓	AIA ↑
ER	44.21 ± 1.12	21.45 ± 0.88	46.92 ± 0.94
DER++	49.82 ± 0.75	16.32 ± 0.54	52.14 ± 0.61
RRR	51.34 ± 1.04	14.92 ± 0.91	53.88 ± 1.02
EPR	52.65 ± 0.82	13.51 ± 0.44	55.07 ± 0.73
C^3L	59.12 ± 0.64	10.36 ± 0.37	62.43 ± 0.92
C^3L-PG	58.83 ± 0.95	10.78 ± 0.62	62.11 ± 0.88

The performance gap between C^3L and pixel-based methods confirms that semantic guidance is a primary driver for stability. By grounding the model in persistent parts, we achieve superior retention and learning dynamics, independent of the foundation model used for discovery.

H.2 Semantic Depth Over Object Segmentation

We examined the impact of mask granularity in Appendix E (Table 11) to further emphasize the contribution of our semantic anchoring approach.

We implemented a baseline using the exact **same SAM 3 foundation model** but prompted it for object-level segmentation (masking the entire object) rather than our fine-grained concept-level segmentation (masking specific attributes like "beak" or "wing"). The results demonstrate that concept-level supervision yields superior performance compared to object-level masks. This suggests that semantic guidance is also a driver of our results, rather than merely the presence of foundation-model-derived segmentation masks.

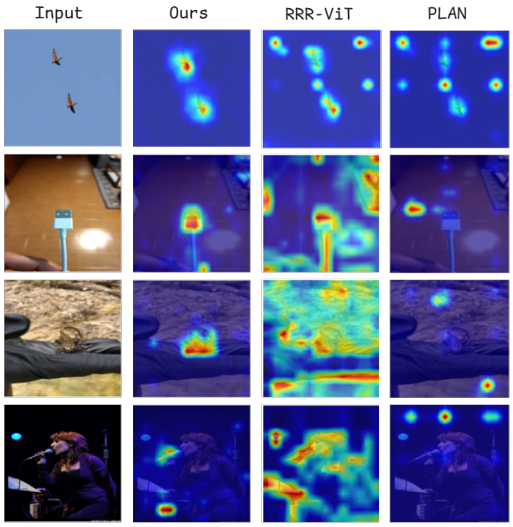


Figure 3: Qualitative comparison of explanation maps produced by C^{3}L and baseline methods before and after continual learning. Explanation maps produced using ViT-B w.r.t. the classes (top to bottom): ‘lorikeet’, ’usb cable’, ‘common newt, Triturus vulgaris’, and ‘accordion, piano accordion’.

H.3 Foundation Models as Automated Annotators

Our framework utilizes GPT-4o-mini and SAM 3 to automate the generation of semantic supervision. We leverage these foundation models in an offline preprocessing stage to decompose images into semantic regions. We view this process as automated annotation leveraging foundation model priors, distinct from knowledge distillation in that we do not transfer class predictions or latent representations. Importantly, this automated concept discovery and grounding stage is as an offline pre-processing stage.

In standard knowledge distillation, a student model is typically trained to mimic the soft targets (logits) or align with the latent feature space of a teacher model. In contrast, C^{3}L does not access the class probability distributions or internal embeddings of the foundation models. Instead, these models function as an automated framework to generate binary spatial masks ($M_{concept}$). These masks provide the "where" (spatial attention targets), but the model must autonomously learn the "what" (discriminative features) from the training data labels. This allows us to inject human-interpretable priors without leaking instance-specific predictive information.

I Qualitative Visualizations

Figure 3 presents qualitative examples (ImageNet pretrained data) of explanation maps generated by C^{3}L and baseline methods before and after continual learning, using a ViT-B model pretrained on ImageNet and trained on the ObjectNet dataset. C^{3}L preserves attention on semantically meaningful object regions (e.g., discriminative parts and attributes), whereas baseline methods exhibit explanation drift toward background regions and texture cues.

J Failure Cases

J.1 Concept Discovery Failures

While automated concept discovery achieves quite high retention rate, some generated concepts fail filtering:

Too Abstract. Concepts like “graceful posture", “aggressive demeanor", or "natural habitat" describe semantic properties but lack clear visual grounding. SAM 3 cannot reliably localize such concepts, leading to noisy masks.

Too Rare. Concepts appearing in $< 15\%$ of class instances (e.g., “open beak" for birds, “raised paw" for dogs) fail occurrence rate criteria. While visually grounded when present, their rarity makes them unsuitable for class-wide regularization.

Ambiguous. Concepts like “medium size" or “dark color" lack clear spatial boundaries. For instance, SAM 3 may produce inconsistent masks for "dark color" depending on lighting.

J.2 Segmentation Failures

SAM 3 rarely produces noisy or incorrect masks:

Small or Occluded Features. Fine-grained concepts covering $< 2\%$ of image area (e.g., “eye-ring" on small birds, “toe claws" on animals) are missed or segmented with low IoU. This primarily affects fine-grained datasets like CUB.

Concept Ambiguity. Polysemous concepts (e.g., “fluffy texture" can describe fur, clouds, or cotton) may segment unintended regions. While validation filters some cases, occasional false positives remain.

K Limitations and Future Work

Limitations. While C^3L advances continual learning through semantic grounding, several limitations warrant consideration. First, our approach relies on foundation models for automated concept generation. Although such models are increasingly used to guide learning in related domains [79–81] and have driven recent breakthroughs in many, they may also introduce failure modes. For instance, spatial grounding may fail if language models produce non-visual or ambiguous concepts. We mitigate this through occurrence rate and coverage validation, but more sophisticated concept quality metrics could improve robustness. Second, C^3L adds computational overhead for the concept and explanation map preprocessing procedure. Yet, this cost is modest relative to the performance gains. Third, our evaluation focuses on vision tasks. Extending concept-based regularization to other modalities (language, audio, multimodal) requires domain-specific concept discovery mechanisms. Fourth, the reference saliency used by $\mathcal{L}_{\text{stab}}$ is model-derived: while its magnitude is free to decay under crowding-out, the relative spatial pattern of background relevance at buffer insertion is preserved, so spurious background cues already present at that point are attenuated rather than fully removed.

Broader impact. Beyond technical contributions, our work has important implications for deploying continual learning systems in safety-critical domains. By maintaining interpretable reasoning patterns alongside predictive accuracy, C^3L enables practitioners to monitor whether models learn reliably or develop hidden vulnerabilities through spurious correlations. This transparency is crucial for applications like medical diagnosis, autonomous systems, and content moderation where understanding *how* decisions are made is as important as ensuring they are correct. Moreover, our automated framework demonstrates that semantic grounding need not require extensive human supervision, making interpretable continual learning more accessible to practitioners without domain expertise in explainable AI.

Future directions. This work opens several promising avenues for future research. First, extending concept-level regularization to domain-incremental and task-free (online) continual learning settings may better address scenarios with gradually shifting input distributions or unknown task boundaries. Second, exploring concept compositionality, namely, whether concepts learned for one class can be reused or recombined for related classes in subsequent tasks, could enable more efficient and scalable knowledge accumulation. Third, integrating concept-based regularization with parameter-efficient adaptation techniques for large vision–language models may combine the benefits of semantic grounding with efficient fine-tuning. Fourth, developing theoretical frameworks that characterize when and why concept-level regularization outperforms pixel-level approaches would provide principled guidance for the design of explanation-aware continual learning methods. Additionally, leveraging “around” prompts [15] to selectively preserve informative background regions, rather than suppressing all background content, may further reduce reliance on spurious correlations. Investigating whether concept-level regularization can support the *detection* of distribution shifts, by monitoring changes in concept activation patterns, could enable adaptive systems that identify novel contexts and trigger

human intervention when needed. Finally, we believe that extending this approach to knowledge distillation presents a promising direction; in scenarios where a teacher model relies heavily on spurious correlations, concept-guided regularization could actively intervene during distillation, allowing the student to extract valuable predictive knowledge while strictly suppressing the inheritance of the teacher's flawed reasoning patterns.

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 abstract and introduction clearly outline our proposed method C^3L and the theoretical and empirical claims we make regarding the mitigation of explanation drift in continual learning.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   Justification: We fully discuss limitations of the work and failure cases on Appendices K and J.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

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]

Justification: Our paper focuses on an empirical methodology and does not introduce formal mathematical theorems or proofs; however, we do formulate an approximate relevance conservation constraint in Section 3.3.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

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: Section B details all reproducibility information needed. We also provided the full code in the supplementary.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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: The code is fully readable and we provide a Readme.md file in it in which we explain how to use it.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

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: Detailed experimental settings, including datasets, baselines, hyperparameters, and training protocols, are provided in Section 4.1 and comprehensively expanded in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

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: We report the mean and standard deviation across 5 independent runs for our experiments and assess statistical significance using paired t-tests, as stated in Section 4.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

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: We detail the compute hardware (NVIDIA A100 GPUs, single-GPU training) in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

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: We have reviewed the NeurIPS Code of Ethics and confirm that our research and methodology conform to these guidelines.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We discuss our work's broader impact on Appendix K.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

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]

    Justification: Our work deals with continual learning models for computer vision, thus we believe it does not pose such risk.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

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: All existing datasets, foundation models, and baselines used in our work are properly credited and referenced in the main text and appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: Our code is released and is well documented.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

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: [Yes]

    Justification: We describe our use of an LLM (GPT-4o-mini) for automated concept discovery as a core component of our methodology in Section 3.2 and Appendix B.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.