Harnessing Textual Semantic Priors for Knowledge Transfer and Refinement in CLIP-Driven Continual Learning

Lingfeng He[1], **De Cheng**[1*], **Huaijie Wang**[1], **Nannan Wang**[1]

[1]Xidian University

Abstract

Continual learning (CL) aims to equip models with the ability to learn from a stream of tasks without forgetting previous knowledge. With the progress of vision-language models like Contrastive Language-Image Pre-training (CLIP), their promise for CL has attracted increasing attention due to their strong generalizability. However, the potential of rich textual semantic priors in CLIP in addressing the stability–plasticity dilemma remains underexplored. During backbone training, most approaches transfer past knowledge without considering semantic relevance, leading to interference from unrelated tasks that disrupt the balance between stability and plasticity. Besides, while text-based classifiers provide strong generalization, they suffer from limited plasticity due to the inherent modality gap in CLIP. Visual classifiers help bridge this gap, but their prototypes lack rich and precise semantics. To address these challenges, we propose Semantic-Enriched Continual Adaptation (SECA), a unified framework that harnesses the anti-forgetting and structured nature of textual priors to guide semantic-aware knowledge transfer in the backbone and reinforce the semantic structure of the visual classifier. Specifically, a Semantic-Guided Adaptive Knowledge Transfer (SG-AKT) module is proposed to assess new images' relevance to diverse historical visual knowledge via textual cues, and aggregate relevant knowledge in an instance-adaptive manner as distillation signals. Moreover, a Semantic-Enhanced Visual Prototype Refinement (SE-VPR) module is introduced to refine visual prototypes using inter-class semantic relations captured in class-wise textual embeddings. Extensive experiments on multiple benchmarks validate the effectiveness of our approach. Code is available in the supplementary materials.

Introduction

Continual learning (CL) aims to equip models with the ability to learn continuously from a stream of tasks (Lu et al. 2024), a key requirement for adapting to evolving open-world scenarios. A fundamental challenge is *the stability-plasticity dilemma* (Kim and Han 2023; Chen et al. 2023), which requires models to preserve previously learned knowledge (stability) while adapting to new information (plasticity). We investigate the challenging rehearsal-free Class-Incremental Learning (CIL) setting, where the model must continually learn new classes without rehearsal and predict over all seen categories without task identities.

*Corresponding author

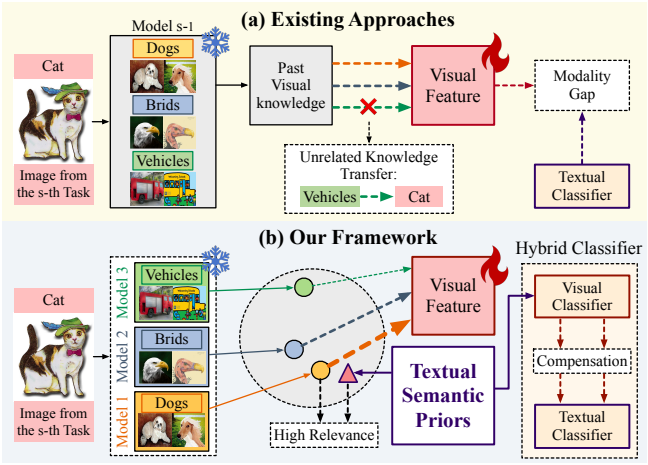

Figure 1: (a) Existing methods exhibit limited stability-plasticity trade-offs due to the unrelated knowledge transfer in backbone and the modality gap in classifier. (b) Our SECA leverage textual priors to (1) prioritize transferring relevant knowledge (Dog classes) when learning new classes (Cat), and (2) inject semantic relations into the visual classifier to bridge the modality gap, improving this trade-off.

Recent advances in vision-language models, such as Contrastive Language-Image Pre-training (CLIP) (Radford et al. 2021), provide strong zero-shot capabilities and serve as promising backbones for CL. While several methods (Zhou et al. 2025; Huang et al. 2024; Jha, Gong, and Yao 2024) leverage CLIP mainly as a powerful backbone, the potential of its textual semantic priors for improving the stability–plasticity trade-off remains underexplored. To retain old knowledge in backbones, most approaches (Kirkpatrick et al. 2017; Kang, Park, and Han 2022; Asadi et al. 2023) rely on regularization or distillation to enforce consistency with model after the most recent task. However, such non-selective knowledge retention can lead to semantic interference, where unrelated or noisy knowledge from previous tasks hampers adaptation to new tasks and distorts earlier feature spaces. As shown in Fig. 1(a), when learning a new class like 'cat', prior knowledge from similar classes such as 'dog' or 'bird' can be helpful, while unrelated classes like 'vehicle' introduce interference that degrades learning. As for classifiers, although the text-only designs (Huang et al. 2024; Jha, Gong, and Yao

2024) offer strong generalization, they limit plasticity due to the inherent modality gap in CLIP. Prior studies (Liang et al. 2022; Fahim, Murphy, and Fyshe 2024) attributes this gap to a geometric separation in the shared embedding space, where visual and textual embeddings remain distinctly apart in the feature space. This modality gap is further exacerbated in CL, as task-isolated training prevents joint calibration, reducing the text-based classifier's ability to adapt to new classes.

To mitigate semantic interference in the backbone, we consider it essential to prioritize the transfer of semantically relevant historical knowledge. Since the text branch provides consistent semantics throughout continual adaptation, and CLIP inherently offers an aligned visual-textual space, we assume that *textual cues offer reliable guidance for determining which past visual knowledge to transfer and which to suppress*, as shown in Fig.1(b). Moreover, to overcome the modality gap during classification, a straightforward solution lies in hybrid classifiers that support both cross-modality and intra-modality matching. While visual classifiers based on raw prototypes often suffer from semantic inaccuracies due to limited data and class imbalance in real-world CL, we propose to inject the stable and structured semantics from the textual branch into the visual side, enabling a semantic-enriched visual classifier to bridge the modality gap.

Building on the above analysis, we propose Semantic-Enriched Continual Adaptation (SECA), a unified framework that harnesses the anti-forgetting and structured nature of textual semantic priors in CLIP. Specifically, we propose a Semantic-Guided Adaptive Knowledge Transfer (SG-AKT) module for selective transfer in the backbone. It exploit the textual embeddings of new images as semantic vectors to assess their relevance to a pool of adapters that encapsulate historical visual knowledge. Guided by the relevance scores, past representations from the adapter pool are aggregated in an instance-adaptive manner and serve as teacher signals for selective distillation. Besides, a Semantic-Enhanced Visual Prototype Refinement (SE-VPR) module is introduced to enable a powerful visual-side classifier. It models the inter-class semantic relationships encoded in class-wise textual embeddings and uses them to refine the structure of coarse CLIP visual prototypes, aligning them with the relational topology of the textual space. Combined with the text classifier, it forms a hybrid classification paradigm that compensates for the modality gap and enhances plasticity.

Our main contributions are summarized as follows:

- We propose Semantic-Guided Adaptive Knowledge Transfer (SG-AKT), a novel instance-adaptive distillation approach that uses textual semantics to guide the selective transfer of historical knowledge, mitigating knowledge interference in the visual backbone;
- We propose Semantic-Enhanced Visual Prototype Refinement (SE-VPR), which injects inter-class textual semantics into visual prototypes to build a powerful visual-side classifier, effectively compensating for the modality gap;
- Extensive experiments demonstrate that our SECL outperforms existing methods and provides new insights into harnessing the potential of textual priors for CL.

Related Work

Continual Learning. Continual Learning (CL) aims to incrementally acquire new knowledge while preserving performance on previously learned tasks. Traditional CL methods typically learn models from scratch and can be categorized into three types: (a) replay-based methods (Bonicelli et al. 2022; Aljundi et al. 2019a,b), which store exemplar samples to jointly train with new tasks; (b) regularization-based methods (Jung et al. 2020; Zenke, Poole, and Ganguli 2017; Aljundi et al. 2018; Wang et al. 2025), which constrain parameter updates through additional regularization terms; and (c) expansion-based methods (Hung et al. 2019; Li et al. 2019), which dynamically allocate new parameters for future tasks to improve model plasticity. With the advent of large-scale pre-trained models (Dosovitskiy 2020; Radford et al. 2021), recent studies (Wang et al. 2022b; Lu et al. 2024; Liang and Li 2024; He et al. 2025) have explored Parameter-Efficient Fine-Tuning (PEFT) as a promising solution for CL, enabling task adaptation without full model retraining. In particular, CLIP-based continual learning (Huang et al. 2024; Zhou et al. 2025; Jha, Gong, and Yao 2024; Menabue et al. 2024) has attracted increasing interest due to its strong zero-shot generalization. STAR (Menabue et al. 2024) treats the pre-trained CLIP as a powerful task-id selector. RAPF (Huang et al. 2024) introduces a lightweight learnable projector and a selective pseudo-feature replay approach, enhances fine-grained class discrimination. PROOF (Zhou et al. 2025) improves image-text alignment via a learnable cross-attention module and alleviates forgetting through a weight averaging strategy.

Parameter-Efficient Fine-Tuning (PEFT) for Vision-Lanugage Models. Recent vision-language models (VLMs) such as CLIP (Radford et al. 2021) and BLIP (Li et al. 2022) enable strong zero-shot transfer but suffer from domain shift on downstream tasks. To address this, PEFT methods adapt VLMs by updating only a small set of parameters. Prompt tuning-based methods ($e.g.$, CoOp (Lester, Al-Rfou, and Constant 2021), CoCoOp (Zhou et al. 2022), ProDA (Lu et al. 2022), KgCoOp (Yao, Zhang, and Xu 2023)) learn soft prompts to condition the text branch, while adapter-based approaches (e.g., CLIP-Adapter (Gao et al. 2024), Tip-Adapter (Zhang et al. 2021)) introduce lightweight projection matrices or cache-based classifiers to improve model performance on downstream tasks. Other PEFT strategies include low-rank tuning (LoRA (Hu et al. 2021)), prefix tuning (Li and Liang 2021), and Mixture-of-Experts (MoE (Jacobs et al. 1991)). Despite tuning only a small subset of parameters, these methods have demonstrated effectiveness comparable to full-model fine-tuning.

Methodology

Problem Definition. In continual learning (CL), the model is trained sequentially on a series of tasks $\mathcal{D} = \{\mathcal{D}^1, \cdots, \mathcal{D}^s, \cdots, \mathcal{D}^S\}$, where the task $\mathcal{D}^s = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{|\mathcal{D}^s|}$ in s-th stage consists of image-label pairs from a distinct distribution. The label set are non-overlapping, $i.e.$, $\mathcal{Y}^s \cap \mathcal{Y}^{s'} = \varnothing$ for any $s \neq s'$, where $\mathcal{Y}^s$ denotes the label space of $\mathcal{D}^s$. The goal is to incrementally learn all S

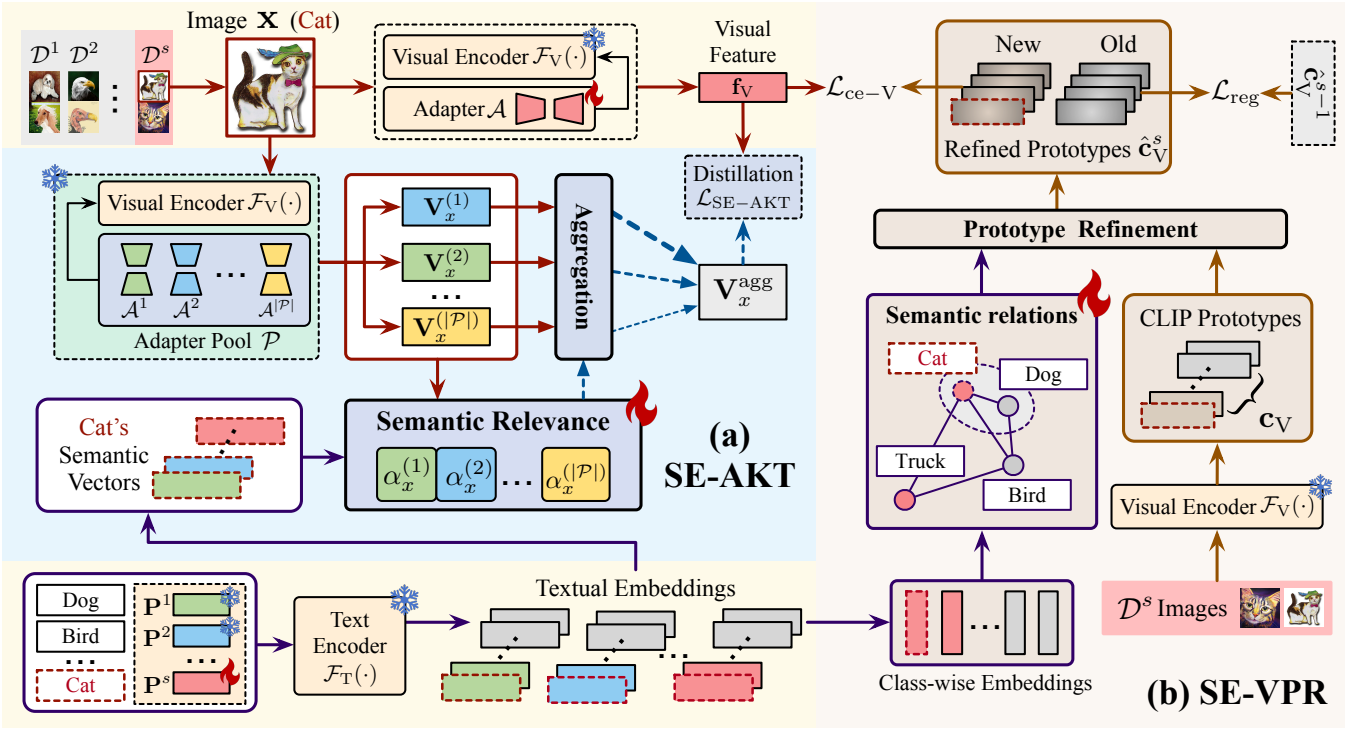

Figure 2: Overall framework. Our SECA is composed of two novel components: (a) The SG-AKT module that utilizes textual semantic vectors to aggregates relevant visual representations from a pool of historical adapters for distillation; (b) The SE-VPR module that leverages inter-class semantic relationships to refine CLIP prototypes, constructing a powerful visual-side classifier.

tasks while maintaining model performance on images from all seen classes $\mathcal{Y}^1 \cup \cdots \cup \mathcal{Y}^S$ during inference.

Overall framework. The overall framework of SECA is illustrated in Fig.2. Our SECA is built upon CLIP, which includes a visual encoder $\mathcal{F}_{\mathrm{V}}(\cdot)$ and a text encoder $\mathcal{F}_{\mathrm{T}}(\cdot)$. For efficient adaptation, we inserts learnable textual prompts and visual adapters into both CLIP's visual and text encoders. The proposed SECA mainly consists of two components: (1) Semantic-Guided Adaptive Knowledge Transfer (SG-AKT, Fig. 2(a)), an instance-adaptive distillation approach that utilizes textual cues to estimate semantic relevance and selectively aggregate past knowledge. The aggregated representations act as teacher signals for distillation. (2) Semantic-Enhanced Visual Prototype Refinement (SE-VPR, Fig. 2(b)), a module that leverages inter-class semantic relations to refine the structure of visual prototypes, which are then used as a visual-side classifier to support hybrid classification.

Hybrid Parameter-Efficient Fine-Tuning Baseline

CLIP Fine-Tuning with Learnable Textual Prompts. The CLIP architecture comprises a visual encoder $\mathcal{F}_{\mathrm{V}}(\cdot)$ and a text encoder $\mathcal{F}_{\mathrm{T}}(\cdot)$. For the s-th task $\mathcal{D}^s$, we introduce a learnable task-specific textual prompt $\mathbf{P}^s \in \mathbb{R}^{M \times d_{\mathrm{T}}}$, where M is the number of prompt tokens and d_{T} is the embedding dimension of the text encoder. Given a class label y, the input is a concatenation of the learnable prompt and its tokenized class name CLASS_y:

$$\mathbf{E}_y^s = [\mathbf{P}^s]_1[\mathbf{P}^s]_2 \cdots [\mathbf{P}^s]_M[\mathrm{CLASS}_y], \tag{1}$$

where $[\mathbf{P}^s]_m$ is the m-th learnable token in $\mathbf{P}^s$. This sequence is then fed into the frozen text encoder $\mathcal{F}_{\mathrm{T}}(\cdot)$ to generate the textual feature for y: $\mathcal{F}_{\mathrm{T}}(y; \mathbf{P}^s) = \mathcal{F}_{\mathrm{T}}(\mathbf{E}_y^s) \in \mathbb{R}^{d_{\mathrm{T}}}$.

CLIP Fine-Tuning with Visual Adapters. To adapt the visual encoder $\mathcal{F}_{\mathrm{V}}(\cdot)$ to downstream tasks, we introduce a set of task-shared lightweight adapters (Chen et al. 2022) into the visual backbone. At the s-th task, we denote the adapter set as $\mathcal{A} = \{\mathcal{A}_l\}_{l=1}^{L}$, where each $\mathcal{A}_l$ is inserted alongside the l-th transformer block. Given an input $\hat{\mathbf{h}}_l$ of the Feed-Forward Network (FFN) at layer l, the output is computed as:

$$\mathbf{h}_{l+1} = \hat{\mathbf{h}}_l + \mathrm{FFN}(\hat{\mathbf{h}}_l) + \mathcal{A}_l(\hat{\mathbf{h}}_l). \tag{2}$$

Given an image $\mathbf{x} \in \mathcal{D}^s$, its final visual feature is denoted as $\mathcal{F}_{\mathrm{V}}(\mathbf{x}; \mathcal{A}) \in \mathbb{R}^{d_{\mathrm{V}}}$, where d_{V} is the dimension of the representation from visual encoder.

Hybrid Fine-Tuning Baseline. The probability of $\mathbf{x}$ being classified to class y is computed as:

$$p^s(y \mid \mathbf{x}) = \frac{\exp(\langle \mathcal{F}_{\mathrm{V}}(\mathbf{x}; \mathcal{A}), \mathcal{F}_{\mathrm{T}}(y; \mathbf{P}^s) \rangle / \tau)}{\sum_{y \in \mathcal{Y}^s} \exp(\langle \mathcal{F}_{\mathrm{V}}(\mathbf{x}; \mathcal{A}), \mathcal{F}_{\mathrm{T}}(y; \mathbf{P}^s) \rangle / \tau)}, \tag{3}$$

where τ is a temperature factor and $\langle \cdot, \cdot \rangle$ denotes the cosine similarity between ℓ_2-normalized visual and textual features. Given a data point $(\mathbf{x}, y) \in \mathcal{D}^s$, the learnable PEFT modules are trained through a standard cross-entropy loss:

$$\mathcal{L}_{\mathrm{ce-T}} = -y \log p^s(y \mid \mathbf{x}), \tag{4}$$

This formulation keeps the pre-trained CLIP frozen while enabling efficient representation adaptation in both visual and textual branches.

Semantic-Guided Adaptive Knowledge Transfer

To enable selective knowledge transfer, we begin by extracting semantic vectors for new images, which are then used to assess their relevance to past representations from a pool of historical adapters. The resulting relevance scores guide the instance-adaptive aggregation of past representations, and the aggregated features serve as teacher signals for distillation.

Knowledge Aggregation and Distillation. To continually record learned visual knowledge, we maintain a pool of historical adapters, which serves as a repository of knowledge from past tasks. At the end of each task, the trained adapter is cached in a adapter pool $\mathcal{P} = \{\mathcal{A}^1, \mathcal{A}^2, \cdots, \mathcal{A}^{|\mathcal{P}|}\}$ with a fixed size $|\mathcal{P}|$. To maintain a constant pool size, we monitor adapter utility score and prune the past adapter with the highest utility when incorporating a new one.

When a new task arrives ($i.e.$, $\mathcal{D}^s$), given a data point $(\mathbf{x}, y) \in \mathcal{D}^s$, we extract the visual knowledge $\mathbf{V}_x$ of $\mathbf{x}$ from the adapter pool $\mathcal{P}$ by encoding $\mathbf{x}$ through all adapters in $\mathcal{P}$:

$$\mathbf{V}_x^{(p)} = \mathcal{F}_{\mathrm{V}}(\mathbf{x}; \mathcal{A}^p), \quad \mathbf{V}_x = [\mathbf{V}_x^{(1)}, \cdots, \mathbf{V}_x^{(|P|)}], \tag{5}$$

where $\mathbf{V}_x^{(p)} \in \mathbb{R}^{d_{\mathrm{V}}}$ is the visual representation of $\mathbf{x}$ encoded by the p-th adapter $\mathcal{A}^p \in \mathcal{P}$. $\mathbf{V}_x$ collects diverse visual representations of $\mathbf{x}$ from different previous models.

In parallel, we collect the textual semantic vectors of class label y associated with image $\mathbf{x}$. To be specific, we combine the ground-truth word embedding CLASS_y with textual prompts from the new and all previous tasks:

$$\mathbf{S}_y^{(p)} = \mathcal{F}_{\mathrm{T}}(y; \mathbf{P}^p), \quad \mathbf{S}_y = [\mathbf{S}_y^{(1)}, \cdots, \mathbf{S}_y^{(s)}], \tag{6}$$

where $\mathbf{S}_y^{(p)} \in d_{\mathrm{T}}$ is the text feature obtained by concatenating the word embedding CLASS_y with the p-th task's prompt $\mathbf{P}^p$. Each textual feature provides a task-specific semantic perspective of class y. This scheme extracts comprehensive and abundant semantic embeddings of class y.

During training, consider the image $\mathbf{x}$, we assess its semantic relevance to representations from the adapter pool using the associated semantic vectors $\mathbf{S}_y$. To be specific, we introduce two learnable semantic projectors, $\mathbf{W}_{\mathrm{S}} \in \mathbb{R}^{d_{\mathrm{T}} \times d_{\mathrm{V}}}$ and $\mathbf{W}_{\mathrm{V}} \in \mathbb{R}^{d_{\mathrm{V}} \times d_{\mathrm{V}}}$, to map both $\mathbf{S}_y$ and $\mathbf{V}_x$ into a shared semantic space. Afterwards, the relevance score between $\mathbf{x}$ and adapter $\mathcal{A}^p$ is formulated as follows:

$$\alpha_x^{(p)} = \frac{1}{s} \sum_{i=1}^{s} [\phi(\mathbf{S}_y^{(i)}) \mathbf{W}_{\mathrm{S}}]^\top [\phi(\mathbf{V}_x^{(p)}) \mathbf{W}_{\mathrm{V}}], \tag{7}$$

where $\phi(\cdot)$ denotes the LayerNorm (Ba, Kiros, and Hinton 2016) operation to to stabilize training. $\alpha_x^{(p)}$ represents the relevance of its ground-truth class y to learned visual knowledge from p-th adapter in shared semantic space. The relevance scores then used to guide the aggregation of visual knowledge $\mathbf{V}^{(x)}$ from the adapter pool:

$$\mathbf{V}_x^{\mathrm{agg}} = \sum_{p=1}^{|\mathcal{P}|} \frac{\exp(\lambda \alpha_x^{(p)})}{\sum_{i=1}^{|\mathcal{P}|} \exp(\lambda \alpha_x^{(i)})} \mathbf{V}_x^{(p)}, \tag{8}$$

where the scaling factor λ is introduced to promote smoother feature aggregation. Such an aggregation paradigm adaptively assigns higher weights to knowledge from past adapters that are closely matched to the textual semantic vectors for each instance, prioritizing aggregating relevant knowledge.

Two semantic projectors $\mathbf{W}_{\mathrm{S}}$ and $\mathbf{W}_{\mathrm{V}}$ are optimized by a cross-entropy loss $\mathcal{L}_{\mathrm{agg}}$, which guarantee the alignment between visual representations $\mathbf{V}_x$ and textual vectors $\mathbf{S}_y$:

$$\mathcal{L}_{\mathrm{agg}} = -y \log p^s(y \mid \mathbf{V}_x^{\mathrm{agg}}). \tag{9}$$

The aggregated representation $\mathbf{V}_x^{\mathrm{agg}}$ serves as a teacher representation to distill relevant knowledge into the current model. The distillation is implemented by minimizing the KL divergence between the predictions of the aggregated feature $\mathbf{V}_x^{\mathrm{agg}}$ and feature $\mathbf{f}_{\mathrm{V}} = \mathcal{F}_{\mathrm{V}}(\mathbf{x}, \mathcal{A})$ from the current adapter:

$$\mathcal{L}_{\text{SG-AKT}} = \sum_{y \in \mathcal{Y}^s} \mathrm{sg}\left(p^s(y \mid \mathbf{V}_x^{\mathrm{agg}}, \tau')\right) \log \frac{\mathrm{sg}\left(p^s(y \mid \mathbf{V}_x^{\mathrm{agg}}, \tau')\right)}{p^s(y \mid \mathbf{f}_{\mathrm{V}}, \tau') + \epsilon}. \tag{10}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operation, τ' is a temperature factor used to soften the probability distribution, similar to the role of τ in Eq. 3, and ϵ is a small constant added for numerical stability.

Adapter Pool Management via Utility Score. To effectively manage the fixed capacity of the adapter pool, we adopt a pruning strategy: when incorporating a new adapter, we remove the existing one with the highest utility score. We define the utility score of the p-th adapter as U^p, which represents the adapter's accumulated relevance to the current task. It is initially set to a uniform value of $\frac{1}{|\mathcal{P}|}$. During the aggregation process, the utility scores are updated using a momentum-based strategy based on $\alpha_x^{(p)}$:

$$\mathrm{U}^p \leftarrow \mu \mathrm{U}^p + (1 - \mu) \alpha_x^{(p)}, \quad p \in \{1, \cdots, |\mathcal{P}|\}, \tag{11}$$

where μ is a momentum factor. After training, a higher utility score indicates the adapter's knowledge has been sufficiently transferred to the latest model, making it a suitable candidate for removal to prune redundant knowledge while preventing the pool size from growing linearly with the number of tasks.

Semantic-Enhanced Visual Prototype Refinement

To bridge the modality gap, we construct a semantic-guided visual-side classifier for collaborative inference with the textual classifier. This is achieved by modeling inter-class semantic relationships via textual embeddings, which are then used to refine the structure of coarse visual prototypes, aligning them with semantic structures in the text branch.

Prototype Refinement. Consider the s-th task, for each seen class within and before the s-th task (with class label set $\mathcal{Y}^{1:s} = \mathcal{Y}^1 \cup \cdots \cup \mathcal{Y}^s$), we obtain a semantic embedding: Given a class k and the task-specific prompt $\mathbf{P}^s$, we encode the class name using CLIP's text encoder: $\mathbf{Z}_k = \mathcal{F}_{\mathrm{T}}(k; \mathbf{P}^s)$.

To enable flexible modeling of inter-class relationships, we introduce a learnable projector $\mathbf{H}_{\mathrm{proj}} \in \mathbb{R}^{d_{\mathrm{T}} \times d_{\mathrm{T}}}$ to project

the class-wise embeddings into a trainable and more expressive latent space. We model the inter-class affinity matrix $\mathbf{M}$, which captures semantic similarities between different classes, based on the projected class-wise embeddings:

$$\mathbf{M}_{k,j} = \exp(-\gamma \|\phi(\mathbf{Z}_k)\mathbf{H}_{\text{proj}} - \phi(\mathbf{Z}_j)\mathbf{H}_{\text{proj}}\|_2^2), \quad (12)$$

where $\mathbf{M}_{k,j}$ represents the affinity score between the k-th and j-th classes, and γ is a scaling factor to smooth the distribution of affinity values. Subsequently, the affinity scores are used to refine the structure of coarse visual prototypes:

$$\hat{\mathbf{c}}_{\text{V},k} = \sum_{j \in \mathcal{Y}^{1:s}} \frac{\mathbf{M}_{k,j}}{\sum_{i \in \mathcal{Y}^{1:s}} \mathbf{M}_{k,i}} \mathbf{c}_{\text{V},k}, \quad (13)$$

where $\mathbf{c}_{\text{V},k}$ denotes the coarse CLIP prototype of the k-th class. It is calculated by averaging the CLIP visual encoder features of all images belonging to class k at the beginning of its corresponding task:

$$\mathbf{c}_{\text{V},k} = \frac{1}{N_k} \sum_{(\mathbf{x},y) \in \mathcal{D}^{1:s}} \mathcal{F}_{\text{V}}(\mathbf{x}) \mathbb{1}(y = k), \quad (14)$$

where N_k is the number of images of class k. $\mathbb{1}(y = k)$ is an indicator function that equals 1 if $y = k$, and 0 otherwise.

The refined visual prototypes $\hat{\mathbf{c}}_{\text{V}}^s = [\hat{\mathbf{c}}_{\text{V},1}^s, \cdots, \hat{\mathbf{c}}_{\text{V},|\mathcal{Y}^{1:s}|}^s]$ serve as a visual-side classifier. For $(\mathbf{x}, y) \in \mathcal{D}^s$, the probability of predicting $\mathbf{x}$ as class $y \in \mathcal{Y}^s$ is computed as:

$$\hat{p}_{\text{V}}(y \mid \mathbf{x}) = \frac{\exp(\langle \mathcal{F}_{\text{V}}(\mathbf{x}; \mathcal{A}), \hat{\mathbf{c}}_{\text{V},y}^s \rangle / \tau)}{\sum_{y \in \mathcal{Y}^s} \exp(\langle \mathcal{F}_{\text{V}}(\mathbf{x}; \mathcal{A}), \hat{\mathbf{c}}_{\text{V},y}^s \rangle / \tau)}. \quad (15)$$

The projector $\mathbf{H}_{\text{proj}}$ is trained with a classification loss $\mathcal{L}_{\text{ce}-\text{V}}$ to instruct the inter-class relationship modeling:

$$\mathcal{L}_{\text{ce}-\text{V}} = -y \log \hat{p}_{\text{V}}(y \mid \mathbf{x}). \quad (16)$$

Prototype Consistency Regularization. Furthermore, to prevent $\mathbf{H}_{\text{proj}}$ from overfitting to new classes, we introduce a prototype regularization loss $\mathcal{L}_{\text{reg}}$ that promotes the temporal stability of visual prototypes for previously learned classes ($\mathcal{Y}^{1:s-1}$). This loss encourages the updated prototypes associated with old tasks to remain close to their earlier versions:

$$\mathcal{L}_{\text{reg}} = \frac{1}{|\mathcal{Y}^{1:s-1}|} \sum_{k \in \mathcal{Y}^{1:s-1}} \|\hat{\mathbf{c}}_{\text{V},k}^s - \hat{\mathbf{c}}_{\text{V},k}^{s-1}\|_2^2, \quad (17)$$

where $\hat{\mathbf{c}}_{\text{V},k}^{s-1}$ is the refined visual prototype after the $(s-1)$-th task of the previous class k. By injecting textual semantics into visual prototypes and preserving temporal consistency, SE-VPR establishes a powerful visual-side classifier to compensate for the modality gap.

Training Objective and Inference Strategy

Training Objective. Our framework optimizes the SG-AKT and the SE-VPR modules, including the following terms:

$$\mathcal{L} = \mathcal{L}_{\text{ce}-\text{T}} + \underbrace{\mathcal{L}_{\text{agg}} + \beta \mathcal{L}_{\text{SG}-\text{AKT}}}_{\text{SG}-\text{AKT}} + \underbrace{\mathcal{L}_{\text{ce}-\text{V}} + \mathcal{L}_{\text{reg}}}_{\text{SE}-\text{VPR}}, \quad (18)$$

where β is a trade-off hyper-parameter controlling the contribution of $\mathcal{L}_{\text{SG}-\text{AKT}}$. Due to the growing accumulation of old task knowledge during continual adaptation, we set β as a task-dependent variable that increases with tasks.

Inference Strategy. During inference, after the s-th task, given an image $\mathbf{x}$, its final prediction $\hat{y}$ is in a hybrid form, combining the prediction $\hat{p}_{\text{V}}$ of the refined visual classifier $\hat{\mathbf{c}}_{\text{V}}$ and predictions from all task-specific text classifiers:

$$\hat{y} = \arg\max_y \left(\hat{p}_{\text{V}}(y \mid \mathbf{x}, \tau') + \frac{1}{s} \sum_{i=1}^{s} p^i(y \mid \mathbf{x}, \tau') \right), \quad (19)$$

where τ' is the temperature factor for inference, which is set to match the one used in $\mathcal{L}_{\text{SG}-\text{AKT}}$ to ensure consistency between training and inference.

Experiments

Datasets. We evaluate our method on three representative CIL benchmarks: ImageNetR (Hendrycks et al. 2021a), ImageNetA (Hendrycks et al. 2021b), and CIFAR-100 (Krizhevsky, Hinton et al. 2009). ImageNetR comprises 30,000 images spanning 200 classes, covering a wide range of visual styles. ImageNetA contains 7,500 challenging real-world adversarial examples images across 200 classes. CIFAR100 is a commonly-used dataset in continual learning, which consists of 60000 32×32 images of 100 classes. We adopt both 10-split and 20-split settings, where the class set is evenly divided into 10 or 20 non-overlapping tasks.

Evaluation Metrics. Following prior PEFT-based CIL works (Tan et al. 2024; Gao, Cen, and Chang 2024; Huang et al. 2024), we adopt two standard evaluation metrics: (1) The last session accuracy (Last): the final accuracy over all classes after completing the last task, and (2) Average accuracy (Avg), the average accuracy across all incremental tasks. All experiments are conducted using three random seeds, and we report the mean performances.

Implementation Details. All experiments are conducted on a single NVIDIA RTX 3090 GPU. We adopt CLIP (Radford et al. 2021) with a ViT-B/16 (Dosovitskiy 2020) visual encoder pre-trained by OPENAI as our backbone. Our model is optimized using Adam (Kingma and Ba 2014) with an initial learning rate of 0.001. For the training schedule, each task is trained for 10 epochs on ImageNetR and ImageNetA, and 3 epochs on CIFAR100. The batch size is set to 64 for ImageNetR and ImageNetA and 100 for CIFAR100. The capacity of the frozen adapter pool $\mathcal{P}$ in SG-AKT is set to 5, and the temperature factor τ' in $\mathcal{L}_{\text{SG}-\text{AKT}}$ is set to 0.05.

SECA++: An Enhanced Variant of SECA. Following existing PEFT-based CIL methods (Tan et al. 2024; Huang et al. 2024; Lu et al. 2024), we further implement an enhanced

Table 1: Experimental results on ImageNetR and ImageNetA. We report the averaged results over 3 trials. VPT-NSP† and RAPF† denote the feature replay-free version of VPT-NSP (Lu et al. 2024) and RAPF (Huang et al. 2024), respectively. The highest results are in **bold**, and the second best results are underlined.

Method	Venue	Backbone	ER/FR	10S-ImageNetR		10S-ImageNetA		20S-ImageNetR		20S-ImageNetA	
				Last ↑	Avg ↑	Last ↑	Avg ↑	Last ↑	Avg ↑	Last ↑	Avg ↑
L2P (Wang et al. 2022b)	CVPR-22	ViT-21k	✗	72.34	77.36	44.04	51.24	69.64	75.28	40.48	49.62
CODA (Smith et al. 2023)	CVPR-23	ViT-21k	✗	73.31	78.47	52.08	63.92	69.96	75.34	44.62	54.86
RanPAC (McDonnell et al. 2023)	NeurIPS-23	ViT-21k	✗	77.90	82.91	62.40	67.58	–	–	–	–
Adam-adapter (Zhou et al. 2024)	IJCV-24	ViT-21k	✗	65.79	72.42	48.81	58.84	57.42	64.75	48.65	59.55
SSIAT (Tan et al. 2024)	CVPR-24	ViT-21k	✓	79.38	83.63	62.43	70.83	75.67	82.30	59.16	68.45
VPT-NSP (Lu et al. 2024)	NeurIPS-24	ViT-21k	✗	77.95	83.44	53.83	63.93	75.69	81.87	49.81	61.41
DIA (Li et al. 2024)	CVPR-25	ViT-21k	✓	79.03	85.61	59.78	70.43	76.32	83.51	–	–
ACMap (Fukuda, Kera, and Kawamoto 2024)	CVPR-25	ViT-21k	✗	73.50	79.50	56.19	65.19	–	–	–	–
ZS-CLIP	–	CLIP	✗	74.93	81.56	47.33	58.35	74.93	82.09	47.33	59.36
L2P (Wang et al. 2022b)	CVPR-22	CLIP	✗	75.98	81.67	47.86	59.35	68.78	76.87	47.54	59.77
DualPrompt (Wang et al. 2022a)	CVPR-22	CLIP	✗	75.77	82.01	48.18	59.05	69.41	77.07	48.05	60.22
CODA (Smith et al. 2023)	CVPR-22	CLIP	✗	67.52	78.00	50.24	64.32	64.53	75.23	49.95	65.08
Adam-adapter (Zhou et al. 2024)	IJCV-24	CLIP	✗	71.35	78.65	59.35	68.56	68.75	76.71	58.55	67.72
RAPF† (Huang et al. 2024)	ECCV-24	CLIP	✗	73.23	82.20	45.54	60.67	71.28	81.66	43.85	58.54
RAPF (Huang et al. 2024)	ECCV-24	CLIP	✓	79.62	86.28	55.37	67.32	80.28	85.58	49.85	65.28
CLAP (Jha, Gong, and Yao 2024)	NeurIPS-24	CLIP	✓	79.98	85.77	58.66	69.35	79.12	85.03	55.84	67.72
VPT-NSP† (Lu et al. 2024)	NeurIPS-24	CLIP	✗	77.45	83.60	47.14	59.04	77.52	83.94	47.07	59.88
VPT-NSP (Lu et al. 2024)	NeurIPS-24	CLIP	✓	82.48	87.94	61.42	71.76	82.06	88.09	60.70	72.57
PROOF (Zhou et al. 2025)	T-PAMI-25	CLIP	✓	77.25	82.69	55.67	65.50	77.05	82.83	54.05	64.53
SECA (Ours)	–	CLIP	✗	83.18	88.58	65.09	74.45	82.25	88.19	62.80	73.02
SECA++ (Ours)	–	CLIP	✓	**83.41**	**88.75**	**65.77**	**74.65**	**83.02**	**88.62**	**64.41**	**74.64**

Table 2: Experimental results on CIFAR100. All results are based on the CLIP model with a ViT-B/16 vision backbone.

Method	ER/FR	10S-CIFAR100		20S-CIFAR100	
		Last ↑	Avg ↑	Last ↑	Avg ↑
ZS-CLIP	✗	66.68	75.15	66.68	75.93
L2P (Wang et al. 2022b)	✗	73.08	81.90	68.67	79.18
DualPrompt (Wang et al. 2022a)	✗	72.51	81.45	69.91	79.74
CODA (Smith et al. 2023)	✗	62.25	76.98	41.98	69.78
SLCA (Zhang et al. 2023)	✓	67.58	80.53	66.84	78.96
Adam-adapter (Zhou et al. 2024)	✗	65.50	75.76	58.12	70.18
RAPF (Huang et al. 2024)	✓	79.04	86.19	79.26	86.87
CLAP (Jha, Gong, and Yao 2024)	✓	78.21	86.13	77.35	86.08
PROOF (Zhou et al. 2025)	✓	76.29	84.88	76.13	85.12
SECA (Ours)	✗	79.79	86.70	77.73	85.35
SECA++ (Ours)	✓	**81.59**	**87.80**	**80.07**	**87.11**

variant, denoted as SECA++, which incorporates Gaussian sampling-based feature replay to alleviate inter-task confusion during the training of multi-modal classifier, particularly for semantically confusing or fine-grained classes.

Comparison with the State-of-the-Arts

We compare SECA with recent state-of-the-art PEFT-based approaches, including prompt-based methods (Wang et al. 2022b; Lu et al. 2024; Smith et al. 2023), adapter-based methods (Li et al. 2024; Tan et al. 2024; Huang et al. 2024), the zero-shot CLIP (ZS-CLIP) and other representative techniques built upon either ViT-B/16 pretrained on ImageNet-21K (Russakovsky et al. 2015) (ViT-21K) or CLIP-pretrained backbones. Evaluations are conducted on ImageNetR, ImageNetA and CIFAR100, and the results are shown in Tab.**??** and Tab.2. 'ER/FP' indicates the use of experience replay (Rolnick et al. 2019) or distributional sampling-based feature replay. Notably, most CLIP-based CIL methods incorporate such replay strategies into their frameworks. For a fair comparison, we reproduce the results of two representative methods, RAPF (Huang et al. 2024) and VPT-NSP (Lu et al. 2024), using a CLIP backbone under a replay-free setting. These variants are denoted as RAPF† and VPT-NSP†.

The results demonstrate the superiority of our method. Specifically, under the 10-split setting, our replay-free SECA surpasses the strongest VPT-NSP (the replay-based version) by 0.70%, 3.67% on ImageNet-R and ImageNet-A in terms of Last accuracy, respectively. Our feature replay-enhanced version, SECA++, further boosts the performance, outperforming VPT-NSP by 0.93% and 4.35% in Last accuracy on 10S-ImageNetR and 10S-ImageNetA, respectively.

To sum up, our SECA has two main advantages: (1) It enables seamless knowledge injection into each layer within the visual encoder via layer-wise adapters combined with a adaptive knowledge transfer mechanism, rather than merely attaching tunable modules after the frozen backbone; (2) It is compatible with replay-based strategies, allowing further performance improvements when integrated with lightweight feature replay. ***More evaluations can be found in Appendix .B.***

Ablation Studies

The effectiveness of components. We conduct ablation studies on components of our training pipeline, including the hybrid PEFT baseline (H-PEFT), SG-AKT and SE-VPR. The results are shown in Tab.3. The **H-PEFT** (Idx 2) achieves a substantial improvement over ZS-CLIP (Idx 1), with +5.64% and +8.45% gains in Last accuracy on 10S-ImageNetR and 10S-ImageNetA. Incorporating **SG-AKT** (Idx 3) brings further gains (e.g., +2.13% on ImageNetA), effectively improving representation learning in the visual encoder by selectively transferring past knowledge. Moreover, **SE-VPR** (Idx 4) form a hybrid classification paradigm by leveraging textual semantics to refine visual prototypes, resulting in notable performance improvements (e.g., +4.65% on ImageNetA)

Table 3: Ablation studies of each component in our approach on the three benchmarks.

Idx	H-PEFT	SG-AKT	SE-VPR	10S-ImageNetA		10S-CIFAR100		10S-ImageNetR	
				Last ↑	Avg ↑	Last ↑	Avg ↑	Last ↑	Avg ↑
1	-	-	-	47.33	58.35	67.19	76.23	74.93	81.56
4	✓	-	-	$55.78_{\pm 1.36}$	$67.97_{\pm 0.38}$	$73.97_{\pm 0.65}$	$82.75_{\pm 0.28}$	$80.57_{\pm 0.15}$	$87.04_{\pm 0.35}$
5	✓	✓	-	$57.91_{\pm 0.55}$	$68.56_{\pm 0.66}$	$75.93_{\pm 0.09}$	$84.16_{\pm 0.21}$	$81.36_{\pm 0.49}$	$87.55_{\pm 0.29}$
6	✓	-	✓	$62.62_{\pm 0.65}$	$73.15_{\pm 0.25}$	$77.13_{\pm 0.56}$	$84.91_{\pm 0.43}$	$81.30_{\pm 0.55}$	$87.63_{\pm 0.12}$
7	✓	✓	✓	$\mathbf{65.09}_{\pm 0.48}$	$\mathbf{74.45}_{\pm 0.76}$	$\mathbf{79.79}_{\pm 0.42}$	$\mathbf{86.70}_{\pm 0.16}$	$\mathbf{83.18}_{\pm 0.34}$	$\mathbf{88.58}_{\pm 0.07}$

Table 4: Ablation studies of different distillation strategies on 10S-ImageNetA and 10S-CIFAR100.

Method	10S-ImageNetA		10S-CIFAR100	
	Last ↑	Avg ↑	Last ↑	Avg ↑
Methods without SE-VPR.				
Seq.	$55.78_{\pm 1.36}$	$67.97_{\pm 0.38}$	$73.97_{\pm 0.65}$	$82.75_{\pm 0.28}$
CLIP-KD	$55.01_{\pm 0.68}$	$67.00_{\pm 0.27}$	$71.21_{\pm 0.61}$	$80.98_{\pm 0.50}$
Vanilla	$57.05_{\pm 0.75}$	$67.93_{\pm 0.47}$	$74.55_{\pm 0.26}$	$83.26_{\pm 0.32}$
Avg-KD	$56.90_{\pm 0.11}$	$68.02_{\pm 0.76}$	$75.05_{\pm 0.78}$	$83.65_{\pm 0.33}$
SG-AKT (Ours)	$\mathbf{57.91}_{\pm 0.55}$	$\mathbf{68.56}_{\pm 0.66}$	$\mathbf{75.93}_{\pm 0.09}$	$\mathbf{84.16}_{\pm 0.21}$
Methods with SE-VPR.				
Seq.	$62.62_{\pm 0.65}$	$73.15_{\pm 0.25}$	$77.13_{\pm 0.56}$	$84.91_{\pm 0.43}$
CLIP-KD	$62.39_{\pm 0.77}$	$73.01_{\pm 1.17}$	$73.30_{\pm 0.80}$	$83.76_{\pm 0.18}$
Vanilla	$63.51_{\pm 0.71}$	$73.89_{\pm 0.69}$	$76.42_{\pm 0.44}$	$85.75_{\pm 0.37}$
Avg-KD	$64.32_{\pm 0.59}$	$74.08_{\pm 0.82}$	$78.98_{\pm 0.51}$	$86.47_{\pm 0.30}$
SG-AKT (Ours)	$\mathbf{65.09}_{\pm 0.48}$	$\mathbf{74.45}_{\pm 0.76}$	$\mathbf{79.79}_{\pm 0.42}$	$\mathbf{86.70}_{\pm 0.16}$

compared to text-only classifier.

SG-AKT v.s. other Distillation Strategies. We compare our SG-AKT with four alternative training strategies in Tab.4: (1) **Seq.** denotes sequential tuning without any distillation loss; (2) **CLIP-KD** distills from the frozen CLIP model as a global teacher; (3) **Vanilla** distills knowledge from the most recent task model, following prior works (Kang, Park, and Han 2022; Asadi et al. 2023); (4) **Avg-KD** is a simplified variant of SG-AKT that aggregates adapter pool features by averaging, without text-aware attention. Tab. 4 shows that SG-AKT consistently outperforms all variants, with +0.77% and +1.58% Last-acc improvements over Avg-KD and Vanilla on 10S-ImageNetA, respectively. Although these two designs are effective for knowledge retention, they fall short compared to SG-AKT, highlighting the effectiveness of leveraging textual information to prioritize semantically relevant transfer.

SE-VPR v.s. other Classifier Designs. We compare our SE-VPR module with four alternative designs for constructing the classifier in Tab.5: (1) **Only Text** denotes using the text encoder alone for classification without visual classifiers; (2) **Centroid (CLIP)** computes class centroids using pretrained frozen CLIP prototypes as the visual-side classifier; (3) **Centroid (adapted)** leverages the fine-tuned CLIP visual encoder to extract class centroids; (4) **Linear** trains an additional linear classifier on top of the adapted visual features. The results indicate the necessity of incorporating a visual-side classifier to bridge the modality gap. It also show the superiority of our SE-VPR, which achieves the best performance compared to other variants, yielding +1.93% and +1.49% gains in Last accuracy over Centroid (adapted) on 10S-ImageNetA and 10S-CIFAR100.

Table 5: Ablation studies of different classifier designs on 10S-ImageNetA and 10S-CIFAR100 datasets.

Method	10S-ImageNetA		10S-CIFAR100	
	Last ↑	Avg ↑	Last ↑	Avg ↑
Only Text	$57.91_{\pm 0.55}$	$68.56_{\pm 0.66}$	$75.93_{\pm 0.09}$	$84.16_{\pm 0.21}$
Centroid (CLIP)	$58.55_{\pm 0.57}$	$68.30_{\pm 1.17}$	$75.77_{\pm 0.56}$	$83.88_{\pm 0.42}$
Centroid (Adapted)	$63.16_{\pm 0.30}$	$72.89_{\pm 0.98}$	$78.30_{\pm 0.89}$	$86.15_{\pm 0.66}$
Linear	$51.07_{\pm 3.34}$	$62.76_{\pm 3.19}$	$74.20_{\pm 0.50}$	$82.98_{\pm 0.59}$
SE-VPR (Ours)	$\mathbf{65.09}_{\pm 0.48}$	$\mathbf{74.45}_{\pm 0.76}$	$\mathbf{79.79}_{\pm 0.42}$	$\mathbf{86.70}_{\pm 0.16}$

Hyper-Parameter Analysis

Analysis of the Pool Size $|\mathcal{P}|$ and the Temperature Factor τ'. We analyze the sensitivity of two key hyper-parameters in SG-AKT: the size $|\mathcal{P}|$ of the adapter pool and the temperature factor τ' for knowledge transfer. The results are shown in Fig. 3. The performance constantly improves with larger pool size and saturates when $|\mathcal{P}| \geq 5$, where the results closely match those of using all previous adapters (denoted as 'ALL'). This confirms that our utility-score-based pool management effectively prunes redundant knowledge without compromising accuracy. For the temperature factor τ', a relatively high τ' yields better performance, suggesting that smoother teacher labels benefits knowledge trransfer. Based on the results, we set $|\mathcal{P}| = 5$ and $\tau' = 20.0$ for all datasets.

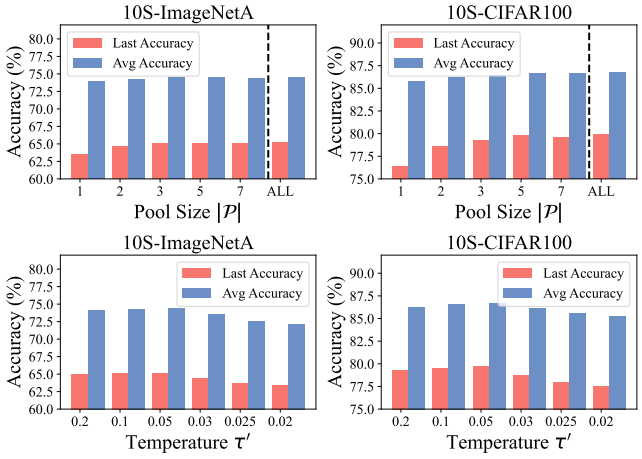

Figure 3: Model performances under different pool sizes $|\mathcal{P}|$ and temperatures τ'.

Conclusion

In this paper, we focus on the potential of textual semantics in addressing the stability–plasticity dilemma in continual learning (CL), and propose Semantic-Enriched Continual Adaptation (SECA), a unified framework for class-incremental

learning with CLIP. By assessing the new image's semantic relevance to historical knowledge using textual semantics, our Semantic-Guided Adaptive Knowledge Transfer (SG-AKT) selectively transfers relevant past knowledge to the current model, suppressing the transfer of unrelated or conflicting knowledge. Furthermore, by injecting inter-class semantic relations into visual prototypes, our Semantic-Enhanced Visual Prototype Refinement (SE-VPR) builds a powerful visual-side classifier to bridge the modality gap. We believe SECA provides valuable insights into text-grounded continual learning by demonstrating how textual semantics enable better stability–plasticity trade-off. In the future, we aim to further explore adaptive transfer and multi-modal synergy in CL.

References

Aljundi, R.; Babiloni, F.; Elhoseiny, M.; Rohrbach, M.; and Tuytelaars, T. 2018. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, 139–154.

Aljundi, R.; Belilovsky, E.; Tuytelaars, T.; Charlin, L.; Caccia, M.; Lin, M.; and Page-Caccia, L. 2019a. Online continual learning with maximal interfered retrieval. *Advances in neural information processing systems*, 32.

Aljundi, R.; Lin, M.; Goujaud, B.; and Bengio, Y. 2019b. Gradient based sample selection for online continual learning. *Advances in neural information processing systems*, 32.

Asadi, N.; Davari, M.; Mudur, S.; Aljundi, R.; and Belilovsky, E. 2023. Prototype-sample relation distillation: towards replay-free continual learning. In *International conference on machine learning*, 1093–1106. PMLR.

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bonicelli, L.; Boschini, M.; Porrello, A.; Spampinato, C.; and Calderara, S. 2022. On the effectiveness of lipschitz-driven rehearsal in continual learning. *Advances in Neural Information Processing Systems*, 35: 31886–31901.

Chen, Q.; Shui, C.; Han, L.; and Marchand, M. 2023. On the stability-plasticity dilemma in continual meta-learning: Theory and algorithm. *Advances in Neural Information Processing Systems*, 36: 27414–27468.

Chen, S.; Ge, C.; Tong, Z.; Wang, J.; Song, Y.; Wang, J.; and Luo, P. 2022. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35: 16664–16678.

Dosovitskiy, A. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fahim, A.; Murphy, A.; and Fyshe, A. 2024. It's Not a Modality Gap: Characterizing and Addressing the Contrastive Gap. *arXiv preprint arXiv:2405.18570*.

Fukuda, T.; Kera, H.; and Kawamoto, K. 2024. Adapter Merging with Centroid Prototype Mapping for Scalable Class-Incremental Learning. *arXiv preprint arXiv:2412.18219*.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2024. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595.

Gao, Z.; Cen, J.; and Chang, X. 2024. Consistent Prompting for Rehearsal-Free Continual Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28463–28473.

He, L.; Cheng, D.; Ma, Z.; Wang, H.; Zhang, D.; Wang, N.; and Gao, X. 2025. CKAA: Cross-subspace Knowledge Alignment and Aggregation for Robust Continual Learning. *arXiv preprint arXiv:2507.09471*.

Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T.; Parajuli, S.; Guo, M.; et al. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8340–8349.

Hendrycks, D.; Zhao, K.; Basart, S.; Steinhardt, J.; and Song, D. 2021b. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15262–15271.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, L.; Cao, X.; Lu, H.; and Liu, X. 2024. Class-Incremental Learning with CLIP: Adaptive Representation Adjustment and Parameter Fusion. arXiv:2407.14143.

Hung, C.-Y.; Tu, C.-H.; Wu, C.-E.; Chen, C.-H.; Chan, Y.-M.; and Chen, C.-S. 2019. Compacting, picking and growing for unforgetting continual learning. *Advances in neural information processing systems*, 32.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1): 79–87.

Jha, S.; Gong, D.; and Yao, L. 2024. Clap4clip: Continual learning with probabilistic finetuning for vision-language models. *arXiv preprint arXiv:2403.19137*.

Jung, S.; Ahn, H.; Cha, S.; and Moon, T. 2020. Continual learning with node-importance based adaptive group sparse regularization. *Advances in neural information processing systems*, 33: 3647–3658.

Kang, M.; Park, J.; and Han, B. 2022. Class-incremental learning by knowledge distillation with adaptive feature consolidation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16071–16080.

Kim, D.; and Han, B. 2023. On the stability-plasticity dilemma of class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20196–20204.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3045–3059. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, J.; Wang, S.; Qian, B.; He, Y.; Wei, X.; and Gong, Y. 2024. Dynamic Integration of Task-Specific Adapters for Class Incremental Learning. *arXiv preprint arXiv:2409.14983*.

Li, X.; Zhou, Y.; Wu, T.; Socher, R.; and Xiong, C. 2019. Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In *International conference on machine learning*, 3925–3934. PMLR.

Li, X. L.; and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4582–4597. Online: Association for Computational Linguistics.

Liang, V. W.; Zhang, Y.; Kwon, Y.; Yeung, S.; and Zou, J. Y. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35: 17612–17625.

Liang, Y.-S.; and Li, W.-J. 2024. InfLoRA: Interference-Free Low-Rank Adaptation for Continual Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23638–23647.

Lu, Y.; Liu, J.; Zhang, Y.; Liu, Y.; and Tian, X. 2022. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5206–5215.

Lu, Y.; Zhang, S.; Cheng, D.; Xing, Y.; Wang, N.; Wang, P.; and Zhang, Y. 2024. Visual Prompt Tuning in Null Space for Continual Learning. *arXiv preprint arXiv:2406.05658*.

McDonnell, M. D.; Gong, D.; Parvaneh, A.; Abbasnejad, E.; and Van den Hengel, A. 2023. Ranpac: Random projections and pre-trained models for continual learning. *Advances in Neural Information Processing Systems*, 36: 12022–12053.

Menabue, M.; Frascaroli, E.; Boschini, M.; Sangineto, E.; Bonicelli, L.; Porrello, A.; and Calderara, S. 2024. Semantic residual prompts for continual learning. In *European Conference on Computer Vision*, 1–18. Springer.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rolnick, D.; Ahuja, A.; Schwarz, J.; Lillicrap, T.; and Wayne, G. 2019. Experience replay for continual learning. *Advances in neural information processing systems*, 32.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.

Smith, J. S.; Karlinsky, L.; Gutta, V.; Cascante-Bonilla, P.; Kim, D.; Arbelle, A.; Panda, R.; Feris, R.; and Kira, Z. 2023. 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*, 11909–11919.

Tan, Y.; Zhou, Q.; Xiang, X.; Wang, K.; Wu, Y.; and Li, Y. 2024. Semantically-Shifted Incremental Adapter-Tuning is A Continual ViTransformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23252–23262.

Wang, H.; Cheng, D.; He, L.; Li, Y.; Li, J.; Wang, N.; and Gao, X. 2025. EKPC: Elastic Knowledge Preservation and Compensation for Class-Incremental Learning. *arXiv preprint arXiv:2506.12351*.

Wang, Z.; Zhang, Z.; Ebrahimi, S.; Sun, R.; Zhang, H.; Lee, C.-Y.; Ren, X.; Su, G.; Perot, V.; Dy, J.; et al. 2022a. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*, 631–648. Springer.

Wang, Z.; Zhang, Z.; Lee, C.-Y.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022b. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 139–149.

Yao, H.; Zhang, R.; and Xu, C. 2023. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6757–6767.

Zenke, F.; Poole, B.; and Ganguli, S. 2017. Continual learning through synaptic intelligence. In *International conference on machine learning*, 3987–3995. PMLR.

Zhang, G.; Wang, L.; Kang, G.; Chen, L.; and Wei, Y. 2023. Slca: Slow learner with classifier alignment for continual learning on a pre-trained model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19148–19158.

Zhang, R.; Fang, R.; Zhang, W.; Gao, P.; Li, K.; Dai, J.; Qiao, Y.; and Li, H. 2021. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*.

Zhou, D.-W.; Cai, Z.-W.; Ye, H.-J.; Zhan, D.-C.; and Liu, Z. 2024. Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need. *International Journal of Computer Vision*, 1–21.

Zhou, D.-W.; Zhang, Y.; Wang, Y.; Ning, J.; Ye, H.-J.; Zhan, D.-C.; and Liu, Z. 2025. Learning without forgetting for vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16816–16825.