REVISITING CONTINUAL LEARNING FOR MULTIMODAL RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval is a core capability of vision-language models (VLMs), yet continual adaptation for retrieval remains underexplored, with existing evaluations largely restricted to narrow domains and a single training protocol. We introduce SCOPE, a reproducible framework for continual multimodal retrieval spanning seven heterogeneous visual domains, three training protocols, and both task-specific and global-gallery evaluation. Using SCOPE, we find that conclusions about continual learning (CL) methods depend strongly on the evaluation setting: method rankings change across training protocols, task-specific galleries conceal cross-task retrieval failures, and simple fine-tuning remains competitive with dedicated CL methods. Our analysis further shows that preserving local neighbourhood structure is insufficient to preserve fine-grained retrieval rankings. As a reference method for SCOPE, we introduce Adapter Routing and Consolidation (ARC), which limits cross-task interference through low-rank adapters and selective consolidation. ARC consistently outperforms the evaluated methods across SCOPE settings. Together, these results establish SCOPE as a comprehensive evaluation framework and highlight the importance of diverse training protocols and gallery settings for reliably assessing progress in continual multimodal retrieval.

1 INTRODUCTION

Retrieval is a core functionality of vision-language models (VLMs) such as CLIP (Radford et al., 2021), underlying the real-world deployment of search, recommendation, and indexing systems. Despite its immense practical importance, the problem of continually updating multimodal models for retrieval has received limited attention. Only a small number of works explicitly address continual cross-modal retrieval (Liu et al., 2025; Cui et al., 2024; Ni et al., 2023; Li et al., 2025), while the majority of research on continual learning (CL) for VLMs focuses on classification-oriented settings, i.e., class-incremental learning (CIL) (Wang et al., 2021b; Jha et al., 2024; Yu et al., 2024; Huang et al., 2024; Lu et al., 2025; Zheng et al., 2023).

However, continual multimodal retrieval imposes requirements beyond those typically considered in class-incremental learning: sequential updates must preserve a globally consistent embedding space in which similarities remain comparable across tasks (Wang et al., 2021a). Even small representation shifts can therefore substantially alter global rankings, while catastrophic forgetting manifests as distortions of relative similarities and retrieval orderings (Cui et al., 2024; Ni et al., 2023) (see Figure 1).

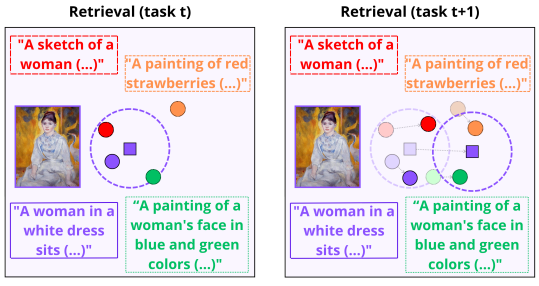


Figure 1: Small shifts in the embedding space caused by continual adaptation can substantially alter retrieval rankings, even if the overall neighbourhood structure appears stable.

Detecting failures in retrieval requires retrieval-specific metrics and task sequences that reveal how learning new data affects rankings across previously seen domains. However, existing work on continual retrieval often lacks sufficiently broad training protocols and domain diversity, so such evaluations fail to capture retrieval-specific challenges and reliably assess the true

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

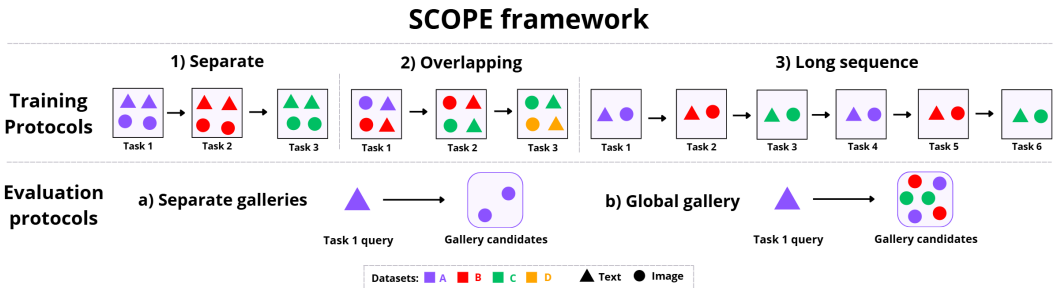


Figure 2: Overview of our **SCOPE** framework for continual retrieval training and evaluation.

utility of CL approaches for retrieval. To address this gap, we introduce SCOPE, a new evaluation framework for continual multimodal retrieval. Our framework consists of three training protocols that address different challenges in continual retrieval: separate sequences of heterogeneous, non-overlapping datasets spanning various visual domains; tasks with overlapping domains to simulate more complex data distributions; and long sequences in which similar domains recur. In addition to the commonly used task-specific evaluation, where each query is matched only against candidates from the same task, we advocate also including a more challenging global-gallery evaluation, where candidates from all tasks compete directly in a shared retrieval space.

We use SCOPE to conduct a systematic analysis of knowledge transfer, interference, and robustness across common CL methods used for retrieval. Notably, we find that conclusions about these methods depend strongly on the training and evaluation protocol. In particular, methods that perform well under one continual protocol do not necessarily retain their advantage under others, while simple fine-tuning remains a competitive baseline. The differences become particularly pronounced under global-gallery evaluation, which exposes cross-task retrieval failures hidden by task-specific galleries. To better understand these failures, we analyse the learned representations and find that preserving stable local neighbourhoods is insufficient for reliable retrieval: a method may retain similar neighbours across continual updates while still allowing hard negatives to outrank the correct match. This reveals a gap between preserving local embedding geometry and maintaining the fine-grained rankings required for retrieval.

Since our analysis suggests that limiting cross-task interference can help preserve the inter-embedding rankings, we introduce Adapter Routing and Consolidation (ARC), a simple parameter-isolation baseline that learns task-specific low-rank adaptations and selectively consolidates compatible ones. ARC consistently outperforms the evaluated CL methods across SCOPE settings, providing a strong reference point for evaluating future methods. Together, our results demonstrate the importance of thorough evaluation of the continual retrieval models across diverse training protocols and gallery settings in order to draw reliable conclusions about their practical feasibility.

Our main contributions can be summarized as follows:

- We introduce **SCOPE**, a principled evaluation framework for continual multimodal retrieval spanning seven heterogeneous visual domains, three complementary training protocols, and both task-specific and global gallery evaluation.
- We present a **systematic empirical analysis** of existing CL methods under SCOPE, showing that their effectiveness for continual retrieval is highly sensitive to the training and evaluation setting, with common CL strategies providing limited and inconsistent gains over simple fine-tuning.
- We introduce **Adapter Routing and Consolidation (ARC)** as a simple parameter-isolation method which directly limits cross-task interference while selectively consolidating compatible adaptations. Despite it's simplicity, we find that ARC consistently outperforms the evaluated CL methods across all SCOPE variants.

2 Related work

Continual Retrieval Benchmarks. Historically, CL for VLMs has primarily relied on CIL benchmarks derived from standard classification datasets, which are inherently classification-oriented and therefore poorly capture retrieval-specific challenges. Several works have introduced evaluation settings for continual retrieval (Wang et al., 2021a; Ni et al., 2023; Cui et al., 2024), but each captures only a subset of the challenges arising in continual multimodal retrieval. Wang et al. (2021a) consider three sequential tasks restricted to natural images, limiting both domain diversity and sequence complexity. Ni et al. (2023) study a streaming setting spanning natural and e-commerce images, but do not consider structured task curricula or cross-domain robustness. Cui et al. (2024) broaden domain coverage, yet similarly omit structured task ordering and OOD evaluation. TiC-CLIP (Garg et al., 2024) evaluates retrieval over temporal data streams, but targets large-scale continual pretraining rather than lightweight, reproducible downstream adaptation; similarly, FOMO-in-FLUX (Udandarao et al., 2024) primarily studies continual pretraining dynamics. More recently, C-CLIP (Liu et al., 2025) proposed a dedicated framework for continual retrieval, but its use of closed-source implementation limits reproducibility and broader adoption. Taken together, existing benchmarks cover important aspects of continual retrieval, but no single evaluation setting combines diverse domains, structured task sequences, OOD evaluation, and practical reproducibility. This motivates a comprehensive evaluation framework, which we introduce in Section 3.

CL for VLMs. Most work on CL for VLMs focuses on classification, evaluating class-incremental accuracy and the preservation of zero-shot capabilities. Popular approaches mitigate forgetting through distillation from the pre-trained model (Zheng et al., 2023), prompt-based adaptation (Wang et al., 2021b; Lu et al., 2025; Li et al., 2025), probabilistic fine-tuning (Jha et al., 2024), representation adjustment and parameter fusion (Huang et al., 2024), or mixture-of-experts (Yu et al., 2024). However, continual retrieval poses a distinct challenge: rather than preserving discrete decision boundaries, it requires maintaining a coherent shared embedding space and the relative similarities between queries and candidates across tasks (Wang et al., 2021a). Retrieval-specific methods address this challenge through similarity preservation (Mod-X (Ni et al., 2023)), knowledge rectification (DKR (Cui et al., 2024)), and contrastive knowledge consolidation (C-CLIP (Liu et al., 2025)); however, these approaches update a shared model and primarily rely on regularisation or distillation to control representation drift, leaving successive tasks able to interfere with earlier representations. Parameter-efficient adaptation offers an alternative by isolating task-specific updates through LoRA (Hu et al., 2022), adapters (Pfeiffer et al., 2020), or expert routing (Yu et al., 2024), while model-merging approaches consolidate independently learned adaptations through task-vector arithmetic (Ilharco et al., 2023; Marczak et al., 2024) or low-rank core-space merging (Panariello et al., 2025). Beyond task-incremental adaptation, FOMO-in-Flux (Udandarao et al., 2024) studies continual multimodal pretraining under evolving data distributions. While these works highlight a range of strategies for continual adaptation, their effectiveness for retrieval has not been systematically examined across different training protocols and settings.

3 A Principled Framework for Continual Retrieval Evaluation

3.1 Problem Setup: Continual Multimodal Retrieval

Multimodal Dual-Encoder Models. We assess cross-modal similarity between image and text pairs using a dual-encoder model consisting of an image encoder $f(\cdot;\theta)$ and a text encoder $g(\cdot;\phi)$. Given an image $\mathbf{v}$ and text $\mathbf{c}$, the model produces embeddings $\mathbf{z} = f(\mathbf{v};\theta)$ and $\mathbf{h} = g(\mathbf{c};\phi)$, and their alignment is measured via cosine similarity $s(\mathbf{v},\mathbf{c}) = \text{sim}(\mathbf{z},\mathbf{h})$.

Cross-Modal Retrieval. Given a query set $\mathcal{Q} = \{q_i\}_{i=1}^N$ and a gallery $\mathcal{R} = \{r_j\}_{j=1}^N$ with unique paired annotations, the retrieval system ranks gallery items by similarity to each query. We evaluate performance using standard Recall@K (R@K) in both Image-to-Text (I2T) and Text-to-Image (T2I) directions (formal definitions in Section F).

Continual Learning Setup. We consider a sequence of T tasks, with $t = 0$ denoting the pretrained initialization. At each task t, the encoders are updated to θ^t and ϕ^t, yielding $f^t(\mathbf{v}) = f(\mathbf{v};\theta^t)$ and $g^t(\mathbf{c}) = g(\mathbf{c};\phi^t)$. After each task, the model is evaluated on all previously seen tasks. While we track performance throughout training, our primary metric is the Average Final R@K (formally

Table 1: Comparison of existing retrieval benchmarks for CL across the five SCOPE design principles (see Section G for further discussion).

Benchmark	**S**pecificity	**C**urriculum	**O**OD	**P**racticality	**E**xtensive Diversity
Wang et al. (2021a) (2021)	✓	✗	✗	✓	✗
Ni et al. (2023) (2023)	✗	✗	✗	✓	✗
Cui et al. (2024) (2024)	✓	✗	✗	✓	✓
Garg et al. (2024) (2024)	✗	✓	✓	✗	✓
Liu et al. (2025) (2025)	✗	✗	✓	✗	✓
SCOPE	✓	✓	✓	✓	✓

defined in Section F), which measures per-task retention after the full training sequence is complete, reported separately for I2T and T2I directions.

3.2 Framework Design Principles

We identify five key requirements for a principled continual retrieval evaluation framework: semantic specificity, a difficulty-calibrated curriculum, OOD robustness, practical reproducibility, and extensive domain diversity. Existing training and evaluation protocols address these requirements only partially, leaving no unified setup that captures all five. To bridge this gap, we introduce **SCOPE** (**S**pecificity, **C**urriculum, **O**OD robustness, **P**ractical reproducibility, and **E**xtensive domain diversity), an evaluation framework built around these design principles to enable rigorous and reproducible assessment of continual multimodal retrieval.

Semantic Specificity. Many common datasets (e.g., Oxford Pets (Parkhi et al., 2012) used by Liu et al. (2025)) suffer from pair redundancy, where multiple images might match a generic caption or vice versa. Therefore, we select datasets with high semantic precision and distinct targets to ensure cross-modal retrieval metrics (I2T and T2I) provide reliable performance signals.

Curriculum. Since task ordering heavily impacts CL performance (Masana et al., 2020), we use the frozen backbone's zero-shot performance as a proxy for task complexity, inspired by class-incremental works such as Menabue et al. (2024). This creates a structured progression from in-distribution to challenging OOD domains, allowing us to evaluate plasticity degradation while providing a modular way to integrate future tasks.

OOD Robustness. Robust methods should leverage sequentially acquired knowledge to enhance generalization to datasets outside the training sequences. Therefore, we evaluate the model's capacity for OOD retrieval on commonly used held-out, broad-domain benchmarks such as **COCO** (Lin et al., 2014)), a standard proxy for general cross-modal transfer in the retrieval literature (Radford et al., 2021; Garg et al., 2024). To provide a more complete method assessment, we also compute classification accuracy on common datasets such as **DomainNet** (Peng et al., 2019), **EuroSAT** (Helber et al., 2019), **CIFAR100** (Krizhevsky & Hinton, 2009), and **ImageNet** (Deng et al., 2009).

Practical Reproducibility. SCOPE is designed around practical fine-tuning scenarios while remaining computationally tractable for the broader research community. We use publicly available datasets and release the complete codebase[1], evaluation protocols, and hyperparameter configurations, enabling standardized and reproducible comparisons.

Extensive Domain Diversity. To minimize overlap with pre-training data, we propose a sequence of heterogeneous domains: natural images (**Flickr30K** (Young et al., 2014)), AI-generated content (**Lexica-SD** (yuwan0, 2024), **KreaM** (hahminlew, 2023)), artwork (**WikiArt** (AterMors, 2024)), specialized distributions such as cartoons (**Flintstones** (Kapuriya & Buitelaar, 2025; Kapuriya, 2025)), sketches (**Sketch** (Chowdhury et al., 2022; zoheb, 2025)), and medical imaging (**ROCOv2** (Rückert et al., 2024)). This diversity introduces substantial distribution shifts and helps prevent overly optimistic performance assessments.

[1] https://anonymous.4open.science/r/scope-framework-4608

3.3 Training and Evaluation Protocols

We define three complementary training protocols that expose different forms of distribution shift, interference, and knowledge reuse, as illustrated in Figure 2. The protocols differ in how the available datasets are partitioned into tasks. In the **separate** protocol, the setting most commonly investigated in the literature, each full dataset constitutes a single task, such that task boundaries coincide with dataset boundaries and distribution shifts between consecutive tasks are explicit. In the **overlapping** protocol, each dataset is partitioned into k parts, and each task is constructed by combining one such part with data from another dataset. This creates controlled overlap between domains across task boundaries and allows us to examine whether CL methods can exploit shared distributions across tasks (see Section D.2). Finally, in the **long sequence** protocol, each dataset is similarly partitioned into k parts, but each part constitutes a separate task. This increases the number of sequential updates and allows the same domain to recur later in the stream, stressing long-term stability and knowledge reuse (see Section D.3).

We further consider two complementary gallery settings for evaluation, illustrated in Figure 2. The **task-specific gallery** evaluates each task independently against same-task candidates; since it is commonly used in the literature, this serves us as the default across all three protocols. Additionally, we introduce a **global gallery** evaluation, in which candidates from all tasks are pooled into a single retrieval space. This directly tests whether sequential adaptation preserves a coherent embedding space, as cross-task representation drift that remains hidden under task-specific evaluation can become visible once examples from different domains compete directly.

3.4 ARC: A Simple Baseline for Continual Retrieval

As illustrated in Figure 1, continual adaptation can alter fine-grained retrieval rankings even when the broader embedding structure remains largely preserved. This motivates isolating adaptations for distinct data distributions while still allowing related knowledge to be shared. We propose **Adapter Routing and Consolidation (ARC)**, which (i) partitions each incoming task into K regions in the frozen backbone space, (ii) learns an independent low-rank expert for each region, and (iii) selectively consolidates compatible experts across tasks.

Intra-Task Expert Construction. We start from a frozen pretrained multimodal model with image encoder $f(\cdot;\theta^0)$ and text encoder $g(\cdot;\phi^0)$. For each incoming task t, we first encode all training pairs with the frozen backbone. For pair $(\mathbf{v}_i, \mathbf{c}_i)$, let $\mathbf{z}_i$ and $\mathbf{h}_i$ denote the normalized image and text embeddings. We cluster the joint representations

$$\mathbf{u}_i = \mathrm{norm}\left(\frac{\mathbf{z}_i + \mathbf{h}_i}{2}\right) \tag{1}$$

using spherical K-means, yielding K clusters within the task.

Each cluster e defines an expert consisting of an image prototype $\mathbf{p}^e_{\mathbf{v}}$, a text prototype $\mathbf{p}^e_{\mathbf{c}}$, and an independent pair of LoRA adapters $(\Delta\theta^e, \Delta\phi^e)$. The prototypes are the normalized mean image and text embeddings of samples assigned to the cluster. The corresponding expert is

$$f^e(\mathbf{v}) = f(\mathbf{v};\theta^0 + \Delta\theta^e), \qquad g^e(\mathbf{c}) = g(\mathbf{c};\phi^0 + \Delta\phi^e). \tag{2}$$

Samples are routed to their closest within-task prototype and each expert is trained only on its assigned samples. Thus, a task may contain multiple specialized adaptations rather than being represented by a single adapter.

Prototype-Based Routing. At inference time, all retained experts from previously observed tasks form a shared expert pool. For an image-text candidate pair $(\mathbf{v}_i, \mathbf{c}_j)$, we compute its compatibility with expert e using frozen-backbone representations:

$$S^e_{ij} = \frac{1}{2}\left[\mathrm{sim}(\mathbf{z}_i, \mathbf{p}^e_{\mathbf{v}}) + \mathrm{sim}(\mathbf{h}_j, \mathbf{p}^e_{\mathbf{c}})\right], \qquad e^*_{ij} = \arg\max_e S^e_{ij}. \tag{3}$$

Both elements of the pair are then encoded with expert e^*_{ij}, and their cosine similarity defines the retrieval score. Since routing uses the frozen backbone and stored prototypes, it does not require a task identifier at inference time.

Prototype-Guided Consolidation. Keeping every cluster expert independently would prevent interference but cause the expert pool to grow continually. After learning task t, we therefore compare each newly created expert e with experts retained from previous tasks. Their compatibility is

$$C(e, e') = \frac{1}{2}\left[\mathrm{sim}(\mathbf{p}_{\mathbf{v}}^{e}, \mathbf{p}_{\mathbf{v}}^{e'}) + \mathrm{sim}(\mathbf{p}_{\mathbf{c}}^{e}, \mathbf{p}_{\mathbf{c}}^{e'})\right]. \tag{4}$$

Let $e' = \arg\max_{\bar{e}} C(e, \bar{e})$. If $C(e, e') > \beta$, we consolidate the new expert into e' using CoreSpace (Panariello et al., 2025); otherwise, it is retained as a separate expert. Consolidation is performed only after the current task has been learned, so newly created experts cannot interfere during acquisition. ARC therefore follows an **isolate-then-consolidate** strategy: intra-task clustering captures heterogeneous modes within each task, while cross-task consolidation controls growth and promotes reuse between compatible modes.

4 Experiments

We use the framework described in Section 3 to evaluate **ARC** (Section 3.4) alongside a diverse set of CL approaches for VLMs, spanning simple baselines, standard CL approaches, and retrieval-oriented methods. We consider **Zero-shot (ZS)** performance of the pretrained backbone and **Fine-tuning (FT)**, where the model is sequentially updated without an explicit mechanism for mitigating forgetting. We further include standard CL approaches applicable to retrieval models: **EWC** (Kirkpatrick et al., 2017), **L2P** (Wang et al., 2021b), and **Task Arithmetic (TA)** (Ilharco et al., 2023). As retrieval-specific methods, we evaluate **Mod-X** (Ni et al., 2023), **DKR** (Cui et al., 2024), and **C-CLIP** (Liu et al., 2025). We additionally include **MoE** (Yu et al., 2024), a parameter-isolation approach originally developed for continual classification, which we adapt to the retrieval setting.

Experimental Setup We study the impact of training and evaluation protocol design through two complementary analyses. First, fixing task-specific evaluation, we compare the separate-dataset, overlapping-domain, and long-sequence protocols to examine sensitivity to the structure of the continual training stream. Second, using the separate-dataset and overlapping-domain protocols, we compare task-specific and global-gallery evaluation to isolate the effect of gallery composition. Unless explicitly stated otherwise, all methods use CLIP ViT-B/16 as the initial backbone and adapt it according to their respective CL strategies. We report Average Final Recall@1 (as in Equation (11)) to measure cross-modal retrieval performance. All methods are trained using AdamW for 20 epochs with a batch size of 256. For ARC, we keep the pretrained backbone fixed and learn task-specific LoRA adapters inserted into both the visual and text encoders, together with the routing and consolidation procedure described in Section 3.4. Given the breadth of our evaluation across methods, training protocols, datasets, and gallery settings, we run each configuration with a single random seed and report the resulting performance. We provide implementation details in Section A and share the exact method-specific hyperparameters in the open-source code.

4.1 A single training protocol is not enough to assess CL methods in retrieval

CL methods may behave differently under different task partitions and degrees of distribution shift. Therefore, we first examine whether the relative performance of CL methods remains consistent across the continual training protocols introduced in Section 3.3. Concretely, we train and evaluate the methods across three protocols: separate-dataset, overlapping-domain, and long-sequence. We present the results in Table 2, which consists of average T2I and I2T R@1 at the end of the training on the full task sequence, alongisde the ranking of each method. We provide the full per-task results for the three protocols in Sections D.1 to D.3.

While specific CL methods perform well under particular protocols, their relative performance varies considerably across settings, suggesting that different methods may favour different CL scenarios. Notably, more sophisticated CL strategies do not consistently outperform simple fine-tuning across the evaluated protocols. This highlights how evaluation under only the commonly used separate-dataset protocol may provide an incomplete picture of a method's effectiveness. In contrast, our proposed method, ARC, performs consistently well across the evaluated protocols. We hypothesize that this robustness stems from isolating task-specific adaptations before selectively consolidating compatible ones, thereby limiting interference between dissimilar tasks while still allowing knowledge reuse across related distributions; we investigate this mechanism further in Section 4.4. Overall,

Table 2: Final Recall@1 across training protocols with **separate galleries** evaluation protocol. We report average I2T and T2I retrieval performance for each protocol, alongside intra-protocol ranks based on the average of I2T and T2I. While commonly used methods from literature might perform well on specific settings, their overall ranking across diverse training protocols exposed in SCOPE varies, and our proposed ARC consistently outperforms the evaluated methods.

	Separate			Overlap			Long			
Method	**I2T R@1**	**T2I R@1**	**Rank**	**I2T R@1**	**T2I R@1**	**Rank**	**I2T R@1**	**T2I R@1**	**Rank**	**Avg. Rank (↓)**
FT	38.7	36.2	5	39.6	42.1	4	44.6	43.6	5	4.7
TA	36.0	34.1	7	34.0	34.8	7	37.4	36.3	7	7.0
EWC	40.2	38.4	3	40.2	42.2	2	46.6	46.3	2	2.3
Mod-X	38.5	36.8	4	39.5	40.8	5	46.7	45.9	3	4.0
C-CLIP	35.9	34.3	6	35.0	36.1	6	41.3	41.0	6	6.0
L2P	27.4	27.2	9	25.7	30.1	9	32.1	33.3	8	8.7
DKR	30.1	30.9	8	31.8	33.9	8	30.6	33.1	9	8.3
MoE	42.2	39.9	2	40.4	41.6	3	44.9	43.6	4	3.0
ARC	**46.5**	**45.6**	**1**	**45.3**	**46.6**	**1**	**50.2**	**49.6**	**1**	**1.0**

Table 3: Average final Recall@1 under **task-specific and global-gallery evaluation**. For each setting, methods are ranked separately by their average Text→Image and Image→Text R@1, and the two ranks are averaged; ties receive the mean of the tied positions. Evaluation with a task-specific gallery provides an incomplete view of CL methods: approaches that perform well in this setting may not retain their relative performance under global-gallery evaluation.

	Separate Protocol						Overlapping Protocol ($n = 2$)					
	Task-Specific			Global Gallery			Task-Specific			Global Gallery		
Method	**T→I**	**I→T**	**Rank**	**T→I**	**I→T**	**Rank**	**T→I**	**I→T**	**Rank**	**T→I**	**I→T**	**Rank**
ZS	25.8	27.4	9.8	25.0	27.3	6.5	26.9	24.7	10.0	17.9	17.0	10.0
FT	36.2	38.7	4.5	29.7	28.4	3.0	42.1	39.6	3.5	29.5	28.8	2.5
TA	34.1	36.0	6.5	28.2	26.0	7.5	34.8	34.0	7.0	23.7	24.7	7.0
EWC	38.4	40.2	3.0	28.5	26.3	5.3	42.2	40.2	2.5	28.1	26.3	4.0
Mod-X	36.8	38.5	4.5	28.1	26.3	7.3	40.8	39.5	5.0	29.1	29.3	2.5
C-CLIP	34.3	35.9	6.5	28.2	26.4	5.8	36.1	35.0	6.0	25.6	26.1	5.0
L2P	27.2	27.4	9.3	24.1	26.0	9.3	30.1	25.7	9.0	20.7	17.1	9.0
DKR	30.9	30.1	8.0	30.1	29.1	2.0	33.9	31.8	8.0	24.8	23.5	7.5
MoE	39.9	42.2	2.0	28.4	25.1	7.5	41.6	40.4	3.0	24.9	24.6	6.5
ARC	**45.6**	**46.5**	**1.0**	**37.0**	**33.7**	**1.0**	**46.6**	**45.3**	**1.0**	**34.2**	**33.6**	**1.0**

these results suggest that sensitivity to the training protocol is an important dimension that may be overlooked when evaluating methods under a single CL setup.

4.2 Task-specific galleries can mask cross-task retrieval failures

We next examine whether the gallery construction used at evaluation time affects the conclusions drawn about continual retrieval methods. Common task-specific evaluation restricts each query to candidates from the same task. While this measures within-task retrieval quality, it removes competition from examples encountered in other tasks and can therefore hide cross-task interference. In contrast, a global gallery pools candidates from all encountered tasks, requiring the model to maintain a coherent retrieval space across the entire continual sequence. We compare these two evaluation settings under both the separate and overlapping training protocols in Table 3.

The change in gallery composition substantially alters the relative performance of the evaluated methods. Therefore, similarly to the training protocols evaluated earlier, the global gallery not only makes retrieval harder, but also changes which methods appear effective. These rank reversals expose a limitation of task-specific evaluation, where a method can perform well when distinguishing candidates within each task while failing to preserve the relative similarities required for ranking examples across tasks. Such failures remain largely invisible when the gallery is partitioned by task, but become apparent once candidates from all encountered tasks compete within a single ranking. Consequently, task-specific evaluation alone can give an incomplete picture of continual retrieval performance and obscure differences in how well methods control cross-task interference. ARC re-

mains the best-performing method under both task-specific and global-gallery evaluation. Its strong global-gallery performance indicates that pair-dependent expert routing produces scores that remain sufficiently comparable across tasks for effective global retrieval.

4.3 Results transferability across backbones

In Figure 3, we further examine whether the relative performance of continual retrieval methods is sensitive to the choice of CLIP backbone. We observe noticeable changes in method rankings across ViT-B/32, ViT-B/16, and ViT-L/14. For example, MoE ranks second with both ViT-B backbones but drops to fifth with ViT-L/14, while the relative ordering of FT and Mod-X also changes across backbones. These results suggest that backbone choice constitutes an additional source of variation when comparing continual retrieval methods, complementing the sensitivity to training and evaluation protocols observed in our main experiments. Consequently, comparisons based on a single backbone may not fully characterize the relative behavior of different methods.

Figure 3: Final Recall@1 across different CLIP backbones under the separate-dataset protocol. Rank is computed from the average of T→I and I→T R@1 within each backbone.

Method	ViT-B/32			ViT-B/16			ViT-L/14		
	T→I	I→T	Rank	T→I	I→T	Rank	T→I	I→T	Rank
FT	33.4	31.1	3	36.2	38.7	4	44.0	45.5	3
Mod-X	32.9	30.6	4	36.8	38.5	3	44.0	45.8	2
C-CLIP	32.1	29.2	5	34.3	35.9	5	41.0	42.3	4
MoE	29.3	34.4	2	39.9	42.2	2	35.8	40.4	5
ARC	**41.1**	**41.7**	**1**	**45.6**	**46.5**	**1**	**50.3**	**48.7**	**1**

4.4 Understanding Continual Retrieval Performance

Additional metrics The preceding results demonstrate that CL methods can vary substantially depending on the training and evaluation protocols, as well as on the choice of the backbone. We next examine how retrieval performance relates to changes in the learned representation space. Beyond aggregate Recall@K, we report mean reciprocal rank (MRR), averaged over I2T and T2I retrieval, and characterize representation stability using neighbour overlap and cross-modal asymmetry.

In particular, the **hard-negative violation rate** measures how often the highest-scoring negative candidate scores at least as high as the correct I2T match under T2I retrieval. For a text query $\mathbf{c}_i$ with ground-truth image $\mathbf{v}_i$, let

$$s_i^- = \max_{j \neq i} s(\mathbf{c}_i, \mathbf{v}_j) \tag{5}$$

denote the hardest negative similarity. We compute the violation rate HNV as

$$\mathrm{HNV} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left[s_i^- \geq s(\mathbf{c}_i, \mathbf{v}_i)\right]. \tag{6}$$

Under the single-positive retrieval setting considered here, the hard-negative violation rate corresponds to the complement of T2I R@1 (up to ties). We use this formulation to directly relate top-1 retrieval failures to changes in local neighbourhood structure.

We complement this with **neighbour overlap**, which measures how many of a text query's top-K retrieved images are preserved between consecutive CL steps under text-to-image retrieval. Let $S_t(\mathbf{c}_i)$ denote the set of top-K retrieved image indices for text query $\mathbf{c}_i$ at step t. For each query, we then compute neighbour overlap NO as

$$\mathrm{NO}(\mathbf{c}_i) = \frac{|S_t(\mathbf{c}_i) \cap S_{t-1}(\mathbf{c}_i)|}{K}. \tag{7}$$

Together, violation rate and neighbour overlap allow us to examine how local neighbourhood stability relates to fine-grained retrieval outcomes. Under the single-positive retrieval setting considered here, the hard-negative violation rate corresponds to the complement of T2I R@1 (up to ties), and therefore captures whether the correct match is outranked by at least one negative. Combined with R@5, this allows us to contrast neighbourhood preservation, top-1 ordering, and broader retrieval quality: a model may retain similar local neighbourhoods while still allowing hard negatives to overtake the correct match, without necessarily displacing the correct match beyond the top ranks. Full definitions of all metrics are provided in Section F.

Table 4: Final retrieval and cross-modal alignment metrics for selected CL methods under the separate training protocol with task-specific galleries. R@1, R@5, MRR, and COCO R@1 are averaged over I2T and T2I; Acc. is the mean zero-shot accuracy over ImageNet, CIFAR100, EuroSAT, and DomainNet. Violation rate and neighbour overlap (k=10) are computed for T2I retrieval; neighbour overlap is undefined for ZS. Arrows indicate whether a higher (↑) or lower (↓) metric is desirable.

	Retrieval				Cls.	Local Alignment		Structure
Method	R@1 ↑	R@5 ↑	MRR ↑	COCO R@1 ↑	Acc. ↑	Violation Rate ↓	Asymmetry Gap ↓	Neighbour Overlap ↑
ZS	26.6	43.2	0.30	42.8	61.8	0.74	1.57	–
FT	37.5	57.8	0.40	51.3	62.3	0.64	2.51	0.60
C-CLIP	35.1	53.6	0.39	50.8	**63.5**	0.66	1.66	0.80
Mod-X	37.7	59.0	0.41	50.0	59.1	0.63	1.79	0.52
MoE	41.1	62.2	0.49	**52.2**	61.8	0.60	2.33	0.85
ARC	**46.0**	**68.5**	**0.52**	51.9	56.8	**0.55**	**0.97**	**0.88**

As shown in Figure 4, neighbour overlap does not closely track retrieval performance across methods. C-CLIP, for example, maintains relatively high neighbour overlap despite exhibiting more frequent hard-negative violations and weaker R@5 than MoE and ARC. Conversely, MoE and ARC combine similarly high neighbourhood preservation with substantially stronger retrieval performance. These results indicate that methods with comparable stability of their retrieved neighbourhoods can still differ considerably in the rankings within those neighbourhoods.

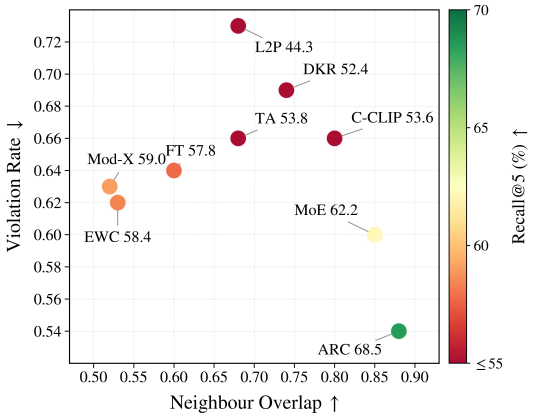


Figure 4: Relation between neighbourhood preservation, hard-negative violations, and retrieval performance (R@5).

Table 4 extends this analysis across a broader set of retrieval and representation-level metrics. ARC achieves the strongest continual retrieval performance among the evaluated methods, with the highest R@1, R@5, and MRR, while also exhibiting the highest neighbour overlap and lowest hard-negative violation rate. However, these gains do not uniformly extend to the other evaluated capabilities.

We additionally observe that stronger continual retrieval performance does not necessarily translate to better zero-shot generalization. Despite achieving the strongest retrieval performance on the continual tasks, ARC does not perform best on either held-out COCO retrieval or zero-shot classification. This highlights a further distinction between optimizing retrieval along the continual training sequence and preserving the broader capabilities of the pretrained model.

5 Conclusions

We introduce SCOPE, a principled framework for evaluating continual multimodal retrieval across diverse domains, training protocols, and gallery settings. Our results show that conclusions about CL methods depend strongly on the evaluation setting: method rankings vary across protocols, while simple fine-tuning remains competitive. We further show that preserving local neighbourhood structure is insufficient for reliable retrieval, motivating ARC, a simple parameter-isolation baseline that consistently outperforms the evaluated methods across SCOPE settings.

Limitations and future work. Given the breadth of our experimental evaluation, we use a single random seed per configuration, limiting our ability to assess run-to-run variability. Our evaluation also focuses primarily on CLIP models and staged CL. Future work should explore other VLM families, online and longer CL settings, and adaptive ARC consolidation.

AI USE STATEMENT

In this work, we used generative AI tools to assist with writing and code implementation. For writing, AI assistance was used to refine phrasing and improve clarity in parts of the manuscript. For code, AI tools were used to assist with implementation, adaptation, and general development tasks. All AI-assisted text was reviewed and edited by the authors, and all AI-assisted code was manually reviewed and tested for correctness. The authors take full responsibility for the final content of the work, including all text, code, results, and other artifacts produced with the assistance of AI.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide anonymous source code at `https://anonymous.4open.science/r/scope-framework-4608/README.md`. Hyperparameter configurations and experimental setup details are embedded in the scripts provided in the repository.

REFERENCES

AterMors. wikiart_recaption. https://huggingface.co/datasets/AterMors/wikiart_recaption, 2024. Hugging Face dataset.

Pinaki Nath Chowdhury, Aneeshan Sain, Ayan Kumar Bhunia, Tao Xiang, Yulia Gryaditskaya, and Yi-Zhe Song. Fs-coco: Towards understanding of freehand sketches of common objects in context. In *European conference on computer vision*, pp. 253–270. Springer, 2022.

Zhenyu Cui, Yuxin Peng, Xun Wang, Manyu Zhu, and Jiahuan Zhou. Continual vision-language retrieval via dynamic knowledge rectification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 11704–11712, 2024.

Jia Deng, R. Socher, Li Fei-Fei, Wei Dong, Kai Li, and Li-Jia Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, volume 00, pp. 248–255, 06 2009. doi: 10.1109/CVPR.2009.5206848. URL https://ieeexplore.ieee.org/abstract/document/5206848/.

Saurabh Garg, Mehrdad Farajtabar, Hadi Pouransari, Raviteja Vemulapalli, Sachin Mehta, Oncel Tuzel, Vaishaal Shankar, and Fartash Faghri. Tic-clip: Continual training of clip models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

hahminlew. kream-product-blip-captions. https://huggingface.co/datasets/hahminlew/kream-product-blip-captions, 2023. Hugging Face dataset.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217 – 2226, 2019. doi: 10.1109/JSTARS.2019.2918242. URL https://ieeexplore.ieee.org/document/8736785.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Linlan Huang, Xusheng Cao, Haori Lu, and Xialei Liu. Class-incremental learning with clip: Adaptive representation adjustment and parameter fusion. In *European Conference on Computer Vision*, pp. 214–231. Springer, 2024.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023.

Saurav Jha, Dong Gong, and Lina Yao. CLAP4CLIP: Continual learning with probabilistic finetuning for vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Janak Kapuriya. Flintstonessv_plus_plus. https://huggingface.co/datasets/Janak12/FlintstonesSV_Plus_Plus, 2025. Hugging Face dataset.

Janak Kapuriya and Paul Buitelaar. Flintstonessv++ : Improving story narration using visual scene graph. In *Text2Story@ECIR*, 2025. URL https://api.semanticscholar.org/CorpusID:279053465.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Yukun Li, Guansong Pang, Wei Suo, Chenchen Jing, Yuling Xi, Lingqiao Liu, Hao Chen, Guoqiang Liang, and Peng Wang. Coleclip: Open-domain continual learning via joint task prompt and vocabulary learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(8): 15137–15151, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Wenzhuo Liu, Fei Zhu, Longhui Wei, and Qi Tian. C-clip: Multimodal continual learning for vision-language model. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, pp. 46461–46477, 2025.

Haodong Lu, Xinyu Zhang, Kristen Moore, Jason Xue, Lina Yao, Anton van den Hengel, and Dong Gong. Continual learning on CLIP via incremental prompt tuning with intrinsic textual anchors. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Daniel Marczak, Bartlomiej Twardowski, Tomasz Trzcinski, and Sebastian Cygert. MAGMAX: leveraging model merging for seamless continual learning. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXV*, Lecture Notes in Computer Science, pp. 379–395. Springer, 2024.

Marc Masana, Bartlomiej Twardowski, and Joost van de Weijer. On class orderings for incremental learning. *CoRR*, abs/2007.02145, 2020.

Martin Menabue, Emanuele Frascaroli, Matteo Boschini, Enver Sangineto, Lorenzo Bonicelli, Angelo Porrello, and Simone Calderara. Semantic residual prompts for continual learning. In *European Conference on Computer Vision*, 2024.

Zixuan Ni, Longhui Wei, Siliang Tang, Yueting Zhuang, and Qi Tian. Continual vision-language representation learning with off-diagonal information. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, 2023.

Aniello Panariello, Daniel Marczak, Simone Magistri, Angelo Porrello, Bartłomiej Twardowski, Andrew D. Bagdanov, Simone Calderara, and Joost van de Weijer. Accurate and efficient low-rank model merging in core space. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012.

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 (ICCV)*, October 2019.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. AdapterHub: A framework for adapting transformers. In Qun Liu and David Schlangen (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 46–54, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.7. URL https://aclanthology.org/2020.emnlp-demos.7/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Johannes Rückert, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Cynthia S. Schmidt, Sven Koitka, Obioma Pelka, Asma Ben Abacha, Alba G. Seco de Herrera, Henning Müller, Peter A. Horn, Felix Nensa, and Christoph M. Friedrich. Rocov2: Radiology objects in context version 2, an updated multimodal image dataset. *Scientific Data*, 11(1), 2024. ISSN 2052-4463.

Vishaal Udandarao, Karsten Roth, Sebastian Dziadzio, Ameya Prabhu, Mehdi Cherti, Oriol Vinyals, Olivier Hénaff, Samuel Albanie, Zeynep Akata, and Matthias Bethge. A practitioner's guide to real-world continual multimodal pretraining. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 133801–133845. Curran Associates, Inc., 2024.

Kai Wang, Luis Herranz, and Joost van de Weijer. Continual learning in cross-modal retrieval . In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE Computer Society, 2021a.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer G. Dy, and Tomas Pfister. Learning to prompt for continual learning. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 139–149, 2021b.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2, 2014.

Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23219–23230, June 2024.

yuwan0. lexica-stable-diffusion-v1-5. `https://huggingface.co/datasets/yuwan0/lexica-stable-diffusion-v1-5`, 2024. Hugging Face dataset.

Zangwei Zheng, Mingyuan Ma, Kai Wang, Ziheng Qin, Xiangyu Yue, and Yang You. Preventing zero-shot transfer degradation in continual learning of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19125–19136, October 2023.

zoheb. sketch-scene. `https://huggingface.co/datasets/zoheb/sketch-scene`, 2025. Hugging Face dataset.

A Implementation details

Global training protocol. Unless otherwise stated, all methods are trained with the same optimization, data, and evaluation protocol. We use CLIP ViT-B/16 initialized from the pretrained checkpoint, train for 20 epochs per task, and use a training batch size of 256. Validation and zero-shot evaluation use batch size 512. We use 32 data-loader workers and a maximum text length of 77 tokens. Retrieval evaluation is performed on the same task sequence and validation datasets for all methods.

Table 5: Global hyperparameters shared by all methods.

Hyperparameter	Value
Backbone	CLIP ViT-B/16
Epochs per task	20
Training batch size	256
Validation batch size	512
Zero-shot batch size	512
Max text length	77
Number of workers	32
Seed	42

B ARC Method Details

Training procedure. Figure 5 provides an overview of ARC training. At the beginning of each task t, we keep the pre-trained CLIP backbone frozen and partition the current task into K regions in the frozen representation space. Specifically, all image-text training pairs are encoded using the frozen backbone, their joint representations are clustered using spherical K-means, and each resulting cluster defines a separate region of the task.

For each cluster, we initialize a new set of LoRA parameters for both the image and text encoders. Each adapter is trained only on the examples assigned to its corresponding cluster, while the shared backbone parameters remain unchanged. This isolates cluster-specific updates from adaptations learned previously and prevents new training from directly overwriting previously acquired parameters. Thus, when $K > 1$, a single task may give rise to multiple specialized experts.

After training, we summarize each cluster using separate image and text prototypes. Specifically, the examples assigned to a cluster are encoded using the frozen backbone, independently of the newly trained adapter. We ℓ_2-normalize the resulting embeddings and average them separately for the image and text modalities, yielding image and text prototypes $\mathbf{p}_{\mathbf{v}}^{e}$ and $\mathbf{p}_{\mathbf{c}}^{e}$ for expert e. Because all prototypes are computed using the same frozen backbone, prototypes created at different stages of continual training remain directly comparable.

Each newly trained cluster expert is then compared with the existing expert pool. For every existing expert e', we compute the average cosine similarity between the corresponding image and text prototypes, as defined in Section 3.4. If the similarity to the closest existing expert exceeds the consolidation threshold β, the new LoRA update is consolidated with that expert using CoreSpace (Panariello et al., 2025). Otherwise, the new adapter is retained as a separate expert. Its associated image and text prototypes are stored in the prototype memory and subsequently used for routing.

Consequently, each task is first decomposed into K independently adapted regions, and consolidation is performed only after the corresponding experts have been trained. Previous experts are never modified by gradient updates from a new task.

Inference and adapter routing. Figure 6 illustrates inference with the resulting expert pool. Routing is performed in the frozen CLIP representation space and does not require task identity. When $K > 1$, the expert pool may contain multiple experts originating from different regions of the same task, in addition to experts retained or consolidated across different tasks.

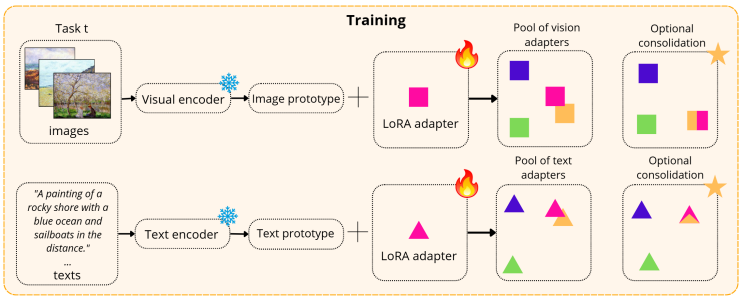


Figure 5: Overview of the ARC training procedure, illustrated for the special case $K = 1$ for visual clarity. In the general case $K > 1$, the incoming task is first partitioned into K clusters in the frozen-backbone representation space, and an independent image–text LoRA adaptation and corresponding prototype pair are learned for each cluster. Each newly learned expert is then either retained or consolidated with the most similar existing expert when their prototype similarity exceeds the consolidation threshold. The corresponding prototypes are stored for task-agnostic routing at inference time.

Given an image–text candidate pair $(\mathbf{v}_i, \mathbf{c}_j)$, we first compute their normalized frozen-backbone representations. For each available expert e, we compute the cosine similarity of the image representation to the expert's image prototype and of the text representation to its text prototype. We average these two similarities to obtain a pair-specific routing score,

$$S_{ij}^{e} = \frac{1}{2}\left[\mathrm{sim}(\mathbf{z}_i, \mathbf{p}_{\mathbf{v}}^{e}) + \mathrm{sim}(\mathbf{h}_j, \mathbf{p}_{\mathbf{c}}^{e})\right]. \tag{8}$$

We select the expert with the highest routing score,

$$e_{ij}^{*} = \arg\max_{e} S_{ij}^{e}, \tag{9}$$

and use this expert to encode both elements of the candidate pair. The retrieval score for $(\mathbf{v}_i, \mathbf{c}_j)$ is then given by the cosine similarity between the resulting expert-specific image and text embeddings. Consequently, different candidate pairs within the same retrieval gallery may be evaluated using different experts.

Routing is therefore task-agnostic: no task identifier is provided at inference time, and expert selection depends only on similarity in the common frozen-backbone representation space. Moreover, routing remains well-defined after consolidation because related adaptations share the corresponding consolidated expert.

For task-specific evaluation, queries are ranked only against candidates belonging to the corresponding evaluation task. For global-gallery evaluation, candidates from all tasks are pooled into a single gallery. Importantly, ARC does not assign a single adapter to the query and use it for all gallery candidates. Instead, routing is performed independently for every query–candidate pair. Thus, a query may be compared with different gallery candidates using different experts, depending on which expert obtains the highest pair-specific routing score. This allows cross-task candidates to compete directly without assuming either the task identity or a shared adapter assignment for the entire gallery.

B.1 Inference Costs

ARC introduces additional inference overhead compared with a standard dual encoder due to its pair-dependent expert routing. Routing itself is lightweight, requiring only similarity computations between frozen-backbone embeddings and stored prototypes. However, exact scoring may require evaluating a query under multiple experts. In practice, query–candidate pairs assigned to the same expert can be batched, while gallery embeddings can be precomputed for each retained expert, trading additional storage for online computation. The resulting cost therefore depends primarily on the number of experts retained after consolidation. More efficient query-level or approximate routing represents a promising direction for scaling ARC to very large galleries.

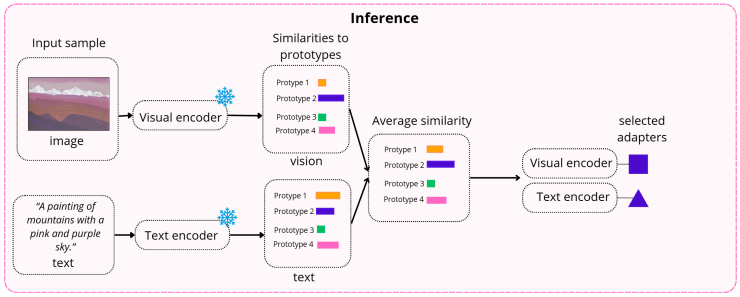


Figure 6: Overview of ARC inference, illustrated for the special case $K = 1$ for visual clarity. In the general case $K > 1$, each task may contribute multiple cluster-specific experts and prototype pairs to the expert pool. For each image–text candidate pair, frozen-backbone representations are compared with the stored prototype pairs to select the expert with the highest pair-specific routing score. The selected expert's LoRA parameters are then activated to compute the retrieval score. Expert selection does not require task identity at inference time and may differ across candidates for the same query.

C Compute Resources

All experiments were conducted on either a single NVIDIA A100 GPU or NVIDIA B200 GPU card. Training each method on the main continual retrieval benchmark with 7 tasks required approximately 4 GPU hours per run, while the extended CL sequences consisting of 14 tasks required around 8 GPU hours.

D Full Experimental Results

D.1 Separate Datasets Sequence

Table 6: Training protocol: **separate**. Evaluation protocol: **common continual learning**. Cross-modal **Recall@1** at the end of continual training on our proposed evaluation framework for CLIP ViT-B/16. Surprisingly, commonly used CL approaches often fail to improve upon fine-tuning baseline, which highlights how continual retrieval presents its own set of challenges that cannot be addressed simply by adapting CIL methods. Our novel method, **ARC**, shows robust performance and outperforms the previous approaches by a sizable margin.

	Text → Image								Image → Text							
	Flickr	**Lexica**	**WikiArt**	**KreaM**	**Flints**	**Sketch**	**ROCO**	**Avg.**	**Flickr**	**Lexica**	**WikiArt**	**KreaM**	**Flints**	**Sketch**	**ROCO**	**Avg.**
ZS	62.3	52.3	22.6	20.0	16.6	5.2	1.8	25.8	82.0	52.0	20.8	20.2	11.1	4.2	1.5	27.4
FT	73.5	63.0	37.1	26.7	38.3	8.3	6.5	36.2	88.5	64.8	38.5	28.4	35.4	8.5	6.9	38.7
TA	73.2	62.5	35.0	24.1	32.8	8.2	3.2	34.1	88.7	64.9	35.2	24.4	27.5	7.8	3.6	36.0
EWC	75.4	62.0	36.3	31.6	42.3	12.3	8.6	38.4	89.5	62.5	36.5	33.3	38.8	12.4	8.6	40.2
Mod-X	73.5	61.0	36.4	27.6	40.1	9.4	9.4	36.8	88.5	60.9	36.8	28.9	36.9	8.6	9.3	38.5
C-CLIP	73.3	61.9	34.3	25.1	32.6	8.9	3.7	34.3	88.3	65.7	34.1	26.1	26.7	7.2	3.3	35.9
L2P	66.3	51.5	24.3	18.4	20.7	6.0	2.9	27.2	83.2	51.1	20.6	14.9	15.5	4.2	2.5	27.4
DKR	62.9	45.0	24.7	24.5	36.6	9.9	12.5	30.9	78.0	38.1	20.4	22.1	33.1	7.1	12.2	30.1
MoE	79.22	63.18	43.62	32.69	43.79	7.80	8.95	39.9	92.10	64.65	45.84	34.21	42.49	7.25	9.05	42.2
ARC	79.92	76.31	48.44	41.76	47.99	13.20	11.32	**45.6**	93.10	74.07	49.03	40.45	45.08	12.10	11.87	**46.5**

D.2 Overlapping Domains Sequence

Extension to higher domain overlap. To illustrate the extensibility of the proposed evaluation framework, Table 7 also provides an example $n = 3$ curriculum, in which each continual task contains data from three domains. The experiments in this paper focus on the $n = 2$ setting; we do not report results for $n = 3$ due to computational constraints.

Table 7: Composition of the overlapping-domain CL curricula. Subscripts denote disjoint dataset partitions.

Task	$n = 2$	$n = 3$
1	Flickr_1 + Lexica_1	Flickr_1 + Lexica_1 + WikiArt_1
2	Lexica_2 + WikiArt_1	Lexica_2 + WikiArt_2 + KreaM_1
3	WikiArt_2 + KreaM_1	WikiArt_3 + KreaM_2 + Flints_1
4	KreaM_2 + Flints_1	KreaM_3 + Flints_2 + Sketch_1
5	Flints_2 + Sketch_1	Flints_3 + Sketch_2 + ROCO_1
6	Sketch_2 + ROCO_1	Sketch_3 + ROCO_2 + Flickr_2
7	ROCO_2 + Flickr_2	ROCO_3 + Flickr_3 + Lexica_3

Table 8: Cross-modal **Recall@1** at the end of continual training under the $n = 2$ **overlapping** evaluation setting for CLIP ViT-B/16. Each continual task contains data from two domains.

	Text → Image								**Image → Text**							
	Task 1	**Task 2**	**Task 3**	**Task 4**	**Task 5**	**Task 6**	**Task 7**	**Avg.**	**Task 1**	**Task 2**	**Task 3**	**Task 4**	**Task 5**	**Task 6**	**Task 7**	**Avg.**
ZS	60.0	30.6	27.6	26.9	16.6	3.6	23.0	26.9	64.4	30.3	26.5	23.6	10.7	3.3	13.9	24.7
FT	71.5	46.2	42.0	44.0	37.7	16.9	36.6	42.1	71.4	44.8	41.0	40.8	37.0	15.8	26.8	39.6
TA	67.8	44.0	37.1	31.9	28.7	6.1	27.7	34.8	71.9	44.8	38.2	31.2	26.9	6.9	18.0	34.0
EWC	71.9	47.6	42.5	43.7	37.7	16.1	36.1	42.2	73.6	46.4	42.6	41.2	36.3	15.4	25.9	40.2
MOD-X	72.6	48.5	41.7	40.7	34.8	13.4	33.9	40.8	75.0	48.5	42.6	39.3	34.5	12.9	24.1	39.5
C-CLIP	69.8	45.6	38.7	34.2	28.5	6.9	29.2	36.1	73.8	47.0	39.9	32.6	25.4	7.3	19.0	35.0
L2P	65.3	34.7	29.5	25.2	21.1	6.5	28.2	30.1	62.4	30.1	25.6	19.9	18.2	5.9	17.7	25.7
DKR	67.0	40.0	34.0	31.5	29.0	9.0	26.8	33.9	70.0	39.0	32.5	29.0	26.0	8.0	18.1	31.8
Moe	74.78	53.78	45.35	38.37	34.08	11.69	33.16	41.6	76.90	55.29	46.28	36.12	34.08	11.87	22.57	40.4
ARC	77.5	56.6	51.7	50.0	40.0	15.0	35.4	**46.6**	80.5	57.1	52.4	49.3	38.3	14.3	25.5	**45.3**

D.3 RECURRING DIFFICULTY SEQUENCE

Extension to longer sequences. To illustrate the extensibility of the long sequence protocol, Table 9 additionally defines a 3-cycle, 21-task sequence. The experiments in this paper focus on the 2-cycle, 14-task setting; we do not report results for the 3-cycle variant due to computational constraints.

Table 9: Task composition for the recurring-difficulty CL sequences. Each dataset is partitioned into two or three disjoint subsets, and the difficulty-ordered domain sequence is repeated across partitions. Subscripts denote the corresponding dataset partition.

Task	2-cycle sequence	3-cycle sequence
Task 1	Flickr_1	Flickr_1
Task 2	Lexica_1	Lexica_1
Task 3	WikiArt_1	WikiArt_1
Task 4	KreaM_1	KreaM_1
Task 5	Flints_1	Flints_1
Task 6	Sketch_1	Sketch_1
Task 7	ROCO_1	ROCO_1
Task 8	Flickr_2	Flickr_2
Task 9	Lexica_2	Lexica_2
Task 10	WikiArt_2	WikiArt_2
Task 11	KreaM_2	KreaM_2
Task 12	Flints_2	Flints_2
Task 13	Sketch_2	Sketch_2
Task 14	ROCO_2	ROCO_2
Task 15	–	Flickr_3
Task 16	–	Lexica_3
Task 17	–	WikiArt_3
Task 18	–	KreaM_3
Task 19	–	Flints_3
Task 20	–	Sketch_3
Task 21	–	ROCO_3

Table 10: Cross-modal **Recall@1** at the end of continual training on a 14-task recurring-difficulty sequence using CLIP ViT-B/16. Each dataset is divided into two disjoint partitions. The curriculum first traverses all seven domains using the first partition and then revisits the same domain order using the second partition: Flickr_1, Lexica_1, ..., ROCO_1, Flickr_2, Lexica_2, ..., ROCO_2. Averages are computed across all 14 subtasks.

Text → Image															
Method	**Flickr_1**	**Lexica_1**	**WikiArt_1**	**KreaM_1**	**Flints_1**	**Sketch_1**	**ROCO_1**	**Flickr_2**	**Lexica_2**	**WikiArt_2**	**KreaM_2**	**Flints_2**	**Sketch_2**	**ROCO_2**	**Avg.**
ZS	62.8	60.1	26.4	29.2	25.4	7.6	2.5	62.3	60.0	26.9	28.3	24.0	8.8	2.5	30.5
FT	74.7	72.6	46.7	36.9	52.3	11.6	9.1	75.6	74.3	47.1	35.4	51.2	14.3	8.8	43.6
TA	71.8	62.8	39.4	27.6	39.3	10.0	3.6	71.7	62.3	38.5	26.4	38.9	12.2	3.8	36.3
EWC	73.9	68.7	46.8	41.9	57.5	17.2	15.2	76.5	70.8	47.6	40.8	57.7	17.8	15.4	46.3
Mod-X	75.5	73.2	48.6	39.6	56.2	13.9	12.4	76.6	75.5	49.4	38.8	54.5	16.0	12.2	45.9
C-CLIP	74.7	70.6	43.6	32.8	46.9	11.7	5.9	75.2	71.4	44.2	32.1	46.1	13.0	6.1	41.0
L2P	67.6	62.3	31.6	26.9	32.1	7.6	4.0	68.9	62.9	32.7	25.9	30.1	9.9	4.0	33.3
DKR	57.7	46.6	24.5	28.8	44.5	12.8	15.7	59.2	46.3	25.4	28.5	46.1	12.2	15.4	33.1
MoE	77.2	64.9	48.9	38.8	55.3	9.3	10.6	78.5	65.2	49.6	37.9	53.5	10.0	10.3	43.6
ARC	79.2	80.6	54.7	45.0	59.7	16.7	11.0	80.0	79.4	54.9	45.7	58.7	17.9	10.6	**49.6**
Image → Text															
Method	**Flickr_1**	**Lexica_1**	**WikiArt_1**	**KreaM_1**	**Flints_1**	**Sketch_1**	**ROCO_1**	**Flickr_2**	**Lexica_2**	**WikiArt_2**	**KreaM_2**	**Flints_2**	**Sketch_2**	**ROCO_2**	**Avg.**
ZS	72.9	59.2	26.6	27.0	14.6	6.1	2.3	74.9	59.6	26.8	25.8	16.2	6.8	2.2	30.1
FT	81.7	73.6	49.2	37.1	48.5	12.9	9.3	82.9	72.5	49.4	36.5	48.5	12.3	9.8	44.6
TA	79.3	66.2	38.6	27.1	36.4	10.7	4.6	80.0	63.4	38.9	27.0	35.3	11.6	5.1	37.4
EWC	82.3	67.7	46.4	41.5	55.3	17.5	14.8	82.6	68.7	47.0	42.2	54.8	16.9	15.2	46.6
Mod-X	82.8	73.0	50.4	40.0	51.0	15.1	12.3	83.7	73.1	50.2	39.8	54.9	14.5	12.6	46.7
C-CLIP	82.2	71.5	44.2	34.1	41.0	10.6	6.1	82.1	72.0	44.5	33.8	39.1	10.5	6.5	41.3
L2P	76.3	61.2	29.6	23.0	22.6	6.9	3.2	79.6	60.5	30.2	22.5	23.1	7.6	3.2	32.1
DKR	64.9	39.6	20.0	25.8	38.0	10.9	14.6	65.3	39.3	20.8	26.1	39.1	10.0	14.5	30.6
MoE	84.9	67.4	51.3	39.4	51.7	8.1	9.8	87.1	68.6	51.8	39.6	51.0	8.7	9.9	44.9
ARC	85.9	80.4	55.6	46.0	58.2	15.6	10.5	87.2	79.1	55.5	46.1	56.5	15.6	10.6	**50.2**

D.4 RECALL@5 FOR THE MAIN TABLE

Table 11: Cross-modal retrieval performance on our proposed evaluation framework measured by **Recall@5** at the end of continual training for CLIP ViT-B/16.

	Text → Image								Image → Text							
Method	**Flickr**	**Lexica**	**WikiArt**	**KreaM**	**Flints**	**Sketch**	**ROCOv2**	**Avg.**	**Flickr**	**Lexica**	**WikiArt**	**KreaM**	**Flints**	**Sketch**	**ROCOv2**	**Avg.**
ZS	85.7	74.6	39.9	45.2	42.0	15.9	4.5	44.0	96.7	73.5	40.2	40.7	27.0	13.8	4.3	42.3
FT	92.7	83.1	61.8	52.8	75.1	22.8	15.4	57.7	97.3	82.9	63.2	53.9	69.6	22.2	16.4	57.9
EWC	90.6	76.6	55.0	56.3	77.6	27.6	25.2	58.4	96.9	74.0	53.6	56.4	74.6	27.4	24.9	58.3
Mod-X	92.4	80.7	61.0	55.1	76.7	25.2	20.5	58.8	97.3	80.2	61.6	55.9	73.8	25.4	20.5	59.2
C-CLIP	92.3	82.7	58.1	49.4	66.1	21.2	8.7	54.1	97.9	83.2	58.0	49.5	53.8	20.1	9.0	53.1
L2P	88.2	73.4	44.7	40.6	49.4	15.8	7.4	45.6	97.0	73.9	40.3	34.1	35.5	13.1	6.9	43.0
DKR	85.6	68.2	46.4	49.9	72.5	26.1	26.9	53.7	93.2	60.1	39.9	47.6	67.8	22.6	26.2	51.1
TA	92.4	83.2	59.1	48.6	65.5	21.1	8.7	54.1	97.5	83.2	59.5	47.5	57.1	20.0	9.6	53.5
MoE	94.9	83.2	69.5	62.0	82.3	20.3	19.8	61.7	99.1	83.8	71.3	63.8	80.7	20.0	20.1	62.7
ARC	95.7	91.0	75.1	74.4	85.8	34.6	24.9	**68.8**	99.4	90.5	75.8	73.3	82.6	29.8	25.3	**68.1**

D.5 ZERO-SHOT RETRIEVAL

Table 12: Zero-shot retrieval performance on COCO (Lin et al., 2014) during continual fine-tuning. We report Recall@1 for I2T and T2I, together with performance difference Δ.

Direction	Method	Task ID 0	1	2	3	4	5	6	7	Δ
I2T	Fine-tuning	52.4	59.4	59.3	59.7	60.4	60.2	60.0	61.0	+8.6
	EWC	52.4	62.7	62.8	61.7	61.5	60.7	60.1	59.9	+7.5
	Mod-X	52.4	61.4	61.9	60.5	60.9	60.2	59.6	59.0	+6.6
	C-CLIP	52.4	57.7	58.2	58.8	59.0	59.7	60.8	59.8	+7.4
	L2P	52.4	52.2	53.0	53.5	55.0	52.2	53.6	53.4	+1.0
	DKR	52.4	62.2	62.3	59.4	58.6	57.8	54.9	45.4	-6.9
	TA	52.4	57.4	57.7	58.5	58.9	59.6	59.4	59.7	+7.3
	MoE	52.4	61.2	61.3	61.0	61.1	61.2	61.1	60.8	+8.4
	ARC	52.4	62.3	62.3	62.1	62.2	62.1	60.9	60.9	+8.5
T2I	Fine-tuning	33.2	41.6	41.3	41.5	41.9	41.9	41.7	41.6	+8.4
	EWC	33.2	44.3	43.8	43.7	43.4	42.3	42.5	42.4	+9.2
	Mod-X	33.2	43.3	42.9	42.3	42.6	41.5	41.8	40.9	+7.7
	C-CLIP	33.2	38.9	39.3	40.9	41.0	41.6	41.9	41.7	+8.5
	L2P	33.2	37.1	36.9	37.8	37.7	37.8	37.4	37.1	+3.9
	DKR	33.1	44.2	43.1	41.7	40.7	38.4	37.2	31.4	-1.7
	TA	33.2	38.7	38.9	40.6	40.5	40.6	40.8	41.0	+7.8
	MoE	33.2	42.9	43.0	43.7	43.6	43.5	43.3	43.5	+10.3
	ARC	33.2	43.6	43.6	43.6	43.6	43.6	42.8	42.8	+9.6

D.6 ZERO-SHOT CLASSIFICATION

Table 13: Zero-shot accuracy on ImageNet during continual fine-tuning.

Method	0	1	2	3	4	5	6	7	Δ
Fine-tuning	68.1	68.8	68.8	67.4	67.8	65.9	65.6	66.5	-1.6
EWC	68.1	67.9	67.6	66.6	66.1	63.7	62.5	63.1	-5.0
Mod-X	68.1	68.6	68.3	66.2	66.4	63.2	62.7	63.2	-4.9
C-CLIP	68.1	69.2	69.1	68.9	68.7	67.9	67.7	67.6	**-0.5**
L2P	68.1	65.2	66.5	66.5	67.4	64.8	65.9	66.3	-1.8
DKR	68.1	68.0	66.5	63.7	62.2	56.6	53.9	48.2	-19.9
TA	68.1	69.0	68.9	68.5	68.5	67.6	67.3	67.0	-1.1
MoE	68.1	68.1	68.1	68.1	68.1	68.1	68.1	68.1	0.0
ARC	68.1	64.5	64.5	64.5	64.5	64.5	64.4	64.3	-3.8

Table 14: Zero-shot accuracy on CIFAR100 during continual fine-tuning.

Method	0	1	2	3	4	5	6	7	Δ
Fine-tuning	68.4	68.7	68.4	69.7	69.9	68.8	69.6	71.4	+3
EWC	68.4	68.2	68.5	69.2	67.7	66.9	65.4	68.3	-0.1
Mod-X	68.4	69.0	68.8	69.5	69.1	67.0	67.8	67.7	-0.7
C-CLIP	68.4	69.7	69.7	70.7	70.6	71.1	72.0	72.7	**+4.3**
L2P	68.4	65.9	66.8	66.6	68.0	63.5	66.8	64.4	-4.0
DKR	68.4	68.4	67.3	67.5	65.4	60.2	59.7	45.2	-23.2
TA	68.4	68.7	68.6	69.3	69.4	69.4	69.7	70.1	+1.7
MoE	68.4	68.4	68.4	68.4	68.4	68.4	68.4	68.4	0.0
ARC	68.4	66.1	66.1	66.1	66.1	66.1	63.5	63.5	-4.9

Table 15: Zero-shot accuracy on EuroSAT during continual fine-tuning.

Method	0	1	2	3	4	5	6	7	Δ
Fine-tuning	54.0	54.3	56.9	52.6	56.5	50.6	51.1	55.8	+1.8
EWC	54.0	51.4	56.1	53.1	56.2	47.4	47.7	55.0	+1.0
Mod-X	54.0	51.5	56.3	49.7	54.9	43.9	46.8	50.4	-3.6
C-CLIP	54.0	53.5	56.4	56.1	59.2	57.1	58.0	56.9	**+2.9**
L2P	54.0	48.3	51.4	52.1	59.3	53.7	52.5	54.9	+0.1
DKR	54.0	50.9	54.9	50.2	47.3	33.4	33.8	32.8	-21.2
TA	54.0	54.1	56.4	55.4	55.6	52.1	53.8	53.6	-0.4
MoE	54.0	53.9	53.9	53.9	53.9	53.9	53.9	53.9	-0.1
ARC	54.0	52.6	52.6	52.6	52.6	52.6	47.3	47.3	-6.7

Table 16: Zero-shot accuracy on DomainNet during continual fine-tuning.

Method	0	1	2	3	4	5	6	7	Δ
Fine-tuning	56.8	56.8	56.8	56.1	56.1	56.1	56.3	55.6	-1.2
EWC	56.8	56.1	55.6	55.5	55.0	55.0	55.0	54.3	-2.5
Mod-X	56.8	56.6	56.2	55.5	55.3	55.3	55.4	55.0	-1.8
C-CLIP	56.8	56.9	56.8	56.6	56.6	56.7	56.9	56.7	**-0.1**
L2P	56.8	56.1	56.5	56.3	56.8	55.7	56.4	54.6	-2.2
DKR	56.8	56.3	55.4	54.5	53.1	52.0	50.6	45.9	-10.9
TA	56.8	56.9	57.0	56.7	56.7	56.5	56.6	56.2	-0.6
MoE	56.8	56.8	56.8	56.8	56.8	56.8	56.8	56.8	0.0
ARC	56.8	54.7	54.7	54.7	54.7	54.7	52.1	52.0	-4.8

E ARC ABLATIONS

Consolidation threshold. Table 17 studies the effect of the consolidation threshold τ, which controls the trade-off between sharing and specialization. A lower threshold leads to more aggressive consolidation: $\tau = 0.7$ retains only 4 of the 14 task-specific experts, corresponding to a 71.4% reduction in the expert pool. Increasing the threshold to 0.8 and 0.9 retains 5 and 7 experts, respectively, while progressively improving retrieval performance. Thus, τ provides an explicit accuracy-compactness trade-off: aggressive consolidation substantially limits expert-pool growth, whereas more conservative consolidation preserves additional task-specific capacity and achieves higher retrieval accuracy.

Table 17: Ablation on the adapter merging threshold τ in ARC under the $n = 2$ overlapping-domain evaluation setting for CLIP ViT-B/16. We report the average Recall@1 across all seven continual tasks and the number of experts retained after consolidation. Without consolidation, the sequence results in 14 task-specific experts.

Merging threshold τ	0.70	0.80	0.90
T→I	41.6	44.0	**46.6**
I→T	39.8	42.8	**45.3**
Average	40.7	43.4	**46.0**
Retained experts	4	5	7
Pool reduction	71.4%	64.3%	50.0%

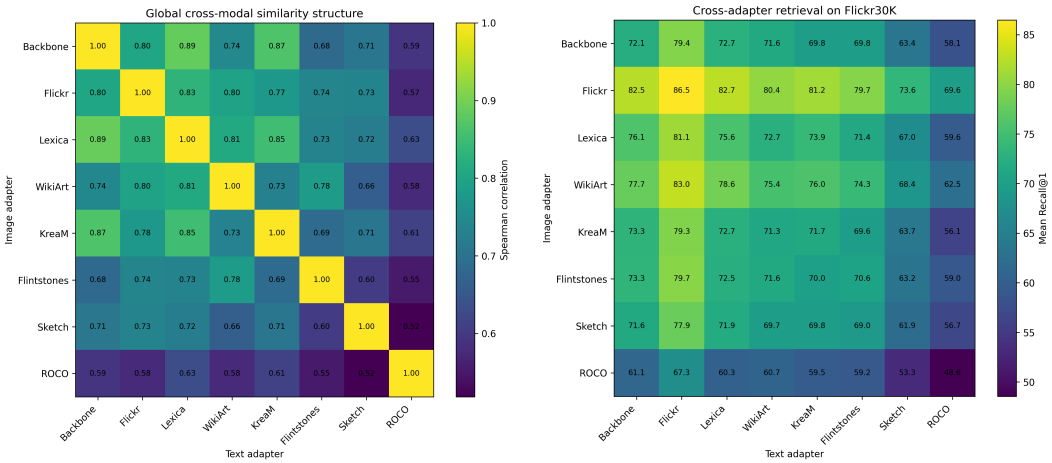
Number of adapters per task. Table 18 examines the effect of the number of adapters k introduced per task. Performance improves substantially from $k = 1$ to $k = 2$, with average R@1 increasing from 42.2 to 46.0. However, increasing k further does not yield additional gains and instead progressively reduces retrieval performance. We hypothesize that this degradation arises because increasing k partitions the available task data across more adapters, reducing the amount of

training data observed by each individual adapter. Thus, while a small amount of within-task specialization is beneficial, excessive specialization may leave individual adapters insufficiently trained. Based on this trade-off, we use $k = 2$ in our main experiments.

Table 18: Ablation on the number of adapters per task k in ARC under the $n = 2$ overlapping-domain evaluation setting for CLIP ViT-B/16. We report the average Recall@1 across the seven overlapping tasks at the end of continual training.

Adapters per task k	1	2	4	6	8
T→I	42.8	**46.6**	45.0	43.9	43.4
I→T	41.5	**45.3**	42.7	41.4	40.8
Average	42.2	**46.0**	43.9	42.7	42.1

E.1 ARC ANALYSIS



(a) Correlation of cross-modal similarity structures.

(b) Cross-adapter mean Recall@1.

Figure 7: **Cross-adapter compatibility on Flickr30K. Left:** Spearman correlation between similarity matrices. **Right:** retrieval using different image and text adapters.

F METRICS DEFINITIONS

Recall@K. Let $\text{rank}(q_i, r_i)$ denote the one-based rank of the ground-truth reference r_i among all gallery items sorted by decreasing similarity to query q_i. Recall@K is defined as:

$$\text{Recall@K} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[\text{rank}(q_i, r_i) \leq K]. \tag{10}$$

We apply this to both I2T retrieval, where images are queries and texts are the gallery ($\mathcal{Q} = \{\mathbf{v}_i\}_{i=1}^N$, $\mathcal{R} = \{\mathbf{c}_j\}_{j=1}^N$), and T2I retrieval, where texts are queries and images are the gallery ($\mathcal{Q} = \{\mathbf{c}_i\}_{i=1}^N$, $\mathcal{R} = \{\mathbf{v}_j\}_{j=1}^N$).

Average Final R@K. To evaluate overall retention at the end of the training sequence, we compute the Average Final R@K separately for each retrieval direction $m \in \{\text{I2T}, \text{T2I}\}$:

$$A_T^m = \frac{1}{T} \sum_{k=1}^{T} R_{T,k}^m, \tag{11}$$

where $R_{T,k}^m$ is the R@K on task k evaluated using the final encoders f^T and g^T, i.e., after training on all T tasks. In our tables, we report per-task results $R_{T,k}^m$ for each task k alongside their average A_T^m.

Mean Reciprocal Rank. Mean Reciprocal Rank (MRR) measures the inverse of the one-based rank of the correct match. For T2I retrieval, where text $\mathbf{c}_i$ is the query and $\mathbf{v}_i$ is the ground-truth image, this is:

$$\mathrm{MRR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\mathrm{rank}(\mathbf{c}_i, \mathbf{v}_i)},$$

where $\mathrm{rank}(\mathbf{c}_i, \mathbf{v}_i) \in \{1, \ldots, N\}$ is the one-based rank of $\mathbf{v}_i$ among all gallery images sorted by decreasing similarity to $\mathbf{c}_i$. We report symmetric MRR by averaging the T2I and I2T scores.

Hard-negative violation rate. For a text query $\mathbf{c}_i$ with ground-truth image $\mathbf{v}_i$, let

$$s_i^+ = s(\mathbf{c}_i, \mathbf{v}_i)$$

denote the positive similarity, and let

$$s_i^- = \max_{j \neq i} s(\mathbf{c}_i, \mathbf{v}_j)$$

denote the hardest-negative similarity. The violation rate is

$$\mathrm{HNV} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{s_i^- \geq s_i^+\}.$$

This measures the fraction of text queries for which at least one negative image scores no lower than the ground-truth match. Lower values indicate stronger local alignment. We report HNV for the T2I direction.

Mean top-k negative margin. To assess local separation in the embedding space beyond the single hardest negative, let $\mathcal{N}_k(i)$ contain the indices of the k highest-scoring negative images for text query $\mathbf{c}_i$:

$$\mathcal{N}_k(i) = \mathrm{TopK}_{j \neq i}\ s(\mathbf{c}_i, \mathbf{v}_j).$$

The mean top-k negative margin is

$$\mathrm{Margin@}\,k = \frac{1}{N}\sum_{i=1}^{N}\left(s_i^+ - \frac{1}{k}\sum_{j \in \mathcal{N}_k(i)} s(\mathbf{c}_i, \mathbf{v}_j)\right).$$

We use $k = 10$. Larger values indicate better average separation between the correct match and its strongest competitors.

I2T/T2I asymmetry gap. To quantify directional imbalance between retrieval modes, we compute

$$\mathrm{Gap@}\,K = \left|\mathrm{R@K_{I2T}} - \mathrm{R@K_{T2I}}\right|.$$

We report this at both $K = 1$ and $K = 5$.

Top-k neighbour overlap. To assess retrieval stability across CL checkpoints, let $\mathcal{S}_k^t(i)$ denote the set of top-k retrieved image indices for text query $\mathbf{c}_i$ at step t, and $\mathcal{S}_k^{t-1}(i)$ the corresponding set after the previous task. The neighbour overlap is

$$\mathrm{Overlap@}\,k = \frac{1}{N}\sum_{i=1}^{N}\frac{|\mathcal{S}_k^t(i) \cap \mathcal{S}_k^{t-1}(i)|}{k}.$$

Higher overlap indicates greater local neighbourhood stability throughout continual adaptation. We use $k = 10$, computed for the T2I direction.

Linear CKA. To evaluate global alignment between image and text representations, we compute linear centered kernel alignment (CKA). Let

$$X \in \mathbb{R}^{N \times d_x}, \quad Y \in \mathbb{R}^{N \times d_y}$$

denote mean-centered feature matrices for N paired samples. Linear CKA is defined as

$$\mathrm{CKA}(X, Y) = \frac{\|X^\top Y\|_F^2}{\|X^\top X\|_F \, \|Y^\top Y\|_F}.$$

G SUMMARY OF RELATED WORK EVALUATION PROTOCOLS

Table 19: Comparison of Continual Retrieval Benchmarks, evaluated each suite against our proposed desiderata: **Div.** (Domain Diversity), **Spec.** (Semantic Specificity), **Curr.** (Calibrated Curriculum), **Rob.** (OOD Robustness) **Acc.** (Accessibility).

Benchmark	Div.	Spec.	Curr.	Rob.	Acc.
Wang et al. (Wang et al., 2021a)	✗ Natural images only (COCO/VG)	✓ High-information target pairs	✗ Only 3 sequential tasks	✗ No zero-shot/OOD testing	✓ Small-scale and accessible
Ni et al. (Ni et al., 2023)	✗ Limited to Natural/E-commerce	✗ High overlap in Flickr/COCO	✗ Streaming focus; no curriculum	✗ No cross-domain stability eval	✓ Lightweight fine-tuning
Cui et al. (Cui et al., 2024)	✓ Diverse (Natural, e-commerce, etc.)	✓ Precise semantic targets	✗ No structured task ordering	✗ Lacks OOD classification	✓ Publicly available
Garg et al. (Garg et al., 2024)	✓ Diverse although the domain boundaries are not clearly defined.	✗ Significant web-scale noise	✓ Temporal-based ordering	✓ Extensive OOD suite	✗ Prohibitive compute costs
Liu et al. (Liu et al., 2025)	✓ Diverse (Flickr, WikiArt, etc.)	✗ Redundant pairs (e.g., Oxford Pets)	✗ Arbitrary/Random task sequence	✓ HaVG & Cls.	✗ Computationally tractable but no code
Our Framework	✓ 7 heterogeneous domains (Med, Art, etc.)	✓ Curated high-density captions	✓ ZS-calibrated difficulty order	✓ Multi-dataset Ret. & Cls.	✓ Moderate scale; Open-source