Beyond Classification: Continual Learning for Multimodal Retrieval

Alicja Dobrzeniecka [1] **Filip Szatkowski** [2 3] **Sebastian Cygert** [1] **Szymon Łukasik** [1] **Bartłomiej Twardowski** [3 4]

Abstract

While retrieval is a core function of vision-language models, continually updating these models for retrieval tasks remains critically underexplored. Existing work often treats continual retrieval as a byproduct of class-incremental learning (CIL), applying off-the-shelf methods within narrow evaluation schemes that obscure retrieval-specific failure modes and overestimate performance. To address this, we introduce a principled evaluation framework for continual multimodal retrieval spanning diverse visual domains, and systematically evaluate common approaches within this setting. Our empirical analysis shows that standard CIL methods fail to yield meaningful gains in this more realistic and challenging scenario. To tackle this problem, we propose Dynamic Adapter Routing (DAR), a novel approach based on prototypes, LoRA adapters and model merging, which outperforms existing methods by 8%. We hope our work highlights the unique challenges of continual retrieval and encourages further research in this direction.

1. Introduction

Retrieval is a core functionality of vision-language models (VLMs) such as CLIP (Radford et al., 2021), serving as the primary interface for deploying these models in real-world search, recommendation, and indexing systems. Despite its immense practical importance, the problem of continually updating multimodal models for retrieval has received relatively 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 (Wang et al., 2021b; Jha et al., 2024; Yu et al., 2024; Huang et al., 2024; Lu et al., 2025; Zheng et al., 2023), particularly class-incremental learning (CIL).

As a result, existing approaches fail to capture the unique dynamics and challenges inherent to multimodal retrieval. Instead of learning discrete decision boundaries, retrieval requires maintaining a globally consistent embedding space where images and text stay aligned across all tasks (Wang et al., 2021a). This creates unique failure modes: representation drift can ruin global rankings even if a model learns a new task perfectly, and catastrophic forgetting distorts relative similarities rather than causing simple misclassifications (Cui et al., 2024; Ni et al., 2023).

Ensuring the progress of the field, therefore, requires dedicated methods and realistic, well-designed evaluation settings. Unfortunately, existing works suffer from limited scale and domain diversity, failing to present sufficiently challenging scenarios. This in turn produces overly optimistic assessments, ultimately obscuring the true utility of CL approaches for retrieval.

To address this, we introduce a new evaluation framework for continual multimodal retrieval. Our framework includes sequences of heterogeneous, non-overlapping datasets that span various visual domains. This design ensures a sufficiently challenging evaluation, which allows us to compare the methods in a more principled way. Additionally, we propose to assess the model performance on out-of-distribution (OOD) classification and retrieval data, and demonstrate that improving both in- and out-of-distribution results remains challenging for the commonly used continual methods.

We use our framework to conduct a systematic analysis of knowledge transfer, interference and robustness under realistic distribution shifts across common CIL methods and retrieval-oriented approaches from literature. Surprisingly, we find that commonly used CL strategies fail to deliver consistent improvements in our more challenging setup.

Motivated by our findings, we introduce a novel approach

We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2024/017872 [1]NASK National Research Institute, Warsaw, Poland [2]Warsaw University of Technology, Warsaw, Poland [3]IDEAS Research Institute, Warsaw, Poland [4]Universitat Autonoma de Barcelona, Barcelona, Spain. Correspondence to: Alicja Dobrzeniecka <alicja.dobrzeniecka@nask.pl>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

for continual multimodal retrieval that explicitly addresses the above-mentioned challenges. Our Dynamic Adapter Routing (DAR) method maintains a bank of lightweight, task-specific adapters, which are selected by using a prototype-based routing mechanism. To improve robustness of our method we introduce the adaptive merging of multiple adapters, which allows the model to interpolate smoothly between the tasks at inference time. These dynamic adapters mitigate representation drift and cross-task interference directly, thereby improving retrieval performance by around 8% in heterogeneous and evolving environments, while retaining flexibility for OOD scenarios.

Overall, we position continual retrieval as a distinct problem that demands dedicated evaluation protocols and methods designed around embedding-space consistency. Our results highlight the limitations of current approaches and provide a stronger benchmark for future work.

2. Improving Continual Retrieval Evaluation

A multimodal model consists of an image encoder f_θ^I and a text encoder f_θ^T, which map inputs to a shared embedding space. Given an image x and text y, we obtain embeddings $z_x = f_\theta^I(x)$ and $z_y = f_\theta^T(y)$. Cross-modal similarity is computed using cosine similarity:

$$s(x, y) = \frac{\langle z_x, z_y \rangle}{\|z_x\| \, \|z_y\|}.$$

Retrieval is performed by ranking candidates according to $s(x, y)$. We evaluate performance using Recall@K for both image-to-text (I2T) and text-to-image (T2I) retrieval.

We consider a continual learning setting in which a multimodal model is trained over a sequence of tasks $T_1, \ldots, T_K$. Each task T_t corresponds to a dataset $\mathcal{D}_t = \{(x_i, y_i)\}_{i=1}^{N_t}$ of aligned image-text pairs. After learning task T_t, the model is evaluated on all previously seen tasks without access to their data. In contrast to classification, retrieval performance depends on relative similarities in the embedding space rather than explicit decision boundaries. Consequently, even small representation shifts can alter nearest-neighbor relationships and degrade ranking quality across tasks.

To address these challenges, we introduce a benchmark comprising seven diverse, non-overlapping datasets from various visual domains, such as natural images, synthetic images, artwork, cartoons, sketches and medical data. The task sequence is defined by ordering the datasets according to the zero-shot performance of a pre-trained model. This approach was previously introduced in the CIL evaluation framework in (Menabue et al., 2024). This forms a curriculum that progresses from easier in-domain data to more challenging out-of-distribution domains. This provides a controlled and reproducible approach to studying adaptation under increasing distribution shifts. This is in contrast to previous studies where task ordering was often arbitrary or unspecified (Liu et al., 2025). In addition to evaluating in-domain retrieval performance across all tasks, we assess generalisation by conducting zero-shot classification and zero-shot retrieval on held-out datasets. This provides a more comprehensive understanding of the impact of continual updates on specialisation and generalisation in multimodal models.

3. Dynamic Adapter Routing (DAR)

We propose Dynamic Adapter Routing (DAR), a novel adaptation framework for continual multimodal learning that operates on top of a pretrained multimodal embedding model, making it applicable to CLIP-like architectures.

Our method learns a bank of task-specific LoRA adapters, with one adapter per training dataset while maintaining a prototype memory that anchors each task in a shared backbone feature space. At inference time, DAR routes each sample to the most relevant adapter based on similarity to stored prototypes in a task-agnostic way. To improve knowledge transfer and performance for OOD samples that can benefit from more adapters, we employ model merging based on its cosine similarity to the existing prototypes.

3.1. Model and Adapter Design

We adopt a frozen pre-trained multimodal model and introduce task-specific LoRA adapters into both the visual and text backbone. During training, only the LoRA parameters associated with the active task are updated, while all backbone parameters remain fixed. We apply LoRA to the attention output projection and the two feed-forward (MLP) projections: `attn.out_proj`, `mlp.c_fc`, and `mlp.c_proj`. We use a single adapter per task, which results in a bank of lightweight task experts that share a common backbone representation. After the training we keep the adapters frozen, and train new task adapters based on the original pretrained backbone.

3.2. Prototype Memory and Margin of Similarity Score

To enable task-agnostic routing, we create a prototype memory to summarise each task within the shared backbone feature space. After training each task, we compute single image and text prototypes by averaging normalised backbone features over the corresponding task dataset.

At inference time, each sample is routed based on a similarity margin. Specifically, we calculate the cosine similarity between each sample and the prototypes associated with each adapter. The margin is defined as the difference between the top-1 and top-2 similarity scores.

Table 1. Cross-modal retrieval performance on our proposed evaluation framework measured by Recall@1 at the end of continual training for CLIP ViT-B/16. Surprisingly, commonly used CL approaches often fail to improve upon finetuning baseline, highlighting how continual retrieval presents its own set of challenges that cannot be addressed simply by adapting CIL methods. Our proposed method tailored to retrieval, DAR, shows robust performance and outperforms the alternatives by a sizable margin.

	Text → Image								Image → Text							
Method	**Flickr**	**Lexica**	**WikiArt**	**KreaM**	**Flints**	**Sketch**	**ROCOv2**	**Avg.**	**Flickr**	**Lexica**	**WikiArt**	**KreaM**	**Flints**	**Sketch**	**ROCOv2**	**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
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
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
Ours	80.0	77.6	50.0	43.0	48.3	15.8	12.4	**46.7**	92.7	76.7	50.1	44.0	46.6	15.5	12.2	**48.3**

If the margin is smaller than a predefined threshold, we interpret this as ambiguity between tasks and merge the adapters. The idea is that when the similarity scores are close, the sample may benefit from combining knowledge from multiple adapters rather than relying on a single one.

3.3. Continual Adapter Merging

To combine adapters, we leverage the Core Space merging framework (Panariello et al., 2026), which enables efficient and well-aligned merging of low-rank updates. In contrast to prior work, where merging is performed as a static, model-level operation, we apply merging conditioned on prototype-based routing. Specifically, for each input, we compute similarities to task prototypes and select the top-k adapters. For ambiguous samples, identified via a small similarity margin, we merge the selected adapters using temperature-scaled softmax weights derived from these similarities. This transforms model merging into a per-sample adaptive mechanism, improving robustness to distribution shifts and task ambiguity.

4. Experiments

We build our continual learning setup on top of a frozen CLIP ViT-B/16 backbone, which serves as a strong pretrained multimodal representation. To evaluate adaptation under realistic distribution shifts, we construct a sequence of heterogeneous datasets spanning diverse visual domains.

Importantly, the task order is not arbitrary. Instead, we sort datasets according to the zero-shot performance of the pretrained CLIP model, yielding a curriculum that progresses from in-domain to increasingly out-of-distribution data. This allows us to systematically study how models behave as the difficulty of the retrieval problem grows.

The resulting benchmark covers the following domains: **Natural images:** Flickr30K (Young et al., 2014); **AI-generated (general):** Lexica-SD (yuwan0, 2024); **AI-generated (fashion):** KreaM (hahminlew, 2023); **Artwork:** WikiArt (Ater-Mors, 2024); **Cartoons:** Flintstones (Kapuriya & Buitelaar, 2025; Kapuriya, 2025); **Sketches:** Sketch (Chowdhury et al., 2022; zoheb, 2025); **Medical images:** ROCOv2 (Rückert et al., 2024).

We compare our proposed **DAR** with a diverse set of CL strategies for VLMs, which includes regularization-based, parameter-efficient, and task-aware approaches. We employ simple baselines, such as **Zero-shot (ZS)** performance of the backbone and **Fine-tuning (FT)**, where the model is sequentially updated without explicit mechanisms to prevent forgetting. Moreover, we include standard CL approaches applicable to retrieval models, such as **EWC** (Kirkpatrick et al., 2017) and **L2P** (Wang et al., 2021b) adjusted for retrieval, and **Task Arithmetic** (Ilharco et al., 2023). Finally we also evaluate methods that were created directly for retrieval task such as **Mod-X** (Ni et al., 2023) and **C-CLIP** (Liu et al., 2025).

4.1. Implementation details

We implement DAR using task-specific LoRA adapters with rank 16, scaling factor 32, and dropout 0.1, inserted into both visual and text encoders of a frozen CLIP backbone. Prototype representations are L2-normalized, and routing decisions combine image and text similarities with equal weighting. For ambiguous samples, we select the top-2 adapters and trigger merging when the similarity margin falls below 0.05. Models are trained for 20 epochs using AdamW with a learning rate of 1e-4, batch sizes of 256 (train) and 512 (validation), and a fixed random seed of 42.

4.2. Main results for training tasks

In Table 1, we report Recall@1 for both I2T and T2I retrieval after training CLIP with common CL methods and DAR in our framework. Surprisingly, under more challenging evaluation setting, naive fine-tuning is a strong baseline which outperforming or matching very close to several dedicated CL methods such as C-CLIP and Mod-X. In contrast, DAR provides a strong improvement over the baseline, outperforming finetuning by approximately +10% for both I2T

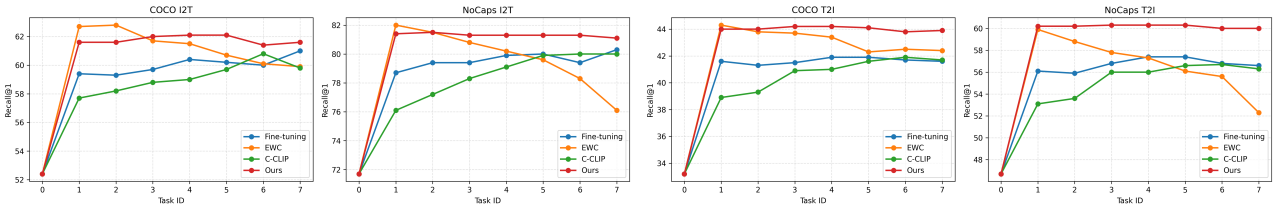

Figure 1. (left) Image-to-Text and (right) Text-to-Image Recall@1 for continual learning methods on COCO and NoCaps

and T2I. Overall, our results demonstrate that existing methods fail to consistently address the challenges of continual retrieval under a more demanding evaluation protocol, while highlighting the effectiveness of our proposed approach.

4.3. Out-of-distribution evaluations

We first visualize zero-shot retrieval performance on held-out datasets in Figure 1, where DAR consistently achieves the strongest results, with gains of up to +12.8 Recall@1 on both COCO and NoCaps. This indicates that, despite continual updates, DAR preserves and even improves the global alignment structure required for retrieval under distribution shift. Full quantitative results are provided in Tables Tables 13 and 14 in the appendix. Additional results in the appendix show that this improvement comes with a trade-off in zero-shot classification performance, highlighting a tension between specialization required for good retrieval and classification generalization under continual training.

5. Related Work

CL for VLMs has been studied across several directions, including cross-modal retrieval, continual adaptation of CLIP-like models, and large-scale multimodal pretraining. Early work on continual cross-modal retrieval (Wang et al., 2021a) highlights that preserving a globally consistent embedding space is critical, distinguishing retrieval from standard classification-based CL. Subsequent methods such as DKR (Cui et al., 2024) and Mod-X (Ni et al., 2023) focus on mitigating representation drift in retrieval settings. Another line of research studies the continual adaptation of vision-language models under domain shifts, with approaches such as ZSCL (Zheng et al., 2023) and TiC-CLIP (Garg et al., 2024) investigating maintaining zero-shot and downstream performance during continual updates. In particular, TiC-CLIP introduces time-continuous benchmarks based on web-scale data streams, where the distribution evolves naturally over time. C-CLIP framework (Liu et al., 2025) addresses multimodal continual learning with both downstream and zero-shot evaluation and proposes a novel method for that. Other approaches explore parameter-efficient adaptation, including expert-style adapters with learned routing for continual VLM classification (Yu et al., 2024), large-scale continual multimodal pretraining (Udandarao et al., 2024), or structured vision-language reasoning (Smith et al., 2023). While Yu et al. (2024) is closely related in employing expert-style adapters, it focuses on continual classification rather than retrieval. A direct comparison is an interesting direction for future work. Unlike prior studies that primarily address temporal distribution shifts or evaluate on fixed sequences with limited domain diversity, our work specifically targets heterogeneous domain shifts across distinct tasks and rigorously assesses robustness to varying task orderings. We present a detailed comparison between the evaluation settings in previous works and DAR in Table 15.

6. Conclusions

We study continual multimodal retrieval as a setting that differs structurally from standard CIL. In this setting, performance depends on preserving relative embedding alignment rather than learning decision boundaries.

We introduce a heterogeneous evaluation framework and demonstrate that widely used continual learning methods offer limited or inconsistent improvements in this more realistic scenario, often performing similarly to straightforward fine-tuning. To address this issue, we propose DAR: a new method that is based on prototype-guided adapter selection and input-dependent merging. DAR improves retrieval performance consistently across domains and under distribution shift, outperforming existing approaches on both in-domain and held-out benchmarks.

Overall, our results suggest that continual retrieval requires evaluation protocols and methods that are specifically designed to preserve embedding-space consistency. A key area for future research is understanding how to balance specialisation and generalisation in continually updated multimodal systems. Overcoming this challenge is essential for the deployment of continual multimodal retrieval systems in dynamic, real-world environments.

Impact Statement

This paper presents work aimed at advancing the field of machine learning. While there are many potential societal consequences of our work, we do not feel that any of these need to be highlighted specifically here.

References

Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., and Anderson, P. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8948–8957, 2019.

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

Chowdhury, P. N., Sain, A., Bhunia, A. K., Xiang, T., Gryaditskaya, Y., and Song, Y.-Z. Fs-coco: Towards understanding of freehand sketches of common objects in context. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pp. 253–270, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-20073-1. doi: 10.1007/978-3-031-20074-8_15. URL https://doi.org/10.1007/978-3-031-20074-8_15.

Cui, Z., Peng, Y., Wang, X., Zhu, M., and Zhou, J. Continual vision-language retrieval via dynamic knowledge rectification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(10):11704–11712, Mar. 2024. doi: 10.1609/aaai.v38i10.29054. URL https://ojs.aaai.org/index.php/AAAI/article/view/29054.

Garg, S., Farajtabar, M., Pouransari, H., Vemulapalli, R., Mehta, S., Tuzel, O., Shankar, V., and Faghri, F. Tic-clip: Continual training of clip models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=TLADT8Wrhn.

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

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

Huang, L., Cao, X., Lu, H., and Liu, X. Class-incremental learning with clip: Adaptive representation adjustment and parameter fusion. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LIV*, pp. 214–231, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72948-5. doi: 10.1007/978-3-031-72949-2_13. URL https://doi.org/10.1007/978-3-031-72949-2_13.

Ilharco, G., Ribeiro, M. T., Wortsman, M., Schmidt, L., Hajishirzi, H., and Farhadi, A. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023.

Jha, S., Gong, D., and Yao, L. CLAP4CLIP: Continual learning with probabilistic finetuning for vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=rF1YRtZfoJ.

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

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

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL https://www.pnas.org/doi/abs/10.1073/pnas.1611835114.

Li, Y., Pang, G., Suo, W., Jing, C., Xi, Y., Liu, L., Chen, H., Liang, G., and Wang, P. 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. doi: 10.1109/TNNLS.2025.3547882.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Liu, W., Zhu, F., Wei, L., and Tian, Q. C-clip: Multimodal continual learning for vision-language model. In Yue, Y., Garg, A., Peng, N., Sha, F., and Yu, R. (eds.), *International Conference on Learning Representations*, volume 2025, pp. 46461–46477, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/72fb9ab442fc60b7ae5d53bf6b478273-Paper-Conference.pdf.

Lu, H., Zhang, X., Moore, K., Xue, J., Yao, L., van den Hengel, A., and Gong, D. Continual learning on CLIP via incremental prompt tuning with intrinsic textual anchors. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=YJnjkzKq5Y.

Marczak, D., Magistri, S., Cygert, S., Twardowski, B., Bagdanov, A. D., and van de Weijer, J. No task left behind: Isotropic model merging with common and task-specific subspaces. In *Forty-second International Conference on Machine Learning*, 2025.

Menabue, M., Frascaroli, E., Boschini, M., Sangineto, E., Bonicelli, L., Porrello, A., and Calderara, S. Semantic residual prompts fornbsp;continual learning. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXI*, pp. 1–18, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-73029-0. doi: 10.1007/978-3-031-73030-6_1. URL https://doi.org/10.1007/978-3-031-73030-6_1.

Ni, Z., Wei, L., Tang, S., Zhuang, Y., and Tian, Q. Continual vision-language representation learning with off-diagonal information. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Panariello, A., Marczak, D., Magistri, S., Porrello, A., Twardowski, B., Bagdanov, A. D., Calderara, S., and van de Weijer, J. Accurate and efficient low-rank model merging in core space. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=y1z7SAS8q8.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Rückert, J., Bloch, L., Brüngel, R., Idrissi-Yaghir, A., Schäfer, H., Schmidt, C. S., Koitka, S., Pelka, O., Abacha, A. B., G. Seco de Herrera, A., Müller, H., Horn, P. A., Nensa, F., and Friedrich, C. M. Rocov2: Radiology objects in context version 2, an updated multimodal image dataset. *Scientific Data*, 11(1), June 2024. ISSN 2052-4463. doi: 10.1038/s41597-024-03496-6. URL http://dx.doi.org/10.1038/s41597-024-03496-6.

Smith, J. S., Cascante-Bonilla, P., Arbelle, A., Kim, D., Panda, R., Cox, D., Yang, D., Kira, Z., Feris, R., and Karlinsky, L. Construct-vl: Data-free continual structured vl concepts learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14994–15004, June 2023.

Udandarao, V., Roth, K., Dziadzio, S., Prabhu, A., Cherti, M., Vinyals, O., Hénaff, O., Albanie, S., Akata, Z., and Bethge, M. A practitioner's guide to real-world continual multimodal pretraining. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 133801–133845. Curran Associates, Inc., 2024. doi: 10.52202/079017-4252. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/f1a6a2cdc7e65dbb4579e78f97cd2665-Paper-Datasets_and_Benchmarks_Track.pdf.

Wang, K., Herranz, L., and van de Weijer, J. Continual learning in cross-modal retrieval . In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3623–3633, Los Alamitos, CA, USA, June 2021a. IEEE Computer Society. doi: 10.1109/CVPRW53098.2021.00402. URL https://doi.ieeecomputersociety.org/10.1109/CVPRW53098.2021.00402.

Wang, Z., Zhang, Z., Lee, C.-Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J. G., and Pfister, T. Learning to prompt for continual learning. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 139–149, 2021b. URL https://api.semanticscholar.org/CorpusID:245218925.

Yadav, P., Tam, D., Choshen, L., Raffel, C. A., and Bansal, M. Ties-merging: Resolving interference when merging models. *Advances in neural information processing systems*, 36:7093–7115, 2023.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. doi: 10.1162/tacl_a_00166. URL https://aclanthology.org/Q14-1006/.

Yu, J., Zhuge, Y., Zhang, L., Hu, P., Wang, D., Lu, H., and He, Y. 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.

Zheng, Z., Ma, M., Wang, K., Qin, Z., Yue, X., and You, Y. 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. Ablations

We conducted ablation studies to estimate both the upper bound of achievable performance and the effectiveness of a minimal LoRA (Hu et al., 2022) configuration. For the upper bound, we performed task-aware inference, obtaining results that closely match those achieved with the prototypes (48.5 for I2T and 47.2 for T2I on average). In contrast, using a single global LoRA leads to a substantial performance drop, reducing scores to 36.3 for I2T and 34.3 for T2I.

Table 2. Generalization of DAR across CLIP backbones. Results show that DAR consistently outperforms prior continual learning methods on ViT-B/32, ViT-B/16, and ViT-L/14.

Method	ViT-B/32		ViT-B/16		ViT-L/14	
	T→I	I→T	T→I	I→T	T→I	I→T
FT	33.4	31.1	36.2	38.7	44.0	45.5
Mod-X	32.9	30.6	36.8	38.5	44.0	45.8
C-CLIP	32.1	29.2	34.3	35.9	41.0	42.3
DAR	**42.3**	**43.7**	**47.5**	**49.3**	**52.8**	**53.7**

Table 3. LoRA rank ablation.

	T→I	I→T
Full-FT	36.2	38.7
LoRA-FT (r=16)	36.0	34.2
DAR - LoRA (r=4)	46.9	48.3
DAR - LoRA (r=8)	**47.5**	**49.3**
DAR - LoRA (r=16)	47.0	48.5
DAR - LoRA (r=32)	47.1	48.5

Table 4. DAR routing strategies.

	T→I	I→T
Random	36.7	32.7
Oracle	47.1	48.6
Image-only	46.9	48.4
Text-only	43.0	45.0
Image+Text (max)	46.6	48.2
Image+Text (avg)	46.9	48.5
DAR	**47.5**	**49.3**

Table 5. Model merging ablation.

	T→I	I→T
No merging	46.8	48.6
Avg (uniform)	46.8	48.2
TA (Ilharco et al., 2023)	46.9	48.4
DARE-TIES (Yadav et al., 2023)	46.8	48.3
ISO-C (Marczak et al., 2025)	46.9	48.4
CoreSpace (Panariello et al., 2026)	**47.5**	**49.3**

Table 6. Adaptive merging ablation.

		ID		OOD	
k	γ	T→I	I→T	T→I	I→T
1	–	46.8	48.6	51.8	71.3
2	0.00	46.9	48.5	52.0	71.2
2	0.01	46.9	48.5	52.0	71.2
2	0.05	**47.5**	**49.3**	52.2	71.4
2	0.10	46.8	48.4	**52.4**	71.6
3	0.05	46.9	48.3	52.3	71.5
4	0.05	46.9	48.3	52.3	**71.7**

B. Additional quantitative results for the main experimental suite

Table 7. 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
DAR	95.5	91.6	76.5	78.1	87.8	39.5	28.8	**71.1**	99.5	91.0	77.4	77.3	86.2	39.7	28.8	**71.4**

Table 8. Cross-modal retrieval performance on **a reverse of our main evaluation framework** measured by Recall@1 at the end of continual training for CLIP ViT-B/16.

	Text → Image								**Image → Text**							
Method	**ROCOv2**	**Sketch**	**Flints**	**KreaM**	**WikiArt**	**Lexica**	**Flickr**	**Avg.**	**ROCOv2**	**Sketch**	**Flints**	**KreaM**	**WikiArt**	**Lexica**	**Flickr**	**Avg.**
ZS	1.8	5.3	16.6	22.0	20.6	53.3	62.3	26.0	1.5	4.2	11.1	20.2	20.8	53.0	82.0	27.5
FT	5.1	8.9	35.4	27.1	38.1	68.8	75.9	37.1	5.4	8.1	31.6	27.9	38.2	68.8	90.6	38.7
EWC	7.0	10.1	40.1	32.7	41.0	68.3	79.6	39.8	6.2	8.6	32.1	31.7	37.4	68.5	92.3	39.5
Mod-X	6.0	9.5	39.0	29.3	39.8	69.6	77.9	38.7	6.4	8.5	32.5	29.7	38.8	69.0	91.9	9.6
C-CLIP	3.2	8.2	30.9	25.5	34.8	65.0	74.1	34.5	3.0	6.6	24.4	26.1	33.5	67.7	88.8	35.7
L2P	2.2	5.9	22.3	20.6	26.8	55.3	70.2	29.0	1.7	5.0	16.9	18.6	25.6	51.5	86.4	29.4
DKR	6.9	10.0	41.2	33.5	43.4	69.4	80.3	40.7	6.0	8.0	32.4	31.4	39.4	69.3	93.4	40.0
TA	3.2	8.4	32.5	24.3	34.9	62.5	73.1	34.1	3.7	7.9	27.5	24.5	35.2	65.0	88.3	36.0
DAR	9.7	13.6	45.1	39.8	42.7	74.5	75.1	**42.9**	10.0	13.5	43.4	39.1	43.6	74.5	88.1	**44.6**

C. Additional zero-shot classification results

Table 9. 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
Merging-TA	68.1	69.0	68.9	68.5	68.5	67.6	67.3	67.0	-1.1
Ours	68.1	64.4	64.4	64.4	64.4	64.4	64.4	64.4	-3.7

Table 10. 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
Merging-TA	68.4	68.7	68.6	69.3	69.4	69.4	69.7	70.1	+1.7
Ours	65.8	65.8	65.8	65.8	65.8	68.0	68.0	68.0	+2.2

Table 11. 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
Merging-TA	54.0	54.1	56.4	55.4	55.6	52.1	53.8	53.6	-0.4
Ours	54.0	46.6	46.6	46.6	46.6	46.6	54.0	53.8	-0.02

Table 12. 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
Merging-TA	56.8	56.9	57.0	56.7	56.7	56.5	56.6	56.2	-0.6
Ours	56.8	54.7	54.7	54.8	54.8	54.8	54.5	54.5	-2.3

D. Additional zero-shot retrieval results

Table 13. Zero-shot retrieval performance on COCO2014 during continual fine-tuning. We report Recall@1 for Image-to-Text (I2T) and Text-to-Image (T2I), together with performance difference Δ.

		Task ID								Δ
Direction	Method	0	1	2	3	4	5	6	7	
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
	Merging-TA	52.4	57.4	57.7	58.5	58.9	59.6	59.4	59.7	+7.3
	Ours	52.4	61.6	61.6	62.0	62.1	62.1	61.4	61.6	**+9.2**
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
	Merging-TA	33.2	38.7	38.9	40.6	40.5	40.6	40.8	41.0	+7.8
	Ours	33.2	44.0	44.0	44.2	44.2	44.1	43.8	43.9	**+10.7**

Table 14. Zero-shot retrieval performance on NoCaps during continual fine-tuning. We report Recall@1 for Image-to-Text (I2T) and Text-to-Image (T2I), together with performance difference Δ.

		Task ID								Δ
Direction	Method	0	1	2	3	4	5	6	7	
I2T	Fine-tuning	71.7	78.7	79.4	79.4	79.9	80.0	79.4	80.3	+8.6
	EWC	71.7	82.0	81.5	80.8	80.2	79.6	78.3	76.1	+4.4
	Mod-X	71.7	81.1	80.9	80.4	81.0	80.2	80.0	79.8	+8.1
	C-CLIP	71.7	76.1	77.2	78.3	79.1	79.9	80.0	80.0	+8.3
	L2P	71.7	72.0	72.7	73.8	74.9	71.9	73.2	73.6	+0.9
	Merging-TA	71.7	76.2	76.6	78.1	78.4	79.0	78.9	79.7	+8.0
	Ours	71.7	81.4	81.5	81.3	81.3	81.3	81.3	81.1	**+9.4**
T2I	Fine-tuning	46.7	56.1	55.9	56.8	57.4	57.4	56.8	56.6	+9.9
	EWC	46.7	59.9	58.8	57.8	57.3	56.1	55.6	52.3	+5.6
	Mod-X	46.7	58.2	57.8	57.8	58.2	57.4	57.1	55.9	+9.2
	C-CLIP	46.7	53.1	53.6	56.0	56.0	56.6	56.7	56.3	+9.6
	L2P	46.7	51.3	51.1	52.5	52.4	53.0	52.2	51.3	+4.6
	Merging-TA	46.7	53.0	53.2	55.1	55.3	55.3	55.3	55.6	+8.9
	Ours	46.7	60.2	60.2	60.3	60.3	60.3	60.0	60.0	**+13.3**

E. Comparing other frameworks

Table 15. Summary of training and evaluation datasets across selected vision-language CL papers.

Paper	Training Data	Training Domains Approximation (no. unique vs total)	Eval ID	Eval OOD Classification	Eval OOD Retrieval
Our framework	**Retrieval:** Flickr30K, LexicaSD, WikiArt, Kream, Flintstones, Sketch, ROCOv2	7/7: Natural images, synthetic, art, fashion, cartoons, sketches, medical	Same datasets	ImageNet, CIFAR100, EuroSAT, DomainNet	COCO (Lin et al., 2014), NoCaps (Agrawal et al., 2019)
C-CLIP (Liu et al., 2025)	**Retrieval:** Flickr30K, COCO, Pets, Lexica, Simpsons, WikiArt, Kream, Sketch	6/8: Natural images, synthetic, cartoons, art, e-commerce, sketches	Same datasets	CIFAR100, ImageNet, Flowers, DTD, Food101, Stanford Cars	HAVG
DKR (Cui et al., 2024)	**Retrieval:** MS-COCO, Flickr30K, IAPR TC-12, EC, RSICD	4/5: Natural images, remote sensing, e-commerce/products, general web images	Same datasets	–	Train on EC splits → test on unseen MS-COCO, Flickr30K
TIC-CLIP (Garg et al., 2024)	**Retrieval:** TIC-DataComp, TIC-YFCC, TIC-RedCaps	2/3: Large-scale web data, diverse real-world domains	Classification: TIC-DataComp-Net; Retrieval: TIC-DataComp-Retrieval, TIC-YFCC Retrieval and TIC-RedCaps	Over 28 datasets: ImageNet, Food101, MNIST, Oxford-Flowers, Stanford Cars, SUN-97, Oxford-Pet, ObjectNet, and more.	Flickr30k
Cross-modal Retrieval (Wang et al., 2021a)	**Retrieval:** Sequential Visual Genome (SeViGe), Sequential MS-COCO (SeCOCO)	1/2: Natural images	Same datasets	–	–
Mod-X (Ni et al., 2023)	**Retrieval:** COCO (initial training) + Flickr30K (streaming); also ECommerce-T2I in extended experiments	2/3: Natural images, e-commerce/products	Same datasets	–	Train on COCO/Flickr → test on unseen ECommerce-T2I