Continual Vision-Language Retrieval via Dynamic Knowledge Rectification

Zhenyu Cui[1], Yuxin Peng[1*], Xun Wang[2], Manyu Zhu[2], Jiahuan Zhou[1]

[1]Wangxuan Institute of Computer Technology, Peking University
[2]ByteDance Inc
cuizhenyu@stu.pku.edu.cn, {pengyuxin, jiahuanzhou}@pku.edu.cn, {wangxun.2, zhumanyu}@bytedance.com

Abstract

The recent large-scale pre-trained models like CLIP have aroused great concern in vision-language tasks. However, when required to match image-text data collected in a streaming manner, namely Continual Vision-Language Retrieval (CVRL), their performances are still limited due to the catastrophic forgetting of the learned old knowledge. To handle this issue, advanced methods are proposed to distil the affinity knowledge between images and texts from the old model to the new one for anti-forgetting. Unfortunately, existing approaches neglect the impact of incorrect affinity, which prevents the balance between the anti-forgetting of old knowledge and the acquisition of new knowledge. Therefore, we propose a novel framework called Dynamic Knowledge Rectification (DKR) that simultaneously achieves incorrect knowledge filtering and rectification. Specifically, we first filter the incorrect affinity knowledge calculated by the old model on the new data. Then, a knowledge rectification method is designed to rectify the incorrect affinities while preserving the correct ones. In particular, for the new data that can only be correctly retrieved by the new model, we rectify them with the corresponding new affinity to protect them from negative transfer. Additionally, for those that can not be retrieved by either the old or the new model, we introduce paired ground-truth labels to promote the acquisition of both old and new knowledge. Extensive experiments on several benchmark datasets demonstrate the effectiveness of our DKR and its superiority against state-of-the-art methods.

Introduction

In recent years, the pre-trained CLIP model (Radford et al. 2021) has become a milestone for vision-language retrieval, where an image is retrieved via a text description of its contents or vice versa. Owing to the large-scale pre-trained dataset, CLIP has demonstrated its promising ability to tackle new datasets in a zero-shot manner but still suffers from inadequate performance (Gu et al. 2021; Xu et al. 2022; Wu et al. 2023). Hence, various works adopt fine-tuning and adapter strategies to improve the performance of the pre-trained CLIP model on new datasets (Gao et al. 2023a; Luo et al. 2022; Zhang et al. 2022). However, when facing a practical and crucial scenario where a series of datasets come one by one, the above strategies always fail to handle all the datasets well. Their performance can be significantly degraded on the learned old datasets due to the well-known catastrophic forgetting challenge (De Lange et al. 2021). Therefore, in this paper, we focus on adapting CLIP to new datasets while maintaining its performance on old datasets, known as the task of Continual Vision-Language Retrieval (CVLR).

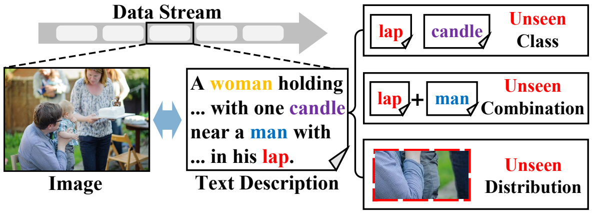
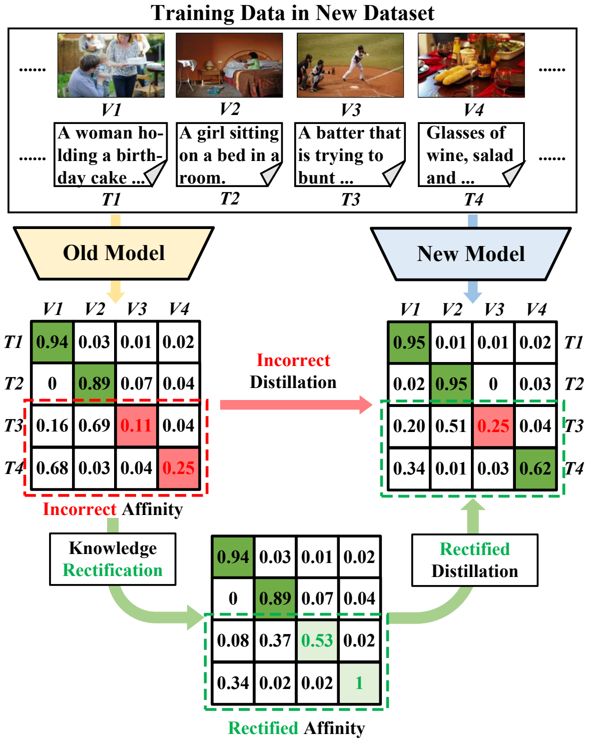


Figure 1: Challenges in CVLR. In addition to the unseen class label concerned by CIL, CVLR involves the unseen class combination and the unseen class distribution.

So far, most continual learning works focus on how to achieve the anti-forgetting of class information of the images in the learned old datasets (Masana et al. 2022) which is also known as class incremental learning (CIL). However, the existing CIL methods can not be directly used to handle the CVLR task. As shown in Fig. 1, on the one hand, images and text descriptions in the CVLR task do not have the necessary category label information for CIL, which prevents correctly matching images with corresponding descriptions. On the other hand, CVLR exhibits a more complicated scenario where the text description usually meets the unseen class label, the unseen class combination, and the unseen class distribution. These challenges leave the CVLR task still an unsolved problem.

To tackle the above challenges, knowledge distillation has become an effective solution to CVLR task (Wang, Herranz, and van de Weijer 2021; Srinivasan et al. 2022; Dong et al. 2021; Ni et al. 2023), which involves transferring old knowledge learned by the old model to the new model. Recent works (Dong et al. 2021; Ni et al. 2023) regarded the affinity relationship between image-text pairs calculated by the old model as old knowledge and developed knowledge distillation to alleviate the catastrophic forgetting problem. How-

[*]Corresponding author.


Training Data in New Dataset
......
V1
V2
V3
V4
......
A woman ho- lding a birth- day cake ...
A girl sitting on a bed in a room.
A batter that is trying to bunt ...
Glasses of wine, salad and ...
......
T1
T2
T3
T4
Old Model
New Model
Incorrect Distillation
Incorrect Affinity
Knowledge Rectification
Rectified Distillation
Rectified Affinity

	V1	V2	V3	V4
T1	0.94	0.03	0.01	0.02
T2	0	0.89	0.07	0.04
T3	0.16	0.69	0.11	0.04
T4	0.68	0.03	0.04	0.25

V1	V2	V3	V4	
0.95	0.01	0.01	0.02	T1
0.02	0.95	0	0.03	T2
0.20	0.51	0.25	0.04	T3
0.34	0.01	0.03	0.62	T4

0.94	0.03	0.01	0.02
0	0.89	0.07	0.04
0.08	0.37	0.53	0.02
0.34	0.02	0.02	1

Figure 2: Existing CVLR methods fail to deal with the incorrect affinity calculated by the old model, thereby limiting the acquisition of both old and new knowledge. In contrast, we rectify those incorrect affinities and distil them into a new model to solve the above problem.

ever, as shown in Fig. 2, these existing methods neglect the impact of incorrect affinity when distilling old knowledge, thereby limiting the acquisition of both old and new knowledge. Although the latest research (Ni et al. 2023) preliminarily alleviated this problem by filtering the incorrect affinity, it inevitably leads to the loss of relevant old knowledge. As a result, existing methods can hardly perform well on both old and new datasets.

Inspired by the above observation, we propose a novel framework for the CVLR task, named Dynamic Knowledge Rectification (DKR). As shown in Fig. 2, the core idea of DKR is to rectify the incorrect affinity calculated by the old model while maintaining the correct one. To this end, a dynamic knowledge filtering and rectification module is designed to strike the balance between the anti-forgetting of old knowledge and the acquisition of new knowledge. Specifically, we first compute an old affinity and a new affinity based on the deep embedding extracted by the old model and the new model, respectively. Then, for the new data that can be correctly matched by the old model, DKR directly maintains the old affinity to avoid forgetting of the old knowledge. Besides, for those that can only be correctly retrieved by the new model, DKR exploits the new affinity to rectify the incorrect part in the old affinity to avoid the negative transfer of the old knowledge. In addition, for those that not be retrieved by either the old or the new model, DKR introduces additional knowledge, i.e., the paired ground-truth label, to promote the acquisition of both old and new knowledge. Then, we combine the rectified results and exploit the knowledge distillation to constrain the learning process of the new model. In summary, the main contributions of this paper are as follows:

1) To alleviate the catastrophic forgetting problem in the continual vision-language retrieval task, we propose a novel Dynamic Knowledge Rectification (DKR) framework to dynamically filter and rectify incorrect old knowledge, which strikes the balance between the anti-forgetting of old knowledge and the acquisition of new knowledge.

2) A knowledge rectification method is designed to protect the new model from negative transfer and promote the acquisition of both old and new knowledge.

3) Extensive experiments on five real-world vision-language retrieval benchmark datasets demonstrate the superiority of our proposed DKR against the state-of-the-art methods.

Related Work

Vision-Language Retrieval

Matching the same semantics between images and texts is the key to vision-language retrieval. Most existing works calculated the pairwise affinity by extracting and mapping vision and language features into a common embedding space. According to different interaction patterns, existing image-text retrieval methods can be roughly categorized into two branches: 1) Corss-modal interaction methods. These methods focus on inferring and aligning the pairwise relationship across cross-modal entities (Chen et al. 2020a,b; Li et al. 2020). IMRAM (Chen et al. 2020a) designed an iterative image-text retrieval framework to capture correspondences by stacking cross-attention neural networks. Although achieving some progress, the redundant computational cost of calculating the similarity between images and text limits their practicality (Chen et al. 2021). 2) Intra-modal representation methods. To tackle the above limitations, methods of this category employ independent representation networks for vision and language modalities, thus improving the inference efficiency (Li et al. 2019; Wang et al. 2020; Chen et al. 2021). Wang (Wang et al. 2020) further exploit a consensus-aware visual-semantic embedding model to incorporate the commonsense knowledge share between both modalities. However, the above methods are typically designed for retrieval in specific datasets.

Recently, large-scale multi-modal pre-trained models (Li et al. 2021; Radford et al. 2021; Li et al. 2023) have aroused great concern in the community. CLIP (Radford et al. 2021) greatly improves the performance of the vision-language retrieval task through large-scale self-supervised contrastive learning. These models are usually built with extremely large-scale vision-language datasets (Ordonez, Kulkarni, and Berg 2011; Sharma et al. 2018; Schuhmann et al. 2021)), and show impressive results in multiple downstream tasks (Wang et al. 2022; Chowdhury, Zhuang, and Wang 2022; Luo et al. 2022). However, their performance can be greatly degraded when it comes to sequentially given training datasets due to the catastrophic forgetting problem, which limits their applicability in real scenarios.

Continual Learning

Continual Learning (CL) (De Lange et al. 2021) aims to enable intelligent systems to continuously learn to adapt to new datasets while preserving old knowledge learned from old datasets. Most CL methods (Liu et al. 2022; Qiu et al. 2023; Gao et al. 2023b) are designed for class-incremental learning, namely CIL, i.e. categories in the new task never appeared in the old task. Qiu (Qiu et al. 2023) devised a causal intervened learning strategy to eliminate the causal path that causes the task-induced bias which resulted in the catastrophic forgetting problem. Gao (Gao et al. 2023b) transferred diverse knowledge from both task-specific and task-general knowledge to the current task to balance the stability and the plasticity of the old knowledge. However, CIL methods are not suitable for the CVLR task when considering the lack of class label information for training and the complicated scenario. To address the above challenge, some methods employed knowledge distillation strategies (Wang, Herranz, and van de Weijer 2021; Srinivasan et al. 2022; Dong et al. 2021; Ni et al. 2023) for generalized uses. Dong (Dong et al. 2021) proposed to distillate relation knowledge between paired samples to prevent the forgetting of the structural knowledge. Ni (Ni et al. 2023) maintained the multimodal common representation space by aligning the contrastive matrices to alleviate the spatial disorder when learning new knowledge.

Different from these methods, we propose a dynamic knowledge rectification strategy for CVLR. It can dynamically filter and rectify the incorrect old knowledge, and distil it to the new model in a unified way to strike the balance between the anti-forgetting of old knowledge and the acquisition of both old and new knowledge.

Proposed Method

Problem Definition and Notations

In this paper, we focus on the challenging Continual Vision-Language Retrieval (CVLR) task, which assumes the vision-language data comes in a streaming manner. Specifically, given a sequentially collected vision-language datasets $\mathcal{D} = \{\mathcal{D}^{(1)}, \mathcal{D}^{(2)}, ..., \mathcal{D}^{(S)}\}$, where S denotes the total length of the sequence. The s-th dataset $\mathcal{D}^{(s)} = \{(v_i^{(s)}, t_i^{(s)})\}_{i=1}^{N^{(s)}}$ consists of $N^{(s)}$ input images $v_i^{(s)}$ and corresponding text descriptions $t_i^{(s)}$. Each $\mathcal{D}^{(s)}$ is randomly divided into a training set $\mathcal{D}_{train}^{(s)}$ and a testing set $\mathcal{D}_{test}^{(s)}$. Our goal is to train an encoding model $f(x) : x \to e_x$ to map the input sample x to a deep embedding e_x and maximize the similarity between the embedding of v_i and t_i. Notably, the previous $s-1$ datasets are completely unavailable when training on the s-th dataset, while all s testing sets are jointly used to evaluate the overall retrieval performance of $f(x)$ to all datasets.

Overview

As shown in Fig. 3, our proposed DKR method consists of a CLIP-based image encoder, a CLIP-based text encoder, and a Dynamic Knowledge Rectification (DKR) module. The encoders are first initialized with the pre-trained parameters, which are then sequentially trained on S datasets. For the beginning of the training stage s, we preserve a copy of the encoders as the old model to represent old knowledge. Then, we use the old model and the new model to calculate the affinity between each two image-text samples in the s-th dataset, which forms an old affinity matrix $\mathcal{M}^{(s-1)}$ and a new affinity matrix $\mathcal{M}^{(s)}$, respectively. Sequentially, the DKR module filters the incorrect affinity in $\mathcal{M}^{(s-1)}$, and dynamically rectifies them with the correct new affinity and paired Ground-Truth (GT) label to form a rectified old knowledge matrix $\mathcal{M}^{(T)}$. Finally, DKR distils the knowledge in $\mathcal{M}^{(T)}$ to $\mathcal{M}^{(s)}$.

During the training phase, we employ an original contrastive loss $\mathcal{L}_{clip}$ to train our model on $\mathcal{D}^{(1)}$. When training the following steps ($s >= 2$), we add our customized knowledge distillation loss $\mathcal{L}_{JS}$ to alleviate the catastrophic forgetting problem. During the inference phase, only the trained encoders are retained to validate the CVLR performance on all S datasets.

CLIP-based Retrieval Baseline

In this section, we present the CLIP-based retrieval baseline of our DKR.

Considering that the vision-language retrieval task is mainly to match paired images and texts, we employ the CLIP-based encoders to learn the discriminative deep embedding. Formally, let $f_v(x)$ and $f_t(x)$ be the encoders for images and texts, respectively. The deep embedding $\boldsymbol{z}_v^i$ and $\boldsymbol{z}_t^i$ can be formulated as:

$$\begin{cases} \boldsymbol{z}_v^i = f_v(v_i) \\ \boldsymbol{z}_t^i = f_t(t_i). \end{cases} \tag{1}$$

To maximize the affinity of the matched image-text pairs, a contrastive loss $\mathcal{L}_{clip}$ is employed to pull the distance between the paired samples while pushing the distance of unpaired samples. The $\mathcal{L}_{clip}$ can be expressed as follows:

$$\begin{cases} \mathcal{L}_v^i = -log(\frac{e^{\langle \boldsymbol{z}_v^i, \boldsymbol{z}_t^i \rangle / \tau}}{\sum_{k=1}^N e^{\langle \boldsymbol{z}_v^i, \boldsymbol{z}_t^k \rangle / \tau}}) \\ \mathcal{L}_t^i = -log(\frac{e^{\langle \boldsymbol{z}_t^i, \boldsymbol{z}_v^i \rangle / \tau}}{\sum_{k=1}^N e^{\langle \boldsymbol{z}_t^i, \boldsymbol{z}_v^k \rangle / \tau}}) \\ \mathcal{L}_{clip}^s = \sum_{i=1}^N (\mathcal{L}_v^i + \mathcal{L}_t^i)/2, \end{cases} \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product of two embedding, N is the batch size, and $\tau \in \mathcal{R}^+$ is a learnable temperature parameter.

Dynamic Knowledge Rectification

Although the CLIP-based retrieval baseline shows impressive performance on new datasets, it is still challenging to maintain its performance on old datasets. Therefore, we propose a Dynamic Knowledge Rectification (DKR) module to alleviate the catastrophic forgetting problem. Different from the common class incremental learning, which can use the

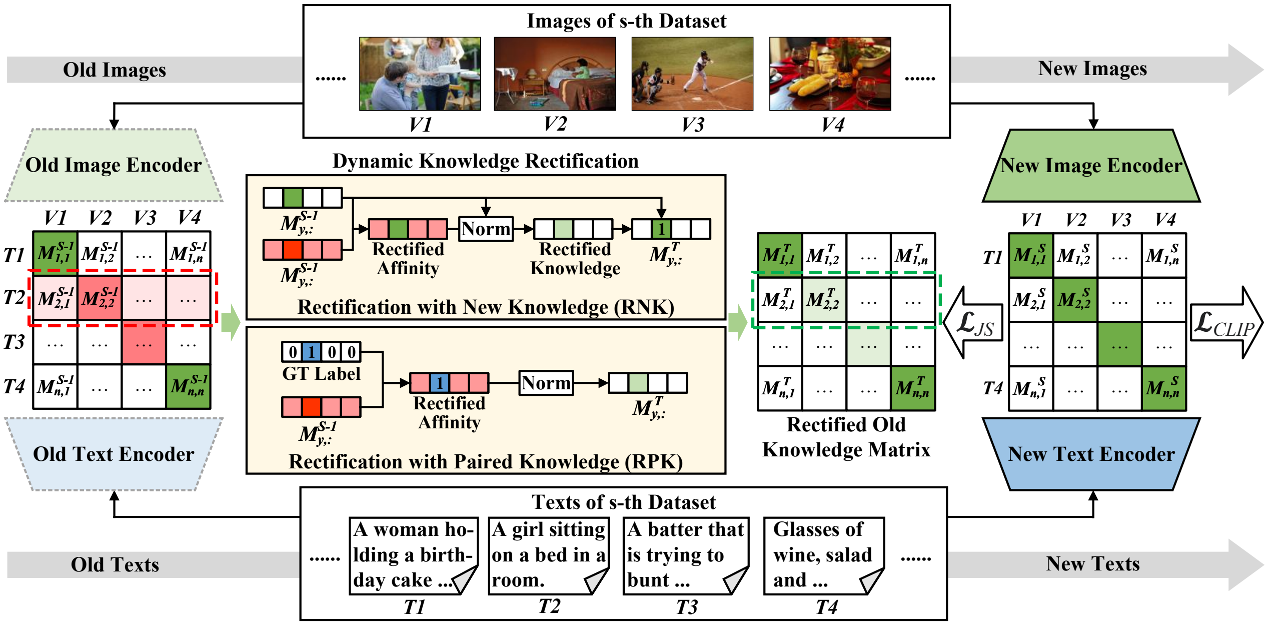


Figure 3: The framework of our proposed Dynamic Knowledge Rectification (DKR) model. DKR consists of three components: a CLIP-based image encoder, a CLIP-based text encoder, and a Dynamic Knowledge Rectification (DKR) module. The encoders embed input data (images and texts) and form two affinity matrices for old and new knowledge. The DKR module dynamically rectifies the incorrect old knowledge and distils it into the new model.

class prediction results to represent old knowledge, continual retrieval does not have specific class label information. Therefore, we first give a definition to the old knowledge space of the last training stage $(s-1)$ by an affinity matrix $\mathcal{M}^{(s-1)} \in \mathbb{R}^{N\times N}$. For convenience, we take the text-to-image affinity matrix as an example. Then, $\mathcal{M}^{(s-1)}$ can be calculated as follows:

$$\mathcal{M}^{(s-1)}_{i,j} = \frac{e^{\langle \boldsymbol{z}_t^i, \boldsymbol{z}_v^j\rangle/\tau}}{\sum_{k=1}^{N} e^{\langle \boldsymbol{z}_t^i, \boldsymbol{z}_v^k\rangle/\tau}}, (i,j \in [1,N]). \tag{3}$$

Notably, $\boldsymbol{z}$ in Eq. 3 is calculated using the model trained after the $(s-1)$-th stage. Then, we impose a knowledge distillation $\mathcal{L}_{JS}$ based on Kullback-Leibler (KL) divergence to transfer knowledge from the old model to the new model, which can be calculated as follows:

$$\mathcal{L}_{KL}(\mathcal{M}^{(x)}, \mathcal{M}^{(y)}) = \sum \mathcal{M}^{(x)} log(\frac{\mathcal{M}^{(x)}}{\mathcal{M}^{(y)}}), \tag{4}$$

$$\begin{aligned} \mathcal{L}_{JS}(s-1,s) =& \mathcal{L}_{KL}(\mathcal{M}^{(s-1)}, \frac{\mathcal{M}^{(s-1)}+\mathcal{M}^{(s)}}{2}) \\ &+ \mathcal{L}_{KL}(\mathcal{M}^{(s)}, \frac{\mathcal{M}^{(s-1)}+\mathcal{M}^{(s)}}{2}), \end{aligned} \tag{5}$$

Note that the weight parameter in the old model is kept constant during the distillation.

However, the distillation in Eq. 5 only holds if the old knowledge is correct. When the old knowledge cannot correlate paired samples in the new dataset, it will naturally lead to negative transfer and prevent knowledge acquisition. Formally, we define elements on the diagonal of $\mathcal{M}$ without the largest affinity as the incorrect knowledge, ie: $i \neq \arg\max(\mathcal{M}_{i,1}, ..., \mathcal{M}_{i,N})$, while the others as the correct knowledge. Next, we elaborate on the proposed two strategies to rectify the incorrect old knowledge.

Rectification with New Knowledge

To prevent the aforementioned negative transfer problem, we replace the incorrect affinity in $\mathcal{M}^{(s-1)}$ with the corresponding correct affinity calculated by the new model. In order to protect the old feature space as much as possible, we only correct the affinity values on the diagonal as follows:

$$\mathcal{M}^{(T)}_{i,i} = \mathcal{M}^{(s)}_{i,i}, i \neq \arg\max(\mathcal{M}^{(s-1)}_{i,1}, ..., \mathcal{M}^{(s-1)}_{i,N}). \tag{6}$$

To promote the transfer of the rectified old knowledge to new knowledge, we use the newly introduced affinity to constrain the off-diagonal old affinity, which eliminates the gap between the old and new knowledge as follows:

$$\mathcal{M}^{(T)}_{i,j} = \mathcal{M}^{(s-1)}_{i,j} \times \frac{1-\mathcal{M}^{(s)}_{i,i}}{1-\mathcal{M}^{(s-1)}_{i,i}}, i \neq j. \tag{7}$$

Although the rectified knowledge calculated by Eq. 6 and Eq. 7 achieve anti-forgetting by exploiting regularized old off-diagonal affinity, it meanwhile ignores the acquisition of new knowledge. Therefore, we further enhance the diagonal affinity by:

$$\mathcal{M}^{(T)}_{i,i} = 1, \tag{8}$$

which is crucial to balance the adaptation to the new dataset.

Based on the above strategy, DKR achieves anti-forgetting by maintaining the relative relationship between affinities in the old model. Nevertheless, such a strategy is still undesirable when it comes to incorrect new knowledge. For example, when training on a new dataset that differs significantly from the old datasets, the new model can hardly output the correct affinity in time. In this case, such rectification with new knowledge is ineffective and even harmful due to the new model subject to accumulating erroneous knowledge.

Rectification with Paired Knowledge

When encountering both erroneous old and new knowledge, the model will significantly deviate from the learned knowledge and finally lead to performance degradation. To tackle the above issue, we use the known paired GT label to realize the rectification.

Derived from Eq. 6, we replace the incorrect affinity in the old knowledge with the paired label which is correct for both old and new datasets as follows:

$$\mathcal{M}^{(T)}_{i,i} = 1, i \neq \arg\max(\mathcal{M}^{(s,s-1)}_{i,1}, ..., \mathcal{M}^{(s,s-1)}_{i,N}). \tag{9}$$

Then, we constrain the learning of the current stage based on the correlation between the off-diagonal affinity and the label information in the old feature space, which can be formulated as follows:

$$\mathcal{M}^{(T)}_{i,j} = \begin{cases} \dfrac{1}{1 + \sum_{k \neq i}^{N} \mathcal{M}^{(s-1)}_{i,k}}, i = j \\ \dfrac{\mathcal{M}^{(s-1)}_{i,j}}{1 + \sum_{k \neq i}^{N} \mathcal{M}^{(s-1)}_{i,k}}, i \neq j. \end{cases} \tag{10}$$

Finally, the complete DKR is shown in Alg. 1. We combine the above strategies to obtain the rectified old knowledge matrix $\mathcal{M}^T$ and develop Eq. 5 to achieve the balance between the anti-forgetting of the old dataset and the adaptation to the new dataset. Particularly, $\mathcal{L}_{js}$ consists of both text-to-image distillation and image-to-text distillation. We sum up the above two distillation losses to achieve the balance in the vision-language retrieval task.

Objective Function

Finally, the total objective function of our proposed DKR is calculated as follows:

$$\mathcal{L} = \mathcal{L}_{clip} + \lambda \mathcal{L}_{JS}(T, s), \tag{11}$$

where λ is a hyperparameter for training.

Experiments

In this section, we conduct extensive experiments to validate the effectiveness of our proposed DKR.

Datasets and Evaluations

We conduct the validation experiments on five real-world benchmark image-text retrieval datasets: 1) **MS-COCO Caption**: MS-COCO Caption (MS-COCO) (Lin et al. 2014) is a widely used image caption dataset. It contains 80K training images and 5K testing images, where each image has five captions. 2) **Flickr30K**: Flickr30K (Young et al. 2014) contains 31,783 images from the Flickr website, and each image is annotated by 5 sentences. We use 30K images as the training set and the rest 1K images as the testing set. 3) **IAPR TC-12**: IAPR TC-12 (Grubinger et al. 2006) consists of 20,000 images with corresponding captions collected around the world. We use 15K images for training and the rest 5K images for testing. 4) **ECommerce-T2I**: ECommerce-T2I (EC) (Yang et al. 2021) is a large-scale e-commerce products retrieval dataset. It contains 90K images for training and 5K images for testing, where each image is annotated with one sentence. 5) **RSICD**: RSICD (Lu et al. 2017) is a remote sensing image retrieval dataset, which contains 10,921 images of 30 scenes. We use 9,828 images for training and the rest 1,093 images for testing.

Algorithm 1: Dynamic Knowledge Rectification (DKR)

Input: Old affinity matrix $\mathcal{M}^{(s-1)}$, New affinity matrix $\mathcal{M}^{(s)}$.
Parameter: Batch size N.
Output: Rectified affinity matrix $\mathcal{M}^{s-1}$.

```
for $i = 1$ to $N$ do
    if $i = \arg\max(\mathcal{M}^{(s-1)}_{i,1}, ..., \mathcal{M}^{(s-1)}_{i,N})$ then
        $\mathcal{M}^T_i = \mathcal{M}^{(s-1)}_i$;
    else if $i = \arg\max(\mathcal{M}^{(s)}_{i,1}, ..., \mathcal{M}^{(s)}_{i,N})$ then
        $\mathcal{M}^T_{i,j} = \mathcal{M}^{(s-1)}_{i,j} \times \frac{1 - \mathcal{M}^{(s)}_{i,i}}{1 - \mathcal{M}^{(s-1)}_{i,i}}, i \neq j$;
        $\mathcal{M}^T_{i,i} = 1$;
    else
        $\mathcal{M}^{s-1}_{i,i} = 1$;
        $\mathcal{M}^T_i = \frac{\mathcal{M}^{s-1}_i}{1 + \sum_{k \neq i}^{N} \mathcal{M}^{(s-1)}_{i,k}}$
    end if
end for
return $\mathcal{M}^T$
```

Considering that the CVLR data may come from a specific dataset or different datasets, we verify the effectiveness of our DKR under two settings. **Setting-1**: To evaluate the effectiveness of our proposed DKR on different datasets, we conducted experiments on five sequential given datasets (i.e., MS-COCO → Flickr30K → IAPR TC-12 → EC → RSICD). **Setting-2**: To further evaluate the performance on the specific dataset, we follow the benchmark in (Ni et al. 2023), which randomly and uniformly divides the EC dataset into 5 sub-datasets, and sequentially train on these 5 sub-datasets. The test dataset includes Flickr30K, MS-COCO, and EC.

We employ the widely used evaluation metrics in cross-modal retrieval, Recall at Top K (R@K), to evaluate and compare our method with the existing methods under the same setting for fair comparisons.

Method	MS-COCO			Flickr30K			IAPR TC-12			EC			RSICD			Average		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Joint	53.7	78.4	86.4	80.9	95.6	98.1	53.6	82.6	90.5	18.6	43.7	57.0	12.0	31.3	45.6	43.7	66.3	75.5
Zero-Shot	39.8	64.4	74.5	68.3	89.0	94.2	36.0	65.8	76.6	11.0	27.7	37.9	5.2	15.7	26.3	32.1	52.5	61.9
SFT	39.0	65.5	75.3	69.7	90.1	94.3	47.0	77.4	87.0	14.9	37.1	50.1	12.8	33.5	47.9	36.7	60.7	70.9
EWC	39.9	66.6	76.5	70.0	90.6	94.5	47.3	77.3	87.0	15.0	37.2	50.3	13.3	32.5	47.5	37.1	60.8	71.1
LwF	47.0	72.9	82.1	76.2	93.6	97.1	**54.6**	**83.6**	**91.2**	17.0	40.2	53.8	14.0	**35.0**	**49.7**	41.7	65.0	74.8
ERL	48.3	73.5	82.6	77.3	93.6	96.7	51.0	80.5	88.8	18.2	42.1	**55.5**	14.1	33.8	48.2	41.8	64.7	74.3
AFC	48.3	73.7	82.6	77.7	93.6	96.7	51.2	80.7	89.0	**18.3**	42.2	55.4	**14.5**	33.3	47.7	42.0	64.7	74.3
Mod-X	49.7	73.8	82.6	77.8	93.7	97.1	51.1	80.7	89.0	18.2	**42.3**	**55.5**	**14.5**	33.5	47.8	41.9	64.8	74.3
Ours	**51.4**	**76.0**	**84.4**	**79.3**	**94.6**	**97.4**	53.9	83.1	90.8	18.0	42.0	54.8	14.1	34.3	49.2	**43.3**	**66.0**	**75.3**

Table 1: Comparison with state-of-the-art methods on different datasets (Training order: MS-COCO→Flickr30K→IAPR TC-12→EC→RSICD. The best results are bolded and the second-best results are underlined.)

Method	Image-Text Retrieval						Text-Image Retrieval						Average	
	Flickr30K		MS-COCO		EC		Flickr30K		MS-COCO		EC			
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
Joint	64.5	88.6	39.8	64.8	23.5	50.8	46.9	73.1	22.2	44.5	23.5	50.6	36.7	62.1
Zero-Shot	77.7	94.5	50.1	74.6	11.3	27.6	58.9	83.5	30.2	55.6	10.1	25.5	39.7	60.2
SFT	63.4	87.2	36.8	61.5	16.6	40.7	44.4	71.0	20.6	42.6	15.8	40.5	32.9	57.3
EWC	64.0	87.8	37.7	64.3	16.2	40.0	44.8	72.4	20.7	44.1	16.5	42.0	33.3	58.4
Mod-X	73.1	92.1	47.1	70.5	20.1	44.8	55.6	79.9	27.9	51.0	20.0	44.8	40.6	63.9
Ours	**78.5**	**95.7**	**51.7**	**75.4**	**20.4**	**46.2**	**58.7**	**83.6**	**29.7**	**54.2**	**20.2**	**45.3**	**43.2**	**66.7**

Table 2: Comparison with state-of-the-art methods on the specific dataset. (Training order: EC-1→EC-2→EC-3→EC-4→EC-5. The best results are bolded and the second-best results are underlined.)

Implementation Details

The proposed DKR is implemented in PyTorch with NVIDIA V100 GPUs. We use CLIP(ViT-Based/32) (Radford et al. 2021) with the pre-trained weight on the large-scale open-world datasets in (OpenAI. 2021) as our backbone. Input images are resized to 224 × 224. Each task is trained with 35 epochs with a batch size of 280. We use Adam optimizer with (β_1,β_2)=(0.9, 0.99) and weight decay of 0.2 to update the whole CLIP. The initial learning rate is set to 1e-6 with 20% warm-up iterations, and a cosine-decay learning rate scheduler is also used to update the whole framework. The hyperparameter λ is set to 1.0 and 0.1 for Setting-1 and Setting-2, respectively.

Comparison with State-of-the-art Methods

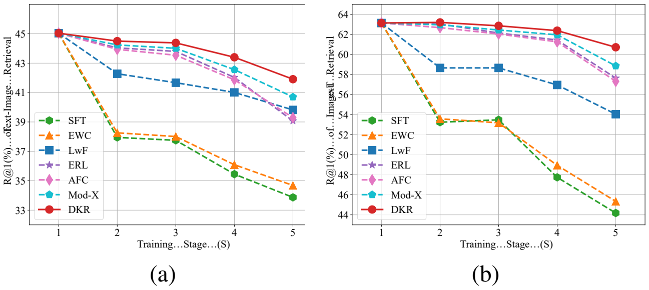
In this section, we compare our DKR to five continual learning methods that do not rely on class label information: EWC (Elastic Weight Consolidation) (Kirkpatrick et al. 2017), LwF (Learning Without Forgetting) (Li and Hoiem 2017), AFC (Adaptive Feature Consolidation) (Kang, Park, and Han 2022), ERL (Exemplars Relation Loss) (Dong et al. 2021) and Mod-X (Maintain Off-diagonal information matriX) (Ni et al. 2023). In addition, we report three basic training strategies: Joint, Zero-Shot, and SFT. Joint represents that all data are available at any time, which is an upper bound. Zero-shot represents directly testing the original pre-trained CLIP. SFT represents sequentially training and testing the model without any anti-forgetting strategies.

Comparison on Different Datasets Tab. 1 summarizes the results of our DKR on five different datasets. Compared with existing methods, DKR ranks first on all average metrics of I2T and T2I and achieves 43.4%, 66.0%, and 75.3% on R@1, R@5 and R@10, respectively. In particular, on the MS-COCO dataset, DKR outperforms Mod-X by 1.7%, 2.2%, and 1.8% on R@1, R@5 and R@10, which illustrates that our DKR can effectively alleviate the catastrophic forgetting problem. Note that existing methods traded expensive performance degradation in old datasets (e.g., MS-COCO and Flickr30K) for limited improvements in new datasets (e.g., EC and RSICD). Fig. 4 illustrates the R@1 accuracy on MS-COCO after each training step. It can be seen that our DKR achieves the best anti-forgetting performance, which significantly outperforms existing methods. The above results illustrate the superiority of our DKR against state-of-the-art methods.

Comparison on the EC Dataset As shown in Tab. 2, our DKR achieves the best retrieval performance on the EC dataset. Specifically, DKR achieves the highest retrieval performance on the EC dataset, i.e. achieves 20.4% and 20.2% on R@1 in Image-Text Retrieval (I2T) and Text-Image Retrieval tasks (I2T). Meanwhile, our DKR significantly suppresses Mod-X in the old datasets (Flickr30K and MS-COCO) on all criteria. The above results illustrate that DKR can effectively preserve the old knowledge for anti-forgetting. Note that although DKR is not directly trained on Flickr30K and MS-COCO, it can even achieve higher results than CLIP (Zero-Shot) on the Image-Text Retrieval task. This is because our DKR method can achieve positive transfer for both old and new knowledge by properly filtering and rectifying the incorrect knowledge.

Base	RNK	RPK	RNK*	MS-COCO	Flickr30k	IAPR TC-12	EC	RSICD	Average
✓				49.0	77.9	51.9	15.8	11.1	41.2
✓	✓			50.7	78.7	53.4	17.3	12.7	42.6
✓		✓		49.4	78.1	52.8	16.7	12.7	41.9
✓		✓	✓	51.1	79.0	53.8	17.2	12.2	42.9
✓	✓	✓		**51.4**	**79.3**	**53.9**	**18.0**	**14.1**	**43.3**

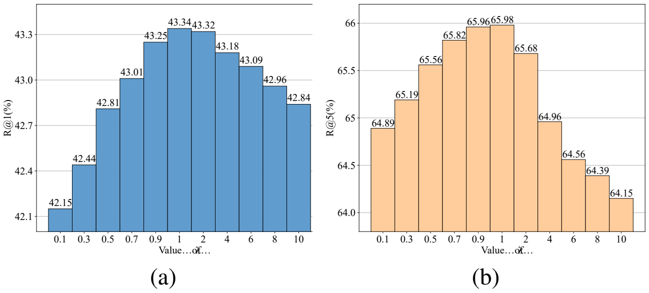
Table 3: Ablation studies of each component of DKR under Setting-1, where average R@1 accuracy is reported. (The best results are bolded and the second-best results are underlined.)



(a)

(b)

Figure 4: The retrieval performance of different methods in each training stage under Setting-1 on MS-COCO, where Text-Image and Image-Text retrieval results are shown in (a) and (b), respectively.



(a)

(b)

Figure 5: The effects of hyperparameter λ under Setting-1, where averaged R@1 and R@5 accuracy (%) are shown in (a) and (b), respectively.

Ablation Study

In this section, we conduct ablation studies under Setting-1 to evaluate the effectiveness of each component of DKR and the effect of the hyperparameter.

hyperparameter Study We first evaluate the effect of the hyperparameter λ in Eq. 11 under Setting-1. The average R@1 and R@5 accuracy on all five sub-datasets are shown in Fig. 5. With λ increasing, the average R@1 and R@5 keep improving before λ arrives at 1.0. This is because a slight λ (< 1) cannot alleviate the catastrophic forgetting of old knowledge, while an excessive λ (> 1) will limit the acquisition of new knowledge. It can be seen that our DKR strikes a balance of the above two states when λ=1 in a general scene.

Effectiveness of Rectification with New Knowledge To evaluate the effectiveness of our Rectification with New Knowledge (RNK), we introduce two variants to carefully evaluate the impact of the new knowledge, i.e. RNK (full version) and RNK* (RNK w/o Eq. 8). As shown in Tab. 3, compared to the Base, RNK significantly improves the average performance by 1.4%. When combined with RPK, our final model achieves a significant improvement of 2.1%. This shows that by rectifying the incorrect old knowledge with new knowledge, DKR significantly improves the anti-forgetting of old knowledge. In addition, compared to the other variant (RNK*), RNK gains an improvement on new datasets (+0.8% on EC and +1.9% on RSICD). These results indicate that the enhancement to new knowledge (w.r.t. Eq. 8) improves the performance on new datasets while maintaining the anti-forgetting of old knowledge.

Effectiveness of Rectification with Paired Knowledge We compare the proposed Rectification with Paired Knowledge (RPK) to the basic knowledge distillation strategy without any rectification (Base). As shown in Tab. 3, compared with the Base, RPK brings an improvement by 0.7% on average R@1 accuracy on all datasets. When combined with RNK, the performance on each dataset is further improved by 0.7% on average R@1 accuracy. It indicates that by rectifying with additional knowledge, RPK promotes the acquisition of both old and new knowledge and thus improving the overall performance, verifying its high effectiveness.

The above results verify that our DKR is essential for striking the balance between the anti-forgetting of old knowledge and the acquisition of new knowledge.

Conclusion

In this paper, we proposed a novel framework for the Continual Vision-Language Retrieval (CVLR) task, called Dynamic Knowledge Rectification (DKR), which alleviates the catastrophic forgetting problem by dynamically filtering and rectifying incorrect old knowledge. Specifically, we designed a knowledge rectification method to achieve the anti-forgetting of the old knowledge and the acquisition of both old and new knowledge. Extensive experiments on five vision-language retrieval benchmark datasets demonstrate that our DKR achieves state-of-the-art performance.

Ackowledgements

This work was supported by the National Natural Science Foundation of China (61925201, 62132001, 62376011).

References

Chen, H.; Ding, G.; Liu, X.; Lin, Z.; Liu, J.; and Han, J. 2020a. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12655–12663.

Chen, J.; Hu, H.; Wu, H.; Jiang, Y.; and Wang, C. 2021. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15789–15798.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020b. Uniter: Universal image-text representation learning. In *European conference on computer vision*, 104–120. Springer.

Chowdhury, J. R.; Zhuang, Y.; and Wang, S. 2022. Novelty controlled paraphrase generation with retrieval augmented conditional prompt tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 10535–10544.

De Lange, M.; Aljundi, R.; Masana, M.; Parisot, S.; Jia, X.; Leonardis, A.; Slabaugh, G.; and Tuytelaars, T. 2021. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7): 3366–3385.

Dong, S.; Hong, X.; Tao, X.; Chang, X.; Wei, X.; and Gong, Y. 2021. Few-shot class-incremental learning via relation knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1255–1263.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2023a. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 1–15.

Gao, X.; He, Y.; Dong, S.; Cheng, J.; Wei, X.; and Gong, Y. 2023b. DKT: Diverse Knowledge Transfer Transformer for Class Incremental Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24236–24245.

Grubinger, M.; Clough, P.; Müller, H.; and Deselaers, T. 2006. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International workshop ontoImage*, volume 2.

Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2021. Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. In *International Conference on Learning Representations*.

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

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

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv:2301.12597.

Li, K.; Zhang, Y.; Li, K.; Li, Y.; and Fu, Y. 2019. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4654–4662.

Li, W.; Gao, C.; Niu, G.; Xiao, X.; Liu, H.; Liu, J.; Wu, H.; and Wang, H. 2021. UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2592–2607.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, 121–137. Springer.

Li, Z.; and Hoiem, D. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12): 2935–2947.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, X.; Yi, J.; Cheung, Y.-m.; Xu, X.; and Cui, Z. 2022. OMGH: Online manifold-guided hashing for flexible cross-modal retrieval. *IEEE Transactions on Multimedia*.

Lu, X.; Wang, B.; Zheng, X.; and Li, X. 2017. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4): 2183–2195.

Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; and Li, T. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508: 293–304.

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

Ni, Z.; Wei, L.; Tang, S.; Zhuang, Y.; and Tian, Q. 2023. Continual Vision-Language Representaion Learning with Off-Diagonal Information. arXiv:2305.07437.

OpenAI. 2021. CLIP. https://github.com/openai/CLIP. Accessed: 2023-08-16.

Ordonez, V.; Kulkarni, G.; and Berg, T. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24.

Qiu, B.; Li, H.; Wen, H.; Qiu, H.; Wang, L.; Meng, F.; Wu, Q.; and Pan, L. 2023. CafeBoost: Causal Feature Boost To Eliminate Task-Induced Bias for Class Incremental Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16016–16025.

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

Schuhmann, C.; Kaczmarczyk, R.; Komatsuzaki, A.; Katta, A.; Vencu, R.; Beaumont, R.; Jitsev, J.; Coombes, T.; and Mullis, C. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. In *NeurIPS Workshop Datacentric AI*, FZJ-2022-00923. Jülich Supercomputing Center.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2556–2565.

Srinivasan, T.; Chang, T.-Y.; Pinto Alva, L.; Chochlakis, G.; Rostami, M.; and Thomason, J. 2022. Climb: A continual learning benchmark for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 35: 29440–29453.

Wang, H.; Zhang, Y.; Ji, Z.; Pang, Y.; and Ma, L. 2020. Consensus-aware visual-semantic embedding for image-text matching. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, 18–34. Springer.

Wang, K.; Herranz, L.; and van de Weijer, J. 2021. Continual learning in cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3628–3638.

Wang, Z.; Zhang, Z.; Lee, C.-Y.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 139–149.

Wu, W.; Luo, H.; Fang, B.; Wang, J.; and Ouyang, W. 2023. Cap4Video: What Can Auxiliary Captions Do for Text-Video Retrieval? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10704–10713.

Xu, M.; Zhang, Z.; Wei, F.; Lin, Y.; Cao, Y.; Hu, H.; and Bai, X. 2022. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, 736–753. Springer.

Yang, A.; Lin, J.; Men, R.; Zhou, C.; Jiang, L.; Jia, X.; Wang, A.; Zhang, J.; Wang, J.; Li, Y.; et al. 2021. M6-t: Exploring sparse expert models and beyond. arXiv:2105.15082.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. 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.

Zhang, R.; Zhang, W.; Fang, R.; Gao, P.; Li, K.; Dai, J.; Qiao, Y.; and Li, H. 2022. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European Conference on Computer Vision*, 493–510. Springer.