Continual Vision-Language Representation Learning with Off-Diagonal Information

Zixuan Ni [1 *] **Longhui Wei** [2] **Siliang Tang** [1 †] **Yueting Zhuang** [1] **Qi Tian** [2 †]

Abstract

Large-scale multi-modal contrastive learning frameworks like CLIP typically require a large amount of image-text samples for training. However, these samples are always collected continuously in real scenarios. This paper discusses the feasibility of continual CLIP training using streaming data. Unlike continual learning based on self-supervised learning methods for pure images, which is empirically robust against catastrophic forgetting, CLIP's performance degeneration in the continual setting is significant and non-neglectable. By analyzing the changes in the model's representation space during continual CLIP training from a spatial geometry perspective, we explore and summarize these spatial variations as **Spatial Disorder (SD)**, which can be divided into **Intra-modal Rotation** and **Inter-modal Deviation**. Moreover, we empirically and theoretically demonstrate how SD leads to a performance decline for CLIP on cross-modal retrieval tasks. To alleviate SD, we propose a new continual vision-language representation learning framework **Mod-X**: **M**aintain **o**ff-**d**iagonal information-matri**X**. By selectively aligning the off-diagonal information distribution of contrastive matrices, the Mod-X improves the capability of the multi-modal model by maintaining the multi-modal representation space alignment on the old data domain during continuously fitting the new training data domain. Experiments on commonly used datasets with different scales and scopes have demonstrated the effectiveness of our method.

[*]Work done when interning at Huawei Cloud.;[†]Corresponding author. [1]Zhejiang University [2]Huawei Cloud. Correspondence to: Siliang Tang <siliang@zju.edu.cn>, Qi Tian <tian.qi1@huawei.com>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

1. Introduction

Recently, multi-modal pre-trained models such as CLIP (Radford et al., 2021) have attracted much attention. By utilizing these pre-trained models, many works have achieved new progress in downstream tasks such as image classification, semantic segmentation, object detection, speech recognition (Wei et al., 2022; Wang et al., 2021b; Li et al., 2022; Baevski et al., 2020), etc. Although the CLIP model has strong generalization in open-world data, as mentioned in its original paper (Radford et al., 2021), the ability to match image-text samples that are not in its training data distribution is still weak. The natural idea to alleviate this problem is to scale up the training data that covers different data domains. However, it is impractical to train infinite data distribution with limited hardware resources at once.

To address the above problems, this paper mainly explores the feasibility of continuously training the CLIP model through streaming data, a training paradigm that follows Continual Learning (CL) (McCloskey & Cohen, 1989). Traditional supervised continual learning has been proven to suffer from catastrophic forgetting (Rebuffi et al., 2017; Kirkpatrick et al., 2017): The model's performance on old tasks drops significantly as training phases rising. Recently, some works (Ni et al., 2021b; Hu et al., 2021) have validated that self-supervised models based on pure images like SimCLR (Chen et al., 2020) and BarlowTwins (Zbontar et al., 2021) do not suffer from severe catastrophic forgetting during continual training. Some works (Madaan et al., 2021; Thai et al., 2021) conjecture that the reason is that the contrastive loss is not directly affected by the supervised signal, and the self-supervised framework does not have a Softmax function to amplify the influence of labels.

However, the performance of CLIP with a continual training setting is clearly different from the self-supervised continual training, which only uses images, though they both utilize contrastive loss. There is a significant degradation of multi-modal retrieval results with continual CLIP training compared with joint training (the experiment results are shown in Section 3 and 5). By analyzing the changes in the model's representation space during continual CLIP training from a spatial geometry perspective, we explore and summarize these spatial variations as **Spatial Disorder (SD)**, which

can be divided into **Intra-modal Rotation** and **Inter-modal Deviation**. The intra-modal rotation represents the representation space of the single-modal feature extractor (vision or language) within the CLIP rotates around the center of the high-dimensional sphere. The inter-modal deviation represents the shift of representation alignment of different modal extractors (vision and language) to the same entities during continual training. Moreover, we demonstrate how intra-modal rotation and inter-modal deviation lead to a performance decline for CLIP on cross-modal retrieval tasks in both empirically and theoretically (in Section 3).

To alleviate this SD in continual CLIP training, we propose a simple yet effective framework Mod-X: **M**aintain **o**ff-**d**iagonal information-matri**X**. Unlike contrastive loss (Oord et al., 2018) only focuses on widening the similarity gap between positive and negative sample pairs, the Mod-X framework pays more attention to representation space alignment. The elements in the contrastive matrix represent the similarity between visual and textual entities, which also refer to the included angle between visual and textual representation vectors when the length of vectors is **1**. The angle distribution between the vectors represents the inherent representation space structure of the model under the current samples. By selectively aligning the distribution of the off-diagonal elements, Mod-X preserves the spatial relationships between modals of various old entities while fitting the current vision-language data during continual training. The experiments (in Section 4, 5 and Appendix B) on commonly used datasets with different scales and scopes show that our Mod-X framework improves the capability of the multi-modal model by maintaining the multi-modal representation space alignment on the old data domain during continuously fitting the new training data domain. The contributions of this paper are summarized as follows:

- We discuss the feasibility of training the CLIP model continuously through streaming data. Empirical experiments demonstrate that continual CLIP training leads to persistent performance degrades on cross-modal retrieval tasks, which is different from the phenomenon of continual learning based on self-supervised learning methods for pure images.
- We explore and summarize the model's spatial variation during continual CLIP training as Spatial Disorder, which can be divided into intra-modal rotation and inter-modal deviation. Furthermore, we demonstrate how spatial disorder leads to a performance decline for CLIP on cross-modal retrieval tasks in both empirically and theoretically (in Section 3).
- We propose a simple yet effective continual CLIP training framework **Mod-X** that alleviates space disorder during continual CLIP training by selectively aligning contrastive matrices' off-diagonal information. Experiments (in Section 5 and Appendix B) on commonly used datasets with different scales and scopes have evaluated the effectiveness of our method.

2. Related Work

Continual Learning. Continual learning (CL) (Thrun, 1995), or incremental learning, mainly focuses on supervised tasks. In addition to the vision-based tasks (De Lange et al., 2021; Kj et al., 2021; Cha et al., 2021; Ahn et al., 2021), some works discuss language-based tasks (Biesialska et al., 2020; Sun et al., 2019). We can summarize the existing continual learning methods into three categories: regularization (Kirkpatrick et al., 2017; Ahn et al., 2019; Ni et al., 2021a), replay (Rebuffi et al., 2017; Rolnick et al., 2019; Wang et al., 2021a), and architecture (Thai et al., 2021; Ni et al., 2021b; Hu et al., 2021; Madaan et al., 2021). However, traditional supervised continual learning methods are limited by labels and unsuitable for self-supervised or unsupervised situations.

In unsupervised and self-supervised single-modal continual training, the latest work (Thai et al., 2021; Ni et al., 2021b; Hu et al., 2021; Madaan et al., 2021) has drawn some conclusions different from those of supervised continual learning. However, only a few pieces (Srinivasan et al., 2022; Fan et al., 2022) focus on incremental multi-modal learning. However, (Srinivasan et al., 2022) did not propose a new method to alleviate the catastrophic forgetting problem in multimodal continual learning. Instead, it provided baselines for the state-of-the-art supervised unimodal continual learning methods when applied to some single-model multimodal tasks. And (Fan et al., 2022) discussed continuous updates in multimodal graphs, which is far from our intended goal of continuously updating multimodal pre-trained models. Because of the cooperation between different modalities, continual multi-modal pre-training shows different performance and complex problems from single-modal continual training.

Visual-Language Representational Learning. Vision-language representation learning based on contrastive loss (Oord et al., 2018), such as CLIP (Radford et al., 2021), has attracted a lot of attention in various fields (Radford et al., 2021; Li et al., 2021; Andonian et al., 2022). And the pre-trained model performs surprisingly well on downstream tasks (Shu et al., 2022; Wang et al., 2022; Chowdhury et al., 2022). At the same time, the large-scale image-text datasets, e.g., Laion400M (Schuhmann et al., 2021) and Conceptual Captions (Sharma et al., 2018), have played a key role in multimodal pre-training. Although large-scale open-world datasets contain various samples, the pre-trained model still loses the ability to perfectly match image-text sample pairs that are not in its training data domain (Radford et al., 2021).

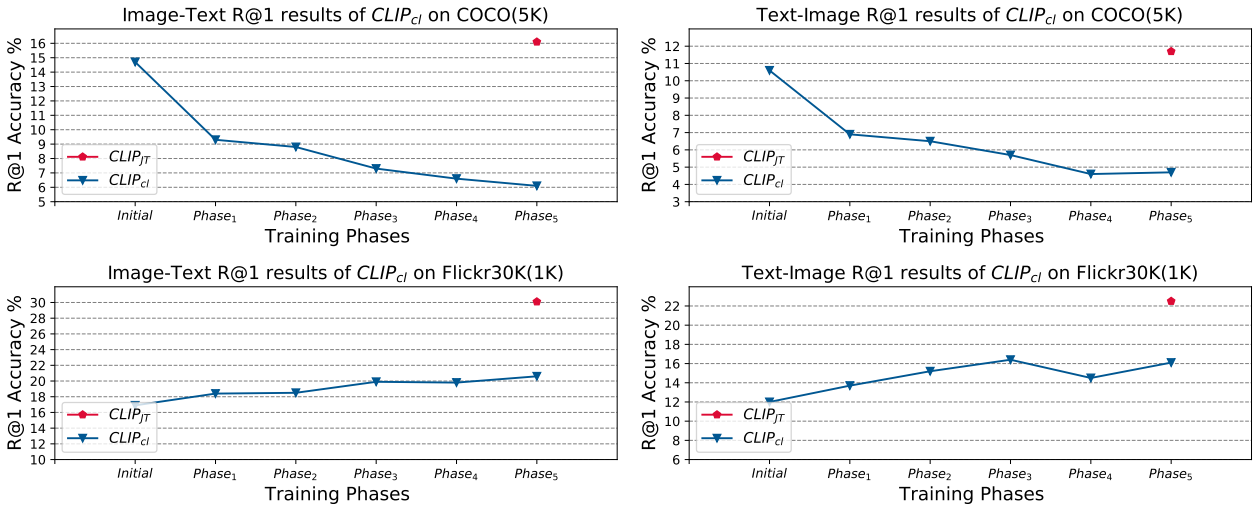


Figure 1. The multi-modal retrieval $R@1$ results of CLIP$_t$ $(0 \leq t \leq 5)$ on test sets COCO (5K) and Flickr30k (1K). The two sub-figures on the left show the Image-Text retrieval $R@1$ performance of CLIP$_t$ on the continual training phase t. The initial training phase represents the performance of CLIP$_0$. The rights show the Text-Image $R@1$ results of CLIP$_t$ on the continual training phase t. The pentagon points (CLIP$_{jt}$) show the results of the CLIP under joint training, which is an upper bound for continual CLIP training (CLIP$_{ct}$).

3. Spatial Disorder in Continual CLIP

This section mainly aims to explore the characteristics of the CLIP model while training continually. By analyzing the changes in the model's representation space from a spatial geometry perspective during continual CLIP training, we explore and summarize these spatial variations as Spatial Disorder (SD), which can be divided into intra-modal rotation and inter-modal deviation. Then, we demonstrate how intra-modal rotation and inter-modal deviation lead to a performance decline for CLIP on cross-modal retrieval tasks in both empirically and theoretically, respectively.

Exploration Setup. To ensure the controllability of the exploration, we train a CLIP$_0$ model from scratch on the COCO dataset (Lin et al., 2014) based on the OpenAI source code (OpenAI) and use it as the initial state (start) of continual CLIP training. After that, we divide the Flickr30K dataset (Young et al., 2014) into five sub-datasets $\{D_1, D_2, D_3, D_4, D_5\}$ uniformly and randomly to simulate streaming data. Then we train the CLIP$_0$ based on these sub-datasets sequentially. We name this pure continual training without other operations as CLIP$_{ct}$. After finishing five training phases, we obtain the model CLIP$_5$. For comparison with CLIP$_5$, we joint training a CLIP$_{jt}$ model using joint dataset COCO and Flickr30K, as the upper bound of the CLIP$_5$. The hyper-parameters for all training phases are kept the same, and detailed settings of CLIP model and training hyper-parameters can be seen in Appendix B.1.

3.1. The Performance of Continual CLIP Training

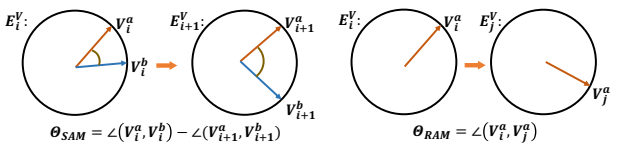
We show the $R@1$ retrieval results of CLIP$_t$ $(0 \leq t \leq 5)$ on the test set COCO(5K) and Flickr30K(1K) in Figure 1. By comparing the performances of the CLIP$_0$ (initial phase) and CLIP$_{jt}$ on Flickr30K(1K), we can find that the retrieval performance of CLIP$_{jt}$ (red point) is significantly better than that of CLIP$_0$ (initial) which is not trained on Flickr30K. This phenomenon shows that **the performance of the CLIP model is affected by the training data domain**, which is consistent with the conclusion of the paper (Radford et al., 2021). Besides this, it can be clearly seen that the multi-modal retrieval performance of the CLIP$_{ct}$ on the COCO(5K) declines continually with the rising of training phases. The final Image-Text R@1 result of CLIP$_5$ on COCO(5K) plummeted from the initial 14.7% to 6.1%, and the Text-Image results dropped from 10.6% to 4.7%. The gap with CLIP$_{jt}$ reached 10.0% and 7.0%, respectively. On the other hand, CLIP$_{ct}$ exhibits a slow and erratic increase in multi-modal retrieval $R@1$ results on the test set Flickr30K(1K). Although the results between CLIP$_{ct}$ and CLIP$_{jt}$ on the Image-Text $R@1$ has been narrowed from the original 13.2% to 9.5% while the Text-Image $R@1$ of CLIP$_{ct}$ has increased from 12.0% to 16.1%, the gap between CLIP$_5$ and CLIP$_{jt}$ is still great.

3.2. The reasons for catastrophic forgetting

In CLIP, the vision and language encoders normalize the final representation vector to a unit vector of length 1 using a dimension-based L_2 norm. This design makes the representation space in vision and language encoders form a high-dimensional unit sphere, respectively. Therefore, we ignore the influence of the representation vectors' length and track their direction changes. We summarize these spatial variations as Spatial Disorder (SD), which can be divided into intra-modal rotation and inter-modal deviation.

3.2.1. THE INTRA-MODAL ROTATION

Firstly, we analyze the directional changes of the representation vectors of model's vision and language extractors during continual CLIP training. Taking the visual representation space as an example, we use the visual encoder E_i^V in CLIP$_i$ to extract the image representations of the test set COCO(5K) and obtain the vision representation vectors sets $V_i = \{V_i^0, ..., V_i^N, ..., V_i^{5K}\}$, where $i = 0, ..., 5$ stands for five different training phases. After that, we take the inner product of each pair of vectors $< V_i^a, V_i^b >$, where a and b are arbitrary indexes in each vector set V_i and perform $arccos$ operation, the inverse trigonometric function of cosine, to obtain their **S**elf-**A**ngle relationship **M**atrix (SAM_i). The $SAM_i^{(a,b)} = arccos(< V_i^a, V_i^b >)$. Any element $SAM_i^{(a,b)}$ in the SAM_i matrix represents the included angle between the sample a and b in the vision encoder E_i^V. By counting the difference value $\theta_{SAM} = \angle(V_i^a, V_i^b) - \angle(V_{i+1}^a, V_{i+1}^b)$ between the corresponding elements in two continual SAM matrix SAM_i and SAM_{i+1} as shown in Figure 2(a), we get the following included angle change distribution in Figure 2(c).



(a) (b)

$\theta_{SAM} \in$	$[0°, 5°]$	$(5°, 10°]$	$(10°, 15°]$	$(15°, 20°]$	$(20°, 180°]$
SAM_{0-1}	54.23%	31.64%	11.28%	2.48%	0.38%
SAM_{1-2}	61.18%	30.21%	7.54%	0.99%	0.07%
SAM_{2-3}	61.44%	30.00%	7.50%	0.98%	0.07%
SAM_{3-4}	55.33%	31.75%	10.57%	2.07%	0.28%
SAM_{4-5}	50.51%	32.14%	13.17%	3.50%	0.68%
SAM_{0-5}	42.94%	31.12%	16.73%	6.66%	2.55%

(c)

Figure 2. The sub-figure on the (a) shows a schematic diagram of computing θ_{SAM}. The sub-figure on the (b) shows a schematic diagram of computing θ_{RAM}. The table on the bottom (c) shows the distribution of the change of the included angle between any two samples in different training phases' vision representation space. And $SAM_{i-j} = |SAM_i - SAM_j|$.

From Figure 2(c), we can find that 80% of the angle changes between any two vision representation vectors are between 0 and 10 degrees in continual training phases, while only 20% are above 10 degrees. Moreover, less than 1% of the angle changes are above 20 degrees. That angle changes between 15-20 degrees also only account for about 5% of all image pairs. Therefore, we conclude that **the topology of the visual representation of the CLIP$_{ct}$ changes slowly during the continual CLIP training.** In Appendix A.3, we reached the same empirical conclusion by comparing the representation quality of vision encoders.

In addition to discussing the change in the included angle between sample pairs in the visual representation space, by taking the inner product of the same sample's vision representation vector from different training phases' vision encoder E_i^V, we use the $arccos$ operation to compute the rotation angles $\theta_{RAM} = \angle(V_i^a, V_j^a)$ of each test sample a in vision encoder E_i^V and E_j^V and get the **R**otation **A**ngle **M**atrix RAM$_{(i,j)}$. The $RAM_{(i,j)}^a = arccos(< V_i^a, V_j^a >)$. The schematic diagram can be seen in Figure 2(b). By counting the distribution of rotation angles, we get the following rotation angle distribution Table 1.

As shown in Table 1, we can find that the direction of the same sample in the visual representation space of different training phases has changed greatly. Only less than 0.4% samples are rotated within 20 degrees in the continual CLIP training, while the samples rotated within 20-25 degrees are at most less than 9%, and the samples of 25 degrees and above account for more than 90%. We speculate that **the vision representation space of CLIP$_{ct}$ has undergone a large rotation around the high-dimensional sphere center during the continual training.** After analyzing the language representation space, we reach the same conclusion as the vision representation space. Detailed SAM and RAM of language encoders can be viewed in Appendix A.2.

According to our analysis of the geometric changes of the single-modal encoder's representation space during continual CLIP training, we conclude that: **During the continual CLIP training, the representation space in the CLIP$_{ct}$ is significantly rotated. The topology of the representation space is slightly rotated compared with the rotation of the whole representation space.** We name this phenomenon **Intra-modal Rotation**.

$\theta_{RAM} \in$	$[0°, 15°]$	$(15°, 20°]$	$(20°, 25°]$	$(25°, 30°]$	$(30°, 180°]$
$RAM_{(0,1)}$	0.00%	0.20%	4.62%	22.68%	72.50%
$RAM_{(1,2)}$	0.00%	0.40%	8.30%	34.11%	57.20%
$RAM_{(2,3)}$	0.00%	0.30%	8.40%	34.29%	57.01%
$RAM_{(3,4)}$	0.00%	0.00%	1.89%	17.11%	81.00%
$RAM_{(4,5)}$	0.00%	0.00%	0.00%	3.20%	96.81%
$RAM_{(0,5)}$	0.00%	0.00%	0.00%	0.21%	99.80%

Table 1. The table on the bottom shows the rotation angle distribution of the same samples in different training phases.

3.2.2. THE INTER-MODAL DEVIATION

Although the topology of the single-modal representation space changes during continual training, this slight rotation should not be the main reason for the significant degradation of CLIP's multi-modal retrieval performance in continual training. To this end, we conduct a thought experiment: it

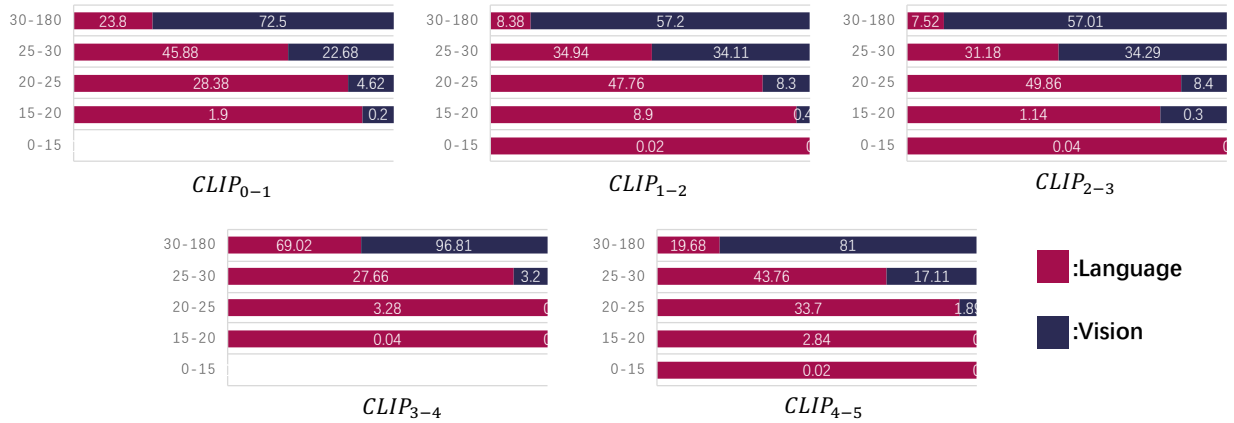


Figure 3. The comparison of the rotation distributions of the vision encoder and langugae encoder during continual CLIP training. CLIP$_{i-j}$ refers to the CLIP's continual training from training phase i to j. The values under the same color represent the proportion of test samples to total samples in each rotation angle interval of the same modality.

is known that the representation spaces of vision and language encoders exhibit significant spatial rotations during continual training. Now we assume that the topology of the single-modal representation space is completely fixed during continual training. Therefore, if the CLIP$_{ct}$'s performance on multi-modal retrieval tasks does not degrade during continual training, **the rotations of the two encoders' representation spaces should be synchronized**. However, the fact is the **opposite**. So we think **there is a deviation between the rotation of the vision and language representation spaces**. Based on this suppose, we compare the rotation distributions of vision encoder (Table 1) and language encoder (Appendix A.2) and draw the rotation distribution comparison diagram (Figure 3). The values under the same color represent the proportion of test samples to total samples in each rotation angle interval of the same modality. Comparing the difference in the distribution of rotation angles of the vision and language encoders, we can see that the space rotations of the two encoders are very different in the continual training. The rotation of language representation space is mostly concentrated between 20-30 degrees, while the vision's rotations are mostly between 30-180 degrees. This shows that the rotation of the representation space of the two modal extractors within CLIP$_{ct}$ is not synchronized during the continual training, which verifies our previous deduction: **The unsynchronized rotation of the vision and language representation spaces leads to representation space deviations between the CLIP's modal encoders (vision and language).** We name this phenomenon **Inter-modal Deviation**.

3.2.3. The Relationship between Spatial Disorder and Contrastive Matrix

How do spatial disorder cause the model to mis-align the old sample's vision and language representation? We show

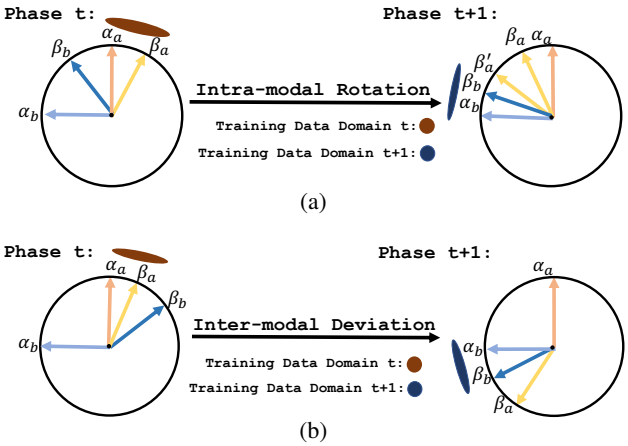


Figure 4. The Schematic illustration of spatial disorder caused by intra-modal rotation and inter-modal deviation. The α is vision representation and β is language representation. The a,b denote different image-text samples.

a schematic here to illustrate this. As shown in Figure 4, the α is vision representation and β is language representation. The a,b denote different image-text samples. For the convenience of illustration, we fix the vision vectors' relative location and rotate the language vectors to represent the unsynchronous rotation of the two modal spaces. When intra-modal rotation happens (Figure 4(a)), β_a in training phase $t+1$ is rotated to β'_a, the modal similarity between a and b shift from ($\beta_a^T\alpha_a > \beta_a^T\alpha_b$) to ($\beta'^T_a\alpha_a < \beta'^T_a\alpha_b$), which break the alignment of the current model to old sample a. The superscript T is a transpose operation that is often used for matrix multiplication. When inter-modal deviation happens (Figure 4(b)), the relative rotation of the representation space breaks the original modal alignment of the sample a, which makes the ($\beta_a^T\alpha_b > \beta_a^T\alpha_a$). From

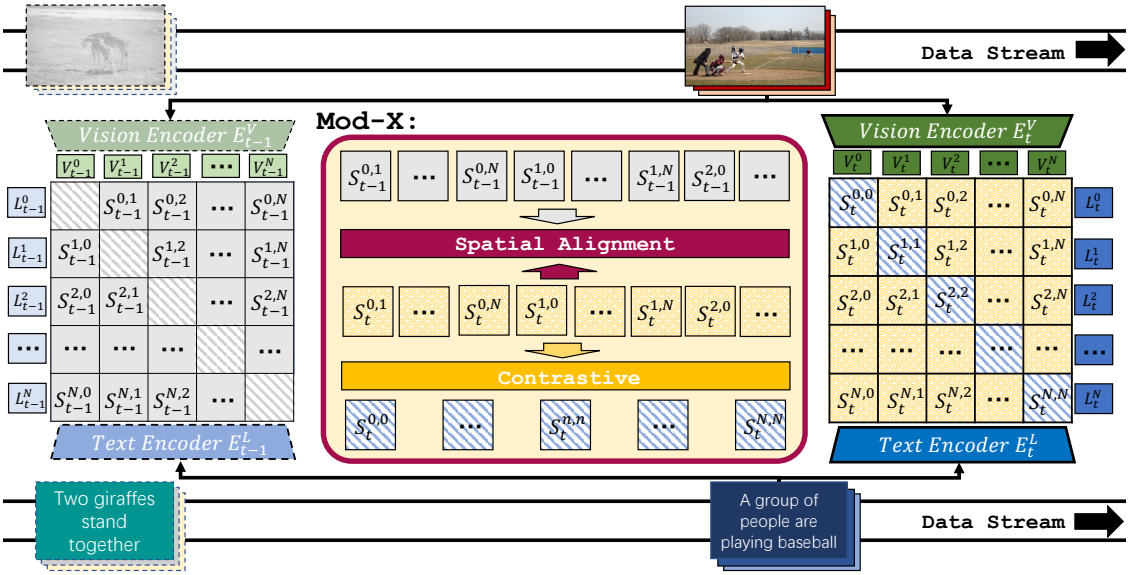


Figure 5. The Mod-X framework mainly consists of two sub-modules. Spatial Alignment helps the current model align the representation space of the old model based on current data. And Contrative helps the model fit the current training data domain.

the perspective of the contrastive matrix, the element $M_{i,j}$ in the i,j position of the contrastive matrix M represents the similarity score of the i'th sample vision embedding and the j'th sample text embedding. Since the length of the representation vector is **1**, the similarity score $M_{i,j}$ also refers to the angle between the i'th sample vision embedding and the j'th sample text embedding. When the angle becomes larger due to spatial disorder, its similarity score within the model becomes smaller, which affects the multi-modal retrieval ability of the model. Because of this, the performance of CLIP$_{ct}$ drops significantly during continual training. Detailed mathematical derivations can be found in Appendix A.1. The value of the diagonal elements in the contrast matrix M represents the angle between different modals of the same sample. The value of the off-diagonal elements represents the angle between the different modals of different samples in the CLIP's representation space. From an overall perspective, **the similarity distribution of the contrastive matrix M is equivalent to the structure of the representation space of the model.**

4. Alleviating Spatial Disorder

4.1. General continual CLIP training Setting

Suppose we have used training dataset D_0 got a pre-trained model CLIP$_0$. And there is another vision-language dataset D. We split D into N sub-datasets $\{D_1, ..., D_N\}$, randomly and evenly, to simulate a stream of data and $D_t = \{(v_t^0, l_t^0), ..., (v_t^n, l_t^n)\}$ denotes the training data in the training phase t, where $t \in \{1, 2, ..., N\}$. Then, we train the model CLIP$_0$ using this sub-datasets sequentially. The enocded l_2 normalized embeddings of vision and text is $V_t^i = E_V^t(v_t^i)$ and $L_t^i = E_L^t(l_t^i)$. When the model CLIP$_t$ is trained during the training phase t using training data D_t, the previous sub-datasets $\{D_0, D_1, ..., D_{t-1}\}$ are no longer available. The joint training represents that training a CLIP$_{jt}$ from scratch using all data $D_{jt} = D_0 \cup D$.

4.2. Mod-X: Maintain off-diagonal information-matrix

To alleviate spatial disorder of the CLIP$_{ct}$ model during continual training. We propose a simple but effective new training framework: Maintain off-diagonal information-matrix (Mod-X). It boots the current CLIP model to retain the spatial alignment to past samples by distilling the contrastive matrix's off-diagonal information which is constructed by the model before and after continual training based on the current training data. The entire training framework is shown in Figure 5, where the $S^{i,j}$ means cosine similarity score of the i'th sample's vision embedding and the j'th sample's text embedding. The Contrastive module in Figure 5 is a traditional InfoNCE loss (Baevski et al., 2020) , which inherits from CLIP (Radford et al., 2021). In the following, we mainly introduce our Spatial Alignment module.

4.3. Spatial Alignment

The diagonal elements in CLIP's contrastive matrix represent the similarity of the visual and language information of the current sample. The off-diagonal elements represent the similarity between the vision and language representation of the current sample and other samples. As mentioned in Section 3.2.3, the distribution of the elements in the contrastive

matrix represents the spatial distribution of representations between modalities of the model. Therefore, we feel out the old model's representation space through the old model's contrastive matrix on the current training data. Then, selectively distill the old model's spatial distribution while training the current model. We construct contrastive matrix M_{t-1} and M_t using the last and current model CLIP$_{t-1}$ and CLIP$_t$ based on current sub-dataset D_t.

$$M_{t-1}^{i,j} = CLIP_{t-1}(D_t) = s(E_V^{t-1}(v_t^i), E_L^{t-1}(l_t^j)) \quad (1)$$

$$M_t^{i,j} = CLIP_t(D_t) = s(E_V^t(v_t^i), E_L^t(l_t^j)) \quad (2)$$

Where the $s(a, b) = a^T b$ is the cosine similarity function. However, the last model's representation space for current data is not totally correct. For those misunderstood sample information (diagonal elements are not the largest in the current retrieval), we use the corresponding similarity information of the current model to replace them, thereby removing their influence during continual distillation.

$$M_{t-1}^{(i,:)} = M_t^{(i,:)}; \quad if \quad max(M_{t-1}^i) \neq i \quad (3)$$

After that, we align the information matrix M_{t-1} and M_t using Kullback-Leibler Divergence (Csiszár, 1975).

$$L_{KL}^t(M_t, M_{t-1}) = -\sum M_{t-1} ln(\frac{M_t}{M_{t-1}}) \quad (4)$$

The final training loss can be written as L_{Mod-X}, and α is a hyper-parameter.

$$L_{Mod-X}^t = L_{InfoNCE}^t + \alpha L_{KL}^t \quad (5)$$

5. Experiments

5.1. Datasets

In the experiments, we use three different training datasets varying in scope and domain to evaluate the effectiveness of our Mod-X framework. **MS COCO Captions** (Lin et al., 2014): MS COCO Captions (COCO) is a widely used image caption dataset. It contains 80K training images, 30K validation images, and 5K testing images (COCO(5K)). **Flickr30K** (Young et al., 2014): Flickr30K contains 30K training images and 1K test samples (Flickr30K(1K)) collected from Flickr, together with 5 reference sentences provided by human annotators. **ECommerce-T2I** (Yang et al., 2021) is a text-to-image e-commerce dataset that contains 90k training images and 5k testing images set (EC(5K)). Each image corresponds to a text description, and the description for each sample in the training and test sets does not repeat. Many detailed training settings and experiments (CC12M) can be viewed in Appendix B.

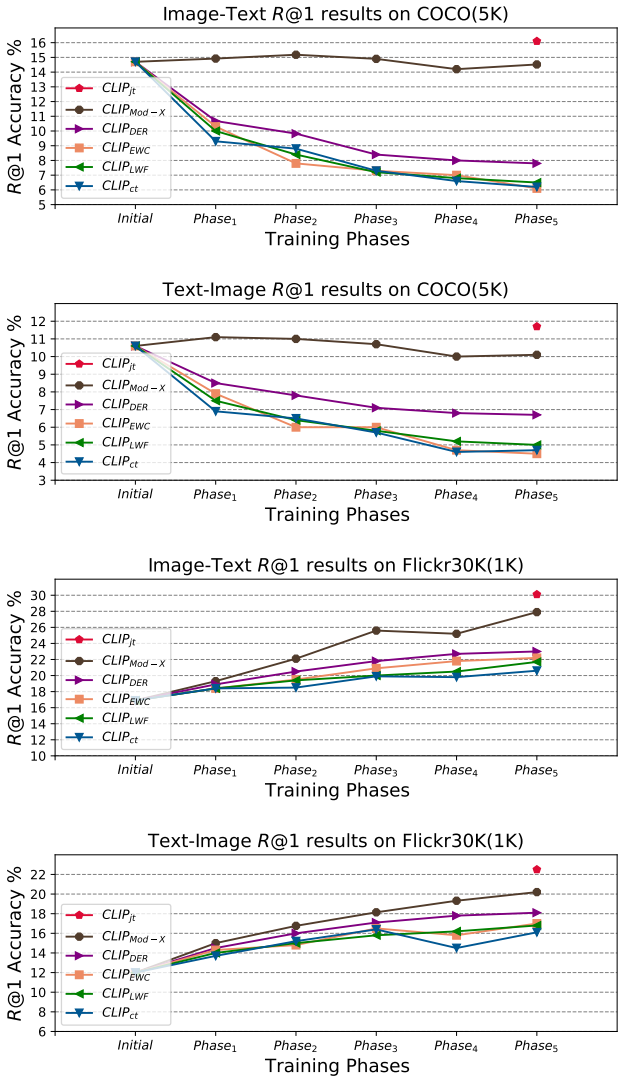


Figure 6. The multi-modal retrieval $R@1$ results of the different training strategies on COCO(5K) and Flickr30K(1K). The first two sub-figures show the retrieval $R@1$ performance on previous COCO dataset. The last two sub-figures show the $R@1$ results on current training domain Flickr30K.

5.2. The performance in Exploratory Experiments

Firstly, we follow the setup of the exploratory experiments described in Section 3.1, comparing the results of our Mod-X framework with CLIP$_{ct}$, CLIP$_{EWC}$ and CLIP$_{jt}$ in the Flickr30K dataset. The CLIP$_{ct}$ means training CLIP continually without any other operation. The CLIP$_{jt}$ is training CLIP model in the joint dataset of COCO and Flickr30K, which is an upper bound for continual CLIP training. Since label information is not used in CLIP training, recent supervised continual training methods like iCaRL (Rebuffi et al., 2017), PodNet (Douillard et al., 2020), and Dyn (Yan et al., 2021) cannot be reproduced in such experimental settings. We compared our Mod-X with the typical continual learning strategy such as DER (Buzzega et al., 2020)

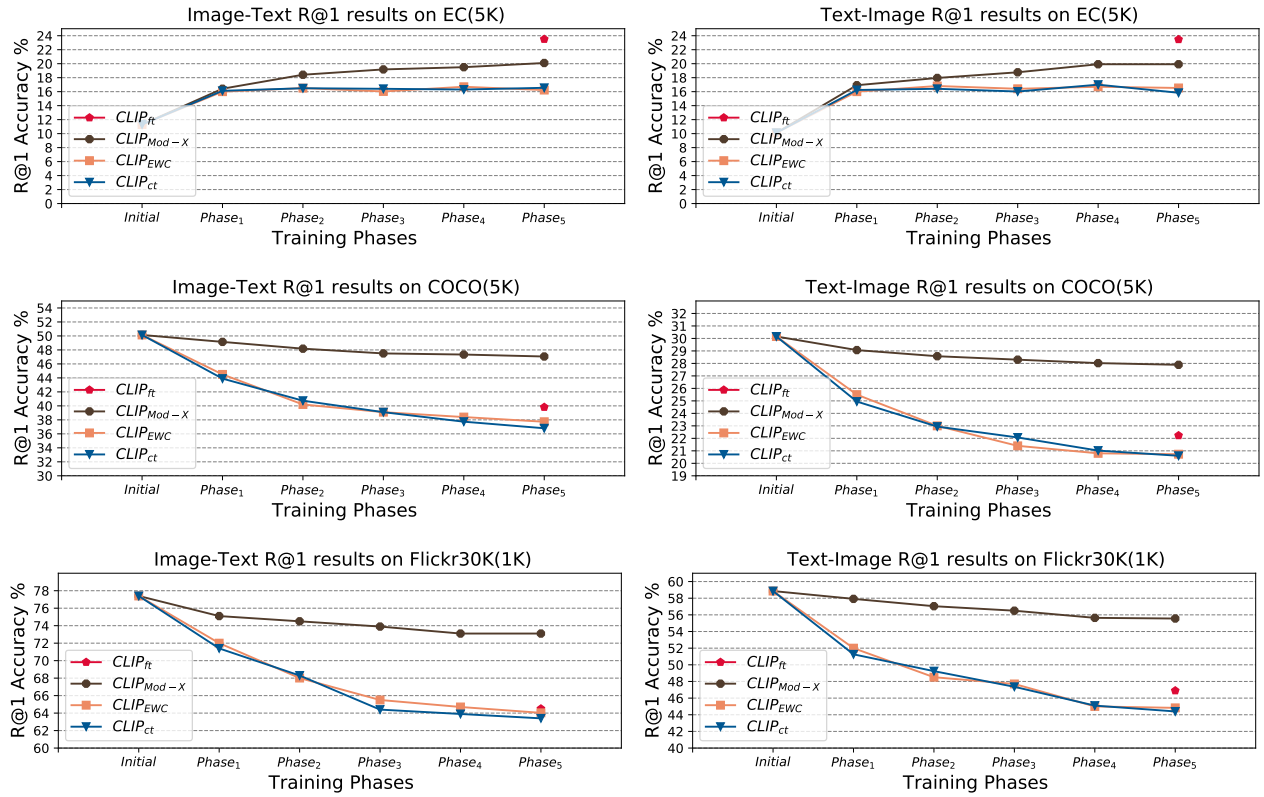


Figure 7. The retrieval $R@1$ performance of different training strategies in each training phase on EC(5K), COCO(5K) and Flickr30K(1K). The CLIP$_{ft}$ is the fine-tuning results of CLIP$_{initial}$ on the full ECommerce-T2I dataset as an upper bound on CLIP$_{ct}$.

(CLIP$_{DER}$), EWC (Kirkpatrick et al., 2017) (CLIP$_{EWC}$) and LWF (Li & Hoiem, 2017) (CLIP$_{LWF}$). Notably, to make LWF and DER work properly within the CLIP framework, we reproduced them using contrastive loss replaced their cross-entropy loss. The replay buffer size of DER is set to 3000. Figure 6 shows the effect of our framework Mod-X (CLIP$_{Mod-X}$) and the performance of other training strategies at each training phase. At each training phase, the $R@1$ results of CLIP$_{Mod-X}$ on COCO(5K) did not show a significant drop, and the gap with the initial accuracy (*Initial*) remained at $\pm 1\%$. Additionally, by comparing the retrieval performance of the CLIP$_{ct}$ and CLIP$_{Mod-X}$ on the current training data domain (Flickr30K), it can be found that the CLIP$_{Mod-X}$ is also significantly better than CLIP$_{ct}$ in continual fitting the current data domain. The low performance of CLIP$_{EWC}$ and CLIP$_{LWF}$ also shows that continual multi-modal training is more complex than single-modal supervised training. Due to the use of old training samples in memory buffer, the performance of DER is slightly better than LWF. However, its performance is still far from that of Mod-X. In Table 2, we presents the R@1 performance of the Mod-X framework in the final phase with a memory buffer size of 3000 or 5000. Before training on the Flickr30K, we randomly saved 1000 training samples from the COCO dataset into the memory buffer. Afterward,

Methods	COCO(I2T/T2I)	Flickr30K (I2T/T2I)
$CLIP_{jt}$	16.1 / 11.7	30.1 / 22.5
$CLIP_{Mod-X}$ with 5000	15.3 / 11.0	29.1 / 21.7
$CLIP_{Mod-X}$ with 3000	15.0 / 10.8	28.5 / 21.0
$CLIP_{Mod-X}$	14.5 / 10.1	27.9 / 20.2

Table 2. The table show the R@1 performance of the Mod-X framework in the final phase with a buffer size of 3000 or 5000. Since memory buffer strategy stores the model's knowledge from the input, it remains effective in the Mod-X framework which does not limit the input form.

based on the size of the buffer, an equal number of current training samples were randomly selected and stored in the memory buffer after each continual training phase which is similar to previous works (Rebuffi et al., 2017; Buzzega et al., 2020; Douillard et al., 2020). Since memory buffer strategy stores the model's knowledge from the input, it remains effective in the Mod-X framework which does not limit the input form. We can see from the results that as the number of old samples used increases, the performance of the Mod-X becomes closer to joint training.

Beside of this, in order to show the performance of our Mod-X in high semantics correlations data sets, we adopt an approximate strategy to simulate class incremental setting

Methods	COCO(I2T/T2I)	Flickr30K (I2T/T2I)
$CLIP_{Mod-X}$	14.5 / 10.1	27.9 / 20.2
$CLIP_{Mod-X}$ with cls	13.8/9.8	27.4/19.7
$CLIP_{ct}$	6.2/4.7	20.6/16.1
$CLIP_{ct}$ with cls	6.3/4.7	18.1/15.7

Table 3. The table show the R@1 performance of the Mod-X framework in a class incremental setting. The results demonstrate that continuous training with class incremental setting did not have a heavy impact on the effectiveness of the Mod-X.

in Flickr30K. Considering that the image labels are not available in Flickr30K, we used a pre-trained Imagenet1K model to automatically label the Flickr30K training data and divided it into 5 subsets, with each subset containing 200 classes. The R@1 results in the final phase have shown in Table 3, where the "cls" means "class incremental setting". From the results, it seems that continuous training with class incremental setting did not have a heavy impact on the effectiveness of the Mod-X. In Appendix B.2, we show the spatial alignment in CLIP$_{Mod-X}$ and CLIP$_{ct}$, which demonstrates that the Mod-X framework can alleviate the spatial disorder well during continual CLIP training.

5.3. The performance on special domain dataset ECommerce-T2I

To illustrate that the Mod-X framework is not only applicable to similar data domains, in this section, we compare the performance of different continual training frameworks on a specific e-commerce dataset ECommerce-T2I. We set the CLIP$_{vit32}$ with ViT-Base/32 vision encoder as the initial model, pre-trained using large-scale open-world datasets in (OpenAI). To simulate streaming data, we are dividing the entire ECommerce-T2I into five sub-datasets uniformly and randomly. Since the entire CLIP$_{vit32}$ pre-training dataset is not available, we use the fine-tuning results of CLIP$_{vit32}$ on the entire ECommerce-T2I dataset (CLIP$_{ft}$) as an upper bound on CLIP$_{ct}$.

The multi-modal retrieval $R@1$ results of CLIP$_{Mod-X}$, CLIP$_{ct}$, CLIP$_{EWC}$ and CLIP$_{ft}$ in each training phase are shown in Figure 7. Comparing the $R@1$ performance of the CLIP$_{jt}$ and CLIP$_{ct}$ on the EC(5K) test set, it's clear that the training of the CLIP model is affected by the training data domain: The final $R@1$ results of CLIP$_{jt}$ on EC(5K) is 8.8% (Image-Text) and 9.9% points (Text-Image) higher than CLIP$_{ct}$. However, the retrieval results of CLIP$_{jt}$ on COCO(5K) and Flickr30K(1K) have dropped by more than 10% points (comparing with CLIP$_{vit32}$ (*Initial*)) on average, which means that the performance of fine-tuning (one phase continual training) CLIP is also affected by the data domain. This is also verified by observations that the $R@1$ performance of CLIP$_{ct}$ performs lower than CLIP$_{jt}$. On the contrary, the CLIP$_{Mod-X}$ obtained after continual training by the Mod-X framework only has a tie drop of 3.3% points in the $R@1$ retrieval results on COCO(5K) and Flickr30K(1K). What's more, the performance of the CLIP$_{Mod-X}$ on EC(5K) outperformed CLIP$_{ct}$ by 3.5% (Image-Text R@1) and 4.2% points (Text-Image R@1), respectively. The overall $R@1$ trend of CLIP$_{EWC}$ is similar to that of CLIP$_{ct}$ but is more unstable than CLIP$_{ct}$. All of this shows that Mod-X framework not only preserves the inter-modal spatial structure of old samples during the continual training but also improves the fitting ability of the CLIP in the current training data domain.

6. Conclusion

This paper discusses the feasibility of continuously training the CLIP model through streaming data. Then, by tracking the directional changes of the representation vectors in the continuously updated CLIP model, we explore and summarize these spatial variations as Spatial Disorder (SD), which can be divided into Intra-modal Rotation and Inter-modal Deviation. Moreover, we demonstrate how intra-modal rotation and inter-modal deviation lead to a performance decline for CLIP on cross-modal retrieval tasks in both empirically and theoretically. To alleviate the spatial disorder, we propose a simple yet effective continual learning framework Mod-X: **M**aintain **o**ff-**d**iagonal information-matri**X**. The experiments (in Section 4, 5 and Appendix B) on commonly used datasets with different scales and scopes have illustrated the effectiveness of our method.

Social Impacts

The goal of continual learning is to help the model adapt to new data domains using only new data, without forgetting its past performance on old data domains. For example, in the fashion field, as fashion trends change, a image-text matching model fitted with old data will gradually become less suitable for the current fashion data. Therefore, by using Mod-X framework, the model can be updated to fit new image-text data using only the current fashion data while preventing the forgetting of past image-text knowledge. Beside of this, catastrophic forgetting, arises due to different reasons under various scenarios. In this work, our analysis methods and perspectives on continual image-text pretraining provide new ideas and approaches for future research on different continual learning tasks.

Acknowledgements

This work has been supported in part by the Zhejiang NSF (LR21F020004) and the NSFC (No. 62272411). We are grateful to Jiacheng Li and Xin He for their technical assistance. We also appreciate Haizhou Shi and Juncheng Li for their help in writing this paper.

References

Ahn, H., Cha, S., Lee, D., and Moon, T. Uncertainty-based continual learning with adaptive regularization. *Advances in neural information processing systems*, 32, 2019.

Ahn, H., Kwak, J., Lim, S., Bang, H., Kim, H., and Moon, T. Ss-il: Separated softmax for incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 844–853, 2021.

Andonian, A., Chen, S., and Hamid, R. Robust cross-modal representation learning with progressive self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16430–16441, 2022.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

Biesialska, M., Biesialska, K., and Costa-Jussa, M. R. Continual lifelong learning in natural language processing: A survey. *arXiv preprint arXiv:2012.09823*, 2020.

Buzzega, P., Boschini, M., Porrello, A., Abati, D., and Calderara, S. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020.

Cha, S., Yoo, Y., Moon, T., et al. Ssul: Semantic segmentation with unknown label for exemplar-based class-incremental learning. *Advances in Neural Information Processing Systems*, 34:10919–10930, 2021.

Changpinyo, S., Sharma, P., Ding, N., and Soricut, R. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3558–3568, 2021.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Chowdhury, J. R., Zhuang, Y., and Wang, S. Novelty controlled paraphrase generation with retrieval augmented conditional prompt tuning. *arXiv preprint arXiv:2202.00535*, 2022.

Csiszár, I. I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, pp. 146–158, 1975.

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

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Douillard, A., Cord, M., Ollion, C., Robert, T., and Valle, E. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *European Conference on Computer Vision*, pp. 86–102. Springer, 2020.

Fan, Z., Wei, Z., Chen, J., Wang, S., Li, Z., Xu, J., and Huang, X. A unified continuous learning framework for multi-modal knowledge discovery and pre-training. *arXiv preprint arXiv:2206.05555*, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Hu, D., Yan, S., Lu, Q., Lanqing, H., Hu, H., Zhang, Y., Li, Z., Wang, X., and Feng, J. How well does self-supervised pre-training perform with streaming data? In *International Conference on Learning Representations*, 2021.

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

Kj, J., Rajasegaran, J., Khan, S., Khan, F. S., and Balasubramanian, V. N. Incremental object detection via meta-learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Li, J., HE, X., Wei, L., Qian, L., Zhu, L., Xie, L., Zhuang, Y., Tian, Q., and Tang, S. Fine-grained semantically aligned vision-language pre-training. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 7290–7303. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/2fb4be70fc9668e9ec2c71b34fb127d4-\Paper-Conference.pdf.

Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., and Yan, J. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021.

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

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.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Madaan, D., Yoon, J., Li, Y., Liu, Y., and Hwang, S. J. Representational continuity for unsupervised continual learning. In *International Conference on Learning Representations*, 2021.

McCloskey, M. and Cohen, N. J. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Ni, Z., Shi, H., Tang, S., Wei, L., Tian, Q., and Zhuang, Y. Revisiting catastrophic forgetting in class incremental learning, 2021a.

Ni, Z., Tang, S., and Zhuang, Y. Self-supervised class incremental learning. *arXiv preprint arXiv:2111.11208*, 2021b.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

OpenAI. Clip. https://github.com/openai/CLIP.

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

Rebuffi, S.-A., Kolesnikov, A., Sperl, G., and Lampert, C. H. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T., and Wayne, G. Experience replay for continual learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. 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)*, pp. 2556–2565, 2018.

Shu, M., Nie, W., Huang, D.-A., Yu, Z., Goldstein, T., Anandkumar, A., and Xiao, C. Test-time prompt tuning for zero-shot generalization in vision-language models. *arXiv preprint arXiv:2209.07511*, 2022.

Srinivasan, T., Chang, T.-Y., Alva, L. L. P., Chochlakis, G., Rostami, M., and Thomason, J. Climb: A continual learning benchmark for vision-and-language tasks. *arXiv preprint arXiv:2206.09059*, 2022.

Sun, F.-K., Ho, C.-H., and Lee, H.-Y. Lamol: Language modeling for lifelong language learning. *arXiv preprint arXiv:1909.03329*, 2019.

Thai, A., Stojanov, S., Rehg, I., and Rehg, J. M. Does continual learning= catastrophic forgetting? *arXiv preprint arXiv:2101.07295*, 2021.

Thrun, S. A lifelong learning perspective for mobile robot control. In *Intelligent robots and systems*, pp. 201–214. Elsevier, 1995.

Wang, L., Yang, K., Li, C., Hong, L., Li, Z., and Zhu, J. Ordisco: Effective and efficient usage of incremental unlabeled data for semi-supervised continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5383–5392, 2021a.

Wang, X., Zhang, R., Shen, C., Kong, T., and Li, L. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3024–3033, 2021b.

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

Wei, L., Xie, L., Zhou, W., Li, H., and Tian, Q. Mvp: Multimodality-guided visual pre-training. *arXiv preprint arXiv:2203.05175*, 2022.

Yan, S., Xie, J., and He, X. Der: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3014–3023, 2021.

Yang, A., Lin, J., Men, R., Zhou, C., Jiang, L., Jia, X., Wang, A., Zhang, J., Wang, J., Li, Y., Zhang, D., Lin, W., Qu, L., Zhou, J., and Yang, H. M6-T: exploring sparse expert models and beyond. *CoRR*, abs/2105.15082, 2021.

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.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021.

A. Appendix to Section 3

A.1. The Theoretical Demonstrate that Inter-modal Deviation and Intra-modal Rotation lead to a decline in CLIP's multimodal retrieval performance

Inter-modal Deviation and Intra-modal Rotation can influence the CLIP's sample contrastive matrix, but this does not necessarily lead to errors in multimodal retrieval results. Unless the similarity of the visual language representation of the model for the same sample is smaller than that between different samples. In there, we abstract this problem and give the theoretical conditions that the Intra-modal Rotation and Inter-modal Deviation leads to a performance decline for CLIP on cross-modal retrieval tasks.

There has N image-text pairs $\{(\alpha_1,\beta_1),(\alpha_2,\beta_2),(\alpha_3,\beta_3),...,(\alpha_i,\beta_i),...,(\alpha_N,\beta_N)\} \in R^{W\times W}$. Through function $\mathcal{M}(\alpha)$ and $\mathcal{Q}(\beta)$, $\mathcal{M} \neq \mathcal{Q}$, the Euclidean space A and B of images and texts are formed.

$$\begin{aligned} A &= span\{\mathcal{M}(\alpha_1), \mathcal{M}(\alpha_2), \mathcal{M}(\alpha_2), ..., \mathcal{M}(\alpha_i), ..., \mathcal{M}(\alpha_N)\} \\ B &= span\{\mathcal{Q}(\beta_1), \mathcal{Q}(\beta_2), \mathcal{Q}(\beta_2), ..., \mathcal{Q}(\beta_i), ..., \mathcal{Q}(\beta_N)\} \end{aligned} \tag{6}$$

The $\mathcal{M}(\alpha_i)$,$\mathcal{Q}(\beta_j) \in R^D$ and $\|\mathcal{M}(\alpha_i)\| = 1$, $\|\mathcal{Q}(\beta_j)\| = 1$, $i, j = 1, 2, 3, ..., N$. $< \mathcal{M}(\alpha_i), \mathcal{Q}(\beta_j) >$ is the cosine between $\mathcal{M}(\alpha_i)$ and $\mathcal{Q}(\beta_j)$, $j = 1, 2, 3, ..., N$.

Suppose: $\exists(\alpha_a, \beta_a), (\alpha_b, \beta_b) \in \{(\alpha_i, \beta_j), i, j = 1, 2, 3, ..., N\}$ and $a \neq b$ makes:

$$\begin{aligned} < \mathcal{M}(\alpha_a), \mathcal{Q}(\beta_a) > &= \underset{\beta_i=\beta_a}{\arg\max} < \mathcal{M}(\alpha_a), \mathcal{Q}(\beta_i) > \\ < \mathcal{M}(\alpha_b), \mathcal{Q}(\beta_b) > &< \underset{\beta_j\neq\beta_b}{\arg\max} < \mathcal{M}(\alpha_b), \mathcal{Q}(\beta_j) > \end{aligned} \tag{7}$$

A.1.1. How does Inter-modal Deviation affect CLIP's multimodal retrieval performance?

Prove: There is a rotation matrix pair ($\mathcal{A}$,$\mathcal{B}$) that not only keeps the A and B topology unbiased and makes the

$$\begin{aligned} < \mathcal{M}'(\alpha_a), \mathcal{Q}'(\beta_a) > &< \underset{\beta_i\neq\beta_a}{\arg\max} < \mathcal{M}'(\alpha_a), \mathcal{Q}'(\beta_i) > \\ < \mathcal{M}'(\alpha_b), \mathcal{Q}'(\beta_b) > &= \underset{\beta_j=\beta_b}{\arg\max} < \mathcal{M}'(\alpha_b), \mathcal{Q}'(\beta_j) > \end{aligned} \tag{8}$$

where the $\mathcal{M}' = \mathcal{A}(\mathcal{M})$ and $\mathcal{Q}' = \mathcal{B}(\mathcal{Q})$, $\mathcal{A} \neq \mathcal{B}$. And the space A and B can be written as A' and B':

$$\begin{aligned} A' &= \mathcal{A}(A) = span\{\mathcal{M}'(\alpha_1), \mathcal{M}'(\alpha_2), \mathcal{M}'(\alpha_2), ..., \mathcal{M}'(\alpha_i), ..., \mathcal{M}'(\alpha_N)\} \\ B' &= \mathcal{B}(B) = span\{\mathcal{Q}'(\beta_1), \mathcal{Q}'(\beta_2), \mathcal{Q}'(\beta_2), ..., \mathcal{Q}'(\beta_i), ..., \mathcal{Q}'(\beta_N)\} \end{aligned} \tag{9}$$

Solution: the Equ.7 can be written as:

$$\begin{aligned} < \mathcal{M}(\alpha_a), \mathcal{Q}(\beta_a) > - < \mathcal{M}(\alpha_a), \mathcal{Q}(\beta_i) >> 0, \forall\beta_i \in \beta, i \neq a \\ < \mathcal{M}(\alpha_b), \mathcal{Q}(\beta_b) > - < \mathcal{M}(\alpha_b), \mathcal{Q}(\beta_j) >< 0, \exists\beta_j \in \beta, j \neq b \end{aligned} \tag{10}$$

hence:

$$\begin{aligned} \mathcal{M}(\alpha_a)^T\mathcal{Q}(\beta_a) - \mathcal{M}(\alpha_a)^T\mathcal{Q}(\beta_i) > 0, \forall\beta_i \in \beta, i \neq a \\ \mathcal{M}(\alpha_b)^T\mathcal{Q}(\beta_j) - \mathcal{M}(\alpha_b)^T\mathcal{Q}(\beta_b) > 0, \exists\beta_j \in \beta, j \neq b \end{aligned} \tag{11}$$

because the rotation matrix pair ($\mathcal{A}$,$\mathcal{B}$) can be seen as a rotation matrix $\mathcal{R}(\theta^D)$, where the θ^D is a rotation angle between AB and $A'B'$. Hence, when applying this rotation matrix $\mathcal{R}(\theta^D)$, the Equ.11 can be written as:

$$\begin{aligned} \mathcal{M}(\alpha_a)^T\mathcal{R}(\theta^D)\mathcal{Q}(\beta_a) - \mathcal{M}(\alpha_a)^T\mathcal{R}(\theta^D)\mathcal{Q}(\beta_i) < 0, \exists\beta_i \in \beta, i \neq a \\ \mathcal{M}(\alpha_b)^T\mathcal{R}(\theta^D)\mathcal{Q}(\beta_j) - \mathcal{M}(\alpha_b)^T\mathcal{R}(\theta^D)\mathcal{Q}(\beta_b) < 0, \forall\beta_j \in \beta, j \neq b \end{aligned} \tag{12}$$

Because the rotation matrix satisfies that the inner product of itself is 1. So, Equ 12 can be written as:

$$\mathcal{M}(\alpha_a)^T\mathcal{R}(\theta^D)(\mathcal{Q}(\beta_a)-\mathcal{Q}(\beta_i))<0, \exists\beta_i\in\beta, i\neq a, \mathcal{R}(\theta^D)^T\mathcal{R}(\theta^D)=\mathcal{I} \tag{13}$$

$$\mathcal{M}(\alpha_b)^T\mathcal{R}(\theta^D)(\mathcal{Q}(\beta_j)-\mathcal{Q}(\beta_b))<0, \forall\beta_j\in\beta, j\neq b, \mathcal{R}(\theta^D)^T\mathcal{R}(\theta^D)=\mathcal{I} \tag{14}$$

For example, when $\mathcal{R}(\theta^D)=-\mathcal{I}$, then $\mathcal{R}(\theta^D)^T\mathcal{M}(\alpha_a)=-\mathcal{M}(\alpha_a)$ the equ 15 and 16 will hold. So, rotation matrices ($\mathcal{A}$,$\mathcal{B}$) that makes Equ.8 true exists.

A.1.2. How does Intra-modal Rotation affect CLIP's multimodal retrieval performance?

Since intra-modal rotation just requires the length of representation vectors after rotation is 1 and does not require that the intra-modal representation space is invariant, it is a more general case of inter-modal deviation. This means that all rotation matrixes that satisfy A.1.1 can also satisfy Intra-modal Rotation. Different from intra-modal deviation, the inner product of the mapping matrix $\mathcal{P}$ does not require to be 1. So, we rewrite the Equ 15 and 16 to:

$$\mathcal{M}(\alpha_a)^T(\mathcal{Q}(\beta_a)-\mathcal{Q}(\beta_i))\mathcal{P}<0, \exists\beta_i\in\beta, i\neq a, \tag{15}$$

$$\mathcal{M}(\alpha_b)^T(\mathcal{Q}(\beta_j)-\mathcal{Q}(\beta_b))\mathcal{P}<0, \forall\beta_j\in\beta, j\neq b, \tag{16}$$

Any mapping matrix $\mathcal{P}$ that rotation the direction of $(\mathcal{Q}(\beta_j)-\mathcal{Q}(\beta_b))$ by more than 90 degrees.

A.2. Detailed SAM and RAM Distribution of Language Encoders

The topology of the language representation space does not change significantly during the continual CLIP training. But the whole language representation space, like the vision representation space, has a large rotation around the center of the high-dimensional sphere during the continual training. The angle change distribution Table 8(a) and rotation angle distribution Table 8(b) are shown below.

$\theta_{SAM}\in$	$[0°, 5°]$	$(5°, 10°]$	$(10°, 15°]$	$(15°, 20°]$	$(20°, 180°]$
SAM_{0-1}	64.43%	28.49%	6.23%	0.78%	0.07%
SAM_{1-2}	71.54%	24.89%	3.35%	0.22%	0.01%
SAM_{2-3}	71.36%	25.01%	3.40%	0.22%	0.01%
SAM_{3-4}	67.30%	27.27%	4.93%	0.48%	0.03%
SAM_{4-5}	58.84%	30.70%	8.77%	1.50%	0.20%
SAM_{0-5}	55.39%	31.60%	10.52%	2.15%	0.33%

(a) The include angle distribution (SAM) in text representation space.

$\theta_{RAM}\in$	$[0°, 15°]$	$(15°, 20°]$	$(20°, 25°]$	$(25°, 30°]$	$(30°, 180°]$
$RAM_{(0,1)}$	0.00%	1.94%	28.38%	45.88%	23.80%
$RAM_{(1,2)}$	0.02%	8.90%	47.76%	34.94%	8.38%
$RAM_{(2,3)}$	0.04%	1.14%	49.86%	31.18%	7.52%
$RAM_{(3,4)}$	0.02%	2.84%	33.70%	43.76%	19.68%
$RAM_{(4,5)}$	0.00%	0.04%	3.28%	27.66%	69.02%
$RAM_{(0,5)}$	0.00%	0.00%	0.00%	1.12%	98.88%

(b) The rotation angle distribution (RAM) in text representation space.

Figure 8. Detailed SAM and RAM Distribution of Language Encoders.

By observing the table in Table 8, we can find that more than 88% of the angle change between any two language representation vectors in the language representation space are between 0 and 10 degrees in the process of continual CLIP training, while only 20% are above 10 degrees. Moreover, less than 0.2% of the angle changes is above 20 degrees. Those angle change between 15-20 degrees also only account for about 1.5% of all images pairs. Similar to the visual representation space, the direction of the same sample in the language representation space of different training phases also has changed greatly. However, unlike most of the rotations in the vision representation space, which are distributed over 30 degrees, in the language space, the rotations in the representation space are mostly distributed between 20 and 30 degrees. **Because of**

this difference, the representation alignment of the CLIP for different modalities of the same sample deviates during the continual training.

A.3. The representation quality of vision encoders during continual CLIP training

In Section 3, based on the distribution Table 2(c), we inference that the topology of the visual representation of the CLIP$_{ct}$ changes slowly during the continual CLIP training. Due to the topology of the representation space is correlated with the quality of the model's representation, so we use the linear probe evaluation method, commonly used in self-supervision (Oord et al., 2018; He et al., 2020), to detect the quality of the model's vision encoders to verify our suppose.By fixing the vision encoder, retrain a single Linear layer, which is connected behind the vision encoder, based on the ImageNet (Deng et al., 2009) training set and evaluate its top-1 accuracy on the ImageNet test set to represent the vision encoder's representation quality. As shown in Figure 9, we calculate the vision encoders' linear evaluation in each training phase in explore experiment 3.

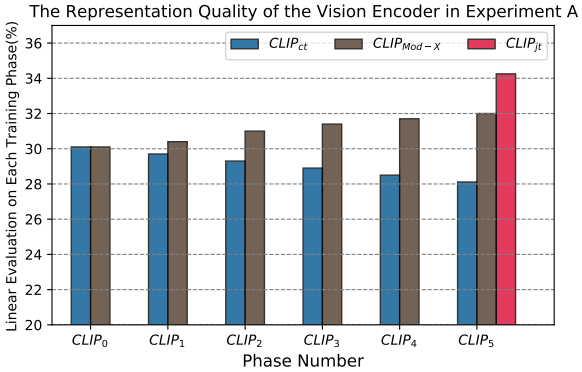


Figure 9. The representation quality of visual encoders in each training phase in explore experiment 3.

Observing the changing trends in the linear evaluation accuracy of each training phase, we can find that the representation quality of the vision encoder in CLIP$_{cl}$ gradually decreases as the training phase increases. The top-1 accuracy in the ImageNet test set dropped from 30.1% to 28.1%, which is consistent with our conjecture 3.2.1. Compared to the decline in multimodal retrieval, the decrease in the quality of visual representations appears to be negligible. In addition, by comparing the results of CLIP$_{Mod-X}$ and CLIP$_{jt}$, we can find that our Mod-X framework can not only help the model fit new image-text samples but also improve the representation quality of the modal encoders within the CLIP. The top-1 accuracy of the vision encoder in CLIP$_{Mod-X}$ improved from 30.1% to 32.0%. All of this also illustrates that the quality of the extractor representation is not precisely positively correlated with the decline in multimodal retrieval performance of CLIP model. Alignment of the representation between the different modals is also critical.

B. Appendix to Section 5

B.1. Detailed Experiment Setting

In exploration experiments 3 and Experiment 5.2, we use RN50 (He et al., 2016) as the vision encoder. In experiment 5.3 we use Vit-32/B as the vision encoder. The language encoder in all experiments is a transformer-based architecture which follows modification proposed in CLIP (OpenAI). In all experiments, the input images are resized to 224 × 224 and the input texts are tokenized by WordPiece with a maximum length of 77. We utilize AdamW (Loshchilov & Hutter, 2017) optimizer and a cosine annealing learning rate schedule with warmup which is consistent with (OpenAI). All of the experiments are conducted on 8 NVIDIA V100 GPUS.

In exploration experiment 3 and Experiment 5.2, we use the hyper-parameters as be shown in table 3(a). Since the experiment 5.3 based on the pre-training model ViT-32/B in (OpenAI), we set a smaller learning rate from 5e-4 to 1e-6. And other hyper-parameters is consistent with Experiment 5.2 and CLIP (OpenAI).

(a)

Hyperparameter	Value
Batch size	280
Vocabulary size	49408
Training epochs	35
Initial temperature τ	0.07
α	20
Weight decay	0.2
Warm-up iterations (%)	20
Learning rate	$5e^{-4}$
Adam β_1	0.9
Adam β_2	0.99
Adam ϵ	$1e^{-8}$

(b)

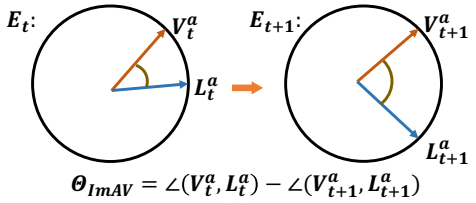
Hyperparameter	Value
Batch size	280
Vocabulary size	49408
Training epochs	35
Initial temperature τ	0.07
α	20
Weight decay	0.2
Warm-up iterations (%)	20
Learning rate	$1e^{-6}$
Adam β_1	0.9
Adam β_2	0.99
Adam ϵ	$1e^{-8}$

Table 4. Table (a) is the hyper-parameter in exploration experiment (Section 3) and Experiment 5.2. Table (b) is the hyper-parameter in Experiment 5.3).

B.2. The Relationship Between Contrastive Matrix, Intra-modal Rotation, Inter-modal Deviation and Mod-X

From a detailed point of view, the element $M_{i,j}$ in the i,j position of the contrastive matrix M is the similarity score of the i'th sample vision embedding and the j'th sample text embedding. Since the length of the representation vector is **1**, the similarity score $M_{i,j}$ also refers to the angle between the i'th sample vision embedding and the j'th sample text embedding. Greater similarity means a smaller angle. Therefore, the value of the diagonal elements in the contrast matrix M represents the angle between different modals of the same sample. The value of the off-diagonal elements represents the angle between the different modals of different samples in the CLIP's representation space. Through our exploration (in section 3), the Intra-modal Rotation and the Inter-modal Deviation affect these angles or similarity scores. From an overall perspective, **the similarity distribution of the contrastive matrix M is equivalent to the structure of the representation space of the model.** Our Mod-X framework attempts to distill the similarity distribution of off-diagonal elements identical to aligning the model's representation space structure, which reduces the influence of spatial disorder during continual CLIP training.

To better illustrate the relationship between the model's representation space and the model's similarity performance, we add a more direct statistical analysis, **inter-modal angle variation distribution**. Based on the settings in section 3, in the training phase t, we compare the change of angle distribution between modalities for the training samples retrieved correctly in the training phase $t-1$. A schematic diagram of inter-modal angle variation θ_{ImAV} is shown in Figure 10(a), where the sample a refers to the training sample that can be retrieved correctly by model CLIP$_{t-1}$ in training phase $t-1$. The V is the vision representation and L is the language representation. Inter-modal angle variation distribution table can be seen in Figure 10(b).



(a)

$\theta_{ImAV} \in$	$[0°, 5°]$	$(5°, 10°]$	$(10°, 15°]$	$(15°, 20°]$	$(20°, 180°]$
$ImAV_{(0,1)}$	44.78%	31.54%	14.62%	8.26%	0.81%
$ImAV_{(1,2)}$	50.37%	28.48%	16.58%	4.57%	0.00%
$ImAV_{(2,3)}$	49.70%	24.22%	20.53%	5.13%	0.42%
$ImAV_{(3,4)}$	46.25%	30.12%	18.53%	4.81%	0.29%
$ImAV_{(4,5)}$	43.36%	32.83%	19.81%	3.82%	0.18%
$ImAV_{(0,5)}$	31.98%	33.62%	24.37%	10.01%	0.02%

(b)

Figure 10. The sub-figure on the left shows a schematic diagram of computing θ_{ImAV}. The table on the right shows the distribution of the included angle change between the vision and language representation of the samples in CLIP$_{ct}$, which were correctly retrieved in the previous training phases.

As shown in Figure 10(b), during the continual training, the samples that were correctly retrieved in the past have apparent changes in the angle between the modalities as the training phases go up. Only less than 50% of the samples change within 5 degrees in the continual training, and about 30% of the samples have a change of 5-10 degrees. However, more than

20% of the samples change their included angle by more than 10 degrees during the training process. This shows that the inter-modal spatial alignment (similarity performance) of the CLIP$_{ct}$ is affected by spatial disorder.

To illustrate our Mod-X framework indeed alleviates the spatial disorder between sample's modalities during continual training, we show the inter-modal angle variation distribution of the CLIP$_{Mod-X}$ in Experiment 5.2 in Table 5.

$\theta_{ImAM} \in$	$[0°, 5°]$	$(5°, 10°]$	$(10°, 15°]$	$(15°, 20°]$	$(20°, 180°]$
$ImAM_{(0,1)}$	88.66%	7.81%	2.56%	0.97%	0.00%
$ImAM_{(1,2)}$	91.79%	4.01%	3.20%	0.00%	0.00%
$ImAM_{(2,3)}$	90.70%	9.02%	0.24%	0.04%	0.01%
$ImAM_{(3,4)}$	92.13%	6.20%	1.61%	0.06%	0.00%
$ImAM_{(4,5)}$	91.91%	7.71%	0.38%	0.00%	0.00%
$ImAM_{(0,5)}$	87.81%	10.87%	1.12%	0.20%	0.00%

Table 5. The table shows the distribution of the included angle change between the vision and language representation of the sample in the CLIP$_{Mod-X}$ in Experiment 5.2.

Comparing the Figure 10(b) and Table 5, it can be found that the CLIP$_{Mod-X}$ well maintains the inter-modal spatial alignment of the correctly retrieved samples during the continual CLIP training. On average, 90% of the correctly retrieved samples have an angle change of less than 5 degrees in continual training, and the samples with an angle change of more than 15 degrees account for less than 1% of all samples. All of this shows that the Mod-X framework does mitigates the spatial disorder during continual CLIP training by preserving the inter-modal spatial alignment of the samples retrieved correctly in the past during the continual training.

B.3. Validation of Inter-modal Deviation on ECommerce-T2I dataset

In section 3, we discuss the representational space variation of CLIP$_{ct}$ under the open-world dataset COCO(Lin et al., 2014) and Flickr30K(Young et al., 2014). In there, following the explore settings of the section 3.2.2, we compare the rotation distribution of the representation space of the vision and language extractors of CLIP$_{ct}$ under the specific e-commerce text to image dataset ECommerce-T2I (Yang et al., 2021) (Experiment 5.3) By evaluating the rotation distribution of the modal's representation space at various training phases on the COCO(5K) testset, we drawn the rotation distribution comparison diagram in Figure 11.

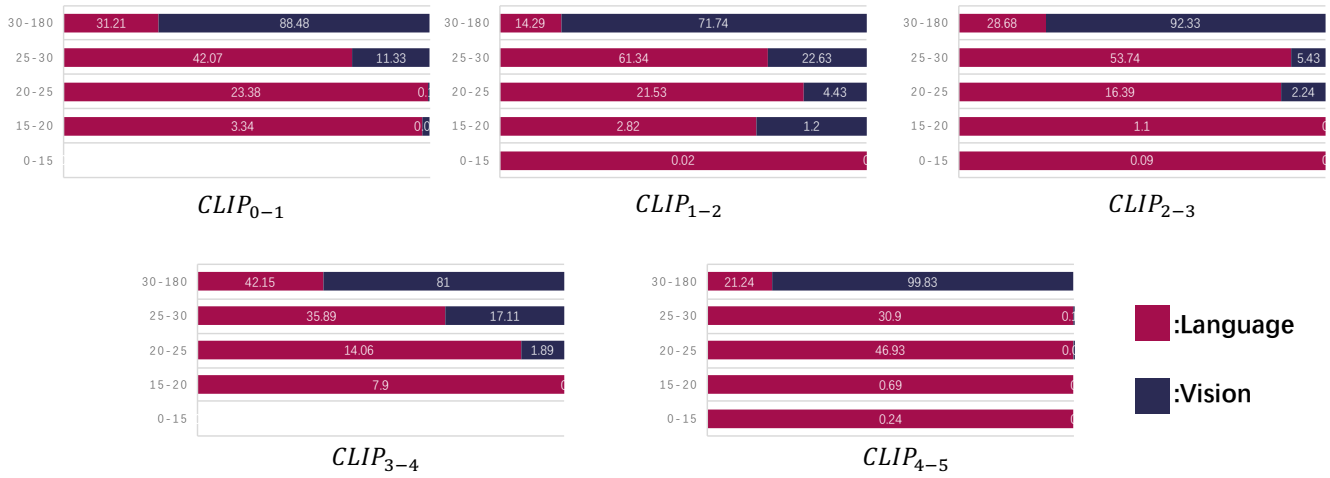


Figure 11. The comparison of the rotation distributions of the vision encoder and langugae encoder during continual CLIP training on ECommerce-T2I dataset. CLIP$_{i-j}$ refers to the CLIP's continual training from training phase i to j. The values under the same color represent the proportion of test samples to total samples in each rotation angle interval of the same modality.

From Figure 11, we can find that when the CLIP is trained on a specific data domain, the rotation of visual representation

space becomes more severe, among which more than 70% of the samples have more than 30 degrees of rotation in the visual space, which is higher than that of the open-world dataset. Although the rotation of more than 30 degrees in the language space has also seen a large proportional increase than the open-world dataset, it is still significantly out of sync with the rotation in the visual space. Most samples are rotated within 30 degrees in language space. Through this validation, we show that inter-modal deviation (rotational asynchrony) of the representation space of different modal encoders persists during the continual CLIP training on a specific data domain.

B.4. The sensitivity of hyper-parameter α

In this section, we discuss the effect of different α on the final performance of the CLIP$_{Mod-X}$ based on the settings of Experiment 5.2. Table 6 presents the final retrieval results of the CLIP$_{Mod-X}$ model with $\alpha = 10, 15, 20, 25, 30$.

Pretraining Dataset	Model	Image-Text Retrieval(%)						Text-Image Retrieval(%)					
		Flickr30K(1K)			COCO(5K)			Flickr30K(1K)			COCO(5K)		
		$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$
COCO	CLIP$_0$	16.9	37.0	46.2	14.7	34.2	47.0	12.0	30.0	41.0	10.6	29.6	41.0
	CLIP$_{ct}$	20.6	42.8	56.4	6.2	17.8	26.1	16.1	38.5	50.4	4.7	14.3	21.8
	$\alpha = 10$	25.7	50.4	60.3	11.6	28.4	30.9	17.3	40.2	54.6	7.9	20.9	34.7
	$\alpha = 15$	28.1	54.3	66.7	14.0	32.8	45.4	20.7	45.8	58.0	9.7	26.0	36.4
	$\alpha = 20$	**27.9**	**53.4**	**64.4**	**14.5**	**34.0**	**46.1**	**20.2**	**45.0**	**57.2**	**10.1**	**26.4**	**37.4**
	$\alpha = 25$	26.6	52.8	62.3	14.5	34.8	46.7	20.2	44.7	57.0	10.0	27.7	38.1
	$\alpha = 30$	25.5	51.7	61.8	14.7	35.0	47.1	18.4	42.8	55.5	10.2	27.0	38.3
COCO+F30K	CLIP$_{jt}$	30.1	55.9	60.1	16.1	38.1	51.9	22.5	48.5	59.6	11.7	30.9	42.7

Table 6. The final multimodal retrieval performance of different α on continual CLIP$_{Mod-X}$ training in the Experiment 5.2.

From the table, we can find that although different α affects the performance of the CLIP$_{Mod-X}$, **different α does not significantly affect the effectiveness of the Mod-X framework.** The performance of CLIP$_{Mod-X}$ is better than CLIP$_{ct}$ under different α. As α increases, the CLIP$_{Mod-X}$ better maintains its retrieval ability on past COCO samples. The Image-Text R@1 and Text-Image R@1 on COCO(5K) remain around 14.5% and 10.0%. However, an excessively large α also limits the model's ability to fit new datasets. With the value of α increased from 20 to 30, the Image-Text R@1 and Text-Image R@1 of the CLIP$_{Mod-X}$ on the Flickr30k(1K) drops from 27.9% and 20.2% to 25.2% and 18.4%.

B.5. The detailed performance of different training strategies at final training phase in Experiment 5.2

Pretraining Dataset	Model	Image-Text Retrieval(%)						Text-Image Retrieval(%)					
		Flickr30K(1K)			COCO(5K)			Flickr30K(1K)			COCO(5K)		
		$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$
COCO	CLIP$_0$	16.9	37.0	46.2	14.7	34.2	47.0	12.0	30.0	41.0	10.6	29.6	41.0
	CLIP$_{ct}$	20.6	42.8	56.4	6.2	17.8	26.1	16.1	38.5	50.4	4.7	14.3	21.8
	CLIP$_{EWC}$	22.2	43.1	57.0	6.1	17.2	26.5	17.0	39.1	51.2	4.5	13.9	22.0
	CLIP$_{Mod-X}$	**27.9**	**53.4**	**64.4**	**14.5**	**34.0**	**46.1**	**20.2**	**45.0**	**57.2**	**10.1**	**26.4**	**37.4**
COCO+F30K	CLIP$_{jt}$	30.1	55.9	60.1	16.1	38.1	51.9	22.5	48.5	59.6	11.7	30.9	42.7

Table 7. The final multimodal retrieval performance of the different continual CLIP training strategies in the Experiment 5.2.

From the results in the Table 7, it is clear that our method CLIP$_{Mod-X}$ maintains its multimodal retrieval results on COCO(5K) after completing continual training on Flickr30K. The gap between CLIP$_0$ and CLIP$_{Mod-X}$ is just 0.2% points in image-text retrieval $R@1$ and 0.5% points in text-image retrieval $R@1$ on COCO(5K). At the same time, the retrieval results of the CLIP$_{Mod-X}$ on the test set Flickr30K(1K) are also affected by the training domain and have a significant increase. The $R@1$ performance of the CLIP$_{Mod-X}$ in image-text retrieval rise from 16.9% (in CLIP$_0$) to 27.9%. And the $R@1$ results in text-image retrieval increase from 12.0% (in CLIP$_0$) to 20.2%. The performance gap between CLIP$_{Mod-X}$ and CLIP$_{jt}$ on the Flickr30K is only at most 2.3% points. Conversely, due to the model's spatial disorder in continual training, the performance of CLIP$_{ct}$ on COCO(5K) drops significantly. In addition, although the performance of CLIP$_{ct}$ on Flickr30K(1K) has improved, it is still far from the upper bound CLIP$_{jt}$. From the above experimental results, although CLIP$_{EWC}$ improves the accuracy of continual CLIP training on Flickr30K(1K), it does not preserve the

model's understanding in past samples (COCO(5K)). According to the above comparisons, we can conclude that our Mod-X framework can not only maintain the representation alignment on old samples during continual CLIP learning but also improve the model's fitting ability to the current training data domain.

B.6. The detailed performance of different training strategies at final training phase in Experiment 5.3

In table 8, we show the performance of different training strategies at final training phase in Experiment 5.3. Comparing the CLIP$_{Mod-X}$'s $R@1$ and $R@5$ results with others in different datasets, we can find that CLIP$_{vit32}$ model that have not been trained on ECommerce-T2I dataset have poor multimodal retrieval capabilities on EC(5K) dataset (11.3% and 10.1%). When fine-tuning the CLIP$_{vit32}$ on ECommerce-T2I, the $R@1$ and $R@5$ performance of all training strategies improves. Different from other strategies, our Mod-X framework improves the model's multimodal retrieval ability to the current training data domain while maintaining its performance to the previous data domain (Flickr30K and COCO).

Model	Image-Text Retrieval(%)						Text-Image Retrieval(%)					
	Flickr30k(1K)		COCO(5K)		EC(5K)		Flickr30k(1K)		COCO(5K)		EC(5K)	
	$R@1$	$R@5$	$R@1$	$R@5$	$R@1$	$R@5$	$R@1$	$R@5$	$R@1$	$R@5$	$R@1$	$R@5$
CLIP$_{vit32}$	77.7	94.5	50.1	74.6	11.3	27.6	58.9	83.5	30.2	55.6	10.1	25.5
CLIP$_{ct}$	63.4	87.2	36.8	61.5	16.6	40.7	44.4	71.0	20.6	42.6	15.8	40.5
CLIP$_{EWC}$	64.0	87.8	37.7	64.3	16.2	40.0	44.8	72.4	20.7	44.1	16.5	42.0
CLIP$_{Mod-X}$	**73.1**	**92.1**	**47.1**	**70.5**	**20.1**	**44.8**	**55.6**	**79.9**	**27.9**	**51.0**	**20.0**	**44.8**
CLIP$_{ft}$	64.5	88.6	39.8	64.8	23.5	50.8	46.9	73.1	22.2	44.5	23.5	50.6

Table 8. The final multimoal retrieval performance of the CLIP$_{ct}$, CLIP$_{Mod-X}$ and CLIP$_{ft}$ based on OpenAI's CLIP$_{vit32}$ on specific e-commerce dataset ECommerce-T2I (Experiment 5.3).

B.7. The performance of the Mod-X when training in CC12M dataset

In this section, we show the performance of different continual training strategies in CC12M (Changpinyo et al., 2021) training dataset. The CC12M training dataset collects about 12M images and their raw descriptions harvested from the alt-text HTML attribute associated with the webscraped images, therefore representing a wider variety of content styles. Due to unavailable URLs, we utilize about 10M examples from this dataset. Firstly, we randomly and evenly split the CC12M dataset into 10 sub-datasets, each containing 1M image-text pairs. Then, we continuously train a CLIP based on these sub-datasets from scratch without any pre-training. **The purpose of this experiment is to demonstrate that our Mod-X framework still excels in large-scale continual pre-training.** In table 9, we show the final retrieval performance of different continual training strategies in COCO(5K) and Flickr30K(1K) test sets. The CLIP$_{ct}$ means continual training without any other operations. The CLIP$_{Mod-X}$ means continual training using our Mod-X framework. And the CLIP$_{jt}$ refers to training CLIP model using the joint dataset CC12M.

Model	Image-Text Retrieval(%)						Text-Image Retrieval(%)					
	Flickr30k(1K)			COCO(5K)			Flickr30k(1K)			COCO(5K)		
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$
CLIP$_{ct}$	35.50	64.80	76.10	17.38	39.24	51.68	24.54	49.96	61.44	12.10	29.60	40.26
CLIP$_{Mod-X}$	**40.40**	**67.90**	**77.40**	**22.06**	**46.12**	**58.14**	**27.74**	**53.88**	**64.66**	**14.22**	**33.68**	**45.02**
CLIP$_{jt}$	58.00	83.90	90.40	34.38	60.30	71.50	43.02	72.34	80.92	22.63	46.44	58.35

Table 9. The final multimoal retrieval performance of the CLIP$_{ct}$, CLIP$_{Mod-X}$ and CLIP$_{jt}$ on COCO(5K) and Flickr30K(1K).

Comparing the final performance of the three training strategies, Mod-X framework (CLIP$_{Mod-X}$) still outperforms CLIP$_{ct}$ in the large-scale pre-training. After continual pre-training, the CLIP$_{Mod-X}$ obtain 40.40% Image-Text R@1 result and 27.74% Text-Image R@1 result on Flickr30K(1K) test set, which surpasses the 35.50% and 24.54% of CLIP$_{ct}$. The results on COCO(5K) are similar to those on Flickr30K(1K). The Image-Text R@1 result of CLIP$_{Mod-X}$ on COCO(5K) is 4.68% points higher than CLIP$_{ct}$ and the Text-Image R@1 result of CLIP$_{Mod-X}$ on COCO(5K) exceeds CLIP$_{ct}$ 2.12% points. The detailed $R@1$ performance of three training strategies at each training phase can be seen in Figure 12.

Beside of this, we compare the performance of the Mod-X (CLIP$_{Mod-X}$), continual learning without other operations

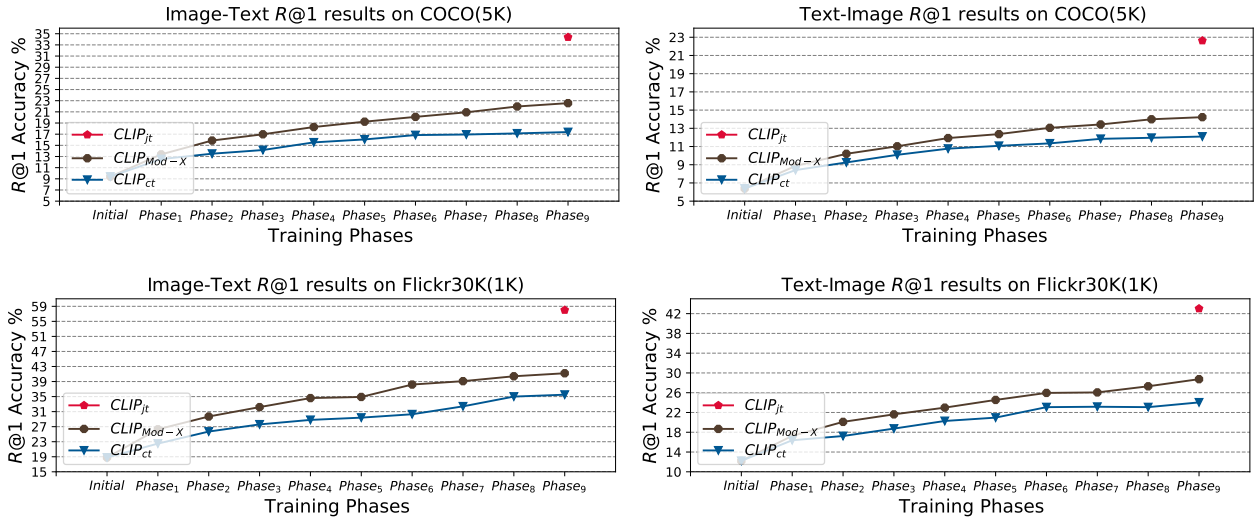


Figure 12. The retrieval performance of different training strategies in each training phase on COCO(5K) and Flickr30K(1K).

(CLIP_{ct}) and baseline joint learning (CLIP_{jt}) on linear probe top-1 accuracy (%) and zero-shot image classification top-1 accuracy(%) at final training phase. The results can be seen in the following Table 10.

Model	Zero-Shot Image Classification(%)						Linear Probe(%)
	Cifar10	Caltech101	Places365	ObjectNet	ImageNet	Average	ImageNet
CLIP_{jt}	73.1	40.4	32.3	10.4	35.7	38.4	47.3
CLIP_{Mod-X}	**71.2**	**35.8**	**28.7**	**8.3**	**29.8**	**34.8**	**41.6**
CLIP_{ct}	64.7	30.2	23.5	6.2	23.4	29.6	35.1

Table 10. The linear probe top-1 accuracy (%) and zero-shot image classification top-1 accuracy(%) at final training phase of the CLIP_{ct}, CLIP_{Mod-X} and CLIP_{jt}.

From the results, we can find that the linear probe performance of CLIP_{Mod-X} on ImageNet is significantly higher than that of CLIP_{ct}. This shows that the representation quality of the model continuously trained by the Mod-X framework (CLIP_{Mod-X}) is better than that of pure continual training (CLIP_{ct}). Comparing the zero-shot top-1 average results of the model on multiple classification datasets, it can be found that the representation generalization performance of the CLIP_{Mod-X} is also significantly better than that of CLIP_{ct}. All of this shows that our Mod-X framework indeed improves the representation space quality of the CLIP model during continual training, which provides a good baseline for future continual self-supervised pre-training works.

B.8. The performance of Mod-X when continual training the OpenAI's CLIP on COCO and Flickr30K dataset

We set the CLIP$_{vit32}$ as the initial model, which is consistant with experiment 5.3, and divide the joint-dataset (COCO and Flickr30K) into five sub-datasets uniformly and randomly to simulate streaming data. Because the pre-training datasets of CLIP$_{vit32}$ are not available, we train CLIP$_{vit32}$ on the joint-dataset to get the model CLIP$_{ft}$ as an upper bound for the performance of continual training. We apply our framework Mod-X in this setting and compare the final multimodal retrieval results with CLIP$_{ct}$, which is just continual training without any other operations, in Table 11.

Model	Image-Text Retrieval(%)						Text-Image Retrieval(%)					
	Flickr30k(1K)			COCO(5K)			Flickr30k(1K)			COCO(5K)		
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$
CLIP$_{vit32}$	77.7	94.5	98.3	50.1	74.6	83.0	58.9	83.5	90.1	30.2	55.6	66.7
CLIP$_{ct}$	85.6	97.3	98.8	59.7	83.2	90.2	71.2	91.5	94.9	43.5	70.9	80.6
CLIP$_{Mod-X}$	**86.9**	**97.7**	**99.3**	**62.1**	**85.6**	**91.7**	**73.4**	**92.9**	**96.2**	**46.2**	**73.5**	**82.6**
CLIP$_{ft}$	86.3	97.2	99.1	63.6	86.4	92.3	72.7	92.6	96.3	46.3	73.1	82.3

Table 11. The final multimoal retrieval performance of the CLIP$_{ct}$, CLIP$_{Mod-X}$ and CLIP$_{ft}$ based on OpenAI's CLIP$_{vit32}$ with VIT-B/32 vision encoder.

The performance of our framework Mod-X is still better than CLIP$_{ct}$ on all of the evaluation settings. Comparing the $R@1$ results on the test set Flickr30K(1K), we can find that CLIP$_{Mod-X}$ not only surpasses the initial results (CLIP$_{vit32}$) but also 1.3% points and 2.2% points higher than CLIP$_{ct}$. The results on COCO(5K) also illustrate that our framework not only resists the cognitive disorder of the model but also fits the new data domain better than CLIP$_{ct}$. The $R@1$ results of CLIP$_{Mod-X}$ on COCO(5K) surpasses the CLIP$_{ct}$ by 2.4% and 2.7% points, respectively.