Embracing Language Inclusivity and Diversity in CLIP through Continual Language Learning

Bang Yang[1,2], **Yong Dai**[2], **Xuxin Cheng**[1], **Yaowei Li**[1,2], **Asif Raza**[1], **Yuexian Zou**[1*]

[1] ADSPLAB, School of ECE, Peking University, Shenzhen, China
[2] Pengcheng Laboratory, Shenzhen, China
{yangbang, chengxx, ywl, asifraza151, zouyx}@pku.edu.cn, chd-dy@foxmail.com

Abstract

While vision-language pre-trained models (VL-PTMs) have advanced multimodal research in recent years, their mastery in a few languages like English restricts their applicability in broader communities. To this end, there is an increasing interest in developing multilingual VL models via a joint-learning setup, which, however, could be unrealistic due to expensive costs and data availability. In this work, we propose to extend VL-PTMs' language capacity by continual language learning (CLL), where a model needs to update its linguistic knowledge incrementally without suffering from catastrophic forgetting (CF). We begin our study by introducing a model dubbed CLL-CLIP, which builds upon CLIP, a prevailing VL-PTM that has acquired image-English text alignment. Specifically, CLL-CLIP contains an expandable token embedding layer to handle linguistic differences. It solely trains token embeddings to improve memory stability and is optimized under cross-modal and cross-lingual objectives to learn the alignment between images and multilingual texts. To alleviate CF raised by covariate shift and lexical overlap, we further propose a novel approach that ensures the identical distribution of all token embeddings during initialization and regularizes token embedding learning during training. We construct a CLL benchmark covering 36 languages based on MSCOCO and XM3600 datasets and then evaluate multilingual image-text retrieval performance. Extensive experiments verify the effectiveness of CLL-CLIP and show that our approach can boost CLL-CLIP, e.g., by 6.7% in text-to-image average Recall@1 on XM3600, and improve various state-of-the-art methods consistently. Our code and data are available at https://github.com/yangbang18/CLFM.

Introduction

Large-scale vision-language pre-trained models (VL-PTMs) such as CLIP (Radford et al. 2021), Flamingo (Alayrac et al. 2022), and BLIP-2 (Li et al. 2023a) have made great strides in multimodal research (Gan et al. 2022; Chen et al. 2023a). Nevertheless, the majority of the current literature is biased toward a few languages, predominantly English, making it a barrier to the widespread adoption and accessibility of VL-PTMs across different linguistic communities. Considering that we are living in a world with roughly 7,000 languages, it is indispensable to strive for greater language inclusivity and diversity in VL-PTMs.

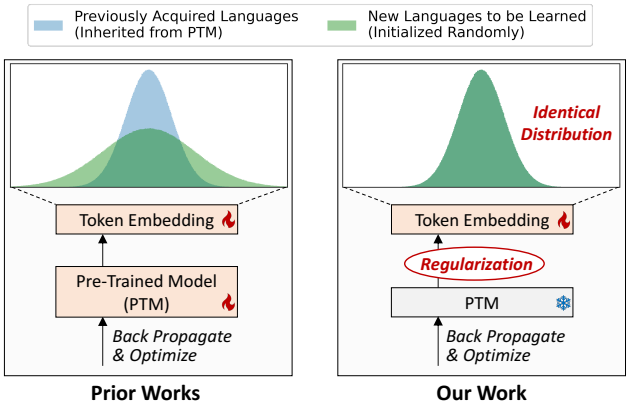

Figure 1: For continual language learning, prior works in NLP (Garcia et al. 2021; Huang et al. 2022) train full model parameters to learn a new language, with new token embeddings initialized randomly without considering the distribution of prior ones. Our work requires the least amount of components to be trained (i.e., the token embedding layer) and targets token embedding initialization and regularization to avert catastrophic forgetting. Note that our frozen vision PTM is not plotted for clarity.

To endow VL-PTMs with an ability to understand multilingual contexts, there is an increasing interest in developing multilingual VL-PTMs via a joint-learning setup (Zhou et al. 2021; Zhang, Hu, and Jin 2022; Chen et al. 2023b; Li et al. 2023b), which has shown remarkable performance in tasks like multilingual image-text retrieval. However, two critical issues plague the joint learning. One is the high computational cost and inflexibility of learning new knowledge, as we need to re-train models on new data alongside all previous data. Another one is that data is not always available during the learning cycle due to privacy and other factors. Alternatively, *continual language learning* (CLL), also known as *lifelong language learning*, is a more practical setup to extend PTMs' language capacity with low costs and high flexibility. The goal of CLL is to consolidate multilingual performance into a single, parameter- and memory-constrained model, ensuring that this model can

*Corresponding Author.

evolve under *non-stationary* data streams without suffering from *catastrophic forgetting* (McCloskey and Cohen 1989). While CLL has been extensively studied in natural language processing (NLP) (Biesialska, Biesialska, and Costa-jussà 2020; Escolano, Costa-Jussà, and Fonollosa 2021; Zhang et al. 2022; M'hamdi, Ren, and May 2023), the effective integration of VL-PTMs with CLL is still under-explored and it presents distinctive challenges like leveraging visual information to aid in language learning.

In this paper, we study the multilingual acquisition of VL-PTMs in the CLL setup. We begin our study by selecting CLIP (Radford et al. 2021), a prevailing VL-PTM that can correlate images and English texts into the same latent space, as our backbone. Next, we propose a model dubbed CLL-CLIP to incrementally learn new languages. Specifically, our model contains an expandable token embedding layer to handle linguistic differences. Such design is crucial to prevent our model from encountering a high portion of *out-of-vocabulary* tokens. During training, CLL-CLIP keeps all pre-trained components frozen except its token embedding layer to retain previously acquired knowledge and is optimized under cross-modal and cross-lingual objectives to learn the alignment between images and multilingual texts.

Next, we propose a CLL approach that targets **T**oken **E**mbedding **I**nitialization and **R**egularization (TEIR) to alleviate catastrophic forgetting (CF). Figure 1 differentiates our TEIR from prior approaches in NLP (Garcia et al. 2021; Huang et al. 2022). In particular, to reduce CF raised by *covariate shift* (Shimodaira 2000; Ioffe and Szegedy 2015), our approach ensures the *identical distribution* of all token embeddings during initialization. To mitigate CF caused by the *lexical overlap* (Pfeiffer et al. 2021), our approach regularizes token embedding learning based on the number of times that tokens appear in the tasks they have already learned by CLL-CLIP. Our insight is that if a token is common in previously learned tasks, its embedding update should be penalized to avoid task interference.

To evaluate the effectiveness of our CLL-CLIP model and TEIR approach, we first construct a benchmark covering 36 languages based on MSCOCO (Chen et al. 2015) and XM3600 (Thapliyal et al. 2022) datasets. We then reproduce various state-of-the-art (SOTA) continual learning and parameter-efficient fine-tuning methods based on our CLL-CLIP model on this benchmark. Extensive experiments verify the effectiveness of CLL-CLIP and show that TEIR can boost CLL-CLIP, e.g., by 6.7% in text-to-image average Recall@1 on XM3600, and improve the performance of SOTA methods consistently.

Our main contributions are as follows. (1) To the best of our knowledge, we present the first systematic study on enhancing the language capacity of dual-stream VL-PTMs through continual language learning. (2) We design a model named CLL-CLIP for this challenging setup and introduce a novel approach called TEIR that underscores the initialization and regularization of token embeddings to mitigate catastrophic forgetting. (3) We construct a CLL benchmark for evaluating image-text retrieval across 36 languages. Extensive experiments verify the effectiveness of our CLL-CLIP and TEIR and demonstrate the generality of TEIR on various SOTA methods.

Related Work

Multilingual VL Pre-Training As monolingual visual-language pre-training models (VL-PTMs) continue to evolve, an increasing amount of effort is directed toward enhancing the adaptability of these models for multilingual scenarios via pre-training. M^3P (Ni et al. 2021) and UC^2 (Zhou et al. 2021) adopt a BERT-like single-stream architecture (Devlin et al. 2019) for pre-training, yet they diverge in their data augmentation strategies. M^3P uses word-level augmentation to obtain code-switched VL pairs, whereas UC^2 utilizes translation engines to transform English image captions into other languages. In contrast, MURAL (Jain et al. 2021), M-CLIP (Carlsson et al. 2022), MLA (Zhang, Hu, and Jin 2022), and mCLIP (Chen et al. 2023b) build their model on a dual-stream model like CLIP for better efficiency on retrieval tasks. These models use the same data augmentation strategy as UC^2, but MURAL and mCLIP additionally consider annotated translation pairs. Besides retrieval tasks, recent encoder-decoder-based PaLI (Chen et al. 2023c) and WS-mVLP (Li et al. 2023b) have shown their superiority in multilingual VL generation tasks. However, all the above methods develop multilingual VL-PTMs via a joint-learning setup and thus suffer from high costs and inflexibility of learning new languages. In this paper, we focus on endowing dual-stream VL-PTMs with a multilingual understanding ability via a more practical and flexible setup, i.e., continual language learning.

Continual Learning (CL) The core aspiration of CL is to enable machines to mimic the strong adaptability of humans to continually acquire, update, organize, and exploit knowledge (Wang et al. 2023). The computer vision (CV) community has witnessed significant advances in CL, which can be mainly divided into four categories. Specifically, *regularization*-based methods penalize changes to model parameters or predictions (Kirkpatrick et al. 2017; Lee et al. 2019; Ahn et al. 2021); *rehearsal*-based methods store historical data or features to retain previously acquired knowledge (Chaudhry et al. 2019; Buzzega et al. 2020; Cha, Lee, and Shin 2021); *architecture*-based methods assign isolated parameters for different tasks (Yoon et al. 2018; Li et al. 2019; Ke, Liu, and Huang 2020); *prompt*-based methods add parameter-efficient modules into frozen PTMs to harness their power (Wang et al. 2022a,b; Smith et al. 2023; Gao et al. 2023). The success of CL in CV inspires related research in NLP (Biesialska, Biesialska, and Costa-jussà 2020; Wu et al. 2022; M'hamdi, Ren, and May 2023). In particular, a line of research studies on how to add new languages to pre-trained neural machine translation models. One attempt is to add and train language-specific components, like encoder/decoder (Escolano, Costa-Jussà, and Fonollosa 2021) and adapter (Berard 2021). Another attempt proposes to substitute models' vocabulary dynamically (Garcia et al. 2021; Huang et al. 2022). In this paper, we differentiate our work from prior ones in NLP in Figure 1. Unlike those regularization methods that need to estimate parameter importance by feeding data into the model, our approach only requires the

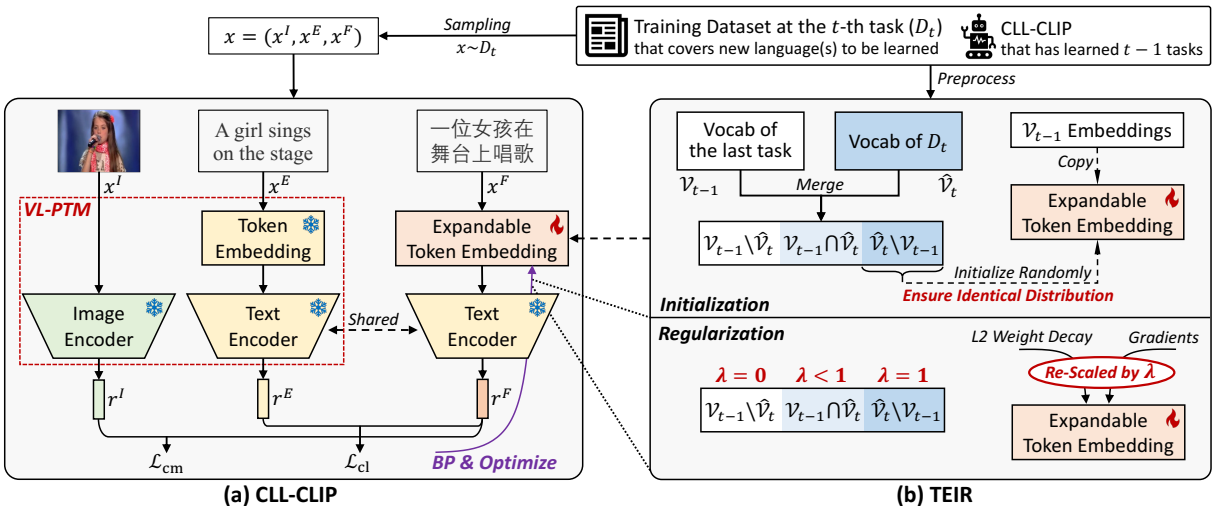

Figure 2: Overview of our proposals. (a): CLL-CLIP builds upon a two-tower VL-PTM (i.e., CLIP), keeps all pre-trained components frozen, and contains an expandable and trainable token embedding layer for continual language learning. (b): Our TEIR approach eases catastrophic forgetting by underscoring the initialization and regularization of token embeddings.

lexical statistics of data. By contrast with the CL of CLIP in visual recognition (Ding et al. 2022; Thengane et al. 2022), we value the CL of CLIP in language acquisition.

Approach

In our continual language learning (CLL) setting, a model needs to sequentially learn T tasks, each with its corresponding training dataset $D_t (t \in [1, T])$ that covers non-overlapping subsets of languages. After training a model parameterized by ϕ_t on D_t, the goal of CLL is to ensure the model can perform well in previous t tasks. To achieve that, we propose CLL-CLIP and TEIR, as introduced next.

CLL-CLIP

Architecture As shown in Figure 2(a), our model builds upon CLIP to avoid from-scratch training and contains an expandable token embedding layer parameterized by $\boldsymbol{\theta}_t$ to vectorize multilingual texts. In particular, CLIP consists of a vision encoder, a text encoder, and a token embedding layer mainly for English[1]. Let denote their parameters as $\boldsymbol{\Omega}_{ve}$, $\boldsymbol{\Omega}_{te}$, and $\boldsymbol{\Omega}_{emb}$, respectively. Then parameters of our model at the t-th task are $\phi_t = \{\boldsymbol{\Omega}_{ve}, \boldsymbol{\Omega}_{te}, \boldsymbol{\Omega}_{emb}, \boldsymbol{\theta}_t\}$, where $\boldsymbol{\Omega}_{emb}$ can be discarded during inference. We keep all CLIP parameters $\boldsymbol{\Omega}_*$ frozen and solely train $\boldsymbol{\theta}_t$. This choice is in line with the research on efficient VL pre-training (Zhai et al. 2022; Zhang, Hu, and Jin 2022) and also benefits the preservation of previously acquired knowledge during continual learning (Wang et al. 2022b; Smith et al. 2023).

Vocab Substitution Let denote the vocab corresponding to $\boldsymbol{\theta}_t$ as $\mathcal{V}_t$. Before training, $\mathcal{V}_0$ is identical to CLIP's vocab, and $\boldsymbol{\theta}_0 = \boldsymbol{\Omega}_{emb}$. For $t \in [1, T]$, $\mathcal{V}_t$ needs to be dynamically updated to accommodate the lexicon of new languages. Thus, we first adopt the same BPE procedure (Sennrich, Haddow, and Birch 2016) as CLIP to build vocab $\hat{\mathcal{V}}_t$ from D_t and then follow (Garcia et al. 2021) to obtain $\mathcal{V}_t$ by merging $\mathcal{V}_{t-1}$ and $\hat{\mathcal{V}}_t$, i.e., $\mathcal{V}_t = \mathcal{V}_{t-1} \cup \hat{\mathcal{V}}_t$. There are two issues to be noted: (1) the embedding initialization of $\hat{\mathcal{V}}_t \setminus \mathcal{V}_{t-1}$ (new tokens that only exist in $\hat{\mathcal{V}}_t$) and (2) the sub-optimal nature of $\mathcal{V}_t$ due to lacking comprehensive text statistics. We will address (1) in TEIR and discuss (2) in later experiments.

Training Objectives Each training sample for our CLL-CLIP is a triplet $x = (x^I, x^E, x^F)$ that includes an image x^I, a text in *native* language x^E (i.e., *English* text), and a *foreign* text x^F. At the t-th task, we obtain global representations of the triplet x as follows:

$$\begin{aligned} \boldsymbol{r}^I &= g(x^I; \boldsymbol{\Omega}_{ve}), \\ \boldsymbol{r}^E &= g(x^E; \boldsymbol{\Omega}_{te}, \boldsymbol{\Omega}_{emb}), \\ \boldsymbol{r}^F &= g(x^F; \boldsymbol{\Omega}_{te}, \boldsymbol{\theta}_t), \end{aligned} \tag{1}$$

where $g(\cdot)$ indicates the feed-forward transformation. We suggest training CLL-CLIP with cross-modal and cross-lingual objectives, i.e., $\mathcal{L}_{\text{cm}}$ and $\mathcal{L}_{\text{cl}}$, so that CLL-CLIP can correlate $\boldsymbol{r}^I$ with $\boldsymbol{r}^F$ based on the already acquired knowledge, i.e., the alignment between $\boldsymbol{r}^I$ and $\boldsymbol{r}^E$. Following CLIP, we implement $\mathcal{L}_{\text{cm}}$ as InfoNCE-based image-text contrast (van den Oord, Li, and Vinyals 2018):

$$\begin{aligned} \mathcal{L}_{\text{cm}} &= \frac{1}{2}\left(\mathcal{L}^{I\to F}_{\text{InfoNCE}} + \mathcal{L}^{F\to I}_{\text{InfoNCE}}\right), \\ \mathcal{L}^{Y\to Z}_{\text{InfoNCE}} &= -\frac{1}{K}\sum_{k=1}^{K}\log\frac{\exp(\langle \boldsymbol{r}^Y_k, \boldsymbol{r}^Z_k\rangle/\tau)}{\sum_{l=1}^{K}\exp(\langle \boldsymbol{r}^Y_k, \boldsymbol{r}^Z_l\rangle/\tau)}, \end{aligned} \tag{2}$$

where K denotes the batch size, $\langle\cdot,\cdot\rangle$ the cosine similarity, and τ a temperature hyper-parameter. Motivated by

[1]We separate token embeddings from CLIP for clarity, and the text encoder means the rest of the components (positional embeddings, Transformer blocks, projection head) to obtain text features.

(Reimers and Gurevych 2020), we implement $\mathcal{L}_{\text{cl}}$ as the mean-square error between paired text features:

$$\mathcal{L}_{\text{cl}} = \frac{1}{2K}\sum_{k=1}^{K}||\boldsymbol{r}_k^E - \boldsymbol{r}_k^F||_2^2, \tag{3}$$

where $||\cdot||_2$ denotes L2-norm. The overall training objective of CLL-CLIP can be formulated as follows:

$$\mathcal{L} = \gamma_1 \cdot \mathcal{L}_{\text{cm}} + \gamma_2 \cdot \mathcal{L}_{\text{cl}}, \tag{4}$$

where γ_* are hyper-parameters to balance two losses.

TEIR

As shown in Figure 2(b), the key of TEIR is how we treat $\mathcal{V}_{t,old} = \mathcal{V}_{t-1} \backslash \hat{\mathcal{V}}_t$, $\mathcal{V}_{t,\cap} = \mathcal{V}_{t-1} \cap \hat{\mathcal{V}}_t$, and $\mathcal{V}_{t,new} = \hat{\mathcal{V}}_t \backslash \mathcal{V}_{t-1}$ differently to mitigate catastrophic forgetting (CF).

Initialization Language models building on Transformer (Vaswani et al. 2017) typically initialize token embeddings with a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ with zero mean ($\mu = 0$) and a pre-defined variance σ^2. Let denote CLL-CLIP's token embeddings after training on D_t as $\boldsymbol{\theta}_t^*$. With the assumption that $\boldsymbol{\theta}_{t-1}^* \sim \mathcal{N}(\mu_{t-1}, \sigma_{t-1}^2)$, the focus now becomes how we initialize $\boldsymbol{\theta}_t$ properly. Following (Garcia et al. 2021), $\boldsymbol{\theta}_t$ inherits pre-trained embeddings of $\mathcal{V}_{t-1}$ from $\boldsymbol{\theta}_{t-1}^*$ to preserve previously acquired linguistic knowledge. Instead of initializing embeddings of $\mathcal{V}_{t,new}$ with a fixed distribution $\mathcal{N}(\mu, \sigma^2)$, we suggest $\mu = \mu_{t-1}$ and $\sigma = \sigma_{t-1}$ to ensure the *identical distribution* of new and prior token embeddings. By doing so, our approach alleviates the feature drift (a.k.a. covariate shift) problem, which is a potential factor to arise CF (Ramasesh, Dyer, and Raghu 2021).

Regularization Although *lexical overlap* is beneficial for transfer learning (Pfeiffer et al. 2021), learning the embeddings of $\mathcal{V}_{t,\cap}$ without constraints will cause interference to the performance of previous tasks that contain lexically overlapping tokens. Let denote token statistics till the t-th task as $\boldsymbol{c}_t \in \mathbb{R}^{|\mathcal{V}_t|}$, where $c_{t,j}$ is the number of times that the j-th token appears in prior $t-1$ tasks and $c_{1,j}$ is initialized as 1. To overcome CF raised by lexical overlap, we re-scale the rate of L2 weight decay β and gradients $\nabla\mathcal{L}(\boldsymbol{\theta}_t)$ w.r.t. token embeddings $\boldsymbol{\theta}_t$ as follows, with standard stochastic gradient descent (SGD) with L2 weight decay as an example:

$$\boldsymbol{\theta}_{t,j} \leftarrow (1 - \alpha\beta\lambda_{t,j})\boldsymbol{\theta}_{t,j} - \alpha\lambda_{t,j}\nabla\mathcal{L}(\boldsymbol{\theta}_{t,j}) \tag{5}$$

where α is a learning rate, and $\lambda_{t,j}$ is defined as:

$$\lambda_{t,j} = \begin{cases} 0, & \text{if token}_j \in \mathcal{V}_{t,old} \\ 1/(c_{t,j}+1), & \text{if token}_j \in \mathcal{V}_{t,\cap} \\ 1, & \text{if token}_j \in \mathcal{V}_{t,new} \end{cases} \tag{6}$$

For sophisticated optimizers with momentum, the scaling operation is still applied on β and $\nabla\mathcal{L}(\boldsymbol{\theta}_t)$ directly. As indicated by Equation (5) and (6), we keep token embeddings unrelated to the t-th task intact, penalize embedding learning of $\mathcal{V}_{t,\cap}$, while updating embeddings of $\mathcal{V}_{t,new}$ as usual. This method averts task interference and ensures the effective learning of text features ($\boldsymbol{r}^F$), leading to a better trade-off between memory stability and learning plasticity.

	MSCOCO_{36}	XM3600
# Train/Val/Test Images	113,287/5,000/5,000	-/-/3,600
# Languages	1 + 35	36
# Captions per Language	616,767	≈7260

Table 1: Dataset Statistics. # means "The number of". MSCOCO_{36} is obtained by translating the English captions of MSCOCO into the other 35 languages in XM3600 via Google Translator, following (Thapliyal et al. 2022).

Experiments

Experimental Settings

Benchmark We build a CLL benchmark based on **MSCOCO** (Chen et al. 2015) and **XM3600** (Thapliyal et al. 2022) to evaluate the effectiveness of our proposals. Here are the reasons: (1) MSCOCO is a popular VL benchmark and it contains high-quality image-English caption pairs. (2) XM3600 consists of image-caption pairs in 36 languages[2] spoken by geographically-diverse people. This dataset covers the most diverse languages to our best knowledge. (3) The multi-lingual VL benchmark IGLUE (Bugliarello et al. 2022) varies in both task types and languages, making it hard to justify the effect of linguistic differences. As shown in Table 1, we use Google Translator[3] for data augmentation and thus obtain a multilingual dataset named MSCOCO_{36}. We train models on MSCOCO_{36} based on the Karpathy split (Karpathy and Fei-Fei 2015). Then, we report *in-domain* and *out-of-domain* results on MSCOCO_{36} and XM3600, respetively.

Tasks and Task Order We treat each language as a task and thus obtain $T = 36$ tasks. Models are trained on the English task first and then the rest 35 tasks in a random order.

Metrics Let $a_{j,i}$ ($j \geq i$) denotes Recall@1 (a popular metric in information retrieval) on the i-th task after training on the j-th task. In line with the continual learning research (Wang et al. 2023), we compute two metrics:

- Average Recall: $\mathbf{AR}_j = \frac{1}{j}\sum_{i=1}^{j} a_{j,i}$, a composite metric for a model's learning capacity and memory stability.
- Forgetting: $\mathbf{F}_j = \frac{1}{j-1}\sum_{i=1}^{j-1} \max_{k\in[1,j-1]}(a_{k,i} - a_{j,i})$, whose lower value means less catastrophic forgetting.

Unless otherwise specified, we report the-end $\mathbf{AR}_T$ and $\mathbf{F}_T$ performance in *percentile* and omit the subscript.

Implementation Details We follow (Zhang, Hu, and Jin 2022; Yang et al. 2023) to adopt the ViT-B/16 variant of CLIP as the backbone. We follow OpenCLIP (Ilharco et al. 2021) and set the initial temperature of $\mathcal{L}_{\text{cm}}$ to 0.07. We

[2]The 36 languages are Arabic, Bengali, Czech, Danish, German, Greek, English, Spanish, Farsi, Finnish, Filipino, French, Hebrew, Hindi, Croatian, Hungarian, Indonesian, Italian, Japanese, Korean, Maori, Dutch, Norwegian, Polish, Portuguese, Cusco Quechua, Romanian, Russian, Swedish, Swahili, Telugu, Thai, Turkish, Ukrainian, Vietnamese, and Chinese-Simplified.

[3]https://translate.google.com/

Setting	Model	MSCOCO_{36} (In-Domain)				XM3600 (Out-of-Domain)			
		Image-to-Text		Text-to-Image		Image-to-Text		Text-to-Image	
		AR (↑)	F (↓)	AR (↑)	F (↓)	AR (↑)	F (↓)	AR (↑)	F (↓)
Joint Learning	CLL-CLIP	**53.3**	-	**31.4**	-	50.7	-	37.1	-
	M-CLIP (2022)	42.7	-	25.9	-	**53.6**	-	**41.1**	-
	PaLI (2023c)	-	-	-	-	36.0	-	28.5	-
Continual Learning	CLL-CLIP	29.6	23.2	15.2	15.6	26.4	23.1	17.6	18.4
	with TEIR	**38.3** (+8.7)	**14.7** (+8.5)	**20.5** (+5.3)	**10.5** (+5.1)	**35.0** (+8.6)	**15.3** (+7.8)	**24.3** (+6.7)	**12.5** (+5.9)
	oEWC (2018)	37.0	15.7	19.3	11.3	32.3	17.2	21.8	14.1
	with TEIR	**40.2** (+3.2)	**12.7** (+3.0)	**21.6** (+2.3)	**9.3** (+2.0)	**36.7** (+4.4)	**13.4** (+3.8)	**25.6** (+3.8)	**11.2** (+2.9)
	ER (2019)	34.1	17.9	17.8	12.3	29.0	20.0	19.4	16.0
	with TEIR	**39.3** (+5.2)	**12.8** (+5.1)	**21.5** (+3.7)	**8.8** (+3.5)	**35.4** (+6.4)	**13.9** (+6.1)	**24.7** (+5.3)	**11.2** (+4.8)
	DER (2020)	37.6	14.6	19.5	10.6	31.6	17.4	21.0	14.4
	with TEIR	**42.7** (+5.1)	**9.4** (+5.2)	**23.4** (+3.9)	**6.9** (+3.7)	**38.3** (+6.7)	**10.9** (+6.5)	**26.7** (+5.7)	**9.3** (+5.1)
	MLA† (2022)	35.9	20.9	18.4	15.0	30.7	21.8	20.6	18.1
	with TEIR	**46.0** (+10.1)	**11.2** (+9.7)	**25.2** (+6.8)	**8.6** (+6.4)	**41.1** (+10.4)	**12.3** (+9.5)	**29.0** (+8.4)	**10.7** (+7.4)
	P-Tuning† (2022)	30.1	23.9	15.0	16.3	24.9	23.9	16.4	19.3
	with TEIR	**41.1** (+11.0)	**13.3** (+10.6)	**22.2** (+7.2)	**9.6** (+6.7)	**35.5** (+10.6)	**13.8** (+10.1)	**25.4** (+9.0)	**11.5** (+7.8)
	LoRA† (2022)	31.8	22.5	16.2	15.9	28.0	22.7	18.7	18.9
	with TEIR	**41.6** (+9.8)	**12.9** (+9.6)	**22.8** (+6.6)	**9.7** (+6.2)	**38.0** (+10.0)	**13.9** (+8.8)	**27.0** (+8.3)	**11.7** (+7.2)
	DualPrompt (2022a)	28.4	23.6	14.1	15.8	25.5	22.9	16.4	18.4
	with TEIR	**38.3** (+9.9)	**14.0** (+9.6)	**19.7** (+5.6)	**10.6** (+5.2)	**35.3** (+9.8)	**14.1** (+8.8)	**23.6** (+7.2)	**12.1** (+6.3)
	CodaPrompt (2023)	28.9	22.6	14.4	15.2	24.6	22.2	15.9	17.6
	with TEIR	**41.4** (+12.5)	**9.7** (+12.9)	**22.3** (+7.9)	**7.1** (+8.1)	**36.7** (+12.1)	**9.3** (+12.9)	**25.3** (+9.4)	**7.9** (+9.7)

Table 2: Retrieval performance on MSCOCO_{36} and XM3600. †: Task identity is needed during inference. All results are reproduced by ourselves except that of PaLI. Note that PaLI is not optimized for image-text retrieval, but we draw its results from (Chen et al. 2023c) for completeness. The numbers in brackets indicate the absolute improvements brought by our approach.

search the hyperparameters γ_1 and γ_2 in Equation (4) from values $\{1, 0.1, 0.01\}$ and set $\gamma_1 = 0.01$ and $\gamma_2 = 1$ based on the AR metric on the validation set. For models without TEIR, we initialize new token embeddings with $\mathcal{N}(0, 0.02^2)$ following OpenCLIP and set $\forall t, \forall j, \lambda_{t,j} = 1$ (Equation (6)). For each task, we set the vocab size to 10K. We use batches of 128 samples and AdamW (Loshchilov and Hutter 2019) with L2 weight decay of 0.05 to train models for 3 epochs. We set the learning rate fixed to 5e-5 after 10% warm-up iterations. The model achieving the highest summation of Recall@{1, 5, 10} on the current-task validation set is selected for training on the next task. We conduct experiments in PyTorch on a single NVIDIA V100 card and every run of an experiment takes less than 20 hours.

Comparing Methods We reproduce the following SOTA continual learning (CL) and parameter-efficient fine-tuning (PEFT) methods for comparisons: (1) regularization-based online Elastic Weight Consolidation (**oEWC**) (Schwarz et al. 2018) that penalizes the changes in model parameters; (2) rehearsal-based **ER** (Chaudhry et al. 2019) that stores historical training samples for current-task learning; (3) rehearsal- and regularization-based **DER** (Buzzega et al. 2020) that stores features of previously learned samples for knowledge distillation; (4) architecture-based **MLA** (Zhang, Hu, and Jin 2022), **P-Tuning** (Liu et al. 2022), and **LoRA** (Hu et al. 2022) that inserts task-specific adapters (Houlsby et al. 2019), learnable prompt tokens, and decomposed matrices into frozen PTMs, respectively. (5) prompt-based **DualPrompt** (Wang et al. 2022a) and **CodaPrompt** (Smith et al. 2023) that rely on a key-query mechanism to generate proper prompts for frozen PTMs. We reproduce all the above methods in the text branch of CLL-CLIP with the aforementioned implementation details.

Main Results

Table 2 provides retrieval results of different models. Specifically, joint-learning models CLL-CLIP and M-CLIP (Carlsson et al. 2022) respectively achieve the highest AR scores on MSCOCO_{36} and XM3600. As the joint-learning setting covers all languages at the (pre-)training stage, its results can be regarded as the *upper bound* of CL models. When learning different languages incrementally, all CL models experience different levels of forgetting. Notably, our TEIR can consistently boost all CL models across all metrics and datasets, e.g., with absolute improvements ranging from 3.7% to 10.2% in text-to-image AR on XM3600. The im-

Setting	Initialization	Regularization		Oracle Vocab	MSCOCO_{36} (In-Domain)				XM3600 (Out-of-Domain)			
	Identical Distribution	Gradient	Weight Decay		Image-to-Text		Text-to-Image		Image-to-Text		Text-to-Image	
					AR (↑)	F (↓)	AR (↑)	F (↓)	AR (↑)	F (↓)	AR (↑)	F (↓)
(1): CLL-CLIP				×	29.6	23.2	15.2	15.6	26.4	23.1	17.6	18.4
(2)	√			×	32.4	21.9	16.8	15.1	29.9	22.5	20.5	18.2
(3)		√		×	31.9	20.3	16.9	13.5	28.4	20.2	19.5	16.1
(4)			√	×	33.3	19.2	17.0	13.5	30.0	19.0	19.9	15.5
(5)		√	√	×	37.2	14.9	19.7	10.6	33.3	15.4	22.8	12.6
(6): (1) + TEIR	√	√	√	×	**38.3**	**14.7**	**20.5**	**10.5**	**35.0**	**15.3**	**24.3**	**12.5**
(7)	√	√	√	√	**42.4**	**10.5**	**23.2**	**7.9**	**38.4**	**12.0**	**27.1**	**10.0**

Table 3: Ablation study on MSCOCO_{36} and XM3600. By default, we dynamically substitute the model's vocab when new languages arrive, whereas setting (7) requires the accessibility of corpora of all languages to construct a task-shared vocab.

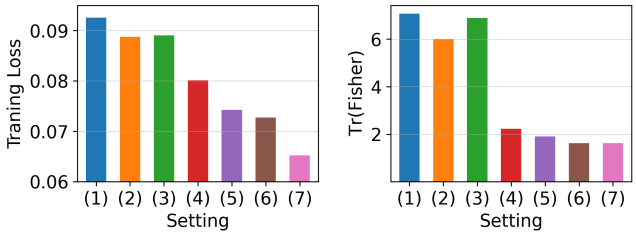

Figure 3: Convergence analysis for different settings in Table 3, focusing (left) the training loss and (right) the Fisher eigenvalues. Lower values respectively indicate closer to global minima and the convergence to flatter minima.

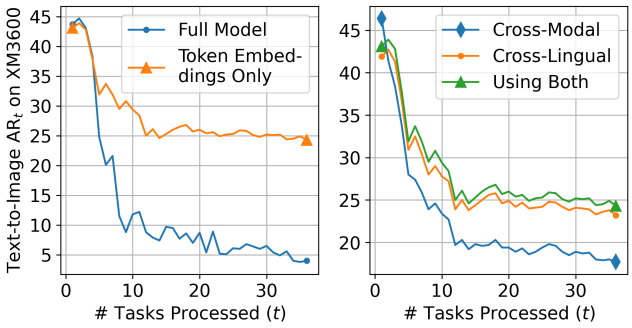

Figure 4: Analysis of CLL-CLIP's core designs: (left) trainable components and (right) training objectives.

proved performance demonstrates the generality of TEIR across various CL and PEFT methods, proves the validity of our approach to maintaining acquired language skills, and highlights the importance of proper token embedding initialization and regularization.

Ablations and Additional Analyses

In the following, we delve deeper into our proposals via ablation studies and additional analyses, with "CLL-CLIP with TEIR" as the default model unless otherwise specified.

Effect of Initialization Table 3(1,2) shows that ensuring identical distribution of new and prior token embeddings during initialization improves AR and F metrics by large margins. Compared with setting (5), setting (6) can still improve the model's learning capacity without sacrificing memory stability. These results suggest the importance of addressing the covariate shift problem in CLL.

Effect of Regularization Table 3(3,4) shows that imposing constraints on gradients or L2 weight decay when updating token embeddings can effectively mitigate the catastrophic forgetting problem of CLL-CLIP. So, it is crucial to penalize the embedding learning of lexically overlapping tokens and keep unrelated token embeddings intact. Moreover, the superiority of setting (5) against (3,4) indicates the complementary nature of these two strategies. Since our regularization method solely relies on the lexical statistics of the data, it incurs negligible additional costs, e.g., the training time of settings (1,6) is 11.1 and 11.3 hours, respectively.

Effect of Vocab Substitution Strategy We stick to the principle of continual learning and thus dynamically substitute the model's vocab when new languages arrive. In contrast, if we are allowed to access corpora of all languages, we can build an *oracle* vocab and only need to substitute the model's vocab at the beginning. As shown in Table 3(6,7), using the oracle vocab contributes to a boost in performance. Since we employ BPE to construct vocab in this work, the improvements confirm BPE's capacity to learn more accurate merging operations of sub-word units from extensive text statistics. Therefore, the exploration of refined vocab substitution strategies is a valuable avenue in future studies.

Effect of TEIR on Model Convergence We consider the model at the end of training and measure the property of the training minima of settings (1-7) in Table 3. Firstly, we calculate the average loss across all training samples of MSCOCO_{36}. As depicted in Figure 3(left), the training loss of settings (2-7) is lower than that of (1), illustrating that TEIR facilitates the convergence of CLL-CLIP towards a *global* minimum. Furthermore, we compute the trace of the empirical Fisher information matrix w.r.t. all training samples of MSCOCO_{36} and treat it as a proxy for Hessian eigenvalues following (Chaudhari et al. 2017; Kirkpatrick et al.

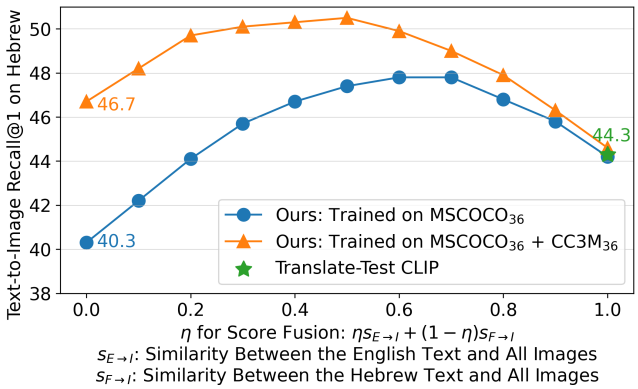

Figure 5: Translate-test performance on Hebrew data in XM3600. Although translate-test CLIP is a strong pipeline system, our model can process foreign texts directly ($\eta = 0$) or achieve better retrieval performance via score fusion when translations are available ($\eta > 0$).

2017; Buzzega et al. 2020). As depicted in Figure 3(right), settings (2-7) produce lower eigenvalues than (1), revealing that TEIR helps CLL-CLIP converge to a *flatter* minimum.

Effectiveness of CLL-CLIP We here ablate the core designs of CLL-CLIP, including the trainable components and training objectives. As shown in Figure 4(left), training the full model obtains dramatically degrades as more tasks are processed. Instead, our proposal of solely training token embeddings can preserve knowledge effectively. In Figure 4(right), we can find that the cross-lingual objective is more efficient than the cross-modal objective to align images with multi-lingual texts, and leveraging both of them can achieve better results. This observation indicates the potential of utilizing text-only pairs for CLL.

Comparison with Translate-Test CLIP To enable the original CLIP to understand multilingual texts, one intuitive approach is translating foreign texts into English, which is known as *Translate Test* in the literature (Bugliarello et al. 2022; Li et al. 2023b). In Figure 5, we compare our model with the translate-test CLIP under different η for score fusion. Specifically, we can see that translate-test CLIP is a strong pipeline system that can achieve 44.3 text-to-image Recall@1 on Hebrew. In contrast, our CLL model can process Hebrew texts directly ($\eta = 0$). Encouragingly, our single model can even surpass the translate-test CLIP (46.7 vs. 44.3) when including an additional augmented CC3M dataset (Sharma et al. 2018) for training. Therefore, continual language learning presents a viable avenue to evade the computation costs of translation and the error accumulation problem of a translation-based pipeline system. Moreover, when translations are available ($\eta > 0$), our model can simultaneously measure image-English text and image-Hebrew text similarities to achieve score fusion, leading to better retrieval performance compared with the case $\eta = 0$.

Comparisons with Multilingual VL-PTMs In Table 4, we compare CLL models with multilingual VL-PTMs on Multi30K (Elliott et al. 2016). As we can see, although the joint-learning MLA model performs generally the best among the four languages, it obtains inferior performance in Czech compared with "MLA with TEIR" which has learned 36 languages in a continual learning manner. Given the gap between joint learning and continual learning, there is much room for improving CLL models.

Setting	Model	en	de	fr	cs	Avg.
Learn in English	CLIP (2021)	86.3	38.4	48.9	8.1	45.4
Joint Learning (<10 languages)	M^3P (2021)	57.9	36.8	27.1	20.4	35.6
	UC^2 (2021)	66.6	62.5	60.4	55.1	61.2
	MLA (2022)	**86.4**	**80.8**	**80.9**	72.9	**80.3**
(>60 languages)	M-CLIP (2022)	84.1	79.1	77.5	**76.3**	79.3
Continual Learning (36 languages)	CLL-CLIP	75.1	36.2	46.5	57.6	53.8
	with TEIR	**82.5**	**48.6**	**60.4**	**66.5**	**64.5**
	MLA (2022)	73.8	42.7	52.7	64.6	58.4
	with TEIR	**82.4**	**58.1**	**68.0**	**74.1**	**70.7**

Table 4: Zero-shot image-text retrieval results (averaged over recall@{1,5,10} on two directions) on Multi30K under English (en), German (de), French (fr), and Czech (cs).

Conclusion

In this paper, we present to our best knowledge the first systematical study on extending the language capacities of dual-stream vision-language pre-trained models (VL-PTMs) under the practical continual language learning setting. We introduce a CLL-CLIP model and a TEIR approach to learn the alignment between images and multilingual texts while mitigating catastrophic forgetting raised by the covariate shift and lexical overlap problems. To comprehensively validate our proposals, we construct a benchmark spanning 36 languages and conduct evaluations on multilingual image-text retrieval. Through a series of experiments and analyses, we verify the effectiveness of CLL-CLIP and TEIR and gain insights into their inner workings. We hope our research can serve as a basis to enhance the accessibility of VL-PTMs across different linguistic communities.

Limitations This paper focuses exclusively on the continual language learning of CLIP-like VL-PTMs, emphasizing evaluations for image-text retrieval. Nonetheless, we posit that our ideas hold the potential to be adaptable to encoder-decoder-based VL-PTMs and generation tasks like visual captioning (Yang, Cao, and Zou 2023). We leave it to our future study. Moreover, TEIR requires current-task text statistics to compute Equation (6), making it difficult to handle the challenges posed by, e.g., *boundary-free continual learning* (Aljundi, Kelchtermans, and Tuytelaars 2019).

Acknowledgements

This paper was partially supported by NSFC (No. 6217600 8), the project of Pengcheng Laboratory (PCL2023A08), and Shenzhen Science and Technology Research Program (No. GXWD20201231165807007-20200814115301001).

References

Ahn, H.; Kwak, J.; Lim, S.; Bang, H.; Kim, H.; and Moon, T. 2021. SS-IL: Separated Softmax for Incremental Learning. In *ICCV*, 844–853.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; Ring, R.; Rutherford, E.; Cabi, S.; Han, T.; Gong, Z.; Samangooei, S.; Monteiro, M.; Menick, J.; Borgeaud, S.; Brock, A.; Nematzadeh, A.; Sharifzadeh, S.; Binkowski, M.; Barreira, R.; Vinyals, O.; Zisserman, A.; and Simonyan, K. 2022. Flamingo: A Visual Language Model for Few-Shot Learning. In *NeurIPS*, 23716–23736.

Aljundi, R.; Kelchtermans, K.; and Tuytelaars, T. 2019. Task-Free Continual Learning. In *CVPR*, 11246–11255.

Berard, A. 2021. Continual Learning in Multilingual NMT via Language-Specific Embeddings. In *WMT*, 542–565.

Biesialska, M.; Biesialska, K.; and Costa-jussà, M. R. 2020. Continual Lifelong Learning in Natural Language Processing: A Survey. In *COLING*, 6523–6541.

Bugliarello, E.; Liu, F.; Pfeiffer, J.; Reddy, S.; Elliott, D.; Ponti, E. M.; and Vulić, I. 2022. IGLUE: A Benchmark for Transfer Learning across Modalities, Tasks, and Languages. In *ICML*, 2370–2392.

Buzzega, P.; Boschini, M.; Porrello, A.; Abati, D.; and CALDERARA, SIMONE. 2020. Dark Experience for General Continual Learning: A Strong, Simple Baseline. In *NeurIPS*, 15920–15930.

Carlsson, F.; Eisen, P.; Rekathati, F.; and Sahlgren, M. 2022. Cross-Lingual and Multilingual CLIP. In *LREC*, 6848–6854.

Cha, H.; Lee, J.; and Shin, J. 2021. Co2L: Contrastive Continual Learning. In *ICCV*, 9516–9525.

Chaudhari, P.; Choromanska, A.; Soatto, S.; LeCun, Y.; Baldassi, C.; Borgs, C.; Chayes, J.; Sagun, L.; and Zecchina, R. 2017. Entropy-SGD: Biasing Gradient Descent Into Wide Valleys. In *ICLR*, 1–19.

Chaudhry, A.; Rohrbach, M.; Elhoseiny, M.; Ajanthan, T.; Dokania, P. K.; Torr, P. H. S.; and Ranzato, M. 2019. On Tiny Episodic Memories in Continual Learning. arxiv:1902.10486.

Chen, F.-L.; Zhang, D.-Z.; Han, M.-L.; Chen, X.-Y.; Shi, J.; Xu, S.; and Xu, B. 2023a. VLP: A Survey on Vision-Language Pre-Training. *Mach. Intell. Res.*, 20(1): 38–56.

Chen, G.; Hou, L.; Chen, Y.; Dai, W.; Shang, L.; Jiang, X.; Liu, Q.; Pan, J.; and Wang, W. 2023b. mCLIP: Multilingual CLIP via Cross-Lingual Transfer. In *ACL*, 13028–13043.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollar, P.; and Zitnick, C. L. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. arxiv:1504.00325.

Chen, X.; Wang, X.; Changpinyo, S.; Piergiovanni, A. J.; Padlewski, P.; Salz, D.; Goodman, S.; Grycner, A.; Mustafa, B.; Beyer, L.; Kolesnikov, A.; Puigcerver, J.; Ding, N.; Rong, K.; Akbari, H.; Mishra, G.; Xue, L.; Thapliyal, A.; Bradbury, J.; Kuo, W.; Seyedhosseini, M.; Jia, C.; Ayan, B. K.; Riquelme, C.; Steiner, A.; Angelova, A.; Zhai, X.; Houlsby, N.; and Soricut, R. 2023c. PaLI: A Jointly-Scaled Multilingual Language-Image Model. In *ICLR*, 1–33.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, 4171–4186.

Ding, Y.; Liu, L.; Tian, C.; Yang, J.; and Ding, H. 2022. Don't Stop Learning: Towards Continual Learning for the CLIP Model. arxiv:2207.09248.

Elliott, D.; Frank, S.; Sima'an, K.; and Specia, L. 2016. Multi30K: Multilingual English-German Image Descriptions. In *ACL Workshop: VL*, 70–74.

Escolano, C.; Costa-Jussà, M. R.; and Fonollosa, J. A. R. 2021. From Bilingual to Multilingual Neural-based Machine Translation by Incremental Training. *J. Assoc. Inf. Sci. Technol*, 72(2): 190–203.

Gan, Z.; Li, L.; Li, C.; Wang, L.; Liu, Z.; and Gao, J. 2022. Vision-Language Pre-Training: Basics, Recent Advances, and Future Trends. *Found. Trends Comput. Graph. Vis.*, 14(3–4): 163–352.

Gao, Q.; Zhao, C.; Sun, Y.; Xi, T.; Zhang, G.; Ghanem, B.; and Zhang, J. 2023. A Unified Continual Learning Framework with General Parameter-Efficient Tuning. In *ICCV*, 11483–11493.

Garcia, X.; Constant, N.; Parikh, A.; and Firat, O. 2021. Towards Continual Learning for Multilingual Machine Translation via Vocabulary Substitution. In *NAACL-HLT*, 1184–1192.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; Laroussilhe, Q. D.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-Efficient Transfer Learning for NLP. In *ICML*, 2790–2799.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*, 1–13.

Huang, K.; Li, P.; Ma, J.; and Liu, Y. 2022. Entropy-Based Vocabulary Substitution for Incremental Learning in Multilingual Neural Machine Translation. In *EMNLP*, 10537–10550.

Ilharco, G.; Wortsman, M.; Carlini, N.; Taori, R.; Dave, A.; Shankar, V.; Namkoong, H.; Miller, J.; Hajishirzi, H.; Farhadi, A.; and Schmidt, L. 2021. OpenCLIP. Zenodo.

Ioffe, S.; and Szegedy, C. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *ICML*, 448–456.

Jain, A.; Guo, M.; Srinivasan, K.; Chen, T.; Kudugunta, S.; Jia, C.; Yang, Y.; and Baldridge, J. 2021. MURAL: Multimodal, Multitask Representations Across Languages. In *EMNLP*, 3449–3463.

Karpathy, A.; and Fei-Fei, L. 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *CVPR*, 3128–3137.

Ke, Z.; Liu, B.; and Huang, X. 2020. Continual Learning of a Mixed Sequence of Similar and Dissimilar Tasks. In *NeurIPS*, 18493–18504.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; Hassabis, D.; Clopath, C.; Kumaran, D.; and Hadsell, R. 2017. Overcoming Catastrophic Forgetting in Neural Networks. *PNAS*, 114(13): 3521–3526.

Lee, K.; Lee, K.; Shin, J.; and Lee, H. 2019. Overcoming Catastrophic Forgetting With Unlabeled Data in the Wild. In *ICCV*, 312–321.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. BLIP-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models. In *ICML*, 19730–19742.

Li, X.; Zhou, Y.; Wu, T.; Socher, R.; and Xiong, C. 2019. Learn to Grow: A Continual Structure Learning Framework for Overcoming Catastrophic Forgetting. In *ICML*, 3925–3934.

Li, Z.; Fan, Z.; Chen, J.; Zhang, Q.; Huang, X.; and Wei, Z. 2023b. Unifying Cross-Lingual and Cross-Modal Modeling Towards Weakly Supervised Multilingual Vision-Language Pre-Training. In *ACL*, 5939–5958.

Liu, X.; Ji, K.; Fu, Y.; Tam, W.; Du, Z.; Yang, Z.; and Tang, J. 2022. P-Tuning: Prompt Tuning Can Be Comparable to Fine-Tuning Across Scales and Tasks. In *ACL*, 61–68.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *ICLR*, 1–18.

McCloskey, M.; and Cohen, N. J. 1989. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. *Psychol. Learn. Motiv.*, 24: 109–165.

M'hamdi, M.; Ren, X.; and May, J. 2023. Cross-Lingual Continual Learning. In *ACL*, 3908–3943.

Ni, M.; Huang, H.; Su, L.; Cui, E.; Bharti, T.; Wang, L.; Zhang, D.; and Duan, N. 2021. M3P: Learning Universal Representations via Multitask Multilingual Multimodal Pre-Training. In *CVPR*, 3976–3985.

Pfeiffer, J.; Vulić, I.; Gurevych, I.; and Ruder, S. 2021. UNKs Everywhere: Adapting Multilingual Language Models to New Scripts. In *EMNLP*, 10186–10203.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 8748–8763.

Ramasesh, V. V.; Dyer, E.; and Raghu, M. 2021. Anatomy of Catastrophic Forgetting: Hidden Representations and Task Semantics. In *ICLR*, 1–31.

Reimers, N.; and Gurevych, I. 2020. Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation. In *EMNLP*, 4512–4525.

Schwarz, J.; Czarnecki, W.; Luketina, J.; Grabska-Barwinska, A.; Teh, Y. W.; Pascanu, R.; and Hadsell, R. 2018. Progress & Compress: A Scalable Framework for Continual Learning. In *ICML*, 4528–4537.

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Neural Machine Translation of Rare Words with Subword Units. In *ACL*, 1715–1725.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-Text Dataset For Automatic Image Captioning. In *ACL*, 2556–2565.

Shimodaira, H. 2000. Improving Predictive Inference under Covariate Shift by Weighting the Log-Likelihood Function. *J. Stat. Plan. Inference*, 90(2): 227–244.

Smith, J. S.; Karlinsky, L.; Gutta, V.; Cascante-Bonilla, P.; Kim, D.; Arbelle, A.; Panda, R.; Feris, R.; and Kira, Z. 2023. CODA-Prompt: COntinual Decomposed Attention-Based Prompting for Rehearsal-Free Continual Learning. In *CVPR*, 11909–11919.

Thapliyal, A. V.; Pont Tuset, J.; Chen, X.; and Soricut, R. 2022. Crossmodal-3600: A Massively Multilingual Multimodal Evaluation Dataset. In *EMNLP*, 715–729.

Thengane, V.; Khan, S.; Hayat, M.; and Khan, F. 2022. CLIP Model Is an Efficient Continual Learner. arxiv:2210.03114.

van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation Learning with Contrastive Predictive Coding. arxiv:1807.03748.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention Is All You Need. In *NeurIPS*, 5998–6008.

Wang, L.; Zhang, X.; Su, H.; and Zhu, J. 2023. A Comprehensive Survey of Continual Learning: Theory, Method and Application. arxiv:2302.00487.

Wang, Z.; Zhang, Z.; Ebrahimi, S.; Sun, R.; Zhang, H.; Lee, C.-Y.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022a. DualPrompt: Complementary Prompting for Rehearsal-Free Continual Learning. In *ECCV*, 631–648.

Wang, Z.; Zhang, Z.; Lee, C.-Y.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022b. Learning to Prompt for Continual Learning. In *CVPR*, 139–149.

Wu, T.; Caccia, M.; Li, Z.; Li, Y.-F.; Qi, G.; and Haffari, G. 2022. Pretrained Language Model in Continual Learning: A Comparative Study. In *ICLR*, 1–17.

Yang, B.; Cao, M.; and Zou, Y. 2023. Concept-Aware Video Captioning: Describing Videos With Effective Prior Information. *IEEE Trans. Image Process.*, 32: 5366–5378.

Yang, B.; Liu, F.; Wu, X.; Wang, Y.; Sun, X.; and Zou, Y. 2023. MultiCapCLIP: Auto-Encoding Prompts for Zero-Shot Multilingual Visual Captioning. In *ACL*, 11908–11922.

Yoon, J.; Yang, E.; Lee, J.; and Hwang, S. J. 2018. Lifelong Learning with Dynamically Expandable Networks. In *ICLR*, 1–11.

Zhai, X.; Wang, X.; Mustafa, B.; Steiner, A.; Keysers, D.; Kolesnikov, A.; and Beyer, L. 2022. LiT: Zero-Shot Transfer With Locked-Image Text Tuning. In *CVPR*, 18123–18133.

Zhang, H.; Zhang, S.; Xiang, Y.; Liang, B.; Su, J.; Miao, Z.; Wang, H.; and Xu, R. 2022. CLLE: A Benchmark for Continual Language Learning Evaluation in Multilingual Machine Translation. In *EMNLP*, 428–443.

Zhang, L.; Hu, A.; and Jin, Q. 2022. Multi-Lingual Acquisition on Multimodal Pre-Training for Cross-Modal Retrieval. In *NeurIPS*, 29691–29704.

Zhou, M.; Zhou, L.; Wang, S.; Cheng, Y.; Li, L.; Yu, Z.; and Liu, J. 2021. UC2: Universal Cross-Lingual Cross-Modal Vision-and-Language Pre-Training. In *CVPR*, 4153–4163.