Universiteit
Leiden
The Netherlands

Exploring deep learning for multimodal understanding

Lao, M.

Citation
Lao, M. (2023, November 28). *Exploring deep learning for multimodal understanding*. Retrieved from https://hdl.handle.net/1887/3665082

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/3665082

Note: To cite this publication please use the final published version (if applicable).

Chapter 7

Multi-Domain Lifelong Visual Question Answering via Self-Critical Distillation

In the previous chapters, we analyze VQA algorithms from both model performance and robustness aspects in the stationary training process, which is fixed by the choice of a given dataset. In real-world scenarios, these methods are often inefficient because VQA systems are always supposed to extend their knowledge and meet the ever-changing demands of users. In this chapter, we turn to focus on **RQ5** to put VQA algorithm into the practical lifelong learning setting.

In this chapter, we introduce a new and challenging multi-domain lifelong VQA task, dubbed MDL-VQA, which encourages the VQA model to continuously learn across multiple domains while mitigating the forgetting on previously-learned domains. Furthermore, we propose a novel replay-free Self-Critical Distillation (SCD) framework tailor-made for MDL-VQA, which alleviates forgetting issue via transferring previous-domain knowledge from teacher to student models. First, we propose to introspect the teacher's understanding over original and counterfactual samples, thereby creating informative instance-relevant and domain-relevant knowledge for logits-based distillation. Second, on the side of feature-based distillation, we propose to introspect the reasoning behavior of student model to establish the harmful domain-specific knowledge acquired in current domain, and further leverage the metric learning strategy to encourage student to learn useful knowledge in new domain. Through blending such two components, our SCD enhances the VQA model's stability to anti-forgetting while keeping its plasticity to learn newly-coming knowledge. Extensive experiments demonstrate that our SCD framework outperforms state-of-the-art competitors on the proposed MDL-VQA with different training orders.

This chapter is based on the following publication:

- **Lao, M.**, Pu, N., Liu, Y., Zhong, Z., Bakker E.M., Sebe, N., and Lew, M. S. "Multi-Domain Lifelong Visual Question Answering via Self-Critical Distillation." ACM International Conference on Multimedia, 2023.

7.1 Introduction

Visual Question Answering (VQA) [30, 48] aims at answering textual questions conditioned on given images, which requires intricate vision-language reasoning. With the flourishing developments of large-scale pre-trained models [39, 314, 315] and cross-modal learning techniques [73, 74, 191], current VQA models have achieved state-of-the-art performance over various datasets [30, 48, 49, 57]. Despite the tremendous success, their training process always learns through a stationary domain that is fixed by the choice of a given dataset. However, this limitation violates many practical scenarios where the data is continuously increasing from different domains. In real-world applications, the VQA systems are always expected to constantly acquire and update their knowledge, thereby catering to the evolving demands from users.

To empower AI machines with the capacity of acquiring new knowledge from sequentially arriving tasks with less forgetting[316] of previously learned tasks, lifelong learning [119, 317, 318] has gained extensive research interests, and inspired considerable delicate and efficient approaches [319, 320, 321, 322, 323] in both CV[157] and NLP[324] communities. However, accomplishing lifelong learning in vision-language tasks is still challenging, especially in the fields of multimodal understanding and reasoning [84, 325]. In terms of VQA task, the work in [137] is the first attempt to explore simple class-incremental learning in the diagnostic dataset. Likewise, the method by [138] introduces a function- and scene-incremental settings on the realistic GQA dataset [49], and reduces the forgetting problem by replaying scene graphs. However, in contrast to these settings that focus on inner-domain incremental VQA within a single dataset, we note that the domain-incremental setting is more practical yet under-explored in VQA tasks, as different sequential tasks are typically composed of samples represented by different visual/textual domains, and heterogeneous label spaces.

To explore the setting, we propose a novel yet practical VQA task, namely Multi-Domain Lifelong VQA (MDL-VQA). This task requires VQA models to accumulate informative knowledge from sequentially-arrived domains, while alleviating forgetting the knowledge learned from previous domains. The challenges of MDL-VQA are mainly three-fold. 1) Severe Domain Shift: as depicted in Fig. 7.1(a) and Fig. 8.2, MDL-VQA embraces five datasets with vastly different domains in visual inputs, accompanied with non-negligible domain shift in textual representations. 2) Label-Space Variations: the label spaces (i.e., answer candidates) in different domains are inconsistent. Some general answers (e.g. *yes*, *one* and *red*) typically coexists in several or all domains. Meanwhile, a certain number of answers are only involved in one specific domain. 3) Data Privacy: we highlight the data privacy issue in MDL-VQA, because the training data might be privacy-protected in some domains. Thus, the training process can use only current domain data, without storing and replaying any instances from previous domains.

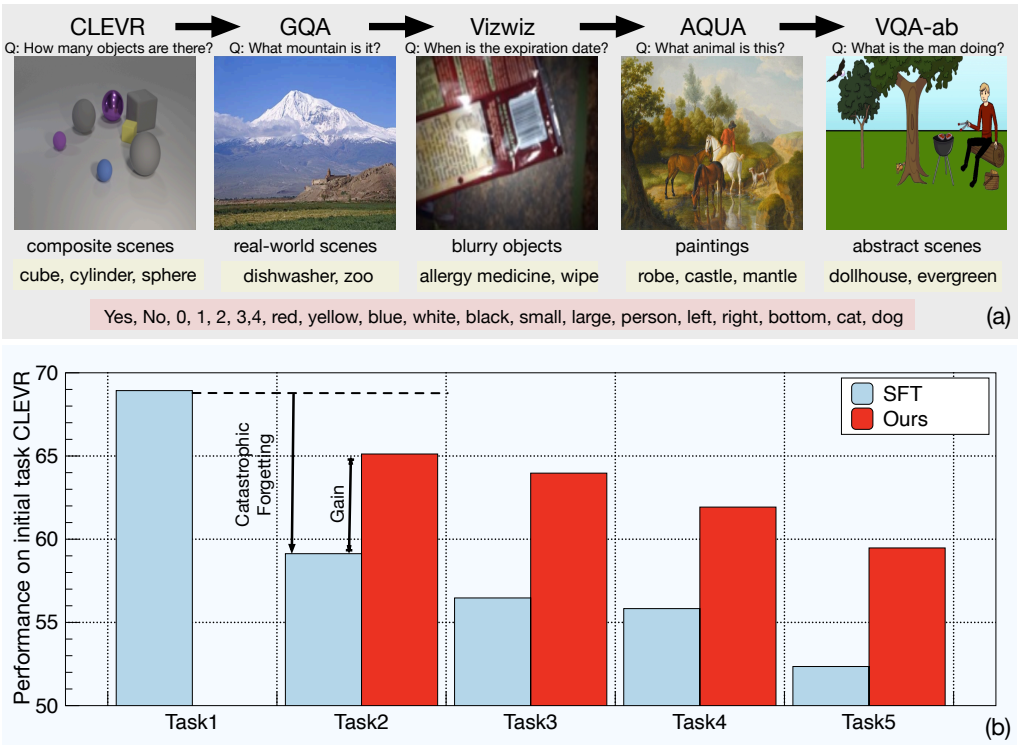

Figure 7.1: (a): Our MDL-VQA benchmark involves five tasks in different domains: CLEVR [57], GQA [49], Vizwiz [326], VQA-ab [48], AQUA [202]. Unlike standard domain-incremental learning, the the label spaces for different domains are inconsistent, where words in red shading denote some general answers coexist in several domains, and in yellow shading are domain specific. (b) The performance of the initial task (CLEVR) during the sequential training, where VQA model encounters the problem of catastrophic forgetting (SFT). In contrast, our approach remarkably alleviate the forgetting.

To address these challenges in the MDL-VQA task, we propose a novel Self-Critical Distillation (SCD) to overcome the forgetting issue without data storage. SCD is built on the teacher-student framework, and jointly introspects teacher and student based on their understanding with respect to different instances, so as to self-critically adjust the transfer of old knowledge and the acquirement of new knowledge. Specifically, SDC is implemented on both logits-level (SCDL) and feature-level (SCDF), by addressing the following two self-reflection questions for both self-critical teacher and student.

In SCDL, the frozen teacher needs to consider the question "*what is the informative old knowledge which is expected to deliver from the teacher to the student?*". To tackle it, we introspect the discrimination ability of teacher model over counterfactual samples, and then create instance- and domain-relevant knowledge for adaptive knowledge transfer. In SCDF, the student should introspect about "*What is the useless knowledge in new domain and which should be neglected when reviewing the old knowledge from the teacher?*". To achieve this, we propose to model the irrelevant knowledge by introspecting the student's reasoning behavior, and exploit

metric learning to prevent the student from acquiring useless yet domain-specific knowledge on current task. Fig. 7.1(b) demonstrates the capacity of our SCD strategy to mitigate forgetting after incrementally learning across five domains. Overall, our contributions are summarized as:

- We explore a new yet practical VQA task, namely MDL-VQA, which considers VQA problem under a multi-domain lifelong learning scenario. Correspondingly, we propose a benchmark to evaluate the model's lifelong learning ability.
- We propose a novel data-free SCD approach on both sides of logits- and features-level distillation, so as to not only transfer informative previous-domains knowledge, but also accumulate useful knowledge in currently-learned domain.
- Extensive experimental results demonstrate that our SCD framework outperforms other competitors and achieves a promising performance on our MDL-VQA benchmark.

7.2 Related Works

7.2.1 Multi-Domain Learning in Visual Question Answering

In recent years, increasing amount of datasets [30, 48, 57, 202, 326] with diverse visual and textual domains have been proposed to facilitate VQA research. Therefore, a longstanding research topic, multi-domain learning [327] has become an attractive yet practical topic in VQA community, where most of related works are focused on the model robustness against domain shift. [328] reveals that most methods normally perform poorly on either natural or composite dataset, and propose a conceptually simple RAMEN model for adapting to complex reasoning required in two domains. [329, 330] design delicate feature-learning strategies to enhance domain adaptation across different datasets. [164] analyzes domain shifts between nine widely-used VQA datasets and improve domain robustness via an unsupervised method. Apart from domain adaption based on visual information, the generalization of VQA models on different linguistic domains is also crucial, especially due to models' brittleness to the language variations [193].

In contrast to existing works adapting source knowledge to a target domain in general, our work equips multi-domain learning into the lifelong learning setting with sequentially arriving domains, and emphasises on retaining old domains performance while adapting to any upcoming domain.

7.2.2 Lifelong Learning in Vision-Language Tasks

Lifelong learning a.k.a. continual learning [119, 317, 318], has been extensively explored in CV and NLP tasks, where the mainstream research settings could be

divided into 1) class- or task-incremental learning, in which models are required to learn to classify a growing number or group of classes sequentially from a single domain in general, and 2) domain-incremental learning, where a model continually learns to solve tasks typically crossing different domains, whereas sharing the same label space. Inspired by the significant progress in vision-language learning, several works explore the lifelong learning in the perceptual-level multimodal tasks, such as cross-modal retrieval [331] and image captioning [135, 332]. For the VQA task requiring both perceptual- and reasoning-level understanding, [137] is the first attempt to exploit a simple class-incremental learning setting for lifelong VQA, where samples in question types '*wh-*' and '*yes/no*' are tested under different sequence. [138] proposes a CLOVE benchmark to establish the scene- and function-incremental learning through splitting the GQA dataset in natural visual domain, and mitigate the forgetting problem by replaying scene graphs. Moreover, [139] introduces an attractive CLiMB benchmark, where models are expected to continually learn crossing different multi-modal reasoning tasks, including VQA.

Compared with these VQA lifelong learning benchmarks, our proposed MDL-VQA is to overcome forgetting under continual tasks with multi-domain representations. However, it may not be considered as a standard domain-incremental learning in computer vision tasks, since the label spaces are not consistent across different domains. It is also not suitable to classify MDL-VQA as the class- or task-incremental learning, since some general or high-frequency answers would be involved in the whole training process.

7.2.3 Knowledge Distillation for Overcoming Forgetting

The common strategies [317] to alleviate catastrophic forgetting in lifelong learning are three-fold: 1) Rehearsal methods explicitly retrain on a limited subset of stored samples while training on new tasks. 2) Parameter isolation methods typically assign new branches with different model parameters for new tasks, while freezing previous task parameters. 3) Regularization-based methods tend to conduct extra regularization incorporated in the loss function, thereby solidifying previous knowledge when learning on new data. For lifelong VQA, due to the potential problem derived from data privacy and constrained computation resource, regularization-based approaches would be more valuable and practical, among which the data-focused Knowledge Distillation (KD) [333] has drawn widespread research interest. The technical route of KD, is to transfer learned knowledge from a frozen teacher model to a to-be-trained student model when new data are used only, which is re-introduced by LwF [334] in lifelong image classification. Apart from standard classifier-based KD characterizing the differences between the teacher and the student through metrics such Kullback-Leibler (KL) divergence, increasing number of advanced KD methods [335, 336, 337] have been presented to overcome forgetting issues in various lifelong learning tasks.

Although directly applying these methods on MDL-VQA can mitigate forgetting problem to some extent, such a way neglects the nature of cross-modality reasoning of VQA tasks. To this end, we first analyse the properties of reliance on shortcut learning and the reasoning behaviors implied among pair-wise instance interactions of attention modules, and then propose a novel Self-Critical Distillation (SCD) tailor-made for MDL-VQA. The proposed SCD framework leverages the comprehensive analysis results to selectively transfer knowledge while depressing the negative impact of the irrelevant learned knowledge for learning on current domain.

7.3 Multi-Domain Lifelong Visual Question Answering

7.3.1 Problem Definition

In the MDL-VQA task, a unified VQA architecture is required to learn T domains in an incremental fashion. Suppose we have a series of datasets $\mathcal{D} = \left\{D^{(t)}\right\}_{t=1}^{T}$. The data in the t-th domain is comprised of train and test split, which is represented as $D^{(t)} = \left\{D_{tr}^{(t)}, D_{te}^{(t)}\right\}$. Note that, only $D_{tr}^{(t)}$ is available at the t-th training step, and the data from previous domains are not available any more. Specifically, the dataset $D^{(t)} = \{(\mathbf{v}_i, \mathbf{q}_i, \mathbf{a}_i)\}_{i=1}^{\left|D^{(t)}\right|}$ contains $\left|D^{(t)}\right|$ triplets and each triplet consists of an image $\mathbf{v} \in \mathcal{V}^{(t)}$, a question in natural language $\mathbf{q} \in \mathcal{Q}^{(t)}$ and the ground-truth answer $\mathbf{a} \in \mathcal{A}^{(t)}$. The multi-domain setting is implemented by $\mathcal{V}^{(i)} \neq \mathcal{V}^{(j)}$, $\mathcal{Q}^{(i)} \neq \mathcal{Q}^{(j)}$ and $\mathcal{A}^{(i)} \neq \mathcal{A}^{(j)}$, where $\forall i, j \in \{1, .., T\}$ and $j \neq k$. It is noteworthy that, although $\mathcal{A}^{(i)}$ and $\mathcal{A}^{(j)}$ are different, they may share few common answer candidates (e.g., "Yes", "No" and the numbers shown in Fig. 7.1).

7.3.2 Analyses of Multi-Domain Lifelong VQA Benchmark

In this paper, we reorganize five popular VQA datasets to build a new multi-domain lifelong VQA benchmark, in which each dataset servers as a domain with domain-specific visual scenes. As illustrated in Fig 7.1, these domains include real-world scenes (GQA [49]), abstract scenes (VQA-ab [48]), synthetic scenes (CLEVR [57]), paintings (AQUA [202]) and blur-object scenes (Vizwiz [326]). Similar to conventional multi-modal or cross-domain learning, the MDL-VQA model surfers from inevitable domain shifts, which increases the difficulties of MDL-VQA tasks due to additional domain gaps.

To explicitly investigate and expose this problem, we propose to measure the visual and textual correlations among the five domains via Maximum Mean Discrepancy

(MMD). Formally, the MMD between $\mathcal{D}^{(i)}$ and $\mathcal{D}^{(j)}$ is given as:

$$\begin{aligned}\mathrm{MMD}\left(\mathcal{D}^{(i)}, \mathcal{D}^{(j)}\right) &= \left\|\mathbb{E}_{X\sim\mathcal{D}^{(i)}}[\varphi(X)] - \mathbb{E}_{Y\sim\mathcal{D}^{(j)}}[\varphi(Y)]\right\|_{\mathcal{H}} \\ &= \frac{1}{\left|D^{(i)}\right|^2}\sum_{i=1}^{\left|D^{(i)}\right|}\sum_{j=1}^{\left|D^{(i)}\right|} k\left(\mathbf{x}_i, \mathbf{x}_j\right) + \frac{1}{\left|D^{(j)}\right|^2}\sum_{i=1}^{\left|D^{(j)}\right|}\sum_{j=1}^{\left|D^{(j)}\right|} k\left(\mathbf{y}_i, \mathbf{y}_j\right) \\ &- \frac{2}{\left|D^{(i)}\right|\left|D^{(j)}\right|}\sum_{i=1}^{\left|D^{(i)}\right|}\sum_{j=1}^{\left|D^{(j)}\right|} k\left(\mathbf{x}_i, \mathbf{y}_j\right),\end{aligned} \tag{7.1}$$

where k denotes the RBF kernel. Specifically, we randomly select 5000 VQA samples in each dataset, and attempt to acquire multimodal representation. For visual features, we exploit pre-trained ResNet [58] to extract visual inputs, and obtain a 2048-D high-level representation for each image. For textual representation, because current VQA models are prone to be brittle to linguistic variations [193], we follow the work [338] to extract 20 low-level features: question length, prepositions, number of conjunctions, pronouns, etc.

By analysing the pairwise MMD comparisons among the five datasets in Fig. 7.2, we find that the domain shifts are severe among every pair of datasets. Especially, the question domain gap between CLEVR and other four datasets is largely remarkable, as CLEVR involves more complex linguistic expressions to test the reasoning abilities of VQA models. Thus, it is important to prevent the model from learning only domain-specific knowledge across different domains.

	CLEVR	VQA-ab	AUQA	Vizwiz	GQA
CLEVR	-	0.75	0.45	0.77	0.51
VQA-ab	0.63	-	0.67	0.24	0.37
AUQA	0.71	0.36	-	0.73	0.34
Vizwiz	0.53	0.42	0.51	-	0.39
GQA	0.51	0.41	0.57	0.52	-

Figure 7.2: Visual and Textual domain gaps, where green shading is MMD over 20-D syntax statistics, and the blue is MMD over ResNet-101 2048-D features.

7.3.3 Baseline Approach

Considering the data privacy issues in MDL-VQA, we introduce replay-free regularization-based lifelong learning approaches as our baseline approaches, including logits-based, feature-based and correlation-based knowledge distillation. These methods often combine two learning objectives for model training. One is to acquire knowledge in the current domain, and the other aims at maintaining the old knowledge learned from previous domains. Formally, at the t-th domain, the overall objective is given by:

$$\mathcal{L}(t) = \mathcal{L}_{new} + \lambda\mathcal{L}_{old}, \tag{7.2}$$

where λ controls the contribution of the $\mathcal{L}_{old}$ that preserves the old knowledge, and the $\mathcal{L}_{new}$ allows the model to acquiring knowledge in new domains.

Based on the formulation, $\mathcal{L}_{new}$ is implemented by a standard cross-entropy loss in our MLD-VQA tasks. Specifically, we define a classification-based VQA model as $f(\cdot;\theta,\phi)$, comprised of a multimodal fusion encoder $m(\cdot;\theta)$ with parameters θ and a classifier $c(\cdot;\phi)$ parameterized by ϕ. Given a newly-coming domain $D^{(t)}$ at the t-th training step, we minimize the standard cross-entropy loss to acquire knowledge in the current domain, which is formulated as:

$$\mathcal{L}_{ce} = -\sum_{(\mathbf{v},\mathbf{q},\mathbf{a})\in D^{(t)}} \log(\boldsymbol{\sigma}(f(\mathbf{v},\mathbf{q};\theta,\phi^{(t)}))[\mathbf{a}], \tag{7.3}$$

where $\sigma(\cdot)$ refers to the $softmax$ function, and $\phi^{(t)}$ is the classifier specialized for the domain $D^{(t)}$.

On the other hand, in distillation-based lifelong learning approaches, the $\mathcal{L}_{old}$ considers the types of knowledge which is efficient for transferring knowledge.

For logits-based knowledge distillation [334], by feeding a training sample into the to-be-trained student model $f\left(\mathbf{v},\mathbf{q};\theta,\phi^{(k)}\right)$, its output logits can be represented by $\mathbf{z}^{(k)} = \left[z_1, z_2, \ldots, z_{|\mathcal{A}^{(k)}|}\right] \in \mathbb{R}^{1\times|\mathcal{A}^{(k)}|}$,where z_i is the logit of the i-th class and $|\mathcal{A}^k|$ refers to the number of classes in the label space of the k-th task ($1 \leq k < t$). Then, we can use $softmax$ function to compute its classification probabilities $\mathbf{p}^{(k)}$ by

$$p_i = \frac{\exp\left(z_i\right)}{\sum_{j=1}^{|\mathcal{A}^{(k)}|}\exp\left(z_j\right)}, \tag{7.4}$$

where p_i represents the probability of the i-th class in $\mathbf{p}^{(k)}$. Meanwhile, by analogy, we can obtain the probabilities $\hat{\mathbf{p}}^{(k)}$ through feeding the same training instance into the teacher model $f\left(\mathbf{v},\mathbf{q};\hat{\theta},\hat{\phi}^{(k)}\right)$. Concretely, $\hat{\theta}$ and $\hat{\phi}^{(k)}$ for the k-th learned task are copied from θ as well as ϕ before current-step training, respectively. Finally, we adopt the common Kullback-Leibler Divergence [339] constraint as distillation loss between teacher and student models:

$$\mathcal{L}_{lkd}(\hat{\mathbf{p}}^{(k)},\mathbf{p}^{(k)}) = \mathrm{KL}\left(\hat{\mathbf{p}}^{(k)}\|\mathbf{p}^{(k)}\right), \tag{7.5}$$

where $\mathcal{L}_{lkd}(\hat{\mathbf{p}}^{(k)},\mathbf{p}^{(k)})$ denotes the loss function of logits-based knowledge distillation for the k-th previous task, based on the given input sample. Practically, we can consider previous domain of $k = t-1$, to avolid the linearly-increased usage of computation sources in long-sequence lifelong learning.

For feature-based knowledge distillation [340], we can intuitively utilize the output feature $\hat{\boldsymbol{f}} \in \mathbb{R}^{1\times M}$ extracted from the multimodal fusion encoder $m(v,q;\hat{\theta})$ in frozen teacher as the knowledge learned from previous tasks. M refers to the dimension of

the intermediate feature. Then, feature-based knowledge distillation employs Mean Square Error [341] (MSE) to distill the knowledge from teacher into student:

$$\mathcal{L}_{fkd}\left(\hat{\boldsymbol{f}}, \boldsymbol{f}\right) = \left\|\hat{\boldsymbol{f}} - \boldsymbol{f}\right\|_2^2, \tag{7.6}$$

where $\boldsymbol{f}$ is the corresponding feature obtained from the currently-trained student model $m(v, q; \theta)$.

Correlation-based knowledge distillation [336] focuses on transferring the knowledge about semantic correlation of features in a training batch. Based on L2-normalized outer products [342], we can obtain the pairwise similarities $\hat{G}$ and G of the mini-batch features yielded from teacher and student, respectively. Sequentially, the correlation-based knowledge distillation loss for a given training batch is defined as :

$$\mathcal{L}_{ckd}\left(\hat{G}, G\right) = \frac{1}{b^2}\sum\left\|\hat{G} - G\right\|_2^2, \tag{7.7}$$

where b implies the batch size.

We adapt the above-mentioned methods into our MDL-VQA benchmark and evaluate their effects in Tab 7.2. We find that most of these methods achieve unsatisfactory performances, since they overlook the inherent reasoning mechanism of the VQA task. We analyse and discuss the drawbacks below.

7.3.4 Limitations

As for a practical and flexible regularization-based approach, knowledge distillation strategies could be easily deployed into any sequentially-leaning process to handle the forgetting problem. However, unlike other tasks in the incremental fashion, we suggest that VQA models may encountered two important challenges in the process of knowledge transferring: (1) For logits-based distillation depicted in Equ. (7.5), the old knowledge from previous domains is obtained by feeding the training samples in current domain into the frozen teacher model. However, due to the over-reliance of language shortcut learning [85] in VQA model, when the old model meets the new data with visual domain shift, the teacher is prone to establish the old knowledge only relying on the language questions from current-domain samples. In this case, the question-dominated old knowledge is typically irrelevant to overcome the forgetting ratio in previous domain, as it may lose some useful semantic information of current input. Moreover, the negative effect of question-dominated knowledge would be more serious in our MDL-VQA, because VQA model can easily capture the correlations between question types and general answers co-existed in different datasets. (2) For student network learning knowledge from a new domain, it is inevitable to acquire the domain-specific knowledge (e.g. visual appeal and linguistic

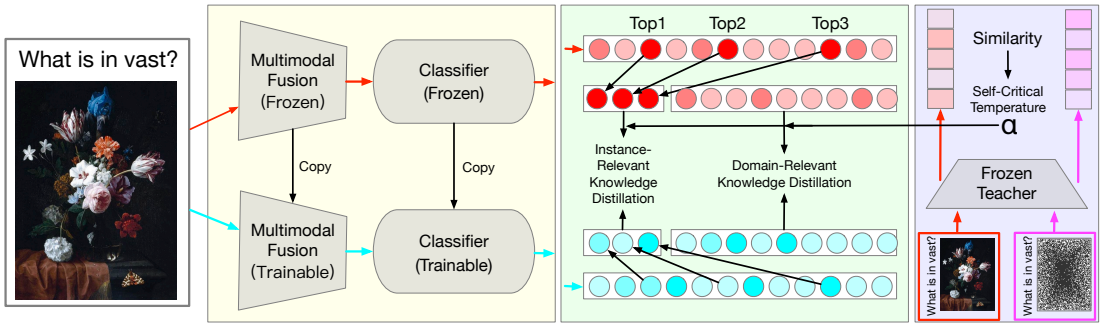

Figure 7.3: Illustration of the proposed Logits-level SCD, where the training samples in current domain is fed into both frozen teacher and to-be-trained student networks (yellow region). In the green region, the long vectors in red and blue refer to the original predictions yielded from teacher and student respectively, which are separated into instance- and domain-relevant knowledge based on high-response classes in teacher. The region in purple depicts the process of obtaining self-critical temperatures α through introspecting the teacher about original (red) and counterfactual (pink) samples.

styles), which is not only pointless to understanding visual concepts for question answering, but also accelerate the process of forgetting previous knowledge.

7.4 Self-Critical Distillation

In this section, we attempt to break the aforementioned limitations, and propose a Self-Critical Distillation (SCD) to improve the anti-forgetting efficiency from dual-level knowledge transferring.

7.4.1 Logits-level SCD

Logits-level SCD (SCDL) seeks to introspect the reasoning process of teacher model and transfer informative knowledge to student, thereby alleviating the first limitation described in Section 7.3.4. Specifically, SCDL first introspects the discrimination capacity of frozen teacher between counterfactual training sample and its original counterpart, to decomposes the logits knowledge into instance-relevant knowledge (IRK) and domain-relevant knowledge (DRK). Then, the teacher separately transfers the two types of knowledge to the student with adaptive temperature generated by the introspection.

Intuitively, IRK more likely refers to the high-response classes in the answer prediction, which involves the information about potential correct answers to each training samples. On the other hand, the classes with lower predicting probabilities can be regarded as DRK, including the semantic relationships of different answer candidates. Thus, we decompose the original answer prediction $\hat{\mathbf{p}}^{(k)}(\tau)$ from the frozen teacher model into aforementioned two types of knowledge, based on their responses over different answer candidates illustrated in Fig. 7.3. Specifically, we denote the

IRK as $\hat{\mathbf{p}}_I^{(k)}(\tau) = [p_a, \ldots, p_b, p_{\setminus[a,\ldots,b]}]$, where $\forall p_i \in [p_a, \ldots, p_b]$ is the possibilities of Top-C high-response classes. $p_{\setminus[a,\ldots,b]}$ refers to the summations of low-responded probabilities:

$$p_{\setminus[a,\ldots,b]} = \frac{\sum_{k=1,k\notin[a,\ldots,b]}^{|\mathcal{A}^{(k)}|} \exp\left(z_k/\tau\right)}{\sum_{j=1}^{|\mathcal{A}^{(k)}|} \exp\left(z_j/\tau\right)}. \tag{7.8}$$

Then, we define the DRK as the $\hat{\mathbf{p}}_D^{(k)}(\tau) \in \mathbb{R}^{1\times(|\mathcal{A}^{(k)}|-C)}$, where C is the number of classes with high-responded probabilities in classes $[a, \ldots, b]$. Concretely, we compute the probabilities in $\hat{\mathbf{p}}_D^{(k)}(\tau)$ by taking only the low-responded classes into account. The i-th element q_i of $\hat{\mathbf{p}}_D^{(k)}(\tau)$ can be formulated as:

$$q_i = \frac{\exp\left(z_i/\tau\right)}{\sum_{j=1,j\notin[a,\ldots,b]}^{|\mathcal{A}^{(k)}|} \exp\left(z_j/\tau\right)}. \tag{7.9}$$

Based on the separated knowledge and Equ. (7.5), we derive the separated logits-based KD $\mathcal{L}_{slkd}$, which separately transfers IRK and DRK from teacher to student:

$$\mathcal{L}_{slkd} = \tau^2 \mathrm{KL}\left(\hat{\mathbf{p}}_I^{(k)}(\tau) \| \mathbf{p}_I^{(k)}(\tau)\right) + \tau^2 \mathrm{KL}\left(\hat{\mathbf{p}}_D^{(k)}(\tau) \| \mathbf{p}_D^{(k)}(\tau)\right). \tag{7.10}$$

Furthermore, we propose a self-critical temperature to adaptivaly adjust the knowledge transfer of IRK and DRK. To obtain the adaptive temperature, we seek to quantify the teacher's reliance on textual information to create the old knowledge in previous domain, which is achieved by introspecting teacher's understanding about discriminating the original and counterfactual samples. To be specific, in contrast to the multimodal feature yielded from the original VQA instance $\hat{\mathbf{f}}$ as $m(\mathbf{v}, \mathbf{q}; \hat{\theta})$, its counterfactual logits $\hat{\mathbf{b}}$ is computed by replacing the raw image input $\mathbf{v}$ into the zero-padding counterparts $\mathbf{o}$ as $m(\mathbf{o}, \mathbf{q}; \hat{\theta})$. Ultimately, by reformulating the Equ. (7.9) with knowledge-specific temperatures (α and β), the SCDL loss is:

$$\mathcal{L}_{scdl} = \alpha^2 \mathrm{KL}\left(\hat{\mathbf{p}}_I^{(k)}(\alpha) \| \mathbf{p}_I^{(k)}(\alpha)\right) + \beta^2 \mathrm{KL}\left(\hat{\mathbf{p}}_D^{(k)}(\beta) \| \mathbf{p}_D^{(k)}(\beta)\right), \tag{7.11}$$

$$\alpha = max(\frac{\hat{\mathbf{f}} \cdot \hat{\mathbf{b}}}{\|\hat{\mathbf{f}}\|\|\hat{\mathbf{b}}\|}, 0) \cdot \tau_{max}, \quad \beta = \tau_{max} - \alpha, \tag{7.12}$$

where the self-critical temperature α for IRK is determined by the maximum temperature setting τ_m, accompanied with the *cosine similarity* between teachers' features derived from original ($\hat{\mathbf{f}}$) and counterfactual ($\hat{\mathbf{b}}$) samples.

Discussion: Through the comparison of the output logits yielded from counterfactual and original samples, the teacher can introspect itself about whether it forms old knowledge by understanding both visual and textual information for input sample (lower cosine similarity). Otherwise, it may extract spurious class-related knowledge overwhelmingly from question input. If the old knowledge is dominated by question with higher value of α, the teacher would create the more smoothed IRK with relatively high temperature, while turning to establish more informative DRK, as the overused question information is typically involved more 'dark knowledge' about semantic correlations among different classes.

7.4.2 Feature-level SCD

Compared with SCDL that handles high-level semantic information, the intermediate feature typically covers knowledge across a wide range of semantic levels, from superficial visual/linguistic styles to the question-related visual concepts. However, when the student learns from the data in newly-arrived domain, it unavoidably extracts harmful domain-specific knowledge (DSK), such as the low-level information irrelevant for question answering. This behavior plays a negative role on maintaining crucial knowledge for question answering in previous domains. To mitigate this issue, we present a feature-level SCD (SCDF), whose idea is first to model the negative DSK by introspecting students' reasoning behavior that is suggested by the instance interactions in attention modules, and propose a **metric learning** strategy to promote student to bypass the deleterious effect from such knowledge when reviewing the old knowledge from previous domain.

Given a training sample, we assume that it is sophisticated to directly recognize the DSK, such as visual and linguistic styles in current domain. Thus, we turn to model such useless knowledge via removing the indispensable visual/textual components from the original sample. To this end, we firstly identify crucial question words as well as image regions by introspecting the attention maps in student network, which reveals how the student network understands and reasons over different visual/textual components for answer prediction. In VQA task, the widely-used attention mechanism is Multi-Head Attention (MHA) approach [62, 343] equipped in state-of-the-art transformer-based VQA models [73, 74, 344]. Hence, we average attention maps existing the final layer of MHA as the attention weights for different components in each image-question training pair. From the statistics of attention weights (see Fig. 7.4), among more than 200 visual and textual components in a VQA instance, merely several visual/textual components are crucial to deduce the correct answer.

Based on the observation, we propose to intervene the original VQA sample by removing its components (e.g. question words and visual regions) with Top-K attention weights, where the corrupted image-question pair is denoted as $(\dot{\boldsymbol{v}}, \dot{\boldsymbol{q}}$. Then, we feed $\dot{v}$ and $\dot{q}$ into the currently-trained student model to obtain the $\dot{\boldsymbol{f}} = m(\dot{v}, \dot{q}; \theta)$.

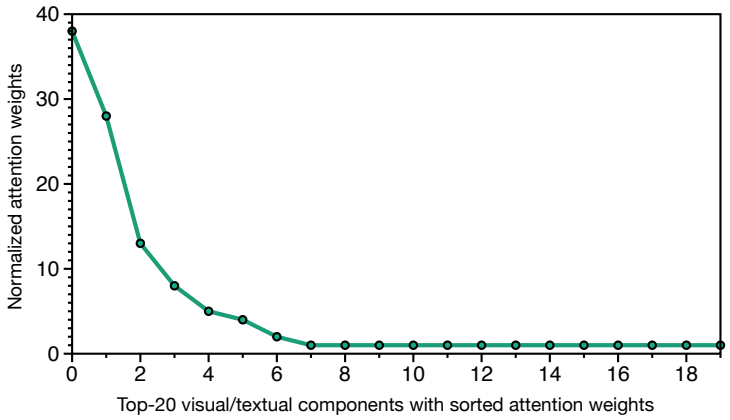

Figure 7.4: The Distribution of sorted attention weights based on 10,000 VQA samples from five datasets, which is generated by ViLT model [344].

We assume that the $\dot{\boldsymbol{f}}$ is pointless for VQA task, since it has lost reasoning cues for question answering. Meanwhile, the rest part still maintains the DSK. In order to prevent the student from over-exploiting useless information in current domain when reviewing the old-domain knowledge, we utilize metric learning to implement our SCDF. Specifically, the anchor/positive feature in metric learning is the intermediate feature yielded from the student/teacher model ($\boldsymbol{f}/\hat{\boldsymbol{f}}$), whereas the negative is the corrupted feature $\dot{\boldsymbol{f}}$ from self-criticism of reasoning behaviour. The SCDF loss is given by:

$$\mathcal{L}_{scdf}\left(\hat{\boldsymbol{f}}, \boldsymbol{f}\right) = \max\left(\|\hat{\boldsymbol{f}} - \boldsymbol{f}\|^2 - \|\dot{\boldsymbol{f}} - \boldsymbol{f}\|^2, 0\right). \tag{7.13}$$

Meanwhile, we also propose to substitute the correlation-based KD (Equ. (7.7)) by metric learning, which is formulated by:

$$\mathcal{L}_{scdc}\left(\hat{G}, G\right) = \max\left(\|G - \hat{G}\|^2 - \|G - \dot{G}\|^2, 0\right), \tag{7.14}$$

where $\dot{G}$ is the similarities of corrupted features within a mini-batch.

Discussion: From the conceptual example in Fig. 7.5, through the self-criticism from reasoning behaviour of attention map, the corrupted current-domain samples in purpose usually represent the uninformative (e.g. no question intention with related visual cues) but domain-specific (e.g. keeping the cartoon style) information. The output feature in red from teacher typically involves the associated knowledge and scenarios from previous domain, even though the input picture is described in abstract domain. The metric learning in feature-level SCD aims at narrowing the semantic distance between samples with consistent domain-invariant concepts, and meanwhile weakening the negative impact from domain-specific biases.

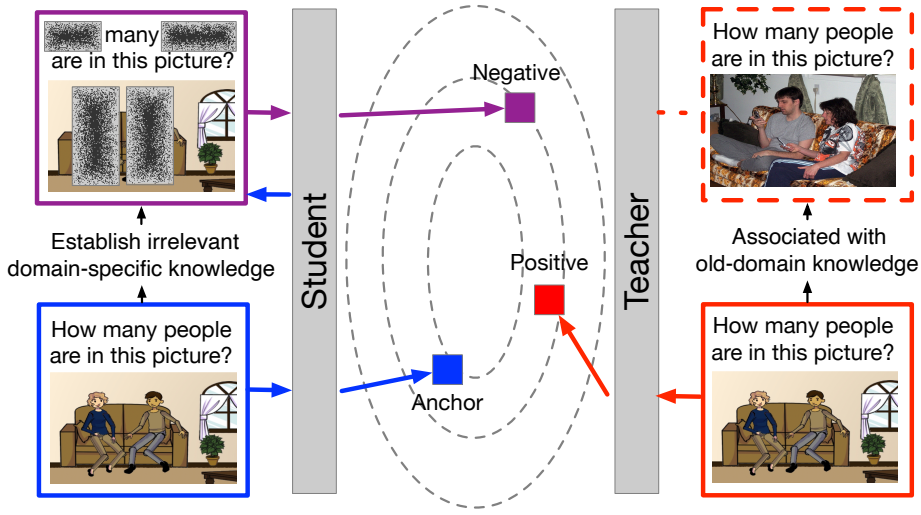

Figure 7.5: Conceptual illustration of Feature-level SCD, where we suppose the currently-trained samples are in abstract domain, and the training data utilized in the previous task is in realistic domain. The samples in red, and blue refer to the intermediate features extracted from teacher and student models, respectively.

7.4.3 Optimization

Ultimately, our proposed Self-Critical Distillation (SCD) can be achieved by the proposed dual-level distillation strategies. For the overall objective at the training step t ($t \geq 2$), we train the parameters of whole VQA model with classifiers in all involved tasks $\{\theta, \phi^{(1)}, \ldots, \phi^{(t)}\}$ on dataset $\mathcal{D}^t$. The overall loss function is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ce}} + \lambda_l \mathcal{L}_{scdl} + \lambda_f (\mathcal{L}_{scdf} + \mathcal{L}_{scdc}), \tag{7.15}$$

where $\mathcal{L}_{scdl}$ and $(\mathcal{L}_{scdf} + \mathcal{L}_{scdc})$ denote the loss terms of logits- and feature-level SCD defined in Equ. (7.10), (7.12) and (7.13), respectively. They enforce VQA model to remember the old knowledge from previous domains and avoid forgetting. λ_l and λ_f are the weighting factors to adjust the contributions between dual-level distillations.

7.5 Experiments

7.5.1 Implementation Details

Training Strategy. The whole training process includes two stages: initial model training and lifelong training. In the first stage, the initial model $f(\cdot; \theta, \phi^{(1)})$ (multi-modal fusion encoder $m(\cdot; \theta)$ and the classifier $c(\cdot; \phi^{(1)})$ for the first task) is trained to converge on the classes in the label space of the first dataset $D^{(1)}$, using cross entropy loss only. It is optimized by the AdamW optimizer [345] with a learning rate of 10^{-4} and weight decay of 10^{-2}. The total number of training epoch across all datasets is set to 10. We warm up the learning rate in the first epoch, and linearly

decay it to zero in the remaining of training epochs. In the second stage of lifelong learning, we employ the loss function depicted in Equ. (7.18) to train the VQA model with new classifiers, where the settings including learning rate and training epochs are the same as those in the first training stage.

Network Architecture. We use a pre-trained Vision-Language Transformer (ViLT) [344] as the backbone multimodal fusion encoder. Unlike other pre-trained vision-language models [346, 347] that build upon region-level features extracted from Faster R-CNN [246], ViLT directly operates on image patches without using any convolutional layers, which is suitable for image representation across diverse domains in our MDL-VQA benchmark. For the classifier for each task, it is comprised of double-layers of MLP with LayerNorm [348] and ReLU activation function. For the hyper-parameter setting of our SCD approach, we select the threshold $\delta = 0.15$, trade-off factors $\lambda_l = 1$ and $\lambda_l = 0.5$, which are validated in Tab. 7.5. The number of visual/textual components to be removed in our feature-level SCD is set to 10, which is based on the distribution illustrated in Fig. 7.4.

Table 7.1: The statistics of datasets in the MDL-VQA. '*' denotes the modification of random sampling from the raw datasets.

	Train	Test	Label	Frequent Answers
GQA*	93786	12946	1657	no, yes, left, right, man, white
CLEVR*	69852	10000	28	no, yes, 1, 0, small, rubber, metal
VQA-AB	59074	29476	426	yes, no , 2, 1, red, 3, white, blue
AQUA	29568	1508	453	person, people, building, church
Vizwiz	20524	4320	3648	unanswerable, unsuitable, no, yes

7.5.2 Evaluation Metrics

For each dataset involved in MDL-VQA, we determine the ground-truth answer for each sample via the soft voting of ten annotated answers, following by the same rule in VQA-v2 dataset [48]. To quantitatively validate the efficiency of related strategies to alleviate forgetting problem, we exploit the Average Accuracy [335] and Average Forgetting [349] as evaluation metric in our MDL-VQA.

Average Accuracy: Suppose that, after sequential learning across t domains, $acc_t^{(i)}$ denotes the model accuracy obtained from the test set $D_{te}^{(i)}$, whose related train split $D_t^{(i)}$ was learned in the i-th stage ($i \leq t$). The average accuracy $\overline{s}_t$ at the t-th stage is defined as $\overline{s}_t = \frac{1}{t}\sum_{i=1}^{t} acc_t^{(i)}$.

Average Forgetting: is to quantify the forgetting ratio $\overline{f}_t$ after learning the t-th domain ($t \geq 2$). Specifically, the ratio for a particular task (e.g. dataset i) is determined by the difference between the maximum accuracy $acc_{max}^{(i)}$ gained throughout the lifelong training process in the past, and the accuracy of the currently-trained

model. Then, the forgetting ratios $\overline{f}_t$ for all previous $t-1$ domains is defined as:

$$\overline{f}_t = \frac{1}{t-1}\sum_{i=1}^{t-1}\left(\max_{l\in\{1,\ldots,t-1\}} acc_{max}^{(i)} - acc_t^{(i)}\right), \forall i < t. \tag{7.16}$$

7.5.3 Datasets

In our proposed MDL-VQA benchmark, we exploit five VQA datasets where the images are represented in various visual domains, including artistic, abstract, real-world, sythetic and blurred-objects scenes. To be specific, *AQUA* dataset [202] aims to ask questions about artworks, where the artistic images are obtained from SemArt [350] dataset. *VQA-abstract* [48] contains the images of abstract/cartoon scenes. *GQA* [49] is a large-scale dataset to test multiple reasoning skills through compositional questions, where images are described in high-quality real-world scenes. *Vizwiz* [326] is proposed to help visually-impaired people, which is involved the images about blurred objects. It focuses on validating VQA models about the perceptual understanding of visual objects. In contrast, *CLEVR* [57] is a diagnostic dataset with sythetic images, which emphasises on the model capacities of spatial and logical reasoning. The numbers of train/test VQA samples and the labels, accompanied with frequent answer candidates for five datasets are depicted in Tab. 7.1. It is noteworthy that, the amount of train/test samples in original GQA and CLEVR datasets are considerably larger than those in other datasets (e.g. the number of training data in raw GQA is 45 times that of Vizwiz). As a result, we randomly select 10% of samples in these two datasets in our benchmark. From Tab. 7.1, the sampling operation alleviates the original extreme imbalance of VQA samples for multi-domain learning, but also maintains the diversity of data volumes among different datasets.

7.5.4 Performance Evaluation

The capacity of approaches to alleviate forgetting, is mainly reflected in its effectiveness of mitigating the accuracy degradation on previous domains. In this section, we validate our SCD on the MDL-VQA benchmarks against prevalent yet competitive strategies. The fundamental solution (FST) is to fine-tune model with newly-arrived datasets without reviewing old knowledge. EWC [351] and ALASSO [352] refer to the prior-focused regularization methods, which focuses on penalizing network parameters in sequential training. FKD [340], SPD [342], LWF [334] and IRG [353] are data-focused knowledge distillation approaches firstly deployed in the scenarios of model compression, among which LWF, FKD, SPD [342] denote our baseline strategies described in Equ. (7.5), (7.6) and (7.7). ECD [335], DKD [336] and MBP [337] are the advanced knowledge distillation strategies tailored to class-incremental lifelong learning. Last but not least, we define the Reference method served as the

Table 7.2: Non-forgetting evaluation in MDL-VQA benchmark. We test model after sequentially training on all datasets in five domains (t=5) started from the synthetic (CLEVER) dataset. **Best** numbers are in bold, without considering the reference approach.

Method	CLEVR→GQA→Vizwiz→AQUA→VQA-ab					$\overline{s}$	$\overline{f}$
SFT	52.35	47.98	36.57	76.88	75.01	57.76	10.46
EWC [351]	52.14	48.05	36.68	77.13	74.92	57.78	10.40
ALASSO [352]	52.43	48.28	36.24	77.37	75.19	57.90	10.32
FKD [340]	54.77	50.45	37.56	78.12	75.26	59.23	8.68
SPD [342]	53.97	49.16	37.88	77.89	**76.15**	59.01	9.19
LWF [334]	55.43	50.74	37.93	78.36	74.82	59.46	8.29
IRG [353]	56.32	51.02	37.14	77.97	75.43	59.58	8.29
ECD [335]	54.30	49.86	37.99	77.94	75.67	59.15	8.88
DKD [336]	55.54	51.15	37.83	78.25	**76.15**	59.79	8.21
MBP [337]	57.58	51.24	40.77	77.87	74.69	60.43	7.04
Ours	**59.48**	**52.47**	**43.41**	**79.44**	74.90	**61.94**	**5.30**
Reference	68.93	59.21	46.10	81.38	75.34	66.19	-

upper-bound, which directly trains the to-be-evaluated dataset on the pretrained ViLT [344] model, without sequential learning.

We conduct two training orders with different initial datasets (*i.g.,* synthetic CLEVER and real-world GQA). The results are summarized as follows. First, SFT, EWC and ALASSO encounters significant forgetting in our long-sequence training settings. In the comparisons of the three strategies, LWF achieves superior efficacy of reducing forgetting by transferring logits-based knowledge, whereas feature-correlation based SPD obtains better accuracy when learning the last task. Among the state-of-the-art KD specialized for lifelong learning, our proposed SCD occupies the first place on both metrics of average accuracy and average forgetting. It should be noted that, similar to our SCD, the competitive MBP method jointly considered logits-based and feature-based distillations, and further improves them by protecting model's ranking behaviors. SCD acts as a more effective strategy to eliminate forgetting, since our SCD enhances the effectiveness to acquire new and review old knowledge.

Furthermore, to validate the robustness of the aforementioned strategies against order variations, we propose to fix the initial tasks in two orders, and alter the original sequences among last four domains. We mainly compare SCD with the basic SFT and the typical KD approaches LWF (logits-based KD), FKD (feature-based distillation), SPD (correlation-based distillation), and MBP (Logits- and feature-level KD). Specifically, the process of accuracy degradation in the first task under different orders is illustrated in Fig. 7.6. From the accuracy obtained from the last-step training, we can notice that, when training a group of datasets with different orders, the degree of the forgotten old knowledge is typically different. In comparison, our SCD demonstrates stable improvements in terms of alleviating forgetting against different sequences, and outperforms the SFT by approximately 8% averaged from four depicted orders.

Table 7.3: Non-forgetting evaluation in MDL-VQA benchmark. We test model after sequentially training on all datasets in five domains (t=5) started from the real-world (GQA) dataset, respectively. **Best** numbers are in bold, without considering the reference approach.

Method	GQA→Vizwiz→AQUA→VQA-ab→CLEVR					$\overline{s}$	$\overline{f}$
SFT	42.20	34.06	67.24	62.24	68.25	54.80	14.07
EWC [351]	41.68	33.99	67.38	62.77	68.15	54.79	14.05
ALASSO [352]	42.41	34.39	67.19	62.89	68.44	55.07	13.78
FKD [340]	44.98	36.03	69.14	64.47	68.01	56.53	11.85
SPD [342]	42.52	35.11	69.04	63.99	68.91	55.91	12.84
LWF [334]	45.98	38.65	69.02	65.63	67.91	57.44	10.69
IRG [353]	46.18	39.90	68.76	65.33	68.50	57.73	10.46
ECD [335]	42.82	35.21	68.45	64.23	68.44	55.83	12.83
DKD [336]	46.20	37.74	68.84	65.82	**70.20**	57.76	10.86
MBP [337]	47.73	39.85	73.56	68.15	67.95	59.44	8.19
Ours	**50.11**	**40.98**	**74.98**	**69.52**	68.14	**60.74**	**6.61**
Reference	59.21	46.10	81.38	75.34	68.93	66.19	-

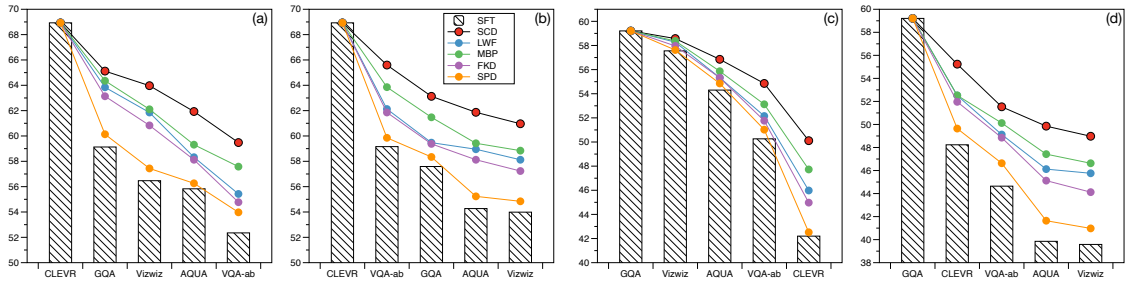

Figure 7.6: Non-forgetting evaluation against order variations. (a)/(b) and (c)/(d) illustrate the trend of accuracy computed from initial tasks (CLEVR and GQA) against different four-task sequences, where we use shadowed bar to represent the SFT, and lines with different colors for KD-based approaches.

(1) **Efficacy of Different Components**: We first analyze the effectiveness of different components in our SCD. Specifically, the experimental results are reported in Tab. 7.4, which are obtained from the first order involved in Tab. 7.2, as well as a two-task sequence (VQA-ab→AQUA). Based on fine-tuning, independently exploiting logits- and feauture-level SCD could effectively reduce the forgetting, where the logits-level SCD yields remarkable performance for reviewing old knowledge, and correlation-based SCD (case (d)) performs better on the plasticity when acquiring new knowledge. In case (e), through blending dual-level knowledge, our complete SCD cooperatively overcomes forgetting from the perspectives of label prediction and intermediate representation, thereby obtaining further improvements on both average accuracy and forgetting.

7.5.5 Ablation Study

(2) **Logits-level SCD vs Logits-based KD**: In this subsection, we make detailed comparisons between Logits-level SCD (SCDL) and the standard logits-based KD

Table 7.4: Ablation study under the setup of a five-task sequence order1 (CLEVR→GQA→Vizwiz→AQUA→VQA-ab) and a two-task sequence (VQA-ab→AQUA). We verify the improvements by progressively adding components in our SCD.

Configurations					Order1		VQA-ab→AQUA	
Case	$\mathcal{L}_{\text{ce}}$	$\mathcal{L}_{scdl}$	$\mathcal{L}_{scdf}$	$\mathcal{L}_{scdc}$	$\overline{s}$	$\overline{f}$	VQA-ab	AQUA
(a)	✓				57.76	10.46	65.34	80.31
(b)	✓	✓			60.45	6.96	70.86	79.74
(c)	✓		✓		59.18	8.53	68.19	80.44
(d)	✓			✓	58.43	9.61	66.14	80.77
(e)	✓		✓	✓	59.60	8.37	68.37	80.46
(e)	✓	✓	✓	✓	61.94	5.30	71.97	79.92

(LKD) (Equ. (7.5)) with different manually-defined temperatures ($T = 1, 2, 3$). We conduct the comparative experiments under double-task sequences (VQA-ab→AQUA), where the performance on the former and latter implies the plasticity and stability, respectively. Moreover, to validate the effectiveness of our self-critical temperature α in Equ. (7.11), we also take the SCD counterpart that creating the instance-aware and domain-aware knowledge with the same temperatures into the comparison. From the comparative results in Fig. 7.7, We can see that our approach is consistently superior to the standard KD under various setting of trade-off factor λ_l with different temperatures. Meanwhile, even though our method without self-critical temperature α slightly surpasses the standard KD, but still performs worse than the complete Logits-level SCD on both plasticity and stability on previous and current domains, respectively. It verifies that our knowledge-separated operation is beneficial to overcome forgetting in the MDL-VQA with label-space variations, and the self-critical temperature can further promotes the teacher to transfer more informative knowledge.

(3) **Feature-level SCD vs Feature-based KD**: We compare standard feature-based KD (FKD+CKD in Equ. (7.6) and (7.7)) with our feature-level SCD, and the counterparts that corrupting components randomly (RAND) in Fig. 7.8. We can notice that our self-criticism of model behaviour (attention) for sample corruption is indispensable, as the random-removing counterpart fails to attain any accuracy boost. On the contrary, benefiting from well-established DSK and metric learning, our method fulfils significant improvements for anti-forgetting on previous VQA-ab dataset, and meanwhile maintains the plasticity on the AQUA dataset when acquiring new knowledge.

Trade-off factors λ_l and λ_f: we first jointly discuss the trade-off factors λ_l and λ_f in the total loss function (Equ. (7.14)), which not only control the equilibrium with the cross-entropy function, but also dynamically adjust dual-level SCD to review different old knowledge. We dynamically adjust the value of λ_l and λ_f in the reasonable range of {0, 0.1, 0.25, 0.5, 1, 2}, respectively. The experiments

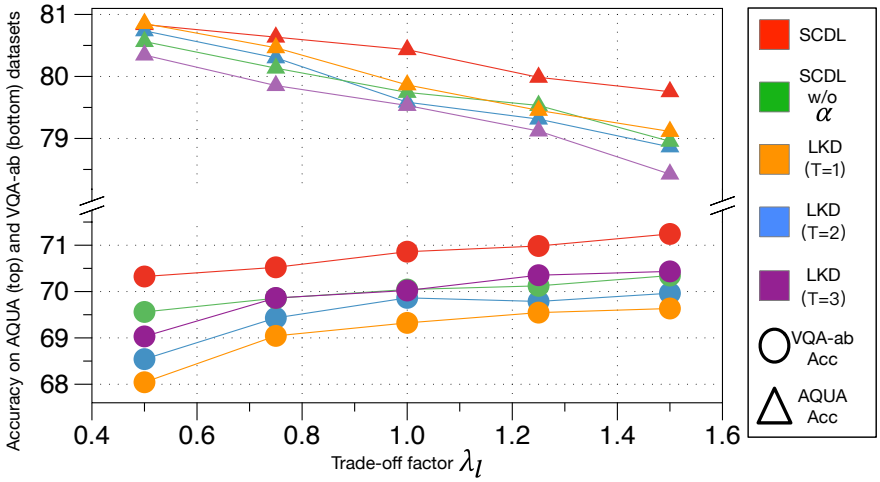

Figure 7.7: The comparisons of baseline LKD with manually-defined temperature T, our SCDL, and the counterpart without self-critical temperature α under various settings of factor λ_l.

are carried out under the first five-domain order in Tab. 7.3. As shown in Tab. 7.5, increasing the impact of λ_l from 0 to 1 would consistently boost the efficacy of reducing forgetting. If we fix the λ_l (e.g. $\lambda_l = 1$), introducing the Feature-level SCD is beneficial to the forgetting problem, which leads to a further improvement by 1.5% when $\lambda_f = 0.5$. Based on the observation, the optimal setting is $\lambda_l = 1$ and $\lambda_f = 0.5$, where dual-level distillations in our SCD are mutually complementary when reviewing old knowledge.

Table 7.5: COMPARISON OF THE AVERAGE ACCURACY WITH diverse SETTINGS OF trade-off factors λ_l and λ_f UNDER FIVE-domain sequence CLEVER→GQA→Vizwiz→AQUA→VQA-ab.

λ_f	λ_l					
	0	0.1	0.25	0.5	1	2
0	57.75	58.37	59.12	59.84	60.45	59.97
0.1	58.11	58.31	58.87	59.65	61.43	60.13
0.25	59.60	59.67	59.98	60.10	61.67	60.35
0.5	59.98	60.13	60.24	60.89	**61.94**	60.15
1	59.43	60.45	60.35	59.75	59.13	58.84
2	58.35	58.68	58.57	58.23	57.91	57.81

The number of high-responsed classes Top-C for logits-based SCD: Then, we analyze the hyper-parameters of C to identify the number of high-responsed classes for our logits-based SCD (SCDL), which acts as a crucial role on separating instance- and domain-aware knowledge. From the results in Tab. 7.6, when the C increasing from 0 to 3, our method reaches it highest performance, which reveals that top-3 predictive answers could better cover the semantic of ground-truth answer for input VQA sample. On the contrary, the other extreme setting of threshold ($C \geq 4$)

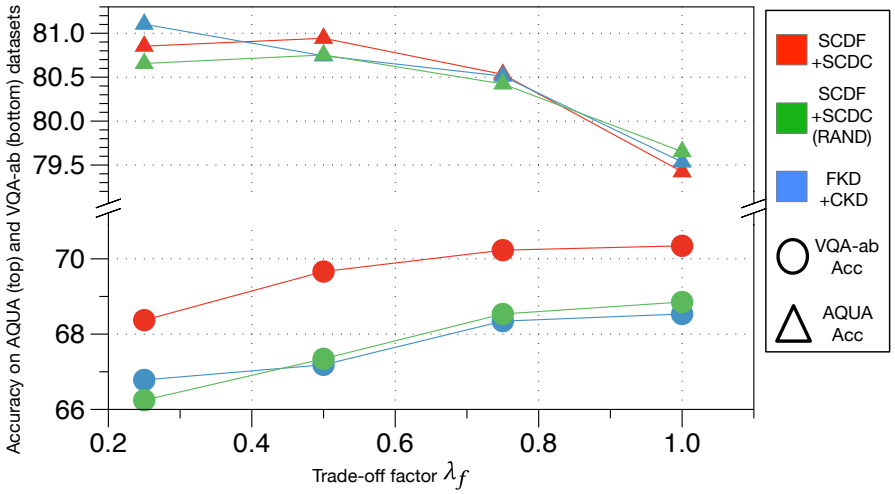

Figure 7.8: The comparisons of baseline Feature-based KD, our Feature-level SCD, and the counterpart with random removing components under various settings of trade-off factor λ_f.

would also impair the both representations of instance- and domain-level knowledge. Finally, we select $C = 3$ as the optimal setting in our experiments.

Table 7.6: Comparison of the average accuracy with dynamic settings of Top-C in logits-level SCD (SCDL) under five-domain sequence CLEVER→GQA→Vizwiz→AQUA→VQA-ab.

Method	Top-C							
	1	2	3	4	5	7	10	15
SCDL	58.1	59.8	**60.5**	60.3	60.0	59.5	59.1	58.9

The number of to-be-removed components Top-K for feature-based SCD: Finally, we experimentally validate the hyper-parameters of K to remove a specific number of visual/textual components with highest attention weights for our feature-based SCD, which acts as a crucial role on formulating the harmful domain-specific knowledge. In Tab. 7.7, our method achieves the best performance when removing Top-10 important components based on attention weights ($K = 10$), which is consistent to the observation in Fig 7.4. However, when considering more to-be-removed components (e.g., $K > 15$), the improvements caused by feature-level SCD would be impaired, since it reduces the difficulty for student model to distinguish the domain-specific knowledge from the useful knowledge learned in current domain.

7.5.6 Qualitative Results

Fig. 7.9 reveals the qualitative results of VQA samples in different domains, when employed in lifelong learning. Generally, thanks to remarkable performance of attention mechanism, the corrupted samples can be roughly considered as the uninfor-

Table 7.7: Comparison of the average accuracy with dynamic settings of Top-K in feature-level SCD (SCDF+SCDC) under five-domain sequence CLEVER→GQA→Vizwiz→AQUA→VQA-ab.

Method	Top-K							
	3	5	7	10	15	20	50	100
SCDF+SCDC	57.5	58.5	59.3	**60.0**	59.6	58.9	58.0	57.8

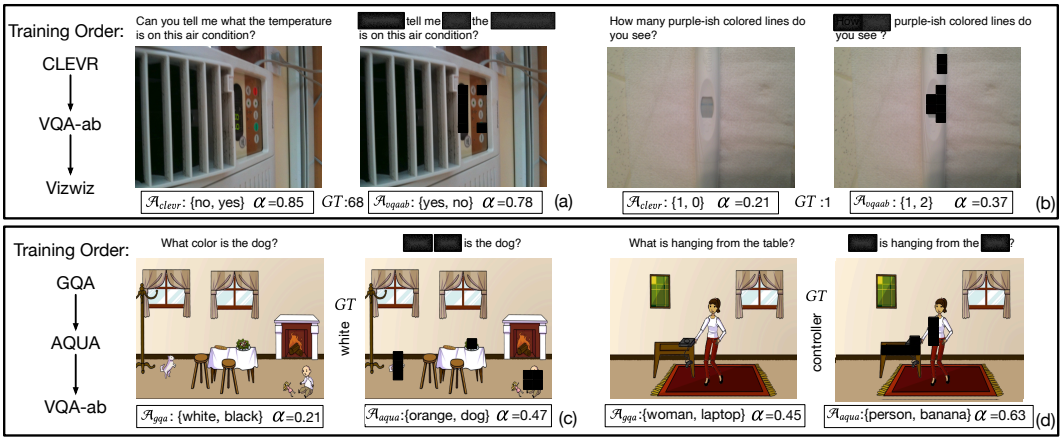

Figure 7.9: Qualitative analysis. Case (a)/(b) and (c)/(d) belong to the Vizwiz and VQA-ab datasets, respectively. The self-critical temperatures α generated by the model that is training on the third domain, in the order of CLEVR→VQA-ab→Viwiz and GQA→AQUA→VQA-ab. Each case involves the original image-question pair (left), its related corrupted counterpart (right) and the ground truth answer (GT). $\mathcal{A}_{gqa}$ denotes the Top-2 high-response classes (instance-relevant knowledge) in the previous label space.

mative domain-specific counterparts, since the majority of important question words with related image regions are removed. For the samples (b) and (c) grounded by general answers ('*1*' and '*white*'), their labels typically co-exist in the instance-aware knowledge of previous domains, even without annotated information. In the sample (a), through introspecting from counterfactual sample, the high value of self-critical temperature α can smooth the spurious instance-aware knowledge ('*yes*' and '*no*') when reviewing previous domains.

7.6 Conclusion

In this paper, we introduce a new yet practical VQA task, coined Multi-Domain Lifelong VQA (MDL-VQA). To solve this task, we propose a Self-Critical Distillation (SCD) framework to allow the VQA model to introspect its learned knowledge and further reduce forgetting ratio while efficiently learning on new data. According to this idea, we propose the counterfactual sample based introspection for rectifying logit-based knowledge distillation, and the reasoning behavior introspection to

filter the negative knowledge transferred by the feature-based distillation. Extensive experiments demonstrate that our SCD can significantly improve the model's anti-forgetting ability and outperforms other competitors by large margins on MDL-VQA.

Future work: From the benchmark side, we attempt to add the function-incremental setting into our MDL-VQA benchmarks, where each sequential tasks may focus on a specific sub-task with related answer cantidates (e.g. 'red', 'blue', 'black', and 'white' for the color identification). From the side of anti-forgetting strategy, we seek to explore more efficient introspection strategy beyond attention mechanism to analyze the model's reasoning behavior in the feature space.