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

Mingrui Lao
m.lao@liacs.leidenuniv.nl
Leiden University
The Netherlands

Nan Pu*
nan.pu@unitn.it
University of Trento
Italy

Yu Liu
liuyu8824@dlut.edu.cn
Dalian University of Technology
China

Zhun Zhong
zhunzhong007@gmail.com
University of Nottingham
United Kingdom

Erwin M. Bakker
erwin@liacs.leidenuniv.nl
Leiden University
The Netherlands

Nicu Sebe
niculae.sebe@unitn.it
University of Trento
Italy

Michael S. Lew
m.s.k.lew@liacs.leidenuniv.nl
Leiden University
The Netherlands

ABSTRACT

Visual Question Answering (VQA) has achieved significant success over the last few years, while most studies focus on training a VQA model on a stationary domain (e.g., a given dataset). In real-world application scenarios, however, 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 paper, 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. Extensive experiments demonstrate that SCD framework outperforms state-of-the-art competitors with different training orders.

CCS CONCEPTS

• **Computing methodologies** → *Computer vision.*

*Corresponding Author.

MM '23, October 29-November 3, 2023, Ottawa, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612121

KEYWORDS

Lifelong Learning, Visual Question Answering, Knowledge Distillation, Multi-Domain Learning

ACM Reference Format:
Mingrui Lao, Nan Pu, Yu Liu, Zhun Zhong, Erwin M. Bakker, Nicu Sebe, and Michael S. Lew. 2023. Multi-Domain Lifelong Visual Question Answering via Self-Critical Distillation. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, 12 pages. https://doi.org/10.1145/3581783.3612121

1 INTRODUCTION

Visual Question Answering (VQA) [4, 18] aims to answer textual questions conditioned on given images, which requires intricate vision-language reasoning. With the flourishing developments of large-scale pre-trained models [23, 35, 65] and cross-modal learning techniques [3, 53, 62], current VQA models have achieved state-of-the-art performance over various datasets [4, 18, 24, 27]. 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, VQA systems are expected to constantly acquire and update their knowledge, thereby catering to users' demands.

To empower AI machines with the capacity of acquiring new knowledge from sequentially arriving tasks with less forgetting [41] of previously learned tasks, lifelong learning [11, 34, 44] has gained extensive research interests, and inspired considerable delicate and efficient approaches [12, 22, 25, 46, 52] in both CV [57] and NLP [43] communities. However, accomplishing lifelong learning in vision-language tasks is still challenging, especially in the fields of multi-modal reasoning [50, 63]. In terms of VQA task, the work in [19] is the first attempt to explore simple class-incremental learning in the diagnostic dataset. Likewise, the method by [33] introduces a function- and scene-incremental settings on the realistic GQA dataset [24], 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 argue that the domain-incremental setting is more practical yet under-explored

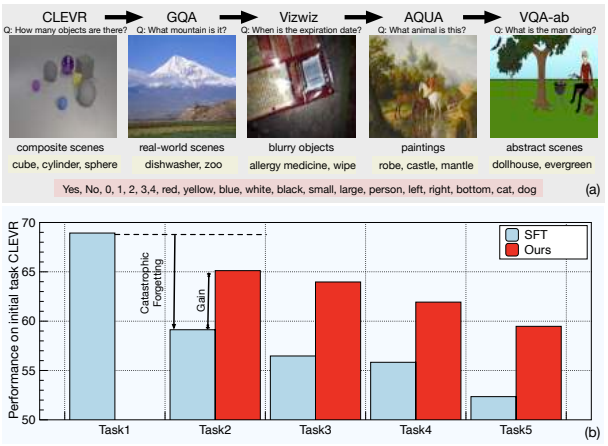

Figure 1: (a): MDL-VQA involves five tasks in different domains: CLEVR [27], GQA [24], Vizwiz [20], VQA-ab [18], AQUA [15]. The label spaces for different domains are inconsistent, where words in red shading denote some general answers coexist in several domains, and in yellow shading are domain specific. (b) The performance of the initial task (CLEVR) during the sequential fine-tuning (SFT), where VQA model encounters the catastrophic forgetting. In contrast, our method remarkably reduces the forgetting extent.

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. 1(a) and Fig. 8 (Appendix), 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 is typically collected and privacy-protected in some specific domains. Thus, the training process can use only current domain data, without storing and replaying any instances from previous domains.

To address these challenges in the MDL-VQA setting, 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, SCD is implemented on both logits-level (SCDL) and feature-level (SCDF), by addressing the following two self-critical questions for both 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. 1(b) demonstrates the capacity of our SCD to mitigate forgetting after incrementally learning across five domains. Overall, our contributions are summarized as:

- We explore a new yet practical VQA setting, 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 experiments show that SCD outperforms other competitors and achieves promising results on MDL-VQA.

2 RELATED WORKS

2.1 Lifelong Learning in Vision-Language Tasks

Lifelong learning [11, 34, 44] has been extensively explored in CV 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 [58] and image captioning [13, 60]. For the VQA task requiring high-level reasoning, [19] is the first attempt to exploit a simple class-incremental setting for VQA, where samples in question types '*wh-*' and '*yes/no*' are tested under different sequence. [33] proposes a CLOVE benchmark to establish the scene- and function-incremental learning through splitting the GQA dataset in natural visual domain.Moreover, [51] introduces a CLiMB benchmark, where models continually learn crossing different multi-modal reasoning tasks, including VQA.

Unlike these lifelong VQA benchmarks, our MDL-VQA aims to overcome forgetting issues on multiple distinct domains, as analysed in Fig. 8 (Appendix). Moreover, in contrast to standard domain-incremental learning where the label spaces are consistent across different domains, the domains in MDL-VQA include domain-specific and domain-shared answers, thereby leading to more challenges.

2.2 Multi-Domain Learning in VQA

In recent years, increasing amount of datasets [4, 15, 18, 20, 27] with diverse visual and textual domains have been proposed to facilitate VQA research. Therefore, a longstanding research topic, multi-domain learning [5] has become an attractive yet practical topic in VQA community, where most of related works focus on the

model robustness against domain shift. [49] reveals that most methods perform poorly on either natural or composite dataset, and proposes a conceptually simple RAMEN model for adapting to complex reasoning required in two domains. [6, 59] design delicate feature-learning strategies to enhance domain adaptation across different datasets. [64] analyzes domain shifts between nine widely-used VQA datasets and improve domain robustness via an unsupervised method. Except visual domain adaption, the generalization of VQA models on different linguistic domains is also crucial, especially due to models' brittleness to the language variations [48].

Different from the existing works that adapt source knowledge to a target domain, our work introduces multi-domain learning into lifelong VQA tasks, and emphasises on retaining old domains performance while adapting to any upcoming domain.

2.3 Anti-Forgetting Knowledge Distillation

The common strategies [11] 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) [17] has drawn widespread research interest. KD aims 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 [36] 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 [8, 9, 38] 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.Thus, we analyse the properties of reliance on shortcut learning and the reasoning behaviors implied among pair-wise instance interactions of attention modules, and further propose a novel Self-Critical Distillation (SCD) framework, which leverages the comprehensive analysis results to selectively transfer knowledge while depressing the negative impact of the irrelevant learned knowledge for learning on current domain.

3 MULTI-DOMAIN LIFELONG VQA

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/domains $\mathcal{D} = \left\{D^{(t)}\right\}_{t=1}^{T}$. The data in the t-th domain is comprised of train and test splits, $D^{(t)} = \left\{D_{tr}^{(t)}, D_{te}^{(t)}\right\}$. Note that, only $D_{tr}^{(t)}$ is available at the t-th training step. 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 $j \neq k$. 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. 1).

3.2 Baseline Approach

Considering the data privacy issues in MDL-VQA, we exploit three replay-free knowledge distillation (KD) approaches, based on logits [36], feature [61] and correlation [9], as our baseline. These methods often combine two learning objectives for model training. One ($\mathcal{L}_{new}$) is to acquire knowledge in the current domain, and the other ($\mathcal{L}_{old}$) aims at maintaining the old knowledge learned from previous domains. At the t-th domain, the objective is:

$$\mathcal{L}(t) = \mathcal{L}_{new} + \lambda \mathcal{L}_{old}, \tag{1}$$

where λ controls the contribution of the $\mathcal{L}_{old}$.

In our baseline, $\mathcal{L}_{new}$ is implemented by a standard cross-entropy loss in our MLD-VQA tasks. 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 loss to acquire knowledge in the current domain by:

$$\mathcal{L}_{ce} = -\sum_{(\mathbf{v},\mathbf{q},\mathbf{a}) \in D^{(t)}} \log(\sigma(f(\mathbf{v}, \mathbf{q}; \theta, \phi^{(t)}))[\mathbf{a}], \tag{2}$$

where $\sigma(\cdot)$ is the $softmax$ function, and $\phi^{(t)}$ is the classifier specialized for the domain $D^{(t)}$. Moreover, the $\mathcal{L}_{old}$ needs to consider the knowledge type which is efficient for transferring knowledge.

For **logits-based KD** [36], 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, the classification probabilities $\mathbf{p}^{(k)}(\tau)$ is calculated by:

$$p_i = \frac{\exp\left(z_i/\tau\right)}{\sum_{j=1}^{|\mathcal{A}^{(k)}|} \exp\left(z_j/\tau\right)}, \tag{3}$$

where p_i represents the probability of the i-th class in $\mathbf{p}^{(k)}$, and τ is the temperature to scale the smoothness of two distributions. Analogously, 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 (KL) Divergence [28] to constrain the teacher and the student. Given input sample, the loss function is:

$$\mathcal{L}_{lkd}(\hat{\mathbf{p}}^{(k)}(\tau), \mathbf{p}^{(k)}(\tau)) = \tau^2 \mathrm{KL}\left(\hat{\mathbf{p}}^{(k)}(\tau) \| \mathbf{p}^{(k)}(\tau)\right). \tag{4}$$

Practically, we set $k = t - 1$ to avoid the linearly-increased usage of computation sources in long-sequence lifelong learning.

For **feature-based KD** [61], 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 is regarded as the knowledge learned from previous tasks. M refers to the dimension of the intermediate feature. Then, feature-based KD employs Mean Square Error [2] (MSE) to distill the knowledge from teacher into student:

$$\mathcal{L}_{fkd}\left(\hat{f}, f\right) = \left\|\hat{f} - f\right\|_2^2, \tag{5}$$

where f is the corresponding feature from the student model $m(v, q; \theta)$.

Correlation-based KD [9] focuses on transferring the knowledge about semantic correlation of features in a training batch. Based on L2-normalized outer products [54], we can obtain the pairwise similarities $\hat{G}$ and G of the mini-batch features yielded from teacher and student, respectively. The correlation-based KD loss is given by:

$$\mathcal{L}_{ckd}\left(\hat{G}, G\right) = \frac{1}{b^2}\sum\left\|\hat{G} - G\right\|_2^2, \tag{6}$$

where b implies the batch size. We adapt the above-mentioned methods into our MDL-VQA setting and evaluate their effects in Tab. 1. We find that most of these methods achieve unsatisfactory performances, since they overlook the inherent reasoning mechanism of VQA. We analyse and discuss the drawbacks below.

3.3 Limitations

As for a practical and flexible regularization-based approach, KD could be easily deployed into any sequentially-leaning process to handle the forgetting. However, unlike other tasks in the incremental fashion, we suggest that VQA models may encountered two important challenges in the process of knowledge transferring: (i) For logits-based distillation depicted in Eq. (4), 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 [1, 26, 31, 32, 42] 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. (ii) For student learning knowledge from a new domain, it is inevitable to acquire the domain-specific knowledge (e.g. visual and linguistic styles), which is not only pointless to understanding visual concepts for question answering, but also accelerate the process of forgetting previous knowledge.

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.

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 Sec. 3.3. 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. 2. 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:

$$p_{\setminus[a,\ldots,b]} = \frac{\sum_{k=1,k\notin[a,\ldots,b]}^{|\mathcal{A}^{(k)}|}\exp\left(z_k/\tau\right)}{\sum_{j=1}^{|\mathcal{A}^{(k)}|}\exp\left(z_j/\tau\right)}. \tag{7}$$

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:

$$q_i = \frac{\exp\left(z_i/\tau\right)}{\sum_{j=1,j\notin[a,\ldots,b]}^{|\mathcal{A}^{(k)}|}\exp\left(z_j/\tau\right)}. \tag{8}$$

Based on the separated knowledge and Eq. (4), we derive the separated logits-based KD $\mathcal{L}_{slkd}$, which separately transfers IRK and DRK from teacher to student:

$$\mathcal{L}_{slkd} = \tau^2\mathrm{KL}\left(\hat{\mathbf{p}}_I^{(k)}(\tau)\|\mathbf{p}_I^{(k)}(\tau)\right) + \tau^2\mathrm{KL}\left(\hat{\mathbf{p}}_D^{(k)}(\tau)\|\mathbf{p}_D^{(k)}(\tau)\right). \tag{9}$$

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 $m(\mathbf{o}, \mathbf{q}; \hat{\theta})$. Finally, by reformulating the Equ. (9) with knowledge-specific temperatures (α and β), the SCDL loss is:

$$\mathcal{L}_{scdl} = \alpha^2\mathrm{KL}\left(\hat{\mathbf{p}}_I^{(k)}(\alpha)\|\mathbf{p}_I^{(k)}(\alpha)\right) + \beta^2\mathrm{KL}\left(\hat{\mathbf{p}}_D^{(k)}(\beta)\|\mathbf{p}_D^{(k)}(\beta)\right), \tag{10}$$

$$\alpha = max(\frac{\hat{\mathbf{f}}\cdot\hat{\mathbf{b}}}{\|\hat{\mathbf{f}}\|\|\hat{\mathbf{b}}\|}, 0)\cdot\tau_{max}, \quad \beta = \tau_{max} - \alpha, \tag{11}$$

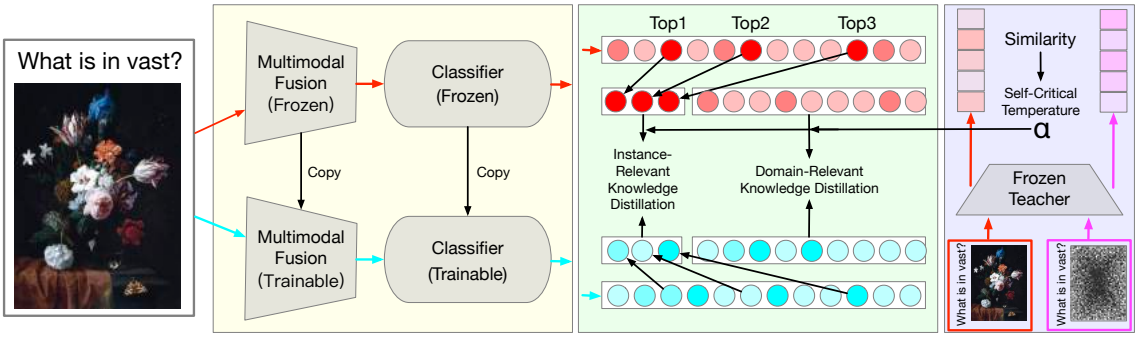

Figure 2: Illustration of Logits-level SCD, where training samples in current domain is fed into both teacher and student networks (yellow region). In the green region, the long vectors in red and blue denote the raw predictions yielded from teacher and student respectively, which are separated into dual-level knowledge based on high-response classes. The purple region depicts how to compute the temperature α via introspecting the teacher about raw (red) and counterfactual (pink) samples.

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.

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) [55, 56] equipped in competitive transformer-based VQA models [29, 53, 62]. 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. 9 in Appendix), among more than 200 visual and textual components in a VQA instance, merely several visual/textual components are crucial.

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 $(\dot{v}, \dot{q})$. Then, we feed $\dot{v}$ and $\dot{q}$ into the currently-trained student model to obtain the $\dot{f} = m(\dot{v}, \dot{q}; \theta)$. We assume that the $\dot{f}$ 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:

$$\mathcal{L}_{scdf}\left(\hat{\boldsymbol{f}}, \boldsymbol{f}\right) = \max\left(\|\hat{\boldsymbol{f}} - \boldsymbol{f}\|^2 - \|\dot{\boldsymbol{f}} - \boldsymbol{f}\|^2, 0\right). \tag{12}$$

Meanwhile, we also propose to substitute the correlation-based KD (Eq. (6)) by metric learning, which is formulated by:

$$\mathcal{L}_{scdc}\left(\hat{G}, G\right) = \max\left(\|G - \hat{G}\|^2 - \|G - \dot{G}\|^2, 0\right), \tag{13}$$

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

Discussion: From the conceptual example in Fig. 3, through the self-criticism from reasoning behaviour of attention, 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 to narrow the semantic distance between samples with consistent domain-invariant concepts, and meanwhile weaken the negative impact from domain-specific biases.

Table 1: Non-forgetting evaluation in MDL-VQA benchmark. We test model after sequentially training on all datasets in five domains (t=5), where we select two different orders starting from synthetic (CLEVER) and real-world (GQA) datasets.

Method	CLEVR→GQA→Vizwiz→AQUA→VQA-ab					$\bar{s}$	$\bar{f}$	GQA→Vizwiz→AQUA→VQA-ab→CLEVR					$\bar{s}$	$\bar{f}$
SFT	52.35	47.98	36.57	76.88	75.01	57.76	10.46	42.20	34.06	67.24	62.24	68.25	54.80	14.07
EWC [30]	52.14	48.05	36.68	77.13	74.92	57.78	10.40	41.68	33.99	67.38	62.77	68.15	54.79	14.05
ALASSO [45]	52.43	48.28	36.24	77.37	75.19	57.90	10.32	42.41	34.39	67.19	62.89	68.44	55.07	13.78
FKD [61]	54.77	50.45	37.56	78.12	75.26	59.23	8.68	44.98	36.03	69.14	64.47	68.01	56.53	11.85
SPD [54]	53.97	49.16	37.88	77.89	**76.15**	59.01	9.19	42.52	35.11	69.04	63.99	68.91	55.91	12.84
LWF [36]	55.43	50.74	37.93	78.36	74.82	59.46	8.29	45.98	38.65	69.02	65.63	67.91	57.44	10.69
IRG [37]	56.32	51.02	37.14	77.97	75.43	59.58	8.29	46.18	39.90	68.76	65.33	68.50	57.73	10.46
ECD [8]	54.30	49.86	37.99	77.94	75.67	59.15	8.88	42.82	35.21	68.45	64.23	68.44	55.83	12.83
DKD [9]	55.54	51.15	37.83	78.25	**76.15**	59.79	8.21	46.20	37.74	68.84	65.82	**69.20**	57.76	10.86
MBP [38]	57.58	51.24	40.77	77.87	74.69	60.43	7.04	47.73	39.85	73.56	68.15	67.95	59.44	8.19
Ours	**59.48**	**52.47**	**43.41**	**79.44**	74.90	**61.94**	**5.30**	**50.11**	**40.98**	**74.98**	**69.52**	68.14	**60.74**	**6.61**
Reference	68.93	59.21	46.10	81.38	75.34	66.19	-	59.21	46.10	81.38	75.34	68.93	66.19	-

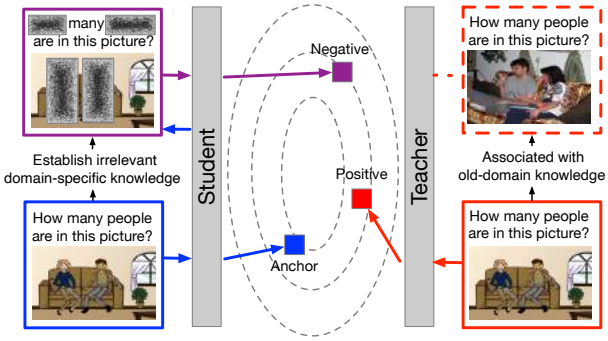

Figure 3: Conceptual illustration of Feature-level SCD, where we suppose the currently-trained samples are in abstract domain, and the training data utilized in the previous task is in realistic domain. The samples in red, and blue refer to the intermediate features extracted from teacher and student.

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:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ce}} + \lambda_l \mathcal{L}_{scdl} + \lambda_f (\mathcal{L}_{scdf} + \mathcal{L}_{scdc}), \tag{14}$$

where $\mathcal{L}_{scdl}$ and $(\mathcal{L}_{scdf} + \mathcal{L}_{scdc})$ denote the loss terms of logits- and feature-level SCD defined in Equ. (10), (12) and (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.

5 EXPERIMENTS

5.1 Implementation Details

For the training of lifelong learning, we optimize VQA models by the AdamW optimizer [39] 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. We set the minibatch size as 32, which is evenly distributed on two GPUs. For network architecture, We use a pre-trained Vision-Language Transformer (ViLT) [29] as the backbone multimodal fusion encoder. Unlike other pre-trained vision-language models [10, 40] that build upon region-level features extracted from Faster R-CNN [47], 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. We select the trade-off factors $\lambda_l = 1$ and $\lambda_l = 0.5$. We set the number of high-response classes as the IRK is $C = 3$, and $\tau_{max} = 4$. The number of visual/textual components to be removed in feature-level SCD is set to 10. All the hyper-parameters are validated in Appendix.

Table 2: 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

5.2 Datasets

To comprehensively evaluate the MDL-VQA models, we propose a MDL-VQA benchmark. We exploit five VQA datasets where the images are represented in various visual domains, including artistic, abstract, real-world, synthetic and blurred-objects scenes. To be specific, *AQUA* dataset [15] aims to ask questions about artworks, where the artistic images are obtained from SemArt [14] dataset. *VQA-abstract* [18] contains the images of abstract/cartoon scenes. *GQA* [24] is a large-scale dataset to test multiple reasoning skills through compositional questions, where images are described in high-quality real-world scenes. *Vizwiz* [20] 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* [27] is a diagnostic dataset with synthetic 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. 2.

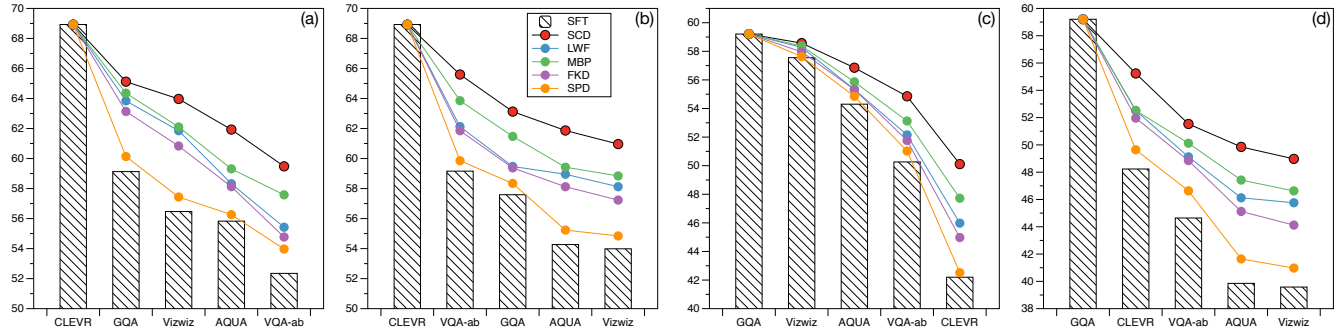

Figure 4: Evaluation against order variations. (a)/(b) and (c)/(d) illustrate the trend of accuracy computed from initial tasks against 4 different sequences. We use shadowed bar to represent the SFT, and lines with different colors for KD-based approaches.

5.3 Performance Evaluation

We validate the SCD on our MDL-VQA benchmark against the competitive methods in Tab. 1, with the evaluate metrics of average accuracy $\bar{s}$ and forgetting $\overline{f}$ (details in Appendix). The SFT is to sequentially train VQA model with newly-arrived datasets without reviewing old knowledge. EWC [30] and ALASSO [45] refer to the prior-focused regularization methods, which focuses on penalizing network parameters in sequential training. FKD [61], SPD [54], LWF [36] and IRG [37] are data-focused KD approaches firstly deployed in the scenarios of model compression, among which LWF, FKD, SPD [54] denote our baseline strategies described in Eqs. (4) to (6). ECD [8], DKD [9] and MBP [38] are the advanced KD strategies tailored to class-incremental lifelong 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 jointly considered logits- 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 it 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 raw 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. 4. 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 3: 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).

Configurations					Order1		VQA-ab→AQUA	
Case	$\mathcal{L}_{ce}$	$\mathcal{L}_{scdl}$	$\mathcal{L}_{scdf}$	$\mathcal{L}_{scdc}$	$\bar{s}$	$\overline{f}$	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

5.4 Ablation Study

(1) **Efficacy of Different Components** We first analyze the effectiveness of different components in our SCD. Specifically, the experimental results are reported in Tab.3, which are obtained from the first order involved in Tab. 1, 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.

(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 (LKD) (Eq. (4)) 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. (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. 6, 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

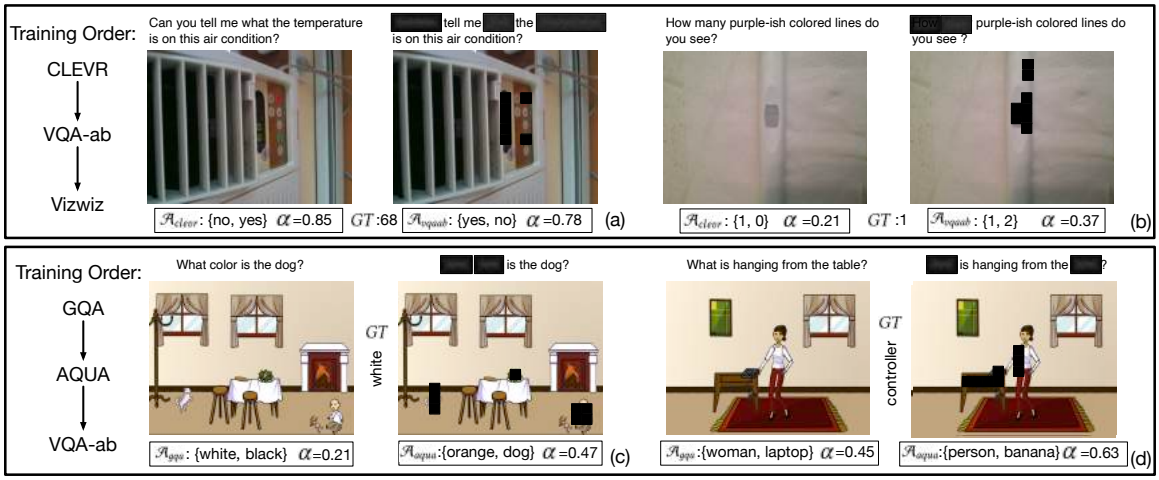

Figure 5: Case (a)/(b) and (c)/(d) belong to Vizwiz and VQA-ab datasets. The self-critical temperatures α generated by the model that is training on the third domain. Each case involves the raw image-question pair (left), its related corrupted counterpart (right) and ground truth (GT). $\mathcal{A}$ is the Top-2 high-response classes (instance-relevant knowledge) in the previous domains.

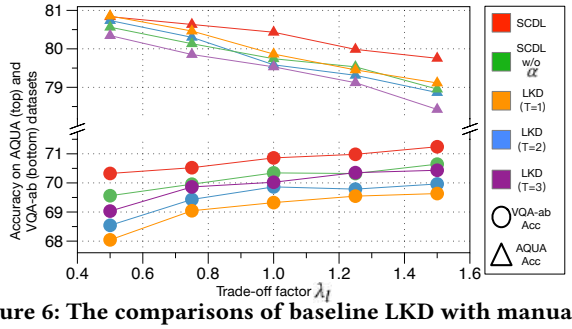

Figure 6: The comparisons of baseline LKD with manually-defined temperature T, SCDL, and the counterpart without self-critical temperature α under various settings of λ_l.

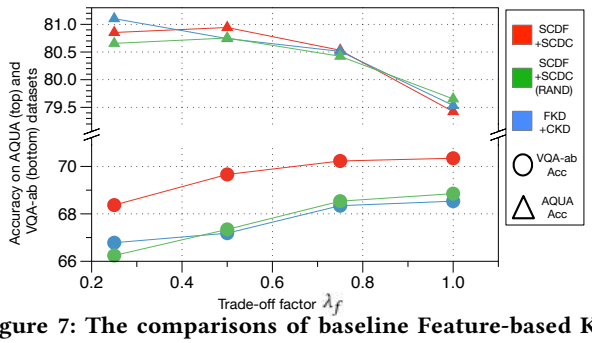

Figure 7: The comparisons of baseline Feature-based KD, our Feature-level SCD, and the counterpart with random removing components under various settings of λ_f.

operation is beneficial to overcome forgetting in the MDL-VQA with label-space variations, and the self-critical temperature can further promotes teacher to transfer more informative knowledge.

(3) **Feature-level SCD vs Feature-based KD**: We compare standard feature-based KD (FKD+CKD in Eqs. (5) and (6)) with feature-level SCD, and the counterparts that corrupting components randomly (RAND) in Fig. 7. We 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. In contrast, 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 AQUA dataset when acquiring new knowledge.

5.5 Qualitative Results

Fig. 5 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 uninformative 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 annotations. In case (a), the high value of self-critical temperature α can smooth the spurious instance-aware knowledge ('*yes*' and '*no*') when reviewing previous domains.

6 CONCLUSION

In this paper, we introduce a new yet practical VQA task, coined Multi-Domain Lifelong 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, we propose the counterfactual sample based introspection for rectifying logit-based distillation, and the reasoning behavior introspection to filter the negative knowledge transferred by the feature-based distillation. Extensive experiments show that SCD remarkably improves model's anti-forgetting ability and outperforms other competitors by large margins on MDL-VQA. Moving forward, we will explore a learnable loss weight to coordinate dual-level distillations.

7 ACKNOWLEDGMENTS

This work was supported mainly by the LIACS Media Lab at Leiden University, and partially by the China Scholarship Council and the NSF of China under grant 62102061.

REFERENCES

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4971–4980.

[2] David M Allen. 1971. Mean square error of prediction as a criterion for selecting variables. *Technometrics* 13, 3 (1971), 469–475.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6077–6086.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.

[5] Rodrigo Berriel, Stephane Lathuillere, Moin Nabi, Tassilo Klein, Thiago Oliveira-Santos, Nicu Sebe, and Elisa Ricci. 2019. Budget-aware adapters for multi-domain learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 382–391.

[6] Wei-Lun Chao, Hexiang Hu, and Fei Sha. 2018. Cross-dataset adaptation for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5716–5725.

[7] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. 2018. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European conference on computer vision (ECCV)*. 532–547.

[8] Wei Chen, Yu Liu, Nan Pu, Weiping Wang, Li Liu, and Michael S Lew. 2021. Feature estimations based correlation distillation for incremental image retrieval. *IEEE Transactions on Multimedia* 24 (2021), 1844–1856.

[9] Wei Chen, Haoyang Xu, Nan Pu, Yu Liu, Mingrui Lao, Li Liu, Weiping Wang, and Michael S Lew. 2022. Lifelong Fine-grained Image Retrieval. *IEEE Transactions on Multimedia* (2022).

[10] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Learning universal image-text representations. (2019).

[11] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. 2021. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence* 44, 7 (2021), 3366–3385.

[12] Cyprien de Masson D'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. *Advances in Neural Information Processing Systems* 32 (2019).

[13] Riccardo Del Chiaro, Bartłomiej Twardowski, Andrew Bagdanov, and Joost Van de Weijer. 2020. Ratt: Recurrent attention to transient tasks for continual image captioning. *Advances in Neural Information Processing Systems* 33 (2020), 16736–16748.

[14] Noa Garcia and George Vogiatzis. 2018. How to read paintings: semantic art understanding with multi-modal retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 0–0.

[15] Noa Garcia, Chentao Ye, Zihua Liu, Qingtao Hu, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Teruko Mitamura. 2020. A dataset and baselines for visual question answering on art. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 92–108.

[16] Katy Gero, Chris Kedzie, Jonathan Reeve, and Lydia Chilton. 2019. Low-Level Linguistic Controls for Style Transfer and Content Preservation. *arXiv preprint arXiv:1911.03385* (2019).

[17] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision* 129 (2021), 1789–1819.

[18] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6904–6913.

[19] Claudio Greco, Barbara Plank, Raquel Fernández, and Raffaella Bernardi. 2019. Psycholinguistics meets continual learning: Measuring catastrophic forgetting in visual question answering. *arXiv preprint arXiv:1906.04229* (2019).

[20] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3608–3617.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[22] Wenpeng Hu, Zhou Lin, Bing Liu, Chongyang Tao, Zhengwei Tao, Jinwen Ma, Dongyan Zhao, and Rui Yan. 2018. Overcoming catastrophic forgetting for continual learning via model adaptation. In *International conference on learning representations*.

[23] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. 2021. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12976–12985.

[24] Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6700–6709.

[25] Khurram Javed and Martha White. 2019. Meta-learning representations for continual learning. *Advances in neural information processing systems* 32 (2019).

[26] Jingjing Jiang, Ziyi Liu, Yifan Liu, Zhixiong Nan, and Nanning Zheng. 2021. X-ggm: Graph generative modeling for out-of-distribution generalization in visual question answering. In *Proceedings of the 29th ACM international conference on multimedia*. 199–208.

[27] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2901–2910.

[28] James M Joyce. 2011. Kullback-leibler divergence. In *International encyclopedia of statistical science*. 720–722.

[29] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*. PMLR, 5583–5594.

[30] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences* 114, 13 (2017), 3521–3526.

[31] Mingrui Lao, Yanming Guo, Yu Liu, Wei Chen, Nan Pu, and Michael S Lew. 2021. From superficial to deep: Language bias driven curriculum learning for visual question answering. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3370–3379.

[32] Mingrui Lao, Nan Pu, Yu Liu, Kai He, Erwin M Bakker, and Michael S Lew. 2023. COCA: COllaborative CAusal Regularization for Audio-Visual Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 12995–13003.

[33] Stan Weixian Lei, Difei Gao, Jay Zhangjie Wu, Yuxuan Wang, Wei Liu, Mengmi Zhang, and Mike Zheng Shou. 2023. Symbolic replay: Scene graph as prompt for continual learning on vqa task. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 1250–1259.

[34] Timothée Lesort, Vincenzo Lomonaco, Andrei Stoian, Davide Maltoni, David Filliat, and Natalia Díaz-Rodríguez. 2020. Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. *Information fusion* 58 (2020), 52–68.

[35] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*. Springer, 121–137.

[36] Zhizhong Li and Derek Hoiem. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence* 40, 12 (2017), 2935–2947.

[37] Yufan Liu, Jiajiong Cao, Bing Li, Chunfeng Yuan, Weiming Hu, Yangxi Li, and Yunqiang Duan. 2019. Knowledge distillation via instance relationship graph. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7096–7104.

[38] Yu Liu, Xiaopeng Hong, Xiaoyu Tao, Songlin Dong, Jingang Shi, and Yihong Gong. 2022. Model behavior preserving for class-incremental learning. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[39] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[40] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* 32 (2019).

[41] Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*. Vol. 24. Elsevier, 109–165.

[42] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12700–12710.

[43] Daniel W Otter, Julian R Medina, and Jugal K Kalita. 2020. A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems* 32, 2 (2020), 604–624.

[44] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. 2019. Continual lifelong learning with neural networks: A review. *Neural networks* 113 (2019), 54–71.

[45] Dongmin Park, Seokil Hong, Bohyung Han, and Kyoung Mu Lee. 2019. Continual learning by asymmetric loss approximation with single-side overestimation. In

Proceedings of the IEEE/CVF international conference on computer vision. 3335–3344.

[46] Nan Pu, Wei Chen, Yu Liu, Erwin M Bakker, and Michael S Lew. 2021. Lifelong person re-identification via adaptive knowledge accumulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7901–7910.

[47] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).

[48] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. 2019. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6649–6658.

[49] Robik Shrestha, Kushal Kafle, and Christopher Kanan. 2019. Answer them all! toward universal visual question answering models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10472–10481.

[50] Jingkuan Song, Yuyu Guo, Lianli Gao, Xuelong Li, Alan Hanjalic, and Heng Tao Shen. 2018. From deterministic to generative: Multimodal stochastic RNNs for video captioning. *IEEE transactions on neural networks and learning systems* 30, 10 (2018), 3047–3058.

[51] Tejas Srinivasan, Ting-Yun Chang, Leticia Pinto Alva, Georgios Chochlakis, Mohammad Rostami, and Jesse Thomason. 2022. Climb: A continual learning benchmark for vision-and-language tasks. *Advances in Neural Information Processing Systems* 35 (2022), 29440–29453.

[52] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 8968–8975.

[53] Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* (2019).

[54] Frederick Tung and Greg Mori. 2019. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1365–1374.

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[56] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418* (2019).

[57] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, Eftychios Protopapadakis, et al. 2018. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience* 2018 (2018).

[58] Kai Wang, Luis Herranz, and Joost van de Weijer. 2021. Continual learning in cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3628–3638.

[59] Yiming Xu, Lin Chen, Zhongwei Cheng, Lixin Duan, and Jiebo Luo. 2019. Open-ended visual question answering by multi-modal domain adaptation. *arXiv preprint arXiv:1911.04058* (2019).

[60] Shipeng Yan, Lanqing Hong, Hang Xu, Jianhua Han, Tinne Tuytelaars, Zhenguo Li, and Xuming He. 2022. Generative negative text replay for continual vision-language pretraining. In *European Conference on Computer Vision*. Springer, 22–38.

[61] Lu Yu, Vacit Oguz Yazici, Xialei Liu, Joost van de Weijer, Yongmei Cheng, and Arnau Ramisa. 2019. Learning metrics from teachers: Compact networks for image embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2907–2916.

[62] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6281–6290.

[63] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. 2018. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems* 29, 12 (2018), 5947–5959.

[64] Mingda Zhang, Tristan Maidment, Ahmad Diab, Adriana Kovashka, and Rebecca Hwa. 2021. Domain-robust vqa with diverse datasets and methods but no target labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7046–7056.

[65] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 13041–13049.

APPENDICES

.1 Dataset 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. 1 (a), these domains include real-world scenes (GQA [24]), abstract scenes (VQA-ab [18]), synthetic scenes (CLEVR [27]), paintings (AQUA [15]) and blur-object scenes (Vizwiz [20]).

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:

$$\begin{aligned}\mathrm{MMD}\left(\mathcal{D}^{(i)},\mathcal{D}^{(j)}\right) &= \left\|\mathbb{E}_{X\sim\mathcal{D}^{(i)}}\left[\varphi(X)\right]-\mathbb{E}_{Y\sim\mathcal{D}^{(j)}}\left[\varphi(Y)\right]\right\|_{\mathcal{H}} \\ &= \frac{1}{|D^{(i)}|^2}\sum_{i=1}^{|D^{(i)}|}\sum_{j=1}^{|D^{(i)}|}k\left(\mathbf{x}_i,\mathbf{x}_j\right)+\frac{1}{|D^{(j)}|^2}\sum_{i=1}^{|D^{(j)}|}\sum_{j=1}^{|D^{(j)}|}k\left(\mathbf{y}_i,\mathbf{y}_j\right) \\ &-\frac{2}{|D^{(i)}||D^{(j)}|}\sum_{i=1}^{|D^{(i)}|}\sum_{j=1}^{|D^{(j)}|}k\left(\mathbf{x}_i,\mathbf{y}_j\right),\end{aligned} \tag{15}$$

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 [21] 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 [48], we follow the work [16] to extract 20 low-level features: question length, prepositions, number of conjunctions, pronouns, etc.

By analysing the MMD comparisons among the five datasets in Fig. 8, we find that domain shifts are severe among every pair of datasets. Especially, the question domain gap between CLEVR and other four datasets is remarkable, as CLEVR involves more complex linguistic expressions to test the VQA reasoning abilities. Thus, it is important to prevent the model from learning only DSK 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 8: Visual and textual domain gaps measured by Maximum Mean Discrepancy (MMD). The green and blue shadings are MMD over the textual syntax statistics and visual features, respectively. Domain shifts are severe among every pair of datasets.

.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 [18]. To quantitatively validate the efficiency of related strategies to alleviate forgetting problem, we exploit the Average Accuracy [8] and Average Forgetting [7] as evaluation metric in our MDL-VQA.

Average Accuracy. Suppose that, after sequential learning across t domains, $acc_t^{(i)}$ is 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 ($i \le t$). The average accuracy $\bar{s}_t$ at the t-th stage is defined as $\bar{s}_t = \frac{1}{t}\sum_{i=1}^{t} acc_t^{(i)}$.

Average Forgetting is to quantify the forgetting ratio $\overline{f}_t$ after learning the t-th domain ($t \ge 2$). 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:

$$\overline{f}_t = \frac{1}{t-1}\sum_{i=1}^{t-1}\left(\max_{l\in\{1,\ldots,t-1\}} acc_{max}^{(i)} - acc_t^{(i)}\right), \forall i < t. \tag{16}$$

.3 Distribution of attention weights

Fig. 4 visualizes the distribution of sorted attention weights generated from ViLT model [29] mentioned in Section 4.2, which evenly considers 10,000 VQA samples from five datasets in MDL-VQA benchmark. The extreme long-tail distribution demonstrates that, among more than 200 components in a VQA instance, merely several visual/textual components are crucial to deduce the correct answer. As a result, in our proposed Feature-level SCD, we select to remove Top-10 visual/textual component with highest attention weights in our proposed feature-level SCD, and the corrupted samples maintained with less-attended components can be regarded as the domain-specific knowledge.

.4 Experiments for hyper-parameters

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. (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 are carried out

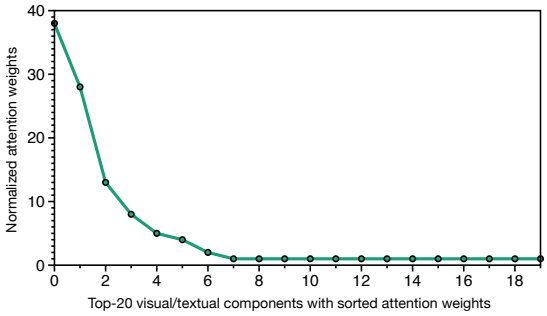

Figure 9: The Distribution of sorted attention weights based on 10,000 VQA samples from five datasets as depicted in Section 4.2.

under the first five-domain order in Tab. 1. As shown in Tab. 4, 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 4: Comparison of the average accuracy with diverse setting 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. 5, 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. In contrast, the other extreme setting of threshold ($C \geq 4$) would also impair the both representations of instance- and domain-level knowledge. Hence, we select $C = 3$ in SCD.

Table 5: Comparison of the average accuracy with dynamic settings of Top-C in logits-level SCD (SCDL).

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. 6, 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. 9. 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.

Table 6: 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