

One VLM to Keep it Learning: Generation and Balancing for Data-free Continual Visual Question Answering

Deepayan Das¹ Davide Talon² Massimiliano Mancini¹
 Yiming Wang² Elisa Ricci^{1,2}

¹University of Trento ²Fondazione Bruno Kessler

{deepayan.das, massimiliano.mancini, e.ricci}@unitn.it

{dtalon, ywang}@fbk.eu

Abstract

Vision-Language Models (VLMs) have shown significant promise in Visual Question Answering (VQA) tasks by leveraging web-scale multimodal datasets. However, these models often struggle with continual learning due to catastrophic forgetting when adapting to new tasks. As an effective remedy to mitigate catastrophic forgetting, rehearsal strategy uses the data of past tasks upon learning new task. However, such strategy incurs the need of storing past data, which might not be feasible due to hardware constraints or privacy concerns. In this work, we propose the first data-free method that leverages the language generation capability of a VLM, instead of relying on external models, to produce pseudo-rehearsal data for addressing continual VQA. Our proposal, named as GaB, generates pseudo-rehearsal data by posing previous task questions on new task data. Yet, despite being effective, the distribution of generated questions skews towards the most frequently posed questions due to the limited and task-specific training data. To mitigate this issue, we introduce a pseudo-rehearsal balancing module that aligns the generated data towards the ground-truth data distribution using either the question meta-statistics or an unsupervised clustering method. We evaluate our proposed method on two recent benchmarks, i.e. VQACL-VQAv2 and CLOVE-function benchmarks. GaB outperforms all the data-free baselines with substantial improvement in maintaining VQA performance across evolving tasks, while being on-par with methods with access to the past data. Code and models are available at <https://github.com/Deepayan137/GaB>.

1. Introduction

Visual Question Answering (VQA) has been a prominent research challenge since its introduction a decade

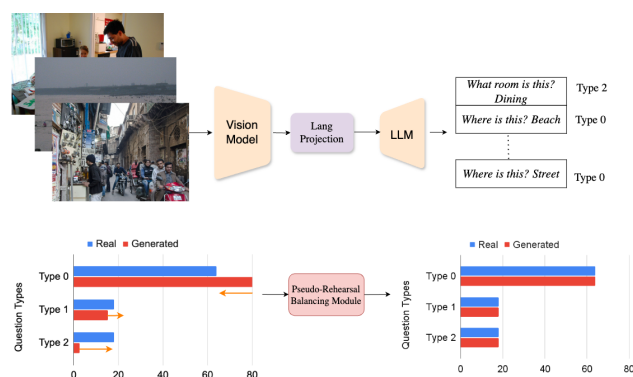


Figure 1. Key ideas. **Top:** we explore the language generation capability of VLMs to synthesize pseudo-rehearsal data of previous tasks to mitigate catastrophic forgetting in continual VQA. **Bottom:** as pseudo-rehearsal data tends to skew towards particular question types, we further propose a pseudo-rehearsal balancing module to align such skewed distribution towards the ground-truth meta-statistics, effectively improving the task performance while avoiding forgetting.

ago [4], aiming to mimic the real-world scenarios in which an agent can interpret and reason about the visual content, e.g. the object attributes, their relations, and common-sense [13, 16, 40, 57] through natural language questions. One of the main challenges of VQA is to bridge the multimodal gap as both visual and textual understanding are required, and their knowledge representation should be integrated. Recent Vision Language Models (VLMs), e.g., BLIP-2 [23] and LLaVa [26], have greatly enhanced such multimodal understanding thanks to their web-scale pre-training with image-text pairs. VLMs in addressing the VQA task, have demonstrated a good generalization ability, yet they require continuous model updates to cope with new knowledge and to handle new types of reasoning skills [40, 54]. The main challenge in updating VQA models in such continual learning setup, a.k.a. VQACL [57], is to avoid the famous

catastrophic forgetting problem, i.e., without hindering the performance of previously learned tasks upon learning new ones [14, 15, 40, 46, 57]. One common remedy to catastrophic forgetting is via data rehearsal, i.e., replay of the data samples of prior tasks during the model updates for new tasks [5, 8, 32]. However, such prior data might not be available due to privacy or proprietary concerns [45]. *Can we address VQACL without the access to data samples of previous tasks?* Prior methods leverage pseudo-rehearsal by generating data samples of prior task [22, 36, 40]. Such pseudo-rehearsal methods are capable of mitigating catastrophic forgetting in a data-free manner. However, existing methods leverage additional models to produce rehearsal samples [22, 36] or computationally demanding adversarial samples [40].

In the era of VLMs, can we exploit their generative language model to generate data samples? In this work, we explore this direction, instead of relying on additional models. We propose **GaB**, the first data-free method that leverages the language module of a VLM to address VQACL by generating question-answer pairs to mimic old tasks. Specifically, we generate questions and pseudo-answers about previous tasks by conditioning on the current visual data at hand. Crucially, the lack of diversity of questions related to the same task leads to generation collapse to data frequently occurring questions, inhibiting the VQA model capabilities (as shown in Fig. 1). To mitigate this problem, we balance the generated multimodal data: we initially learn to categorize the questions that are posed during learning new tasks and later categorize generated data accordingly. Hence, the rehearsal buffer is re-sampled ensuring that training data and generated question type distributions align.

We evaluate GaB on continual learning VQA benchmarks, namely VQACL-VQAv2 [57] and CLOVE-function [22] and compare the proposed approach with respect to state-of-the-art and continual learning baselines. Results show a large margin improvement on all considered settings among data-free methods.

To summarize, our contributions are three-fold:

- We propose the first data-free method for VQACL, leveraging the language model inherent in VLMs to generate pseudo-rehearsal data, achieving the best performance among data-free methods on VQACL-VQAv2 [57] and CLOVE-function [22].
- We propose to jointly learn a task-specific projection module for generating question-answer pairs conditioned on images of the current task, to then synthesize data for the rehearsal buffer.
- We propose a novel and effective pseudo-rehearsal balancing module to mitigate the skewed distribution of the VLM-generated question types.

2. Related Work

In the following, we review related work on continual learning and visual question answering.

Continual Learning. Continual learning (CL) addresses the problem of learning new knowledge while avoiding forgetting old concepts, a problem commonly known as catastrophic forgetting [11, 20]. Depending on the strategy to mitigate forgetting, CL approaches can be divided into different categories. Regularization-based methods incorporate additional regularization terms into the objective functions when learning a new task [2, 18, 20, 25, 35, 55]. Instead, rehearsal-based techniques replay previous tasks samples alongside new task data [5, 8, 32, 43, 44]. However, storing data may be limiting in settings with privacy or copyright concerns. Architecture-based methods [3, 21, 24, 33, 34, 49, 50, 53] allocate different model parameters for different tasks. Here we follow the second family of approaches, addressing forgetting by replaying knowledge. However, we exploit the capabilities of VLMs to generate pseudo question-answer pairs of previous tasks. Contrary to previous work, we employ a data-free rehearsal strategy that do not revert the past task images [9, 39, 52] but mimic the old language question-answer pairs on incoming new visual samples.

Continual Learning and VLMs. Vision Language foundation models [1, 23, 31] provide general knowledge that can serve as the basis for specialized models. By building on the features transferability and the initial off-the-shelf strong performance, recent work has investigated the adaptation of VLMs for the continual setting [38, 40, 51, 54, 56]. While most of the work has focused on maintaining the zero-shot generalization capabilities of foundation VLMs [51, 54], multimodal continual learning approaches have relied on model expansion [48, 61], caption and label matching [6], uncertainty [17] and prompting [47]. In contrast to [19] building on VLMs for pseudo-labels refinement, our approach leverages the textual generation of question-answer pairs on new data. After the initial expansion of reply heads, the model replays previous questions, avoiding estimating the task identity at inference time [48, 61].

Continual Learning and VQA. Despite recent advancements in the classic VQA setting [1, 10, 23, 26], limited work concentrates on the multi-modal continual learning VQA setting where a model is sequentially trained to address different skills [15, 22, 29, 30, 40, 41, 57, 58, 60]. VQACL [57] leverages previous tasks samples and the question-type, question-object separation. Instead, REMIND [15] builds on latent replay samples as a form of compressed training data storage. Thanks to scene-graphs provided from a pre-trained model, [22] maintains a symbolic replay memory collecting

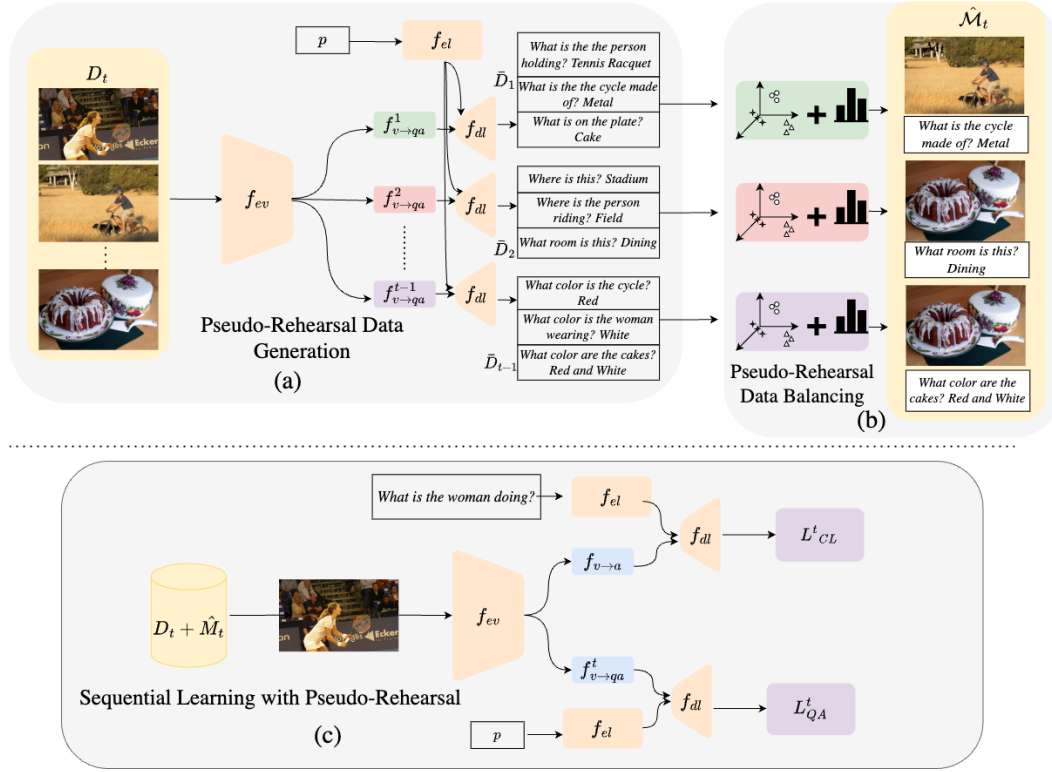


Figure 2. **Architecture of the proposed data-free method GaB for addressing VQACL.** (a) At task t , past task-specific projection heads $f_{v \rightarrow qa}^s, s = 1 \dots t-1$ are used to generate pseudo-rehearsal data with question-answer pairs about old tasks on current task images. (b) Pseudo-rehearsal samples undergo a balancing process through a module designed to ensure that under-represented question types are adequately represented. Finally (c) uses the pseudo-rehearsal data to mitigate the forgetting in the sequential learning scenario and data D_t to learn the current question-answer projector $f_{v \rightarrow qa}^t$.

scene-graph and question-answer pairs. We set apart from replay-based approaches [15, 22, 58] that collect statistics or a coreset of previous tasks samples. Similarly to [40] we build on incoming data to generate pseudo-labeled replay samples, however, by building on VLM language generation capabilities to replicate previous tasks questions in a self-training manner, we avoid the large number of expensive adversarial negative samples. While automatic question generation has been explored for question-answer rehearsal [21, 42], previous work costly generate context for previous data samples [42] or requires past rehearsal data to access conditioning answer types [21]. On the contrary, our pseudo-samples strategy leverage current task visual data as augmentation strategy to replay previous tasks.

3. Method

In this section we present GaB, our novel data-free approach for continual VQA. We start with a formal definition of the VQACL task with the necessary notations. We then provide an overview of the proposed method, followed by a detailed description on each key element, including

the *pseudo-rehearsal data generation*, the *pseudo-rehearsal data balancing*, and the *sequential learning with pseudo-rehearsal*.

Problem formulation. Continual Learning for Visual Question Answering is generally defined as learning on a sequence of T VQA tasks querying about different set of properties. Let $D_t, t = 1 \dots T$, be the only dataset available at task t . Each dataset is defined as $D_t = \{d_t^1, d_t^2, \dots, d_t^{N_t}\}$, where d_t^i consists of an image-question-answer triplet (x_t^i, q_t^i, a_t^i) and N_t is the number of samples of the t -th task. Here, x_t^i denotes the image, q_t^i the associated question, and a_t^i is the correct answer. The goal of the VQA model f is to correctly predict the answer a_t^i from the associated visual input x_t^i with question q_t^i , while mitigating the catastrophic forgetting of old tasks $s, s < t$. In the following, when the task or the sample are clear from the context, we simplify the notation by dropping the associated indexes, e.g., x, q, a .

Addressing VQACL with data rehearsal. A common and effective strategy for addressing catastrophic forgetting

is rehearsal [5, 8, 32]. Rehearsal-based CL follows a simple idea: a buffer stores samples of old tasks $s < t$. This buffer is then merged with the current task data D_t and used to train the model, reducing forgetting by replaying old knowledge.

Let $\mathcal{M}_t$ be a memory buffer of size M , containing data for tasks $s = 1 \dots t-1$. Without loss of generality, we assume that $\mathcal{M}_t$ is balanced across tasks, i.e., it contains the same amount of samples from each seen dataset $D_1 \dots D_{t-1}$. A model can achieve high performance on the VQA task via fine-tuning and, depending on the content of $\mathcal{M}_t$, avoid forgetting. Once trained for t , the rehearsal strategy updates $\mathcal{M}_t$ to build the buffer for the next learning task $t+1$.

GaB Overview. As rehearsal-based methods incur the need for accessing real data in $\mathcal{M}_t$, which might not be feasible due to *e.g.*, security or privacy constraints, we propose a data-free rehearsal strategy. Our method GaB, as shown in Fig. 2, leverages the language-generation capabilities within a VLM to synthesize pseudo-rehearsal data, *i.e.*, question-answer pairs from prior tasks on current task images. Specifically, we consider f as a VLM which is composed of multiple components: an image encoder f_{ev} and a text encoder f_{el} which project images and text into their respective embedding spaces; A projector $f_{v \rightarrow l}$ projects visual features in the language embedding space; Finally, a language decoder f_{dl} decodes the embeddings into natural language answers $\hat{a}$. With f_{dl} , VLMs have the inherent ability of generating text given a visual input and a textual prompt. This property enables us to build a data-free rehearsal buffer, without storing any data of old tasks.

At task t , GaB on one side of the model sequentially learns the task t -specific projector, with the goal to reconstruct D_t question-answer pairs. On the other, old tasks projectors are employed for the generation of previous question-answer pairs given the current task images (Section 3.1). To align the distribution of generated and real data, we employ a re-sampling strategy (Section 3.2) that leverages the data question type meta-statistics to re-sample pseudo-questions and obtain an aligned replay buffer distribution. Finally, the model leverages the available pseudo-rehearsal data for sequential learning (Section 3.3). Notice that GaB requires task IDs only during training to manage independent and task-specific generation of rehearsal data. A single, shared visual question-answering head is trained using a replay strategy, facilitating backward transfer by integrating knowledge from new tasks, with no task IDs.

3.1. Pseudo-rehearsal data generation

In our continual learning framework, we employ the same VLM model trained for VQA to generate the pseudo-rehearsal data for mitigating the catastrophic forgetting problem. Similarly to the continual learning for the VQA head,

we learn to pose questions and associated answers, by reconstructing the textual information of the data samples.

Learning to generate question-answers. At task t we train a task-specific projection module $f_{v \rightarrow qa}^t$ to generate question answer pairs (q_t^i, a_t^i) given input image x_t^i . Specifically, let f_{qa}^t represent the use of projector $f_{v \rightarrow qa}^t$ within f and operator $\circ$ denote the language concatenation. We can define the question-answer generation objective as:

$$\mathcal{L}_{QA}^t = \frac{1}{N_t} \sum_{(x,a,q) \in D_t} \mathcal{L}_{lm}(f_{qa}^t(x,p), q \circ a), \quad (1)$$

where p is a prompt used to query the model to generate question-answer pairs and the language modelling loss is defined as:

$$\mathcal{L}_{lm}(\hat{y}, y) = \text{CE}(\hat{y}, y), \quad (2)$$

evaluating the token-level cross-entropy loss between predicted text $\hat{y}$ and ground truth one y . With Eq. (1) we obtain a projection module that allows us to generate question and answer pairs for task t given *any* arbitrary image x . Previously learned ones $f_{v \rightarrow qa}^s$ are kept frozen, $s = 1 \dots t-1$.

Construct the pseudo-rehearsal buffer. Now that we have a model that can generate question-answer pairs for each seen task t and arbitrary images, we can use the learned projectors and the images of the current task dataset D_t to generate questions. Specifically, we can generate a set of pseudo question-answer pairs $\bar{D}_i$ for each task $i < t$ as:

$$\bar{D}_i = \{(x_t, \text{split}_\gamma(f_{qa}^i(x_t, p))) \mid (x_t, q_t, a_t) \in D_t\}, \quad (3)$$

where $\text{split}_\gamma(\cdot)$ means we divide the generated text containing the concatenated question and answer based on the position of the symbol "??". Note that q_t, a_t are not used when producing the samples.

From the set of generated datasets $\bar{D}_1 \dots \bar{D}_{t-1}$ we can obtain a buffer $\bar{\mathcal{M}}_t$ by performing balanced sampling across each task. For instance, if the size of $\mathcal{M}_t$ is M , we can sample $M/(t-1)$ random triplets from each pseudo dataset. However, this strategy does not take into account the inherent biases of the dataset on the type of questions and the fact that generative models not only reflect but also exaggerate such biases [28, 59]. In the following, we describe how we can populate $\bar{\mathcal{M}}_t$ by taking into account the ideal distribution of question types across each dataset.

3.2. Pseudo-rehearsal data balancing

Generated data might not reflect the original question-answer distribution and concentrate on most present type of questions/answers, hindering the model generalization capabilities. A significant challenge arises due the tendency of the question generation module to overfit when the training data

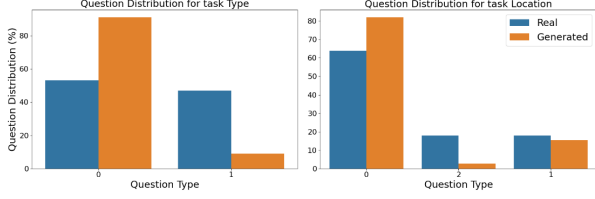


Figure 3. Data-bias present in real vs generated data. Generated questions are heavily skewed towards a certain type of questions. **Left:** question distribution for task *type*. **Right:** question distribution for task *location*.

is unbalanced towards certain question types. This can result in generated question-answer pairs that are skewed towards specific categories, limiting the model’s ability to generalize across the diverse array of questions in the datasets. As can be seen from Fig. 3-left for the VQACL-VQAv2 benchmark on the *type* task, we observe significant disparities in the distribution of question types. On real data, the questions beginning with “what kind” (type 0) constitute 55.1% and those starting with “what type” (type 1) make up 44.9%. However, in the generated data, the distribution is heavily skewed, with “what kind” questions soaring to 90.94%, and “what type” questions drastically reduced to 9.06%. Similarly, the generated questions for the task *location* (Fig. 3-right) under-represent questions of type 2, concentrating to the most frequent type 1 ones. This skew illustrates the potential for substantial shift in question generation, underscoring the need for mechanisms to ensure a more uniform distribution of question types to enhance model robustness.

Mitigating the skewed distribution in generation. We tackle the generated data shift by employing a post-generation alignment with respect to ground truth data. Let $\mathcal{Q}$ be the space of questions. In an initial step, we define a partitioning function $c_t : \mathcal{Q} \rightarrow 1 \dots K_t$ mapping questions $q_t^i \in \mathcal{Q}$ associated to task t into the associated question type $k, k = 1 \dots K_t$. Hence we use the same partitioning function $c_t(\cdot)$ to categorize ground truth data D_t and the generated questions $\bar{D}_t$. We denote as D_t^k the subset of questions in D_t associated to type k , i.e., $D_t^k = \{c_t(q) = k | (x, q, a) \in D_t\}$ and with $\bar{D}_t^k$ the generated counterpart.

From the partitioning, we obtain the expected (ground truth) distribution of question type k as the cardinality of D_t^k w.r.t. the other types, i.e., $p_t^k = |D_t^k| / \sum_{j=1}^{K_t} |D_t^j|$. We can now populate the memory buffer following the obtained distribution. Specifically, let us assume we are at task t and we aim to allocate $\hat{m} = \hat{M} / (t - 1)$ number of samples per task, with $\hat{M} < M$. Given the past task s and the set question type $k, k = 1 \dots K_s$, we take a subset $\hat{\mathcal{M}}_s^k$ of random samples from $\bar{D}_s^k$ whose cardinality is $|\hat{\mathcal{M}}_s^k| = \lceil p_s^k \cdot \hat{m} \rceil$. The final buffer is the union of all old task and types specific buffers, i.e., $\hat{\mathcal{M}}_t = \bigcup_{s=1}^{t-1} \bigcup_{k=1}^{K_s} \hat{\mathcal{M}}_s^k$. In this way, $\hat{\mathcal{M}}_t$ follows the

expected question type distributions, and can be employed as substitute memory buffer without the need for (i) storing past data (just their meta-statistics) and (ii) biasing the model towards specific question types.

Partitioning function. A key component of our approach is the partitioning function $c_j(\cdot)$ used to split real and generated questions into different types. In this paper, we implement two different variants of the partitioning function, denoted as *GaB-classifier* and *GaB-clustering*. GaB-classifier leverages auxiliary meta-information on the question types that is usually available for continual VQA benchmarks. In an initial step, a classifier is trained on task data to predict question types, later on its predictions allow to partition generated data in a coherent manner. As such meta-information can be unavailable, we also presents GaB-clustering which clusters data in an unsupervised manner and uses the obtained centroids to apply a coherent partitioning of the generated data buffer. We evaluate both variants in Sec. 4.1, as well their performance w.r.t. the buffer size in Sec. 4.2.

3.3. Sequential learning with pseudo-rehearsal

At each task t , the VLM jointly learns to replicate question-answer pairs of the current task while training the VQA head to correctly answer asked questions. The data-free pseudo rehearsal buffer avoids forgetting of previous tasks. The objective with the pseudo-rehearsal for continual VQA is:

$$\mathcal{L}_{CL}^t = \frac{1}{N_t + \hat{M}} \sum_{(x, q, a) \in R_t} \mathcal{L}_{lm}(f(x, q), a), \quad (4)$$

where $R_t = D_t \cup \hat{\mathcal{M}}_t$ with $\hat{\mathcal{M}}_1 = \emptyset$, and f denotes the learning VQA model using the answer language projector $f_{v \rightarrow a}$. By accounting for the loss regarding question-answer generation $\mathcal{L}_{QA}^t, t = 1 \dots T$ (defined in Eq. 1), the final loss function at task t can be expressed as:

$$\mathcal{L}^t = \mathcal{L}_{CL}^t + \mathcal{L}_{QA}^t. \quad (5)$$

4. Experiments

We evaluate GaB on two benchmark datasets that feature VQACL, in comparison to state-of-the-art methods. In the following, we first introduce the datasets, the evaluation protocol and implementation details. Then, we discuss the main comparison conducted on the two datasets. Finally, we provide the ablation study to justify our method design.

Datasets. We consider two benchmark datasets, namely VQACL-VQAv2 [57] and CLOVE-function [22]. VQACL-VQAv2 builds on VQA v2 [13] and it constructs a dual-level task sequence where target objects (i.e., the subject of the

questions) and performed tasks are nested together. At the outer level, $T = 10$ tasks are defined based on different skills, *e.g.*, location recognition, general recognition, counting, commonsense, etc. Within each task, skills are applied to a sequence of 5 different object groups, leading to an inner continual learning setting based on target objects. We focus on the standard testing evaluation, where test data comes from the same distribution as training data. CLOVE-function benchmark gathers data from GQA [16], CRIC [12] and TextVQA [37], defining a sequence of 6 tasks about different properties: objects, attributes, relations, logical, knowledge, and scene-text. Note that for the scene-text task, the method proposed in [22] leverages specialized backbone and features for the OCR queries. Thus, to avoid unfair comparison, we leave out the scene-text task, and hence, consider a setting with the remaining $T = 5$ tasks. We refer to the Supp. Mat. for more details.

Performance metrics. Following prior works [40, 57], we evaluate the approach in terms of final model performance and forgetting of previous tasks. Specifically, we measure the average performance (AP) at the end of the CL sequence and the average forgetting (AF) [7, 27] that probes the average model degradation due to learning of subsequent tasks. Formally, let $a_{i,j}$ be the performance of the model trained on task i and evaluated on task j . We define the average performance as $AP = \frac{1}{T} \sum_t a_{T,t} (\uparrow)$, while the average forgetting is evaluated as $AF = \frac{1}{T-1} \sum_t \max_{t \leq i \leq T-1} a_{i,t} - a_{T,t} (\downarrow)$.

Implementation details. We implement our method with BLIP-2 [23], a VLM pre-trained to bridge the vision-language modality gap. Note that our approach can generalize to other VLMs with similar functional components, such as LLaVa [26] or MiniGPT-4 [62]. During adaptation, we freeze the model and limit fine-tuning to language projection layers only. We use Adam optimizer with a fixed learning rate of 1×10^{-4} where the best validation checkpoint is resumed for continual training. The training follows the same hyperparameter settings as respective benchmarks. We consider an initial generated buffer of size $M = 20k$, later downsampled to $\hat{M} = 5k$, where samples are balanced according to features extracted from a frozen `all-mpnet-base-v2`. While GaB-classifier classifies questions according to the available meta-information, for GaB-clustering we select the number of clusters to be 7 and 10 for VQACL-VQAv2 and CLOVE-function respectively, according to validation performance. See Supp. for more implementation details.

4.1. Comparison

Baselines. For fair comparison, we implement all baselines with the BLIP-2 [23] backbone as GaB. We consider the off-the-shelf application of the pre-trained model (`Zero-shot`)

Table 1. The continual learning performance in terms of AP and AF on the VQACL-VQAv2 dataset. All methods use BLIP-2 as their backbone. Best data-free method is highlighted in bold.

Method	Data-free	AP ($\uparrow$)	AF ($\downarrow$)
Zero-shot	-	40.84	-
Rehearsal	$\times$	46.94	4.21
Multi-task	$\times$	49.55	-
VQACL [57]	$\times$	49.80	1.18
Seq-FT	$\checkmark$	41.29	15.98
EWC [20]	$\checkmark$	40.62	9.78
MAS [2]	$\checkmark$	40.27	10.98
L2P [50]	$\checkmark$	40.15	16.56
LAMOL* [42]	$\checkmark$	28.26	6.81
GaB-classifier	$\checkmark$	47.65	3.61
GaB-clustering	$\checkmark$	48.40	1.40

and basic learning strategies, such as naive sequential learning (`Seq-FT`) and the joint training that considers all data available at once (`Multi-task`). In addition, we compare with continual learning regularization strategies, including EWC [20] and MAS [2], a rehearsal strategy replaying previous tasks data (`Rehearsal`) and state-of-the-art prompting strategy L2P [50]. We further adapted the text-based LAMOL (denoted with * for the adaptation) to the data-free VQACL setting, where a single language head leverages current task images to jointly train for question-answer generation and VQA. Finally, we compare GaB with SOTA continual VQA strategies such as VQACL [57] on the VQACL-VQAv2, and SGP [22] on CLOVE-function.

Discussion. Table 1 presents the results on the VQACL-VQAv2 benchmark dataset. Notably, the Zero-shot BLIP-2 shows competitive with continual learning baselines, demonstrating the inherent effectiveness of the BLIP-2 backbone for the task. Our proposed GaB-clustering, which incorporates balancing the rehearsal buffer using clustering, achieves an AP of 48.40, slightly outperforming our classifier-based variant GaB-classifier with 47.65 AP. Intuitively, the classifier controls the alignment of the generated data based on the single-labelled property of the question. On the contrary, clustering-based grouping allows for accounting for different relevant aspects of the questions. Surprisingly, these results surpass the rehearsal strategy involving real data samples. The recent method VQACL and the Multi-task baseline are only marginally better than GaB, despite having access to past tasks real data. Our methods effectively reduce knowledge loss, as demonstrated by the lower AF rate of 1.40 for GaB-clustering compared to the 4.21 AF rate of the Rehearsal method. Under the same memory constraint, LAMOL* achieves a low 28.26% AP, indicating its unsuitability for the setting. Further inspection on generated questions shows that the single language projection layer fails to effectively rehearse question generation

Table 2. The continual learning performance in terms of AP and AF on the CLOVE-function dataset. Best data-free method in bold.

Method	Backbone	Data-free	AP ($\uparrow$)	AF ($\downarrow$)
Zero-shot	BLIP-2	-	24.47	-
Rehearsal	BLIP-2	$\times$	41.82	3.14
Multi-task	BLIP-2	$\times$	32.26	-
SGP [22]	UniVQA	$\checkmark$	27.72	38.16
Seq-FT	BLIP-2	$\checkmark$	22.70	22.19
EWC [20]	BLIP-2	$\checkmark$	23.90	10.14
MAS [2]	BLIP-2	$\checkmark$	26.30	6.46
L2P [50]	BLIP-2	$\checkmark$	16.86	16.02
LAMOL* [42]	BLIP-2	$\checkmark$	15.26	10.46
GaB-classifier (Ours)	BLIP-2	$\checkmark$	38.34	2.85
GaB-clustering (Ours)	BLIP-2	$\checkmark$	36.57	4.61

for old tasks. In line with recent findings [40], continual learning strategies are ineffective in managing the trade-off between adaptation to the different tasks and preservation of original knowledge, as showed by the lower performance with respect to the zero-shot evaluation. On one hand, L2P mitigates the catastrophic forgetting on the vision side only, without accounting for the multimodal nature of the problem. On the other, regularization approaches fail to estimate the relevance of weights. We hypothesize the text generation head does not provide a suitable space to estimate how knowledge is stored. Even the best performing regularization method, EWC, only achieves 40.62 in terms of AP.

Table 2 reports the results on *CLOVE-function* benchmark. CLOVE-function raises different challenges with respect to VQACL-VQAv2, such as the more diverse set of tasks, the compositional nature of questions, and the needed auxiliary information required for the knowledge task. Overall, we observe similar trends as in Table 1. Large pre-trained models benefit the final task performance in the continual learning setting, as evidenced by the strong performance of the zero-shot BLIP-2 model, compared with the state-of-the-art specialized architecture UniVQA. GaB achieves 38.34 in terms of AP, with the cluster variant being closely behind at 36.57 AP. Both variants of our GaB approach achieve the best performance among data-free methods, only lagging behind the real rehearsal method by a mere 3%. Our model significantly bridges the gap between rehearsal and data-free approaches, requiring a minimal reliance on real data from past tasks. Moreover, our model demonstrates a strong capability in handling diverse and complex question types, typical of CLOVE-function, as demonstrated by the 32.26 AP of the Multitask baseline. We conjecture the negative transfer across tasks hinders the model generalization capabilities. GaB is effective especially in challenging relational and logic-based queries (see the Supp. Mat. for extended results on all tasks). Interestingly, while EWC achieves a performance similar to the Seq-FT baseline at 22.70 AP, MAS yields a 26.60 AP proving the most effective approach among the classic continual learning strategies. L2P instead

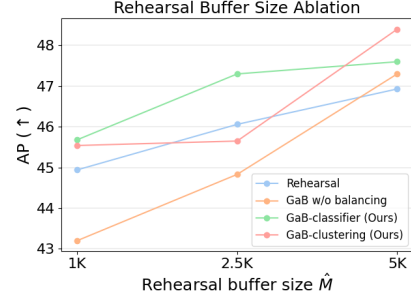


Figure 4. Analysis of balancing strategies at varying rehearsal buffer sizes (1k, 2.5k, and 5k samples) on the VQACL-VQAv2 benchmark in terms of AP.

confirms the unsuitability of the approach to the multimodal setting [40].

4.2. Ablation studies

We study the different components and design choices of GaB, by analyzing the impact of various balancing strategies, question category conditioning, pseudo-labelling and the images used to populate the memory. We perform the ablation with VQACL-VQAv2 and analyze the results below.

Balancing strategies and the buffer size. To assess the impact of our balancing strategy, we experiment with a variant, referred as GaB w/o balancing, where we remove the balancing component. We also analyze the impact of varying sizes of the rehearsal buffer on the variants of our methods. As depicted in Figure 4, both the GaB-classifier and GaB-clustering methods improve performance with increasing buffer sizes. The results show that our balancing strategies significantly outperform the scenarios where no balancing is applied, the gain is consistent across all considered buffer sizes. Notably, with the small-sized buffers, GaB-classifier and GaB-clustering yields a similar performance, providing a benefit with respect to the rehearsal strategy with real data. As the buffer size increases to $\hat{M} = 5k$ samples, GaB-clustering emerges as the most effective, achieving the highest AP, highlighting the scalability of the clustering approach under conditions of larger data availability.

Question category conditioning. GaB generates questions freely, based on the image content. As we discussed in Sec. 3.2, this results in a skewed distribution towards more prevalent question types and we need to fix this issue via a post-generation balancing strategy. A different direction is to avoid this problem altogether by conditioning the question generation process on target question types. We study this alternative strategy, and compare the proposed approach with GaB-conditioning where the language model is trained by explicitly prompting to generate

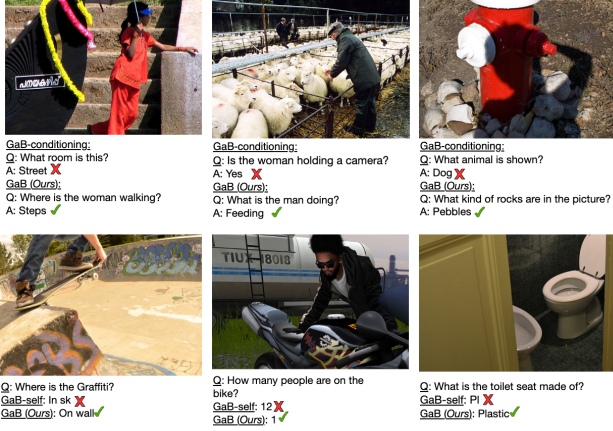


Figure 5. Qualitative visualization of the generated pseudo-rehearsal data on the VQACL-VQAv2 benchmark. **Top row:** analysis on generation conditioning for questions balancing. **Bottom row:** analysis on the pseudo-strategy used for answer generation.

questions based on their type, $p = \text{"Question type"} \langle \text{type} \rangle$: Question: $\langle \text{question} \rangle$ Answer: $\langle \text{answer} \rangle$ ". During inference, question generation is similarly driven by the classifier partitioning function to ensure a balanced distribution. Table 3 presents the results of this conditioning approach in comparison to our methods involving only post-generation processing. We observe that the results of GaB-conditioning are inferior to ours, going from 47.65 to 44.07 in terms of AP. Upon further inspection, we find that forcing the model to produce questions related to certain types can lead to false assumptions, such as forcing to ask a question type that is not relevant for the given image (see Fig. 5 for illustration).

Pseudo-labelling strategy. In our method, rehearsal generation simultaneously produces question-answer pairs based on image inputs. Alternatively, we explored using our trained answer generation model to respond to these questions separately in a self-training manner (GaB-self). As detailed in Table 3, this method resulted in a significant decrease in AP by approximately 2.65, likely due to the inferior quality of the pseudo-answers, which are often noisy or incomplete. These issues may cause confirmation bias during fine-tuning. For illustrative examples, see the bottom row of Fig. 5.

Using past visual images. In this study, we evaluate the effectiveness of generating question-answer (QA) pairs with images from the current versus previous tasks. Our goal is to discern why our rehearsal strategy, which uses current task images for QA generation, surpasses traditional methods utilizing real-data. In Table 3, we demonstrate with GaB-pastimages that applying our QA generation strategy to past task images results in a performance decrease of

Table 3. Analysis on the design choices of GaB in terms of AP and AF on the VQACL-VQAv2 dataset.

Method	AP($\uparrow$)	AF($\downarrow$)
<i>Generation conditioning</i>		
GaB-conditioning	44.07	6.45
<i>Pseudo strategy</i>		
GaB-self	45.01	4.76
<i>Rehearsal on Task data</i>		
GaB-pastimages	38.78	4.07
GaB-classifier (Ours)	47.65	3.61

about 10% compared to using current task images, as shown in GaB-classifier.

We believe that using current task images for QA pair generation offers enhanced contextual relevance, similar to data augmentation. This approach not only provides more context by posing multiple questions on the same images but also helps the model bridge knowledge gaps across tasks, effectively serving as a form of positive regularization. Consequently, the model better assimilates new information with previously acquired knowledge, improving its ability to maintain relevant capabilities across different tasks.

5. Conclusion

We presented GaB, a novel data-free method for VQACL building on top of the inherent generative capability of VLMs. With current task images, GaB synthesizes data samples of previous VQA tasks by learning to generate visually conditioned question-answer pairs. We tackled the model generation issue where the questions are skewed to the most-occurring questions, by balancing the generated data according to real data statistics or patterns. Results on standard benchmarks, including VQACL-VQAv2 and CLOVE-function, prove the benefit of GaB against existing data-free methods. As GaB depends on the data generated for rehearsal, noisy or ambiguous pseudo-samples can hinder the model performance. Future work will focus on improving generated data quality, introducing techniques such as hallucination mitigation and automated quality control.

6. Acknowledgment

The authors acknowledge the CINECA award under the ISCRA initiative for the availability of high-performance computing resources and support. Deepayan Das is supported by the PRIN project "B-FAIR" (Prot.2022EXF3HX) and the PAT project "AI@TN". This work was supported by the projects EU Horizon ELIAS (No. 101120237), AI4TRUST (No.101070190), FAIR Future AI Research (PE00000013), funded by NextGeneration EU, and carried out in the Vision and Learning joint laboratory of Fondazione Bruno Kessler and the University of Trento, Italy

References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 2022. [2](#)
- [2] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision*, 2018. [2](#), [6](#), [7](#)
- [3] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. Expert gate: Lifelong learning with a network of experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. [2](#)
- [4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2015. [1](#)
- [5] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in Neural Information Processing Systems*, 2020. [2](#), [4](#)
- [6] Xusheng Cao, Haori Lu, Linlan Huang, Xialei Liu, and Ming-Ming Cheng. Generative multi-modal models are good class incremental learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. [2](#)
- [7] Arslan Chaudhry, Marc’Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. In *International Conference on Learning Representations*, 2018. [6](#)
- [8] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K. Dokania, Philip H. S. Torr, and Marc’Aurelio Ranzato. On tiny episodic memories in continual learning, 2019. [2](#), [4](#)
- [9] Yoojin Choi, Mostafa El-Khamy, and Jungwon Lee. Dual-teacher class-incremental learning with data-free generative replay. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. [2](#)
- [10] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 2024. [2](#)
- [11] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999. [2](#)
- [12] Difei Gao, Ruiping Wang, Shiguang Shan, and Xilin Chen. Cric: A vqa dataset for compositional reasoning on vision and commonsense. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5561–5578, 2022. [6](#)
- [13] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. [1](#), [5](#)
- [14] Claudio Greco, Barbara Plank, Raquel Fernández, and Raffaella Bernardi. Psycholinguistics meets continual learning: Measuring catastrophic forgetting in visual question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3601–3605, 2019. [2](#)
- [15] Tyler L Hayes, Kushal Kafle, Robik Shrestha, Manoj Acharya, and Christopher Kanan. Remind your neural network to prevent catastrophic forgetting. In *Proceedings of the European Conference on Computer Vision*, 2020. [2](#), [3](#)
- [16] Drew A Hudson and Christopher D Manning. 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*, 2019. [1](#), [6](#)
- [17] Saurav Jha, Dong Gong, and Lina Yao. Clap4clip: Continual learning with probabilistic finetuning for vision-language models. *arXiv preprint arXiv:2403.19137*, 2024. [2](#)
- [18] Heechul Jung, Jeongwoo Ju, Minju Jung, and Junmo Kim. Less-forgetting learning in deep neural networks. *arXiv preprint arXiv:1607.00122*, 2016. [2](#)
- [19] Junsu Kim, Yunhoe Ku, Jiyeon Kim, Junuk Cha, and Seungryul Baek. Vlm-pl: Advanced pseudo labeling approach for class incremental object detection via vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. [2](#)
- [20] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. [2](#), [6](#), [7](#)
- [21] Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Information maximizing visual question generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. [2](#), [3](#)
- [22] Stan Weixian Lei, Difei Gao, Jay Zhangjie Wu, Yuxuan Wang, Wei Liu, Mengmi Zhang, and Mike Zheng Shou. Symbolic replay: Scene graph as prompt for continual learning on vqa task. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. [2](#), [3](#), [5](#), [6](#), [7](#)
- [23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023. [1](#), [2](#), [6](#)
- [24] Xilai Li, Yingbo Zhou, Tianfu Wu, Richard Socher, and Caiming Xiong. Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In *International Conference on Machine Learning*, 2019. [2](#)
- [25] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017. [2](#)
- [26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 2024. [1](#), [2](#), [6](#)
- [27] David Lopez-Paz and Marc’Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in Neural Information Processing Systems*, 2017. [6](#)
- [28] Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. In *Proceedings of the 2023*

- AAAI/ACM Conference on AI, Ethics, and Society, 2023. 4
- [29] Mavina Nikandrou, Lu Yu, Alessandro Suglia, Ioannis Konstantas, and Verena Rieser. Task formulation matters when learning continually: A case study in visual question answering. *arXiv preprint arXiv:2210.00044*, 2022. 2
- [30] Zi Qian, Xin Wang, Xuguang Duan, Pengda Qin, Yuhong Li, and Wenwu Zhu. Decouple before interact: Multi-modal prompt learning for continual visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2
- [31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 2
- [32] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 2, 4
- [33] Anurag Roy, Vinay K Verma, Sravan Voonna, Kripabandhu Ghosh, Saptarshi Ghosh, and Abir Das. Exemplar-free continual transformer with convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2
- [34] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016. 2
- [35] Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *International Conference on Machine Learning*, 2018. 2
- [36] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in Neural Information Processing Systems*, 2017. 2
- [37] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 6
- [38] Vaibhav Singh, Rahaf Aljundi, and Eugene Belilovsky. Controlling forgetting with test-time data in continual learning. *arXiv preprint arXiv:2406.13653*, 2024. 2
- [39] James Smith, Yen-Chang Hsu, Jonathan Balloch, Yilin Shen, Hongxia Jin, and Zsolt Kira. Always be dreaming: A new approach for data-free class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2
- [40] James Seale Smith, Paola Cascante-Bonilla, Assaf Arbelle, Donghyun Kim, Rameswar Panda, David Cox, Diyi Yang, Zsolt Kira, Rogerio Feris, and Leonid Karlinsky. Construct-vl: Data-free continual structured vl concepts learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 3, 6, 7
- [41] Tejas Srinivasan, Ting-Yun Chang, Leticia Pinto Alva, Georgios Chochlakis, Mohammad Rostami, and Jesse Thomason. Climb: A continual learning benchmark for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 2022. 2
- [42] Fan-Keng Sun, Cheng-Hao Ho, and Hung-Yi Lee. Lamol: Language modeling for lifelong language learning. *International Conference on Learning Representations*, 2019. 3, 6, 7
- [43] Yu-Ming Tang, Yi-Xing Peng, and Wei-Shi Zheng. Learning to imagine: Diversify memory for incremental learning using unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2
- [44] Rishabh Tiwari, Krishnateja Killamsetty, Rishabh Iyer, and Pradeep Shenoy. Gcr: Gradient coreset based replay buffer selection for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2
- [45] Paul Voigt and Axel Bussche. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. 01 2017. 2
- [46] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2
- [47] Runqi Wang, Xiaoyue Duan, Guoliang Kang, Jianzhuang Liu, Shaohui Lin, Songcen Xu, Jinhu Lü, and Baochang Zhang. Attriclip: A non-incremental learner for incremental knowledge learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2
- [48] Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. S-prompts learning with pre-trained transformers: An occam’s razor for domain incremental learning. *Advances in Neural Information Processing Systems*, 2022. 2
- [49] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *Proceedings of the European Conference on Computer Vision*, 2022. 2
- [50] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 6, 7
- [51] Yicheng Xu, Yuxin Chen, Jiahao Nie, Yusong Wang, Huiping Zhuang, and Manabu Okumura. Advancing cross-domain discriminability in continual learning of vision-language models. *arXiv preprint arXiv:2406.18868*, 2024. 2
- [52] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deep-inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2
- [53] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. In *International Conference on Learning Representations*, 2018. 2
- [54] Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vi-*

- sion and Pattern Recognition, 2024. 1, 2
- [55] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, 2017. 2
 - [56] Wenxuan Zhang, Paul Janson, Rahaf Aljundi, and Mohamed Elhoseiny. Overcoming general knowledge loss with selective parameter finetuning. 2023. 2
 - [57] Xi Zhang, Feifei Zhang, and Changsheng Xu. Vqacl: A novel visual question answering continual learning setting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 5, 6
 - [58] Yao Zhang, Haokun Chen, Ahmed Frikha, Yezi Yang, Dennis Krompass, Gengyuan Zhang, Jindong Gu, and Volker Tresp. Cl-crossvqa: A continual learning benchmark for cross-domain visual question answering. *arXiv preprint arXiv:2211.10567*, 2022. 2, 3
 - [59] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Conference on Empirical Methods in Natural Language Processing*, 2017. 4
 - [60] Junhao Zheng, Qianli Ma, Zhen Liu, Binqun Wu, and Huawei Feng. Beyond anti-forgetting: Multimodal continual instruction tuning with positive forward transfer. *arXiv preprint arXiv:2401.09181*, 2024. 2
 - [61] Da-Wei Zhou, Yuanhan Zhang, Jingyi Ning, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Learning without forgetting for vision-language models. *arXiv preprint arXiv:2305.19270*, 2023. 2
 - [62] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*. 6