VQACL: A Novel Visual Question Answering Continual Learning Setting

Xi Zhang[1,2], Feifei Zhang[4], Changsheng Xu[1,2,3]

[1]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences

[2]School of Artificial Intelligence, University of Chinese Academy of Sciences [3]Peng Cheng Laboratory

[4]School of Computer Science and Engineering, Tianjin University of Technology

zhangxi2019@ia.ac.cn, feifeizhang@email.tjut.edu.cn, csxu@nlpr.ia.ac.cn

Abstract

Research on continual learning has recently led to a variety of work in unimodal community, however little attention has been paid to multimodal tasks like visual question answering (VQA). In this paper, we establish a novel VQA Continual Learning setting named VQACL, which contains two key components: a dual-level task sequence where visual and linguistic data are nested, and a novel composition testing containing new skill-concept combinations. The former devotes to simulating the ever-changing multimodal datastream in real world and the latter aims at measuring models' generalizability for cognitive reasoning. Based on our VQACL, we perform in-depth evaluations of five well-established continual learning methods, and observe that they suffer from catastrophic forgetting and have weak generalizability. To address above issues, we propose a novel representation learning method, which leverages a sample-specific and a sample-invariant feature to learn representations that are both discriminative and generalizable for VQA. Furthermore, by respectively extracting such representation for visual and textual input, our method can explicitly disentangle the skill and concept. Extensive experimental results illustrate that our method significantly outperforms existing models, demonstrating the effectiveness and compositionality of the proposed approach. The code is available at https://github.com/zhangxi1997/VQACL.

1. Introduction

Continual learning [43] has recently gained a lot of attention in the deep learning community because it enables models to learn continually on a sequence of non-stationary tasks and is close to the human learning process [2, 36]. However, the vibrant research in continual learning mainly focuses on unimodal tasks such as image classification [37, 46, 51] and sequence tagging [4, 48], and the demand of multimodal tasks is ignored. In recent years, the volume of multimodal data has grown tremendously [8, 56, 57]. For example, tens of millions of texts, images, and videos are uploaded to social media platforms every day, such as Facebook and Twitter. To cope with such constantly emerging real-world data, a practical AI system should be capable of continually learning from multimodal sources while alleviating forgetting previously learned knowledge.

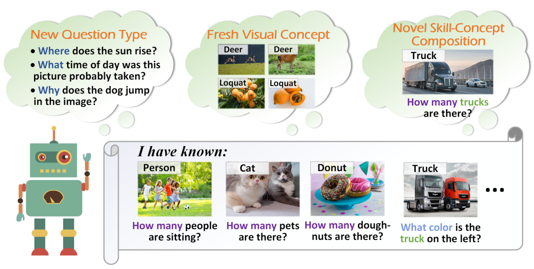


Figure 1. The illustration of real-world scenario for VQA system, which may continuously receive new types of questions, fresh visual concepts, and novel skill-concept compositions.

Visual Question Answering (VQA) is a typical multimodal task and has drawn increasing interest over the past few years [12, 49, 60], which can automatically generate a textual answer given a question and an image. To deal with ever-changing questions and visual scenes in real life, applying continual learning to VQA is essential. However, it is not easy to set up a suitable continual learning setting for this task. We identify that two vital issues need to be considered. First, the VQA input comes from both vision and linguistic modalities, thus the task setting should simultaneously tackle continuous data from both modalities in a holistic manner. For example, as shown in Fig. 1, the AI system might deal with new types of questions (e.g., *Where ...*, *Why ...*) as well as fresh visual concepts (e.g., *Loquat*, *Deer*). Second, compositionality [24], a vital property of cognitive reasoning, should be considered in the VQA continual learning. The compositionality here denotes the model's generalization ability towards novel combinations of **reasoning skills** (i.e., question type) and **visual concepts** (*i.e.*, image object). As illustrated in Fig. 1, if the system has been trained with question type *Count* (e.g., *How many*) with a variety of objects (e.g., *Person, Cat, and Dount*), as well as another question type (e.g., *What color*) about a new object (e.g., *Truck*). Then, it is expected to answer a novel question like *'How many trucks are there?'*, even if the composition of skill *Count* and concept *Truck* has yet to be seen. Such ability is very crucial when deploying a model in the real world because it is infeasible to view all possible skill-concept compositions. Remarkably, several works has addressed continual learning with VQA [14, 16, 25]. However, they still apply a classic unimodal-like continual learning setting for the task by devising a set of VQA tasks simply based on question type or image scene, which ignores above two crucial issues: handling continuous multimodal data simultaneously and testing model's compositionality.

To achieve these two keypoints, in this paper, we propose a novel generative VQA continual learning setting named VQACL based on two well-known datasets: VQA v2 [13] and NExT-QA [49]. Specifically, as shown in Fig. 2(a), our VQACL setting consists of a dual-level task sequence. In the outer level, we set up a sequence of linguistic-driven tasks to evaluate models' ability for the ever-changing question types. Moreover, to process the continuously shifted visual contents, for each outer level task, we further construct a series of randomly ordered visual-driven subtasks according to image object categories in the inner level. Such dual-level setting is similar to the cognitive process of children, who master a skill by trying it on various objects. For example, when learning to recognize colors, a baby usually asks all the things surrounding him '*what color*' they are. Besides, to evaluate models' compositionality, we construct a novel composition split. As shown in Fig 2(b), we remove a visual-driven subtask from each task in the outer level during training and utilize it for testing. In this way, the testing data contain novel skill-concept combinations that are not seen at the training time. In conclusion, on the one hand, our VQACL setting requires models to perform effective multimodal knowledge transfer from old tasks to new tasks while mitigating catastrophic forgetting [31]. On the other hand, the model should be capable of generalizing to novel compositions for cognitive reasoning.

Using the proposed VQACL setting, we establish an initial set of baselines by adapting several well-known and state-of-the-art continual learning methods [1, 3, 7, 22, 45] from image classification to the generative VQA tasks. The baselines are implemented on an advanced vision-and-language transformer [9] without pre-training. After benchmarking these baseline models, we find that few of them can do well in the novel composition testing, which limits their wide applications in practice. To enhance the model's compositionality, it is critical to learn an excellent representation that is discriminative for seen skills or concepts, and is generalizable to novel skill-concept compositions. To achieve it, recent static VQA methods [27, 47, 59] always first learn joint representations for visual and textual inputs, and then utilize contrastive learning to implicitly disentangle the skill and concept within the joint feature. However, such implicit disentangling makes existing models still dogged by the interference between the skill and concept, leading to suboptimal generalization results. Moreover, the complex contrastive sample building process makes these works tough to be applied to continual learning.

Inspired by above discussions, we propose a novel representation learning method for VQACL, which introduces a sample-specific (SS) and a sample-invariant (SI) feature to learn better representations that are both discriminative and generalizable. To explicitly decouple the reasoning skills and visual concepts, we learn the SS and SI representation for visual and textual input separately. Specifically, the SS feature for each modality is learned through a transformer encoder that stacks multiple self-attention layers, which can encode the most attractive and salient contents into the SS feature to make it discriminative. For the SI feature, we resort to prototype learning to aggregate the object class or question type information into it. Because the category knowledge is stable and representative across different scenarios, the SI feature can possess strong generalizability. Besides, to fit the continual learning setting, we constantly update the SI feature in training. In this way, it can capture new typical knowledge while retaining historical experience, helping alleviate the forgetting problem. In conclusion, combining the SS and SI features, we can obtain the representation that is conducive to the model's compositional discriminability and generalizability.

In summary, the major contributions of our work are threefold: (1) We introduce a new continual learning setting VQACL to simulate real-world generative VQA. It can not only simultaneously tackle the continuous data from vision and linguistic modality, but also test models' compositionality for cognitive reasoning. (2) We propose a simple but effective representation learning method for continual VQA, which novelly deploys a discriminative sample-specific feature and a generalizable sample-invariant feature to alleviate forgetting and enhance the models' composition ability. (3) We re-purpose and evaluate five well-established methods on our VQACL, and observe that they struggle to obtain satisfactory results. Remarkably, our model consistently achieves the best performance, demonstrating the effectiveness and compositionality of our approach.

2. Related Work

2.1. Visual Question Answering

Visual question answering (VQA) has gained much attention in AI, which requires co-reasoning over both visual and textual input to automatically generate a correct answer. These years, various approaches have been proposed for this task [50, 52, 54, 55, 58–61], which mainly focus on exploiting attention mechanism and multimodal fusion techniques. Recently, mirroring the success of language transformers [19], vision-language transformers have achieved remarkable success in VQA [9, 35, 62]. For example, Cho et al. [9] propose a generative transformer to do VQA, which performs answer generation based on image objects and question words. Nevertheless, most existing methods are designed without explicitly considering the generalization ability, thus having limited *compositionality*. As discussed in [20, 24], *compositionality* is an ability to systematically understand and generalize to novel combinations of known components, which is critical for cognitive reasoning.

In recent years, researchers have begun to explore the composition issue in VQA [17,18,27,47,59]. For example, Johnson et al. [18] study the composition of visual attributes (e.g., color, size) and objects (e.g., cube, cylinder) and propose a dataset for compositional reasoning. More similar to us, Whitehead et al. [47] also investigate the composition of reasoning skills and visual concepts, and leverage contrastive learning to implicitly disentangle the skill and concept in a joint feature to enhance the model's compositionality. However, the implicit decoupling may lead to suboptimal generalization performance, and the contrastive sample building process is complex. In contrast, our work explicitly decouples the skill and concept through separately learning sample-specific and sample-invariant features for the textual and visual input, which can make the learned representation more discriminative and generalizable. Besides, these existing VQA models perform offline training and ignore the demand for tackling continuous multimodal data in practice. Differently, we apply continual learning to VQA and train the model with a sequential series of tasks, which is more consistent with real world applications.

2.2. Continual Learning

Continual learning aims to train a single model that can incrementally update knowledge with a new stream of tasks while preserving previously learned information [10]. The major challenge is to learn without catastrophic forgetting [10]: the model's performance on previously learned tasks should not significantly degrade over time. To overcome the challenge, existing continual learning algorithms can be categorized into regularization, rehearsal, and architectural methods. Specifically, the regularization methods [1, 22, 33, 40] impose a regularization constraint to the objective to limit parameter changes. The rehearsal-based methods [3,6,7,42,44] store some training examples of previous tasks in a memory buffer, and retrain the model on old data to review past knowledge. Differently, the architectural approaches [26, 39, 53] dynamically expand the network to learn specific parameters for each task. Although these methods have shown remarkable results in unimodal tasks such as image classification and sequence tagging, their use within multimodal tasks remains under-explored.

Recently, a number of work has shown interest in multimodal continual learning [11, 14, 32, 41]. For example, Del et al. [11] consider continual image captioning with LSTM-based models. More similarly, several works [14, 16, 25] introduce a continual VQA setting that is composed of a sequence of tasks with different question types, and [25] also designs a setting containing VQA tasks with different image scenes. However, they cannot simultaneously tackle multimodal continuous data and ignore the essential composition generalization issue for VQA. Differently, we propose the VQACL, a more challenging and realistic setting for generative VQA continual learning. Specifically, our VQACL consists of a dual-level task sequence to tackle the multimodal data, where the outer level setups sequential linguistic-driven tasks with different question types and the inner level builds serial visual-driven subtasks with shifting object categories. Besides, we design a novel composition testing to further evaluate the model's compositionality. Based on the VQACL, we also propose a rehearsal-based representation learning method to boost the continual VQA performance and alleviate the forgetting problem.

3. VQA Continual Learning Setting

In this section, we introduce our proposed generative VQA Continual Learning setting (VQACL), which aims to test the learning algorithm's ability to adapt to a sequentially arriving datastream of VQA.

3.1. Problem Definition

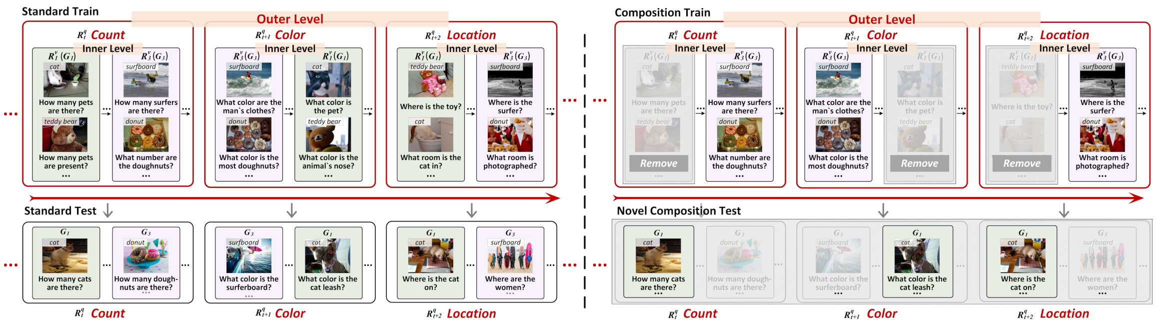
In our work, we formulate the VQA as a generation task, which aims to generate textual answers automatically given an image and a question. Unlike traditional offline training that the model can visit entire training data, we focus on a continual learning setup, where the model visits a non-stationary stream of the data. Specifically, we optimize a single neural network over a sequence of VQA tasks, and search the parameters that can maximize the average VQA performance. Each VQA task contains its own training data.

We do continual VQA on two standard datasets: VQA v2 [13], an image QA dataset with 1.1 Million pairs of real-world images and human-written questions; and NExT-QA [49], a video QA dataset with $52K$ manually annotated question-answer pairs. In the following, we introduce the building details of the continual learning setting for VQA.

3.2. VQACL

Continual VQA comes with two unique requirements: (1) the setting should be capable of tackling continuous data from both vision and linguistic modality; and (2) the setup is expected to evaluate models' generalizability on novel skill-concept composition. Informed by these issues, we design the VQACL setting as follows.

Dual-level task sequence. Inspired by the cognitive process of baby, we design a dual-level task sequence, where the visual and textual data are nested to construct continuous datastream. Specifically, the standard training and testing process is shown in Fig. 2(a). In the **outer** level, we define a series of linguistic-driven tasks $\{R_1^q, ..., R_T^q\}$, where T denotes the number of the task, and each task corresponds



(a) Standard process

(b) Composition process

Figure 2. The proposed VQACL setting. (a) Standard training and testing. (b) Composition training and novel composition testing. The data covered with the gray box denotes that it is removed.

to learning a specific reasoning skill. For example, for the '*Count*' task illustrated in Fig. 2(a), its training data mainly contain the examples that can teach the model how to count, such as '*How many*' and '*What number is*'. According to the question types in the dataset, we define $T = 10$ for VQA v2 and $T = 8$ for NExT-QA. Detailed information can be found in the supplementary material. In the **inner** level, each linguistic-driven task is further composed of a sequence of visual-driven subtasks $\{R_1^v, ..., R_K^v\}$, where each subtask R_k^v contains the images from object group G_k. Specifically, we uniformly partition all the object classes $\{c_i\}_{i=1}^C$ into K parts to obtain the $\{G_k\}_{k=1}^K$, which are then randomly assigned to different visual subtasks. For both VQA v2 and NExT-QA, the K is set to 5, and the class number C is set to 80 according to COCO [28].

Novel Composition Testing. Compositionality is an important property in cognitive reasoning, which is crucial in real-world scenarios. To this end, in VQA continual learning, besides the common stability-plasticity issue, our VQACL also focuses on measuring the model's compositionality of the reasoning skill (e.g., *Count*, *Color*) and visual concept (e.g., *Cat*, *Surfboard*). To achieve it, based on the standard process in Fig. 2(a), a composition training and testing process is built and shown in Fig. 2(b). Specifically, we randomly remove a visual-driven subtask R_k^v from each linguistic-driven task during training and utilize it as the novel compositions for testing. As a result, our testing data involve unseen combinations that consist of image objects in R_k^v and each question type. Besides, to guarantee that the elements contained in the novel compositions have been seen before, we train the model in the first linguistic-driven task with all the visual objects. To make the testing results more convincing, we perform K-fold object independent cross-validation. In detail, we repeat the above process for K times and each time remove a different visual-driven subtask, so that all the objects could fairly appear.

In conclusion, under our VQACL setting, the model requires to not only minimize the forgetting of multimodal tasks seen earlier in training, but also facilitate generalizable knowledge transfer to improve performance on constantly emerged skills, concepts, and skill-concept compositions.

3.3. Evaluation Metrics

In the VQACL setting, we use two standard continual learning metrics [5, 6, 29]: Final Average performance (i.e., *AP*) and Average Forgetting (i.e., *Forget*). Specifically, the *AP* is the average performance of the model for all learned tasks, which shows the model's capability when continually learning new tasks. Suppose $a_{i,j}$ is the testing performance on task R_i^q when the model completes learning task R_j^q, $AP = \frac{1}{T}\sum_{t=1}^{T} a_{t,T}$. Besides, the *Forget* measures performance degradation in subsequent tasks and is defined by $Forget = \frac{1}{T-1}\sum_{t=1}^{T-1} max_{z\in\{t,...,T-1\}}(a_{t,z} - a_{t,T})$. For a fair comparison, we compute $a_{i,j}$ in NExT-QA following [49], and use Wu-Palmer similarity (WUPS) [30] to evaluate the quality of generated answer. In VQA v2, following [9], we leverage the percentage of correctly answered questions as the $a_{i,j}$.

4. Proposed Method

4.1. Overall Architecture

We present a simple but effective representation learning approach to enhance the model's compositionality in our VQA continual learning setting (VQACL). In our method, for both vision and linguistic modality, a sample-specific (SS) and a sample-invariant (SI) feature are introduced to help learn a discriminative and generalizable representation for the VQACL. The architecture of our model is shown in Fig. 3, which adopts a transformer encoder-decoder network [9] as the backbone and includes a prototype learning module. Besides, following the common rehearsal methods [7, 29], to alleviate the catastrophic forgetting in continual learning, we also construct a memory buffer $\mathcal{M}$, which stores randomly selected training examples from each past task. As shown in Fig. 3, given an image V and a question Q from either the current task or memory $\mathcal{M}$, we first extract the visual embeddings E^v and textual embeddings E^q. Then, E^v and E^q are fed into the transformer encoder to capture attractive and salient contents in V and Q, thus making the output features discriminative. The features are then adopted as the visual and textual sample-specific

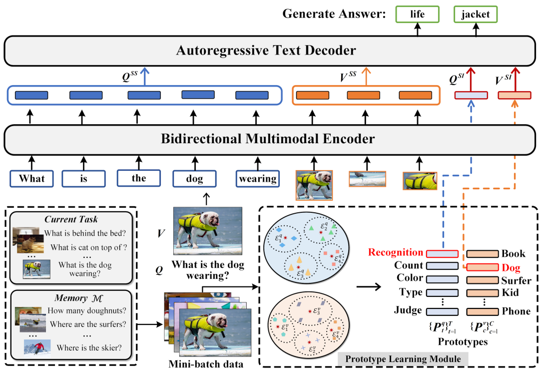


Figure 3. The overall architecture of our proposed method, which incorporates a transformer backbone, a memory buffer, and a prototype learning module.

features V^{SS} and Q^{SS} in our method. In the prototype-learning module, we learn and update prototypes of different question types and object classes. Since the prototype can aggregate typical category information that is robust to novel data, we select suitable visual and textual prototypes as the sample-invariant feature V^{SI} and Q^{SI} based on the V and Q. Finally, the vectors V^{SS}, Q^{SS}, V^{SI}, and Q^{SI} are combined and fed into the text decoder to generate the answer. A conventional negative log-likelihood loss is leveraged to optimize the whole network.

4.2. Visual and Textual Embedding

Given n object regions for image V, each region is encoded as a sum of three types of features: (1) region feature, (2) region bounding box coordinates, and (3) region $id \in \{1, ..., n\}$. Specifically, the region feature and bounding box coordinates are encoded by a linear layer, and the region id is encoded with learned embeddings [19]. In this way, we obtain visual embedding $E^v \in \mathbb{R}^{n\times d}$ for V, where d is the dimension of the embedding. For question Q, we first tokenize it to words and then encode them as textual embedding $E^q \in \mathbb{R}^{m\times d}$ through an embedding layer, where m is the number of words.

4.3. Sample-specific and Sample-invariant Representation Learning

A VQA model with good compositionality should be equipped with two capabilities: discriminative for seen question types or image objects, and generalizable to novel combinations of them. We believe that the key is to perform effective representation learning. To achieve this, we propose a simple but effective representation learning method by leveraging a sample-specific and a sample-invariant feature. In this way, the representation learned by our method contains not only prominent contents of the input, but also representative category knowledge.

Sample-specific Feature. To learn a discriminative SS feature, we utilize a bidirectional multimodal encoder $Enc(\cdot)$ that consists of a stack of transformer blocks. Specifically, each transformer block contains a multi-head self-attention layer and a fully-connected layer with residual connections, which helps capture the most attractive and prominent feature of the input. Formally, the SS feature $Q^{ss} \in \mathbb{R}^{n\times d}$ and $V^{ss} \in \mathbb{R}^{m\times d}$ for the question and image are encoded as:

$$Q^{ss}, V^{ss} = Enc(E^q, E^v). \tag{1}$$

Sample-invariant Feature. For the SI feature, we hope it contains typical reasoning knowledge for a type of question, or common attribute information for a class of image, which is invariant across different domains and can be adapted to novel scenarios. To achieve it, we design a prototype learning module to construct prototypes for different kinds of questions and objects, and each prototype aggregates representative category information of corresponding training examples. Specifically, we first initialize a set of question prototypes $\{P_t^q\}_{t=1}^T$ and object prototypes $\{P_c^v\}_{c=1}^C$, where $P_t^q, P_c^v \in \mathbb{R}^d$, and T and C denote the number of question types and object classes in our VQACL. Then, to fit the continual learning setting, the prototypes are constantly updated based on the mini-batch data from the current task or memory $\mathcal{M}$. Taking the update of P_t^q as an example, we first compute the expectation $\mathcal{E}_t$ over all the questions that belong to the t-th question type as follows:

$$\mathcal{E}_t = \frac{1}{j}\sum_{i=1}^{j} Pool(Enc(E_t^{q,i})), \tag{2}$$

where j denotes the number of questions with type t in the current mini-batch, $E_t^{q,i}$ represents the textual embedding of the i-th question with type t, and $Pool(\cdot)$ represents the mean pooling operation. Then, the expectation $\mathcal{E}_t$ is leveraged to refresh the prototype as follows:

$$P_t^q = (1-\alpha)\mathcal{E}_t + \alpha P_t^q, \tag{3}$$

where α is the parameter to adjust the updated degree. With the above strategy, on the one hand, we can update the prototype with the latest information to make it more representative, thus enhancing the feature's generalization ability. On the other hand, the prototype retains the knowledge of historical data, which helps mitigate the forgetting for continual learning. After that, given a question, we can obtain its SI feature Q^{SI} by looking up a suitable prototype from $\{P_t^q\}_{t=1}^T$ based on its specific feature Q^{SS}. Formally, $Q^{SI} \in \mathbb{R}^d$ can be selected by solving following objective:

$$Q^{SI} = \arg\max_{P_t^q} \cos(th(Q^{SS}), th(P_t^q)),\ t = 1, ..., T, \tag{4}$$

where $th(\cdot)$ is the hyperbolic tangent function, and $cos(\cdot,\cdot)$ denotes the cosine similarity. In this way, Q^{SI} can contain essential skill knowledge of the corresponding question type. Similar to Q^{SI}, the visual SI feaute $V^{SI} \in \mathbb{R}^d$ for the image V can be learned through Eq. (2-4) with the different input and a new parameter β in Eq. (3).

4.4. Text Decoder and Objective Function

Similar to $Enc(\cdot)$, the text decoder $Dec(\cdot)$ is also a stack of transformer blocks, where each block has an additional

Table 1. Model performance on VQA v2 and NExT-QA with the VQACL setting. #Mem: memory size; Standard Test: standard testing; Novel Comp. Test: novel composition testing; *AP*: Final Average Performance (%); *Forget*: Average Forgetting (%).

Methods	VQA v2					NExT-QA				
	#Mem	Standard Test		Novel Comp. Test		#Mem	Standard Test		Nove Comp. Test	
		AP (↑)	*Forget* (↓)	*AP* (↑)	*Forget* (↓)		*AP* (↑)	*Forget* (↓)	*AP* (↑)	*Forget* (↓)
Joint	-	51.64	-	51.10	-	-	35.92	-	36.24	-
Vanilla	None	14.49	30.80	11.79	27.16	None	11.97	26.14	12.59	28.04
EWC [22]	None	15.77	30.62	12.83	28.16	None	13.01	24.06	11.91	27.44
MAS [1]	None	20.56	11.16	23.90	6.24	None	18.04	10.07	21.12	10.09
ER [7]	5000	36.99	5.99	33.78	5.76	500	30.55	4.91	32.20	5.57
DER [3]	5000	35.35	8.62	31.52	8.59	500	26.17	5.12	21.56	12.68
VS [45]	5000	34.03	8.79	32.96	5.78	500	28.13	4.45	29.47	6.14
Ours	5000	**38.77**	**3.96**	**35.40**	**4.90**	500	**32.27**	**3.00**	**34.22**	**3.80**

cross-attention layer. Given the previously generated tokens $Y_{<j}$ and the extracted SS and SI feature, the decoder predicts the probability of future text tokens as follows:

$$P_\theta(Y_j|Y_{<j},Q,V) = Dec(Y_{<j},Q^{SS},V^{SS},Q^{SI},V^{SI}). \quad (5)$$

In Eq. (5), we utilize the representations that simultaneously involve discriminative sample-specific content and generalizable sample-invariant knowledge to perform continual learning in VQA. Finally, we train our model parameters θ by minimizing the negative log-likelihood of label text Y tokens as follows:

$$\mathcal{L} = -\sum_{j=1}^{|Y|} logP_\theta(Y_j|Y_{<j},Q,V). \quad (6)$$

5. Experimental Results

5.1. Implementation Details

We construct the proposed model according to Fig. 3. Specifically, to obtain the visual embedding, for the image in VQA v2, we use a Faster R-CNN [38] trained on Visual Genome [23] to extract 36 region features. For the video in NExT-QA, we adopt the clip-level motion feature captured by inflated 3D RexNeXt-101 [15] as the region feature and $n = 16$. In the transformer backbone, we stack 12 blocks for $Enc(\cdot)$ and $Dec(\cdot)$, and the attention layer in each block further has 12 attention heads. The embedding dimension d is set as 768, and the size of the memory buffer $\mathcal{M}$ is set as $5,000$ for VQA v2 and 500 for NExT-QA according to the volume of the datasets. In the prototype learning module, we set α and β as 0.5 and 0.3, respectively. During the model learning, we train each task for 3 epochs with a batch size of 80. Adam [21] is adopted as the optimizer, and the initial learning rate is 10^{-4}. We implement our proposed method based on PyTorch [34].

5.2. Experimental Results in the VQACL setting

The proposed VQACL setting enables a comprehensive analysis of models' continual learning capacity and compositional generalization ability. In this section, we investigate and evaluate five well-established and state-of-the-art continual learning methods in both the standard testing and novel composition testing to verify the effectiveness of our approach, including two regularization methods (EWC [22], MAS [1]) and three rehearsal approaches (ER [7], DER [3], and VS [45]). Besides, we also provide a lower bound (Vanilla) that simply performs gradient update without any countermeasure for the forgetting in the VQACL setting, and an upper bound (Joint) that trains all tasks jointly. For a fair comparison, all the methods are realized using official codes and added to the same transformer backbone introduced in Section 5.1 as our method.

Performance Analysis of Standard Testing. The orange parts in Table 1 provide the model performance on the standard continual learning test set of VQA v2 and NExT-QA. From the results, we can draw the following conclusions: (1) compared with other continual learning approaches, our proposed method consistently achieves the best in terms of both *AP* and *Forget*. Take a closer look at the results, on the VQA v2 and NExT-QA, our model respectively exceeds the rehearsal methods (ER [7], DER [3], and VS [45]) from 1.78% to 4.74% and 1.72% to 6.1% on the *AP*, and achieves 2.03% to 4.83% and 1.45% to 2.12% reduction in terms of the *Forget*. The results demonstrate the superiority of our proposed representation learning method in VQA continual learning. (2) Through the comparison between the regularization methods (EWC [22], MAS [1]) and the rehearsal methods, we observe that the former lags significantly behind the latter on the *AP*. This may be because that in the regularization methods, the regularization constraint that designed for reducing forgetting limits the model's ability to adapt to new tasks. (3) Compared with the offline training model JOINT, the models trained in the VQACL setting largely underperform on both VQA v2 and NExT-QA. This indicates that catastrophic forgetting is prevalent in VQA continual learning, demonstrating the difficulty of our VQACL. (4) Among the compared rehearsal methods, ER [7] achieves the best performance in most cases. This is

Table 2. Fine-grained VQA performance *AP* (%) on the *Novel* and *Seen* skill-concept compositions of VQA v2 and NExT-QA. $+\Delta$ denotes the improvement of our method over the baseline ER [7].

Dataset	Method	Group-1		Group-2		Group-3		Group-4		Group-5		Avg	
		Novel	*Seen*	*Novel*	*Seen*	*Novel*	*Seen*	*Novel*	*Seen*	*Novel*	*Seen*	*Novel*	*Seen*
VQA v2	DER [3]	30.80	29.89	32.19	33.24	34.88	34.08	29.60	30.90	30.14	32.56	31.52	32.13
	VS [45]	33.35	33.87	33.18	32.21	34.50	33.84	31.29	33.98	32.46	33.87	32.96	33.55
	ER [7]	34.52	37.03	33.40	35.55	34.79	34.20	33.86	35.02	32.34	35.91	33.78	35.54
	Ours	**36.12**	**37.99**	**35.39**	**36.92**	**36.26**	**35.16**	**34.85**	**35.64**	**34.36**	**36.28**	**35.40**	**36.40**
	$+\Delta$	1.60	0.96	1.99	1.37	1.47	0.96	0.99	0.62	2.02	0.37	1.62	0.86
NExT-QA	DER [3]	27.56	26.09	26.14	24.54	23.53	26.43	9.30	9.79	21.26	23.74	21.56	21.38
	VS [45]	31.42	30.88	29.17	31.26	25.23	26.10	30.01	29.10	31.54	31.79	29.47	29.83
	ER [7]	31.86	34.51	32.36	35.08	29.50	34.30	33.57	33.30	33.71	32.91	32.20	34.02
	Ours	**35.50**	**35.54**	**33.97**	**35.91**	**31.34**	**35.62**	**34.08**	**33.57**	**36.71**	**33.46**	**34.22**	**34.82**
	$+\Delta$	3.64	1.03	1.61	0.83	1.84	1.32	0.51	0.27	3.00	0.55	2.02	0.80

in contrast to the results in continual learning on unimodal tasks, where DER [3] and VS [45] achieve state-of-the-art results. We think it may be caused by the discrepancy between different continual learning settings.

Performance Analysis of Novel Composition Testing. The blue parts in Table 1 show the comparison results in the novel composition testing, which can measure models' skill-concept compositionality for cognitive reasoning. From the results, we can see that our method obtains the best generalization performance, and outperforms the other continual learning models with clear improvements on both VQA v2 (i.e., 1.62% to 22.57% for *AP*) and NExT-QA (i.e., 2.02% to 22.31% for *AP*), which demonstrates the effectiveness of our proposed method.

We illustrate more fine-grained results in Table 2. Specifically, the results shown in each column mean that the corresponding object group is removed during training. For example, *Group-1* represents that the visual-driven subtask with object group G_1 is omitted. With such training setting, we conduct two types of testing: the *Novel* illustrated in Table 2 represents evaluating the model on novel skill-G_1 compositions, and the *Seen* denotes the testing on seen skill-$G_{2,3,4,5}$ combinations. The average performance across all groups is provided in the last column. From Table 2, we can find that our approach consistently achieves the highest performance for both novel compositions and seen ones. Besides, to better understand the improvement compared with existing methods, we illustrate the improvement over the state-of-the-art method ER [7] in the last line ($+\Delta$) in Table 2. From the results, we can observe that the improvement on *Novel* is much higher than that on *Seen*, which indicates that our method can really enhance the model's compositional generalizability. It may benefit from the learned discriminative sample-specific feature and generalizable sample-invariant feature. In addition, by comparing the results in *Novel* and *Seen*, we find that most continual learning methods obtain lower performance on *Novel* than *Seen*, which implies that compositional generalization is quite challenging for VQA models, and establishing a novel composition testing is rewarding.

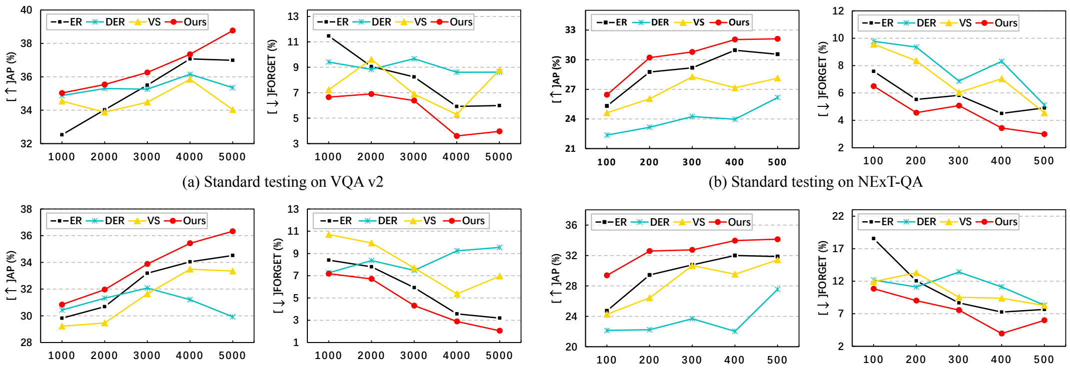
5.3. Ablation Study and Analysis

Effect of Each Component. To investigate the effectiveness of each component in our method, we design several ablated versions and the results are shown in Table 3. Specifically, in Line 1 and Line 2, the variant *Ours w/o* SS and *Ours w/o* SI respectively delete the SS feature (i.e., Q^{SS}, V^{SS}) and SI feature (i.e., Q^{SI}, V^{SI}) in Eq. (5). The comparison between *Ours* and these two models suggests that both the SS and SI feature can effectively boost the VQA continual learning and improve the model's generalization ability. Besides, we find that *Ours w/o* SS gets a quite low performance, which is an unsurprising result because the SI feature only contains category information and lacks detailed contents of the input. In addition, in Line 3, the variant *Ours*_QV^{SI} replaces the Q^{SI} and V^{SI} in *Ours* with a single SI feature that fuses the visual and textual input. Compared with *Ours*, the *Ours*_QV^{SI} obtains a clear performance decrease, which indicates that disentangling the skill and concept is critical for VQA, especially for the model's compositionality. Finally, our full model shown in the last line outperforms all the variants, demonstrating the effectiveness of our representation learning approach.

Sensitive Analysis on Memory Size. Fig. 4 illustrates the model performance on standard and novel composition testing with different memory sizes. From Fig. 4, we can observe that our method always achieves the best perfor-

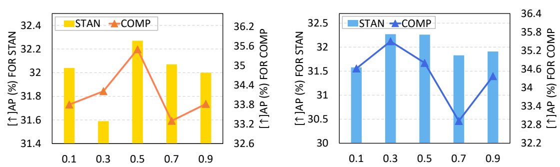
Table 3. Ablation study in both standard testing (Standard) and novel composition testing of Group-1 (Composition).

Method	Standard		Composition	
	AP	*Forget*	*AP*	*Forget*
Ours w/o SS	15.07	11.79	15.49	13.23
Ours w/o SI	30.55	4.91	31.86	7.67
Ours_QV^{SI}	31.88	3.06	32.35	9.47
Ours	**32.27**	**3.00**	**35.50**	**4.45**



(a) Standard testing on VQA v2
(b) Standard testing on NExT-QA
(c) Novel composition testing on VQA v2
(d) Novel composition testing on NExT-QA

Figure 4. Sensitivity analysis on the memory size in both standard and novel composition testing (Group-1) of VQA v2 and NExT-QA.



(a) Parameter α
(a) Parameter β

Figure 5. Parameter variation with different α and β. STAN: standard testing; COMP: novel composition testing (Group-1).

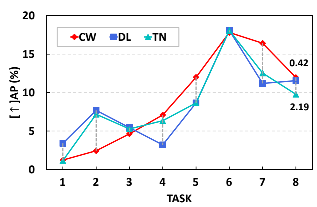


Figure 6. Effect of the linguisitic task order on NExT-QA.

mance, regardless of how many examples are stored. The result indicates the efficacy of the proposed method for continual VQA. Besides, when the memory is larger, the performance of all continual learning methods can obtain clear improvements in most cases, suggesting that more replayed data helps mitigate the forgetting problem. However, as shown in Fig. 4(a) and Fig. 4(c), the performance of VS [45] and DER [3] tends to decrease with larger memory sizes. We think it may be due to that the VS and DER overfit to the data stored in the memory.

Impact of hyperparameter. We investigate the influence of two important parameters involved in our method, i.e., α and β in Eq. (3). Specifically, we train models with $\alpha, \beta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, and the results are depicted in Fig. 5. From the figure, considering the model's performance in both standard and novel composition testing, we find that $\alpha = 0.5$ and $\beta = 0.3$ works the best. Therefore, we set $\alpha = 0.5$ and $\beta = 0.3$ in our experiments.

Effect of Task Order. Fig. 6 provides the performance of the Vanilla model with three different task orders, which respectively adopt *Causal Why (CW)*, *Descriptive Location (DL)*, and *Temporal Next (TN)* as the first linguistic-driven task. Each line in Fig. 6 illustrates the *AP* on the tasks observed so far. From the figure, we find that the task order causes the model performance to vary from 0.42% to 2.19% in terms of the *AP* for the last task, which suggests that the impact of the order is not significant and our VQACL setting is robust to the task order. Besides, among the three sequences, the one beginning with *TN* achieves the worst final performance. This may be because that the task about temporal relationships requires a higher-order reasoning ability.

6. Conclusion

In this paper, we propose and analyze VQACL, a generative VQA continual learning setting. To meet real-world requirements, our VQACL constructs a dual-level task sequence where the vision and linguistic input are nested to cope with continuous multimodal data, and builds a novel composition test to evaluate modes' compositionality. Besides, we design a novel rehearsal representation learning method for the VQACL by extracting sample-specific and sample-invariant features, which can effectively deal with the forgetting problem and is beneficial to improve the composition ability of the model. In experiments, we evaluate five well-known continual learning approaches in our VQACL setting and provide extensive analysis. The comparison between these methods and our approach demonstrates the effectiveness and generalizability of the proposed model. In the future, we hope the VQACL would open a new avenue for the community and contribute to the development of new generative VQA models. We also plan to apply our method to relevant tasks, such as visual dialog and image captioning.

7. Acknowledgement

This work was supported by National Key Research & Development Plan of China (No. 2020AAA0106200), National Natural Science Foundation of China (No. 62036012, 62002355, 61720106006, 62102415, 62106262, 62072455, 62202331 and 62206200), Tianjin Natural Science Foundation (22JCYBJC00030), Beijing Natural Science Foundation (L201001, 4222039).

References

[1] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *ECCV*, pages 139–154, 2018. 2, 3, 6

[2] Pablo Alvarez and Larry R Squire. Memory consolidation and the medial temporal lobe: a simple network model. *Proceedings of the national academy of sciences*, 91(15):7041–7045, 1994. 1

[3] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *NeurIPS*, 33:15920–15930, 2020. 2, 3, 6, 7, 8

[4] Giuseppe Castellucci, Simone Filice, Danilo Croce, and Roberto Basili. Learning to solve nlp tasks in an incremental number of languages. In *ACL*, pages 837–847, 2021. 1

[5] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *ECCV*, pages 532–547, 2018. 4

[6] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. In *ICLR*, 2019. 3, 4

[7] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and M Ranzato. Continual learning with tiny episodic memories. *arXiv preprint arXiv:1902.10486*, 2019. 2, 3, 4, 6, 7

[8] Mengyuan Chen, Junyu Gao, Shicai Yang, and Changsheng Xu. Dual-evidential learning for weakly-supervised temporal action localization. In *ECCV*, pages 192–208. Springer, 2022. 1

[9] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *ICML*, pages 1931–1942, 2021. 2, 3, 4

[10] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *TPAMI*, 44(7):3366–3385, 2021. 3

[11] Riccardo Del Chiaro, Bartłomiej Twardowski, Andrew Bagdanov, and Joost Van de Weijer. Ratt: Recurrent attention to transient tasks for continual image captioning. *NeurIPS*, 33:16736–16748, 2020. 3

[12] Radhika Dua, Sai Srinivas Kancheti, and Vineeth N Balasubramanian. Beyond vqa: Generating multi-word answers and rationales to visual questions. In *CVPR*, pages 1623–1632, 2021. 1

[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 *CVPR*, pages 6904–6913, 2017. 2, 3

[14] Claudio Greco, Barbara Plank, Raquel Fernández, and Raffaella Bernardi. Psycholinguistics meets continual learning: Measuring catastrophic forgetting in visual question answering. In *ACL*, pages 3601–3605, 2019. 2, 3

[15] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, pages 6546–6555, 2018. 6

[16] Tyler L Hayes, Kushal Kafle, Robik Shrestha, Manoj Acharya, and Christopher Kanan. Remind your neural network to prevent catastrophic forgetting. In *ECCV*, pages 466–483. Springer, 2020. 2, 3

[17] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709, 2019. 3

[18] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pages 2901–2910, 2017. 3

[19] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019. 2, 5

[20] Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, et al. Measuring compositional generalization: A comprehensive method on realistic data. In *ICLR*, 2019. 3

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 6

[22] 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, 3, 6

[23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017. 6

[24] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017. 1, 3

[25] 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 *AAAI*, 2023. 2, 3

[26] Xilai Li, Yingbo Zhou, Tianfu Wu, Richard Socher, and Caiming Xiong. Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In *ICML*, pages 3925–3934. PMLR, 2019. 3

[27] Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. Learning to contrast the counterfactual samples for robust visual question answering. In *EMNLP*, pages 3285–3292, 2020. 2, 3

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 4

[29] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *NeurIPS*, 30, 2017. 4

[30] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. *NeurIPS*, 27, 2014. 4

[31] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989. 2

[32] Mavina Nikandrou, Lu Yu, Alessandro Suglia, Ioannis Konstas, and Verena Rieser. Task formulation matters when learning continually: A case study in visual question answering. *arXiv preprint arXiv:2210.00044*, 2022. 3

[33] Inyoung Paik, Sangjun Oh, Taeyeong Kwak, and Injung Kim. Overcoming catastrophic forgetting by neuron-level plasticity control. In *AAAI*, volume 34, pages 5339–5346, 2020. 3

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, volume 32, 2019. 6

[35] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, pages 18082–18091, 2022. 2

[36] Roger Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285, 1990. 1

[37] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, pages 2001–2010, 2017. 1

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015. 6

[39] 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. 3

[40] 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 *ICML*, pages 4528–4537. PMLR, 2018. 3

[41] Tejas Srinivasan, Ting-Yun Chang, Leticia Leonor Pinto Alva, Georgios Chochlakis, Mohammad Rostami, and Jesse Thomason. Climb: A continual learning benchmark for vision-and-language tasks. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 3

[42] Yu-Ming Tang, Yi-Xing Peng, and Wei-Shi Zheng. Learning to imagine: Diversify memory for incremental learning using unlabeled data. In *CVPR*, pages 9549–9558, 2022. 3

[43] Sebastian Thrun. A lifelong learning perspective for mobile robot control. In *Intelligent robots and systems*, pages 201–214. Elsevier, 1995. 1

[44] Rishabh Tiwari, Krishnateja Killamsetty, Rishabh Iyer, and Pradeep Shenoy. Gcr: Gradient coreset based replay buffer selection for continual learning. In *CVPR*, pages 99–108, 2022. 3

[45] Timmy ST Wan, Jun-Cheng Chen, Tzer-Yi Wu, and Chu-Song Chen. Continual learning for visual search with backward consistent feature embedding. In *CVPR*, pages 16702–16711, 2022. 2, 6, 7, 8

[46] Zhen Wang, Liu Liu, Yajing Kong, Jiaxian Guo, and Dacheng Tao. Online continual learning with contrastive vision transformer. In *ECCV*, pages 631–650, 2022. 1

[47] Spencer Whitehead, Hui Wu, Heng Ji, Rogerio Feris, and Kate Saenko. Separating skills and concepts for novel visual question answering. In *CVPR*, pages 5632–5641, 2021. 2, 3

[48] Tongtong Wu, Massimo Caccia, Zhuang Li, Yuan-Fang Li, Guilin Qi, and Gholamreza Haffari. Pretrained language model in continual learning: A comparative study. In *ICLR*, 2021. 1

[49] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, pages 9777–9786, 2021. 1, 2, 3, 4

[50] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. Video as conditional graph hierarchy for multi-granular question answering. AAAI, 2022. 2

[51] Mengqi Xue, Haofei Zhang, Jie Song, and Mingli Song. Meta-attention for vit-backed continual learning. In *CVPR*, pages 150–159, 2022. 1

[52] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, pages 21–29, 2016. 2

[53] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. In *ICLR*, 2018. 3

[54] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *CVPR*, pages 6281–6290, 2019. 2

[55] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, pages 6720–6731, 2019. 2

[56] Feifei Zhang, Mingliang Xu, and Changsheng Xu. Geometry sensitive cross-modal reasoning for composed query based image retrieval. *TIP*, 31:1000–1011, 2021. 1

[57] Feifei Zhang, Ming Yan, Ji Zhang, and Changsheng Xu. Comprehensive relationship reasoning for composed query based image retrieval. In *ACM MM*, pages 4655–4664, 2022. 1

[58] Xi Zhang, Feifei Zhang, and Changsheng Xu. Explicit cross-modal representation learning for visual commonsense reasoning. *TMM*, 24:2986–2997, 2021. 2

[59] Xi Zhang, Feifei Zhang, and Changsheng Xu. Multi-level counterfactual contrast for visual commonsense reasoning. In *ACM MM*, pages 1793–1802, 2021. 2, 3

[60] Yifeng Zhang, Ming Jiang, and Qi Zhao. Query and attention augmentation for knowledge-based explainable reasoning. In *CVPR*, pages 15576–15585, 2022. 1, 2

[61] Yiyi Zhou, Tianhe Ren, Chaoyang Zhu, Xiaoshuai Sun, Jianzhuang Liu, Xinghao Ding, Mingliang Xu, and Rongrong Ji. Trar: Routing the attention spans in transformer for visual question answering. In *ICCV*, pages 2074–2084, 2021. 2

[62] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *CVPR*, pages 8746–8755, 2020. 2