Ask and Remember: A Questions-Only Replay Strategy for Continual Visual Question Answering

Imad Eddine Marouf[1] Enzo Tartaglione[1] Stéphane Lathuilière[1,2]
Joost van de Weijer[3]

[1] LTCI, Télécom-Paris, Institut Polytechnique de Paris, France
[2]Inria, LJK, Univ. Grenoble Alpes, France, [3]Universitat Autónoma de Barcelona, Spain

Abstract

*Continual Learning in Visual Question Answering (VQACL) requires models to acquire new visual-linguistic skills (plasticity) while preserving previously learned knowledge (stability). The inherent multimodality of VQACL exacerbates this challenge, as models must balance stability across visual and textual domains while adapting to novel objects and reasoning tasks. Existing methods, primarily designed for unimodal settings, often fall short in addressing this dual requirement. In this work, we present QUestion-only replay with Attention Distillation (*QUAD*), a novel approach for VQACL that leverages only past task questions for regularization. By eliminating the need to store visual data,* QUAD *not only reduces memory overhead, but also alleviates privacy concerns. Our method introduces a Question-only Replay mechanism that selectively reuses prior task questions to counteract overfitting to the answer space of the current task, addressing the problem out of answer set. Complementing this, we propose Attention Consistency Distillation to enforce both intra-modal and inter-modal attention consistency across tasks, preserving essential visual-linguistic associations. Extensive experiments on VQAv2 and NExT-QA demonstrate that* QUAD *significantly outperforms state-of-the-art methods, achieving robust performance in continual VQA. Code is available at: https://github.com/IemProg/QUAD.* [1]

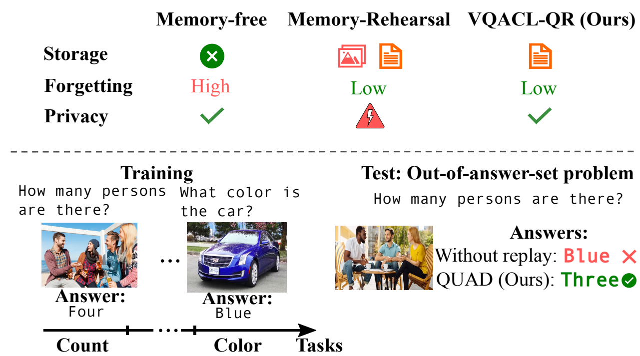

Figure 1. Comparison of continual learning methods for Visual Question Answering (VQA) in terms of storage, forgetting, and privacy. **Memory-free** methods ensure privacy but suffer from high forgetting. **Memory-rehearsal** methods reduce forgetting but raise privacy concerns by storing sensitive images (*e.g.* people's identities, car plates). Our approach **QUAD**, only stores questions and avoids image storage, preserving privacy while achieving low forgetting by leveraging question-based regularisation to effectively solve the out-of-answer set problem.

1. Introduction

Continual learning (CL) allows models to lean new skills while retaining prior knowledge, effectively mitigating catastrophic forgetting (CF) [57, 58]. This capability is essential in dynamic real-world settings, where models must continuously adapt to evolving data while preserving previously acquired knowledge [95]. Despite significant advances in CL, most research has focused on unimodal tasks (*e.g.* image classification) [26, 38, 56, 80, 81, 97]. However, real-world applications often require multimodal learning that integrates complex visual and textual reasoning. Visual Question Answering (VQA) stands as a representative multimodal task, requiring models to jointly interpret visual content and natural language. For instance, answering questions like "*What color is the car on the left?*" or "*How many trucks are visible?*" requires robust object recognition and a nuanced grasp of the linguistic query structure [10, 70, 83, 88, 95].

The emerging field of Visual Question Answering Continual Learning (VQACL) [95] focuses on enabling models to improve VQA performance iteratively by learning from a sequence of tasks without catastrophic forgetting. Un-

[1]Work done during an internship at Computer Vision Center (CVC), Universitat Autonoma de Barcelona.

like static pretrained multimodal models [1, 44, 50] that rely on massive datasets and heavy computation, VQACL offers a more efficient and scalable alternative by allowing incremental adaptation without costly re-training [61, 95]. Research in VQACL is further driven by the broader scientific goal of developing multimodal agents that continuously improve their reasoning throughout deployment in a multimodal environments. A key challenge in VQACL is balancing *stability*—the preservation of past knowledge—and *plasticity*—the ability to learn new information—across both visual and linguistic modalities [5, 95]. This dual-modality requirement, combined with the need for generalisation ability, introduces additional complexity in learning. Models must retain visual and linguistic knowledge across tasks, while also generalizing to novel objects and unseen question types. For example, a model that masters counting vehicles should be able to transfer this counting capability to novel object categories, such as bicycles [27].

To achieve this balance, continual VQA methods rely on memory-based replay, where previously seen examples are stored and revisited to reinforce prior knowledge and mitigate forgetting [8, 9, 77, 95]. In VQACL, this necessitates storing full image-question pairs, which demands substantial memory resources, often amounting to thousands of samples per task (*e.g.* 5000 samples in VQACL [95]). Storing visual data poses two core challenges. *First*, it incurs high computational and storage overhead, which can be infeasible in resource-constrained real-world applications. *Second*, more importantly, it also raises serious privacy concerns. Visual data often contains sensitive and personally identifiable information, especially in domains like healthcare, finance, and surveillance [49, 74], where stringent regulations such as GDPR [22, 91] govern data usage and storage. In contrast, textual data—such as questions—is typically generic and non-identifiable, thereby posing minimal privacy risks. Memory-free methods [2, 38, 41, 93], though eliminating storage and privacy concerns by avoiding data retention entirely, often deliver suboptimal performance in multimodal settings like VQACL. This tension between memory-rehearsal and memory-free approaches prompts a key question: *Is it truly necessary to store visual data, or could retaining only past questions suffice to mitigate forgetting?* To explore these questions, we propose a novel intermediate setting: VQACL with Question-only Rehearsal (VQACL-QR) (Fig. 1). In this framework, only questions from past tasks are stored—eliminating the need for visual data—and offering a practical, privacy-preserving solution for continual VQA.

To address VQACL-QR, we propose **Q**UESTION-ONLY REPLAY WITH **A**TTENTION **D**ISTILLATION (**QUAD**), a novel replay framework that balances stability and plasticity using only past questions—no visual data required. Our approach introduces two key contributions, each targeting specific challenges in continual VQA. *First*, we introduce a **Question-only Replay** mechanism that leverages stored questions from previous tasks to regularize the current model through a strategic selection process that mitigates the *out-of-answer-set problem*, where models overfit to the current task's answer space and consequently misanswer previous queries (see Fig. 2). *Second*, we propose **Attention Consistency Distillation**, a novel strategy that preserves attention patterns across tasks. It preserves intra-modal (text–text, image–image) and inter-modal (text–image) attention consistency, ensuring the model maintains focus on relevant regions even as it adapts to new information. Extensive experiments on standard VQACL benchmarks (VQAv2 and NExT-QA) demonstrate that QUAD achieves state-of-the-art performance, surpassing both memory-free approaches and rehearsal methods. Notably, QUAD surpasses prior methods that rely on image storage, demonstrating that *storing only questions is sufficient to mitigate forgetting*, validating the practicality of the Question-only Rehearsal (VQACL-QR) setting.

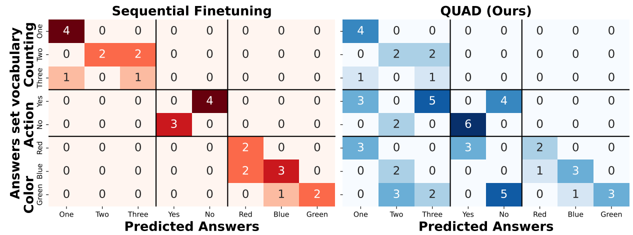

Figure 2. **Out-of-Answer-Set Problem in Sequential Finetuning.** Confusion matrices compare Sequential Finetuning (left) and QUAD (right) across three *sequentially* trained tasks: *Counting, Action, and Color* (y-axis: answers set vocabulary, x-axis: predicted answers). Diagonal values indicate model predictions on the current task, while off-diagonal shifts reveal predictions on previous tasks. Sequential Finetuning exhibits, misclassifying past-task questions with responses from the latest task (e.g., counting questions answered with 'Yes/No'). QUAD mitigates forgetting, preserving prior knowledge while still adapting to new tasks.

2. Related work

Visual Question Answering (VQA) is the task of responding to natural language queries by interpreting visual content [5, 45, 60]. Recent approaches leverage vision-language models (VLMs) built on transformer architectures [10, 19, 43, 71, 86] alongside pre-trained language models [88, 90]. For example, Cho et al. [11] introduced a generative transformer model that integrates visual and textual modalities for VQA. Many approaches enhance generalization by leveraging compositionality—a cornerstone of cognitive reasoning [35, 42]. For instance, Johnson et al. [33] explored the composition of visual attributes by creating a dataset designed for compositional reasoning, while

Whitehead et al. [84] used contrastive learning to enhance compositionality, disentangling reasoning skills from visual concepts. Despite these advances, implicit decomposition strategies can hinder generalization, and crafting effective contrastive samples remains challenging.

Continual Learning (CL) seeks to develop frameworks that incrementally assimilate new data while mitigating catastrophic forgetting. This is a fundamental challenge for many deep learning methods due to catastrophic forgetting (CF) [57]. CL methods are broadly categorized into *knowledge distillation*-based approaches, which constrain mappings between successive models to prevent forgetting [16, 31, 34, 46, 54, 66]; *optimisation*-based, which adjust gradient updates to minimize interference with prior tasks [7, 52, 69, 89] and *representation*-based strategies that learn robust, adaptable features [17, 18, 20, 21]. Recently, *prompt*-based methods have emerged [72, 78, 81, 82], employing visual prompts with pre-trained transformers in CL scenarios [51]. However, since most techniques are designed for Class-Incremental Learning (CIL) in unimodal settings, they fall short when applied to multimodal data and overlook the critical issue of compositional generalization in VQA [61, 95].

Continual VQA. Recent studies have explored multimodal continual learning in VQA [13, 28, 61, 73]. However, prior works like Greco et al. [28] primarily analyze forgetting dynamics without proposing dedicated solutions, often overlooking the role of pre-trained models [59]. The VQACL benchmark [95] evaluated conventional continual learning approaches such as [8, 9, 77, 95]. Other studies have investigated specific facets of continual VQA, including question diversity [28], compositionality [95] and domain adaptation [96]. In contrast, our work introduces a novel question-only replay mechanism with attention distillation for continual VQA that is both memory-efficient, and privacy-conscious, eliminating the need for stored prototypes as employed in the VQACL method [95][2].

3. Question-only replay with Attention Distillation (QUAD)

3.1. Setting Overview

The VQACL setting [95] is designed to assess the capability of a model to adapt to a sequence of tasks, each involving both visual and linguistic inputs, in a continual learning environment. It approaches VQA as a generative task, where the objective is to generate textual answers given an image and a corresponding question [23, 95]. The model encounters a non-stationary stream of data, requiring it to learn and adapt incrementally over time without revisiting prior data. We consider a sequence of T tasks, denoted as $\mathcal{T}^1, \mathcal{T}^2, \dots, \mathcal{T}^T$. Each task $\mathcal{T}^t$ is characterized by a set of image-question-answer triplets (x^t, q^t, y^t), where $x^t \in \mathcal{X}^t$ denotes the image, $q^t \in \mathcal{Q}^t$ represents the question, and $y^t \in \mathcal{Y}^t$ corresponds to the answer.[3] The challenge is to train a model ϕ that can effectively learn the current task $\mathcal{T}^t$ while retaining the knowledge from all previous tasks $\{\mathcal{T}^1, \mathcal{T}^2, \dots, \mathcal{T}^{t-1}\}$.

In VQACL, the sequence of tasks is organized as a *series of L macro-tasks*, each comprising K sub-tasks, resulting in a total of $T = L \times K$ tasks. Each macro-task is designed to develop specific reasoning skills such as counting, color identification, or object recognition (*i.e.* linguistic task). For example, in a counting task, the model primarily engages with questions like "*How many objects are there?*" or "*What number is shown?*".

Each linguistic macro-task is further divided into visually-driven sub-tasks. Formally, each macro-task $\mathcal{T}^t$ is split into K visually-driven sub-tasks, $\{\mathcal{S}_1^t, \mathcal{S}_2^t, \dots, \mathcal{S}_K^t\}$, which are learned sequentially. These sub-tasks $\mathcal{S}_k^t$ are constructed by grouping the C distinct visual object categories, $\{c_i\}_{i=1}^C$, into K sets. This hierarchical structure mirrors the continuous nature of the visual and linguistic data streams that the model processes. The VQACL setting introduces two unique challenges for continual learning models. *1) Knowledge Retention*: as the model progresses through the sequence of tasks, it must retain knowledge from earlier tasks to perform well on future tasks, where both visual and linguistic modalities must be preserved. *2) Generalization to Novel Compositions*: this setting also evaluates the model's ability to generalize to novel combinations of visual concepts and reasoning skills that it has not encountered during training. This aspect is crucial for real-world applications where new object-skill combinations are frequently encountered. Details of task sequences, object groupings, and novel composition testing is in the Appendix.

3.2. Overview

We propose **QUAD**, a novel approach for VQACL-QR framework that avoids image storage by relying solely on previously encountered questions (see Fig. 3). Inspired by prior work in continual learning [14, 47], we adopt a regularisation framework in which the overall learning objective $\mathcal{L}_{\text{VQACL}}$ is composed of two main components:

$$\mathcal{L}_{\text{VQACL}} = (1-\lambda)\mathcal{L}_{\text{Plasticity}} + \lambda\mathcal{L}_{\text{Stability}}. \tag{1}$$

The plasticity term $\mathcal{L}_{\text{Plasticity}}$, drives the model to adapt to the current task $\mathcal{T}^t$, while the weighting factor $\lambda > 0$ balances the trade-off between the two loss terms. Following common practice in VQA [4, 65, 95], we implement the plasticity loss using cross-entropy to compare the network's prediction for an input image-question pair (x^t, q^t), against

[2] VQACL represents both the setting and the approach for continual learning in Visual Question Answering, as described in [95].

[3] The sample index is omitted for clarity.

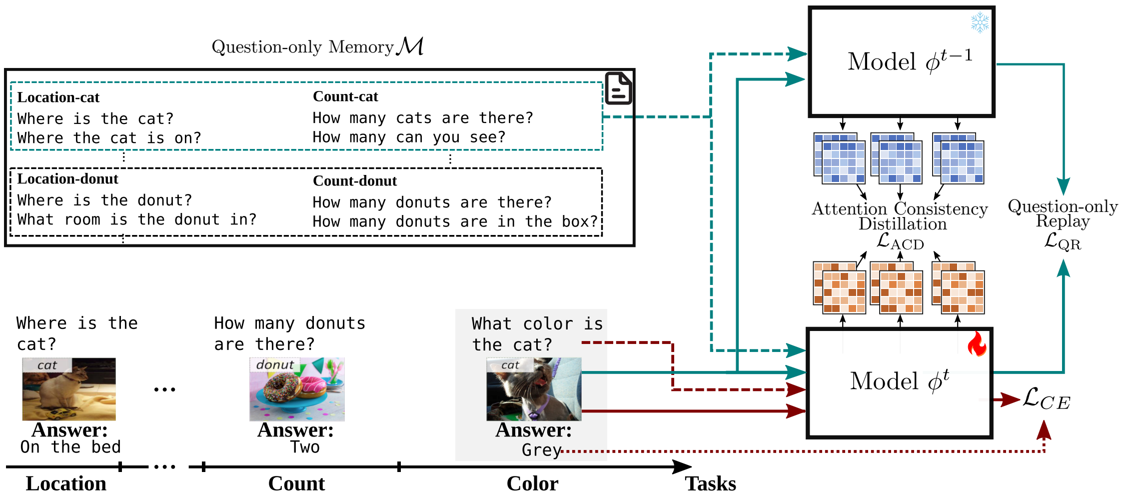

Figure 3. **Overview of QUAD**. QUAD is composed of three components that jointly promote stability and plasticity in VQACL setting. **(1) Question-Only Memory** ($\mathcal{M}$) stores questions from past tasks, without visual data. **(2) Question-only replay** ($\mathcal{L}_{\text{QR}}$) leverages answers generated by the previous model θ^{t-1} for new image-question pairs, encouraging the current model θ^t to retain past knowledge. **(3) Attention Consistency Distillation** ($\mathcal{L}_{\text{ACD}}$) aligns the self-attention maps between θ^{t-1} and θ^t to maintain focus on relevant visual-linguistic relationships. The task-specific loss ($\mathcal{L}_{\text{CE}}$) is applied solely to current task samples, promoting adaptation to new data.

the ground-truth answer y^t:

$$\mathcal{L}_{\text{Plasticity}} = \mathbb{E}_{(x^t,q^t,y^t)\sim\mathcal{T}^t}\mathcal{L}_{\text{CE}}\left[\phi(x^t,q^t),y^t\right]. \tag{2}$$

The second loss component, the stability term $\mathcal{L}_{\text{Stability}}$, mitigates catastrophic forgetting. In the standard VQACL, where images from previous tasks are stored in memory $\mathcal{M}$, the stability term is computed analogously using cross-entropy, averaging the loss over triplets $(x^m, q^m, y^m) \in \mathcal{M}$. However, under the VQACL-QR setting, storage of images x^m from past tasks in the memory $\mathcal{M}$ is disallowed. To overcome this limitation, we design a novel stability loss $\mathcal{L}_{\text{Stability}}$ tailored for VQACL-QR. Our approach integrates two complementary losses—question-only replay $\mathcal{L}_{\text{QR}}$ and attention distillation $\mathcal{L}_{\text{ACD}}$—to effectively compensate for the absence of past task images, ensuring robust knowledge retention within the constraints of the VQACL-QR framework. Thus, the stability term is expressed as:

$$\mathcal{L}_{\text{Stability}} = \mathcal{L}_{\text{QR}} + \mathcal{L}_{\text{ACD}}. \tag{3}$$

Next, we describe the implementation of these loss components without storing past task images.

3.3. Question-only Replay

To enhance knowledge retention in continual VQA, we propose a questions-only replay strategy that replays stored questions by pairing them with current task images to simulate past tasks. For each image from the current task x^t, we pair it with a question q^m sampled from the memory. By combining stored questions with new images, our approach encourages the model to recall and reinforce previously acquired knowledge. Inspired by prior distillation-based works [14, 47], we use the model from the previous task, ϕ^{t-1} to generate answers for each new image-question pair (x^t, q^m). The generated answers act as soft pseudo-labels for the current model ϕ^t, enforcing consistency with prior knowledge. The loss is defined as:

$$\mathcal{L}_{\text{QR}} = \mathbb{E}_{x^t\sim\mathcal{T}^t}\mathbb{E}_{q^m\sim\mathcal{M}}\mathcal{L}_{\text{CE}}\left[\phi^t(x^t,q^m),\phi^{t-1}(x^t,q^m)\right]. \tag{4}$$

Notably, we employ the network's output as soft pseudo-labels without applying the argmax operator. Using argmax would force the network to align exclusively with the most probable class; in contrast, soft pseudo-labeling allows for a more nuanced alignment with the full output distribution of ϕ^{t-1} [30, 76, 94].

These image-question pairs (x^t, q^m) expose the model to a wide range of visual-question combinations, thereby improving the model's retention of prior knowledge. Importantly, this strategy enables the model to maintain versatility in answering diverse question types, even when predictions exhibit some uncertainty. Without this regularization, the model becomes vulnerable to the *out-of-answer-set problem*, wherein overfitting to the current task's answer space results in erroneous responses for questions from prior tasks (see Fig. 2). For example, a model trained on color recognition might erroneously return a color name when confronted with a counting question from a previous task. Notably, this phenomenon closely resembles class recency bias in class-incremental learning (CIL), where models disproportionately favor newly learned classes over previously encountered ones [55, 68].

Question Selection for QUAD. Randomly pairing current task images x^t with past questions q^m from memory can partially mitigate forgetting but can produce semantically incoherent image-question pairs. For instance, if the current macro-task $\mathcal{T}^t$ focuses on counting, naive pairing might associate an image of cars with the question "*How many cows are in this image?*", leading to incoherent supervision and diminished performance. To address this, we introduce a targeted question selection strategy that prioritizes questions related to the object categories c_i in the current visually-driven subtask $\mathcal{S}^l$. This ensures that replayed examples remain contextually meaningful, reinforcing relevant knowledge while preventing misleading associations. By aligning stored questions with the current task's visual concepts, QUAD facilitates effective adaptation across linguistic and visual subtasks, improving continual learning performance.

Example: Suppose the current visually-driven subtask $\mathcal{S}^t$ involves learning to count cars. The question selection strategy prioritizes memory questions relevant to cars, such as "*What's the color of the car?*", ensuring that replay maintains associations between visual and linguistic queries.

3.4. Attention Consistency Distillation

While question-only replay using pseudo-labeling ensures output consistency, it fails to constrain internal representations, causing self-attention drift and disrupting alignment with previously learned tasks [24, 63, 79]. In contrast, feature distillation across multiple layers [14, 34, 63, 87] offers stronger regularization by aligning internal representations; however, it often restricts plasticity due to its rigid constraints. These methods enforce layer-wise alignment, which becomes problematic when encountering new image-question pairs (x^t, q^m) that were not present in previous training. Despite their semantic relevance, the model lacks prior exposure to these pairs, making strict layer-wise regularization overly restrictive. Our VQA model [61, 95] encodes image and text tokens within a unified transformer sequence, allowing self-attention to dynamically capture intra-modal (text-to-text, image-to-image) and inter-modal (text-to-image) dependencies. However, sequential fine-tuning gradually shifts self-attention [24, 79], causing the model to focus on different visual regions than those relevant to past tasks, ultimately degrading performance on prior knowledge [24, 75]. This drift arises because pseudo-labeling alone fails to stabilize internal representations, necessitating explicit regularization.

To mitigate this, we introduce *attention consistency distillation* (ACD), which aligns the self-attention maps of the current model ϕ^t and its predecessor ϕ^{t-1}. By preserving intra-modal and inter-modal relationships, this regularization stabilizes attention patterns while maintaining flexibility. Distilling only attention maps guides the model to attend to task-relevant visual regions for question answering while preserving adaptability to novel input patterns.

For an image-question pair (x^t, q^m) with $x^t \sim \mathcal{T}^t, q^m \sim \mathcal{M}$: we denote A_k^t and A_k^{t-1} as the corresponding self-attention maps in an attention head k for the current and previous model. Thus, we define the self-attention consistency loss $\mathcal{L}_{\text{ACD}}$ via cross-entropy, ensuring that attention distributions remain aligned across tasks:

$$\mathcal{L}_{\text{ACD}} = \mathbb{E}_{x^t \sim \mathcal{T}^t} \mathbb{E}_{q^m \sim \mathcal{M}} \mathbb{E}_{k \sim \mathcal{K}_\phi} \mathcal{L}_{\text{CE}} \left[A_k^t(x^t, q^m), A_k^{t-1}(x^t, q^m) \right], \tag{5}$$

where $\mathcal{K}_\phi$ represents all attention heads across layers of ϕ.

Unlike prior L1-based approaches [14, 63], which impose uniform penalties and operate directly on the raw query-key products, ACD leverages cross-entropy on normalized attention maps (after softmax operation) to preserve the probabilistic structure of attention distributions. This gives greater emphasis to highly attended regions while low-attended areas remain flexible, preventing over-penalization of less critical regions. Asymmetric regularization (e.g., ReLU+L1) [63] partially addresses this issue by prioritizing reductions in attended regions but lacks a probabilistic alignment mechanism, making it less effective in maintaining structured dependencies in multimodal learning. In contrast, ACD offers an importance-weighted alignment that preserves essential cross-modal associations without hindering adaptation. The effectiveness of ACD is empirically demonstrated in Sec. 4, and further discussed in the Appendix.

4. Experiments

4.1. Experimental setup

Implementation Details. To ensure a fair comparison, we adopt the protocol of [95] for both feature extraction and training across datasets. For the visual embeddings, we use a Faster R-CNN [67] trained on the Visual Genome dataset [40] extracting 36 region-based object features per image in the VQAv2 dataset. For videos in the NExT-QA dataset, we extract clip-level motion features using an inflated 3D ResNeXt-101 [29], setting $n = 16$ regions per clip. A two-layer MLP with GELU activation adapts these features for input into the transformer backbone. Our transformer backbone, based on T5 [64], consists of 12 blocks for both encoder and decoder modules, each containing 12 attention heads. The embedding dimension d is tailored to the task-specific requirements. Training is conducted for 3 epochs per task, with a batch size of 80. We utilize the Adam optimizer [36] with an initial learning rate of 10^{-4}. λ is set to 0.5 in all experiments. All implementations are based on PyTorch [62]. Importantly, unlike [95], QUAD does not store any visual or question prototypes.

Table 1. Model performance on VQAv2 and NExT-QA. #Mem: memory size; Type #Mem: type of memory used, where ❓ and 🖼 denote storing both questions and images in the memory buffer; Standard Test: standard testing; Novel Comp. Test: novel composition testing; AP: Final Average Performance (%); Forget: Average Forgetting (%). Best results are in **bold**, second-best are <u>underlined</u>.

Methods	VQAv2						NExT-QA					
	#Mem	Type #Mem	Standard Test		Novel Comp. Test		#Mem	Type #Mem	Standard Test		Novel Comp. Test	
			AP (↑)	Forget (↓)	AP (↑)	Forget (↓)			AP (↑)	Forget (↓)	AP (↑)	Forget (↓)
Joint	-	None	51.64	-	51.10	-	-	None	35.92	-	36.24	-
Vanilla	None	None	14.92	30.80	11.79	27.16	None	None	12.68	25.94	12.59	28.04
EWC [37]	None	None	15.77	30.62	12.83	28.16	None	None	13.01	24.06	11.91	27.44
MAS [2]	None	None	20.56	11.16	23.90	6.24	None	None	18.04	10.07	21.12	10.09
ER [9]	5000	❓/🖼	36.99	5.99	33.78	5.76	500	❓/🖼	30.55	4.91	32.20	5.57
DER [8]	5000	❓/🖼	35.35	8.62	31.52	8.59	500	❓/🖼	26.17	5.12	21.56	12.68
VS [77]	5000	❓/🖼	34.03	8.79	32.96	5.78	500	❓/🖼	28.13	4.45	29.47	6.14
VQACL [95]	5000	❓/🖼	<u>37.46</u>	<u>6.96</u>	<u>35.40</u>	<u>4.90</u>	500	❓/🖼	<u>30.86</u>	<u>4.12</u>	**33.85**	**3.80**
QUAD (Ours)	5000	❓	**39.25**	**4.91**	**40.00**	**3.81**	500	❓	**31.70**	**2.91**	<u>33.21</u>	<u>4.16</u>

Evaluation Metrics. We utilise two established metrics for continual learning [6, 53, 95]: final average performance (AP), and average forgetting ($Forget$). The AP metric reflects the model's overall performance across all learnt tasks, highlighting its ability to consistently acquire new tasks. Let $a_{i,j}$ represent the performance of the model on task $\mathcal{T}^i$ after it has completed learning this task $\mathcal{T}^j$. Then, AP is calculated as: $AP = \frac{1}{T}\sum_{t=1}^{T} a_{t,T}$. Additionally, the $Forget$ metric serves as a proxy for knowledge retention, quantifying performance degradation on prior tasks as new tasks are learned. It is computed as: $Forget = \frac{1}{T-1}\sum_{t=1}^{T-1} \max_{z\in\{t,\dots,T-1\}}(a_{t,z} - a_{t,T})$. To ensure a fair comparison, For NExT-QA, following [85, 95], we compute $a_{i,j}$ using Wu-Palmer Similarity (WUPS) to evaluate answer quality. For the VQAv2 dataset, as described in [95], we use the percentage of correctly answered questions as the value for $a_{i,j}$.

Baselines. We benchmark QUAD against five established continual learning methods spanning regularization and rehearsal-based strategies. This includes two regularization-based methods: Elastic Weight Consolidation (EWC) [38] and Memory Aware Synapses (MAS) [2], as well as three rehearsal-based approaches: Experience Replay (ER) [9], Dark Experience Replay (DER) [8], Virtual Sample (VS) [77], and VQACL [95] (see details in Appendix). We additionally report lower and upper performance bounds to frame the results: a lower bound (Vanilla) using naive finetuning without forgetting mitigation, and an upper bound (Joint) that trains on all tasks simultaneously. Following the evaluation protocol in VQACL [95], we employ two key evaluation strategies: standard testing, which evaluates the models' performance on previously encountered task types, assessing their ability to retain learnt knowledge, and novel composition testing, which challenges the models with previously unseen combinations of visual and linguistic elements, probing their capacity for compositional generalization. This twofold evaluation reveals how well each method balances retention and adaptation (details in Appendix). Additional experiments using pretrained models are included in the Supp Mat.

4.2. Main results

Performance Analysis On Standard Setting. Tab. 1 shows a detailed comparison of continual learning approaches in the VQACL setting, with our proposed QUAD achieving superior performance across both standard and novel composition tests. QUAD consistently outperforms other methods in AP and forgetting demonstrating robust knowledge retention. On VQAv2, QUAD attains an AP of 39.25% in standard testing, outperforming the best rehearsal-based approach (VQACL) by 1.79%. Similarly, in NExT-QA, QUAD achieves an AP of 31.70%, exceeding competing methods by 0.84% to 4.73%. The forgetting rates for QUAD are also the lowest, with 4.91% and 2.91% for VQAv2 and NExT-QA, respectively, underscoring its capacity for stable knowledge retention.

In novel composition testing, QUAD demonstrates strong generalization, achieving top AP scores of 40.00% on VQAv2 and 33.85% on NExT-QA. Notably, despite not storing images, our approach maintains high performance in novel settings, with a minimal performance gap between standard and novel composition tests (0.75% for VQAv2 and 0.64% for NExT-QA). This narrow gap indicates that distilling knowledge using only seen questions effectively reinforces visual-linguistic associations, enabling the model to generalise well even in unfamiliar contexts. This performance underscores QUAD's capacity to construct robust, adaptable representations, showing that question-only distillation can successfully support compositional reasoning without the need for visual memory.

Performance Analysis of Novel Composition Testing. Tab. 2 provides a detailed comparison of model performance on novel and seen skill-concept compositions across VQAv2 and NExT-QA datasets. QUAD consistently sur-

Table 2. Fine-grained VQA performance AP (%) on the Novel and Seen skill-concept compositions of VQAv2 and NExT-QA. +Δ denotes the improvement of our method over the state of the art. Best results are in **bold**, second-best are underlined.

Dataset	Method	Memory	Group-1		Group-2		Group-3		Group-4		Group-5		Avg	
			Novel	Seen	Novel	Seen	Novel	Seen	Novel	Seen	Novel	Seen	Novel	Seen
NExT-QA	DER [8]	❓/🖼	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 [77]	❓/🖼	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 [9]	❓/🖼	31.86	34.51	32.36	35.08	29.50	34.30	33.57	33.30	33.71	32.91	32.20	34.02
	VQACL [95]	❓/🖼	32.86	31.47	31.98	35.58	31.79	35.70	35.04	34.12	37.62	34.92	**33.85**	34.35
	QUAD (Ours)	❓	33.42	33.92	32.02	35.42	31.78	36.36	32.98	33.34	37.84	34.06	33.21	**34.62**
VQAv2	DER [8]	❓/🖼	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 [77]	❓/🖼	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 [9]	❓/🖼	34.52	37.03	33.40	35.55	34.79	34.20	33.86	35.02	32.34	35.91	33.78	35.54
	VQACL [95]	❓/🖼	36.12	37.99	35.39	36.92	36.26	35.16	34.85	35.64	34.36	36.28	35.40	36.40
	QUAD (Ours)	❓	39.19	41.06	38.40	39.50	43.15	39.19	40.01	40.72	39.20	40.62	**40.00**	**40.21**

passes previous approaches. On VQAv2, QUAD shows significant gains over VQACL, with an average improvement of 4.60% on novel compositions and 3.81% on seen compositions. This result indicates enhanced compositional generalisation, as evidenced by the smaller gap between novel and seen performance compared to other methods. For NExT-QA, QUAD maintains competitive performance, with slight gains over VQACL on seen groups (average +0.27%) in a dataset that presents unique challenges, such as temporal and causal reasoning. For such tasks, storing images may be necessary to maintain a comprehensive task understanding. These consistent gains highlight QUAD's robustness in compositional reasoning, validating its effectiveness for continual VQA.

5. Ablation Study and Analysis

① **QUAD Components.** Tab. 3 shows the impact of questions-only replay and attention distillation in QUAD. We evaluate: (1) *Question-only replay alone*, (2) *attention distillation alone*, and (3) *the combination of both*.

Using question-only replay mechanism achieves moderate performance, with AP of 30.72% on VQAv2 and 29.04% on NExT-QA, indicating effective knowledge preservation across tasks. In contrast, attention distillation alone yields lower AP scores (13.34% on VQAv2 and 13.24% on NExT-QA) with high forgetting scores (32.08% on VQAv2 and 24.56% on NExT-QA), suggesting limited task adaptation. Combining $\mathcal{L}_{\text{QR}}$ and $\mathcal{L}_{\text{ACD}}$ achieves the best results, with AP scores of 39.25% on VQAv2 and 31.70% on NExT-QA, and the lowest forgetting rates.

② **Attention Distillation.** Tab. 4 demonstrates the effectiveness of our proposed $\mathcal{L}_{\text{ACD}}$ within QUAD. Unlike prior methods such as Attn-dist (L1) and Asym-Attn [63], which impose L1 or ReLU+L1 losses on raw attention scores, our approach applies cross-entropy over normalized attention maps. QUAD outperforms previous methods, achieving an AP of 39.25% and Forgetting of 4.91% on VQAv2, and an AP of 31.70% with Forgetting of 2.91% on NExT-QA.

Table 3. Ablation study of QUAD components.

$\mathcal{L}_{\text{QR}}$	$\mathcal{L}_{\text{ACD}}$	Memory Type	VQAv2		NExT-QA	
			AP (↑)	Forget (↓)	AP (↑)	Forget (↓)
✓		❓	30.72	13.74	29.04	4.58
	✓	❓	13.34	32.08	13.24	24.56
✓	✓	❓	**39.25**	**4.91**	**31.70**	**2.91**

Table 4. Comparison of attention distillation methods.

Method	VQAv2		NExT-QA	
	AP (↑)	Forget (↓)	AP (↑)	Forget (↓)
$\mathcal{L}_{\text{QR}}$ + Attn-dist (L1)	34.56	7.91	30.14	5.78
$\mathcal{L}_{\text{QR}}$ + Asym-Attn [63]	38.15	5.57	31.18	4.13
QUAD (Ours)	**39.25**	**4.91**	**31.70**	**2.91**

These results highlight the benefit of distributional consistency in attention maps over unstructured alignment.

③ **Analysing Plasticity/Stability.** Fig. 4 compares how different methods manage forgetting and adaptation on the VQAv2 dataset. The sequential fine-tuning baseline (left) exhibits uniformly low off-diagonal scores, signaling a complete failure to retain knowledge from earlier tasks. This severe forgetting stems from overfitting to the current task's answer space, a phenomenon we call the *out-of-answer-set problem*, it occurs when the model overfits to the current task's answer space, preventing it from correctly responding to questions from earlier tasks. By contrast, question-only replay $\mathcal{L}_{\text{QR}}$ (center) noticeably improves retention through our question-only replay mechanism, especially in tasks like "commonsense" and "count". However, it remains less effective in complex reasoning tasks like "causal" and "subcategory".

Our full method, QUAD (right), achieves the best balance: it sustains high diagonal accuracy while boosting off-diagonal retention. reflecting both adaptability and long-term memory. For example, QUAD preserves 62.6% accuracy on the 'judge' task, compared to only 35.0% with replay alone—highlighting the value of attention consistency distillation. Notably, tasks like 'type' that rely heavily on visual semantics remain challenging under the question-

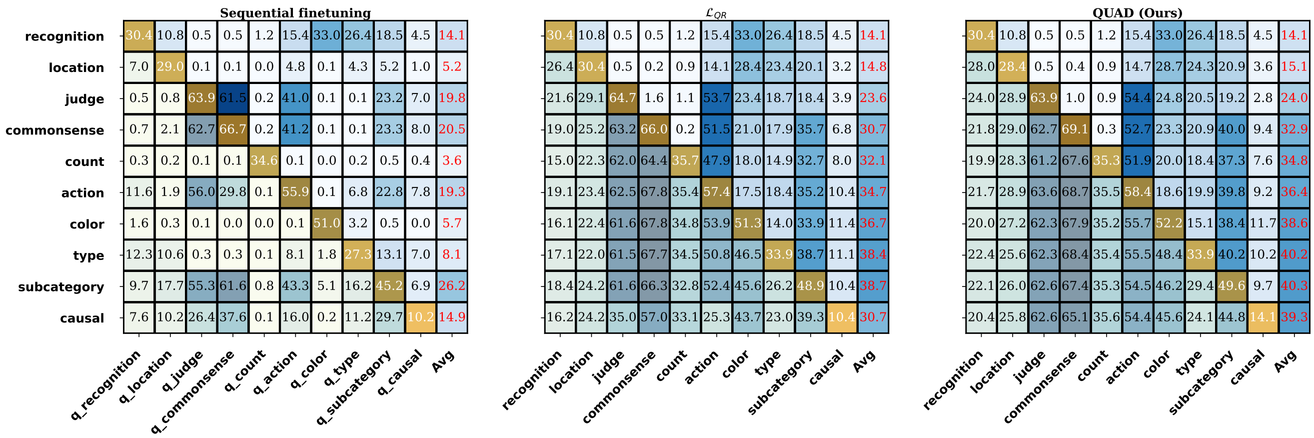

Figure 4. **Plasticity/Stability analysis on VQAv2.** Each matrix shows the performance of a model trained on tasks (rows) and evaluated on tasks **sequentially** (columns). The diagonal (highlighted in orange) represents in-domain performance, while off-diagonal elements indicate cross-domain generalization. Higher values (darker colors) suggest better retention and plasticity. The progression from 'Sequential finetuning', to $\mathcal{L}_{QR}$ and then to our full method, QUAD, highlights the improvement in retaining knowledge across sequential tasks.

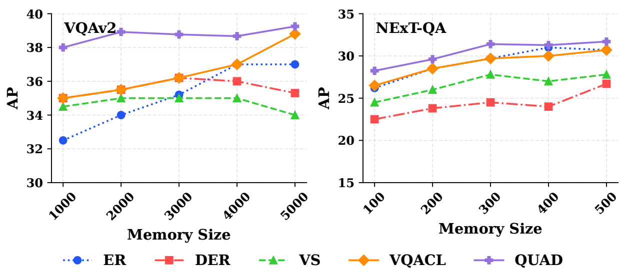

Figure 5. **Sensitivity analysis to memory size.** Our method, QUAD, consistently achieves higher AP than baselines, demonstrating strong scalability on VQAv2 and stable performance on NExT-QA, especially as memory size increases.

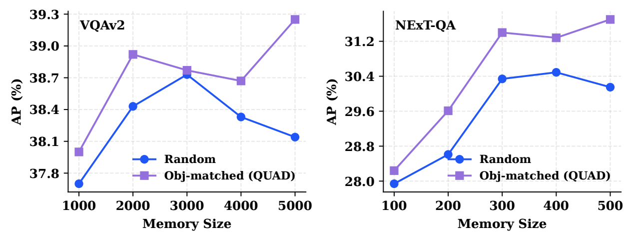

Figure 6. **Question selection strategy in QUAD**. Figure shows AP for random (blue) and object-matched (purple) question selection across memory sizes on VQAv2, and NExT-QA. Object-matched selection consistently outperforms random selection.

only setting, reaffirming the need for richer visual grounding in some categories. Nevertheless, QUAD excels in conceptually driven tasks such as "commonsense", showing its effectiveness even with reduced supervision.

4 **Sensitivity to Memory Size.** Fig. 5 shows how different continual learning methods respond to varying memory budgets. Across all memory sizes, QUAD consistently outperforms baselines (ER, DER, VS, VQACL), demonstrating the effectiveness of our distillation strategy. On VQAv2, QUAD exhibits robust scalability, with AP increasing steadily from 1K to 5K samples. In constrained storage scenarios, QUAD maintains competitive performance by leveraging question-only distillation. On the more visually complex NExT-QA benchmark, QUAD still leads, though with narrower margins. This reflects the greater challenge of retaining visual-semantic alignment using question-only signals.

5 **Effectiveness of Object-Matched Question Selection.** Fig. 6 highlights the advantage of object-matched question selection compared to random pairing. We analyze its impact across memory sizes on VQAv2 and NExT-QA, where it consistently improves AP as memory grows. This targeted selection ensures semantic relevance between the question and visual context, which helps reinforce meaningful cross-task associations. As a result, the model adapts effectively to new tasks while retaining prior knowledge.

6. Conclusion

In this work, we introduced **Q**UESTION-ONLY REPLAY WITH **A**TTENTION **D**ISTILLATION (QUAD), a novel questions-only replay framework for continual VQA. Unlike conventional methods that store both images and questions, QUAD addresses storage and privacy concerns by retaining only past task questions. This design enables effective regularization without storing sensitive visual data, making it highly practical for privacy-conscious applications. Comprehensive evaluations on VQAv2 and NExT-QA show that QUAD consistently outperforms both memory-free and memory-rehearsal methods, achieving state-of-the-art performance. Surprisingly, our method, without any image exemplars, outperforms previous methods, which do require image storage.

Acknowledgements. This paper is supported by the French National Research Agency (ANR) in the framework of the JCJC project "BANERA under Grant ANR-24-CE23-4369, and was funded by the European Union's Horizon Europe research and innovation program under grant agreement No. 101120237 (ELIAS). It was also partially funded by Hi!PARIS Center on Data Analytics and Artificial Intelligence and was granted access to the HPC resources of IDRIS under the allocation AD011013860 made by GENCI.

References

[1] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick Von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. Pixtral 12b, 2024. 2

[2] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *ECCV*, 2018. 2, 6, 7

[3] Andrea Amaduzzi, Giuseppe Lisanti, Samuele Salti, and Luigi Di Stefano. Looking at words and points with attention: a benchmark for text-to-shape coherence. *arXiv preprint arXiv:2309.07917*, 2023. 3

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 3

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 2

[6] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *ECCV*, 2018. 6

[7] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*, 2018. 3

[8] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet Kumar Dokania, Philip H. S. Torr, and Marc'Aurelio Ranzato. Continual learning with tiny episodic memories. *CoRR*, abs/1902.10486, 2019. 2, 3, 6, 7

[9] 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, 3, 6, 7

[10] Feng Cheng, Xizi Wang, Jie Lei, David Crandall, Mohit Bansal, and Gedas Bertasius. Vindlu: A recipe for effective video-and-language pretraining. In *CVPR*, 2023. 1, 2

[11] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. 2021. 2

[12] Deepayan Das, Davide Talon, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. One vlm to keep it learning: Generation and balancing for data-free continual visual question answering. IEEE, 2025. 4

[13] Riccardo Del Chiaro, Bartł omiej Twardowski, Andrew Bagdanov, and Joost van de Weijer. Ratt: Recurrent attention to transient tasks for continual image captioning. In *NeurIPS*, 2020. 3

[14] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. Learning without memorizing. In *CVPR*, 2019. 3, 4, 5, 1, 2

[15] Sayanton V Dibbo. Sok: Model inversion attack landscape: Taxonomy, challenges, and future roadmap. In *2023 IEEE 36th Computer Security Foundations Symposium (CSF)*. IEEE, 2023. 1

[16] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *ECCV*, 2020. 3

[17] Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. Dytox: Transformers for continual learning with dynamic token expansion. In *CVPR*, 2022. 3

[18] Beyza Ermis, Giovanni Zappella, Martin Wistuba, Aditya Rawal, and Cedric Archambeau. Memory efficient continual learning with transformers. *NeurIPS*, 2022. 3

[19] Clayton Fields and Casey Kennington. Vision language transformers: A survey. *arXiv preprint arXiv:2307.03254*, 2023. 2

[20] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020. 3

[21] Qiankun Gao, Chen Zhao, Yifan Sun, Teng Xi, Gang Zhang, Bernard Ghanem, and Jian Zhang. A unified continual learning framework with general parameter-efficient tuning. *arXiv preprint arXiv:2303.10070*, 2023. 3

[22] General Data Protection Regulation GDPR. General data protection regulation. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC*, 2016. 2

[23] Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. Exploring the frontier of vision-language models: A survey of current methodologies and future directions, 2024. 3

[24] Nathan Godey, Éric de la Clergerie, and Benoît Sagot. Anisotropy is inherent to self-attention in transformers. *arXiv preprint arXiv:2401.12143*, 2024. 5

[25] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org. 2

[26] Dipam Goswami, Yuyang Liu, Bartł omiej Twardowski, and Joost van de Weijer. Fecam: Exploiting the heterogeneity of class distributions in exemplar-free continual learning. In *NeurIPS*, 2023. 1

[27] 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*, Florence, Italy, 2019. Association for Computational Linguistics. 2

[28] 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*, 2019. 3

[29] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, 2018. 5

[30] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 4

[31] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *CVPR*, 2019. 3

[32] Tianyu Huai, Jie Zhou, Xingjiao Wu, Qin Chen, Qingchun Bai, Ze Zhou, and Liang He. Cl-moe: Enhancing multimodal large language model with dual momentum mixture-of-experts for continual visual question answering. In *CVPR*, 2025. 4

[33] 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*, 2017. 2

[34] Minsoo Kang, Jaeyoo Park, and Bohyung Han. Class-Incremental Learning by Knowledge Distillation with Adaptive Feature Consolidation. In *CVPR*, 2022. 3, 5

[35] 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. *arXiv preprint arXiv:1912.09713*, 2019. 2

[36] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[37] 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. 6

[38] 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. 1, 2, 6, 7

[39] Chris Kolb, Tobias Weber, Bernd Bischl, and David Rügamer. Deep weight factorization: Sparse learning through the lens of artificial symmetries. *ICLR*, 2025. 2

[40] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 5

[41] Canasai Kruengkrai and Junichi Yamagishi. Mitigating the diminishing effect of elastic weight consolidation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4568–4574, Gyeongju, Republic of Korea, 2022. International Committee on Computational Linguistics. 2

[42] 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:e253, 2017. 2

[43] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*, 2021. 2

[44] 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 *ICML*, 2023. 2

[45] Li Li, Jiawei Peng, Huiyi Chen, Chongyang Gao, and Xu Yang. How to configure good in-context sequence for visual question answering. In *CVPR*, 2024. 2

[46] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 3

[47] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018. 3, 4

[48] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'a r, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 5

[49] Chi Liu, Tianqing Zhu, Jun Zhang, and Wanlei Zhou. Privacy intelligence: A survey on image privacy in online social networks. *ACM Computing Surveys*, 2022. 2

[50] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 2

[51] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9): 1–35, 2023. 3

[52] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *NeurIPS*, 2017. 3

[53] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning, 2022. 6

[54] Yichen Lu, Mei Wang, and Weihong Deng. Augmented geometric distillation for data-free incremental person reid. In *CVPR*, 2022. 3

[55] Zheda Mai, Ruiwen Li, Hyunwoo Kim, and Scott Sanner. Supervised contrastive replay: Revisiting the nearest class mean classifier in online class-incremental continual learning. In *CVPR*, 2021. 4, 5

[56] Imad Eddine Marouf, Subhankar Roy, Enzo Tartaglione, and Stéphane Lathuilière. Weighted ensemble models are strong continual learners. In *ECCV*, 2024. 1

[57] James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995. 1, 3

[58] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*. 1989. 1

[59] Sanket Vaibhav Mehta, Darshan Patil, Sarath Chandar, and Emma Strubell. An empirical investigation of the role of pre-training in lifelong learning. *arXiv preprint arXiv:2112.09153*, 2021. 3

[60] Nandita Naik, Christopher Potts, and Elisa Kreiss. Context-vqa: Towards context-aware and purposeful visual question answering. In *ICCV*, 2023. 2

[61] 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. 2, 3, 5

[62] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*. 2019. 5

[63] Francesco Pelosin, Saurav Jha, Andrea Torsello, Bogdan Raducanu, and Joost van de Weijer. Towards exemplar-free continual learning in vision transformers: an account of attention, functional and weight regularization. In *CVPR*, 2022. 5, 7, 1, 2

[64] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. 2020. 5

[65] Sahithya Ravi, Aditya Chinchure, Leonid Sigal, Renjie Liao, and Vered Shwartz. Vlc-bert: Visual question answering with contextualized commonsense knowledge. 2023. 3

[66] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, 2017. 3

[67] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 2015. 5

[68] Grzegorz Rypeść, Sebastian Cygert, Tomasz Trzcinski, and Bartłomiej Twardowski. Task-recency bias strikes back: Adapting covariances in exemplar-free class incremental learning. *NeurIPS*, 2025. 4, 5

[69] Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *ICLR*, 2021. 3

[70] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *ECCV*, 2022. 1

[71] Kunyu Shi, Qi Dong, Luis Goncalves, Zhuowen Tu, and Stefano Soatto. Non-autoregressive sequence-to-sequence vision-language models. In *CVPR*, 2024. 2

[72] James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *CVPR*, 2023. 3

[73] 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. *NeurIPS*, 2022. 3

[74] Miao Tian, Yushu Zhang, Yongming Zhang, Xiangli Xiao, and Wenying Wen. A privacy-preserving image retrieval scheme with access control based on searchable encryption in media cloud. *Cybersecurity*, 2024. 2

[75] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*, 2019. 5

[76] Nidhi Vyas, Shreyas Saxena, and Thomas Voice. Learning soft labels via meta learning, 2021. 4

[77] Timmy S. T. Wan, Jun-Cheng Chen, Tzer-Yi Wu, and Chu-Song Chen. Continual learning for visual search with backward consistent feature embedding. In *CVPR*, 2022. 2, 3, 6, 7, 8

[78] Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. *NeurIPS*, 2022. 3

[79] Zihan Wang, Jiuxiang Gu, Jason Kuen, Handong Zhao, Vlad Morariu, Ruiyi Zhang, Ani Nenkova, Tong Sun, and Jingbo Shang. Learning adaptive axis attentions in fine-tuning: Beyond fixed sparse attention patterns. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 916–925, Dublin, Ireland, 2022. Association for Computational Linguistics. 5

[80] Zhen Wang, Liu Liu, Yiqun Duan, Yajing Kong, and Dacheng Tao. Continual learning with lifelong vision transformer. In *CVPR*, 2022. 1

[81] 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 *ECCV*, 2022. 1, 3

[82] 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 *CVPR*, 2022. 3

[83] Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Unified coarse-to-fine alignment for video-text retrieval. In *ICCV*, 2023. 1

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

[85] 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. 6

[86] Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. Video graph transformer for video question answering. In *ECCV*, 2022. 2

[87] Yangyang Xu, Yibo Yang, and Lefei Zhang. Multi-task learning with knowledge distillation for dense prediction. In *ICCV*, 2023. 5

[88] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *NeurIPS*, 2022. 1, 2

[89] Enneng Yang, Li Shen, Zhenyi Wang, Shiwei Liu, Guibing Guo, and Xingwei Wang. Data augmented flatness-aware gradient projection for continual learning. In *ICCV*, 2023. 3

[90] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *arXiv preprint arXiv:2305.06988*, 2023. 2

[91] Razieh Nokhbeh Zaeem and K. Suzanne Barber. The effect of the gdpr on privacy policies: Recent progress and future promise. *ACM Trans. Manage. Inf. Syst.*, 2020. 2

[92] Jiliang Zhang, Chen Li, Jing Ye, and Gang Qu. Privacy threats and protection in machine learning. In *Proceedings of the 2020 on Great Lakes Symposium on VLSI*, 2020. 1

[93] Wenxuan Zhang, Paul Janson, Rahaf Aljundi, and Mohamed Elhoseiny. Overcoming generic knowledge loss with selective parameter update. In *CVPR*, 2024. 2

[94] Xiao Zhang, Yixiao Ge, Yu Qiao, and Hongsheng Li. Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification. In *CVPR*, 2021. 4

[95] Xi Zhang, Feifei Zhang, and Changsheng Xu. Vqacl: A novel visual question answering continual learning setting. In *CVPR*, 2023. 1, 2, 3, 5, 6, 7, 8

[96] Yanzhe Zhang, Xuezhi Wang, and Diyi Yang. Continual sequence generation with adaptive compositional modules. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3653–3667, Dublin, Ireland, 2022. Association for Computational Linguistics. 3

[97] Da-Wei Zhou, Hai-Long Sun, Han-Jia Ye, and De-Chuan Zhan. Expandable subspace ensemble for pre-trained model-based class-incremental learning. In *CVPR*, 2024. 1