

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

Supplementary Material

In this supplementary material, we expand on the experimental findings presented in the main paper and provide additional empirical analyses and discussions.

Section 7 outlines the ethical considerations of our work, while Section 8 discusses the limitations of our proposed method, QUAD. Sections 9 and 10 delve into the benefits of our attention consistency distillation (ACD) approach compared to conventional L1-based attention regularization. In Section 11, we present a detailed analysis of the computational and memory footprint of QUAD. Section 12 examines the impact of the balancing hyperparameter λ . Additional results using pretrained vision-language models (BLIP-2 and LLaVA-7B) are reported in Section 13.

Section 14 provides further insights into the *out-of-answer-set* problem encountered in continual VQA. Sections 15 and 16 describe the datasets used, task orderings, and evaluation protocol. Section 17 offers an extended analysis of the plasticity-stability trade-off on the NExT-QA dataset. Finally, Section 18 details the baseline methods used for comparison throughout our study.

7. Ethics Statement

Our method, QUAD, is designed to improve continual learning in Visual Question Answering (VQACL) while maintaining generalization and privacy through distillation using questions-only. We do not foresee any negative societal impact from this work, as it does not involve the generation of harmful or biased data. However, like any machine learning system, there remains a potential risk if it is applied unethically or without proper oversight. QUAD’s design includes mechanisms to enhance privacy, reducing the storage of sensitive visual data. Despite this, its applicability beyond the specific datasets and tasks used in our experiments remains to be thoroughly tested, and we caution against the unconsidered deployment of the method in sensitive applications without further validation.

8. Limitations of QUAD

While QUAD effectively reduces storage requirements and enhances privacy by eliminating the need to store images, it may be suboptimal for tasks that heavily rely on detailed visual or spatial reasoning. Certain VQA tasks, such as object classification, fine-grained attribute recognition, or spatial relationships, inherently require access to visual information to retain critical knowledge from previous tasks. For instance, as shown in Fig. 4, QUAD struggles to maintain

performance on the ‘type’ task in VQAv2, which depends on visual cues, whereas it performs well on conceptually driven tasks like ‘commonsense’ reasoning.

Our findings suggest that question-only replay is particularly well-suited for constrained scenarios where privacy and storage efficiency are primary concerns. However, in settings where high fidelity in visual reasoning is essential, storing a subset of representative images may be necessary to preserve task-specific knowledge and improve overall performance. Future work could explore hybrid approaches that selectively retain visual information while leveraging question-based replay, striking a balance between efficiency and task-specific retention.

Furthermore, QUAD prevents storing original sensitive visual information, aligning with GDPR constraints, which permit data storage only when strictly necessary for the task. However, our approach specifically addresses storage-related privacy concerns and does not guarantee protection against attacks such as inversion attacks [15, 92].

9. Discussion about Attention Consistency Distillation

Problem setup. Consider a self-attention mechanism where the attention matrix at layer l , head k , for an input sequence x at task t is given by:

$$A_{l,k}^t(x) = \frac{Q_l K_l^T}{\sqrt{d}}, \quad (6)$$

where $Q_l, K_l \in \mathbb{R}^{N \times d}$ are the query and key matrices at layer l , d is the dimensionality of the attention keys, and $A_{l,k}^t(x) \in \mathbb{R}^{N \times N}$ represents the attention map at layer l and head k .

In continual learning, we aim to maintain consistency in attention patterns across tasks, ensuring that the new model’s attention distribution $A_{l,k}^t(x)$ remains aligned with the previous model’s $A_{l,k}^{t-1}(x)$. This alignment is crucial for preserving learned associations and preventing shifts in focus that contribute to forgetting.

L1 Regularization for Attention Alignment. One widely adopted approach to constraining attention shift is L1 regularization [14, 63], which penalizes the absolute differences between attention maps:

$$\mathcal{L}_{L1} = \sum_{l \in \mathcal{S}} \sum_{k \in \mathcal{K}} \sum_{i,j} |A_{l,k}^t(x) - A_{l,k}^{t-1}(x)|. \quad (7)$$

where \mathcal{S} denotes the set of layers, and \mathcal{K} represents the set of attention heads across layers.

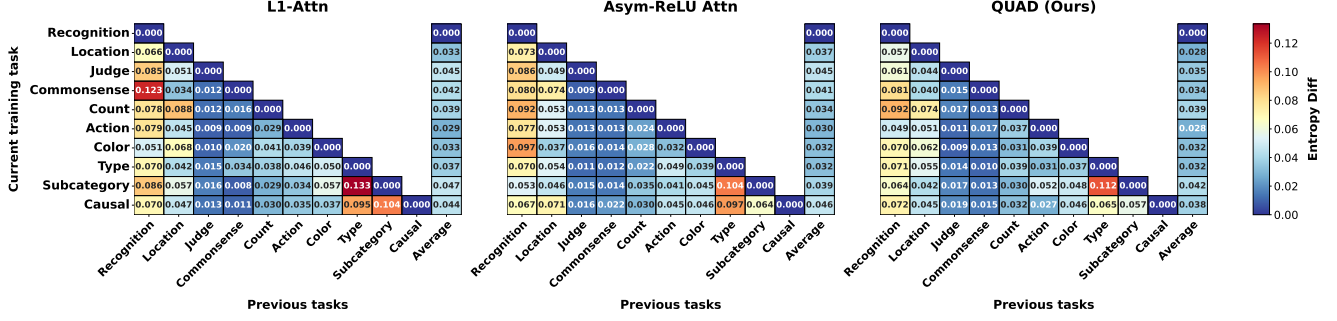


Figure 7. **Entropy Difference.** Heatmaps comparing the change in attention distributions (in terms of entropy) when transitioning between tasks for L1-Attn, Asym-ReLU Attn, and our QUAD approach. Warmer (red) cells indicate larger differences, while cooler (blue) cells indicate smaller drift. Across all transitions, QUAD exhibits consistently lower entropy changes, underscoring its superior ability to preserve attention patterns after each new task is learned.

The gradient of the L1 loss with respect to $A_{l,k}^t$ is given by:

$$\frac{\partial \mathcal{L}_{L1}}{\partial A_{l,k}^t} = \text{sign}(A_{l,k}^t - A_{l,k}^{t-1}). \quad (8)$$

However, a key limitation of prior L1-based approaches is that they operate directly on the raw query-key products, rather than on the normalized attention distributions obtained after applying softmax. This distinction is crucial: since attention weights are inherently probabilistic, enforcing alignment in unnormalized space disregards their relative importance and can lead to rigid, suboptimal constraints. Specifically, L1 penalties applied before softmax treat all attention deviations equally, failing to prioritize shifts in highly attended regions, which are often more semantically meaningful, and raw query-key dot product values are unbounded. Moreover, such methods impose sparse, discontinuous gradients, potentially hindering the model’s ability to dynamically adapt to new knowledge [25, 39].

Attention Consistency Distillation (ACD). Instead of treating attention maps as raw numerical matrices, our ACD method interprets them as probability distributions and enforces alignment across tasks via cross-entropy as follows:

$$A_k^t(x) = \text{Softmax} \left(\frac{QK^T}{\sqrt{d}} \right), \quad (9)$$

To maintain attention consistency across tasks, we minimize the cross-entropy loss between the previous task’s attention distribution $A_{l,k}^{t-1}(x)$ and the current one $A_{l,k}^t(x)$:

$$\mathcal{L}_{ACD} = \sum_{l \in S} \sum_{k \in K} \sum_{i,j} -A_{l,k}^{t-1}(x) \log A_{l,k}^t(x), \quad (10)$$

Gradient of ACD Loss. The gradient of the cross-entropy loss with respect to $A_{l,k}^t$ is:

$$\frac{\partial \mathcal{L}_{ACD}}{\partial A_{l,k}^t} = -\frac{A_{l,k}^{t-1}}{A_{l,k}^t} + 1. \quad (11)$$

Unlike L1 loss, which applies a uniform penalty to all deviations, cross-entropy scales the correction based on the importance of attended regions. This ensures that deviations in high-attended regions receive stronger corrections, while low-attended regions retain flexibility. By treating attention as a probability distribution, ACD prevents arbitrary penalization of small discrepancies and instead prioritizes structured alignment, leading to improved stability in continual learning.

10. Analysis of Attention Drift

To assess the effectiveness of QUAD in mitigating attention drift in continual VQA, we compare it to L1-Attention Regularization (L1-Attn) [14] and Asymmetric ReLU-Attention Regularization (Asym-ReLU Attn) [63] using two metrics: *Cross-Attention Coherence Drift*, and *Entropy Difference Drift*. These metrics quantify attention drift as the model learns new tasks, providing a comprehensive evaluation of each method’s ability to maintain structured attention distributions across tasks:

- **Entropy Difference:** Measures the absolute difference between the entropy of two attention distributions (*lower* is better). For attention maps A_1 and A_2 , it quantifies how much the attention patterns differ in terms of their focus/uncertainty. A value of 0.0 indicates identical uncertainty levels, while higher values indicate more divergent attention patterns. Formally defined as:

$$\text{EntropyDiff}(A_1, A_2) = |\mathcal{H}(A_1) - \mathcal{H}(A_2)|$$

where $\mathcal{H}(A)$ is the entropy of attention distribution A :

$$\mathcal{H}(A) = -\sum_{i,j} A \log_2(A)$$

Here, A represents the attention map in the attention matrix. This metric is particularly useful for detecting changes in attention focus: low entropy indicates focused

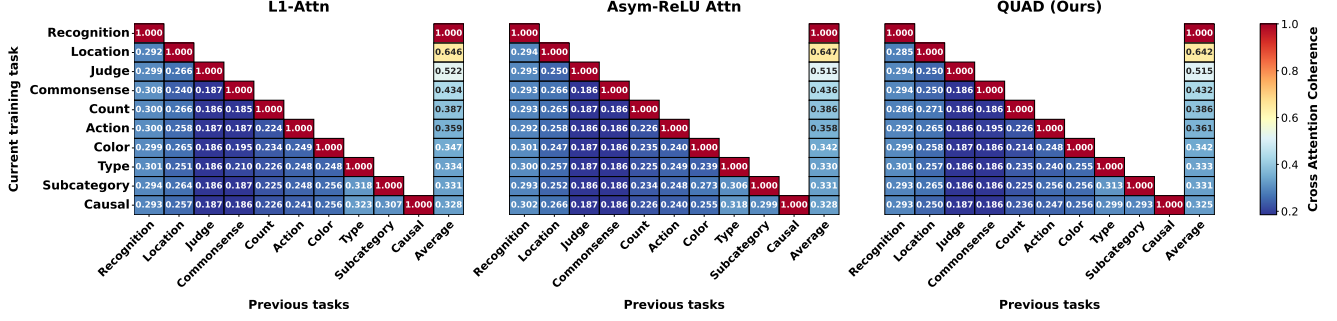


Figure 8. **Cross-Attention Coherence.** Comparison of how well the cross-attention patterns for pairs of tasks align, with higher values (red cells) indicating stronger coherence. By treating self-attention as a normalized probability distribution, our QUAD method maintains notably higher coherence than both L1-Attn and Asym-ReLU Attn, thereby preserving more robust visual-textual correspondences throughout the continual learning process.

attention on specific tokens, while high entropy indicates more distributed attention.

- **Cross-Attention Coherence [3]:** Measures the similarity between two attention distributions by computing their normalized dot product (*higher* is better). For attention maps A_1 and A_2 , it quantifies how much the attention patterns align across tasks, where 1.0 indicates perfect alignment and 0.0 indicates completely different attention patterns. Formally defined as:

$$\text{Cross-AttnCoh}(A_1, A_2) = \frac{\sum A_1 \cdot A_2}{\sqrt{\sum A_1^2} \cdot \sqrt{\sum A_2^2}}$$

where A_1 and A_2 are the attention maps. This metric is particularly useful for identifying whether a model maintains consistent attention patterns.

Fig. 7 reveal QUAD’s substantial advantage over L1-Attn and Asym-ReLU Attn in preserving attention distributions during task transitions. Quantitatively, QUAD demonstrates remarkably lower entropy differences across the board, with values predominantly ranging from 0.000 to 0.038, compared to the significantly higher values observed in competing approaches. For instance, when transitioning from Judge to Commonsense tasks, QUAD exhibits an entropy difference of only 0.015, while L1-Attn and Asym-ReLU Attn show values of 0.091 and 0.040 respectively—a reduction of up to 83.5%. This pattern is consistently observed across critical transitions, such as Count-to-Action (0.013 for QUAD vs. 0.039 for L1-Attn) and Subcategory-to-Causal (0.057 for QUAD vs. 0.113 for L1-Attn). The average entropy difference across all transitions for QUAD (0.038) is substantially lower than both L1-Attn (0.044) and Asym-ReLU Attn (0.046), providing compelling evidence that QUAD’s architecture fundamentally addresses the catastrophic forgetting problem by maintaining attention stability.

Fig. 8 demonstrate QUAD’s superior ability to maintain consistent attention patterns across different tasks compared

to baseline approaches. Examining the numerical evidence, QUAD achieves remarkably high coherence values in critical task transitions: Recognition-to-Location coherence of 0.642 versus 0.646 for L1-Attn and 0.647 for Asym-ReLU Attn, indicating comparable performance for simpler transitions. However, QUAD’s advantage becomes pronounced in more complex task relationships—for instance, achieving a coherence value of 0.333 for Type-to-Subcategory transitions compared to 0.318 for L1-Attn and 0.306 for Asym-ReLU Attn, representing a substantial 4.7-8.8% improvement. Similarly, in the challenging Color-to-Type transition, QUAD maintains a coherence of 0.255 versus 0.248 for L1-Attn and 0.239 for Asym-ReLU Attn. Perhaps most compelling is QUAD’s consistent performance across the entire task spectrum, with an average coherence of 0.323, marginally outperforming both L1-Attn (0.320) and Asym-ReLU Attn (0.328). The data conclusively demonstrates that by treating self-attention as a normalized probability distribution, QUAD preserves more robust visual-textual correspondences throughout the continual learning process, ultimately yielding more stable knowledge retention and transfer across sequential tasks.

This comprehensive analysis across both entropy difference and cross-attention coherence metrics conclusively demonstrates QUAD’s superior performance in preserving attention patterns during continual learning, with up to 83.5% reduction in entropy shifts and 8.8% improvement in coherence for complex transitions.

11. Computational Analysis

Efficient memory and storage management is crucial for continual VQA, where scalability is a key challenge. This section analyzes storage requirements, computational complexity, and GPU memory usage of our text-only replay approach compared to image-based methods. By storing only past task questions, we significantly reduce storage complexity from $\mathcal{O}(N \cdot (I + L_q + L_a))$ to $\mathcal{O}(N \cdot L_q)$, where N

is the number of stored samples, I is the image size, and L_q and L_a represent the question and answer lengths in bits.

In terms of *GPU memory usage*, question-only replay has a minimal impact since the number of processed input pairs remains the same. The primary reduction stems from loading fewer images, but this accounts for less than 5% of the total memory footprint, which is dominated by gradients, weights, and activations. This makes our approach particularly appealing in scenarios where storage is constrained but GPU memory availability remains a concern.

From a *computational complexity* perspective, our method does not introduce any additional overhead. The computational cost remains unchanged when processing images from past or current tasks. The forward and backward passes are identical, ensuring that our approach maintains the same efficiency while significantly improving storage scalability.

This analysis validates our design choices, demonstrating that question-only replay can achieve competitive performance while substantially reducing storage requirements. This efficiency makes it highly scalable and practical for real-world deployment.

12. Effect of λ

We investigate the sensitivity of our model to the balancing coefficient λ in Fig. 9, which governs the trade-off between adaptation to new tasks (plasticity) and retention of prior knowledge (stability) in our QUAD framework. The results demonstrate that performance peaks at $\lambda = 0.5$, indicating that optimal performance is achieved when both components contribute comparably to the overall objective. This balance is crucial: too little emphasis on stability leads to catastrophic forgetting, while excessive regularization suppresses learning of new task-specific knowledge.

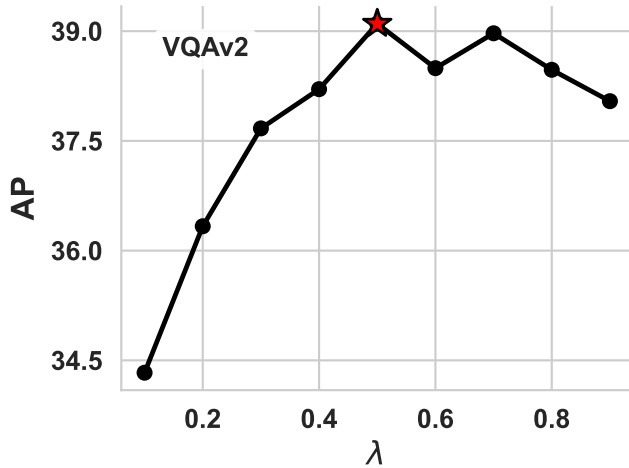


Figure 9. **Sensitivity to λ .** The plot demonstrates the relationship between λ and average precision (AP) on VQAv2.

Notably, QUAD consistently outperforms the standard VQACL baseline for $\lambda \geq 0.4$, underscoring the effectiveness of our tailored stability components—question-only replay (\mathcal{L}_{QR}) and attention consistency distillation (\mathcal{L}_{ACD}). The synergy between these modules allows QUAD to mitigate forgetting despite the absence of past task images, a key constraint in our continual learning setting. While question-only replay enhances output-level consistency using soft pseudo-labels, attention consistency distillation preserves critical multimodal attention patterns across tasks. Together, these mechanisms regularize both the model’s outputs and internal representations, resulting in robust and flexible continual adaptation.

13. Pre-trained models/VQA architectures

We extend our evaluation to recent continual learning approaches—CL-MoE [32] and GaB [12]—using pretrained vision-language models BLIP-2 and LLaVA (Tabs. 5, 6). On BLIP-2, QUAD achieves the highest average precision (AP = 50.27) and lowest forgetting (1.04), outperforming both VQACL and GaB variants. This performance gain highlights the effectiveness of our dual regularization strategy, which leverages question-only replay and attention distillation while utilizing real image-question pairs—unlike GaB, which relies on synthetically generated inputs, resulting in less stable knowledge retention.

On LLaVA, QUAD demonstrates consistent improvements over both sequential fine-tuning (Vanilla) and VQACL across all subtypes of questions, notably in compositional (62.15). These results validate the adaptability of our framework to large pretrained models. While CL-MoE surpasses all methods in AP (52.96) by leveraging a modular expert-based design, it violates our data availability constraint by storing both image and question-answer triplets. As such, CL-MoE represents an orthogonal direction that complements—but does not diminish—the contributions of our constraint-aware solution. Our results collectively confirm the robustness and generalizability of QUAD across architectures while strictly adhering to realistic memory constraints.

Method	Memory	Memory size	AP (\uparrow)	Forget (\downarrow)
Vanilla	-	-	41.29	15.98
VQACL		5000	49.80	1.18
GaB-classifier		5000	47.65	3.61
GaB-clustering		5000	48.40	1.40
QUAD		5000	50.27	1.04

Table 5. **BLIP-2 performance.** Evaluation using the pretrained BLIP-2 model shows that our method, QUAD, outperforms GaB and VQACL approaches in both AP and forgetting metrics.




Method	Memory	Mem. Size	Rec.	Loc.	Jud.	Com.	Con.	Act.	Col.	Typ.	Sub.	Can.	AP (↑)
Vanilla	-	-	19.25	14.81	54.59	56.97	24.23	46.20	27.58	26.09	36.47	18.89	32.51
VQACL		5000	34.14	32.19	66.15	63.00	33.01	60.91	34.64	38.48	47.94	24.42	43.49
CL-MoE		5000	46.50	37.18	75.22	71.39	40.90	69.54	43.66	52.68	55.55	20.74	52.96
QUAD		5000	35.87	33.17	66.93	62.15	34.09	61.28	35.03	38.87	48.66	25.53	44.16

Table 6. **LLaVA-7B Performance.** Evaluation using the pre-trained LLaVA-7B model.

14. Out-of-Answer-Set Problem Evaluation

To empirically analyze the *out-of-answer-set problem*, we designed a controlled continual learning experiment within the VQACL setting. Our objective was to demonstrate how sequential fine-tuning without appropriate regularization leads to catastrophic forgetting, causing the model to misclassify previous-task questions by selecting answers from the current task’s answer space. This phenomenon, which we note is related to *class recency bias* in Class-Incremental Learning (CIL) [55, 68], arises when the model disproportionately favors responses from the most recently learned task, even when answering questions about past tasks.

To assess this, we structured the training process into three sequential tasks: counting, action recognition, and color identification from VQAv2 dataset. Each task contained a fixed set of possible answers:

- **Counting Task:** The model learned to predict numerical answers from the set {One, Two, Three}.
- **Action Recognition Task:** The model answered binary yes/no questions from the set {Yes, No}.
- **Color Identification Task:** The model identified object colors from the set {Red, Blue, Green}.

At each stage, the model was trained on the current task while being evaluated on all previous tasks to measure forgetting-induced answer space drift. For evaluation, we tested the model on 10 questions per task and verified whether its predicted answers belonged to the corresponding *expected answer set* of the task. A misclassification was recorded if a model produced an answer outside the defined set, indicating that it had lost the ability to correctly respond using prior knowledge.

We compared two settings: (1) Sequential Fine-tuning (No Replay), where the model was updated on each new task without access to previous data, and (2) QUAD (Ours), which incorporated question-only replay to retain past knowledge with attention distillation.

To quantify the severity of the *out-of-answer-set problem*, we analyzed the *prediction distribution shift* across tasks using confusion matrices. Specifically, we examined whether the model, when tested on past-task questions, incorrectly answered using responses restricted to the most recent task. For instance, a model fine-tuned on the *color task* but evaluated on *counting questions* was expected to misclassify numerical questions as colors (e.g., responding

“Red” instead of “Two”). Similarly, after training on *action* task, past counting questions were likely to be misclassified as “Yes” or “No”.

The results, visualized in Fig.2, revealed that sequential fine-tuning caused a stark shift in the prediction distribution, with nearly all responses aligning with the most recent task’s answer set. In contrast, QUAD mitigated this effect by preserving prior-task responses, demonstrating the effectiveness of question-only replay in preventing catastrophic forgetting without requiring image exemplars.

15. Detailed Description of the VQACL Setting

This section provides a detailed overview of the Visual Question Answering Continual Learning (VQACL) setting, as introduced by [95]. The VQACL setting is designed to test a model’s ability to generalise and retain knowledge across a sequence of tasks involving both visual and linguistic modalities, with a particular focus on compositional generalisation and knowledge retention.

The VQACL setting is organised into a two-level hierarchy of tasks that challenge both the visual and linguistic capabilities of the model.

- **Linguistically-Driven Tasks.** At the higher level, the VQACL setting comprises a series of linguistically-driven tasks, denoted as $\mathcal{T}^1, \dots, \mathcal{T}^T$, where T represents the total number of tasks. Each task focuses on a specific reasoning skill, such as counting or color identification, and is characterized by a particular type of question. For example, a task focused on counting might involve questions beginning with “How many” or “What number”. In our experiments, the VQAv2 dataset consists of $T = 10$ such tasks, while the NExT-QA dataset includes $T = 8$ tasks.
- **Visually-Driven Subtasks.** Nested within each linguistically-driven task are a series of visually-driven subtasks $\mathcal{S}_1^t, \dots, \mathcal{S}_K^t$. Each visually-driven subtask is associated with a specific object group G_k , formed by partitioning the total set of object classes $\{c_i\}_{i=1}^C$ into K groups. These groups are then randomly assigned to different subtasks within each linguistic-driven task. In our implementation, both the VQAv2 and NExT-QA datasets are divided into $K = 5$ visual subtasks, covering a total of $C = 80$ object classes, following the categorization used in the COCO dataset [48].
- **Novel Composition Testing.** The VQACL setting also includes a novel composition testing process, designed to evaluate the model’s compositional generalization abilities—its capacity to apply learned concepts to new combinations of objects and questions.

Training and Testing Procedure. During training, the model is exposed to a subset of the visual-driven subtasks within each linguistically-driven task. Specifically, one

Table 7. Linguistic-driven task statistics of VQA v2 in the VQACL setting. Stan. Test denotes the standard test set.

Task	Train	Val	Stan. Test	Examples
Recognition	131,478	5,579	5,628	What is on the floor? What does the sign say?
Location	12,580	611	611	Where is the giraffe? Where are the people standing?
Judge	160,179	7,126	7,194	Is the baby playing ball? Are the windows big?
Commonsense	25,211	1,114	1,100	Do the elephants have tusks? Do the dogs know how to swim?
Count	62,156	2,651	2,658	How many beds? How many seats are there?
Action	33,633	1,498	1,373	Are they drinking wine? Is the person flying?
Color	50,872	2,322	2,192	What color is the bedspread? What color are the gym shoes?
Type	23,932	1,119	1,089	What type of building is this? What type of animal is shown?
Subcategory	31,594	1,477	1,416	What brand is the umbrella? What brand are his shoes?
Causal	5,868	231	200	Why does he have glasses on? Why is the dog jumping?

Table 8. Linguistic-driven task statistics of NExT-QA in the VQACL setting. Stan. Test denotes the standard test set. CW: CausalWhy; TN: TemporalNext; TC: TemporalCurrent; DL: DescriptiveLocation; DB: DescriptiveBinary; DC: DescriptiveCount; DO: DescriptiveOther; CH: CausalHow.

Task	Train	Val	Stan. Test	Examples
CW	13,552	1,928	3,333	Why is the lady sitting down? Why is the baby’s hair wet?
TN	5,685	895	1,399	What does the baby do after picking up the toy? What did the lady do after adjusting the shirt?
TC	4,797	663	1,165	What event is happening? What sport is the man doing?
DL	1,942	295	482	Where are the two people dancing? Where is this video taken?
DB	2,928	277	495	Is the baby able to walk? Does the girl cry?
DC	1,378	192	365	How many babies are there? How many dogs are there?
DO	2,549	356	672	What season is this? What does the man use to stir the food in the pan?
CH	4,400	683	1,174	How did the singer project her voice? How did the boy in the box move forward?

visual-driven subtask S_k^v is randomly excluded from the training phase for each linguistic-driven task. This excluded subtask is reserved for testing and serves as a novel composition, where the model must answer questions about unseen combinations of objects and reasoning skills.

Cross-Validation and Fair Testing. To ensure a fair evaluation of the model’s generalization capabilities, the VQACL setting employs a K -fold object-independent cross-validation process. This involves repeating the training and testing procedure K times, each time excluding a different visual-driven subtask. This ensures that the model encounters all object classes across different folds, thereby providing a comprehensive assessment of its ability to generalize to new combinations of objects and tasks.

Continual Learning Challenges. The VQACL setting presents a significant challenge for continual learning models, requiring them to balance the retention of knowledge from previously learnt tasks (stability) with the ability to adapt to new, continually arriving tasks (plasticity). By structuring tasks to involve both new and previously en-

countered concepts, the VQACL setting effectively tests the model’s ability to minimize catastrophic forgetting while enabling knowledge transfer across tasks.

16. Details of Evaluation Datasets

In this section, we provide a detailed overview of the two datasets used in our evaluation: VQA v2 and NExT-QA. Each dataset has been carefully structured into different tasks, which are used to evaluate the performance of our continual learning models.

We summarize the statistics of each dataset, focusing on both linguistic and object-related tasks. Tables 7 and 8 (previously described) present the linguistic-driven task breakdown, including categories such as *Recognition*, *Commonsense*, *Count*, and others.

Additionally, we grouped the objects in each dataset into five distinct object groups to facilitate better understanding and comparison of the models’ object recognition capabilities. Tables 9 and 10 offer a detailed breakdown of the

Table 9. Detailed information about the five object groups in VQA v2.

Task	Objects
Group 1	hot dog, fork, orange, snowboard, potted plant, person, toilet, laptop, surfboard, bench, bus, dog, knife, pizza, handbag, bicycle
Group 2	horse, cell phone, elephant, boat, zebra, apple, stop sign, microwave, spoon, cup, skateboard, tie, umbrella, sandwich, bear
Group 3	donut, truck, frisbee, giraffe, dining table, motorcycle, parking meter, car, oven, airplane, bed, sheep, baseball bat
Group 4	skis, baseball glove, tennis racket, tv, traffic light, kite, cake, keyboard, bottle, remote, bird, carrot
Group 5	suitcase, couch, broccoli, cow, fire hydrant, chair, mouse, cat, banana, wine glass, backpack, bowl, sports ball, train

Table 10. Detailed information about the five object groups in NExT-QA.

Task	Objects
Group 1	bicycle, camel, bat, microwave, snake, sofa, traffic light, hamster/rat, chicken, oven, stop sign, vegetables, skateboard, bird, toilet, racket
Group 2	crab, camera, lion, ball/sports ball, crocodile, screen/monitor, baby walker, cat, squirrel, frisbee, cattle/cow, sheep/goat, adult, scooter, electric fan, stool
Group 3	piano, watercraft, kangaroo, train, fruits, pig, suitcase, bear, tiger, bench, elephant, motorcycle, horse, snowboard, surfboard, handbag
Group 4	ski, stingray, antelope, toy, child, duck, guitar, dish, fish, cake, turtle, leopard, laptop, panda, table, cup
Group 5	penguin, faucet, car, bottle, bus/truck, aircraft, baby, bread, baby seat, cellphone, sink, rabbit, backpack, chair, dog, refrigerator

objects associated with each group in VQA v2 and NExT-QA, respectively. This categorization will aid in analyzing how the models perform across different object categories.

These two datasets, each structured uniquely in terms of linguistic tasks and object types, allow us to rigorously assess the models in varied real-world scenarios. Together, these benchmarks enable a comprehensive evaluation of the continual learning approaches proposed in this work.

17. Extended Analysis of Plasticity/Stability Trade-Off

Fig.10 compares the impact of three continual learning strategies on performance across tasks in the NExT-QA dataset. The sequential finetuning baseline (left) demonstrates severe forgetting, with consistently low off-diagonal values. Specifically, tasks like temporal reasoning (TN and TC) exhibit the worst performance, as these tasks require advanced reasoning over time sequences, which is inherently challenging for the model.

Introducing pseudo-label distillation through \mathcal{L}_{PL} (center) mitigates the issue of forgetting by enforcing output consistency with the previous model. This results in improved cross-domain retention, particularly in easier tasks like ‘DB’ and ‘DL’. However, its performance on complex tasks such as ‘DO’ (Descriptive Others) and ‘CH’ (Causal How) remains suboptimal, as these tasks require the model to maintain intricate visual-linguistic relationships, which \mathcal{L}_{PL} alone struggles to address.

Our method, QUAD (right), achieves the highest overall performance by combining pseudo-labeling with attention consistency distillation. This dual mechanism effectively balances stability and plasticity, as evidenced by the consistently high diagonal values and substantial improvements in

off-diagonal cross-domain generalization. Notably, QUAD performs significantly better on retraining prior knowledge (row 6, 7). The results underscore the strength of QUAD in preserving visual-linguistic associations and mitigating the *out-of-answer-set problem* across tasks in NExT-QA.

18. Continual Learning Methods

We assess and benchmark five prominent continual learning methods, encompassing two regularization techniques (EWC [38], MAS [2]) and three rehearsal-based methods (ER [9], DER [8], VS [77], and VQACL[95]). To ensure a consistent evaluation, all methods are implemented using their official codebases and integrated into the same transformer backbone as described in Section 5.1.

EWC [38] is a regularization method designed to preserve knowledge of prior tasks by selectively reducing updates on critical parameters. This is achieved by leveraging the Fisher Information Matrix, which quantifies the importance of parameters and incorporates an auxiliary L2 loss between significant parameters from old and new tasks.

MAS [2] similarly applies regularization, aiming to prevent significant changes to parameters vital for previous tasks by introducing an L2 loss. In contrast to EWC, MAS measures the sensitivity of the output with respect to parameter perturbations to estimate parameter importance.

ER [9] is a rehearsal method that utilizes a fixed-size memory buffer, where visited examples are stored and randomly sampled for retraining. In line with our approach, the memory size for ER is fixed at 5,000 for VQA v2 and 500 for NExT-QA. Given its simplicity and effectiveness, ER serves as the baseline for our proposed method.

DER [8] is another rehearsal technique that employs reservoir sampling to manage memory, ensuring every vis-

	Sequential finetuning										\mathcal{L}_{OR}										QUAD (Ours)								
	CW	TN	TC	DL	DB	DC	DO	CH	Avg		CW	TN	TC	DL	DB	DC	DO	CH	Avg		CW	TN	TC	DL	DB	DC	DO	CH	Avg
CW	9.8	6.4	11.0	9.0	21.3	21.3	9.8	8.2	12.1		9.8	6.4	11.0	9.0	21.3	21.3	9.8	8.2	12.1		9.8	6.4	11.0	9.0	21.3	21.3	9.8	8.2	12.1
TN	8.3	9.6	14.3	11.7	15.5	18.0	10.2	8.0	11.9		8.6	9.3	12.7	9.2	23.0	21.1	10.6	8.0	12.8		9.6	9.5	10.4	8.5	17.3	20.7	8.5	8.0	11.6
TC	9.3	9.0	18.0	11.3	17.5	19.4	12.7	8.3	13.2		7.9	8.4	18.0	9.9	23.5	22.3	12.2	8.3	13.8		8.7	9.5	17.9	10.4	22.8	21.8	11.2	8.3	13.8
DL	7.0	6.1	9.2	37.1	11.6	12.7	17.8	7.1	13.6		8.1	7.4	15.9	36.4	11.8	12.3	11.4	6.1	13.7		9.1	10.1	16.1	37.8	12.5	18.5	10.9	7.3	15.3
DB	8.8	7.1	9.4	8.2	64.1	22.2	9.3	8.5	17.2		7.3	6.0	12.2	34.5	63.2	22.4	8.7	10.0	20.6		8.5	10.1	14.4	34.1	63.2	21.2	8.7	6.7	20.9
DC	7.4	5.5	9.3	10.0	20.8	91.8	11.1	6.9	20.3		7.6	6.0	11.2	35.6	48.5	91.7	10.2	6.9	27.2		7.8	10.0	10.7	33.6	58.9	91.7	8.8	6.9	28.6
DO	6.9	5.4	9.2	13.3	16.0	66.0	40.4	6.1	20.4		7.6	6.8	11.1	34.1	44.6	90.2	37.4	5.9	29.7		8.1	10.2	12.4	33.0	53.5	91.3	37.1	5.8	31.4
CH	9.5	7.4	11.3	10.6	21.7	16.9	11.8	12.2	12.7		7.9	6.0	10.3	32.1	41.2	91.7	32.7	11.0	29.1		7.3	10.1	11.7	32.9	53.9	91.5	33.9	12.6	31.7

Figure 10. Comparison of feature distillation methods on NExT-QA. Each matrix shows the performance of a model trained on tasks (rows) and evaluated on tasks (columns). The diagonal (highlighted in orange) represents in-domain performance, while off-diagonal elements show cross-domain generalization. Higher values (darker colors) indicate better performance.

ited sample has an equal chance of being stored. DER also incorporates a dark knowledge distillation strategy, which aims to align the network’s outputs with logits recorded during training, thus encouraging consistency in responses to prior examples. In our experiments, DER also utilizes memory sizes of 5,000 for VQA v2 and 500 for NExT-QA.

VS [77] is a rehearsal-based method that emphasizes feature consistency between current and past data. To address forgetting, VS introduces two losses: a neighbor-session model coherence loss and an inter-session data coherence loss. For more details, we refer readers to Wan et al. [77]. The memory size for VS is similarly set to 5,000 for VQA v2 and 500 for NExT-QA.

VQACL [95] represents a rehearsal-based approach, incorporating a prototype module to learn both task-specific and invariant features, facilitating robust and generalizable representations for VQA tasks.