Supplementary File of Re-evaluating Continual VQA: Toward Fair and Robust Evaluation for Multimodal Continual Learning

Zijian Gao[1,2]†, Zicheng Sun[3]†, Xingxing Zhang[4]*, Kele Xu[1,2]*, Huaimin Wang[1,2]
[1]College of Computer Science and Technology, National University of Defense Technology
[2]State Key Laboratory of Complex & Critical Software Environment
[3]School of Computer Science and Technology, Beijing Jiaotong University
[4]School of Computer Science, Tsinghua University
{gaozijian19, xukelele, hmwang}@nudt.edu.cn, zichengsun@bjtu.edu.cn, xxzhang1993@gmail.com

Algorithm 1 MaDQ Training Procedure

Input: Pre-trained model W^0, epochs E, EMA momentum α
Data: Continual tasks $\{\mathcal{T}^t\}_{t=1}^n$
Output: Adapted model $\{W^t\}$

Initialize: classification head h_{cls}, IQM head h_{IQM}, LoRA-w $\{A, B\}$, LoRA-m $\{\bar{A}, \bar{B}\}$, memory buffer $\mathcal{M}$
for $t = 1$ to n **do**
  for $e = 1$ to E **do**
    while triplets in $\mathcal{T}^t$ not exhausted **do**
      Sample (x^t, q^t, a^t) and replay $q^i \sim \mathcal{M}$
      $W \leftarrow W^0 + BA$
      Compute $\phi^t(x^t, q^t)$ and $\mathcal{L}_{\text{TSA}}$ (Eq. 3)
      if $t = 1$ **then**
        Update $\{A, B\}$ and h_{cls} w.r.t. $\mathcal{L}_{\text{TSA}}$
      else
        Compute $\phi^t(x^t, q^i)$, $\phi^{t-1}(x^t, q^i)$
        Construct $q^j \in \{q^t, q^i\}$ with label y
        Compute $\psi^t(x^t, q^j)$ and $\psi^{t-1}(x^t, q^j)$
        Compute losses $\mathcal{L}_{\text{APD}}, \mathcal{L}_{\text{IQM}}, \mathcal{L}_{\text{MCD}}$
        Update $\{A, B\}$ via total loss (Eq. 2)
    EMA update: $\bar{A} \leftarrow \alpha\bar{A} + (1-\alpha)A$, $\bar{B} \leftarrow \alpha\bar{B} + (1-\alpha)B$
    Frozen model: $W^t \leftarrow W^0 + \bar{B}\bar{A}$
  Add $\mathcal{T}^t$ questions to memory buffer $\mathcal{M}$

1. Training Procedure and Implementation Details

In this section, we present the training procedure in Algorithm 1 and implementation details of our benchmark and experimental setups.

*Corresponding authors. † Equal Contribution.

Model Architecture: Our implementation is built upon the official BLIP [9] and ZAF [4] repositories. All experiments initialize from the BLIP model architecture and its released pre-trained weights. The image encoder f_ν is a ViT-B/16, while the question encoder f_τ is a 12-layer BERT with a hidden size of 768. The answer decoder f_ω is also a 12-layer Transformer. Following prior work, the classification head h_{cls} is trained only on the first task $\mathcal{T}^1$ and then kept frozen for all subsequent tasks to avoid label-space entanglement. For all baselines, we follow the implementation protocol established in ZAF [4], which has shown strong performance in SVLC settings. This ensures consistent comparison across methods under our continual VQA framework.

Details on Baselines: In implementing LwF [11], we adapt it to the LoRA-based fine-tuning setting rather than updating the full model. For each incremental task, we retain the LoRA weights learned from the previous task and use them for the distillation process. For ZAF [4], we follow a wild-data strategy aligned with the original design. Specifically, when evaluating on VQA v2, we use GQA as the wild data; conversely, when evaluating on GQA, we use VQA v2 as the wild data. For MoE-Adapters [12], originally designed for image classification with task-id–dependent routing, we adapt the method to our VL task where task boundaries are unavailable. To this end, we employ a single task-dependent router shared across all tasks. Following the original work, we use 14 experts, and the router selects the top-4 experts during both training and inference.

Benchmark Details: Our proposed Uco-VQA benchmark is built upon two widely used visual question answering datasets: VQA v2 [5] and GQA [7]. For VQA v2, following the protocol of VQACL [13], we divide the dataset into 8 incremental learning tasks according to question semantics: `location`, `commonsense`, `type`, `action`, `color`,

Table 1. Task Division of VQA v2 [5].

Task	Question Type
Judge	are there, are there any, are these, is that a, is there, is there a, is this, is this a, is this an
Location	what room is, where are the, where is the
Action	are they, is he, is the man, is the person, is the woman, is this person, what is the man, what is the person, what is the woman
Color	what color, what color are the, what color is, what color is the, what is the color of the
Commonsense	can you, could, do, do you, does the, does this
Type	what kind of, what type of
Recognition	what are, what are the, what does the, what is, what is in the, what is on the, what is the, what is the name, what is this
Count	how many, how many people are, how many people are in, what number is

Table 2. Task Division of GQA [7].

Task	Detailed Semantic Annotations
Material	twoSameMaterial, materialChoose, materialVerify, material, materialVerifyC, twoSameMaterialC
Position	positionVerifyC, positionVerify, positionQuery
Action	activityChoose, activity, activityWho
Color	directOf, directWhich
Logical	diffAnimals, comparativeChoose, twoCommon, twoSameC, sameAnimalsC, sameGenderC, sameAnimals, twoDifferentC, sameGender, twoSame, twoDifferent, diffAnimalsC
Object	category, categoryThisChoose, objThisChoose, categoryThis, place, placeChoose
Relation	relS, relVerifyCr, relChooser, sameRelate, categoryRelOChoose, categoryRelS, relVerify, positionChoose, dir, relVerifyCo, relO, categoryRelO

`count`, `recognition`, and `judge`. For GQA, leveraging its fine-grained structural annotations, we group samples into 7 semantic tasks: `action`, `material`, `logical`, `object`, `color`, `position`, and `relation`. The full partitioning criteria are detailed in Table 1 and Table 2. For constructing our Uco-VQA continual splits, we also incorporate VQA-CP v2 [1] and GQA-OOD [8], where we retain the same task granularity (8 for VQA and 7 for GQA) by mapping question IDs from our standard splits.

2. Extended Results

In this section, we provide some extended results for the main text.

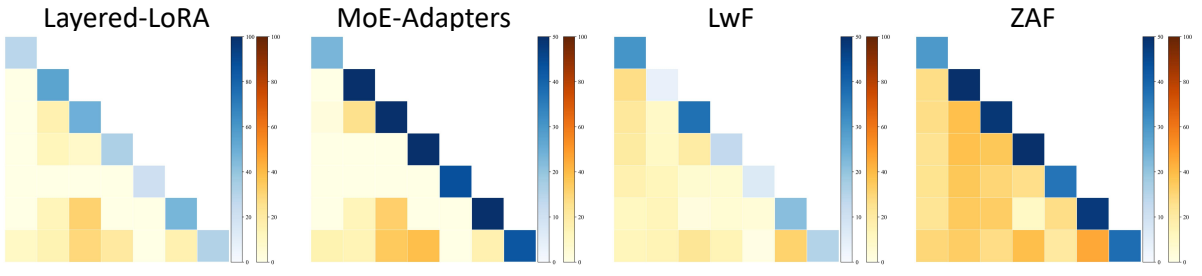
2.1. Extended Evidence of Spurious Anti-Forgetting Performance

In the main text, we compare continual learning methods on both the original GQA dataset and the debiased GQA v2 (SS) dataset. Figure 1 further presents the performance heatmaps of representative approaches on the original GQA benchmark [13] and our proposed GQA v2 (SS). Consistent with the main findings, these heatmaps demonstrate that anti-forgetting performance is substantially overestimated when the answer space remains unmodified. The seemingly strong knowledge retention exhibited by many methods is largely attributable to inter-task overlap in the answer vocabulary, which allows models to reuse memorized answer priors rather than truly learning task-specific visual–semantic knowledge, thereby artificially inflating their resistance to forgetting.

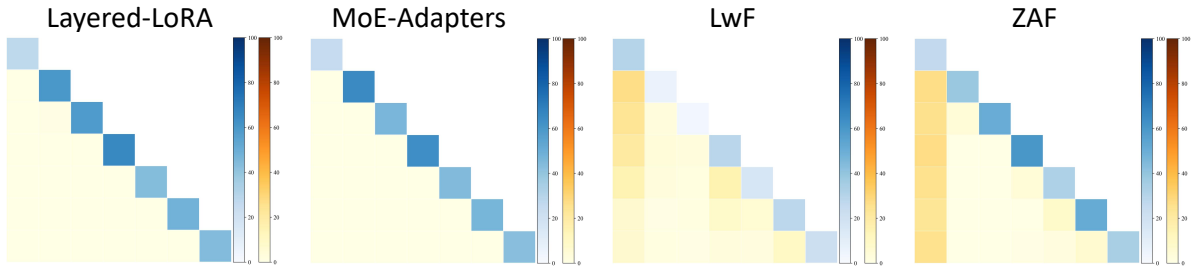
Meanwhile, Figure 2 compares model performance when trained on the original VQA v2 benchmark and on our proposed VQA v3 benchmark. Again consistent with the conclusions in the main text, models trained on VQA v2 achieve markedly higher FAA and CAA scores, together with lower FFM. However, this apparently superior performance is primarily driven by shared answer tokens across tasks, enabling models to rely on overlapping lexical priors instead of preserving genuine task-specific knowledge. This results in an exaggerated perception of anti-forgetting ability. Notably, MaDQ maintains comparable FAA/CAA on both benchmarks while consistently achieving low FFM, indicating stronger knowledge retention and enhanced robustness to distributional shifts.

2.2. More Results on Large Language Model

In the main text, we report results when replacing BLIP with the larger vision–language model BLIP2 [10] on the VQA



(a) GQA Dataset



(b) GQA v2 Dataset

Figure 1. Performance heatmaps on the GQA and GQA v2 (SS) benchmarks using four representative continual learning methods. Each heatmap visualizes the accuracy of a model trained up to task i and evaluated on task j. Darker diagonal blocks indicate strong within-task learning, while the decay of off-diagonal values reflects catastrophic forgetting. Compared to the original GQA dataset, the proposed GQA v2 (SS) split produces substantially lower off-diagonal scores—revealing that the seemingly strong knowledge retention observed on the original benchmark is largely attributed to inter-task overlaps in answer vocabulary, rather than genuine resistance to forgetting.

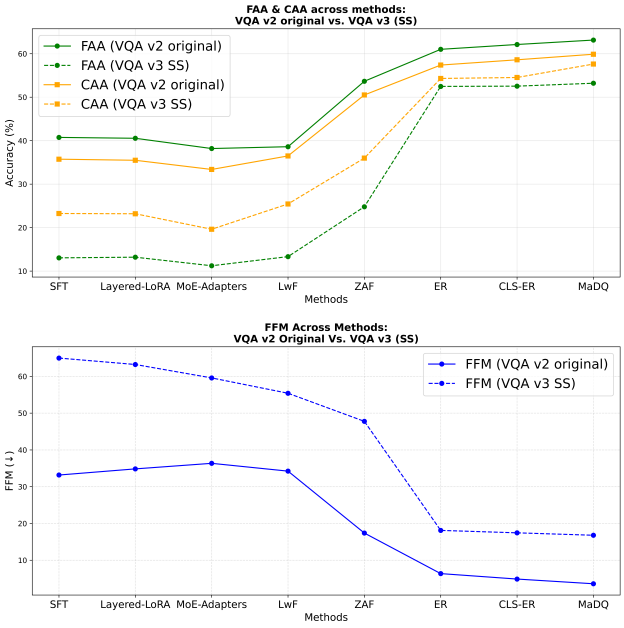


Figure 2. Comparison of CL methods on the original VQA v2 and the debiased VQA v3 (SS). (**Top**) FAA and CAA; (**Bottom**) FFM.

v3 (SS) dataset. Here, we further present BLIP2-based performances on the GQA v2 (SS) dataset. As shown in Table 3, we observe a consistent trend: BLIP2-based models achieve noticeably lower FAA and CAA than their BLIP counterparts. Despite this, our MaDQ* still remains competitive,

Table 3. Performance comparison on GQA v2 (SS) using the larger vision–language model BLIP2 [10].

Metric	SFT	Layered-LoRA	LwF	ER	CLS-ER	MaDQ*
FAA (↑)	11.85	12.32	26.04	30.89	34.08	31.69
CAA (↑)	24.43	25.73	30.18	35.87	36.94	36.62
FFM (↓)	41.98	40.04	5.73	8.51	4.89	6.27

Table 4. Performance comparison across different α values and baseline methods on GQA v2 (SS) dataset.

Metric	EMA Parameters α				
	0.75	0.80	0.85	0.90	0.95
FAA (↑)	40.57	42.16	43.51	37.41	33.48
CAA (↑)	42.32	43.51	43.00	33.91	31.34
FFM (↓)	9.80	9.72	6.96	4.41	3.45

matching ER and CLS-ER without storing ground-truth answers. Together, these findings confirm that our lightweight replay-and-distillation strategy is both effective and generally applicable.

2.3. Hyperparameter Analysis

Table 4 reports the performance of MaDQ on GQA v2 (SS) under different EMA coefficients, with $\alpha \in [0.75, 0.95]$, and compares it to existing baselines. When α lies between *0.75* and *0.85*, MaDQ consistently matches or surpasses

all competitors across metrics, demonstrating strong adaptability to this hyperparameter. In contrast, for larger values (*0.90–0.95*), MaDQ exhibits reduced forgetting, but the overly large EMA coefficient slows parameter updates and harms plasticity, leading to a noticeable drop in overall performance; consequently, both FAA and CAA decrease significantly. Overall, MaDQ remains robust over a broad range of α, while very large EMA coefficients may require more training epochs to recover plasticity.

2.4. Task Order Analysis

Table 5 compares the performance of various CL methods on GQA v2 (PS) under three distinct task orders: $action \rightarrow material \rightarrow logical \rightarrow object \rightarrow color \rightarrow position \rightarrow relation$, $relation \rightarrow position \rightarrow color \rightarrow object \rightarrow logical \rightarrow material$, and $relation \rightarrow material \rightarrow action \rightarrow color \rightarrow logical \rightarrow position \rightarrow object$. Our method, MaDQ, exhibits consistent robustness across all task orders, significantly outperforming other CL methods. These results highlight MaDQ's strong adaptability and effectiveness to task permutations.

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4971–4980, 2018. 2

[2] Elahe Arani, Fahad Sarfraz, and Bahram Zonooz. Learning fast, learning slow: A general continual learning method based on complementary learning system. In *International Conference on Learning Representations*, 2022. 5

[3] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019. 5

[4] Zijian Gao, Xingxing Zhang, Kele Xu, Xinjun Mao, and Huaimin Wang. Stabilizing zero-shot prediction: A novel antidote to forgetting in continual vision-language tasks. In *Advances in Neural Information Processing Systems*, pages 128462–128488, 2024. 1, 5

[5] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 1, 2

[6] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021. 5

[7] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 1, 2

[8] Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. Roses are red, violets are blue... but should vqa expect them to? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2776–2785, 2021. 2

[9] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1

[10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2, 3

[11] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 1, 5

[12] Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1

[13] Xi Zhang, Feifei Zhang, and Changsheng Xu. Vqacl: A novel visual question answering continual learning setting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19102–19112, 2023. 1, 2

Table 5. Results of various CL methods across three distinct task orders with GQA v2 (PS).

Method	Task Order 1			Task Order 2			Task Order 3		
	FAA (↑)	CAA (↑)	FFM (↓)	FAA (↑)	CAA (↑)	FFM (↓)	FAA (↑)	CAA (↑)	FFM (↓)
Joint Learning	44.93	-	-	44.93	-	-	44.93	-	-
SFT [6]	5.48	15.88	45.63	4.20	16.56	48.95	9.64	16.94	43.27
LwF [11]	9.09	14.68	10.76	6.80	12.27	6.46	8.43	12.57	10.02
ZAF [4]	11.29	21.76	34.11	9.70	22.73	38.34	12.24	23.48	37.37
ER [3]	39.45	42.93	13.40	40.59	39.97	9.24	42.98	40.23	5.56
CLS-ER [2]	39.57	42.97	11.87	40.45	40.54	8.25	41.88	39.89	5.54
MaDQ (Ours)	**40.72**	**45.18**	**9.37**	**41.00**	**40.77**	**6.77**	**43.32**	**41.39**	**4.66**

Table 6. Impact of architectures on continual VQA (GQA v2). MaDQ* denotes the variant for frozen-LLM models.

Arch.	**Method**	**Standard Splits (SS)**			**Proposed Splits (PS)**		
		FAA(↑)	*CAA*(↑)	*FFM*(↓)	*FAA*(↑)	*CAA*(↑)	*FFM*(↓)
BLIP (E2E)	SFT	13.05	23.25	64.98	7.41	14.74	37.99
	GaB [7]	26.54	31.69	46.99	13.15	27.94	29.94
	ER [36]	52.48	54.31	18.11	32.24	33.64	9.27
	CLS-ER [3]	52.54	54.53	17.45	34.32	32.58	7.56
	MaDQ	**53.21**	**57.62**	**16.79**	**36.15**	**37.82**	**4.54**
BLIP-2 (Frozen)	SFT	7.07	15.64	34.77	6.47	6.82	13.20
	GaB [7]	17.45	23.38	19.37	10.51	12.90	8.41
	ER [36]	23.46	31.28	14.23	14.27	16.51	4.38
	CLS-ER [3]	24.45	31.71	13.77	14.71	17.20	4.13
	MaDQ*	21.48	30.13	14.80	–	–	–