VQACL: A Novel Visual Question Answering Continual Learning Setting

Xi Zhang[1,2], Feifei Zhang[4], Changsheng Xu[1,2,3]
[1]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences [3]Peng Cheng Laboratory
[4]School of Computer Science and Engineering, Tianjin University of Technology
zhangxi2019@ia.ac.cn, feifeizhang@email.tjut.edu.cn, csxu@nlpr.ia.ac.cn

In this supplementary material, we will further detail the following aspects omitted in the main paper.

- Section A: Construction details about the proposed VQACL setting.
- Section B: More details about the compared continual learning algorithms, including EWC [9], MAS [2]), ER [6], DER [3], and VS [16].
- Section C: More quantitative and qualitative experiment results, which further justify the effectiveness of the proposed method.

A. More details about the VQACL Setting

we establish a novel VQA Continual Learning setting named VQACL, which contains two key components: a dual-level task sequence where visual and linguistic data are nested, and a novel composition testing containing new skill-concept combinations.

Dual-Level Task Definition. In the outer level, according to the question type annotation, we define ten linguistic-driven tasks for the VQA v2 dataset, including *Recognition, Location, Judge, Commonsense, Count, Action, Color, Type, Subcategory*, and *Causal*. For the NExT-QA dataset, eight linguistic-driven tasks are defined in our VQACL setting (i.e., *CausalWhy (CW), TemporalNext (TN), TemporalCurrent (TC), DescriptiveLocation (DL), DescriptiveBinary (DB), DescriptiveCount (DC), DescriptiveOther (DO)*, and *CausalHow (CH)*). Detailed statistics are provided in Table 1 and Table 2. In the inner level, we define five object groups and randomly assign them to different visual-driven tasks. In VQA v2 and NExT-QA, the object groups are obtained by uniformly partitioning the object classes annotated in [10] and [18], respectively. Detailed partition results are provided in Table 3 and Table 4.

Novel Composition Testing. In VQA v2, we find that for some types of questions (e.g., *Judge*, *Commonsense*), their ground-truth answers are mostly 'Yes' or 'No', which are too simple and not suitable for the compositional generalizability evaluation. Therefore, for the novel composition testing in VQA v2, we select six representative linguistic-driven tasks: *Location, Count, Action, Color, Type*, and *Subcategory*. For the novel composition testing in NExT-QA, we consider the *TN, TC, DL, DB, DC, DO*, and *CH* tasks.

Evaluation Metric. As mentioned in Section 3.3 in the main paper, in NExT-QA, we use Wu-Palmer similarity (WUPS) [11] following [18] to compute the model performance a in our VQACL setting. Specifically, the WUPS score is adopted to semantically evaluate the generated answer. Given an image and a question, suppose the predicted answer is $R = \{r_1, r_2, ..., r_i, ...\}$ and the corresponding ground-truth answer is $Y = \{y_1, y_2, ..., y_i, ...\}$, where r_i and y_i respectively denote the i-th words of the predicted and ground-truth answers. Then, the WUPS score computes the similarity between two answers as follows:

$$WUPS(R,Y) = \min\Big\{ \prod_{r\in R} \max_{y\in Y} WUP(r,y), \prod_{y\in Y} \max_{r\in R} WUP(y,r) \Big\} \times 100. \tag{1}$$

In Eq. (1), $WUP(r,y)$ calculates the Wu-Parlmer similarity [8, 17] of two words based on their depth in the taxonomy [4, 12]: WUP(r,y)=2depth(*lcs*) / (depth(r)+depth(y)), where *lcs* is the least common ancestor of the words r and y. If two words are semantically closer, they would be in same or nearer depths in the hierarchy and share more common ancestors, thus getting a higher WUP score.

For VQA v2, following [7], we leverage the percentage of correctly answered questions as the a. Specifically, a question is considered to be answered correctly only if the predicted answer is exactly the same as the ground-truth answer.

B. Continual Learning Methods

In our VQACL setting, we investigate and evaluate five well-established and state-of-the art continual learning methods, including two regularization methods (EWC [9], MAS [2]) and three rehearsal-based approaches (ER [6], DER [3], and VS [16]). For a fair comparison, all the

Table 1. 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 2. 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 baby do after picking up the toy? What did lady do after adjusting 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?

methods are implemented using official codes and added to the same transformer backbone introduced in Section 5.1 in the main paper as our method. Specifically,

EWC [9] is a regularization method and remembers old tasks by selectively slowing down learning on the parameters that are important for these tasks. To achieve it, EWC uses the Fisher Information Matrix [13] to estimate the importance of each parameter, and adds an auxiliary L_2 loss between the important parameters learned from the new task and old tasks.

MAS [2] is also a regularization method and discourages big changes in parameters that are important for previous tasks through an additional L_2 loss. To estimate the importance of a parameter, MAS measures how sensitive the predicted output function is to a change in this parameter.

ER [6] is a rehearsal approach and randomly stores visited examples in a fix-sized memory called the replay memory. At each training step, it randomly samples these stored examples for retraining. Consistent with our method, the memory size of ER is set to 5,000 for VQA v2 and 500 for NExT-QA. Since ER is well-established and simple to implement, we utilize it as the baseline of our proposed approach.

DER [3] belongs to rehearsal methods and adopts reservoir sampling [15] to decide examples to store and replace from the replayed memory. Specifically, the reservoir algorithm ensures each visited example has the same probability to be stored in the memory. Based on the memory, DER designs a dark experience based knowledge distillation strategy to match the network's output logits sampled throughout the training process, which encourages the network to mimic its original responses for past examples. In our experiments, the memory size is set to 5,000 for VQA v2 and 500 for NExT-QA.

VS [16] is a rehearsal method and considers the feature compatibility between the ongoing and previous data. To model the feature consistency and mitigate the forgetting, it designs a neighbor-session model coherence loss and an inter-session data coherence loss. We suggest readers to check Wan et al. [16] for more details about these two losses. As in our method, the memory size of VS is set to 5,000 for VQA v2 and 500 for NExT-QA.

C. Experimental Results

C.1. More Fine-grained Results in the VQACL

Table 5 and Table 6 respectively provide fine-grained model performance on the standard continual learning test of VQA v2 and NExT-QA. Specifically, the results shown in each column denote the model's final performance on the corresponding linguistic-driven task, and the Final Average

Table 3. 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, fir hydrant, chair, mouse, cat, banana, wine glass, backpack, bowl, sports ball, train

Table 4. 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

Performance (AP) across all tasks is provided in the last column. From Table 5 and Table 6, we can observe that our approach achieves the highest performance on most tasks and outperforms other continual learning methods with clear improvements, especially on the *Location* (i.e., 6.66% to 19.95%) and *Type* (i.e., 3.05% to 15.38%) task in VQA v2, and *DB* (i.e., 2.75% to 40.93%) and *DO* (i.e., 2.07% to 23.38%) task in NExT-QA. In conclusion, the detailed results further demonstrate the superiority of our proposed method, which may benefit from the learned discriminative sample-specific feature and generalizable sample-invariant feature.

C.2. Backward Transfer (BWT) Analysis

BWT [3, 5] is the influence of learning a task on the performance of previous tasks, defined by BWT$= a_{t,T} - a_{t,t}$, where $a_{t,T}$ and $a_{t,t}$ respectively denotes the testing performance on the t-th task when the model completed learning the final T-th task and the t-th task, $t = \{1, ..., T-1\}$. We analyze BWT for different rehearsal methods (ER [6], DER [3], VS [16]) on the standard test set of VQA v2 with 5,000 memory size, and the results are illustrated in Fig. 1. From the figure, we can observe that the compared approaches have large negative BWT in the VQACL setting, which means they suffer from severe forgetting problem. In contrast, our model even achieves positive BWT on the 2-th and 4-th tasks, indicating that the learning on new tasks can boost the performance of previous tasks in our method. The results further demonstrate the effectiveness of the proposed representation learning approach.

C.3. Qualitative Results

In the VQACL setting, Fig. 2 presents some qualitative examples in both the standard and novel composition test set of VQA v2, which are predicted by our method and the baseline model that without the sample-specific and sample-variant features. For the standard testing shown in Fig 2(a), we can observe that the baseline tends to predict some words that are unrelated to the question input. For example, it incorrectly generates the word '*happy*' for the first example and the '*hammer time*' for the second example. Actually, we find that these words often appear

Table 5. The VQA performance (%) on the standard test set of VQA v2 with the VQACL setting. The memory size in the rehearsal methods is 5,000. The results of our method are highlighted in bold.

Method	*Recognition*	*Location*	*Judge*	*Commonsense*	*Count*	*Action*	*Color*	*Type*	*Subcategory*	*Causal*	*AP*
Joint	26.70	24.29	64.94	66.30	34.80	57.89	49.32	30.95	46.02	11.95	51.64
Vanilla	7.39	4.94	22.29	32.30	0.71	12.14	12.10	10.69	27.29	15.10	14.49
EWC [9]	6.73	8.43	27.22	47.10	0.14	12.40	1.76	10.98	31.05	11.85	15.77
MAS [2]	30.81	8.07	25.50	4.00	31.90	32.39	26.24	24.75	19.85	2.75	20.56
ER [6]	18.64	21.36	61.27	64.17	30.29	52.84	43.39	23.31	42.75	11.85	36.99
DER [3]	14.55	13.83	62.88	65.16	30.96	51.19	40.51	19.04	42.87	12.55	35.35
VS [16]	15.66	19.21	59.86	66.16	27.28	47.79	32.32	20.44	41.38	10.20	34.03
Ours	**20.47**	**28.02**	**62.55**	**68.61**	**29.35**	**50.66**	**44.45**	**26.36**	**44.65**	**12.60**	**38.77**

Table 6. The VQA performances (%) on the standard test set of NExT-QA with the VQACL setting. The memory size in the rehearsal methods is 500. The results of our method are highlighted in bold.

Method	*CW*	*TN*	*TC*	*DL*	*DB*	*DC*	*DO*	*CH*	*AP*
Joint	10.84	11.62	19.47	34.96	66.73	91.73	40.25	11.76	35.92
Vanilla	7.80	7.78	10.63	10.26	16.55	18.62	11.24	12.91	11.97
EWC [9]	8.71	9.19	11.79	9.54	20.22	17.03	14.05	13.56	13.01
MAS [2]	5.14	1.09	6.45	4.57	14.68	89.86	16.97	5.53	18.04
ER [6]	7.15	7.71	15.11	21.81	52.86	90.39	35.36	13.98	30.55
DER [3]	0.83	8.32	14.41	31.89	31.06	91.17	20.30	11.39	26.17
VS [16]	6.72	7.39	11.47	19.61	49.72	88.02	31.65	10.43	28.13
Ours	**7.48**	**10.32**	**13.39**	**30.52**	**55.61**	**90.72**	**37.43**	**12.68**	**32.27**

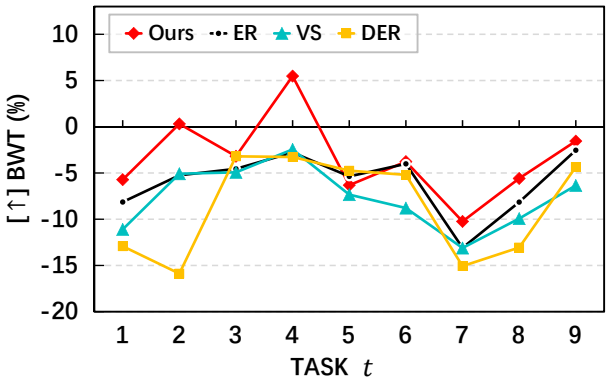


Figure 1. Backward Transfer (BWT) analysis for each linguistic-driven task in the VQACL setting.

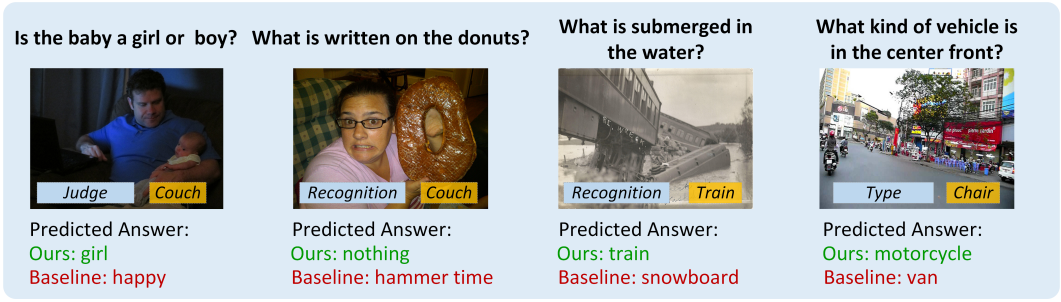
in the last *Causal* task of the continual learning sequence. The results indicate that the baseline is prone to ignore the old experience and focus on the newest learned knowledge in continual VQA, that is, suffering from the catastrophic forgetting problem. In contrast, our method can generate correct answers, even though the examples are from previously learned tasks (e.g., *Recognition* and *Judge*). The superior performance demonstrates the effectiveness of our approach in VQA continual learning. Besides, for the novel composition testing shown in Fig 2(b), the baseline tends to exploit the well-known language bias [1, 14] and choose the incorrect high-frequency answers (e.g., *2*, *red*), which shows that it cannot do well in the compositional generalization. Differently, our model consistently makes correct predictions, which indicates that the proposed model can successfully generalize to novel skill-concept compositions. It may be attributed to our effective sample-specific features and generalizable sample-invariant features.

References

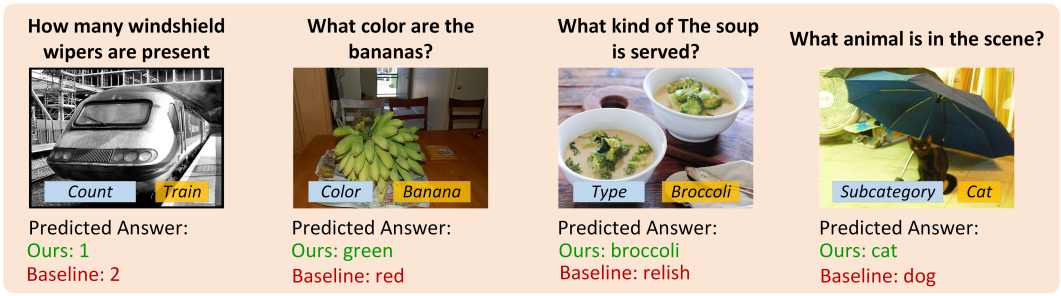
[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*, pages 4971–4980, 2018. 4

[2] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *ECCV*, pages 139–154, 2018. 1, 2, 4

[3] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *NeurIPS*, 33:15920–15930, 2020. 1, 2, 3, 4



(a) Standard Testing



(b) Novel Composition Testing

Figure 2. Qualitative examples on VQA v2 in the VQACL setting. For each sample, we show the prediction results of our proposed method and the baseline model, which does not deploy the sample-specific and sample-invariant features. The correct answer is colored with green, while the wrong one is colored with red.

[4] Carol Chapelle. *The encyclopedia of applied linguistics*. Wiley-Blackwell Malden, MA, 2013. 1

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

[6] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and M Ranzato. Continual learning with tiny episodic memories. *arXiv preprint arXiv:1902.10486*, 2019. 1, 2, 3, 4

[7] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *ICML*, pages 1931–1942, 2021. 1

[8] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*, pages 2712–2719, 2013. 1

[9] 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, 4

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 1

[11] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. *NeurIPS*, 27, 2014. 1

[12] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 1

[13] Razvan Pascanu and Yoshua Bengio. Revisiting natural gradient for deep networks. *arXiv preprint arXiv:1301.3584*, 2013. 2

[14] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. *NeurIPS*, 31, 2018. 4

[15] Jeffrey S Vitter. Random sampling with a reservoir. *TOMS*, 11(1):37–57, 1985. 2

[16] Timmy ST Wan, Jun-Cheng Chen, Tzer-Yi Wu, and Chu-Song Chen. Continual learning for visual search with backward consistent feature embedding. In *CVPR*, pages 16702–16711, 2022. 1, 2, 3, 4

[17] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *ACL*, pages 133–138, 1994. 1

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