Attention-Logit Steering to Compositional Generalization for Continual VQA

Suyoung Yang[1]

Department of Artificial Intelligence, Yonsei University
Seoul, Republic of Korea
suyoung425@yonsei.ac.kr

Abstract. Continual visual question answering (VQA) requires learning new question types while retaining previously acquired visual grounding and reasoning. Existing methods mainly reduce parameter interference, but do not explicitly control how sequential adaptation changes the coupling between a question and its visual evidence. We propose Q-STEER, a task-agnostic framework that separates plasticity from retention through a shared question-conditioned controller. For plasticity, probe-based MoE-LoRA self-expansion activates new expert slots only when the current capacity is insufficient. For retention, a KL-based late-layer attention-drift signal conditions a low-rank correction to attention logits; the drift is diagnostic rather than an attention-matching loss. On continual VQA v2, Q-STEER obtains 53.68 average performance and 4.51 average forgetting, and reaches 51.00/51.48 on Novel/Seen compositional splits. Experiments on a heterogeneous visual-instruction stream, together with seed, ablation, grounding, sensitivity, and overhead analyses, support the effectiveness of drift-aware steering and on-demand expansion.

Keywords: Continual Learning · Visual Question Answering · Multimodal Large Language Models · Compositional Generalization

1 Introduction

Multimodal large language models (MLLMs) provide a strong basis for visual question answering (VQA) by combining visual representations with instruction-following language models [15,18]. In deployment, however, question types, visual domains, and concept combinations arrive sequentially. A continual VQA model must therefore learn new capabilities without forgetting earlier ones, and a task-agnostic model must do so without a task identifier at inference time [10, 28].

Most continual-learning methods explain forgetting as parameter interference. Regularization constrains updates to important parameters [2, 12], replay revisits previous examples [3, 4], and parameter-isolation or expansion methods allocate separate capacity [19, 21, 27]. These approaches are valuable, but VQA introduces an additional failure mode: a model may preserve an answer prior or reasoning routine while losing the association between the question and the relevant visual evidence. We call this failure *question–evidence coupling drift*.

This distinction matters for compositional generalization. As illustrated in Fig. 1, sequential adaptation can redirect the evidence used for answering from the queried object to an irrelevant visual region, causing an incorrect prediction even when the relevant visual concept remains familiar.

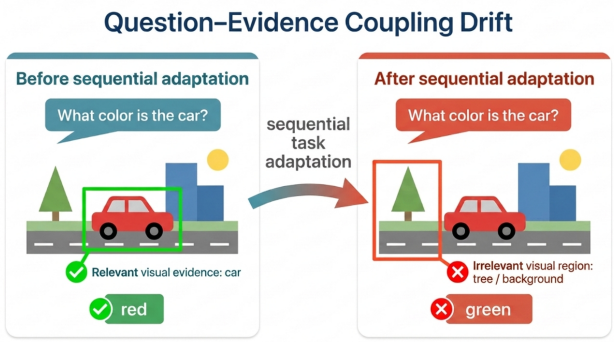

Fig. 1: Conceptual illustration of question–evidence coupling drift. Sequential adaptation redirects answer-relevant evidence from the queried object to an irrelevant region, changing the prediction from "red" to "green." The highlighted regions are schematic rather than measured attention maps.

VQA models can exploit language priors when grounding is weak, and their robustness drops when answer priors change or image evidence becomes necessary [1, 7, 22]. In continual VQA, later tasks can change both answer behavior and cross-modal routing. Consequently, a model may retain individual concepts yet fail on an old or novel skill–concept composition because it no longer binds the linguistic cue to the correct image region. Forgetting should therefore be controlled along two axes: response-behavior adaptation and cross-modal coupling retention.

We propose **Q-STEER**. The overall architecture is shown in Fig. 2. A shared question-conditioned, drift-aware controller coordinates two complementary mechanisms. First, a probe-based MoE-LoRA bank supplies plasticity by activating new expert slots only when the current capacity is insufficient. Second, Input-conditioned Coupling Retention (ICR) supplies retention by adding a low-rank correction to late-layer attention logits. The correction is conditioned on the KL divergence between current and consolidated attention summaries. Importantly, KL drift is not optimized as a loss and does not force shifted tasks to imitate old attention; it only informs routing and steering.

Our contributions are threefold. First, we formulate question–evidence coupling drift as a distinct source of forgetting in task-agnostic continual VQA. Second, we introduce a shared controller that combines factorized old/new MoE-LoRA routing, probe-based self-expansion, and drift-aware attention-logit steering. Third, we provide empirical evidence on continual VQA v2 and a heteroge-

neous visual-instruction stream, including three-seed evaluation, compositional testing, drift and component ablations, grounding proxies, hyperparameter sensitivity, expert utilization, and computational overhead.

2 Related Work

Continual Learning and Continual VQA. Regularization, replay, and parameter isolation are the principal strategies for catastrophic forgetting [2–4, 12, 19, 21, 27]. Self-expansion methods additionally allocate lightweight capacity when existing modules are insufficient [25]. VQACL established a question-type continual VQA protocol [28]; CL-MoE introduced modular expert routing [10]; and QUAD combined questions-only replay with attention distillation [20]. These methods preserve behavior through constraints, replay, or modularity. Our focus is complementary: we use attention drift as an input-dependent signal for preserving question–evidence coupling without an attention-matching objective.

Parameter-Efficient Adaptation, Experts, and Grounding. Adapters and LoRA enable efficient adaptation of frozen backbones [8, 9]; prefix and prompt tuning provide related lightweight alternatives [14, 16]. Sparse mixture-of-experts models scale capacity through conditional computation [5,6,11,13,23]. We use preallocated LoRA experts for continual capacity growth and factorize old/new routing to reduce direct competition. Our attention steering is motivated by VQA work showing that reliance on question priors rather than relevant image evidence harms robustness [1, 7, 22].

3 Method

3.1 Setting and Overview

We consider a sequence of tasks $\{D_t\}_{t=1}^{T}$, where each example is an image–question–answer tuple (I, Q, Y). The model is trained sequentially and receives no task identifier at inference time. The MLLM backbone remains frozen; trainable parameters consist of LoRA experts, a controller H_ϕ, and low-rank ICR bases.

Given $x = (I, Q)$, the controller produces old/new routing logits, a shared interpolation gate, and ICR coefficients:

$$H_\phi(c(x), d(x)) = \big(\{z^l_{\text{old}}, z^l_{\text{new}}\}_{l\in\mathcal{L}}, \lambda, \{\alpha^l, \gamma^l\}_{l\in\mathcal{L}_{\text{late}}}\big). \tag{1}$$

Here $\lambda(x) \in [0, 1]$ controls the mixture between carried-over and newly activated experts and also scales ICR. The same controller is used for every task.

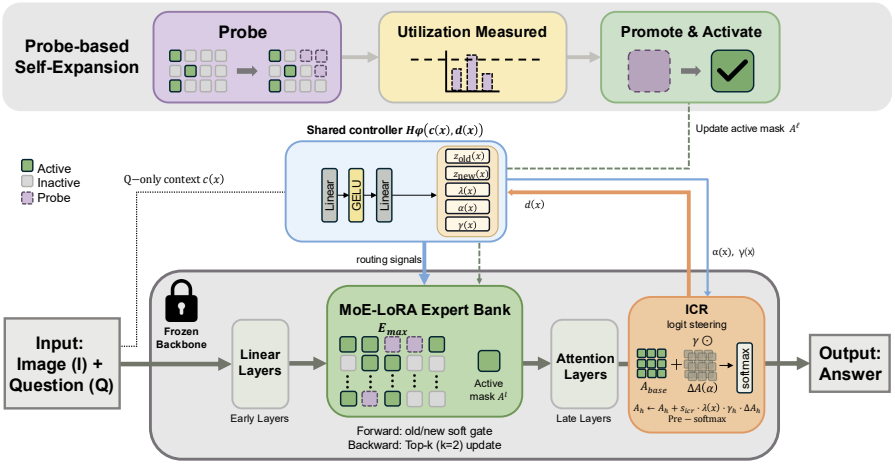

Fig. 2: Overview of Q-STEER. The probe stage measures the utilization of candidate expert slots and promotes selected slots by updating the active mask. During main training and task-agnostic inference, the shared controller uses the question-only context and attention-drift signal to coordinate factorized old/new MoE-LoRA routing and low-rank ICR steering, while the multimodal backbone remains frozen.

3.2 Question-Conditioned Drift Signal

To prevent teacher-forced answers from leaking into the controller, we pool question-token representations only:

$$c(x) = \mathrm{MeanPool}\left(\{h_j : j \in \mathcal{T}_Q\}\right). \tag{2}$$

A prompt-only diagnostic pass then summarizes, at each controlled late layer l, attention from the answer-start query position to visual tokens. Let $p^l_{\mathrm{cur}}(x)$ be the current distribution and p^l_{ref} a dataset-level reference consolidated from completed tasks. We define

$$d^l(x) = \mathrm{KL}\big(p^l_{\mathrm{cur}}(x) \,\|\, p^l_{\mathrm{ref}}\big), \qquad d(x) = \frac{1}{|\mathcal{L}_{\mathrm{late}}|} \sum_{l \in \mathcal{L}_{\mathrm{late}}} d^l(x). \tag{3}$$

For task t, the reference is finalized after task $t-1$, frozen throughout task t, and updated only after task t. For the first task, it is initialized from the frozen backbone. Both distributions are smoothed and renormalized before KL computation.

The drift is *diagnostic*: it is neither a loss nor a distillation target. Training uses standard answer cross-entropy, so a shifted task is free to adopt a different attention pattern. Drift only calibrates the controller's routing and steering decisions.

3.3 Factorized MoE-LoRA and Self-Expansion

For a frozen transform W_0^l and token representation u, an expandable layer outputs

$$y^l = W_0^l u + \sum_{e=1}^{E_{\max}} \tilde{g}_e^l(x) B_e^l A_e^l u, \tag{4}$$

where each (A_e^l, B_e^l) is a rank-r_{lora} LoRA expert. Let $\mathcal{O}^l$ denote carried-over active slots and $\mathcal{N}^l$ newly promoted slots. We normalize the two banks separately:

$$p_{\text{old}}^l = \text{softmax}(\text{mask}(z_{\text{old}}^l, \mathcal{O}^l)), \tag{5}$$

$$p_{\text{new}}^l = \text{softmax}(\text{mask}(z_{\text{new}}^l, \mathcal{N}^l)), \tag{6}$$

and, when $\mathcal{N}^l \neq \emptyset$, use

$$\tilde{g}^l = (1 - \lambda(x)) p_{\text{old}}^l + \lambda(x) p_{\text{new}}^l. \tag{7}$$

If no slot is promoted, the new branch is bypassed. Inactive slots are assigned $-\infty$ before softmax. The forward mixture is soft, while gradients are restricted to the top-K routed experts to reduce expert interference.

Each layer contains $E_{\max}$ preallocated slots, of which E_{init} are initially active. At each post-initial task, a short probe temporarily exposes candidate inactive slots. Their utilization is

$$\text{Act}_p^l = \mathbb{E}_{x \sim D_t}\big[\,\text{mean}_{\text{tokens}}\, \tilde{g}_p^l(x)\big]. \tag{8}$$

A probe slot is promoted when its utilization exceeds

$$\text{Thr}^l = \text{mean}(\text{Act}^l) - \beta_{\text{thr}}\, \text{std}(\text{Act}^l). \tag{9}$$

Thus expansion changes only the active mask, avoiding runtime parameter-shape changes.

3.4 Drift-Aware Attention-Logit Steering

For attention head h at late layer l, ICR constructs the low-rank bias

$$\Delta A_h^l(x) = \big(Q_h^l U_{q,h}^l\big)\, \text{diag}(\alpha^l(x)) \big(K_h^l U_{k,h}^l\big)^\top, \tag{10}$$

and injects it before softmax:

$$A_h^l \leftarrow A_h^l + s_{\text{icr}} \lambda(x) \gamma_h^l(x) \Delta A_h^l(x). \tag{11}$$

The correction is applied only to answer-query positions. This provides input-dependent control over late cross-modal routing while leaving the backbone weights unchanged.

3.5 Training, Inference, and Retention View

The objective is standard autoregressive cross-entropy over answer tokens. On the first task, the initial experts, controller, and ICR bases are trained. For later tasks, the probe stage promotes candidate slots; previously consolidated experts are then frozen while newly promoted experts, the controller, and ICR bases are updated. After each task, new experts and the attention reference are consolidated. Inference computes $c(x)$ and $d(x)$ from the prompt, then routes and steers without a task label.

The shared gate gives a concise sufficient condition for retaining an old prediction. Let $z_{\text{old}}(x)$ be logits from the old-only path, $z(x)$ the final logits, and $m_{\text{old}}(x)$ the old-path answer margin. Let $\Psi(x)$ collect the layer-wise MoE and ICR perturbation magnitudes weighted by the corresponding local Lipschitz constants. Under a local Lipschitz assumption for the late-layer MoE and attention perturbations, the detailed derivation in the supplement yields

$$\|z(x) - z_{\text{old}}(x)\|_\infty \le \lambda(x)\Psi(x), \qquad m(x) \ge m_{\text{old}}(x) - 2\lambda(x)\Psi(x). \tag{12}$$

Hence $\lambda(x)\Psi(x) < m_{\text{old}}(x)/2$ preserves the old prediction. For a fixed interpolation gate $\lambda(x)$, factorized normalization also removes the direct competition path from new-bank logits to old-bank routing weights. For old expert i and new expert j,

$$\left.\frac{\partial \tilde{g}^l_{\text{old},i}}{\partial z^l_{\text{new},j}}\right|_{\lambda(x)} = 0, \tag{13}$$

whereas a single concat-softmax gives a nonzero negative cross-derivative. This decoupling is local to the gate composition; the shared controller itself remains trainable across tasks. These are sufficient design properties, not universal guarantees.

4 Experiments

4.1 Setup

Benchmarks and Metrics. The main benchmark follows the VQACL continual VQA-v2 protocol [7,28], whose ten tasks are Recognition, Localization, Judgment, Commonsense, Counting, Action, Color, Type, Subcategory, and Causal. Images primarily originate from COCO [17]. We also evaluate a heterogeneous Continual Visual Instruction Tuning (CVIT) stream following SMoLoRA [26], covering ScienceQA, TextVQA, Flickr30k, ImageNet, GQA, and VQAv2. CVIT aggregates native task metrics, so its absolute scores are meaningful only within that protocol.

Let $a_{i,t}$ denote performance on task t after training through task i. We report

$$AP = \frac{1}{T}\sum_{t=1}^{T} a_{T,t}, \qquad AF = \frac{1}{T-1}\sum_{t=1}^{T-1}\left(\max_{t\le i\le T} a_{i,t} - a_{T,t}\right). \tag{14}$$

All AF values in the paper and supplement were recomputed from stored task-wise performance matrices with an independently verified implementation of Eq. (14). This max-over-history convention measures forgetting from each task's best historical score, so all methods are compared under the same AF definition. Novel and Seen scores evaluate held-out and observed skill–concept compositions, respectively.

Baselines and Fairness. We compare Vanilla, EWC [12], MAS [2], ER [4], DER [3], VS [24], VQACL [28], CL-MoE [10], and QUAD [20]. VS denotes the baseline labeled VS in prior VQACL-style evaluations, corresponding to Wan et al.'s backward-consistent feature-embedding continual-learning method [24]; we keep the VS label for consistency with those evaluations. For VQA-v2, all methods use the same LLaVA-style initialization, task order, answer set, evaluator, and task-agnostic inference protocol; only the adaptation mechanism differs. The CVIT comparison follows the shared protocol and baselines of SMoLoRA.

Implementation. The backbone is frozen. Unless noted otherwise, we control the last six transformer layers, set $r_{\text{lora}} = r = 8$, $s_{\text{icr}} = 0.3$, $E_{\text{init}} = 8$, $E_{\max} = 32$, $\beta_{\text{thr}} = 0.5$, use a probe budget of 128 samples per task, and apply top-K backward updates with $K = 2$. Attention softmax is computed in fp32. Training and inference overhead include the prompt-only drift pass; training additionally includes the short probe stage.

4.2 Continual VQA-v2 Results

Table 1: Performance (%) on continual VQA v2. Cou. denotes counting.

Method	Question type										$AP \uparrow$	$AF \downarrow$
	Rec.	Loc.	Jud.	Com.	Cou.	Act.	Col.	Typ.	Sub.	Cau.		
Vanilla	19.25	14.81	54.59	56.97	24.23	46.20	27.58	26.09	36.47	18.89	32.51	20.69
EWC	28.12	23.02	61.50	61.08	26.13	54.29	23.65	32.25	44.97	17.83	37.28	15.27
MAS	31.54	22.09	60.85	46.32	32.48	56.47	30.05	35.69	42.73	18.83	37.71	14.91
ER	29.31	25.74	63.46	65.78	31.92	58.39	45.17	34.55	46.24	18.96	41.95	10.20
DER	26.95	21.43	64.88	66.17	31.01	55.92	44.60	32.85	47.09	20.74	41.16	11.28
VS	28.48	24.09	61.37	67.20	29.56	54.64	33.98	32.91	45.82	19.89	39.79	12.70
VQACL	34.14	32.19	66.15	63.00	33.01	60.91	34.64	38.48	47.94	24.42	43.49	9.10
CL-MoE	33.39	28.63	65.29	63.33	33.05	60.84	33.25	38.98	46.95	22.58	42.63	10.56
QUAD	35.87	33.17	66.93	62.15	34.09	61.28	35.03	38.87	48.66	25.53	44.16	8.97
Q-STEER	**48.53**	**39.82**	**77.36**	**73.54**	**44.06**	**71.26**	**46.05**	**55.00**	**57.57**	23.61	**53.68**	**4.51**

Table 1 shows that Q-STEER improves AP by 9.52 points over QUAD and reduces AF by 4.46 points. It also improves AP/AF over VQACL by 10.19/4.59 points. The largest gains occur on several grounding-intensive and compositional question types, while causal questions remain challenging. Across three seeds, Q-STEER obtains 53.62±0.28 AP, 4.55±0.15 AF, 50.96±0.22 Novel, and

51.43±0.24 Seen, showing that the AF improvement remains well outside seed variation.

4.3 Compositional Generalization

Table 2: Fine-grained VQA performance (%) on Novel and Seen skill–concept compositions.

Group	Split	DER	VS	ER	VQACL	CL-MoE	QUAD	Q-STEER
G1	Novel	39.71	38.44	40.58	42.37	41.83	43.18	**51.72**
	Seen	40.63	39.21	41.67	43.28	42.57	43.85	**52.20**
G2	Novel	38.86	37.63	39.74	41.69	40.66	41.73	**50.86**
	Seen	39.94	38.52	40.86	42.74	41.92	42.96	**51.37**
G3	Novel	39.32	38.19	40.11	41.33	41.21	42.84	**50.52**
	Seen	39.77	39.03	41.28	42.16	42.18	43.51	**51.03**
G4	Novel	38.41	37.88	39.22	40.97	39.94	41.52	**49.72**
	Seen	39.18	38.47	40.17	41.88	41.36	42.37	**50.57**
G5	Novel	39.39	38.76	40.49	42.56	41.48	43.07	**52.19**
	Seen	39.78	39.39	40.62	42.99	42.44	43.52	**52.23**
Avg.	Novel	39.14	38.18	40.03	41.78	41.02	42.47	**51.00**
	Seen	39.86	38.92	40.92	42.61	42.09	43.24	**51.48**

Table 2 shows that Q-STEER consistently outperforms all baselines across all five evaluation groups, rather than obtaining its gain from a single favorable composition split. On average, Q-STEER improves Novel and Seen performance over QUAD by 8.53 and 8.24 points, respectively. The particularly large improvement on Novel compositions supports the central hypothesis that retaining question–evidence coupling helps the model recombine previously learned skills and visual concepts, rather than merely preserving task-specific answer statistics.

4.4 Heterogeneous Stream and Efficiency

Table 3: Robustness on the heterogeneous CVIT stream. Dataset columns use native metrics.

Method	SQA	TVQA	F30k	INet	GQA	VQAv2	$AP \uparrow$	$AF \downarrow$
SeqLoRA	59.21	50.80	20.99	20.30	49.98	64.41	44.28	48.73
MoE-LoRA	58.09	53.30	22.82	22.61	51.80	65.15	45.63	49.94
DoRA	52.03	47.37	27.97	26.18	46.05	57.33	42.82	34.11
C-LoRA	55.58	38.64	59.05	22.81	40.93	51.65	44.78	18.83
Replay	66.06	47.78	24.21	25.66	46.53	58.59	44.81	26.88
EWC	53.60	49.07	20.38	20.48	50.11	64.63	43.05	52.47
EWC+TIR	66.94	45.76	29.49	21.68	46.90	58.80	44.93	26.38
Eproj	63.45	53.18	**151.41**	20.63	45.30	57.32	65.22	14.43
SMoLoRA	80.50	58.30	146.63	94.28	52.42	65.96	83.02	6.50
Q-STEER	**82.10**	**59.40**	148.30	**96.10**	**54.60**	**70.40**	**85.15**	**4.33**

On the heterogeneous CVIT stream, Q-STEER achieves the best aggregate performance and retention, improving AP from 83.02 to 85.15 and reducing AF from 6.50 to 4.33 relative to SMoLoRA (Tab. 3). It also obtains the best native score on five of the six constituent datasets. Flickr30k uses a retrieval-style native score that can exceed 100, which explains the scale difference from the other columns. These results provide a robustness check across science reasoning, OCR VQA, image–text understanding, classification, compositional grounding, and standard VQA, rather than a claim of transfer across different backbone families.

The added cost comes from the controller, ICR bases, prompt-only diagnostic pass, and one-time task probe. Inactive slots are masked from decoding FLOPs, and Q-STEER stores no replay samples. The resulting overhead is moderate relative to CL-MoE (Tab. 4).

Table 4: Computational overhead.

Method	Add-on Params	FLOPs	Mem.	Train	Infer.
VQACL	0.0M	+0.0%	+5.8%	+9.7%	+0.0%
CL-MoE	40.0M	+5.4%	+7.6%	+13.2%	+5.0%
Q-STEER	45.6M	+6.5%	+8.2%	+15.6%	+6.1%

4.5 Ablations and Grounding Analysis

Table 5: Ablation study on continual VQA v2. Factorized control removes both the old/new expert split and the shared interpolation gate.

Variant	$AP \uparrow$	$AF \downarrow$	Novel ↑	Seen ↑
Full Q-STEER	**53.68**	**4.51**	**51.00**	**51.48**
constant-$d(x)$	52.87	5.15	49.65	50.38
cosine drift	53.14	4.91	50.12	50.82
w/o factorized control	48.73	7.62	44.91	45.62
w/o ICR	51.94	5.90	47.88	49.26
w/o expansion	50.96	6.47	47.11	47.85
w/o active mask	51.28	6.28	47.54	48.12
w/o top-K	50.41	6.89	46.35	46.98

Replacing KL drift with a constant or cosine signal weakens both retention and compositional performance, showing that input-dependent KL drift provides more useful control information than a fixed signal or an alternative similarity measure. Removing ICR increases AF from 4.51 to 5.90, whereas removing expansion lowers AP from 53.68 to 50.96, supporting their respective retention and plasticity roles. Removing factorized control causes the largest degradation,

indicating that separate normalization of old and new expert banks is important for limiting routing interference. Active pre-softmax masking and sparse top-K backward updates also contribute to stable continual adaptation (Tab. 5).

For a validation-level grounding proxy, Evidence-Mass measures generated-answer attention assigned to the top 10% of visual patches most similar to question object or attribute nouns. Attention entropy measures how diffusely the answer-position attention is distributed over visual tokens.

Table 6: Grounding proxy analysis and sensitivity to the ICR steering scale. EM denotes average Evidence-Mass.

Setting	s_{icr}	EM ↑	Entropy ↓	AP ↑	AF ↓
w/o ICR	–	0.421	2.31	51.94	5.90
ICR	0.1	0.456	2.20	52.76	5.22
ICR	0.3	**0.487**	**2.07**	**53.68**	**4.51**
ICR	0.5	0.473	2.14	52.98	4.96

As shown in Tab. 6, the default ICR scale $s_{\text{icr}} = 0.3$ increases Evidence-Mass from 0.421 to 0.487 and reduces attention entropy from 2.31 to 2.07 relative to the model without ICR. Both weaker steering at $s_{\text{icr}} = 0.1$ and stronger steering at $s_{\text{icr}} = 0.5$ produce lower AP and higher AF, indicating that the correction strength involves a selectivity–stability trade-off. The question-type results in Fig. 3 show that the grounding improvement is consistent across recognition, localization, counting, color, action, and causal questions. The exact grounding proxy definition and evaluation details are provided in the supplement.

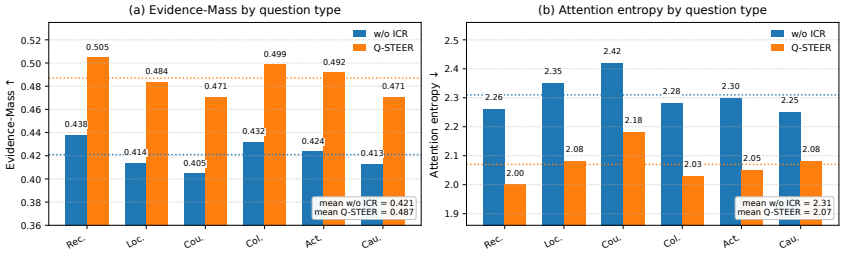

Fig. 3: Grounding proxy analysis by question type. Q-STEER increases Evidence-Mass and reduces attention entropy across recognition, localization, counting, color, action, and causal questions.

This proxy does not establish causal explanation, but it supports the interpretation that ICR concentrates answer-position attention on question-relevant visual evidence.

Sensitivity and expansion dynamics are summarized in Tab. 7 and Fig. 4.

Table 7: Sensitivity to the late-layer intervention window and expansion hyperparameters. The default configuration is highlighted in bold.

Category	Setting	$AP \uparrow$	$AF \downarrow$	Novel ↑	Seen ↑
Late layers	1	50.72	6.52	47.11	47.63
	2	51.86	5.79	48.54	49.02
	4	53.21	4.83	50.37	50.89
	6	**53.68**	**4.51**	**51.00**	**51.48**
	8	52.94	4.96	50.11	50.63
β_{thr}	0.25	52.47	5.34	49.36	49.92
	0.50	**53.68**	**4.51**	**51.00**	**51.48**
	0.75	53.12	4.86	50.24	50.91
E_{init}	4	52.55	5.26	49.51	49.88
	8	**53.68**	**4.51**	**51.00**	**51.48**
	12	53.18	4.74	50.67	50.84
E_{max}	16	52.31	5.49	49.08	49.63
	32	**53.68**	**4.51**	**51.00**	**51.48**
Probe budget	64	52.88	4.97	49.91	50.58
	128	**53.68**	**4.51**	**51.00**	**51.48**
	256	53.56	4.57	50.93	51.31

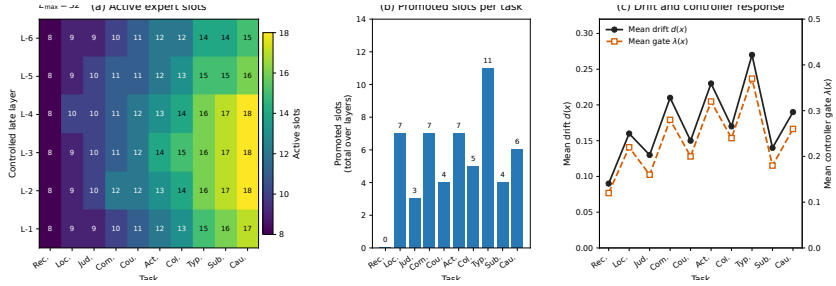

Fig. 4: Expansion dynamics and controller response. The bank expands sparsely and never exceeds 18 of 32 active slots per layer; drift–gate trends are descriptive rather than causal.

Table 7 shows that performance improves as the controlled late-layer window increases from one to six layers, while extending the intervention to eight layers slightly reduces selectivity. The expansion parameters exhibit a capacity–interference trade-off: insufficient initial or maximum capacity degrades performance, while the default threshold and probe budget provide the strongest AP, AF, and compositional results. Figure 4 shows that active expert slots grow gradually but remain well below the preallocated maximum of 32 slots. At most 18 slots are active in any controlled layer, leaving at least 14 unused slots. Expansion is task-dependent rather than uniform, and the task-level controller gate exhibits a trend similar to the measured attention drift.

5 Limitations

Q-STEER preallocates expert slots, which simplifies mask-based expansion but incurs parameter storage even for inactive experts. The prompt-only drift pass

and per-task probe add compute. Attention drift and Evidence-Mass are proxies for coupling quality rather than causal explanations, and the sufficient margin condition in Eq. (12) is not a universal retention guarantee. Finally, the shared gate may be suboptimal when expert usage and attention correction should vary independently. Future work should consider compressible expert banks, richer drift estimators, causal grounding tests, and separate but coordinated routing and steering controls.

6 Conclusion

We introduced Q-STEER, a task-agnostic continual VQA framework that assigns new capacity to new response behavior while steering late-layer attention to preserve question–evidence coupling. Probe-based MoE-LoRA expansion, factorized routing, and KL-conditioned ICR jointly improve average performance, forgetting, and compositional generalization. The results, seed study, heterogeneous-stream evaluation, ablations, grounding proxies, sensitivity analysis, and overhead measurements consistently support the proposed separation of plasticity and retention.

References

1. Agrawal, A., Batra, D., Parikh, D., Kembhavi, A.: Don't just assume; look and answer: Overcoming priors for visual question answering. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4971–4980 (2018)
2. Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., Tuytelaars, T.: Memory aware synapses: Learning what (not) to forget. In: Eur. Conf. Comput. Vis. pp. 139–154 (2018)
3. Buzzega, P., Boschini, M., Porrello, A., Abati, D., Calderara, S.: Dark experience for general continual learning: A strong, simple baseline. In: Adv. Neural Inform. Process. Syst. vol. 33, pp. 15920–15930 (2020)
4. Chaudhry, A., Rohrbach, M., Elhoseiny, M., Ajanthan, T., Dokania, P.K., Torr, P.H.S., Ranzato, M.: On tiny episodic memories in continual learning. arXiv preprint arXiv:1902.10486 (2019)
5. Du, N., Huang, Y., Dai, A.M., Tong, S., Lepikhin, D., Xu, Y., Krikun, M., Zhou, Y., Yu, A.W., Firat, O., Zoph, B., Fedus, L., Bosma, M., Zhou, Z., Wang, T., Wang, E., Webster, K., Pellat, M., Robinson, K., Meier-Hellstern, K., Duke, T., Dixon, L., Zhang, K., Le, Q.V., Wu, Y., Chen, Z., Cui, C.: GLaM: Efficient scaling of language models with mixture-of-experts. In: Int. Conf. Mach. Learn. Proceedings of Machine Learning Research, vol. 162, pp. 5547–5569 (2022)
6. Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. J. Mach. Learn. Res. **23**(120), 1–39 (2022)
7. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6904–6913 (2017)
8. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for NLP. In: Int. Conf. Mach. Learn. Proceedings of Machine Learning Research, vol. 97, pp. 2790–2799 (2019)

9. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: Int. Conf. Learn. Represent. (2022)
10. Huai, T., Zhou, J., Wu, X., Chen, Q., Bai, Q., Zhou, Z., He, L.: CL-MoE: Enhancing multimodal large language model with dual momentum mixture-of-experts for continual visual question answering. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 19608–19617 (2025)
11. Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., de las Casas, D., Hanna, E.B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L.R., Saulnier, L., Lachaux, M.A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T.L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., El Sayed, W.: Mixtral of experts. arXiv preprint arXiv:2401.04088 (2024)
12. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., Hadsell, R.: Overcoming catastrophic forgetting in neural networks. Proc. Natl. Acad. Sci. USA **114**(13), 3521–3526 (2017)
13. Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., Chen, Z.: GShard: Scaling giant models with conditional computation and automatic sharding. In: Int. Conf. Learn. Represent. (2021)
14. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. In: Conference on Empirical Methods in Natural Language Processing. pp. 3045–3059 (2021)
15. Li, J., Li, D., Savarese, S., Hoi, S.C.H.: BLIP-2: Bootstrapping language–image pre-training with frozen image encoders and large language models. In: Int. Conf. Mach. Learn. Proceedings of Machine Learning Research, vol. 202, pp. 19730–19742 (2023)
16. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. In: Annual Meeting of the Association for Computational Linguistics. pp. 4582–4597 (2021)
17. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: Eur. Conf. Comput. Vis. pp. 740–755 (2014)
18. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Adv. Neural Inform. Process. Syst. vol. 36 (2023)
19. Mallya, A., Lazebnik, S.: PackNet: Adding multiple tasks to a single network by iterative pruning. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7765–7773 (2018)
20. Marouf, I.E., Tartaglione, E., Lathuilière, S., van de Weijer, J.: Ask and remember: A questions-only replay strategy for continual visual question answering. In: Int. Conf. Comput. Vis. pp. 18078–18089 (2025)
21. Rusu, A.A., Rabinowitz, N.C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., Hadsell, R.: Progressive neural networks. arXiv preprint arXiv:1606.04671 (2016)
22. Selvaraju, R.R., Lee, S., Shen, Y., Jin, H., Ghosh, S., Heck, L., Batra, D., Parikh, D.: Taking a HINT: Leveraging explanations to make vision and language models more grounded. In: Int. Conf. Comput. Vis. pp. 2591–2600 (2019)
23. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q.V., Hinton, G.E., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In: Int. Conf. Learn. Represent. (2017)

24. Wan, T.S.T., Chen, J.C., Wu, T.Y., Chen, C.S.: Continual learning for visual search with backward consistent feature embedding. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 16702–16711 (2022)
25. Wang, H., Lu, H., Yao, L., Gong, D.: Self-expansion of pre-trained models with mixture of adapters for continual learning. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10087–10098 (2025)
26. Wang, Z., Che, C., Wang, Q., Li, Y., Shi, Z., Wang, M.: SMoLoRA: Exploring and defying dual catastrophic forgetting in continual visual instruction tuning. In: Int. Conf. Comput. Vis. pp. 177–186 (2025)
27. Yoon, J., Yang, E., Lee, J., Hwang, S.J.: Lifelong learning with dynamically expandable networks. In: Int. Conf. Learn. Represent. (2018)
28. Zhang, X., Zhang, F., Xu, C.: VQACL: A novel visual question answering continual learning setting. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 19102–19112 (2023)