HyLoVQA: Dynamic Hypernetwork-Generated Low-Rank Adaptation for Continual Visual Question Answering

Yiran Wang[1] , **Chenyi Xiong**[1] , **Ziyue Qin**[1] , **Miao Zhang**[1] , **Kui Xiao**[1] , **Zhifei Li**[1,2,3*]
[1]School of Computer Science, Hubei University, Wuhan 430062, China
[2]Hubei Key Laboratory of Big Data Intelligent Analysis and Application (Hubei University), Wuhan 430062, China
[3]Key Laboratory of Intelligent Sensing System and Security (Hubei University), Ministry of Education, Wuhan 430062, China
{yiranwang, xiongchenyi, qinziyue}@stu.hubu.edu.cn, {zhangmiao, xiaokui, zhifei1993}@hubu.edu.cn

Abstract

Continual Visual Question Answering (VQA) requires learning from non-stationary streams of visual inputs and questions while preserving past knowledge. Most prior methods adapt by updating a largely shared parameter set. This often leads to cross-level task interference, hindering accurate adaptation to the current task and object. To address this limitation, we propose HyLoVQA. It maintains a drift-resilient memory bank of anchors. The bank stores the content of visual objects and textual tasks, and they are updated using current input features. Conditioned on retrieved anchors, a hypernetwork generates lightweight Low-Rank Adaptation (LoRA) adapters. This ensures parameter efficiency, allowing the model to adapt to each task and object dynamically. Additionally, we formulate an alignment loss that aligns semantic discrepancies in the feature space with functional changes in the parameter space, thereby constraining LoRA adapters to remain focused on the current task and object. Extensive experiments on VQA v2 and NExT-QA under both standard and compositional settings demonstrate the superiority of HyLoVQA over prior state-of-the-art methods. The code is available at https://github.com/HubuKG/HyLoVQA.

1 Introduction

Visual Question Answering (VQA) is a central task in multimodal intelligence [Antol *et al.*, 2015; Johnson *et al.*, 2017]. Given an image (or video) and a natural-language question, a VQA model predicts the answer. Solving VQA requires grounding language in visual evidence and reasoning over the scene in a question-dependent way. This often calls for multi-step, compositional inference rather than direct recognition alone. Over time, the data distribution and question patterns may change, motivating robustness under shift and continual adaptation [Gama *et al.*, 2014; Parisi *et al.*, 2019; Hayes *et al.*, 2020].

*Corresponding author.

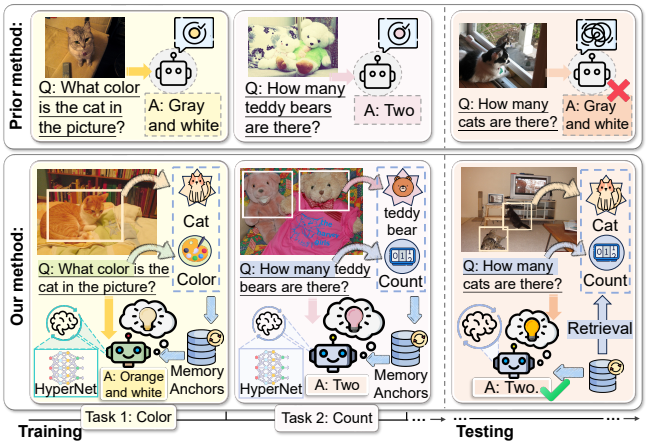

Figure 1: Motivation and overview. Top: Prior methods use shared-parameter continual updates, which induce task/object ambiguity (e.g., *count* → *color*). Bottom: HyLoVQA retrieves Memory Anchors and uses a HyperNet to generate LoRA adapters conditioned on the current task and object, reducing drift-induced interference.

Continual learning (CL) offers a natural framework for continual VQA by learning from non-stationary data streams. It aims to acquire new knowledge while retaining prior capabilities. A central challenge in CL is the stability–plasticity trade-off: models must remain stable enough to avoid forgetting while staying plastic enough to rapidly integrate novel information. Classic CL methods include regularization, rehearsal, and architectural methods for unimodal learning. When applied to VQA, these ideas face added pressure from multimodal grounding and compositional reasoning. Shifts can arise not only in vision or language, but also in their cross-modal co-occurrences and required reasoning, which makes retention and updating more difficult than in unimodal classification. Recent continual VQA work adapts CL principles to multimodal grounding and compositional shifts; VQACL, for example, evaluates unseen skill–concept compositions [Zhang *et al.*, 2023].

Despite recent progress, most continual VQA methods still follow classic CL recipes. They adapt by updating the main model and mitigate forgetting with regularization

or rehearsal [Sun *et al.*, 2023; Greco *et al.*, 2019]. Regularization improves stability but can reduce plasticity under large shifts. Rehearsal depends on the memory budget and how well stored samples remain representative as the stream evolves [Kirkpatrick *et al.*, 2017a; Li *et al.*, 2019]. These strategies can retain performance on past data. However, most existing methods adapt by updating a largely shared parameter set across the stream. This often induces cross-level task interference, where previously learned tasks disturb the current one. They also struggle to dynamically adapt to the current task and object. Here, the task is the capability required by the question (e.g., *count*), and the object is the grounded entity in the input (e.g., *a cat*). In this paper, a sample refers to a VQA instance, which can be viewed as a combination of a textual task and a visual object. As illustrated in the top of Figure 1, these methods can introduce task and object ambiguity, leading to incorrect answers.

To address this limitation, we propose HyLoVQA for continual VQA. It ensures parameter efficiency, enabling the model to adapt to each task and object dynamically while reducing drift-induced interference, as shown in Figure 1. **First**, we maintain a Drift-Resilient Memory Anchor Bank. It stores compact anchors that capture the content of visual objects and textual tasks, and the anchors are updated using current input features to remain stable under non-stationary streams. **Second**, we introduce a Hypernetwork-Generated LoRA Module, where a hypernetwork generates lightweight LoRA adapters from the retrieved anchors. This ensures parameter efficiency and allows the model to dynamically adapt to each task and object, while mitigating interference from shared backbone updates. **Third**, we propose a Semantic–Functional Alignment loss that aligns semantic discrepancy (feature space) to functional change (parameter space). It avoids updates that deviate from the current task and object.

In this paper, we make the following contributions:

- We develop a drift-resilient memory bank that stores anchor representations of visual objects and textual tasks, and updates them online using current input features.
- We generate lightweight LoRA adapters using a hypernetwork conditioned on retrieved anchors, enabling parameter-efficient and dynamic adaptation to each task and object.
- We formulate an alignment objective that couples feature-space semantic discrepancies with parameter-space functional changes, thereby constraining LoRA adapters to stay focused on the current task and object.
- We evaluate HyLoVQA on VQA v2 and NExT-QA under both standard and novel compositional test settings. On standard tests, HyLoVQA achieves 45.41% AP / 2.82% AF on VQA v2 and 42.43% AP / 2.07% AF on NExT-QA.

2 Related Work

2.1 Visual Question Answering

Visual Question Answering (VQA) asks a model to answer a natural-language question about an image[Goyal *et al.*, 2017]. Early pipelines paired CNN image encoders with RNN question encoders and used attention for visual grounding[Ren *et al.*, 2015; Devlin *et al.*, 2019]. Fusion evolved from concatenation to bilinear pooling, with Multimodal Compact Bilinear pooling as an efficient high-order design[Kim *et al.*, 2018; Theeuwes, 2010]. Object-centric representations further advanced VQA via region-level attention, where Bottom-Up and Top-Down attention over detected regions became a widely used baseline[Chen *et al.*, 2020; Gómez-Pérez and Ortega, 2020]. Recently, Transformer-based architectures and large-scale vision-language pretraining have dominated VQA, improving multimodal representation learning and reasoning in the pretrain-then-finetune paradigm[Liang *et al.*, 2020].

However, as VQA applications expand to more complex and dynamic scenarios, the assumptions of a fixed task and a stationary data distribution become fragile. For traditional VQA models, high scores on standard benchmarks may not transfer to VQA settings with distribution shifts [Whitehead *et al.*, 2021; Li *et al.*, 2019]. Meanwhile, incomplete or noisy images and open-ended question formulations further amplify the limitations of fixed training setups. Therefore, VQA requires models that can adapt as tasks and data distributions keep changing, while avoiding forgetting previously learned knowledge. This has motivated continual learning for VQA to mitigate catastrophic forgetting and support long-term adaptation [Zhang *et al.*, 2021].

2.2 Continual Learning in VQA

Continual Learning (CL) studies learning from sequential data streams whose distribution shifts over time [Wan *et al.*, 2022a; Lopez-Paz and Ranzato, 2017]. A central challenge is the stability–plasticity trade-off [Kim *et al.*, 2024]. Models must acquire new knowledge while retaining old knowledge. CL has been widely explored in vision and language, and recent efforts extend it to multimodal settings. To mitigate catastrophic forgetting, most methods fall into three categories: regularization, rehearsal, and architectural approaches. Regularization methods constrain updates to parameters that are important for past tasks, such as MAS and NPC [Aljundi *et al.*, 2018; Paik *et al.*, 2020]. Rehearsal methods replay stored samples to stabilize training, such as ER and DER [Chaudhry *et al.*, 2019; Chaudhry *et al.*, 2018; Wan *et al.*, 2022b]. Architectural methods are commonly studied in unimodal continual learning[Rusu *et al.*, 2016].

Recent continual VQA methods tackle evolving visual concepts and language. VQACL proposes a dual-level task stream and compositional testing. QUAD and ProtoGroup reduce memory via attention consistency and multi-prototype grouping. PROOF mitigates interference with expandable projections, while CLT-VQA targets long-tailed continual VQA via prototype balancing and feature alignment[Marouf *et al.*, 2025; Zhang *et al.*, 2025b; Zhou *et al.*, 2025; Zhang *et al.*, 2025a]. Yet these approaches typically keep a fixed backbone, so interference from shared parameters can still accumulate over time. In contrast, our work reduces cross-level task interference and improves accurate adaptation to the current task and object. By preserving past knowledge while maintaining parameter efficiency, it provides a more reliable and adaptive solution for long-term continual VQA.

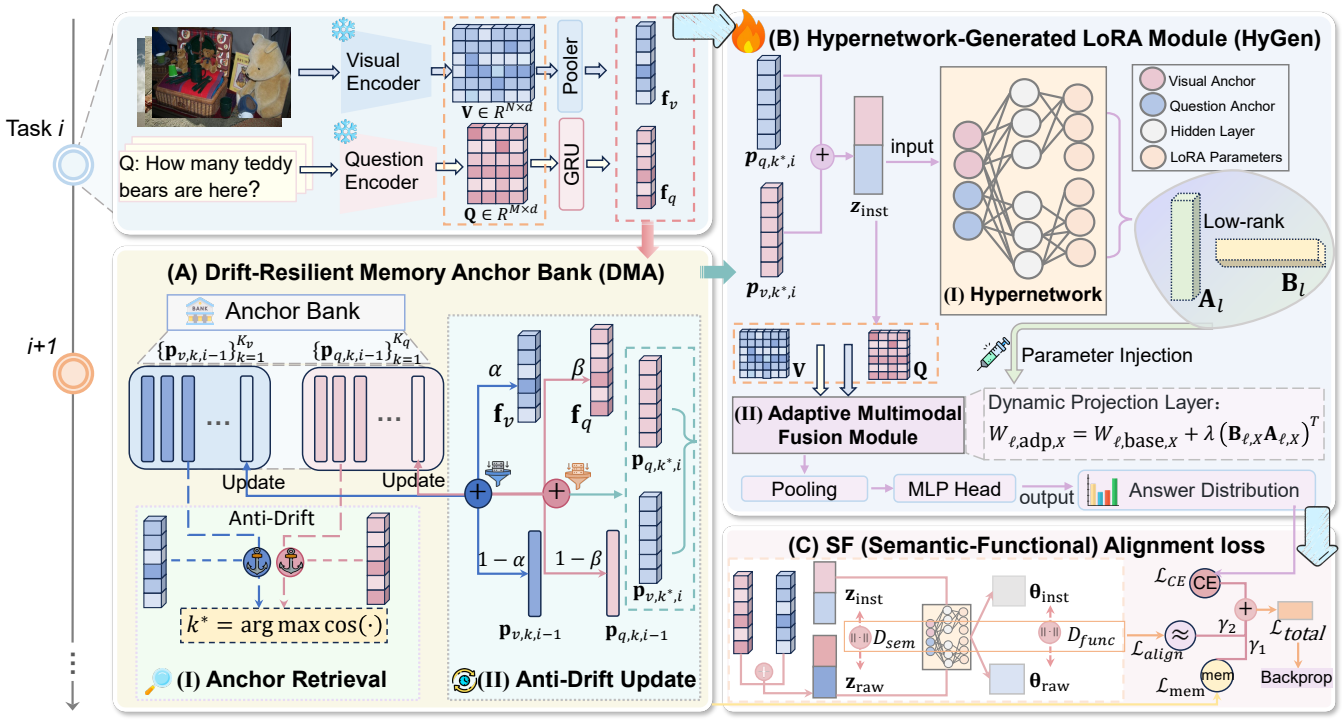

Figure 2: Overview of HyLoVQA. (A) Drift-Resilient Memory Anchor Bank stores compact anchors for visual objects and textual tasks and updates them with current input features for stability. (B) Hypernetwork-Generated LoRA Module generates anchor-conditioned lightweight LoRA adapters for parameter-efficient, task/object-specific adaptation with reduced interference on the shared backbone. (C) Semantic–Functional alignment loss constrains training by aligning semantic discrepancy (feature space) with functional change (parameter space), avoiding LoRA adapters that deviate from the current sample.

3 Methodology

We propose HyLoVQA, a lightweight continual learning framework (Figure 2) for visual question answering, addressing dynamic adaptation, memory robustness under drift, and stable knowledge updates. By integrating a Drift-Resilient Memory Anchor Bank, a Hypernetwork-Generated LoRA Module, and a Semantic–Functional alignment loss, HyLoVQA enables parameter-efficient, task/object-specific adaptation while reducing interference on the shared backbone. Module details are described below.

3.1 Drift-Resilient Memory Anchor Bank

The Drift-Resilient Memory Anchor Bank (DMA) stores compact, modality-specific anchors that summarize visual objects and textual tasks. For each incoming sample, DMA retrieves the most relevant visual and question anchors to form an anchor-conditioned context, which then conditions the hypernetwork to generate LoRA adapters. This yields an explicit long-term memory for continual VQA.

Multimodal representations are constructed from the input image and question as follows. Given an image I, N salient regions are extracted and each region is embedded into a d-dimensional token by linearly projecting its RoI feature $\mathbf{f}_{j,\mathrm{roi}} \in \mathbb{R}^{d_{\mathrm{roi}}}$, adding a projected normalized bounding box $\mathbf{b}_j = [x_1, y_1, x_2, y_2] \in \mathbb{R}^4$ (each coordinate in $[0,1]$), and a learnable region-id embedding $\mathbf{e}_{j,\mathrm{id}} \in \mathbb{R}^d$. The resulting region token is denoted by $\mathbf{v}_j \in \mathbb{R}^d$. Stacking all region tokens yields $\mathbf{V} \in \mathbb{R}^{N\times d}$, and a global visual summary is computed by mean pooling:

$$\mathbf{f}_v = \frac{1}{N}\sum_{j=1}^{N} \mathbf{v}_j. \tag{1}$$

For a question Q with M tokens, a text encoder outputs token features $\mathbf{Q} \in \mathbb{R}^{M\times d}$, and a global question summary is obtained as

$$\mathbf{f}_q = \mathrm{Pool}(\mathbf{Q}), \tag{2}$$

where $\mathrm{Pool}(\cdot)$ denotes a pooling operator.

Within task i, a sample denotes an image–question pair. DMA updates the anchor bank online per sample, and the within-task index is omitted by using $\mathcal{P}_{v,i}$ and $\mathcal{P}_{q,i}$ to denote the current bank. At the beginning of task i, anchors are inherited from the previous task (for i=1, $\mathcal{P}_{v,0} = \{\mathbf{f}_v^{(k)}\}_{k=1}^{K_v}$ and $\mathcal{P}_{q,0} = \{\mathbf{f}_q^{(k)}\}_{k=1}^{K_q}$):

$$\begin{cases} \mathcal{P}_{v,i-1} = \{\mathbf{p}_{v,k,i-1}\}_{k=1}^{K_v} \\ \mathcal{P}_{q,i-1} = \{\mathbf{p}_{q,k,i-1}\}_{k=1}^{K_q} \end{cases}. \tag{3}$$

For each incoming sample, the most relevant anchors are retrieved by cosine similarity $\cos(\mathbf{a},\mathbf{b}) = \frac{\mathbf{a}^\top \mathbf{b}}{\|\mathbf{a}\|_2 \|\mathbf{b}\|_2}$:

$$\begin{cases} k_v^* = \arg\max\limits_{k\in\{1,\dots,K_v\}} \cos(\mathbf{f}_v, \mathbf{p}_{v,k,i-1}) \\ k_q^* = \arg\max\limits_{k\in\{1,\dots,K_q\}} \cos(\mathbf{f}_q, \mathbf{p}_{q,k,i-1}) \end{cases}. \tag{4}$$

Only the retrieved anchors are updated using momentum:

$$\begin{cases} \mathbf{p}_{v,k_v^*,i} \leftarrow \alpha\, \mathbf{p}_{v,k_v^*,i-1} + (1-\alpha)\, \mathbf{f}_v \\ \mathbf{p}_{q,k_q^*,i} \leftarrow \beta\, \mathbf{p}_{q,k_q^*,i-1} + (1-\beta)\, \mathbf{f}_q \end{cases}, \tag{5}$$

where $\alpha, \beta \in [0, 1)$ are sample-adaptive coefficients. For all non-retrieved anchors, they are kept unchanged, e.g., $\mathbf{p}_{v,k,i} = \mathbf{p}_{v,k,i-1}$ for $k \neq k_v^*$ and $\mathbf{p}_{q,k,i} = \mathbf{p}_{q,k,i-1}$ for $k \neq k_q^*$, yielding the updated bank $\mathcal{P}_{v,i}$ and $\mathcal{P}_{q,i}$.

Finally, the instance context is defined as

$$\mathbf{z}_{\text{inst},i} = \left[\mathbf{p}_{v,k_v^*,i}\,;\, \mathbf{p}_{q,k_q^*,i}\right]. \tag{6}$$

To further mitigate representation drift in continual learning, we introduce a memory consistency loss. We use $\mathcal{L}_{\text{mem}}$ to denote the aggregate of modality-specific memory terms:

$$\mathcal{L}_{\text{mem}} = \left(1 - \cos\left(\mathbf{f}_v, \mathbf{p}_{v,k_v^*,i}\right)\right) + \left(1 - \cos\left(\mathbf{f}_q, \mathbf{p}_{q,k_q^*,i}\right)\right). \tag{7}$$

In the next section 3.2, it is detailed how $\mathbf{z}_{\text{inst},i}$ conditions the hypernetwork to generate LoRA adapters.

3.2 Hypernetwork-Generated LoRA Module

The Hypernetwork-Generated LoRA Module (HyGen) enables continual VQA adaptation by generating lightweight LoRA adapters for a largely shared backbone. A hypernetwork produces these updates conditioned on the anchors retrieved from DMA, allowing sample-wise, dynamic adaptation under non-stationary streams. This design adapts the model by injecting Hypernetwork-Generated LoRA adapters into a shared backbone, maintaining parameter efficiency while reducing interference across tasks.

The construction starts from generating LoRA factors for each incoming sample within task i, conditioned on its instance representation $\mathbf{z}_{\text{inst},i}$. Let $H_\phi(\cdot)$ denote a hypernetwork with parameters ϕ. To enable layer-specific generation, we additionally introduce a learned layer embedding $\mathbf{e}_\ell \in \mathbb{R}^{d_\ell}$. For each fusion layer $\ell \in \{1, \ldots, L\}$, we generate LoRA factors for cross-attention projections as

$$\{\mathbf{A}_{\ell,X}, \mathbf{B}_{\ell,X}\}_{X \in \{Q,K,V\}} = H_\phi([\mathbf{z}_{\text{inst},i}; \mathbf{e}_\ell]), \tag{8}$$

where $X \in \{Q, K, V\}$ denotes the projection type (query/key/value), $\mathbf{A}_{\ell,X} \in \mathbb{R}^{r \times d_{\text{in}}}$, $\mathbf{B}_{\ell,X} \in \mathbb{R}^{d_{\text{out}} \times r}$, and the rank satisfies $r \ll \min(d_{\text{in}}, d_{\text{out}})$ (typically $d_{\text{in}} = d_{\text{out}} = d$).

Given the generated factors, the frozen cross-attention projections are augmented to obtain sample-adaptive projections. We represent each sequence as a matrix whose rows are token embeddings, and apply linear projections by right-multiplication. Let $\mathbf{W}_{\ell,\text{base},X} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$ denote the frozen base projection. The sample-adaptive projection is formed as

$$\mathbf{W}_{\ell,\text{adp},X} = \mathbf{W}_{\ell,\text{base},X} + \lambda\, (\mathbf{B}_{\ell,X}\mathbf{A}_{\ell,X})^\top, \tag{9}$$

where $\lambda \geq 0$ controls the adaptation strength.

In parallel to the parameter adaptation, the retrieved anchors are injected as memory-guided global context tokens by prepending them to the original sequences (concatenation along the sequence-length dimension). Here $\mathbf{Q} \in \mathbb{R}^{n_q \times d}$ and $\mathbf{V} \in \mathbb{R}^{n_v \times d}$ denote the input token sequences on the query side and value side of cross-attention, respectively (note that $\mathbf{V}$ here denotes a sequence, not the value projection). We compute

$$\begin{cases} \mathbf{g}_v = \mathbf{W}_g \mathbf{p}_{v,k_v^*,i} \\ \hat{\mathbf{V}} = [\mathbf{g}_v; \mathbf{V}] \end{cases}, \quad \begin{cases} \mathbf{g}_q = \mathbf{W}_g \mathbf{p}_{q,k_q^*,i} \\ \hat{\mathbf{Q}} = [\mathbf{g}_q; \mathbf{Q}] \end{cases}, \tag{10}$$

where $\mathbf{p}_{v,k_v^*,i} \in \mathbb{R}^d$ and $\mathbf{p}_{q,k_q^*,i} \in \mathbb{R}^d$ are the retrieved anchor embeddings for task i, and $\mathbf{W}_g \in \mathbb{R}^{d \times d}$ is learnable. Accordingly, $\mathbf{g}_v, \mathbf{g}_q \in \mathbb{R}^d$ are global context tokens.

With the augmented sequences and sample-adaptive projections in place, the model performs adaptive multimodal fusion and outputs the answer distribution. At layer ℓ, we compute cross-attention with adaptive projections as

$$\begin{cases} \mathbf{S}_\ell = \dfrac{\left(\hat{\mathbf{Q}}\mathbf{W}_{\ell,\text{adp},Q}\right)\left(\hat{\mathbf{V}}\mathbf{W}_{\ell,\text{adp},K}\right)^\top}{\sqrt{d_k}} \\ \mathbf{H}_\ell = \text{Softmax}(\mathbf{S}_\ell)\left(\hat{\mathbf{V}}\mathbf{W}_{\ell,\text{adp},V}\right) \end{cases}, \tag{11}$$

where d_k is the key dimension used for scaling (typically $d_k = d/h$ for h attention heads), and $\text{Softmax}(\cdot)$ is applied row-wise. For brevity, we omit the standard multi-head reshape and per-head projections in the notation.

After L fusion layers, we pool the fused sequence into $\mathbf{e}_{\text{out}}$ and predict an answer distribution $\pi \in [0, 1]^C$ via an MLP with Softmax. We construct a soft target distribution $\mathbf{t} \in [0, 1]^C$ from multiple human annotations.

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{C} t_c \log \pi_c. \tag{12}$$

We optimize the model by minimizing $\mathcal{L}_{\text{CE}}$.

3.3 Semantic–Functional Alignment loss

The Semantic–Functional (SF) Alignment loss stabilizes anchor-conditioned adaptation by enforcing consistency between semantic and functional changes. Specifically, it matches the semantic distance between the current input features and the retrieved anchors with the functional change measured in parameter space between the corresponding Hypernetwork-Generated LoRA adapters. The loss discourages inconsistent parameter updates and improves long-term robustness.

For task i, $\mathbf{z}_{\text{raw},i}$ is constructed by concatenating $\mathbf{f}_v$ and $\mathbf{f}_q$. The semantic discrepancy between the anchor-conditioned context and $\mathbf{z}_{\text{raw},i}$ is measured using cosine distance:

$$D_{\text{sem}}\left(\mathbf{z}_{\text{inst},i}, \mathbf{z}_{\text{raw},i}\right) = 1 - \cos(\mathbf{z}_{\text{inst},i}, \mathbf{z}_{\text{raw},i}). \tag{13}$$

To measure functional discrepancy, the hypernetwork outputs across all layers are vectorized, where $\text{vec}(\cdot)$ concatenates all generated matrices into a single vector:

$$\begin{cases} \theta_{\text{inst}} = \text{vec}(H_\phi(\mathbf{z}_{\text{inst},i})) \\ \theta_{\text{raw}} = \text{vec}(H_\phi(\mathbf{z}_{\text{raw},i})) \end{cases}, \tag{14}$$

where m is the total number of generated scalar parameters after vectorization, and a normalized ℓ_2 distance in parameter space is defined as:

$$D_{\text{func}}\left(\theta_{\text{inst}}, \theta_{\text{raw}}\right) = \frac{\|\theta_{\text{inst}} - \theta_{\text{raw}}\|_2}{\sqrt{m}}. \tag{15}$$

Methods	VQA v2					NExT-QA				
	Standard Test		Novel Comp. Test		#Mem	Standard Test		Novel Comp. Test		#Mem
	AP(↑)	AF(↓)	AP(↑)	AF(↓)		AP(↑)	AF(↓)	AP(↑)	AF(↓)	
Vanilla	14.49	30.80	11.79	27.16	5000	11.97	26.14	12.59	28.04	500
EWC (PNAS'17)	15.77	30.62	12.83	28.16	5000	13.01	24.06	11.91	27.44	500
MAS (ECCV'18)	20.56	11.16	23.90	6.24	5000	18.04	10.07	21.12	10.09	500
ER (MTLR'19)	36.99	5.99	33.78	5.76	5000	30.55	4.91	32.20	5.57	500
DER (NeurIPS'20)	35.35	8.62	31.52	8.59	5000	26.17	5.12	21.56	12.68	500
VS (CVPR'22)	34.03	8.79	32.96	5.78	5000	28.13	4.45	29.47	6.14	500
VQACL (CVPR'23)	38.77	3.96	35.40	4.90	5000	32.27	3.00	34.22	3.80	500
QUAD (ICCV'25)	39.25	4.91	40.00	3.81	5000	31.70	2.91	33.21	4.16	500
ProtoGroup (ICASSP'25)	39.81	2.87	36.90	4.06	5000	35.29	2.15	35.79	3.17	500
CLT-VQA (ICCV'25)	40.69	2.11	38.46	3.51	5000	35.77	2.09	35.86	3.64	500
HyLoVQA (Ours)	**45.41**	**2.82**	**44.96**	**2.62**	**5000**	**42.43**	**2.07**	**43.05**	**3.10**	**500**

Table 1: Comparison on VQA v2 and NExT-QA under Standard and Novel Composition tests.

Dataset	Methods	Group-1		Group-2		Group-3		Group-4		Group-5		Avg	
		Novel	Seen	Novel	Seen	Novel	Novel	Seen	Seen	Novel	Seen	Novel	Seen
VQA v2	ER	34.52	37.03	33.40	35.55	34.79	34.20	33.86	35.02	32.34	35.91	33.78	35.54
	DER	30.80	29.89	32.19	33.24	34.88	34.08	29.60	30.90	30.14	32.56	31.52	32.13
	VS	33.35	33.87	33.18	32.21	34.50	33.84	31.29	33.98	32.46	33.87	32.96	33.55
	VQACL	36.12	37.99	35.39	36.92	36.26	35.16	34.85	35.64	34.36	36.28	35.40	36.40
	QUAD	39.19	41.06	38.40	39.50	43.15	39.19	40.01	40.72	39.20	40.62	40.00	40.21
	ProtoGroup	36.23	36.85	35.76	37.91	38.70	39.57	36.74	37.85	37.65	38.19	36.90	37.78
	CLT-VQA	37.78	37.91	38.95	38.11	36.55	38.68	39.72	39.94	36.88	38.78	38.46	39.51
	HyLoVQA	**40.15**	**41.90**	**41.05**	**41.56**	**45.49**	**44.90**	**45.55**	**45.33**	**44.98**	**45.26**	**44.96**	**45.30**
	+Δ	0.96	0.84	2.10	2.06	2.34	5.33	5.54	4.61	5.78	4.64	4.96	5.09
NExT -QA	ER	31.86	34.51	32.36	35.08	29.50	34.30	33.57	33.30	33.71	32.91	32.20	34.02
	DER	27.56	26.09	26.14	24.54	23.53	26.43	9.30	9.79	21.26	23.74	21.56	21.38
	VS	31.42	30.88	29.17	31.26	25.23	26.10	30.01	29.10	31.54	31.79	29.47	29.83
	VQACL	35.50	35.54	33.97	35.91	31.34	35.62	34.08	33.57	36.71	33.46	34.22	34.82
	QUAD	33.42	33.92	32.02	35.42	31.78	36.36	32.98	33.34	37.84	34.06	33.21	34.62
	ProtoGroup	35.79	36.18	34.56	36.83	32.90	36.58	35.21	35.60	36.92	34.40	35.79	36.48
	CLT-VQA	36.14	36.46	35.98	37.03	38.71	37.32	36.44	36.92	36.39	35.89	35.86	36.79
	HyLoVQA	**40.23**	**41.56**	**41.77**	**42.21**	**42.97**	**43.28**	**42.39**	**43.96**	**43.15**	**43.08**	**43.05**	**43.97**
	+Δ	4.09	5.10	5.79	5.18	+4.26	+5.96	5.95	7.04	+5.31	7.19	+7.19	7.18

Table 2: Fine-grained VQA performance AP (%) on the Novel and Seen skill-concept compositions of VQA v2 and NExT-QA.

Dual-Space Consistency is used as an auxiliary objective: $\mathcal{L}_{\text{align}}$ encourages agreement between the anchor-conditioned semantic shift ($\mathbf{z}_{\text{inst}}$ vs. $\mathbf{z}_{\text{raw}}$) and the induced change in parameter space:

$$\mathcal{L}_{\text{align}} = \left(D_{\text{sem}}\left(\mathbf{z}_{\text{inst},i}, \mathbf{z}_{\text{raw},i}\right) - D_{\text{func}}\left(\theta_{\text{inst}}, \theta_{\text{raw}}\right)\right)^2. \tag{16}$$

The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \gamma_1 \mathcal{L}_{\text{mem}} + \gamma_2 \mathcal{L}_{\text{align}}, \tag{17}$$

where γ_1 and γ_2 control the relative weights of $\mathcal{L}_{\text{mem}}$ and $\mathcal{L}_{\text{align}}$, respectively. We minimize $\mathcal{L}_{\text{total}}$ by backpropagation, updating all trainable modules while keeping the base backbone projections frozen.

4 Experiments

To comprehensively evaluate HyLoVQA in continual VQA, we conduct experiments addressing six research questions:

- **RQ1:** How does HyLoVQA compare to baselines on Standard and Novel Composition tests, overall and fine-grained?
- **RQ2:** What is the contribution of each component to overall effectiveness and forgetting?
- **RQ3:** How effective are HyLoVQA's sample-specific LoRA adapters, and what evidence supports this?
- **RQ4:** What is the impact of DMA capacity on retention and forgetting across tasks?
- **RQ5:** What is HyLoVQA's sensitivity to key hyperparameters such as anchor momentum?
- **RQ6:** What failure modes of prior methods are revealed by case studies, and how does HyLoVQA address them?

Method			VQA v2		NExT-QA	
DMA	HyGen	Align.	AP (↑)	AF (↓)	AP (↑)	AF (↓)
✓	✗	✗	41.77$_{(-3.64)}$	3.46$_{(+0.64)}$	39.14$_{(-3.29)}$	3.79$_{(+1.72)}$
✓	✓	✗	44.96$_{(-0.45)}$	2.95$_{(+0.13)}$	41.65$_{(-0.78)}$	2.52$_{(+0.45)}$
✓	✓	✓	**45.41**	**2.82**	**42.43**	**2.07**

Table 3: Ablation study of HyLoVQA on VQA v2 and NExT-QA. Align. indicates adding the alignment loss term $\mathcal{L}_{align}$ to the training objective.

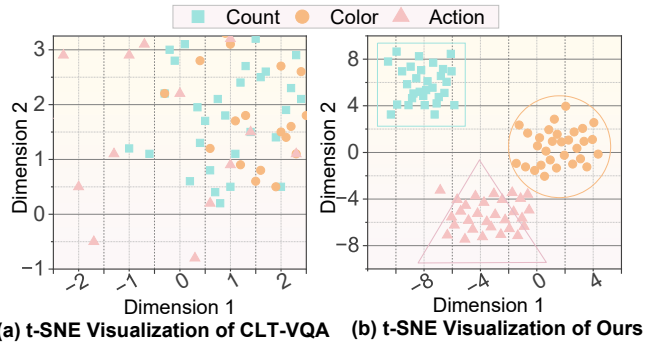

Figure 3: t-SNE visualization of sample feature representations on VQA v2 across task categories.

4.1 Experimental Settings

Datasets. We evaluate on VQA v2 (200k+ COCO images and 1.1M QA pairs [Lin *et al.*, 2014]) and NExT-QA, a video QA benchmark requiring temporal and causal reasoning [Xiao *et al.*, 2021]. Following the standard continual VQA protocol, VQA v2 is split into 10 question-type tasks (*recognition, location, judge, commonsense, count, action, color, type, subcategory, causal*), and NExT-QA into 8 tasks (*CW, TN, TC, DL, DB, DC, DO, CH*).

Evaluation Metrics. We report Final Average Performance (AP) and Average Forgetting (AF), which measure overall performance after continual learning and forgetting on earlier tasks, respectively[Lee and Toutanova, 2018].

Baselines. We compare HyLoVQA with sequential fine-tuning (Vanilla), EWC[Kirkpatrick *et al.*, 2017b], MAS[Aljundi *et al.*, 2018], ER[Chaudhry *et al.*, 2019], DER[Chaudhry *et al.*, 2018], VS[Wan *et al.*, 2022b], VQACL[Zhang *et al.*, 2023], QUAD[Marouf *et al.*, 2025], ProtoGroup[Zhang *et al.*, 2025b] and CLT-VQA[Zhang *et al.*, 2025a]. All methods are evaluated under the same task stream, backbone, and metrics for fair comparison.

Implementation Details. Evaluation follows two paradigms: *Standard* testing on seen task–object combinations and *Novel Composition* testing on held-out (unseen) combinations. For visual tokens, we use a Faster R-CNN[Ren *et al.*, 2015] pretrained on Visual Genome to extract N=36 RoI features per image on VQA v2, and an inflated 3D ResNeXt-101[Hara *et al.*, 2018] to extract N=16 clip-level motion features on NExT-QA. We use Adam [Adam and others, 2014] with a learning rate of 3e−5, gradient clipping of 5, and a warmup ratio of 0.1; γ_1 and γ_2 are tuned from {0.1, 1, 10}. We implement our proposed method based on PyTorch.

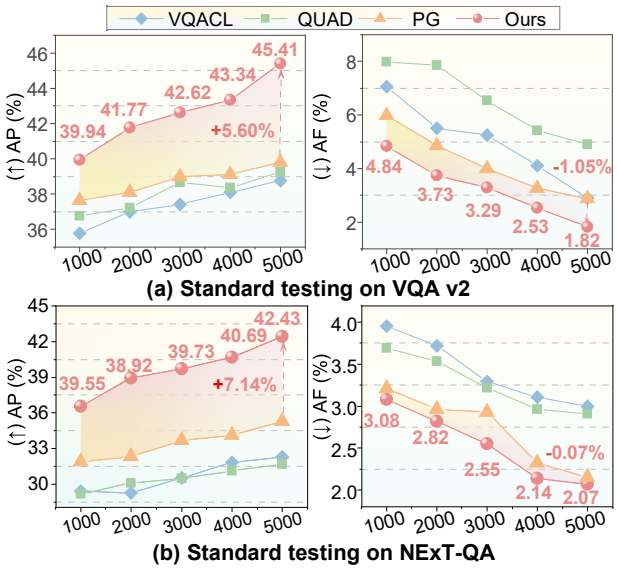

Figure 4: Memory size sensitivity analysis under standard testing on two datasets: (a) VQA v2 and (b) NExT-QA.

4.2 RQ1: Main Results

As shown in Table 1, HyLoVQA achieves the best overall performance on both datasets under Standard and Novel Composition evaluation. On VQA v2, HyLoVQA attains 45.41% AP/2.82% AF (Standard) and 44.96% AP/2.62% AF (Novel), yielding about +4.72% AP and +4.96% AP over the strongest baseline, respectively, while maintaining low forgetting. On NExT-QA, HyLoVQA reaches 42.43% AP/2.07% AF (Standard) and 43.05% AP/3.10% AF (Novel), improving AP by about +6.66% and +7.19% over the best baseline in the two settings. Table 2 further breaks down the Novel Composition setting under the group-removed protocol in Fig.2 (AP only), where we hold out one object group G_i during training and evaluate Novel compositions involving G_i versus Seen compositions over the remaining groups. HyLoVQA achieves the best AP on both Novel and Seen splits (VQA v2: 44.96%/45.30%; NExT-QA: 43.05%/43.97%), confirming consistent gains across group splits and stronger compositional generalization.

4.3 RQ2: Ablation Study

Table 3 shows that each component consistently contributes to continual VQA under the Standard Test setting. The values in parentheses denote the change relative to the full model (DMA+HyGen+Align.). Starting from DMA only, performance drops and forgetting increases (41.77% AP/3.46% AF on VQA v2; 39.14%/3.79% on NExT-QA). Adding HyGen improves both AP and AF (44.96%/2.95% on VQA v2; 41.65%/2.52% on NExT-QA), confirming that anchor-conditioned, hypernetwork-generated LoRA adapters enhance instance-specific adaptation and alleviate interference. Adding Align. further yields the best trade-off (45.41%/2.82% on VQA v2; 42.43%/2.07% on NExT-QA) by keeping LoRA adapters aligned with the current task and object and reducing off-target adaptation.

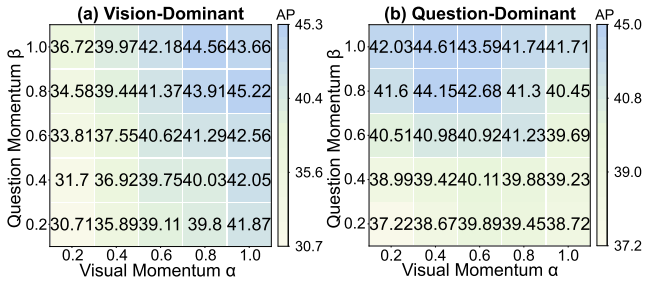

Figure 5: Sensitivity to modality-aware anchor momentum.

4.4 RQ3: Representation Stability (t-SNE)

This section examines whether HyLoVQA can efficiently adapt to the current sample input under continual distribution shift, where each sample consists of a visual object and a textual task. As shown in Figure 3, we visualize sample features on VQA v2 and color points by task category (*Count/Color/Action*). Baseline features exhibit substantial overlap across categories, while HyLoVQA yields more compact intra-category clusters and clearer inter-category separation. This indicates reduced cross-task representation entanglement, which is closely associated with forgetting and negative transfer in continual VQA. We attribute this behavior to the DMA providing stable anchor references and HyGen performing anchor-conditioned adaptation, consistent with the higher AP and lower AF in Table 1.

4.5 RQ4: Memory Capacity Robustness

To evaluate memory capacity robustness, we vary the anchor memory bank size from 1k to 5k and report AP/AF under Standard testing on two datasets: VQA v2 and NExT-QA (Figure 4). HyLoVQA consistently outperforms strong baselines (VQACL, QUAD, and ProtoGroup) across all capacities, achieving a better AP–AF trade-off without requiring a large replay buffer. On VQA v2 (Figure 4a), increasing the memory from 1k to 5k improves AP from 39.94% to 45.41% while reducing AF from 4.84% to 1.82%. On NExT-QA (Figure 4b), AP increases from 39.55% to 42.43% and AF decreases from 3.08% to 2.07% as the memory grows from 1k to 5k. Overall, HyLoVQA benefits steadily from additional memory across datasets, indicating effective and robust utilization of the anchor memory bank.

4.6 RQ5: Hyperparameter Sensitivity

We evaluate the sensitivity of the modality-aware anchor update momentum on NExT-QA. Figure 5 reports AP (%) over a grid of visual and question momentum (α, β) and shows clear task-specific preferences, highlighting the flexibility of our DMA design. On Vision-Dominant tasks (Figure 5(a)), performance concentrates in the high-α region and peaks at $(\alpha, \beta) = (1.0, 0.8)$ with 45.22% AP. In contrast, Question-Dominant tasks (Figure 5(b)) favor larger β, achieving the best AP of 44.61% at $(\alpha, \beta) = (0.4, 1.0)$. Overall, the optimal region shifts consistently with the dominant modality, validating that DMA can adapt its anchor/prototype update dynamics to diverse reasoning needs via modality-aware momentum control.

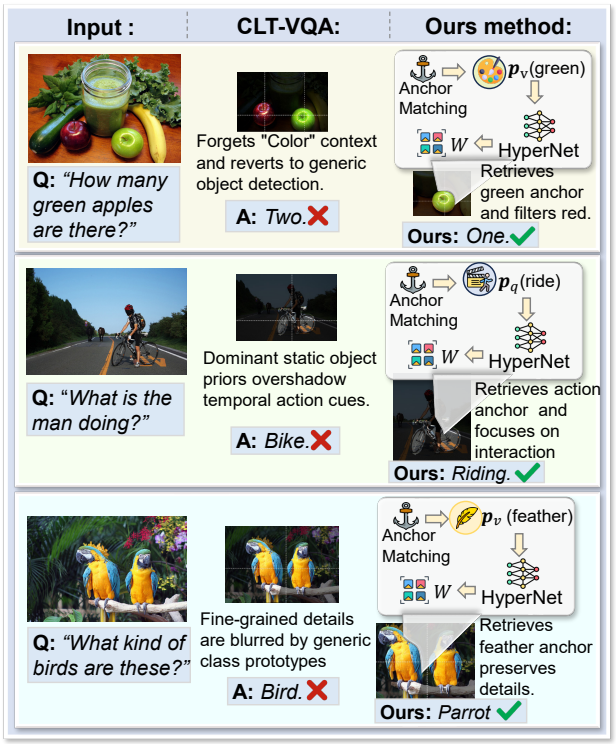

Figure 6: Qualitative case studies: CLT-VQA vs. HyLoVQA.

4.7 RQ6: Qualitative Case Study

Figure 6 illustrates failure modes of prior methods and how HyLoVQA addresses them. CLT-VQA often forgets sample-specific constraints and falls back to generic shortcuts, e.g., ignoring color-conditioned counting, relying on static object priors in action queries, or merging fine-grained categories into coarse labels. HyLoVQA retrieves relevant anchors from DMA and uses HyGen to generate anchor-conditioned LoRA adapters that steer cross-attention toward task- and object-specific cues: retrieving $\boldsymbol{p}_v$ (*green*) to suppress distractors, retrieving $\boldsymbol{p}_q$ (*ride*) to emphasize interaction regions for action recognition, and retrieving fine-grained visual anchors to preserve discriminative feather details. As a result, HyLoVQA preserves the intended constraints under drift and produces answers consistent with the queried attributes and interactions (e.g., *one*, *riding*, and *parrot*).

5 Conclusion

We propose HyLoVQA, a lightweight continual VQA framework that mitigates cross-level task interference while enabling dynamic adaptation to each incoming sample's task and object. With a drift-resilient memory anchor bank and a hypernetwork that generates task/object-specific LoRA adapters, HyLoVQA supports dynamic and parameter-efficient adaptation and better preserves prior knowledge. A semantic–functional alignment loss further keeps updates focused on the current task and object. Overall, HyLoVQA highlights memory-guided parameter generation as a simple and general route to balance stability and plasticity in continual VQA under non-stationary and compositional shifts.

Acknowledgments

This work was supported in part by the Natural Science Foundation of Hubei Province of China (No. 2025AFB653), the Open Fund of Hubei Key Laboratory of Big Data Intelligent Analysis and Application, Hubei University (No. 2025BDIAA01), and the National Natural Science Foundation of China (No. 62207011, 62407013, 62377009).

References

[Adam and others, 2014] Kingma DP Ba J Adam et al. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 1412(6), 2014.

[Aljundi *et al.*, 2018] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision*, pages 139–154, 2018.

[Antol *et al.*, 2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[Chaudhry *et al.*, 2018] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European conference on computer vision*, pages 532–547, 2018.

[Chaudhry *et al.*, 2019] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, P Dokania, P Torr, and M Ranzato. Continual learning with tiny episodic memories. In *Workshop on Multi-Task and Lifelong Reinforcement Learning*, 2019.

[Chen *et al.*, 2020] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Proceedings of the European conference on computer vision*, pages 104–120, 2020.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 4171–4186, 2019.

[Gama *et al.*, 2014] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37, 2014.

[Gómez-Pérez and Ortega, 2020] José Manuél Gómez-Pérez and Raúl Ortega. ISAAQ - mastering textbook questions with pre-trained transformers and bottom-up and top-down attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 5469–5479, 2020.

[Goyal *et al.*, 2017] 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.

[Greco *et al.*, 2019] Claudio Greco, Barbara Plank, Raquel Fernández, and Raffaella Bernardi. Psycholinguistics meets continual learning: Measuring catastrophic forgetting in visual question answering. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 3601–3605, 2019.

[Hara *et al.*, 2018] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

[Hayes *et al.*, 2020] Tyler L Hayes, Kushal Kafle, Robik Shrestha, Manoj Acharya, and Christopher Kanan. Remind your neural network to prevent catastrophic forgetting. In *Proceedings of the European conference on computer vision*, pages 466–483, 2020.

[Johnson *et al.*, 2017] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

[Kim *et al.*, 2018] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *Advances in neural information processing systems*, 31, 2018.

[Kim *et al.*, 2024] Junsu Kim, Yunhoe Ku, Jihyeon Kim, Junuk Cha, and Seungryul Baek. Vlm-pl: Advanced pseudo labeling approach for class incremental object detection via vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4170–4181, 2024.

[Kirkpatrick *et al.*, 2017a] 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.

[Kirkpatrick *et al.*, 2017b] 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.

[Lee and Toutanova, 2018] JDMCK Lee and K Toutanova. Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 3(8):4171–4186, 2018.

[Li *et al.*, 2019] Xilai Li, Yingbo Zhou, Tianfu Wu, Richard Socher, and Caiming Xiong. Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In *Proceedings of the International conference on machine learning*, pages 3925–3934, 2019.

[Liang *et al.*, 2020] Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. Learning to contrast the counterfactual samples for robust visual question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing*, pages 3285–3292, 2020.

[Lin *et al.*, 2014] 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 *Proceedings of the European conference on computer vision*, pages 740–755, 2014.

[Lopez-Paz and Ranzato, 2017] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

[Marouf *et al.*, 2025] Imad Eddine Marouf, Enzo Tartaglione, Stéphane Lathuilière, and Joost Van De Weijer. Ask and remember: A questions-only replay strategy for continual visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18078–18089, 2025.

[Paik *et al.*, 2020] Inyoung Paik, Sangjun Oh, Taeyeong Kwak, and Injung Kim. Overcoming catastrophic forgetting by neuron-level plasticity control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5339–5346, 2020.

[Parisi *et al.*, 2019] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71, 2019.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[Rusu *et al.*, 2016] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

[Sun *et al.*, 2023] Lingfeng Sun, Haichao Zhang, Wei Xu, and Masayoshi Tomizuka. Efficient multi-task and transfer reinforcement learning with parameter-compositional framework. *IEEE Robotics and Automation Letters*, 8(8):4569–4576, 2023.

[Theeuwes, 2010] Jan Theeuwes. Top–down and bottom–up control of visual selection. *Acta psychologica*, (2):77–99, 2010.

[Wan *et al.*, 2022a] Timmy ST Wan, Jun-Cheng Chen, Tzer-Yi Wu, and Chu-Song Chen. Continual learning for visual search with backward consistent feature embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16702–16711, 2022.

[Wan *et al.*, 2022b] Timmy ST Wan, Jun-Cheng Chen, Tzer-Yi Wu, and Chu-Song Chen. Continual learning for visual search with backward consistent feature embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16702–16711, 2022.

[Whitehead *et al.*, 2021] Spencer Whitehead, Hui Wu, Heng Ji, Rogerio Feris, and Kate Saenko. Separating skills and concepts for novel visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5632–5641, 2021.

[Xiao *et al.*, 2021] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021.

[Zhang *et al.*, 2021] Xi Zhang, Feifei Zhang, and Changsheng Xu. Multi-level counterfactual contrast for visual commonsense reasoning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1793–1802, 2021.

[Zhang *et al.*, 2023] 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.

[Zhang *et al.*, 2025a] Feifei Zhang, Zhihao Wang, Xi Zhang, and Changsheng Xu. Overcoming dual drift for continual long-tailed visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4413–4423, 2025.

[Zhang *et al.*, 2025b] Licheng Zhang, Zhendong Mao, Yixing Peng, Zheren Fu, and Yongdong Zhang. Multi-prototype grouping for continual learning in visual question answering. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2025.

[Zhou *et al.*, 2025] Da-Wei Zhou, Yuanhan Zhang, Yan Wang, Jingyi Ning, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Learning without forgetting for vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.