Vision-language Incremental Learning with Dual Class-individual Memory

Fuhai Chen[1], **Feng Zhang**[1], **XiaoGuang Ma**[2], **Yiyi Zhou**[3], **Jiarong Liu**[1], **Xuri Ge** [4*]

[1]College of Computer and Data Science, Fuzhou University, China
[2]College of Information Science and Engineering Northeastern University, China
[3]Institute of Artificial Intelligence, Xiamen University, China
[4]School of Artificial Intelligence, Shandong University, China
xuri.ge@sdu.edu.cn

Abstract

The emergence of multimodal technologies has propelled Vision-Language Incremental Learning (VLIL) into a research spotlight. Current VLIL approaches predominantly inherit unimodal paradigms, failing to address fundamental distinctions between visual and linguistic modalities. Crucially, the semantic gap between images and text creates divergent learning dynamics: visual data exhibits rich, distributed information while textual representations remain explicit and compact. Consequently, textual elements align with class-specific tasks, whereas individual images inherently span multiple such tasks, creating dual bottlenecks in class-level memory allocation and scene-level knowledge transfer. To overcome these challenges, we propose DCIM (Dual Class-Individual Memory), a novel framework featuring complementary mechanisms for vision-language continual learning. For class-level constraints, our Hierarchical Class Memory Management (HCMM) strategy dynamically allocates memory resources across object categories. It employs forgetting simulation to identify and preserve the most vulnerable samples, ensuring robust long-term knowledge retention. For scene-level adaptation, the Scene Reconstruction Memory(SRM) module captures generalized environmental representations, enabling contextual transfer to novel classes and disambiguation of semantically related concepts within shared scenes.Extensive experiments on two vision-language tasks, i.e., visual question answering (VQA) and Image captioning (IC), demonstrate the effectiveness and excellent generalization ability of our approach, achieving superior performance under continual learning settings.

Code — https://github.com/fofo1117/DCIM

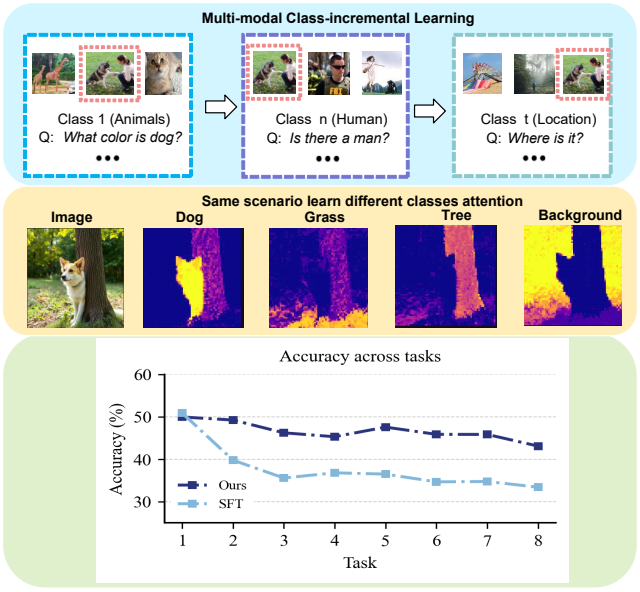

Figure 1: Illustration of challenges in multi-modal class-incremental learning. Top: Sequential learning of new classes (Animals → Humans → Locations) while maintaining knowledge of previous classes. Middle: Attention maps showing how different objects (dog, grass, tree) share similar visual contexts within the same scene, requiring the model to distinguish between coexisting concepts. Bottom: Experimental results demonstrating catastrophic forgetting in baseline methods and the effectiveness of our approach in maintaining performance across tasks.

Introduction

Vision-Language Models (VLMs) aim to align visual and textual modalities for comprehensive multimodal understanding. Powered by large-scale pre-training and cross-modal alignment strategies, recent advances have achieved impressive results on downstream tasks such as Visual Question Answering (VQA) (Antol et al. 2015; Teney, Liu, and van Den Hengel 2017) and Image Captioning (IC) (Anderson et al. 2018) under static benchmark settings. However, these models often struggle to generalize to real-world scenarios, where information evolves and updates continuously over time. In such dynamic environments, an ideal multimodal system should be capable of incrementally acquiring new knowledge while retaining previously learned capabilities. This fundamental challenge is known as *vision-language incremental learning* (VLIL).

Existing methods largely inherit unimodal paradigms and struggle to address the inherent structural asymmetry between visual and textual modalities. Textual features typically exhibit compact and explicit semantic structures with

*Corresponding Authors.

strong continuity and memory retention, resulting in relatively mild forgetting during continual learning. In contrast, visual semantics are spatially distributed, highly context-dependent, and lack a unified structure. When new categories are introduced, multimodal alignment mechanisms often shift visual attention toward novel objects, intensifying the issue of catastrophic forgetting.

Despite the severity of this modality-specific forgetting problem, current VLIL approaches have primarily focused on coarse-grained adaptation scenarios. For instance, continual learning for VQA (Zhang, Zhang, and Xu 2023; Lei et al. 2023; Lao et al. 2023) has concentrated on adapting to new question types (e.g., "how many", "why") or novel scene domains (e.g., office, outdoor), addressing task-level or scene-level forgetting when new tasks emerge. However, in practical applications, the emergence of entirely new question formats or complete domain shifts occurs relatively infrequently. Instead, as shown in Figure 1, the continuous introduction of new object classes within familiar question formats and visual scenes represents the predominant pattern of knowledge evolution, presenting a more fine-grained and challenging forgetting scenario that existing methods fail to adequately address.

Motivation. To this end, this paper investigates two core problems in vision-language incremental learning. First, as shown at the top of Fig.1, the model must maintain its knowledge of previously learned classes while continuously learning new ones. Second, as shown in the middle part of Fig.1, multiple visual elements often coexist within a single scene. This requires the model not only to recall previously learned class-specific knowledge from similar contexts, but also to distinguish between multiple concepts within the same scene, thereby preventing knowledge confusion.

Proposed method. To address the above issues, we propose a novel vision-language incremental learning framework with Dual Class-Individual Memory (DCIM). Contrary to the traditional training splitting approach, we divide the database according to object superclasses in multimodality, each of which contains multiple fine-grained sub-classes. In this way, the DCIM consists of two core components, i.e. a Hierarchical Class Memory Management (HCMM) and a Scene Reconstruction Memory (SRM). HCMM is used for the knowledge memory storage of trained superclasses, which adaptively updates the memory buffer capacity allocation based on forgetting probability for each superclass after training on new superclass data. By boosting the sampling of highly forgettable superclass samples in the class memory space beyond the uniformly allocated training model, the learning of highly forgettable classes is achieved while ensuring the learning of new class. To achieve this, we design a novel forgetting simulation strategy to precisely identify and prioritize samples with the highest risk of being forgotten for efficient memory management.

Additionally, we designed a Scene Reconstruction Memory (SRM) module to create connections between previously acquired class knowledge and newly introduced classes. On the one hand, we perform global scene representation capability by reconstructing the image, and on the other hand, if the reconstruction loss is too high during inference, we introduce less loss image features that are most similar to the current scene for fusion. This facilitates knowledge transfer between object classes that share similar visual contexts, enabling the model to leverage familiar environmental cues for learning new concepts.

Our comprehensive experiments across multiple benchmarks demonstrate significant improvements over state-of-the-art continual learning methods. Overall, our contributions can be summarized as follows:

1. We propose a novel Hierarchical Class Memory Management (HCMM) strategy, which optimizes the utilization of limited buffer space through task-adaptive memory allocation and forgetting simulation, enabling more efficient knowledge retention across incrementally learned classes.
2. We introduce a Scene Reconstruction Memory (SRM) module that enables the model to leverage contextual similarities across different classes through knowledge transfer. This mechanism not only mitigates catastrophic forgetting but also enhances generalization to new class tasks by exploiting shared visual contexts.
3. We establish new standardized benchmarks for class-incremental learning on vision-language tasks (VQA and IC) that better reflect practical, real-world scenarios. Our extensive evaluation on these benchmarks provides a robust platform to facilitate future research in this critical direction.

Related Works

Vision-Language Models

Recent advances in Vision-Language Models (VLMs) (Zhang et al. 2024; Chen et al. 2019; Ge et al. 2024; He et al. 2025) have demonstrated remarkable capabilities in bridging visual and textual modalities across various tasks, including Visual Question Answering (VQA) (Zhang, Zhang, and Xu 2023; Lei et al. 2023) and Image Captioning (IC) (Del Chiaro et al. 2020; Chen et al. 2017, 2018). While modern approaches, particularly those leveraging large-scale pre-training (Huang et al. 2021; Li et al. 2020) and Transformer architectures (Lu et al. 2016), achieve impressive performance on static benchmarks, they face significant challenges when deployed in dynamic, real-world scenarios. The reliance on fixed datasets renders these models vulnerable to catastrophic forgetting when learning new classes or concepts incrementally. This fundamental limitation significantly hinders their practical deployment in evolving real-world environments. Addressing this challenge requires developing specialized frameworks for class-incremental learning in vision-language tasks that can continually adapt to new knowledge while preserving previously acquired capabilities.

Class-incremental Learning

Class-Incremental Learning (CIL) (Zhou et al. 2024; Masana et al. 2022) represents a specific paradigm within Continual Learning (De Lange et al. 2021; Lesort et al. 2020) where models sequentially learn to recognize new

classes across a series of tasks. A key characteristic of CIL is that classes from previous tasks are either completely absent in subsequent tasks or represented by extremely limited exemplars during new task training. The fundamental objective of CIL is to continuously acquire knowledge about new classes while maintaining performance on previously learned classes. Replay-based approaches (Chaudhry et al. 2019; Lopez-Paz and Ranzato 2017; Gao and Liu 2023) maintain a fixed-capacity memory buffer containing a small subset of historical examples to preserve previously acquired knowledge during new task learning. Given the buffer's limited capacity, strategic selection of representative samples becomes crucial for maintaining method effectiveness. For instance, Reservoir Sampling (Vitter 1985)randomly selects original samples for preservation. iCaRL (Rebuffi et al. 2017) employs a herding mechanism based on feature representations to ensure balanced sample selection across classes. MRFA (Zheng et al. 2024) explores the problem of buffer case overfitting from the perspective of feature boundaries. Additionally, numerous outstanding works (Yan et al. 2022; Ye and Bors 2020; Xiang et al. 2019) have leveraged generative models to obtain exemplars. A significant challenge in sample selection lies in our inability to anticipate which samples will be prone to forgetting during future tasks while learning the current task. This challenge becomes particularly pronounced in multi-modal tasks, where the combinations of text and images are infinitely varied, and even text and images with distinctive features may combine to form easily forgettable samples. To address this, we propose a simple yet elegant and effective method for buffer management that maximize knowledge preservation across tasks. Our approach recognizes the unique characteristics of forgetting in multimodal reasoning and implements targeted mechanisms to counteract this phenomenon.

Method

Task Formulation

Incremental Learning in VQA and IC. The vision-language incremental learning addresses the challenge of continually acquiring new visual concepts without forgetting prior knowledge, under stable language patterns. This setting reflects real-world applications where systems such as visual question answering and image captioning continuously encounter new object classes after deployment.

To simulate such dynamics, we organize the dataset based on the object classes referenced in multimodal samples. For example, in VQA, a question like "Is this a dog?" is assigned to the "dog" class, which can be grouped under a higher-level superclass "animal". For IC, samples are similarly categorized according to the entities mentioned in the captions. Detailed specifications are provided in the Supplementary Materials. Formally, given a dataset $D = \{(I_i, Q_i, A_i)\}_{i=1}^{N}$ of N samples, where I_i denotes images, Q_i questions/captions, and A_i answers/descriptions. Then, we split the data into a sequence of T class-incremental tasks. During training on class task t, the model is trained solely on the current class data D_t, which does not appear in previous classes, forming a class-incremental learning scenario. The central challenge lies in learning new knowledge in class-task D_t without forgetting previously learned knowledge.

Benchmark Construction. We establish standardized benchmarks on VQA2.0 (Goyal et al. 2017) and MS-COCO datasets (Chen et al. 2015), partitioning them into 8 sequential class-tasks. To ensure robustness, we evaluate on both forward and reverse task orderings for VQA and IC, respectively. This bidirectional evaluation mitigates order-dependent biases and offers deeper insights into forgetting patterns under different learning sequences.

Overview

Our DCIM framework addresses catastrophic forgetting in class-incremental learning through two complementary memory mechanisms: sample-level memory management and scene-level knowledge preservation. As shown in Figure 2, DCIM consists of three core components: (1) a standard encoder-decoder architecture that processes multimodal inputs (image I, question Q for VQA or None for IC) to generate answers or captions, (2) a Hierarchical Class Memory Management (HCMM) module that dynamically allocates memory buffer M based on class-specific forgetting patterns, and (3) a Scene Reconstruction Memory (SRM) module that captures and preserves visual scene-level knowledge F_{scene} to facilitate cross-class transfer.

Hierarchical Class Memory Management

This module addresses both class-level and sample-level forgetting by implementing a Dynamic Memory Allocation strategy and a Forgetting Simulation Strategy to optimize the allocation of limited memory space and sample selection, thereby enhancing memory buffer quality.

Dynamic Memory Allocation. Traditional approaches (Rebuffi et al. 2017) allocate equal memory space across all previous classes, ignoring their varying forgetting characteristics. Our key insight is that classes exhibit different forgetting patterns—some classes are inherently more stable, while others are prone to catastrophic forgetting. Inspired by this observation, we propose a dynamic allocation strategy that dynamically adjusts memory distribution based on each class's historical forgetting behavior.

Specifically, given a fixed memory buffer of size M_{total}, when learning class t, we need to accommodate new samples while preserving knowledge from classes $\{1, ..., t-1\}$. Our strategy computes class-specific contribution weights based on their forgetting resilience. For each previous class i, we first calculate its average forgetting rate $\bar{F}_i$ over past learning episodes, where the forgetting rate refers to the difference between the best performance when training class i and the performance after training for the current class t. Then, the contribution weight is computed as:

$$w_i = \frac{1/(\bar{F}_i + \epsilon)}{\sum_{j=1}^{t-1}(1/(\bar{F}_j + \epsilon))} \tag{1}$$

Using these weights, we then determine the new memory allocation for each class-task:

$$M_i^{\text{new}} = M_i^{\text{current}} - \left(\frac{M_{\text{total}}}{t}\right) \cdot w_i \tag{2}$$

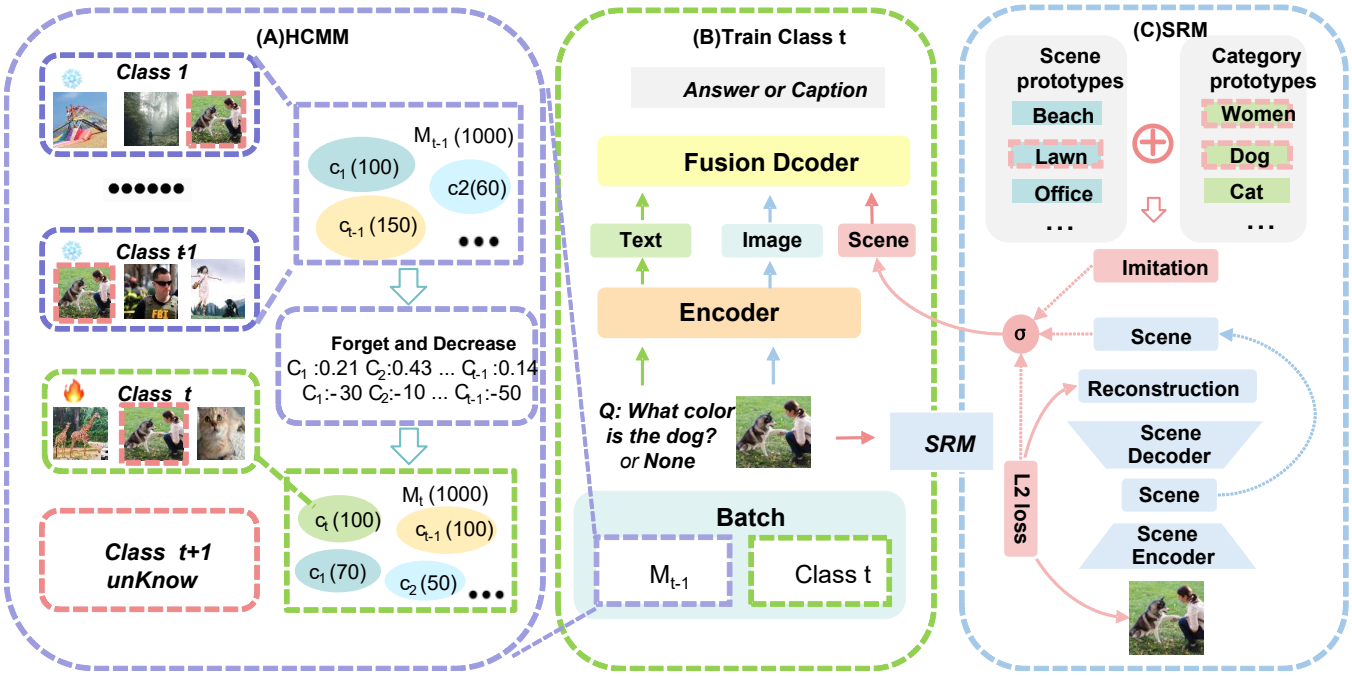

Figure 2: Architecture of the proposed Dual Class-Individual Memory (DCIM) framework for multimodal incremental learning. (A) Hierarchical Class Memory Management (HCMM) dynamically allocates memory buffer across tasks based on forgetting patterns. (B) Standard encoder-decoder processes multimodal inputs to generate answers/captions. (C) Scene Reconstruction Memory (SRM) extracts scene-level features and adaptively combines them with prototypical representations through a gating mechanism for robust cross-task knowledge transfer.

where M_i^{new} and M_i^{current} are the new and current memory allocations for class i. w_i is its contribution weight. $\bar{F}_i$ is its average forgetting rate. M_{total} is the total buffer capacity. And ϵ is a stabilization constant. These formulations ensure that class tasks with lower forgetting rates (more stable) contribute more memory space for reallocation.

Forgetting Simulation Strategy. To optimize memory buffer quality, we need to identify samples that are both learnable and forgetting-prone to fill the memory M. The key challenge lies in predicting which samples will be forgotten when learning future tasks, as we cannot access future data during current training.

Previous methods simulate forgetting through random parameter perturbations or gradient ascent, measuring sample vulnerability under artificial disturbances. However, these approaches poorly approximate real forgetting patterns during continual learning. We propose a more realistic simulation: using existing memory samples from previous tasks to estimate how current samples might be affected by future learning.

For each sample in the current task t, we evaluate its memory priority through a three-stage process. Let $L(\cdot)$ denote the loss function computed on this sample:

- **Stage 1:** Measure initial loss L_{init} on the untrained model.
- **Stage 2:** Measure post-training loss L_{trained} after learning task t.
- **Stage 3:** Measure simulated loss L_{replay} after training on memory buffer samples.

Based on these measurements, we compute two complementary scores:

Learning Score: Quantifies how much the model improved on this sample: $S_{\text{learn}} = L_{\text{init}} - L_{\text{trained}}$. A higher S_{learn} indicates the sample was successfully learned (neither too simple nor too difficult).

Forgetting Score: Measures vulnerability to interference from other tasks: $S_{\text{forget}} = L_{\text{replay}} - L_{\text{trained}}$. A higher S_{forget} indicates the sample is easily forgotten when learning other data.

The final memory priority combines both factors:

$$\text{Priority} = \alpha \cdot S_{\text{learn}} + (1 - \alpha) \cdot S_{\text{forget}} \tag{3}$$

where $\alpha \in [0, 1]$ balances the importance of learnability versus forgetting vulnerability. Samples with higher priority scores are retained in the memory buffer, ensuring we preserve those that are both meaningful to learn and susceptible to catastrophic forgetting.

Scene Reconstruction Memory

To achieve individual-level scene recall across different samples, our model incorporates Scene Reconstruction Memory (SRM) module inspired by the AutoEncoder architecture to extract individual-level scene features from visual representations. This module compresses the visual individual features $F_v \in \mathbb{R}^{d_v}$ into a low-dimensional panoramic vector $F_{\text{scene}} \in \mathbb{R}^{d_s}$ ($d_s \ll d_v$).

Scene Encoder compresses high-dimensional visual features into a compact scene representation:

$$F_{\text{scene}} = \phi_2(W_2 \cdot \text{LN}(\phi_1(W_1 \cdot F_v))) \tag{4}$$

where W_* represents learnable mapping weights, LN denotes layer normalization, and ϕ_1, ϕ_2 are activation functions.

Scene Decoder reconstructs the original visual features from the compressed representation to ensure information preservation:

$$F_{\text{recon}} = \tanh(W_4 \cdot \text{LN}(\phi_3(W_3 \cdot F_{\text{scene}}))) \tag{5}$$

where $\tanh$ activation bounds the reconstructed features to match the normalized input range.

The reconstruction quality is measured by the L2 distance between original and reconstructed features:

$$L_{\text{recon}} = \|F_v - F_{\text{recon}}\|_2^2 \tag{6}$$

This reconstruction loss serves dual purposes: during training, it guides the learning of meaningful scene representations; during inference, it provides a confidence score for the reliability of extracted scene features, which becomes crucial for handling out-of-distribution inputs .

Adaptive Scene Integration

During inference, the model may encounter novel scenes that differ significantly from training data, resulting in unreliable scene reconstructions. To maintain robust performance, we develop an adaptive integration mechanism that dynamically combines reconstructed scene features with learned prototypical representations.

Confidence Gating. We introduce a gating mechanism that evaluates the reliability of reconstructed scene features based on three factors:

$$g = \sigma(W_g \cdot [F_{\text{scene}}; L_{\text{recon}}; \text{sim}(F_{\text{scene}}, P^*)]) \tag{7}$$

Where P^* is the nearest scene prototype, $\text{sim}(\cdot,\cdot)$ computes cosine similarity, and σ is the sigmoid activation. The gate learns to predict reconstruction reliability by jointly considering feature quality, reconstruction error, and similarity to known patterns.

Prototypical Representation. When scene reconstruction is unreliable, we leverage prototypical representations learned during training as robust alternatives. We maintain two complementary types of prototypes:

Category prototypes: For each object category c, we compute the mean visual representation:

$$\bar{F}_c = \frac{1}{N_c} \sum_{i=1}^{N_c} F_{v,i}^{(c)} \tag{8}$$

where $F_{v,i}^{(c)}$ denotes the i-th visual feature from category c.

Scene prototypes: We discover k recurring scene patterns via clustering:

$$\{P_1, P_2, ..., P_k\} = \text{k-means}(\{F_{\text{scene},i}\}_{i=1}^{N}, k) \tag{9}$$

These prototypes capture common visual contexts (e.g., indoor, outdoor, urban environments).

For a given input, we construct a hybrid representation by combining the nearest scene prototype P^* with relevant category information:

$$F_{\text{proto}} = W_{\text{proj}} \cdot [P^*; \bar{F}_{c_1}; \bar{F}_{c_2}; \ldots; \bar{F}_{c_n}] + b_{\text{proj}} \tag{10}$$

where $c_1, ..., c_n$ are the top-n relevant categories based on the current input, and $W_{\text{proj}} \in \mathbb{R}^{d_s \times (d_s + n \cdot d_v)}$ projects the concatenated features back to the scene dimension.

Adaptive Fusion. The final scene representation combines both sources weighted by the confidence gate:

$$F_{\text{scene}}^{\text{final}} = g \cdot F_{\text{scene}} + (1 - g) \cdot F_{\text{proto}} \tag{11}$$

When $g \approx 1$ (high reconstruction confidence), the model relies primarily on the reconstructed scene features. When $g \approx 0$ (low confidence), it falls back to prototypical representations. This adaptive mechanism ensures robust scene understanding even for out-of-distribution inputs.

Experiments

Implementation Details

Model Architecture. We employ VL-T5 (Cho et al. 2021) as our backbone architecture, which consists of 12 stacked encoder and decoder blocks, each with 12 attention heads and an embedding dimension $d = 768$. For visual feature extraction, we utilize Faster R-CNN (Ren et al. 2015) pretrained on Visual Genome (Krishna et al. 2017) to extract 36 region features per image.

Training Configuration. We train each task for 3 epochs with a batch size of 128, using the Adam optimizer (Kingma and Ba 2014) with an initial learning rate of 1×10^{-4} and a linear decay schedule. The memory buffer M maintains a fixed capacity of 1,000 samples throughout all experiments. For the Scene Reconstruction module, we set the scene feature dimension $d_s = 256$ and the reconstruction loss weight $\lambda_{recon} = 0.005$. All experiments are implemented in PyTorch and conducted on two NVIDIA RTX 3090 GPUs.

Task Division Strategies. We evaluate our framework using two task division strategies: (1) **Taxonomy-Driven Division (TDD)**, where classes within each task share taxonomic relationships (e.g., all animals), testing semantic coherence exploitation; (2) **Diverse-Driven Division (DDD)**, where classes are randomly distributed across tasks, testing learning of disparate concepts.

Baseline Methods. To rigorously evaluate our approach, we compare against established continual learning methods: iCaRL (Rebuffi et al. 2017), TAM-CL (Cai, Thomason, and Rostami 2023), ER (Chaudhry et al. 2019), DER (Buzzega et al. 2020), VQACL (Zhang, Zhang, and Xu 2023), GaB (Das et al. 2025), DECO (Luo et al. 2024)

Additionally, we establish performance boundaries with two reference models:

- Sequential Fine-Tuning (SFT): A lower-bound baseline that performs simple gradient updates without any anti-forgetting mechanisms.
- Upper Bound (UB): An ideal scenario where the model has simultaneous access to all task data during training.

Method	Memory Size	VQA2.0-TDD			R-VQA2.0-TDD			VQA2.0-DDD			R-VQA2.0-DDD		
		AP	F	RF	AP	F	RF	AP	F	RF	AP	F	RF
UB	None	51.64	-	-	51.64	-	-	51.64	-	-	51.64	-	-
SFT	None	34.51	19.80	33.84	34.33	19.31	35.54	37.61	16.32	28.67	37.44	16.51	26.94
iCaRL	1000	36.92	14.21	27.94	38.32	13.45	25.34	39.67	11.41	23.41	40.17	10.23	21.36
ER	1000	42.54	6.39	12.97	42.33	6.81	13.17	43.78	5.19	8.93	43.52	5.34	9.12
DER	1000	43.55	5.77	11.76	43.61	5.62	11.62	44.61	3.52	6.07	44.35	3.66	6.17
TAM-CL	1000	43.39	6.56	13.11	43.17	6.69	13.23	44.27	4.21	7.89	43.91	4.48	8.01
VQACL	1000	42.24	7.71	15.93	42.36	7.38	15.61	43.96	4.77	9.81	43.81	4.62	9.51
DECO	1000	43.71	6.14	12.63	43.25	6.89	12.51	44.34	4.62	7.51	44.23	4.25	6.84
GaB	1000	43.59	6.08	12.44	43.36	6.73	12.32	44.59	4.32	6.64	44.21	3.82	6.34
DCIM	1000	44.81	5.94	11.63	44.11	6.24	11.78	45.53	3.86	5.92	45.14	3.91	6.45
DCIM (KD)	1000	**45.03**	**5.32**	**11.41**	**44.56**	**5.27**	**11.26**	**46.14**	**3.28**	**5.64**	**45.49**	**3.46**	**5.78**

Table 1: Performance comparison across VQA-CIL benchmarks. Results show Average Performance (AP), Forgetting Measure (F), and Relative Forgetting (RF) on four settings. Our dual-memory approach consistently outperforms existing methods. Upper bound represents simultaneous training on all tasks (Memory = None), with F and RF metrics marked as ”-” as they don’t apply to non-incremental learning.

Method	Memory Size	COCO-CIL		R-COCO-CIL	
		BLEU-4	F	BLEU-4	F
UB	None	25.7	-	25.7	-
SFT	None	14.5	13.9	13.8	13.7
iCaRL	1000	17.5	5.9	18.3	5.7
ER	1000	20.8	4.1	21.1	3.9
DER	1000	22.3	3.2	22.9	3.5
TAM-CL	1000	21.7	3.4	21.3	3.7
DCIM	1000	**23.6**	**2.4**	**22.9**	**2.7**

Table 2: Performance comparison across IC-CIL benchmarks. We report BLEU-4 for generative quality (higher is better) and the Forgetting measure F (lower is better). All metrics are presented in percentages (%).

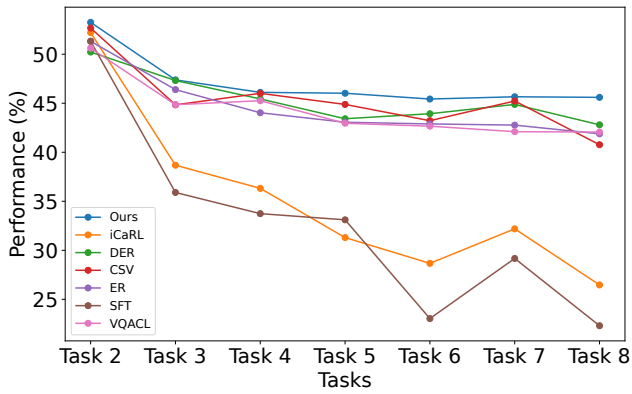

Figure 3: Performance trajectory across sequential tasks on VQA2.0-TDD benchmark. Our approach demonstrates superior initial learning capability at Task 2 and maintains a stable performance trajectory throughout Tasks 3-8, indicating effective resistance to catastrophic forgetting.

Since knowledge distillation is widely adopted in continual learning methods (e.g., DER, TAM-CL) despite its computational overhead, we also present **DCIM (KD)** to ensure fair comparison and demonstrate our framework’s compatibility with this complementary technique.

Evaluation Metrics. Following standard continual learning evaluation protocols (Chaudhry et al. 2018b,a), we employ three key metrics to comprehensively assess model performance:

Average Performance (AP): Quantifies the mean performance across all previously encountered tasks after completing the current task.

Forgetting Measure (F): Measures knowledge loss by quantifying performance degradation for each task compared to when it was initially learned. Lower F values indicate reduced catastrophic forgetting.

Relative Forgetting Measure (RF): Addresses limitations in the standard forgetting measure by normalizing forgetting relative to initial performance. This metric penalizes methods that achieve low forgetting by compromising initial learning capability, providing a more balanced assessment of the stability-plasticity trade-off in continual learning.

Performance Evaluation

We evaluate our method by comparing it against previous state-of-the-art approaches, with performance on the VQA and IC tasks reported in Table 1 and Table 2 , respectively. The results show that our DCIM framework consistently leads on the most critical metric, Average Performance (AP), while also surpassing competitors on the majority of other metrics. To ensure a fair comparison with methods that leverage knowledge distillation, we also present DCIM (KD). This enhanced version further extends our performance lead, achieving a comprehensive state-of-the-art result across all evaluated metrics. Furthermore, Figure 3, which plots the accuracy trajectory during the continual learning process, visually confirms the consistent and superior performance of DCIM over time.

Method	Memory	Speed
DER	22.4	2.98
TAM-CL	20.6	2.80
VQACL	13.4	1.83
ER	11.6	1.62
DCIM	12.8	1.78
DCIM (KD)	21.9	2.86

Table 3: Training resource consumption. Comparison of memory usage (GB) and speed (s/iter, batch=128) across different methods.

Model		Module			Metric		
		DMA	FSS	SRM	AP(%)↑	F(%)↓	RF(%)↓
Number	1	×	×	×	42.54	6.39	12.97
	2	✓	×	×	43.32	6.84	13.48
	3	×	✓	×	43.36	5.89	13.65
	4	×	×	✓	44.41	**5.64**	**11.62**
	5	✓	✓	×	43.61	5.91	11.81
	6	✓	×	✓	44.61	6.11	13.87
	7	×	✓	✓	44.70	5.99	13.46
	8	✓	✓	✓	**44.81**	5.94	11.78

Table 4: Ablation studies. This table examines the effectiveness of each component in our dual-memory approach. We evaluate the impact of Dynamic Memory Allocation (DMA), Forgetting Simulation Strategy (FSS), and Scene Reconstruction Memory (SRM) on model performance.

Table 3 presents the resource consumption during training, highlighting that the performance gains from knowledge distillation often come at a significant resource cost. Our standard DCIM method not only achieves superior results compared to KD-based methods like DER and TAM-CL but does so while remaining lightweight and efficient. This demonstrates our framework's strong balance of performance and efficiency, with the flexibility to use knowledge distillation for peak results.

Ablation Study

To evaluate the individual and combined contributions of our proposed components, we conducted a comprehensive ablation study. The analysis revealed that each component—Dynamic Memory Allocation (DMA), Forgetting Simulation Strategy (FSS), and Scene Reconstruction (SR)—provides a meaningful performance enhancement over the baseline. Most notably, the SR module delivered the single largest improvement in average performance (+1.87%) and the most significant reduction in forgetting (-0.75%), highlighting the critical impact of scene-level knowledge transfer. Furthermore, the study confirmed strong synergistic effects, with combinations of components outperforming the sum of their individual contributions. The pairing of FSS and SR was particularly effective, demonstrating that our sample selection strategy and scene-reconstruction mechanism are highly complementary. These results underscore that while all components are beneficial, the scene-level knowledge transfer enabled by SR is the most crucial element for mitigating catastrophic forgetting in our VQA-CIL framework.

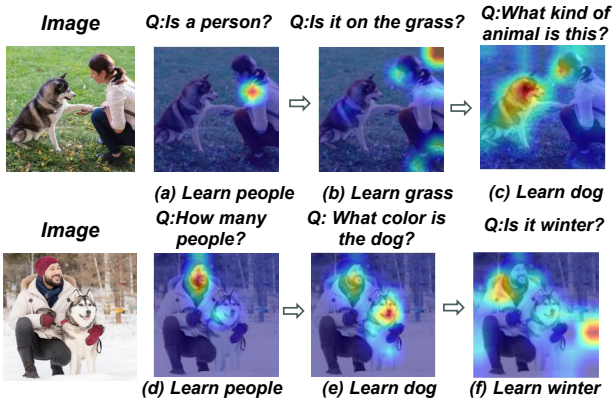

Figure 4: Attention visualization during class-incremental learning.

Qualitative Analysis

Figure 4 preesents GradCAM (Selvaraju et al. 2017) attention visualizations that illustrate how our model's attention mechanisms evolve during incremental learning. The visualizations reveal two key behaviors that confirm the effectiveness of our approach. First, the model retains and distributes its attention across previously learned concepts when a new class is introduced within the same scene, rather than completely shifting its focus (Figure 4, a-c). Second, and more importantly, the model actively leverages contextual features from known classes to learn new ones. This is clearly demonstrated in the bottom row (d-f), where the model utilizes features from a person's clothing (a down jacket) to help identify the concept of "winter." This observed cross-task knowledge transfer empirically validates our approach's ability to facilitate learning through shared visual contexts. By establishing meaningful connections between related concepts instead of learning them in isolation, our model mitigates catastrophic forgetting and enhances its performance on new tasks. The attention patterns particularly support our hypothesis that the Scene Reconstruction mechanism provides an effective bridge for knowledge transfer across incrementally learned tasks.

Conclusions

In this paper, We presented DCIM, a novel framework for Vision-Language Incremental Learning that addresses the challenge of learning new object classes within shared visual contexts. Through Hierarchical Class Memory Management (HCMM) for optimized buffer utilization and Scene Reconstruction Memory (SRM) for cross-class knowledge transfer, our approach significantly outperforms existing methods on VQA and IC benchmarks. The strong performance demonstrates that leveraging scene-level contextual similarities is crucial for multimodal continual learning, providing a promising direction for real-world applications where models must continuously adapt to emerging object categories.

Acknowledgments

This work was supported by the Fujian Provincial Department of Education Youth Project, China (grant JZ230006) , the Funding of Fuzhou University for Scientific Research (grant XRC-23119) , the Engineering Research Center of Big Data Intelligence, Ministry of Education , China , and the Fujian Key Laboratory of Network Computing and Intelligent Information Processing (Fuzhou University).

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077–6086.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

Buzzega, P.; Boschini, M.; Porrello, A.; Abati, D.; and Calderara, S. 2020. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33: 15920–15930.

Cai, Y.; Thomason, J.; and Rostami, M. 2023. Task-attentive transformer architecture for continual learning of vision-and-language tasks using knowledge distillation. *arXiv preprint arXiv:2303.14423*.

Chaudhry, A.; Dokania, P. K.; Ajanthan, T.; and Torr, P. H. 2018a. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European conference on computer vision (ECCV)*, 532–547.

Chaudhry, A.; Ranzato, M.; Rohrbach, M.; and Elhoseiny, M. 2018b. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*.

Chaudhry, A.; Rohrbach, M.; Elhoseiny, M.; Ajanthan, T.; Dokania, P.; Torr, P.; and Ranzato, M. 2019. Continual learning with tiny episodic memories. In *Workshop on Multi-Task and Lifelong Reinforcement Learning*.

Chen, F.; Ji, R.; Ji, J.; Sun, X.; Zhang, B.; Ge, X.; Wu, Y.; Huang, F.; and Wang, Y. 2019. Variational structured semantic inference for diverse image captioning. *Advances in Neural Information Processing Systems*, 32.

Chen, F.; Ji, R.; Su, J.; Wu, Y.; and Wu, Y. 2017. Structcap: Structured semantic embedding for image captioning. In *Proceedings of the 25th ACM international conference on Multimedia*, 46–54.

Chen, F.; Ji, R.; Sun, X.; Wu, Y.; and Su, J. 2018. Groupcap: Group-based image captioning with structured relevance and diversity constraints. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1345–1353.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Cho, J.; Lei, J.; Tan, H.; and Bansal, M. 2021. Unifying Vision-and-Language Tasks via Text Generation. In *ICML*.

Das, D.; Talon, D.; Mancini, M.; Wang, Y.; and Ricci, E. 2025. One vlm to keep it learning: Generation and balancing for data-free continual visual question answering. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 5635–5645. IEEE.

De Lange, M.; Aljundi, R.; Masana, M.; Parisot, S.; Jia, X.; Leonardis, A.; Slabaugh, G.; and Tuytelaars, T. 2021. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7): 3366–3385.

Del Chiaro, R.; Twardowski, B.; Bagdanov, A.; and Van de Weijer, J. 2020. Ratt: Recurrent attention to transient tasks for continual image captioning. *Advances in Neural Information Processing Systems*, 33: 16736–16748.

Gao, R.; and Liu, W. 2023. Ddgr: Continual learning with deep diffusion-based generative replay. In *International Conference on Machine Learning*, 10744–10763. PMLR.

Ge, X.; Xu, S.; Chen, F.; Wang, J.; Wang, G.; An, S.; and Jose, J. M. 2024. 3SHNet: Boosting image–sentence retrieval via visual semantic–spatial self-highlighting. *Information Processing & Management*, 61(4): 103716.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. 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*, 6904–6913.

He, Y.; Fu, J.; Zheng, K.; Xu, S.; Chen, F.; Li, J.; Jose, J. M.; and Ge, X. 2025. Double-Filter: Efficient Fine-tuning of Pre-trained Vision-Language Models via Patch&Layer Filtering. In *Forty-second International Conference on Machine Learning*.

Huang, Z.; Zeng, Z.; Huang, Y.; Liu, B.; Fu, D.; and Fu, J. 2021. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12976–12985.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123: 32–73.

Lao, M.; Pu, N.; Liu, Y.; Zhong, Z.; Bakker, E. M.; Sebe, N.; and Lew, M. S. 2023. Multi-domain lifelong visual question answering via self-critical distillation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 4747–4758.

Lei, S. W.; Gao, D.; Wu, J. Z.; Wang, Y.; Liu, W.; Zhang, M.; and Shou, M. Z. 2023. Symbolic replay: Scene graph as prompt for continual learning on vqa task. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1250–1259.

Lesort, T.; Lomonaco, V.; Stoian, A.; Maltoni, D.; Filliat, D.; and Díaz-Rodríguez, N. 2020. Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. *Information fusion*, 58: 52–68.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, 121–137. Springer.

Lopez-Paz, D.; and Ranzato, M. 2017. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29.

Luo, Y.; Zhao, S.; Wu, H.; and Lu, Z. 2024. Dual-enhanced coreset selection with class-wise collaboration for online blurry class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23995–24004.

Masana, M.; Liu, X.; Twardowski, B.; Menta, M.; Bagdanov, A. D.; and Van De Weijer, J. 2022. Class-incremental learning: survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5): 5513–5533.

Rebuffi, S.-A.; Kolesnikov, A.; Sperl, G.; and Lampert, C. H. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2001–2010.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Teney, D.; Liu, L.; and van Den Hengel, A. 2017. Graph-structured representations for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.

Vitter, J. S. 1985. Random sampling with a reservoir. *ACM Transactions on Mathematical Software (TOMS)*, 11(1): 37–57.

Xiang, Y.; Fu, Y.; Ji, P.; and Huang, H. 2019. Incremental learning using conditional adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6619–6628.

Yan, S.; Hong, L.; Xu, H.; Han, J.; Tuytelaars, T.; Li, Z.; and He, X. 2022. Generative negative text replay for continual vision-language pretraining. In *European Conference on Computer Vision*, 22–38. Springer.

Ye, F.; and Bors, A. G. 2020. Learning latent representations across multiple data domains using lifelong VAEGAN. In *European Conference on Computer Vision*, 777–795. Springer.

Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2024. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8): 5625–5644.

Zhang, X.; Zhang, F.; and Xu, C. 2023. Vqacl: A novel visual question answering continual learning setting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19102–19112.

Zheng, B.; Zhou, D.-W.; Ye, H.-J.; and Zhan, D.-C. 2024. Multi-layer rehearsal feature augmentation for class-incremental learning. In *Forty-first International Conference on Machine Learning*.

Zhou, D.-W.; Wang, Q.-W.; Qi, Z.-H.; Ye, H.-J.; Zhan, D.-C.; and Liu, Z. 2024. Class-incremental learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.