SCoT: Similarity-guided, Conflict-aware Task Consolidation for Continual VQA

Anand Patel[1], Moloud Abdar[2], and Biplab Banerjee[1]

[1] Indian Institute of Technology Bombay, India
[2] The University of Queensland, Australia
{anandp2307, m.abdar1987, getbiplab}@gmail.com

Abstract. Continual learning in visual question answering (VQACL) requires a single vision–language model to acquire new multimodal reasoning skills from a task stream while retaining prior capabilities. However, naïve sequential finetuning suffers from catastrophic forgetting. Existing continual VQA methods primarily rely on replay or parameter regularization but largely overlook how task-specific updates accumulate and interact in parameter space, particularly whether successive updates are synergistic or conflicting across layers. To address this, we introduce **SCoT** (**S**imilarity-guided **Co**nflict-aware **T**ask Consolidation), a continual learning framework that represents each task as a parameter update relative to a pretrained anchor model and integrates tasks through layer-wise parameter-space reasoning. For each layer, SCoT measures alignment between incoming and accumulated task vectors, removes only destructive components via conditional projection when conflicts arise, and adaptively modulates consolidation strength using similarity-guided weighting. This preserves beneficial transfer while suppressing harmful interference, enabling stable yet adaptive continual learning. Experiments on VQAv2 and NExT-QA demonstrate strong continual VQA performance, reducing forgetting to near-zero (0.07 and -1.90) while achieving rare positive backward transfer (+5.64 and +6.97), outperforming strong continual-learning and task-vector baselines. Project page: https://anand-patel05.github.io/SCoT

Keywords: Continual Learning · Visual Question Answering · Multimodal Reasoning · Catastrophic Forgetting · Positive Backward Transfer

1 Introduction

Modern vision–language models (VLMs) deliver strong visual question answering (VQA) by learning rich multimodal representations and compositional reasoning from large-scale pretraining [4, 11, 26, 31]. However, in deployment, these models are rarely trained once and frozen [56]. They must continually acquire new capabilities, particularly new question types, reasoning patterns, domains, or datasets over time while remaining reliable on earlier skills. Enabling such continual VQA remains a fundamental challenge because naive sequential fine-tuning rapidly induces catastrophic forgetting [39], disrupts cross-modal alignment [75, 76], and can erode the general-purpose reasoning inherited from pretraining [33].

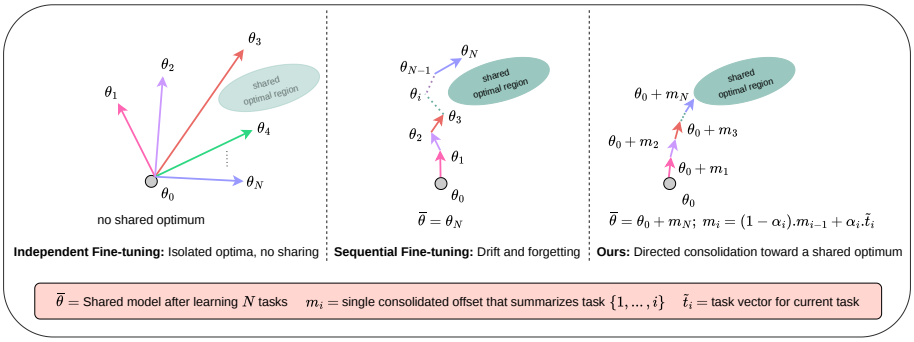


Fig. 1: Independent fine-tuning yields task-specific optima but requires storing separate model parameters for each task, leading to high memory overhead. Sequential fine-tuning suffers catastrophic forgetting due to task interference. Our method steers learning in parameter space toward a shared optimum under continual learning constraints, by geometrically consolidating task directions into a single shared model.

To address this challenge, existing continual VQA methods [2, 5, 9, 24, 37, 58, 78] largely adapt techniques from unimodal continual learning. They rely on replay buffers (small subset of data from past training), regularization, or architectural heuristics that constrain training dynamics in activation or loss space [6, 47, 60, 63, 65]. While effective to some extent, these approaches treat each task update as an opaque optimization step and provide limited control over how task knowledge accumulates and interact in the shared parameter space. As a result, they struggle to exploit positive transfer between related task and often trade plasticity for stability in a brittle manner. This motivates a key question: Can we go beyond preventing forgetting to actively enable beneficial knowledge transfer by explicitly modeling whether successive task updates are synergistic or conflicting in parameter space?

A natural route to explicit control over continual updates is to operate directly in parameter space using task vectors [21], where each task is represented as a displacement from a pretrained anchor and continual learning becomes consolidation of these displacements over time. This view has motivated a range of model-merging approaches that combine task-specific updates through weight averaging [22, 62], Fisher-weighted merging [38], continual model merging [54, 68], or sign- and magnitude-based heuristics [66, 71], enabling exemplar-free knowledge integration. However, most such methods target offline merging of independently trained models and assume access to a pool of completed task models; thereby escaping the harder continual-learning regime, where tasks arrive sequentially with data and must be learned, integrated, and deployed on the fly.

More recent approaches adapt task-vector consolidation to continual settings. MagMax [36] retains the maximum-magnitude parameter update per element, while DuET [40] introduces layer-wise magnitude-driven mixing with directional consistency regularization. However, both remain magnitude-centric and do not explic-

itly model alignment between successive task updates–failing to resolve directional conflicts, where incoming updates partially negate consolidated knowledge, and behaving conservatively in ways that limit both forward and backward transfer. This motivates consolidation strategies that reason explicitly about task update geometry to preserve synergistic components while suppressing destructive interference.

We propose to reinterpret continual multimodal learning as geometric consolidation in parameter space (Fig. 1) and introduce SCoT, a continual learning framework for multimodal transformers. SCoT explicitly models how an incoming task vector interacts with accumulated knowledge per layer. It first measures directional alignment between incoming and consolidated task vectors; when conflicts arise, it removes only the destructive component via conditional geometric projection, rather than employing uniform shrinking or magnitude-based mixing. It then applies a similarity-guided weighting scheme that modulates the consolidation strength, integrating compatible tasks more aggressively while conservatively merging incompatible ones. This similarity-guided, conflict-aware layer-wise consolidation respects the hierarchical structure of transformers, enabling individual layers to balance stability and plasticity independently.

This formulation enables a fundamentally different continual learning behavior. Instead of competing for shared parameters, tasks can reinforce one another when aligned and be selectively decoupled when conflicting. As a result, learning becomes constructive rather than competitive, enabling the continual model to refine shared representations over time. Our approach, therefore, not only reduces forgetting to near-zero (Tab. 1), but also achieves positive backward transfer–a rare regime in continual learning (Fig. 4). Remarkably, on VQAv2, SCoT surpasses even joint multi-task training (Fig. 2), demonstrating that continual learning with proper consolidation can exceed the traditional upper bound for continual learning.

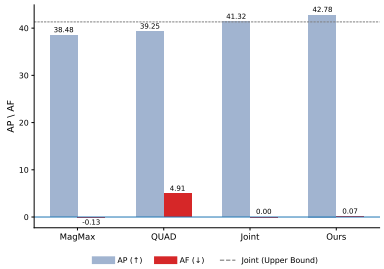


Fig. 2: Surpassing the Upper Bound. On VQAv2, SCoT effectively eliminates forgetting (AF ≈ 0) and, remarkably, outperforms **Joint Training**—the traditional upper bound for continual learning.

Our **major contributions** are as follows:

1. We introduce a task-vector perspective for continual learning that anchors continual adaptation around a pretrained multimodal backbone.
2. We propose SCoT, a continual learning framework that explicitly controls how task updates accumulate and interact in parameter space.
3. We demonstrate a consistently superior stability–plasticity trade-off across diverse benchmarks, including VQAv2 and NExT-QA, achieving near-zero forgetting (0.07 and -1.90) and rare positive backward transfer (+5.64 and +6.97).

2 Related Work

2.1 Visual Question Answering. VQA [3] requires jointly reasoning over images and text to answer natural-language queries. Early systems relied on attention-based multimodal fusion [69, 72, 76], while modern VLMs unify visual and linguistic representations within pretrained architectures [11], with generative VQA treating answering as open-vocabulary text generation. A long-standing challenge is compositional generalization–the systematic recombination of known concepts [20, 23, 29] which prior work addresses via factorized representations and disentangled objectives [10, 77] but primarily under offline, fixed-dataset assumptions that do not extend to non-stationary task streams.

2.2 Continual Learning. CL [56] aims to incrementally acquire knowledge from a task sequence while preserving prior capabilities, but suffers from catastrophic forgetting [39]. Classic approaches include regularization [2, 24, 43, 51], rehearsal [5, 8, 9, 55, 57], and architectural heuristic [27, 49, 70], with more recent variants emphasizing gradient-interference control [8, 34, 50, 67], robust representations [12, 14–16], and prompt-based adaptation for pretrained transformers [52, 58, 59, 61]. Yet most CL methods target unimodal class-incremental learning; in multimodal or generative regimes, they often struggle to preserve cross-modal alignment and compositional structure, essential for VQA.

2.3 Continual VQA. Continual VQA extends CL to multimodal reasoning [41, 53, 78], where both visual and linguistic distributions drift over time. VQACL [78] formalized a generative continual VQA benchmark and showed that standard CL techniques [2, 5, 9, 24, 58] are insufficient, proposing rehearsal with dual-level task organization and prototype memory. QUAD [37] improved efficiency and privacy via question-only replay and attention distillation. Despite progress, these approaches mainly operate in activations or gradient space and treat each task update as an opaque outcome of optimization, offering limited control over how task knowledge accumulates or conflicts in parameter space.

2.4 Task Vectors and Model Merging. Task vectors [21] encode task knowledge as weight displacements from a pretrained anchor, enabling parameter arithmetic for composition and multi-task adaptation [38, 62, 66]. However, most merging approaches, including continual variants, operate offline, consolidating pools of independently trained models post-hoc [13, 54, 68], bypassing the harder continual regime where tasks must be learned and integrated on the fly. MagMax [36] and DuET [40] are closest to this regime, but remain magnitude-driven–ignoring geometric relations between task directions and behaving conservatively in ways that limit both forward and backward transfer. Bridging the gap between offline model merging and online continual learning, while preserving the geometric structure of parameter space, remains a largely unexplored direction.

3 VQA Continual Learning Setting

We study continual learning for generative VQA, where a model parameterized by θ defines a conditional distribution over answer sequences $Y = (Y_1, \ldots, Y_{|Y|})$ given

an image–question pair (V, Q):

$$P_\theta(Y \mid V, Q) = \prod_{k=1}^{|Y|} P_\theta(Y_k \mid Y_{<k}, V, Q)\,. \tag{1}$$

Following VQACL [78], we adopt [11] and train it using the standard autoregressive cross-entropy objective. Unlike offline training, data arrive sequentially as a stream of N tasks $\{T_i\}_{i=1}^{N}$, where each task T_i provides triplets $(V, Q, Y) \sim \mathcal{D}_i$. At step i, the learner updates θ on T_i with a small replay buffer and is evaluated on all tasks $\{T_1, \ldots, T_i\}$.

This protocol organizes the task stream hierarchically: macro-tasks correspond to distinct linguistic reasoning skills (e.g., recognition, color, counting), and each macro-task is further split into a sequence of visually driven sub-tasks by partitioning object categories into disjoint groups, inducing coupled linguistic and visual shifts. This setting poses two coupled challenges: (i) retention–prevent catastrophic forgetting of previously acquired visual grounding and language reasoning under sequential updates; and (ii) compositional generalization–transfer a learned skill to novel visual contexts (e.g., applying counting to unseen object categories), requiring the model to preserve and recombine reasoning primitives across tasks.

4 Methodology

Continual learning requires integrating new task updates without corrupting previously acquired knowledge. To explicitly control how such updates accumulate, we propose SCoT, a parameter-space consolidation framework built upon task vectors [21]. Instead of relying on implicit regularization or gradient-based constraints [24, 37, 78], SCoT treats each task update as an explicit displacement in weight space and consolidates these updates with controlled transfer and interference mitigation.

Let θ_0 denote a pretrained multimodal backbone. For each task T_i in the sequential stream (§3), we fine-tune the current shared model on T_i, extract the task-induced displacement relative to θ_0, and integrate it into a single consolidated offset that initializes the subsequent task. Fig. 3 illustrates this continual geometric consolidation process.

SCoT consists of four key components. First, we represent each task as an anchor-relative task vector, which decouples task-specific learning from consolidation and enables explicit geometric reasoning in parameter space (§4.1). Second, we regulate the stability–plasticity trade-off by mapping task similarity to bounded consolidation weights, promoting aggressive merging for related tasks and conservative integration otherwise (§4.2). Third, we mitigate destructive interference by measuring directional alignment between incoming and consolidated updates, isolating and removing only conflicting components via conditional projection (§4.3). Finally, recognizing that compositional sensitivity is non-uniform across layers, we introduce depth-aware stabilization that caps consolidation strength in empirically identified composition-sensitive layers (§4.4).

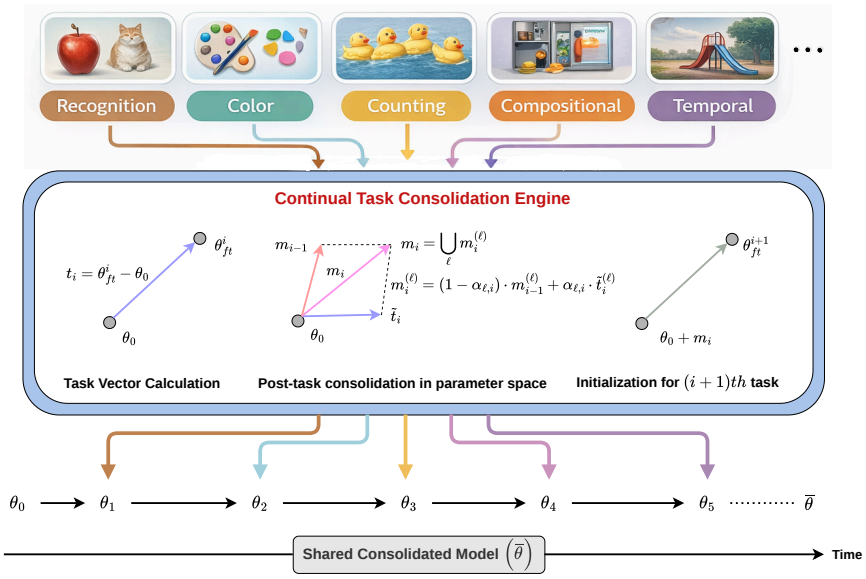


Fig. 3: SCoT performs continual task consolidation directly in parameter space. After fine-tuning on task T_i, a task vector t_i is computed relative to the initial parameters θ_0. Task directions are integrated into a single consolidated offset m_i using conflict-aware, layer-wise consolidation to mitigate task interference. The shared model $\bar{\theta}_i = \theta_0 + m_i$ aggregates knowledge from tasks $\{1, \ldots, i\}$ and initializes learning for the next task.

4.1 Task Vectors as the Continual State

Task vectors [21] have recently emerged as a principled way to represent task-specific knowledge as parameter displacements, enabling downstream merging and composition [62, 66]. We reframe this perspective for the continual setting by maintaining a cumulative consolidated offset over sequentially arriving tasks.

Let θ_0 be a pretrained backbone. For each task T_i in the stream, we initialize the model with the previous consolidated parameters $\bar{\theta}_{i-1}$ and fine-tune to obtain θ_i^{ft}. We represent the resulting knowledge as a task vector:

$$t_i \;=\; \theta_i^{\text{ft}} - \theta_0, \tag{2}$$

which captures the cumulative displacement from a fixed, pretrained anchor θ_0 for all tasks in the stream. In the continual regime, we maintain a single cumulative consolidated offset m_i that summarizes tasks $\{1, \ldots, i\}$ and recover the shared model as:

$$\bar{\theta}_i \;=\; \theta_0 + m_i. \tag{3}$$

This anchor-relative view decouples what the task changes (the vector t_i) from how tasks are accumulated (the update of m_i), making the task interaction amenable

to explicit geometric control in weight space. We formalize this geometric control in the following subsections.

4.2 Similarity-guided Layer-wise Consolidation

The stability–plasticity dilemma [18] lies at the heart of continual learning–a model must be plastic enough to absorb new knowledge yet stable enough to retain what it has already learned. Classical regularization methods such as EWC [24] and Synaptic Intelligence [74] address this by assigning per-parameter importance weights that penalize changes to consolidated knowledge. In the merging literature, Fisher-weighted averaging [38] similarly scales contributions by parameter sensitivity. However, computing and maintaining such importance estimates requires second-order statistics that are prohibitively expensive to recompute as the task stream scales. Furthermore, existing merging methods often treat all layers uniformly, ignoring evidence that task-specific information is distributed non-uniformly across the depth of multimodal transformers [42, 46].

Motivated by this, we derive a natural importance signal at the granularity of individual layers directly from the first-order geometry of the parameter space. For each layer ℓ, we compute the cosine alignment $s_{\ell,i}$ between the incoming task vector t_i^ℓ and the consolidated state m_{i-1}^ℓ:

$$s_{\ell,i} \;=\; \frac{\langle t_i^{(\ell)},\, m_{i-1}^{(\ell)} \rangle}{\|t_i^{(\ell)}\|\,\|m_{i-1}^{(\ell)}\| + \epsilon}, \tag{4}$$

where ϵ is a small numerical stability constant. We map this alignment to a bounded layer-wise consolidation weight $[\alpha_{min}, \alpha_{max}]$, so that layers highly aligned with prior knowledge absorb new task more aggressively, while a weakly related layer should integrate it conservatively to protect the existing structure.

$$\alpha_{\ell,i} \;=\; \mathrm{clip}\left(\alpha_{\min} + (\alpha_{\max} - \alpha_{\min})\frac{s_{\ell,i}+1}{2}\right). \tag{5}$$

The consolidated offset is updated by treating the integration as a data-driven weighted interpolation in the cumulative displacement space:

$$m_i^{(\ell)} = (1-\alpha_{\ell,i})m_{i-1}^{(\ell)} + \alpha_{\ell,i}\boldsymbol{\Delta}_i^{(\ell)}, \tag{6}$$

where $\boldsymbol{\Delta}_i^{(\ell)} = t_i^{(\ell)}$ represents the current cumulative displacement. This formulation ensures that $m_i^{(\ell)}$ moves toward the current task-optimized state without double-counting prior knowledge, as the interpolation effectively scales only the task-specific novelty relative to the previous state. However, similarity weighting alone does not prevent updates from directly opposing prior knowledge.

4.3 Conflict-aware Layer-wise Projection

Catastrophic forgetting in neural networks is fundamentally a problem of destructive interference–when updates for a new task overwrite parameter directions that

encoded prior knowledge [24, 39]. Gradient projection methods [50, 73] have addressed this during training by constraining weight updates to lie in the null space of previous task gradients. However, such approaches require access to training dynamics and are incompatible with post-hoc task vectors. To resolve this, we use the alignment $s_{\ell,i}$ from Sec. 4.2 to detect and correct conflicts at the granularity of the individual layers.

From Eq. (4), when $s_{\ell,i} < 0$, the incoming update counteracts the consolidated direction at layer ℓ. Rather than shrinking the entire update, we apply a surgical conditional projection that removes only the conflicting component:

$$\tilde{t}_i^{(\ell)} = \begin{cases} t_i^{(\ell)} - \dfrac{\langle t_i^{(\ell)},\, m_{i-1}^{(\ell)} \rangle}{\|m_{i-1}^{(\ell)}\|^2 + \epsilon}\, m_{i-1}^{(\ell)}, & s_{\ell,i} < 0, \\ t_i^{(\ell)}, & \text{otherwise.} \end{cases} \tag{7}$$

This is the minimum-norm modification of $t_i^{(\ell)}$ that enforces $\langle \tilde{t}_i^{(\ell)}, m_{i-1}^{(\ell)} \rangle \geq 0$. By substituting $\mathbf{\Delta}_i^{(\ell)} = \tilde{t}_i^{(\ell)}$ in Eq. (6), SCoT preserves aligned and orthogonal components (which support transfer) while excising only the destructive direction responsible for forgetting.

4.4 Depth-aware Stabilization for Compositional Generalization

While similarity-guided, conflict-aware consolidation improves standard retention, we observe a larger forgetting gap in compositional evaluation (Tab. 3). Prior work has established that compositional reasoning in transformers is not uniformly distributed across layers, but the precise location of composition-sensitive layers varies with architecture. Liu [32] shows that causal language models concentrate compositional processing in early layers while masked models concentrate it in late layers, and Li et al. [28] similarly identify middle-layer MHSA modules as critical in decoder-only LLMs. This implies that no universal assumption about which layers are compositionally sensitive can be made–the location of these layers is an emergent property of the specific architecture and training regime.

In our encoder-decoder VLM setting, we identify these layers empirically through a systematic ablation of layer groups (Tab. 4). Our analysis reveals that stabilizing late encoder and early decoder layers consistently yields the strongest compositional performance across both VQAv2 and NExT-QA, suggesting that in our encoder-decoder VLM, compositional structure is formed at the encoder-decoder interface, where high-level encoded representations are first composed into reasoning outputs. Motivated by this finding, we introduce a depth-aware stabilization that selectively caps the consolidation strength specifically at this interface. We retain the similarity-guided $\alpha_{\ell,i}$ from Sec. 4.2 but apply a depth-dependent cap α_ℓ^{cap}:

$$\alpha_{\ell,i} \leftarrow \min(\alpha_{\ell,i},\, \alpha_\ell^{\text{cap}}), \quad \ell \in \mathcal{L}_{\text{comp}}, \tag{8}$$

where $\mathcal{L}_{\text{comp}}$ denotes the late encoder and early decoder layers identified via Tab. 4. This refinement improves compositional AP and reduces compositional forgetting

(Tab. 3), with a minor trade-off on standard-task adaptation–a cost outweighed by consistent compositional gains across both benchmarks.

Viewed **geometrically**, SCoT performs projected proximal consolidation in task-vector space. Task integration is modeled as a layer-wise weighted interpolation between the consolidated state and the incoming task displacement, where the mixing weight ($\alpha_{\ell,i}$) is a monotone function of cosine similarity. This enables aggressive merging of aligned task directions while conservatively integrating weakly related ones. When conflicts arise, a least-perturbation projection removes only the destructive components. Finally, depth-aware stabilization anchors consolidation at composition-sensitive layers, providing targeted protection against compositional forgetting. A formal theoretical analysis of this geometric framework is provided in the Suppl. Material.

5 Experiments

Implementation Details. To ensure fair comparison, we follow the experimental protocol of [78] for feature extraction and training across datasets. For visual embeddings, we use a Faster R-CNN model [48] pretrained on the Visual Genome dataset [25], extracting 36 region-based object features per image for VQAv2. For videos in NExT-QA, clip-level motion features are extracted using an inflated 3D ResNeXt-101 network [19], with $n = 16$ regions per clip. These visual features are projected into the transformer space using a two-layer MLP with GELU activation. Our multimodal backbone is based on the T5 architecture [45], consisting of 12 encoder and 12 decoder blocks, each with 12 attention heads and an embedding dimension $d = 768$. Following [78], the memory buffer $\mathcal{M}$ is set to 5000 samples for VQAv2 and 500 for NExT-QA, reflecting dataset scale. Training is performed for 3 epochs per task with a batch size of 80 using the Adam optimizer [1] and an initial learning rate of 10^{-4}. For task-vector consolidation, we employ similarity-guided adaptive plasticity (Sec. 4.2) with $\alpha_{\min} = 0.08$ and $\alpha_{\max} = 0.45$ in all experiments. To preserve compositional structure during continual learning, we apply depth-aware stabilization (Tab. 4) by constraining updates in composition-critical layers, particularly late encoder layers, which are strongly stabilized using $\alpha_{\min}$, and early decoder layers use a moderate cap ($1.5 \times \alpha_{\min}$), while remaining layers are allowed full adaptation up to $\alpha_{\max}$. These settings were selected to balance adaptation performance with compositional stability. All implementations are developed in PyTorch [44].

Evaluation Metrics. We adopt two standard continual learning metrics [7, 8, 34], namely Final Average Performance (AP) and Average Forgetting (AF). The AP metric reflects overall performance across all learned tasks, indicating the model's ability to accumulate knowledge consistently. Let $a_{i,j}$ denote the performance on task T_i after completing training on task T_j. The AP is defined as $\text{AP} = \frac{1}{N}\sum_{i=1}^{N} a_{i,N}$. The Average Forgetting metric measures knowledge retention by quantifying performance degradation on earlier tasks as new tasks are learned, computed as $\text{AF} = \frac{1}{N-1}\sum_{i=1}^{N-1} \max_{i \le j \le N-1}(a_{i,j} - a_{i,N})$. To further analyze the stability–plasticity trade-off, we also report Backward Transfer (BWT), which captures how learning new

Table 1: Model performance on VQAv2 and NExT-QA. #Mem: memory size; Std. Test: standard testing; Comp. Test: novel composition testing; AP: Final Average Performance (%); AF: Average Forgetting (%).

Methods	VQAv2					NExT-QA				
	#Mem	Std. Test		Comp. Test		#Mem	Std. Test		Comp. Test	
		AP (↑)	AF (↓)	AP (↑)	AF (↓)		AP (↑)	AF (↓)	AP (↑)	AF (↓)
Joint	–	41.32	–	40.88	–	–	35.92	–	36.24	–
Vanilla	None	14.92	30.80	11.79	27.16	None	12.68	25.94	12.59	28.04
EWC [24]	None	15.77	30.62	12.83	28.16	None	13.01	24.06	11.91	27.44
MAS [2]	None	20.56	11.16	23.90	6.24	None	18.04	10.07	21.12	10.09
ER [9]	5000	36.99	5.99	33.78	5.76	500	30.55	4.91	32.20	5.57
DER [5]	5000	35.35	8.62	31.52	8.59	500	26.17	5.12	21.56	12.68
VS [58]	5000	34.03	8.79	32.96	5.78	500	28.13	4.45	29.47	6.14
VQACL [78]	5000	37.46	6.96	35.40	4.90	500	30.86	4.12	33.85	3.80
QUAD [37]	5000	39.25	4.91	40.00	3.81	500	31.70	2.91	33.21	4.16
Ours	5000	**42.78**	**0.07**	**42.11**	**2.14**	500	**32.28**	**-1.90**	**33.94**	**2.04**

tasks affects learned tasks (Fig. 4). It is defined as $\text{BWT} = \frac{1}{N-1}\sum_{i=1}^{N-1}(a_{i,N} - a_{i,i})$, where positive BWT indicates beneficial backward transfer (improved performance on earlier tasks after subsequent training), while negative BWT reflects forgetting. For NExT-QA [64], we compute $a_{i,j}$ using Wu-Palmer Similarity (WUPS) [35] to evaluate the quality of generated answers, whereas for VQAv2 [17], following [78], we use the percentage of correctly answered questions as $a_{i,j}$.

Baselines. We benchmark our method against a diverse set of established continual learning approaches. These include two regularization methods: EWC [24] and MAS [2], and three rehearsal-based approaches: ER [9], DER [5], and VS [58]. We also include VQACL [78] and QUAD [37], which are specifically designed for continual VQA and represent as strong baselines. In addition, we report two reference benchmarks to contextualize performance: a lower bound, Vanilla, obtained by naive sequential fine-tuning, and a multitask reference, Joint, obtained by simultaneous training on the full task set under identical experimental conditions (e.g., same backbone, optimizer, and total training epochs). Following [78], we employ two testing strategies: standard testing to evaluate knowledge retention, and novel composition testing to probe generalization to unseen combinations. This dual evaluation reveals how effectively each method balances stability and adaptation. Additional experiments and analysis are included in the Suppl. Material.

Performance Analysis on Standard Testing. As shown in Tab. 1, SCoT achieves superior accuracy and near-zero forgetting across both benchmarks. Remarkably, on VQAv2, SCoT (AP = 42.78) surpasses even Joint Training (AP = 41.32)—the traditional upper bound. This suggests that sequential geometric consolidation avoids the simultaneous gradient interference inherent in joint training, reaching a more synergistic global optimum. Similarly, on NExT-QA, SCoT achieves AP = 32.28 with negative forgetting (AF = -1.90), indicating significant positive backward transfer. These results demonstrate that structure-preserving consolidation effectively

Table 2: Fine-grained VQA performance AP (%) on Novel and Seen skill-concept compositions of VQAv2 and NExT-QA.

Dataset	Method	Group-1		Group-2		Group-3		Group-4		Group-5		Avg	
		Novel	Seen	Novel	Seen	Novel	Seen	Novel	Seen	Novel	Seen	Novel	Seen
VQAv2	DER [5]	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 [58]	33.35	33.87	33.18	32.21	34.50	33.84	31.29	33.94	32.46	33.87	32.96	33.55
	ER [9]	34.52	37.03	33.40	35.55	34.79	34.20	33.86	35.02	32.34	35.91	33.78	35.54
	VQACL [78]	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 [37]	39.19	41.06	38.40	39.50	43.15	39.19	40.01	40.72	39.20	40.62	40.00	40.21
	Ours	41.29	43.57	42.20	43.69	42.71	43.42	43.21	42.94	41.14	44.43	**42.11**	**43.61**
NExT-QA	DER [5]	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 [58]	31.42	30.88	29.17	31.26	25.23	26.10	30.01	29.10	31.54	31.79	29.47	29.83
	ER [9]	31.86	34.51	32.36	35.08	29.50	34.30	33.57	33.30	33.71	32.91	32.20	34.02
	VQACL [78]	32.86	31.47	31.98	35.58	31.79	35.70	35.04	34.12	37.62	34.92	33.85	34.35
	QUAD [37]	33.42	33.92	32.02	35.42	31.78	36.36	32.98	33.34	37.84	34.06	33.21	**34.62**
	Ours	33.20	34.91	32.81	34.50	34.36	34.88	33.84	33.17	35.50	33.73	**33.94**	34.24

resolves the stability–plasticity dilemma, enabling shared representations to refine rather than degrade over time (Fig. 4).

Performance Analysis of Novel Composition Testing. Results on the compositional evaluation (Tab. 1 and Tab. 2) further highlight the effectiveness of our approach in maintaining compositional generalization under continual learning. On VQAv2, our method achieves the highest performance (AP = 42.11) with moderate forgetting, indicating that compositional knowledge is largely preserved despite sequential task updates. On NExT-QA, it remains competitive (AP = 33.94) while maintaining controlled forgetting, reflecting robust transfer to unseen skill–concept combinations. The fine-grained results in Tab. 2 show consistent improvements across most compositional groups, particularly on novel combinations, suggesting that depth-aware stabilization (Tab. 4) helps preserve intermediate structural representations critical for compositional reasoning. Overall, these results confirm that the proposed consolidation strategy supports both continual knowledge retention and compositional generalization.

6 Ablation Study and Analysis

Effect of each component Tab. 3 presents a progressive ablation of SCoT's four components. C_1 alone–uniform task vector averaging, achieves moderate AP (39.23/27.35) with negative forgetting, but the strongly negative Comp AF on NExT-QA (-5.67) reveals over-conservative integration that suppresses adaptation on its more diverse task distribution. Adding C_2 (similarity-guided consolidation) yields the largest single gain–standard AP improves by +3.39 on VQAv2 and +4.80 on NExT-QA, confirming that adaptive consolidation strength is the most critical factor for stability–plasticity balance. However, increased plasticity from C_2 raises forgetting (AF: 0.35 → 1.42 on VQAv2), motivating C_3. Adding C_3 (conflict-aware projection) directly addresses this–AF drops substantially (1.42 → 0.92 on VQAv2, 1.79 → 0.49 on NExT-QA) while maintaining strong AP, suggesting that

Table 3: Ablation study of model components. Std: standard testing, Comp: novel compositional testing. Components: C_1 Task Vector as the Continual State; C_2 Similarity-guided Layer-wise Consolidation; C_3 Conflict-aware Layer-wise Projection; C_4 Depth-aware Stabilization for Compositional Generalization.

C_1	C_2	C_3	C_4	VQAv2				NExT-QA			
				Std		Comp		Std		Comp	
				AP(↑)	AF(↓)	AP(↑)	AF(↓)	AP(↑)	AF(↓)	AP(↑)	AF(↓)
✓	✗	✗	✗	39.23	-0.35	38.51	-0.28	27.35	-1.79	28.44	-5.67
✓	✓	✗	✗	42.62	1.42	41.08	2.99	32.15	1.79	34.00	3.33
✓	✓	✓	✗	43.02	0.92	41.31	2.95	32.39	0.49	33.75	3.07
✓	✓	✓	✓	42.78	0.07	42.11	2.14	32.28	-1.90	33.94	2.04

explicit interference control provides additional robustness against the forgetting introduced by stronger consolidation. The marginal Comp AP dip on NExT-QA (34.00 → 33.75) reflects occasional over-projection of beneficial components, which C_4 corrects. Finally, C_4 (depth-aware stabilization) delivers the best compositional performance (42.11/33.94) with near-zero forgetting (0.07/-1.90). The minor standard AP trade-off (43.02 → 42.78) reflects a deliberate stability bias at composition-sensitive layers–a cost outweighed by consistent compositional gains and dramatically reduced forgetting across both benchmarks.

Analysis of Depth-aware Stabilization. Table 4 ablates encoder–decoder layer group combinations under standard (Std) and compositional (Comp) settings. Two trends emerge. First, encoder depth is the dominant factor–upgrading from early to late encoder layers yields consistent gains in both standard and compositional AP across VQAv2 (40.38 → 42.11) and NExT-QA (29.34 → 33.94), consistent with prior findings that compositionally sensitive layers vary by architecture [28, 32]. Second, early decoder stabilization best complements late encoder stabilization *i.e.* the late+early combination achieves the strongest compositional performance with competitive forgetting (AF(↓) 2.14/2.04), identifying the encoder–decoder interface as the most composition-sensitive region. Deeper decoder stabilization offers diminishing returns and risks reduced plasticity. We therefore adopt late encoder+early decoder as $\mathcal{L}_{\text{comp}}$ (Sec. 4.4).

Analysing Stability/Plasticity. Fig. 4 compares forgetting and adaptation on VQAv2 using transfer matrices and BWT scores. Sequential fine-tuning suffers severe forgetting ($BWT = -29.44$), with off-diagonal entries collapsing after each new task. QUAD reduces this effect ($BWT = -4.76$), but its sparse off-diagonal structure indicates limited backward transfer. In contrast, our method achieves rare positive backward transfer ($BWT = +5.64$) [30, 34], showing that conflict-aware, geometry-driven consolidation improves stability without sacrificing plasticity. Additional NExT-QA analysis is provided in the supplementary material, where SCoT also achieves positive BWT (+6.97).

Sensitivity to Memory Size. Fig. 5a evaluates performance across varying replay buffer sizes on VQAv2 and NExT-QA. Our method consistently outperforms all baselines across all memory budgets, with the gap widening as memory increases. On VQAv2, our AP grows monotonically from 1K to 5K while DER [5] and VS [58]

Table 4: Depth-aware stabilization ablation across encoder–decoder layer groups (early: 0–3, mid: 4–7, late: 8–11) on VQAv2 and NExT-QA. Late-encoder+early-decoder stabilization achieves the strongest overall performance across datasets, indicating that stabilization near the encoder–decoder interface is particularly effective.

Encoder	Decoder	VQAv2				NExT-QA			
		Std		Comp		Std		Comp	
		AP(↑)	AF(↓)	AP(↑)	AF(↓)	AP(↑)	AF(↓)	AP(↑)	AF(↓)
early	late	38.50	-0.52	40.52	0.74	24.35	1.03	28.30	1.79
early	mid	38.40	-0.72	39.90	0.45	24.87	1.18	28.77	0.70
early	early	38.45	-0.29	40.38	0.92	24.54	1.21	29.34	-0.21
mid	late	42.60	0.35	42.42	2.13	31.86	-0.97	33.82	1.62
mid	mid	42.70	0.40	41.97	2.42	32.09	-0.95	33.12	2.74
mid	early	42.52	0.77	41.64	2.47	32.39	-1.59	33.50	2.31
late	late	42.12	0.56	41.88	1.76	32.39	-1.11	33.56	2.57
late	mid	42.33	0.92	42.05	2.04	31.87	-1.32	33.60	2.73
late	early	42.78	0.07	42.11	2.14	32.28	-1.90	33.94	2.04

plateau or degrade, suggesting they fail to exploit additional replay capacity. On NExT-QA, our method similarly maintains stable improvement while others exhibit saturation or fluctuation. These results confirm that our consolidation mechanism scales gracefully with memory, translating increased replay capacity into consistent performance gains.

Impact of Hyperparameters. Tab. 5 studies sensitivity to the consolidation weight range (Sec. 4.2). The narrowest range $[0.05, 0.30]$ leads to under-adaptation, yielding the lowest AP (41.22/29.09) and negative forgetting due to excessive stability. The range $[0.08, 0.45]$ provides the best stability–plasticity trade-off, achieving the highest AP (42.78/32.28) with near-zero forgetting (0.07/-1.90). Increasing the range to $[0.15, 0.60]$ maintains strong AP (42.99/32.06) but introduces noticeable forgetting (1.07/1.68), indicating greater plasticity. The widest range $[0.30, 0.80]$ further increases forgetting (2.10/2.59) despite reasonable AP. Overall, α serves as a continuous stability–plasticity controller, with $[0.08, 0.45]$ yielding the best balance.

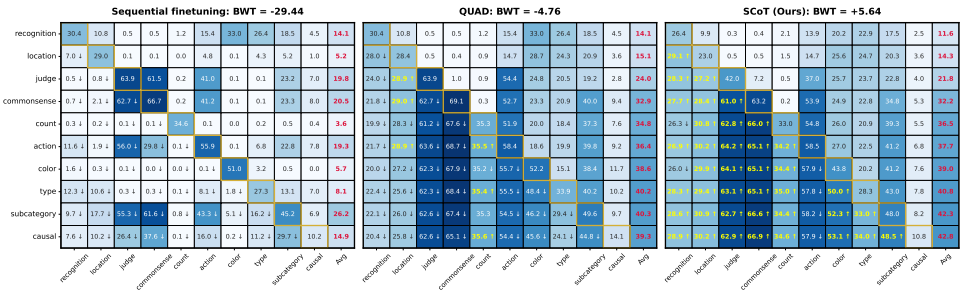
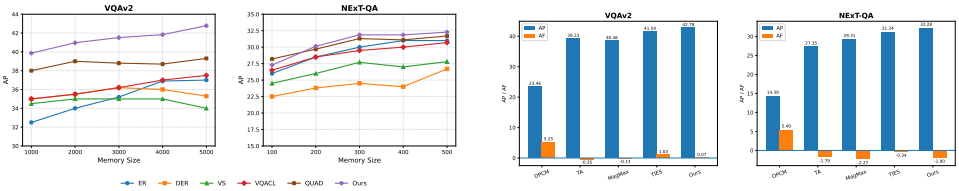


Fig. 4: Stability–Plasticity analysis on VQAv2. Performance across training tasks (rows) and evaluation tasks (columns) illustrates a sharp contrast in learning behavior. While baselines exhibit catastrophic forgetting (blue off-diagonals) as knowledge degrades under sequential updates, SCoT demonstrates constructive learning.



(a) Sensitivity analysis to memory size.

(b) Comparison with model-merging methods.

Fig. 5: Ablation and comparative evaluation. Left: Performance sensitivity to memory size. Right: Comparison against representative model-merging baselines.

Table 5: Sensitivity to consolidation weight range $[\alpha_{min}, \alpha_{max}]$

α Range	VQAv2		NExT-QA	
	AP (↑)	AF (↓)	AP (↑)	AF (↓)
$[0.05 - 0.30]$	41.22	-0.41	29.09	-1.72
$[0.08 - 0.45]$	42.78	0.07	32.28	-1.90
$[0.15 - 0.60]$	42.99	1.07	32.06	1.68
$[0.30 - 0.80]$	42.14	2.10	32.13	2.59

Comparison with Model-Merging Algorithms. Fig. 5b compares our method against representative merging strategies adapted to continual learning. TA [21] and MagMax [36] achieve strong stability (negative AF) but sacrifice plasticity, limiting final AP. TIES [66] improves alignment through global heuristics but remains layer-agnostic. OPCM [54] enforces strong orthogonality, suppressing beneficial task interactions and resulting in lower AP and higher forgetting (AF = 5.49) on NExT-QA. In contrast, our method achieves the highest AP (42.78/32.28) with near-zero forgetting (0.07/-1.90), demonstrating a superior stability–plasticity trade-off over existing merging strategies.

Role of Replay. To verify that SCoT's gains are not merely due to stored samples, Tab. 6A compares replay-based and replay-free settings. Even without replay, SCoT preserves prior knowledge through weight-space consolidation, outperforming TIES by +2.60 AP and MagMax by +6.28 AP, while replay further strengthens the stability–plasticity trade-off.

Table 6: Role of replay, class-incremental (CIL), and domain-incremental (DIL) learning.

Tab. 6A: Role of Replay (VQAv2)

Method	w/ Replay		w/o Replay	
	AP(↑)	AF(↓)	AP(↑)	AF(↓)
MagMax	38.48	**-0.13**	25.16	**8.48**
TIES	41.54	1.03	28.84	11.86
SCoT	**42.78**	0.07	**31.44**	8.93

Tab. 6B: CIL

	CIFAR-100		
Method	/5	/10	/20
MagMax	**82.68**	**79.57**	**76.92**
TIES	81.40	79.09	76.30
SCoT	81.90	78.44	76.62

Tab. 6C: DIL

	DomainNet
Method	Accuracy
MagMax	**71.20**
TIES	69.68
SCoT	70.86

Generalization beyond VQACL setting. While our main experiments follow the VQACL protocol for fair comparison, SCoT is not tied to VQA-specific architectures or data. Since it consolidates task knowledge directly in weight space, it is task-agnostic. Tab. 6B and Tab. 6C demonstrate this generality on ViT-B, where SCoT remains competitive on both CIFAR-100 class-incremental and DomainNet domain-incremental settings.

Validation on a Recent Backbone. Tab. 7A shows that SCoT improves over strong baselines on LLaVA-7B, both with and without replay. Without replay, SCoT improves over TIES by +2.96 AP while reducing forgetting; with 5k replay samples, it achieves the best AP and near-zero AF, outperforming TIES by +2.02 AP. The depth-aware ablation in Tab. 7B further shows that applying $\mathcal{L}_{\text{comp}}$ to mid-decoder layers gives the best stability–plasticity trade-off.

Table 7: Results and layer ablation on LLaVA-7B.

Tab. 7A: VQAv2 on LLaVA-7B ($\checkmark = 5k$)

Method	Replay	AP(↑)	AF(↓)
Vanilla	✗	32.51	20.64
QUAD	✓	44.16	1.73
TIES	✗	40.21	4.82
TIES	✓	46.30	0.95
SCoT (Ours)	✗	43.17	2.31
SCoT (Ours)	✓	**48.32**	**0.13**

Tab. 7B: Layer ablation on LLaVA-7B

$\mathcal{L}_{\text{comp}}$	AP(↑)	AF(↓)
Early dec.	44.53	0.28
Mid dec.	**48.32**	**0.13**
Late dec.	46.24	0.21

7 Conclusions

In this paper, we introduce SCoT, a similarity-guided and conflict-aware task consolidation framework for continual VQA that operates directly in parameter space using anchor-relative task vectors. By modeling layer-wise task alignment, removing destructive interference, and applying depth-aware stabilization, SCoT achieves a strong stability–plasticity balance in sequential multimodal learning. Experiments on VQAv2 and NExT-QA show that SCoT reduces forgetting to near-zero, enables positive backward transfer, and improves compositional generalization over existing continual VQA baselines. These results highlight geometric parameter-space consolidation as a promising direction for continual multimodal reasoning.

Acknowledgements

The authors gratefully acknowledge the support and resources provided by the IITB-HRTI Centre of Excellence (CoE) on Computer Vision, Multimodal AI, and Geo-Intelligence, which contributed to this research.

References

1. Adam, K.D.B.J., et al.: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 **1412**(6) (2014)
2. Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., Tuytelaars, T.: Memory aware synapses: Learning what (not) to forget. In: Proceedings of the European conference on computer vision (ECCV). pp. 139–154 (2018)
3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
4. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond (2023)
5. Buzzega, P., Boschini, M., Porrello, A., Abati, D., Calderara, S.: Dark experience for general continual learning: a strong, simple baseline. Advances in neural information processing systems **33**, 15920–15930 (2020)
6. Castellucci, G., Filice, S., Croce, D., Basili, R.: Learning to solve nlp tasks in an incremental number of languages. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 837–847 (2021)
7. Chaudhry, A., Dokania, P.K., Ajanthan, T., Torr, P.H.: Riemannian walk for incremental learning: Understanding forgetting and intransigence. In: Proceedings of the European conference on computer vision (ECCV). pp. 532–547 (2018)
8. Chaudhry, A., Ranzato, M., Rohrbach, M., Elhoseiny, M.: Efficient lifelong learning with a-gem. arXiv preprint arXiv:1812.00420 (2018)
9. Chaudhry, A., Rohrbach, M., Elhoseiny, M., Ajanthan, T., Dokania, P., Torr, P., Ranzato, M.: Continual learning with tiny episodic memories. In: Workshop on Multi-Task and Lifelong Reinforcement Learning (2019)
10. Chen, S., Zhao, Q.: Divide and conquer: Answering questions with object factorization and compositional reasoning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6736–6745 (2023)
11. Cho, J., Lei, J., Tan, H., Bansal, M.: Unifying vision-and-language tasks via text generation. In: International Conference on Machine Learning. pp. 1931–1942. PMLR (2021)
12. Douillard, A., Ramé, A., Couairon, G., Cord, M.: Dytox: Transformers for continual learning with dynamic token expansion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9285–9295 (2022)
13. Dziadzio, S., Udandarao, V., Roth, K., Prabhu, A., Akata, Z., Albanie, S., Bethge, M.: How to merge your multimodal models over time? In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 20479–20491 (2025)
14. Ermis, B., Zappella, G., Wistuba, M., Rawal, A., Archambeau, C.: Memory efficient continual learning with transformers. Advances in Neural Information Processing Systems **35**, 10629–10642 (2022)
15. Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B.: Sharpness-aware minimization for efficiently improving generalization. arXiv preprint arXiv:2010.01412 (2020)
16. Gao, Q., Zhao, C., Sun, Y., Xi, T., Zhang, G., Ghanem, B., Zhang, J.: A unified continual learning framework with general parameter-efficient tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11483–11493 (2023)

17. 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: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6904–6913 (2017)
18. Grossberg, S.: Nonlinear neural networks: Principles, mechanisms, and architectures. Neural networks **1**(1), 17–61 (1988)
19. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6546–6555 (2018)
20. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6700–6709 (2019)
21. Ilharco, G., Ribeiro, M.T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., Farhadi, A.: Editing models with task arithmetic. arXiv preprint arXiv:2212.04089 (2022)
22. Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., Wilson, A.G.: Averaging weights leads to wider optima and better generalization. arXiv preprint arXiv:1803.05407 (2018)
23. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2901–2910 (2017)
24. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)
25. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**(1), 32–73 (2017)
26. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J.: Building machines that learn and think like people. Behavioral and brain sciences **40**, e253 (2017)
27. Li, X., Zhou, Y., Wu, T., Socher, R., Xiong, C.: Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In: International conference on machine learning. pp. 3925–3934. PMLR (2019)
28. Li, Z., Jiang, G., Xie, H., Song, L., Lian, D., Wei, Y.: Understanding and patching compositional reasoning in llms. arXiv preprint arXiv:2402.14328 (2024)
29. Liang, Z., Jiang, W., Hu, H., Zhu, J.: Learning to contrast the counterfactual samples for robust visual question answering. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). pp. 3285–3292 (2020)
30. Lin, S., Yang, L., Fan, D., Zhang, J.: Beyond not-forgetting: Continual learning with backward knowledge transfer. Advances in Neural Information Processing Systems **35**, 16165–16177 (2022)
31. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
32. Liu, J.: Layer specialization underlying compositional reasoning in transformers. arXiv preprint arXiv:2510.17469 (2025)
33. Liu, Y., Hong, Q., Huang, L., Gomez-Villa, A., Goswami, D., Liu, X., van de Weijer, J., Tian, Y.: Continual learning for vlms: A survey and taxonomy beyond forgetting. arXiv preprint arXiv:2508.04227 (2025)

34. Lopez-Paz, D., Ranzato, M.: Gradient episodic memory for continual learning. Advances in neural information processing systems **30** (2017)
35. Malinowski, M., Fritz, M.: A multi-world approach to question answering about real-world scenes based on uncertain input. Advances in neural information processing systems **27** (2014)
36. Marczak, D., Twardowski, B., Trzciński, T., Cygert, S.: Magmax: Leveraging model merging for seamless continual learning. In: European Conference on Computer Vision. pp. 379–395. Springer (2024)
37. 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: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18078–18089 (2025)
38. Matena, M.S., Raffel, C.A.: Merging models with fisher-weighted averaging. Advances in Neural Information Processing Systems **35**, 17703–17716 (2022)
39. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: The sequential learning problem. In: Psychology of learning and motivation, vol. 24, pp. 109–165. Elsevier (1989)
40. Monga, M., Chudasama, V., Wasnik, P., Banerjee, B.: Duet: Dual incremental object detection via exemplar-free task arithmetic. arXiv preprint arXiv:2506.21260 (2025)
41. Nikandrou, M., Yu, L., Suglia, A., Konstas, I., Rieser, V.: Task formulation matters when learning continually: A case study in visual question answering. arXiv preprint arXiv:2210.00044 (2022)
42. Ortiz-Jimenez, G., Favero, A., Frossard, P.: Task arithmetic in the tangent space: Improved editing of pre-trained models. Advances in Neural Information Processing Systems **36**, 66727–66754 (2023)
43. Paik, I., Oh, S., Kwak, T., Kim, I.: Overcoming catastrophic forgetting by neuron-level plasticity control. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 5339–5346 (2020)
44. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
45. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research **21**(140), 1–67 (2020)
46. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision transformers see like convolutional neural networks? Advances in neural information processing systems **34**, 12116–12128 (2021)
47. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2001–2010 (2017)
48. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
49. 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)
50. Saha, G., Garg, I., Roy, K.: Gradient projection memory for continual learning. arXiv preprint arXiv:2103.09762 (2021)

51. Schwarz, J., Czarnecki, W., Luketina, J., Grabska-Barwinska, A., Teh, Y.W., Pascanu, R., Hadsell, R.: Progress & compress: A scalable framework for continual learning. In: International conference on machine learning. pp. 4528–4537. PMLR (2018)
52. Smith, J.S., Karlinsky, L., Gutta, V., Cascante-Bonilla, P., Kim, D., Arbelle, A., Panda, R., Feris, R., Kira, Z.: Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11909–11919 (2023)
53. Srinivasan, T., Chang, T.Y., Pinto Alva, L., Chochlakis, G., Rostami, M., Thomason, J.: Climb: A continual learning benchmark for vision-and-language tasks. Advances in Neural Information Processing Systems **35**, 29440–29453 (2022)
54. Tang, A., Yang, E., Shen, L., Luo, Y., Hu, H., Du, B., Tao, D.: Merging models on the fly without retraining: A sequential approach to scalable continual model merging. arXiv preprint arXiv:2501.09522 (2025)
55. Tang, Y.M., Peng, Y.X., Zheng, W.S.: Learning to imagine: Diversify memory for incremental learning using unlabeled data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9549–9558 (2022)
56. Thrun, S.: A lifelong learning perspective for mobile robot control. In: Intelligent robots and systems. pp. 201–214. Elsevier (1995)
57. Tiwari, R., Killamsetty, K., Iyer, R., Shenoy, P.: Gcr: Gradient coreset based replay buffer selection for continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 99–108 (2022)
58. Wan, T.S., Chen, J.C., Wu, T.Y., Chen, C.S.: Continual learning for visual search with backward consistent feature embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16702–16711 (2022)
59. Wang, Y., Huang, Z., Hong, X.: S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. Advances in Neural Information Processing Systems **35**, 5682–5695 (2022)
60. Wang, Z., Liu, L., Kong, Y., Guo, J., Tao, D.: Online continual learning with contrastive vision transformer. In: European Conference on Computer Vision. pp. 631–650. Springer (2022)
61. Wang, Z., Zhang, Z., Ebrahimi, S., Sun, R., Zhang, H., Lee, C.Y., Ren, X., Su, G., Perot, V., Dy, J., et al.: Dualprompt: Complementary prompting for rehearsal-free continual learning. In: European conference on computer vision. pp. 631–648. Springer (2022)
62. Wortsman, M., Ilharco, G., Gadre, S.Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A.S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., et al.: Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In: International conference on machine learning. pp. 23965–23998. PMLR (2022)
63. Wu, T., Caccia, M., Li, Z., Li, Y.F., Qi, G., Haffari, G.: Pretrained language model in continual learning: A comparative study. In: International conference on learning representations (2022)
64. Xiao, J., Shang, X., Yao, A., Chua, T.S.: Next-qa: Next phase of question-answering to explaining temporal actions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9777–9786 (2021)
65. Xue, M., Zhang, H., Song, J., Song, M.: Meta-attention for vit-backed continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 150–159 (2022)
66. Yadav, P., Tam, D., Choshen, L., Raffel, C.A., Bansal, M.: Ties-merging: Resolving interference when merging models. Advances in Neural Information Processing Systems **36**, 7093–7115 (2023)

67. Yang, E., Shen, L., Wang, Z., Liu, S., Guo, G., Wang, X.: Data augmented flatness-aware gradient projection for continual learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5630–5639 (2023)
68. Yang, E., Tang, A., Shen, L., Guo, G., Wang, X., Cao, X., Zhang, J.: Continual model merging without data: Dual projections for balancing stability and plasticity. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
69. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 21–29 (2016)
70. Yoon, J., Yang, E., Lee, J., Hwang, S.J.: Lifelong learning with dynamically expandable networks. arXiv preprint arXiv:1708.01547 (2017)
71. Yu, L., Yu, B., Yu, H., Huang, F., Li, Y.: Language models are super mario: Absorbing abilities from homologous models as a free lunch. In: Forty-first International Conference on Machine Learning (2024)
72. Yu, Z., Yu, J., Cui, Y., Tao, D., Tian, Q.: Deep modular co-attention networks for visual question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6281–6290 (2019)
73. Zeng, G., Chen, Y., Cui, B., Yu, S.: Continual learning of context-dependent processing in neural networks. Nature Machine Intelligence **1**(8), 364–372 (2019)
74. Zenke, F., Poole, B., Ganguli, S.: Continual learning through synaptic intelligence. In: International conference on machine learning. pp. 3987–3995. Pmlr (2017)
75. Zhang, F., Xu, M., Xu, C.: Geometry sensitive cross-modal reasoning for composed query based image retrieval. IEEE Transactions on Image Processing **31**, 1000–1011 (2021)
76. Zhang, X., Zhang, F., Xu, C.: Explicit cross-modal representation learning for visual commonsense reasoning. IEEE Transactions on Multimedia **24**, 2986–2997 (2021)
77. Zhang, X., Zhang, F., Xu, C.: Multi-level counterfactual contrast for visual commonsense reasoning. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1793–1802 (2021)
78. Zhang, X., Zhang, F., Xu, C.: Vqacl: A novel visual question answering continual learning setting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19102–19112 (2023)