DUALMEM-EVO: DUAL-MEMORY EVOLUTION FOR REPLAY-FREE CONTINUAL ADAPTATION OF VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) have demonstrated remarkable transferability, yet their deployment in dynamic environments requires continuous adaptation to emerging tasks without access to previous training data. This setting poses a fundamental challenge: how to acquire new knowledge while preserving previously learned semantic representations. Existing continual adaptation approaches typically rely on either shared adaptation modules, which suffer from severe interference, or isolated task experts, which limit knowledge transfer across tasks. In this work, we propose DualMem-Evo, a dual-memory continual adaptation framework that views lifelong learning as a process of knowledge evolution rather than static knowledge preservation. Our framework introduces two complementary memories: a semantic evolution memory that continuously accumulates transferable cross-task knowledge, and an episodic expert memory that preserves task-specific expertise. To enable effective memory reuse under a strict no-replay constraint, we develop a Prototype-Guided Expert Evolution mechanism that dynamically retrieves, interacts with, and consolidates historical expert memory modules through cross-modal semantic prototypes and low-rank memory fusion. Furthermore, we introduce functional memory separation and stability-aware semantic regulation to maintain representation diversity while preventing semantic drift during incremental adaptation. The empirical results demonstrate that our framework achieves the state-of-the-art performance.

1 INTRODUCTION

Vision-language models (VLMs), such as CLIP, learn aligned visual-semantic representations from large-scale image-text corpora Radford et al. (2021); Kim et al. (2022), achieving strong zero-shot transfer across tasks including open-world recognition and multimodal understanding Li et al. (2026); Zhao et al. (2025). However, real-world deployment requires continuous adaptation to evolving concepts, domains, and user needs while preserving prior knowledge Ye & Bors (2023). This challenge motivates **multi-task incremental learning (MTIL)** for VLMs, where tasks arrive sequentially without access to previous training data Song et al. (2026).

Unlike conventional offline multi-task learning, continual adaptation requires balancing *knowledge accumulation* Ke et al. (2021) and *knowledge preservation* Chaudhry et al. (2021). A continual VLM must acquire new semantic knowledge while maintaining stable representations from previous tasks. However, shared adaptation modules often suffer from task interference and catastrophic forgetting due to parameter competition Parisi et al. (2019), whereas independent task-specific experts preserve old knowledge but fail to leverage task correlations, leading to fragmented knowledge utilization Kang et al. (2022).

Recent parameter-efficient continual learning methods based on low-rank adaptation (LoRA) provide a promising direction for continual VLMs learning by introducing lightweight task-specific updates while freezing the pretrained backbone Hu et al. (2022); Gao et al. (2026). Nevertheless, current approaches mainly treat incremental adaptation as either parameter preservation Qiu et al. (2026) or expert selection Wang et al. (2025), overlooking an essential property of lifelong learning: **memories should evolve rather than remain isolated snapshots**. This characteristic also aligns with the way the brain processes information, in which some knowledge represents shared semantic structures that can continuously benefit future tasks, while other knowledge captures task-specific

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

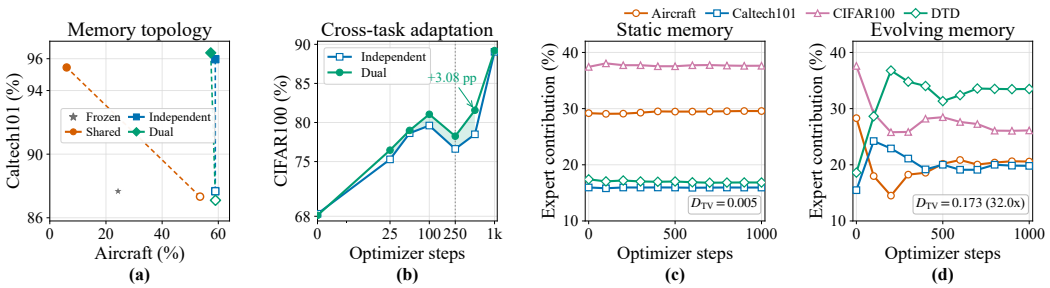

Figure 1: (a) Memory-topology comparison on Aircraft and Caltech101. Hollow and filled markers denote the states after learning Aircraft and after subsequently learning Caltech101, respectively. (b) Adaptation to CIFAR100 after Caltech101; the shaded region highlights the gain from dual memory, the dotted line marks the 250-step warm-up boundary, and "pp" denotes percentage points. (c–d) Historical-expert contributions during EuroSAT training with fixed and adaptive memory fusion. For the normalized expert-contribution distribution $\mathbf{p}^{(s)}$ at step s, $D_{\mathrm{TV}} = \frac{1}{2}\sum_k |p_k^{(1000)} - p_k^{(0)}|$. Detailed experimental settings and contribution normalization are provided in Appendix D.1.

expertise that should remain specialized Shi et al. (2025). Therefore, an effective continual vision-language learner should not choose between sharing and isolation. Instead, it should enable different forms of knowledge to coexist, interact, and evolve throughout the learning process.

The empirical analyses in Figure 1 reveal two requirements for effective continual adaptation. Figure 1(a) shows that shared adaptation is vulnerable to cross-task interference, whereas isolated experts better preserve task-specific knowledge. Figure 1(b) further indicates that combining shared and episodic memories facilitates subsequent adaptation compared with strict expert isolation. Figures 1(c) and 1(d) show that, unlike static aggregation, evolving memory dynamically reallocates historical-expert contributions as new evidence arrives. Together, these observations motivate a dual-memory system that protects episodic expertise while continuously evolving reusable semantic knowledge. Additional functional-separation and feature-level analyses are provided in Appendix D.

Motivated by this brain learning mechanism Shi et al. (2025), which process information through cortical side and hippocampal side, we propose **DualMem-Evo**, a dual-memory continual adaptation framework that models lifelong learning as a process of **knowledge evolution**. Our key insight is that effective continual vision-language adaptation requires two complementary forms of memory: a *semantic evolution memory* that continuously accumulates transferable cross-task knowledge, and an *episodic memory* that preserves task-specific discriminative expertise. Instead of forcing all tasks to share a single adaptation space or maintaining fully independent experts, DualMem-Evo enables shared knowledge and specialized knowledge to collaboratively evolve.

Specifically, the proposed DualMem-Evo framework formulates the LoRA parameters as the learnable memory blocks, where semantic memory is continuously inherited and updated across incremental sessions, while episodic memories are independently maintained as task experts. To enable effective knowledge reuse without storing historical samples, we develop a Prototype-Guided Expert Evolution Mechanism (PGEEM). By constructing compact cross-modal semantic prototypes from frozen CLIP representations, historical experts are treated as an evolving memory bank rather than static parameter snapshots. Different from conventional nearest-neighbor retrieval Ye & Bors (2026), PGEEM involves a centered semantic routing strategy that identifies both positively correlated and complementary knowledge directions, allowing the model to exploit diverse historical memorized knowledge. The selected expert memory blocks are further integrated through adaptive low-rank memory consolidation, producing a compact and reusable memory representation.

Furthermore, we empirically observe that encouraging semantic memory and expert memory to encode complementary functional representations can balance specialization and stability, leading to better performance, as shown in Table 3 from Appendix D.2. Here, stability refers to retaining previously acquired knowledge. We implement this goal by introducing a Stability-Regularized Evolution mechanism that involves a marginal contribution based memory separation objective to promote representation diversity and a stability-aware semantic evolution regulation loss to preserve the continuity of semantic memory across sessions. Together, these designs enable continual knowledge accumulation while mitigating catastrophic forgetting under strict no-replay constraints.

Our contributions are summarized as follows: (1) We propose **DualMem-Evo**, a dual-memory continual adaptation framework for vision-language models, which explicitly separates lifelong semantic evolution from task-specific expertise preservation through complementary memory mechanisms; (2) We introduce a Prototype-Guided Expert Evolution Mechanism that transforms historical experts from isolated snapshots into an adaptive memory bank. By leveraging centered semantic routing and low-rank consolidation, the framework enables effective knowledge transfer and compact replay-free memory reuse; (3) We propose a functional memory separation objective and a stability-aware semantic evolution regularization strategy, which jointly improve representation diversity and prevent semantic drift during continual adaptation; (4) Extensive experiments on MTIL benchmarks demonstrate that DualMem-Evo achieves superior performance compared with existing parameter-efficient and expert-based continual learning methods, showing strong capability in balancing knowledge acquisition, retention, and efficient memory utilization.

2 Related Work

Continual Adaptation of Vision–Language Models. Pre-trained VLMs such as CLIP Radford et al. (2021) provide strong zero-shot transferability, which can deteriorate during continual adaptation Zheng et al. (2023). Existing approaches preserve knowledge through prediction distillation Li & Hoiem (2018), weight interpolation Wortsman et al. (2022), or rehearsal. ZSCL Zheng et al. (2023) combines reference-data distillation with parameter averaging, while GIFT Wu et al. (2025) uses synthetic image–text pairs and adaptive weight consolidation. More recently, NuSA-CL Jo & Kim (2026) constrains low-rank updates to approximate null spaces of model weights without replay. DualMem-Evo instead organizes adaptation into evolving semantic memory and protected episodic experts, using only current-session data.

Parameter-Efficient and Expert-Based (PEEB) methods. PEEB adapt frozen backbones through lightweight modules such as LoRA Hu et al. (2022). L2P Wang et al. (2022b), DualPrompt Wang et al. (2022a), and CODA-Prompt Smith et al. (2023) use learnable prompts to retain and retrieve task knowledge. For continual VLMs, MoE-Adapters Yu et al. (2024), DIKI Tang et al. (2024), and IAP Fu et al. (2025) enable input-dependent expert or prompt adaptation. DPW Jang et al. (2026) introduces token-level prefix weighting and residual adapter modulation to balance adaptation and knowledge preservation. Our focus is prototype-guided historical-expert collaboration and low-rank consolidation within a dual-memory architecture.

Complementary Memory Systems. Complementary learning systems theory McClelland et al. (1995) links continual learning to fast episodic and slow semantic memory. DualNet Pham et al. (2021), CLS-ER Arani et al. (2022), BiMeCo Nie et al. (2023), and corticohippocampal hybrid networks Shi et al. (2025) instantiate complementary learners or interacting memories. HiCL Kapoor et al. (2026) combines hippocampal pattern separation, prototype-gated experts, and replay-based consolidation, while HippoTune Chen et al. (2026) performs iterative associative retrieval within Transformer layers. These studies motivate complementary memory organization and active knowledge reuse, providing a foundation for studying parameterized semantic and episodic memories in continual VLM adaptation.

3 Methodology

3.1 Problem Definition

We consider the problem of Multi-domain Task Incremental Learning (MTIL) for vision–language models, where a pretrained model is required to continuously adapt to a sequence of emerging tasks while preserving previously acquired knowledge. Unlike conventional multi-task learning that jointly optimizes all tasks with simultaneous access to the complete training data, MTIL assumes that tasks arrive sequentially and training data from previous tasks become inaccessible after each learning session. This setting reflects realistic deployment scenarios of vision–language models, where new visual concepts, domains, or application requirements continuously appear after the initial model deployment. A model is presented with a sequence of T tasks $\{\mathcal{D}_0, \mathcal{D}_1, \ldots, \mathcal{D}_{T-1}\}$, where task t provides a training set $\mathcal{D}_t = \{(\mathbf{x}_i^{(t)}, y_i^{(t)})\}_{i=1}^{N_t}$, where $y_i^{(t)} \in \mathcal{C}_t$ and N_t denotes the number of training samples. $C_t = |\mathcal{C}_t|$ is the number of classes in task t. At session/task t, only $\mathcal{D}_t$ is accessible while all prior data $\{\mathcal{D}_j\}_{j<t}$ are unavailable, enforcing a strict no-replay constraint. During evaluation, we follow the standard MTIL protocol, where prediction is performed within the task-specific class space Zheng et al. (2023).

3.2 DUAL-MEMORY CONTINUAL ADAPTATION FRAMEWORK

We address continual learning on CLIP under the MTIL protocol, where a model sequentially adapts to T tasks $\{\mathcal{D}_0, \mathcal{D}_1, \ldots, \mathcal{D}_{T-1}\}$ without revisiting past training data. Inspired by the complementary nature of semantic accumulation and episodic specialization, we introduce a dual-memory continual adaptation paradigm, where a shared evolutionary memory and a task-specific expert memory jointly drive lifelong knowledge acquisition, as shown in Fig. 2. The key insight is that lifelong adaptation requires two complementary memory mechanisms: (1) **Semantic Evolution Memory**: continuously accumulates transferable cross-task knowledge and maintains a globally shared representation space; (2) **Episodic Expert Memory**: preserves task-specific discriminative knowledge, preventing destructive interference between heterogeneous tasks.

Instead of forcing all tasks to compete within a single adaptation space, our framework allows shared knowledge and task-specific expertise to evolve cooperatively. The semantic memory provides stable knowledge accumulation, while the expert memory provides selective specialization. Their interaction forms a continual knowledge evolution cycle consisting of: memory accumulation, expert retrieval, knowledge consolidation, and positive knowledge transfer.

We instantiate DualMem-Evo on CLIP by freezing the pretrained backbone and inserting two LoRA-based memory adapters into the MLP layers of both image and text Transformers. At session t, the semantic adapter is denoted by $G_t = \{(\mathbf{A}_g^i, \mathbf{B}_g^i)\}_{i=1}^L$, while $T_t = \{(\mathbf{A}_t^i, \mathbf{B}_t^i)\}_{i=1}^L$ denotes the current episodic adapter, where L is the number of adapted linear layers. G_t is initialized from the preceding semantic memory G_{t-1}, whereas T_t is initialized for the current session. After optimization and consolidation, T_t is stored as the frozen episodic expert $\mathcal{E}_t$, while G_t is preserved to initialize G_{t+1}. Given representation $\mathbf{z}$, the output of its i-th layer is defined as:

$$\mathbf{z}' = \mathbf{z}\mathbf{W}^i + s_g\, \mathbf{z}\mathbf{A}_g^i\mathbf{B}_g^i + s_t\, \mathbf{z}\mathbf{A}_t^i\mathbf{B}_t^i, \tag{1}$$

where $\mathbf{W}^i \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$ is the weight matrix of the i-th representation layer. d_{in} and d_{out} denote the input and output dimensions, respectively. $\mathbf{A}_g^i \in \mathbb{R}^{d_{\text{in}} \times r_g}$, $\mathbf{B}_g^i \in \mathbb{R}^{r_g \times d_{\text{out}}}$ are the semantic memory blocks implemented using the LoRA matrices, and $\mathbf{A}_t^i \in \mathbb{R}^{d_{\text{in}} \times r_t}$, $\mathbf{B}_t^i \in \mathbb{R}^{r_t \times d_{\text{out}}}$ are the episodic memory blocks implemented using the LoRA matrices. The scaling factors $s_g = \alpha/\sqrt{r_g}$ and $s_t = \alpha/\sqrt{r_t}$ normalize activation magnitudes across different rank choices, suppressing amplitude drift when ranks vary, where α is the LoRA scaling factor. $r_g \leq \min(d_{\text{in}}, d_{\text{out}})$ and $r_t \leq \min(d_{\text{in}}, d_{\text{out}})$ denote the ranks of the semantic and episodic adapters, respectively.

3.3 PROTOTYPE-GUIDED EXPERT EVOLUTION AND MEMORY CONSOLIDATION

As the number of incremental tasks increases, maintaining and exploiting historical knowledge becomes increasingly challenging. A naive strategy that aggregates all previous experts introduces irrelevant knowledge interference, while completely isolating historical expert memory modules prevents beneficial knowledge transfer across related tasks. Therefore, an effective continual adaptation system should not only preserve historical expertise but also dynamically identify, reuse, and consolidate valuable knowledge according to the evolving task distribution.

To this end, we introduce a Prototype-Guided Expert Evolution Mechanism (PGEEM), which regards historical task experts as an evolving memory bank rather than static parameter snapshots. The proposed mechanism consists of three stages: (1) semantic memory modelling through cross-modal prototypes, (2) selective expert routing through centered affinity estimation, and (3) low-rank memory consolidation for compact future reuse; unlike conventional nearest-neighbor expert retrieval that only considers positive similarity Ye & Bors (2026), our method models expert relationships from a deviation perspective, allowing both transferable and complementary knowledge to participate in memory evolution.

Cross-modal semantic memory modelling. To enable replay-free expert retrieval, we store compact semantic signatures instead of historical training samples. After completing task j, the corresponding expert $\mathcal{E}_j$ is associated with a cross-modal prototype representation constructed from the frozen pretrained CLIP encoders. Specifically, for each category c in task j, we define:

$$\mathbf{p}_c^{(j)} = \text{Norm}\left(\frac{\bar{\mathbf{u}}_c + \bar{\mathbf{v}}_c}{2}\right), \tag{2}$$

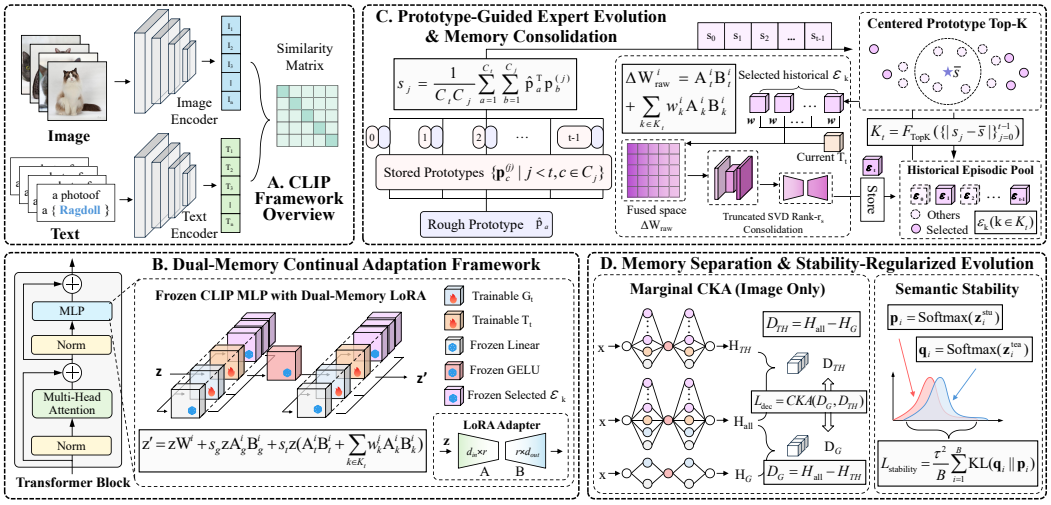

Figure 2: **Overview of the proposed DualMem-Evo framework.** (A) Frozen CLIP encoders provide a reference vision–language space for constructing cross-modal semantic prototypes. (B) Each adapted MLP linear layer combines the evolving semantic memory G_t, the current episodic memory T_t, and the selected frozen historical experts $\{\mathcal{E}_k\}_{k\in\mathcal{K}_t}$. (C) PGEEM retrieves correlated and complementary historical memories using centered prototype affinities, learns their layer-wise contributions, and consolidates the resulting task-side update into a fixed-rank episodic expert $\mathcal{E}_t$ through truncated SVD. (D) Marginal CKA reduces functional redundancy between semantic and episodic contributions, while KL distillation from the frozen G_{t-1} teacher to the trainable G_t student stabilizes semantic evolution. Snowflake and flame symbols denote frozen and trainable components, respectively.

where $\mathrm{Norm}(\cdot)$ denotes L2 normalization and $\bar{\mathbf{u}}_c$ is the visual prototype computed as:

$$\bar{\mathbf{u}}_c = \frac{1}{N_c}\sum_{n=1}^{N_c} \mathrm{Norm}\left(f_I^0(\mathbf{x}_{c,n})\right), \tag{3}$$

where $\mathbf{x}_{c,n}$ denotes the n-th sample from the c-th category and N_c is the total number of samples belonging to the category c. $f_I^0(\cdot)$ denotes the output of the frozen CLIP image encoder. The semantic prototype $\bar{\mathbf{v}}_c$ is obtained by averaging textual embeddings generated from multiple prompt templates:

$$\bar{\mathbf{v}}_c = \frac{1}{P}\sum_{p=1}^{P} \mathrm{Norm}\left(f_T^0(\mathrm{prompt}_p(c))\right), \tag{4}$$

where $f_T^0(\cdot)$ represents the frozen CLIP text encoder and where P is the number of prompt templates shared by all classes within the current dataset. $\mathrm{prompt}_p(c)$ denotes the p-th template instantiated with class name c. The resulting prototype jointly captures visual appearance and linguistic semantics, providing a compact task identity without storing any historical images. Compared with image-only task statistics He et al. (2025), the cross-modal prototype provides a more stable semantic memory information because it inherits CLIP's aligned vision-language representation space.

Centered semantic memory routing. When a new task t arrives, we first estimate its semantic location in the memory space. For each current class a, a lightweight image prototype is calculated using $M = 5$ randomly sampled examples, expressed as:

$$\hat{\mathbf{p}}_a = \frac{1}{M}\sum_{m=1}^{M} \mathrm{Norm}\left(f_I^0(\mathbf{x}_{a,m})\right), \qquad M = 5, \tag{5}$$

The affinity between the current task and historical expert $\mathcal{E}_j$ is defined as:

$$s_j = \frac{1}{C_t C_j}\sum_{a=1}^{C_t}\sum_{b=1}^{C_j} \hat{\mathbf{p}}_a^{\top}\mathbf{p}_b^{(j)}. \tag{6}$$

Rather than directly selecting experts with the largest similarity scores, we measure the relative semantic deviation among all historical memories:

$$\tilde{s}_j = s_j - \bar{s}, \qquad \bar{s} = \frac{1}{t}\sum_{j=0}^{t-1} s_j, \qquad t > 0. \tag{7}$$

and select the informative memory directions:

$$\mathcal{K}_t = F_{\mathrm{TopK}}\left(\{|\tilde{s}_j|\}_{j=0}^{t-1}\right), \tag{8}$$

where $\mathcal{K}_t$ denotes the selected expert indexes and F_{TopK} returns the indices of the $\min(K, t)$ largest absolute centered scores. The intuition behind centered routing is that experts with large deviations represent distinctive knowledge directions. Experts with positive deviations provide transferable semantic patterns, while experts with negative deviations provide complementary or contrastive information that may help the model establish discriminative boundaries.

Therefore, unlike conventional similarity-based retrieval that only favors close neighbors, our routing strategy discovers both cooperative and complementary memories. For each selected expert $\mathcal{E}_k$, the collaboration coefficient at every adapted layer is initialized as $w_k^i = \tilde{s}_k$. Therefore, an expert with below-average affinity receives a negative initial coefficient, providing a signed subtractive direction that can be further adjusted during optimization.

Expert memory evolution through adaptive memory interaction. During session t, selected historical expert memory modules participate in the forward process together with the current task expert memory module. The updated representation becomes:

$$\mathbf{z}' = \mathbf{z}\mathbf{W}^i + s_g \mathbf{z}\mathbf{A}_g^i\mathbf{B}_g^i + s_t \mathbf{z}\Big(\mathbf{A}_t^i\mathbf{B}_t^i + \sum\nolimits_{k\in\mathcal{K}_t} w_k^i \mathbf{A}_k^i\mathbf{B}_k^i\Big), \tag{9}$$

where w_k^i is a trainable layer-wise collaboration coefficient for the frozen historical expert $\mathcal{E}_k$. Although initialized with the same centered affinity across layers, these coefficients are optimized independently in different layers and modalities. This allows the contribution of each historical expert to adapt to the representation hierarchy, while its LoRA factors remain frozen. After optimization, the learned coefficients define the evolved task-side update, which is consolidated into a compact episodic expert.

Low-rank memory consolidation. Although adaptive expert memory evolution enables knowledge transfer, directly storing all historical expert memory information leads to continuously increasing memory consumption. Moreover, repeatedly preserving redundant expert directions may introduce conflicting knowledge. To address this issue, PGEEM performs low-rank memory consolidation to compress the evolved expert trajectory into a compact representation. Specifically, after expert memory module interaction, the merged task-side update before scaling is defined as:

$$\Delta\mathbf{W}_{\mathrm{raw}}^i = \mathbf{A}_t^i\mathbf{B}_t^i + \sum\nolimits_{k\in\mathcal{K}_t} w_k^i \mathbf{A}_k^i\mathbf{B}_k^i. \tag{10}$$

Since LoRA adaptations lie in a low-dimensional subspace, we decompose the evolved update using truncated singular value decomposition:

$$\Delta\mathbf{W}_{\mathrm{raw}}^i \approx \mathbf{U}_{r_s}\mathbf{\Sigma}_{r_s}\mathbf{V}_{r_s}^\top, \tag{11}$$

where $\mathbf{U}_{r_s}$ and $\mathbf{V}_{r_s}$ contain the top-r_s left and right singular vectors, respectively. The consolidated LoRA factors are reconstructed as:

$$\widetilde{\mathbf{A}}_t^i = \mathbf{U}_{r_s}\mathbf{\Sigma}_{r_s}^{1/2}, \qquad \widetilde{\mathbf{B}}_t^i = \mathbf{\Sigma}_{r_s}^{1/2}\mathbf{V}_{r_s}^\top. \tag{12}$$

The original task expert memory module is then replaced by the compressed expert representation $\mathcal{E}_t = \{\widetilde{\mathbf{A}}_t^i, \widetilde{\mathbf{B}}_t^i\}_{i=1}^L$. By retaining only dominant singular directions, the consolidation process achieves: **(i) Memory Compression**, where multiple expert trajectories are merged into a fixed-rank representation; **(ii) Knowledge Filtering**, where low-energy directions caused by noise or conflicting transfer are removed; and **(iii) Memory Absorption**, where adaptive expert interactions are permanently integrated into future reusable memories.

3.4 Memory Separation and Stability-Regularized Evolution

Although semantic and episodic expert memories serve distinct roles, their adaptation branches may converge to similar representation spaces, weakening complementarity and limiting the model's ability to preserve general knowledge while acquiring task-specific expertise. Moreover, continuous updates to semantic evolution memory can cause representation drift and, without proper regulation, overwrite existing semantic structures and lead to catastrophic forgetting.

To address these challenges, we introduce two complementary mechanisms: (1) **memory separation alignment**, which encourages different memory components to encode distinct representation geometries; and (2) **stability-aware semantic evolution regulation**, which constrains the evolution of shared memory during incremental adaptation. Together, these mechanisms maintain a balance between knowledge specialization and lifelong knowledge stability.

Table 1: Comparison with SOTA methods on the MTIL benchmark (Order-I) using the "Transfer", "Average", and "Last" scores (%). We label the best methods with **bold** styles. The lines with colored background correspond to the results of our proposed method.

	Method	Aircraft	Caltech101	CIFAR100	DTD	EuroSAT	Flowers	Food	MNIST	OxfordPet	Cars	SUN397	Average
CLIP	Zero-shot	24.3	88.4	68.2	44.6	54.9	71.0	88.5	59.4	89.0	64.7	65.2	65.3
	Full Fine-tune	62.0	95.1	89.6	79.5	98.9	97.5	92.7	99.6	94.7	89.6	81.8	89.2
	LoRA Fine-tune	62.2	97.2	89.3	79.7	98.9	97.8	92.4	99.6	94.3	90.9	81.2	89.4
Transfer	LwF	-	74.5	56.9	39.1	51.1	52.6	72.8	60.6	75.1	30.3	55.9	56.9
	WiSE-FT	-	73.5	55.6	35.6	41.5	47.0	68.3	53.9	69.3	26.8	51.9	52.3
	ZSCL	-	86.0	67.4	45.4	50.4	69.1	87.6	61.8	86.8	60.1	66.8	68.1
	MoE-Adapters	-	87.9	68.2	44.4	49.9	70.7	88.7	59.7	89.1	64.5	65.5	68.9
	DIKI	-	92.9	69.0	43.2	48.2	67.4	85.2	63.0	87.9	63.8	66.2	68.7
	IAP	-	**93.0**	68.7	44.0	47.0	70.4	85.9	63.5	89.7	**66.2**	63.3	69.2
	GIFT	-	88.5	69.8	**46.0**	49.4	68.5	87.1	**69.9**	88.9	57.7	67.7	69.3
	ConDU(LoRA)	-	88.1	68.9	45.7	**57.0**	**71.3**	88.8	61.2	89.3	65.1	**67.8**	70.3
	DPW	-	92.9	69.0	45.3	54.2	71.1	86.2	63.8	89.3	65.7	66.5	**70.4**
	KeepLoRA+	-	85.9	**69.9**	44.6	53.7	70.9	**88.9**	65.4	**90.8**	63.0	66.1	69.9
	DualMem-Evo	-	87.4	68.1	**46.0**	56.3	71.2	**88.9**	65.1	88.9	64.6	67.2	**70.4**
Average	LwF	36.3	86.9	72.0	59.0	73.7	60.0	73.6	74.8	80.0	37.3	58.1	64.7
	WiSE-FT	26.7	86.5	64.3	57.1	65.7	58.7	71.1	70.5	75.8	36.9	54.6	60.7
	ZSCL	45.1	92.0	80.1	64.3	79.5	81.6	89.6	75.2	88.9	64.7	68.0	75.4
	MoE-Adapters	50.2	91.9	83.1	69.4	78.9	84.0	89.1	73.7	89.3	67.7	66.9	76.7
	DIKI	45.1	95.5	83.1	64.8	79.9	83.5	87.0	76.2	89.6	67.0	67.1	76.3
	IAP	45.9	95.8	83.3	66.5	79.5	84.8	87.5	76.6	91.0	69.2	64.5	76.8
	GIFT	51.9	93.9	81.4	67.7	80.3	82.8	89.3	**80.6**	90.3	63.1	68.9	77.3
	ConDU(LoRA)	51.9	94.9	84.4	69.8	81.1	84.4	90.0	77.3	89.5	69.0	**69.3**	78.3
	DPW	**60.4**	96.2	84.5	70.8	82.4	**85.8**	87.7	76.8	90.9	**69.4**	67.6	79.3
	KeepLoRA+	58.4	**96.5**	84.4	67.8	82.1	84.5	**90.7**	77.8	**91.9**	67.5	67.6	79.0
	DualMem-Evo	59.6	95.5	**85.2**	**71.1**	**83.0**	84.7	90.5	77.7	90.4	**69.4**	68.5	**79.6**
Last	LwF	26.3	87.5	71.9	66.6	79.9	66.9	83.8	**99.6**	92.1	66.1	80.4	74.6
	WiSE-FT	27.2	90.8	68.0	68.9	86.9	74.0	87.6	**99.6**	92.6	77.8	81.3	77.7
	ZSCL	40.6	92.2	81.3	70.5	94.8	90.5	91.9	98.7	93.9	85.3	80.2	83.6
	MoE-Adapters	49.8	92.2	86.1	78.1	95.7	94.3	89.5	98.1	89.9	81.6	80.0	85.0
	DIKI	45.2	95.7	86.3	72.9	98.0	97.0	89.2	99.4	94.2	81.6	76.6	85.1
	IAP	46.8	96.1	86.7	75.2	98.1	97.0	89.6	99.4	94.7	82.8	76.7	85.7
	GIFT	47.9	95.6	82.8	75.1	97.3	94.2	91.7	99.2	94.2	87.0	80.9	86.0
	ConDU(LoRA)	48.9	95.2	87.8	78.5	96.3	95.2	91.7	97.6	93.0	85.3	78.8	86.2
	DPW	**60.5**	96.6	87.9	**80.4**	**98.5**	**98.1**	89.4	**99.6**	**95.1**	86.3	78.8	88.3
	KeepLoRA+	57.3	**97.6**	87.2	76.5	98.4	95.7	**92.6**	99.5	94.7	87.2	**83.2**	88.2
	DualMem-Evo	59.9	96.4	**89.1**	**80.4**	98.2	95.8	92.4	**99.6**	94.4	**90.6**	81.9	**89.0**

Marginal contribution alignment for memory separation. The objective of memory separation is not to force parameter-level orthogonality, but rather to encourage different memories to capture complementary functional information. Therefore, we measure the similarity between memories through their conditional marginal contributions rather than their raw activations. Specifically, considering the residual interaction among semantic memory, current expert memory, and historical expert memory, we define three forward paths through the image encoder:

Table 2: Ablation of the dual-memory architecture and its regulation on MTIL Order-I. Single-memory variants use rank 128, while dual-memory variants use $r_g = r_t = 64$, resulting in comparable total LoRA ranks.Best results are marked in **bold**.

Configuration	Transfer	Avg.	Last
Semantic memory only ($r = 128$)	69.43	70.35	74.97
Episodic memory only ($r = 128$)	70.11	79.12	88.50
Unregulated dual memory ($r_g = r_t = 64$)	69.92	77.81	86.60
Regulated dual memory w/o PGEEM	70.15	79.36	88.78
DualMem-Evo	**70.37**	**79.59**	**88.95**

$$\mathbf{H}_{\text{all}} = f_I^{(G,T,H)}(\mathbf{x}), \qquad \mathbf{H}_G = f_I^{(G,0,0)}(\mathbf{x}), \qquad \mathbf{H}_{TH} = f_I^{(0,T,H)}(\mathbf{x}), \tag{13}$$

where $f_I^{(G,T,H)}$ denotes the complete forward path containing semantic memory, current expert memory, and historical expert memory. $f_I^{(G,0,0)}$ represents the path containing only semantic evolution memory. $f_I^{(0,T,H)}$ represents the path containing current and historical expert memories without semantic memory. The conditional marginal contributions are approximated as:

$$\mathbf{D}_G = \mathbf{H}_{\text{all}} - \mathbf{H}_{TH}, \qquad \mathbf{D}_{TH} = \mathbf{H}_{\text{all}} - \mathbf{H}_G. \tag{14}$$

Although residual interactions between adaptation branches are inherently nonlinear, the subtraction-based marginal representation provides an effective first-order approximation of each

memory component's functional contribution. Given a mini-batch containing B samples, we denote $\mathbf{X} = \mathbf{D}_G \in \mathbb{R}^{B\times d}$ and $\mathbf{Y} = \mathbf{D}_{TH} \in \mathbb{R}^{B\times d}$ as the representation matrices using Eq. (14). Both representations are centered along the batch dimension:

$$\mathbf{X}^c = \mathbf{X} - \mathbf{1}\boldsymbol{\mu}_X^\top, \qquad \mathbf{Y}^c = \mathbf{Y} - \mathbf{1}\boldsymbol{\mu}_Y^\top, \tag{15}$$

where $\boldsymbol{\mu}_X = \frac{1}{B}\sum_{i=1}^{B}\mathbf{X}_i$ and $\boldsymbol{\mu}_Y = \frac{1}{B}\sum_{i=1}^{B}\mathbf{Y}_i$. $\mathbf{X}_i$ and $\mathbf{Y}_i$ denote the i-th sample of $\mathbf{X}$ and $\mathbf{Y}$, respectively. The corresponding sample-level similarity structures are represented by Gram matrices $\mathbf{K}_X = \mathbf{X}^c(\mathbf{X}^c)^\top, \mathbf{K}_Y = \mathbf{Y}^c(\mathbf{Y}^c)^\top$. We formulate the memory separation objective using Centered Kernel Alignment (CKA):

$$\mathcal{L}_{\text{dec}} = \text{CKA}(\mathbf{D}_G, \mathbf{D}_{TH}) = \frac{\langle \mathbf{K}_X, \mathbf{K}_Y\rangle_F}{\max(\|\mathbf{K}_X\|_F\|\mathbf{K}_Y\|_F, \zeta)}. \tag{16}$$

where $\langle\cdot,\cdot\rangle_F$ denotes the Frobenius inner product and ζ is a small constant for numerical stability.

Minimizing $\mathcal{L}_{\text{dec}}$ encourages semantic memory and expert memory to encode different sample-relationship structures. Unlike explicit orthogonality constraints, this functional separation does not restrict parameter optimization and therefore allows each memory component to adapt flexibly. Importantly, the separation objective is only applied to the image encoder, where visual representation diversity provides the most informative signal for distinguishing memory functions.

Stability-aware semantic memory evolution. The semantic evolution memory is responsible for accumulating transferable knowledge across tasks. However, continuous updating also makes it the main source of catastrophic forgetting. Conventional old-knowledge distillation commonly relies on stored or generated reference samples, which are unavailable under our no-replay protocol. Instead, we introduce an intra-session stability regulation strategy that constrains semantic memory evolution using only current-task data. At the beginning of session $t > 0$, we preserve a frozen copy of the previous semantic memory G_{t-1} as a session-local teacher. The current semantic memory G_t acts as the student and is optimized together with the current expert memory.

Different from conventional distillation that transfers complete predictions, our objective is applied exclusively to the semantic memory pathway. This design prevents task experts from being constrained by historical semantic structures and preserves their specialization capability. For an input image x_i and category c, teacher and student produce normalized visual and textual representations $\{\bar{\mathbf{h}}_i^{\text{tea}}, \bar{\mathbf{t}}_c^{\text{tea}}\}$ and $\{\bar{\mathbf{h}}_i^{\text{stu}}, \bar{\mathbf{t}}_c^{\text{stu}}\}$, respectively. The temperature-scaled similarity logits are:

$$\mathbf{z}_{i,c}^{\text{tea}} = \frac{\gamma(\bar{\mathbf{h}}_i^{\text{tea}})^\top\bar{\mathbf{t}}_c^{\text{tea}}}{\tau}, \mathbf{z}_{i,c}^{\text{stu}} = \frac{\gamma(\bar{\mathbf{h}}_i^{\text{stu}})^\top\bar{\mathbf{t}}_c^{\text{stu}}}{\tau}, \tag{17}$$

where γ is the frozen CLIP logit scale and τ denotes the distillation temperature. The teacher and student distributions are obtained as $\mathbf{q}_i = \text{Softmax}(\mathbf{z}_i^{\text{tea}})$ and $\mathbf{p}_i = \text{Softmax}(\mathbf{z}_i^{\text{stu}})$, respectively, where the logit vectors are given by $\mathbf{z}_i^{\text{tea}} = [z_{i,c}^{\text{tea}}]_{c\in\mathcal{C}_t}$ and $\mathbf{z}_i^{\text{stu}} = [z_{i,c}^{\text{stu}}]_{c\in\mathcal{C}_t}$. The stability regulation objective is formulated as:

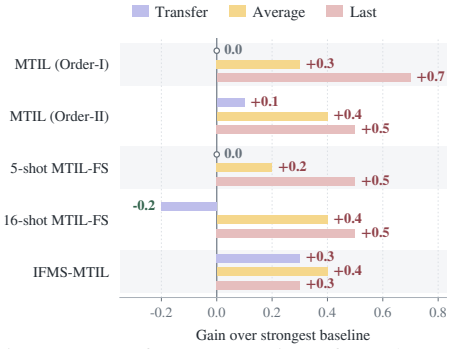

Figure 3: Performance gains of DualMem-Evo over the strongest competing result for each metric and protocol.

$$\mathcal{L}_{\text{stability}} = \frac{\tau^2}{B}\sum_{i=1}^{B}\text{KL}(\mathbf{q}_i\|\mathbf{p}_i) = \frac{\tau^2}{B}\sum_{i=1}^{B}\sum_{c=1}^{C_t} q_{i,c}\log\frac{q_{i,c}}{p_{i,c}} \quad . \tag{18}$$

The temperature factor τ^2 compensates for gradient attenuation caused by softened probability distributions. This objective does not attempt to reconstruct inaccessible historical samples. Instead, it preserves local semantic consistency between consecutive semantic memories, preventing excessive representation drift while allowing continuous adaptation.

Overall optimization objective. The final training objective of DualMem-Evo combines task adaptation, memory separation, and semantic stability regulation, defined as:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{dec}}\mathcal{L}_{\text{dec}} + \lambda_{\text{stability}}\mathcal{L}_{\text{stability}}, \tag{19}$$

where $\mathcal{L}_{\text{CE}}$ denotes the cross-entropy loss for optimizing discrimination on the current task, λ_{dec} controls the strength of memory separation, and $\lambda_{\text{stability}}$ balances semantic memory stability. During optimization, only the semantic evolution memory, current episodic expert memory, and expert routing coefficients are updated. The pretrained CLIP backbone and historical expert memories remain frozen. Through the joint optimization of adaptation, separation, and stabilization objectives, our framework achieves a continual knowledge evolution process that simultaneously supports: **(i) lifelong semantic accumulation**, **(ii) task-specific expertise preservation**, and **(iii) efficient replay-free memory reuse**.

4 Experiments

We evaluate DualMem-Evo across MTIL (full-shot, few-shot, and IFMS-MTIL) and CIL settings. MTIL reports Transfer, Average, and Last for generalization, overall performance, and retention, while CIL uses Average and Last all-seen-class accuracy. Details are in Appendix A.

4.1 Main Results

Full-shot MTIL. On Order-I, DualMem-Evo matches the highest reported Transfer of 70.4 and achieves the best Average and Last of 79.6 and 89.0, exceeding DPW by 0.3 and 0.7 points, respectively (Table 1). It obtains the best or tied-best Average and Last on four datasets each, indicating benefits across multiple tasks. Under the substantially different Order-II sequence, DualMem-Evo leads all three aggregate metrics by 0.1, 0.4, and 0.5 points, respectively (Table 4), supporting its robustness across task orders.

Few-shot MTIL. Under the 5-shot protocol, DualMem-Evo ties the best Transfer score of 70.3 while improving Average and Last by 0.2 and 0.5 points (Table 5). Under the distinct 16-shot CoOp protocol, it improves Average and Last by 0.4 and 0.5 points, respectively, with only a 0.2-point Transfer trade-off (Table 6). These results show that the transfer–retention balance remains effective under limited supervision.

Table 3: Ablation study of different components on MTIL Order-I. Δ denotes the change from the unregulated dual-memory baseline. Best results are marked in **bold**.

PGEEM	$\mathcal{L}_{\text{dec}}$	$\mathcal{L}_{\text{stability}}$	Transfer	Δ	Avg.	Δ	Last	Δ
			69.92	-	77.81	-	86.60	-
	✓	✓	70.15	+0.23	79.36	+1.55	88.78	+2.18
✓		✓	69.89	-0.03	79.28	+1.47	88.81	+2.21
✓	✓		70.33	+0.41	77.87	+0.06	86.11	-0.49
✓	✓	✓	**70.37**	**+0.45**	**79.59**	**+1.78**	**88.95**	**+2.35**

IFMS-MTIL. When identifier-based module lookup is prohibited, DualMem-Evo exceeds the strongest compatible baseline by 0.3, 0.4, and 0.3 points on Transfer, Average, and Last, respectively (Table 7). This result demonstrates effective instance-wise memory selection without relying on task or dataset identifiers. Figure 3 summarizes the gains across MTIL settings, while complete comparisons and per-session accuracy matrices are provided in Appendix C.

CIL. Beyond MTIL, DualMem-Evo ranks first in both Average and Last accuracy in five of the six CIL configurations on CIFAR-100 and TinyImageNet, and remains second on the fine-grained 50-step CIFAR-100 protocol (Tables 8 and 9). These results demonstrate that the dual-memory architecture generalizes beyond MTIL without historical-image replay or CIL-specific changes to its core training architecture.

4.2 Ablation Study

Tables 2 and 3 show that neither a single memory nor an unregulated dual-memory architecture achieves a satisfactory transfer–retention balance. Episodic memory supplies most of the retention capability, whereas semantic memory alone remains vulnerable to continual interference. Functional separation principally protects Transfer, while removing semantic stability regulation reduces Last by 2.84 points relative to the complete model. PGEEM further raises all three metrics when added to the regulated dual-memory model. Their combination yields the strongest and most balanced result.

5 Conclusion

We propose DualMem-Evo, a dual-memory continual adaptation framework for vision-language models. By combining semantic evolution memory, episodic expert memory, and stability-aware knowledge regulation, our method enables effective knowledge transfer while reducing forgetting.

AI USE STATEMENT

Generative AI tools were used for language editing. All technical claims, implementations, experiments, and reported numbers were verified by the authors.

REFERENCES

Elahe Arani, Fahad Sarfraz, and Bahram Zonooz. Learning fast, learning slow: A general continual learning method based on complementary learning system. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=uxxFrDwrE7Y.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101: Mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Arslan Chaudhry, Albert Gordo, Puneet Dokania, Philip Torr, and David Lopez-Paz. Using hindsight to anchor past knowledge in continual learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 6993–7001, 2021.

Yanxi Chen, Xiuxing Li, Yuyang Han, Zhuo Wang, Qing Li, Ziyu Li, Xiang Li, Chen Wei, and Xia Wu. HippoTune: A hippocampal associative loop-inspired fine-tuning method for continual learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8a53dd81165df5cbca5a7b4e12b616cc-Abstract-Conference.html.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

Li Deng. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 2012.

Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. 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.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2004.

Hao Fu, Hanbin Zhao, Jiahua Dong, Henghui Ding, Chao Zhang, and Hui Qian. IAP: Improving continual learning of vision-language models via instance-aware prompting. *arXiv preprint arXiv:2503.20612*, 2025.

Haoyuan Gao, Zicong Zhang, Yuqi Wei, Linglan Zhao, Guilin Li, Yexin Li, Bo Wang, Linghe Kong, and Weiran Huang. Enhanced continual learning of vision-language models with model fusion. In *International Conference on Learning Representations*, volume 2026, pp. 113856–113880, 2026.

Jiangpeng He, Zhihao Duan, and Fengqing Zhu. Cl-lora: Continual low-rank adaptation for rehearsal-free class-incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 30534–30544. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02843. URL https://openaccess.thecvf.com/content/CVPR2025/html/He_CL-LoRA_Continual_Low-Rank_Adaptation_for_Rehearsal-Free_Class-Incremental_Learning_CVPR_2025_paper.html.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019.

Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 831–839, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Hyeonseo Jang, Hyuk Kwon, and Kibok Lee. Enhancing continual learning of vision-language models via dynamic prefix weighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18042–18052, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Jang_Enhancing_Continual_Learning_of_Vision-Language_Models_via_Dynamic_Prefix_Weighting_CVPR_2026_paper.html.

Yujin Jo and Taesup Kim. Memory-free continual learning with null space adaptation for zero-shot vision-language models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ad217e0c7fecc71bdf48660ad6714b07-Abstract-Conference.html.

Haeyong Kang, Rusty John Lloyd Mina, Sultan Rizky Hikmawan Madjid, Jaehong Yoon, Mark Hasegawa-Johnson, Sung Ju Hwang, and Chang D Yoo. Forget-free continual learning with winning subnetworks. In *International Conference on Machine Learning*, pp. 10734–10750. PMLR, 2022.

Kushal Kapoor, Wyatt Mackey, Yiannis Aloimonos, and Xiaomin Lin. HiCL: Hippocampal-inspired continual learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 22518–22526, 2026. doi: 10.1609/aaai.v40i27.39411. URL https://ojs.aaai.org/index.php/AAAI/article/view/39411.

Zixuan Ke, Bing Liu, Nianzu Ma, Hu Xu, and Lei Shu. Achieving forgetting prevention and knowledge transfer in continual learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2426–2435, 2022.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Ya Le and Xuan Yang. Tiny ImageNet visual recognition challenge. Technical report, Stanford University, 2015.

Jindong Li, Yongguang Li, Yali Fu, Jiahong Liu, Yixin Liu, Menglin Yang, and Irwin King. Clip-powered domain generalization and domain adaptation: A comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Z. Li and D. Hoiem. Learning without forgetting. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018.

Mao-Lin Luo, Zi-Hao Zhou, Yi-Lin Zhang, Yuanyu Wan, Min-Ling Zhang, and Tong Wei. KeepLoRA: Continual learning with residual gradient adaptation. In *International Conference on Learning Representations*, 2026.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457, 1995.

Xing Nie, Shixiong Xu, Xiyan Liu, Gaofeng Meng, Chunlei Huo, and Shiming Xiang. Bilateral memory consolidation for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16026–16035, 2023.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.

G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

Quang Pham, Chenghao Liu, and Steven Hoi. Dualnet: Continual learning, fast and slow. *Advances in Neural Information Processing Systems*, 34, 2021.

Zihuan Qiu, Yi Xu, Chiyuan He, Fanman Meng, Linfeng Xu, Qingbo Wu, and Hongliang Li. Mingle: Mixture of null-space gated low-rank experts for test-time continual model merging. *Advances in Neural Information Processing Systems*, 38:143841–143878, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. iCaRL: Incremental classifier and representation learning. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2001–2010, 2017.

Qianqian Shi, Faqiang Liu, Hongyi Li, Guangyu Li, Luping Shi, and Rong Zhao. Hybrid neural networks for continual learning inspired by corticohippocampal circuits. *Nature Communications*, 16:1272, 2025.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. 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.

Xiang Song, Yuhang He, Lin Peng, and Yihong Gong. Multi-task unified domain incremental learning with domain difference adapters. *IEEE Transactions on Image Processing*, 35:3893–3908, 2026.

Longxiang Tang, Zhuotao Tian, Kai Li, Chunming He, Hantao Zhou, Hengshuang Zhao, Xiu Li, and Jiaya Jia. Mind the interference: Retaining pre-trained knowledge in parameter efficient continual learning of vision-language models. In *European Conference on Computer Vision*, pp. 346–365, 2024.

Huiyi Wang, Haodong Lu, Lina Yao, and Dong Gong. Self-expansion of pre-trained models with mixture of adapters for continual learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10087–10098. IEEE, 2025.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European conference on computer vision*, pp. 631–648. Springer, 2022a.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 139–149, 2022b.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7959–7971, 2022.

Bin Wu, Wuxuan Shi, Jinqiao Wang, and Mang Ye. Synthetic data is an elegant GIFT for continual vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2813–2823, 2025.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

Shipeng Yan, Jiangwei Xie, and Xuming He. DER: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3014–3023, 2021.

Fei Ye and Adrian G Bors. Lifelong variational autoencoder via online adversarial expansion strategy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10909–10917, 2023.

Fei Ye and Adrian G Bors. Online task-free continual learning via expansible vision transformer. *Pattern Recognition*, 169:111730, 2026.

Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23219–23230, 2024.

Zihao Zhao, Yuxiao Liu, Han Wu, Mei Wang, Yonghao Li, Sheng Wang, Lin Teng, Disheng Liu, Zhiming Cui, Qian Wang, et al. Clip in medical imaging: A survey. *Medical Image Analysis*, 102:103551, 2025.

Zangwei Zheng, Mingyuan Ma, Kai Wang, Ziheng Qin, Xiangyu Yue, and Yang You. Preventing zero-shot transfer degradation in continual learning of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19125–19136, 2023.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5871–5880, 2021.

A Detailed Experimental Setup

A.1 Benchmarks and Evaluation Protocols

MTIL and full-shot orders. We follow the Multi-domain Task Incremental Learning (MTIL) benchmark introduced by ZSCL Zheng et al. (2023). It contains 11 image-classification datasets and 1,201 classes: Aircraft Maji et al. (2013), Caltech101 Fei-Fei et al. (2004), CIFAR100 Krizhevsky & Hinton (2009), DTD Cimpoi et al. (2014), EuroSAT Helber et al. (2019), Flowers102 Nilsback & Zisserman (2008), Food101 Bossard et al. (2014), MNIST Deng (2012), Oxford-IIIT Pet Parkhi et al. (2012), Stanford Cars Krause et al. (2013), and SUN397 Xiao et al. (2010). Each dataset constitutes one incremental session. Only the current session's training data are accessible, and no historical images are retained or replayed. After every session, the model is evaluated on the native test split of every dataset, including both learned and future tasks. For Order-I, training proceeds through Aircraft, Caltech101, CIFAR100, DTD, EuroSAT, Flowers102, Food101, MNIST, Oxford-IIIT Pet, Stanford Cars, and SUN397. For Order-II, we use the task-order variation defined in the original MTIL benchmark Zheng et al. (2023) and subsequently adopted by KeepLoRA Luo et al. (2026): Stanford Cars, Food101, MNIST, Oxford-IIIT Pet, Flowers102, SUN397, Aircraft, Caltech101, DTD, EuroSAT, and CIFAR100. Both full-shot orders use 1,000 optimization iterations per session, following the established MTIL protocol Zheng et al. (2023); Yu et al. (2024); Gao et al. (2026).

5-shot MTIL-FS. The 5-shot protocol is a data-scarce MTIL variant adopted by MoE-Adapters Yu et al. (2024) and ConDU Gao et al. (2026). It retains all 11 datasets and the Order-I training sequence, but restricts the training set to exactly five images per class while leaving the native test sets unchanged. Each session is trained for 500 iterations, and the same no-replay constraint and session-wise evaluation procedure as full-shot MTIL are used.

16-shot MTIL-FS. We strictly follow the modified 16-shot protocol introduced by DIKI Tang et al. (2024) and used by IAP Fu et al. (2025). Relative to MTIL Order-I, this protocol removes EuroSAT, MNIST, and Oxford-IIIT Pet and retains eight datasets in the following order: Aircraft, Caltech101, CIFAR100, DTD, Flowers102, Food101, Stanford Cars, and SUN397. Exactly 16 training images per class are selected using the official CoOp partitions Zhou et al. (2022); because CIFAR100 is not included in the CoOp repository, its subset is sampled with the protocol seed of 42, as in DIKI Tang et al. (2024). Every session is trained for 10 epochs. The batch sizes for training and evaluation are 128 and 256, respectively Tang et al. (2024); Fu et al. (2025).

IFMS-MTIL. Identifier-Free Module-Selection MTIL (IFMS-MTIL) retains the datasets, training order, and session-wise evaluation matrix of full-shot MTIL Order-I, while imposing an identifier-free module-selection constraint at inference. For each evaluation dataset, its target label space $\mathcal{C}_j$ and text templates are provided to construct the corresponding zero-shot classifier. However, the task identity, dataset name, and learned/future status cannot be used to select a task-specific expert, adapter, prompt, classifier head, or checkpoint; nor may $\mathcal{C}_j$ be treated as an exact key for retrieving such a module. A method must therefore use either a shared model or input-driven module selection. The protocol does not additionally constrain whether identifier-free routing is performed per batch or per sample. DualMem-Evo adopts the latter strategy, making each prediction independent of the composition of the test batch.

Class-incremental learning. We follow the CIL protocols adopted by MoE-Adapters Yu et al. (2024) and GIFT Wu et al. (2025) on CIFAR-100 Krizhevsky & Hinton (2009) and TinyImageNet Le & Yang (2015). For CIFAR-100, the 100 classes are partitioned into 10, 20, or 50 sessions, corresponding to 10, 5, or 2 new classes per session. For TinyImageNet, the first session contains 100 base classes, after which the remaining 100 classes are divided into 5, 10, or 20 increments containing 20, 10, or 5 classes, respectively. Evaluation is performed over all classes observed so far without access to the task identity. DualMem-Evo is trained at each stage using only images from the newly introduced classes and retains no historical training images, whereas baseline results follow their respective original protocols.

A.2 Evaluation Metrics

Let $A \in \mathbb{R}^{T\times T}$ denote the accuracy matrix, where $A_{i,j}$ is the top-1 accuracy on dataset j after completing training session i. For each dataset $j > 0$, its pre-learning transfer score is the average

of the cells above its diagonal. The three reported metrics Zheng et al. (2023) are

$$\text{Transfer} = \frac{1}{T-1}\sum_{j=1}^{T-1}\left(\frac{1}{j}\sum_{i=0}^{j-1}A_{i,j}\right), \tag{20}$$

$$\text{Average} = \frac{1}{T^2}\sum_{i=0}^{T-1}\sum_{j=0}^{T-1}A_{i,j}, \tag{21}$$

$$\text{Last} = \frac{1}{T}\sum_{j=0}^{T-1}A_{T-1,j}. \tag{22}$$

The first dataset has no pre-learning session and is excluded from Transfer. Transfer measures retained zero-shot generalization before a task is learned, Average summarizes performance over the complete learning trajectory, and Last measures final retention after the last session. We use $T = 11$ for full-shot, 5-shot, and IFMS-MTIL, and $T = 8$ for 16-shot MTIL-FS. All results are reported as percentages.

For CIL, let a_s denote the top-1 accuracy over all classes observed after training stage s, and let S be the total number of training stages. We report

$$\text{Avg.} = \frac{1}{S}\sum_{s=0}^{S-1}a_s, \tag{23}$$

$$\text{Last} = a_{S-1}. \tag{24}$$

The initial stage is included in Avg. For CIFAR-100, $S \in \{10, 20, 50\}$. For TinyImageNet, $S \in \{6, 11, 21\}$ because its 100-class base session is followed by 5, 10, or 20 incremental stages.

A.3 Implementation Details

Common training protocol. All MTIL experiments use CLIP ViT-B/16 Radford et al. (2021) and the dataset splits, class vocabularies, text templates, and session-wise matrix evaluation adopted by prior MTIL studies Zheng et al. (2023); Yu et al. (2024); Tang et al. (2024); Fu et al. (2025); Gao et al. (2026); Luo et al. (2026). The pretrained CLIP backbone and historical task memories are not optimized by data from later sessions, and no historical training images are replayed. Full-shot Order-I, Order-II, and IFMS-MTIL use 1,000 iterations per session; 5-shot MTIL-FS uses 500 iterations per session; and 16-shot MTIL-FS uses 10 epochs per session with its protocol-matched batch sizes. Method-specific optimization settings are held fixed within each protocol and do not alter the data splits, task order, training budget, or evaluation procedure.

Inference and routing. Under the standard MTIL protocol used for Order-I, Order-II, 5-shot, and 16-shot experiments, a learned dataset is evaluated using the current semantic memory together with its corresponding episodic expert. For a future dataset, the available episodic experts and frozen CLIP are evaluated over the provided target label space, and their predictions are selected and fused without using target labels or future training images. IFMS-MTIL applies this identifier-free candidate evaluation to every matrix cell. Specifically, each available episodic expert is combined with the current semantic memory, while frozen CLIP serves as an additional candidate. Their predictions over the known target label space are scored using normalized prediction confidence and image–text consistency, after which the top-ranked candidates are fused with temperature-scaled softmax weights. DualMem-Evo performs this procedure independently for each test sample. For CIL, the classifier covers all classes observed so far; candidate episodic experts are retrieved using frozen-CLIP class prototypes, re-ranked according to their owned-class margins, and fused with the semantic and frozen-CLIP anchor. No test labels or task identities are used for routing.

A.4 Comparison Methods and Result Provenance

Our comparison set contains zero-shot CLIP and independently adapted Full Fine-tune and LoRA Fine-tune, together with LwF Li & Hoiem (2018), WiSE-FT Wortsman et al. (2022), ZSCL Zheng et al. (2023), L2P Wang et al. (2022b), DualPrompt Wang et al. (2022a), MoE-Adapters Yu et al. (2024), DIKI Tang et al. (2024), IAP Fu et al. (2025), GIFT Wu et al. (2025), ConDU

(LoRA) Gao et al. (2026), DPW Jang et al. (2026), and KeepLoRA and KeepLoRA+ Luo et al. (2026). KeepLoRA+ denotes the prototype-augmented extension reported in the KeepLoRA paper. LoRA Fine-tune is our own baseline, obtained by fine-tuning CLIP with LoRA under the data split and training budget of each corresponding protocol. Unless otherwise stated, the remaining baseline values are transcribed from the original paper or from a later table with an identical experimental protocol. The Order-II ConDU (LoRA) result is our reproduction in the same data environment with a carefully tuned configuration. For IFMS-MTIL, we reuse the reported Order-I results of methods whose official inference implementations already satisfy identifier-free module selection. MoE-Adapters, DIKI, and IAP retain their original batch-wise routing strategies, whereas KeepLoRA uses a single continually merged model. For CIL, we additionally compare with iCaRL Rebuffi et al. (2017), UCIR Hou et al. (2019), DER Yan et al. (2021), PASS Zhu et al. (2021), and DyTox/DyTox+ Douillard et al. (2022), following the result provenance used by GIFT Wu et al. (2025).

B Additional Theoretical Analysis

Theoretical analysis of low-rank memory consolidation. The proposed consolidation strategy is based on the observation that the knowledge evolution trajectory induced by expert collaboration lies in a compact low-dimensional subspace. We provide a theoretical justification from the perspective of low-rank approximation.

Let the evolved task-side adaptation before consolidation be denoted as:

$$\Delta \mathbf{W}^i_{\text{raw}} = \mathbf{A}^i_t \mathbf{B}^i_t + \sum_{k \in \mathcal{K}_t} w^i_k \mathbf{A}^i_k \mathbf{B}^i_k. \tag{25}$$

Although $\Delta \mathbf{W}^i_{\text{raw}}$ contains information from multiple experts, each expert update is constrained to a low-dimensional LoRA subspace. Therefore, the accumulated update can be interpreted as a mixture of several low-rank semantic directions.

Let the singular value decomposition of $\Delta \mathbf{W}^i_{\text{raw}}$ be:

$$\Delta \mathbf{W}^i_{\text{raw}} = \sum_{m=1}^{r} \sigma_m \mathbf{u}_m \mathbf{v}_m^\top, \tag{26}$$

where r is the rank of the merged update and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$ are singular values.

We approximate the merged memory using the top-r_s singular components:

$$\Delta \mathbf{W}^i_s = \sum_{m=1}^{r_s} \sigma_m \mathbf{u}_m \mathbf{v}_m^\top. \tag{27}$$

According to the Eckart–Young–Mirsky theorem, the truncated solution $\Delta \mathbf{W}^i_s$ is the optimal rank-r_s approximation under both Frobenius norm and spectral norm. Therefore, for any rank-r_s matrix $\mathbf{M}$:

$$\|\Delta \mathbf{W}^i_{\text{raw}} - \Delta \mathbf{W}^i_s\|_F \leq \|\Delta \mathbf{W}^i_{\text{raw}} - \mathbf{M}\|_F. \tag{28}$$

Moreover, the approximation error is explicitly bounded by the discarded singular spectrum:

$$\|\Delta \mathbf{W}^i_{\text{raw}} - \Delta \mathbf{W}^i_s\|_F^2 = \sum_{m=r_s+1}^{r} \sigma_m^2. \tag{29}$$

Equation 29 indicates that consolidation only removes low-energy directions associated with small singular values. Since singular values measure the contribution of each adaptation direction, the preserved subspace contains the dominant knowledge variations produced by expert interaction, whereas insignificant or conflicting directions are naturally filtered out.

Proposition 1 (Knowledge-preserving consolidation). Assume that the output perturbation induced by the merged LoRA update is Lipschitz continuous with respect to the adaptation matrix, i.e., for any two updates $\mathbf{W}_1$ and $\mathbf{W}_2$:

$$\|f(\mathbf{x}; \mathbf{W}_1) - f(\mathbf{x}; \mathbf{W}_2)\|_2 \leq L \|\mathbf{W}_1 - \mathbf{W}_2\|_F, \tag{30}$$

where L is the local Lipschitz constant of the frozen CLIP backbone.

Then the representation deviation introduced by consolidation satisfies:

$$\|f(\mathbf{x}; \Delta\mathbf{W}^i_{\text{raw}}) - f(\mathbf{x}; \Delta\mathbf{W}^i_s)\|_2 \leq L\sqrt{\sum_{m=r_s+1}^{r} \sigma_m^2}. \tag{31}$$

Proof. By the Lipschitz continuity assumption:

$$\begin{aligned} &\|f(\mathbf{x}; \Delta\mathbf{W}^i_{\text{raw}}) - f(\mathbf{x}; \Delta\mathbf{W}^i_s)\|_2 \\ &\leq L\|\Delta\mathbf{W}^i_{\text{raw}} - \Delta\mathbf{W}^i_s\|_F. \end{aligned} \tag{32}$$

Using the optimal truncated SVD approximation property:

$$\|\Delta\mathbf{W}^i_{\text{raw}} - \Delta\mathbf{W}^i_s\|_F = \sqrt{\sum_{m=r_s+1}^{r} \sigma_m^2}, \tag{33}$$

which directly gives the conclusion. □

The proposition demonstrates that when the merged expert trajectory has a rapidly decaying singular spectrum, the compressed memory can preserve almost identical functional behavior with significantly fewer parameters.

Proposition 2 (Memory complexity reduction). Assume that N historical experts are maintained, each represented by a rank-r_t LoRA adaptation. Storing all experts requires:

$$\mathcal{O}(Nr_t(d_{\text{in}} + d_{\text{out}})) \tag{34}$$

parameters for each layer.

After consolidation, the merged expert is represented by a fixed rank r_s approximation, requiring:

$$\mathcal{O}(r_s(d_{\text{in}} + d_{\text{out}})) \tag{35}$$

parameters.

Therefore, the compression ratio is:

$$\rho = \frac{Nr_t}{r_s}. \tag{36}$$

When $Nr_t \gg r_s$, consolidation achieves substantial memory reduction while preserving the dominant adaptation subspace.

Discussion. The above analysis reveals three properties of the proposed consolidation mechanism. First, the Eckart–Young–Mirsky optimality ensures that the compressed expert preserves the maximum possible adaptation energy under a fixed rank budget. Second, the singular spectrum provides an implicit importance criterion, allowing the model to retain transferable knowledge directions and discard noisy or conflicting updates. Third, the fixed-rank representation bounds the memory footprint of each consolidated expert, enabling scalable lifelong adaptation under the strict no-replay constraint.

C Supplementary Experimental Results

C.1 Complete Comparisons

This section provides the complete comparisons across the evaluated MTIL and CIL protocols. In addition to the aggregate metrics, the MTIL tables report dataset-wise results, allowing us to examine whether the overall gains arise from a few favorable tasks or remain consistent across heterogeneous visual domains.

Full-shot MTIL Order-I. Table 1 reports the standard Order-I comparison. DualMem-Evo matches the highest Transfer score of 70.4 and improves Average and Last over DPW by 0.3 and 0.7 points,

respectively. The improvement is distributed across the sequence: DualMem-Evo obtains the best or tied-best Average on CIFAR100, DTD, EuroSAT, and Cars, and the best or tied-best Last on CIFAR100, DTD, MNIST, and Cars. These results indicate that its aggregate advantage is supported by gains on multiple domains rather than by a single favorable dataset. In particular, the stronger Average and Last scores, without sacrificing Transfer, are consistent with the intended balance between continual adaptation and retention.

Table 4: Comparison with SOTA methods on the MTIL benchmark (Order-II) using the "Transfer", "Average", and "Last" scores (%). We label the best methods with **bold** styles. The lines with colored background correspond to the results of our proposed method.

	Method	Cars	Food	MNIST	OxfordPet	Flowers	SUN397	Aircraft	Caltech101	DTD	EuroSAT	CIFAR100	Average
CLIP	Zero-shot	64.7	88.5	59.4	89.0	71.0	65.2	24.3	88.4	44.6	54.9	68.2	65.3
	Full Fine-tune	89.6	92.7	99.6	94.7	97.5	81.8	62.0	95.1	79.5	98.9	89.6	89.2
	LoRA Fine-tune	90.9	92.4	99.6	94.3	97.8	81.2	62.2	97.2	79.7	98.9	89.3	89.4
Transfer	LwF	-	87.8	58.5	71.9	46.6	57.3	12.8	81.4	34.5	34.5	46.8	53.2
	WiSE-FT	-	87.2	57.6	67.0	45.0	54.0	12.9	78.6	35.5	28.4	44.3	51.1
	ZSCL	-	88.3	57.5	84.7	68.1	64.8	21.1	88.2	45.3	55.2	68.2	64.1
	MoE-Adapters	-	88.8	59.5	89.1	69.9	64.4	18.1	86.9	43.7	54.6	68.2	64.3
	DIKI	-	85.8	59.8	89.1	71.8	62.6	24.3	93.3	42.7	46.8	67.8	64.4
	IAP	-	85.7	59.4	89.1	71.3	62.7	24.4	**94.0**	43.8	49.0	68.6	64.9
	GIFT	-	88.3	**63.4**	88.1	70.8	**67.7**	22.8	90.4	**46.7**	51.8	68.8	65.9
	ConDU(LoRA)	-	88.3	59.4	89.0	70.7	67.5	23.7	88.5	45.5	**55.8**	**71.0**	65.9
	DPW	-	85.9	63.3	89.4	72.2	64.2	**24.6**	**94.0**	44.6	49.4	69.1	65.7
	KeepLoRA+	-	**89.1**	58.1	**90.7**	**72.4**	65.4	24.0	88.9	44.0	52.7	70.2	65.5
	DualMem-Evo	-	88.6	61.2	89.1	71.2	66.1	24.2	88.4	46.1	54.7	70.7	**66.0**
Average	LwF	49.0	77.0	92.1	85.9	66.5	67.2	20.9	84.7	44.6	45.5	50.5	62.2
	WiSE-FT	52.6	79.3	91.9	83.9	63.4	65.2	23.3	83.7	45.4	40.0	48.2	61.5
	ZSCL	81.7	91.3	91.9	91.0	82.9	72.5	33.6	89.7	53.3	62.8	69.9	74.6
	MoE-Adapters	84.9	89.9	89.3	91.4	86.2	72.2	33.4	89.4	53.3	61.4	69.9	74.7
	DIKI	81.9	88.9	92.1	92.8	87.7	70.3	34.3	94.2	51.5	56.1	69.5	74.5
	IAP	82.5	89.2	92.3	93.2	88.0	70.4	34.3	94.4	52.4	57.9	70.2	75.1
	GIFT	83.2	90.8	92.6	92.8	85.8	74.1	36.0	92.1	54.7	60.0	70.4	75.7
	ConDU(LoRA)	88.1	92.1	91.1	91.6	81.4	74.5	38.9	91.3	54.0	**63.3**	**72.5**	76.3
	DPW	86.6	88.9	**92.8**	93.6	**88.5**	71.5	**40.4**	**94.9**	54.2	58.3	70.9	76.4
	KeepLoRA+	88.0	**92.4**	91.9	**93.9**	87.4	**75.2**	39.2	92.0	52.8	60.9	71.8	76.9
	DualMem-Evo	**90.3**	92.3	92.5	93.2	86.8	74.4	39.8	91.1	**55.3**	62.6	72.4	**77.3**
Last	LwF	34.6	69.6	99.3	88.7	61.1	72.5	32.5	88.1	65.6	90.9	87.9	71.9
	WiSE-FT	35.6	76.9	99.5	89.1	62.1	71.8	27.8	90.8	67.0	85.6	87.6	72.2
	ZSCL	78.2	91.1	97.6	92.5	87.4	78.2	45.0	92.3	72.7	96.2	86.3	83.4
	MoE-Adapters	84.1	88.5	94.0	91.8	94.1	77.8	50.4	93.3	77.1	87.7	86.6	84.1
	DIKI	81.9	89.2	99.4	94.3	96.8	76.7	46.3	95.9	74.8	98.3	86.6	85.5
	IAP	82.5	88.6	99.4	94.9	97.7	76.9	46.1	96.1	74.7	98.0	86.6	85.9
	GIFT	81.0	90.2	98.6	94.0	91.5	78.6	51.7	94.6	75.6	95.4	86.6	85.3
	ConDU(LoRA)	85.7	92.4	97.6	92.2	85.5	79.9	56.2	95.9	75.8	96.4	87.6	85.9
	DPW	86.6	89.2	**99.6**	**95.1**	**97.9**	78.4	**59.4**	96.4	79.8	**98.4**	87.9	88.1
	KeepLoRA+	87.4	92.5	99.3	95.0	96.0	**83.2**	56.9	**97.5**	76.9	98.0	88.0	88.2
	DualMem-Evo	**90.4**	**92.6**	99.5	94.7	95.7	81.3	58.5	96.0	**80.0**	98.1	**89.0**	**88.7**

Full-shot MTIL Order-II. The substantially reordered sequence in Table 4 changes which tasks are encountered early and which historical memories are available to later sessions. Nevertheless, DualMem-Evo achieves the best Transfer, Average, and Last scores of 66.0, 77.3, and 88.7, exceeding the strongest preceding results by 0.1, 0.4, and 0.5 points. The dataset-wise results show particularly strong Average and Last performance on Cars and DTD, while the final CIFAR100 accuracy also remains the highest. Together with Order-I, this result shows that the benefit of the proposed memory evolution is not tied to one specific task ordering.

5-shot MTIL-FS. Under the substantially reduced supervision of Table 5, DualMem-Evo ties the highest Transfer score of 70.3 and improves the best prior Average and Last from 72.7 and 77.4 to 72.9 and 77.9. Its gains are most visible on DTD, EuroSAT, MNIST, and Cars, where it achieves the best Average, and it additionally obtains the best final accuracy on Aircraft. Although individual competitors remain stronger on several datasets, DualMem-Evo provides the most favorable aggregate balance, indicating that the dual-memory organization remains effective when each class provides only five training examples.

Table 5: Comparison with SOTA methods on the 5-shot MTIL-FS benchmark using the "Transfer", "Average", and "Last" scores (%). We label the best methods with **bold** styles. The lines with colored background correspond to the results of our proposed method.

	Method	Aircraft	Caltech101	CIFAR100	DTD	EuroSAT	Flowers	Food	MNIST	OxfordPet	Cars	SUN397	Average
CLIP	Zero-shot	24.3	88.4	68.2	44.6	54.9	71.0	88.5	59.4	89.0	64.7	65.2	65.3
	Full Fine-tune	30.6	93.5	76.8	65.1	91.7	92.9	83.3	96.6	84.9	65.4	71.3	77.5
	LoRA Fine-tune	38.6	96.0	75.9	65.5	88.9	94.4	88.7	97.6	90.6	75.1	72.9	80.4
Transfer	LwF	-	72.1	49.2	35.9	44.5	41.1	66.6	50.5	69.0	19.0	51.7	50.0
	WiSE-FT	-	77.6	60.0	41.3	39.4	53.0	76.6	58.1	75.5	37.3	58.2	57.7
	ZSCL	-	84.0	68.1	44.8	46.8	63.6	84.9	61.4	81.4	55.5	62.2	65.3
	MoE-Adapters	-	87.9	68.2	44.1	48.1	64.7	88.8	**69.0**	**89.1**	64.5	65.1	68.9
	ConDU(LoRA)	-	88.1	**68.5**	**45.6**	**56.4**	**71.2**	**89.0**	64.0	88.8	**64.9**	**66.4**	**70.3**
	DualMem-Evo	-	**88.3**	67.8	**45.6**	55.8	70.6	88.7	66.9	88.2	64.7	66.2	**70.3**
Average	LwF	23.5	77.4	43.5	41.7	43.5	52.2	54.6	63.4	68.0	21.3	52.6	49.2
	WiSE-FT	32.0	87.7	61.0	55.8	68.1	69.3	76.8	71.5	77.6	42.0	59.3	63.7
	ZSCL	28.2	88.6	66.5	53.5	56.3	73.4	83.1	56.4	82.4	57.5	62.9	64.4
	MoE-Adapters	30.0	89.6	73.9	58.7	69.3	79.3	**88.1**	76.5	89.1	65.3	65.8	71.4
	ConDU(LoRA)	**32.4**	**92.1**	**75.4**	58.8	75.1	**82.9**	87.3	74.0	**89.3**	65.1	**67.0**	72.7
	DualMem-Evo	32.3	91.5	74.1	**61.0**	**77.1**	81.6	85.5	**77.4**	87.9	**66.2**	66.8	**72.9**
Last	LwF	22.1	58.2	17.9	32.1	28.1	66.7	46.0	84.3	64.1	31.5	60.1	46.5
	WiSE-FT	30.8	88.9	59.6	60.3	80.9	81.7	77.1	94.9	83.2	62.8	70.0	71.9
	ZSCL	26.8	88.5	63.7	55.7	60.2	82.1	82.6	58.6	85.9	66.7	70.4	67.4
	MoE-Adapters	30.1	89.3	74.9	64.0	82.3	89.4	**87.1**	89.0	89.1	69.5	**72.5**	76.1
	ConDU(LoRA)	31.8	**92.4**	**76.7**	63.4	86.8	**91.8**	85.6	93.9	**90.3**	68.1	70.9	77.4
	DualMem-Evo	**32.6**	91.9	75.6	**66.7**	**89.3**	91.0	81.5	**95.8**	87.3	**72.8**	72.1	**77.9**

Table 6: Comparison with SOTA methods on the 16-shot MTIL-FS benchmark using the "Transfer", "Average", and "Last" scores (%). We label the best methods with **bold** styles. The lines with colored background correspond to the results of our proposed method.

	Method	Aircraft	Caltech101	CIFAR100	DTD	Flowers	Food	Cars	SUN397	Average
CLIP	Zero-shot	24.8	92.9	68.4	43.8	71.4	85.8	65.8	62.6	64.4
	Full Fine-tune	62.0	96.2	89.6	79.5	97.5	92.7	89.6	81.8	86.1
	LoRA Fine-tune	52.9	97.4	81.2	72.7	98.4	86.9	87.6	76.6	81.7
Transfer	ZSCL	-	87.3	67.7	45.4	67.8	**86.6**	59.7	63.4	68.3
	L2P	-	66.7	54.3	30.6	47.3	71.5	54.6	52.4	53.9
	DualPrompt	-	78.8	64.4	32.0	51.7	77.5	49.4	51.3	57.9
	DIKI	-	92.7	68.8	44.1	70.0	86.2	65.1	65.5	70.3
	IAP	-	**93.2**	**68.9**	44.5	**71.4**	85.5	**66.1**	65.0	**70.9**
	DualMem-Evo	-	91.7	68.1	**46.6**	70.6	85.9	65.6	**66.2**	70.7
Average	ZSCL	33.5	90.5	74.7	58.5	79.7	**87.7**	64.8	64.8	69.3
	L2P	30.2	84.5	70.1	51.9	69.6	77.1	60.0	55.2	62.3
	DualPrompt	36.5	89.5	72.5	52.7	72.3	80.8	56.1	54.2	64.3
	DIKI	41.3	95.3	76.5	58.5	82.2	86.4	68.2	66.6	71.9
	IAP	42.5	94.8	77.6	59.3	**82.5**	86.5	69.6	65.4	72.5
	DualMem-Evo	**44.2**	**95.6**	**78.3**	**63.7**	77.9	85.8	**70.0**	**67.4**	**72.9**
Last	ZSCL	27.7	90.9	74.4	64.7	90.2	**89.2**	80.6	74.6	74.0
	L2P	30.2	87.1	75.4	64.7	91.9	86.4	76.1	74.7	73.3
	DualPrompt	36.5	91.0	75.1	65.1	92.9	86.2	76.2	74.2	74.7
	DIKI	41.3	95.6	79.0	67.3	94.4	86.8	77.6	74.4	77.1
	IAP	42.5	95.8	78.6	68.1	**95.3**	87.5	79.2	74.8	77.7
	DualMem-Evo	**44.2**	**96.2**	**81.8**	**73.9**	85.0	85.6	**83.2**	**75.9**	**78.2**

16-shot MTIL-FS. Table 6 considers the distinct eight-dataset CoOp protocol. DualMem-Evo obtains 70.7 Transfer, only 0.2 points below IAP, while improving Average and Last by 0.4 and 0.5 points. It achieves the highest Average on six of the eight datasets, with especially clear gains on Aircraft and DTD, and leads the final accuracy on the same six datasets. The remaining trade-off is concentrated on Flowers and Food rather than being broadly distributed across the sequence. Thus, the improvement in adaptation and retention is obtained with only a minor reduction in pre-learning transfer.

Table 7: Comparison with SOTA methods on the IFMS-MTIL benchmark using the "Transfer", "Average", and "Last" scores (%). We label the best methods with **bold** styles. The lines with colored background correspond to the results of our proposed method.

	Method	Aircraft	Caltech101	CIFAR100	DTD	EuroSAT	Flowers	Food	MNIST	OxfordPet	Cars	SUN397	Average
Transfer	ZSCL	-	86.0	67.4	45.4	50.4	69.1	87.6	61.8	86.8	60.1	66.8	68.1
	MoE-Adapters	-	87.9	68.2	44.4	49.9	**70.7**	88.7	59.7	89.1	64.5	65.5	68.9
	DIKI	-	92.9	69.0	43.2	48.2	67.4	85.2	63.0	87.9	63.8	66.2	68.7
	IAP	-	**93.0**	68.7	44.0	47.0	70.4	85.9	63.5	89.7	**66.2**	63.3	69.2
	GIFT	-	88.5	**69.8**	**46.0**	49.4	68.5	87.1	**69.9**	88.9	57.7	**67.7**	69.3
	KeepLoRA	-	84.6	68.7	45.9	54.3	70.1	87.7	64.8	**90.3**	59.5	64.1	69.0
	DualMem-Evo	-	87.2	67.8	45.7	**56.1**	69.5	**88.8**	59.4	88.8	65.0	67.4	**69.6**
Average	ZSCL	45.1	92.0	80.1	64.3	79.5	81.6	89.6	75.2	88.9	64.7	68.0	75.4
	MoE-Adapters	50.2	91.9	83.1	**69.4**	78.9	84.0	89.1	73.7	89.3	67.7	66.9	76.7
	DIKI	45.1	95.5	83.1	64.8	79.9	83.5	87.0	76.2	89.6	67.0	67.1	76.3
	IAP	45.9	**95.8**	83.3	66.5	79.5	**84.8**	87.5	76.6	91.0	**69.2**	64.5	76.8
	GIFT	51.9	93.9	81.4	67.7	80.3	82.8	89.3	**80.6**	90.3	63.1	**68.9**	77.3
	KeepLoRA	**55.6**	95.7	83.2	65.6	**82.2**	82.0	89.5	77.4	**91.5**	63.9	65.8	77.5
	DualMem-Evo	54.9	94.5	**83.8**	68.8	81.7	81.5	**90.3**	73.2	90.3	68.9	68.5	**77.9**
Last	ZSCL	40.6	92.2	81.3	70.5	94.8	90.5	91.9	98.7	93.9	85.3	80.2	83.6
	MoE-Adapters	49.8	92.2	86.1	**78.1**	95.7	94.3	89.5	98.1	89.9	81.6	80.0	85.0
	DIKI	45.2	95.7	86.3	72.9	98.0	**97.0**	89.2	99.4	94.2	81.6	76.6	85.1
	IAP	46.8	96.1	86.7	75.2	**98.1**	**97.0**	89.6	99.4	**94.7**	82.8	76.7	85.7
	GIFT	47.9	95.6	82.8	75.1	97.3	94.2	91.7	99.2	94.2	**87.0**	80.9	86.0
	KeepLoRA	53.2	**96.8**	85.7	71.4	**98.1**	90.8	91.4	**99.6**	94.5	83.1	**82.0**	86.1
	DualMem-Evo	**53.9**	95.2	**87.3**	77.0	95.6	91.5	**92.0**	97.4	94.2	86.5	79.8	**86.4**

Table 8: Class-incremental learning results (%) on CIFAR-100. "Avg." averages accuracy over all incremental stages, while "Last" reports the final-stage accuracy. Best results are in **bold**.

Method	10 steps		20 steps		50 steps	
	Avg.	Last	Avg.	Last	Avg.	Last
UCIR	58.66	43.39	58.17	40.63	56.86	37.09
DER	74.64	64.35	73.98	62.55	72.05	59.76
DyTox+	74.10	62.34	71.62	57.43	68.90	51.09
CLIP Zero-shot	74.47	65.92	75.20	65.74	75.67	65.94
LwF	65.86	48.04	60.64	40.56	47.69	32.90
iCaRL	79.35	70.97	73.32	64.55	71.28	59.07
ZSCL	82.15	73.65	80.39	69.58	79.92	67.36
MoE-Adapters	85.21	77.52	83.72	76.20	**83.60**	**75.24**
GIFT	85.11	77.70	82.11	73.73	80.81	71.29
DualMem-Evo	**86.60**	**78.39**	**85.27**	**76.61**	82.87	73.84

IFMS-MTIL. Table 7 evaluates methods without identifier-based module lookup. Under this restriction, DualMem-Evo reaches 69.6 Transfer, 77.9 Average, and 86.4 Last, improving the strongest compatible results by 0.3, 0.4, and 0.3 points. The method attains strong results across both future and learned tasks, including the highest Transfer on EuroSAT and Food and the highest final accuracy on Aircraft, CIFAR100, and Food. These results show that the proposed input-driven memory selection remains competitive without using task identities, dataset identifiers, or class vocabularies as exact keys for retrieving task-specific modules.

CIL on CIFAR-100. Table 8 evaluates increasingly fine-grained class increments. DualMem-Evo ranks first for both Avg. and Last under the 10-step and 20-step settings, improving the strongest preceding results by 1.39/0.69 and 1.55/0.41 points, respectively. Under the more demanding 50-step setting, where each session introduces only two classes, it ranks second to MoE-Adapters, with gaps of 0.73 in Avg. and 1.40 in Last. It nevertheless remains ahead of GIFT by 2.06 and 2.55 points. The results suggest that extremely small class increments constitute the most challenging CIL regime for the current architecture, while its advantage remains clear under moderate sequence granularities.

CIL on TinyImageNet. The advantage is more consistent on TinyImageNet, as shown in Table 9. DualMem-Evo ranks first in both metrics for all three protocols. Relative to the strongest preceding

Table 9: Class-incremental learning results (%) on TinyImageNet. "Avg." averages accuracy over all incremental stages, while "Last" reports the final-stage accuracy. Best results are in **bold**.

Method	5 steps		10 steps		20 steps	
	Avg.	Last	Avg.	Last	Avg.	Last
UCIR	50.30	39.42	48.58	37.29	42.84	30.85
PASS	49.54	41.64	47.19	39.27	42.01	32.93
DyTox	55.58	47.23	52.26	42.79	46.18	36.21
CLIP Zero-shot	69.62	65.30	69.55	65.59	69.49	65.30
LwF	60.97	48.77	57.60	44.00	54.79	42.26
iCaRL	77.02	70.39	73.48	65.97	69.65	64.68
ZSCL	80.27	73.57	78.61	71.62	77.18	68.30
MoE-Adapters	81.12	76.81	80.23	76.35	79.96	75.77
GIFT	81.16	77.04	80.20	75.51	79.32	74.87
DualMem-Evo	**82.96**	**78.10**	**82.50**	**77.47**	**81.83**	**76.57**

result, its Avg./Last gains are 1.80/1.06, 2.27/1.12, and 1.87/0.80 points for the 5-step, 10-step, and 20-step settings, respectively. It maintains the highest Average and Last accuracy across all three evaluated incremental settings. These results demonstrate that the dual-memory architecture transfers beyond multi-domain MTIL and remains effective in an all-seen-class CIL evaluation without historical-image replay.

C.2 Detailed Performance Matrices

Table 10: Accuracy of DualMem-Evo on the MTIL benchmark with order-I. Each row represents the performance on every dataset of the model trained after the corresponding task. Transfer, Average, and Last metrics are shown.

	Aircraft	Caltech101	CIFAR100	DTD	EuroSAT	Flowers	Food	MNIST	OxfordPet	Cars	SUN397	
Transfer		87.4	68.1	46.0	56.3	71.2	88.9	65.1	88.9	64.6	67.2	70.4
Aircraft	58.3	87.4	67.8	46.2	55.2	70.9	88.8	60.6	88.9	64.7	65.5	
Caltech101	59.4	96.3	68.3	45.7	55.8	71.2	88.9	62.3	88.5	64.8	66.3	
CIFAR100	59.7	96.1	89.1	46.2	56.3	71.3	89.0	65.3	89.1	64.7	67.6	
DTD	59.7	96.2	89.1	80.4	58.0	71.3	89.1	68.6	89.0	64.8	68.0	
EuroSAT	59.7	96.0	89.1	80.4	98.2	71.2	88.9	67.7	89.0	64.7	67.9	
Flowers	59.9	96.4	89.1	80.4	98.2	96.0	88.9	66.5	88.8	64.6	67.2	
Food	59.8	96.4	89.0	80.5	98.3	95.9	92.4	64.8	88.8	64.7	67.4	
MNIST	60.0	96.4	89.0	80.5	98.2	95.9	92.4	99.6	88.8	64.5	67.2	
OxfordPet	59.8	96.4	89.0	80.5	98.2	96.0	92.4	99.6	94.5	64.3	67.2	
Cars	59.8	96.4	89.1	80.3	98.2	95.8	92.4	99.6	94.4	90.6	67.2	
SUN397	59.9	96.4	89.1	80.4	98.2	95.8	92.4	99.6	94.4	90.6	81.9	89.0
Average	59.6	95.5	85.2	71.1	83.0	84.7	90.5	77.7	90.4	69.4	68.5	79.6

The complete session-wise matrices are reported in Tables 10–14. Each entry records the accuracy on an evaluation dataset after a particular training session. The cells above the diagonal determine Transfer, the complete matrix determines Average, and the final row determines Last. Beyond making the aggregate metrics reproducible, these matrices expose how future-task generalization and learned-task retention evolve throughout the sequence.

Full-shot sequences. Tables 10 and 11 show that the performance of a learned task generally remains stable during subsequent adaptation. In Order-I, all learned datasets except Aircraft remain within 0.2 points of their immediate post-training accuracy in the final row, while Aircraft improves from 58.3 to 59.9. In Order-II, every learned dataset remains within 0.3 points of its diagonal value. The reordered sequence therefore changes the absolute Transfer and Average scores but does not

Table 11: Accuracy of DualMem-Evo on the MTIL benchmark with order-II. Each row represents the performance on every dataset of the model trained after the corresponding task. Transfer, Average, and Last metrics are shown.

	Cars	Food	MNIST	OxfordPet	Flowers	SUN397	Aircraft	Caltech101	DTD	EuroSAT	CIFAR100	
Transfer		88.6	61.2	89.1	71.2	66.1	24.2	88.4	46.1	54.7	70.7	66.0
Cars	90.1	88.6	60.8	88.9	71.1	65.7	24.4	88.7	45.7	55.7	68.7	
Food	90.3	92.7	61.5	89.4	71.3	66.5	24.4	88.6	45.7	55.1	70.7	
MNIST	90.3	92.7	99.5	88.9	71.1	66.2	23.9	88.1	45.7	53.5	70.1	
OxfordPet	90.3	92.7	99.5	94.8	71.4	66.3	24.2	88.3	45.5	53.7	70.2	
Flowers	90.3	92.7	99.5	94.8	95.7	66.0	23.8	88.3	46.4	54.3	70.5	
SUN397	90.4	92.7	99.5	94.9	95.7	81.3	24.2	88.5	46.4	54.9	71.3	
Aircraft	90.4	92.7	99.5	94.9	95.7	81.3	58.6	88.0	46.5	54.6	71.4	
Caltech101	90.5	92.7	99.5	94.8	95.7	81.3	58.8	96.0	46.5	54.6	71.5	
DTD	90.4	92.7	99.5	94.8	95.7	81.3	58.6	95.9	79.8	56.2	71.5	
EuroSAT	90.4	92.7	99.5	94.8	95.6	81.3	58.8	95.9	79.8	98.1	71.5	
CIFAR100	90.4	92.6	99.5	94.7	95.7	81.3	58.5	96.0	80.0	98.1	89.0	88.7
Average	90.3	92.3	92.5	93.2	86.8	74.4	39.8	91.1	55.3	62.6	72.4	77.3

Table 12: Accuracy of DualMem-Evo on the 5-shot MTIL-FS benchmark. Each row represents the performance on every dataset of the model trained after the corresponding task. Transfer, Average, and Last metrics are shown.

	Aircraft	Caltech101	CIFAR100	DTD	EuroSAT	Flowers	Food	MNIST	OxfordPet	Cars	SUN397	
Transfer		88.3	67.8	45.6	55.8	70.6	88.7	66.9	88.2	64.7	66.2	70.3
Aircraft	31.8	88.3	67.8	44.9	54.9	70.7	88.6	62.3	88.7	64.9	65.3	
Caltech101	32.2	92.1	67.8	45.7	56.6	70.7	88.7	65.0	87.8	64.9	65.9	
CIFAR100	32.2	92.0	75.6	46.1	56.3	70.6	88.8	68.9	88.3	64.8	66.6	
DTD	32.4	91.8	75.5	66.7	55.4	70.9	88.8	68.7	88.2	64.8	67.2	
EuroSAT	32.3	91.8	75.5	66.7	89.3	70.2	88.7	68.4	88.1	64.9	66.8	
Flowers	32.2	91.8	75.5	66.7	89.3	90.7	88.7	68.2	88.1	64.7	66.3	
Food	32.4	91.8	75.6	66.9	89.4	90.8	81.6	66.7	88.2	64.6	66.3	
MNIST	32.6	91.7	75.5	66.8	89.3	90.8	81.6	95.8	88.3	64.6	66.0	
OxfordPet	32.2	91.8	75.5	66.6	89.3	90.8	81.5	95.8	87.1	64.3	65.9	
Cars	32.4	91.8	75.5	66.8	89.1	90.9	81.6	95.8	87.2	72.8	65.9	
SUN397	32.6	91.9	75.6	66.7	89.3	91.0	81.5	95.8	87.3	72.8	72.1	77.9
Average	32.3	91.5	74.1	61.0	77.1	81.6	85.5	77.4	87.9	66.2	66.8	72.9

introduce substantial backward degradation, supporting the role of episodic memory in protecting acquired knowledge.

Few-shot sequences. The matrices in Tables 12 and 13 exhibit a similar pattern under limited supervision. In the 5-shot setting, the lower diagonal values reflect the difficulty of learning from only five images per class, yet all final accuracies remain within 0.8 points of their corresponding post-training values. Under the 16-shot protocol, this range further narrows to 0.5 points. Hence, the main limitation in the few-shot regimes is the reduced initial adaptation capacity rather than progressive forgetting after a task has been learned.

Table 13: Accuracy of DualMem-Evo on the 16-shot MTIL-FS benchmark. Each row represents the performance on every dataset of the model trained after the corresponding task. Transfer, Average, and Last metrics are shown.

	Aircraft	Caltech101	CIFAR100	DTD	Flowers	Food	Cars	SUN397	
Transfer		91.7	68.1	46.6	70.6	85.9	65.6	66.2	70.7
Aircraft	43.7	91.7	67.7	46.7	70.2	85.9	65.4	65.6	
Caltech101	44.1	96.1	68.4	46.4	70.2	85.9	65.5	65.9	
CIFAR100	44.1	96.1	81.7	46.6	70.5	85.9	65.7	66.3	
DTD	44.2	96.2	81.7	73.9	71.3	86.0	65.4	66.4	
Flowers	44.2	96.2	81.7	73.9	85.2	85.8	65.7	66.3	
Food	44.4	96.2	81.8	73.8	85.3	85.6	65.9	66.2	
Cars	44.3	96.3	81.8	74.0	85.1	85.6	83.3	66.5	
SUN397	44.2	96.2	81.8	73.9	85.0	85.6	83.2	75.9	78.2
Average	44.2	95.6	78.3	63.7	77.9	85.8	70.0	67.4	72.9

Table 14: Accuracy of DualMem-Evo on the IFMS-MTIL benchmark. Each row represents the performance on every dataset of the model trained after the corresponding task. Transfer, Average, and Last metrics are shown.

	Aircraft	Caltech101	CIFAR100	DTD	EuroSAT	Flowers	Food	MNIST	OxfordPet	Cars	SUN397	
Transfer		87.2	67.8	45.7	56.1	69.5	88.8	59.4	88.8	65.0	67.4	69.6
Aircraft	55.5	87.2	67.6	45.6	55.4	70.7	88.7	59.8	89.0	64.8	65.5	
Caltech101	55.7	95.7	68.0	45.7	55.8	69.2	88.8	63.3	88.3	64.8	66.3	
CIFAR100	56.1	95.3	87.7	45.7	56.3	69.3	88.9	53.1	88.9	64.9	67.7	
DTD	55.4	95.2	87.5	77.9	56.8	69.2	88.8	60.1	89.0	65.0	68.0	
EuroSAT	55.4	95.3	87.5	77.8	96.9	69.3	88.8	60.7	89.1	64.9	68.0	
Flowers	54.9	95.2	87.3	77.8	96.7	91.7	88.8	59.4	88.8	65.1	67.6	
Food	54.6	95.2	87.3	77.6	96.7	91.7	92.1	59.2	88.8	65.1	67.8	
MNIST	54.0	95.2	87.2	77.3	96.1	91.5	92.0	97.4	88.7	65.1	67.6	
OxfordPet	54.0	95.2	87.2	77.3	96.1	91.5	92.0	97.4	94.4	64.9	67.4	
Cars	54.1	95.3	87.4	77.5	96.2	91.4	92.1	97.3	94.5	87.0	67.6	
SUN397	53.9	95.2	87.3	77.0	95.6	91.5	92.0	97.4	94.2	86.5	79.8	86.4
Average	54.9	94.5	83.8	68.8	81.7	81.5	90.3	73.2	90.3	68.9	68.5	77.9

Identifier-free module selection. Table 14 shows that eliminating identifier-based expert lookup introduces additional variation but does not cause severe performance collapse. The largest diagonal-to-final change is a 1.6-point decrease on Aircraft, while the remaining learned datasets change by at most 1.3 points. Relative to standard Order-I, identifier-free inference reduces Transfer, Average, and Last by 0.8, 1.7, and 2.6 points, respectively. This quantifies the cost of removing exact module lookup, while the stable lower-triangular region shows that the input-driven router still preserves most previously acquired knowledge.

D Supplementary Explanatory Analyses

Figure 1 motivates the design of dual and evolving memory. We further examine the evolution of historical-expert contributions during training, the functional separation between semantic

and episodic memories, feature stability across sessions, and historical-expert selection and reuse throughout the task sequence.

D.1 Controlled Analyses of Dual and Evolving Memory

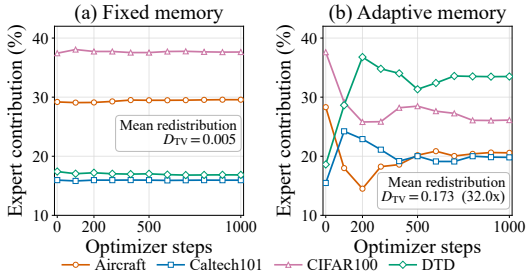

Figure 4: **Evolution of historical-expert contributions during EuroSAT adaptation.** Fixed memory maintains an almost unchanged utilization pattern, whereas adaptive memory reallocates contributions among historical experts during early optimization and then gradually stabilizes.

Experimental settings. The controlled analyses in Figure 1 follow the no-replay MTIL Order-I protocol Zheng et al. (2023) and use a frozen CLIP ViT-B/16 backbone. All variants are trained for 1,000 iterations with a batch size of 64 and a 250-step warm-up. For comparable adapter capacity, the shared-only and independent-expert variants use rank 128, whereas the dual-memory variant uses $r_g = r_t = 64$. The functional-separation and semantic-stability objectives are disabled in these experiments to isolate the effects of memory topology and memory evolution.

Figures 1(a–b) use controlled two-task sequences. Figure 1(a) compares frozen CLIP, shared-only adaptation, independent experts, and dual memory on Aircraft→Caltech101. Figure 1(b) compares independent and dual memory while learning CIFAR100 after Caltech101, with accuracy evaluated at matched optimization steps. The annotated value reports the pointwise Dual-minus-Independent accuracy difference at the marked checkpoint.

Figures 1(c–d) use the first five Order-I tasks: Aircraft, Caltech101, CIFAR100, DTD, and EuroSAT. During EuroSAT adaptation, the contributions of the four preceding experts are measured every 100 steps using the same fixed 32-sample probe from the current-session training stream. Fixed fusion keeps the historical-expert coefficients unchanged, whereas adaptive fusion optimizes them during training; all remaining settings are identical.

Let $u_k^{(s)} \geq 0$ denote the functional contribution of historical expert k at step s, computed as the summed mean ℓ_2 norm of its residual outputs across the instrumented image and text LoRA layers. We normalize it as

$$p_k^{(s)} = \frac{u_k^{(s)}}{\sum_{j \in \mathcal{K}} u_j^{(s)}}, \qquad \sum_{k \in \mathcal{K}} p_k^{(s)} = 1, \tag{37}$$

where $\mathcal{K}$ contains the four historical experts. The curves report $100 p_k^{(s)}\%$, and the redistribution between the initial and final contribution distributions is measured by

$$D_{\mathrm{TV}}\left(\mathbf{p}^{(0)}, \mathbf{p}^{(S)}\right) = \frac{1}{2} \sum_{k \in \mathcal{K}} \left| p_k^{(S)} - p_k^{(0)} \right|, \qquad S = 1000. \tag{38}$$

To test whether historical experts function as an evolving memory bank rather than static parameter snapshots, we track the four preceding experts while adapting to EuroSAT. Fixed fusion produces an almost unchanged contribution distribution, yielding $D_{\mathrm{TV}} = 0.0054$. In contrast, PGEEM-driven adaptive fusion reorganizes expert usage as new evidence arrives and yields $D_{\mathrm{TV}} = 0.1733$ before converging to a stable mixture. Figure 4 therefore shows that historical knowledge is selectively recombined rather than uniformly or permanently fused.

D.2 Functional Separation of Dual Memories

Figure 5 measures the marginal CKA between the semantic and episodic contributions. Without explicit separation, the two branches remain highly redundant throughout training and reach a final CKA of 0.847 at the regularized layer. The proposed objective rapidly reduces this similarity and reaches 0.045, while the non-zero final contribution norms confirm that both branches remain active. The layer-wise result localizes the strongest separation at the optimized visual layer while retaining functional interactions elsewhere. This behavior is consistent with the intended functional division between semantic and episodic memories.

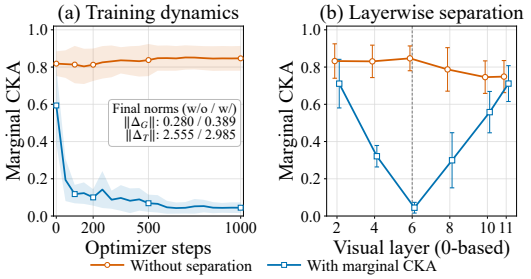

Figure 5: **Functional separation between semantic and episodic memories.** (a) Marginal CKA at the regularized visual layer over five sequential tasks; shaded regions denote the standard deviation across fixed task-specific probes. (b) Final layer-wise CKA, where the dashed line marks the regularized layer. The separation objective reduces layer-6 similarity from 0.847 to 0.045 without collapsing either branch.

D.3 Feature Stability Across Sessions

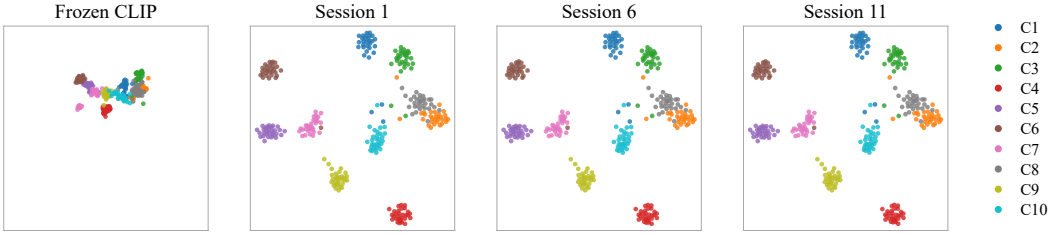

Figure 6: **Feature evolution on Aircraft.** Joint t-SNE visualization of identical test samples under frozen CLIP and the Aircraft episodic expert combined with the evolving semantic memory at Sessions 1, 6, and 11. The persistent class structure provides qualitative evidence that task-specific knowledge is preserved throughout continual learning.

We jointly project the same Aircraft test samples under frozen CLIP and under DualMem-Evo after Sessions 1, 6, and 11. As shown in Figure 6, adaptation transforms the overlapping frozen-CLIP representation into well-separated class clusters after the first task. More importantly, the cluster organization remains visually consistent as ten additional tasks are learned. Because t-SNE is qualitative and does not preserve all global distances, we use this visualization only as supporting evidence; the corresponding quantitative retention behavior is reported by the session-wise matrices in Appendix C.

D.4 Historical Expert Reuse under PGEEM

As shown in Fig. 7, PGEEM first identifies informative historical experts according to their centered prototype affinities and subsequently adapts their functional roles during training. Once the candidate pool exceeds five experts, both positively and negatively centered memories are retained, while their final contributions remain non-uniform and task-dependent, demonstrating selective reuse rather than static or uniform expert fusion.

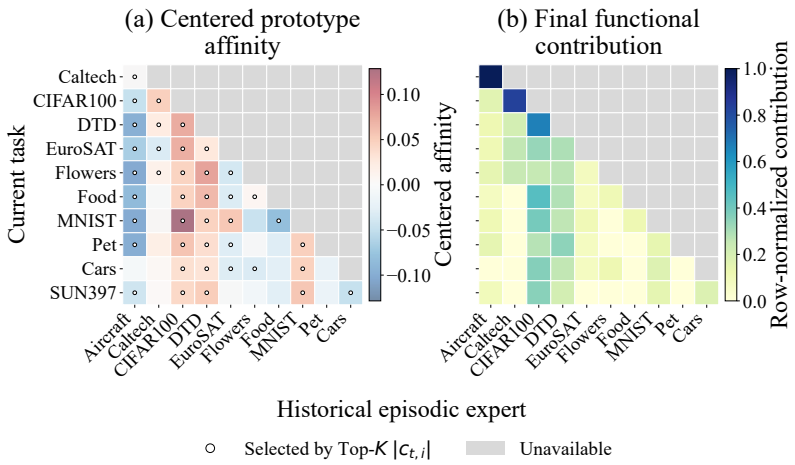

Figure 7: **Historical expert reuse under PGEEM on MTIL Order-I.** (a) Centered prototype affinities, where circles indicate the Top-K experts selected by absolute centered score. (b) Row-normalized magnitudes of their applied LoRA residuals after training and before SVD fusion. Gray cells denote unavailable future experts. Statistics are computed on a fixed class-balanced training subset.

E Parameter–Performance Trade-off and Hyperparameter Analysis

E.1 Trainable Parameters, Auxiliary Data, and Performance

Table 15 compares peak trainable parameters, auxiliary-data usage, and Order-I performance. ZSCL Zheng et al. (2023), MoE-Adapters Yu et al. (2024), and GIFT Wu et al. (2025) respectively use ImageNet and Conceptual Captions, TinyImageNet, and diffusion-generated images, whereas the remaining parameter-efficient methods use only the current MTIL training data. DualMem-Evo matches the best Transfer and achieves the highest Average and Last with 19.7M trainable parameters and no auxiliary data, using 86.9% fewer trainable parameters than full-model approaches and 67.1% fewer than MoE-Adapters. Frozen components and non-parametric statistics are excluded from all parameter counts.

Table 15: Parameter–performance and auxiliary-data comparison on MTIL Order-I. "Params." denotes peak trainable parameters per session, and CC denotes Conceptual Captions. Best metric values are marked in **bold**. Counts follow the original analyses or official configurations; the implementation-derived 149.6M count is used for ZSCL for consistency under CLIP ViT-B/16.

Method	Params. (M)	Auxiliary data	Transfer	Average	Last
LwF	149.6	–	56.9	64.7	74.6
WiSE-FT	149.6	–	52.3	60.7	77.7
ZSCL	149.6	ImageNet + CC	68.1	75.4	83.6
MoE-Adapters	59.8	TinyImageNet	68.9	76.7	85.0
DIKI	1.8	–	68.7	76.3	85.1
IAP	2.4	–	69.2	76.8	85.7
GIFT	149.6	Synthetic images	69.3	77.3	86.0
ConDU (LoRA)	9.8	–	70.3	78.3	86.2
DPW	30.8	–	**70.4**	79.3	88.3
KeepLoRA+	1.2	–	69.9	79.0	88.2
DualMem-Evo	19.7	–	**70.4**	**79.6**	**89.0**

E.2 Hyperparameter Sensitivity

SVD consolidation rank. Figure 8 shows limited sensitivity to the effective SVD rank. Across ranks 16–64, Transfer, Average, and Last vary by at most 0.21, 0.25, and 0.29 percentage points,

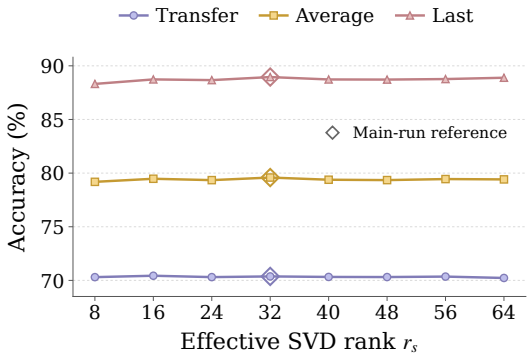

Figure 8: Effective SVD rank sensitivity on MTIL Order-I. Curves show the controlled sweep; hollow diamonds denote the main-run reference at $r_s = 32$.

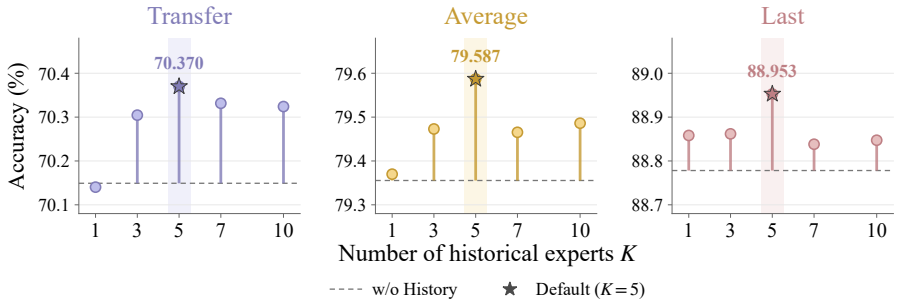

Figure 9: Historical-expert-count sensitivity on MTIL Order-I. Stems show differences from the dashed w/o-History reference, which disables both historical collaboration and SVD consolidation. Stars mark the default $K = 5$; all panels span 0.4 percentage points.

respectively. Rank 8 reduces Average and Last, while larger ranks offer no consistent improvement across all three metrics. These observations support $r_s = 32$ as a balanced consolidation choice.

Number of historical experts. As shown in Figure 9, performance remains relatively stable across $K = 3$–10, with each metric varying by less than 0.13 percentage points. The default $K = 5$ achieves the highest scores on all three metrics, outperforming both the w/o-History baseline and the all-history setting ($K = 10$). These results support selective historical-expert reuse with a moderate retrieval budget.

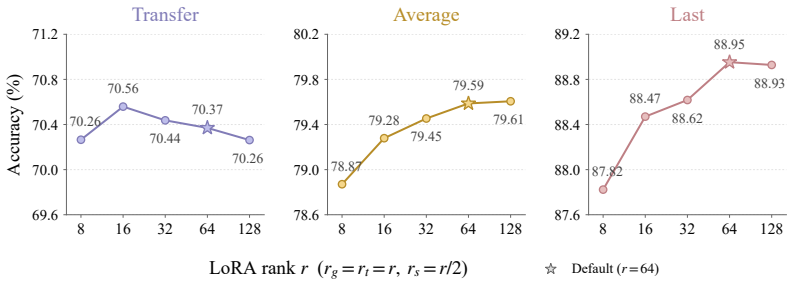

Figure 10: Performance across LoRA ranks on MTIL Order-I, where $r_g = r_t = r$ and $r_s = r/2$. The star denotes the default configuration ($r = 64$).

Rank of LoRA. Figure 10 shows that DualMem-Evo is robust to the LoRA rank. Average and Last generally improve as the rank increases and saturate near $r = 64$, whereas Transfer varies only slightly and peaks at $r = 16$. Notably, $r = 16$ already preserves most of the default configuration's performance and outperforms several state-of-the-art baselines, demonstrating strong parameter efficiency in a compact low-rank space.