D²C-MMCL: DUAL DECOUPLING AND CONSOLIDATION FOR MULTIMODAL CONTINUAL LEARNING IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual learning for Vision-Language Models (VLMs) aims to acquire new downstream knowledge while preserving previously learned knowledge and pretrained zero-shot capabilities. However, existing methods largely overlook the evolving roles of different modalities during continual learning. We observe that the relative contributions of visual and textual modalities are inherently dynamic and task-dependent, varying as new tasks are introduced and knowledge accumulates. Motivated by this observation, we propose **D²C-MMCL** (**D**ual **D**ecoupling and **C**onsolidation for **M**ultimodal **C**ontinual **L**earning), a dual-decoupling and knowledge-consolidation framework that disentangles continual knowledge along modality and task dimensions. By dynamically modeling modality contributions, D²C-MMCL performs modality-aware cross-task consolidation and selectively exploits modality-specific knowledge for different tasks. Extensive experiments across various continual learning settings demonstrate that our method consistently outperforms strong existing approaches, achieving superior continual learning performance while better preserving zero-shot generalization.

1 INTRODUCTION

Real-world data distributions and task requirements are inherently dynamic, requiring models to continually adapt to new tasks and knowledge while time and resource constraints often make repeated training from scratch impractical(Wang et al., 2024; Shi et al., 2025). Continual Learning (CL) addresses this setting by enabling models to acquire new knowledge while retaining previously learned competencies(Zhou et al., 2024; Aghasanli et al., 2025; Sun et al., 2025). In parallel, Vision-Language Models (VLMs) like CLIP(Radford et al., 2021) have become a strong foundation for continual adaptation, owing to the transferable cross-modal representations and zero-shot generalization acquired through large-scale image-text pre-training(Liu et al., 2025b; Feng et al., 2024; Yada et al., 2025; Jin et al., 2025). Sequentially adapting VLMs to downstream tasks can still induce catastrophic forgetting and erode their pretrained zero-shot capability(Yan et al., 2022). Balancing adaptation to new tasks with the retention of prior knowledge and pretrained generalization therefore remains a fundamental challenge in continual learning for VLMs(Gao et al., 2026).

To address these challenges, existing approaches to continual learning for VLMs have primarily focused on two directions: knowledge preservation and parameter-efficient adaptation(Marczak et al., 2024; Huang et al., 2024; Jha et al., 2024). Distillation-based methods regularize continual adaptation by transferring knowledge from the pretrained model or previously learned models, thereby mitigating the forgetting of acquired capabilities(Zheng et al., 2023; Yu et al., 2024b; Wu et al., 2025). However, they often require auxiliary reference data and are sensitive to its choice. Parameter-efficient methods adapt VLMs using lightweight modules, such as prompts and adapters, which accommodate new-task knowledge while minimizing modifications to the pretrained parameters(Bian et al., 2024; Houlsby et al., 2019; Hu et al., 2021). More recent studies have begun to account explicitly for the multimodal nature of VLMs, recognizing that the visual and textual modalities can play distinct roles throughout continual adaptation(Zhou et al., 2025; Huang et al., 2025).

We further observe that the roles of the visual and textual modalities in VLM continual learning are not fixed, but evolve as the model accumulates knowledge over time. As shown in Fig. 1, the

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.

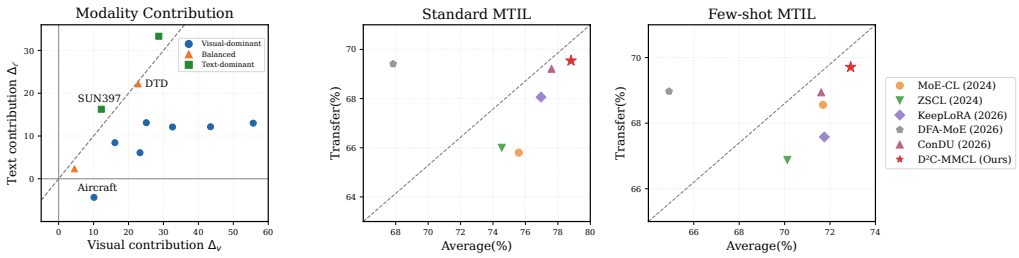

Figure 1: Motivation and overall performance. **Left:** Relative contributions of visual and textual modalities across downstream tasks, showing task-varying modality dependence. **Right:** Average-Transfer performance comparison under full-shot and few-shot settings, where our method achieves a favorable trade-off between continual adaptation and zero-shot generalization.

two modalities exhibit markedly different relative contributions across downstream tasks. Rather than being a static property determined solely by the data distribution of an individual task, this modality imbalance emerges dynamically from the interaction between current task demands and previously acquired knowledge. This observation reveals that modality contributions in continual VLM adaptation are inherently dynamic and task-dependent.

Motivated by this observation, we propose D²C-MMCL (Dual Decoupling and Consolidation for Multimodal Continual Learning), a VLM multimodal continual learning framework that explicitly models the distinct roles of visual and textual modalities during continual adaptation. D²C-MMCL decouples continual knowledge along modality and task dimensions. At modality level, it decomposes task-induced knowledge shifts into visual and textual components and dynamically estimates their relative contributions conditioned on current task and accumulated prior knowledge. Within each modality, knowledge from different tasks is then continually consolidated while preserving task-specific distinctions through task-wise decoupling. At inference time, we selectively reconstruct visual and textual knowledge according to their task-dependent modality contributions. As shown in Fig. 1, our method achieves a favorable Average–Transfer trade-off across different continual learning settings, demonstrating the effectiveness of organizing and utilizing multimodal knowledge according to dynamic modality contributions. Our contributions are summarized as follows:

- We uncover a dynamic and task-dependent modality imbalance in VLM continual learning, where the relative contributions of visual and textual knowledge vary across tasks and evolve throughout the learning process.
- We propose D²C-MMCL, which performs dual decoupling along modality and task dimensions and consolidates cross-task knowledge according to dynamic modality contributions, enabling more effective organization and utilization of modality-specific knowledge.
- Extensive experiments across standard, few-shot and different task-order settings demonstrate consistent improvements over strong existing methods, while striking a favorable balance between robust knowledge retention and zero-shot generalization.

2 RELATED WORK

Continual Learning of VLMs. Current approaches to continual learning in Vision-Language Models (VLMs) primarily follow two directions, distillation-based knowledge preservation and parameter-efficient fine-tuning(Cui et al., 2024; Tang et al., 2024; Li et al., 2025; Park, 2024). Distillation-based methods typically rely on auxiliary data to mitigate representation drift during continual adaptation(Zheng et al., 2024). ZSCL(Zheng et al., 2023) introduces a semantically diverse reference dataset to preserve the pretrained feature space, while Select and Distill (Yu et al., 2024b) employs selective dual-teacher knowledge transfer and GIFT (Wu et al., 2025) constructs synthetic reference data using generative models. In contrast, PEFT-based approaches freeze most pretrained parameters and introduce prompts, adapters, or low-rank updates for task adaptation (Wang et al., 2022a;c;b; Smith et al., 2023). Subsequent studies further explore more flexible adapta-

tion mechanisms, including mixture-of-experts based module selection (Yu et al., 2024a) and label-specific adaptation (Luo et al., 2025). More recent methods constrain the parameter update space to reduce cross-task interference; KeepLoRA (Luo et al., 2026a) and NuSA-CL (Jo & Kim, 2026) restrict new-task updates to the residual subspace and an approximate null space, respectively.

Modality-aware Continual Learning for VLMs. Recent studies have begun to examine the distinct behaviors of the visual and textual modalities during continual learning(Luo et al., 2026b; Liu et al., 2025a; Zhou et al., 2025). DMNSP (Kang et al., 2025) observes that the visual and language branches exhibit different degrees of parameter variation throughout continual adaptation, and accordingly adopts an asymmetric optimization strategy to treat the two modalities differently. MG-CLIP (Huang et al., 2025) instead focuses on the modality gap between visual and textual representations, preserving and compensating for this gap to coordinate cross-modal relationships during continual adaptation. D-MoLE (Ge et al., 2025) further adjusts the update ratio between the language model and the modality encoder according to the task-specific difficulty of each modality. These works collectively demonstrate that visual and textual modalities behave differently in continual learning. Differently, our work examines their relative contributions to task learning and how these contributions shift with the continual-learning state.

3 METHOD

3.1 Problem Formulation

We consider multimodal continual learning for Vision-Language Models (VLMs). Given a pre-trained VLM such as CLIP, which comprises a visual encoder and a text encoder, we denote its parameters by θ^0. The model is required to sequentially learn a stream of downstream tasks $\left\{\tau^1, \tau^2, \ldots, \tau^T\right\}$. Each task $\tau^t = (\mathcal{D}^t, \mathcal{C}^t)$ consists of a training dataset $\mathcal{D}^t$ and corresponding class set $\mathcal{C}^t$. The dataset $\mathcal{D}^t = \{(x_i^t, y_i^t)\}_{i=1}^{N_t}$ contains N_t training samples, where x_i^t denotes the input image and y_i^t denotes the corresponding class label. The class set $\mathcal{C}^t = \left\{c_j^t\right\}_{j=1}^{C_t}$ consists of the class names in task τ^t, with a total of C_t classes. Each class name is combined with a predefined prompt template and encoded by the text encoder into a textual representation, while each input image is encoded by the visual encoder into a visual representation. During continual learning, the model has access only to the current task, while the raw training samples from previous tasks are no longer available. The objective is to continuously adapt the pre-trained VLM to incoming tasks while mitigating catastrophic forgetting on previously learned tasks and preserving its zero-shot capabilities.

3.2 Modality Decoupling

For each downstream task τ^t, we obtain a task expert $\theta^t = (\theta_v^t, \theta_l^t)$ by individually fine-tuning the pre-trained VLM $\theta^0 = \left(\theta_v^0, \theta_l^0\right)$ on the current task. We define the task delta as $\delta^t = \theta^t - \theta^0$, the parameter offset of the task expert τ^t from the pre-trained model, where δ^t captures the task-specific parameter changes induced by fine-tuning on τ^t.

In multimodal continual learning, downstream tasks have heterogeneous dependencies on vision and text, yielding strong task-specific variation in their relative contributions.Some tasks are predominantly driven by visual information, whereas others rely more heavily on textual semantics. To explicitly capture such task-dependent modality differences, we decouple the task delta δ^t along the modality dimension into visual and textual components, denoted by δ_v^t and δ_l^t, respectively. Let the parameters of the visual and text encoders in the pre-trained model and the task expert be denoted by $\theta^0 = \left(\theta_v^0, \theta_l^0\right)$ and $\theta^t = (\theta_v^t, \theta_l^t)$, respectively. Then, $\delta_v^t = \theta_v^t - \theta_v^0$ and $\delta_l^t = \theta_l^t - \theta_l^0$.

3.3 Cross-task Consolidation and Task-wise Decoupling

Modality Reliability Estimation. Following modality decoupling, we further quantify the relative contributions of the visual and textual modalities to the current task, using modality reliability as the metric. For the current task τ^t, the previous-task expert $\widehat{\theta}^{t-1} = \left(\widehat{\theta}_v^{t-1}, \widehat{\theta}_l^{t-1}\right)$ is decoupled from the modality-specific knowledge space $U_m^{1:t-1}$ maintained in the preceding stage via the corresponding task trigger; the detailed procedure is introduced in the paragraph of Task-wise Decoupling. We

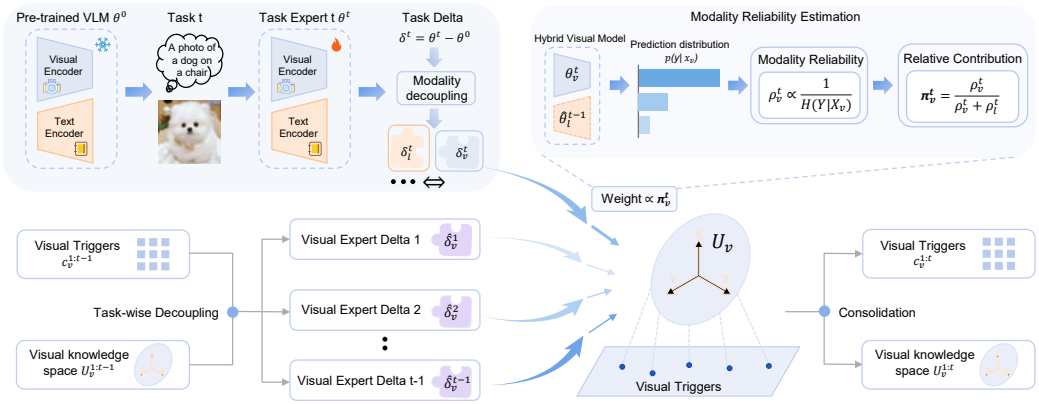

Figure 2: Overview of the D²C-MMCL framework. For each incoming task, a task expert is first obtained by fine-tuning the pre-trained VLM, and its task-induced parameter update is then decoupled into visual and textual modality deltas. Modality Reliability Estimation dynamically evaluates their relative contributions through cross-modal hybrid models and produces modality weights for subsequent consolidation. For each modality, historical task knowledge is task-wise decoupled from the previous knowledge space using task triggers and integrated with the current task delta to update the modality-specific knowledge space and triggers. For clarity, only the visual consolidation branch is illustrated, while the textual branch follows the same procedure.

construct two cross-decoupled hybrid models M_v^t and M_l^t:

$$M_v^t = \left(\theta_v^t,\ \widehat{\theta}_l^{t-1}\right), \qquad M_l^t = \left(\widehat{\theta}_v^{t-1},\ \theta_l^t\right) \tag{1}$$

Each combination preserves the current-task variation in one modality while retaining the complementary modality from the prior continual-learning stage.This formulation situates the assessment of modality contributions within the evolving continual-learning state, capturing modality changes relative to accumulated knowledge rather than static deviations from the fixed pre-trained model.

For modality $m \in \{v, l\}$, let $p_m^t\left(y \mid x_i^t\right)$ denote the predictive distribution of the hybrid model M_m^t for sample x_i^t. Its normalized predictive entropy is defined as:

$$\bar{H}_m^t\left(x_i^t\right) = -\frac{\sum_{c=1}^{C_t} p_m^t\left(y=c \mid x_i^t\right) \log p_m^t\left(y=c \mid x_i^t\right)}{\log C_t}. \tag{2}$$

Here, $\log C_t$ is used to account for class-cardinality effects on entropy range. Modality reliability is defined as the average inverse normalized predictive entropy over the samples of the current task:

$$\rho_m^t = \frac{1}{N_t} \sum_{i=1}^{N_t} \frac{1}{\bar{H}_m^t\left(x_i^t\right)}, \qquad m \in \{v, l\}. \tag{3}$$

We further normalize the reliability scores of the two modalities as

$$\pi_m^t = \frac{\rho_m^t}{\rho_v^t + \rho_l^t}, \quad m \in \{v, l\}. \tag{4}$$

Theoretical analysis. The modality reliability defined above via predictive entropy can be interpreted from the perspective of task-relevant information(Hadizadeh et al., 2024). Let Z_m denote the representation of modality m, and let Y denote the task label. Their mutual information is given by

$$I(Y; Z_m) = H(Y) - H(Y \mid Z_m), \tag{5}$$

which quantifies the label-relevant information contained in the modality representation. Here, $H(Y)$ denotes the intrinsic uncertainty of the task labels, while $H(Y \mid Z_m)$ measures the residual label uncertainty after observing Z_m. For a fixed task, $H(Y)$ is independent of the modality. Therefore, a larger $I(Y; Z_m)$ corresponds to a smaller $H(Y \mid Z_m)$, indicating that the modality representation resolves more uncertainty about the task labels.

This conditional uncertainty is also theoretically related to classification error. For a classification task with C_t classes, Fano's inequality gives

$$H(Y \mid Z_m) \leq h(P_e^m) + P_e^m \log(C_t - 1), \tag{6}$$

where $P_e^m = P(\widehat{Y} \neq Y)$ is the classification error probability based on Z_m, and

$$h(P_e^m) = -P_e^m \log P_e^m - (1 - P_e^m) \log(1 - P_e^m) \tag{7}$$

is the binary entropy function. Since $h(P_e^m) \leq \log 2$, we obtain

$$P_e^m \geq \frac{H(Y \mid Z_m) - \log 2}{\log(C_t - 1)}. \tag{8}$$

Thus higher $H(Y|Z_m)$ raises the error lower bound, establishing a connection between residual label uncertainty conditioned on a modality representation and its task-discrimination reliability.

In practice, $H(Y \mid Z_m)$ depends on the unknown true distribution $p(Y \mid Z_m)$ and cannot be directly computed during training. Instead, the model outputs a class predictive distribution whose concentration reflects predictive uncertainty. Motivated by this correspondence, we use the normalized predictive entropy defined above to approximate the label uncertainty captured by $H(Y \mid Z_m)$, and define modality reliability ρ_m^t as its inverse.

Reliability-weighted Cross-task Consolidation. For each modality $m \in \{v, l\}$, we maintains an evolving modality knowledge space $U_m^{1:t}$ that consolidates the knowledge accumulated up to task τ^t. Given a new task τ^t, the modality decoupling yields the current modality delta δ_m^t, while the historical task-specific deltas $\{\hat{\delta}_m^k\}_{k=1}^{t-1}$ are decoupled from the previous knowledge space $U_m^{1:t-1}$ using their corresponding task triggers. Then we consolidates the historical and current modality deltas to update the knowledge space:

$$U_m^{1:t} = \text{Consolidation}\left(\hat{\delta}_m^1, \ldots, \hat{\delta}_m^{t-1}, \delta_m^t\right). \tag{9}$$

π_m^k characterizes the relative contribution of modality m within task τ^k. Since cross-task consolidation is performed independently for each modality, we further normalize these modality contributions along the task dimension to obtain the cross-task weights $w_m^k = \frac{\pi_m^k}{\frac{1}{t}\sum_{j=1}^{t} \pi_m^j}$, $k = 1, \ldots, t$.

At stage t, we denote the modality deltas involved in the current consolidation as $\delta_m^{k|t} = \begin{cases} \hat{\delta}_m^k, & k < t, \\ \delta_m^t, & k = t. \end{cases}$ We then formulate the update of the modality knowledge space as a reliability-weighted cross-task reconstruction problem:

$$U_m^{1:t} = \arg\min_{\mathbf{U}_m} \sum_{k=1}^{t} w_m^k \left\| \delta_m^{k|t} - \mathbf{U}_m \mathbf{U}_m^T \delta_m^{k|t} \right\|_F^2. \tag{10}$$

A larger w_m^k imposes a stronger reconstruction constraint on task τ^k, thereby encouraging its modality-specific knowledge to be better preserved in the consolidated knowledge space.

After obtaining the updated modality knowledge space $U_m^{1:t}$, we store a task trigger for each task τ^k as $c_m^k = (U_m^{1:t})^T \delta_m^{k|t}$, $k = 1, \ldots, t$. At stage t, D^{2}C maintains, for each modality m,

$$\mathcal{K}_m^{1:t} = \left(U_m^{1:t}, \{c_m^k\}_{k=1}^t\right). \tag{11}$$

Here, $U_m^{1:t}$ captures the modality structure consolidated across tasks, while $\{c_m^k\}_{k=1}^t$ encodes the task-specific coefficients associated with individual tasks. Together, the knowledge space and task triggers enable D^{2}C-MMCL to continually consolidate cross-task modality knowledge while preserving structural distinctions among tasks.

Task-wise Decoupling. For task τ^k, the corresponding trigger c_m^k decouples its task-specific modality delta from the current knowledge space as $\hat{\delta}_m^k = U_m^{1:t} c_m^k$. Applying this operation to both modalities yields the task expert

$$\hat{\theta}^k = \left(\theta_v^0 + \hat{\delta}_v^k,\ \theta_l^0 + \hat{\delta}_l^k\right). \tag{12}$$

3.4 SELECTIVE RECONSTRUCTION

The preceding cross-task consolidation uses modality reliability to regulate how strongly knowledge from different tasks is preserved in each modality space. However, this modulation only affects the construction of the knowledge space and does not explicitly control the contribution of each modality delta to the final task model. To further align inference-time adaptation with the modality dependence of each task, we selectively reconstruct the decoupled modality deltas as

$$\tilde{\delta}_m^k = s(\pi_m^k; \beta)\hat{\delta}_m^k, \qquad m \in \{v, l\}, \tag{13}$$

where $s(\cdot)$ maps the task-specific modality contribution to its reconstruction strength. Selective Reconstruction is applied only at inference time. This separation lets training to focus on consolidating and preserving cross-task modality knowledge, while inference modulates each modality's influence based on the target task's modality dependence. The resulting inference model for task τ^k is

$$\theta_*^k = \left(\theta_v^0 + \tilde{\delta}_v^k,\ \theta_l^0 + \tilde{\delta}_l^k\right). \tag{14}$$

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTING

Datasets and Protocols. We evaluate our method on the Multi-domain Task Incremental Learning (MTIL) benchmark(Zheng et al., 2023), which contains 11 cross-domain tasks with 1,201 classes. We adopt two standard task orders(Xu et al., 2024; Zhang et al., 2026; Park, 2024): Order-I, which arranges the tasks as Aircraft, Caltech101, CIFAR100, DTD, EuroSAT, Flowers, Food, MNIST, OxfordPet, StanfordCars, and SUN397; and Order-II, which uses StanfordCars, Food, MNIST, OxfordPet, Flowers, SUN397, Aircraft, Caltech101, DTD, EuroSAT, and CIFAR100. Experiments are conducted under full-shot and 5-shot settings for Order-I, and full-shot setting for Order-II.

Evaluation Metrics and Baselines. Following the MTIL protocol, we report three metrics: "Transfer", "Average", and "Last"."Transfer" measures the zero-shot generalization to unseen tasks."Average" is the mean accuracy of each task across all evaluation stages, measuring long-term overall performance."Last" reports the accuracy of each task at the final stage, reflecting the retention of previously acquired knowledge.We further average each metric across tasks to provide an overall score.We compare our method with representative VLM continual learning approaches, including distillation-based methods represented by LwF(Li & Hoiem, 2017) and ZSCL(Zheng et al., 2023), parameter-efficient adaptation methods including KeepLoRA(Luo et al., 2026a), MOE-CL(Yu et al., 2024a), and DFA-MoE(Kang et al.) and the model fusion-based method ConDU(Gao et al., 2026).

Implementation details. We employ the CLIP model with ViT-B/16 as our backbone. We perform 1,000 training iterations for each task in the standard MTIL setting and 500 iterations in the few-shot MTIL setting. AdamW is used as the optimizer with a batch size of 32 across all experiments. We adopt a cosine learning-rate schedule with 100 warm-up iterations and apply label smoothing with a factor of 0.2.The learning rate is selected from $\{1 \times 10^{-5}, 5 \times 10^{-5}\}$ according to the task. For Selective Reconstruction, we use the exponential scaling function with $\beta = 0.2$. All experiments are conducted using PyTorch on a single NVIDIA 4090 GPU.

4.2 MAIN RESULTS

Overall Performance. Tables 1 and 2 present the results under standard and few-shot MTIL settings with Order-I, respectively. Under the standard setting, our method achieves 69.54%, 78.81%, and 88.50% on Transfer, Average, and Last. While maintaining zero-shot transfer performance, it shows clearer advantages in Average and Last, indicating better retention of knowledge acquired throughout continual learning. This advantage remains in the more challenging few-shot setting, where our method achieves the best overall results across three metrics, reaching 69.71% Transfer, 72.91% Average, and 78.24% Last, exceeding the best competing performance by 0.74%, 1.16%, and 0.21%. Despite limited supervision, our method maintains strong generalization to unseen tasks while effectively preserving acquired knowledge. As further illustrated by the Average-Transfer comparison

Table 1: Comparison with state-of-the-art methods on the multi-domain task incremental learning benchmark under Order-I in terms of "Transfer", "Average" and "Last" scores (%).

	Method	Aircraft	Caltech101	CIFAR100	DTD	EuroSAT	Flowers	Food	MNIST	OxfordPet	Cars	SUN397	*Average*
Transfer	LwF	–	72.24	52.85	36.60	38.24	47.54	64.74	56.88	71.78	24.74	48.92	51.45
	MOE-CL	–	88.42	57.67	38.67	42.38	67.50	88.51	63.13	89.04	64.71	57.93	65.80
	ZSCL	–	84.10	67.45	46.44	43.93	68.01	86.35	64.33	82.15	55.65	61.64	66.00
	KeepLoRA	–	83.82	69.10	46.21	54.78	68.97	87.54	64.71	89.93	57.50	57.99	68.06
	DFA-MoE	–	87.85	68.59	45.23	54.35	70.12	87.99	66.55	88.72	63.01	61.64	69.41
	ConDU	–	88.54	67.54	44.4	58.29	70.26	87.47	61.08	87.61	63.05	63.85	69.21
	Ours	–	88.08	67.17	43.92	58.13	70.35	87.60	64.29	87.82	64.03	64.02	**69.54**
Average	LwF	37.03	85.17	65.84	54.94	68.12	57.48	64.69	72.37	77.37	32.59	51.23	60.62
	MOE-CL	49.81	92.80	80.74	67.86	75.15	83.73	88.72	75.88	89.04	68.03	59.79	75.59
	ZSCL	43.92	91.38	79.21	65.68	77.09	81.51	88.56	77.57	88.53	62.67	63.87	74.54
	KeepLoRA	56.81	95.84	83.58	66.23	82.50	80.91	89.45	77.44	91.12	62.59	60.13	76.96
	DFA-MoE	23.87	89.29	69.40	44.39	70.26	69.35	88.70	76.57	88.42	63.33	62.64	67.84
	ConDU	56.90	94.90	82.23	66.76	83.48	83.36	89.17	75.07	89.56	67.02	65.26	77.61
	Ours	58.48	95.28	84.30	69.49	83.94	84.84	89.62	77.10	89.70	68.57	65.59	**78.81**
Last	LwF	28.65	87.15	63.80	64.20	81.56	64.99	79.13	99.46	91.28	70.91	74.29	73.22
	MOE-CL	49.80	93.20	85.89	78.78	93.93	97.25	88.97	98.18	89.04	83.06	78.39	85.14
	ZSCL	40.41	91.88	80.80	72.23	94.24	90.10	90.79	98.88	93.81	82.76	77.51	83.04
	KeepLoRA	53.30	96.92	86.30	72.25	98.11	89.91	91.63	99.50	94.11	84.83	82.34	86.29
	DFA-MoE	23.55	89.75	65.56	43.78	73.50	67.30	88.12	90.66	86.05	57.11	72.67	68.91
	ConDU	55.87	95.33	85.05	74.47	97.78	93.95	91.13	99.55	94.77	84.62	79.39	86.54
	Ours	58.36	95.91	88.05	79.10	98.65	96.83	92.04	99.53	94.74	89.01	81.23	**88.50**

Table 2: Comparison with state-of-the-art methods on the multi-domain task incremental learning benchmark under few-shot Order-I in terms of "Transfer", "Average" and "Last" scores (%).

	Method	Aircraft	Caltech101	CIFAR100	DTD	EuroSAT	Flowers	Food	MNIST	OxfordPet	Cars	SUN397	*Average*
Transfer	LwF	–	76.96	54.25	39.20	40.08	50.13	68.36	56.92	75.53	27.90	47.92	53.73
	MOE-CL	–	88.42	68.22	42.64	52.43	68.15	88.52	62.94	89.04	64.71	60.56	68.56
	ZSCL	–	88.42	64.83	46.54	53.22	67.96	86.02	68.56	83.89	54.48	54.85	66.88
	KeepLoRA	–	79.88	67.32	45.86	54.22	70.63	89.32	59.30	91.71	62.20	55.33	67.58
	DFA-MoE	–	87.79	68.57	45.60	54.25	71.03	88.52	59.44	89.04	64.79	60.65	68.97
	ConDU	–	87.90	68.49	45.37	53.78	69.77	88.63	61.60	88.06	63.31	62.46	68.94
	Ours	–	88.25	68.82	45.43	55.61	70.36	88.37	65.56	88.12	64.51	62.10	**69.71**
Average	LwF	26.80	82.47	48.58	45.94	52.20	60.16	58.93	68.62	75.61	30.62	49.67	54.51
	MOE-CL	29.10	91.51	74.53	60.59	75.02	82.26	88.25	72.21	89.03	64.72	61.39	71.69
	ZSCL	30.64	90.96	72.99	56.37	68.90	80.58	86.69	76.83	87.00	60.60	59.76	70.12
	KeepLoRA	40.43	91.02	72.63	59.91	70.00	83.11	89.21	69.53	92.01	64.72	56.71	71.75
	DFA-MoE	24.33	87.79	68.59	45.59	54.26	71.03	88.52	59.45	89.05	64.80	60.65	64.91
	ConDU	33.33	90.29	74.20	58.93	74.59	79.79	88.06	73.22	87.99	63.92	63.35	71.61
	Ours	35.90	90.62	76.69	60.38	76.41	82.43	87.77	75.54	87.72	65.47	63.05	**72.91**
Last	LwF	24.21	82.90	39.83	50.16	50.89	70.71	60.98	88.78	77.98	45.67	67.11	59.93
	MOE-CL	29.10	91.76	75.69	67.29	87.59	93.95	87.90	88.33	89.02	64.79	69.73	76.83
	ZSCL	28.50	91.01	73.76	59.89	74.41	87.71	87.42	92.64	89.04	73.14	72.77	75.48
	KeepLoRA	39.64	91.91	73.83	65.16	78.51	93.35	89.34	87.32	92.80	75.73	70.72	78.03
	DFA-MoE	24.36	87.85	68.58	45.59	54.30	71.02	88.53	59.45	89.07	64.82	60.66	64.93
	ConDU	33.75	90.61	75.60	64.15	86.46	87.69	87.39	93.53	87.98	66.46	72.28	76.90
	Ours	35.97	90.67	78.41	65.96	88.20	92.37	87.09	93.01	86.67	69.80	72.49	**78.24**

in Fig. 1, these results demonstrate consistent performance across different training regimes, with strong zero-shot transfer accompanied by improved continual learning performance.

Table 3: Comparison with state-of-the-art methods on the multi-domain task incremental learning benchmark under Order-II in terms of "Transfer", "Average" and "Last" scores (%).

	Method	Aircraft	Caltech101	CIFAR100	DTD	EuroSAT	Flowers	Food	MNIST	OxfordPet	Cars	SUN397	*Average*
Transfer	LwF	–	87.28	59.95	68.32	43.42	49.02	11.24	77.16	31.28	30.20	40.39	49.83
	MOE-CL	–	88.52	58.50	89.04	70.51	60.59	19.45	88.42	44.06	34.98	60.71	61.48
	ZSCL	–	88.00	57.33	83.93	66.90	60.59	19.89	87.81	44.16	48.84	65.99	62.34
	KeepLoRA	–	88.98	56.62	91.24	71.82	60.46	23.11	89.21	45.25	50.72	68.77	64.62
	DFA-MoE	–	88.58	60.35	88.89	70.38	60.83	23.34	88.13	44.67	41.17	63.85	63.02
	ConDU	–	88.49	61.38	87.61	69.56	62.53	22.01	88.88	44.42	55.30	69.70	64.99
	Ours	–	88.53	62.30	88.20	70.95	63.54	23.16	89.41	45.57	55.65	70.78	**65.81**
Average	LwF	41.50	67.54	92.30	83.10	55.99	58.27	19.83	81.85	41.26	42.05	44.45	57.10
	MOE-CL	83.17	88.87	90.94	89.16	87.18	68.40	32.00	90.70	53.45	45.48	62.41	72.87
	ZSCL	77.94	90.28	91.05	90.64	82.80	69.13	32.00	89.42	52.81	57.55	67.77	72.85
	KeepLoRA	85.55	92.30	91.81	93.64	85.03	72.52	37.00	91.91	53.26	59.45	70.44	75.72
	DFA-MoE	58.00	87.70	88.14	88.26	69.18	67.06	23.71	88.53	44.80	48.41	65.23	66.28
	ConDU	84.81	90.18	92.62	91.95	82.13	70.49	33.74	90.84	52.09	63.00	70.99	74.80
	Ours	89.32	91.70	92.75	92.99	87.76	73.18	39.43	91.89	54.87	63.48	72.36	**77.25**
Last	LwF	21.50	57.82	99.32	84.33	55.08	64.19	31.38	88.19	59.15	92.20	85.09	67.11
	MOE-CL	83.21	88.90	98.07	89.21	96.71	74.86	47.04	94.70	78.51	92.74	79.42	83.94
	ZSCL	73.73	90.11	97.68	92.53	88.34	76.05	44.76	91.88	74.36	95.96	85.52	82.81
	KeepLoRA	84.13	92.23	99.60	94.21	90.68	81.73	51.37	96.61	72.82	98.07	87.42	86.26
	DFA-MoE	58.11	87.54	91.03	88.83	68.86	71.05	24.48	89.00	45.05	78.33	79.02	71.03
	ConDU	83.30	90.27	99.58	93.43	88.83	76.88	47.82	94.07	72.29	97.61	83.94	84.37
	Ours	89.09	92.01	99.52	94.79	97.15	81.20	59.02	96.26	79.63	98.70	88.24	**88.69**

Table 4: Ablation study in terms of "Transfer", "Average", and "Last" scores (%).

	Method	Aircraft	Caltech101	CIFAR100	DTD	EuroSAT	Flowers	Food	MNIST	OxfordPet	Cars	SUN397	*Average*
Average	Ours	58.48	95.28	84.30	69.49	83.94	84.84	89.62	77.10	89.70	68.57	65.59	78.81
	-w/o SR	57.61	94.62	84.00	68.93	83.75	84.32	89.60	77.88	89.58	68.39	65.54	78.57
	-w/o CCTD	56.90	94.90	82.23	66.76	83.48	83.36	89.17	75.07	89.56	67.02	65.26	77.61
Last	Ours	58.36	95.91	88.05	79.10	98.65	96.83	92.04	99.53	94.74	89.01	81.23	88.50
	-w/o SR	57.46	94.82	87.28	77.71	98.56	95.76	91.93	99.52	94.30	88.04	80.88	87.84
	-w/o CCTD	55.87	95.33	85.05	74.47	97.78	93.95	91.13	99.55	94.77	84.62	79.39	86.54

Robustness to Task Order. To further examine whether the effectiveness of our method depends on a specific task sequence, we evaluate it under Order-II while keeping the remaining experimental settings unchanged. As shown in Table 3, our method achieves 65.81%, 77.25%, and 88.69% on Transfer, Average, and Last, respectively, surpassing the strongest competing results by 0.82%, 1.53%, and 2.43%. The consistent gains under Order-II confirm that our method is robust to changes in task order and does not rely on a particular task sequence.

4.3 Analysis

Ablation Study. We conduct ablation experiments to evaluate the contributions of main components of our method. As shown in Table 4, removing the emoving the Cross-task Consolidation and Task-wise Decoupling (CCTD) module leads to a clear performance degradation, reducing the Average and Last scores from 78.81% and 88.50% to 77.61% and 86.54%. This confirms that CCTD plays a central role in organizing and consolidating multimodal knowledge throughout continual learning. Removing Selective Reconstruction (SR) also consistently decreases performance, yielding 78.57% Average and 87.84% Last. The consistent improvements brought by Selective Reconstruction further demonstrate the effectiveness of calibrating modality-specific knowledge at inference time.

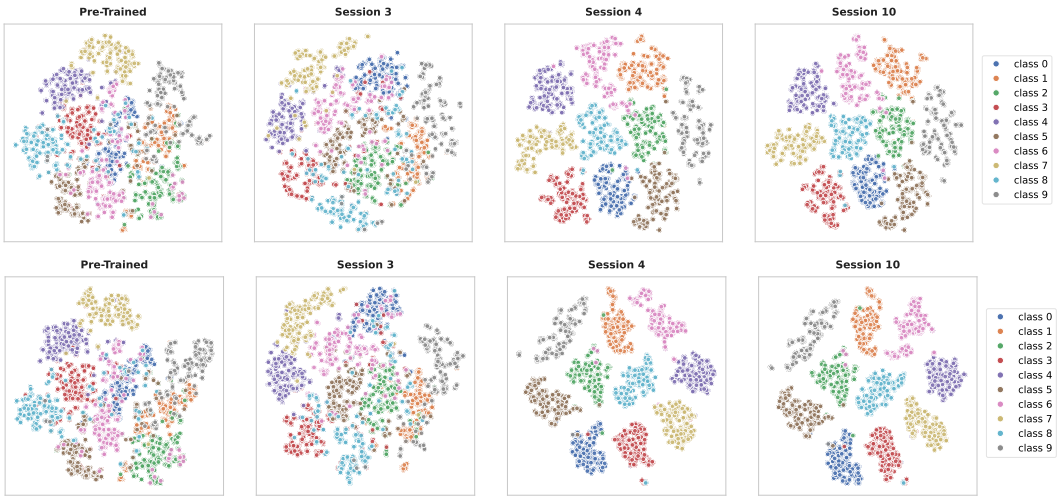

Figure 3: t-SNE visualization of EuroSAT features across different continual learning stages. The top and bottom rows correspond to ConDU and our method.

Visualization of Feature Evolution. To examine the evolution and preservation of task-specific representations, we compare our method with ConDU by visualizing EuroSAT feature space (Task 4 in Order-I) at four representative stages: the pre-trained model, Session 3 before learning EuroSAT, Session 4 immediately after learning EuroSAT, and Session 10, the final stage of continual learning. As shown in Fig. 3, the feature distributions are relatively mixed before EuroSAT is introduced, indicating that the model has not yet formed clear task-specific discriminative structures for this task. After learning EuroSAT at Session 4, our method produces more compact and better-separated feature clusters than ConDU and preserves this discriminative structure more effectively after subsequent tasks, demonstrating stronger retention of task-specific representations.

Hyperparameter Sensitivity Analysis. We study the effect of β in Selective Reconstruction under different scaling functions, measured by the average Last accuracy across all tasks. As shown in Fig. 4, all variants perform best with a moderate β, around 0.2, while larger values gradually degrade performance due to excessive suppression of modality-specific knowledge. The similar results across different functions in the effective range indicate that Selective Reconstruction is not sensitive to a particular functional form, with the exponential variant showing relatively stable performance over a broader range of β.

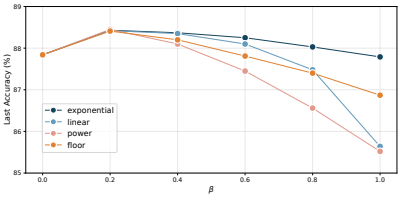

Figure 4: Sensitivity of Selective Reconstruction to β under different scaling functions $s(\pi_m^k)$.

5 Conclusion

In this work, we investigate multimodal continual learning for vision-language models from the perspective of dynamic modality contributions. We show that the relative roles of the visual and textual modalities are not fixed across tasks, but evolve with task demands and previously accumulated knowledge. Based on this observation, we propose D²C-MMCL, which decouples task-induced knowledge along both modality and task dimensions and consolidates cross-task knowledge within each modality according to its estimated contribution. Selective Reconstruction further calibrates modality-specific knowledge at inference time to construct task-adaptive models. Extensive experiments under different continual learning settings demonstrate that our method consistently improves both zero-shot transfer and knowledge retention, while remaining robust to task order, limited supervision, and different reconstruction configurations. These results highlight the importance of explicitly modeling dynamic modality contributions for continual adaptation of VLMs.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, grammatical editing, and improving the clarity and presentation of the manuscript. All technical content, experimental design, mathematical formulations, analyses, and scientific claims were developed and verified by the authors. All AI-assisted text was carefully reviewed and revised by the authors, who take full responsibility for the final content of the paper.

ETHICS STATEMENT

This work focuses on methodological research for continual learning of vision-language models and evaluates the proposed method on established public benchmark datasets. It does not involve human subjects, personally identifiable information, or the collection of sensitive data. We therefore do not identify any specific ethical concerns arising from the experimental setting of this work.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed method, training procedure, evaluation protocol, and implementation settings in the main paper and appendix. The datasets, task orders, evaluation metrics, optimization settings, and hyperparameters used in our experiments are explicitly specified. Additional analyses in the appendix include the scaling functions used for Selective Reconstruction, hyperparameter sensitivity, and further visualization results. These details are provided to facilitate reproduction of the reported results.

REFERENCES

Agil Aghasanli, Yi Li, and Plamen Angelov. Prototype-based continual learning with label-free replay buffer and cluster preservation loss. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 6547–6556. IEEE, 2025.

Ang Bian, Wei Li, Hangjie Yuan, Chengrong Yu, Mang Wang, Zixiang Zhao, Aojun Lu, Pengliang Ji, and Tao Feng. Make continual learning stronger via c-flat. *Advances in Neural Information Processing Systems*, 37:7608–7630, 2024.

Zhenyu Cui, Yuxin Peng, Xun Wang, Manyu Zhu, and Jiahuan Zhou. Continual vision-language retrieval via dynamic knowledge rectification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11704–11712, 2024.

Chen Feng, Georgios Tzimiropoulos, and Ioannis Patras. Clipcleaner: Cleaning noisy labels with clip. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 876–885, 2024.

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.

Chendi Ge, Xin Wang, Zeyang Zhang, Hong Chen, Jiapei Fan, Longtao Huang, Hui Xue, and Wenwu Zhu. Dynamic mixture of curriculum lora experts for continual multimodal instruction tuning. *arXiv preprint arXiv:2506.11672*, 2025.

Hadi Hadizadeh, S Faegheh Yeganli, Bahador Rashidi, and Ivan V Bajić. Mutual information analysis in multimodal learning systems. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 390–395. IEEE, 2024.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 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. *arXiv preprint arXiv:2106.09685*, 2021.

Linlan Huang, Xusheng Cao, Haori Lu, and Xialei Liu. Class-incremental learning with clip: Adaptive representation adjustment and parameter fusion. In *European Conference on Computer Vision*, pp. 214–231. Springer, 2024.

Linlan Huang, Xusheng Cao, Haori Lu, Yifan Meng, Fei Yang, and Xialei Liu. Mind the gap: Preserving and compensating for the modality gap in clip-based continual learning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3777–3786. IEEE, 2025.

Saurav Jha, Dong Gong, and Lina Yao. Clap4clip: Continual learning with probabilistic finetuning for vision-language models. *Advances in neural information processing systems*, 37:129146–129186, 2024.

Hyundong Jin, Hyung Jin Chang, and Eunwoo Kim. Instruction-grounded visual projectors for continual learning of generative vision-language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3466–3476. IEEE, 2025.

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*, volume 2026, pp. 106089–106106, 2026.

Borui Kang, Jinrui Gu, Tao Feng, Qi Fan, Yinghuan Shi, Lei Wang, Wenbin Li, and Yang Gao. Don't forget why you started: Tackling dual forgetting in vision-language continual learning. In *Forty-third International Conference on Machine Learning*.

Borui Kang, Lei Wang, Zhiping Wu, Tao Feng, Yawen Li, Yang Gao, and Wenbin Li. Dynamic multi-layer null space projection for vision-language continual learning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2077–2086. IEEE, 2025.

Yukun Li, Guansong Pang, Wei Suo, Chenchen Jing, Yuling Xi, Lingqiao Liu, Hao Chen, Guoqiang Liang, and Peng Wang. Coleclip: Open-domain continual learning via joint task prompt and vocabulary learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(8): 15137–15151, 2025.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Wenzhuo Liu, Fei Zhu, Longhui Wei, and Qi Tian. C-clip: Multimodal continual learning for vision-language model. In *International Conference on Learning Representations*, volume 2025, pp. 46461–46477, 2025a.

Yuyang Liu, Qiuhe Hong, Linlan Huang, Alexandra Gomez-Villa, Dipam Goswami, Tiantian Peng, Xialei Liu, Joost van de Weijer, and Yonghong Tian. Continual learning for vlms: A survey and taxonomy beyond forgetting. *arXiv preprint arXiv:2508.04227*, 2025b.

Mao-Lin Luo, Zi-Hao Zhou, Tong Wei, and Min-Ling Zhang. Lada: Scalable label-specific clip adapter for continual learning. *arXiv preprint arXiv:2505.23271*, 2025.

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*, volume 2026, pp. 136017–136040, 2026a.

Ruikun Luo, Jiarui Wang, Yuan Gao, Jing Yang, Jieming Yang, Song Wu, Hai Jin, and Xiaoyu Xia. Rlap-clip: Continual multimodal learning with prototype adaptation and difficulty-aware routing. In *International Conference on Learning Representations*, volume 2026, pp. 155978–156003, 2026b.

Daniel Marczak, Bartłomiej Twardowski, Tomasz Trzciński, and Sebastian Cygert. Magmax: Leveraging model merging for seamless continual learning. In *European Conference on Computer Vision*, pp. 379–395. Springer, 2024.

Sejik Park. Learning more generalized experts by merging experts in mixture-of-experts. *arXiv preprint arXiv:2405.11530*, 2024.

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.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, 58(5):1–42, 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 *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11909–11919. IEEE, 2023.

Haopeng Sun, Yingwei Zhang, Lumin Xu, Sheng Jin, Ping Luo, Chen Qian, Wentao Liu, and Yiqiang Chen. Unsupervised continual domain shift learning with multi-prototype modeling. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10131–10141. IEEE, 2025.

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. Springer, 2024.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5362–5383, 2024.

Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. *Advances in Neural Information Processing Systems*, 35:5682–5695, 2022a.

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, 2022b.

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 *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 139–149. IEEE, 2022c.

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

Yicheng Xu, Yuxin Chen, Jiahao Nie, Yusong Wang, Huiping Zhuang, and Manabu Okumura. Advancing cross-domain discriminability in continual learning of vision-language models. *Advances in Neural Information Processing Systems*, 37:51552–51576, 2024.

Yuki Yada, Sho Akiyama, Ryo Watanabe, Yuta Ueno, Yusuke Shido, and Andre Rusli. Improving visual recommendation on e-commerce platforms using vision-language models. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, pp. 975–978, 2025.

Shipeng Yan, Lanqing Hong, Hang Xu, Jianhua Han, Tinne Tuytelaars, Zhenguo Li, and Xuming He. Generative negative text replay for continual vision-language pretraining. In *European Conference on Computer Vision*, pp. 22–38. Springer, 2022.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23219–23230. IEEE, 2024a.

Yu-Chu Yu, Chi-Pin Huang, Jr-Jen Chen, Kai-Po Chang, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. Select and distill: Selective dual-teacher knowledge transfer for continual learning on vision-language models. In *European Conference on Computer Vision*, pp. 219–236. Springer, 2024b.

Hongsheng Zhang, Zhong Ji, Jingren Liu, Yanwei Pang, and Jungong Han. Multi-stage knowledge integration of vision-language models for continual learning. *IEEE Transactions on Image Processing*, 2026.

Mengyu Zheng, Yehui Tang, Zhiwei Hao, Kai Han, Yunhe Wang, and Chang Xu. Adapt without forgetting: Distill proximity from dual teachers in vision-language models. In *European Conference on Computer Vision*, pp. 109–125. Springer, 2024.

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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19068–19079. IEEE, 2023.

Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Class-incremental learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(12):9851–9873, 2024.

Da-Wei Zhou, Yuanhan Zhang, Yan Wang, Jingyi Ning, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Learning without forgetting for vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4489–4504, 2025.

A Appendix

A.1 Additional Visualization Results

To further examine the representation dynamics under different modality-dependence patterns, we provide additional t-SNE visualizations for three representative tasks in Fig. 5: CIFAR100, DTD, and OxfordPet, corresponding to visual-dominant, balanced, and text-dominant tasks, respectively. For each task, we visualize the feature space produced by our method at four representative stages: the pre-trained model, the session immediately before learning the target task, the session immediately after learning it, and the final session after completing the entire task sequence.

Across the three tasks, the visualizations show a consistent evolution of task-specific representations. Before the target task is learned, class distributions are generally less structured and exhibit more overlap. After learning the task, clearer discriminative structures emerge, with improved intra-class compactness and inter-class separation. More importantly, these structures remain largely preserved at the final stage despite subsequent continual learning. The consistent behavior across visual-dominant, balanced, and text-dominant tasks further indicates that our method can effectively acquire and retain task-specific representations under different modality contribution patterns.

A.2 Selective Reconstruction Functions

In Selective Reconstruction, we rescale the reconstructed modality-specific delta according to its estimated modality contribution. For modality m of task k, we define

$$\tilde{\delta}_m^k = s(\pi_m^k; \beta)\hat{\delta}_m^k, \tag{15}$$

where $\hat{\delta}_m^k$ denotes the reconstructed modality delta, $\pi_m^k \in [0, 1]$ is the relative contribution of modality m, and β controls the strength of the scaling. We consider six candidate scaling functions, summarized in Table 5.

These functions induce different scaling behaviors with respect to the modality contribution. The linear function applies direct attenuation, whereas the power and exponential functions provide nonlinear modulation. The floor function further imposes a lower bound on the scaling factor to avoid excessive suppression of one modality. The sigmoid function produces a smooth transition around $\pi = 0.5$, while the cosine function provides a smooth bounded mapping. In our main experiments, we adopt the exponential function with $\beta = 0.2$, which achieves strong and relatively stable performance in our sensitivity analysis.

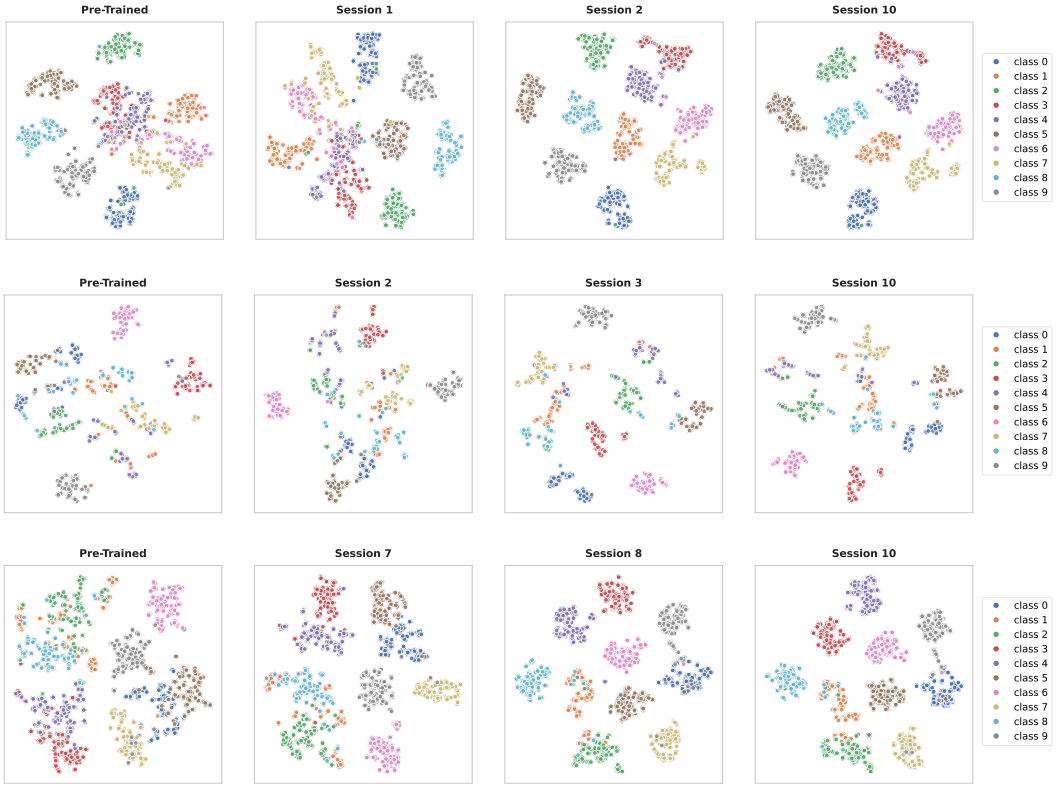

Figure 5: t-SNE visualizations of CIFAR100, DTD, and OxfordPet under our method, representing visual-dominant, balanced, and text-dominant tasks, respectively. For each task, the four columns correspond to the pre-trained model, the session before learning the target task, the session after learning it, and the final session of continual learning.

Table 5: Candidate scaling functions considered for Selective Reconstruction.

Function	**Scaling factor** $s(\pi;\beta)$
Linear	$1-\beta(1-\pi)$
Power	π^{β}
Floor	$0.1+0.9\pi^{\beta}$
Sigmoid	$\sigma(\beta(\pi-0.5))$
Exponential	$\exp[-\beta(1-\pi)]$
Cosine	$0.5+0.5\cos(\pi\beta(1-\pi))$