Mitigate Modality-Asymmetric Forgetting via Stabilizing Visual Representations in CLIP-Based Class-Incremental Learning

Yuanhong Zhang[1,2,3]*, Yanan Chen[1,2]*, Xin Zhang[3], Zhaoyang Wang[1,2], Weizhan Zhang[1,2]†, Muyao Yuan[1,2], Hongjin Niu[1,2], Lan Ma[4], Yuan Gao[4], and Joey Tianyi Zhou[3]

[1] School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, P.R. China
[2] Ministry of Education Key Laboratory of Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, P.R. China
[3] Institute of Advanced Intelligence and Computing (IAIC), Agency for Science, Technology and Research (A*STAR), Singapore 138632, Singapore
[4] China Telecom, Xi'an, Shaanxi, 710000, P.R. China
zhangwzh@xjtu.edu.cn

Abstract. Continual Learning (CL) enables models to learn sequentially while retaining prior knowledge. Recently, CL under the CLIP framework has gained increasing attention for its remarkable cross-task generalization capability. Despite recent progress, existing approaches overlook ***modality-asymmetric forgetting***, where the visual modality undergoes severe degradation during CL. In this paper, we empirically identify this phenomenon and provide evidence that the lack of a stable structural anchor in the visual branch may drive high-level visual features to over-adapt to new tasks, thereby exacerbating catastrophic forgetting. To address this issue, we propose **Visual Anchor-based Structural Transfer (VAST)** to stabilize visual representations during continual learning. Specifically, we integrate a cross-layer representation enhancement module to construct the stable visual anchor, and further employ an information-theoretic structural distillation mechanism to transfer them across tasks, promoting stable adaptation and mitigating catastrophic forgetting. Extensive experiments across multiple benchmarks demonstrate that our method consistently surpasses SOTA approaches without relying on any replay data. Code is available at VAST.

Keywords: Continual Learning · Modality-Asymmetric Forgetting · CLIP

1 Introduction

The goal of continual learning is to enable models to acquire new knowledge from sequential tasks while mitigating catastrophic forgetting [5,46]. A representative and challenging setting is class-incremental learning (CIL), where new classes are

* Equal contribution. † Corresponding author.

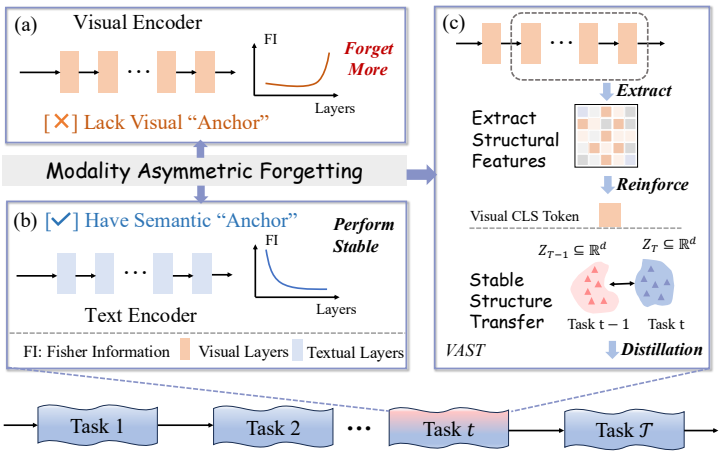


Fig. 1: Motivation of VAST. Left: (a) In CLIP-based continual learning, the visual branch suffers more from catastrophic forgetting; (b) the text branch remains stable as its linguistic embeddings provide a fixed semantic anchor. Right: (c) To address this, VAST extracts cross-layer comparative visual anchors to reinforce visual tokens, facilitating stable knowledge preservation across tasks.

introduced over time and the model must recognize all previously learned categories without task identifiers [23, 24, 37]. Conventional approaches [33, 40] seek to balance plasticity and stability through task-specific training from scratch, but often struggle with scalability and data efficiency. Recent advances therefore turn to pre-trained vision–language models such as CLIP [39], whose large-scale image–text contrastive training empowers them with impressive zero-shot generalization, providing a more robust starting point for continual adaptation [44]. While recent efforts have begun to explore continual learning within the CLIP framework, this paradigm also introduces new modality-specific challenges.

Recent CLIP-based continual learning approaches [21, 23, 25, 34, 51, 60] typically adopt parameter-efficient fine-tuning (PEFT) strategies [16, 17], which insert lightweight modules (e.g., LoRA or Adapter) for textual and/or visual paths to adapt new knowledge while retaining CLIP's zero-shot capability. Unlike conventional single-stream architectures [8, 14], such dual-branch vision-language models require closer examination of how the relationship between modalities evolves during continual learning. Yet, explorations in this direction remain limited. While some attempts (e.g., MG-CLIP [21], DMNSP [25]) have sought to mitigate modality gaps or gradient misalignment, modality-asymmetric forgetting during continual adaptation remains insufficiently studied from a representation-space perspective.

In this paper, we investigate CLIP-based class-incremental learning by examining modality dynamics (see Fig. 1). We observe a critical asymmetry: the visual branch suffers more severe catastrophic forgetting than the textual branch. Our analysis attributes this phenomenon to an inherent modality imbalance.

Specifically, the textual encoder empirically serves as a stable semantic anchor grounded in concept-level linguistic representations, providing a consistent reference structure in the embedding space [21, 25], whereas the visual encoder lacks a comparable protective mechanism. The absence of such stable anchor makes the higher visual layers highly sensitive to parameter updates, manifested as large Fisher Information (see Fig. 2). This behavior causes overfitting to the current classes in the representation space and eventually results in catastrophic forgetting (see Fig. 3). Moreover, existing distillation-based methods [57, 60] fail to correct this drift and instead amplify it over successive tasks.

To address the above limitations, we propose VAST (Visual Anchor-based Structural Transfer), a novel CLIP-based class-incremental learning framework that stabilizes the visual feature space by constructing the visual anchor and transferring their structural knowledge across tasks. In this paper, we view the visual anchor as a relatively stable representation that preserves the geometric structure of the visual feature space across tasks. To construct such anchors, we introduce a cross-layer Visual Representation Reinforcement Module (VRRM), which leverages self-similarity attention to extract stable spatial relations from intermediate visual layers, where structural information is less affected by task-specific adaptation. These locality-aware structural features are fused with high-level cross-modal representations to reinforce the CLS token, stabilizing visual representations against high-layer over-adaptation. Building upon the learned visual anchors, we introduce an information-theoretic distillation loss to align feature distributions across tasks, enabling the model to inherit structured knowledge from previous tasks rather than relying on point-wise matching. Through this dual design, our method alleviates modality-asymmetric forgetting in CLIP-based continual learning. The contributions can be summarized as follows:

- We identify the phenomenon of modality-asymmetric forgetting in CLIP-based continual learning. Our empirical analysis suggests that, compared with the textual branch, the visual branch lacks a comparably stable structural anchor and is therefore more vulnerable to forgetting.
- We propose VAST, a framework that mitigates modality-asymmetric forgetting through two complementary mechanisms: constructing stable visual anchors and transferring them across tasks via information-theoretic structural distillation.
- We conduct extensive experiments across multiple benchmarks, achieving state-of-the-art performance in both accuracy and forgetting rate.

2 Related Work

2.1 Class-Incremental Continual Learning.

Class-incremental continual learning (CIL) [27] has received extensive attention recently. Representative strategies can be broadly categorized as follows: (1) Regularization-based methods [10, 11, 33, 59], which impose explicit constraints on model weights or representations to balance the stability–plasticity trade-off

between old and new tasks. (2) Memory-based methods [35, 40, 47, 54], which replay or approximate previously learned data or feature distributions to preserve prior knowledge. However, their scalability is constrained by the growing memory requirements. (3) Dynamic-based methods [1, 9, 18, 50], which incrementally add new parameters (e.g., neurons, branches, or heads) to accommodate new tasks, often at the cost of increased model complexity. Additionally, with the rise of pre-trained models, parameter-efficient fine-tuning approaches gradually expand a small number of parameters as tasks increase [43, 51, 57]. They show that preserving the stability of pre-trained models enhances generalization and facilitates class-incremental learning on downstream tasks.

2.2 Vision-Language Model-based Class-Incremental Learning

With the emergence of vision–language models such as CLIP, CIL has extended from purely visual domains to multimodal paradigms [44]. To preserve pre-trained knowledge while adapting to new tasks, most methods freeze the backbone and adopt lightweight parameter-efficient modules, including prompts [43, 48, 49, 58] and adapters [4, 21, 51]. Prompt-based methods (e.g., L2P [49], DualPrompt [48], ENGINE [58]) learn task-aware text or visual prompts for dynamic adaptation, whereas adapter-based approaches [34, 51, 55] insert small trainable modules into transformer layers to encode task-specific knowledge. For example, MoE4CL [51] employs mixture-of-experts routing for adaptive learning, while LGVLM [55] trains separate LoRA modules for each incremental task.

A few other studies have begun to explore improving cross-modal consistency in CLIP-based continual learning. For instance, DMNSP [25] adopts a gradient-based strategy to regulate parameter variation, while other studies focus on enhancing cross-modal interaction (e.g., PROOF [60], CLAP4CLIP [23]) or reducing the modality gap (e.g., MG_CLIP [21]) to achieve better alignment between visual and textual features. However, few studies explore how representation spaces evolve during continual learning, particularly under modality imbalance where the visual branch forgets more. In contrast, we investigate the cause of such modality-asymmetric forgetting and introduce structural anchoring to mitigate visual drift throughout continual adaptation.

2.3 Self-correlation Attention in CLIP

Recent training-free CLIP-based methods for dense prediction have increasingly explored modifying the attention mechanism to improve the spatial quality of visual representations. CLIPSurgery [32] first attributes noisy visual responses in CLIP to heterogeneous query-key projections and replaces the original attention with homogeneous self-correlation. Building on this insight, GEM [3] generalizes value-value attention into a self-self attention path, where repeated self-correlation progressively clusters similar tokens. SCLIP [45] leverages pairwise token similarities, such as $QQ^\top$ and $KK^\top$, to encourage spatially covariant features for dense prediction. ClearCLIP [29] attributes part of the noise

in CLIP-based dense prediction to residual connections and adopts self-self attention to obtain cleaner structural features. CLIPtrase [41] restores patch-wise relations by fusing self-correlations from multiple projections. ProxyCLIP [30] and NACLIP [13] further improve spatial coherence through self-similarity priors and key-key attention with locality constraints, respectively. Different from prior methods that mainly focus on dense prediction or feature refinement in static tasks, VAST leverages intermediate-layer structural correlation in class-incremental learning to construct relatively stable visual anchors and mitigate forgetting during continual adaptation.

3 Preliminaries

3.1 Class-Incremental Learning for CLIP

We adopt a class-incremental learning (CIL) setup built upon a pre-trained CLIP model. The objective is to train a model over a sequence of T tasks $\{(\mathcal{C}_1, \mathcal{D}_1), (\mathcal{C}_2, \mathcal{D}_2), \ldots, (\mathcal{C}_T, \mathcal{D}_T)\}$. Each task $t \in [1, T]$ provides a dataset $\mathcal{D}_t = \{(x_i, y_i)\}_{i=1}^{N_t}$ of image-label pairs, where x_i denotes an input image and y_i its label drawn from the task-specific class set $\mathcal{C}_t = \{c_t^1, c_t^2, \ldots, c_t^{K_t}\}$. Here, N_t and K_t denote the number of training samples and classes in task t, respectively. Following prior works [21,23], the class sets are assumed to be mutually disjoint, i.e., $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$ for $i \neq j$.

For each task t, the CLIP model $\mathcal{M}_t$ is initialized with parameters θ_{t-1} from the previous stage and optimized by minimizing the cross-entropy loss over $\mathcal{C}_t$. Specifically, for each input $(x, y) \in \mathcal{D}_t$, CLIP extracts the visual representation V^t from the image and the textual embedding E^t from the corresponding label prompt using its current parameters θ_t. After learning task t, the updated model $\mathcal{M}_t$ should correctly classify samples from all seen categories $C_{1:t} = C_1 \cup C_2 \cup \cdots \cup \mathcal{C}_t$, without using task identifiers. Additionally, we follow the exemplar-free CIL setting [21, 48] in this paper.

3.2 Matrix-based Joint Entropy.

Given feature sets $X = \{x_i\}_{i=1}^n$ and $Y = \{y_i\}_{i=1}^n$, let $G_X, G_Y \in \mathbb{R}^{n \times n}$ denote Gram matrices, where each entry is computed as $(G_X)_{ij} = \kappa(x_i, x_j)$ and $(G_Y)_{ij} = \kappa(y_i, y_j)$ using a positive-definite kernel function $\kappa(\cdot, \cdot)$. Following matrix-based $\alpha-$order Rényi entropy [12], the joint entropy between X and Y is computed as:

$$S_\alpha(G_X, G_Y) = \frac{1}{1-\alpha} \log_2 \left(\mathrm{tr}\left[\left(\frac{G_X \circ G_Y}{\mathrm{tr}(G_X \circ G_Y)} \right)^\alpha \right] \right), \tag{1}$$

where $\circ$ denotes the Hadamard product, and $tr(\cdot)$ is the matrix trace operator. It can be estimated directly from data without making strong assumptions about the underlying distribution [7, 52]. This quantity measures the structural dependency between X and Y: a lower joint entropy indicates stronger alignment and less structural divergence.

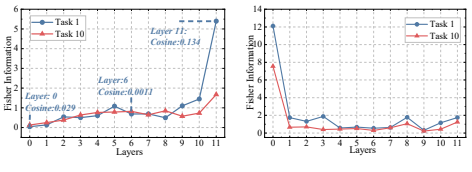


Fig. 2: Layer-wise Fisher information of CLIP with LoRA finetuning on CIFAR100. Left: visual encoder. Right: text encoder. The cosine similarity measures the alignment between layer-wise visual features and the final text embedding.

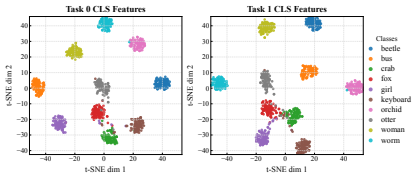


Fig. 3: t-SNE [36] plot of CLS features on Task 0 classes from CIFAR100. Left: model after LoRA finetuning on Task 0. Right: same model finetuned on Task 1 but evaluated on Task 0, showing representational drift and class confusion.

4 Methodology

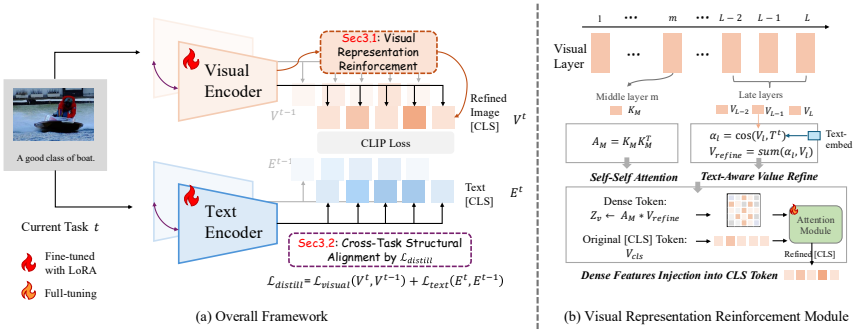
4.1 Empirical Analysis

Modality-Asymmetric Forgetting: A Missing Structural Anchor in the Visual Branch. In this section, we investigate modality-asymmetric forgetting in CLIP-based continual learning from the perspective of cross-modal adaptation dynamics. We begin by examining the Fisher Information (FI) [19,27] of the visual and textual branches during continual adaptation. FI measures the sensitivity of model parameters to gradient updates: higher values indicate greater sensitivity to task-specific updates, which is often associated with less stable representations during continual adaptation [25].

As shown in Fig. 2, the textual encoder exhibits consistently low Fisher Information in higher layers, indicating limited sensitivity to task-specific updates. This relative stability is consistent with the role of pre-trained linguistic representations, which may provide a semantic reference that helps preserve global consistency across tasks. In contrast, the visual encoder shows a sharp rise in Fisher Information in late layers, suggesting that it bears stronger adaptation pressure during continual learning. Correspondingly, we observe increased visual–textual cosine similarity (see Fig. 2), suggesting stronger visual–textual alignment during adaptation. This trend may be accompanied by reduced structural stability in the visual branch.

These observations suggest that the visual encoder lacks a source of structural stability comparable to that of the textual branch during continual adaptation. Motivated by this insight, we introduce the concept of a visual structural anchor, defined as a stable representation that preserves the intrinsic geometry of the visual feature space across tasks. By stabilizing the visual feature space, such an anchor can help reduce representation drift and thereby mitigate catastrophic forgetting during continual learning.

Visual Structural Collapse and the Failure of Knowledge Distillation in CIL. Without such a stabilizing mechanism, the over-adaptation of the visual branch leads to severe structural instability in the representation space. Specifically, the CLS tokens of old classes suffer from significant representational drift,



(a) Overall Framework (b) Visual Representation Reinforcement Module

Fig. 4: Overview of VAST. VAST addresses modality-asymmetric forgetting in CLIP-based continual learning through two complementary designs: (1) a cross-layer Visual Representation Reinforcement Module (VRRM) that injects locality-aware and text-aware features into the visual CLS token to strengthen its structural representation; and (2) a distribution-aware distillation objective that preserves cross-task structural information via matrix-based entropy.

which can be observed as the separation and displacement of clusters in the t-SNE projection (see Fig. 3).

Crucially, existing CL methods fail to address this fundamental structural issue. Parameter-efficient fine-tuning (e.g., LoRA) directly leads to CLS token drift due to the absence of a constraint mechanism. Moreover, knowledge distillation (KD) methods [33, 34, 38] are also insufficient: they merely enforce the new model to imitate the features of the old model. Since the old features have already been shifted and are structurally unstable, KD-based methods essentially transmit or even amplify the existing instability. This highlights a critical research gap: the necessity to explicitly introduce and protect a robust structural visual anchor to stabilize the CLIP-based continual learning process.

4.2 Visual Representation Reinforcement

In this section, we aim to construct a reliable visual anchor by reinforcing the global CLS representation with stable structural and semantic cues. To this end, we propose the Visual Representation Reinforcement Module (VRRM), which consists of three synergistic stages: (1) Structure-Preserving Attention to extract task-invariant spatial information; (2) Text-Guided Value Refinement to align visual features with target semantics; and (3) Global Semantic Reinforcement to fuse these reinforced cues into the CLS token.

Structure-Preserving Attention. To maintain the spatial consistency of visual representations, we introduce a structure-preserving attention mechanism that leverages the stable topology of CLIP's intermediate layers. Unlike task-specialized higher layers, mid-level representations (e.g., layers 5–8) retain richer, task-invariant spatial relations [29], providing a robust geometric foundation.

Let $\mathtt{Q}_M$ and $\mathtt{K}_M$ denote the query and key representations in the M-th transformer block. Standard attention, $softmax(\mathtt{Q}_M\mathtt{K}_M^\top/\sqrt{d})$, typically relies on query-key interactions. In continual learning, however, queries are highly susceptible to task-specific updates, leading to unstable attention patterns and semantic drift [29]. To preserve the intrinsic geometry of the visual space, we introduce Self-Similarity Attention (SSA) to construct a query-independent structural foundation. SSA computes patch relations directly via $A_{self} = \text{softmax}(\mathtt{K}_M\mathtt{K}_M^\top/\sqrt{d})$, yielding a symmetric affinity that promotes spatial consistency and mitigates the impact of task-induced query shifts [3,29]. To further enhance locality-aware information, we compute second-order relations $A_{rel} = A_{self} \times A_{self}$ to propagate affinities across neighboring patches, where $\times$ denotes matrix multiplication [61]. The finalized structure-preserving attention is:

$$A_M = softmax\left(\frac{A_{self} + A_{rel}}{\sqrt{d}}\right), \tag{2}$$

where d is the scaling dimension. By incorporating A_{rel}, we reinforce the connectivity of local semantic clusters, providing a stable spatial topology to ground the visual structural anchor.

Text-Guided Value Refinement. While A_M captures the stable spatial information, the underlying visual content must remain aligned with the target semantic space. We therefore refine the value features by adaptively re-weighting high-level layers $\{V_l\}$ (where $l \in \{9, 10, 11\}$) based on their alignment with the textual embedding E^t. For each layer l, the patch-level features V_l are projected using the frozen value matrix $W_v^{(L)}$ from the final transformer layer ($L = 11$). Their semantic relevance is measured by $s_l = \mathbb{E}_{v\sim V_l}[\cos(v, E^t)]$. Therefore, the refined visual representation V_{refine} is obtained via:

$$V_{refine} = \sum\nolimits_l \alpha_l V_l, \quad \text{where } \alpha_l = \frac{\exp(s_l)}{\sum_{l'} \exp(s_{l'})}. \tag{3}$$

By aggregating the refined value features V_{refine} according to the structure-preserving attention A_M, we generate the reinforced dense features:

$$Z_v = \text{Proj}(A_M \cdot V_{refine}), \tag{4}$$

where $\text{Proj}(\cdot)$ is the frozen projection layer of the CLIP visual encoder that maps visual features into the shared embedding space. It constrains the resulting dense feature by the intrinsic geometry of the intermediate layers.

Global Semantic Reinforcement. To further enhance the global representation, we introduce a lightweight fusion module to integrate the reinforced dense features Z_v into the global CLS token. This process employs an attention-based residual connection:

$$V'_{cls} = V_{cls} + \beta \sum\nolimits_i softmax\big(f_\phi([V_{cls}; z_i])\big)\, z_i, \tag{5}$$

where $z_i \in Z_v$ denotes each dense token, and $f_\phi(\cdot)$ is a two-layer MLP that generates attention weights (see the supplementary material for more details). The β controls the strength of the injection of information.

This fusion mechanism allows the global CLS token to selectively absorb spatially distributed semantic cues from Z_v. The reinforced feature V'_{cls} is then propagated to subsequent tasks as the stabilized representation V^t. This representation serves as the visual structural anchor, as it fuses global semantics with the task-invariant spatial topology of intermediate layers. By effectively preserving the intrinsic geometry of the visual feature space, V^t provides a robust and consistent target for structural distillation.

4.3 Distribution-Aware Structural Distillation

Building upon the stabilized visual representation V^t, we introduce a distribution-aware distillation loss to preserve the structural consistency of feature distributions across tasks. In continual learning, high-level features exhibit Fisher information instability, causing the internal representation subspace to be constantly reshaped. This structural drift is the root of visual forgetting, and capturing it requires a globally structure-sensitive metric for semantic change, which pointwise differences like KL or MSE cannot provide.

Motivated by the information-theoretic representation learning [22, 53, 56], we employ the matrix-based joint entropy [7, 12] as the structural divergence between representations at task $t-1$ and t. Given the visual features extracted from CLIP at task t and $t-1$, denoted as V^t and V^{t-1}, we project the current features V^t into a task-specific embedding space and compute their normalized Gram matrices G_V^t and G_V^{t-1} as described in Sec. 3.2. The visual-side alignment consists of a joint entropy loss and a randomized regularization term:

$$\mathcal{L}_{vis}^{joint} = S_\alpha(G_V^t, G_V^{t-1}), \quad \mathcal{L}_{vis}^{reg} = S_\alpha(G_V^t, \Pi G_V^{t-1} \Pi^\top),$$

Here, $S_\alpha(\cdot,\cdot)$ denotes the matrix-based joint entropy. The $\mathcal{L}_{vis}^{joint}$ directly preserves the visual structural information between tasks. However, to prevent the model from over-fitting to noise-like dependencies within the feature distributions, we introduce $\mathcal{L}_{vis}^{reg}$. This term employs a random permutation matrix Π to disrupt spurious correlations [42], ensuring that the distillation focuses on robust, high-level semantic structures.

Therefore, the final distribution-aware distillation loss for the visual branch is formulated as:

$$\mathcal{L}_{vis} = \mathcal{L}_{vis}^{joint} - \mathcal{L}_{vis}^{reg},$$

Similarly, the text-side loss $\mathcal{L}_{text}$ is defined analogously using textual embeddings (E^t, E^{t-1}). Overall, the distribution-aware distillation loss is:

$$\mathcal{L}_{distill} = \mathcal{L}_{vis} + \mathcal{L}_{text},$$

which jointly enforces cross-task consistency in both the visual and textual branches. Therefore, our overall training objective for task t is:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{distill}, \tag{6}$$

Table 1: Comparison of average and final performance among different methods over diverse datasets. The best-performing results are shown in bold.

Method	CIFAR100				ImageNet-R				ImageNet-1K	
	B10_Inc10		B50_Inc10		B20_Inc20		B100_Inc20		B100_Inc100	
	Last	Avg	Last	Avg	Last	Avg	Last	Avg	Last	Avg
Continual-CLIP [44]	66.66	75.15	66.66	69.66	71.73	78.67	71.73	75.20	67.69	75.55
L2P++ [49]	73.08	81.90	73.27	74.41	64.32	65.55	58.73	60.23	69.60	79.30
DualPrompt [48]	72.51	81.45	76.63	77.83	67.98	68.76	62.87	64.39	69.79	79.39
CODA-Prompt [43]	79.01	85.94	77.83	79.63	71.83	73.14	66.55	68.09	66.96	76.99
PROOF [60]	79.11	86.77	79.73	83.32	78.50	84.11	78.03	81.06	63.01	73.66
RAPF [20]	79.39	85.87	79.81	83.04	79.83	85.65	80.45	82.65	72.80	81.67
ZSCL [57]	77.07	84.71	82.16	85.81	80.42	84.66	81.52	84.24	71.56	80.65
MoE4CL [51]	77.59	85.30	79.60	83.67	81.05	86.14	82.03	84.79	73.41	81.68
MG-CLIP [21]	79.56	86.99	79.72	83.61	82.62	**87.65**	82.78	84.88	73.33	81.71
ENGINE [58]	79.29	86.87	79.22	82.95	80.71	86.26	80.98	83.68	–	–
DMNSP [25]	79.57	87.23	82.39	86.43	81.58	87.05	83.20	85.39	72.87	81.57
VAST (Ours)	**80.40**	**87.40**	**82.48**	**86.62**	**82.83**	87.61	**83.85**	**85.54**	**73.85**	**81.92**

Method	TinyImageNet				Food101				ImageNet-1K	
	B20_Inc20		B100_Inc20		B10_Inc10		B50_Inc10		B200_Inc200	
	Last	Avg	Last	Avg	Last	Avg	Last	Avg	Last	Avg
Continual-CLIP [44]	65.92	74.64	65.92	69.99	86.19	90.73	86.19	88.14	67.69	74.53
L2P++ [49]	70.98	75.38	68.52	72.02	78.13	85.77	73.13	80.42	69.24	72.84
DualPrompt [48]	72.53	76.72	70.98	74.18	78.49	85.95	72.75	80.00	69.44	72.90
CODA-Prompt [43]	75.14	79.05	72.38	76.33	78.77	86.11	74.13	80.98	69.78	73.29
PROOF [60]	58.76	72.03	57.21	66.52	84.73	90.04	84.74	87.52	69.12	71.40
RAPF [20]	73.73	81.51	73.45	77.83	82.35	88.77	82.18	86.82	73.09	81.41
ZSCL [57]	75.87	82.52	77.01	81.31	88.17	91.97	89.20	91.68	74.16	81.90
MoE4CL [51]	76.04	83.02	76.49	80.70	87.77	93.11	88.34	91.70	73.89	82.02
MG-CLIP [21]	75.87	83.56	77.09	81.41	88.19	93.30	87.43	91.26	74.25	81.77
ENGINE [58]	–	–	–	–	83.92	89.77	83.99	86.93	–	–
DMNSP [25]	75.81	82.54	78.05	81.62	87.75	92.74	89.31	91.78	74.55	82.22
VAST (Ours)	**76.05**	**83.98**	**78.72**	**82.19**	**88.71**	**93.42**	**89.81**	**91.86**	**74.87**	**82.43**

where the cross-entropy $\mathcal{L}_{ce}$ is computed within the current task [23, 60], and $\mathcal{L}_{distill}$ measures the distributional consistency between features from task t and $t-1$. The coefficient λ balances task-specific learning and cross-task knowledge preservation, and is set to 0.5.

5 Experiments

5.1 Experimental Setting

Datasets. Following standard CIL protocols [20, 21, 60], we evaluate on five benchmarks: CIFAR100 [28], Tiny ImageNet [31], ImageNet-R [15], Food-101 [2], and ImageNet-1K [6]. We sample 100 classes from CIFAR100 and Food-101, 200 from ImageNet-R and Tiny ImageNet, and 1000 from ImageNet-1K [21,58]. The setting "Bm_Inc$n$" denotes a CIL split where m classes are used initially and n new classes are added in each subsequent stage.

Baselines and Evaluation Metrics. We compare VAST with several methods: (1) zero-shot CLIP baseline (i.e., Continual-CLIP [44]); (2) vision-only

Table 2: Comparison of zero-shot performance among CLIP-based continual learning methods. Bold numbers indicate the best continual learning results, excluding the pre-trained CLIP.

Methods	Datasets			Avg
	Food101	ImageNet-100	ImageNet-1K	
CLIP	86.28±0.06	73.78±0.14	66.01±0.05	75.35±0.08
PROOF	22.88±0.02	68.94±0.02	65.66±0.02	52.49±0.02
ZSCL	82.38±0.10	67.50±0.12	62.43±0.10	70.77±0.10
RAPF	80.69±0.10	69.24±0.09	59.14±0.12	69.69±0.10
DMNSP	81.45±0.04	68.78±0.27	63.07±0.04	71.10±0.12
Ours	**84.67**±0.08	**72.22**±0.06	**65.96**±0.04	**74.28**±0.06

Table 3: Ablation of VRRM and distillation loss $\mathcal{L}_{\text{distill}}$ under B10_Inc10 (CIFAR100) and B20_Inc20 (ImageNet-R), showing complementary gains.

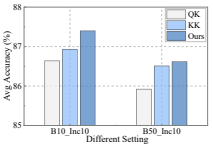
VRRM	$\mathcal{L}_{\text{distill}}$	CIFAR100		ImageNet-R	
		Avg	Last	Avg	Last
✗	✗	84.81±0.21	77.33±0.18	84.01±0.19	79.15±0.22
✓	✗	86.90±0.08	79.35±0.10	87.24±0.07	81.95±0.10
✗	✓	86.96±0.07	78.93±0.11	87.32±0.09	82.32±0.09
✓	✓	**87.40**±0.06	**80.40**±0.08	**87.61**±0.07	**82.83**±0.07

prompt-based approaches (i.e., L2P++ [49], DualPrompt [48], CODA-Prompt [43]); (3) CLIP-based continual learning methods, including the replay-based approaches (i.e., PROOF [60], RAPF [20]) and non-replay approaches (i.e., ZSCL [57], MoE4CL [51], MG-CLIP [21], ENGINE [58]). All experiments adopt OpenAI's pre-trained CLIP ViT-B/16 backbone for fair comparison. The average accuracy after training task t over all seen tasks is denoted as A_t. We report *Avg* as the mean accuracy across all tasks and *Last* as the accuracy after the final task.

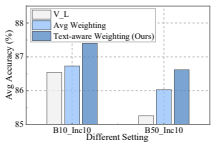
Implementation Details. Our framework is implemented in PyTorch and trained on an NVIDIA A800 GPU. Following previous work [21], we adopt OpenAI's pre-trained CLIP ViT-B/16 as the backbone and insert LoRA [17] modules into both the visual and textual branches for parameter-efficient fine-tuning, with the rank set to 8 by default. Additional results with ViT-B/32 are provided in the Supplementary Materials. Each task is trained for 5 epochs with a batch size of 128, using the Adam optimizer and a cosine learning rate scheduler starting at 0.001. All reported results are averaged over five runs with different random seeds. For VRRM, we compute self-similarity attention at layer 7 and apply text-aware refinement at layers 9–11. The fusion coefficient β in Eq. (5) is set to 0.1, and the weight of $\mathcal{L}_{\text{distill}}$ is 0.5. For distillation, following previous work [52], we set the order parameter of the matrix-based α-order Rényi entropy to $\alpha = 1.01$.

5.2 Main Results

Comparison with State-of-the-Arts. Tab. 1 presents the comparison with state-of-the-art benchmarks. For visual prompt-based methods such as L2P++,

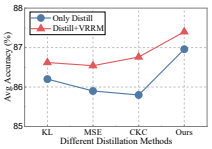


(a) Self-similarity attention variants.

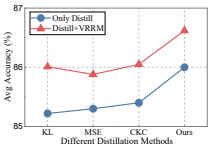


(b) Text-aware aggregation strategies.

Fig. 5: Ablation study of the VRRM components on CIFAR100.



(a) On B10_Inc10 setting.



(b) On B50_Inc10 setting.

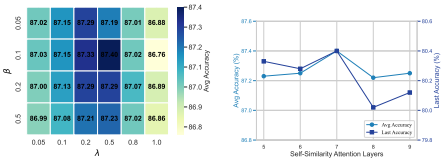
Fig. 6: Effect of different distillation regularization strategies on CIFAR100.

DualPrompt, and CODA-Prompt, their performance is limited by the lack of access to textual information. Across these datasets under two incremental settings, our method outperforms most other approaches. Compared with replay-based methods (i.e., RAPF and PROOF), it achieves a 1.96%–4.48% improvement on average accuracy over ImageNet-R. Moreover, on the large-scale ImageNet-1K dataset, our approach maintains stable and slightly superior performance over baselines. Consistent improvements on both medium- and large-scale datasets demonstrate strong generalization beyond specific settings, achieving a better balance between knowledge retention and adaptation.

Comparison of the Zero-Shot Capability. To ensure fine-tuning does not compromise CLIP's inherent zero-shot capabilities, we evaluate generalization by testing on unseen datasets (Food101, ImageNet-100, and ImageNet-1K) after training on CIFAR-100. As shown in Tab. 2, our approach best preserves CLIP's original performance with minimal degradation. Specifically, it achieves average gains of 3.51% and 3.18% over distillation-based ZSCL [57] and DMNSP [25], respectively. These results also highlight the inherent trade-off in continual learning between task-specific adaptation and zero-shot generalization. While replay-based methods like PROOF [60] and RAPF [20] suffer from significant zero-shot drops due to downstream overfitting, our non-replay method effectively retains pretrained knowledge and minimizes this trade-off compared with existing baselines.

5.3 Ablation Study

Main Components of VAST. Tab. 3 reports the ablation results of the main components. The baseline fine-tunes CLIP with LoRA. VRRM denotes the Visual Representation Reinforcement Module. Adding VRRM improves the LoRA baseline by +2.1% on CIFAR100 and +3.2% on ImageNet-R, with gains on average accuracy. Applying $\mathcal{L}_{distill}$ alone yields similar improvements, suggesting both modules mitigate representation drift. Combining VRRM with $\mathcal{L}_{distill}$ achieves the best results (87.40/80.40 on CIFAR100 and 87.61/82.83 on ImageNet-R, Avg/Last), demonstrating their complementary effect in reducing forgetting.



(a) Effect of β and λ. **(b)** Effect of M.

Fig. 7: Parameter sensitivity analysis of VAST on CIFAR100 B10_Inc10. (a) Impact of fusion coefficient β and distillation weight λ. (b) Influence of the selected self-similarity layer M in the visual encoder.

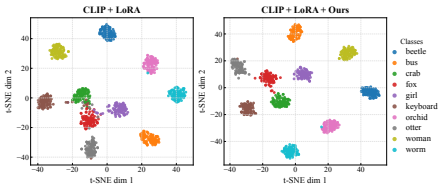


Fig. 8: t-SNE visualization of CLS features on old classes (Task 0) from CIFAR100 after training on Task 1. Left: "CLIP + LoRA". Right: "CLIP + LoRA + Ours".

Impact of VRRM. As shown in Fig. 5, we analyze components of VRRM: self-similarity attention (SSA) and text-aware aggregation under two incremental settings on CIFAR100. For SSA, we compare three variants: (1) QK, the standard query–key attention in the M-th visual layer; (2) KK, which replaces query–key interaction with key–key correlations to build a symmetric self-similarity matrix; and (3) Ours, the second-order self-similarity attention in Eq. (2). Results show that SSA more effectively captures structural dependencies, preserves spatial topology, and produces more stable visual representations across tasks. For text-aware value refinement, we compare (1) using the last-layer feature V_L and (2) uniform average weighting in Eq. (3). Our text-aware weighting achieves the best performance by adaptively emphasizing semantically aligned layers and improving cross-modal stability.

Impact of the Regularization Loss. In Fig. 6, we compare various regularization losses with the proposed loss in VAST, including classical KL and MSE losses from LwF [33] and Mod-X [38], as well as the recent Contrastive Knowledge Consolidation (CKC) loss [34] for CLIP-based continual learning. The results show that while all these losses alleviate forgetting, those emphasizing distributional structure alignment perform notably better. Moreover, our loss shows strong synergy with the proposed VRRM, further enhancing representation stability. Additionally, integrating VRRM into other distillation frameworks also yields consistent gains, confirming that unstable visual structures can impair distillation, whereas stabilizing them is crucial for knowledge preservation.

5.4 Parameter Sensitive Analysis

Impact of self-similarity layer M. We evaluate different layers M of the CLIP visual encoder for constructing self-similarity attention on the CIFAR100 B10_Inc10 task (see Fig. 7b). The results show that most layers yield comparable performance, while layer 7 performs slightly better. This suggests that the middle layer offers a great trade-off between preserving mid-level structural patterns and capturing high-level semantic cues, making it particularly suitable for constructing stable self-similarity representations.

Table 4: Quantitative analysis of visual representation drift from Task 0 to Task 1 (0→1) measured by FDR (inter / intra), where higher inter (↑) and lower intra (↓) indicate better separability.

Dataset	Method	FDR ↑ (0→1)	inter ↑ (0→1)	intra ↓ (0→1)
CIFAR100	LoRA	2.19→2.08↓	0.19→0.18↓	0.09→0.08↓
	Ours	2.48→2.97↑	0.24→0.29↑	0.10→0.10→
TinyImageNet	LoRA	1.75→1.74↓	0.19→0.18↓	0.11→0.10↓
	Ours	2.00→2.35↑	0.23→0.28↑	0.12→0.12→

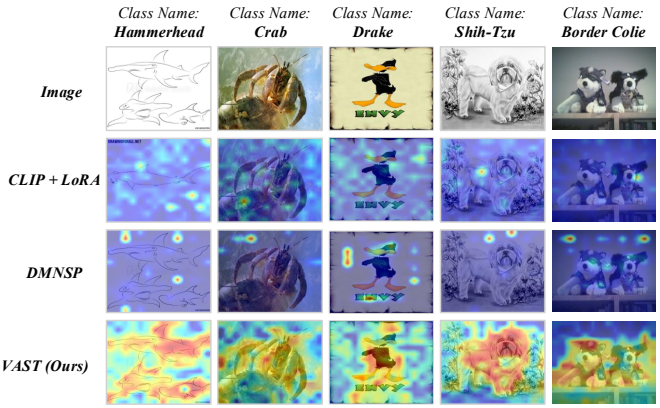


Fig. 9: Visualization of CLS-to-patch attention maps across different methods on ImageNet-R. These maps reveal how the visual CLS token aggregates patch-level features, allowing us to compare how different approaches extract and preserve structural visual information.

Impact of semantic fusion strength β and distillation weight λ. Fig. 7a analyzes two key hyperparameters in VAST: the semantic fusion strength β in Eq. (5), which controls the injection of dense features into the CLS token, and the distillation weight λ in Eq. (6), which balances knowledge preservation across tasks. The results show that performance peaks at $\lambda = 0.5$, while both strong and overly weak distillation lead to degraded accuracy. A similar trend is observed for β, where moderate fusion ($\beta = 0.1$) achieves the best result by integrating spatial cues without overwhelming the global representation.

5.5 Visualization Results

t-SNE Visualization. To evaluate stability, we visualize Task 0 features using t-SNE after learning Task 1. As shown in Fig. 8, while the baseline suffers from severe cluster entanglement and drift, our method maintains compact and well-separated class structures. Beyond qualitative visualization, we provide quantitative evidence using the Fisher Discriminant Ratio (FDR) [26], defined as the ratio of between-class variance to within-class variance . A higher FDR indicates superior feature separability, where class clusters are both compact and well-distinguished. On CIFAR-100, our method improves the FDR from 2.48 to 2.97

after learning the new task, whereas the LoRA baseline degrades from 2.19 to 2.08. Consistent trends on TinyImageNet confirm that our approach effectively preserves feature geometry and mitigates drift.

VRRM Attention Maps for the Visual Anchor. Fig. 9 illustrates a comparison of CLS-to-patch attention maps across different methods, showing how the visual CLS token aggregates patch-level features. The baseline methods and DMNSP exhibit scattered CLS-to-patch activation, often focusing on irrelevant background regions. In contrast, VRRM produces structurally consistent attention maps that align with object contours and salient regions. This focused activation forms a stable visual anchor that effectively preserves the intrinsic geometry and mitigates representation drift during continual learning.

6 Conclusion

In conclusion, we present VAST, a framework for continual CLIP learning that effectively mitigates catastrophic forgetting arising from modality-asymmetric adaptation. The framework identifies the progressive overwriting of visual representations that weakens visual features and disrupts cross-task alignment. To address this, VAST reinforces visual representations by extracting structural information from stable intermediate layers, providing a robust geometric foundation. These reinforced features are then preserved via a distribution-aware structural distillation objective, which maintains the intrinsic consistency of the feature space. By integrating these components, VAST jointly maintains stable and transferable semantics, ensuring long-term knowledge retention for continual learning.

Acknowledgements

This work was supported in part by the New Generation Artificial Intelligence National Science and Technology Major Project No. 2025ZD0123003, the Key Research and Development Project in Shaanxi Province No. 2024PT-ZCK-89, and the Project of China Knowledge Centre for Engineering Science and Technology. This research is also supported by Xin Zhang's A*STAR Career Development Fund (CDF) (Project No. H26-KSR0051). This research is also supported by the Japan Science and Technology Agency (JST) and the Agency for Science, Technology and Research (A*STAR) under the Japan-Singapore Joint Call (Project No. R24I6IR133).

References

1. Aljundi, R., Chakravarty, P., Tuytelaars, T.: Expert gate: Lifelong learning with a network of experts. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3366–3375 (2017)
2. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: European conference on computer vision. pp. 446–461. Springer (2014)
3. Bousselham, W., Petersen, F., Ferrari, V., Kuehne, H.: Grounding everything: Emerging localization properties in vision–language transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3828–3837 (2024)
4. Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., Luo, P.: Adaptformer: Adapting vision transformers for scalable visual recognition. Advances in Neural Information Processing Systems **35**, 16664–16678 (2022)
5. De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., Tuytelaars, T.: A continual learning survey: Defying forgetting in classification tasks. IEEE transactions on pattern analysis and machine intelligence **44**(7), 3366–3385 (2021)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
7. Dong, Y., Gong, T., Yu, S., Li, C.: Optimal randomized approximations for matrix-based rényi's entropy. IEEE Transactions on Information Theory **69**(7), 4218–4234 (2023)
8. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Douillard, A., Ramé, A., Couairon, G., Cord, M.: Dytox: Transformers for continual learning with dynamic token expansion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9285–9295 (2022)
10. Feng, T., Li, W., Zhu, D., Yuan, H., Zheng, W., Zhang, D., Tang, J.: Zeroflow: Overcoming catastrophic forgetting is easier than you think. arXiv preprint arXiv:2501.01045 (2025)
11. Feng, T., Wang, M., Yuan, H.: Overcoming catastrophic forgetting in incremental object detection via elastic response distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9427–9436 (2022)

12. Giraldo, L.G.S., Rao, M., Principe, J.C.: Measures of entropy from data using infinitely divisible kernels. IEEE Transactions on Information Theory **61**(1), 535–548 (2014)
13. Hajimiri, S., Ben Ayed, I., Dolz, J.: Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In: Proceedings of the Winter Conference on Applications of Computer Vision. pp. 5061–5071 (2025)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
15. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8340–8349 (2021)
16. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International conference on machine learning. pp. 2790–2799. PMLR (2019)
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
18. Hu, Z., Li, Y., Lyu, J., Gao, D., Vasconcelos, N.: Dense network expansion for class incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11858–11867 (2023)
19. Huang, C., Wei, Y., Yang, Z., Hu, D.: Adaptive unimodal regulation for balanced multimodal information acquisition. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25854–25863 (2025)
20. Huang, L., Cao, X., Lu, H., Liu, X.: Class-incremental learning with clip: Adaptive representation adjustment and parameter fusion. In: European Conference on Computer Vision. pp. 214–231. Springer (2024)
21. Huang, L., Cao, X., Lu, H., Meng, Y., Yang, F., Liu, X.: Mind the gap: Preserving and compensating for the modality gap in clip-based continual learning. arXiv preprint arXiv:2507.09118 (2025)
22. Jalali, M., Li, C.T., Farnia, F.: An information-theoretic evaluation of generative models in learning multi-modal distributions. Advances in Neural Information Processing Systems **36**, 9931–9943 (2023)
23. Jha, S., Gong, D., Yao, L.: Clap4clip: Continual learning with probabilistic finetuning for vision-language models. Advances in neural information processing systems **37**, 129146–129186 (2024)
24. Jha, S., Gong, D., Zhao, H., Yao, L.: Npcl: Neural processes for uncertainty-aware continual learning. Advances in Neural Information Processing Systems **36**, 34329–34353 (2023)
25. Kang, B., Wang, L., Wu, Z., Feng, T., Li, Y., Gao, Y., Li, W.: Dynamic multi-layer null space projection for vision-language continual learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2077–2086 (2025)
26. Kim, S.J., Magnani, A., Boyd, S.: Robust fisher discriminant analysis. Advances in neural information processing systems **18** (2005)
27. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)
28. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. Technique report (2009)

29. Lan, M., Chen, C., Ke, Y., Wang, X., Feng, L., Zhang, W.: Clearclip: Decomposing clip representations for dense vision-language inference. In: European Conference on Computer Vision. pp. 143–160. Springer (2024)
30. Lan, M., Chen, C., Ke, Y., Wang, X., Feng, L., Zhang, W.: Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In: European Conference on Computer Vision. pp. 70–88. Springer (2024)
31. Le, Y., Yang, X.: Tiny imagenet visual recognition challenge. CS 231N **7**(7), 3 (2015)
32. Li, Y., Wang, H., Duan, Y., Zhang, J., Li, X.: A closer look at the explainability of contrastive language-image pre-training. Pattern Recognition **162**, 111409 (2025)
33. Li, Z., Hoiem, D.: Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence **40**(12), 2935–2947 (2017)
34. Liu, W., Zhu, F., Wei, L., Tian, Q.: C-clip: Multimodal continual learning for vision-language model. In: The Thirteenth International Conference on Learning Representations (2025)
35. Lopez-Paz, D., Ranzato, M.: Gradient episodic memory for continual learning. Advances in neural information processing systems **30** (2017)
36. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008)
37. Masana, M., Liu, X., Twardowski, B., Menta, M., Bagdanov, A.D., Van De Weijer, J.: Class-incremental learning: survey and performance evaluation on image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(5), 5513–5533 (2022)
38. Ni, Z., Wei, L., Tang, S., Zhuang, Y., Tian, Q.: Continual vision-language representation learning with off-diagonal information. In: International Conference on Machine Learning. pp. 26129–26149. PMLR (2023)
39. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
40. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2001–2010 (2017)
41. Shao, T., Tian, Z., Zhao, H., Su, J.: Explore the potential of clip for training-free open vocabulary semantic segmentation. In: European Conference on Computer Vision. pp. 139–156. Springer (2024)
42. Skean, O., Osorio, J.K.H., Brockmeier, A.J., Giraldo, L.G.S.: Dime: Maximizing mutual information by a difference of matrix-based entropies. arXiv preprint arXiv:2301.08164 (2023)
43. Smith, J.S., Karlinsky, L., Gutta, V., Cascante-Bonilla, P., Kim, D., Arbelle, A., Panda, R., Feris, R., Kira, Z.: Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11909–11919 (2023)
44. Thengane, V., Khan, S., Hayat, M., Khan, F.: Clip model is an efficient continual learner. arXiv preprint arXiv:2210.03114 (2022)
45. Wang, F., Mei, J., Yuille, A.: Sclip: Rethinking self-attention for dense vision-language inference. In: European conference on computer vision. pp. 315–332. Springer (2024)

46. Wang, L., Zhang, X., Su, H., Zhu, J.: A comprehensive survey of continual learning: Theory, method and application. IEEE transactions on pattern analysis and machine intelligence **46**(8), 5362–5383 (2024)
47. Wang, L., Zhang, X., Yang, K., Yu, L., Li, C., Hong, L., Zhang, S., Li, Z., Zhong, Y., Zhu, J.: Memory replay with data compression for continual learning. arXiv preprint arXiv:2202.06592 (2022)
48. Wang, Z., Zhang, Z., Ebrahimi, S., Sun, R., Zhang, H., Lee, C.Y., Ren, X., Su, G., Perot, V., Dy, J., et al.: Dualprompt: Complementary prompting for rehearsal-free continual learning. In: European conference on computer vision. pp. 631–648. Springer (2022)
49. Wang, Z., Zhang, Z., Lee, C.Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J., Pfister, T.: Learning to prompt for continual learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 139–149 (2022)
50. Ye, F., Bors, A.G.: Self-evolved dynamic expansion model for task-free continual learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22102–22112 (2023)
51. Yu, J., Zhuge, Y., Zhang, L., Hu, P., Wang, D., Lu, H., He, Y.: 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)
52. Yu, S., Giraldo, L.G.S., Jenssen, R., Principe, J.C.: Multivariate extension of matrix-based rényi's α-order entropy functional. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(11), 2960–2966 (2020)
53. Yuan, M., Zhang, Y., Zhang, W., Ma, L., Gao, Y., Ying, J., Xin, Y.: Infoclip: Bridging vision-language pretraining and open-vocabulary semantic segmentation via information-theoretic alignment transfer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 12240–12248 (2026)
54. Zhai, J.T., Liu, X., Bagdanov, A.D., Li, K., Cheng, M.M.: Masked autoencoders are efficient class incremental learners. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19104–19113 (2023)
55. Zhang, W., Huang, Y., Zhang, W., Zhang, T., Lao, Q., Yu, Y., Zheng, W.S., Wang, R.: Continual learning of image classes with language guidance from a vision-language model. IEEE Transactions on Circuits and Systems for Video Technology (2024)
56. Zhang, Y., Yuan, M., Zhang, W., Gong, T., Wen, W., Ying, J., Shi, W.: Infosam: Fine-tuning the segment anything model from an information-theoretic perspective. arXiv preprint arXiv:2505.21920 (2025)
57. Zheng, Z., Ma, M., Wang, K., Qin, Z., Yue, X., You, Y.: 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)
58. Zhou, D.W., Li, K.W., Ning, J., Ye, H.J., Zhang, L., Zhan, D.C.: External knowledge injection for clip-based class-incremental learning. arXiv preprint arXiv:2503.08510 (2025)
59. Zhou, D.W., Ye, H.J., Zhan, D.C.: Co-transport for class-incremental learning. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1645–1654 (2021)
60. Zhou, D.W., Zhang, Y., Wang, Y., Ning, J., Ye, H.J., Zhan, D.C., Liu, Z.: Learning without forgetting for vision-language models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)

61. Zhou, H., Xiao, S., Zhang, S., Peng, J., Zhang, S., Li, J.: Jump self-attention: Capturing high-order statistics in transformers. Advances in Neural Information Processing Systems **35**, 17899–17910 (2022)