

Dual-Mode Low-Rank Learner with Bridge-Prototype Ensemble for Vision-Language Class-Incremental Learning

Chiyuan He¹, Zihuan Qiu¹, Fanman Meng^{1,✉}, Chao Wang², Liangjiang Chen¹, Linfeng Xu¹, Qingbo Wu¹, Hongliang Li¹

¹University of Electronic Science and Technology of China, Chengdu, China

²Qiyuan Lab, Beijing, China

{cyhe,zihuanqiu,202522011613}@std.uestc.edu.cn

{lfxu,qbwu,hlli}@uestc.edu.cn, w-c15@tsinghua.org.cn

Abstract

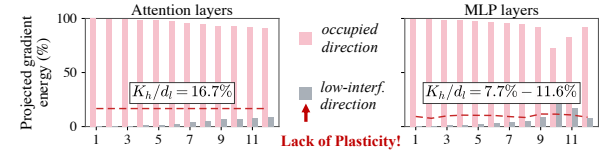
Benefiting from transferable visual-textual alignment, CLIP has been widely adopted for class-incremental learning (CIL). However, existing learners either repeatedly update components shared across tasks, leading to knowledge overwriting, or overly isolate new-task updates, hindering the reuse of CLIP’s transferable knowledge and limiting plasticity. Moreover, the text-based or bimodal classifier designs still fail to effectively integrate complementary information from the visual and textual modalities. To address these challenges, we introduce DuLBE, which couples dual-mode low-rank learning with a bridge-prototype ensemble classifier for exemplar-free CIL. DuLBE allocates two visual low-rank update modes according to the gradient demand and uses gradient routing to coordinate them: a compact and rewritable shared mode is selected from historically occupied visual directions to reuse transferable knowledge, while residual modes provide low-interference channels for task-specific variations. Building on the resulting stable inter-modal structure, we further construct geodesic bridges between visual prototypes and text embeddings on the unit hypersphere, and ensemble reliable bridge prototypes to compensate for the modality-gap limitations of textual decision boundaries. Extensive experiments under multiple settings show that DuLBE achieves state-of-the-art CIL performance while retaining the high parameter efficiency of low-rank tuning.

Introduction

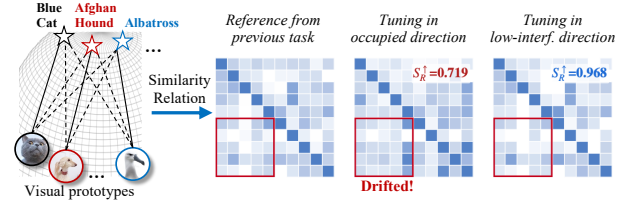
Benefiting from large-scale pre-training over aligned vision-text data, vision-language models (VLMs) like CLIP (Radford et al. 2021) can use natural-language text to organize the continually expanding class space of downstream class-incremental learning (CIL). The central challenge is to balance stability and plasticity: when previous data are unavailable or strictly limited, the model must preserve knowledge from old classes while continually incorporating knowledge of new classes.

Parameter-efficient tuning methods for pre-trained CLIP commonly freeze the backbone and introduce a small number of trainable parameters, such as prompts, adapters, or low-rank factors (Zhou et al. 2022; Gao et al. 2024; Hu et al. 2022). These methods retain the general alignment knowledge acquired during pre-training while adapting the model to downstream tasks. Building on this adaptation paradigm,

✉ Corresponding author: fmmeng@uestc.edu.cn. Preprint. Work in progress.



(a) Layer-wise comparison of projected gradient energy



(b) Structure drift of old visual-textual relations

Figure 1: Motivating observations. (a) New-task cross-modal gradients retain substantial projection energy in historically occupied directions ($K_h = 128$) across the visual tower. (b) Tuning occupied directions causes greater old-class relation-structure drift than tuning low-interference directions. S_R denotes rank-structure similarity to the old-task relation matrix, whose entries are cosine similarities between old-class visual prototypes and corresponding text embeddings.

recent continual learning methods introduce global constraints (Zheng et al. 2023; Yu et al. 2024; Wu et al. 2025; He et al. 2026b), or tailored update strategies (Zhang et al. 2024; Huang et al. 2025) to continually adapt CLIP. Nevertheless, these methods still repeatedly optimize trainable components shared across tasks. Without old-task data, new-task supervision can progressively overwrite the previously acquired knowledge encoded in these shared components.

Another line of work isolates task knowledge to reduce interference from new-task learning, with particular attention to the more fragile visual tower (Zhou et al. 2025a; Li et al. 2026; Kang et al. 2025; Peng et al. 2025; Qiu et al. 2026). Within this line, strategies based on orthogonal update or null-space learning reduce interference by steering new-task updates away from the principal directions occupied by old-task representations. However, our study reveals that avoiding these directions may be overly restrictive in CLIP. As

shown in Figure 1(a), a substantial portion of the new-task cross-modal gradient energy in CIL is projected onto these occupied directions (i.e., the top- K_h principal directions of old-task representations), well above the dimensional baseline K_h/d_l expected for a random K_h -dimensional subspace, where d_l is the input dimension of layer l . We attribute this concentration to the way CLIP organizes visual and textual concepts in a shared general-purpose alignment space, where different classes reuse common visual patterns (e.g., wheel-related features shared by concepts: trucks and bicycles). Excluding these directions would therefore discard transferable write-in signals and restrict alignment plasticity. However, such transferability does not make them freely rewritable. Figure 1(b) shows that tuning in the occupied directions induces substantially greater drift in old-class visual-textual relations than tuning in low-interference directions (i.e., directions selected from the orthogonal complement of their span). Without old data, aligning towards new textual objective can shift these shared visual patterns, thereby disrupting the old semantic relation structure and ultimately leading to forgetting. Taken together, these observations reveal that occupied directions support transfer but require controlled rewriting, whereas low-interference directions enable safer adaptation but offer limited plasticity.

To address this bottleneck, we propose **DuLBE** (**D**ual-mode **L**ow-rank learner with **B**ridge-prototype **E**nsemble), a collaborative CLIP-based CIL framework. DuLBE performs current-task-driven, layer-wise allocation of the low-rank subspace for adaptation in CLIP’s visual tower. At each adapted layer, it decomposes the prospective gradient according to historical visual-mode occupation. Following the principle of maximizing gradient projection energy, DuLBE selects one compact rewritable shared mode from occupied directions and allocates a small set of low-interference residual modes to the remaining task-specific demand. Gradient routing further assigns distinct responsibilities to the two modes: the shared mode reuses a high-demand historical direction under semantic-structure regularization, whereas the residual modes capture task-specific variations with negligible interference to previous tasks. This dual-mode learner transfers previously established cross-modal structures while injecting new knowledge with a small rank overhead. Building on the resulting stable inter-modal structure, we further introduce a bridge-prototype ensemble classifier. It constructs bridges between visual prototypes and text embeddings on the unit hypersphere through geodesic interpolation and uses bridge reliability to refine the decision boundary, flexibly combining textual semantics with adapted visual evidence.

In summary, our contributions include: (1) we propose DuLBE, a dual-mode collaborative low-rank continual learner that allocates rewritable and low-interference sub-modes in CLIP’s visual tower, together with a gradient-routing mechanism for stable yet plastic adaptation. (2) we introduce a bridge-prototype ensemble classifier that leverages stable inter-modal geodesic structures to compensate for the limitations of a single text-based decision boundary. (3) extensive experiments demonstrate that DuLBE achieves state-of-the-art performance with modest trainable overhead.

Related Work

Class-Incremental Learning

Class-incremental learning (CIL) requires a model to learn new classes sequentially while recognizing all seen classes without task identities. Traditional CIL methods mitigate forgetting by storing exemplars, distilling old responses, or allocating task-specific parameters (De Lange et al. 2021; Rebuffi et al. 2017; Li and Hoiem 2017; Douillard et al. 2020). These strategies establish the stability-plasticity trade-off, but they often rely on stored data, explicit old-task supervision, or expanding capacity.

CLIP-based CIL shifts the focus from learning visual representations from scratch to continually optimizing visual-textual alignment. Existing methods mainly exploit this prior from three perspectives. First, external knowledge, textual priors, or language-guided concept descriptions are introduced to enrich class semantics and optimize the continual learning objective (Zhou et al. 2025a; He et al. 2026d; Yin et al. 2024; Huang et al. 2024; Yu et al. 2025). Second, several studies protect the global cross-modal semantic structure by preserving modality-gap properties, relative semantic relations, or geometric consistency during continual adaptation (Yu et al. 2024; He et al. 2026d; Hu et al. 2025; Liang et al. 2025; He et al. 2026b; Gong et al. 2026). Third, interference-isolation methods either expand and freeze task-specific projection heads (Zhou et al. 2025b,a; Liang et al. 2025), or constrain new-task updates in low-interference directions (Li et al. 2026; Kang et al. 2025; Qiu et al. 2025). These studies demonstrate the value of CLIP’s semantic structure, but they either keep updating shared trainable modules or mainly avoid risky directions.

LoRA-Based Continual Learning

Low-Rank Adaptation (LoRA) (Hu et al. 2022) freezes pre-trained weights and parameterizes updates with low-rank factors, which makes it suitable for compact continual adaptation. However, a small-rank update can still interfere with old knowledge if its directions overlap with old-task-sensitive regions. Recent LoRA-based continual learning methods therefore control the update by orthogonalization (Wang et al. 2023), interference-free subspace allocation (Liang and Li 2024; Luo et al. 2026), energy-driven subspace decomposition (He et al. 2026c) or predefined functional LoRA insertion (He, Duan, and Zhu 2025). These methods establish low-rank subspace control as an effective way to balance stability and plasticity. In contrast, DuLBE uses cross-modal gradient demand and mode occupation to decide which shared directions are rewritable and which residual directions should be included.

Methodology

Preliminaries

We consider exemplar-free class-incremental learning (CIL) with a pre-trained CLIP model. The learner receives a sequence of tasks $\{\mathcal{D}_t\}_{t=1}^N$, where task t introduces a disjoint class set $\mathcal{C}_t$ satisfying $\mathcal{C}_t \cap \mathcal{C}_k = \emptyset$ for $t \neq k$. After finishing task t , the model is evaluated over all seen classes

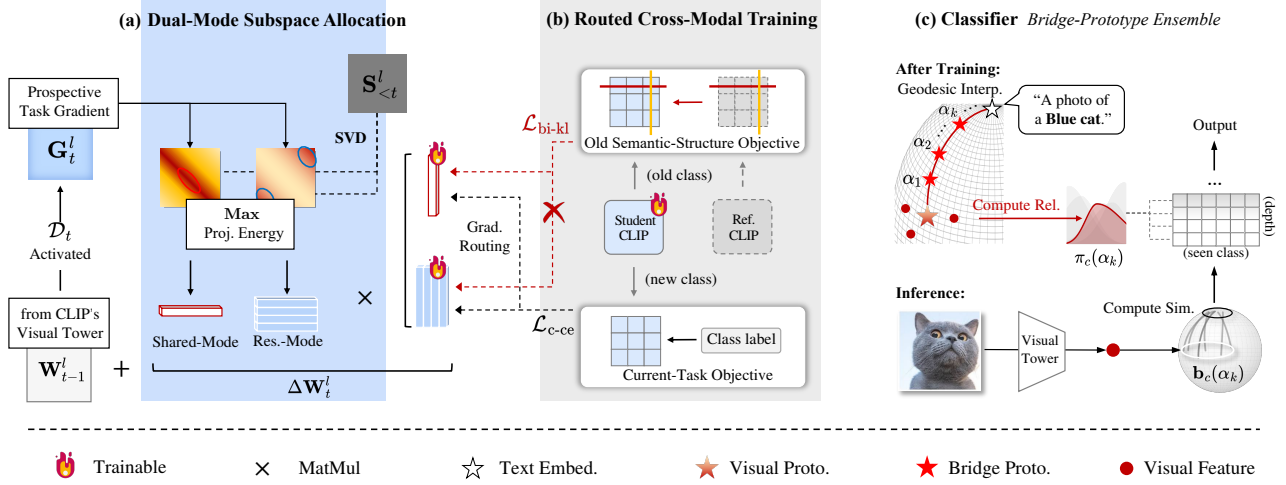


Figure 2: Overview of **DuLBE**. (a) Dual-mode allocation uses the prospective task gradient to form a compact shared mode within historically occupied directions and a low-interference residual mode outside them. (b) Routed cross-modal training directs gradients from the current-task objective to both learners, while routing gradients from the semantic-structure preservation objective only to the shared learner. (c) Classifier: it builds visual-text geodesic bridges, estimates bridge-depth reliability after each task, and ensembles reliable bridge prototypes for inference.

$\mathcal{C}_{\leq t} = \bigcup_{k=1}^t \mathcal{C}_k$ without task identities, and no raw data from previous tasks are retained.

CLIP performs classification by comparing an image representation with text embeddings generated from class-name prompts, e.g., A photo of a {CLASS}. Given the normalized visual feature $\mathbf{z}_i$ of image x_i and class-text embedding $\mathbf{e}_c$, the CLIP logit is $\ell_i(c) = \tau \mathbf{z}_i^\top \mathbf{e}_c$, where τ is the logit scale. This language-defined classifier naturally supports an expanding class space, but it also makes continual visual adaptation delicate: the visual tower must learn new classes while remaining compatible with both the textual decision space and the cross-modal relations formed by previous tasks.

Overview

As shown in Figure 2, DuLBE integrates three components. First, it allocates the visual low-rank update space using the current-task prospective gradient. Within the historically occupied subspace, the direction with the strongest gradient demand forms a compact rewritable shared mode, while dominant directions outside this subspace form low-interference residual modes for task-specific knowledge. Second, the shared and residual mode learners are trained collaboratively through mode-specific gradient routing, which assigns different knowledge roles: the current-task CE updates both learners, while semantic-structure preservation regularizes only the shared learner, encouraging controlled knowledge reuse and residual task-specific plasticity. Third, DuLBE forms a reliability-guided bridge-prototype ensemble by estimating visual-text geodesic bridges after each task and aggregating reliable bridge prototypes for classification.

Gradient-Demand Dual-Mode Subspace Allocation

A single current-task update may overwrite old visual-text relations. Some interference-aware strategies restrict new up-

dates to orthogonal or null-space directions (Wang et al. 2023; Liang and Li 2024), reducing forgetting but also blocking directions that could remain reusable in CLIP’s shared semantic space. Dual-branch designs introduce shared and task-specific components (He, Duan, and Zhu 2025; He et al. 2026c), but their allocation is often predefined rather than determined by the current write-in demand. We instead allocate the visual low-rank subspace from the perspective of the current cross-modal gradient. For each adapted visual layer l , a prospective gradient is computed on current-task data $\mathcal{D}_t$:

$$\mathbf{G}_t^l = \nabla_{\mathbf{W}_{t-1}^l} \mathcal{L}_{c-ce}(\theta_{t-1}; \mathcal{D}_t). \quad (1)$$

This gradient is used only to plan the low-rank update basis, not as a direct parameter update. It indicates which input directions are activated by the current cross-modal CE objective in the layer l .

Historical Visual-Mode Occupation. The historical visual support is estimated from the second-order statistics of layer inputs. For a previous task t' , the cached input tokens of the layer l are row-stacked as $\mathbf{X}_{t'}^l \in \mathbb{R}^{n_{t'}^l \times d_l}$, where $n_{t'}^l$ is the number of cached tokens and d_l is the input dimension. The task-wise and accumulated occupation statistics are:

$$\mathbf{S}_{t'}^l = (\mathbf{X}_{t'}^l)^\top \mathbf{X}_{t'}^l, \quad \mathbf{S}_{\leq t}^l = \sum_{t' < t} \mathbf{S}_{t'}^l. \quad (2)$$

For any unit direction $\mathbf{p}$, the quadratic form $\mathbf{p}^\top \mathbf{S}_{\leq t}^l \mathbf{p} = \sum_{t' < t} \|\mathbf{X}_{t'}^l \mathbf{p}\|_2^2$ measures its accumulated activation energy on previous tasks. Large eigenvalues of $\mathbf{S}_{\leq t}^l$ indicate repeatedly occupied visual modes. Since $\mathbf{S}_{\leq t}^l$ is positive semi-definite, its SVD takes the eigendecomposition form:

$$\mathbf{S}_{\leq t}^l = \mathbf{U}_{\leq t}^l \mathbf{\Lambda}_{\leq t}^l (\mathbf{U}_{\leq t}^l)^\top. \quad (3)$$

The top- K_h eigenvectors form the historical principal support $\mathcal{H}_{< t}^l = \text{span}(\mathbf{U}_{< t}^{l, K_h})$. This support may contain transferable visual directions, but rewriting it as a whole would be too aggressive because many directions encode old visual-text relations irrelevant to the current task.

Rewritable Shared Mode. The shared mode should therefore be a compact portion of the historical support, selected by its current gradient demand. Under a LoRA update, when the down-projection direction is fixed as a unit vector $\mathbf{p}$, the gradient that can be absorbed by the corresponding trainable up-projection is proportional to $\mathbf{G}_t^l \mathbf{p}$. Thus, $\|\mathbf{G}_t^l \mathbf{p}\|_2^2$ measures the write-in energy available along $\mathbf{p}$. We select the historical direction that maximizes this energy:

$$\mathbf{p}_t^{l, S} = \arg \max_{\substack{\mathbf{p} \in \mathcal{H}_{< t}^l \\ \|\mathbf{p}\|_2 = 1}} \|\mathbf{G}_t^l \mathbf{p}\|_2^2. \quad (4)$$

Writing $\mathbf{p} = \mathbf{U}_{< t}^{l, K_h} \mathbf{q}$ with $\|\mathbf{q}\|_2 = 1$ turns the objective into $\mathbf{q}^\top \mathbf{M}_t^l \mathbf{q}$, where $\mathbf{M}_t^l = (\mathbf{U}_{< t}^{l, K_h})^\top (\mathbf{G}_t^l)^\top \mathbf{G}_t^l \mathbf{U}_{< t}^{l, K_h}$. By the Rayleigh-Ritz theorem (Horn and Johnson 2012), the maximizer is the principal eigenvector of $\mathbf{M}_t^l$. Therefore, the shared down-projection basis can be defined as:

$$\mathbf{q}_t^l = \text{eig}_{\max}(\mathbf{M}_t^l), \quad \mathbf{P}_t^{l, S} = \mathbf{U}_{< t}^{l, K_h} \mathbf{q}_t^l. \quad (5)$$

This basis keeps only the strongest overlap between historical occupation and current cross-modal demand. $\mathbf{P}_t^{l, S}$ is the only historical direction allowed to be rewritten.

Low-Interference Residual Modes. After extracting the rewritable shared mode, the remaining gradient demand is used to construct additional learnable channels for task-specific adaptation. These residual modes follow the spirit of interference-avoidance methods that project new updates away from historical or alignment-sensitive subspaces (Wang et al. 2023; Liang and Li 2024; Li et al. 2026). Unlike these methods, our residual modes are not expected to learn the new task alone. Instead, they provide low-interference plasticity that cooperates with the shared mode. Concretely, we remove the historical principal support from the current gradient:

$$\mathbf{G}_{t, R}^l = \mathbf{G}_t^l (\mathbf{I} - \mathbf{U}_{< t}^{l, K_h} (\mathbf{U}_{< t}^{l, K_h})^\top). \quad (6)$$

The residual basis is then taken from the dominant right singular directions of this projected gradient. Specifically, if $\mathbf{G}_{t, R}^l = \mathbf{U}_R^l \Sigma_R^l (\mathbf{V}_R^l)^\top$ with $\mathbf{V}_R^l = [\mathbf{v}_{R, 1}^l, \dots, \mathbf{v}_{R, d_l}^l]$, then we choose:

$$\mathbf{P}_t^{l, R} = [\mathbf{v}_{R, 1}^l, \dots, \mathbf{v}_{R, r_R}^l]. \quad (7)$$

By construction, $\text{span}(\mathbf{P}_t^{l, R}) \subseteq (\mathcal{H}_{< t}^l)^\perp$. Moreover, for any unit residual direction $\mathbf{r} \in \text{span}(\mathbf{P}_t^{l, R})$, the historical occupation energy satisfies $\mathbf{r}^\top \mathbf{S}_{< t}^l \mathbf{r} \leq \lambda_{K_h+1}^l$. Residual modes therefore capture current-task gradient demand outside the historical support, while maintaining low overlap with previously occupied visual modes.

Subspace-Frozen Parameterization. The allocation stage only determines where each branch can write. After allocation, $\mathbf{P}_t^{l, S}$ and $\mathbf{P}_t^{l, R}$ are frozen as the down-projection bases, while only $\mathbf{B}_t^{l, S}$ and $\mathbf{B}_t^{l, R}$ remain trainable. For the

first task, without any historical statistics, the bases are initialized with the dominant right singular vectors of $\mathbf{G}_1^l$. The resulting task- t update is

$$\Delta \mathbf{W}_t^l = \mathbf{B}_t^{l, S} (\mathbf{P}_t^{l, S})^\top + \mathbf{B}_t^{l, R} (\mathbf{P}_t^{l, R})^\top. \quad (8)$$

Routed Cross-Modal Training

Given the allocated bases $\mathbf{P}_t^{l, S}$ and $\mathbf{P}_t^{l, R}$, their up-projections $\mathbf{B}_t^{l, S}$ and $\mathbf{B}_t^{l, R}$ are trained with two routed objectives: a current-task cross-modal objective for new class learning and an old semantic-structure objective for preserving old visual-text relations.

Current-Task Cross-Modal Objective. For a mini-batch $\mathcal{B}$, the student model produces row-stacked normalized image features $\mathbf{Z}_t \in \mathbb{R}^{|\mathcal{B}| \times d}$. The text embeddings of current-task classes are $\mathbf{E}_t = [\mathbf{e}_c]_{c \in \mathcal{C}_t} \in \mathbb{R}^{|\mathcal{C}_t| \times d}$, with batch labels $\mathbf{Y}_\mathcal{B}$. The batch image-text logits and current-task CE loss are:

$$\mathbf{L}_t = \tau \mathbf{Z}_t \mathbf{E}_t^\top, \quad \mathcal{L}_{\text{c-ce}}(\mathcal{B}) = \text{CE}(\mathbf{L}_t, \mathbf{Y}_\mathcal{B}). \quad (9)$$

Old Semantic-Structure Objective. The shared basis $\mathbf{P}_t^{l, S}$ is selected from directions occupied by old tasks, making it the main channel for reusing transferable cross-modal knowledge. However, optimizing it may reinterpret these old modes for new-class texts and distort old visual-text relations. We therefore simply constrain the shared branch using semantic structure induced by old text embeddings. Specifically, before training task t , we freeze a copy of the model obtained after task $t-1$ as the reference teacher. For the same mini-batch $\mathcal{B}$, the reference visual features are denoted as $\bar{\mathbf{Z}}_{t-1}$, and the old-class text embeddings are collected as $\bar{\mathbf{E}}_{< t} = [\mathbf{e}_c]_{c \in \mathcal{C}_{< t}}$. The student and reference logits over old classes can be represented as:

$$\mathbf{L}_t^{< t} = \tau \mathbf{Z}_t \bar{\mathbf{E}}_{< t}^\top, \quad \bar{\mathbf{L}}_{t-1}^{< t} = \tau \bar{\mathbf{Z}}_{t-1} \bar{\mathbf{E}}_{< t}^\top. \quad (10)$$

The old semantic-structure loss is defined as:

$$\mathcal{L}_{\text{bi-kl}}(\mathcal{B}) = \underbrace{\tau_C^2 \text{KL}(\sigma_C(\bar{\mathbf{L}}_{t-1}^{< t} / \tau_C) \parallel \sigma_C(\mathbf{L}_t^{< t} / \tau_C))}_{\text{class-wise structure constraint}} + \underbrace{\tau_I^2 \text{KL}(\sigma_I(\bar{\mathbf{L}}_{t-1}^{< t} / \tau_I) \parallel \sigma_I(\mathbf{L}_t^{< t} / \tau_I))}_{\text{instance-wise structure constraint}}. \quad (11)$$

where $\sigma_C(\cdot)$ and $\sigma_I(\cdot)$ are softmax operators that normalize over old classes for each image and over batch instances for each old class-text, respectively. τ_C, τ_I are temperatures. The class-wise term preserves the relative positioning of images with respect to old text embeddings, while the instance-wise term preserves how each old text embedding organizes the visual distribution within the batch.

Gradient Routing. Gradient routing controls which objective contributes gradients to each mode. The current-task cross-modal objective sends gradients to both branches, since both modes participate in new-class learning. In contrast, the old semantic-structure objective is back-propagated only through the shared branch, which rewrites directions already occupied by old tasks:

$$\begin{aligned} \mathbf{B}_t^{l, S} &\leftarrow \mathbf{B}_t^{l, S} - \eta \nabla_{\mathbf{B}_t^{l, S}} [\mathcal{L}_{\text{c-ce}}(\mathcal{B}) + \lambda \mathcal{L}_{\text{bi-kl}}(\mathcal{B})], \\ \mathbf{B}_t^{l, R} &\leftarrow \mathbf{B}_t^{l, R} - \eta \nabla_{\mathbf{B}_t^{l, R}} \mathcal{L}_{\text{c-ce}}(\mathcal{B}). \end{aligned} \quad (12)$$

Here η is the learning rate and λ controls the strength of old-structure constraint. As a result, the shared mode learns transferable alignment under old semantic constraints, while the residual mode captures task-specific evidence through low-interference directions.

Reliability-Guided Bridge-Prototype Ensemble

Text embeddings offer stable semantic anchors, but they mainly encode high-level semantics and may miss discriminative evidence from the visual side. Existing methods often combine text embeddings and visual prototypes for prediction (Zhou et al. 2025a; Li et al. 2026; He et al. 2026d), but such fusion is usually coarse: the two modal representations are treated as fixed endpoints, and a global fusion rule cannot reflect that different classes may rely on different degrees of textual semantics and visual evidence. We therefore search the *inter-modal* discriminative space on the unit hypersphere by constructing class-wise geodesic bridges and estimating which bridge depths are reliable. For each class c introduced at task t , we define its sample set as $\mathcal{X}_c^t = \{x_i \mid (x_i, y_i) \in \mathcal{D}_t, y_i = c\}$. Its normalized visual prototype is computed as:

$$\mathbf{v}_c = \frac{\sum_{x_i \in \mathcal{X}_c^t} \mathbf{z}_i}{\left\| \sum_{x_i \in \mathcal{X}_c^t} \mathbf{z}_i \right\|_2}. \quad (13)$$

Given $\mathbf{v}_c$ and the normalized text embedding $\mathbf{e}_c$, the bridge depth is discretized as $\{\alpha_1, \alpha_2, \dots, \alpha_{K_b}\}$, and the endpoint angle on the unit hypersphere is $\theta_c = \arccos(\mathbf{v}_c^\top \mathbf{e}_c)$. The bridge prototype at depth α_k is obtained by geodesic interpolation:

$$\mathbf{b}_c(\alpha_k) = \frac{\sin((1 - \alpha_k)\theta_c)\mathbf{v}_c + \sin(\alpha_k\theta_c)\mathbf{e}_c}{\sin \theta_c}. \quad (14)$$

Here $\alpha_k = 0$ and $\alpha_k = 1$ correspond to the visual and textual endpoints, respectively. Compared with linear interpolation (Li et al. 2026), geodesic interpolation provides a smooth transition on the unit hypersphere from the visual semantic coordinate to the textual one, so bridge prototypes at each depth remain semantically comparable across classes despite the modality gap. For a fixed depth α_k , we instantiate a bridge classifier using the bridge prototypes of all seen classes. Its logits for class $j \in \mathcal{C}_{\leq t}$ on image x_i is $\tau \mathbf{z}_i^\top \mathbf{b}_j(\alpha_k)$, and the corresponding softmax probability for class c is denoted by $p_{\alpha_k}(c|x_i)$. The reliability of depth α_k for class c is then measured by the average correct-class confidence on these samples as $\rho_c(\alpha_k) = \frac{1}{|\mathcal{X}_c^t|} \sum_{x_i \in \mathcal{X}_c^t} p_{\alpha_k}(c|x_i)$. The reliability weights are then normalized over all bridge depths by:

$$\pi_c(\alpha_k) = \frac{\exp(\rho_c(\alpha_k)/\beta)}{\sum_{m=1}^{K_b} \exp(\rho_c(\alpha_m)/\beta)}. \quad (15)$$

The temperature β controls how sharply the ensemble concentrates on highly reliable depths. For old classes, the visual prototypes and reliability weights are stored when the classes are first introduced, and no raw old data is kept.

At inference, prediction is made by a reliability-weighted ensemble of bridge-prototype logits:

$$\hat{y} = \arg \max_{c \in \mathcal{C}_{\leq t}} \sum_{k=1}^{K_b} \pi_c(\alpha_k) \tau \mathbf{z}(x)^\top \mathbf{b}_c(\alpha_k). \quad (16)$$

This ensemble improves CIL classification by exploiting discriminative bridge prototypes in the inter-modal semantic region. Its effectiveness relies on a stably evolving inter-modal semantic structure. As shown in Table 3, methods without explicit semantic-structure preservation do not obtain consistent gains from the bridge-prototype ensemble. Thus, the classifier is tightly coupled with the proposed dual-mode low-rank learner. Further evidence and analysis of the bridge-prototype ensemble are provided in Appendix A.3.

Experiments

Experimental Setup

Datasets. We evaluate DuLBE under two standard CLIP-based CIL settings adopted in recent studies. **Setting-A** (Huang et al. 2024) uses OpenAI CLIP (Radford et al. 2021) and includes CIFAR (Krizhevsky 2009), CUB (Wah et al. 2011), ImageNet-R (Hendrycks et al. 2021), and ImageNet100 (Deng et al. 2009). **Setting-B** (Zhou et al. 2025a) uses OpenCLIP (Cherti et al. 2023) pretrained on LAION-400M (Schuhmann et al. 2021) and we include CIFAR, SUN (Xiao et al. 2010), Cars (Krause et al. 2013), and Food (Bossard, Guillaumin, and Van Gool 2014). For each setting, we use the same task splits and data preprocessing as the corresponding protocol.

Implementation Details. All experiments are implemented in PyTorch and conducted on an NVIDIA RTX 4090 GPU. For fair comparison, all methods use the same ViT-B/16 pretrained weights required by each setting: OpenAI CLIP in Setting-A and LAION-400M-pretrained OpenCLIP in Setting-B. The proposed dual-mode low-rank learner is inserted into the transformer blocks of CLIP’s visual tower, including the key/value attention projections and MLP layers. The historical support dimension is $K_h = 128$, with shared rank $r_S = 1$ and residual rank $r_R = 8$. The KL temperatures are set to $\tau_C = 5$ and $\tau_I = 0.1$, respectively, and the semantic-structure loss weight is $\lambda = 0.5$. We optimize the model with Adam (Kingma and Ba 2015) using a learning rate of 1×10^{-3} and a batch size of 32. For the bridge-prototype ensemble, we sample $K_b = 10$ bridge depths and set the reliability temperature to $\beta = 0.05$.

Comparison with the State-of-the-Arts

We compare DuLBE with representative baselines, including the zero-shot ContinualCLIP baseline (Thengane et al. 2022), prompt-based methods DualPrompt (Wang et al. 2022) and CODA (Smith et al. 2023), and recent CLIP-based CIL methods PROOF (Zhou et al. 2025b), RAPF (Huang et al. 2024), SGCL (Yu et al. 2024), CLG-CBM (Yu et al. 2025), ENGINE (Zhou et al. 2025a), DesCLIP (He et al. 2026a), MG-CLIP (Huang et al. 2025), BOFA (Li et al. 2026), LoDA-CLIP (He et al. 2026c), and LfI (Gong et al. 2026). We report the average accuracy over all incremental stages, denoted as $\bar{\mathcal{A}}$, and the final accuracy after the last task, denoted as $\mathcal{A}_l$.

As shown in Table 1, DuLBE achieves the strongest overall performance under both standard settings. In Setting-A, DuLBE ranks first on all four datasets and all reported metrics. The advantage is particularly evident on CIFAR, where

Method	Setting-A (CLIP)								Setting-B (OpenCLIP)							
	CIFAR		CUB		I.N.-R		I.N.100		CIFAR		SUN		Cars		Food	
	$\bar{A}$	A_l	$\bar{A}$	A_l	$\bar{A}$	A_l	$\bar{A}$	A_l	$\bar{A}$	A_l	$\bar{A}$	A_l	$\bar{A}$	A_l	$\bar{A}$	A_l
ContinualCLIP	—	66.7	—	51.2	—	72.0	—	75.4	—	71.4	—	72.1	—	76.4	—	81.9
DualPrompt (ECCV’22)	81.5	72.5	—	—	82.0	75.8	80.7	67.4	81.6	72.4	82.5	74.4	76.3	62.9	84.9	77.3
CODA (CVPR’23)	77.0	62.3	66.6	50.9	78.0	67.5	64.1	34.8	82.4	73.4	83.3	75.7	80.2	66.5	86.2	78.8
PROOF (TPAMI’25)	86.2	76.3	—	—	82.8	77.1	84.7	72.5	86.8	79.1	83.9	77.3	90.7	86.5	90.0	84.7
RAPF (ECCV’24)	86.2	79.0	82.7	76.2	85.6	80.3	<u>87.5</u>	<u>80.2</u>	86.1	78.0	82.1	72.5	82.9	62.9	88.6	81.2
CLG-CBM (CVPR’25)	84.9	76.9	82.9	77.8	—	—	86.0	78.5	—	—	—	—	—	—	—	—
ENGINE (ICCV’25)	82.1	73.1	83.9	76.2	84.4	77.0	—	—	86.9	79.2	85.0	78.5	<u>94.1</u>	90.1	<u>89.8</u>	<u>83.9</u>
DesCLIP (TMM’26)	85.6	78.7	<u>85.1</u>	<u>78.7</u>	86.4	77.8	80.7	72.2	—	—	—	—	—	—	—	—
MG-CLIP (ICCV’25)	<u>87.0</u>	<u>80.6</u>	<u>80.6</u>	<u>72.0</u>	87.6	<u>82.7</u>	87.3	78.4	<u>89.1</u>	<u>82.4</u>	55.9	43.0	88.4	80.7	88.2	82.2
BOFA (AAAI’26)	—	—	—	—	—	—	—	—	86.5	79.3	<u>85.6</u>	<u>78.9</u>	93.8	89.2	89.0	82.7
LoDA-CLIP (ICML’26)	86.8	<u>81.5</u>	84.2	78.1	<u>87.8</u>	82.0	85.9	76.4	88.9	81.7	—	—	94.3	90.6	—	—
DuLBE (Ours)	89.5	84.2	86.1	80.6	88.4	84.3	88.3	80.7	90.0	84.5	86.7	80.3	<u>94.1</u>	<u>90.3</u>	91.3	86.7

Table 1: Main comparison under two standard CLIP-based CIL settings. Setting-A (Huang et al. 2024) uses OpenAI CLIP ViT-B/16, and Setting-B (Zhou et al. 2025a) uses OpenCLIP ViT-B/16 pretrained on LAION-400M. All methods are evaluated using the standard 10-task split without *external training data*. The best results are highlighted in **bold**, and the second-best results are underlined.

DuLBE improves $\bar{A}$ and A_l over the best previous results by 2.5% and 2.7%, respectively. It also obtains clear final-accuracy gains on CUB and ImageNet-R, demonstrating that the proposed dual-mode allocation benefits both general object recognition and fine-grained visual adaptation. In Setting-B, DuLBE achieves the best results on most datasets and shows especially strong retention on CIFAR and Food, with A_l improvements of 2.1% and 2.0% over the second. Broader baseline comparisons and backbone scalability results are reported in Appendix C.1 and C.3, respectively.

We further provide a detailed *efficiency-performance* comparison in Figure 3. Compared with parameter-tuning methods: LoRA-CLIP, MG-CLIP, LfI, LoDA-CLIP, and SGCL under different adaptation budgets, DuLBE achieves a more favorable trade-off. In particular, the *shared-mode-only* variant DuLBE_{r1} uses only about 0.065M trainable parameters in the visual tower, while DuLBE_{r1+8} still requires only about 0.58M parameters after adding the *residual-mode* low-rank learner. Despite these modest parameter budgets, DuLBE remains in the upper-left region of the plot, outperforming higher-rank LoRA variants and full fine-tuning counterparts. This confirms that the improvement of DuLBE does not arise from increasing the adaptation scale, but from allocating the key subspaces for transferable shared knowledge and low-interference task-specific variations. Moreover, we provide detailed analyses of training, inference, and CIL memory overheads in Appendix C.5.

Further Analysis

Effectiveness of Dual-Mode Collaborative Learning.

Table 2 shows that both shared and residual modes improve over the LoRA baseline through complementary roles. The shared mode reuses historically occupied directions but remains sensitive to unconstrained rewriting, while the residual mode captures task-specific variations with lower interference. This distinction is supported by Figure 4, where semantic-structure constraint substantially reduces the drift

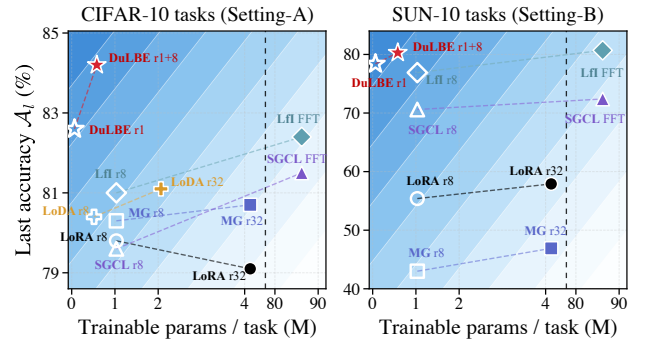


Figure 3: Efficiency-performance comparison with benchmark methods. For a fair comparison, trainable parameters are counted only from the key/value attention projections and MLP layers in CLIP’s visual tower. Here, r denotes the low-rank dimension, and FFT denotes full fine-tuning.

of the shared mode, whereas the residual mode naturally maintains low drift. Simply combining both modes provides limited gains, indicating that mode allocation alone is insufficient for effective collaboration. Regularizing both modes, however, restricts residual plasticity. Routing $\mathcal{L}_{bi-kl}$ only to the shared mode enables controlled knowledge reuse while preserving flexible residual adaptation, yielding the greatest gains of 3.4/5.7% on CIFAR and 5.1/6.8% on CUB in $\bar{A}/A_l$.

Analysis of Bridge-Prototype Ensemble. Table 3 evaluates the bridge-prototype ensemble. Fixed bridge depths (rows 1–3) consistently outperform the text-only classifier, confirming the discriminative value of the inter-modal region. Uniformly combining multiple depths remains effective but is not consistently superior to the best fixed depth, as different classes may favor different bridge regions. Reliability weighting captures this variation and achieves the best performance on both datasets. However, its effect on

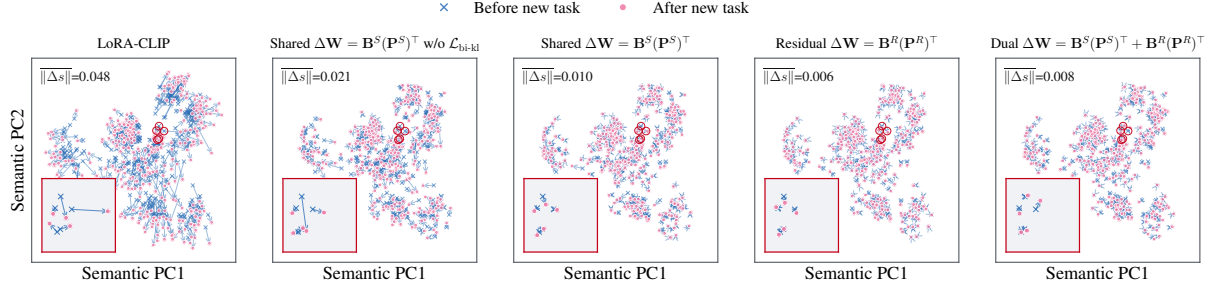


Figure 4: Visualization of old-task semantic-coordinate drift. Each old sample is represented by its cosine-similarity vector to Task-1 class-text embeddings and projected to two dimensions. Arrows show the shift after learning the new task, and $\|\Delta s\|$ denotes the average drift in the original text-semantic coordinate space.

Mode bases	$\nabla_{\mathcal{L}_{c-ce}}$		$\nabla_{\mathcal{L}_{bi-kl}}$		CIFAR (A)		CUB (A)	
	$\mathbf{B}^S$	$\mathbf{B}^R$	$\mathbf{B}^S$	$\mathbf{B}^R$	$\bar{\mathcal{A}} \uparrow$	$\mathcal{A}_l \uparrow$	$\bar{\mathcal{A}} \uparrow$	$\mathcal{A}_l \uparrow$
LoRA* ($r = 32$)	–	–	–	–	(86.1)	(78.5)	(81.0)	(73.8)
Shared $\mathbf{P}^S$	✓	–	✓	×	+2.4	+4.1	+3.2	+4.0
Residual $\mathbf{P}^R$	–	✓	×	×	+2.5	+4.4	+4.7	+5.5
Dual $[\mathbf{P}^S, \mathbf{P}^R]$	✓	✓	×	×	+1.3	+2.4	+1.1	+1.5
Dual $[\mathbf{P}^S, \mathbf{P}^R]$	✓	✓	✓	✓	+2.6	+4.4	+4.2	+5.3
Dual $[\mathbf{P}^S, \mathbf{P}^R]$	✓	✓	✓	×	+3.4	+5.7	+5.1	+6.8

Table 2: Ablation of dual-mode collaborative learning. All values are improvements over the LoRA* baseline, which replaces the low-rank learner with standard LoRA while keeping all other components unchanged.

Method	Text Clf.		B.E. Clf.			
	CUB	Food	α	Rel.	CUB	Food
Ours	67.9	83.7	0.25	×	77.9	85.0
Ours	–	–	0.50	×	77.7	83.9
Ours	–	–	0.75	×	72.9	84.2
Ours	–	–	$\{\alpha_k\}$	×	79.1	84.6
Ours	–	–	$\{\alpha_k\}$	✓	80.6(+12.7)	86.7(+3.0)
LoRA-CLIP _{r8}	63.2	79.6	$\{\alpha_k\}$	✓	73.6(+10.4)	80.0(+0.4)
RAPF	76.2	81.2	$\{\alpha_k\}$	✓	71.9(-4.3)	74.2(-7.0)
LfI _{r8}	79.3	87.1	$\{\alpha_k\}$	✓	75.4(-3.9)	84.0(-3.1)

Table 3: Ablation of the bridge-prototype ensemble on CUB (A) and Food (B). All entries report the last accuracy $\mathcal{A}_l$. B.E. denotes bridge-prototype ensemble and Rel. denotes reliability weighting over bridge depths.

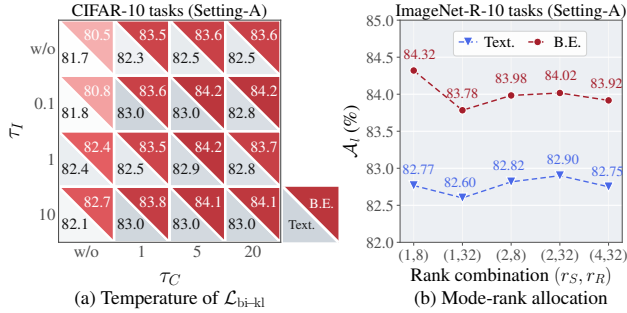


Figure 5: Hyper-parameter analysis of (a) the class-wise and instance-wise temperatures in $\mathcal{L}_{bi-kl}$ and (b) the shared/residual rank allocation.

other learners is less consistent: it improves LoRA-CLIP but degrades LfI and RAPF. Thus, the ensemble is not a universally effective plug-in. Its benefit depends on a compatible inter-modal structure, which DuLBE explicitly stabilizes.

Hyper-parameter Analysis. DuLBE is generally robust to the KL temperatures and mode-rank allocation. As shown in Figure 5, the setting $\tau_C = 5$ and $\tau_I = 0.1$ achieves the best bridge-ensemble accuracy, while nearby choices yield comparable results. Moreover, the compact rank allocation $(r_S, r_R) = (1, 8)$ performs best, and increasing the ranks

brings no consistent benefit. This suggests that a single compact direction is sufficient for knowledge transfer, while a modest residual space provides adequate plasticity. Additionally, the number of geodesic bridge depths K_b and the reliability-smoothing temperature β are further analyzed in Appendix C.4.

Conclusion

We presented DuLBE, an exemplar-free Vision-Language CIL framework that addresses continual adaptation with a dual-mode low-rank learner and a bridge-prototype ensemble classifier. Through subspace allocation and routed objective learning, DuLBE allocates transferable cross-modal knowledge to a compact rewritable shared mode and task-specific variations to low-interference residual modes, enabling controlled knowledge reuse and flexible adaptation. Building on the resulting stable inter-modal structure, the bridge-prototype ensemble classifier exploits discriminative prototypes in the inter-modal semantic region with bridge-depth reliability. Experiments under standard CIL settings show that DuLBE achieves state-of-the-art performance while preserving the parameter efficiency of low-rank adaptation.

References

- Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101: Mining Discriminative Components with Random Forests. In *European Conference on Computer Vision*.
- Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible Scaling Laws for Contrastive Language-Image Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2818–2829.
- De Lange, M.; Aljundi, R.; Masana, M.; Parisot, S.; Jia, X.; Leonardis, A.; Slabaugh, G.; and Tuytelaars, T. 2021. A Continual Learning Survey: Defying Forgetting in Classification Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7): 3366–3385.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.
- Douillard, A.; Cord, M.; Ollion, C.; Robert, T.; and Valle, E. 2020. PODNet: Pooled Outputs Distillation for Small-Tasks Incremental Learning. In *European Conference on Computer Vision*, 86–102.
- Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2024. Clip-adapter: Better vision-language models with feature adapters. *International journal of computer vision*, 132(2): 581–595.
- Gong, Y.; Yu, S.; Al-Nuaimy, W.; and Xiao, J. 2026. Learning from Itself: Mining Internal Knowledge from Vision Language Models for Continual Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10830–10839.
- He, C.; Qiu, Z.; Meng, F.; Xu, L.; Wu, Q.; and Li, H. 2026a. DesCLIP: Robust Continual Learning via General Attribute Descriptions for VLM-Based Visual Recognition. *IEEE Transactions on Multimedia*, 28: 5021–5035.
- He, C.; Qiu, Z.; Meng, F.; Zhang, R.; Xu, L.; Wu, Q.; and Li, H. 2026b. Continual Learning with Vision-Language Models via Semantic-Geometry Preservation. *arXiv:2603.12055*.
- He, J.; Duan, Z.; and Zhu, F. 2025. CL-LoRA: Continual low-rank adaptation for rehearsal-free class-incremental learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 30534–30544.
- He, L.; Cheng, D.; Wang, H.; Yang, X.; Wang, N.; and Gao, X. 2026c. Task-Driven Subspace Decomposition for Knowledge Sharing and Isolation in LoRA-Based Continual Learning. In *International Conference on Machine Learning*.
- He, L.; Cheng, D.; Xu, D.; Wang, H.; and Wang, N. 2026d. Harnessing Textual Semantic Priors for Knowledge Transfer and Refinement in CLIP-Driven Continual Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T.; Parajuli, S.; Guo, M.; Song, D.; Steinhardt, J.; and Gilmer, J. 2021. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Horn, R. A.; and Johnson, C. R. 2012. *Matrix analysis*. Cambridge university press.
- Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.
- Hu, T.; Li, L.; Xie, Z.-H.; and Zhou, D.-W. 2025. Hierarchical Semantic Tree Anchoring for CLIP-Based Class-Incremental Learning. *arXiv preprint arXiv:2511.15633*.
- Huang, L.; Cao, X.; Lu, H.; and Liu, X. 2024. Class-Incremental Learning with CLIP: Adaptive Representation Adjustment and Parameter Fusion. In *European Conference on Computer Vision*, 214–231.
- Huang, L.; Cao, X.; Lu, H.; Meng, Y.; Yang, F.; and Liu, X. 2025. Mind the Gap: Preserving and Compensating for the Modality Gap in CLIP-Based Continual Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Kang, B.; Wang, L.; Wu, Z.; Feng, T.; Li, Y.; Gao, Y.; and Li, W. 2025. Dynamic Multi-Layer Null Space Projection for Vision-Language Continual Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*.
- Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3D Object Representations for Fine-Grained Categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 554–561.
- Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto.
- Li, L.; Hu, T.; Zhou, D.-W.; Yang, J.-Q.; Ye, H.-J.; and Zhan, D.-C. 2026. BOFA: Bridge-Layer Orthogonal Low-Rank Fusion for CLIP-Based Class-Incremental Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 22967–22975.
- Li, Z.; and Hoiem, D. 2017. Learning without Forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12): 2935–2947.
- Liang, G.; Qin, C.; Cheng, D.; Zhang, S.; and Zhang, Y. 2025. Boosting Multi-Modal Alignment: Geometric Feature Separation for Class Incremental Learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 1880–1889.
- Liang, Y.-S.; and Li, W.-J. 2024. InfLoRA: Interference-Free Low-Rank Adaptation for Continual Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23638–23647.
- Luo, M.-L.; Zhou, Z.-H.; Zhang, Y.-L.; Wan, Y.; Wei, T.; and Zhang, M.-L. 2026. KeepLoRA: Continual Learning with Residual Gradient Adaptation. *arXiv preprint arXiv:2601.19659*.
- Peng, T.; Liu, Y.; Yang, S.; Hong, Q.; and Tian, Y. 2025. GNSP: Gradient Null Space Projection for Preserving Cross-Modal Alignment in VLMs Continual Learning. *arXiv:2507.19839*.

- Qiu, Z.; Wang, L.; Cao, Y.; Zhang, R.; Su, B.; Xu, Y.; Meng, F.; Xu, L.; Wu, Q.; and Li, H. 2026. Null-Space Filtering for Data-Free Continual Model Merging: Preserving Stability, Promoting Plasticity. In *International Conference on Learning Representations (ICLR)*. Poster.
- Qiu, Z.; Xu, Y.; He, C.; Meng, F.; Xu, L.; Wu, Q.; and Li, H. 2025. MINGLE: Mixture of Null-Space Gated Low-Rank Experts for Test-Time Continual Model Merging. *Advances in Neural Information Processing Systems*, 38: 143841–143878.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning*, 8748–8763.
- Rebuffi, S.-A.; Kolesnikov, A.; Sperl, G.; and Lampert, C. H. 2017. iCaRL: Incremental Classifier and Representation Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2001–2010.
- Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. arXiv:2111.02114.
- Smith, J. S.; Karlinsky, L.; Gutta, V.; Cascante-Bonilla, P.; Kim, D.; Arbellet, A.; Panda, R.; Feris, R.; and Kira, Z. 2023. 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*, 11909–11919.
- Thengane, V.; Khan, S.; Hayat, M.; and Khan, F. S. 2022. CLIP Model is an Efficient Continual Learner. arXiv:2210.03114.
- Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.
- Wang, X.; Chen, T.; Ge, Q.; Xia, H.; Bao, R.; Zheng, R.; Zhang, Q.; Gui, T.; and Huang, X. 2023. Orthogonal Subspace Learning for Language Model Continual Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 10658–10671. Association for Computational Linguistics.
- Wang, Z.; Zhang, Z.; Ebrahimi, S.; Sun, R.; Zhang, H.; Lee, C.-Y.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022. DualPrompt: Complementary Prompting for Rehearsal-Free Continual Learning. In *European Conference on Computer Vision*, 631–648.
- Wu, B.; Shi, W.; Wang, J.; and Ye, M. 2025. Synthetic data is an elegant gift for continual vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2813–2823.
- Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. SUN Database: Large-Scale Scene Recognition from Abbey to Zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3485–3492.
- Yin, B.; Zhao, J.; Jiang, H.; Hou, N.; Hu, Y.; Beheshti, A.; Yang, M.-H.; and Qi, Y. 2024. Adapter-Enhanced Semantic Prompting for Continual Learning. arXiv:2412.11074.
- Yu, L.; Han, H.; Tao, Z.; Yao, H.; and Xu, C. 2025. Language Guided Concept Bottleneck Models for Interpretable Continual Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14976–14986.
- Yu, L.; Tao, Z.; Goswami, D.; Yao, H.; Twardowski, B.; Van de Weijer, J.; and Xu, C. 2024. Exploiting the Semantic Knowledge of Pre-trained Text-Encoders for Continual Learning. *arXiv preprint arXiv:2408.01076*.
- Zhang, W.; Janson, P.; Aljundi, R.; and Elhoseiny, M. 2024. Overcoming generic knowledge loss with selective parameter update. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24046–24056.
- Zheng, Z.; Ma, M.; Wang, K.; Qin, Z.; Yue, X.; and You, Y. 2023. Preventing zero-shot transfer degradation in continual learning of vision-language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 19125–19136.
- Zhou, D.-W.; Li, K.-W.; Ning, J.; Ye, H.-J.; Zhang, L.; and Zhan, D.-C. 2025a. External Knowledge Injection for CLIP-Based Class-Incremental Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Zhou, D.-W.; Zhang, Y.; Wang, Y.; Ning, J.; Ye, H.-J.; Zhan, D.-C.; and Liu, Z. 2025b. Learning Without Forgetting for Vision-Language Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6): 4489–4504.
- Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9): 2337–2348.

Dual-Mode Low-Rank Learner with Bridge-Prototype Ensemble for Vision-Language Class-Incremental Learning

Appendix

Chiyuan He¹, Zihuan Qiu¹, Fanman Meng^{1,✉}, Chao Wang², Liangjiang Chen¹, Linfeng Xu¹,
Qingbo Wu¹, Hongliang Li¹

¹University of Electronic Science and Technology of China, Chengdu, China

²Qiyuan Lab, Beijing, China

{cyhe,zihuanqiu,202522011613}@std.uestc.edu.cn
{lfxu,qbwu,hlli}@uestc.edu.cn, w-c15@tsinghua.org.cn

Overview

The appendix is organized into four parts. Section A analyzes the main design choices and their supporting evidence. Section B details the experimental protocols and training configurations. Section C reports additional experimental results. Section D gives the complete procedure of proposed DuLBE.

A Method Analysis and Evidence

A.1 Rewritable Shared Mode

This subsection supplements the rewritable *Shared Mode* in the main paper by explaining why DuLBE selects a compact, layer-wise direction instead of making a large-rank historical subspace jointly rewritable. For each adapted layer l , the accumulated historical statistic is decomposed as

$$\mathbf{S}_{< t}^l = \mathbf{U}_{< t}^l \mathbf{\Lambda}_{< t}^l (\mathbf{U}_{< t}^l)^\top. \quad (\text{A.1})$$

The top- K_h eigenvectors $\mathbf{U}_{< t}^{l, K_h}$ define the historical principal support $\mathcal{H}_{< t}^l$. Directions in this support have been repeatedly activated by previous tasks, but only a small subset may also be demanded by current task. Therefore, consider a rank- r shared basis $\mathbf{P}_t^{l, S}(r) \subseteq \text{span}(\mathcal{H}_{< t}^l)$. With the basis fixed, one gradient step on its trainable up-projection gives

$$\begin{aligned} \mathcal{L}_{\text{c-ce}}(\mathbf{W}_{t-1}^l + \Delta \mathbf{W}_t^{l, S}) - \mathcal{L}_{\text{c-ce}}(\mathbf{W}_{t-1}^l) \\ = -\eta \|\mathbf{G}_t^l \mathbf{P}_t^{l, S}(r)\|_F^2 + O(\eta^2). \end{aligned} \quad (\text{A.2})$$

Thus, the projected gradient energy measures how much current-task learning signal can be absorbed through the selected historical directions. To maximize this energy, define

$$\mathbf{M}_t^l = (\mathbf{U}_{< t}^{l, K_h})^\top (\mathbf{G}_t^l)^\top \mathbf{G}_t^l \mathbf{U}_{< t}^{l, K_h}, \quad (\text{A.3})$$

and let $\mu_1^l \geq \dots \geq \mu_{K_h}^l$ be its eigenvalues, with corresponding eigenvectors $\{\mathbf{q}_i^l\}$. By the Ky Fan maximum principle Horn and Johnson (2012), the optimal rank- r shared basis is

$$\begin{aligned} \mathbf{P}_t^{l, S}(r) &= \mathbf{U}_{< t}^{l, K_h} [\mathbf{q}_1^l, \dots, \mathbf{q}_r^l], \\ \|\mathbf{G}_t^l \mathbf{P}_t^{l, S}(r)\|_F^2 &= \sum_{i=1}^r \mu_i^l. \end{aligned} \quad (\text{A.4})$$

✉ Corresponding author: fmmeng@uestc.edu.cn. Preprint. Work in progress.

Therefore, μ_i^l represents the marginal current-task benefit of making the corresponding historical direction rewritable.

Although a larger shared rank captures more gradient energy, it also opens more historically occupied directions to modification. From a sparse-update perspective, this trade-off can be expressed by the analytical surrogate

$$\max_{1 \leq r \leq K_h} \left[\sum_{i=1}^r \mu_i^l - \gamma_l r \right], \quad (\text{A.5})$$

where γ_l denotes the conceptual cost of exposing one additional historical direction to rewriting and is not an extra training hyperparameter. Increasing the shared rank from r to $r+1$ is useful only when its additional gradient benefit μ_{r+1}^l exceeds this rewriting cost. When the projected spectrum is concentrated, the leading direction retains most transferable gradient energy, whereas the remaining directions provide limited benefit while increasing the risk of disturbing old visual-text relations. DuLBE therefore adopts the sparse choice $r_S = 1$:

$$\mathbf{P}_t^{l, S} = \mathbf{U}_{< t}^{l, K_h} \text{eig}_{\max}(\mathbf{M}_t^l). \quad (\text{A.6})$$

This layer-wise construction retains the strongest overlap between historical occupation and current cross-modal demand while avoiding unnecessary high-rank rewriting.

A.2 Residual Modes and Gradient Routing

This subsection supplements the low-interference *Residual Modes* and *Gradient Routing* in the main paper. We analyze two connected aspects: how the residual mode captures task-specific information not covered by the compact shared mode, and why routing the KL-constraint gradient to this space would introduce noise into its plastic exploration.

Residual-mode construction. Because $\mathbf{P}_t^{l, S}$ already represents the most useful direction inside $\mathcal{H}_{< t}^l$, allocating additional residual capacity in the same support would duplicate the shared branch and expose more historical structure to modification. DuLBE therefore projects the prospective gradient onto the complementary space:

$$\begin{aligned} \mathbf{\Pi}_{< t}^{l, \perp} &= \mathbf{I} - \mathbf{U}_{< t}^{l, K_h} (\mathbf{U}_{< t}^{l, K_h})^\top, \\ \mathbf{G}_{t, R}^l &= \mathbf{G}_t^l \mathbf{\Pi}_{< t}^{l, \perp}. \end{aligned} \quad (\text{A.7})$$

This removes gradient components associated with the principal historical support while retaining current-task demand that cannot be absorbed by the shared mode. Let

$$\mathbf{G}_{t,R}^l = \mathbf{A}_t^l \Sigma_t^l (\mathbf{V}_t^l)^\top \quad (\text{A.8})$$

be its singular value decomposition, with singular values $\sigma_1^l \geq \sigma_2^l \geq \dots$. We choose to select

$$\mathbf{P}_t^{l,R} = [\mathbf{v}_{t,1}^l, \dots, \mathbf{v}_{t,r_R}^l]. \quad (\text{A.9})$$

For any orthonormal basis $\mathbf{R}$ satisfying $\mathbf{R}^\top \mathbf{U}_{<t}^{l,K_h} = \mathbf{0}$, we have $\mathbf{G}_t^l \mathbf{R} = \mathbf{G}_{t,R}^l \mathbf{R}$. The variational characterization of singular values therefore gives

$$\begin{aligned} \mathbf{P}_t^{l,R} \in \arg \max_{\substack{\mathbf{R}^\top \mathbf{R} = \mathbf{I}_{r_R} \\ \mathbf{R}^\top \mathbf{U}_{<t}^{l,K_h} = \mathbf{0}}} \|\mathbf{G}_t^l \mathbf{R}\|_F^2, \end{aligned} \quad (\text{A.10})$$

$$\|\mathbf{G}_t^l \mathbf{P}_t^{l,R}\|_F^2 = \sum_{i=1}^{r_R} (\sigma_i^l)^2.$$

Accordingly, the residual basis is not an arbitrary orthogonal component: it is the rank- r_R subspace outside the historical support that retains the largest current-task gradient energy. A gradient step through this basis provides the first-order loss decrease:

$$\begin{aligned} \mathcal{L}_{\text{c-ce}}(\mathbf{W}_{t-1}^l + \Delta \mathbf{W}_t^{l,R}) - \mathcal{L}_{\text{c-ce}}(\mathbf{W}_{t-1}^l) \\ = -\eta \sum_{i=1}^{r_R} (\sigma_i^l)^2 + O(\eta^2). \end{aligned} \quad (\text{A.11})$$

The same construction also limits historical interference. Let $\rho_{K_h+1}^l$ denote the largest eigenvalue of $\mathbf{S}_{<t}^l$ outside its top- K_h principal support. Since $\mathbf{P}_t^{l,R}$ lies in this complementary space, the accumulated response of $\Delta \mathbf{W}_t^{l,R} = \mathbf{B}_t^{l,R} (\mathbf{P}_t^{l,R})^\top$ on historical layer inputs satisfies

$$\sum_{k<t} \|\mathbf{X}_k^l \mathbf{P}_t^{l,R} (\mathbf{B}_t^{l,R})^\top\|_F^2 \leq \rho_{K_h+1}^l \|\mathbf{B}_t^{l,R}\|_F^2. \quad (\text{A.12})$$

Thus, the residual mode simultaneously maximizes the remaining current-task learning signal and bounds its response on historical activations. **This explains why DuLBE can allocate a larger rank r_R to task-specific plasticity while keeping the historically sensitive shared mode compact.**

Gradient routing. Let $\mathbf{g}_t^{l,R} = \nabla_{\mathbf{B}_t^{l,R}} \mathcal{L}_{\text{c-ce}}$ denote the useful current-task gradient in the residual branch. The bidirectional KL objective instead preserves the teacher’s historical semantic structure, and its systematic contribution is handled by the shared branch that rewrites historically occupied directions. Under this intended decomposition, the remaining mini-batch KL component in the residual branch mainly reflects finite-sample variation, imperfect subspace estimation, and nonlinear cross-layer coupling. We denote this component by $\xi_t^{l,R}$ and model it as

$$\mathbb{E}[\xi_t^{l,R}] = \mathbf{0}, \quad \mathbb{E}[\|\xi_t^{l,R}\|_F^2] = (\sigma_t^{l,R})^2. \quad (\text{A.13})$$

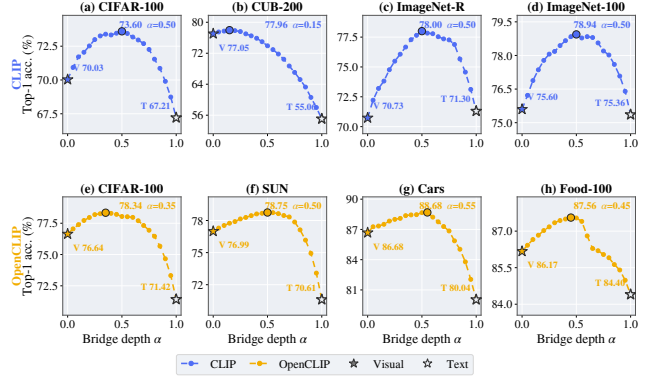


Figure A.1: Dataset-level top-1 accuracy across bridge depths using frozen OpenAI CLIP (top) and OpenCLIP (bottom). Filled and hollow stars denote the visual and textual endpoints, respectively.

If this component were routed to the residual branch, its update can become $\mathbf{B}_{t,+}^{l,R} = \mathbf{B}_t^{l,R} - \eta(\mathbf{g}_t^{l,R} + \lambda \xi_t^{l,R})$.

Assuming that $\mathcal{L}_{\text{c-ce}}$ is locally L_l -smooth with respect to $\mathbf{B}_t^{l,R}$, the expected current-task loss after this noisy update satisfies

$$\begin{aligned} \mathbb{E}[\mathcal{L}_{\text{c-ce}}(\mathbf{B}_{t,+}^{l,R})] &\leq \mathcal{L}_{\text{c-ce}}(\mathbf{B}_t^{l,R}) - \eta \left(1 - \frac{L_l \eta}{2}\right) \|\mathbf{g}_t^{l,R}\|_F^2 \\ &\quad + \frac{L_l \eta^2 \lambda^2}{2} (\sigma_t^{l,R})^2. \end{aligned} \quad (\text{A.14})$$

The final positive term is introduced solely by residual KL noise. It weakens the guaranteed current-task loss decrease and perturbs the dominant task-specific directions selected by Eq. (A.10). Consequently, we choose to route the KL gradient only to $\mathbf{B}_t^{l,S}$, where historical-structure preservation is required, while updating $\mathbf{B}_t^{l,R}$ only with the current-task objective to preserve efficient plastic exploration.

A.3 Analysis of Bridge-Prototype Ensemble

Dataset and class-level evidence. The bridge-prototype ensemble is motivated by the fact that visual and text prototypes provide complementary class information, while their relative importance varies across scenarios and classes. We reveal this kind of effect using frozen CLIP and vary only the bridge depth α , where $\alpha = 0$ and $\alpha = 1$ denote the visual and textual endpoints, respectively. As shown in Figure A.1, all 8 scenarios achieve their best accuracy at an intermediate depth, with α^* ranging from 0.15 to 0.55. The best bridge exceeds the stronger endpoint by 0.91–6.70 percentage points, demonstrating that **the geodesic interior contains useful discriminative representations beyond either modality alone.** Figure A.2 further shows that classes within the same dataset prefer substantially different depths, spanning visual-dominant, intermediate, and text-dominant regions. These results motivate estimating bridge reliability separately for each class instead of adopting a fixed global fusion depth.

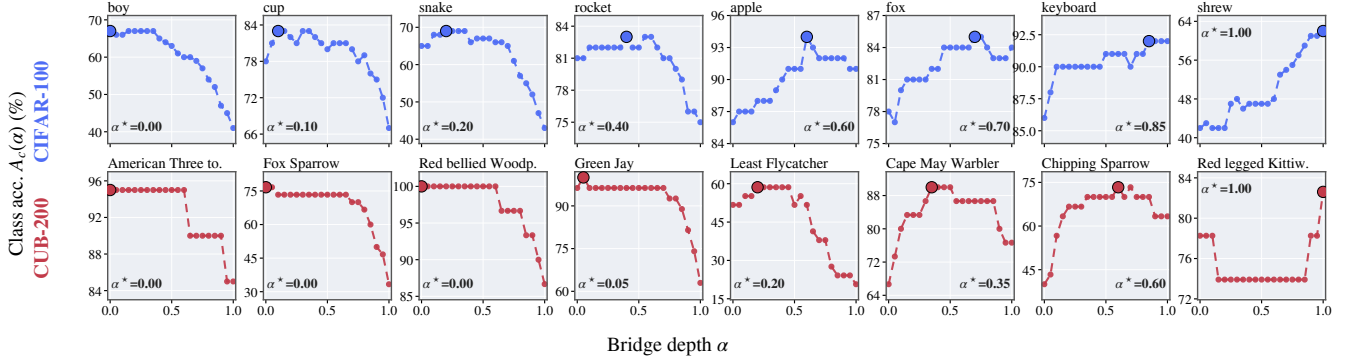


Figure A.2: Class-wise accuracy across bridge depths for representative CIFAR (top) and CUB(bottom) classes. Each class is evaluated against all classes in the corresponding dataset. Enlarged markers indicate the class-specific optimal depths.

Analysis of Geodesic Bridge Interpolation. The visual-text modality gap places visual and textual representations in distinct regions of CLIP’s shared embedding space Liang et al. (2022); Huang et al. (2025). Consequently, a visual prototype and its corresponding text embedding need not be geometrically aligned, although they describe the same class. The visual prototype captures appearance-specific evidence but may inherit specific-dependent bias, whereas the text embedding provides stable semantics but may omit fine-grained visual cues. Thus, the most discriminative class direction need not coincide with either endpoint, but may lie between them, where visual evidence and textual semantics are better integrated. The interior accuracy gains in Figure A.1 and the class-dependent optima in Figure A.2 provide direct evidence for this non-endpoint discriminative region.

Geodesic interpolation provides a calibrated transition between these two modality-specific semantic directions. Let $\theta_c = \arccos(\mathbf{v}_c^\top \mathbf{e}_c)$. The resulting bridge prototype satisfies

$$\|\mathbf{b}_c(\alpha)\|_2 = 1, \quad \angle(\mathbf{v}_c, \mathbf{b}_c(\alpha)) = \alpha\theta_c. \quad (\text{A.15})$$

Therefore, every bridge prototype remains on the CLIP unit hypersphere, and α consistently represents the fraction of semantic displacement from the visual modality toward the textual modality. Prototypes of different classes at the same depth consequently retain comparable value of logits.

For comparison, BOFA Li et al. (2026) adopts ordinary linear interpolation between the two modal endpoints:

$$\begin{aligned} \tilde{\mathbf{b}}_c(\alpha) &= (1 - \alpha)\mathbf{v}_c + \alpha\mathbf{e}_c, \\ \|\tilde{\mathbf{b}}_c(\alpha)\|_2^2 &= 1 - 2\alpha(1 - \alpha)(1 - \cos\theta_c). \end{aligned} \quad (\text{A.16})$$

Its norm, and hence its logit scale, varies with the class-specific modality gap θ_c . Although normalization removes this scale variation, the angular transition remains nonlinear, so the same α represents different semantic progress across classes. In contrast, geodesic interpolation preserves both unit norm and uniform angular depth, making each depth a comparable all-class classifier. Consequently, the reliability weights $\pi_c(\alpha_k)$ can identify the most discriminative inter-modal region (depth) for each class semantics.

B Experimental Details

B.1 Datasets and Protocols

Benchmark settings. We follow the two public protocols commonly used by the corresponding baselines. **Setting-A** is the uniform configurations Huang et al. (2024)¹ with OpenAI CLIP Radford et al. (2021) ViT-B/16. It contains CIFAR-100 Krizhevsky (2009), CUB-200-2011 Wah et al. (2011), ImageNet-R Hendrycks et al. (2021), and 100-class subset of ImageNet Deng et al. (2009). We use the default public class orders. ImageNet-R has no official train/test partition, we therefore use released random split, same as Huang et al. (2024). **Setting-B** is the uniform 10-task configurations Zhou et al. (2025b)² with OpenCLIP ViT-B/16 Cherti et al. (2023) initialized from *laion400m_e32* weights Schuhmann et al. (2021). We choose CIFAR-100, the 300-class SUN-397 Xiao et al. (2010) subset, the 100-class Stanford Cars Krause et al. (2013) subset, and the 100-class Food-101 Bossard, Guillaumin, and Van Gool (2014) subset. The class order is shuffled with seed 1993, same as Zhou et al. (2025b).

Dataset. CIFAR-100 contains 32×32 natural images from 100 object classes, with 500 training and 100 test images per class. CUB-200 and Stanford Cars are fine-grained benchmarks for 200 bird species and 196 car make/model/year categories, respectively. ImageNet-R contains artistic and non-photographic renditions of 200 ImageNet classes, whereas ImageNet-100 retains natural images from 100 ImageNet classes. SUN-397 covers diverse indoor and outdoor scenes, with 50 training and 50 test images per class in its standard partitions. Food-101 contains 750 training and 250 test images for each of 101 dishes. For Setting-B, only the retained classes in Table B.1 enter the CIL stream, and the remaining original classes are never used for training or evaluation.

Text templates. Setting-A follows Huang et al. (2024) and uses the single template $\mathcal{T}_A = \text{“a good photo of a } c\text{.”}$ for every class name c . Setting-B follows the actual configuration of Zhou et al. (2025b), which selects only the first entry of each dataset-specific template list. Thus, CIFAR-100, SUN-397,

¹<https://github.com/linlany/RAPF>

²<https://github.com/LAMDA-CL/ICCV25-ENGINE>

Table B.1: Dataset statistics and class-incremental protocols. “Classes” reports the number used by the protocol and the number in the original benchmark. “Stream” gives the **number of tasks** $\times$ **new classes per task**, and “Training template” reports the text prompt used during optimization. Train/test counts refer only to the classes used in the CIL stream.

Setting	Dataset	Classes	Train	Test	Stream	Training template
A	CIFAR-100	100/100	50,000	10,000	10 $\times$ 10	“a good photo of a <i>c</i> .”
A	CUB-200	200/200	5,994	5,794	10 $\times$ 20	“a good photo of a <i>c</i> .”
A	ImageNet-R	200/200	24,000	6,000	10 $\times$ 20	“a good photo of a <i>c</i> .”
A	ImageNet-100	100/1,000	129,395	5,000	10 $\times$ 10	“a good photo of a <i>c</i> .”
B	CIFAR-100	100/100	50,000	10,000	10 $\times$ 10	“a photo of a <i>c</i> .”
B	SUN-397	300/397	15,000	15,000	10 $\times$ 30	“a photo of a <i>c</i> .”
B	Stanford Cars	100/196	4,114	4,062	10 $\times$ 10	“a photo of a <i>c</i> .”
B	Food-101	100/101	75,000	25,000	10 $\times$ 10	“a photo of <i>c</i> , a type of food.”

and Stanford Cars use “a photo of a *c*.”, while Food-101 uses “a photo of *c*, a type of food.” The exact template for each benchmark is reported in Table B.1.

Image preprocessing. All images are mapped to 224×224 and normalized by the preprocessing pipeline associated with the corresponding pretrained CLIP model. We keep the class subsets, class orders, train/test partitions, and preprocessing fixed to the public configurations of Huang et al. (2024) and Zhou et al. (2025b).

Data access in CIL. DuLBE uses only current-task data, without storing old exemplars or introducing external data. We retain the original resource settings of the compared methods. PROOF Zhou et al. (2025c) stores 20 exemplars per observed class. CLG-CBM Yu et al. (2025), ENGINE Zhou et al. (2025b), and DesCLIP He et al. (2026a) do not replay old images but use externally generated language knowledge, such as LLM-derived class concepts or attribute descriptions. In addition, LfI Gong et al. (2026) uses COCO Captions Chen et al. (2015) as an external reference set during token generation, mixing 512 target samples with 512 COCO image-caption pairs per batch.

B.2 Introduction to Compared Methods

We provide further descriptions of the compared methods in the main. Following the standard CLIP-based CIL evaluation protocol, DualPrompt and CODA-Prompt are applied only to the visual branch of CLIP, while the other methods are implemented according to their original designs.

- **ContinualCLIP Thengane et al. (2022):** ContinualCLIP keeps CLIP frozen and performs zero-shot classification using the text embeddings of all classes observed so far. It requires neither continual optimization nor exemplar replay.
- **DualPrompt Wang et al. (2022a):** DualPrompt learns complementary general and expert prompts for a frozen pretrained transformer. The general prompts encode task-shared knowledge, whereas the expert prompts capture task-specific information.
- **CODA Smith et al. (2023):** CODA decomposes prompts into learnable components and combines them using

input-conditioned attention. This enables end-to-end prompt construction without replaying historical data.

- **PROOF Zhou et al. (2025c):** PROOF freezes the pre-trained visual and text towers and expands task-specific projection layers as new classes arrive. Previously learned projections are fixed, while a cross-modal fusion module integrates projected visual features and visual-textual prototypes.
- **RAPF Huang et al. (2024):** RAPF adopts text semantics to estimate influence of new classes on old learned classes and adaptively adjusts their representations. It further employs decomposed parameter fusion to consolidate task-wise adapter updates and prevent forgetting.
- **SGCL Yu et al. (2024b):** SGCL exploits semantic relations encoded by the pretrained CLIP text tower. It combines semantically guided representation learning with semantically guided knowledge distillation to transfer class relations within and across tasks.
- **CLG-CBM Yu et al. (2025):** CLG-CBM introduces a language-guided concept bottleneck model (CBM) between frozen CLIP features and a linear classifier. Concept alignment improves interpretability, while semantic-guided prototype augmentation generates old-class pseudo-features to alleviate forgetting.
- **ENGINE Zhou et al. (2025b):** ENGINE injects external knowledge through visual and textual branches. The visual branch enriches features through data augmentation, while the textual branch introduces discriminative descriptions. The resulting knowledge is further exploited by re-ranking predictions using abundant text prompts at inference time.
- **DesCLIP He et al. (2026a):** DesCLIP constructs robust vision-general attribute-class associations using general attribute descriptions. An anchor-based filter selects vision-relevant descriptors, which guide visual adaptation and the calibration of class-text embeddings.
- **MG-CLIP Huang et al. (2025):** MG-CLIP treats the intrinsic visual-textual modality gap as an indicator of retained pretrained cross-modal knowledge. It preserves the modality gap to reduce forgetting and compensates using visual learnable classifier in the visual space for

CIL prediction.

- **BOFA Li et al. (2026):** BOFA adapts only CLIP’s cross-modal bridge layer (projection from raw visual space to aligned visual space) and constrains its low-rank updates to a subspace approximately orthogonal to historical features. It further combines stable textual prototypes with adapted visual prototypes using linear Interpolation for classification.
- **LoDA-CLIP He et al. (2026b):** Adopts the official CLIP implementation of LoDA (Low-rank Decomposition and Adaptation) reported in the original paper. LoDA constructs general and isolated LoRA down-projection spaces from layer-wise feature statistics: the former captures directions shared across previous and current tasks, while the latter maximizes the current-to-historical feature energy ratio. After optimizing the up-projections and recalibrating the general update, the official CLIP implementation combines a frozen CLIP branch with the adapted CLIP, fusing their prediction scores at inference.
- **Lfi Gong et al. (2026):** Lfi mines the internal knowledge of CLIP without relying on an external captioning model. It constructs pseudo-captions by optimizing learnable tokens and performs adaptive mutual distillation between the textual classifier and a temporary visual classifier.

B.3 Additional Implementation Details

This subsection supplements the implementation settings reported in the main paper with the dataset-wise training schedules and further details of prospective-gradient estimation.

Training schedule and text adaptation. We train the CIL model on CIFAR for 2 epochs per task under both OpenAI CLIP and OpenCLIP. CUB, ImageNet-100, and ImageNet-R are trained for 4 epochs per task, whereas SUN, Cars, and Food under OpenCLIP are trained for 10 epochs per task. All runs use Adam with a cosine-annealing schedule, without warm-up, weight decay, or mixed-precision training. Following Huang et al. (2025), we adapt the text tower using rank-8 LoRA while keeping its pretrained weights frozen. The text-side LoRA parameters are optimized jointly with the visual learner. Since the text tower serves as a stable semantic anchor, we do not apply dual-mode allocation or historical-structure regularization to its LoRA parameters.

Prospective-gradient estimation. At the beginning of each task, before initializing the new dual-mode branches, we keep the model at its pre-task state and perform one gradient-probing pass over the current-task training data. We use the same cross-modal CE objective and current-class text embeddings as in subsequent training. The mini-batch gradients are accumulated for all adapted visual layers. No optimizer step is performed and no model parameters are modified during this pass. The resulting gradients are used only to construct the shared and residual bases and are then discarded. For the first task, the bases are initialized from the dominant gradient directions. For later tasks, the prospective gradients are combined with the accumulated historical statistics for mode allocation. This procedure uses only current-task samples and introduces no historical replay.

Overall, our optimization epochs do not exceed those of the corresponding standard protocols. Therefore, the performance gains of DuLBE do not rely on an enlarged optimization budget.

B.4 Evaluation Metrics

After learning task t , the model is evaluated on all classes observed so far, $\mathcal{C}_{\leq t}$, without task identity. We denote A_t as the resulting top-1 accuracy, which jointly reflects learning of new classes and retention of previous ones. For the stream of T tasks, we compute $\bar{A} = T^{-1} \sum_{t=1}^T A_t$. It averages accuracy across all incremental stages and measures whether the model performs consistently throughout the learning process. In addition, we define $A_l = A_T$, which evaluates the final model on all benchmark classes after the complete task stream. It summarizes the final balance between learning new classes and preserving old knowledge.

C Additional Experimental Results

C.1 Comparison with Other Baselines

To complement the standard-protocol comparison in the main paper, **we extend the evaluation along two dimensions: a) access to additional training data and b) the choice of continual low-rank adaptation strategy.** All comparisons use Setting-A with the same OpenAI CLIP ViT-B/16 backbone, 10-task class splits, and evaluation metrics as in the main paper.

Comparison with methods using additional training data.

We compare DuLBE with methods that exploit additional training data through either historical replay or external reference datasets. SGCL Yu et al. (2024b), CLAP4CLIP Jha, Gong, and Yao (2024), PROOF Zhou et al. (2025c), and SPU* Zhang et al. (2024) retain 20 historical images per observed class. In contrast, ZSCL Zheng et al. (2023) distills knowledge from the 28k-image validation set of Conceptual Captions Sharma et al. (2018), whereas Lfi Gong et al. (2026) samples 512 reference pairs from COCO Captions Chen et al. (2015) in each token-generation batch. We additionally report the replay-free SPU as a resource-matched baseline. DuLBE uses neither historical exemplars nor external training images and accesses only the current-task data.

Despite using no additional training data, DuLBE achieves the best average and final accuracies on both CIFAR and ImageNet-R, while ranking second on both metrics for CUB. It surpasses the strongest replay-based results on CIFAR by 0.3 and 0.6 points, respectively, and slightly outperforms Lfi on ImageNet-R. On the fine-grained CUB benchmark, DuLBE remains competitive with SGCL while avoiding the storage of old exemplars.

Comparison with LoRA-based methods. We further compare DuLBE with representative LoRA-based continual learning methods: O-LoRA Wang et al. (2023), In-LoRA Liang and Li (2024), CL-LoRA He, Duan, and Zhu (2025), and LoDA He et al. (2026b). We follow the original rank settings and additionally evaluate LoDA with $r = 8$. For a fair comparison, all methods insert LoRA modules into the same key and value projections and MLP layers of the visual

Method	Additional Data	Setting-A (CLIP)					
		CIFAR		CUB		I.N.-R	
		$\bar{\mathcal{A}}$	$\mathcal{A}_l$	$\bar{\mathcal{A}}$	$\mathcal{A}_l$	$\bar{\mathcal{A}}$	$\mathcal{A}_l$
SGCL Yu et al. (2024b)	Replay (20/class)	89.1	82.7	87.1	82.9	86.8	81.9
CLAP4CLIP Jha, Gong, and Yao (2024)	Replay (20/class)	85.1	76.4	85.2	79.9	85.0	79.2
PROOF Zhou et al. (2025c)	Replay (20/class)	86.2	76.3	—	—	82.8	77.1
SPU* Zhang et al. (2024)	Replay (20/class)	<u>89.2</u>	<u>83.6</u>	81.0	70.3	85.7	80.1
ZSCL Zheng et al. (2023)	External (CC val., 28k)	82.4	74.1	71.1	60.6	85.5	79.7
Lfi Gong et al. (2026)	External (COCO, 512/batch)	87.8	82.7	—	—	<u>88.1</u>	84.0
SPU Zhang et al. (2024)	None	84.9	76.4	75.9	68.2	83.1	77.5
DuLBE (Ours)	None	89.5	84.2	<u>86.1</u>	<u>80.6</u>	88.4	84.3

Table C.1: Comparison with methods using additional training data under Setting-A. Replay methods retain 20 historical images per observed class. CC and COCO denote Conceptual Captions and COCO Captions, respectively. SPU* is the replay-based variant, whereas SPU and DuLBE use neither historical exemplars nor external training images. The best results are highlighted in **bold**, and the second-best results are underlined.

Method	CIFAR		I.N.-100	
	$\bar{\mathcal{A}}$	$\mathcal{A}_l$	$\bar{\mathcal{A}}$	$\mathcal{A}_l$
O-LoRA ($r = 8/\text{task}$)	86.2	77.3	83.4	72.0
InfLoRA ($r = 10$)	87.8	80.9	85.4	75.5
CL-LoRA ($r = 10$)	83.9	75.1	83.7	73.4
LoDA ($r = 8$)	86.0	80.4	85.6	74.7
LoDA ($r = 32$)	86.4	81.1	85.8	75.7
DuLBE _{Text.} ($r = 1 + 8$)	<u>89.0</u>	<u>83.0</u>	<u>87.5</u>	<u>78.4</u>
DuLBE_{B.E.} ($r = 1 + 8$)	89.5	84.2	88.3	80.7

Table C.2: Comparison with LoRA-based methods under Setting-A. All methods use identical insertion locations in the visual tower, including the key and value projections and the MLP layers. All LoRA baselines and DuLBE_{Text.} use the same standard CLIP text classifier, whereas DuLBE_{B.E.} uses the complete reliability-guided bridge-prototype ensemble. For O-LoRA, $r = 8$ denotes the rank of each task-wise adapter. Best and second-best results are shown in **bold** and underline, respectively.

tower. Importantly, all reported LoRA baseline results use the same standard CLIP text classifier for prediction, which is also adopted by DuLBE_{Text.} for a classifier-matched comparison. These baselines are formulated for generic pretrained models and are largely agnostic to CLIP’s dual-tower structure. In contrast, DuLBE is specifically designed for CLIP-based continual learning by coordinating visual adaptation with visual-textual knowledge preservation. DuLBE_{B.E.} further reports the complete method with the bridge-prototype ensemble classifier.

C.2 Performance under the Learning-from-Half Protocol

We further evaluate DuLBE under the learning-from-half (LFH) protocol in Setting-B. The initial stage contains half of the benchmark classes and is followed by five equal

incremental stages. Specifically, the 100-class benchmark splits of CIFAR-100, Stanford Cars, and Food-101 use B50 Inc10, while the 300-class SUN-397 split uses B150 Inc30. All results are based on OpenCLIP ViT-B/16 pretrained on LAION-400M, using the same class order and random seed as in the main Setting-B experiments.

Baseline results are taken from the corresponding original papers or matched evaluations under this configuration. PROOF follows its original setting with 20 replay exemplars per observed class, while Lfi uses COCO reference pairs during training. As shown in Table C.3, we achieve the best average and final accuracies on both CIFAR and SUN. Our DuLBE achieve these results without data replay or external reference images, whereas PROOF and Lfi use replay exemplars and COCO reference pairs, respectively. These results demonstrate that DuLBE effectively preserves the knowledge acquired from a large initial task while remaining adaptable to subsequent increments, extending its advantages beyond the uniform 10-task protocol.

C.3 Performance with CLIP ViT-L/14 Backbone

To evaluate the scalability of DuLBE to a larger vision-language backbone, we replace the OpenAI CLIP Radford et al. (2021) ViT-B/16 used in Setting-A with ViT-L/14 and conduct experiments on CIFAR and ImageNet-R. We retain the same uniform 10-task CIL, with 10 and 20 new classes introduced per task, respectively. The CIL protocol remains unchanged.

These baseline results are all taken from the matched ViT-L/14 evaluation reported by Huang et al. (2025). PROOF and CLAP4CLIP use real-data replay, whereas SLCA and RAPF use feature replay. CLAP4CLIP* denotes the exemplar-free variant of CLAP4CLIP. In contrast, DuLBE uses neither replay nor external reference data. This experiment examines whether DuLBE’s dual-mode adaptation and bridge-prototype classification remain effective as the capacity of the CLIP backbone increases. As shown in Table C.4, DuLBE achieves the best results across all four metrics, while its text-classifier variant consistently ranks second. These re-

Method	Setting-B (OpenCLIP)-LFH							
	CIFAR B50 Inc10		SUN B150 Inc30		Cars B50 Inc10		Food B50 Inc10	
	$\bar{\mathcal{A}}$	$\mathcal{A}_l$	$\bar{\mathcal{A}}$	$\mathcal{A}_l$	$\bar{\mathcal{A}}$	$\mathcal{A}_l$	$\bar{\mathcal{A}}$	$\mathcal{A}_l$
ContinualCLIP Thengane et al. (2022)	76.5	71.4	75.0	72.1	78.3	76.4	84.8	81.9
DualPrompt Wang et al. (2022a)	80.1	72.6	79.4	73.0	76.9	67.6	80.0	72.8
CODA-Prompt Smith et al. (2023)	78.7	71.6	80.4	74.2	75.1	64.2	81.0	74.1
PROOF Zhou et al. (2025c)	82.9	78.9	80.7	77.5	90.5	89.5	87.5	84.7
RAPF Huang et al. (2024)	82.2	77.9	78.0	73.1	75.9	63.2	85.5	81.2
ENGINE Zhou et al. (2025b)	83.2	79.5	81.6	78.5	91.6	90.0	86.9	83.9
MG-CLIP Huang et al. (2025)	85.6	81.3	26.6	12.6	84.6	79.6	28.9	12.5
BOFA Li et al. (2026)	83.7	80.1	<u>82.3</u>	79.3	<u>92.3</u>	90.9	86.3	83.7
LfI Gong et al. (2026)	<u>87.5</u>	<u>85.6</u>	81.9	<u>79.8</u>	92.7	91.9	91.9	89.9
DuLBE (Ours)	88.5	85.8	82.9	80.1	92.1	<u>91.0</u>	<u>91.1</u>	<u>88.2</u>

Table C.3: Comparison under the learning-from-half (LFH) protocol in Setting-B. The best results are highlighted in **bold**, and the second-best results are underlined.

Method	CIFAR		I.N.-R	
	$\bar{\mathcal{A}}$	$\mathcal{A}_l$	$\bar{\mathcal{A}}$	$\mathcal{A}_l$
PROOF Zhou et al. (2025c)	89.9	83.6	91.3	87.3
CLAP4CLIP Jha, Gong, and Yao (2024)	87.9	84.9	92.1	88.6
SLCA Zhang et al. (2023)	90.1	84.6	90.0	86.8
RAPF Huang et al. (2024)	90.3	85.3	92.0	88.3
L2P++ Wang et al. (2022b)	85.7	77.9	90.5	86.7
DualPrompt Wang et al. (2022a)	86.6	79.1	90.7	87.1
CODA Smith et al. (2023)	85.8	78.7	89.1	84.6
ContinualCLIP Thengane et al. (2022)	80.5	73.5	87.0	83.1
APER-Adapter Zhou et al. (2025a)	80.2	72.0	89.2	85.4
MOE4CL Yu et al. (2024a)	91.0	85.8	93.3	90.4
CLAP4CLIP* Jha, Gong, and Yao (2024)	74.4	71.6	91.1	87.6
MagMax Marczak et al. (2024)	90.2	86.1	93.2	89.6
MG-CLIP Huang et al. (2025)	91.8	87.0	93.7	91.1
DuLBE_{Text.} (Ours)	<u>93.2</u>	<u>89.1</u>	<u>93.9</u>	<u>91.7</u>
DuLBE_{B.E.} (Ours)	93.4	89.4	94.2	92.3

Table C.4: Comparison using OpenAI CLIP ViT-L/14 under the uniform Setting-A protocol. CIFAR-100 and ImageNet-R are divided into ten tasks with 10 and 20 new classes per task, respectively. Baseline results follow the matched ViT-L/14 evaluation reported by Huang et al. (2025). CLAP4CLIP* denotes its replay-free variant. Text. and B.E. denote the conventional text classifier and bridge-prototype ensemble, respectively. The best results are highlighted in **bold**, and the second-best results are underlined.

sults confirm that the dual-mode learner remains effective with a larger CLIP backbone and that the bridge-prototype ensemble provides further improvements.

C.4 Additional Ablation Studies

We further analyze the bridge-prototype hyperparameters and the insertion locations of the dual-mode low-rank learner.

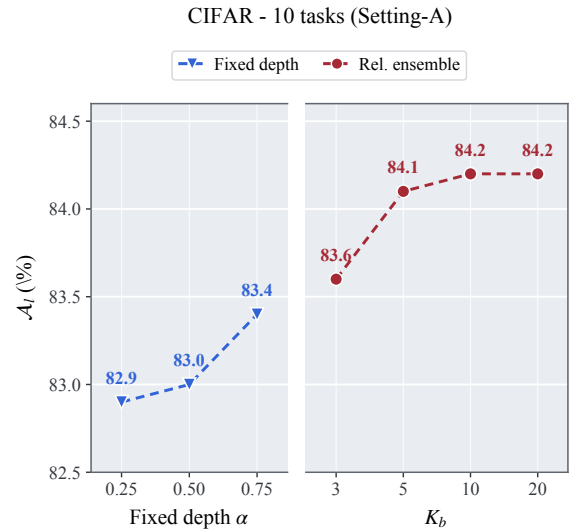


Figure C.1: Effect of fixed bridge depth α and the number of reliability-weighted bridge depths K_b on CIFAR-100 under Setting-A. Blue denotes a fixed-depth classifier, while red denotes the reliability-guided bridge-prototype ensemble.

Bridge-depth discretization and reliability smoothing. The number of bridge depths K_b controls the discretization granularity of the visual-text geodesic. β controls the smoothness of the class-wise reliability weights. We vary both hyperparameters while keeping the trained dual-mode learner and all other settings unchanged.

As shown in Figure C.1, reliability-guided ensembling consistently outperforms the fixed-depth classifiers. Increasing K_b from 3 to 10 improves $\mathcal{A}_l$ from 83.6% to 84.2%, whereas $K_b = 20$ provides no further gain. Thus, $K_b = 10$ provides sufficient bridge granularity without introducing

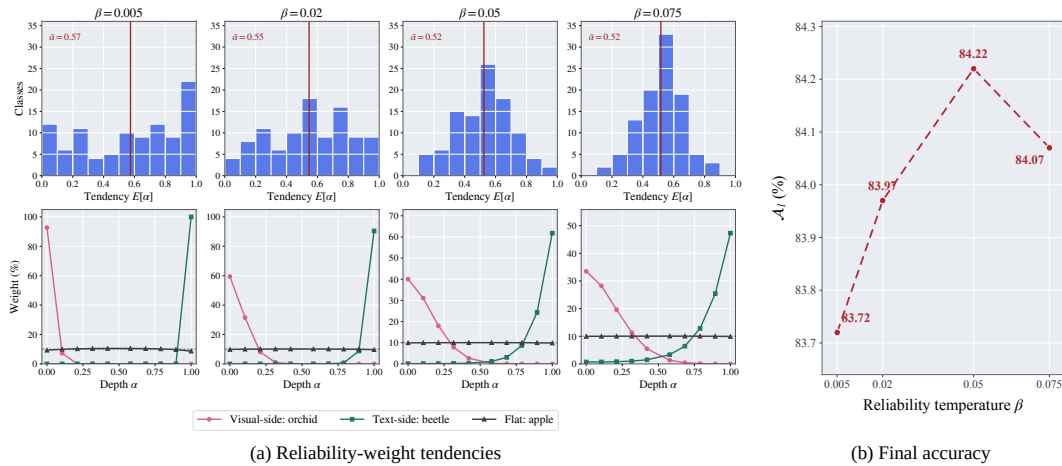


Figure C.2: Effect of the reliability temperature β with $K_b = 10$ on CIFAR-100 under Setting-A. (a) Distributions of the class-wise expected bridge depth $\mathbb{E}_{\pi_c}[\alpha]$ (top) and reliability weights of three representative classes (bottom). The red vertical line denotes the mean expected depth across classes. (b) Final accuracy under different values of β .

redundant prototypes. Figure C.2(a) shows that a very small β produces nearly one-hot depth selection, making the ensemble sensitive to reliability noise. Increasing β smoothly distributes weight over neighboring depths while preserving class-specific preferences: *orchid* favors the visual side, *beetle* favors the textual side, and *apple* remains approximately depth-insensitive. As shown in Figure C.2(b), $\mathcal{A}_t$ increases from 83.72% at $\beta = 0.005$ to 84.22% at $\beta = 0.05$, before slightly decreasing to 84.07% at $\beta = 0.075$. We therefore use $\beta = 0.05$, which provides the best balance between depth selectivity and reliability smoothing.

Insertion locations of the dual-mode low-rank learner.

We compare different insertion locations within each visual Transformer block, including the query (Q), key (K), and value (V) projections and the feed-forward network (MLP). For example, $K, V + \text{MLP}$ applies the dual-mode learner to the key and value projections and the MLP layers. All configurations use $r_S = 1$ and $r_R = 8$, while modules outside the specified locations remain frozen.

As shown in Table C.5, MLP-only adaptation performs

Adapted visual modules	$\tilde{\mathcal{A}} \uparrow$	$\mathcal{A}_t \uparrow$
K, V	88.8	83.0
Q, K, V	89.1	83.9
$K, V + \text{MLP}$ (default)	<u>89.5</u>	84.2
MLP	89.2	83.5
$Q, K, V + \text{MLP}$	89.6	<u>84.1</u>

Table C.5: Effect of the insertion locations of the dual-mode low-rank learner on CIFAR-100 under Setting-A. Here, Q , K , and V denote the query, key, and value projections in visual self-attention, while MLP denotes the feed-forward block. The same per-module ranks are used for all configurations. The best results are highlighted in **bold**, and the second-best results are underlined.

better than adapting only the key and value projections, while combining them provides complementary benefits. Including the query projection brings little additional improvement and slightly weakens final retention. We therefore adopt $K, V + \text{MLP}$ as the default configuration, providing a concise and unified adaptation scheme for both attention and feature transformation.

C.5 Overhead

Training and inference overhead. Table C.6 reports the peak number of parameters optimized within one task, including both the visual and text towers and any learnable auxiliary classifier. Inference GFLOPs are measured per image, following the convention that one ViT-B/16 visual forward costs 17.60 GFLOPs. Class-text embeddings are computed once and cached rather than repeatedly encoded for each test image. Mergeable LoRA updates are folded into the pretrained weights before inference. Additional parameters denote method-specific model parameters and persistent prototype buffers retained beyond one standard CLIP model. The common CLIP text-embedding cache is excluded.

For LoDA-CLIP He et al. (2026b), we follow its official CLIP extension, which combines predictions from a frozen pretrained visual encoder and a separately adapted visual encoder. Its additional parameters therefore include one complete ViT-B/16 visual tower and the stored visual prototypes. This differs from the single-backbone LoDA setting, in which the low-rank updates can be merged without retaining additional inference parameters.

For DuLBE, the trainable-parameter count includes rank-8 text-side LoRA modules applied to the key/value projections and MLP layers. The visual trainable parameters are 0.0645M and 0.5806M for $r = 1$ and $r = 1 + 8$, respectively. At inference, DuLBE retains only the class visual prototypes and reliability weights, requiring $|\mathcal{C}|d + |\mathcal{C}|K_b = 0.0522\text{M}$ additional parameters. The bridge prototypes are generated from these parameters and the cached text embeddings when

Method	Training Stage	Inference Stage		CIFAR (A)	
	Trainable Params (M)	GFLOPs (per sample)	Additional Params (M)	$\bar{\mathcal{A}} \uparrow$	$\mathcal{A}_l \uparrow$
RAPF Huang et al. (2024)	0.26	17.6003	0.262	86.2	79.0
LoRA-CLIP _{r=32} Hu et al. (2022)	6.88	17.6001	0.000	86.2	79.1
LoDA-CLIP _{r=32} [†] He et al. (2026b)	1.77	35.2001	86.244	86.4	81.1
MG-CLIP Huang et al. (2025)	0.54	17.6001	0.051	87.0	80.6
SGCL* Yu et al. (2024b)	7.09	17.6001	0.000	86.4	79.5
Lfi Gong et al. (2026)	149.62	17.6001	0.000	87.8	<u>82.7</u>
DuLBE _{r=1} (Ours)	0.75	17.6005	0.052	<u>88.5</u>	82.6
DuLBE _{r=1+8} (Ours)	1.27	17.6005	0.052	89.5	84.2

Table C.6: Training and inference overhead on CIFAR-100 under Setting-A. Trainable parameters cover the complete method, including trainable components in both CLIP towers. GFLOPs include visual encoding and classification over all 100 classes. SGCL* denotes the replay-free variant. MG-CLIP uses one image encoding and combines a text classifier with a learned visual classifier. DuLBE directly evaluates $K_b = 10$ bridge depths, incurring $K_b|\mathcal{C}|d = 0.000512$ GFLOPs for bridge scoring. Additional parameters include method-specific model parameters and persistent prototype buffers retained beyond the standard CLIP model, while excluding the common text-embedding cache. The best results are in **bold**, and the second-best results are underlined.

needed.

Memory overhead. We analyze two complementary sources of persistent memory: class-level representations and layer-wise historical statistics. Only numerical states retained across tasks are counted. Pretrained CLIP parameters, temporary tensors, and the class-text embedding cache shared by CLIP-based classifiers are excluded.

Representation storage. RAPF Huang et al. (2024) retains a feature mean and a full covariance matrix for each historical class, while CLG-CBM Yu et al. (2025) additionally stores its selected concept representations. Such detailed class distributions support feature-level knowledge preservation, but their storage increases rapidly with the number of classes. ENGINE Zhou et al. (2025b) is more compact because it combines class prototypes with shared GDA statistics. In contrast, DuLBE stores only one visual prototype and a small set of reliability weights per class. Its bridge prototypes are generated when needed and therefore require no additional prototype bank. As shown in Table C.7, DuLBE requires only 0.20 MiB of representation memory while achieving the highest final accuracy under Setting-A.

Layer-statistic storage. The layer-wise statistics of DuLBE are retained only at the locations where the dual-mode low-rank learner is inserted. Each statistic is updated in place, so its storage is independent of the number of tasks and observed classes. We evaluate the full configuration and three compact variants under Setting-B using OpenCLIP ViT-B/16 pretrained on LAION-400M, reporting final accuracy on CIFAR and Food.

Variant A applies the learner only to the key and value projections of all visual blocks, reducing the statistic storage from 486.00 to 27.00 MiB while achieving final accuracies of 83.9% and 85.5%. Variant B further restricts the learner to the last four visual blocks and retains comparable performance with only 9.00 MiB. Variant C adapts only the cross-modal bridge layer and uses the same 2.25 MiB statistic budget

as BOFA Li et al. (2026). Under this matched configuration, DuLBE-C outperforms BOFA by 1.9 and 1.4 points on CIFAR-100 and Food-101, respectively. **Overall, the three variants reduce statistic storage by approximately 18 $\times$, 54 $\times$, and 216 $\times$ while retaining strong final performance, demonstrating that DuLBE can be deployed flexibly under different memory budgets.**

Method	Stored state	MiB $\downarrow$	$\mathcal{A}_l \uparrow$
RAPF	Mean + covariance	100.20	79.0
ENGINE	Prototype + GDA	3.03	73.1
CLG-CBM	Mean + covariance + concepts	102.15	76.9
DuLBE	Prototype + reliability	0.20	84.2

Table C.7: Representation memory and final accuracy on CIFAR-100 under Setting-A.

Config.	Learner location	MiB $\downarrow$	$\mathcal{A}_l \uparrow$	
			CIFAR	Food
Full	K, V + MLP	486.00	84.5	86.7
A	K, V , all blocks	27.00	<u>83.9</u>	<u>85.5</u>
B	K, V , last four	9.00	83.6	85.1
C	Bridge layer	2.25	81.2	84.1
BOFA	Bridge layer	2.25	79.3	82.7

Table C.8: Statistic storage and final accuracy under Setting-B.

C.6 Visualizations

Bridge-depth attribution. We use prototype-conditioned Layer Grad $\times$ Act maps Zhao et al. (2024) to trace the visual evidence associated with each bridge depth. As shown

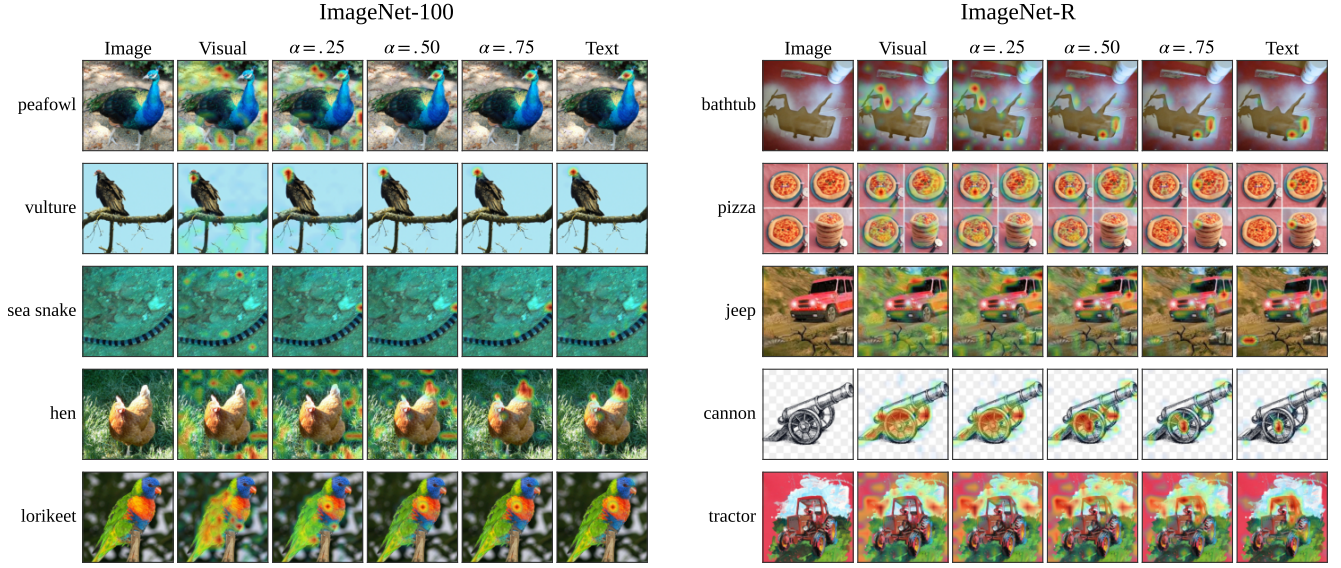


Figure C.3: Prototype-conditioned Layer Grad $\times$ Act maps on ImageNet-100 and ImageNet-R. Columns show the visual prototype, bridge prototypes at $\alpha \in \{0.25, 0.50, 0.75\}$, and the text embedding. Warmer colors indicate larger positive contributions to the image-prototype similarity.

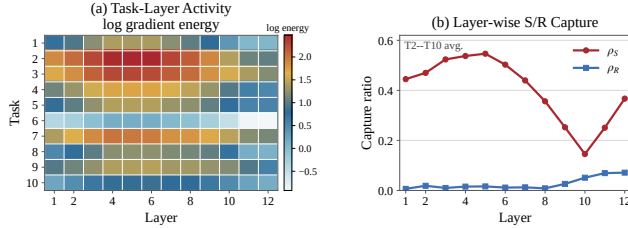


Figure C.4: CIFAR-100 gradient-space diagnostics. (a) Task-layer log expected-gradient energy across visual Transformer layers. (b) Layer-wise expected-gradient capture ratios of the shared and residual spaces, averaged over Task2–Task10 (T2–T10). The two ratios are computed independently and are not complementary.

in Figure C.3, changing the conditioning prototype alters not only the attribution strength but also its spatial distribution. The visual endpoint often captures broader appearance and contextual evidence, whereas the text endpoint tends to emphasize more localized, class-discriminative regions. Intermediate bridge prototypes provide distinct transitions between these patterns, although the transition varies across classes. Similar behavior is observed on both ImageNet-100 and the rendition-style ImageNet-R, suggesting that the intermediate prototypes remain meaningful under visual domain shifts. These observations show that different bridge depths provide complementary decision evidence, supporting class-dependent reliability weighting rather than reliance on a single fixed depth.

Gradient-space alignment. Following the gradient-demand allocation defined in the main paper, Figure C.4 analyzes the prospective cross-modal gradient $\mathbf{G}_t^l$ used to construct the two bases. Panel (a) shows that its energy varies substantially across tasks and visual layers, supporting the layer-wise allocation of low-rank directions rather than using a fixed subspace for the entire visual tower.

For $t \geq 2$, we also measure the gradient energy captured by the shared and residual bases as $\rho_S^l = \|\mathbf{G}_t^l \mathbf{P}_t^{l,S}\|_F^2 / \|\mathbf{G}_t^l\|_F^2$ and $\rho_R^l = \|\mathbf{G}_t^l \mathbf{P}_t^{l,R}\|_F^2 / \|\mathbf{G}_t^l\|_F^2$, respectively. Panel (b) shows that the compact shared basis $\mathbf{P}_t^{l,S}$ captures substantial demand in the early and middle layers, confirming that a single direction selected from the historical support $\mathcal{H}_{<t}^l$ can reuse transferable visual structure. In deeper layers, its capture ratio decreases while the residual ratio increases, indicating stronger demand outside the historical support. This demand is captured by $\mathbf{P}_t^{l,R} \subseteq \text{span}((\mathcal{H}_{<t}^l)^\perp)$ to provide low-interference task-specific plasticity. Since the two ratios measure the energy captured by two compact bases of different ranks, they are computed independently and do not form a complete decomposition of the gradient.

D Complete Procedure

Algorithm D.1 summarizes the complete task-wise procedure of DuLBE.

Algorithm D.1: Procedure of DuLBE

Input: CLIP θ_0 ; task stream $\{\mathcal{D}_t, \mathcal{C}_t\}_{t=1}^N$; adapted layers $\mathcal{V}$; hyperparameters Ω .
Output: θ_N ; stored $\{\mathbf{v}_c, \pi_c(\alpha_k)\}_{c,k}$.

1 **Initialization:** $\theta_0 \leftarrow \theta_0$ and $\mathbf{S}_{<1}^l \leftarrow \mathbf{0}, \forall l \in \mathcal{V}$;
2 **for** task $t = 1, \dots, N$ **do**

Gradient-Demand Dual-Mode Allocation

3 Estimate prospective demand: $\mathbf{G}_t^l \leftarrow \nabla_{\mathbf{W}_{t-1}^l} \mathcal{L}_{c-ce}(\theta_{t-1}; \mathcal{D}_t), \forall l \in \mathcal{V}$;
4 **for each** adapted layer $l \in \mathcal{V}$ **do**
5 Allocate modes: $(\mathbf{P}_t^{l,S}, \mathbf{P}_t^{l,R}) \leftarrow \text{ModeAlloc}(\mathbf{G}_t^l, \mathbf{S}_{< t}^l)$;
6 Freeze $\mathbf{P}_t^{l,S}, \mathbf{P}_t^{l,R}$ and initialize $\mathbf{B}_t^{l,S}, \mathbf{B}_t^{l,R} \leftarrow \mathbf{0}$;

Routed Cross-Modal Training

7 **for each** mini-batch $\mathcal{B} \subset \mathcal{D}_t$ **do**
8 $(\mathbf{g}_{ce}^{l,S}, \mathbf{g}_{ce}^{l,R}) \leftarrow \nabla_{(\mathbf{B}_t^{l,S}, \mathbf{B}_t^{l,R})} \mathcal{L}_{c-ce}, \quad \mathbf{g}_{kl}^{l,S} \leftarrow \mathbb{I}[t > 1] \nabla_{\mathbf{B}_t^{l,S}} \mathcal{L}_{bi-kl}(\bar{\theta}_{t-1})$;
9 Route and update: $\mathbf{B}_t^{l,S} \leftarrow \text{OptStep}(\mathbf{g}_{ce}^{l,S} + \lambda \mathbf{g}_{kl}^{l,S}), \quad \mathbf{B}_t^{l,R} \leftarrow \text{OptStep}(\mathbf{g}_{ce}^{l,R})$;

Reliability-Guided Bridge-Prototype Ensemble

10 **for each** new class $c \in \mathcal{C}_t$ **do**
11 $\mathbf{v}_c \leftarrow \text{Normalize} \left(\sum_{x_i \in \mathcal{X}_c^t} \mathbf{z}_i \right), \quad \mathbf{b}_c(\alpha_k) \leftarrow \text{Slerp}(\mathbf{v}_c, \mathbf{e}_c; \alpha_k)$;
12 $p_{\alpha_k}(j | x_i) \propto \exp(\tau \mathbf{z}_i^\top \mathbf{b}_j(\alpha_k)), \quad \rho_c(\alpha_k) \leftarrow |\mathcal{X}_c^t|^{-1} \sum_{x_i \in \mathcal{X}_c^t} p_{\alpha_k}(c | x_i)$;
13 $\pi_c(\alpha_k) \leftarrow \text{softmax}_k(\rho_c(\alpha_k)/\beta)$; store $\{\mathbf{v}_c, \pi_c(\alpha_k)\}_{k=1}^{K_b}$;
14 Update historical state: $\mathbf{S}_{< t+1}^l \leftarrow \mathbf{S}_{< t}^l + (\mathbf{X}_t^l)^\top \mathbf{X}_t^l, \forall l \in \mathcal{V}$;
15 $\theta_t \leftarrow \text{StopGrad}(\theta_t)$; discard $\mathcal{D}_t$;
16 Evaluate after task t : $s_c(x) \leftarrow \sum_{k=1}^{K_b} \pi_c(\alpha_k) \tau \mathbf{z}(x)^\top \mathbf{b}_c(\alpha_k), \quad \hat{y}(x) \leftarrow \arg \max_{c \in \mathcal{C}_{\leq t}} s_c(x)$;

References

- Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101: Mining Discriminative Components with Random Forests. In *European Conference on Computer Vision*, 446–461.
- Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv preprint arXiv:1504.00325*.
- Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible Scaling Laws for Contrastive Language-Image Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2818–2829.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.
- Gong, Y.; Yu, S.; Al-Nuaimy, W.; and Xiao, J. 2026. Learning from Itself: Mining Internal Knowledge from Vision Language Models for Continual Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10830–10839.
- He, C.; Qiu, Z.; Meng, F.; Xu, L.; Wu, Q.; and Li, H. 2026a. DesCLIP: Robust Continual Learning via General Attribute Descriptions for VLM-Based Visual Recognition. *IEEE Transactions on Multimedia*, 28: 5021–5035.
- He, J.; Duan, Z.; and Zhu, F. 2025. CL-LoRA: Continual Low-Rank Adaptation for Rehearsal-Free Class-Incremental Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 30534–30544.
- He, L.; Cheng, D.; Wang, H.; Yang, X.; Wang, N.; and Gao, X. 2026b. Task-Driven Subspace Decomposition for Knowledge Shar-
- ing and Isolation in LoRA-Based Continual Learning. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*.
- Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T.; Parajuli, S.; Guo, M.; Song, D.; Steinhardt, J.; and Gilmer, J. 2021. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8340–8349.
- Horn, R. A.; and Johnson, C. R. 2012. *Matrix Analysis*. Cambridge University Press, second edition.
- Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.
- Huang, L.; Cao, X.; Lu, H.; and Liu, X. 2024. Class-Incremental Learning with CLIP: Adaptive Representation Adjustment and Parameter Fusion. In *European Conference on Computer Vision*, 214–231.
- Huang, L.; Cao, X.; Lu, H.; Meng, Y.; Yang, F.; and Liu, X. 2025. Mind the Gap: Preserving and Compensating for the Modality Gap in CLIP-Based Continual Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3777–3786.
- Jha, S.; Gong, D.; and Yao, L. 2024. CLAP4CLIP: Continual Learning with Probabilistic Finetuning for Vision-Language Models. In *Advances in Neural Information Processing Systems*, volume 37, 129146–129186.
- Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3D Object Representations for Fine-Grained Categorization. In *Proceedings of*

- the *IEEE International Conference on Computer Vision Workshops*, 554–561.
- Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto.
- Li, L.; Hu, T.; Zhou, D.-W.; Yang, J.-Q.; Ye, H.-J.; and Zhan, D.-C. 2026. BOFA: Bridge-Layer Orthogonal Low-Rank Fusion for CLIP-Based Class-Incremental Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 22967–22975.
- Liang, V. W.; Zhang, Y.; Kwon, Y.; Yeung, S.; and Zou, J. Y. 2022. Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. In *Advances in Neural Information Processing Systems*, volume 35, 17612–17625.
- Liang, Y.-S.; and Li, W.-J. 2024. InfLoRA: Interference-Free Low-Rank Adaptation for Continual Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23638–23647.
- Marczak, D.; Twardowski, B.; Trzciński, T.; and Cygert, S. 2024. MAGMAX: Leveraging Model Merging for Seamless Continual Learning. In *European Conference on Computer Vision*, 379–395.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning*, 8748–8763.
- Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. arXiv:2111.02114.
- Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset for Automatic Image Captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2556–2565.
- Smith, J. S.; Karlinsky, L.; Gutta, V.; Cascante-Bonilla, P.; Kim, D.; Arbelles, A.; Panda, R.; Feris, R.; and Kira, Z. 2023. 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*, 11909–11919.
- Thengane, V.; Khan, S.; Hayat, M.; and Khan, F. S. 2022. CLIP Model is an Efficient Continual Learner. arXiv:2210.03114.
- Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.
- Wang, X.; Chen, T.; Ge, Q.; Xia, H.; Bao, R.; Zheng, R.; Zhang, Q.; Gui, T.; and Huang, X. 2023. Orthogonal Subspace Learning for Language Model Continual Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 10658–10671.
- Wang, Z.; Zhang, Z.; Ebrahimi, S.; Sun, R.; Zhang, H.; Lee, C.-Y.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022a. DualPrompt: Complementary Prompting for Rehearsal-Free Continual Learning. In *European Conference on Computer Vision*, 631–648.
- Wang, Z.; Zhang, Z.; Lee, C.-Y.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022b. Learning to Prompt for Continual Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 139–149.
- Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. SUN Database: Large-Scale Scene Recognition from Abbey to Zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3485–3492.
- Yu, J.; Zhuge, Y.; Zhang, L.; Hu, P.; Wang, D.; Lu, H.; and He, Y. 2024a. 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*, 23219–23230.
- Yu, L.; Tao, Z.; Goswami, D.; Yao, H.; Twardowski, B.; Van de Weijer, J.; and Xu, C. 2024b. Exploiting the Semantic Knowledge of Pre-trained Text-Encoders for Continual Learning. *arXiv preprint arXiv:2408.01076*.
- Yu, Y.; Ko, S.; Liu, H.; Dong, Y.; Wu, X.; and Zhu, Q. 2025. Language Guided Concept Bottleneck Models for Interpretable Continual Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14976–14986.
- Zhang, G.; Wang, L.; Kang, G.; Chen, L.; and Wei, Y. 2023. SLCA: Slow Learner with Classifier Alignment for Continual Learning on a Pre-Trained Model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19148–19158.
- Zhang, W.; Janson, P.; Aljundi, R.; and Elhoseiny, M. 2024. Overcoming Generic Knowledge Loss with Selective Parameter Update. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24046–24056.
- Zhao, C.; Wang, K.; Zeng, X.; Zhao, R.; and Chan, A. B. 2024. Gradient-Based Visual Explanation for Transformer-Based CLIP. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 61072–61091.
- Zheng, Z.; Ma, M.; Wang, K.; Qin, Z.; Yue, X.; and You, Y. 2023. Preventing Zero-Shot Transfer Degradation in Continual Learning of Vision-Language Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 19125–19136.
- Zhou, D.-W.; Cai, Z.-W.; Ye, H.-J.; Zhan, D.-C.; and Liu, Z. 2025a. Revisiting Class-Incremental Learning with Pre-Trained Models: Generalizability and Adaptivity Are All You Need. *International Journal of Computer Vision*, 133(3): 1012–1032.
- Zhou, D.-W.; Li, K.-W.; Ning, J.; Ye, H.-J.; Zhang, L.; and Zhan, D.-C. 2025b. External Knowledge Injection for CLIP-Based Class-Incremental Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3314–3325.
- Zhou, D.-W.; Zhang, Y.; Wang, Y.; Ning, J.; Ye, H.-J.; Zhan, D.-C.; and Liu, Z. 2025c. Learning Without Forgetting for Vision-Language Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6): 4489–4504.