Interference-Aware Continual Vision–Language Learning via Instance-Level Expert Routing

Changyuan Zhang[1]*, Tianxiang Xu[2]*, Fei Shen[3], and Canran Xiao[3,4]†

[1] The University of Hong Kong, Hong Kong, China
[2] Peking University, Beijing, China
[3] National University of Singapore, Singapore
[4] Shenzhen Campus of Sun Yat-sen University, Shenzhen, China
*Equal contribution. †Corresponding author: xiaocr3@mail.sysu.edu.cn.

Abstract. Continual vision–language learning is increasingly required in real deployments where data, skills, and objectives drift, yet models still struggle to learn new tasks without erasing old ones and to act reliably in task-agnostic settings with limited memory. We address this gap by turning cross-task conflict into an instance-level signal that drives how the model adapts over time. We propose an interference-aware adapter routing framework that estimates, for each sample, a metric-consistent projection energy to quantify potential interference, routes to the least interfering experts with capacity balancing, enlarges inter-expert separation via principal-angle packing, and grows LoRA rank only along the residual principal direction when current experts are insufficient. Across classification, structured concept matching, generative VQA, and retrieval, our approach improves over strong baselines. Analyses show that the conflict signal reliably predicts downstream forgetting, packing widens principal angles over time, and rank growth is sparse yet beneficial—especially in mid–high layers.

Keywords: Vision–Language Learning · Continual Learning · LoRA

1 Introduction

Deployed vision–language systems face a moving target [55, 56, 60]: categories, styles, and instructions change continuously, while retaining all historical data is often infeasible due to privacy, cost, and latency constraints [23, 26, 37, 58, 71]. This pressure is especially visible in autonomous-driving and embodied settings, where language-grounded perception, affordance-aware scene parsing, and context-aware prediction must operate across shifting scenes and behaviors [3, 40, 61]. Full retraining cannot keep up with the stream and naive fine-tuning overwrites prior skills [19], which harms reliability in settings like evolving product catalogs, seasonal remote sensing, and signage or OCR in new locales [36, 41, 74]. Continual vision–language learning [45, 54] addresses this regime by acquiring new knowledge on the fly, preserving previously learned abilities, and operating without task identifiers or large replay buffers. Broader studies of foundation-model representation potentials further emphasize that multimodal alignment depends on reusable cross-modal structure, making stable adaptation under drift especially important [30].

Despite rapid progress in continual vision–language learning, a key gap remains. Benchmarks reveal strong order sensitivity and instability in VQA and streaming composition settings [11,17,68, 73]. In LVLM-style VQA, reliability is further complicated by object hallucination, where language outputs can overstate objects not grounded in visual evidence [49]. For CLIP-style VLMs, stability

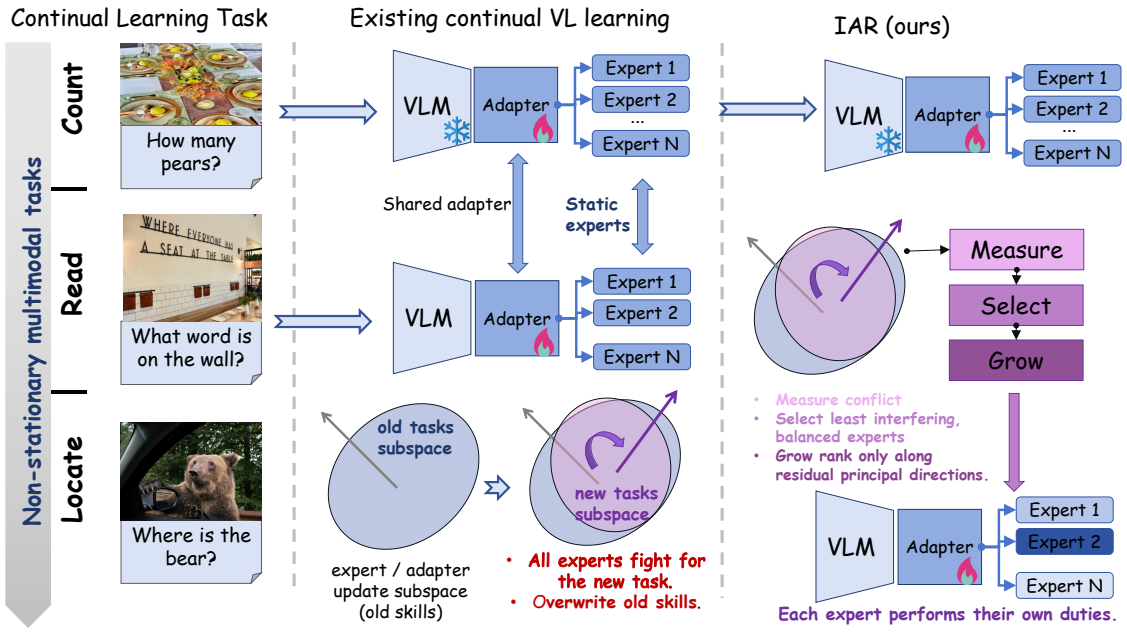

Fig. 1: A continual vision–language model faces a non-stationary stream of multimodal tasks. Existing methods update shared adapters or static experts so all tasks compete in overlapping update subspaces and overwrite old skills, whereas our interference-aware routing (IAR) uses instance-level conflict to measure–select–grow experts, assigning each sample to the least interfering experts and expanding rank only along residual directions.

has been pursued with momentum and geometry preservation [7, 9, 35, 59, 62, 72] and with affinity rectification or task-agnostic evaluation [5, 46, 69]. What is largely absent is a principled, instance-level adaptation rule that explicitly estimates cross-task interference and then uses that signal to drive both expert selection and minimal capacity expansion under drift [6,47]. Consequently, routing is often heuristic or based on coarse distribution cues, and capacity growth is typically pre-scheduled rather than decided per instance.

We study a conflict-aware paradigm for continual VLMs that estimates how a candidate update would interfere with prior competencies and uses this estimate to steer *which experts to use* and *where to add new capacity*, while encouraging experts to remain geometrically well separated. Fig. 1 shows our idea at a high level, our approach turns stability goals into a measure–select–grow loop for instance-level adaptation. Our contributions are as follows:

(**i**) We formulate an instance-level view of interference and show how a conflict signal can drive both expert selection and minimal capacity growth, linking representation geometry with routing decisions.

(**ii**) The framework works in task-agnostic regimes and under limited memory, improving retention without sacrificing adaptation.

(**iii**) Across classification, structured concept matching, generative VQA, and retrieval, our method consistently improves Avg and Last while preserving transfer over the strongest recent baselines.

2 Related Work

Continual Vision–language learning. Early multimodal CL showed the impact of task order and rehearsal on VQA [11] and introduced streaming composition settings [17,24]. For CLIP-style models, methods stabilize representations with momentum and geometry, for example BMU-MoCo [7], ModX [35], CTP [72], ZSCL [62], and large time-continual studies in TiC-CLIP [9]. Retrieval work rectifies teacher affinities [5], while X-TAIL promotes task-agnostic evaluation with adapter fusion [46]. Recent retrieval systems have further explored evidence-guided adaptation, language-controlled quality assessment, and calibration mechanisms for multimodal reasoning and retrieval [21,29]. Adjacent multimodal debiasing work adapts guidance in both textual and visual spaces, highlighting that robustness often requires coordinated treatment of both modalities [50]. These lines largely rely on global regularizers or replay and typically lack a sample-wise routing criterion for conflict-aware adaptation. We introduce a metric-consistent projection energy under Fisher that enables per-sample routing and turns stability objectives into decisions on routing and capacity.

PEFT with prompts, adapters, and routing. Prompt and adapter methods bring lightweight continual adaptation [20, 65, 67, 70]. Recent VLM and MLLM systems add expert routing or residual branches, for example DDAS with MoE-Adapters [48], DIKI with distribution-aware residuals [39], CL-MoE with dual routers [13], and hierarchical decoupling in BranchLoRA and HiDe-LLaVA [12, 52]. Beyond VLM adaptation, routing has also been explored in embodied systems by selecting among heterogeneous off-the-shelf policies according to task representations and execution feedback [4]. These approaches often use heuristic or distributional gates and usually lack an instance-level capacity allocation rule. We differ by using a conflict-aware, sample-wise routing criterion and by growing rank only along the residual principal direction outside the union of experts.

Interference-aware geometry and LoRA. Geometry-driven PEFT encourages orthogonality or preserves topology, for example InfLoRA [25], DMNSP with null-space projection [18], and zero-shot stability or consolidation in ZAF and C-CLIP [8, 27]. Recent multimodal alignment hypotheses also frame alignment as structured representation organization across modalities [31]. Synthetic replay can also be weighted by Fisher information [44]. Robust compositional retrieval methods have also exploited geometric constraints and noise-unlearning objectives to improve robustness under semantic perturbations [22]. Low-rank and information-bottleneck principles have also been used to remove redundancy in efficient perception systems [66]. Causal invariant debiasing provides a complementary view of stability for pretrained models under fine-tuning, but does not decide per-sample expert allocation [64]. Most of these treat geometry as a regularizer and rarely couple it with an instance-level mechanism that links geometry to routing and capacity growth. We measure conflict with Fisher projection energy, enlarge principal angles through subspace packing, and expand LoRA rank only along residual directions.

3 Method: Interference-Aware Adapter Routing (IAR)

We consider continual adaptation of a pre-trained vision–language model under a non-stationary task stream $\{\mathcal{T}_t\}_{t=1}^{T}$, where each task provides samples $(x, y) \sim \mathcal{D}_t$. We *freeze* the backbone parameters and adapt only through a bank of parameter-efficient experts (LoRA/adapters). For each

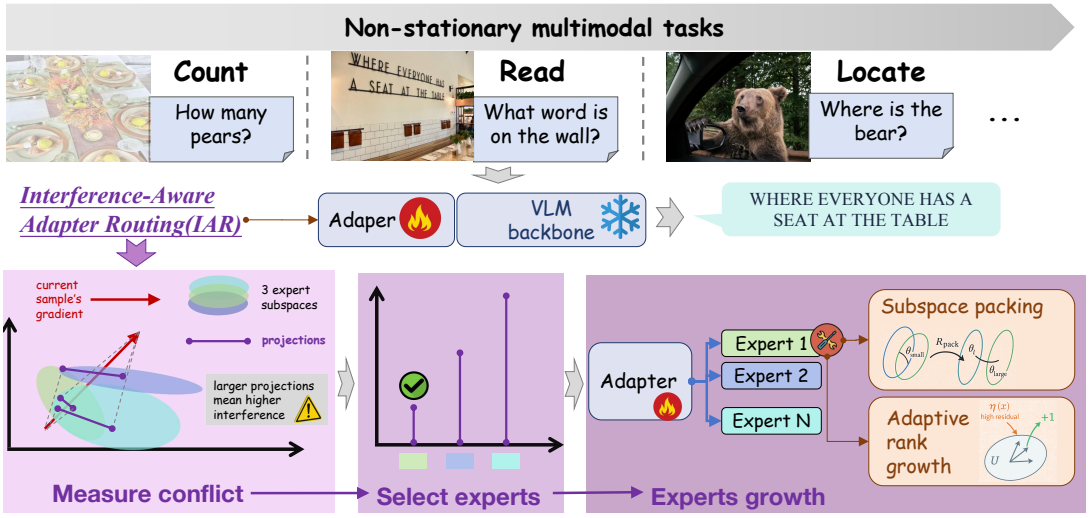

Fig. 2: IAR training pipeline. We freeze the vision–language backbone and train a bank of LoRA/adapter experts. For each sample, a gradient-based conflict estimator computes metric-consistent interference scores for all experts (*measure*), a router selects a small set of least interfering, capacity-balanced experts (*select*), and expert subspaces are regularized and expanded only along residual principal directions when needed (*grow*).

incoming sample, IAR executes a *measure–select–grow* loop: (i) it measures instance-level *interference* by projecting a *probe* update direction onto each expert's historical update subspace under a metric-consistent (Fisher/NTK) whitening; (ii) it routes the sample to a small set of *least-interfering* experts with utilization-aware capacity control; (iii) if the probe direction contains substantial energy outside the union of existing expert subspaces, it grows LoRA rank *minimally* along the leading residual direction. The overall computation graph is illustrated in Fig. 2.

3.1 Preliminaries

Key symbols and operators. We list the operators used throughout: *(i)* $\mathrm{vec}(\cdot)$ stacks a matrix into a column vector (column-major), and $\mathrm{mat}(\cdot)$ is its inverse; *(ii)* $\odot$ is elementwise product; $\|\cdot\|_2$ and $\|\cdot\|_F$ are ℓ_2 and Frobenius norms; *(iii)* $\mathbb{I}\{\cdot\}$ is the indicator (implemented with a straight-through estimator when used for Top-K_r selection); *(iv)* $\mathrm{sg}[\cdot]$ denotes stop-gradient (identity in forward, zero gradient in backward); *(v)* $\mathrm{qf}(\cdot)$ returns the orthonormal factor Q of a QR decomposition; *(vi)* $\mathrm{Top}_{K_r}(s)$ returns the index set of the K_r smallest elements of a score vector s (we use K_r=2).

Layers, experts, and LoRA updates. We index instrumented layers by $\ell \in \mathcal{S}$ and experts by $e \in \{1, \ldots, E\}$. At layer ℓ, the frozen backbone weight is $W_\ell \in \mathbb{R}^{d_\ell^{\mathrm{out}} \times d_\ell^{\mathrm{in}}}$. Expert e contributes a rank-$r_{\ell,e}$ LoRA update $\Delta W_{\ell,e} = A_{\ell,e} B_{\ell,e}$ with $A_{\ell,e} \in \mathbb{R}^{d_\ell^{\mathrm{out}} \times r_{\ell,e}}$ and $B_{\ell,e} \in \mathbb{R}^{r_{\ell,e} \times d_\ell^{\mathrm{in}}}$. Let $a_{\ell,e}^{(k)} \in \mathbb{R}^{d_\ell^{\mathrm{out}}}$ be the k-th column of $A_{\ell,e}$ and $b_{\ell,e}^{(k)} \in \mathbb{R}^{d_\ell^{\mathrm{in}}}$ be the k-th column of $B_{\ell,e}^\top$ so that

$$\Delta W_{\ell,e} = \sum_{k=1}^{r_{\ell,e}} a_{\ell,e}^{(k)} b_{\ell,e}^{(k)\top}. \tag{1}$$

Given an input x, the router outputs sparse mixture weights $\alpha_e(x) \geq 0$ with $\sum_{e=1}^{E} \alpha_e(x) = 1$ and exactly K_r=2 nonzeros. The effective layer weight is

$$W_\ell(x) = W_\ell + \sum_{e=1}^{E} \alpha_e(x)\, \Delta W_{\ell,e}. \tag{2}$$

Loss. We denote the per-sample training loss by $\mathcal{L}(f(x), y)$. For discriminative classification/retrieval, $\mathcal{L}$ is cross-entropy on supervised logits; for generative VQA, $\mathcal{L}$ is next-token negative log-likelihood.

Expert update subspace. We represent an expert's historical update directions at layer ℓ by vectorizing each rank-1 LoRA component:

$$U_{\ell,e} = \Big[\mathrm{vec}\big(a_{\ell,e}^{(1)} b_{\ell,e}^{(1)\top}\big), \ldots, \mathrm{vec}\big(a_{\ell,e}^{(r_{\ell,e})} b_{\ell,e}^{(r_{\ell,e})\top}\big)\Big] \in \mathbb{R}^{(d_\ell^{\mathrm{out}} d_\ell^{\mathrm{in}}) \times r_{\ell,e}}. \tag{3}$$

Initialization and capacity caps. At t=1, we instantiate E experts with initial ranks $r_{\ell,e} = r_{\mathrm{init}}$ for all $\ell \in \mathcal{S}$. LoRA factors are initialized so that $\Delta W_{\ell,e} \approx 0$ (thus the initial model equals the frozen backbone), and a hard cap $r_{\ell,e} \leq r_{\max}$ prevents unbounded growth over long streams.

3.2 Probe Gradient and Metric Whitening

Probe gradient (backbone-only). To avoid circular dependence between routing decisions and the probing signal, we define the probe direction using a *backbone-only* forward (no experts applied):

$$g_\ell(x, y) \;:=\; \nabla_{\mathrm{vec}(W_\ell)}\, \mathcal{L}\big(f_{\boldsymbol{\theta}}(x), y\big) \in \mathbb{R}^{d_\ell^{\mathrm{out}} d_\ell^{\mathrm{in}}}, \tag{4}$$

where $\boldsymbol{\theta}$ denotes frozen backbone parameters. This probe gradient is *only* used to compute routing/growth signals; the backbone is never updated.

Diagonal Fisher metric and whitening. Given a mini-batch $\mathcal{B}_t = \{(x_i, y_i)\}_{i=1}^{N}$ of size N, we maintain a diagonal Fisher (or diagonal NTK [51]) metric $G_\ell^{(t)} = \mathrm{diag}(\mathbf{f}_\ell^{(t)})$ via EMA:

$$\mathbf{f}_\ell^{(t)} = \lambda_F \mathbf{f}_\ell^{(t-1)} + (1 - \lambda_F) \cdot \frac{1}{N} \sum_{i=1}^{N} \big(g_\ell(x_i, y_i) \odot g_\ell(x_i, y_i)\big), \qquad L_\ell^{(t)} = \mathrm{diag}\big(\sqrt{\mathbf{f}_\ell^{(t)} + \epsilon}\big), \tag{5}$$

where $\epsilon > 0$ stabilizes whitening. Since L_ℓ is diagonal, $L_\ell^\top L_\ell = \mathrm{diag}(\mathbf{f}_\ell + \epsilon)$. We work in whitened coordinates $\tilde{g}_\ell = L_\ell g_\ell$.

3.3 Metric-Consistent Interference Energy

A core design goal is to make "interference" a *metric-consistent* quantity rather than a heuristic gate. Under small perturbations δw, the local change of predictive behavior is commonly approximated by the Fisher quadratic form $\delta w^\top G\, \delta w = \|L\, \delta w\|_2^2$ (up to constants). Motivated by this, we measure how much of the whitened probe direction $\tilde{g}_\ell$ lies in each expert's historical update subspace: if a candidate update direction substantially overlaps an expert's subspace in the G-metric, it is more likely to reuse/overwrite directions that previously shaped that expert.

Stabilized energy basis. We maintain an EMA copy $\bar{U}_{\ell,e}$ of $U_{\ell,e}$ for stable interference estimation: $\bar{U}_{\ell,e} \leftarrow \lambda_U \bar{U}_{\ell,e} + (1 - \lambda_U) U_{\ell,e}$. Define the whitened (energy) basis $\tilde{U}_{\ell,e}^{E} = L_\ell\, \mathrm{sg}[\bar{U}_{\ell,e}]$.

$$\mathcal{I}_e(x,y) = \sum_{\ell \in \mathcal{S}} \tilde{g}_\ell(x,y)^\top \tilde{U}^E_{\ell,e} \Big(\tilde{U}^{E\top}_{\ell,e} \tilde{U}^E_{\ell,e} + \varepsilon_p I \Big)^{-1} \tilde{U}^{E\top}_{\ell,e} \tilde{g}_\ell(x,y), \tag{6}$$

with ridge $\varepsilon_p > 0$. This is the squared norm of the (regularized) projection of $\tilde{g}_\ell$ onto $\mathrm{span}(\tilde{U}^E_{\ell,e})$; lower $\mathcal{I}_e(x,y)$ indicates weaker overlap and thus less estimated interference. An efficient computation that avoids forming a $D_\ell \times D_\ell$ projector is given in the supplementary material.

3.4 Min-Interference Routing with Utilization-Aware Capacity Control

We route each sample to the least-interfering experts while discouraging expert collapse via a utilization penalty. Let u_e be the EMA utilization of expert e, updated from routing weights as detailed in the supplementary material, and $\bar{u} = \frac{1}{E}\sum_e u_e$. Define $\rho_e = (u_e/\bar{u})^{\gamma_c}$ with capacity exponent $\gamma_c > 0$. For each sample, we form a routing score and Top-K_r soft gates:

$$s_e(x,y) = \mathcal{I}_e(x,y) + \lambda_C \rho_e, \qquad \alpha_e(x,y) = \frac{\exp\big(-\beta\, s_e(x,y)\big)\, \mathbb{I}\{e \in \mathrm{Top}_{K_r}(s(x,y))\}}{\sum_{j \in \mathrm{Top}_{K_r}(s(x,y))} \exp\big(-\beta\, s_j(x,y)\big)}, \tag{7}$$

where $\beta > 0$ is the temperature and K_r=2 by default. The hard Top_{K_r} selection is treated as a straight-through stop-gradient: gradients flow to the selected experts, while the discrete selection itself is not differentiated.

3.5 Subspace Packing Regularization

If expert subspaces align, routing becomes ambiguous and interference increases over time. We increase separability by maximizing principal angles via a cross-coherence penalty. To ensure gradients reach trainable expert parameters, packing uses the *current* basis $U_{\ell,e}$ and stops gradients only through the whitening factor: $\tilde{U}^P_{\ell,e} = \mathrm{sg}[L_\ell]\, U_{\ell,e}$. Let $\widehat{U}_{\ell,e} = \mathrm{qf}(\tilde{U}^P_{\ell,e})$ be the orthonormal factor from QR. We minimize

$$\mathcal{R}_{\mathrm{pack}} = \sum_{\ell \in \mathcal{S}} \sum_{1 \le i < j \le E} \big\| \widehat{U}^\top_{\ell,i} \widehat{U}_{\ell,j} \big\|_F^2, \tag{8}$$

which pushes the principal angles between expert subspaces toward $\pi/2$ and stabilizes future routing.

3.6 Residual-Guided Rank Growth

When the incoming probe direction lies outside the union of existing expert subspaces, forcing it into current subspaces either underfits new information or increases interference. We grow rank minimally along the leading residual direction.

Let $\tilde{U}^E_{\ell,\cup} = [\tilde{U}^E_{\ell,1}, \ldots, \tilde{U}^E_{\ell,E}]$ and define the union projector $\tilde{P}_{\ell,\cup} = \tilde{U}^E_{\ell,\cup}(\tilde{U}^{E\top}_{\ell,\cup}\tilde{U}^E_{\ell,\cup} + \varepsilon_p I)^{-1}\tilde{U}^{E\top}_{\ell,\cup}$. The whitened residual and residual ratio are

$$\tilde{r}_\ell = (I - \tilde{P}_{\ell,\cup})\, \tilde{g}_\ell, \qquad \eta(x,y) = \frac{\sum_{\ell \in \mathcal{S}} \|\tilde{r}_\ell\|_2^2}{\sum_{\ell \in \mathcal{S}} \|\tilde{g}_\ell\|_2^2}, \qquad \ell^\star = \arg\max_{\ell \in \mathcal{S}} \|\tilde{r}_\ell\|_2^2. \tag{9}$$

After the routed expert update, we compute these residual statistics over the current batch (or phase boundary) and queue candidates with $\eta(x,y) > \tau$, where $\tau \in (0,1)$. Queued growth is applied only

Algorithm 1 IAR training step

Require: Mini-batch $\mathcal{B}_t = \{(x_i, y_i)\}_{i=1}^N$; frozen backbone $\{W_\ell\}$; LoRA experts $\{A_{\ell,e}, B_{\ell,e}\}$; EMA Fisher stats $\{\mathbf{f}_\ell\}$; EMA bases $\{\bar{U}_{\ell,e}\}$; utilization $\{u_e\}$; EMA routing $\{\bar{\alpha}_e\}$

Probe: for each (x_i, y_i) compute $g_\ell(x_i, y_i)$ by Eq. (4) using backbone-only forward
Update $\mathbf{f}_\ell$ and L_ℓ by Eq. (5); set $\tilde{g}_\ell = L_\ell g_\ell$
Update $\bar{U}_{\ell,e} \leftarrow \lambda_U \bar{U}_{\ell,e} + (1-\lambda_U) U_{\ell,e}$; set $\tilde{U}^E_{\ell,e} = L_\ell \,\mathrm{sg}[\bar{U}_{\ell,e}]$
for each sample (x_i, y_i) **do**
    Compute $\mathcal{I}_e(x_i, y_i)$ by Eq. (6) (see the supplementary material for the efficient form)
    Compute gates $\alpha(x_i, y_i)$ by Eq. (7)
end for
Update experts: run gated forward/backward with α and update routed experts to minimize $\mathcal{J}$ in Eq. (11)
Residual statistics: compute $\{\eta_i, \ell_i^\star\}_{i=1}^N$ by Eq. (9); queue candidates with $\eta_i > \tau$
Grow: after the batch/phase, apply queued rank growth by Eq. (10) subject to rank caps
Update utilization u_e, routing EMA $\bar{\alpha}_e$, and $\mathcal{R}_{\text{load}}$ as detailed in the supplementary material

at the batch/phase boundary; after deduplicating selected expert–layer pairs, we expand rank by 1 at layer $\ell^\star$ for $e^\star = \arg\min_e s_e(x, y)$. To obtain a valid matrix-factor direction, we first unwhiten the vectorized residual and only then reshape it. Here $L_{\ell^\star}^{-1}$ acts on the $D_{\ell^\star} = d_{\ell^\star}^{\text{out}} d_{\ell^\star}^{\text{in}}$ dimensional vectorized weight space. We append a new LoRA direction

$$\begin{aligned} r_{\ell^\star} &= L_{\ell^\star}^{-1} \tilde{r}_{\ell^\star}, \qquad R_{\ell^\star} = \mathrm{mat}(r_{\ell^\star}), \\ (a, b) &= \mathrm{topSVD}_1(R_{\ell^\star}), \\ A_{\ell^\star,e^\star} &\leftarrow [A_{\ell^\star,e^\star} \;\; c_a\, a], \qquad B_{\ell^\star,e^\star} \leftarrow \begin{bmatrix} B_{\ell^\star,e^\star} \\ c_b\, b^\top \end{bmatrix}, \qquad c_a c_b = \kappa \,\|\tilde{r}_{\ell^\star}\|_2, \end{aligned} \tag{10}$$

where $\kappa \in (0, 1]$ controls injected energy and the cap $r_{\ell,e} \le r_{\max}$ is enforced by skipping growth once an expert reaches $r_{\max}$.

3.7 Inference-Time Routing and Optimization

Train-time vs. test-time routing. During training, the gates $\alpha(x, y)$ are computed from the gradient-based interference energies (Eq. (6)–(7)) and serve as an oracle supervision signal. At inference, gradients are unavailable, so we distill a lightweight router r_ψ that predicts $\hat{\alpha}(x)$ from frozen-backbone forward features $\phi(x)$: $\hat{\alpha}(x) = r_\psi(\phi(x))$. The router is trained to match oracle gates by KL distillation, as detailed in the supplementary material, and uses the same Top-K_r sparsification as Eq. (7). This removes gradient dependence at deployment; router fidelity and inference overhead are reported in the supplementary material. Algo. 2 summarizes the deployment procedure.

Objective and training step. We optimize routed experts with

$$\mathcal{J} = \mathbb{E}_{(x,y)}\big[\mathcal{L}(f_{\boldsymbol{\theta},\alpha}(x), y)\big] + \mu\, \mathcal{R}_{\text{pack}} + \nu\, \mathcal{R}_{\text{load}}, \tag{11}$$

where $\mathcal{R}_{\text{load}}$ smooths the long-horizon routing distribution, as detailed in the supplementary material. Each iteration executes: (1) compute probe gradients g_ℓ (Eq. (4)) and update $(\mathbf{f}_\ell, L_\ell)$ (Eq. (5)); (2) compute energies $\mathcal{I}_e$ (Eq. (6)) and gates α (Eq. (7)); (3) update routed experts by backpropagating $\mathcal{J}$; (4) compute residual statistics and apply queued rank growth after the batch/phase (Eq. (9)–(10)). Pseudo-code is provided in Algo. 1.

Algorithm 2 IAR inference

Require: Test input x; frozen backbone $\{W_\ell\}$; trained LoRA experts $\{A_{\ell,e}, B_{\ell,e}\}$; distilled router r_ψ; active expert number K_r
Ensure: Prediction $\hat{y}$
Feature extraction: compute frozen-backbone feature $\phi(x)$ by a forward pass
Router prediction: compute dense expert probabilities $p(x) = \mathrm{softmax}(r_\psi(\phi(x))) \in \mathbb{R}^E$
Convert probabilities into routing scores $\hat{s}_e(x) = -\log(p_e(x) + \epsilon_r)$ for $e = 1, \ldots, E$
Select active experts by the same Top-K_r rule: $\mathcal{E}_x = \mathrm{Top}_{K_r}(\hat{s}(x))$
Sparsify and renormalize router weights $\hat{\alpha}_e(x) = p_e(x)\mathbb{I}\{e \in \mathcal{E}_x\} / \sum_{j \in \mathcal{E}_x} p_j(x)$
for each instrumented layer $\ell \in \mathcal{S}$ **do**
    Compose the routed LoRA update $\Delta W_\ell(x) = \sum_{e \in \mathcal{E}_x} \hat{\alpha}_e(x) A_{\ell,e} B_{\ell,e}$
    Use the input-conditioned weight $W_\ell(x) = W_\ell + \Delta W_\ell(x)$
end for
Run the routed model $f_{\boldsymbol{\theta},\hat{\alpha}}(x)$ with the composed adapter weights
if classification or concept matching **then**
    $\hat{y} = \arg\max_c f_{\boldsymbol{\theta},\hat{\alpha}}(x)_c$
else if retrieval **then**
    $\hat{y} = f_{\boldsymbol{\theta},\hat{\alpha}}(x)$ *used as the image/text embedding*
else if generative VQA **then**
    $\hat{y} = \mathrm{Decode}\big(f_{\boldsymbol{\theta},\hat{\alpha}}, x\big)$
end if
return $\hat{y}$

4 Experimental Results

4.1 Experimental Setup

Datasets **(i) Multi-domain classification (MTIL and X-TAIL).** We evaluate on the *Multi-domain Task-Incremental Learning* (MTIL) protocol introduced for VLMs and the more realistic *Cross-domain Task-Agnostic Incremental Learning* (X-TAIL). For MTIL, we follow [62] with 11 domain datasets. For X-TAIL, we follow [46] using 10 domains. We adopt the commonly used 16-shot per domain in cross-domain classification. **(ii) Structured VL concepts (ConStruct-VL).** To measure concept- and relation-level robustness in data-free continual settings, we include ConStruct-VL [38], which streams skills such as color, material, size, spatial relations, action relations, state. Each phase is a binary image–text consistency task (VL matching). **(iii) Generative VQA continual learning (VQACL).** We adopt VQACL [57] on VQA v2 (image QA) and NExT-QA (video QA). VQACL uses a dual-level sequence: an outer loop over reasoning skills (e.g., *Count, Color, Location*, $T = 10$ for VQA v2; $T = 8$ for NExT-QA) and an inner loop over visual sub-tasks obtained by partitioning COCO classes into $K_v = 5$ groups. Following [57], we also report novel composition performance. **(iv) Continual vision–language retrieval (CVLR).** For cross-modal retrieval under continual domain shifts, we use the five-domain sequence from [5]: MS-COCO, Flickr30K, IAPR TC-12, ECommerce-T2I, RSICD. Each step adds a new retrieval domain (image→text and text→image).

Evaluation Metrics For evaluation, we use the following metrics: (i) For **Cross-domain classification (MTIL/X-TAIL)**, following [46, 62], we report **Avg**, **Last**, and **Transfer**; (ii) For **Structured VL concepts**, we report per-task **accuracy** (image–text consistency) and the **macro**

	MTIL			X-TAIL			ConStruct
Method	A	L	T	A	L	T	Acc
LwF [19]	64.3	62.2	70.8	61.5	58.4	74.2	68.7
ZSCL [62]	69.8	66.5	74.2	66.0	63.1	77.0	73.7
Mod-X [35]	68.9	65.0	72.8	65.3	62.0	75.4	74.1
C-CLIP [27]	71.1	68.0	75.0	68.1	66.2	80.0	75.8
ZAF [8]	70.3	67.2	76.1	67.6	65.7	80.4	74.6
DDAS [48]	71.0	68.5	74.0	67.2	65.0	78.6	75.0
RAIL [46]	70.9	68.8	77.4	67.4	65.9	79.2	74.9
DIKI [39]	71.5	69.2	76.0	68.3	66.4	79.6	76.1
GIFT [44]	71.6	69.4	77.2	68.6	66.7	80.7	76.5
LADA [32]	71.9	69.6	76.9	69.0	67.1	80.9	76.6
MG-CLIP [15]	71.2	69.1	76.6	68.2	66.5	80.1	76.0
ENGINE [63]	71.4	68.7	76.7	68.4	66.5	80.3	76.3
ProxyFDA [14]	70.8	67.9	75.9	67.9	66.0	79.9	75.7
DNS [28]	70.7	67.6	75.4	67.5	65.8	79.5	75.2
Ours	**73.0**	**70.7**	**78.1**	**69.8**	**68.0**	**81.2**	**77.3**

Table 1: Classification results. MTIL (11 domains) and X-TAIL (10 domains) report **A** (Avg), **L** (Last), and **T** (Transfer) under 16-shot; ConStruct-VL reports macro accuracy (**Acc**, %). All methods use frozen backbones and matched trainable-parameter budgets (0.8–1.2%); replay-based methods use a common memory budget.

average across phases [38]; (iii) For **VQACL**, we report **AP** and **Forget** as defined in [57], where VQA v2 uses standard VQA accuracy and NExT-QA uses WUPS for generative answers; (iv) For **CVLR**, we report cross-modal retrieval Recall@K (K$\in \{1, 5, 10\}$) for both image→text and text→image retrieval [5], along with **Avg** and **Last** over steps for continual summaries. We additionally report: **utilization imbalance** std($\bar{\alpha}_e$) (lower is better), **added rank** per layer, and **interference–forgetting correlation** $\rho(\mathcal{I}_e, \Delta\text{acc})$.

Implementation For cross-domain classification/retrieval we use CLIP ViT-B/16 by default (ViT-L/14 in a sensitivity study). For VQACL-style generative vision-language tasks, we use LLaVA-1.5-7B with the official vision tower and projector. We use AdamW with learning rate 2×10^{-4} for adapters, router, and MLP, weight decay 0.01, cosine schedule, and no warmup (per [9]). Batch size 64 (classification/retrieval) and 32 (VQACL). Refer to the supplementary material for more details of the experimental setup.

4.2 Main results

As shown in Tab. 1 and Tab.2, IAR raises both accuracy and stability. On **MTIL**, it improves over the best prior by about +1.1 Avg and +1.1 Last; on **X-TAIL** (task-agnostic) it adds a further +0.9 on Last. On **ConStruct-VL**, IAR attains the highest macro accuracy (+0.7 over the strongest competitor). For **VQACL**, it delivers higher accuracy with less forgetting. On VQA v2, compared with the strongest AP baseline CL-MoE, IAR improves AP by +1.1 and reduces forgetting by 0.6; compared with the lowest-forgetting baseline SMoLoRA, it further reduces forgetting by 0.2. On **CVLR**, IAR achieves the top R@K and improves Last R@1 by roughly +1.5. These consistent

Method	VQA v2		NExT-QA		CVLR			
	AP	F↓	AP	F↓	R1	R5	R10	LR1
Vanilla	21.3	14.4	18.5	8.7	19.2	26.6	35.1	17.8
ER [2]	47.8	6.2	31.0	3.4	29.6	57.4	66.8	28.5
DER [1]	48.6	5.9	31.8	3.1	30.3	58.0	67.2	29.1
VS [42]	49.0	5.5	32.2	3.0	30.5	58.3	67.5	29.3
InfLoRA [25]	50.4	4.5	33.4	2.7	30.9	58.8	67.9	30.1
QUAD [34]	49.6	3.8	31.7	2.9	30.6	58.4	67.7	29.4
CL-MoE [13]	<u>51.3</u>	2.9	<u>33.5</u>	2.6	31.4	59.4	68.6	30.3
Adapt-∞ [33]	50.8	2.9	32.9	2.7	–	–	–	–
D-MoLE [10]	51.0	3.0	33.2	2.8	–	–	–	–
SMoLoRA [43]	50.9	<u>2.5</u>	33.4	<u>2.5</u>	–	–	–	–
BCP-MFA [53]	49.9	3.1	33.3	3.0	–	–	–	–
DKR [5]	–	–	–	–	31.9	59.8	69.4	30.7
C-CLIP [27]	50.7	3.4	33.0	2.7	<u>32.5</u>	<u>61.2</u>	<u>70.8</u>	<u>31.1</u>
DNS [28]	–	–	–	–	31.8	60.2	70.2	30.9
Ours	**52.4**	**2.3**	**34.2**	**2.4**	**33.7**	**62.8**	**71.9**	**32.6**

Table 2: VQA and retrieval results. VQACL reports **AP** (%) and **F** (Forgetting, ↓); CVLR (5 domains) reports **R1/R5/R10** (Recall@1/5/10, %) and **LR1** (Last R@1). All methods use frozen backbones and matched trainable-parameter budgets (0.8–1.2%); replay-based methods use a common memory budget; methods marked with dashes do not apply to CVLR.

gains across heterogeneous regimes validate the generality and efficiency of our design under matched budgets.

Variant	Classification / Concepts			VQACL			CVLR	
	MTIL-A↑	X-TAIL-L↑	ConStruct Acc↑	VQAv2 AP↑	VQAv2 Forget↓	NExT-QA AP↑	R@1↑	Last R@1↑
Full IAR	**73.0**	**68.0**	**77.3**	**52.4**	**2.3**	**34.2**	**33.7**	**32.6**
❶	72.3 (-0.7)	67.4 (-0.6)	76.7 (-0.6)	51.8 (-0.6)	2.6 (+0.3)	33.7 (-0.5)	33.1 (-0.6)	32.0 (-0.6)
❷	71.5 (-1.5)	66.6 (-1.4)	75.8 (-1.5)	51.2 (-1.2)	3.0 (+0.7)	33.0 (-1.2)	32.3 (-1.4)	31.1 (-1.5)
❸	72.6 (-0.4)	67.3 (-0.7)	76.5 (-0.8)	52.0 (-0.4)	2.5 (+0.2)	33.9 (-0.3)	33.3 (-0.4)	32.0 (-0.6)
❹	72.6 (-0.4)	67.7 (-0.3)	76.9 (-0.4)	51.8 (-0.6)	2.6 (+0.3)	33.7 (-0.5)	33.4 (-0.3)	32.3 (-0.3)
❺	72.3 (-0.7)	67.4 (-0.6)	76.6 (-0.7)	51.4 (-1.0)	2.7 (+0.4)	33.4 (-0.8)	32.9 (-0.8)	31.9 (-0.7)
❻	72.5 (-0.5)	67.1 (-0.9)	76.8 (-0.5)	52.0 (-0.4)	2.6 (+0.3)	33.8 (-0.4)	33.0 (-0.7)	31.8 (-0.8)
❼	72.7 (-0.3)	67.5 (-0.5)	76.9 (-0.4)	52.1 (-0.3)	2.4 (+0.1)	33.9 (-0.3)	33.4 (-0.3)	32.3 (-0.3)
❽	72.4 (-0.6)	67.2 (-0.8)	76.7 (-0.6)	51.9 (-0.5)	2.6 (+0.3)	33.6 (-0.6)	33.1 (-0.6)	32.0 (-0.6)

Table 3: Single-factor ablations of IAR (higher is better unless ↓).

4.3 Ablation and Analysis

Ablation results We perform single–factor ablations under the same setup as Section 4.1. Variants: ❶ **w/o Fisher whitening** (Euclidean metric for energies); ❷ **w/o router** (uniform adapter mixing); ❸ **w/o subspace packing** (μ=0); ❹ **fixed rank** r=8 (no rank growth); ❺ **no rank growth** (fixed r=4); ❻ **w/o capacity terms** (λ=ν=0); ❼ **Top-**K_r=1 (hard gating); ❽ **w/o EMA basis** (use instantaneous U for routing/packing).

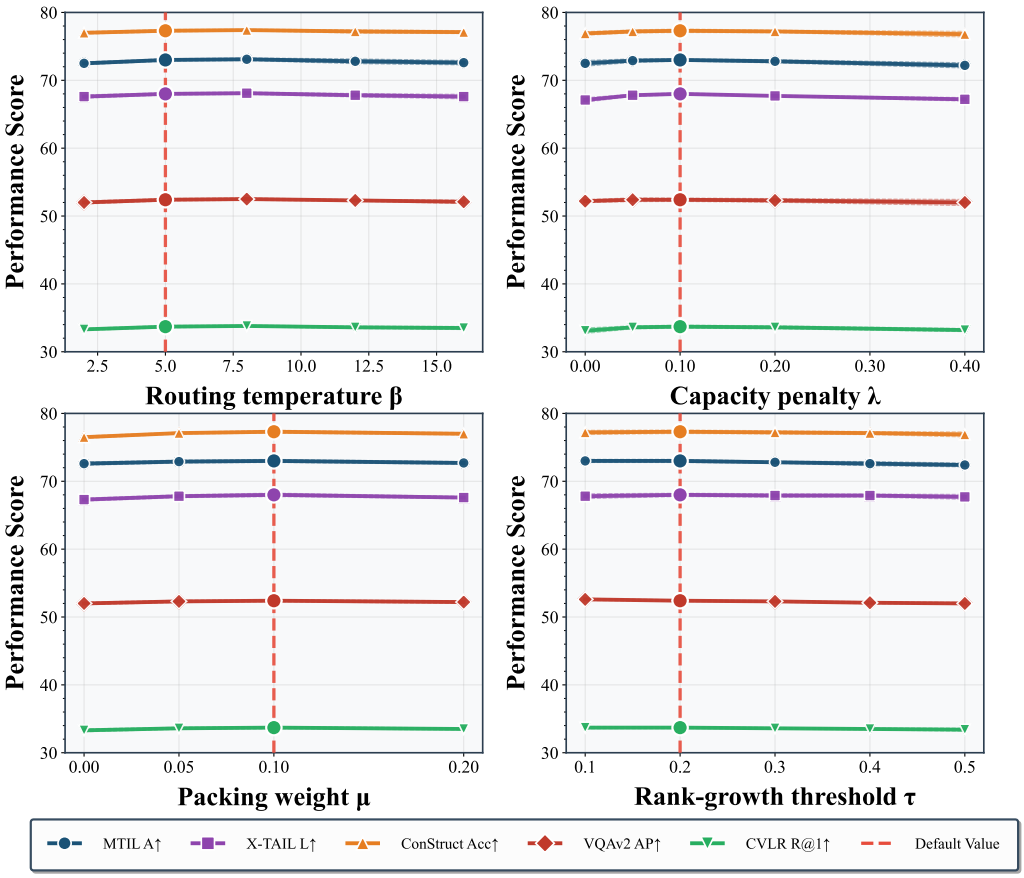

Fig. 3: Sensitivity results (mean±std over 3 seeds). **IAR is stable across wide ranges.**

From Tab. 3, the largest degradation comes from ❷ *w/o router*, confirming the necessity of interference-aware selection, forgetting rises notably on VQA (+0.7). ❶ shows that *metric-consistent* (Fisher/NTK) whitening matters across all regimes, especially for CVLR stability (R@1/LR1 drops 0.6/0.6). ❸ validates *subspace packing*: removing it primarily hurts last-step accuracy (X-TAIL L: -0.7) and concept separation (ConStruct: -0.8). Capacity control ❻ is key for long-horizon balance (X-TAIL L and LR1 drop -0.9/-0.8). Finally, rank mechanisms matter: disabling growth (❺) or fixing rank (❹) reduces adaptation (e.g., VQAv2 AP: -1.0/-0.6), supporting our *residual principal-direction* expansion. Top-K_r=2 soft mixtures perform slightly better than hard Top-1 ❼, and the EMA basis ❽ stabilizes routing/packing under drift.

Parameter Sensitivity We study key hyperparameters in IAR. As shown in Fig. 3, it can be observed that: **(i) Routing temperature** β**.** Performance exhibits a shallow U-shape: too small (soft mixing) or too large (over-peaky) slightly hurts *Last*/R@1. $\beta \in [5, 8]$ is consistently strong, reflecting a good balance between expert diversity and selectivity. **(ii) Capacity penalty** λ**.** λ=0 degrades *Last*, while overly large λ over-regularizes routing. **(iii) Packing weight** μ**.** μ=0 reduces

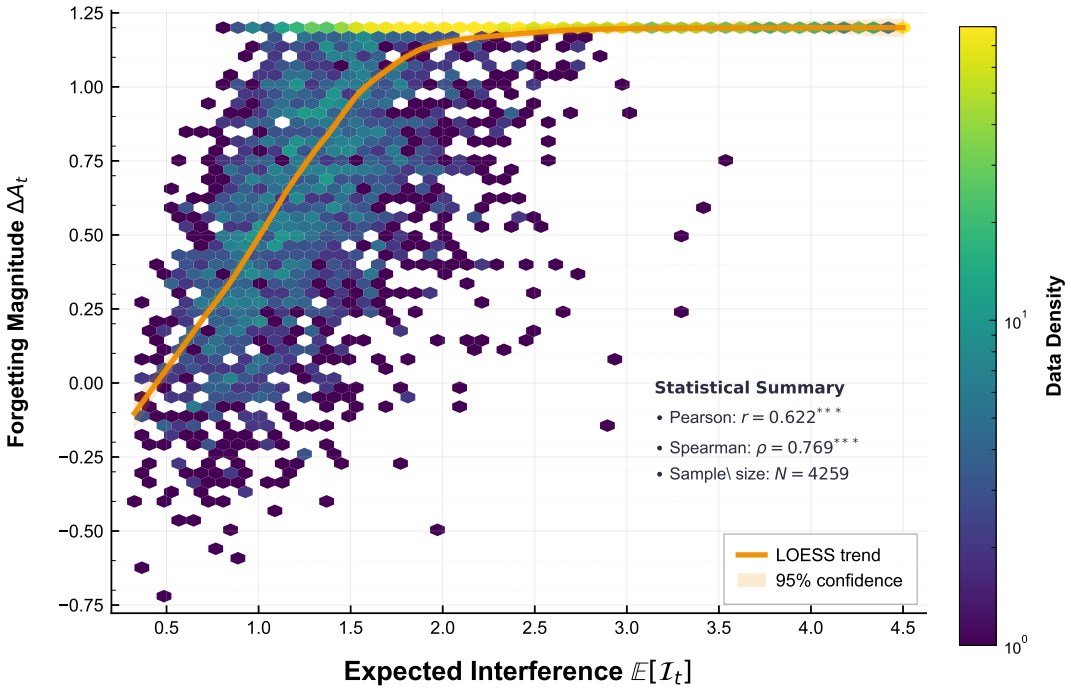

Fig. 4: Interference energy vs. forgetting. Hexbin density of training-time $\mathbb{E}[\mathcal{I}]$ (x-axis) against subsequent old-task Δacc (y-axis; negative is better). The solid curve is LOESS (locally weighted) fit; shaded band is pointwise 95% CI. Inset shows Pearson/Spearman correlations (both significant). A clear positive association and larger variance in the high-energy regime indicate that $\mathbb{E}[\mathcal{I}]$ is a meaningful, discriminative routing signal.

concept separation (lower ConStruct and *Last*). Overly large μ mildly constrains plasticity. A small positive weight (0.05–0.10) best preserves old skills with minimal adaptation cost, aligning with our principal-angle regularization. **(iv) Rank-growth threshold τ.** Lower τ triggers more growth (slightly better AP), whereas higher τ delays capacity expansion (marginally lower AP and *Avg*). The default τ=0.2 balances adaptation (AP) and stability (*Last*), matching our design to grow only when residual energy is significant. **Overall, IAR is robust: variations across reasonable ranges stay within** $\pm$0.3–0.5 **on all metrics**.

Does interference energy predict forgetting? We test whether the interference energy $\mathbb{E}[\mathcal{I}]$ measured during training is a reliable *risk scale* for subsequent old-task degradation. As shown in Fig. 4, higher $\mathbb{E}[\mathcal{I}]$ aligns with larger (worse) Δacc. The LOESS trend is monotonically increasing, and both Pearson/Spearman are significant. Notably, variance expands in the high-energy region—precisely where cautious routing (and, if needed, rank growth) is most beneficial—supporting our use of interference energy as a sample-wise decision metric.

Do principal angles between experts grow over time? We verify that subspace packing increases geometric separation: as training proceeds, the pairwise principal angle $\theta^{(1)}$ between expert subspaces should grow (especially the minimum angle among expert pairs). As shown in Fig. 5, principal angles

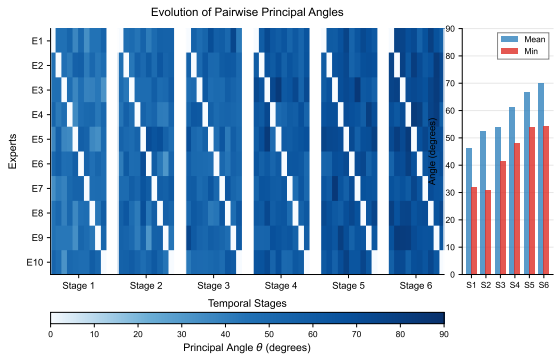

Fig. 5: Principal angles over time. Heatmap shows, for all stages left→right, the pairwise principal angles $\theta_{i,j}^{(1)}$ (degrees) between E experts. The right inset bars summarize per-stage mean and minimum principal angles. We use an E=10 expert bank in this diagnostic experiment.

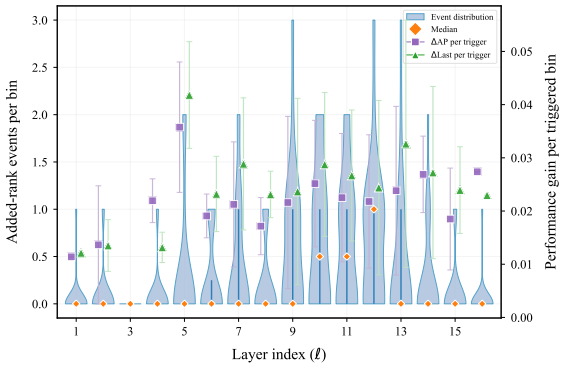

Fig. 6: Residual-triggered rank growth: where and how it helps. Per-layer violin plots of added-rank events per bin (left y-axis), showing that rank growth is sparse overall yet concentrated in *mid–high layers*.

systematically increase across stages; the minimum inter-expert angle accelerates most, indicating that packing primarily pushes apart the most confusable expert pairs, thus stabilizing routing and reducing future interference.

Is residual-triggered rank growth applied only when necessary? We further examine whether residual-guided rank growth behaves as a selective capacity mechanism rather than a simple parameter-increasing heuristic. Ideally, rank growth should satisfy two conditions: first, it should be triggered sparsely and mainly at layers where the incoming update cannot be well represented by existing expert subspaces; second, when triggered, it should bring measurable gains instead of merely adding redundant capacity. To test this, we analyze the layer-wise distribution of growth events and the corresponding per-trigger changes in AP and Last performance. As shown in Fig. 6, rank growth is not uniformly activated across the network. The number of added-rank events remains limited overall and is concentrated in mid–high layers, where vision–language representations are more task-specific and where residual directions are more likely to encode new semantic or reasoning patterns. Meanwhile, the associated ΔAP and ΔLast are consistently positive across layers, with slightly larger gains in the layers where growth is more frequently triggered. This indicates that the residual criterion does not simply expand LoRA capacity everywhere; instead, it identifies directions that are poorly covered by existing experts and injects new rank only when such directions are useful for adaptation. These results support our design choice: IAR adds capacity only when the residual energy is sufficiently strong and primarily at the layers where additional capacity contributes most to continual learning performance.

Case study analysis. As shown in Fig. 7, the left panel contains clear rain cues—visible streaks and a wet ground—so both C-CLIP [27] and IAR correctly answer *Rainy*. This relatively easy case mainly requires detecting local weather-related visual evidence. By contrast, the right panel requires consequence-driven reasoning: despite a merely cloudy sky, the overtopped river, debris, and washed–out road indicate a *Flood*. C-CLIP overweights sky appearance and predicts *Cloudy*,

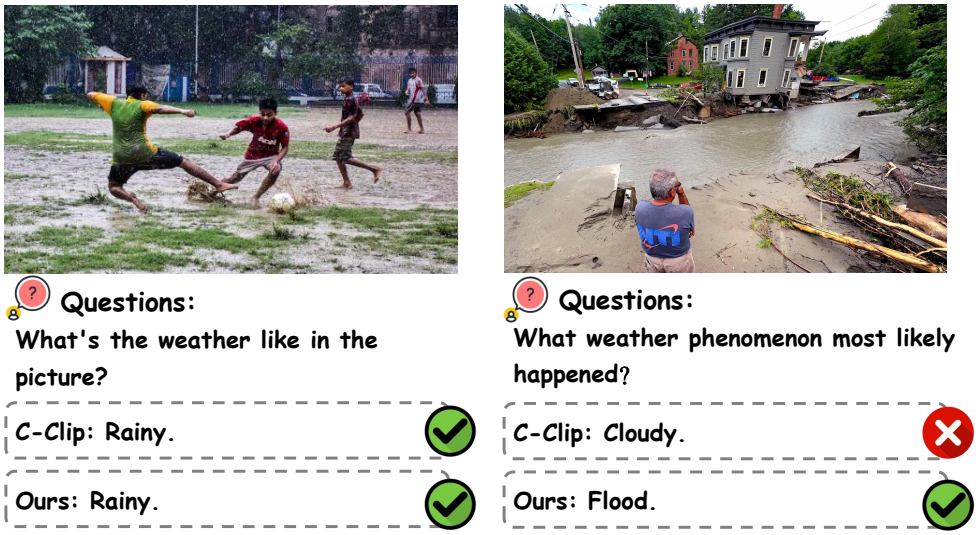

Fig. 7: Qualitative comparison on weather-related VQA. IAR preserves visually grounded reasoning under consequence-driven cases, while C-CLIP overweights superficial sky cues.

while IAR routes to the appropriate experts and answers *Flood*. This suggests that IAR is less tied to the most salient surface cue and can better exploit scene-level evidence when the answer depends on indirect visual consequences.

5 Conclusion

We presented IAR, an interference-aware adapter routing framework for continual vision–language learning under non-stationary task streams. The central idea is to convert cross-task interference from an implicit training failure into an explicit instance-level signal. By measuring metric-consistent projection energy, IAR routes each sample to the least interfering experts, regularizes expert subspaces through principal-angle packing, and grows LoRA capacity only when the incoming update contains substantial residual energy outside existing expert subspaces. This measure–select–grow design provides a unified mechanism for balancing stability and plasticity without updating the frozen backbone or relying on large replay buffers.

Extensive experiments across classification, structured vision–language concept matching, generative VQA, and continual retrieval show that IAR consistently improves both final performance and retention under matched parameter budgets. The analysis further supports the design: interference energy correlates with forgetting, subspace packing increases expert separation over time, and residual-triggered rank growth remains sparse while benefiting adaptation, especially in mid–high layers.

Future work will study how interference-aware routing can be combined with replay and data selection, extended to richer multimodal streams such as video–audio MLLMs, and adapted to resource-constrained edge scenarios [16].

References

1. Buzzega, P., Boschini, M., Porrello, A., Abati, D., Calderara, S.: Dark experience for general continual learning: a strong, simple baseline. Advances in neural information processing systems **33**, 15920–15930 (2020)
2. Chaudhry, A., Rohrbach, M., Elhoseiny, M., Ajanthan, T., Dokania, P., Torr, P., Ranzato, M.: Continual learning with tiny episodic memories. In: Workshop on Multi-Task and Lifelong Reinforcement Learning (2019)
3. Chen, X., Liu, T., Zhao, H., Zhou, G., Zhang, Y.Q.: Cerberus transformer: Joint semantic, affordance and attribute parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19649–19658 (2022)
4. Chen, Y., Cao, Z., Ren, H., Yang, C., Li, W., Wang, S., Wang, Y., Zhang, L., Shao, Y., Zhao, Z., et al.: Roborouter: Training-free policy routing for robotic manipulation. arXiv preprint arXiv:2603.07892 (2026)
5. Cui, Z., Peng, Y., Wang, X., Zhu, M., Zhou, J.: Continual vision-language retrieval via dynamic knowledge rectification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 11704–11712 (2024)
6. Fu, Z., Hu, Y., Yang, Q., Zhang, S., Chen, Z., Li, Z.: Air-know: Arbiter-calibrated knowledge-internalizing robust network for composed image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2658–2670 (2026)
7. Gao, Y., Fei, N., Lu, H., Lu, Z., Jiang, H., Li, Y., Cao, Z.: Bmu-moco: Bidirectional momentum update for continual video-language modeling. Advances in Neural Information Processing Systems **35**, 22699–22712 (2022)
8. Gao, Z., Zhang, X., Xu, K., Mao, X., Wang, H.: Stabilizing zero-shot prediction: A novel antidote to forgetting in continual vision-language tasks. Advances in Neural Information Processing Systems **37**, 128462–128488 (2024)
9. Garg, S., Farajtabar, M., Pouransari, H., Vemulapalli, R., Mehta, S., Tuzel, O., Shankar, V., Faghri, F.: Tic-clip: Continual training of clip models. In: The Twelfth International Conference on Learning Representations (2024)
10. Ge, C., Wang, X., Zhang, Z., Chen, H., Fan, J., Huang, L., Xue, H., Zhu, W.: Dynamic mixture of curriculum lora experts for continual multimodal instruction tuning. In: Forty-second International Conference on Machine Learning (2025)
11. Greco, C., Plank, B., Fernández, R., Bernardi, R.: Psycholinguistics meets continual learning: Measuring catastrophic forgetting in visual question answering. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 3601–3605 (2019)
12. Guo, H., Zeng, F., Xiang, Z., Zhu, F., Wang, D.H., Zhang, X.Y., Liu, C.L.: Hide-llava: Hierarchical decoupling for continual instruction tuning of multimodal large language model. arXiv preprint arXiv:2503.12941 (2025)
13. Huai, T., Zhou, J., Wu, X., Chen, Q., Bai, Q., Zhou, Z., He, L.: Cl-moe: Enhancing multimodal large language model with dual momentum mixture-of-experts for continual visual question answering. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19608–19617 (2025)
14. Huang, C., Seto, S., Pouransari, H., Farajtabar, M., Vemulapalli, R., Faghri, F., Tuzel, O., Theobald, B.J., Susskind, J.M.: Proxy-fda: Proxy-based feature distribution alignment for fine-tuning vision foundation models without forgetting. In: Forty-second International Conference on Machine Learning (2025)
15. Huang, L., Cao, X., Lu, H., Meng, Y., Yang, F., Liu, X.: Mind the gap: Preserving and compensating for the modality gap in clip-based continual learning. arXiv preprint arXiv:2507.09118 (2025)
16. Jiao, R., Zhang, J., Li, C., Hu, L.: Large-kernel spatially parallel feature fusion for monocular 3d perception in autonomous driving. Knowledge-Based Systems **343**, 115998 (2026)
17. Jin, X., Du, J., Sadhu, A., Nevatia, R., Ren, X.: Visually grounded continual learning of compositional phrases. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 2018–2029 (2020)

18. Kang, B., Wang, L., Wu, Z., Feng, T., Li, Y., Gao, Y., Li, W.: Dynamic multi-layer null space projection for vision-language continual learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2077–2086 (2025)
19. Li, Z., Hoiem, D.: Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence **40**(12), 2935–2947 (2017)
20. Li, Z., Chen, Z., Fu, Z., Wang, W., Hu, Y., Guan, W., Nie, L.: Omniego-r^2: A routed reasoning framework for the 1st cross-domain egocross challenge at cvpr 2026. arXiv preprint arXiv:2605.24481 (2026)
21. Li, Z., Hu, Y., Chen, Z., Huang, Q., Qiu, G., Fu, Z., Liu, M.: Retrack: Evidence-driven dual-stream directional anchor calibration network for composed video retrieval. In: AAAI. vol. 40, pp. 23373–23381 (2026)
22. Li, Z., Hu, Y., Chen, Z., Zhang, M., Fu, Z., Nie, L.: Conesep: Cone-based robust noise-unlearning compositional network for composed image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16897–16909 (2026)
23. Liang, G., Wang, Z., Hu, J., Zhou, H., Xue, Z., Zhang, J., Xu, D., Yu, Q.: Render-in-the-loop: Vector graphics generation via visual self-feedback. arXiv preprint arXiv:2604.20730 (2026)
24. Liang, G., Wang, Z., Wang, C., Hu, J., Zhou, H., Liu, J., Zhang, J., Xu, D., Yu, Q.: Vanim: Rendering-aware sparse state modeling for structure-preserving vector animation. arXiv preprint arXiv:2605.01517 (2026)
25. Liang, Y.S., Li, W.J.: Inflora: Interference-free low-rank adaptation for continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23638–23647 (2024)
26. Lin, J., Zhu, C., Kneuertz, P.J., Bai, Y., Xue, Y.: When models learn to ask why: Adaptive causal reasoning for trustworthy medical vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5556–5568 (2026)
27. Liu, W., Zhu, F., Wei, L., Tian, Q.: C-clip: Multimodal continual learning for vision-language model. In: The Thirteenth International Conference on Learning Representations (2025)
28. Liu, X., Xia, X., Ng, S.K., Chua, T.S.: Continual multimodal contrastive learning. arXiv preprint arXiv:2503.14963 (2025)
29. Lu, J., Jenni, S., Kafle, K., Shi, J., Zhao, H., Fu, Y.: Seeing through words: Controlling visual retrieval quality with language models. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=yOEmEXmbV8
30. Lu, J., Wang, H., Xu, Y., Wang, Y., Yang, K., Fu, Y.: Representation potentials of foundation models for multimodal alignment: A survey. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 16669–16684. Association for Computational Linguistics (nov 2025)
31. Lu, J., Wang, H., Yang, K., Zhang, Y., Jenni, S., Fu, Y.: The Indra representation hypothesis for multimodal alignment. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=D2NR5Zq6PG
32. Luo, M.L., Zhou, Z.H., Wei, T., Zhang, M.L.: Lada: Scalable label-specific clip adapter for continual learning. In: Forty-second International Conference on Machine Learning (2025)
33. Maharana, A., Yoon, J., Chen, T., Bansal, M.: Adapt-∞: Scalable continual multimodal instruction tuning via dynamic data selection. In: The Thirteenth International Conference on Learning Representations (2025)
34. Marouf, I.E., Tartaglione, E., Lathuilière, S., van de Weijer, J.: Ask and remember: A questions-only replay strategy for continual visual question answering. arXiv preprint arXiv:2502.04469 (2025), https://arxiv.org/pdf/2502.04469
35. Ni, Z., Wei, L., Tang, S., Zhuang, Y., Tian, Q.: Continual vision-language representation learning with off-diagonal information. In: International Conference on Machine Learning. pp. 26129–26149. PMLR (2023)
36. Shen, F., Jiang, X., He, X., Ye, H., Wang, C., Du, X., Li, Z., Tang, J.: Imagdressing-v1: Customizable virtual dressing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 6795–6804 (2025)

37. Shen, F., Tang, J.: Imagpose: A unified conditional framework for pose-guided person generation. Advances in neural information processing systems **37**, 6246–6266 (2024)
38. Smith, J.S., Cascante-Bonilla, P., Arbelle, A., Kim, D., Panda, R., Cox, D., Yang, D., Kira, Z., Feris, R., Karlinsky, L.: Construct-vl: Data-free continual structured vl concepts learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14994–15004 (2023)
39. Tang, L., Tian, Z., Li, K., He, C., Zhou, H., Zhao, H., Li, X., Jia, J.: Mind the interference: Retaining pre-trained knowledge in parameter efficient continual learning of vision-language models. In: European conference on computer vision. pp. 346–365. Springer (2024)
40. Tian, B., Liu, M., Gao, H.a., Li, P., Zhao, H., Zhou, G.: Unsupervised road anomaly detection with language anchors. In: 2023 IEEE international conference on robotics and automation (ICRA). pp. 7778–7785. IEEE (2023)
41. Tian, S., Chen, G., Li, B., Ma, J., Yu, Z.: Curvature-adaptive consistency flow matching: Autonomous trajectory optimization via reinforcement learning. arXiv preprint arXiv:2606.22394 (2026)
42. Wan, T.S., Chen, J.C., Wu, T.Y., Chen, C.S.: Continual learning for visual search with backward consistent feature embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16702–16711 (2022)
43. Wang, Z., Che, C., Wang, Q., Li, Y., Shi, Z., Wang, M.: Separable mixture of low-rank adaptation for continual visual instruction tuning. arXiv preprint arXiv:2411.13949 (2024)
44. Wu, B., Shi, W., Wang, J., Ye, M.: Synthetic data is an elegant gift for continual vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2813–2823 (2025)
45. Xiao, C., Xu, T., Ma, S., Jiang, Y., Gao, H., Wu, Y.: Reversible primitive–composition alignment for continual vision–language learning. In: The Fourteenth International Conference on Learning Representations (2026)
46. Xu, Y., Chen, Y., Nie, J., Wang, Y., Zhuang, H., Okumura, M.: Advancing cross-domain discriminability in continual learning of vision-language models. Advances in Neural Information Processing Systems **37**, 51552–51576 (2024)
47. Yu, J., Zhou, S., Yang, D., Li, S., Wang, S., Hu, X., Xu, C., Xu, Z., Shu, C., Yuan, Z.: Mquant: Unleashing the inference potential of multimodal large language models via static quantization. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 1783–1792 (2025)
48. Yu, J., Zhuge, Y., Zhang, L., Hu, P., Wang, D., Lu, H., He, Y.: Boosting continual learning of vision-language models via mixture-of-experts adapters. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23219–23230 (2024)
49. Yu, L., Chen, Z., Kuang, P., Feng, Z., Zhou, F., Wang, L., Dobbie, G.: Causally-grounded dual-path attention intervention for object hallucination mitigation in LVLMs. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 36021–36029 (2026)
50. Yu, L., Sun, J., Kuang, P., Zhou, R., Zhou, F., Feng, Z.: Bimodal debiasing for text-to-image diffusion: Adaptive guidance in textual and visual spaces. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 11249–11258 (2025)
51. Yu, Z., Tian, S., Chen, G.: Divergence of neural tangent kernel in classification problems. In: The Thirteenth International Conference on Learning Representations (2025)
52. Zhang, D., Ren, Y., Li, Z.Z., Yu, Y., Dong, J., Li, C., Ji, Z., Bai, J.: Enhancing multimodal continual instruction tuning with branchlora. arXiv preprint arXiv:2506.02041 (2025)
53. Zhang, F., Wang, Z., Zhang, X., Xu, C.: Overcoming dual drift for continual long-tailed visual question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4413–4423 (2025)
54. Zhang, J., Zhao, C., Xiao, C., Duan, R., Mo, W., Gao, H., Wang, W.: Pi-cca: Prompt-invariant cca certificates for replay-free continual multimodal learning. In: The Fourteenth International Conference on Learning Representations (2026)
55. Zhang, J., Song, X., Li, Y., Liang, D., Zhang, Z., Cai, J.: Adaptive dual cross-attention network for multispectral object detection in autonomous driving. Expert Systems with Applications p. 132012 (2026)

56. Zhang, J., Xiang, M., Hu, Y., Hao, W., Lei, L., Yi, K.: Multivariate feature learning and associative spatial information enhancement for snow object detection in autonomous driving. Engineering Applications of Artificial Intelligence **175**, 114672 (2026)
57. Zhang, X., Zhang, F., Xu, C.: Vqacl: A novel visual question answering continual learning setting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19102–19112 (2023)
58. Zhao, Z., Yang, H., Liao, B., Zeng, Y., Yan, S., Gu, Y., Liu, P., Zhou, Y., Li, H., Civera, J.: Advances in global solvers for 3d vision. arXiv preprint arXiv:2602.14662 (2026)
59. Zheng, H., Han, W., Shen, J.: Decoupling fine detail and global geometry for compressed depth map super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 951–960 (2025)
60. Zheng, H., Zhou, Y., Yan, T., Chen, D., Lu, H., Liao, W., He, T., Peng, P., Shen, J.: Clinical cognition alignment for gastrointestinal diagnosis with multimodal llms. arXiv preprint arXiv:2603.20698 (2026)
61. Zheng, X., Wu, L., Yan, Z., Tang, Y., Zhao, H., Zhong, C., Chen, B., Gong, J.: Large language models powered context-aware motion prediction in autonomous driving. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 980–985. IEEE (2024)
62. Zheng, Z., Ma, M., Wang, K., Qin, Z., Yue, X., You, Y.: Preventing zero-shot transfer degradation in continual learning of vision-language models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 19125–19136 (2023)
63. Zhou, D.W., Li, K.W., Ning, J., Ye, H.J., Zhang, L., Zhan, D.C.: External knowledge injection for clip-based class-incremental learning. arXiv preprint arXiv:2503.08510 (2025)
64. Zhou, F., Mao, Y., Yu, L., Yang, Y., Zhong, T.: Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 4227–4241 (2023)
65. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)
66. Zhou, S., Cao, Y., Nie, J., Fu, Y., Zhao, Z., Lu, X., Wang, S.: Comptrack: Information bottleneck-guided low-rank dynamic token compression for point cloud tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 13773–13781 (2026)
67. Zhou, S., Wang, S., Yuan, Z., Shi, M., Shang, Y., Yang, D.: Gsq-tuning: Group-shared exponents integer in fully quantized training for llms on-device fine-tuning. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 22971–22988 (2025)
68. Zhu, C., Lin, J., Tan, G., Zhu, N., Li, K., Wang, C., Li, S.: Advancing ultrasound medical continuous learning with task-specific generalization and adaptability. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 3019–3025. IEEE (2024)
69. Zhu, C., Lin, Y., Chen, S., Wang, Y., Lin, J.: Medeyes: Learning dynamic visual focus for medical progressive diagnosis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 13916–13924 (2026)
70. Zhu, C., Lin, Y., Shao, J., Lin, J., Wang, Y.: Pathology-aware prototype evolution via llm-driven semantic disambiguation for multicenter diabetic retinopathy diagnosis. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 9196–9205 (2025)
71. Zhu, C., Zeng, J., Jiang, J., Lin, J., Wang, Y.: Medsynapse-v: Bridging visual perception and clinical intuition via latent memory evolution (2026), `https://arxiv.org/abs/2604.26283`
72. Zhu, H., Wei, Y., Liang, X., Zhang, C., Zhao, Y.: Ctp: Towards vision-language continual pretraining via compatible momentum contrast and topology preservation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22257–22267 (2023)
73. Zhu, W., Zhang, Y., Jin, X., Zeng, W., Zhang, L.: Knowledge regularized negative feature tuning of vision-language models for out-of-distribution detection. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 3565–3574 (2025)
74. Zhu, W., Zhang, Y., Jin, X., Zeng, W., Zhang, L.: Ants: Adaptive negative textual space shaping for ood detection via test-time mllm understanding and reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20–30 (2026)