COMET: Decoupled Distillation, Routing, and Capacity Control for Task-Agnostic Continual Vision–Language Learning

Seungyoon Woo, Keonhee Park, Gunhee Kim
Seoul National University
{seungyoon.woo, keonhee.park}@vision.snu.ac.kr, gunhee@snu.ac.kr

Abstract

Vision–language models (VLMs) such as CLIP retain zero-shot recognition ability, but their open-world use often requires task-agnostic continual learning: tasks arrive sequentially, past data are unavailable, task identity is unknown, and the backbone cannot be retrained. We argue that this setting fails VLMs through three coupled pressures beyond stability–plasticity tradeoff: (1) modal-consistency drift, (2) representation-space interference, and (3) per-task capacity shortfall. We propose COMET (**CO**llaborative **M**ixture-of-**E**xpert **T**ransfer) to resolve these three issues respectively as follows. A frozen teacher provides only a feature reference on a shared image–text pool; Tri-Affinity Distillation preserves image–image, text–text, and image–text geometry; and Student-Aware Cross-Expert Distillation aligns new experts with compatible prior experts so routing errors degrade gracefully. We design COMET to keep capacity and inference separate: a contextual bandit keeps or merges full per-task experts using post-training signals, while a reconstruction router selects sparse experts at test time without teacher semantics. On the X-TAIL benchmark, we show that each proposed components indeed improves the continual learning performance. COMET shows that continual vision–language adaptation can preserve multimodal geometry, coordinate accumulated experts, and allocate capacity jointly when feature reference, expert shaping, routing, and capacity control are separated by construction. Code is available at `https://github.com/seungyoon-woo/comet`.

1 Introduction

Recent Vision–language models [Liu et al., 2023, 2024, Chen et al., 2024, Wang et al., 2024] such as CLIP [Radford et al., 2021b], trained on web-scale image–text pairs, can recognize objects they were never explicitly taught; this capacity is known as zero-shot transfer. However, in open-world deployment, new tasks can arrive sequentially (chest X-rays [Samir et al., 2023], then satellite imagery [Helber et al., 2019], then microscopy [Iwata et al., 2022]), and the model is expected to be *continually learned* under three desiderata: no access to stored examples from earlier tasks, no knowledge of task identity at inference time, and no budget to retrain the backbone for each new task. Recent task-agnostic image-classification benchmarks such as X-TAIL[Xu et al., 2024] pin down this scenario: during each training phase only the current task's data are available, and at inference the model picks a label from a generalized candidate space containing all classes seen so far and even unseen classes. Because full-backbone updates can be too expensive, the community has focused on *parameter-efficient* continual learning [Houlsby et al., 2019, Li and Liang, 2021, Park et al., 2024]: only small task-specific modules (adapters [Yu et al., 2024a, Luo et al., 2025], prompts, or expert networks) are updated, leaving the backbone fixed. This lighter paradigm can still fail if the stream becomes long, heterogeneous, and open-ended.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

We argue that this failure does not stem from only a stability–plasticity tradeoff [Kirkpatrick et al., 2016, McCloskey and Cohen, 1989], but rather from the simultaneous failure of three axes, each with its own characteristic signature, whereas prior methods typically address only one axis at a time. First, *modal-consistency drift* is the divergence between a model's pre-training and downstream alignments as adaptation accumulates. A CLIP-style model [Radford et al., 2021b] carries two alignments: one from pre-training (behind zero-shot transfer), the other from downstream adaptation accumulated across the stream. Protecting one could silently damage the other; for example, a model well adapted to microscopy may begin to misrank everyday objects it had previously recognized. Second, *representation-space interference* is the incompatibility among independently-trained experts under imperfect routing. Since each is trained in isolation, a routing mistakenly sends an ambiguous input into a module that has never been shaped to handle it, and errors concentrate near expert boundaries. Third, *per-task capacity shortfall* is the inability of a small module to represent feature geometries far from pre-training. A small adapter lacks the representational power for the feature geometry of medical imaging, microscopy, and remote sensing.

Existing method families each make this three-way trade explicit. Weight-interpolation methods (e.g., WiSE-FT [Wortsman et al., 2022], ZSCL [Zheng et al., 2023]) mitigate drift but starve distant domains of capacity, because no convex combination of pre-trained and ask-adapted weights can introduce new feature directions. Mixture-of-experts [Shazeer et al., 2017a] with lightweight adapters (e.g., MoE-Adapters [Yu et al., 2024a], LADA [Luo et al., 2025]) add capacity but never coordinate the experts they accumulate, because each adapter is trained in isolation, leaving no shared geometry Pascanu et al. [2014] for the router to degrade into. Full fine-tuning can recover capacity but lose both alignments, because rewriting the backbone may erase the pre-training geometry itself. Each axis is relieved at the expense of the other two.

We show that the three pressures can be resolved jointly, but only if their mechanisms are kept strictly disjoint. Entangling them is why prior methods [Wortsman et al., 2022, Zheng et al., 2023, Yu et al., 2024b, Xu et al., 2024] trade one failure for another. When the teacher model is also used as the router, its inherent semantic biases can distort capacity allocation: tasks appear similar from its perspective due to pre-training, even if their downstream representation geometries differ substantially. Similarly, a merging rule [Frankle et al., 2020, Neyshabur et al., 2020] that reuses the adaptation loss conflates compatibility with plasticity; it merges two experts whenever they fit the current task equally well, ignoring possible long-term overlap. An adapter that simultaneously absorbs new tasks and corrects past ones eventually fails at both, as its capacity is divided between forward learning and backward retention, with each undermining the other. Across these methods, a common pattern emerges: multiple roles are coupled into a single mechanism, leading to systemic collapse.

COMET (**CO**llaborative **M**ixture-of-**E**xpert **T**ransfer)which enforces this separation by assigning each role to a dedicated component. A frozen vision–language teacher serves only as a task-invariant feature reference: its affinity structure on a shared pool of image–text pairs anchors distillation, and its frozen image encoder fixes the feature space in which routers operate. Neither the router nor the merging criterion uses its predictions. A lightweight contextual-bandit policy [Beygelzimer et al., 2011] serves only as the capacity controller, deciding at each task boundary whether to allocate a new expert or consolidate into an existing one. Using a bandit [Lattimore and Szepesvári, 2020] (rather than a fixed threshold) allows allocation to track the novelty of the incoming stream. It reads only post-training signals that never enter the adaptation loss, such as accuracy on a held-out subset of the reference pool and weight-space similarity to the nearest prior. Its reward is defined as the post-action accuracy on a held-out reference pool. Distillation [Hinton et al., 2015a,b] shapes only the experts. Since allocation does not inherit teacher semantics and compatibility corrections do not entangle with capacity control, the retained expert count grows sublinearly in the task count whenever cumulative novelty is sublinear or stays within a bounded region of the representation space. Each new task's feature geometry tends to lie close to the span of at least one prior expert on the shared reference pool, so consolidation aligns with novelty. This is an empirical regularity rather than a worst-case guarantee: a stream that keeps entering new regions can still force linear growth, even when each per-task shift is bounded.

Three mechanisms realize this decomposition, one per failure axis, each breaking one coupling. **Tri-Affinity Distillation (TAD)** addresses modal-consistency drift by matching image–image, text–text, and image–text similarity structures between student and frozen teacher on the shared pool, anchoring the student to pre-training geometry. **Student-Aware Cross-Expert Distillation (SACED)** addresses

routing-induced interference by aligning each new expert with compatible prior experts on the same pool, so a misrouted sample lands in a still-coherent module, not foreign ones. Since the router is imperfect, we design for graceful degradation rather than assuming perfect routing. **Full per-task experts with bandit-controlled merging** address per-task capacity shortfall: COMET trains a full compact expert per task instead of a shared-backbone adapter [Latif and Zhai, 2024], and the bandit merges when consolidation is justified. The merge operates on already-frozen candidates at task boundaries, avoiding the optimization-time interpolation of weight-ensembling.

Three lines of evidence from X-TAIL experiments validate this decomposition. Axis-wise ablations confirm individual necessity with each mechanism's removal restores only its own failure mode: drift, interference, or shortfall:

- **Modal geometry is preserved.** TAD improves Transfer accuracy by 4.7%p over the no-distillation variant, and a 2.6× smaller student retains the teacher's zero-shot ability while improving downstream accuracy.
- **Routing failures are absorbed, not amplified.** SACED raises accuracy on misrouted samples by 10.8%p without exemplar replay, and 52.5% of misrouted samples are still classified correctly.
- **Capacity matches distant domains at sublinear storage cost.** Full experts under bandit-controlled merging improve Transfer by 18.7%p over the strongest baseline on a four-dataset benchmark spanning medical, microscopy, and histopathology domains. The pool retains roughly 5-8 experts after 10 X-TAIL tasks, illustrating sublinear growth in practice.

These results reframe continual vision–language adaptation: not as a single trade-off to balance, but as three problems to solve jointly: preserving multimodal geometry, coordinating accumulated modules, and allocating sufficient capacity. To the best of our knowledge, this is the first work that establishes training-time role disjointness, across the feature reference, the capacity controller, and the per-expert shaping objective, as the central design principle of a parameter-efficient continual VLM framework. Disjointness is not emergent, it must be enforced by construction.

2 Related Work

Continual Learning Settings. Conventional continual learning [Hadsell et al., 2020, De Lange et al., 2021, Li and Hoiem, 2018, Rebuffi et al., 2017] is typically studied under Task-Incremental Learning (TIL) [Li and Hoiem, 2018] and Class-Incremental Learning (CIL) [Rebuffi et al., 2017] settings, where models are sequentially trained on tasks and evaluated with or without access to task identities at inference, respectively. To relax the assumption that all tasks originate from a single domain, the Multi-domain Task-Incremental Learning (MTIL) [Zheng et al., 2023] considers tasks drawn from diverse domains with domain-specific class distributions. More recently, Cross-domain Task-Agnostic Incremental Learning (X-TAIL) [Xu et al., 2024] further removes task identity information and requires models to perform inference across both seen and unseen domains. The X-TAIL is a particularly challenging benchmark in VLM-based continual learning, as it demands not only the retention of previously learned knowledge but also the preservation of the zero-shot ability.

CLIP-based Continual Learning Methods. Recent continual learning studies [Wortsman et al., 2022, Zheng et al., 2023, Yu et al., 2024a, Xu et al., 2024, Luo et al., 2025] have widely adopted CLIP [Radford et al., 2021b], leveraging its strong representations and zero-shot ability. Conventional approaches fall into two categories: fine-tuning and Parameter-Efficient Fine-Tuning (PEFT) [Houlsby et al., 2019, Hu et al., 2022, Li and Liang, 2021]. Fine-tuning methods [Wortsman et al., 2022, Zheng et al., 2023] directly update the CLIP. WISE-FT [Wortsman et al., 2022] fine-tunes CLIP and interpolates the updated weights with the pre-trained weights to balance adaptation and zero-shot transfer. ZSCL [Zheng et al., 2023] further stabilizes the interpolation via knowledge distillation from both the pre-trained model and a reference dataset, reducing unnecessary shift and preserving zero-shot capability. PEFT methods [Yu et al., 2024a, Xu et al., 2024, Luo et al., 2025], by contrast, keep the backbone frozen and introduce lightweight modules. MoE-Adapters [Yu et al., 2024a] augment CLIP with Mixture-of-Experts adapters and routers, while retaining the CLIP for zero-shot ability on unseen tasks. RAIL [Xu et al., 2024] employs a ridge regression–based adapter paired with a training-free fusion module to preserve zero-shot ability. LADA [Luo et al., 2025] appends class-wise memory vectors to the image encoder via label-specific adapters, enabling efficient adaptation

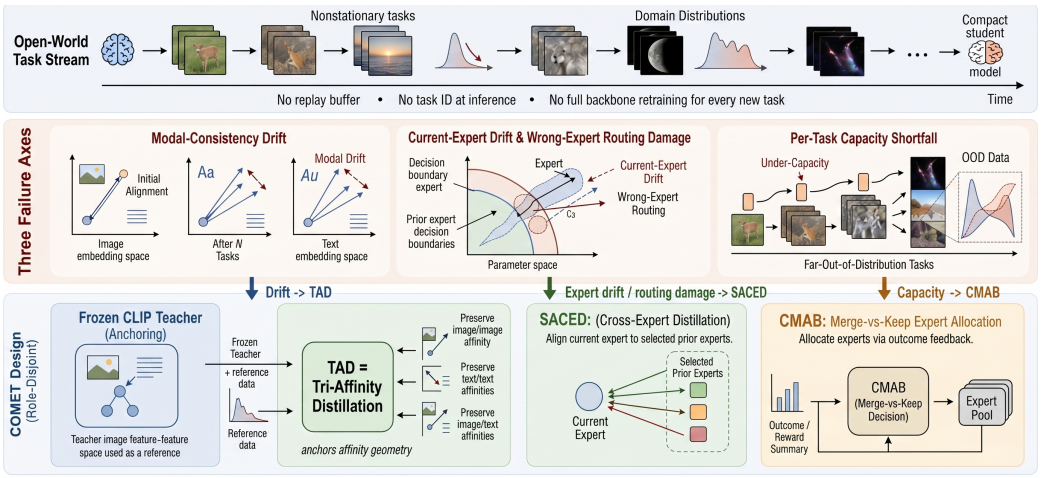


Figure 1: **Overview of COMET.** Targeting task-agnostic continual vision–language learning without replay or full-backbone retraining. It decomposes three failure modes: modal drift, routing-induced expert interference, and capacity shortfall into three role-disjoint remedies: TAD for CLIP-geometry preservation, SACED for cross-expert compatibility, and CMAB for adaptive expert allocation.

without auxiliary parameter selection at inference. Unlike existing methods that fine-tune a CLIP or attach lightweight modules to the pre-trained CLIP, our approach treats each CLIP as a task-specific expert and extends the MoE [Shazeer et al., 2017a] to preserve sufficient plasticity.

3 Preliminary

The Continual Learning Setting. We consider the X-TAIL [Xu et al., 2024], where a model sequentially learns tasks $\mathcal{T} = \{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_K\}$. Each task $\mathcal{T}_k$ has a labeled dataset $\mathcal{D}_k = \{(x_i^{(k)}, c_i^{(k)})\}_{i=1}^{N_k}$ with class labels $c_i^{(k)} \in \mathcal{C}^{(k)}$. At step k, the model trains only on $\mathcal{D}_k$ without access to previous task data. Let $\mathcal{C}_L^{(k)} = \bigcup_{i=1}^{k} \mathcal{C}^{(i)}$ denote the set of seen class labels after learning task k, and let $\mathcal{C}_U$ denote unseen class labels that are not observed during supervised continual training. Following the X-TAIL evaluation protocol, inference is task-agnostic and is performed over the generalized label space $\mathcal{C}_N^{(k)} = \mathcal{C}_L^{(k)} \cup \mathcal{C}_U$. At test time, the task identity is unknown and the model selects a label from both previously seen and unseen classes.

4 Approach: COMET

Figure 1 maps each failure axis to a disjoint mechanism. Given a task stream, COMET incrementally builds a pool of task-specific CLIP experts and isolates each role by construction so that geometry preservation, expert compatibility, and capacity control never share an objective. Accepted experts receive no further gradient updates; they change only when the capacity controller merges a later expert into them. COMET uses a frozen vision–language teacher only as a feature reference: the teacher anchors distillation on a shared image–text reference pool $\mathcal{D}_{\text{ref}}$, and its frozen image encoder supplies the feature space in which routers operate. Neither the router nor the merge rule uses its predictions. A contextual bandit is used only as the capacity controller, deciding at task boundaries whether the newly trained expert should remain separate or be merged with a compatible prior expert (§4.4). The expert itself is shaped by two distillation objectives: Tri-Affinity Distillation (TAD; §4.1) preserves pre-trained multimodal geometry, and Student-Aware Cross-Expert Distillation (SACED; §4.2) aligns the new expert with compatible prior experts so routing errors degrade gracefully. At inference, a lightweight reconstruction-based router selects sparse top-k experts for each input without consulting teacher semantics (§4.5). The remainder of this section follows the data path: anchor the new expert to multimodal geometry, coordinate it with compatible prior experts, decide whether the pool should grow, and route test inputs by reconstruction fit.

Each task follows the same execution order. COMET initializes a new expert from the shared CLIP initialization, trains it on the current task data with task supervision plus the two distillation terms, evaluates post-training compatibility signals on $\mathcal{D}_{\text{ref}}$, and only then invokes the bandit to keep or

merge the expert. The per-task objective is $\mathcal{L}_k = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{TAD}} + \lambda_2 \mathcal{L}_{\text{SACED}}$, where $\mathcal{L}_{\text{task}}$ is the supervised loss on $\mathcal{D}_k$, the two distillation losses use $\mathcal{D}_{\text{ref}}$, and λ_2 balances SACED against the geometry-anchoring TAD term (within TAD itself, λ_1 controls the cross-modal channel. When no compatible prior expert exists, the SACED expert set is empty and the SACED term is zero. Only after this optimization step does the capacity controller act. This ordering prevents the teacher, the router, and the merge policy from becoming hidden parts of the adaptation objective.

4.1 Tri-Affinity Distillation (TAD)

TAD targets modal-consistency drift by preserving the relation structure of the pre-trained CLIP space [Radford et al., 2021a]. Sequential adaptation on non-i.i.d. tasks can move image features, text features, and their cross-modal alignment in different directions, eroding both transferability and zero-shot generalization [Wortsman et al., 2022, Zheng et al., 2023]. Because task labels vary across the stream, logit distillation gives an unstable anchor; TAD instead uses the frozen CLIP teacher only as a label-agnostic feature reference on $\mathcal{D}_{\text{ref}}$ and matches the three CLIP affinity channels—image–image, text–text, and image–text—between the student expert and the teacher. These three constraints preserve the geometry needed for transfer while leaving enough freedom for task-specific adaptation.

Intra-Modality Distillation. We apply intra-modality distillation [Hinton et al., 2015a] separately to image–image and text–text relations so the student keeps the teacher's within-modality geometry. For a reference mini-batch of B image–text pairs from $\mathcal{D}_{\text{ref}}$, let $S_{\text{I}}, T_{\text{I}} \in \mathbb{R}^{B\times d}$ denote the ℓ_2-normalized image embeddings produced by the student and teacher, respectively, and define $S_{\text{T}}, T_{\text{T}} \in \mathbb{R}^{B\times d}$ analogously for text. The image-affinity loss minimizes the discrepancy between the student and teacher pairwise-similarity matrices, $\mathcal{L}_{\text{I}} = \left\| S_{\text{I}} S_{\text{I}}^{\top} - T_{\text{I}} T_{\text{I}}^{\top} \right\|_F^2$, and the text-affinity loss $\mathcal{L}_{\text{T}}$ takes the same form on $(S_{\text{T}}, T_{\text{T}})$. We do not introduce an additional affinity temperature, since a scalar divisor inside this squared loss would only rescale the objective and is therefore equivalent to adjusting the relative loss weight. Matching these within-modality structures inherits the pre-trained CLIP's modality-specific geometry and reduces representation drift during adaptation.

Inter-Modality Distillation. To maintain the shared vision–language embedding space, we align image-to-text and text-to-image similarity distributions between the student expert and the frozen pre-trained CLIP teacher. For each modality pair $(a, b) \in \{(\text{I}, \text{T}), (\text{T}, \text{I})\}$ and model $m \in \{S, T\}$, we compute scaled cross-modal logits $Z^m_{a\to b} = \alpha\, E^m_a (E^m_b)^{\top},\ m \in \{S, T\}$, where α is the logit scaling factor, and $E^S_{\text{I}} = S_{\text{I}}$, $E^S_{\text{T}} = S_{\text{T}}$, $E^T_{\text{I}} = T_{\text{I}}$, and $E^T_{\text{T}} = T_{\text{T}}$. Converting each row to a probability distribution with temperature τ_c, the inter-modality loss uses a teacher-to-student forward KL divergence:

$$\mathcal{L}_{\text{Cross}} = \frac{\tau_c^2}{2B} \sum_{(a,b)\in\{(\text{I},\text{T}),(\text{T},\text{I})\}} \sum_{i=1}^{B} \sum_{\ell=1}^{B} P^T_{a\to b,\, i\ell} \log \frac{P^T_{a\to b,\, i\ell}}{P^S_{a\to b,\, i\ell}}, \tag{1}$$

where $P^m_{a\to b,\, i\cdot} = \text{softmax}(Z^m_{a\to b,\, i\cdot}/\tau_c)$, $\text{sg}(\cdot)$ denotes stop-gradient, and ℓ ranges over the within-batch candidates. The outer summation averages the two cross-modal directions; the KL divergence itself remains asymmetric, with the frozen teacher distribution serving as the target.

The final loss of TAD is $\mathcal{L}_{\text{TAD}} = \mathcal{L}_{\text{I}} + \mathcal{L}_{\text{T}} + \lambda_1 \mathcal{L}_{\text{Cross}}$, where the two intra-modality losses and the bidirectional cross-modal distillation loss keep all three affinity channels tied to the same teacher geometry. Consequently, TAD constrains drift without forcing the student to imitate task-specific teacher logits.

4.2 Student-Aware Cross-Expert Distillation (SACED)

SACED targets representation-space incompatibility among independently trained experts under imperfect routing. Freezing prior experts prevents direct overwriting, but a new expert optimized only for the current task can still become geometrically incompatible with its neighbors, so a misrouted input may land in a module with no shared representational frame. SACED therefore identifies compatible prior experts on a shared reference pool $\mathcal{D}_{\text{ref}}$ [Zheng et al., 2023, Yu et al., 2024b] and aligns the new expert with their attention-level reasoning through a collaborative distillation loss. The frozen teacher is used only as a common reference anchor [Li and Hoiem, 2018, Hinton et al., 2015b, Zagoruyko and Komodakis, 2017a]. The goal is graceful failure: even when the router selects a nearby but non-specialized expert, the activated module remains close to the representation geometry of compatible experts.

Expert Candidate Selection. SACED evaluates previously learned experts on the shared reference pool $\mathcal{D}_{\text{ref}}$, which may be out-of-distribution with respect to the current task, without replaying old task data. A prior expert is rated by a weighted compatibility score combining three signals: (i) co-activation overlap on the top-K_g gating assignments [Shazeer et al., 2017b] with weight 0.5, (ii) top-1 class-prediction agreement [Stanton et al., 2021] with weight 0.3, and (iii) cosine similarity of class-probability distributions [Chen et al., 2025] with weight 0.2. The top-k_{ref} scoring experts above a threshold form $\mathcal{S}_k$, the selected expert set for task k. This keeps cross-expert transfer local to experts likely to share useful representation geometry. If no prior expert passes the threshold, $\mathcal{S}_k = \emptyset$ and the $\mathcal{L}_{SACED}$ is set to zero for the current task.

Teacher-Anchored Attention Alignment. SACED aligns the current expert with selected prior experts using attention-based supervision [Zagoruyko and Komodakis, 2017b], since attention maps provide a label-agnostic signal of representation use [Jia et al., 2021, Li et al., 2021]. To make attention comparisons well-defined, the student expert, the selected prior expert, and the teacher are all evaluated on the same reference input $x_{\text{ref}} \in \mathcal{D}_{\text{ref}}$. This avoids comparing attention maps produced from different images and uses the teacher as a shared reference frame. Given two attention maps A and B, SACED first applies a fixed pooling operation after softmax normalization and then computes a cosine distance:

$$d_{\text{attn}}(A, B) = 1 - \cos(\rho(A), \rho(B)), \quad \text{where } \rho(A) = \text{pool}(\text{softmax}(A)), \tag{2}$$

where $\text{pool}(\cdot)$ denotes the same pooling operation over attention dimensions for all experts, such as averaging over heads, layers, or tokens. The SACED loss then matches the student–teacher attention distance to the selected-prior–teacher attention distance on the same reference input:

$$\mathcal{L}_{\text{SACED}} = \mathbb{E}_{\substack{x_{\text{ref}} \sim \mathcal{D}_{\text{ref}} \\ j \sim \text{Unif}(\mathcal{S}_k)}} \left[\left| d_{\text{attn}}(A_s(x_{\text{ref}}), A_t(x_{\text{ref}})) - \text{sg}(d_{\text{attn}}(A_j(x_{\text{ref}}), A_t(x_{\text{ref}}))) \right| \right], \tag{3}$$

where A_s, A_j, and A_t denote the attention maps of the student expert, the j-th selected prior expert, and the frozen teacher, respectively. The operator $\text{sg}(\cdot)$ denotes stop-gradient, since the selected prior experts and the teacher are fixed during the current task. If $\mathcal{S}_k = \emptyset$, we define $\mathcal{L}_{\text{SACED}} = 0$.

This objective transfers teacher-anchored attention geometry rather than class logits. The student is therefore encouraged to occupy a representation frame compatible with selected prior experts while remaining free to specialize its output predictions for the current task. As a result, SACED reduces the damage caused by inevitable routing mistakes without directly modifying the router.

4.3 Sub-Linear Expert Growth

COMET's expert pool is not intended to grow once every task; its size is controlled by how often the stream presents distributions that are genuinely novel in a fixed teacher-induced representation. Let $\delta_k = \min_{m \le M_{k-1}} d(\psi(\mathcal{D}_k), \psi(E_m))$ be the distance from task k to the current expert pool, where d is a metric in the frozen-teacher feature space and $\delta_1 = +\infty$ because the pool starts empty, and call task k ϵ-novel when $\delta_k > \epsilon$. For any merge-or-keep strategy that adds at most one expert per task and keeps a new expert only on ϵ-novel tasks, the expert count satisfies

$$\mathbb{E}[M_K] \le \sum_{k=1}^{K} \Pr(\delta_k > \epsilon). \tag{4}$$

Hence, if the cumulative novelty is sub-linear, $\sum_{k=1}^{K} \Pr(\delta_k > \epsilon) = o(K)$, then $\mathbb{E}[M_K] = o(K)$. This is a sufficient-condition guarantee rather than a worst-case claim: it covers streams with revisits, finite-cover support, or sublinear cumulative drift, while excluding adversarial streams that continually move into new regions. Appendix A gives the full proof and extends the bound to practical policies with occasional false expansions, including the margin condition needed to translate bandit regret into a sub-linear false-expansion rate. We use the bound to state when expert growth stays sub-linear and rely on experiments to check CMAB against it.

4.4 Expert Merging via Contextual Multi-Armed Bandit

TAD and SACED make experts compatible, but retaining every new expert still increases storage with the stream. COMET therefore uses a contextual multi-armed bandit (CMAB) as a post-training capacity controller. After task k, the new expert E_{new} is either kept or merged into a previous expert,

i.e., $a_k \in \{\text{KEEP}(E_{\text{new}})\} \cup \{\text{MERGE}(E_{\text{new}}, E_c) : E_c \in \mathcal{E}_{<k}\}$. Because this decision is made after expert optimization, capacity control is decoupled from the representation losses.

COMET first forms a small candidate set $\mathcal{C}_k$ using lightweight compatibility signals such as centroid similarity, routing overlap, and validation-based cross-performance. Classification accuracy and rewards are computed on the labeled held-out split $\mathcal{V}_k$ of the current task; the unlabeled reference pool $\mathcal{D}_{\text{ref}}$ is used only for label-free feature, routing, and distribution-agreement signals. For candidate E_c, the context is $\phi_{k,c} = [\text{Acc}_k, m_{k,c}, -d_{k,c}, \kappa_k, -\ell_c]^\top$, where $m_{k,c}$ measures merge compatibility, $d_{k,c}$ estimates degradation risk, κ_k is capacity pressure, and ℓ_c penalizes repeatedly merging into an overloaded expert. The bandit score is

$$B_k(c) = w^\top \phi_{k,c} + \frac{\beta_k}{1 + N_k}, \tag{5}$$

with learned weights w, decaying exploration bonus β_k, and update count N_k. Let $c^\star = \arg\max_{c \in \mathcal{C}_k} B_k(c)$. COMET merges into $E_{c^\star}$ when $B_k(c^\star) > \tau_k$ and otherwise keeps the new expert. Thus, merging is an outcome-adaptive decision over compatibility, risk, capacity pressure, and load, rather than a fixed similarity threshold.

After the selected action is evaluated, both MERGE and KEEP outcomes update the policy. The reward is the validation accuracy of the temporary merged expert minus a small merge cost for MERGE, and the validation accuracy of E_{new} for KEEP. To avoid monotone drift from positive-valued accuracy rewards, CMAB uses a centered advantage $A_k = r_k - \bar{r}_k$ and an action sign $y_k = +1$ for MERGE, $y_k = -1$ for KEEP; the update follows the form $w \leftarrow \Pi(w + \eta_w A_k y_k \phi_{k,c^\star})$ with EMA smoothing. The threshold τ_k is adapted by the same advantage signal, lowering after successful merges and increasing after failed merges or successful keeps. Full candidate construction, reward shaping, threshold update, and hyperparameters are given in Appendix C.1.

4.5 Inference Routing with Reconstruction

COMET routes without task identity and without teacher predictions; the frozen teacher image encoder supplies only the feature space. For each retained expert E_m, we train a lightweight autoencoder router R_m on that expert's task data in the frozen CLIP image-feature space. At inference, each router produces a reconstruction error $r_m(x) = d^{-1}\|R_m(g(x)) - g(x)\|_2^2$, where $g(\cdot)$ is the frozen image feature extractor. To make errors comparable across experts, COMET converts them into calibrated likelihood-ratio scores

$$\ell_m(x) = \log p(\log r_m(x) \mid m) - \log p(\log r_m(x) \mid \text{bg}), \tag{6}$$

where the foreground distribution is estimated from expert-specific training data and the background distribution is estimated from the shared reference pool. The scores are converted into gates by $w_m(x) = \text{softmax}(\ell_m(x)/\tau_g)_m$, and only the top-$k$ experts, with $k \leq 2$, are evaluated. Their class-text similarity logits are combined by gate-weighted averaging. Calibration details and the top-k activation rule are given in Appendix C.2. This preserves COMET's role separation: TAD and SACED shape expert geometry, CMAB controls retained capacity, and the reconstruction router selects experts at test time without semantic task cues.

5 Experiments

5.1 Experimental Setup

Benchmarks. We evaluate COMET on two complementary continual-learning benchmarks. *X-TAIL* [Xu et al., 2024] sequences ten image-classification datasets (Aircraft [Maji et al., 2013], Caltech101 [Fei-Fei et al., 2004], DTD [Cimpoi et al., 2014], EuroSAT [Helber et al., 2019], Flowers [Nilsback and Zisserman, 2008], Food [Bossard et al., 2014], MNIST [Deng, 2012], Oxford-Pet [Parkhi et al., 2012], StanfordCars [Krause et al., 2013], and SUN397 [Xiao et al., 2010]). Each treated as an individual downstream task following [Zheng et al., 2023, Luo et al., 2025]. A second *heterogeneous 4-task* sequence covers medical imaging (Covid19 [Samir et al., 2023]), microscopy (SEM [Iwata et al., 2022]), and histopathology (Kather-Texture [Kather et al., 2016], NCT100K [Kather et al., 2018]); these domains lie far from CLIP pre-training and probe whether COMET continues to adapt under large distribution shift while retaining prior knowledge. We report full-shot results in the main text; 16-shot results are in Appendix P.

Table 1: Comparison of different continual learning methods on X-TAIL full-shot for each task in terms of *Transfer*, *Average*, and *Last* scores (%). The best results are highlighted with **bold** style.

Method	Aircraft	Caltech101	DTD	EuroSAT	Flowers	Food	MNIST	Pets	Cars	Sun397	Average
`TinyCLIP ViT-39M/16@224px`											
Zero-shot	17.0	75.6	39.0	40.6	65.0	77.4	22.0	78.9	47.8	67.2	53.1
`CLIP ViT-B/16@224px`											
Zero-shot	23.8	74.3	36.4	37.4	64.1	83.4	43.9	87.8	65.5	60.8	57.7
Transfer											
LwF [Li and Hoiem, 2018]	–	64.8	29.1	7.6	49.7	72.0	33.3	64.1	29.8	53.7	44.9
WiSE-FT [Wortsman et al., 2022]	–	68.3	30.8	17.4	59.2	82.5	41.8	74.9	48.3	59.7	53.7
ZSCL [Zheng et al., 2023]	–	71.7	32.6	15.2	63.0	85.0	29.6	83.8	59.3	63.2	55.9
MoE-Adapters [Yu et al., 2024a]	–	76.2	30.9	19.2	67.7	**87.8**	**50.2**	**88.4**	63.0	59.8	60.3
LADA [Luo et al., 2025]	–	74.9	35.5	36.5	66.1	83.9	40.3	88.0	**65.3**	59.6	61.1
COMET (Ours)	–	**89.3**	**47.1**	**49.9**	**67.8**	77.5	32.9	79.1	51.3	**68.7**	**62.6**
Average											
LwF	18.3	80.5	42.3	35.1	49.1	80.9	57.5	71.1	37.0	56.3	52.8
WiSE-FT	28.1	80.0	43.8	38.9	62.9	86.6	61.9	79.5	54.9	61.6	59.8
ZSCL	40.3	78.2	50.0	40.6	73.2	**87.6**	56.2	86.3	64.5	64.8	64.2
MoE-Adapters	50.8	78.7	33.8	47.0	69.0	87.5	56.7	88.3	62.2	60.2	63.4
Primal-RAIL [Xu et al., 2024]	48.7	89.7	58.7	74.2	83.8	86.4	**67.3**	88.7	69.3	64.9	73.2
LADA	52.8	93.7	66.8	77.7	85.1	86.3	63.4	**89.8**	**69.6**	61.7	74.7
COMET (Ours)	**72.4**	**95.1**	**71.9**	**83.8**	**86.3**	80.9	59.1	83.1	58.6	**69.9**	**76.1**
Last											
LwF	9.8	75.2	48.4	24.7	50.9	89.4	90.4	86.8	59.5	79.3	61.4
WiSE-FT	31.0	78.3	47.5	33.0	71.0	90.5	88.4	88.1	78.9	78.8	68.5
ZSCL	35.4	79.2	52.9	34.9	80.0	**90.6**	95.4	91.9	84.8	78.6	72.4
MoE-Adapters	50.8	79.0	34.6	58.9	69.9	87.1	66.4	88.1	59.1	64.1	65.8
Primal-RAIL	45.6	94.2	70.9	94.2	96.6	89.0	98.1	93.5	82.0	76.1	84.0
LADA	52.9	**96.4**	76.0	95.2	97.9	88.8	98.0	**94.2**	86.6	**80.7**	86.7
COMET (Ours)	**72.4**	91.5	**77.7**	**98.1**	**98.6**	84.3	**98.6**	92.2	**87.7**	80.2	**88.1**

Evaluation Metrics. We use standard continual learning metrics. *Last* is the final accuracy after the full stream and reflects end-of-stream retention. *Transfer* averages accuracy on unseen tasks across the stream and captures forward transfer. *Average* is the mean accuracy across all evaluation steps.

Implementation Details. The teacher is the frozen CLIP ViT-B/16 [Radford et al., 2021a], and each per-task expert is a TinyCLIP [Wu et al., 2023] with $\sim$ 58M parameters. For each task, we train one expert for $1{,}000$ iterations with AdamW [Loshchilov and Hutter, 2019] at a learning rate of 10^{-5} and a batch size of 64. SACED draws its reference distribution from a fixed pool of $10{,}800$ image–text pairs from the Conceptual Captions validation set [Sharma et al., 2018], which also supplies the contextual-bandit reward. Default loss weights are $\lambda_1{=}0.2$ for the cross-modal TAD term and $\lambda_2{=}1.0$ for SACED; at inference, the autoencoder router activates the top-k experts with $k{\leq}2$. The full training schedule, CMAB hyperparameters, and routing rule are specified in Appendix C.

5.2 Main Results

X-TAIL. We compare COMET against fine-tuning methods (LwF, WiSE-FT, ZSCL [Zheng et al., 2023]) and parameter-efficient methods (MoE-Adapters, LADA [Luo et al., 2025]); per-task and aggregate metrics are reported in Table 1. COMET achieves the best aggregate Transfer (62.6), Average (76.1), and Last (88.1), improving over LADA by 1.5, 1.4, and 1.4 percentage points (pp), respectively. The largest Last improvement over LADA occurs on Aircraft ($+19.5$ pp; 72.4 vs. 52.9). These gains are obtained with a $2.6\times$ smaller per-expert backbone (TinyCLIP 58M vs. CLIP-B/16 150M).

Heterogeneous 4-task sequence. Table 2 reports the same metrics on the medical, microscopy, and histopathology sequence, where pre-trained CLIP priors no longer dominate. COMET improves Transfer by $+18.7$ pp over ZSCL (31.2 vs. 12.5), Average by $+20.2$ pp (72.9 vs. 52.7), and Last by $+3.5$ pp (96.7 vs. 93.2). On NCT100K alone, COMET reaches 40.5 Transfer, whereas every baseline remains below 1.6. These gains show that full per-task experts provide the capacity that lightweight adapters on a frozen CLIP lack, while TAD anchors adaptation to a shared teacher geometry.

Table 2: **Heterogeneous 4-dataset full-shot results.** Transfer, Average, and Last scores (%). The best result in each metric and column is in **bold**.

Method	Cov19	KT	NCT	SEM	Avg.
Zero-shot					
TinyCLIP-ZS	48.6	5.8	17.0	22.3	23.4
CLIP-B-ZS	48.9	16.3	15.6	24.9	26.4
Transfer					
CLIP-FT	–	10.3	0.0	21.5	10.6
WISE-FT	–	12.0	0.1	23.6	11.9
ZSCL	–	12.7	0.1	24.8	12.5
MoE-Adapters	–	10.7	1.6	10.8	7.7
LADA	–	0.0	0.0	22.6	7.5
COMET (Ours)	–	**19.8**	**40.5**	**33.4**	**31.2**
Average					
CLIP-FT	29.4	30.8	46.7	40.4	36.9
WISE-FT	31.1	38.8	44.3	41.9	39.0
ZSCL	30.8	45.1	41.6	41.5	39.8
MoE-Adapters	95.1	75.2	46.5	32.0	51.2
LADA	72.6	73.5	45.7	38.9	52.7
COMET (Ours)	**95.5**	**77.3**	**69.7**	**49.0**	**72.9**
Last					
CLIP-FT	27.4	5.4	87.8	96.5	54.3
WISE-FT	29.1	27.5	78.3	**96.8**	57.9
ZSCL	26.3	43.1	82.1	91.5	60.8
MoE-Adapters	94.8	96.5	91.1	90.2	93.2
LADA	72.7	**97.0**	91.4	88.0	87.3
COMET (Ours)	**95.5**	96.5	**99.0**	95.8	**96.7**

ZS: zero-shot; Cov19: Covid19; KT: Kather-Texture; NCT: NCT 100K.

Table 3: **Ablations and inference cost.** (a) TAD/SACED ablation on X-TAIL; (b) merge policy; (c) inference cost. Best results are in **bold**.

(a) TAD/SACED ablation

TAD	SACED	Trans.	Avg.	Last
Full-shot				
–	–	56.6	73.9	85.6
✓	–	61.3	75.4	87.6
–	✓	59.9	75.1	85.3
✓	✓	**62.6**	**76.1**	**88.1**
16-shot				
–	–	52.7	65.1	73.3
✓	–	57.7	69.2	77.9
–	✓	53.3	66.3	73.8
✓	✓	**60.9**	**72.6**	**83.3**

(b) Merge policy

Variant	M	Stored	Trans.	Avg.	Last
Random	3	261	58.9	64.9	69.1
Merge	1	145	58.4	58.4	58.4
Expand	10	667	61.6	75.8	87.3
COMET	8	551	**62.6**	**76.1**	**88.1**

(c) Inference cost

Method	Active	GFLOPs	Lat.	Img/s
LwF/WISE-FT/ZSCL	~ 150	17.6	166.3	385
MoE-Adapt.	~ 180	18.4	196.5	326
LADA	96–150	**8.6**	176.9	362
COMET	$\sim 145^{\dagger}$	12.2	**107.1**	**598**

M: retained experts; Stored/Active: million parameters; Lat.: milliseconds. Expand/Merge: Always-Expand/Always-Merge; †Top-1 path.

5.3 Ablation Study

We isolate the two distillation losses and the contextual-bandit merging policy on the X-TAIL 10-dataset protocol. Table 3 (a) shows that TAD and SACED are complementary: each improves Transfer, Average, and Last over the base model, and using both gives the strongest results in both full-shot and 16-shot settings. Sub-component decompositions of TAD and SACED, hyperparameter and reference-data sensitivity, and additional analyses are reported in Appendices D–I.

TAD and SACED. TAD alone improves Transfer by $+4.7$ pp ($56.6 \rightarrow 61.3$), Average by $+1.5$ pp, and Last by $+2.0$ pp over the no-distillation variant. SACED alone improves Transfer and Average

by $+3.3$ and $+1.2$ pp, respectively, but slightly reduces Last by 0.3 pp. Combining TAD and SACED yields the strongest results in this ablation: 62.6 Transfer, 76.1 Average, and 88.1 Last. In particular, adding SACED to TAD improves the three metrics by $+1.3$, $+0.7$, and $+0.5$ pp, respectively, supporting the benefit of their combination rather than uniform improvements in isolation.

Expert merging strategy. Holding the teacher, expert architecture, TAD, SACED, router, and merge operator fixed, we vary only the merge-or-keep decision rule. Table 3 (b) shows that CMAB retains 8 experts, compared with 10 for Always-Expand, while improving Transfer, Average, and Last by $+1.0$, $+0.3$, and $+0.8$ pp, respectively. Always-Expand stores more parameters but underperforms CMAB, indicating that simply adding experts does not solve routing and capacity allocation. Always-Merge and random merging degrade sharply, confirming that structured outcome-based merge decisions are necessary. The full seven-variant comparison and cross-sequence robustness analysis are reported in Appendix J.

Routing and robustness. Across the ten task-specific routers, the autoencoder router identifies the correct task for 95.9% of test inputs at top-1 and 99.4% at top-2. After merging, the router sends 2.6% of test samples ($2{,}069$ of $80{,}939$) away from their oracle expert, and SACED turns these errors into graceful degradation: 52.5% of misrouted samples are still correctly classified, and pairwise cross-expert representation similarity rises by CKA $+0.009$ relative to a no-SACED control (Appendix L). Across four task orderings on X-TAIL, Average is 72.8 ± 2.2; per-task forgetting, measured on the three re-ordered runs, is 2.2 ± 0.1 pp. The four key hyperparameters (λ_1, λ_2, the candidate gate θ_{sim}, and the initial merge threshold τ_1) are robust over an order of magnitude, and SACED requires only a few thousand diverse image–text reference pairs. Per-ordering results, hyperparameter sweeps, and reference-data ablations are reported in Appendices E, D, and F, respectively.

5.4 Analysis of Computational Cost

Table 3 (c) reports that COMET achieves the lowest latency and highest throughput because inference uses the teacher vision encoder, the lightweight autoencoder router, and only the selected TinyCLIP expert; the teacher text encoder is not executed. Although the retained pool after X-TAIL stores roughly 551M parameters, the active top-1 path uses only ~ 145M parameters, and the top-2 path remains below 203M. Training adds ~ 2.7 GB peak VRAM because only the current expert receives gradients. A detailed training and storage breakdown is in Appendix O. LwF, WiSE-FT, and ZSCL have identical inference cost because all deploy a single full-CLIP backbone; their differences are only in training objectives.

6 Conclusion

A central question for continual VLMs is how to keep adapting when tasks arrive without replay, task identity, or full-backbone retraining. We argue that this requires separating the roles that prior systems often couple. COMET is a step toward this direction: it uses a frozen teacher only as a multimodal geometry reference, TAD to preserve tri-affinity structure, SACED to reduce routing-induced expert mismatch, CMAB to regulate capacity after training, and reconstruction routing to select sparse experts without teacher semantics. Across X-TAIL and a heterogeneous medical, microscopy, and histopathology stream, this separation improves Average and Last accuracy, matches the best Transfer score, and keeps the retained expert pool compact.

Limitations. COMET's sub-linear storage behavior depends on diminishing cumulative novelty, such as revisits or bounded representation support. Longer streams that keep entering new domains may still grow the expert pool linearly, and we prove no regret guarantee for CMAB. Explicit storage budgets and capacity control with provable guarantees remain directions for future work.

Acknowledgments and Disclosure of Funding

This work was partly supported by the National IT Industry Promotion Agency (NIPA) grant funded by the Korea government (MSIT) (Development of Physical Intelligence Action Models and an Integrated Framework, No. RS-2026-25621677); the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (Fundamental Research on Continual Meta-Learning for Quality Enhancement of Casual Videos and Their 3D Metaverse Transformation, No. RS-2022-II220156; AI Star Fellowship Support Program (Seoul National University), No. RS-2025-25442338); and the IITP and Korea Evaluation Institute of Industrial Technology (KEIT) grants funded by the Korea government (MSIT and MOTIE, respectively) (Ultra-Long-Term Hierarchical Memory and Reasoning Architecture for Next-Generation Omnimodal Agents, No. RS-2026-25524173).

References

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandit algorithms with supervised learning guarantees. In Geoffrey J. Gordon, David B. Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, JMLR Proceedings, pages 19–26. JMLR.org, 2011.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - mining discriminative components with random forests. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI*, volume 8694 of *Lecture Notes in Computer Science*, pages 446–461. Springer, 2014.

I-Chun Chen, Hsu-Shen Liu, Wei-Fang Sun, Chen-Hao Chao, Yen-Chang Hsu, and Chun-Yi Lee. Retraining-free merging of sparse moe via hierarchical clustering. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net, 2025.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 3606–3613. IEEE Computer Society, 2014.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021.

Li Deng. The MNIST database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Process. Mag.*, 29(6):141–142, 2012.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2004, Washington, DC, USA, June 27 - July 2, 2004*, page 178. IEEE Computer Society, 2004.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3259–3269. PMLR, 2020.

Raia Hadsell, Dushyant Rao, Andrei A Rusu, and Razvan Pascanu. Embracing change: Continual learning in deep neural networks. *Trends in cognitive sciences*, 24(12):1028–1040, 2020.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 12(7):2217–2226, 2019.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015a.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015b.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Hiroaki Iwata, Yoshihiro Hayashi, Aki Hasegawa, Kei Terayama, and Yasushi Okuno. Classification of scanning electron microscope images of pharmaceutical excipients using deep convolutional neural networks with transfer learning. *International Journal of Pharmaceutics: X*, 4:100135, 2022. doi: 10.1016/j.ijpx.2022.100135.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 2021.

Jakob Nikolas Kather, Cleo-Aron Weis, Francesco Bianconi, Susanne M. Melchers, Lothar R. Schad, Timo Gaiser, Alexander Marx, and Frank Gerrit Zöllner. Multi-class texture analysis in colorectal cancer histology. *Scientific Reports*, 6:27988, 2016.

Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue, April 2018.

James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796, 2016.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013*, pages 554–561. IEEE Computer Society, 2013.

Ehsan Latif and Xiaoming Zhai. Efficient multi-task inferencing with a shared backbone and lightweight task-specific adapters for automatic scoring. *CoRR*, abs/2412.21065, 2024.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 9694–9705, 2021.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(12): 2935–2947, 2018.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Mao-Lin Luo, Zi-Hao Zhou, Tong Wei, and Min-Ling Zhang. LADA: scalable label-specific CLIP adapter for continual learning. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net, 2025.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, abs/1306.5151, 2013.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing, ICVGIP 2008, Bhubaneswar, India, 16-19 December 2008*, pages 722–729. IEEE Computer Society, 2008.

Keon-Hee Park, Kyungwoo Song, and Gyeong-Moon Park. Pre-trained vision and language transformers are few-shot incremental learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23881–23890, 2024.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 3498–3505. IEEE Computer Society, 2012.

Razvan Pascanu, Yann N. Dauphin, Surya Ganguli, and Yoshua Bengio. On the saddle point problem for non-convex optimization. *CoRR*, abs/1405.4604, 2014.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021a.

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

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5533–5542. IEEE Computer Society, 2017.

Benbakreti Samir, Said Mwanahija, Benbakreti Soumia, and Umut Özkaya. Deep learning for classification of chest x-ray images (covid 19). *CoRR*, abs/2301.02468, 2023.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017a.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017b.

Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A. Alemi, and Andrew Gordon Wilson. Does knowledge distillation really work? In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 6906–6919, 2021.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 7949–7961. IEEE, 2022.

Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi Stephen Chen, Xinggang Wang, Hongyang Chao, and Han Hu. Tinyclip: CLIP distillation via affinity mimicking and weight inheritance. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 21913–21923. IEEE, 2023.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 3485–3492. IEEE Computer Society, 2010.

Yicheng Xu, Yuxin Chen, Jiahao Nie, Yusong Wang, Huiping Zhuang, and Manabu Okumura. Advancing cross-domain discriminability in continual learning of vision-language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 23219–23230. IEEE, 2024a.

Yu-Chu Yu, Chi-Pin Huang, Jr-Jen Chen, Kai-Po Chang, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. Select and distill: Selective dual-teacher knowledge transfer for continual learning on vision-language models. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXVI*, volume 15084 of *Lecture Notes in Computer Science*, pages 219–236. Springer, 2024b.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017a.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017b.

Zangwei Zheng, Mingyuan Ma, Kai Wang, Ziheng Qin, Xiangyu Yue, and Yang You. Preventing zero-shot transfer degradation in continual learning of vision-language models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 19068–19079. IEEE, 2023.

A Sub-Linear Expert Growth: Formal Statement and Proof

This appendix provides the formal scalability argument summarized in Eq. (4). The result is a sufficient-condition guarantee: it identifies regimes in which COMET's expert pool grows sub-linearly with stream length, but it does not claim sub-linear growth for arbitrary adversarial streams. The key quantity is the frequency with which incoming chunks are novel relative to the current expert pool in a fixed teacher-induced representation.

A.1 Setup

Stream and expert pool. Let $\{\mathcal{D}_k\}_{k=1}^{K}$ be a sequence of data chunks arriving online, one per task $\mathcal{T}_k$. No i.i.d. assumption is made on the chunks. Let M_k denote the number of retained experts after processing chunks $1{:}k$, with $M_0 = 0$ because Algorithm 1 starts from an empty pool. The stream and algorithmic decisions may be random; all probabilities and expectations are taken over the joint randomness of the stream, the data samples, and the merge-or-keep policy.

Teacher-induced representation. Let $\psi(\mathcal{D}_k) \in \mathbb{R}^d$ be a fixed summary of chunk $\mathcal{D}_k$ in the frozen teacher feature space. For example, one may use the mean frozen-teacher image embedding

$$\psi(\mathcal{D}_k) = \frac{1}{|\mathcal{D}_k|} \sum_{x \in \mathcal{D}_k} \Phi_v(x),$$

where Φ_v is the frozen CLIP image encoder. For every retained expert E_m, let $\psi(E_m)$ be its fixed representative, such as the teacher-feature centroid of the data on which it was originally trained.

The analysis assumes a genuine metric $d(\cdot, \cdot)$ on this representation space, because the finite-cover proof uses the triangle inequality. Valid choices include Euclidean distance on ℓ_2-normalized embeddings or angular distance $d_{\text{ang}}(u, v) = \arccos(\langle u, v \rangle)$ on the unit sphere. The commonly used cosine dissimilarity $1 - \cos(u, v)$ is useful as a similarity score, but it is not a formal metric and is therefore not used in the proof.

Novelty. For chunk k, define its distance to the current expert pool as

$$\delta_k = \min_{m \in \{1, \ldots, M_{k-1}\}} d(\psi(\mathcal{D}_k), \psi(E_m)) \,. \tag{7}$$

We set $\delta_1 = +\infty$ because the pool is empty when the first chunk arrives, so the first chunk is always novel. A chunk is called ϵ-*novel* if $\delta_k > \epsilon$. Let

$$\mathbb{I}_k = \mathbf{1}\{\delta_k > \epsilon\}$$

be the corresponding novelty indicator. We use little-o notation in the standard sense: $g(K) = o(K)$ means $g(K)/K \to 0$ as $K \to \infty$.

Merge-or-keep assumptions. The core bound applies to any allocation rule satisfying two structural properties:

(P1) the algorithm adds at most one new expert per chunk;

(P2) the algorithm expands only on ϵ-novel chunks, i.e., $M_k > M_{k-1}$ implies $\delta_k > \epsilon$.

These assumptions isolate the role of stream novelty from the details of the merge policy. Later, in Appendix A.5, we relax (P2) to account for practical false expansions caused by an imperfect learned policy.

A.2 Step 1: Expert Growth Is Controlled by Novelty

Under (P1)–(P2), the expert count satisfies

$$M_k \leq M_{k-1} + \mathbb{I}_k. \tag{8}$$

This is deterministic after the stream and decisions are realized. If $\mathbb{I}_k = 0$, then the chunk is not ϵ-novel, so (P2) forbids expansion and $M_k = M_{k-1}$. If $\mathbb{I}_k = 1$, then the chunk is ϵ-novel, and (P1) permits at most one new expert.

Unrolling Eq. (8) gives

$$M_K \leq M_0 + \sum_{k=1}^{K} \mathbb{I}_k = \sum_{k=1}^{K} \mathbb{I}_k. \tag{9}$$

Taking expectations yields

$$\mathbb{E}[M_K] \leq \sum_{k=1}^{K} \mathbb{E}[\mathbb{I}_k] = \sum_{k=1}^{K} \Pr(\delta_k > \epsilon). \tag{10}$$

Equation (10) is the bound reported in the main text as Eq. (4).

A.3 Step 2: Sub-Linear Growth Under Diminishing Novelty

Theorem A.1 (Sub-linear expert growth)**.** *Under (P1)–(P2), if the cumulative novelty satisfies*

$$\sum_{k=1}^{K} \Pr(\delta_k > \epsilon) = o(K), \tag{11}$$

then $\mathbb{E}[M_K] = o(K)$.

Proof. Combining Eq. (10) with Eq. (11),

$$\frac{\mathbb{E}[M_K]}{K} \leq \frac{1}{K} \sum_{k=1}^{K} \Pr(\delta_k > \epsilon) \to 0$$

as $K \to \infty$. Thus, $\mathbb{E}[M_K] = o(K)$. □

Interpretation. The condition in Eq. (11) does not require i.i.d. data. It requires only that, relative to the existing expert pool, genuinely new chunks become rare on average. This holds in several common regimes.

First, if the stream revisits a finite catalogue of underlying distributions, then once one expert has been created for each ϵ-separated region, future chunks are no longer novel. Second, if the set of possible chunk summaries has a finite $(\epsilon/2)$-cover under d, Proposition A.2 gives a deterministic finite bound on the number of expansions. Third, if the stream drifts with sub-linear cumulative path length, e.g., $\sum_{k=1}^{K-1} d(z_{k+1}, z_k) = o(K)$ for $z_k = \psi(\mathcal{D}_k)$, then the number of ϵ-separated exits along the trajectory is also sub-linear.

A constant bounded step size alone is not sufficient in an unbounded space. For example, a stream that moves in a straight line with fixed step size $\beta > 0$ can leave an ϵ-ball every $O(\epsilon/\beta)$ steps and therefore force linear growth. Thus, the sufficient conditions are revisitation, bounded representation support, or asymptotically slowing novelty, not arbitrary bounded local drift.

A.4 Step 3: Finite-Cover Special Case

Proposition A.2 (Finite-cover bound)**.** *Suppose the set of possible chunk summaries* $\mathcal{S} = \{\psi(\mathcal{D}_k) : k \geq 1\}$ *admits a finite* $(\epsilon/2)$*-cover of size* $C_{\epsilon/2}$ *under the metric* d*. That is, there exist anchors* $\{q_j\}_{j=1}^{C_{\epsilon/2}}$ *such that for every* $z \in \mathcal{S}$ *there exists* j *with* $d(z, q_j) \leq \epsilon/2$*. Suppose further that whenever the algorithm creates a new expert on chunk* k*, its representative is set to the chunk summary,* $\psi(E_{\text{new}}) = \psi(\mathcal{D}_k)$*. Then under (P1)–(P2), the total expert count satisfies*

$$M_K \leq C_{\epsilon/2} \quad \textit{for all } K. \tag{12}$$

Proof. Assign each expert E_m, created at step k_m, to any cover index $\phi(E_m) = j$ satisfying $d(\psi(\mathcal{D}_{k_m}), q_j) \leq \epsilon/2$. Such a j exists by the cover assumption.

We show that ϕ is injective. Suppose two distinct experts E_m and $E_{m'}$ with creation times $k_m < k_{m'}$ are assigned to the same cover anchor q_j. Because $\psi(E_m) = \psi(\mathcal{D}_{k_m})$ and both summaries lie within $\epsilon/2$ of q_j, the triangle inequality gives

$$d\big(\psi(\mathcal{D}_{k_{m'}}), \psi(E_m)\big) \leq d\big(\psi(\mathcal{D}_{k_{m'}}), q_j\big) + d(q_j, \psi(E_m)) \leq \epsilon.$$

Since E_m already exists when chunk $k_{m'}$ arrives, this implies $\delta_{k_{m'}} \leq \epsilon$. But (P2) says that a new expert can be created only when $\delta_{k_{m'}} > \epsilon$, a contradiction. Thus, at most one expert is assigned to each cover element. Because $M_0 = 0$, every retained expert was created this way, and Eq. (12) follows. □

A.5 Step 4: Practical Decision Errors

A learned merge-or-keep policy may occasionally keep a new expert on a non-novel chunk. Define the false-expand indicator

$$e_k = \mathbf{1}\{M_k > M_{k-1} \text{ and } \delta_k \leq \epsilon\}.$$

Without exact (P2), Eq. (8) becomes

$$M_k \leq M_{k-1} + \mathbb{I}_k + e_k.$$

Unrolling and taking expectations gives

$$\mathbb{E}[M_K] \leq \sum_{k=1}^{K} \Pr(\delta_k > \epsilon) + \sum_{k=1}^{K} \Pr(e_k = 1). \tag{13}$$

Corollary A.3 (Sub-linear growth with sub-linear false expansions)**.** *If*

$$\sum_{k=1}^{K} \Pr(\delta_k > \epsilon) = o(K) \quad \textit{and} \quad \sum_{k=1}^{K} \Pr(e_k = 1) = o(K),$$

then $\mathbb{E}[M_K] = o(K)$.

Proof. Divide Eq. (13) by K and let $K \to \infty$. □

From regret to false-expansion count. A no-regret bandit bound controls cumulative expected reward gap, not directly the number of wrong actions. To translate regret into a count of false expansions, a margin condition is required.

Let a_k be the action selected at step k, let $a_k^\star$ be a population-optimal action for the same context, and let

$$\Delta_k(a) = \bar{r}_k(a_k^\star) - \bar{r}_k(a) \geq 0$$

be the conditional expected reward gap of action a. Define cumulative regret

$$\mathrm{Reg}_K = \sum_{k=1}^{K} \mathbb{E}[\Delta_k(a_k)]\,.$$

Assume a false-expansion margin:

$$\exists\, \Delta_{\min} > 0 \quad \text{such that} \quad \delta_k \leq \epsilon \text{ and } a_k = \textsc{Keep} \implies \Delta_k(a_k) \geq \Delta_{\min}. \tag{14}$$

Then

$$\Pr(e_k = 1) = \mathbb{E}[e_k] \leq \frac{1}{\Delta_{\min}} \mathbb{E}[\Delta_k(a_k)]\,,$$

and therefore

$$\sum_{k=1}^{K} \Pr(e_k = 1) \leq \frac{\mathrm{Reg}_K}{\Delta_{\min}}. \tag{15}$$

Thus, if $\mathrm{Reg}_K = o(K)$ and Eq. (14) holds, the false-expansion term is sub-linear. Without the margin, a policy may take $O(K)$ suboptimal actions while still having sub-linear regret if the corresponding gaps vanish over time.

Applicability to CMAB. CMAB (Section 4.4) never computes δ_k: it keeps a new expert when no candidate passes the similarity gate or when the best bandit score stays below the threshold. (P2) therefore holds only approximately, and each keep on a non-novel chunk counts as a false expansion e_k. CMAB updates its weights with a projected, EMA-smoothed advantage step, which is not a standard no-regret algorithm, and we do not prove a regret bound for it. Eq. (15) thus names the property CMAB would need for a sub-linear false-expansion term; proving it is left to future work.

Scope. The proof characterizes when expert growth is controlled by stream novelty and decision-error rate. It does not claim sub-linear growth for streams that continually move into new regions of an unbounded representation space. Empirically, COMET's retained expert count saturates at 5–7 experts across X-TAIL orderings, consistent with the sufficient-condition analysis in regimes with revisits, bounded representation support, or decaying novelty.

B Algorithmic Details

Algorithm 1 summarizes the full training and inference procedure. The pseudocode makes three implementation choices explicit. First, the router is trained before expert optimization, so routing is tied to teacher-feature reconstruction rather than class semantics. Second, expert training uses the task loss together with TAD and SACED, while the CMAB capacity controller is invoked only after the expert has been trained. Third, the bandit is updated for both MERGE and KEEP outcomes whenever a candidate set is evaluated.

Algorithm 1 COMET: Collaborative Mixture-of-Expert Transfer

Require: Frozen Teacher $\Phi = (\Phi_v, \Phi_t)$; stream $\{\mathcal{D}_k\}_{k=1}^{K}$; labeled validation splits $\{\mathcal{V}_k\}_{k=1}^{K}$; reference pool $\mathcal{D}_{\mathrm{ref}}$; CMAB state $\pi = (w, \tau, \beta, \bar{r}, N)$
Hyperparameters: $N_R, N_E, \lambda_1, \lambda_2, k_{\mathrm{ref}}, \theta_{\mathrm{sim}}, w_{\max}, \eta_w, \eta_\tau$
Ensure: Retained expert pool $\{E_m\}_{m=1}^{M}$ and router pool $\{R_m\}_{m=1}^{M}$

$M \leftarrow 0$
for $k = 1$ to K **do**
  Initialize new expert E_{new} and router R_{new}
  Router training
  for $n = 1$ to N_R **do**
    Sample image batch $I \sim \mathcal{D}_k$
    $z \leftarrow \textsc{stopgrad}(\Phi_v(I))$
    $\mathcal{L}_R \leftarrow d^{-1}\|R_{\mathrm{new}}(z) - z\|_2^2$
    Update R_{new}
  end for
  Reference-expert selection for SACED
  if $M > 0$ **then**
    Sample $(I^r, T^r) \sim \mathcal{D}_{\mathrm{ref}}$
    $\mathcal{S}_k \leftarrow \textsc{SelectRefs}(E_{\mathrm{new}}, \{E_m\}_{m=1}^{M}, \Phi, I^r, T^r, k_{\mathrm{ref}})$
  else
    $\mathcal{S}_k \leftarrow \emptyset$
  end if
  Expert training
  for $n = 1$ to N_E **do**
    Sample task batch $(I, T) \sim \mathcal{D}_k$ and reference batch $(I^r, T^r) \sim \mathcal{D}_{\mathrm{ref}}$
    $\mathcal{L}_{\mathrm{task}} \leftarrow \textsc{TaskLoss}(E_{\mathrm{new}}, I, T)$
    $\mathcal{L}_{\mathrm{TAD}} \leftarrow \mathrm{TAD}(E_{\mathrm{new}}, \Phi, I^r, T^r)$
    if $\mathcal{S}_k = \emptyset$ **then**
      $\mathcal{L}_{\mathrm{SACED}} \leftarrow 0$
    else
      $\mathcal{L}_{\mathrm{SACED}} \leftarrow \mathrm{SACED}(E_{\mathrm{new}}, \mathcal{S}_k, \Phi, I^r)$
    end if
    $\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{task}} + \mathcal{L}_{\mathrm{TAD}} + \lambda_2 \mathcal{L}_{\mathrm{SACED}}$
    Update E_{new} using $\nabla \mathcal{L}$
  end for
  Merge-or-Keep via CMAB:
  if $M = 0$ **then**
    $M \leftarrow 1$; $E_1 \leftarrow E_{\mathrm{new}}$; $R_1 \leftarrow R_{\mathrm{new}}$
  else
    $\mathcal{C}_k \leftarrow \textsc{Candidates}(E_{\mathrm{new}}, \{E_m\}_{m=1}^{M}, \mathcal{V}_k, \mathcal{D}_{\mathrm{ref}})$
    if $\mathcal{C}_k = \emptyset$ **then**
      $M \leftarrow M + 1$; $E_M \leftarrow E_{\mathrm{new}}$; $R_M \leftarrow R_{\mathrm{new}}$
    else
      for $c \in \mathcal{C}_k$ **do**
        $\phi_{k,c} \leftarrow \textsc{Context}(E_{\mathrm{new}}, E_c, \mathcal{V}_k, \mathcal{D}_{\mathrm{ref}})$
        $B_k(c) \leftarrow w^\top \phi_{k,c} + \beta/(1+N)$
      end for
      $c^\star \leftarrow \arg\max_{c \in \mathcal{C}_k} B_k(c)$
      if $B_k(c^\star) > \tau$ **then**
        $\widetilde{E}_{c^\star} \leftarrow \textsc{Merge}(E_{c^\star}, E_{\mathrm{new}})$
        $r \leftarrow \textsc{MergeReward}(\widetilde{E}_{c^\star}, \mathcal{V}_k)$
        $E_{c^\star} \leftarrow \widetilde{E}_{c^\star}$
        $R_{c^\star} \leftarrow \textsc{MergeRouter}(R_{c^\star}, R_{\mathrm{new}})$
        $a \leftarrow \textsc{Merge}$
      else
        $M \leftarrow M + 1$; $E_M \leftarrow E_{\mathrm{new}}$; $R_M \leftarrow R_{\mathrm{new}}$
        $r \leftarrow \textsc{KeepReward}(E_{\mathrm{new}}, \mathcal{V}_k)$
        $a \leftarrow \textsc{Keep}$
      end if
      $\pi \leftarrow \textsc{UpdateBandit}(\pi, \phi_{k,c^\star}, a, r)$
    end if
  end if
end for

Inference (task-free):
$z \leftarrow \Phi_v(x)$
for $m = 1$ to M **do**
  $r_m(x) \leftarrow d^{-1}\|R_m(z) - z\|_2^2$
  $\ell_m(x) \leftarrow \log p(\log r_m(x) \mid m) - \log p(\log r_m(x) \mid \mathrm{bg})$
end for
$w_m(x) \leftarrow \mathrm{softmax}(\ell_m(x)/\tau_g)_m$
$\mathcal{A}(x) \leftarrow \textsc{TopK}(w(x), k)$ with $k \leq 2$
$\hat{y} \leftarrow \arg\max_{c \in \mathcal{C}_N} \sum_{m \in \mathcal{A}(x)} w_m(x)\, \cos(E_{m,v}(x), E_{m,t}(c))$

C Implementation Details

Architecture. The teacher is the frozen CLIP ViT-B/16 [Radford et al., 2021a], with approximately 87M parameters in the image encoder and 63M parameters in the text encoder. Each expert is a TinyCLIP backbone [Wu et al., 2023], with approximately 39M parameters in the image encoder and 19M parameters in the text encoder. Each task also receives a lightweight autoencoder router with a two-layer encoder and decoder operating in the frozen teacher's image-feature space. The router adds approximately 0.2M parameters per retained expert.

Training schedule. For each task, COMET runs three stages. First, the autoencoder router is trained for 1,000 iterations using normalized reconstruction loss on frozen teacher image features. Second, the new expert is trained for 1,000 iterations with AdamW [Loshchilov and Hutter, 2019], learning rate 10^{-5}, and batch size 64. The training objective is the per-task loss of the main text,

$$\mathcal{L}_k = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{TAD}} + \lambda_2 \mathcal{L}_{\text{SACED}},$$

with λ_1 weighting the cross-modal term inside $\mathcal{L}_{\text{TAD}}$. Third, after expert training, the CMAB capacity controller decides whether the new expert is kept or merged. Retained experts and routers receive no further gradient updates; CMAB changes an expert and its router only when it merges a later expert into them (Algorithm 1).

Loss weights. Inside TAD, the cross-modal weight is $\lambda_1 = 0.2$. The SACED weight is $\lambda_2 = 1.0$. Sensitivity sweeps in Appendix D show that the method is stable over an order of magnitude around these defaults.

Reference and validation signals. The reference pool $\mathcal{D}_{\text{ref}}$ contains 10,800 image–text pairs from the Conceptual Captions validation set [Sharma et al., 2018]. It is used for label-free distillation, feature-centroid estimates, routing agreement, and distribution-level compatibility. Because Conceptual Captions does not provide task classification labels, $\mathcal{D}_{\text{ref}}$ is not used to compute classification accuracy. All classification accuracy, cross-performance, and CMAB reward signals are computed on the labeled held-out split $\mathcal{V}_k$ of the current task.

C.1 CMAB Capacity Controller Details

Candidate generation. After training E_{new}, COMET compares it with each retained expert $E_c \in \mathcal{E}_{<k}$. The candidate compatibility score is

$$\text{sim}(E_{\text{new}}, E_c) = \frac{w_{\text{cross}}\text{CrossPerf}(E_{\text{new}}, E_c) + w_{\text{cent}}\text{CentSim}(E_{\text{new}}, E_c) + w_{\text{gate}}\text{GateCorr}(E_{\text{new}}, E_c)}{w_{\text{cross}} + w_{\text{cent}} + w_{\text{gate}}}. \tag{16}$$

CrossPerf evaluates both experts with the current-task label prompts on $\mathcal{V}_k$ and averages their validation accuracies. CentSim measures cosine similarity between image-feature centroids. GateCorr measures correlation between router/gating responses over a shared unlabeled batch. Experts satisfying $\text{sim}(E_{\text{new}}, E_c) > \theta_{\text{sim}}$ enter the candidate set $\mathcal{C}_k$. If $\mathcal{C}_k = \emptyset$, the new expert is kept and no bandit update is applied, because no valid merge action was evaluated.

Context construction. For each candidate pair (E_{new}, E_c), COMET constructs

$$\phi_{k,c} = [\text{Acc}_k,\ m_{k,c},\ -d_{k,c},\ \kappa_k,\ -\ell_c]^\top. \tag{17}$$

Here Acc_k is the new expert's validation accuracy on $\mathcal{V}_k$. The merge-compatibility feature is

$$m_{k,c} = 0.4 s_{k,c} + 0.3 o_{k,c} + 0.3 q_{k,c},$$

where $s_{k,c}$ is the candidate similarity score, $o_{k,c}$ is routing overlap, and $q_{k,c}$ is angle compatibility for the merge operator. The term $d_{k,c}$ estimates degradation risk, κ_k increases with the current pool size, and ℓ_c increases with the number of tasks already absorbed by candidate expert E_c. The signs in Eq. (17) encode the desired decision direction: high accuracy, high compatibility, and high capacity pressure favor merging, while degradation risk and overloaded targets discourage it.

Bandit score and action. Each candidate receives

$$B_k(c) = w^\top \phi_{k,c} + \frac{\beta_k}{1 + N_k}, \tag{18}$$

where $w \in [0, w_{\max}]^5$ is the learned scoring vector, β_k is a decaying exploration bonus, and N_k is the number of previous updates. Let $c^\star = \arg\max_{c \in \mathcal{C}_k} B_k(c)$. The action is

$$a_k = \begin{cases} \textsc{Merge}(E_{\text{new}}, E_{c^\star}), & B_k(c^\star) > \tau_k, \\ \textsc{Keep}(E_{\text{new}}), & B_k(c^\star) \le \tau_k. \end{cases}$$

Thus, the decision is a learned trade-off among compatibility, risk, capacity pressure, and expert load rather than a fixed similarity threshold.

Reward. For a merge action, the merged expert is constructed temporarily and evaluated on $\mathcal{V}_k$:

$$r_k^{\text{merge}} = \text{Acc}_{\mathcal{V}_k}(\text{Merge}(E_{\text{new}}, E_{c^\star})) - \lambda_{\text{merge}}. \tag{19}$$

For a keep action,

$$r_k^{\text{keep}} = \text{Acc}_{\mathcal{V}_k}(E_{\text{new}}).$$

The small merge cost λ_{merge} discourages merges that do not produce a meaningful validation benefit. Both MERGE and KEEP actions update the policy whenever a candidate set is evaluated.

Advantage-centered update. Because validation accuracy is positive-valued, updating with raw reward and fixed-sign features can cause monotone weight drift. COMET therefore uses a centered advantage. Let $\bar{r}_k$ be an exponential moving average of past rewards:

$$A_k = r_k - \bar{r}_k, \qquad \bar{r}_{k+1} = (1-\rho_r)\bar{r}_k + \rho_r r_k.$$

Let $y_k = +1$ for MERGE and $y_k = -1$ for KEEP. For the selected candidate context $\phi_{k,c^\star}$,

$$\tilde{w} = \Pi_{[0,w_{\max}]}\left(w + \eta_w A_k y_k \phi_{k,c^\star}\right), \qquad w \leftarrow (1-\rho_w)w + \rho_w \tilde{w}. \tag{20}$$

This update is action-aware. Successful merges increase future merge scores for similar contexts, failed merges decrease them, successful keeps make merging less likely in similar contexts, and poor keeps make merging more likely.

Adaptive threshold. The merge threshold is updated with the same advantage signal:

$$\tilde{\tau} = \Pi_{[\tau_{\min},\tau_{\max}]}\left(\tau_k - \eta_\tau A_k y_k\right), \qquad \tau_{k+1} = (1-\rho_\tau)\tau_k + \rho_\tau \tilde{\tau}. \tag{21}$$

Successful merges lower the threshold, failed merges raise it, successful keeps raise it, and suboptimal keeps lower it. The exploration bonus decays as $\beta_{k+1} = \gamma\beta_k$.

CMAB hyperparameters. The similarity gate is $\theta_{\text{sim}} = 0.40$. The merge threshold is initialized as $\tau_1 = 0.65$. We use $\eta_w = 0.05$, $\eta_\tau = 0.05$, $\rho_r = \rho_w = \rho_\tau = 0.10$, $w_{\max} = 10$, $\gamma = 0.95$, and $\lambda_{\text{merge}} = 0.01$. These values are used for all reported experiments unless explicitly varied in Appendix D.

C.2 Inference Routing Details

Router training. For each retained expert E_m, COMET stores a corresponding autoencoder router R_m. The router is trained on frozen CLIP image features from the task that produced the expert. Given $z = \Phi_v(x)$, the training objective is

$$\mathcal{L}_{R_m} = \mathbb{E}_{x\sim\mathcal{D}_m}\left[\frac{1}{d}\|R_m(z) - z\|_2^2\right].$$

Normalizing by feature dimension keeps the reconstruction scale stable across model sizes.

Likelihood-ratio calibration. Raw reconstruction errors are not directly comparable across routers because different routers may have different error scales. COMET therefore calibrates each router using foreground and background error distributions. For router R_m, define

$$r_m(x) = \frac{1}{d}\|R_m(\Phi_v(x)) - \Phi_v(x)\|_2^2, \qquad \tilde{r}_m(x) = \log r_m(x).$$

The foreground distribution $p_m^{\text{fg}}(\tilde{r})$ is estimated from task data associated with expert E_m. The background distribution $p_m^{\text{bg}}(\tilde{r})$ is estimated by passing $\mathcal{D}_{\text{ref}}$ through the same router. At test time, the routing score is the log-likelihood ratio

$$\ell_m(x) = \log p_m^{\text{fg}}(\tilde{r}_m(x)) - \log p_m^{\text{bg}}(\tilde{r}_m(x)). \tag{22}$$

In practice, the distributions are fitted as one-dimensional Gaussians over log reconstruction error, although the method only requires calibrated density estimates.

Top-k activation. The calibrated scores are converted into gate weights by

$$w_m(x) = \text{softmax}(\ell_m(x)/\tau_g)_m.$$

COMET activates the top-1 expert when the margin between the largest and second-largest gate scores exceeds a calibrated threshold; otherwise it activates the top-2 experts. The selected experts compute CLIP-style image–text similarity logits over the candidate label set, and their logits are combined by gate-weighted averaging. This routing mechanism uses no task identity and no teacher semantic prediction at inference.

Role separation. The router, CMAB, and distillation objectives act at different stages. TAD and SACED shape the expert during training. CMAB decides the retained pool after training. The reconstruction router selects the active expert at inference. This separation prevents the teacher from becoming a semantic oracle at test time and keeps routing independent of class-label predictions.

Table 4: Task-order sensitivity on X-TAIL. "Forgetting" is per-task degradation from peak accuracy, averaged over the ten tasks. Forgetting is not available for the default ordering, so its mean is over the three re-ordered runs.

Ordering	Avg	Last	M	Forgetting
Default	76.1	88.1	8	—
EuroSAT-last	72.0	83.3	5	2.3
Random seed 7	71.9	81.0	5	2.2
Random seed 99	71.3	83.8	5	2.1
Mean $\pm$ std	72.8 ± 2.2	84.1 ± 3.0	5.8 ± 1.5	2.2 ± 0.1

Hardware and reporting. All main-text results and appendix measurements are obtained on a single NVIDIA A6000 GPU with 48 GB memory. Latency is measured using CUDA event timing with batch size 64, 50 measured batches, and 5 warmup batches. FLOPs are analytical estimates per sample.

D Hyperparameter Sensitivity

We sweep each of the four key hyperparameters one at a time on X-TAIL with all others held at their default values.

Cross-modal TAD weight λ_1. Across $\lambda_1 \in \{0.2, 0.5, 1.0, 2.0\}$, Average ranges over 74.2–76.1 and Last over 86.6–88.1, all within 2.7 pp of the default.

SACED weight λ_2. Across $\lambda_2 \in \{0.01, 0.5, 1.0, 2.0\}$, Average remains within 74.1–76.2. Even near-zero SACED ($\lambda_2 = 0.01$) preserves TAD's gain, so the two losses are largely separable in practice.

Candidate similarity gate θ_{sim}. Lowering θ_{sim} admits more candidate merge actions into the CMAB decision stage. At the most permissive setting, the pool collapses to $M = 1$ with Average 58.6 and Last 59.2, a -28.9 pp Last drop relative to the default. The gate controls which experts are eligible for consideration, while CMAB learns the final merge-or-keep decision from context features and outcome feedback.

Initial CMAB decision threshold τ_1. Across $\tau_1 \in \{0.20, 0.65, 1.00\}$, Average remains within 74.1–76.2 and Last within 86.1–89.3. The conservative setting $\tau_1 = 0.80$ rejects most candidate merges and reduces Last to 79.8 at $M = 6$, marking the practical upper bound on conservative initialization. During training, the threshold is further adapted online by the advantage-based update in Appendix C.1.

E Task-Order Sensitivity

We re-run X-TAIL under three additional task orderings: *EuroSAT-last* (placing the most dissimilar domain at the end of the stream) and two random permutations (seeds 7 and 99). Table 4 summarizes Average, Last, the final number of retained experts M, and per-task forgetting. Average is 72.8 ± 2.2 across the four orderings, and per-task forgetting is 2.2 ± 0.1 pp across the three re-ordered runs. Frozen experts structurally bound forgetting regardless of order; the contextual bandit adapts the final pool size ($M \in \{5, 6, 7, 8\}$) to the task-similarity structure of each ordering, which explains the moderate Last variance (84.1 ± 3.0). Single-model continual baselines typically swing by 10–15 pp under permutation, so COMET's ± 3.0 Last variance is small.

F Reference-Data Sensitivity

SACED's reference pool is a small slice of Conceptual Captions; we vary both its size and its source domain. On X-TAIL, Average degrades smoothly with size: 1K$\rightarrow$ 68.4, 5K$\rightarrow$ 68.1, 10.8K$\rightarrow$ 76.1, indicating diminishing returns past a few thousand pairs. Substituting ImageNet validation images for Conceptual Captions at matched size drops Average to 68.6 on X-TAIL and to 48.3 on the heterogeneous benchmark. The second drop is the more informative comparison: on far-OOD domains, web-style image–text diversity matters more than scale, so any diverse image–text pool of a few thousand examples is sufficient and curated single-domain alternatives can hurt.

G TAD Sub-Component Decomposition

We decompose TAD into its three constituent losses on the X-TAIL 10-dataset protocol with SACED disabled to isolate TAD's contribution. Table 5 reports leave-one-out and one-only variants. Image affinity $\mathcal{L}_I$ is the

Table 5: TAD sub-component ablation on X-TAIL. "Full TAD" uses all three losses; "Full COMET" adds SACED. SACED is disabled for the sub-component rows.

Variant	$\mathcal{L}_I$	$\mathcal{L}_T$	$\mathcal{L}_{\text{Cross}}$	Avg	Last
Full COMET (TAD+SACED)	✓	✓	✓	**76.1**	**88.1**
Full TAD	✓	✓	✓	75.4	87.6
w/o $\mathcal{L}_I$	×	✓	✓	60.4	63.5
w/o $\mathcal{L}_T$	✓	×	✓	64.5	76.6
w/o $\mathcal{L}_{\text{Cross}}$	✓	✓	×	68.8	80.7
$\mathcal{L}_I$ only	✓	×	×	68.2	78.2
$\mathcal{L}_T$ only	×	✓	×	64.8	75.1
$\mathcal{L}_{\text{Cross}}$ only	×	×	✓	59.9	74.5

strongest single component: trained alone it reaches Average 68.2 and Last 78.2, and its removal causes the largest leave-one-out drop (60.4 and 63.5). Cross-modal affinity $\mathcal{L}_{\text{Cross}}$ is the weakest in isolation but contributes a complementary $+6.6$ pp Average over the two intra-modal losses (no-$\mathcal{L}_{\text{Cross}}$ 68.8 vs. Full TAD 75.4), so all three channels are required for the strongest result. Every partial variant also falls below the no-distillation baseline of Table 3 (a) (73.9 Average, 85.6 Last), while the full objective improves on it. We conjecture that matching only a subset of the affinity relations pins part of the CLIP geometry and lets the unmatched relations absorb the task-driven drift, which distorts the joint image–text structure more than leaving all relations free. Under this reading, the three channels act as one constraint, and the leave-one-out drops measure interactions among channels instead of each channel's additive share.

H Per-Input Drift Analysis of Tri-Affinity Components

The aggregate ablation in Appendix G establishes that all three TAD channels are required for the strongest result; this section probes the same effect at the level of individual inputs to verify that the gain is broadly distributed rather than carried by a few outliers. We compare two variants: *expA*, which enables only the cross-modal affinity distillation loss $\mathcal{L}_{\text{Cross}}$, and *expB*, which enables the full Tri-Affinity objective $\mathcal{L}_I + \mathcal{L}_T + \mathcal{L}_{\text{Cross}}$.

Metric. For a task dataset, we compute the representation of each image x using the model before and after adapting on that task, denoted $h_{\text{before}}(x)$ and $h_{\text{after}}(x)$. We measure per-image drift by cosine shift,

$$\Delta(x) \;=\; 1 - \cos\big(h_{\text{before}}(x),\, h_{\text{after}}(x)\big), \tag{23}$$

and report the distribution of $\Delta(x)$ over task samples using the mean and the 90th percentile ($p90$). We compute $\Delta(x)$ on both (i) raw image features and (ii) their ℓ_2-normalized embeddings.

Results. Figure 2 shows the two tasks with the largest relative reduction in mean shift (SUN397 and Caltech101). In both cases, the expB (blue) histogram is shifted left of expA (red), indicating smaller per-image drift under the full Tri-Affinity objective. On SUN397 Xiao et al. [2010] (Task 09), expB reduces the mean shift from 0.225 to 0.118 (47.5% relative reduction) and reduces $p90$ from 0.311 to 0.164. On Caltech101 Fei-Fei et al. [2004] (Task 01), the mean shift decreases from 0.239 to 0.145 (39.3% relative reduction) and $p90$ decreases from 0.323 to 0.224. Dashed vertical lines mark the mean and dotted lines mark $p90$; both moving left indicates that the improvement is not driven only by a few outliers but reflects reduced drift across a broad portion of samples. The embedding and feature views appear similar because the metric is cosine-based and embeddings are normalized versions of features.

I SACED Selection Criteria

SACED scores each prior expert on a three-signal combination: co-activation overlap (weight 0.5), top-1 prediction agreement (0.3), and class-distribution similarity (0.2), and keeps the top k_{ref} candidates above the threshold. Table 6 compares each criterion in isolation against using all prior experts and against random selection, with all rows trained under the full COMET pipeline so they are directly comparable to full COMET (76.1 Average, 88.1 Last). Co-activation alone matches full COMET on Last (88.4 vs. 88.1) and beats random selection by 0.7 pp, which makes co-activation the dominant signal; the full three-signal criterion adds 1.3 pp Average (76.1 vs. 74.8). Distribution similarity alone underperforms because the signal becomes informative only after co-activation has narrowed the candidate set. On Last, random and all-prior selection both underperform co-activation, so selectivity matters: not every prior expert contributes useful supervision.

J Extended Expert-Merging Strategy Comparison

Table 7 reports the full seven-variant comparison underlying Table 3 (b) in the main text. The conservative fixed threshold $\sigma = 0.9$ retains 9 experts and edges CMAB by 1.1 pp Average (77.2 vs. 76.1) but pays -3.2 pp Last

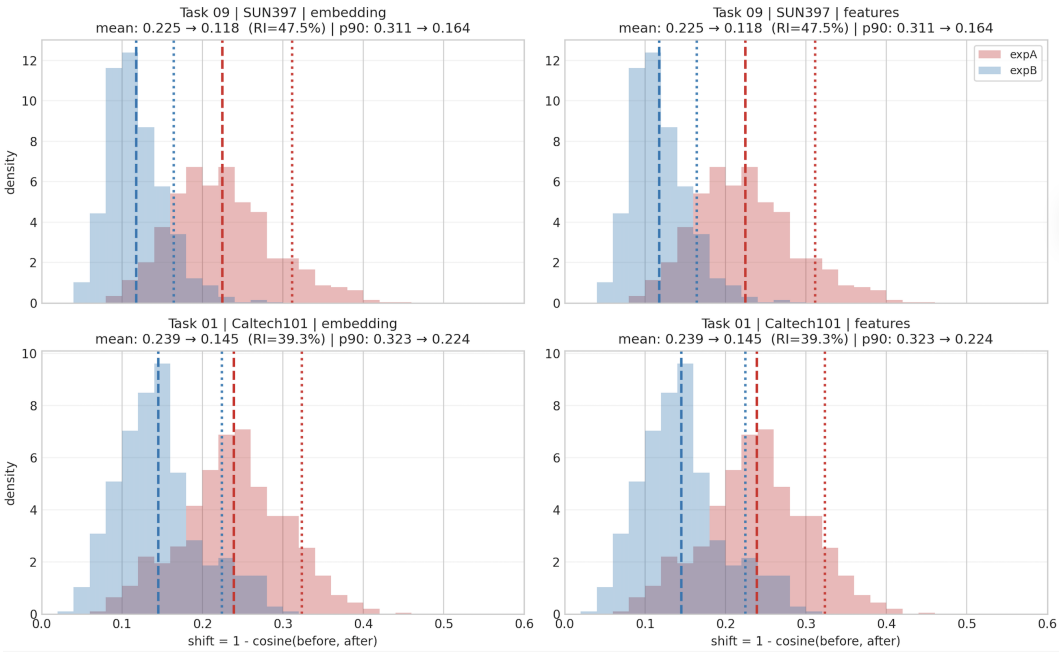


Figure 2: Per-input stability comparison between expA (cross-modal affinity only) and expB (image+text+cross-modal affinity). We plot the distribution of per-image cosine shift $1 - \cos(\text{before}, \text{after})$ on the two tasks (SUN397 and Caltech101), shown for both embeddings and features. Dashed vertical lines indicate the mean shift and dotted lines indicate the 90th-percentile shift ($p90$).

Table 6: SACED selection-criterion ablation on X-TAIL with reference budget k_{ref}=3. The full criterion combines all three signals.

Selection rule	Avg	Last
Full criterion (COMET)	**76.1**	88.1
Co-activation only	74.8	**88.4**
All prior experts	74.4	87.9
Random selection	74.9	87.7
Top-1 agreement only	74.2	87.0
Distribution similarity only	67.8	77.2

and $+10.5\%$ parameters; in continual learning, where Last reflects what the model can still do at the end of the stream, CMAB's trade-off favors the right axis. Aggressive thresholds ($\sigma \in \{0.5, 0.7\}$) collapse the pool to one or two experts and lose 26–29 pp Last. On the heterogeneous 4-task sequence, the best fixed threshold ($\sigma = 0.9$) retains all four experts but achieves only 41.9 Average, whereas CMAB reaches 72.9 Average and 96.7 Last with 3–4 experts. No single fixed threshold is optimal across both sequences; the bandit adapts because its context vector reads accuracy, merge compatibility, degradation risk, capacity pressure, and expert load jointly rather than relying on similarity alone.

A merge-threshold sweep $\tau \in [0.50, 0.80]$ produces final pool sizes $M_K \in \{5, 6, 7, 9, 10\}$ that vary monotonically with τ, confirming smooth control over the storage–accuracy trade-off. EuroSAT consistently merges only at the most permissive setting; Flowers102 merges across $\tau \in [0.50, 0.65]$ at a modest cost of -2.7 pp on its dedicated-expert score; per-task accuracies vary by less than 2 pp across τ on completed tasks.

K Per-Dataset Capacity Analysis

Table 8 explains the per-dataset variation through the lens of teacher and student zero-shot accuracy. On eight of ten X-TAIL tasks, COMET matches or exceeds the CLIP teacher's zero-shot ceiling. Among the five datasets with positive teacher–student capacity gaps, COMET recovers $> 100\%$ of the gap on four, exceeding the teacher despite a $2.6\times$ smaller student. COMET sits below the teacher's zero-shot ceiling only on Food (83.5 vs. 84.4) and Caltech101 (92.9 vs. 93.5), by under one point in each case, margins within run-to-run noise.

Table 7: Extended expert-merging ablation on X-TAIL. $\mathcal{A}/M$ = Average accuracy per stored expert. Fixed σ replaces CMAB with a non-adaptive rule that merges whenever the candidate similarity exceeds σ.

Variant	M	Params	Trans.	Avg	Last	$\mathcal{A}/M$
COMET (CMAB)	**8**	**551M**	**62.6**	**76.1**	**88.1**	**9.5**
Always-Expand	10	667M	61.6	75.8	87.3	7.6
Fixed σ=0.9	9	609M	59.4	77.2	84.9	8.6
Fixed σ=0.7	2	203M	58.6	59.3	61.9	29.7
Fixed σ=0.5	1	145M	58.3	58.4	58.9	58.4
Always-Merge	1	145M	58.4	58.4	58.4	58.4
Random (p=0.5)	3	261M	58.9	64.9	69.1	21.6

Table 8: Per-dataset capacity-gap analysis on X-TAIL. All accuracies use each dataset's own label space, unlike the generalized X-TAIL label space of Table 1. ZS = zero-shot. Recovery $= (\text{COMET}-\text{TinyCLIP ZS})/(\text{CLIP ZS}-\text{TinyCLIP ZS})\times 100\%$ on positive-gap datasets; "–" = capacity gap is non-positive.

Dataset	CLIP ZS	TinyCLIP ZS	Gap	COMET	Recovery
Aircraft	22.6	16.9	+5.7	83.1	$> 100\%$
Caltech101	93.5	93.9	−0.4	92.9	–
DTD	44.5	45.3	−0.8	78.8	–
EuroSAT	36.1	50.7	−14.6	98.5	–
Flowers102	65.7	67.9	−2.2	95.6	–
Food	84.4	78.5	+5.9	83.5	84.7%
MNIST	62.6	38.6	+24.0	98.8	251%
OxfordPet	85.3	79.7	+5.6	92.0	220%
StanfordCars	60.3	51.3	+9.0	90.8	$> 100\%$
SUN397	61.3	69.2	−7.9	80.1	–

Student backbone. Replacing the student with the smaller TinyCLIP-8M backbone causes Last to drop from 88.1 to 61.1 on X-TAIL, with fine-grained domains losing the most (-72.4 pp on EuroSAT and -55.2 pp on Flowers102), confirming that per-dataset variation tracks backbone capacity rather than method failure. MNIST is essentially capacity-insensitive (-1.0 pp), consistent with the observation that handwritten digit classification requires modest representational depth.

Teacher backbone. Replacing the CLIP ViT-B/16 teacher with the smaller CLIP ViT-B/32 (under $\tau = 0.80$) shifts Last from 86.2 to 78.5 (-7.7 pp) and Transfer from 52.7 to 61.5 ($+8.8$ pp), while M decreases from 10 to 9. A weaker teacher therefore lowers the upper bound on per-task accuracy but produces softer supervision that transfers more uniformly, indicating that the teacher's representational ceiling propagates to the experts.

L Routing Accuracy and Cross-Expert Coherence

Across the ten task-specific routers, the autoencoder router identifies the correct task for 95.9% of test inputs at top-1 and 99.4% at top-2; errors concentrate on borderline inputs where the gating margin is low (0.621 vs. 0.977 when correctly routed). Table 9 decomposes the predictions of the retained pool into a four-cell route×predict outcome. After merging, the router sends 2,069 samples (2.6%) away from their oracle expert, and the selected expert still classifies 1,086 of them (52.5%) correctly. On a 7-task subset where with- and without-SACED runs share oracle expert assignments, SACED raises misrouted-sample accuracy from 35.3% to 46.1% (+10.8 pp) and lifts pairwise expert representation similarity (CKA) by +0.009, supporting the cross-expert coherence interpretation that SACED converts hard routing failures into graceful degradation. A random-attention control confirms that the alignment SACED learns is statistically meaningful (Cohen's $d = 0.45$, $p < 0.001$) rather than noise.

M Teacher Attention on Out-of-Distribution Domains

We measure CLS-token attention concentration in the frozen teacher across three OOD domains (Table 10). C_{10} denotes the fraction of total CLS attention captured by the top decile of the $14 \times 14 = 196$ patches (uniform baseline 0.10); H is Shannon entropy over the patch grid in bits (uniform maximum $\log_2 196 \approx 7.61$). Early-layer attention is diffuse on every OOD domain, but by the final layer attention concentrates 66–76% of mass on the top decile, with entropy dropping to 5.1–5.7 bits. This is consistent with COMET's strong heterogeneous-benchmark performance: TAD does not require teacher classification correctness, only stable relative geometry, and SACED compares each expert's attention distance to the teacher with that of a prior expert, so the student never has to reproduce the teacher's attention map.

Table 9: Four-cell decomposition of test predictions on X-TAIL (80,939 samples). Route-OK = autoencoder selected the oracle expert; Predict-OK = expert produced the correct label.

Cell	Count	Share
Route-OK $\wedge$ Predict-OK	66,664	82.4%
Route-OK $\wedge$ Predict-Wrong	12,206	15.1%
Route-Wrong $\wedge$ Predict-OK	1,086	1.3%
Route-Wrong $\wedge$ Predict-Wrong	983	1.2%

Table 10: Frozen teacher attention statistics on three OOD domains, averaged over $N{=}50$ images per dataset. C_{10}: fraction of attention in the top decile of 196 patches (uniform 0.10). H: Shannon entropy in bits (uniform $\log_2 196 \approx 7.61$).

Dataset	Domain	C_{10} (L0)	H (L0)	C_{10} (L11)	H (L11)
COVID-19	Medical	0.338	7.11	**0.663**	5.71
SEM	Microscopy	0.280	7.28	**0.709**	5.48
NCT-CRC	Histopath	0.228	7.39	**0.760**	5.08

N Held-out Zero-Shot Generalization

After completing the full 10-task X-TAIL stream, the same model retains zero-shot accuracy on datasets it never saw during continual learning: 45.6 on CIFAR-100, 71.0 on CIFAR-10, and 60.6 on CUB-200. Because TAD anchors the student's affinity geometry to the frozen teacher rather than to any specific downstream label space, the student inherits a broadly transferable representation that survives the full continual-learning sequence.

O Detailed Computational Cost

Table 11 reports per-task training cost in detail. Among CLIP-based methods, COMET trains $2.6\times$ fewer parameters per task than full-CLIP fine-tuning baselines (LwF, WiSE-FT, ZSCL) and uses below-WiSE-FT VRAM because only the current expert receives gradients while the teacher and at most $k_{\text{ref}} \leq 2$ frozen reference experts are forward-only. PEFT methods (MoE-Adapters, LADA) retain a smaller VRAM footprint because their trainable parameter count is one to two orders of magnitude smaller, but they still require the full frozen CLIP backbone in memory for forward passes.

Storage. With $M = 8$ retained experts after the X-TAIL stream, COMET stores roughly $8 \times 58\,\text{M} + 87\,\text{M} \approx 551\,\text{M}$ parameters (eight experts and the teacher's vision encoder); the eight routers add about 1.6 M. Because at most one expert is active per input at inference, this stored–active ratio of $3.8\times$ is conservative by mixture-of-experts standards (Switch-Base: $\sim 30\times$, Mixtral: $\sim 3.6\times$) and does not enter the deployment compute budget.

P 16-Shot Results

We re-run X-TAIL under a 16-shot per-class budget per task, holding all other hyperparameters at their defaults. Table 12 reports the per-task breakdown across all ten datasets and the three continual-learning metrics. Across the ten tasks, COMET's Average is 72.6 (+9.0 pp over no-distillation) and Last is 83.3 (+10.0 pp over no-distillation), preserving the qualitative ordering of the full-shot table in the main text: TAD and SACED retain their additive contribution at low data budgets. The same trend holds on the heterogeneous 4-task sequence, with COMET continuing to lead all baselines on Average and Last.

Table 11: Per-task training cost on a single NVIDIA A6000 (48 GB), batch size 64, 1,000 iterations. "Trainable" counts only parameters that receive gradients.

Method	Trainable	Peak VRAM	Backbone in Memory
LwF	$\sim$ 150M	$\sim$ 4.1 GB	Full CLIP (trainable)
WiSE-FT	$\sim$ 150M	$\sim$ 2.9 GB	Full CLIP (trainable)
ZSCL	$\sim$ 150M (+150M ref)	$\sim$ 4.1 GB	Full CLIP + frozen ref
MoE-Adapters	2–5M	$\sim$ 1.1 GB	Frozen CLIP
LADA	1–3M	$\sim$ 0.9 GB	Frozen CLIP
COMET (Ours)	$\sim$ 58M	$\sim$ 2.7 GB	Frozen teacher

Table 12: Comparison of different continual learning methods on X-TAIL [Xu et al., 2024] 16-shot for each task in terms of *Transfer*, *Average*, and *Last* scores (%). The best results are highlighted with **bold** style.

Method	Aircraft	Caltech101	DTD	EuroSAT	Flowers	Food	MNIST	Pets	Cars	Sun397	Average
`TinyCLIP ViT-39M/16@224px`											
Zero-shot	17.0	75.6	39.0	40.6	65.0	77.4	22.0	78.9	47.8	67.2	53.1
`CLIP ViT-B/16@224px`											
Zero-shot	23.8	74.3	36.4	37.4	64.1	83.4	43.9	87.8	65.5	60.8	57.7
Transfer											
LwF	–	64.8	29.1	7.6	49.7	72.0	35.6	64.0	29.0	53.7	45.1
WiSE-FT	–	68.3	30.8	17.3	59.2	82.5	41.4	74.9	47.8	59.6	53.5
ZSCL	–	71.7	32.6	15.1	63.0	85.0	29.5	83.6	59.4	63.2	55.9
MoE-Adapters	–	76.2	30.4	37.6	65.4	**87.9**	**53.0**	**88.4**	63.0	59.1	**62.3**
LADA	–	74.9	35.9	36.5	**66.1**	84.2	37.9	88.0	**65.3**	60.1	60.9
COMET (Ours)	–	**85.6**	**45.5**	**44.2**	65.1	79.0	35.9	77.7	47.3	**67.6**	60.9
Average											
LwF	18.2	80.9	42.3	34.0	48.9	80.7	59.6	71.1	35.8	56.3	52.8
WiSE-FT	27.6	80.1	43.3	38.5	63.3	86.5	62.3	79.3	54.6	61.5	59.7
ZSCL	40.1	78.2	49.9	40.0	73.5	**87.7**	56.2	86.2	64.6	64.8	64.1
MoE-Adapters	42.8	77.8	33.1	54.5	67.4	87.1	58.7	87.7	62.1	59.5	63.1
Primal-RAIL	44.9	89.5	55.6	68.9	83.5	85.3	**64.9**	88.1	69.1	64.9	71.5
LADA	48.8	92.0	60.8	**71.3**	84.5	84.1	60.3	**88.9**	**69.2**	61.6	72.1
COMET (Ours)	**66.7**	**96.2**	**68.5**	68.5	**84.7**	78.8	59.4	80.1	55.2	**68.3**	**72.6**
Last											
LwF	10.4	74.6	47.0	23.2	48.7	89.1	93.1	86.8	56.1	**79.3**	60.8
WiSE-FT	29.8	78.5	46.5	31.6	70.9	90.3	90.1	87.5	79.8	78.8	68.4
ZSCL	34.9	78.8	53.1	33.4	80.7	**90.7**	95.7	92.2	84.7	78.6	72.3
MoE-Adapters	42.8	78.0	33.8	61.8	68.8	86.4	67.2	86.1	58.5	63.2	64.7
Primal-RAIL	44.5	94.5	67.4	86.1	96.4	86.9	91.6	**91.6**	80.8	75.7	81.6
LADA	49.3	95.1	68.4	**86.2**	96.8	84.5	93.8	91.3	84.8	74.9	82.5
COMET (Ours)	**66.5**	**97.4**	**73.9**	76.9	**97.8**	78.5	**94.8**	85.5	**86.8**	74.8	**83.3**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes; the abstract and Introduction state the three targeted failure modes and the COMET contributions, and the claims are supported by the method sections and experiments in Sections 4.1–4.5 and Tables 1–3.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes; limitations of the expert-growth argument are discussed in Appendix A, including the dependence on stream novelty and decision errors, and computational/storage trade-offs are discussed in the cost analysis and Appendix O.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Yes; the sub-linear expert-growth result states its assumptions and proof in Appendix A, including Theorem A.1 and Corollary A.3.

   Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes; the experimental setup, datasets, metrics, model choices, training schedule, routing procedure, and hyperparameters are described in Section 4.4, Section 4.5, the Experiments section, Algorithm 1, and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We plan to release the code, reproduction scripts, configuration files, and setup instructions in an open-source GitHub repository by the camera-ready version. Reproducibility is also supported through the public datasets, full algorithmic description, implementation details, hyperparameters, and compute information in Algorithm 1 and Appendix C.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.

- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes; the benchmarks, data sequences, evaluation metrics, model backbones, optimizer, learning rate, batch size, reference pool, loss weights, CMAB settings, and routing rule are specified in the Experiments section and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: No; the current manuscript reports task-order robustness and sensitivity analyses in Appendices E–D, but does not provide seed-level error bars or formal significance tests for all main experimental tables.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Yes; Section *Analysis of Computational Cost* and Appendix O report GPU type, VRAM, FLOPs, latency, throughput, per-task training time, and total training cost.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: Yes; the work uses public datasets and existing pretrained vision-language models, involves no human-subject experiments, and is evaluated as a continual-learning method without collecting private or sensitive data.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: No; the current manuscript focuses on the method and empirical evaluation and does not yet include a dedicated discussion of positive and negative societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: Not applicable; the paper does not introduce or release high-risk assets such as generative models, pretrained language models, scraped datasets, or systems intended for direct deployment.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: No; existing datasets and pretrained models are credited through citations in the Related Work and Experiments sections, but the current manuscript does not yet include a complete license table. The planned open-source GitHub repository for the camera-ready version will include the licenses, terms of use, and attribution information for all external assets used in the experiments.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: Not applicable; the paper proposes a method and does not introduce a new dataset, benchmark, model release, or other new asset in the current submission.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: Not applicable; the paper does not involve crowdsourcing, user studies, or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: Not applicable; the paper does not involve crowdsourcing or human-subject research, so IRB or equivalent review is not required.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: Not applicable; LLMs are not used as an important, original, or non-standard component of the proposed method, experiments, or scientific claims.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.