UNSUPERVISED LONG-TAILED ADAPTATION OF VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Adapting vision-language models to downstream tasks has achieved remarkable success by leveraging pseudo-labels generated from unlabeled data. Existing methods typically assume a uniform unlabeled data distribution, and thus the resulting pseudo-label distribution is likewise uniform. However, real-world data distributions are often long-tailed. To tackle this, we formalize a new scenario termed Unsupervised Long-Tailed Adaptation (ULTA). Under this scenario, existing methods exhibit a contrasting phenomenon: head-class performance drops sharply, which is distinct from supervised long-tailed learning where tail classes suffer the most. In particular, we uncover that the distributional mismatch not only erodes head-class boundaries, but also pushes head samples into confusable classes, reinforcing the model's inherent bias. To address these issues, we propose a novel model called Margin-Aware Refinement with Structural alignment (MARS). Specifically, we mitigate head-class boundary erosion via Boundary-Preserving Alignment, which takes the zero-shot VLM as a fixed visual reference to suppress probability increases that lack visual support in the training targets. Building upon this, we introduce Margin-aware Self-Refinement, which employs a dynamic adjustment strategy to refine tail and confusable classes while preventing prediction bias. Extensive experiments on nine benchmark datasets demonstrate that MARS outperforms state-of-the-art methods, achieving an average accuracy improvement of **4.71** percentage points. **Code is available in the supplementary material.**

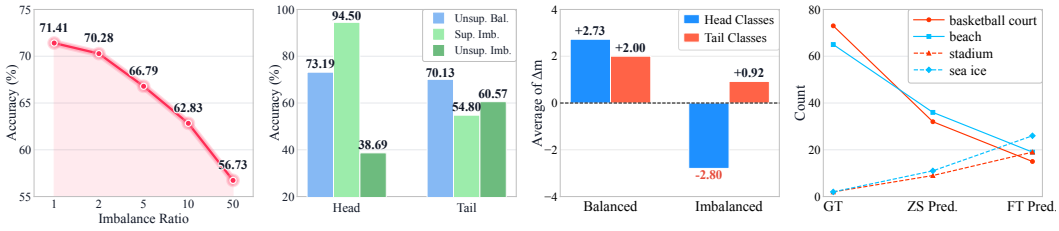
1 INTRODUCTION

Vision-language models (VLMs) (Jia et al., 2021; Li et al., 2022; Xiao et al., 2024), such as CLIP (Radford et al., 2021), have demonstrated remarkable zero-shot capabilities by establishing a shared feature space across visual and textual modalities via web-scale data. Despite the impressive zero-shot capabilities, their performance is limited when transferred to specific downstream tasks due to the domain gap. While fine-tuning on labeled data can bridge this gap (Zhou et al., 2022b; Jia et al., 2022; Ouali et al., 2023; Shi et al., 2024), acquiring sufficient annotations in specialized scenarios (e.g., medical imaging Huang et al., 2021; Wang et al., 2022) is impractical. Consequently, recent studies (Huang et al., 2022; Mirza et al., 2023; Menghini et al., 2023) have shifted their focus toward enhancing performance by exploiting unlabeled data.

These studies (Zhang et al., 2024; Li et al., 2025; Wang et al., 2025) primarily derive supervisory signals from unlabeled data via pseudo-labeling. Those methods usually leverage a uniform prior explicitly or implicitly to yield class-balanced pseudo-labels. However, real-world data exhibits a long-tailed distribution, creating a mismatch with this uniform prior. As evidenced in Fig. 1(a), these methods (e.g., CPL Zhang et al., 2024) suffer from substantial performance degradation on long-tailed data. Notably, we observe an inverted degradation pattern that diverges from supervised long-tailed learning. In supervised long-tailed learning, the tail classes typically suffer the most due to the scarcity of their training samples. While in unsupervised long-tailed scenarios, existing methods lead to **a severe performance drop for head classes** (see Fig. 1(b)).

Empirically, we find that in long-tailed scenarios, the uniform assumption in existing methods forces the model to misclassify many head samples as tail classes. Fitting these "head-as-tail" false samples leads to head-class decision boundary erosion, which causes a significant decrease in the logit margin (i.e., the gap between the ground-truth logit and the highest non-target logit) of head classes (see

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) (b) (c) (d)

Figure 1: **(a)** The overall test accuracy of the existing unsupervised method CPL (Zhang et al., 2024) under different imbalance ratios. **(b)** Comparison of head- and tail-class accuracies across three settings on the RESISC45 dataset: unsupervised learning with balanced data (Unsup. Bal.), supervised learning with imbalanced data (Sup. Imb.), and unsupervised learning with imbalanced data (Unsup. Imb.). **(c)** Comparison of average logit margin changes ($\Delta m = \text{margin}_{\text{Fine-tuned}} - \text{margin}_{\text{Zero-shot CLIP}}$). Under the long-tailed setting, the head-class margin significantly decreases compared to the uniform setting, indicating severe boundary erosion. **(d)** Label counts across three cases: ground-truth (GT), zero-shot predictions (ZS Pred.), and fine-tuned predictions (FT Pred.). We group confusable pairs by color (e.g., blue: "beach" vs. "sea ice") and observe that zero-shot CLIP inherently exhibits bias toward confusable classes, which is further exacerbated by fine-tuning.

Fig. 1(c)). Furthermore, zero-shot CLIP's inherent bias toward confusable classes (e.g., semantically similar pairs like "beach" vs. "sea ice") is exacerbated during fine-tuning (see Fig. 1(d)).

To address these challenges, we propose the Margin-Aware Refinement with Structural alignment (MARS) to tackle **Unsupervised Long-Tailed Adaptation (ULTA)**, a realistic scenario where VLMs are adapted to unlabeled data with a long-tailed distribution. MARS follows a preserve-before-refine design: it first preserves the reliable head-class boundary and then refines the tail and confusable classes. Specifically, to mitigate head-class boundary erosion caused by the erroneous assignment of pseudo-labels, we propose Boundary-Preserving Alignment (BPA). This approach exploits the class-level visual structure of zero-shot CLIP as a fixed reference and suppresses probability increases unsupported by visual evidence before they become training targets, instead of directly fitting the model's own erroneous predictions. Building upon this, we introduce Margin-aware Self-Refinement (MSR) to further improve tail and confusable classes while preserving the well-learned decision boundary. Adopting the aligned model as a teacher, this strategy leverages the estimated class distribution and the sample-specific margin to dynamically adjust the learning focus, distinguishing valid supervision from noise to actively acquire discriminative features.

In summary, our contributions are as follows:

- To our knowledge, we are the first to study unsupervised long-tailed adaptation (ULTA), a realistic yet under-explored scenario where VLMs are adapted to long-tailed unlabeled data.
- We uncover a distinct phenomenon where existing unsupervised methods suffer severe degradation on head classes due to boundary erosion from forced uniform pseudo-labels, and observe aggravated prediction bias in confusable classes. See more analyses in Sections 3.2 and 4.4.
- We propose a novel model, MARS, to tackle these challenges. Specifically, we design Boundary-Preserving Alignment to mitigate boundary erosion by exploiting the visual structure of zero-shot CLIP, and develop Margin-aware Self-Refinement to adaptively balance teacher guidance and pseudo-label supervision, further improving discriminability for tail and confusable classes.
- Comprehensive experiments on nine benchmarks with different imbalance ratios show that MARS achieves state-of-the-art performance with an average improvement of **4.71** percentage points.

2 Related Work

2.1 Prompt Learning in VLMs

Vision-language models (VLMs) (Radford et al., 2021; Jia et al., 2021; Li et al., 2022; Xiao et al., 2024) have achieved remarkable zero-shot generalization by aligning visual and textual modalities.

Prompt learning is a parameter-efficient paradigm for adapting these models to downstream tasks. Specifically, CoOp (Zhou et al., 2022b;a) pioneers this direction by optimizing context vectors within the text encoder. VPT (Jia et al., 2022) extends this concept to the visual modality by injecting learnable prompts into the image encoder. To facilitate cross-modal interaction, MaPLe (Khattak et al., 2023) introduces a multi-modal prompting approach to ensure deep alignment between branches. Alternatively, some works focus on refining the encoder's output features without fine-tuning the backbone. For example, CLIP-Adapter (Gao et al., 2024) employs lightweight residual modules to transform visual and textual features. **Despite their success, these methods heavily rely on labeled data, which severely constrains their applicability in unsupervised scenarios.**

2.2 Learning from Unlabeled Data for VLMs

To mitigate the reliance on labeled data, various strategies leverage the zero-shot capabilities of VLMs to generate pseudo-labels for self-training. For instance, UPL (Huang et al., 2022) and GRIP (Menghini et al., 2023) focus on improving pseudo-label quality via confidence-based filtering or iterative top-k selection. Similarly, CPL (Zhang et al., 2024) assigns each sample a set of candidate pseudo-labels at each iteration, expecting the true label to be captured within the set. Beyond traditional pseudo-labeling, recent studies explore novel perspectives for mining supervisory signals from unlabeled data. UEO (Liang et al., 2024) employs marginal entropy as a unified metric to regularize the decision boundary, minimizing uncertainty for confident predictions while maximizing it for ambiguous samples. TMP (Li et al., 2025) reformulates the adaptation process as a binary verification task instead of the traditional classification task. CAP (Wang et al., 2025) identifies inherent biases and concept confusion in CLIP zero-shot predictions, rectifying these issues by learning a projection layer that encourages balanced predictions across classes. microCLIP (Silva et al., 2026) combines coarse-to-fine token fusion and saliency-aware pooling for unsupervised CLIP adaptation. **However, their reliance on a uniform prior over unlabeled data limits their applicability to long-tailed scenarios.**

3 Methodology

3.1 Problem Formulation and Challenges

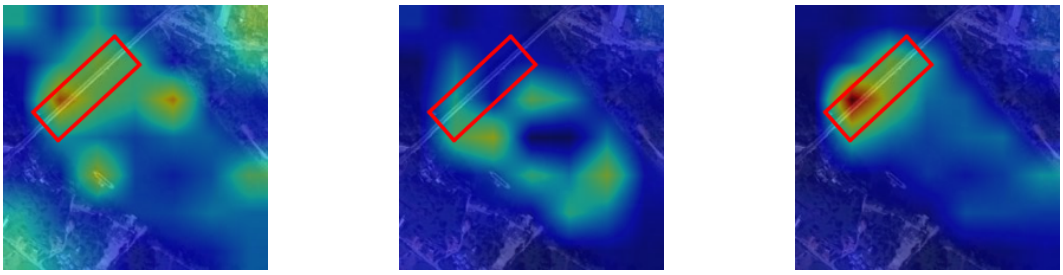
In unsupervised long-tailed adaptation (ULTA), we consider an unlabeled dataset $\mathcal{D}_u = \{x_i\}_{i=1}^N$ of size N belonging to C classes. While one can theoretically define N_c as the number of samples for class c and $\gamma = \max_c N_c / \min_c N_c$ as the imbalance ratio, neither is accessible in practice. The core challenge is that existing methods implicitly assume a uniform prior when generating pseudo-labels, which inevitably leads to severe boundary erosion and prediction bias. Our goal is to achieve robust performance across different imbalance ratios.

3.2 Analysis of Misalignment

Existing unsupervised methods typically rely on a uniform prior to generate pseudo-labels. In long-tailed scenarios, this assumption forces the model to align visual features with incorrect targets, causing a severe misalignment between the learned features and the true semantic information. We analyze the consequences of this misalignment from the perspective of gradient dynamics.

Consider a head-class sample x with normalized visual feature $\mathbf{v} \in \mathbb{R}^d$ and ground-truth label c_h, which is incorrectly assigned a tail-class pseudo-label $\hat{y} = c_t$. In the standard self-training framework, this incorrect pseudo-label is treated as a hard target, forcing the visual feature to align with the textual embedding $\mathbf{t}_{c_t}$. Formally, we minimize the cross-entropy loss $\mathcal{L}_{ce} = -\log p(c_t|\mathbf{v})$, where the probability p is computed via cosine similarity with textual embeddings $\{\mathbf{t}_c\}_{c=1}^C$ scaled by a temperature parameter τ. The gradient of this loss with respect to the visual feature $\mathbf{v}$ is derived as:

$$\nabla_{\mathbf{v}} \mathcal{L}_{ce} = \frac{1}{\tau} \left(\sum_{j=1}^{C} p(j|\mathbf{v}) \cdot \mathbf{t}_j - \mathbf{t}_{c_t} \right). \tag{1}$$



(a) Zero-shot CLIP (b) Baseline (c) Ours

Figure 2: Visualization of attention maps for misassigned samples on RESISC45. We select head-class samples where the baseline method incorrectly assigns tail-class pseudo-labels. **(a)** The model mainly focuses on the foreground object. **(b)** The model is distracted by the irrelevant background. **(c)** The model correctly captures the foreground object while reducing background attention.

Before fitting this incorrect pseudo-label, the model still assigns high confidence to the correct head class c_h, so the probability mass is concentrated on c_h. Consequently, the gradient approximates:

$$\nabla_{\mathbf{v}} \mathcal{L}_{ce} \approx \frac{1}{\tau}(\mathbf{t}_{c_h} - \mathbf{t}_{c_t}). \tag{2}$$

The feature vector $\mathbf{v}$ is updated by moving against the gradient direction (i.e., $\mathbf{v} \leftarrow \mathbf{v} - \eta \nabla_{\mathbf{v}} \mathcal{L}_{ce}$). Consequently, the resulting feature shift satisfies $\Delta \mathbf{v} \propto \mathbf{t}_{c_t} - \mathbf{t}_{c_h}$. This derivation illustrates that the gradient drives the visual feature $\mathbf{v}$ to align with the distant tail embedding $\mathbf{t}_{c_t}$, deviating from its true semantic embedding $\mathbf{t}_{c_h}$ and thereby eroding the decision boundaries of head classes. Meanwhile, to minimize the loss, the model is forced to exploit irrelevant background noise. As visualized in Fig. 2(b), for these mislabeled samples, existing methods (e.g., CPL Zhang et al., 2024) cause the model to ignore the foreground object and focus on the background, thereby degrading feature discriminability.

3.3 Boundary-Preserving Alignment

To mitigate this misalignment, we prevent erroneous probability changes from becoming training targets. In self-training, the targets are generated from the model's own predictions, so an erroneous prediction is fitted again in the next round. We therefore take zero-shot CLIP, which remains fixed during adaptation, as a visual reference, and correct the targets according to whether each probability increase is supported by visual evidence.

Concretely, let x_i^w denote the weak view of x_i. We extract the unit-normalized feature $\mathbf{u}_i$ and the predicted distribution a_i of zero-shot CLIP on x_i^w, and denote the zero-shot class as $c_i = \arg\max_c a_{ic}$. For each class c, we average the features of the samples with $c_i = c$ and ℓ_2-normalize the result to obtain the class-level visual anchor $\boldsymbol{\mu}_c$, and measure the visual support of sample i for class c by $s_{ic} = \mathbf{u}_i^\top \boldsymbol{\mu}_c$. These anchors characterize the class-level visual structure of zero-shot CLIP. Unlike the textual embeddings, which always favor the zero-shot class, they compare each sample with the visual centroids of the zero-shot groups and thus provide complementary evidence.

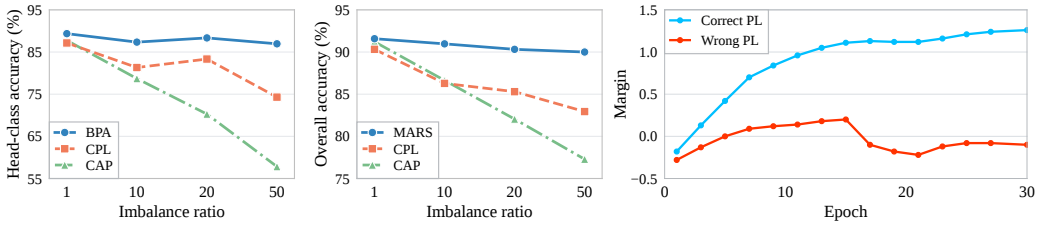
At the start of each alignment epoch, let p_i denote the predicted distribution of the current model on x_i^w and $b_i = \arg\max_c p_{ic}$. For $j \neq c_i$, the probability increase induced by adaptation is $g_{ij} = \max(p_{ij} - a_{ij}, 0)$. We regard this increase as visually unsupported, i.e., $h_{ij} = 1$, if $s_{ic_i} > s_{ij}$, and set $h_{ij} = 0$ otherwise. If $b_i \neq c_i$ and $s_{ib_i} > s_{ic_i}$, the adaptation has likely corrected a zero-shot error, and we set $h_{ij} = 0$ for all j.

Since the verdict of a single sample is noisy and pseudo-label errors tend to recur between the same class pairs (Fig. 1(d)), we aggregate the evidence over the unlabeled data:

$$\rho_{cj} = \frac{\sum_{i:c_i=c} g_{ij} h_{ij}}{\sum_{i:c_i=c} g_{ij}}, \qquad c \neq j. \tag{3}$$

The corrected target then returns the unsupported part of each increase to the zero-shot class:

$$q_i = p_i + \sum_{j \neq c_i} \delta_{ij}(\mathbf{e}_{c_i} - \mathbf{e}_j), \qquad \delta_{ij} = \rho_{c_i j}\, h_{ij}\, g_{ij}, \tag{4}$$



(a) (b) (c)

Figure 3: **(a)** BPA preserves head-class accuracy on OxfordPets as the imbalance ratio increases. **(b)** The complete MARS pipeline maintains stable overall accuracy on OxfordPets across imbalance ratios. **(c)** Evolution of classification margins during training. Samples with correct pseudo-labels (Correct PL) consistently exhibit increasing margins, whereas mislabeled samples (Wrong PL) fluctuate around zero or remain negative.

where $\mathbf{e}_c$ is the one-hot vector for class c. Since $0 \le \delta_{ij} \le g_{ij} \le p_{ij}$, q_i remains a valid distribution. We fix q_i within each epoch and use it to supervise the strong view x_i^s generated by RandAugment (Cubuk et al., 2020):

$$\mathcal{L}_{\text{BPA}} = -\frac{1}{B} \sum_{i \in \mathcal{B}} \sum_{c=1}^{C} q_{ic} \log p(c \mid x_i^s), \tag{5}$$

where $p(c \mid x_i^s)$ is the predicted probability of class c on x_i^s and $\mathcal{B}$ is a minibatch of size B.

Crucially, let $\ell(q_i)$ denote the cross-entropy between target q_i and the softmax of the strong-view logits z_i. Compared with the uncorrected target p_i, the corrected target changes the gradient by

$$\nabla_{z_i} \ell(q_i) - \nabla_{z_i} \ell(p_i) = -\sum_{j \neq c_i} \delta_{ij} (\mathbf{e}_{c_i} - \mathbf{e}_j). \tag{6}$$

Each unsupported increase thus induces a push along $\mathbf{e}_{c_i} - \mathbf{e}_j$ on the logits. Since $z_{ic} = \mathbf{v}^\top \mathbf{t}_c / \tau$ for the visual feature $\mathbf{v}$ of x_i^s, this push moves $\mathbf{v}$ toward $\mathbf{t}_{c_i}$ and away from $\mathbf{t}_j$, counteracting the misaligned gradient in Eq. (2), while supported increases, including those that correct zero-shot errors, are preserved. As shown in Fig. 2(c), BPA restores the attention to the foreground object, and Fig. 3(a) shows that it keeps head-class accuracy stable as the imbalance ratio increases.

3.4 Margin-Aware Self-Refinement

Although boundary-preserving alignment establishes a foundation in the joint feature space, refining the decision boundaries for the tail and confusable classes remains a critical challenge in unsupervised long-tailed scenarios. To address this, instead of directly fitting noisy pseudo-labels that can destroy the boundary, we propose a margin-aware self-refinement strategy that improves the feature space by decoupling well-learned boundary preservation from targeted feature enhancement.

We employ the aligned model from Section 3.3 as a frozen teacher and introduce a learnable adapter as the student. The refinement objective is formulated as follows:

$$\mathcal{L}_{\text{refine}} = \alpha \cdot \mathcal{L}_{\text{KL}}(p^T, p^S) + (1 - \alpha) \cdot \mathcal{L}_{\text{CE}}(\hat{y}, p^S), \tag{7}$$

where p^T and p^S denote the predictions of the frozen teacher and the student, respectively. In this formulation, the Kullback-Leibler (KL) divergence loss $\mathcal{L}_{\text{KL}}$ serves as a safety anchor to preserve the well-learned concepts from the frozen teacher, while the standard cross-entropy loss $\mathcal{L}_{\text{CE}}$ leverages pseudo-label supervision with $\hat{y} = \arg\max_c p_c^T$ to acquire new knowledge. The gating factor $\alpha \in [0, 1]$ plays a critical role in balancing knowledge preservation against the learning of new features.

To adaptively modulate α, we introduce a class-level prior. Specifically, we define the gating factor as $\alpha = (r_{\hat{y}})^\beta$, where β is a hyperparameter. We fix $\beta = 1$ across all datasets, and the sensitivity analysis in Appendix D.1 further shows that our method is robust to this hyperparameter. Here, $r_{\hat{y}} = \frac{\hat{n}_{\hat{y}}}{\max_c(\hat{n}_c)}$ represents the normalized frequency, where $\hat{n}_{\hat{y}}$ denotes the estimated sample count for class $\hat{y}$

accumulated from the teacher's soft predictions. High-frequency predictions primarily correspond to head classes, which require a higher α to maintain the model's well-established boundary. In contrast, low-frequency predictions correspond to tail and confusable classes, which necessitate a lower α to facilitate the acquisition of new knowledge, thereby enhancing feature discriminability.

However, relying solely on class-level priors is insufficient, as it overlooks the quality of individual pseudo-labels within the same class. As illustrated in Fig. 3(c), our empirical observations reveal that correctly pseudo-labeled samples mostly exhibit increasing margins during training, whereas incorrectly pseudo-labeled samples maintain negative margins. Motivated by this, we propose Cumulative Margin Consistency (CMC) to quantify sample quality. Specifically, given the student's logits $z^{(t)}$ at epoch t, the margin is defined as:

$$M^{(t)}(x, \hat{y}) = z_{\hat{y}}^{(t)}(x) - \max_{j \neq \hat{y}} z_j^{(t)}(x). \tag{8}$$

We calculate the exponential moving average of this margin with momentum λ as $\text{CMC}(x, \hat{y})$ and refine α into a sample-specific factor $\hat{\alpha}$ to mitigate the impact of noisy pseudo-labels:

$$\hat{\alpha} = \frac{1}{2}\left(r_{\hat{y}}^{\beta} + (1 - \mathcal{N}(\text{CMC}(x, \hat{y})))\right), \tag{9}$$

where $\mathcal{N}(\cdot)$ is the min-max normalization function. The final refinement objective is formulated as:

$$\mathcal{L}_{\text{refine}} = \hat{\alpha} \cdot \mathcal{L}_{\text{KL}}(p^T, p^S) + (1 - \hat{\alpha}) \cdot \mathcal{L}_{\text{CE}}(\hat{y}, p^S). \tag{10}$$

The resulting stability across imbalance ratios is reflected in Fig. 3(b): the complete MARS pipeline retains its overall accuracy as the long-tailed distribution becomes more severe.

3.5 Model Training and Prediction

Our model consists of Boundary-Preserving Alignment (BPA) and Margin-aware Self-Refinement (MSR). The training process first performs boundary-preserving alignment with corrected weak-view targets, and then conducts margin-aware self-refinement with the frozen teacher to selectively refine the remaining hard classes and samples. The detailed training procedure is provided in Appendix A. Notably, both the teacher and student adapters are appended to the textual encoder. The pre-trained image and text encoders are kept frozen, and only the prompts and adapters are learned.

Upon completion of training, the teacher adapter is discarded. For an unseen test sample x^*, the predicted label y^* is obtained by computing the cosine similarity between the visual feature $\mathbf{v}^*$ extracted by the refined encoder and the pre-computed textual embeddings $\{\mathbf{t}_c\}_{c=1}^{C}$:

$$y^* = \underset{c \in \{1,\ldots,C\}}{\arg\max} \left(\frac{\mathbf{v}^* \cdot \mathbf{t}_c}{\tau}\right). \tag{11}$$

4 Experiments

4.1 Experimental Setting

Datasets. We conduct extensive experiments on nine classification datasets from diverse domains, including CUB (CUB.; Wah et al., 2011), RESISC45 (Res.; Cheng et al., 2017), FGVCAircraft (FA.; Maji et al., 2013), Caltech101 (Cal.; Fei-Fei et al., 2004), EuroSAT (ES.; Helber et al., 2019), Flowers102 (Flw.; Nilsback & Zisserman, 2008), Food101 (Food.; Bossard et al., 2014), OxfordPets (Pets.; Parkhi et al., 2012), and StanfordCars (Cars.; Krause et al., 2013). For the long-tailed settings, the training data is generated by down-sampling the original datasets to follow an exponential decay of different ratios. Specifically, the maximum number of samples per class is set to 100 (or the original class size), while the minimum is constrained to 2. Following the definition in Section 3.1, we set the imbalance ratio $\gamma \in \{10, 20, 50\}$ to simulate different long-tailed scenarios. For the fine-grained analysis, classes are divided into Head, Medium, and Tail groups using a 15/30/55 split by training-set frequency. The details of dataset statistics can be found in Appendix B.

Baselines. We compare our method with seven existing methods, namely, Few Pseudo-Labels (FPL; Menghini et al., 2023), Grow and Refine Iteratively Pseudo-labels (GRIP; Menghini et al.,

Table 1: Comparison of the test accuracy (%) across three seeds on nine benchmark datasets with different imbalance ratios (10, 20, 50). The results are reported as mean_{std} over three random seeds, with the best results presented in bold. Zero-shot CLIP is abbreviated as ZS CLIP.

Imbalance Ratio = 10

Method	CUB.	Res.	FA.	Cal.	ES.	Flw.	Food.	Pets.	Cars.	Avg.
ZS CLIP	$51.82_{0.00}$	$54.48_{0.00}$	$17.58_{0.00}$	$90.30_{0.00}$	$32.88_{0.00}$	$63.67_{0.00}$	$78.80_{0.00}$	$84.36_{0.00}$	$58.16_{0.00}$	$59.12_{0.00}$
FPL (NeurIPS'23)	$51.59_{0.23}$	$62.58_{0.86}$	$19.05_{0.37}$	$90.67_{0.39}$	$55.49_{1.35}$	$61.30_{1.72}$	$78.10_{0.15}$	$86.44_{0.33}$	$58.10_{0.28}$	$62.59_{0.63}$
GRIP (NeurIPS'23)	$48.01_{0.68}$	$61.95_{0.99}$	$18.25_{0.24}$	$88.98_{0.62}$	$58.44_{5.52}$	$64.00_{1.55}$	$73.06_{0.49}$	$84.15_{0.29}$	$50.93_{0.36}$	$60.86_{1.19}$
CPL (ICML'24)	$49.27_{0.69}$	$61.42_{1.22}$	$17.58_{0.32}$	$82.77_{0.37}$	$63.49_{6.24}$	$61.91_{1.04}$	$72.11_{0.23}$	$86.28_{0.30}$	$49.31_{0.59}$	$60.46_{1.22}$
UEO (ICML'24)	$53.56_{0.11}$	$60.07_{0.01}$	$18.73_{0.15}$	$92.74_{0.08}$	$43.68_{0.14}$	$65.71_{0.07}$	$80.67_{0.04}$	$87.94_{0.09}$	$58.03_{0.18}$	$62.35_{0.09}$
TMP (ICML'25)	$18.49_{0.18}$	$39.78_{0.26}$	$10.97_{0.12}$	$82.76_{0.26}$	$16.80_{0.30}$	$13.16_{0.29}$	$59.64_{0.65}$	$33.96_{3.59}$	$19.46_{0.10}$	$32.78_{0.64}$
CAP (ICML'25)	$51.75_{0.55}$	$68.83_{2.66}$	$16.02_{2.07}$	$89.72_{3.51}$	$66.37_{2.67}$	$60.88_{1.73}$	$80.68_{0.37}$	$86.64_{1.41}$	$58.26_{0.43}$	$64.35_{1.71}$
microCLIP (ACL'26)	$49.88_{0.59}$	$67.98_{0.87}$	$17.28_{0.87}$	$93.01_{0.12}$	$51.76_{1.29}$	$67.96_{1.34}$	$79.94_{0.15}$	$88.77_{0.94}$	$58.94_{0.34}$	$63.95_{0.24}$
MARS (Ours)	$\mathbf{54.84_{0.35}}$	$\mathbf{68.85_{0.41}}$	$\mathbf{20.24_{0.27}}$	$\mathbf{94.06_{0.51}}$	$\mathbf{74.37_{0.29}}$	$\mathbf{68.95_{1.00}}$	$\mathbf{81.52_{0.17}}$	$\mathbf{90.97_{0.42}}$	$\mathbf{62.38_{0.52}}$	$\mathbf{68.47_{0.20}}$

Imbalance Ratio = 20

Method	CUB.	Res.	FA.	Cal.	ES.	Flw.	Food.	Pets.	Cars.	Avg.
ZS CLIP	$51.82_{0.00}$	$54.48_{0.00}$	$17.58_{0.00}$	$90.30_{0.00}$	$32.88_{0.00}$	$63.67_{0.00}$	$78.80_{0.00}$	$84.36_{0.00}$	$58.16_{0.00}$	$59.12_{0.00}$
FPL (NeurIPS'23)	$51.04_{0.26}$	$60.03_{0.75}$	$18.46_{0.21}$	$90.30_{0.09}$	$51.89_{1.32}$	$60.87_{1.52}$	$77.48_{0.15}$	$86.77_{0.33}$	$57.51_{0.41}$	$61.60_{0.56}$
GRIP (NeurIPS'23)	$47.63_{0.75}$	$59.27_{1.74}$	$17.20_{0.20}$	$87.75_{0.18}$	$60.01_{2.57}$	$62.26_{0.91}$	$70.86_{0.27}$	$81.85_{0.67}$	$49.65_{0.93}$	$59.61_{0.91}$
CPL (ICML'24)	$48.46_{0.39}$	$58.45_{0.32}$	$16.81_{0.23}$	$79.99_{3.02}$	$64.35_{0.34}$	$60.63_{0.94}$	$70.14_{0.30}$	$85.30_{0.31}$	$47.83_{0.61}$	$59.11_{0.72}$
UEO (ICML'24)	$53.20_{0.09}$	$58.69_{0.11}$	$17.28_{1.01}$	$92.75_{0.06}$	$43.15_{0.03}$	$66.13_{0.04}$	$80.48_{0.09}$	$87.61_{0.10}$	$57.82_{0.12}$	$61.90_{0.18}$
TMP (ICML'25)	$15.61_{0.52}$	$32.71_{0.73}$	$8.10_{0.05}$	$69.82_{0.67}$	$12.27_{1.69}$	$9.51_{0.62}$	$50.46_{0.45}$	$22.03_{1.43}$	$15.11_{0.14}$	$26.18_{0.70}$
CAP (ICML'25)	$49.89_{0.57}$	$59.94_{3.42}$	$13.47_{1.82}$	$88.15_{4.36}$	$62.07_{0.75}$	$59.96_{0.62}$	$78.88_{1.49}$	$82.05_{3.28}$	$56.08_{0.49}$	$61.17_{1.87}$
microCLIP (ACL'26)	$49.11_{0.38}$	$66.47_{0.58}$	$16.70_{0.23}$	$92.89_{0.20}$	$52.65_{4.81}$	$66.53_{1.55}$	$79.46_{0.19}$	$87.98_{0.75}$	$56.89_{0.78}$	$63.19_{0.54}$
MARS (Ours)	$\mathbf{53.80_{0.82}}$	$\mathbf{68.42_{0.58}}$	$\mathbf{20.45_{0.84}}$	$\mathbf{94.04_{0.15}}$	$\mathbf{72.17_{3.73}}$	$\mathbf{67.30_{0.86}}$	$\mathbf{81.52_{0.26}}$	$\mathbf{90.32_{0.26}}$	$\mathbf{61.18_{0.86}}$	$\mathbf{67.69_{0.52}}$

Imbalance Ratio = 50

Method	CUB.	Res.	FA.	Cal.	ES.	Flw.	Food.	Pets.	Cars.	Avg.
ZS CLIP	$51.82_{0.00}$	$54.48_{0.00}$	$17.58_{0.00}$	$90.30_{0.00}$	$32.88_{0.00}$	$63.67_{0.00}$	$78.80_{0.00}$	$84.36_{0.00}$	$58.16_{0.00}$	$59.12_{0.00}$
FPL (NeurIPS'23)	$50.60_{0.32}$	$58.07_{0.51}$	$17.95_{0.66}$	$89.67_{0.75}$	$49.53_{2.32}$	$60.09_{2.26}$	$76.78_{0.35}$	$86.41_{0.54}$	$56.76_{0.29}$	$60.65_{0.89}$
GRIP (NeurIPS'23)	$47.57_{0.28}$	$56.48_{0.98}$	$16.47_{0.24}$	$86.52_{0.47}$	$53.84_{3.69}$	$60.99_{1.23}$	$68.99_{0.73}$	$81.14_{0.80}$	$49.26_{0.84}$	$57.92_{1.03}$
CPL (ICML'24)	$47.88_{0.43}$	$56.64_{0.28}$	$16.43_{0.23}$	$77.73_{2.75}$	$59.76_{6.11}$	$60.58_{0.93}$	$68.24_{0.19}$	$82.94_{0.10}$	$46.71_{0.80}$	$57.43_{1.31}$
UEO (ICML'24)	$52.46_{0.06}$	$56.96_{0.17}$	$17.49_{0.52}$	$92.77_{0.06}$	$42.37_{0.05}$	$66.01_{0.07}$	$80.24_{0.05}$	$86.99_{0.03}$	$57.23_{0.17}$	$61.39_{0.13}$
TMP (ICML'25)	$12.86_{0.05}$	$26.21_{0.12}$	$5.53_{0.37}$	$60.45_{0.74}$	$11.33_{1.64}$	$8.21_{0.25}$	$38.63_{0.52}$	$14.31_{1.35}$	$12.49_{0.29}$	$21.11_{0.59}$
CAP (ICML'25)	$49.11_{0.27}$	$57.29_{1.31}$	$12.77_{1.70}$	$86.66_{5.99}$	$52.88_{1.60}$	$58.43_{0.68}$	$77.89_{3.03}$	$77.29_{5.61}$	$53.10_{0.30}$	$58.38_{2.28}$
microCLIP (ACL'26)	$48.11_{0.60}$	$63.73_{0.79}$	$15.36_{0.42}$	$92.25_{0.51}$	$48.16_{3.79}$	$64.88_{0.68}$	$79.22_{0.29}$	$87.93_{0.43}$	$55.47_{0.83}$	$61.68_{0.53}$
MARS (Ours)	$\mathbf{54.29_{0.68}}$	$\mathbf{68.16_{0.47}}$	$\mathbf{20.05_{0.26}}$	$\mathbf{93.18_{0.08}}$	$\mathbf{69.22_{5.60}}$	$\mathbf{68.13_{1.00}}$	$\mathbf{81.24_{0.47}}$	$\mathbf{90.00_{0.79}}$	$\mathbf{60.55_{0.46}}$	$\mathbf{67.20_{0.90}}$

2023), Candidate Pseudo-Labeling (CPL; Zhang et al., 2024), Universal Entropy Optimization (UEO; Liang et al., 2024), True-false labels via Multimodal Prompt retrieving (TMP; Li et al., 2025), Concept-Adaptive Pseudo-labeling (CAP; Wang et al., 2025), and Coarse-Fine Token Fusion (microCLIP; Silva et al., 2026). For fair comparison, we standardize the visual backbone to OpenAI CLIP ViT-B/32 and use the same training dataset and evaluation metrics. All other settings follow the official recommendations.

Implementation details. Following the standard protocol, we set the multi-modal prompts as MaPLe (Khattak et al., 2023), with a depth of 9 and a context length of 2, initialized with the template "a photo of a ". The weak view follows the standard CLIP preprocessing, while the strong view is generated by RandAugment (Cubuk et al., 2020). We fix $\beta = 1$ and the momentum coefficient ($\lambda = 0.95$) across all datasets. For a comprehensive evaluation of different VLMs, we also report results using OpenCLIP ViT-B/32 (LAION-2B) (Cherti et al., 2023) and SigLIP ViT-B/16 (Zhai et al., 2023). Additional reproduction details can be found in Appendix C.

4.2 Main Results

Our method achieves SOTA classification accuracy. Table 1 demonstrates that our method outperforms both zero-shot CLIP and existing unsupervised methods by a substantial margin across nine benchmark datasets under varying imbalance ratios (i.e., 10, 20, 50). Specifically, the performance of our proposed method exceeds the previous state-of-the-art method by an average of **4.12**, **4.50**, and **5.52** percentage points (pp) across three seeds under imbalance ratios of 10, 20, and 50, respectively. In the most challenging scenario with an imbalance ratio of 50, our method demonstrates remarkable robustness. While CAP suffers a 5.97 pp drop as the ratio increases from 10 to 50, our method

Table 4: Ablation results of Boundary-Preserving Alignment (BPA) and Margin-aware Self-Refinement (MSR) with different imbalance ratios. The improvement is indicated by ↑.

BPA	MSR	Imbalance Ratio = 10			Imbalance Ratio = 20			Imbalance Ratio = 50			Avg.
		Res.	Flw.	ES.	Res.	Flw.	ES.	Res.	Flw.	ES.	
✗	✗	54.48	63.67	32.88	54.48	63.67	32.88	54.48	63.67	32.88	50.34
✓	✗	67.05	65.23	57.26	64.69	63.98	61.92	63.34	64.20	50.68	62.04 ↑11.70
✗	✓	61.50	68.42	52.16	61.71	67.33	52.26	58.65	66.92	52.18	60.13 ↑ 9.79
✓	✓	**69.43**	**69.41**	**74.40**	**68.52**	**67.73**	**74.22**	**68.66**	**67.98**	**73.68**	**70.45** ↑20.11

exhibits high stability with a minimal decline of only 1.27 pp. Notably, at an imbalance ratio of 50, four of the seven baselines even fall below zero-shot CLIP on average, whereas MARS still improves over it by **8.08** pp. Moreover, under an imbalance ratio of 50, our method improves RESISC45 accuracy over the best existing method by **4.43** pp. Furthermore, pairwise t-tests at a 0.05 significance level using three corresponding seed results show that our method achieves significant improvements in **84.7%** of cases (160 out of 189), further confirming its effectiveness in ULTA. Detailed results can be found in Appendix D.2.

Our method maintains performance stability and robustness under a uniform distribution. Using the same comparison protocol as Section 4.1, we further evaluate our method on three datasets (Caltech101, Flowers102, and OxfordPets) in a uniform setting. As reported in Table 2, MARS outperforms existing unsupervised methods. Specifically, MARS achieves the highest average accuracy of 86.27%, surpassing the leading baseline by 0.30 pp. This validates that our method enhances feature discriminability without compromising robustness, ensuring reliable performance across unknown distributions.

Table 2: Comparison of test accuracy (%) under the balanced setting.

Method	Cal.	Flw.	Pets.	Avg.
ZS CLIP	90.30	63.67	84.36	79.44
FPL	92.08	67.47	87.41	82.32
CPL	92.37	68.20	90.32	83.63
UEO	92.98	66.82	89.13	82.98
CAP	94.36	72.35	91.22	85.98
microCLIP	94.12	72.11	89.92	85.38
MARS	**94.89**	**72.35**	**91.58**	**86.27**

Our method is effective across different VLMs. To verify the universality of our method, we extend our evaluation to different vision-language models. As illustrated in Table 3, our method consistently improves performance on OpenCLIP across four benchmarks. Specifically, it achieves an average accuracy gain of **8.79** pp over the best-performing baseline for OpenCLIP across the three imbalance ratios. These results indicate that the proposed model serves as a generalized solution that enhances the robustness of various VLMs against long-tailed distributions. Results with imbalance ratios of 10 and 50 and additional results on SigLIP can be found in Appendix D.3.

Table 3: Test accuracy (%) using OpenCLIP with an imbalance ratio of 20. Zero-shot OpenCLIP is abbreviated as ZS OCLIP.

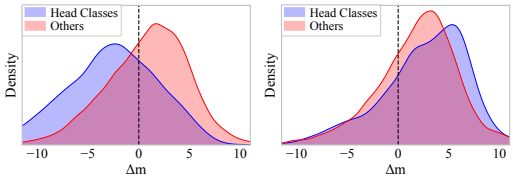
Method	Res.	Flw.	ES.	Pets.	Avg.
ZS OCLIP	63.90	71.43	51.74	90.62	69.42
FPL	60.59	66.04	50.32	87.54	66.12
GRIP	63.25	62.08	61.24	80.84	66.85
CPL	62.68	63.41	64.44	84.60	68.78
UEO	59.57	70.35	52.06	88.88	67.72
TMP	41.07	13.60	10.00	47.75	28.11
CAP	63.25	61.68	63.30	80.54	67.19
microCLIP	66.55	66.16	58.56	89.21	70.12
MARS	**73.12**	**74.22**	**80.88**	**92.40**	**80.16**

4.3 Ablation Studies

All components contribute to the proposed method. As shown in Table 4, both Boundary-Preserving Alignment (BPA) and Margin-aware Self-Refinement (MSR) individually result in noticeable performance improvements over zero-shot CLIP across all settings. Specifically, BPA and MSR yield substantial average improvements of **11.70** pp and **9.79** pp, respectively. Compared with MSR, BPA exhibits a larger improvement. A possible reason is that BPA mitigates feature drift induced by noisy pseudo-labels and effectively prevents severe boundary erosion, which is the primary cause of performance degradation. Moreover, the combination of the two components consistently yields the highest accuracy, with a total boost of **20.11** pp, highlighting their complementary effects on improving feature discriminability. The full model also stays stable across imbalance ratios, dropping only **0.97** pp from 10 to 50. We further examine the impact of visual encoder capacity with a ViT-B/16 backbone in Appendix D.4, where MARS outperforms the strongest baseline by **3.82** pp on average.

4.4 More Analyses

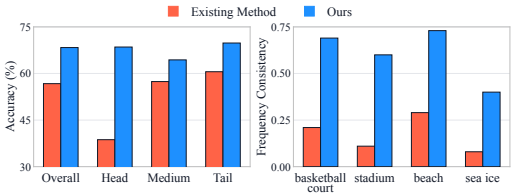
Does boundary-preserving alignment (BPA) effectively reduce head-class boundary erosion? To quantify the impact of BPA, we visualize the density curve of logit margin changes (Δm) for CPL and our method in Fig. 4. The negative shift in the head-class margin reveals that CPL enforces a uniform constraint at the cost of eroding the head-class boundary. In contrast, our method rectifies this issue by ensuring that the margin distributions for all classes are shifted to the positive region ($\Delta m > 0$), preventing boundary erosion. BPA also reduces head-to-tail misassignment: on RESISC45 with $\gamma = 50$, CPL increases the number of head-class samples predicted as tail classes from 428 to 1,060, whereas BPA reduces it to 392. More datasets and a sample-level margin analysis can be found in Appendix D.6.



(a) Existing method (b) Ours

Figure 4: Density curve of logit margin changes (Δm). (a) Existing methods cause a negative shift for head classes. (b) Our method maintains positive margin changes.

Does margin-aware self-refinement (MSR) effectively improve tail and confusable classes without compromising head classes? As shown in Fig. 5, to verify whether MSR effectively addresses this dilemma, we analyze the performance across different class groups and the frequency consistency (i.e., the alignment between predicted and ground-truth class distributions) in confusable classes. Specifically, our method achieves substantial accuracy gains in the tail and medium groups while preserving performance on head classes, confirming that discriminability is enhanced without knowledge forgetting. Furthermore, the significantly higher frequency consistency in confusable classes demonstrates that MSR effectively rectifies prediction errors caused by the model's inherent bias. More fine-grained results and analyses can be found in Appendix D.5.



(a) (b)

Figure 5: Evaluation of margin-aware self-refinement on RESISC45. (a) Accuracy across head, medium, and tail groups. (b) Frequency consistency on confusable classes.

4.5 More Challenging Settings

Our method remains effective on large-scale long-tailed benchmarks. To further evaluate the scalability and robustness of our method, we run additional experiments on large-scale long-tailed benchmarks (Liu et al., 2019; Yang et al., 2022; Van Horn et al., 2018; 2021) (Places-LT, ImageNet-LT, and an iNaturalist2018 subset, with imbalance ratios of 996, 256, and 500). As shown in Table 5, MARS achieves the best performance on all three benchmarks, improving the average accuracy by **2.51** pp. Notably, it improves the best baseline by **3.16** pp on iNaturalist2018. Base-to-new generalization results are provided in Appendix D.7.

Table 5: Comparison of test accuracy (%) on large-scale long-tailed datasets.

Method	Places-LT	ImageNet-LT	iNaturalist2018	Avg.
ZS CLIP	38.33	59.71	69.33	55.79
FPL	39.31	60.05	69.50	56.29
GRIP	36.64	59.92	66.60	54.39
CPL	38.38	55.77	65.30	53.15
UEO	39.71	60.23	70.17	56.70
TMP	32.18	48.32	42.17	40.89
CAP	37.86	61.34	70.00	56.40
microCLIP	40.39	58.62	69.00	56.00
MARS	**41.35**	**62.96**	**73.33**	**59.21**

5 Conclusion

In this paper, we investigate unsupervised long-tailed adaptation (ULTA) of vision-language models. We find that imposing uniform constraints erodes the head-class boundary while reinforcing prediction bias in confusable classes. To address this, we propose margin-aware refinement with structural alignment (MARS), which integrates boundary-preserving alignment to maintain the head-class boundary with margin-aware self-refinement to enhance the discriminability of tail and confusable classes. Extensive experiments across nine benchmarks with varying imbalance ratios validate the effectiveness of MARS, with an average improvement of **4.71** percentage points.

AI USE STATEMENT

Generative AI tools were used solely to assist with language editing and polishing of the manuscript. All scientific content, analyses, and conclusions were developed and verified by the authors.

REPRODUCIBILITY STATEMENT

We provide the methodological and experimental details necessary to reproduce our results in the paper and supplementary materials. The supplementary materials also include the source code.

REFERENCES

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 446–461, 2014.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829, 2023.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 18613–18624, 2020.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 178–178, 2004.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision (IJCV)*, 132(2):581–595, 2024.

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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3942–3951, 2021.

Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 4904–4916, 2021.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 709–727, 2022.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19113–19122, 2023.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 554–561, 2013.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 12888–12900, 2022.

Zhongnian Li, Jinghao Xu, Peng Ying, Meng Wei, and Xinzheng Xu. Learning from true-false labels via multi-modal prompt retrieving. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 36652–36666, 2025.

Jian Liang, Lijun Sheng, Zhengbo Wang, Ran He, and Tieniu Tan. Realistic unsupervised clip fine-tuning with universal entropy optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 29667–29681, 2024.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2537–2546, 2019.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Cristina Menghini, Andrew Delworth, and Stephen Bach. Enhancing clip with clip: Exploring pseudolabeling for limited-label prompt tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 60984–61007, 2023.

Muhammad Jehanzeb Mirza, Leonid Karlinsky, Wei Lin, Horst Possegger, Mateusz Kozinski, Rogerio Feris, and Horst Bischof. Lafter: Label-free tuning of zero-shot classifier using language and unlabeled image collections. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5765–5777, 2023.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, pp. 722–729, 2008.

Yassine Ouali, Adrian Bulat, Brais Matinez, and Georgios Tzimiropoulos. Black box few-shot adaptation for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15488–15500, 2023.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3498–3505, 2012.

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 *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Jiang-Xin Shi, Chi Zhang, Tong Wei, and Yu-Feng Li. Efficient and long-tailed generalization for pre-trained vision-language model. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 2663–2673, 2024.

Sathira Silva, Eman Ali, Chetan Arora, and Muhammad Haris Khan. microclip: Unsupervised clip adaptation via coarse-fine token fusion for fine-grained image classification. In *Findings of the Association for Computational Linguistics (ACL)*, pp. 33277–33294, 2026.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8769–8778, 2018.

Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12884–12893, 2021.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical report, California Institute of Technology, 2011.

Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 33536–33549, 2022.

Yuchen Wang, Xuefeng Bai, Xiucheng Li, Weili Guan, Liqiang Nie, and Xinyang Chen. Handling imbalanced pseudolabels for vision-language models with concept alignment and confusion-aware calibrated margin. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 62357–62372, 2025.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4818–4829, 2024.

Lu Yang, He Jiang, Qing Song, and Jun Guo. A survey on long-tailed visual recognition. *International Journal of Computer Vision (IJCV)*, 130(7):1837–1872, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, 2023.

Jiahan Zhang, Qi Wei, Feng Liu, and Lei Feng. Candidate pseudolabel learning: Enhancing vision-language models by prompt tuning with unlabeled data. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 60004–60020, 2024.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 130(9):2337–2348, 2022b.

TABLE OF CONTENTS FOR THE APPENDIX

A Training Process of MARS 14

B Dataset Details 14

C Implementation Details 14

D More Experiment Results and Analyses 15

D.1 Sensitivity of Hyperparameters 15

D.2 Statistical Significance Analysis 15

D.3 Comparison with Different VLMs 15

D.4 Comparison with Different Visual Backbones 16

D.5 Fine-grained Results across Head, Medium, and Tail Groups 16

D.6 Head-to-Tail Misassignment Analysis 17

D.7 Base-to-new Generalization 17

A TRAINING PROCESS OF MARS

Algorithm 1 provides the complete training procedure of MARS.

Algorithm 1 Training Process of MARS

Input: Unlabeled dataset $\mathcal{D}_u$, alignment epochs T_{align}, and maximum epochs T_{max}.
Output: Trained prompts $\theta_{visual}, \theta_{textual}$ and student textual adapter ϕ_S.
Initialize prompts $\theta_{textual}, \theta_{visual}$ and teacher textual adapter ϕ_T.
Cache zero-shot features and predictions on the weak view, and construct class anchors.
// Boundary-Preserving Alignment (Sec. 3.3)
for epoch = 1 **to** T_{align} **do**
    For all $x_i \in \mathcal{D}_u$, obtain frozen a_i and current p_i on the same weak view.
    Compute visual conditions h_{ij}, pairwise coefficients ρ_{cj} by Eq. (3), and targets by Eq. (4).
    With fixed q_i, update $\theta_{textual}, \theta_{visual}, \phi_T$ on strong-view minibatches via Eq. (5).
end for
// Margin-aware Self-Refinement (Sec. 3.4)
Freeze teacher's predictions, θ_{visual} and ϕ_T; compute class prior $r_{\hat{y}}$ and initialize ϕ_S.
for epoch = $T_{align} + 1$ **to** T_{max} **do**
    Obtain visual feature $\mathbf{v}$ via frozen θ_{visual} and textual embedding $\mathbf{t}$ via $\theta_{textual} \rightarrow \phi_S$;
    Update CMC and compute gating factor $\hat{\alpha}$ by Eq. (9);
    Compute $\mathcal{L}_{\text{refine}}$ and update $\theta_{textual}, \phi_S$ via gradient descent;
end for

B DATASET DETAILS

We evaluate our method on a diverse set of benchmark datasets to demonstrate its robustness across different domains. Table 6 summarizes the detailed statistics of these datasets, including the class count and the number of samples in training and test splits. Note that we include datasets with both balanced and imbalanced test sets to conduct a comprehensive evaluation.

Table 6: Statistics for benchmark datasets used in main experiments. The rightmost column indicates whether the testing set is balanced.

Dataset	Classes	Train	Test	Balanced
CUB (Wah et al., 2011)	200	5,994	5,794	✗
RESISC45 (Cheng et al., 2017)	45	6,300	25,200	✓
FGVCAircraft (Maji et al., 2013)	100	6,667	3,333	✓
Caltech101 (Fei-Fei et al., 2004)	100	5,777	2,465	✗
EuroSAT (Helber et al., 2019)	10	22,000	5,000	✓
Flowers102 (Nilsback & Zisserman, 2008)	102	2,040	6,149	✗
Food101 (Bossard et al., 2014)	101	70,700	30,300	✓
OxfordPets (Parkhi et al., 2012)	37	3,680	3,669	✗
StanfordCars (Krause et al., 2013)	196	8,144	8,041	✗

C IMPLEMENTATION DETAILS

We rerun all baseline methods from their official public codebases under a unified protocol. Specifically, we standardize the visual backbone to OpenAI CLIP ViT-B/32 and use the same training dataset, data split, and evaluation metrics for all methods. All other settings follow the recommendations in the corresponding official implementations, such as the training schedule and optimization hyperparameters. For MARS, in Stage 1, we use a batch size of 64 and train the model for 10 epochs with the SGD optimizer, optimizer momentum 0.9, initial learning rate 0.0035, weight decay 5e-4, and a linear-warmup cosine learning-rate schedule. In Stage 2, we train the student model for 30 epochs. We fix $\beta = 1$ and the momentum coefficient $\lambda = 0.95$ for all datasets. In the balanced setting of Table 2, we further set the imbalance ratio to 1 (cap the maximum number of training samples

per class at 100). We conduct the experiments on a single NVIDIA RTX 4090 GPU using PyTorch v2.8.0.

D More Experiment Results and Analyses

D.1 Sensitivity of Hyperparameters

In the previous experiments, we fix $\beta = 1$ across datasets. We evaluate the sensitivity of this exponent in Fig. 6 on Caltech101 and RESISC45.

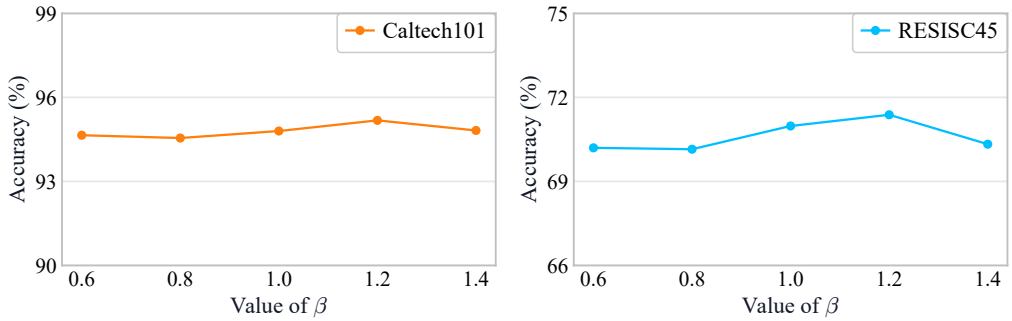


Figure 6: Sensitivity to β on Caltech101 (left) and RESISC45 (right).

D.2 Statistical Significance Analysis

To further validate the statistical significance of the improvements, we conduct two-sided paired t-tests at a 0.05 significance level against seven baseline methods using the three corresponding seed results. As shown in Table 7, our method achieves significant improvements in **84.7%** of cases (160 out of 189), confirming its effectiveness under different imbalance ratios.

Table 7: Statistical significance of performance differences assessed with a two-sided paired t-test at a 0.05 significance level over three corresponding seeds, reported as win/tie/loss counts. Imbalance Ratio is abbreviated as IR.

Method	IR = 10	IR = 20	IR = 50	Total
FPL	8 / 1 / 0	9 / 0 / 0	9 / 0 / 0	26 / 1 / 0
GRIP	9 / 0 / 0	8 / 1 / 0	8 / 1 / 0	25 / 2 / 0
CPL	8 / 1 / 0	8 / 1 / 0	8 / 1 / 0	24 / 3 / 0
UEO	9 / 0 / 0	6 / 3 / 0	7 / 2 / 0	22 / 5 / 0
TMP	9 / 0 / 0	9 / 0 / 0	9 / 0 / 0	27 / 0 / 0
CAP	4 / 5 / 0	7 / 2 / 0	7 / 2 / 0	18 / 9 / 0
microCLIP	6 / 3 / 0	7 / 2 / 0	5 / 4 / 0	18 / 9 / 0
Total	53 / 10 / 0	54 / 9 / 0	53 / 10 / 0	160 / 29 / 0

D.3 Comparison with Different VLMs

To verify the generalization capability of our approach, we complement the OpenCLIP results in Table 3 with imbalance ratios of 10 and 50 (Table 8), and further replace CLIP with SigLIP (Table 9). It can be observed that our method maintains a significant performance advantage over previous baselines on both VLMs, demonstrating that our strategy is effective across different VLMs. Notably, with SigLIP, our method surpasses the best-performing baseline by **5.10** pp on RESISC45 under the imbalance ratio of 50.

Table 8: Comparison of test accuracy (%) using OpenCLIP with imbalance ratios of 10 and 50. Zero-shot OpenCLIP is abbreviated as ZS OpenCLIP.

Method	Imbalance Ratio = 10				Imbalance Ratio = 50				Avg.
	Res.	Flw.	ES.	Pets.	Res.	Flw.	ES.	Pets.	
ZS OpenCLIP	63.90	71.43	51.74	90.62	63.90	71.43	51.74	90.62	69.42
FPL	64.05	66.43	50.10	88.03	59.27	65.18	49.98	86.92	66.25
GRIP	64.63	64.29	62.08	82.47	57.94	66.12	60.20	81.22	67.37
CPL	66.66	63.20	64.18	86.67	59.65	63.70	60.42	81.38	68.23
UEO	62.49	70.16	53.54	88.66	58.77	70.22	51.44	89.02	68.04
TMP	45.98	16.34	10.08	60.70	34.51	11.79	10.00	35.43	28.10
CAP	67.60	65.95	65.36	88.28	58.32	61.67	50.36	73.45	66.37
microCLIP	66.44	67.52	59.30	89.62	64.19	65.46	57.52	88.96	69.88
MARS	**72.10**	**75.64**	**76.74**	**92.12**	**70.10**	**73.52**	**72.34**	**91.77**	**78.04**

Table 9: Comparison of test accuracy (%) using SigLIP with varying imbalance ratios (10, 20, 50). Zero-shot SigLIP is abbreviated as ZS SigLIP.

Method	Imbalance Ratio = 10				Imbalance Ratio = 20				Imbalance Ratio = 50				Avg.
	Res.	Flw.	ES.	Pets.	Res.	Flw.	ES.	Pets.	Res.	Flw.	ES.	Pets.	
ZS SigLIP	59.23	83.46	35.74	93.10	59.23	83.46	35.74	93.10	59.23	83.46	35.74	93.10	67.88
FPL	61.14	82.96	37.12	92.80	60.04	83.36	35.14	92.45	60.23	83.48	35.62	92.83	68.10
GRIP	62.36	83.72	45.38	91.52	61.73	83.66	42.62	91.71	60.62	83.72	34.72	91.11	69.41
CPL	65.56	81.33	46.74	90.92	62.76	81.22	52.14	88.91	60.50	82.50	44.24	88.72	70.46
UEO	59.65	83.48	40.10	93.08	59.55	83.46	39.94	93.10	59.40	83.46	39.74	93.08	69.00
TMP	19.25	7.82	16.98	28.75	17.05	6.05	15.36	20.58	14.27	3.85	16.84	13.44	15.02
CAP	72.62	84.48	50.18	86.83	69.36	85.15	46.86	84.73	63.03	83.59	47.64	82.03	71.38
microCLIP	69.64	85.41	56.02	93.49	68.00	84.45	48.94	92.61	65.59	82.45	46.62	92.31	73.79
MARS	**75.15**	**89.98**	**68.36**	**94.99**	**74.03**	**88.96**	**71.92**	**95.28**	**70.69**	**89.15**	**61.46**	**94.88**	**81.24**

D.4 Comparison with Different Visual Backbones

We further investigate the scalability of our model by upgrading the visual backbone to a higher-capacity ViT-B/16 encoder. The results presented in Table 10 indicate that MARS maintains a performance advantage over existing baselines under all imbalance ratios. Specifically, our method achieves the highest average accuracy of **74.41%**, outperforming the strongest baseline by **3.82** pp on average, with consistent gains across all four datasets. This validates that our approach is not only robust across different VLMs but also capable of effectively leveraging the enhanced features provided by the stronger vision backbone.

Table 10: Comparison of test accuracy (%) using ViT-B/16 as the visual backbone with imbalance ratios of 10, 20, and 50.

Method	Imbalance Ratio = 10				Imbalance Ratio = 20				Imbalance Ratio = 50				Avg.
	Res.	Cub.	Flw.	Pets.	Res.	Cub.	Flw.	Pets.	Res.	Cub.	Flw.	Pets.	
ZS CLIP	62.25	55.32	71.15	88.20	62.25	55.32	71.15	88.20	62.25	55.32	71.15	88.20	69.23
FPL	65.02	56.81	66.74	89.40	63.02	55.50	65.08	89.29	61.38	55.35	65.88	88.85	68.53
GRIP	65.47	54.17	69.18	86.26	61.22	53.86	69.52	85.88	58.41	52.12	67.67	84.08	67.32
CPL	66.72	54.98	67.23	88.80	62.83	53.27	66.08	86.24	58.16	53.33	67.18	84.06	67.41
UEO	61.62	57.18	70.92	89.62	60.75	56.59	70.73	89.59	60.15	56.61	70.60	89.45	69.48
TMP	46.50	20.04	13.21	43.66	38.52	16.07	9.73	29.93	31.21	12.72	7.25	18.86	23.98
CAP	69.87	56.29	69.34	89.18	65.60	56.61	69.18	84.71	63.89	54.88	67.83	83.37	69.23
microCLIP	69.01	53.78	72.48	90.54	67.75	51.66	71.10	90.30	66.87	51.19	71.56	90.84	70.59
MARS	**72.32**	**59.84**	**74.87**	**93.65**	**70.51**	**58.56**	**75.77**	**93.24**	**69.26**	**58.28**	**74.14**	**92.45**	**74.41**

D.5 Fine-grained Results across Head, Medium, and Tail Groups

To complement the group-level analysis, we report detailed head, medium, and tail results on RESISC45 and EuroSAT in Table 11. The table shows that our method improves not only overall

accuracy but also strongly restores head-class performance while maintaining medium- and tail-class performance. For example, on RESISC45, the head accuracy improves by **18.65** pp over UEO and **32.57** pp over CAP.

Table 11: Comparison of test accuracy (%) across head, medium, and tail groups on RESISC45 and EuroSAT with an imbalance ratio of 50.

Method	Res.				ES.			
	Overall	Head	Medium	Tail	Overall	Head	Medium	Tail
FPL	57.61	46.59	58.52	59.72	55.51	23.40	60.20	58.52
CPL	56.73	38.69	57.40	60.57	58.82	13.00	65.53	63.10
UEO	56.97	51.90	54.43	59.83	42.42	6.80	44.67	47.23
CAP	58.78	37.98	59.88	63.05	50.38	26.40	60.47	49.33
MARS	**68.70**	**70.55**	**67.01**	**69.11**	**75.68**	**78.60**	**76.60**	**74.73**

D.6 Head-to-Tail Misassignment Analysis

To examine the head-class degradation in Section 3.2, we track how true head-class test samples are predicted after adaptation. H-to-H and H-to-T denote the numbers of these samples misclassified as head and tail classes, respectively. As shown in Table 12, CPL reduces the head accuracy on RESISC45 from 69.20% to 38.69% and increases the H-to-T count from 428 to 1,060, while H-to-H decreases from 429 to 164. BPA restores the head accuracy to 72.89% and reduces the H-to-T count to 392 without sacrificing the tail accuracy. OxfordPets shows the same pattern.

Table 12: Head/tail accuracy (%) and prediction flow of true head-class samples on RESISC45 and OxfordPets with an imbalance ratio of 50.

Dataset	Method	Head Acc.	Tail Acc.	H-to-H	H-to-T
RESISC45	ZS CLIP	69.20	56.84	429	428
	CPL	38.69	60.57	164	1,060
	BPA	**72.89**	**65.37**	210	**392**
OxfordPets	ZS CLIP	86.75	91.47	28	18
	CPL	80.12	92.14	17	56
	BPA	**88.14**	**92.24**	30	**16**

We further analyze all true head-class samples that CPL predicts as tail classes on Caltech101 and Food101 with an imbalance ratio of 50. For each sample x with ground-truth label y, we compute the logit margin change relative to zero-shot CLIP, $\Delta m = m_{\text{Adapted}}(x, y) - m_{\text{Zero-shot CLIP}}(x, y)$, where $m(x, y) = z_y(x) - \max_{j \neq y} z_j(x)$. As shown in Table 13, CPL yields strongly negative margin changes on these samples, whereas BPA turns them positive. Moreover, BPA no longer predicts most of these samples as tail classes and recovers the correct label for 94.12% of them on Caltech101 and 61.11% on Food101.

Table 13: Average logit margin changes relative to zero-shot CLIP and repairs over all true head-class samples predicted as tail classes by CPL on Caltech101 and Food101 with an imbalance ratio of 50. "No longer tail" counts the samples that BPA no longer predicts as tail classes, and "Exact repairs" counts those that BPA predicts correctly.

Dataset	Cases	CPL Δm	BPA Δm	No longer tail	Exact repairs
Caltech101	204	-3.88	$+2.28$	199 (97.55%)	192 (94.12%)
Food101	1,579	-4.41	$+1.04$	1,204 (76.25%)	965 (61.11%)

D.7 Base-to-new Generalization

Beyond standard closed-set classification, we further evaluate base-to-new generalization. Following the standard 1:1 base/new split, we train only on unlabeled long-tailed base-class samples and evaluate on the base and new test sets using the harmonic mean. As shown in Table 14, our method consistently

outperforms previous methods on both datasets. Specifically, on RESISC45, MARS improves the best prior harmonic mean from **58.92%** to **62.16%**, yielding a gain of **3.24** pp. These results validate that our method preserves strong cross-class generalization under long-tailed adaptation.

Table 14: Base-to-new generalization on RESISC45 and OxfordPets (imbalance ratio=50).

Method	Res.			Pets.		
	Base	New	H	Base	New	H
ZS CLIP	55.36	53.56	54.45	84.77	83.95	84.36
FPL	69.05	51.39	58.92	83.68	86.63	85.13
GRIP	64.15	42.65	51.23	77.67	85.23	81.28
CPL	66.33	45.27	53.81	82.46	85.68	84.04
UEO	60.02	49.66	54.35	85.06	88.59	86.79
CAP	66.90	44.71	53.60	82.24	78.36	80.25
microCLIP	67.76	44.71	53.87	80.75	88.93	84.64
MARS	**69.59**	**56.16**	**62.16**	**89.10**	**89.49**	**89.29**