IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 March 2025, accepted 7 April 2025, date of publication 16 April 2025, date of current version 30 April 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3561581

RESEARCH ARTICLE

LM-CLIP: Adapting Positive Asymmetric Loss for Long-Tailed Multi-Label Classification

CHRISTOPH TIMMERMANN, SEUNGHYEON JUNG, MISO KIM, AND WOOJIN LEE
Graduate School of Computer Science and Artificial Intelligence, Dongguk University, Jung-gu, Seoul 04620, Republic of Korea

Corresponding author: Woojin Lee (wj926@dgu.ac.kr)

This work was supported in part by Stockfolio Inc., and the Ministry of Science and Information and Communication Technology (MSIT), South Korea, through the Information Technology Research Center (ITRC) Support Program, under Grant IITP-2025-2020-0-01789; in part by the Artificial Intelligence Convergence Innovation Human Resources Development, Supervised by the Institute for Information and Communications Technology Planning and Evaluation (IITP), under Grant IITP-2025-RS-2023-00254592; and in part by the National Research Foundation of Korea (NRF) Grant funded by Korean Government under Grant RS-2025-00556289.

ABSTRACT Accurate multi-label image classification is essential for real-world applications, especially in scenarios with long-tailed class distributions, where some classes appear frequently while others are rare. This imbalance often leads to biased models that struggle to accurately recognize underrepresented classes. Existing methods either trade off performance between head and tail classes or rely on image captions, limiting adaptability. To address these limitations, we propose *LM-CLIP*, a novel framework built around a unified loss function. Our *Balanced Asymmetric Loss (BAL)* extends traditional asymmetric loss by emphasizing the gradients of rare positive samples where the model is uncertain, mitigating bias toward dominant classes. This is complemented by a contrastive loss that pushes negative samples further from the decision boundary, creating a more optimal embedding space even in long-tailed scenarios. These loss functions together ensure balanced performance across all classes. Our framework is built on pre-trained models utilizing textual and visual features from millions of image-text pairs. Furthermore, we incorporate a dynamic sampling strategy that prioritizes rare classes based on their occurrence, which ensures effective training without compromising overall performance. Experiments conducted on VOC-MLT and COCO-MLT benchmarks demonstrate the effectiveness of our approach, achieving +4.66% and +8.14% improvements in mean Average Precision (mAP) over state-of-the-art methods. Our code is publicly available at https://github.com/damilab/lm-clip.

INDEX TERMS Long-tailed learning, multi-label classification, CLIP, vision-language models, contrastive learning, class imbalance, loss functions, asymmetric loss, balanced asymmetric loss, imbalanced sampling.

I. INTRODUCTION

In recent years, image recognition has gained widespread popularity, largely due to advancements in computer vision. Vision-Language Models (VLMs) employing contrastive representation learning, notably Contrastive Language-Image Pre-Training (CLIP) [1], improve image recognition accuracy by using extensive large datasets of image-text pairs for pre-training, especially in zero-shot and few-shot learning [2], [3], [4], [5]. Furthermore, there has been a rise in models using CLIP for multi-label recognition tasks [6], [7], [8], [9], [10], which is crucial for real-world applications.

Despite their success, existing VLMs face significant challenges when applied to long-tailed datasets, where the distribution of labels is highly imbalanced [11]. Fine-tuning with cross-entropy loss, or even more advanced methods like Focal Loss [12] and Asymmetric Loss (ASL) [13] are prone to overfitting on head classes due to the disproportionate focus on common classes in the dataset, leaving tail classes underrepresented. Due to this, balancing performance across head and tail classes in multi-label classification tasks remains an unresolved issue.

The associate editor coordinating the review of this manuscript and approving it for publication was Wenming Cao.

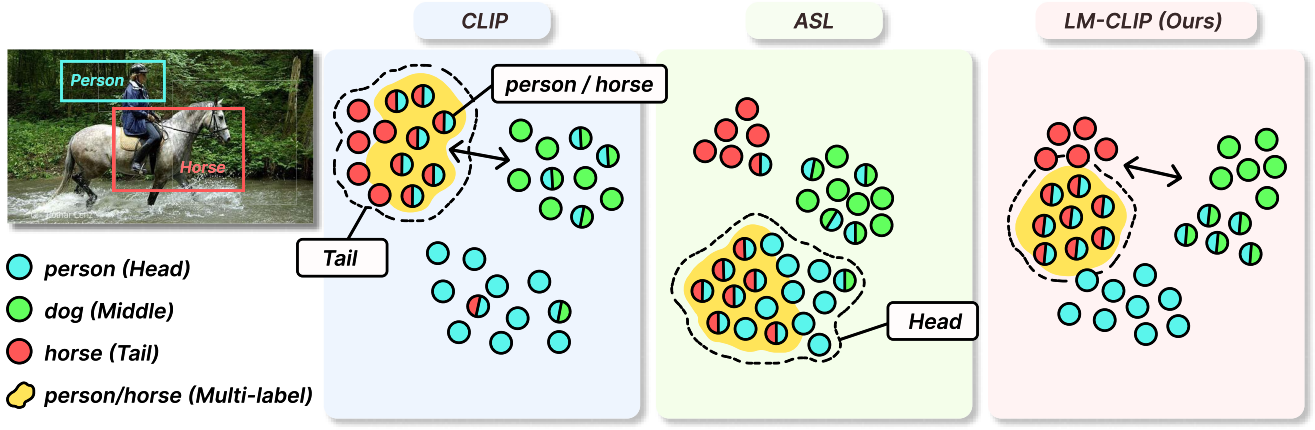


FIGURE 1. Visualization of the effects of loss functions on the embedding space. We represent the embeddings of samples for each label using different colors, and for cases with multi-labels, we divide the colors in half. Specifically, for samples with both 'horse' and 'person' labels, their embedding regions are highlighted in yellow. The CLIP-loss tends to recognize multi-labels as tail classes, whereas the ASL tends to recognize multi-labels as head classes. Our model aims to correctly classify the multi-labeled data by learning appropriate representations without being biased towards either head or tail classes.

However, it is particularly noteworthy that CLIP-based models tend to predict multi-labels as tail classes more frequently in long-tailed multi-label tasks. We explore the CLIP-loss, a method that treats every pair except the positive pair as negatives, therefore handling each sample as distinct. While it may lead to reduced performance for common head classes, it could also prevent excessive grouping with high-frequency classes, enhancing performance for tail classes. This trend is also observed in studies using datasets like Visual Object Classes Multi-Long-Tailed (VOC-MLT) and Common Objects in Context Multi-Long-Tailed (COCO-MLT) [14], where head class performance tends to decline, suggesting that the dataset characteristics might also contribute to this outcome [15], [16], [17].

Binary Cross-Entropy (BCE) computes the loss for each label separately using binary classification and then combines these losses to classify multiple labels per image effectively [18], while Focal Loss [12], with a focusing parameter, aims to down-weight contributions from easy negatives. However, it encounters challenges in handling rare easy positive samples due to the potential elimination of their gradients. To address this, ASL [13] introduces *Asymmetric Focusing*, where the focusing parameter is asymmetrically designed to differentiate between positive and negative losses, with a focus on hard negatives. However, in highly imbalanced scenarios, Asymmetric Focusing tends to ignore rare positives, which leads to a tendency to predict multi-labels as head classes more frequently.

Beyond the exploration of loss functions, recently, there have been methods that use prompt learning like CoOp [19], CoCoOp [20], DualCoOp [21], Tal-DPT [22] for multi-labels, and LMPT [17] for long-tailed multi-labels to efficiently adapt pre-trained CLIP embeddings for image classification. However, the current state-of-the-art, LMPT, depends on natural language image captions for training [17]. Our method outperforms them in long-tailed scenarios relying solely on actual ground truth labels provided in the dataset.

Figure 1 visually summarizes the characteristics of existing loss functions. In the case of CLIP, the contrastive loss treats all data as distinct, resulting in reasonable prediction performance for the less frequent tail classes but showing lower performance for the head classes. On the other hand, using supervised loss functions like ASL tends to classify data into the more numerous head classes, leading to decreased performance on the tail classes. In this paper, we propose a methodology that achieves high performance in classifying multi-label data without being biased towards either head or tail classes and does not rely on captions like LMPT [17].

We introduce *Long-tailed Multi-label Contrastive Language-Image Pretraining (LM-CLIP)* to adapt positive ASL for long-tailed scenarios while also utilizing the strengths of contrastive loss. First, we propose a novel *Balanced Asymmetric Loss (BAL)* that refines the positive loss of ASL to enhance prediction performance for positive classes by introducing class weighting. Our approach to class weighting for rare positives and unbalanced sampling is grounded in class occurrence, particularly in accurately classifying rare classes. Subsequently, our loss function ensures balanced attention to each class during training. In combination with contrastive loss, the model learns a balanced representation where both head and tail classes are well-grouped, effectively embedding multi-labels in positions that span between head and tail classes, as illustrated in Figure 1.

This makes LM-CLIP well-suited for long-tailed multi-label classification tasks, which commonly arise in real-world domains such as medical imaging [23] and environmental monitoring [24], where datasets have both positive/negative label imbalance and skewed class distributions.

LM-CLIP is also practical for deployment in scenarios where image captions or text annotations are unavailable or unreliable. By avoiding dependence on noisy captions, we maintain adaptability and robustness across diverse domains. Despite requiring fine-tuning and hyperparameter selection, our method scales especially well with larger vision-language backbones, delivering consistent improvements in classification performance across head, middle, and tail classes.

We achieve state-of-the-art performance on two multi-label long-tailed datasets. The total mean Average Precision (mAP) on VOC-MLT and COCO-MLT reaches up to 93.82% and 77.87% respectively. Additionally, we propose a new metric to quantitatively demonstrate that our model learns good representations by bringing positive samples closer and pushing negative samples further apart. This metric also validates that our model effectively classifies multi-label data. Our findings are further supported by visualizations and analysis of various samples.

II. RELATED WORK

A. CLIP ON MULTI-LABEL CLASSIFICATION

VLMs like CLIP [1] have significantly improved image recognition by learning contrastive representations from large-scale image-text pairs. CLIP's contrastive loss optimizes embeddings by treating every image-text pair as distinct, even if multiple samples share the same or overlapping classes. This approach helps in open-set recognition but is suboptimal for multi-label classification, where different samples may share common labels. As a result, CLIP tends to focus on the most dominant label while overlooking secondary ones. Fine-tuning addresses this limitation by explicitly modeling label relationships, allowing better capturing shared attributes across samples. By integrating a supervised loss, fine-tuning ensures that instances with overlapping labels are positioned more meaningfully in the embedding space, leading to improved multi-label classification performance.

Recent studies have introduced CLIP-based models to enhance multi-label classification by utilizing the features of CLIP, even with limited annotations and zero-shot challenges. CLIP-Driven Unsupervised Learning (CDUL) [6] applies CLIP to unsupervised multi-label image classification. It overcomes the typical single-label limitation by using global and local alignments to better capture multiple labels within an image. Without labeled data, CDUL generates pseudo-labels based on CLIP's predictions and updates them through training. CLIP-Decoder [7] takes a dual-modal approach, fusing vision and language features by integrating transformer decoder layers. Instead of relying solely on predefined text prompts, CLIP-Decoder uses a customized prompting approach where templates are designed and tested for each class. This optimizes zero-shot multi-label classification by better aligning visual and textual features. They incorporate visual embeddings directly into the transformer decoder layers, ensuring that textual predictions are grounded in image content.

B. FINE-TUNING CLIP

When adapting large VLMs to tasks with limited or imbalanced data, a key challenge is catastrophic forgetting, where performance on unseen classes deteriorates as the model overfits to the few available training samples. CLIP-CITE [25] shows how a simple fine-tuning strategy rapidly decreases training loss yet plateaus on test accuracy, thus confirming that fully fine-tuning CLIP on limited data risks losing its broader generalization capabilities. To combat this, CLIP-CITE introduces a supervised visual-text alignment objective plus knowledge distillation from the original CLIP, thereby reducing catastrophic forgetting and improving downstream performance. Less Overfitting for Better Generalization (LOBG) [26] tackles this by filtering out fine-grained foreground details, thereby reducing overemphasis on specific regions of the image that lead to overfitting. We address catastrophic forgetting and overfitting in long-tailed scenarios by combining a supervised loss with CLIP's contrastive pre-training objective. These approaches demonstrate that carefully controlling how fine-tuning interacts with a pretrained representation can dramatically reduce overfitting and maintain strong generalization.

C. CLIP PROMPT TUNING

Prompt tuning has been widely explored to adapt CLIP for classification tasks while avoiding full model fine-tuning. Instead of modifying CLIP's encoders or projection, these methods learn specific optimized text prompts for each class and/or dataset. Context Optimization (CoOp) [19] first automates prompt engineering for vision-language models, addressing the need for manually designed prompts for classification. It introduces learnable text tokens that are placed along the class tokens as context, which are optimized through cross-entropy classification loss. Conditional Context Optimization (CoCoOp) [20] extends this through image-conditional prompts, which dynamically adapt based on the input image. DualCoOp [21] further refines prompt learning by incorporating both positive and negative contexts in prompts.

These CoOp-based methods use a fixed number of labeled training examples per class, usually 1 to 16. Due to this, there is no need to handle unbalanced distributions in the classification loss. However, this limits the amount of data that can be used for training. Considering this, it is evident that these models are influenced by the characteristics of the data they utilize. Further research is needed to develop methods that can handle various types of data distributions, including long-tailed scenarios.

D. LONG-TAILED DISTRIBUTION LEARNING

In the domain of multi-label classification with long-tailed distributions, various loss functions have been studied to

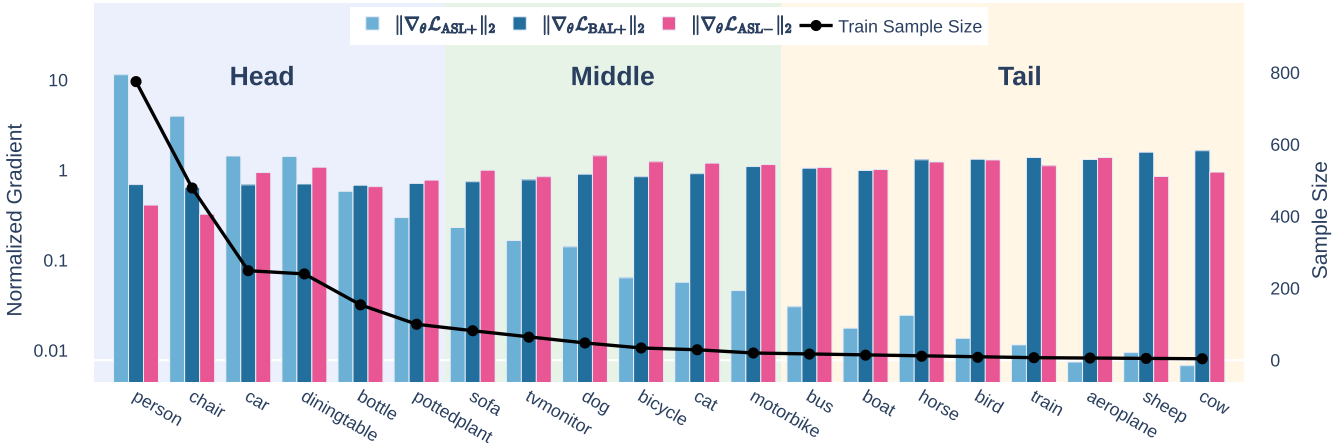


FIGURE 2. Gradient analysis of the positive and negative components of ASL and the proposed loss function (BAL). To investigate the effects of the positive and negative loss of ASL and BAL per class using the same model (Pre-trained CLIP), we calculate the mean of the gradient norm per class over one epoch. We have normalized the size of the gradient norm. The gradient norm of negative loss ($\|\nabla_\theta \mathcal{L}_{ASL-}\|_2$) is similar across all classes. In contrast, for the positive loss, ASL ($\|\nabla_\theta \mathcal{L}_{ASL+}\|_2$) yields very small gradients for training samples corresponding to tail classes. However, our positive BAL loss ($\|\nabla_\theta \mathcal{L}_{BAL+}\|_2$) is reweighted per class, resulting in a balanced gradient norm across most classes.

enhance performance. Probability Guided Loss (PG Loss) [16] refines multi-label classification by adjusting predicted probabilities to prevent overfitting and scaling gradients based on class probabilities, thus better balancing negative and positive labels across diverse datasets.

More recently, Distributionally Robust (DR) Loss [27] refines the Log-Sum-Exp Pairwise (LSEP) loss [28] by incorporating class-wise computation and a negative gradient constraint (NGC). Their method effectively reduces overconfidence in negative labels while preserving the advantages of ranking-based multi-label classification. Unlike standard LSEP, which tends to rely excessively on label co-occurrence, DR Loss enhances feature separation between classes, leading to improved performance on long-tailed datasets. This work is particularly relevant to our study as it demonstrates the effectiveness of loss function modifications in balancing head and tail class performance, which aligns with our approach of optimizing ASL for long-tailed distributions.

There has also been research focused on representation learning to address long-tailed distributions, including studies like the visual-linguistic long-tailed recognition framework (VL-LTR) [29], and those that explore different sampling methods [30].

In label-aware learning, methods to adjust label importance dynamically have been explored. Label Distribution Guided Hashing (LADH) [31] introduces a feature-induced label distribution mechanism that assigns different weights to labels based on their significance in multi-modal data. This approach allows multi-label hashing models to better capture label correlations and avoid treating all labels equally, an issue often present in traditional supervised hashing methods.

Similarly, in long-tailed multi-label classification, rare classes are underrepresented due to imbalanced occurrences. LADH's concept of label-weighted hashing aligns with strategies that aim to reweight loss functions based on class frequency to balance head and tail classes. Our BAL ensures a positive label gradient balance, similar to LADH's semantic enhancement strategy. While LADH focuses on cross-modal retrieval, the idea of label distribution-based optimization is relevant to multi-label classification settings in long-tailed scenarios.

E. PROMPT LEARNING FOR LONG-TAILED DISTRIBUTIONS

Prompt learning has also been adapted to address long-tailed problems on multi-label classification tasks. Prompt tuning on long-tailed multi-label visual recognition (LMPT) [17], based on CoOp [19], introduces a class-specific embedding loss function. This loss function incorporates class-aware soft margins and re-weighting, leveraging textual descriptions to establish semantic relationships between head and tail classes. This approach is effective at handling imbalanced class distributions, with its ability to enhance recognition of tail classes by improving their relationship with the more frequent head classes. However, LMPT has several limitations. Notably, it underperforms on head classes compared to prior models due to its focus on the tail classes. Moreover, LMPT requires a natural language caption for each training image, which may not always be feasible or efficient in certain contexts.

1) RELATIONS TO OUR METHOD

Our proposed method takes the fine-tuning approach and represents a significant improvement over previous methods [12], [13], [16]. While Distribution-Balanced Loss (DBL) [14] and Probability-Guided Loss (PG Loss) [16] address the class imbalance in a similar way, they do not account for the natural positive/negative label imbalance. Our loss benefits from the foundation provided by ASL [13]. It is designed to ensure that both positive and negative labels

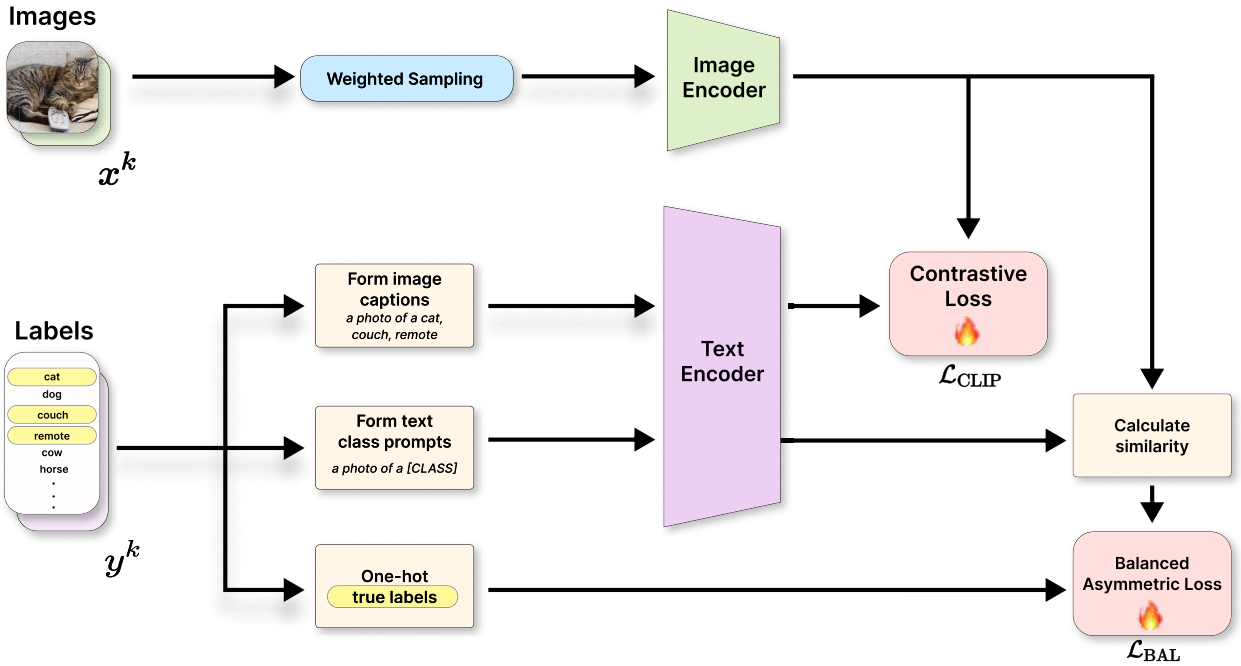


FIGURE 3. Overview of our architecture. ***LM-CLIP*** leverages the pre-trained joint embedding space for images and text of CLIP. $\mathcal{L}_{\text{CLIP}}$ maximizes the cosine similarity of the image and caption embeddings. We combine it with our proposed loss function $\mathcal{L}_{\text{BAL}}$ used on the similarity scores and true labels.

contribute effectively to training, which is important for rare tail classes. In addition to ASL, we incorporate the CLIP contrastive loss into our training. When fine-tuning CLIP, it acts as a form of regularization, stabilizing the process and reducing overfitting, particularly on head classes. This is crucial in long-tailed problems where the model may tend to overfit the more frequent head classes, neglecting the tail classes. Label smoothing further enhances this regularization effect by softening the target labels, thus preventing the model from becoming overly confident in its predictions [32]. We further enhance the robustness of our model by incorporating weighted sampling based on class occurrence, a method shown to be effective in handling long-tailed distributions [33].

While LMPT [17] achieves competitive performance in long-tailed multi-label scenarios, our approach offers an advantage by addressing its key shortcomings, namely the reliance on image captions and the underperformance on head classes.

III. LM-CLIP

A. PROBLEM SETTINGS

The long-tailed multi-label classification problem setting in this paper is as follows: Let N represent the total number of samples in the dataset D, which is divided into training and testing sets D_{train} and D_{test}. The number of classes is denoted by C. Training samples are defined as $(x^k, y^k) \in D_{\text{train}}$ for $k \in \{1, \ldots, N_{\text{train}}\}$ where x^k is an input image, $y^k = [y_1^k, \ldots, y_C^k] \in \{0, 1\}^C$ is the multi-label ground truth vector. During testing, samples are defined as $(x^k, y^k) \in D_{\text{test}}$ for $k \in \{1, \ldots, N_{\text{test}}\}$. Our task is to predict the presence or absence of given classes on each test image.

The key challenge in this setting is the severe imbalance across classes, where some labels appear frequently while others are rare. This imbalance requires careful loss function design and sampling strategies to ensure fair training of all classes.

B. ASYMMETRIC LOSS

The choice of ASL over conventional loss functions such as BCE or Focal loss [12] is based on the need to explicitly balance positive and negative contributions in a multi-label setting [13].

The original ASL is defined as:

$$\text{ASL} = \begin{cases} \mathcal{L}_{\text{ASL+}} = (1-p)^{\gamma_+} \log(p) \\ \mathcal{L}_{\text{ASL-}} = (p_m)^{\gamma_-} \log(1-p_m) \end{cases} \tag{1}$$

$$\mathcal{L}_{\text{ASL}} = -y\mathcal{L}_{\text{ASL+}} - (1-y)\mathcal{L}_{\text{ASL-}} \tag{2}$$

where p is the probability of positives, p_m is the shifted probability of negatives defined as $p_m = \max(p - m, 0)$, and γ_+ and γ_- are hyperparameters for positive and negative asymmetric focusing. The parameter m is a probability shift hyperparameter.

It is applied between the one-hot encoded ground truth class labels and class probabilities, calculated via dot similarity between image and class prompt embeddings. ASL focuses on hard negatives while preserving gradients from rare positives. A probability shift allows very easy negative samples to be discarded from the training loss. [13] However,

TABLE 1. mAP performance comparison of various methods on VOC-MLT and COCO-MLT for RN-50, ViT-B/16, and ViT-L/14 visual encoders. ✓ indicates only the ground truth labels of the dataset were used. All ViT-L/14, CLIP pre-trained, and LM-CLIP results are reproduced by us, all others are taken from [16], [17], and [27]. RN-50 methods above the line are vision-only, below are vision-language models.

Methods	Only uses label info.	VOC-MLT [14] Total↑	Head↑	Middle↑	Tail↑	COCO-MLT [14] Total↑	Head↑	Middle↑	Tail↑
RN-50									
Focal Loss [12]	✓	73.88	69.41	81.43	71.56	49.46	49.80	54.77	42.14
DR Loss [27]	✓	78.01	72.19	83.83	77.69	54.10	50.27	57.00	53.76
DB Focal [14]	✓	78.94	73.22	84.18	79.30	53.55	51.13	57.05	51.06
PG Loss [16]	✓	80.37	73.67	83.83	82.88	54.43	51.23	57.42	53.40
LTML [15]	✓	81.44	**75.68**	85.53	82.69	56.90	**54.13**	60.59	54.47
CLIP pre-trained [1]	N/A	83.83	64.20	87.37	95.91	58.54	38.65	60.98	**72.82**
CoOp [34]	✓	81.34	65.10	81.54	93.37	54.94	38.06	56.67	67.51
CoCoOp [20]	✓	78.63	64.33	80.51	87.94	46.02	36.02	50.57	48.82
DualCoOp [21]	✓	81.03	66.45	80.53	92.33	53.11	40.48	55.20	62.11
TaI-DPT [22]		83.75	66.27	85.17	94.57	56.23	40.52	58.40	66.09
LMPT [17]		85.44	66.62	88.11	**97.86**	**58.97**	41.87	61.60	69.60
LM-CLIP (Ours)	✓	**86.14**	70.66	**88.70**	95.83	58.48	40.18	**61.74**	70.28
ViT-B/16									
CLIP pre-trained [1]	N/A	85.84	67.24	88.65	97.67	63.87	42.11	67.52	78.21
CoOp [34]	✓	86.02	67.71	88.79	97.67	60.68	41.97	63.18	73.85
CoCoOp [20]	✓	84.47	64.58	87.82	96.88	61.49	39.81	64.63	76.42
LMPT [17]		87.88	72.10	89.26	98.49	66.19	44.89	69.80	79.08
LM-CLIP (Ours)	✓	**92.54**	**82.84**	**94.42**	**98.50**	**74.33**	**63.12**	**78.09**	**79.23**
ViT-L/14									
CLIP pre-trained [1]	N/A	85.53	65.81	88.84	97.83	65.86	43.21	68.79	81.93
CoOp [34]	✓	84.71	68.69	84.23	97.08	67.41	46.93	70.22	81.72
CoCoOp [20]	✓	83.97	67.64	85.45	95.12	58.60	36.98	58.66	77.54
LMPT [17]		82.70	65.13	80.12	97.81	65.37	42.60	67.97	81.98
LM-CLIP (Ours)	✓	**93.82**	**86.28**	**94.83**	**98.72**	**77.87**	**63.34**	**83.12**	**83.73**

on the other hand, ASL can overemphasize common positive samples if the dataset is long-tailed.

Figure 2 presents an analysis of the gradient norms ($\|\nabla_\theta \mathcal{L}_{\text{ASL}+}\|_2$ and $\|\nabla_\theta \mathcal{L}_{\text{ASL}-}\|_2$) of each loss per class during the training of one epoch. Upon examination, we observe that the negative loss of ASL ($\|\nabla_\theta \mathcal{L}_{\text{ASL}-}\|_2$) yields gradients of similar magnitude across all classes. However, in the case of long-tailed data, the gradient norm of the positive loss ($\|\nabla_\theta \mathcal{L}_{\text{ASL}+}\|_2$) on tail classes is significantly smaller. Consequently, as shown in Table 2, ASL exhibits low performance on tail classes.

C. BALANCED ASYMMETRIC LOSS

ASL is effective at emphasizing hard negatives while ignoring easy negatives, which solves the natural problem of few positive and many negative samples. However, it tends to underweight rare positive samples, as seen in Figure 2, leading to poor performance on tail classes.

To address this, we propose *Balanced Asymmetric Loss (BAL)*, an intuitive solution aiming to improve performance on tail classes. Since ASL focuses on adjusting the loss for negative samples, we introduce novel modifications to the positive loss component. We incorporate label smoothing and class weighting for positive samples, aimed at providing more gradient value on rare classes in the training set and achieving better accuracy on long-tailed datasets. Our *Balanced Asymmetric Loss* is formulated as:

$$\text{BAL} = \begin{cases} \mathcal{L}_{\text{BAL}+} = -\mathbf{p_w}(1-p)^{\gamma_+}\log(p) \\ \mathcal{L}_{\text{ASL-}} = (p_m)^{\gamma_-}\log(1-p_m) \end{cases} \tag{3}$$

$$\mathcal{L}_{\text{BAL}} = -y_{\text{ls}}\mathcal{L}_{\text{BAL}+} - (1-y)L_{\text{ASL-}} \tag{4}$$

where y_{ls} is the label-smoothed ground truth $y_{\text{ls}} = (1-\epsilon)y + \epsilon/C$ with ϵ as the smoothing amount hyperparameter. $\mathbf{p_w}$ is a vector of weights for all classes C, $\mathbf{p_w} = [w_1, \ldots w_C]$. The weight for a class c is defined as:

$$w_c = \frac{\mu_c}{\max(\mu)} \quad \mu_c = (N/N_c)^s \tag{5}$$

where N is the total number of training samples, N_c the number of positive samples of class c, and s a weighting exponent hyperparameter used to control the strength of class weighting. $\max(\mu)$ denotes the maximum class weight ($\mu = [\mu_1, \ldots \mu_C]$). These modifications adjust the significance of rare positive classes within the loss function, which is essential for multi-label long-tailed datasets, where positive samples are inherently limited due to the multi-label structure [13], and even more so for minority classes. Figure 2 demonstrates that compared to the positive loss of ASL ($\|\nabla_\theta \mathcal{L}_{\text{ASL}+}\|_2$), the gradient norm of our positive loss function ($\|\nabla_\theta \mathcal{L}_{\text{BAL}+}\|_2$) is uniform across all classes and even shows a slight increase towards the tail classes, depending on the hyperparameter s.

TABLE 2. mAP performance comparison of the components of our architecture. The used image encoder backbone is ViT-B/16. No dataset captions were used. ✓ denotes strategies used during training for each result. "avg.Δ" is average performance improvement over the baseline (pre-trained CLIP).

CLIP Loss	ASL Loss	Weighted Sampling	Label Smoothing	Loss Weighting	VOC-MLT [14] Total↑	Head↑	Medium↑	Tail↑	avg.Δ	COCO-MLT [14] Total↑	Head↑	Medium↑	Tail↑	avg.Δ
					85.84	67.24	88.65	97.67		63.87	42.11	67.52	78.21	
	✓				89.93	81.13	90.99	95.73	+4.60	62.36	54.89	65.45	64.85	-1.04
✓					91.02	79.10	92.98	98.50	+5.55	71.37	57.51	74.67	79.23	+7.77
	✓		✓	✓	91.41	80.05	94.73	97.43	+6.06	74.94	64.18	80.03	77.68	**+11.28**
✓	✓				91.93	82.35	93.33	98.06	+6.57	71.61	59.86	75.76	76.46	+8.00
✓	✓	✓			92.35	82.82	94.10	98.18	+7.01	71.94	61.48	75.83	76.00	+8.39
✓	✓	✓	✓		92.26	82.11	94.39	98.27	+6.91	74.33	63.07	78.92	78.17	+10.70
✓	✓	✓	✓	✓	92.54	82.84	94.42	98.50	**+7.23**	74.33	63.12	78.09	79.23	+10.77

We utilize label smoothing (denoted by y_{ls}) to distribute some probability mass from positive classes to negative classes, resulting in less confident predictions [32], especially on the rare tail classes.

In summary, our loss function retains the negative component of ASL while modifying the positive component to mitigate the effects of class imbalance, hard labels and to reduce overfitting, which is particularly important for long-tailed datasets and large models.

D. CLASS-WEIGHTED SAMPLING

Instead of uniformly sampling from the training dataset, we oversample rare class samples with replacement. We calculate a weight for each sample in D_{train} based on class weights, so the weight of a sample k is given as $w_k = \frac{\mu_k}{\max(\mu)}$ $\mu_k = \sum_{c \in y_+^k} w_c$, where y_+^k is the set of positive classes of sample k, and w_c is the calculated weight of class c. $\max(\mu)$ denotes the maximum sample weight and is used to normalize them to be between 0 and 1.

Class-weighted sampling improves training of tail classes but also reduces overfitting on head classes. This shift from head to tail samples can be controlled through the exponent hyperparameter s in Equation 5.

E. LM-CLIP

We use the pre-trained joint embedding space for images and text of CLIP, which allows us to leverage its rich semantic understanding for our classification task.

For each training sample, we build a text caption in the format "a photo of a [CLASS1], [CLASS2], ..." using the ground truth labels y^k. Class prompts for inference are formed in the same way as in CLIP ("a photo of a [CLASS]"). Images x^k are encoded through the image encoder and projected into the same latent space as captions and class prompts. All parts of the encoders are kept unchanged and unfrozen during training.

CLIP's loss $\mathcal{L}_{\text{CLIP}}$ optimizes embeddings by treating every sample as distinct, which can be beneficial in open-set recognition but does not account for shared labels in multi-label classification, so we combine it with our proposed loss function $\mathcal{L}_{\text{BAL}}$. The two loss functions are balanced by a hyperparameter λ:

$$\mathcal{L}_{\text{LM-CLIP}} = \mathcal{L}_{\text{CLIP}} + \lambda \mathcal{L}_{\text{BAL}} \tag{6}$$

$\mathcal{L}_{\text{CLIP}}$ maximizes the cosine similarity of the matching image and caption embeddings in the batch. For this, a symmetric cross-entropy loss [1] is applied to the similarity scores, aligning each image-text pair with its corresponding match while simultaneously decreasing the similarity with all other samples in the batch. The overall architecture of our method is shown in Figure 3.

IV. EXPERIMENTS

A. DATASETS

We conduct experiments on two long-tailed multi-label image classification datasets: VOC-MLT and COCO-MLT. These datasets are subsets introduced in [14] derived from established benchmarks, PascalVOC [35] and MS-COCO [36], respectively. To emulate real-world scenarios with long-tailed class distributions, the datasets are sampled using a pareto distribution probability density function [14].

1) VOC-MLT

The VOC-MLT dataset [14] comprises 1,142 training images and 4,592 test images containing 20 classes. The class distribution ranges from a maximum of 775 images per class to a minimum of 4 images per class. The ratio of head, medium, and tail classes is 6:6:8.

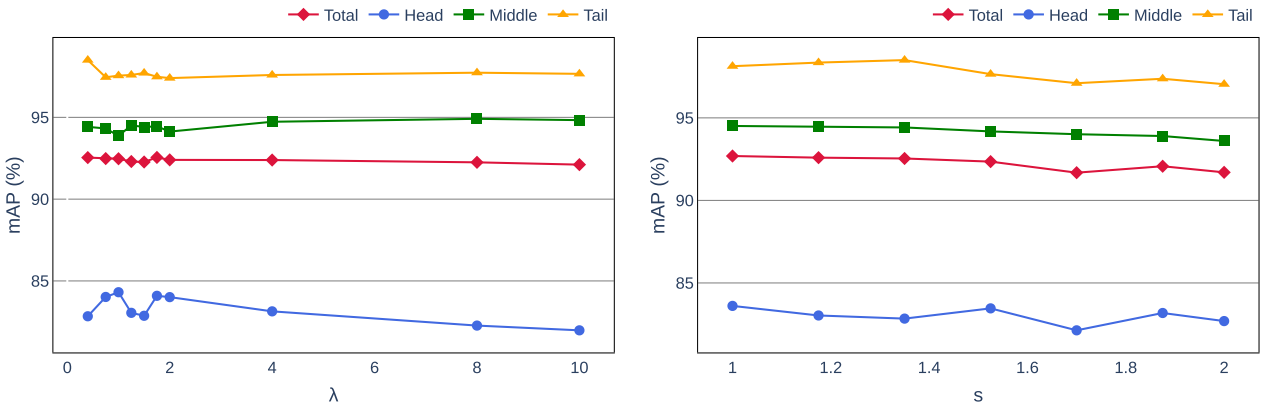
2) COCO-MLT

The COCO-MLT dataset [14] consists of 1,909 training images and 5,000 test images containing 80 classes. The class distribution varies from a maximum of 1,128 images per class to a minimum of 6 images per class. The ratio of head, medium, and tail classes is 22:33:25.

B. EXPERIMENTAL SETTINGS

1) EVALUATION METRICS

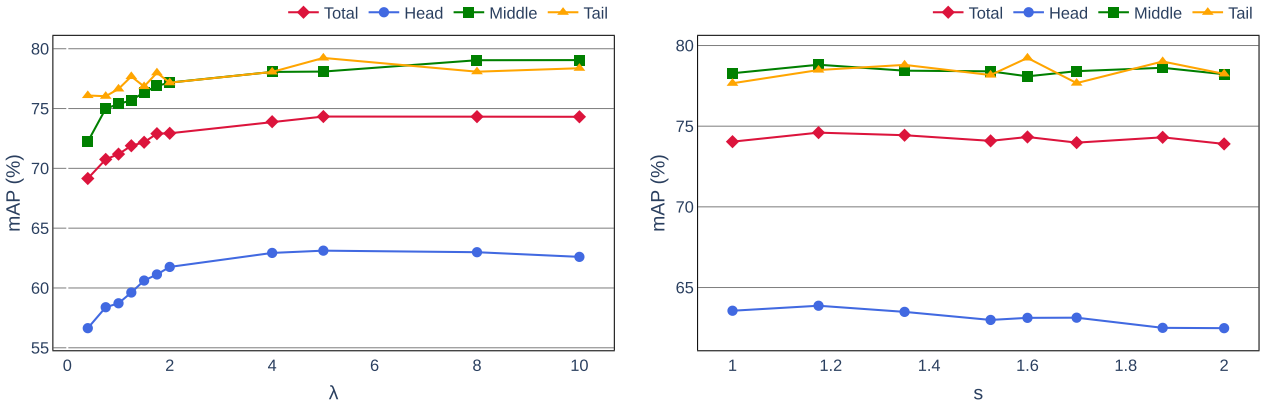
Our primary evaluation metric is the mean Average Precision (mAP) computed across all classes. Additionally, we analyze the performance of classes grouped into head, middle, and tail categories based on the number of samples per class. All the classes are categorized into three groups based on the number of training samples per class: head classes contain more than 100 samples, medium classes have between 20 and 100 samples, and tail classes have fewer than 20 samples.



(a) Effect of λ hyperparameter on mAP.

(b) Effect of s hyperparameter on mAP.

FIGURE 4. Comparison of mAP total, head, middle and tail with different λ and s on VOC-MLT. The used vision encoder is ViT-B/16.



(a) Effect of λ hyperparameter on mAP.

(b) Effect of s hyperparameter on mAP.

FIGURE 5. Comparison of mAP total, head, middle and tail with different λ and s on COCO-MLT. The used vision encoder is ViT-B/16.

2) IMPLEMENTATION DETAILS

We utilize the pre-trained CLIP model by OpenAI [1] with either ResNet-50 (RN-50), ViT-B/16, or ViT-L/14 as the image encoder backbone. The hyperparameter λ is empirically chosen to balance the influence of contrastive loss and our $\mathcal{L}_{\text{BAL}}$. Sample selection during training is conducted with replacement, with oversampling of rare classes and undersampling of common classes. This sampling factor is influenced by the hyperparameter exponent s, which is adjusted for each dataset.

3) DATA AUGMENTATION

During training, images are processed in batches, resized to a square of size 224×224, and then normalized in the same manner as during CLIP's pretraining [1]. We employ data augmentation by randomly flipping training images horizontally with a probability of 0.5 during sampling.

4) COMPUTE DETAILS

All experiments were performed on a machine with a Ryzen 5 5600X CPU, one RTX 4090, and 32GB RAM. The operating system used is Ubuntu 22.04.4 LTS. All required Python packages and their versions are listed in the environment.yml file in the public repository.

C. HYPERPARAMETER DETAILS

During hyperparameter search, the search space was as follows:

$8 \leq$ Batch size ≤ 64 , $2.0 \leq \text{ASL}\gamma_- \leq 12.0$, $0.1 \leq \lambda \leq 10.0$, $1.0 \leq s \leq 2.0$, ϵ: $0.01 \leq$ Label smoothing $\epsilon \leq 0.1$

For both datasets and image encoders, we use AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and weight decay $\lambda_{\text{wd}} = 0.01$. We utilize a cosine annealing learning rate scheduler with $T_{\max} = 50$. Batch size is 8. For ASL parameters, we use $\gamma_+ = 0$ and $p_m = 0.05$ in all experiments.

D. EXPERIMENTAL RESULTS

We compare our method with existing approaches on both VOC-MLT and COCO-MLT datasets, presenting mAP results for head, middle, and tail classes in Table 1. For fair comparison, we utilize the same data augmentation techniques as LMPT [17] and the default CLIP class prompt

TABLE 3. Hyperparameter settings for VOC-MLT and COCO-MLT for RN-50 and all vision transformers.

Hyperparameter	VOC-MLT [14]		COCO-MLT [14]	
	RN-50	ViT	RN-50	ViT
λ	0.2	0.4	6.0	5.0
s	0.7	1.35	1.35	1.6
Label smoothing ϵ	0.01	0.1	0.1	0.1
ASLγ_-	12.0	9.8	9.8	9.8
Learning rate α	$2e-6$	$1e-6$	$1e-7$	$1e-6$

"a photo of a [CLASS]". Compared to LMPT using the ViT-B/16 visual encoder, our model improves the total mAP by +4.66% and +8.14% on VOC-MLT and COCO-MLT respectively. Our method is the new *state-of-the-art*, without depending on captions for training images.

Not requiring dataset captions is particularly effective for VOC-MLT, given the lower quality captions generated through an image captioning model by the authors of LMPT [17]. They also note that the manually annotated COCO-MLT captions contain numerous errors, including missing or inconsistent label words. Because our model performs well even when trained solely on true labels, it demonstrates greater flexibility for real-world applications.

An important trend in our results is that LM-CLIP benefits significantly from larger models with more parameters, whereas prompt tuning methods CoOp [34], CoCoOp [20], and LMPT [17] degrade as model size increases. This pattern is particularly evident for the ViT-L/14 image encoder, where LM-CLIP achieves the highest total mAP of 93.82% on VOC-MLT and 77.87% on COCO-MLT, outperforming all previous methods by a large margin. In contrast, prompt tuning shows diminishing returns as the model size grows, likely due to reliance on a limited set of learned prompts, which struggle to leverage the full capacity of larger architectures.

E. ABLATION STUDY

In this section, we conduct an in-depth ablation analysis to understand the effectiveness of the individual strategies of our architecture, as in Table 2. We compare our individual *Balanced Asymmetric Loss* against CLIP and ASL, as these are fundamental elements of our model. Furthermore, Distribution-Guided Loss (DB-Loss) is also compared against, as it follows a similar class-aware weighting strategy [14].

CLIP contrastive loss yields better performance than ASL, especially on rare tail classes. However, our proposed loss function (ASL with label smoothing and weighting) improves ASL's performance. The better mAP on the tail for CLIP-loss leads to the conclusion that the hyperparameter λ in $\mathcal{L}_{\text{CLIP}} + \lambda\mathcal{L}_{\text{BAL}}$ can be used to shift performance from head to tail classes.

1) EXPERIMENTS ON SINGLE/MULTI-LABEL SAMPLES

In Table 4, we validate the loss function's performances specifically on samples with single/multiple labels. This comparison is helpful as it resembles real-world scenarios where images often contain multiple labels, and understanding how each component performs under these conditions helps us assess their effectiveness in handling complex, multi-label classification tasks. By focusing on multi-label samples, we aim to highlight the practical applicability and robustness of our *Balanced Asymmetric Loss* function in realistic, challenging environments.

The results show that our proposed BAL outperforms CLIP-loss, ASL, and DB-Loss on multi-label samples, even without additional strategies. While performance declines on single-label images compared to other methods, our combined approach is able to mostly mitigate this.

2) DETAILED METRICS BY CLASS

To further analyze the classification performance of our approach, we present a breakdown of class-wise evaluation metrics Recall, Precision, F1-score, and Average Precision in Figure 6. It indicates that while Precision and Average Precision remain relatively stable across all classes, Recall shows a significant drop for head classes. In contrast, middle and tail classes demonstrate higher recall scores. This suggests that our loss function and class-weighted sampling strategy effectively prioritize rare classes, but at the cost of a reduced ability to correctly retrieve positive samples for frequent head classes.

The lower recall in head classes suggests that the model is more conservative in predicting these frequent labels. While this indicates a shift in model behavior, head classes still achieve strong overall performance, with high precision and average precision scores demonstrating robust classification reliability. Minor adjustments to the weight exponent s in Equation 5 and the loss balancing parameter λ in 6 can fine-tune this further, depending on the application requirements.

3) ANALYSIS OF SAMPLE DISTANCE TO CLASS CENTROIDS

To quantify the effectiveness of our BAL loss function in yielding better text-image embeddings for multi-label classification, we introduce some metrics calculated from the embedding space:

The centroid of class i is defined as $c_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_j$, where N_i represents the number of samples in class i, and x_j the embedding of the j-th image with the positive class label. The positive centroid distance of class i is defined as $D_{i+} = \frac{1}{N_{i+}}\sum_{k=1}^{N_{i+}} \|x_k - c_i\|$, where N_{i+} represents the number of samples in class i, and x_k the embedding of the k-th image with the positive class label. It measures the mean distance between the calculated centroid of each class in latent space and their corresponding samples. This metric should be low, indicating dense clustering of related samples.

Conversely, a high negative class centroid distance suggests well-separated class clusters. The negative centroid distance of class i is defined as $D_{i-} = \frac{1}{N_{i-}}\sum_{k=1}^{N_{i-}} \|x_k - c_i\|$, where N_{i-} represents the number of samples *not* in class i,

TABLE 4. Detailed evaluation metrics of pre-trained CLIP and fine-tuned CLIP with different loss functions. The used image encoder backbone is ViT-B/16. mAP is calculated on only multi-label samples and only single-label samples respectively. We also show the positive over negative sample to centroid distance ratio in embedding space.

Methods	mAP only multi-label samples				mAP only single-label samples				P/D sample to centroid distance			
	Total↑	Head↑	Middle↑	Tail↑	Total↑	Head↑	Middle↑	Tail↑	Total↓	Head↓	Middle↓	Tail↓
VOC-MLT [14]												
CLIP pre-trained [1]	84.31	69.00	87.34	93.51	93.88	85.83	95.24	98.89	0.8046	0.9064	0.7440	0.7019
CLIP-Loss [1]	89.69	81.32	91.73	94.45	97.56	94.60	97.82	**99.58**	0.7790	0.8769	0.7109	0.6836
DB-Loss [14]	85.02	75.15	87.23	90.76	95.77	91.16	95.60	99.37	0.7809	0.8839	0.7095	0.6918
ASL [13]	87.87	83.47	90.99	88.84	96.87	93.91	97.01	98.99	0.7994	0.8901	0.7454	0.6987
BAL (Ours)	91.26	85.19	93.74	93.96	93.65	84.32	96.83	98.27	**0.6146**	**0.7210**	**0.5323**	**0.4963**
LM-CLIP (Ours)	**91.69**	**85.39**	**93.96**	**94.73**	**98.39**	**96.70**	**98.67**	99.50	0.7177	0.8323	0.6477	0.5951
COCO-MLT [14]												
CLIP pre-trained [1]	62.57	44.06	66.45	73.74	69.52	37.31	**76.42**	88.76	0.8200	0.8963	0.7966	0.6830
CLIP-Loss [1]	70.56	60.29	74.30	74.68	**72.00**	**44.73**	76.01	90.72	0.7847	0.8611	0.7507	0.6357
DB-Loss [14]	65.44	55.28	69.50	69.02	70.36	42.00	75.37	88.71	0.7948	0.8659	0.7694	0.6725
ASL [13]	64.47	60.44	68.61	62.55	66.07	35.56	72.45	84.50	0.8237	0.8902	0.7967	0.6971
BAL (Ours)	**74.13**	**67.35**	**79.11**	73.53	68.57	38.83	76.08	84.83	**0.7182**	**0.7919**	**0.6774**	**0.5884**
LM-CLIP (Ours)	73.88	66.34	77.81	**75.32**	68.55	37.27	72.46	**90.90**	0.7553	0.8327	0.7228	0.6112

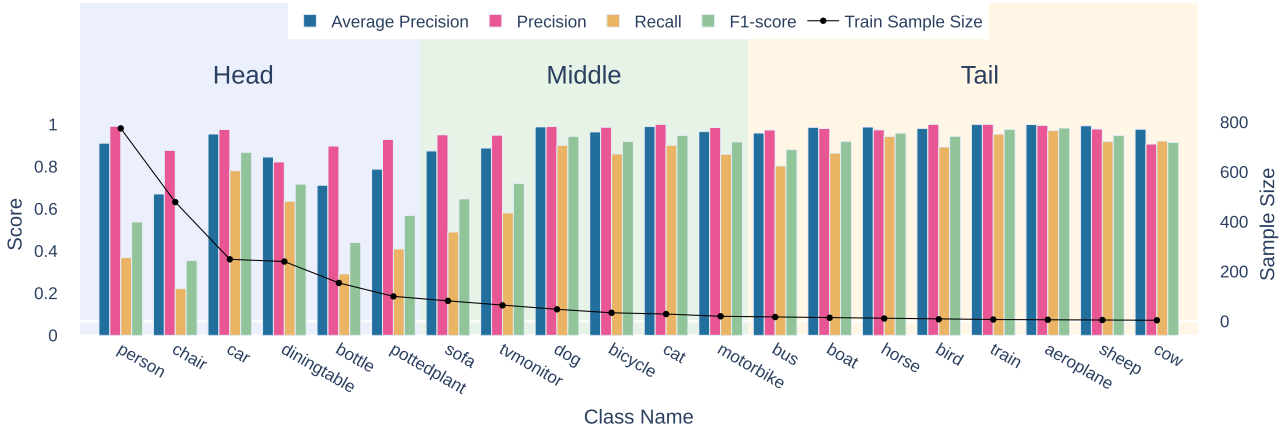


FIGURE 6. Average Precision, Precision, Recall, and F1-score per class of LM-CLIP on VOC-MLT. The used vision encoder is ViT-B/16. We use a threshold of 0.5 after a softmax on the prediction scores to calculate precision/recall. This graph is omitted for COCO-MLT, as it cannot be reasonably plotted for 80 classes.

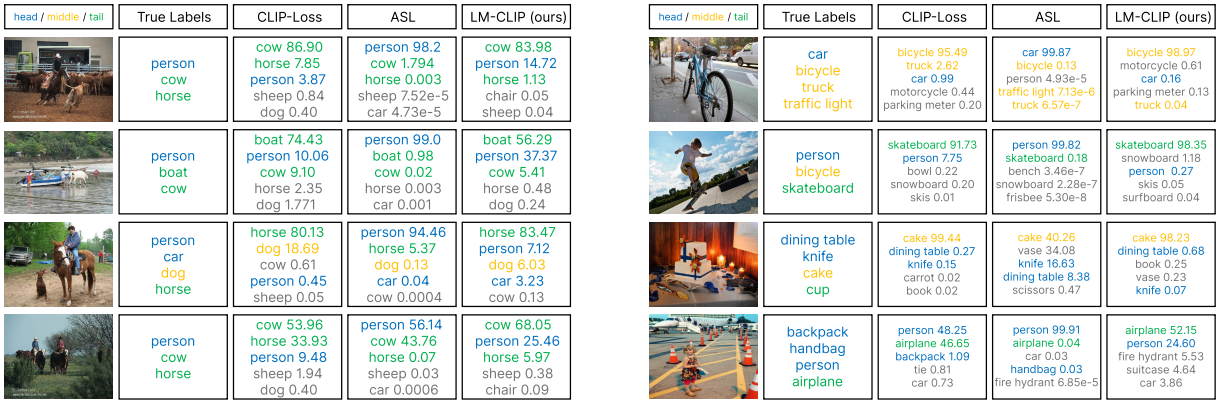


FIGURE 7. Example prediction results. For multi-label images, prediction probabilities across the classes were generated using a softmax function. The CLIP-loss demonstrates superior performance in predicting tail classes, while the ASL excels in predicting head classes. LM-CLIP, however, provides robust predictions for both head and tail classes, indicating balanced performance across the spectrum.

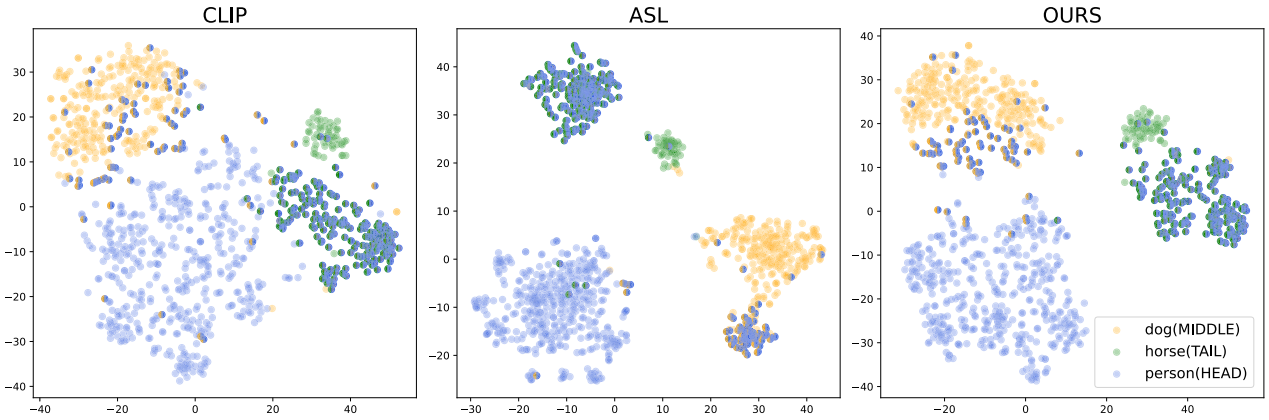


FIGURE 8. Visualized embedding space for the classes "dog", "horse", and "person" on VOC-MLT. Each point represents an image sample. For multi-label samples, points are split to reflect each associated class. Our method combines the strengths of CLIP and ASL: It preserves CLIP's ability to separate negative samples while also achieving the tight clustering of positive samples seen in ASL, effectively capturing multi-label relationships.

and x_k the embedding of the k-th image with the negative class label.

Following this, we calculate positive over negative ratio. This is a combined metric of positive and negative mean class centroid distance, for class i defined as $R_i = \frac{D_{i+}}{D_{i-}}$. A low ratio indicates both dense clustering and good separation of classes. Table 4 suggests that our *Balanced Asymmetric Loss* achieves the best overall embedding space for multi-label classification, with a high distance between class clusters and low distance between same-class samples.

4) EMBEDDING SPACE

In addition to numerically confirming the strengths of our model, we also visualize the embedding space through t-SNE [37] to qualitatively confirm how the images were actually classified. We validated our results by comparing the semantic image representations of different models for the same classes. To verify that our model accurately classifies both head and tail classes, as well as samples with multiple labels, we visualized the three classes that satisfy these conditions. Figure 8 illustrates the embedding space created by "dog", "horse", and "person" classes in VOC-MLT. When solely employing CLIP-loss, all mismatched samples are considered negatives. This approach enhances multi-label classification; however, it does not lead to optimal classification performance. In contrast, ASL achieves tight clustering of positive classes. Our proposed model combines the strengths of both methods, as demonstrated by its superior numerical performance.

5) BALANCED ASYMMETRIC LOSS FOR PROMPT TUNING

In Table 5, we present the results of using our proposed BAL for prompt tuning. We replace LMPT's Distribution-Balanced Loss with class-aware soft margin and re-weighting with our loss function. It is important to note that the image captions are not needed when using BAL, as only LMPT's soft margin hinge embedding loss requires them [17]. These results demonstrate that BAL is also effective at prompt tuning, outperforming all other loss functions when training LMPT with only ground truth labels needed.

TABLE 5. mAP performance of different loss functions when training LMPT. All results are reproduced by us using LMPT's published code. [17].

Loss functions	Total↑	Head↑	Middle↑	Tail↑
VOC-MLT [14]				
BCE	85.35	67.25	87.53	97.30
MLS	85.35	67.25	87.53	97.30
Focal Loss	84.91	65.83	87.30	97.43
CB Loss	83.56	63.50	85.27	97.31
R-BCE-Focal	85.48	67.00	87.87	97.53
ASL	86.04	67.92	**88.66**	97.67
DB Focal	85.41	68.02	86.79	97.41
BAL (Ours)	**86.24**	**68.66**	88.57	**97.67**
COCO-MLT [14]				
BCE	63.77	41.92	66.99	78.75
MLS	63.77	41.92	66.99	78.75
Focal Loss	65.26	**44.77**	68.98	78.38
CB Loss	64.88	43.73	68.70	78.46
R-BCE-Focal	65.04	43.89	69.08	78.32
ASL	63.84	42.25	67.33	78.23
DB Focal	63.58	43.49	66.81	77.00
BAL (Ours)	**65.32**	43.88	**69.16**	**79.12**

V. DISCUSSION

Our proposed LM-CLIP framework introduces *Balanced Asymmetric Loss* and integrates it with contrastive learning to enhance long-tailed multi-label classification. In this section, we analyze the implications of our findings, highlight key comparisons with existing techniques, and outline the strengths and limitations of our approach.

A. IMPLICATIONS

LM-CLIP significantly improves classification performance on long-tailed datasets without requiring image captions, making it ideal for real-world applications where text annotations are scarce. A key example is long-tailed multi-label

disease classification, such as the chest X-ray-long-tailed (CXR-LT) [23] challenge, where rare but critical diseases must be accurately identified. Our approach mitigates the dominance of frequent classes while handling the inherent positive/negative imbalance in multi-label data.

A key contribution is integrating contrastive loss with a supervised multi-label loss. Contrastive loss enhances generalization by separating non-matching pairs, while BAL improves intra-class similarity. Their combination balances class representation, preventing overly confident predictions while ensuring tail-class recognition. This hybrid approach presents a promising direction for future long-tailed multi-label classification research.

B. COMPARISON TO EXISTING LITERATURE

Our work extends prior research on long-tailed multi-label classification. Compared to ASL [13], BAL handles tail class recognition by ensuring rare positive samples contribute sufficiently during training. ASL mainly focuses on hard negatives, while BAL introduces class-aware weighting to balance head and tail class contributions. Label smoothing reduces overconfidence by redistributing label probability mass [32]. When combined with loss adjustment by inverse class frequency, it enables more effective training in long-tailed settings.

Compared to LMPT [17], which enhances classification through natural language captions, LM-CLIP achieves superior mAP without relying on additional textual data. This eliminates inconsistencies from external captions, reinforcing LM-CLIP's effectiveness and generalization ability.

C. LIMITATIONS

While our approach demonstrates impressive results, it is not without limitations. One notable downside is the necessity to fine-tune the entire model, which can be more computationally expensive than prompt tuning. Hyperparameter search is also challenging due to the sensitivity of λ and s. Furthermore, while CLIP provides strong priors, it is trained on internet data, which may introduce biases that affect generalization and fairness [38].

D. CONCLUSION

In this paper, we propose LM-CLIP, a novel approach for long-tailed multi-label image classification that overcomes key limitations of previous methods. Unlike prior work that struggles to balance head and tail class performance or depends on captions, LM-CLIP integrates *Balanced Asymmetric Loss* with contrastive learning to achieve a more robust classification framework. We modify the positive term of ASL to address biases when applied to unbalanced data. Our BAL combined with CLIP-loss is effective in enhancing performance across both head and tail classes. We believe that our model sets a robust benchmark for multi-label classification tasks and provides additional understanding of the challenges involved in long-tailed issues.

REFERENCES

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2021, pp. 8748–8763.

[2] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "CLIP-adapter: Better vision-language models with feature adapters," 2021, *arXiv:2110.04544*.

[3] R. Zhang, R. Fang, W. Zhang, P. Gao, K. Li, J. Dai, Y. Qiao, and H. Li, "Tip-adapter: Training-free CLIP-adapter for better vision-language modeling," 2021, *arXiv:2111.03930*.

[4] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2022, pp. 12888–12900.

[5] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, K.-W. Chang, and J. Gao, "Grounded language-image pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10955–10965.

[6] R. Abdelfattah, Q. Guo, X. Li, X. Wang, and S. Wang, "CDUL: CLIP-driven unsupervised learning for multi-label image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 1348–1357.

[7] M. Ali and S. Khan, "Clip-decoder: Zeroshot multilabel classification using multimodal clip aligned representations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4675–4679.

[8] Z. Novack, S. Garg, J. McAuley, and Z. C. Lipton, "CHiLS: Zero-shot image classification with hierarchical label sets," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2023, pp. 26342–26362.

[9] K. Saito, K. Sohn, X. Zhang, C.-L. Li, C.-Y. Lee, K. Saenko, and T. Pfister, "Prefix conditioning unifies language and label supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2861–2870.

[10] S. Xu, Y. Li, J. Hsiao, C. Ho, and Z. Qi, "Open vocabulary multi-label classification with dual-modal decoder on aligned visual-textual features," 2022, *arXiv:2208.09562*.

[11] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep long-tailed learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10795–10816, Sep. 2023.

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[13] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor, "Asymmetric loss for multi-label classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 82–91.

[14] T. Wu, Q. Huang, Z. Liu, Y. Wang, and D. Lin, "Distribution-balanced loss for multi-label classification in long-tailed datasets," in *16th Eur. Conf. Comput. Vis.*, Jan. 2020, pp. 162–178.

[15] H. Guo and S. Wang, "Long-tailed multi-label visual recognition by collaborative training on uniform and re-balanced samplings," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15084–15093.

[16] D. Lin, "Probability guided loss for long-tailed multi-label image classification," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 2, pp. 1577–1585.

[17] P. Xia, D. Xu, M. Hu, L. Ju, and Z. Ge, "LMPT: Prompt tuning with class-specific embedding loss for long-tailed multi-label visual recognition," 2023, *arXiv:2305.04536*.

[18] . U. R. Dr. A, "Binary cross entropy with deep learning technique for image classification," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 4, pp. 5393–5397, Aug. 2020.

[19] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, pp. 2337–2348, 2022.

[20] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16816–16825.

[21] X. Sun, P. Hu, and K. Saenko, "DualCoOp: Fast adaptation to multi-label recognition with limited annotations," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2022, pp. 30569–30582.

[22] Z. Guo, B. Dong, Z. Ji, J. Bai, Y. Guo, and W. Zuo, "Texts as images in prompt tuning for multi-label image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2808–2817.

[23] G. Holste et al., "Towards long-tailed, multi-label disease classification from chest X-ray: Overview of the CXR-LT challenge," *Med. Image Anal.*, vol. 97, Oct. 2024, Art. no. 103224.

[24] X. Shao, H. Chen, F. Zhao, K. Magson, J. Chen, P. Li, J. Wang, and J. Sasaki, "Multi-label classification for multi-temporal, multi-spatial coral reef condition monitoring using vision foundation model with adapter learning," 2025, *arXiv:2503.23012*.
[25] M. Liu, B. Li, and Y. Yu, "Fully fine-tuned CLIP models are efficient few-shot learners," 2024, *arXiv:2407.04003*.
[26] C. Ding, X. Gao, S. Dong, Y. He, Q. Wang, A. Kot, and Y. Gong, "LOBG: Less overfitting for better generalization in vision-language model," 2024, *arXiv:2410.10247*.
[27] D. X. Lin, T. Peng, R. Chen, X. C. Xie, X. Qin, and Z. Cui, "Distributionally robust loss for long-tailed multi-label image classification," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2024, pp. 417–433.
[28] Y. Li, Y. Song, and J. Luo, "Improving pairwise ranking for multi-label image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1837–1845.
[29] C. Tian, W. Wang, X. Zhu, J. Dai, and Y. Qiao, "VL-LTR: Learning class-wise visual-linguistic representation for long-tailed visual recognition," in *Proc. Eur. Conf. Comput. Vis.*, Jan. 2022, pp. 73–91.
[30] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling representation and classifier for long-tailed recognition," 2019, *arXiv:1910.09217*.
[31] F. Lei, C. Zhang, H. Li, Y. Gao, and C. Chen, "Label distribution guided hashing for cross-modal retrieval," *ACM Trans. Knowl. Discovery from Data*, vol. 19, no. 1, pp. 1–23, Jan. 2025.
[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[33] T. Ma, S. Geng, M. Wang, J. Shao, J. Lu, H. Li, P. Gao, and Y. Qiao, "A simple long-tailed recognition baseline via vision-language model," 2021, *arXiv:2111.14745*.
[34] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, Sep. 2022.
[35] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009.
[36] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *13th Eur. Conf. Comput. Vis.*, Zurich, Switzerland, Jan. 2014, pp. 740–755.
[37] L. V. D. Maaten and G. E. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, Jan. 2008.
[38] I. Alabdulmohsin, X. Wang, A. Steiner, P. Goyal, A. D'Amour, and X. Zhai, "CLIP the bias: How useful is balancing data in multimodal learning?" 2024, *arXiv:2403.04547*.

CHRISTOPH TIMMERMANN received the B.S. degree in computer science from Furtwangen University, Germany, in 2023. He is currently pursuing the master's degree in artificial intelligence with Dongguk University, Republic of Korea. He was a member of the Data Analysis and Machine Intelligence Laboratory, Dongguk University. His research interests include multi-modal models, long-tailed learning, and zero-shot learning.

SEUNGHYEON JUNG received the B.S. degree from the Department of Electronics and Electrical Engineering, Dongguk University, in 2024. He is currently pursuing the M.S. degree with the Department of Computer Science and Artificial Intelligence, Dongguk University, Seoul, Republic of Korea. His research interests include multi-modal, fairness, and human pose estimation.

MISO KIM received the B.S. degree from the Department of Statistics and Software, Dongguk University, Seoul, South Korea, in 2024. She is currently pursuing the M.S. degree with the Department of Computer Science and Artificial Intelligence, Dongguk University. Her research interests include multi-modal and safe AI.

WOOJIN LEE received the B.S. degree in information and industrial engineering from Yonsei University, Seoul, Republic of Korea, in 2015, and the Ph.D. degree in industrial engineering from Seoul National University, in 2020. He is currently an Assistant Professor with the College of AI Convergence, Dongguk University, Seoul. His research interests include robustness and fairness in deep learning.