Dual-Phase Visual-Language Pretraining and Adaptation for Long-Tailed Multi-Label Recognition

Yongcheng Li[1], Xuekuan Wang[1], Zhifei Zhang[1*], Cairong Zhao[1*]

[1]Tongji University
liyongcheng@tongji.edu.cn

Abstract

Long-Tailed Multi-Label Recognition (LTML) is a critical yet challenging task due to two core issues: the severe scarcity of training samples for rare "tail" classes, and the complex co-occurrence patterns among labels that often lead to biased models. To address this, we propose DP-VLPA, a novel Dual-Phase Visual-Language Pretraining and Adaptation framework. In the first phase, our Structured Tail-Aware Generation (STAG) module employs a Large Language Model (LLM) to create detailed descriptions that explicitly emphasize tail classes and their contextual relationships, providing a strong and less-biased feature foundation. In the second adaptation phase, we ensure this knowledge is applied effectively. A Dynamic Query Reweighting (DQR) mechanism forces the model to attend to crucial tail-class evidence. Simultaneously, a Co-occurrence-Aware (COA) loss explicitly teaches the model the statistical dependencies between labels, correcting for co-occurrence biases. Extensive experiments on VOC-LT and COCO-LT datasets demonstrate state-of-the-art performance, achieving mAP scores of 90.72% and 74.42% respectively - surpassing previous best methods by 2.84% and 8.23%.

Code — https://github.com/dz1104/DP-VLPA

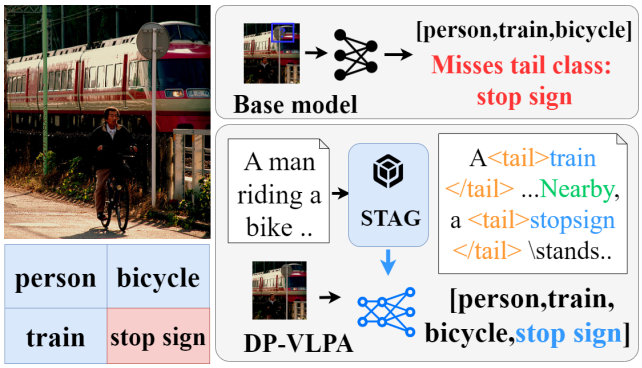

Figure 1: **Our core motivation**. (above) Existing methods suffer from information poverty, failing to recognize tail classes (e.g., 'stop sign') from limited visual cues. (below) We propose a new paradigm of knowledge injection. By leveraging an LLM to generate rich, contextual narratives that highlight tail classes, our model is fundamentally equipped with the necessary knowledge for accurate recognition.

Introduce

Humans can easily recognize multiple objects in an image. For example, at a beach, we can spot not only the common 'person' but also the rare 'surfboard' which might be carrying, because of understanding the context and knowing that certain objects often appear together. However, AI systems struggle with this task and this challenge is known as Long-Tailed Multi-Label Recognition (LTML) problem (Liu et al. 2019; Cui et al. 2019; Zhang et al. 2023b; Liu et al. 2021; Liu and Tsang 2015; Misra et al. 2016). The core of this challenge is twofold: First a severe data imbalance where 'head' classes like 'person' are abundant while 'tail' classes like 'surfboard' are rare. Second, objects in an image often have complex relationships with each other (Rawlekar et al. 2024). We argue that the true bottleneck for tail class recognition is not merely the scarcity of samples, but the profound 'information poverty' that results from it. A model trained on only a few images with 'surfboard' cannot learn the rich context that humans know, such as the fact that surfboards are found at beaches with waves and people. As a result, the model learns a weak visual representation that does not generalize well. This paper is fundamentally motivated by this insight: to truly advance LTML, we should move beyond addressing data imbalance and instead focus on remedying the underlying poverty of semantic information for tail classes.

Previous efforts to address the LTML problem (Zhang et al. 2023b) can be broadly categorized into data balance, data augmentation, and module improvement. Data balance methods like re-sampling (Nitesh 2002; Estabrooks, Jo, and Japkowicz 2004; Liu, Wu, and Zhou 2008; Zhang and Pfister 2021) and re-weighting (Cui et al. 2019; Lin et al. 2017; Tan et al. 2020; Wang et al. 2021) adjust class distributions or loss functions to mitigate imbalance. Data augmentation (Perez and Wang 2017; Wang, Ramanan, and Hebert 2017; Chu et al. 2020) enhances model performance by introducing additional data, while module improvement (Zhang et al. 2023a; Jin et al. 2023; Cui et al.

*Corresponding authors.

2022) refines model structures to improve tail class performance. However, these methods share one fundamental problem that is limiting the information already present in the training set. Their core mechanism involves amplifying the training signal from a few existing tail-class images, rather than introducing new, diverse semantic information. Consequently, a model may learn to recognize specific instances, but it may fail to learn a general concept of the tail class. Thus, the model still lacks rich contextual features for tail classes, and its learned representations remain weak. This problem becomes worse when multiple objects appear together. In multi-label datasets, head classes are not only more frequent but also appear with many different objects. To minimize training loss, the model learns to focus on these frequent head classes and their common partners. This strategy makes the model biased. It tends to suppress or ignore the less common but correct patterns involving tail classes. As a result, the model cannot understand the true relationships between objects, leading to wrong and logically inconsistent predictions, especially for the tail classes.

The fundamental mechanism of Vision-Language Models (VLMs) (Xia et al. 2023; Zhou et al. 2022a; Zhao et al. 2024) offers a paradigm shift for tackling the Long-Tailed Multi-Label Recognition (LTML) challenge. Unlike traditional models confined to the visual modality, VLMs project both images and their corresponding text into a shared latent space, where they are aligned based on semantic similarity. This principle means that a well-formed text embedding can serve as a powerful proxy for its visual counterpart. This simple idea creates a huge opportunity for LTML: we can make up for the lack of rare images, which are difficult and costly to collect, by using rich text descriptions, which are much easier to get or create, as shown in Figure 1. However, simply using this opportunity is not as easy as it sounds. The main research question changes from "how to re-balance the small number of images?" to "what kind of text is actually useful for augmentation?" A simple class name like "surfboard" does not provide enough context. On the other hand, a general description that doesn't target specific objects might give the model a weak or confusing signal. To truly enrich the model's understanding, the augmenting text must be both semantically rich—detailing an object's appearance and interactions—and explicitly tail-aware—purposefully emphasizing the rare classes.

In this work, we introduce **DP-VLPA**, a framework built upon a synergistic, dual-phase learning paradigm. Our core philosophy is that a truly robust solution requires decoupling general knowledge acquisition from biased task adaptation. In the first phase, we solve the 'information poverty' problem of tail classes with our Structured Tail-Aware Generation (STAG) strategy. Instead of passively using the simple captions often found in VL dataset, STAG employs a Large Language Model as a knowledge engine to create rich, contextual descriptions that explicitly emphasize tail classes and their interactions. This process provides a strong and less-biased foundation of knowledge before it even sees the imbalanced training labels. In the second phase, during adaptation, our proposed module of Dynamic Query Reweighting (DQR) forces the model to pay more attention to the evidence for tail classes, fighting its natural tendency to ignore them. Simultaneously, a Co-occurrence-Aware (COA) loss instills the missing relational logic by explicitly modeling label dependencies. Together, these parts work in synergy: STAG provides the necessary knowledge, while DQR and COA make sure this knowledge is used correctly during prediction. Our integrated framework fixes the core problems of information scarcity and biased learning.

In summary, the contributions of our work are three-fold:

- A New Methodology for Knowledge-Driven Pretraining. This is a two-step process: STAG strategy provides a blueprint for using LLMs to generate targeted, high-quality text that enriches the information available for rare classes. QCL provides a mechanism for the model to perform fine-grained alignment, learning to connect specific visual regions to their corresponding descriptions.
- A Practical Method for Long-Tailed Adaptation. We propose an effective adaptation strategy for applying powerful, pre-trained models to real-world, biased data. Our framework includes two key components: DQR teaches the model how to actively seek out and prioritize evidence for rare classes during inference. COA helps the model learn the statistical relationships between objects, leading to more logically consistent predictions.
- We conduct the extensive experiments on VOC-LT and COCO-LT, outperforming the state-of-the-art method by 2.84 / 8.23 % and zero-shot CLIP by 4.95 / 14.25%.

Related Works

Long-Tailed Multi-Label Recognition

Real-world data often exhibits a long-tail distribution, leading to performance degradation on tail classes, which presents significant challenges for many computer vision tasks. To address the long-tail problem, several solutions have been proposed: re-sampling (Nitesh 2002; Estabrooks, Jo, and Japkowicz 2004; Liu, Wu, and Zhou 2008; Zhang and Pfister 2021) adjusts class distributions by up- or down-sampling, re-weighting (Cui et al. 2019; Lin et al. 2017; Tan et al. 2020; Wang et al. 2021) modifies training loss with class-specific weights, and information augmentation (Perez and Wang 2017; Wang, Ramanan, and Hebert 2017; Chu et al. 2020) adds extra data to improve model performance. Module improvement (Zhang et al. 2023a; Jin et al. 2023; Cui et al. 2022) refines model architecture to better handle tail classes. Deep hashing methods learn compact binary codes for efficient retrieval (Shen et al. 2015, 2018).

In addition, multi-label recognition research has concentrated on modeling label dependencies. Early works utilize graph-based models like GNNs (Wei et al. 2015; Liu and Tsang 2015) to explicitly capture co-occurrence statistics. More recent approaches leverage Transformers (Zhou et al. 2022a; Xia et al. 2023), using attention mechanisms to implicitly learn the relationships between different class-specific queries. However, a common limitation underlies these approaches: they are confined to existing visual information. They either re-balance the data statistics or refine the model architecture, but do not fundamentally address

the scarcity of rich, semantic knowledge for tail classes. Our work tackles this issue from a different perspective by introducing external knowledge through generated text.

Vision-Language Models for Recognition

Vsision-language pretraining (VLP) frameworks, such as CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021), have demonstrated significant advances by learning transferable representations through large-scale image-text alignment. Building on this paradigm, advanced methods (Kim, Son, and Kim 2021; Li et al. 2022; Yang et al. 2022) enhance cross-modal interactions via cross-attention mechanisms and auxiliary objectives (e.g., image-text matching and masked language modeling), enabling fine-grained alignments between visual and textual semantics (Goel et al. 2022; Yao et al. 2021; Gao et al. 2022). Notably, BLIP2 (Li et al. 2023) introduces a lightweight querying transformer to bridge frozen image encoders and large language models, facilitating bidirectional modality fusion. Meanwhile, SPARC (Bica et al. 2024) extends CLIP with instance-level alignment for dense visual concepts. However, these methods typically rely on existing or generically generated captions and weakly correlate with textual descriptions of instances. Our work is motivated by this gap. We propose that the key is not to passively use existing text, but to actively generate structured, tail-aware knowledge to directly enrich the model's understanding at the source.

Method

Overall Framework

Our proposed framework, DP-VLPA, is designed to systematically address the core challenges of Long-Tailed Multi-Label (LTML) recognition. We start with a dataset $\mathcal{D} = \{(x_i, y_i, t_i^{orig})\}_{i=1}^{N}$, where each sample consists of an image x_i, its multi-label annotation $y_i \in \{0,1\}^C$, and an original, often simple, caption t_i^{orig}. Our method operates in two synergistic phases to transform this raw data into highly effective predictions, as illustrated in Figure 2. The first phase, **Knowledge Injection**, tackles the problem of information poverty for tail classes. By feeding both the labels y_i and the original caption t_i^{orig} into STAG, we guide a Large Language Model to produce a new, rich textual description t_i, especially, the relationship among multiple objects. Then, we design a visual-language based model to learn from these enhanced text-image pairs (x_i, t_i). The model features a set of K learnable visual queries, $\mathbf{Q} \in \mathbb{R}^{K \times D}$, which are processed through a Q-Former to extract key visual information, yielding fine-grained image features $\mathbf{Q}_{image}$. This pretraining phase aligns these visual features with the detailed textual knowledge, fundamentally enriching the model's understanding of rare objects before any classification occurs.

The second phase, **Adaptive Fine-tuning**, adapts the knowledge-rich model to the specific classification task, with a strong focus on tail-class attentiveness and co-occurrence modeling. In this phase, we introduce class-specific semantic embeddings, $\mathbf{E}_{class} \in \mathbb{R}^{C \times D}$, derived by encoding textual prompts like "a photo of a [CLASS]". The model is then fine-tuned using the ground-truth labels y_i with an objective that incorporates two core innovations. First, a Dynamic Query Reweighting (DQR) mechanism is employed to dynamically amplify the influence of visual queries that are critical for identifying tail classes. Second, a Co-occurrence Aware (COA) loss explicitly teaches the model the statistical dependencies between labels. By integrating these components, DP-VLPA not only learns what tail classes look like but also how to actively look for them and understand their context within a scene.

Enriching Tail Representation via Knowledge Injection

To combat the fundamental problem of information poverty for tail classes, our first phase focuses on injecting rich, structured knowledge into the model. This is a two-step process: first, we generate knowledge-rich textual descriptions using **Structured Tail-Aware Generation (STAG)**, and second, we ensure the model deeply absorbs this knowledge through a multi-level Visual-Language Alignment objective.

The STAG module transforms sparse labels and simple captions into detailed narratives. By providing an LLM with an image's labels (with tail classes explicitly marked) and its original caption, we use a rule-constrained prompt to guide the generation of a new, enhanced text t_i. This prompt ensures the resulting text prioritizes describing tail classes, their interactions with other objects, and their contextual roles. For example, it turns 'person, handbag(tail), ...' into a story like "*A man sits at a table, placing the handbag(tail)...*", effectively creating a knowledge-rich training signal that is absent in the original dataset. The specific process of the stage module is shown in Figure 3, and we have provided a detailed introduction to our data augmentation strategy in the supplementary materials.

With these enhanced descriptions, we then align the model's visual understanding with this textual knowledge using a dual-level contrastive learning approach. For a given image-text pair (x_i, t_i), the model outputs a set of K visual query features $\mathbf{Q}_{image} = \{q_1^i, ..., q_K^i\}$ and a set of word token features $\mathbf{T}_{feature} = \{t_{cls}^i, t_1^i, ..., t_L^i\}$, of which the first token is global '[CLS]' token t_{cls}^i.

First, a **Global Contrastive Learning (GCL)** loss establishes a coarse, scene-level alignment. We compute the overall similarity score $s_{i,j}$ between an image i and a text j by taking the maximum similarity found between any of the image's queries and the text's global '[CLS]' feature:

$$s_{i,j} = \max_{k \in \{1,...,K\}} \left(sim(q_k^i, t_{\text{cls}}^j) \right) \tag{1}$$

where $sim()$ denotes the cosine similarity. This score is then used within a standard InfoNCE loss, $\mathcal{L}_{GCL}$, to pull matching image-text pairs together in the embedding space.

However, global alignment is insufficient for the fine-grained distinctions required by multi-label tasks. We therefore introduce a **Query-based Fine-grained Contrastive Learning (QCL)** loss to create a direct correspondence between individual words and specific visual regions. The core

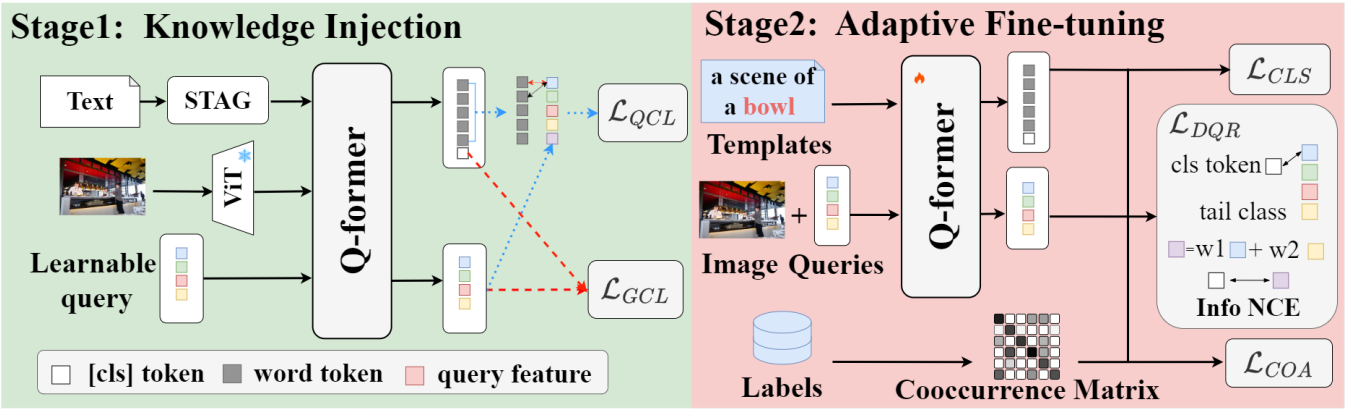

Figure 2: Our DP-VLPA framework consists of two synergistic phases. The initial Knowledge Injection phase leverages a Large Language Model (LLM) via our STAG module to generate tail-aware descriptions, and then pre-trains the model to align visual features with this rich textual knowledge. Subsequently, the Adaptive Fine-tuning phase adapts the model for multi-label classification, employing a Dynamic Query Reweighting (DQR) loss to enhance tail-class attention and a Co-occurrence Aware (COA) loss to model label dependencies.

idea here is the creation of "text-guided queries". For each word token t_j^i in a description, we dynamically synthesize a corresponding visual query q_j^t. This is achieved by computing a weighted average of all original visual queries $\mathbf{Q}_{image}$, where the weights are derived from the normalized similarity between the word t_j^i and each query q_k^i:

$$q_j^t = \sum_{k=1}^{K} q_k^i \cdot \text{weight}(t_j^i, q_k^i) \tag{2}$$

This process yields a set of new queries $\mathbf{Q}_i^t$, each tailored to a specific word in the description. We then enforce a fine-grained alignment by applying a symmetric InfoNCE loss, $\mathcal{L}_{QCL}$, between the original word embeddings $\mathbf{T}_{feature}$ and this new set of synthesized text-guided queries $\mathbf{Q}_i^t$. The total pretraining objective combines both alignment strategies, ensuring knowledge is absorbed at both macro and micro levels:

$$\mathcal{L}_{pretrain} = \mathcal{L}_{GCL} + \mathcal{L}_{QCL} \tag{3}$$

This multi-level alignment process effectively transfers the structured, tail-aware knowledge from the generated text into the visual encoder's feature space, creating a robust foundation for the downstream adaptation phase.

Adaptive Fine-Tuning for Long-Tailed Recognition

Having equipped the model with rich semantic knowledge, the adaptation phase fine-tunes it to master the specific challenges of LTML: the model's inattentiveness to tail classes and its ignorance of label co-occurrence. We address these two intertwined problems through a unified objective that synergistically combines a standard classification loss with two novel, collaborative mechanisms.

First, to combat the model's tendency to be dominated by head classes, we introduce **Dynamic Query Reweighting** (DQR). This mechanism acts as an adaptive attention amplifier, forcing the model to focus on the often subtle visual evidence of tail classes. For each image, DQR identifies the visual queries most correlated with the present tail classes and constructs a dynamic weight matrix W. This matrix assigns significantly higher weights to these tail-critical queries, with the weight intensity modulated by the class rarity ($w \propto 1/\sqrt{n_c}$). This reweighting is then integrated into an enhanced global contrastive loss, $\mathcal{L}_{GCL-DQR}$, by transforming the overall image-text similarity score into a weighted average of per-query similarities:

$$\hat{s}_{i,j} = \frac{\sum_{k=1}^{K} W_{i,k} \cdot s_{i,j,k}}{\sum_{k=1}^{K} W_{i,k}} \tag{4}$$

Here, $s_{i,j,k}$ is the base similarity between the k-th query of image i and the text of image j. By using this reweighted score $\hat{s}_{i,j}$ within the InfoNCE loss, DQR effectively compels the model to prioritize learning from tail-relevant visual information.

Second, beyond recognizing individual objects, a robust model should understand their logical relationships. To instill this contextual common sense, we introduce the **Co-occurrence Aware** (COA) loss. This loss explicitly teaches the model the statistical co-occurrence patterns observed in the training data (e.g., 'person' and 'backpack' often appear together). We start with a pre-computed co-occurrence matrix C_{ij} representing these real-world probabilities. The core idea of COA is to enforce that the similarity structure of the model's visual queries, $\text{Sim}(\mathbf{Q}, \mathbf{Q})$, mirrors these real-world statistics. It achieves this by first calculating an "expected" query similarity matrix, Q_{co}, based on the class co-occurrence matrix C. The COA loss then minimizes the distance between the model's actual query similarities and this expected structure:

$$\mathcal{L}_{COA} = \beta \cdot ||Q_{co} - \mathbf{Q} \cdot \mathbf{Q}^{\mathbf{T}}||_F^2 \tag{5}$$

This loss acts as a powerful regularizer, encouraging the feature representations of co-occurring classes to be geometrically closer.

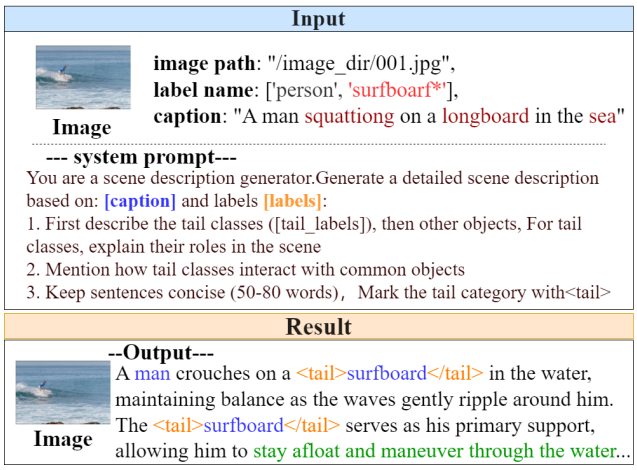

Figure 3: Detailed Explanation of Structured Tail-Aware Generation (STAG)

Finally, the total training objective holistically integrates a standard multi-label classification loss, $\mathcal{L}_{CLS}$, with our two tailored mechanisms. The classification logits L_c for each class are derived by taking the maximum similarity between the visual queries $\mathbf{Q}_{image}$ and the corresponding class embedding from $\mathbf{E}_{class}$. The total loss is:

$$\mathcal{L} = \mathcal{L}_{CLS} + \lambda_1 \mathcal{L}_{GCL-DQR} + \lambda_2 \mathcal{L}_{COA} \quad (6)$$

This joint optimization framework achieves a powerful synergy. DQR enhances the model's ability to "see" rare objects, while COA enhances its ability to "reason" about their relationships. Together, they transform a generic vision-language model into a specialist, finely tuned for the intricate challenges of LTML.

Experiments

Experimental settings

Dataset. We conduct experiments on two long-tailed multi-label classification datasets: VOC-LT and COCO-LT (Wu et al. 2020). They are artificially derived from two multi-label visual recognition benchmarks: PascalVOC (Everingham et al. 2015) and MS-COCO (Lin et al. 2014), respectively. **VOC-LT** is sampled from the PascalVOC2012 train-val set, with categories following a Pareto distribution (Arnold 2014). The training set includes 1,142 images across 20 classes, with 4 to 775 samples per class. These classes are divided into three groups: head class with more than 100 samples, medium class with 20 to 100 samples, and tail class with fewer than 20 samples, resulting in a 6:6:8 ratio. The test set contains 4,952 images, the same as the PascalVOC2007 test set. Using a similar method. **COCO-LT** is sampled from MS-COCO proposed in 2017 years. The training set contains 1,909 images and 80 classes, with the number of samples per class ranging from 6 to 1,128. The class division is similar to that of VOC-LT, with head, medium, and tail classes in a ratio of 22:33:25. The test set contains 5,000 images, corresponding to the MS-COCO 2017 test set.

Figure 4: Visualization of Cross-Attention Maps for Tail-Class Queries. Our method successfully and precisely focuses on the target tail-class objects.

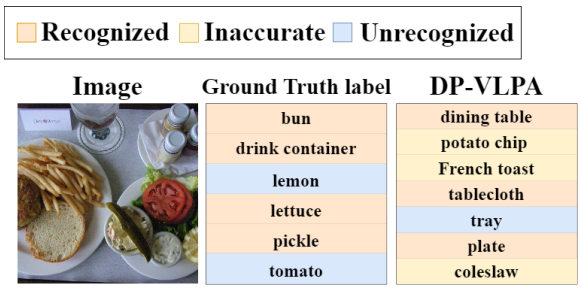

Figure 5: Prediction from DP-VLPA model on LVIS dataset. Orange indicates successfully identified objects, blue represents objects present in the image but either unrecognized or misclassified, and yellow denotes approximate targets.

Implementation Details We use CLIP ViT-B/16 (Radford et al. 2021) as the visual encoder with frozen parameters and the Q-former from BLIP2 (Li et al. 2023) as the backbone, initialized with pre-trained BERT (Lee and Toutanova 2018) parameters. Besides, we use the AdamW optimizer with β_1 = 0.9, and a weight decay of 0.05. And we use a cosine learning rate decay with a peak learning rate of 1e-4 and a linear warmup of 2k steps, pre-training on the LLM-augmented dataset for 30 epochs, using one NVIDIA GeForce RTX 4090 GPU. For reproducibility, we use a fixed random seed (torch.seed(42)) for all experiments.

Evaluation Metrics. We use evaluation metrics similar to (Xia et al. 2023), employing mean average precision (mAP) to assess the performance of long-tailed multi-label visual recognition across all categories, as well as the average performance for head, medium, and tail classes separately.

Comparison with State-of-the-art Methods

To evaluate the effectiveness of our proposed method, we compare mAP performance with the previous state-of-the-art (SOTA) methods on two long-tail datasets, including vision-based methods such as ERM, a smooth version of Re-Weighting (RW) using theinverse proportion to the square root of class frequency, Re-Sampling (RS) (Shen, Lin, and Huang 2016), Focal Loss (Lin et al. 2017), ML-GCN (Chen et al. 2019), OLTR (Liu et al. 2019), LDAM (Cao et al. 2019), Class-Balanced (CB) Focal (Cui et al. 2019), BBN (Zhou et al. 2020), Distribution-Balanced (DB) Focal (Wu et al. 2020) and LTML (Guo and Wang

Datasets	VOC-LT				COCO-LT			
Methods	total	head	medium	tail	total	head	medium	tail
visual models								
ERM	70.86	68.91	80.20	65.31	41.27	48.48	49.06	24.25
RW	74.70	67.58	82.81	73.96	42.27	48.62	45.80	32.02
main Focal Loss (Lin et al. 2017)	73.88	69.41	81.43	71.56	49.46	49.80	54.77	42.14
RS (Shen, Lin, and Huang 2016)	75.38	70.95	82.94	73.05	46.97	47.58	50.55	41.70
ML-GCN (Chen et al. 2019)	68.92	70.14	76.41	62.39	44.24	44.04	48.36	38.96
OLTR (Liu et al. 2019)	71.02	70.31	79.80	64.95	45.83	47.45	50.63	38.05
LDAM (Cao et al. 2019)	70.73	68.73	80.38	69.09	40.53	48.77	48.38	22.92
CB Focal (Cui et al. 2019)	75.24	70.30	83.53	72.74	49.06	47.91	53.01	44.85
BBN (Zhou et al. 2020)	73.37	71.31	81.76	68.62	50.00	49.79	53.99	44.91
DB Focal (Wu et al. 2020)	78.94	73.22	84.18	79.30	53.55	51.13	57.05	51.06
LTML (Guo and Wang 2021)	81.44	75.68	85.53	82.69	56.90	54.13	60.59	54.47
vision-language models								
CLIP (Radford et al. 2021)	85.77	66.52	88.93	97.83	60.17	38.52	65.06	72.28
CoOp (Zhou et al. 2022b)	86.02	67.71	88.79	97.67	60.68	41.97	63.18	73.85
CoCoOp (Zhou et al. 2022a)	84.47	64.58	87.82	96.88	61.49	39.81	64.63	76.42
LMPT (Xia et al. 2023)	87.88	72.10	89.26	**98.49**	66.19	44.89	69.80	79.08
SPARC† (Bica et al. 2024)	86.04	75.43	90.28	90.83	65.06	56.69	70.22	65.62
DP-VLPA(GPT4o)	90.17	**79.46**	92.53	96.42	74.29	**67.90**	76.92	76.44
DP-VLPA(DeepSeek)	**90.72**	78.87	**92.82**	98.04	**74.42**	65.09	**76.97**	**79.26**

Table 1: mAP performance of the proposed method and comparison methods. We will divide the models used for comparison into two categories: visual models and visual language models.

STAG	QCL	DQR	COA	VOC-LT					COCO-LT				
				total	head	medium	tail	△mAP	total	head	medium	tail	△mAP
				82.84	66.64	84.34	93.86	／	48.23	47.55	47.59	49.69	／
✓				83.21	70.77	80.68	94.26	0.37	64.01	51.83	65.12	72.43	15.78
	✓			83.62	69.26	81.70	95.84	0.78	64.07	55.73	63.11	70.89	15.84
✓	✓			84.65	72.49	85.11	93.43	1.81	70.45	62.98	74.60	71.56	22.22
✓	✓	✓		88.19	75.40	89.47	96.82	5.35	72.21	64.96	75.48	74.30	23.98
✓	✓	✓	✓	90.72	78.87	92.82	98.04	7.88	74.42	65.09	76.97	79.26	26.19

Table 2: Ablation analysis on different components of the proposed network.

Dataset	**Queries**	**Total**	**Head**	**Medium**	**Tail**
VOC-LT	64 Queries	89.48	**80.82**	90.82	94.97
	32 Queries	90.72	78.87	**92.82**	98.04
	16 Queries	**90.92**	80.52	91.65	**98.18**
COCO-LT	64 Queries	65.44	52.24	65.97	76.37
	32 Queries	**74.42**	**65.09**	**76.97**	**79.26**
	16 Queries	69.30	64.95	73.05	68.19

Table 3: mAP performance of the proposed method with different numbers of queries.

Analysis	**Key Results**
Quality Assessment	Hallucination: 7.50%, Acc: 8.23/10
Tail Enhancement	Mention Freq. ↑ **4.30**× (0.44→1.88)

Table 4: Quantitative validation of the STAG module.

2021), as well as vision-language model-based approaches such as CLIP (Radford et al. 2021), CoOp (Zhou et al. 2022b), CoCoOp (Zhou et al. 2022a) , and LMPT (Xia et al. 2023). Additionally, since the original code is not available, we independently implement the SPARC described in the (Bica et al. 2024), and pre-trained SPARC on the vision-language dataset.

As shown in tab:tab1, our proposed DP-VLPA achieves 90.72% mAP on VOC-LT and 74.42% on COCO-LT, outperforming pure vision state-of-the-art method of LTML (81.44%/56.90%) by 9.28% and 17.52%, respectively. Notably, even zero-shot CLIP surpasses LTML (85.77%/60.17% vs 81.44%/56.90%), highlighting the critical role of language modality. Compared to LMPT, DP-VLPA achieves absolute gains of 2.84%/8.23% on VOC/COCO-LT. While tail-class improvements are moderate (e.g., +1.9% on COCO-LT), DP-VLPA significantly boosts head/medium classes (e.g., +6.8% head on VOC-LT) without performance trade-offs which is due to our STAG strategy, STAG enriches the entire semantic context by describing interactions between tail classes and their common counterparts. This improves the model's understanding of all involved classes, not just the tail.

Since the official code for SPARC is not available, we manually reproduced the method and trained it on VOC-LT and COCO-LT for multi-label classification. As an approach that enhances CLIP's fine-grained alignment, our reproduced SPARC improved the overall mAP over CLIP by

4.89% on COCO-LT. This result indicates that incorporating a fine-grained loss significantly boosts performance on multi-label classification tasks.

Ablation Analysis

Components Analysis We conducte ablation studies to analyze the effectiveness of each module in DP-VLPA, with the results presented in tab:tab2. The baseline model(only GCL) performs poorly. Adding our fine-grained alignment loss (QCL) to the GCL baseline yields a massive +15.84% mAP gain on COCO-LT, as it grounds text to specific image regions, preventing small, rare objects from being overlooked. More critically, incorporating STAG-generated descriptions delivers the single largest boost to tail-class performance (+22.22% on COCO-LT), directly validating our core hypothesis that targeted textual knowledge is crucial for overcoming information poverty. The adaptation phase then refines this knowledge base. The DQR module improves tail-class mAP (+3.39% on VOC-LT) by focusing attention on tail-specific evidence via query re-weighting. Furthermore, the COA loss boosts performance, particularly on the complex COCO-LT dataset (+4.96% tail mAP), by modeling co-occurrence logic. The full DP-VLPA framework's success stems from a clear synergy: STAG injects knowledge, QCL grounds it, and DQR with COA guide its effective application.

Impact of Query Quantity We compare the performance of the DP-VLPA model with different numbers(16/32/64) of queries, as shown in tab:tab3. The 32-query variant balances performance in both datasets, while the 16-query model achieves the highest VOC-LT mAP (90.92%) but suffers a 5.12% drop in COCO-LT. Increasing queries to 64 reduces mAP to 89.48% (VOC-LT) and 65.44% (COCO-LT). Since each query encodes distinct image regions, an insufficient number of queries hinders the model's information extraction capability, particularly for complex datasets. Conversely, excessive queries cause information dispersion, resulting in less distinct features. Therefore, determining the optimal number of queries is essential.

Stability Analysis with Different Seeds To verify the robustness of our framework, we repeated our full experiment on COCO-LT with five different random seeds. Our method achieves a highly consistent mAP of 74.41% ± 0.12%, demonstrating that the reported performance is stable and not a result of random chance.

Qualitative Analysis

Visual analysis. To intuitively demonstrate the effectiveness of our approach, we visualize the cross-attention maps derived from query vectors most aligned with specific tail classes. As shown in Figure 4, the baseline model without Dynamic Query Reweighting (DQR) fails to locate inconspicuous tail objects like "surfboard"; its attention is instead captured by dominant head classes or scattered across the background. In contrast, our full model, DP-VLPA, leverages DQR to precisely highlight these tail objects, even when small or occluded. This provides direct visual evidence that DQR effectively steers the model's focus toward under-represented categories.

Validity verification of STAG module We validate our STAG module, with results in tab:tab4. We first assess the generation quality using an external vision-language model (Qwen-VL-Max) as a judge, which report a low hallucination rate of 7.50% and a high accuracy score of 8.23/10, confirming the high fidelity of our generated text. Second, we confirm the STAG's core function of amplifying tail-class signals. A frequency analysis over 49,205 tail-class instances reveal that our method increase the mention frequency of tail classes by a remarkable 4.30 times compared to the original captions (from 0.438 to 1.883 mentions per instance). This proves that STAG effectively enriches sparse tail-class information, a key factor for our model's performance.

Open World Multi-label Recognition

We test our model's open-world generalization on LVIS (1,203 classes) using our COCO-trained model (80 classes). Figure 5 shows strong performance, which we attribute to STAG. Training on rich sentences rather than just labels allows the model to learn concepts beyond its training set, such as correctly identifying 'tablecloth' and 'plate'—objects absent from groundtruth label. This proves STAG injects broader world knowledge. Expected fine-grained errors occur (e.g., 'French fries' vs. 'potato chips'), as COCO lacks such detail. Nevertheless, these results show our framework's strong potential extends beyond long-tail recognition to more general visual understanding.

Conclusion

In this paper, we tackle the "information poverty" of tail classes in Long-Tailed Multi-Label Recognition (LTML) by shifting from data re-balancing to active knowledge injection. Our proposed DP-VLPA framework first uses a Large Language Model to generate rich, tail-aware descriptions (STAG). Then, in an adaptation phase, our DQR and COA mechanisms ensure this knowledge is effectively applied. DP-VLPA achieves the state-of-the-art results on VOC-LT and COCO-LT, with 90.72% and 74.42% mAP respectively. Our work validates structured knowledge injection as a powerful strategy for long-tailed problems and opens new avenues for future research.

Acknowledgments

This work was supported by National Natural Science Fund of China (No. U25A20527, 62473286).

References

Arnold, B. C. 2014. Pareto distribution. *Wiley StatsRef: Statistics Reference Online*, 1–10.

Bica, I.; Ilić, A.; Bauer, M.; Erdogan, G.; Bošnjak, M.; Kaplanis, C.; Gritsenko, A. A.; Minderer, M.; Blundell, C.; Pascanu, R.; et al. 2024. Improving fine-grained understanding in image-text pre-training. *arXiv preprint arXiv:2401.09865*.

Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; and Ma, T. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32.

Chen, Z.-M.; Wei, X.-S.; Wang, P.; and Guo, Y. 2019. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5177–5186.

Chu, P.; Bian, X.; Liu, S.; and Ling, H. 2020. Feature space augmentation for long-tailed data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, 694–710. Springer.

Cui, J.; Liu, S.; Tian, Z.; Zhong, Z.; and Jia, J. 2022. Reslt: Residual learning for long-tailed recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9268–9277.

Estabrooks, A.; Jo, T.; and Japkowicz, N. 2004. A multiple resampling method for learning from imbalanced data sets. *Computational intelligence*, 20(1): 18–36.

Everingham, M.; Eslami, S.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2015. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1): 98–136.

Gao, Y.; Liu, J.; Xu, Z.; Zhang, J.; Li, K.; Ji, R.; and Shen, C. 2022. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *Advances in neural information processing systems*, 35: 35959–35970.

Goel, S.; Bansal, H.; Bhatia, S.; Rossi, R.; Vinay, V.; and Grover, A. 2022. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35: 6704–6719.

Guo, H.; and Wang, S. 2021. Long-tailed multi-label visual recognition by collaborative training on uniform and re-balanced samplings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15089–15098.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 4904–4916. PMLR.

Jin, Y.; Li, M.; Lu, Y.; Cheung, Y.-m.; and Wang, H. 2023. Long-tailed visual recognition via self-heterogeneous integration with knowledge excavation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 23695–23704.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, 5583–5594. PMLR.

Lee, J.; and Toutanova, K. 2018. Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 3(8).

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, W.; and Tsang, I. 2015. On the optimality of classifier chain for multi-label classification. *Advances in Neural Information Processing Systems*, 28.

Liu, W.; Wang, H.; Shen, X.; and Tsang, I. W. 2021. The emerging trends of multi-label learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(11): 7955–7974.

Liu, X.-Y.; Wu, J.; and Zhou, Z.-H. 2008. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2): 539–550.

Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; and Yu, S. X. 2019. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2537–2546.

Misra, I.; Lawrence Zitnick, C.; Mitchell, M.; and Girshick, R. 2016. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2930–2939.

Nitesh, V. C. 2002. SMOTE: synthetic minority oversampling technique. *J Artif Intell Res*, 16(1): 321.

Perez, L.; and Wang, J. 2017. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Rawlekar, S.; Bhatnagar, S.; Srinivasulu, V. P.; and Ahuja, N. 2024. Improving Multi-label Recognition using Class Co-Occurrence Probabilities. *arXiv preprint arXiv:2404.16193*.

Shen, F.; Shen, C.; Liu, W.; and Shen, H. T. 2015. Supervised Discrete Hashing. In *IEEE Conference on Computer Vision and Pattern Recognition*, 37–45.

Shen, F.; Xu, Y.; Liu, L.; Yang, Y.; Huang, Z.; and Shen, H. T. 2018. Unsupervised Deep Hashing with Similarity-Adaptive and Discrete Optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12): 3034–3044.

Shen, L.; Lin, Z.; and Huang, Q. 2016. Relay backpropagation for effective learning of deep convolutional neural networks. In *European conference on computer vision*, 467–482. Springer.

Tan, J.; Wang, C.; Li, B.; Li, Q.; Ouyang, W.; Yin, C.; and Yan, J. 2020. Equalization loss for long-tailed object recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11662–11671.

Wang, T.; Zhu, Y.; Zhao, C.; Zeng, W.; Wang, J.; and Tang, M. 2021. Adaptive class suppression loss for long-tail object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3103–3112.

Wang, Y.-X.; Ramanan, D.; and Hebert, M. 2017. Learning to model the tail. *Advances in neural information processing systems*, 30.

Wei, Y.; Xia, W.; Lin, M.; Huang, J.; Ni, B.; Dong, J.; Zhao, Y.; and Yan, S. 2015. HCP: A flexible CNN framework for multi-label image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(9): 1901–1907.

Wu, T.; Huang, Q.; Liu, Z.; Wang, Y.; and Lin, D. 2020. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *European Conference on Computer Vision*, 162–178. Springer.

Xia, P.; Xu, D.; Ju, L.; Hu, M.; Chen, J.; and Ge, Z. 2023. Lmpt: Prompt tuning with class-specific embedding loss for long-tailed multi-label visual recognition. *arXiv preprint arXiv:2305.04536*.

Yang, J.; Duan, J.; Tran, S.; Xu, Y.; Chanda, S.; Chen, L.; Zeng, B.; Chilimbi, T.; and Huang, J. 2022. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15671–15680.

Yao, L.; Huang, R.; Hou, L.; Lu, G.; Niu, M.; Xu, H.; Liang, X.; Li, Z.; Jiang, X.; and Xu, C. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

Zhang, W.; Liu, C.; Zeng, L.; Ooi, B.; Tang, S.; and Zhuang, Y. 2023a. Learning in imperfect environment: Multi-label classification with long-tailed distribution and partial labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1423–1432.

Zhang, Y.; Kang, B.; Hooi, B.; Yan, S.; and Feng, J. 2023b. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 10795–10816.

Zhang, Z.; and Pfister, T. 2021. Learning fast sample re-weighting without reward data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 725–734.

Zhao, C.; Wang, Y.; Jiang, X.; Shen, Y.; Song, K.; Li, D.; and Miao, D. 2024. Learning Domain Invariant Prompt for Vision-Language Models. *IEEE Transactions on Image Processing*, 33: 1348–1360.

Zhou, B.; Cui, Q.; Wei, X.-S.; and Chen, Z.-M. 2020. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9719–9728.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.