A APPENDIX

A.1 RELATED WORKS

A.1.1 MULTI-MODAL LARGE LANGUAGE MODEL

Large Language Models (LLMs) have recently significantly impacted the field of natural language processing. Through alignment techniques such as supervised learning and reinforcement learning with human feedback, LLMs can effectively generalize to perform a wide range of tasks, even with limited training data. A remarkable application of LLM is ChatGPT, which presents an amazing ability to interact with humans. OpenAI's ChatGPT and GPT4 are prime examples of the impact that AI can have, and there have been extensive open-source efforts to replicate their success, such as OPT Zhang et al. (2022), BLOOM Scao et al. (2022), PALM Chowdhery et al. (2022), LLaMA Touvron et al. (2023).

Multi-modal large language models have further promoted the development of the vision-language models Radford et al. (2021); Li et al. (2022d); Alayrac et al. (2022); Li et al. (2023); Zhu et al. (2023); Liu et al. (2023a); Chen et al. (2023); Yang et al. (2024; 2025b). CLIP Radford et al. (2021) was introduced to separately extract features from the visual encoder and the text encoder, and combine them using contrastive learning. CLIP supports a variety of downstream tasks, including image retrieval, image classification tasks and especially zero-shot classification tasks. But, it cannot generate detailed captions based on images due to the lack of a text decoder. In contrast, our model primarily addresses the concept drift issue within multi-modal large language models, since an image-grounded text decoder is employed to generate text based on the images. Besides, CLIP requires a large-scale and high-quality WIT dataset to be driven, that contains 37.6 million entity image-text samples with 11.5 million unique images across 108 Wikipedia languages. Whereas, our method is validated under the extended ImageNet-LT, which consists of only 115.8K imbalanced images-text pairs.

Building on CLIP, GLIP Li et al. (2022d) was developed to learn object-level, language-aware, and semantic-rich visual representations, unifying object detection and phrase grounding for pre-training. Different from the contrastive method, Flamingo Alayrac et al. (2022) aligned a pre-trained vision encoder and language model using gated cross-attention, demonstrating impressive few-shot learning capabilities. BLIP2 Li et al. (2023) was subsequently introduced, and it employed a Flan-T5 Chung et al. along with a Q-Former to effectively align visual features with the language model. MiniGPT4 Zhu et al. (2023), the most recent development in the field is the PaLM-E model, which features 562 billion parameters and is designed to integrate real-world continuous sensor modalities into an LLM, thereby establishing a connection between real-world perceptions and human languages. Based on Visual Fundamental Models like BLIP mentioned above, Visual ChatGPT adopts ChatGPT as the central component for interacting with users. It integrates multiple visual foundation models and utilizes prompt engineering, also known as Prompt Manager, to instruct ChatGPT about the usage, input-output format, and capabilities of each foundation model. This enables ChatGPT to determine how to invoke these models to fulfill the user's requirements. Besides, GPT-4V(ision) OpenAI (2023) and GPT-4O(mni) have recently shown unprecedented ability in understanding and processing an arbitrary mix of input images and texts.

A.1.2 LONG-TAILED OPEN WORLD

In vision tasks, significant efforts have been devoted to mitigating the challenges posed by the long-tailed open world. Two prominent research directions have emerged: long-tailed classification under open-world settings, exemplified by approaches like OLTR++ Liu et al. (2019; 2022b), LUNA Cai et al. (2022), DALC Wang et al. (2023), Open-sampling Wei et al. (2022) and TLC Li et al. (2022a), and OOD detection in long-tailed recognition, as seen in methods such as PASCL Wang et al. (2022), EAT Wei et al. (2024). OLTR++ Liu et al. (2019; 2022b) proposed an ensemble algorithm, consisting of dynamic meta-embedding to improve the recognition of tail categories and active learning for open categories detection. LUNA Cai et al. (2022) presented a distribution-sensitive loss to weigh more on the tail classes and a local-density-based metric to measure the novelty of OOD samples. DALC Wang et al. (2023) designed an active distribution optimization algorithm for clustering, querying and classification to balance the classification bias. Open-sampling Wei et al. (2022) rebalances class priors by sampling labels from a complementary distribution for each open-set

instance, mitigating class imbalance. TLC Li et al. (2022a) utilizes the Dempster-Shafer Evidence Theory in a multi-expert framework for uncertainty estimation of tail and OOD samples. PASCL Wang et al. (2022) applied supervised contrastive learning to explicitly boost the model to distinguish between tail-class in-distribution samples and OOD samples. EAT Wei et al. (2024) introduces abstention classes for clear decision boundaries and augmenting tail classes with context-rich OOD data to focus on discriminative features. MCM Ming et al. (2022) pioneers the integration of vision language models into OOD detection, enabling zero-shot OOD by aligning visual features with text concepts through a proposed maximum concept matching approach.

In addition, more and more VL methods have gained attention in the long-tail domain, such as LPT DONG et al. (2023), BALLAD Ma et al. (2021), Decoder Wang et al. (2024), VL-LTR Tian et al. (2022) and LIFT Shi et al. (2024). However, most of them pay attention to the fine-tuning of the vision language model under long-tailed scenarios. They directly use the pre-trained CLIP model, which is pre-trained using the high-quality and large-scale WIT dataset. In contrast, we are more concerned about the impact of long-tail open data on the whole model training from pre-training onwards, including pre-training and fine-tuning.

Additionally, in the domain of the language model, Kandpal et al. (2023) corroborates that large language models (LLMs) also struggle to learn long-tailed knowledge. While larger models are better at absorbing long-tailed knowledge, they estimate that current models must be scaled by many orders of magnitude to reach competitive performance. Besides, Raunak et al. alleviates the long-tail problem in neural machine translation by quantifying token classification and sequence generation, and introduces an anti-focus loss that incorporates beam search inductive biases to better adapt model training to conditional text generation.

A.1.3 Concept Drift

In the review Lu et al. (2019), the algorithms related to concept drift are categorized into three groups: error rate-based, data distribution-based and multiple hypothesis-based. Our proposed algorithm belongs to the distribution-based concept drift detection and adaptation method. Distribution-based concept drift algorithms not only accurately detect drift through explicit distributions but also analyze the drift to identify its happening timing, location, and severity.

Besides, RBM-IM Korycki & Krawczyk proposes a novel trainable concept drift detector based on Restricted Boltzmann Machine, to solve the concept drift in multi-class imbalanced data streams. Meanwhile, DDG-DA Li et al. initially trains a predictor to estimate future data distribution with concept drift, utilizes this information to create training samples, and subsequently trains models on the generated data. Furthermore, CALMID Liu et al. (2021) proposes a comprehensive active learning method for multiclass imbalanced streaming data with concept drift, including an ensemble classifier, a drift detector, and a variable threshold uncertainty strategy. Subsequently, DES-ICD Jiao et al. (2024) is a dynamic ensemble selection method for imbalanced data streams with concept drift. It considers the local performances of base classifiers and addresses class imbalance using a novel synthetic minority oversampling technique. Moreover, GOOD Gui et al. (2022) develops a graph OOD benchmark, which explicitly distinguishes between covariate and concept shifts and designs data splits that accurately capture these different shifts. Beyond that, ResilientCL Yang et al. (2025a) introduces a causal framework that integrates concept drift adaptation with structural causal modeling. By decoupling spurious correlations via causal graphs and enforcing counterfactual invariance, it addresses distributional biases in streaming training data. Besides, Liu et al. (2022a; 2023b; 2024) propose a multi-view uncertainty framework that addresses concept drift across heterogeneous data streams through set-valued prediction generation, effectively consolidating probabilistic outputs into deterministic categorical representations.

Remark A.1. Differences: Concept Drift vs. Data Drift (Covariate Drift) *Data drift entails changes solely in the distribution of inputs $P(x)$, while concept drift involves alterations in both input and output distributions, i.e., $P(x)$ and $P(y)$, leading to changes in the decision boundary. Furthermore, data drift predominantly stems from internal factors like data collection and processing, whereas concept drift typically arises from external factors, reflecting real-world changes.*

A.1.4 HYPERSPHERICAL DISTRIBUTION MODELLING

The Bayesian estimation of the vMF mixture model with variational inference is addressed in Taghia et al.. The learning task in VI consists of the optimization of the variational posterior distribution. Besides, a deep metric learning model for image classification and retrieval is presented in Zhe et al., which utilizes the vMF distribution to define the loss function and introduces an effective alternative learning algorithm by updating class centers. The model captures global information in the embedding space and approximates the class distribution during training, leading to improved performance in image tasks. Kobayashi extends the vMF distribution to regularize the intra-class feature distribution for imbalanced, small-scale and noisy data. Yang et al. (2023) focus on using hyperspherical embedding to alleviate the crowding problem arisen by the imbalanced data. Ming et al. (2023) utilizes hyperspherical embeddings for OOD detection in representation learning, consisting of two losses, a dispersion loss to increase angular distances between different class prototypes, and a compactness loss to ensure samples are closer to their respective class prototypes. Besides, H-SRDC Tang et al. enhances intra-class compactness by combining target data clustering with a domain-shared classifier and cluster centroid learning, enhancing deep clustering by minimizing Kullback-Leibler divergence between network predictions and an auxiliary distribution.

A.2 THE T-DISTRIBUTED DISTRIBUTION ON HYPERSPHERE

A.2.1 DIRECTIONAL STATISTICS

Directional statistics primarily focus on the distribution of eigenvector angles, while neglecting the impact of eigenvector module lengths. Given the unit feature vector $X_{ij} \in \mathbb{S}^{d-1}$, where $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : ||x||_2 = 1\}$ denotes the $(d-1)$-dimensional hyperspherical set. A key idea in directional distribution is the tangent-normal decomposition. Any unit vector x can be decomposed as:

$$x = t\mu + (1-t^2)^{\frac{1}{2}} v, t \in [-1, 1], \tag{9}$$

with $v \in \mathbb{S}^{d-2}$ a tangent to $\mathbb{S}^{d-1}$ at μ Mardia & Jupp (2000); De Cao & Aziz (2020), where v and t are independent and v is uniform on $\mathbb{S}^{d-2}$. Thus, the intersection of $\mathbb{S}^{d-1}$ with the hyperplane through $t\mu$ and normal to μ is a $(d-2)$-dimensional sphere of radius $\sqrt{1-t^2}$, that t has density as following:

$$p_T(t; d) \propto (1-t^2)^{\frac{d-3}{2}}, t \in [-1, 1]. \tag{10}$$

Therefore, through the marginal density p_T and p_v, we can estimate the density of the entire spherical distribution. One prominent instance is the von Mises-Fisher distribution (vMF) Banerjee et al. (2005), which can be interpreted as a probability distribution over the cosine similarity between a unit vector x and a fixed mean direction μ, following the density:

$$p_X(x; \mu, \kappa) \propto \exp(\kappa \mu^T x), \tag{11}$$

where $\kappa \geqslant 0$ denotes the concentration and exp represents the exponential function. Therefore, combined with the Eq. 9 and Eq. 10, the density of vMF is:

$$\begin{aligned} p(x) &= C_X(\kappa, d)^{-1} \exp(\kappa \mu^T x), \quad x \sim \text{vMF}(\mu, \kappa) \\ C_X(\kappa, d) &= \frac{(2\pi)^{d/2} I_{d/2-1}(\kappa)}{\kappa^{d/2-1}}, \end{aligned} \tag{12}$$

where I_m denotes the modified Bessel function of the first kind at order m.

A.2.2 DERIVATION OF THE T-DISTRIBUTED DISTRIBUTION ON HYPERSPHERE

Given the unit feature vector $X_{ij} \in \mathbb{S}^{d-1}$, where $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : ||x||_2 = 1\}$ denotes the $(d-1)$-dimensional hyperspherical set. The proposed T-distribution metric on hypersphere (Thp) follows the density:

$$p_X(x) \propto \frac{2}{\kappa(1-\mu^T x)}, \tag{13}$$

where $x \in \mathbb{S}^{d-1}$, direction $\mu in \mathbb{S}^{d-1}$ and concentration $\kappa \in \mathbb{R}_{\geq 0}$. Let T bet a random variable that denotes the dot-product $t = \mu^T x$, then $T = 2Z - 1$, with $Z \sim \text{Beta}(\alpha, \beta)$, where $\alpha = \frac{d-1}{2}$ and $\frac{d-3}{2}$.

Proof. Given Eq. 10, the marginal distribution of the dot-product t is

$$t \propto \frac{2}{\kappa(1-t)}(1-t^2)^{\frac{d-3}{2}}. \tag{14}$$

So, its normalizer is:

$$\begin{aligned} N_T(\kappa, d) &= \int_{\mathbb{S}^{d-1}} \frac{2}{\kappa(1-t)}(1-t^2)^{\frac{d-3}{2}}\mathrm{d}t \\ &= \int_{-1}^{1} \frac{1}{\kappa(1-t)}(1+t)^{\frac{d-3}{2}}(1-t)^{\frac{d-3}{2}}\mathrm{d}t \\ &= \frac{1}{\kappa}\int_{-1}^{1}(1+t)^{\frac{d-3}{2}}(1-t)^{\frac{d-5}{2}}\mathrm{d}t. \end{aligned} \tag{15}$$

Given the useful integral function:

$$\int (1+x)^a(1-x)^b\mathrm{d}x = 2^{a+b+1}B_{\frac{x+1}{2}}(a+1, b+1) + C. \tag{16}$$

So, its normalizer is:

$$\begin{aligned} N_T(\kappa, d) &= \frac{1}{\kappa}2^{d-3}(B_1(\frac{d-1}{2}, \frac{d-3}{2}) - B_0(\frac{d-1}{2}, \frac{d-3}{2})) \\ &= \frac{1}{\kappa}2^{d-3}B(\frac{d-1}{2}, \frac{d-3}{2}). \end{aligned} \tag{17}$$

The Beta function:

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \tag{18}$$

So, the normalizer is

$$N_T(\kappa, d) = \frac{1}{\kappa}2^{\alpha+\beta-1}\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \tag{19}$$

where, $\alpha = \frac{d-1}{2}$ and $\beta = \frac{d-3}{2}$. It follows that the probability density function of the marginal distribution of the dot product is,

$$\begin{aligned} p_T(t; \kappa, d) &= N_T(\kappa, d)^{-1}\frac{2}{\kappa(1-t)}(1-t^2)^{\frac{d-3}{2}} \\ &= N_T(\kappa, d)^{-1}\frac{2}{\kappa}(1+t)^{\frac{d-3}{2}}(1-t)^{\frac{d-5}{2}} \\ &= N_T(\kappa, d)^{-1}\frac{2}{\kappa}(2z)^{\frac{d-1}{2}-1}(2-2z)^{\frac{d-3}{2}-1} \\ &= \frac{2}{\kappa}B(\alpha, \beta)^{-1}z^{\alpha-1}(1-z)^{\beta-1}, \end{aligned} \tag{20}$$

where, $\alpha = \frac{d-1}{2}$ and $\beta = \frac{d-3}{2}$. □

Due to the surface area of the hyper-sphere $\mathbb{S}^{d-1}$ is:

$$A_{d-1} = \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})}. \tag{21}$$

The T-distributed spherical distribution is expressed via the tangent normal decomposition as a joint distribution between $T \sim p_T t; \kappa, d$ and $V \sim \mathcal{U}(\mathbb{S}^{d-2})$. Since $T \perp\!\!\!\perp V$, the Thp normalizer $N_x(p, k)$ is the product of the normalizer of $p_T(t; \kappa, d)$ and the uniform distribution on $\mathbb{S}^{d-2}$ is:

$$\begin{aligned} N_X(\kappa, d) &= N_T(\kappa, d) \cdot A_{d-2} \\ &= 2^{\alpha+\beta-1}B(\alpha, \beta)\frac{2\pi^{\beta}}{\kappa\Gamma(\beta)} \\ &= \frac{2^{\alpha+\beta}\pi^{\beta}}{\kappa}\frac{\Gamma(\alpha)}{\Gamma(\alpha+\beta)}. \end{aligned} \tag{22}$$

Thus,

$$p_X(x; \mu, \kappa) = N_X(\kappa, d)^{-1}\frac{2}{\kappa(1-\mu^T x)}. \tag{23}$$

A.3 IMPLEMENTATION DETAILS

For our language-guided image tokenizer, we leverage the strengths of both BERT Devlin et al. (2019b) and ViT as our text encoder, text decoder and visual encoder, respectively.

We employ ViT-Bae as our visual encoder, which consists of 12 transformer encoder layers and an FFN intermediate size of 3,072. The input image size is set to 384×384, with a patch size of 16×16. The hidden dimensions of the ViT-Base are 768, with 12 attention heads. And, the number of parameters is about 86 million. Besides, we also use ResNeXt-50 to perform ablation experiments. In addition, ResNeXt-50 has 16 residual blocks with 50 layers. Each block has 3 convolutional layers with the kernel size of 3×3, the stride of 1 and the padding of 1. The batch normalization and max pooling are utilized to connect the convolutional layers. The classification head hidden dimensions are 2,048.

Additionally, BERT as the language model in our vision-language model, has 12 transformer layers with 768 hidden dimensions and 3,078 intermediate dimensions. The number of attention heads is 12, with the input sequence length of 512. It has approximately 110 million parameters.

In terms of the pre-training progress, the hyperparameters are presented in Table 7. We utilize the AdamW optimizer, which is configured with a cosine annealing schedule as the learning policy. The initial learning rate is set to 2×10^{-5}, and the AdamW optimizer is employed with hyperparameters $\beta = (0.9, 0.98)$. Additionally, we set the weight decay to 0.05 and the dropout rate to 0.1. During the first 1,000 warm-up steps, the learning rate increases to 2×10^{-5}, and subsequently decays to 10^{-7}. Unless otherwise specified, the pre-training of our vision language model consists of 800,000 steps, executed on 2×2 NVIDIA A100 GPUs. And the pre-training experiments are conducted in the manner of different stages, namely gradual drifts with long-tailed data and sudden drifts with OOD data. It is mainly to compare with different methods with the same setup.

Table 7: The training hyperparameters of our vision language model.

Pre-training	
Training Steps	400,000
Warmup Steps	1,000
Optimizer	AdamW
Learning Rate	1e-4
Learning Rate Decay	Cosine
Adam β	(0.9, 0.98)
Weight Decay	0.05
Batch Size	50

Fine-tuning	
Training Steps	18,000
Warmup Steps	0
Optimizer	AdamW
Learning Rate	2e-5
Learning Rate Decay	Cosine
Adam β	(0.9, 0.98)
Weight Decay	0.05
Batch Size	400

While in the fine-tuning on the downstream task of classification, the initial learning rate is reduced to 10^{-6} without the warmup. The visual encoder and text decoder are frozen out of the training. Thus, the batch size can be increased to 400. The fine-tuning consists of 18,000 steps, executed on 2×2 NVIDIA A100 GPUs. Other training parameters are the same as the pre-training. Besides, under the only fine-tuning settings, the image encoder and the text encoder are frozen with the CLIP pre-trained parameters, while the image-grounded text decoder is trained during the fine-tuning.

When evaluating the performance of our VL model under the long-tailed open world, we use the top-1 accuracy metric on the downstream classification task. In particular, the categories are split into three groups: many-shot (with more than 100 training samples), medium-shot (with 20-100 training samples), and few-shot (with fewer than 20 training samples). The Top-1 accuracies are computed for each group to evaluate the performance of mitigating the bias introduced by the long-tail distribution, respectively. Furthermore, in order to assess the capability of detecting the OOD drift, we employ two metrics: FPR95 which measures the false positive rate of OOD samples when the true positive rate of ID samples reaches 95%, and AUROC providing the area under the receiver operating characteristic curve. Besides, cosine distance is exploited to measure the distances between features and centers in the feature space of the VL model.

Caption: *The image depicts a [mask], also known as a [mask], sitting on a branch of a tree. The [mask] is holding a leaf in its mouth, which suggests that it might be eating or chewing on the plant. This behavior is typical of [mask]s, as they primarily feed on bamboo shoots, leaves, fruits, and insects. In the wild, [mask]s are found in the mountainous regions of southern and southwestern China, Myanmar, and India.*
Annotation: *lesser panda, red panda, panda, bear cat, cat bear, Ailurus fulgens*

(a) Sample in Training Set

Caption: *The picture depicts a young man sitting on a bench, holding a [mask] in his hand. This suggests that he is either playing the [mask] or contemplating playing it. The [mask] is a musical instrument that is commonly associated with blues and folk music, and it can be used to create melodic and rhythmic sounds. The presence of the [mask] in the image adds a musical element to the scene.*

Annotation: *harmonica, mouth organ, harp, mouth harp*

(b) Sample in Test Set

Caption: *The main object in the picture is an open suitcase, which is a type of luggage. It is red in color and appears to be medium-sized. The suitcase is located on the floor of a room. The suitcase is partially filled with clothing items, including shirts, pants, and socks. It appears that the suitcase is still in the process of being packed or unpacked, as some items are visible on top of the suitcase while others are spilling out of it. The suitcase is open, allowing easy access to the clothing items inside. Overall, the picture provides a glimpse into the process of preparing for a trip or organizing one's belongings.*

(c) Sample in Open Set

Figure 4: Samples of OpenMMlo in training set, test set and open set.

A.4 Building Multi-modal Long-Tailed OOD Datasets Group OpenMMlo

Figure 4 showcases the samples utilized for training and validation in our study. To intuitively verify the impact of long-tail open-world scenarios on multi-modal large language models, we employ classification as our downstream task. When matching images and texts, we strategically mask words that are directly related to category names. This approach ensures the accuracy and reliability of

our experimental results. As depicted in Figure 4, comprehensive descriptions of the image are provided through long-form text, encompassing details such as size, position, color, relationships, and other relevant information about the objects present in the image. This ensures a detailed and information-rich depiction of the visual content. We have publicly released the datasets used for training and validation, as well as the original unmasked datasets.

ACKNOWLEDGMENT

The work was supported by the Australian Research Council (ARC) under Laureate project FL190100149.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, Suvrit Sra, and Greg Ridgeway. Clustering on the Unit Hypersphere using von Mises-Fisher Distributions. *Journal of Machine Learning Research*, 6(9):1345–1382, 2005.

Jiarui Cai, Yizhou Wang, and Jenq-Neng Hwang. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 112–121, 2021.

Jiarui Cai, Yizhou Wang, Hung-Min Hsu, Jenq-Neng Hwang, Kelsey Magrane, and Craig S. Rose. LUNA: Localizing Unfamiliarity Near Acquaintance for Open-Set Long-Tailed Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1):131–139, 2022. ISSN 2374-3468. doi: 10.1609/aaai.v36i1.19887.

Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al. Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*, 2023.

Eulrang Cho, Jooyeon Kim, and Hyunwoo J. Kim. Distribution-Aware Prompt Tuning for Vision-Language Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22004–22013, 2023.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling Instruction-Finetuned Language Models. URL http://arxiv.org/abs/2210.11416.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing Textures in the Wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014.

Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric Contrastive Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 715–724, 2021.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N. Fung, and Steven Hoi. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *Advances in Neural Information Processing Systems*, 36:49250–49267, 2023.

Nicola De Cao and Wilker Aziz. The power spherical distribution. *arXiv preprint arXiv:2006.04437*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019a.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019b. doi: 10.18653/v1/N19-1423.

Bowen DONG, Pan ZHOU, Shuicheng YAN, and Wangmeng ZUO. LPT: Long-tailed prompt tuning for image classification. pp. 1–20, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021. URL http://arxiv.org/abs/2010.11929.

Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. GOOD: A Graph Out-of-Distribution Benchmark. *Advances in Neural Information Processing Systems*, 35:2059–2073, 2022.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning Deep Representation for Imbalanced Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5375–5384, 2016.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual Prompt Tuning. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 709–727. Springer Nature Switzerland, 2022. ISBN 978-3-031-19827-4. doi: 10.1007/978-3-031-19827-4_41.

Botao Jiao, Yinan Guo, Shengxiang Yang, Jiayang Pu, and Dunwei Gong. Reduced-space multi-stream classification based on multi-objective evolutionary optimization. *IEEE Transactions on Evolutionary Computation*, 2022.

Botao Jiao, Yinan Guo, Dunwei Gong, and Qiuju Chen. Dynamic ensemble selection for imbalanced data streams with concept drift. *IEEE Transactions on Neural Networks and Learning Systems*, 35 (1):1278–1291, 2024. doi: 10.1109/TNNLS.2022.3183120.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large Language Models Struggle to Learn Long-Tail Knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 15696–15707. PMLR, 2023.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling Representation and Classifier for Long-Tailed Recognition. In *Eighth International Conference on Learning Representations (ICLR)*, 2019.

Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy Anchor Loss for Deep Metric Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3238–3247, 2020.

Takumi Kobayashi. T-vMF Similarity For Regularizing Intra-Class Feature Distribution. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6612–6621. IEEE. ISBN 978-1-66544-509-2. doi: 10.1109/CVPR46437.2021.00655.

Lukasz Korycki and Bartosz Krawczyk. Concept Drift Detection from Multi-Class Imbalanced Data Streams. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pp. 1068–1079. doi: 10.1109/ICDE51399.2021.00097.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009. URL http://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf.

Bolian Li, Zongbo Han, Haining Li, Huazhu Fu, and Changqing Zhang. Trustworthy Long-Tailed Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6970–6979, 2022a.

Jun Li, Zichang Tan, Jun Wan, Zhen Lei, and Guodong Guo. Nested Collaborative Learning for Long-Tailed Visual Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6949–6958, 2022b.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. 34:9694–9705, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022c.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training, 2022d.

Wendi Li, Xiao Yang, Weiqing Liu, Yingce Xia, and Jiang Bian. DDG-DA: Data Distribution Generation for Predictable Concept Drift Adaptation. 36(4):4092–4100. ISSN 2374-3468. doi: 10.1609/aaai.v36i4.20327.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023a.

Wei Liu, Xiaodong Yue, Yufei Chen, and Thierry Denoeux. Trusted multi-view deep learning with opinion aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7585–7593, 2022a.

Wei Liu, Yufei Chen, Xiaodong Yue, Changqing Zhang, and Shaorong Xie. Safe multi-view deep classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8870–8878, 2023b.

Wei Liu, Yufei Chen, and Xiaodong Yue. Building trust in decision with conformalized multi-view deep classification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7278–7287, 2024.

Weike Liu, Hang Zhang, Zhaoyun Ding, Qingbao Liu, and Cheng Zhu. A comprehensive active learning method for multiclass imbalanced data streams with concept drift. *Knowledge-Based Systems*, 215:106778, 2021. ISSN 0950-7051. doi: 10.1016/j.knosys.2021.106778.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-Scale Long-Tailed Recognition in an Open World. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2532–2541. IEEE, 2019. ISBN 978-1-72813-293-8. doi: 10.1109/CVPR.2019.00264.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Open Long-Tailed Recognition In A Dynamic World. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–15, 2022b. ISSN 1939-3539. doi: 10.1109/TPAMI.2022.3200091.

Jie Lu, Anjin Liu, Fan Dong, Feng Gu, João Gama, and Guangquan Zhang. Learning under Concept Drift: A Review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12):2346–2363, 2019. ISSN 1558-2191. doi: 10.1109/TKDE.2018.2876857.

Teli Ma, Shijie Geng, Mengmeng Wang, Jing Shao, Jiasen Lu, Hongsheng Li, Peng Gao, and Yu Qiao. A Simple Long-Tailed Recognition Baseline via Vision-Language Model, 2021. URL http://arxiv.org/abs/2111.14745.

K. V. Mardia and Peter E. Jupp. *Directional Statistics*. Wiley Series in Probability and Statistics. J. Wiley, 2000. ISBN 978-0-471-95333-3.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into Out-of-Distribution Detection with Vision-Language Representations. *Advances in Neural Information Processing Systems*, 35:35087–35102, 2022.

Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection?, 2023. URL http://arxiv.org/abs/2203.04450.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, volume 2011, pp. 7. Granada, Spain, 2011.

OpenAI. Gpt-4v(ision) system card, 2023. URL https://cdn.openai.com/papers/GPTV_System_Card.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

Vikas Raunak, Siddharth Dalmia, Vivek Gupta, and Florian Metze. On Long-Tailed Phenomena in Neural Machine Translation. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3088–3095. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.276.

Jiawei Ren, Cunjun Yu, shunan sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and hongsheng Li. Balanced Meta-Softmax for Long-Tailed Visual Recognition. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4175–4186. Curran Associates, Inc., 2020.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein. Imagenet large scale visual recognition challenge. 115(3):211–252, 2015a.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015b. doi: 10.1007/s11263-015-0816-y.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

Vikash Sehwag, Mung Chiang, and Prateek Mittal. SSD: A Unified Framework for Self-Supervised Outlier Detection. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=v5gjXpmR8J.

Jiang-Xin Shi, Tong Wei, Zhi Zhou, Jie-Jing Shao, Xin-Yan Han, and Yu-Feng Li. Long-Tail Learning with Foundation Model: Heavy Fine-Tuning Hurts. In *Forty-First International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=ccSSKTz9LX.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-Distribution Detection with Deep Nearest Neighbors. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 20827–20840. PMLR, 2022.

Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. CSI: Novelty Detection via Contrastive Learning on Distributionally Shifted Instances. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11839–11852. Curran Associates, Inc., 2020.

Jalil Taghia, Zhanyu Ma, and Arne Leijon. Bayesian Estimation of the von-Mises Fisher Mixture Model with Variational Inference. 36(9):1701–1715. ISSN 1939-3539. doi: 10.1109/TPAMI.2014.2306426.

Hui Tang, Xiatian Zhu, Ke Chen, Kui Jia, and C. L. Philip Chen. Towards Uncovering the Intrinsic Data Structures for Unsupervised Domain Adaptation Using Structurally Regularized Deep Clustering. 44(10):6517–6533. ISSN 1939-3539. doi: 10.1109/TPAMI.2021.3087830.

Changyao Tian, Wenhai Wang, Xizhou Zhu, Jifeng Dai, and Yu Qiao. VL-LTR: Learning Class-wise Visual-Linguistic Representation for Long-Tailed Visual Recognition. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 73–91. Springer Nature Switzerland, 2022. ISBN 978-3-031-19806-9. doi: 10.1007/978-3-031-19806-9_5.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. DeiT III: Revenge of the ViT. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 516–533. Springer Nature Switzerland, 2022. ISBN 978-3-031-20053-3. doi: 10.1007/978-3-031-20053-3_30.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 2008.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The INaturalist Species Classification and Detection Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8769–8778, 2018.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning Robust Global Representations by Penalizing Local Predictive Power. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/3eefceb8087e964f89c2d59e8a249915-Abstract.html.

Haotao Wang, Aston Zhang, Yi Zhu, Shuai Zheng, Mu Li, Alex J. Smola, and Zhangyang Wang. Partial and Asymmetric Contrastive Learning for Out-of-Distribution Detection in Long-Tailed Recognition. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 23446–23458. PMLR, 2022.

Min Wang, Lei Zhou, Qian Li, and An-an Zhang. Open world long-tailed data classification through active distribution optimization. *Expert Systems with Applications*, 213:119054, 2023. ISSN 0957-4174. doi: 10.1016/j.eswa.2022.119054.

Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed Recognition by Routing Diverse Distribution-Aware Experts. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=D9I3drBz4UC.

Yidong Wang, Zhuohao Yu, Jindong Wang, Qiang Heng, Hao Chen, Wei Ye, Rui Xie, Xing Xie, and Shikun Zhang. Exploring Vision-Language Models for Imbalanced Learning. 132(1):224–237, 2024. ISSN 1573-1405. doi: 10.1007/s11263-023-01868-w.

Hongxin Wei, Lue Tao, Renchunzi Xie, Lei Feng, and Bo An. Open-Sampling: Exploring Out-of-Distribution data for Re-balancing Long-tailed datasets. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 23615–23630. PMLR, 2022.

Tong Wei, Bo-Lin Wang, and Min-Ling Zhang. EAT: Towards Long-Tailed Out-of-Distribution Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(14):15787–15795, 2024. ISSN 2374-3468. doi: 10.1609/aaai.v38i14.29508.

Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R. Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, Taylan Cemgil, S. M. Ali Eslami, and Olaf Ronneberger. Contrastive Training for Improved Out-of-Distribution Detection. URL http://arxiv.org/abs/2007.05566.

Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated Residual Transformations for Deep Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1492–1500, 2017.

Pingmei Xu, Krista A. Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R. Kulkarni, and Jianxiong Xiao. TurkerGaze: Crowdsourcing Saliency with Webcam based Eye Tracking. URL http://arxiv.org/abs/1504.06755.

Zhengzhuo Xu, Ruikang Liu, Shuo Yang, Zenghao Chai, and Chun Yuan. Learning Imbalanced Data With Vision Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15793–15803, 2023.

Xiaoyu Yang, Yufei Chen, Xiaodong Yue, Shaoxun Xu, and Chao Ma. T-distributed Spherical Feature Representation for Imbalanced Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):10825–10833, 2023. ISSN 2374-3468. doi: 10.1609/aaai.v37i9.26284.

Xiaoyu Yang, Lijian Xu, Hao Sun, Hongsheng Li, and Shaoting Zhang. Enhancing visual grounding and generalization: A multi-task cycle training approach for vision-language models. *arXiv preprint arXiv:2311.12327*, 2024. URL https://arxiv.org/abs/2311.12327.

Xiaoyu Yang, Jie Lu, and En Yu. Causal-informed contrastive learning: Towards bias-resilient pre-training under concept drift. *arXiv preprint arXiv:2502.07620*, 2025a. URL https://arxiv.org/abs/2502.07620.

Xiaoyu Yang, Lijian Xu, Hongsheng Li, and Shaoting Zhang. One leaf reveals the season: Occlusion-based contrastive learning with semantic-aware views for efficient visual representation. *arXiv preprint arXiv:2411.09858*, 2025b. URL https://arxiv.org/abs/2411.09858.

En Yu, Jie Lu, Bin Zhang, and Guangquan Zhang. Online boosting adaptive learning under concept drift for multistream classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16522–16530, 2024.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop. URL http://arxiv.org/abs/1506.03365.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

Xuefei Zhe, Shifeng Chen, and Hong Yan. Directional statistics-based deep metric learning for image classification and retrieval. 93:113–123. ISSN 0031-3203. doi: 10.1016/j.patcog.2019.04.005.

Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16489–16498, 2021.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to Prompt for Vision-Language Models. 130(9):2337–2348, 2022. ISSN 1573-1405. doi: 10.1007/s11263-022-01653-1.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023.