Supplementary Material for "DALIP: Distribution Alignment-based Language-Image Pre-Training for Domain-Specific Data"

Junjie Wu[1†], Jiangtao Xie[2†], Zhaolin Zhang[1], Qilong Wang[1*], Qinghua Hu[1], Peihua Li[2], Sen Xu[3,4]
[1]Tianjin University [2]Dalian University of Technology [3]Yancheng Institute of Technology [4]Harbin Engineering University

S1. Complexity Analysis

To analyze the computational efficiency of our DALIP, we compare it with the original CLIP [2] in terms of training time per batch, inference time per image, and convergence speed. All comparisons are conducted on 8 NVIDIA A100 GPUs with batch sizes of 2048 and 256 for training and testing, respectively. As shown in Table S2, our DALIP increases training time by 0.03 seconds per batch over the original CLIP. However, as shown in Figure S1, DALIP has a faster convergence speed than the original CLIP. Notably, DALIP tuned within about 20 epochs achieves comparable results with the original CLIP tuned within 40 epochs, which helps the models reach an expected result by using fewer training epochs and reduces training time. For inference, DALIP brings an additional 0.93 ms per image, which is affordable for practical applications. In conclusion, DALIP can achieve a better trade-off between efficiency and effectiveness.

Models	Qwen2-VL-7B	Qwen2-VL-72B	InternVL2-8B
Acc. (%)	91	77	87
Latency (s)	0.1	1.0	0.2

Table S1. Comparison of Caption Generation Quality and Generation Time. The captions are generated using various open-source VLLMs and evaluated by GPT-4o to assess their accuracy and identify potential hallucinations. Latency is measured as the average time required to produce a single caption.

S2. More Examples on Generated Descriptions

As illustrated in Fig. S2, we show more examples on the generation of precise and detailed plant descriptions by prompting Qwen2VL-7B with Latin and common names, images, and customized instructions. Clearly, generated descriptions are different from those utilized by CLIP.

	Training Time (s)	Inference Time (ms)
CLIP	1.64	4.02
DALIP$_{\mathrm{MP}}$	1.80	5.21
DALIP$_{\mathrm{MBDC}}$ (Ours)	1.67	4.95

Table S2. Computational complexity of CLIP, DALIP$_{\mathrm{MP}}$ and DALIP$_{\mathrm{MBDC}}$ in terms of training time per batch (s) and inference time per image (ms).

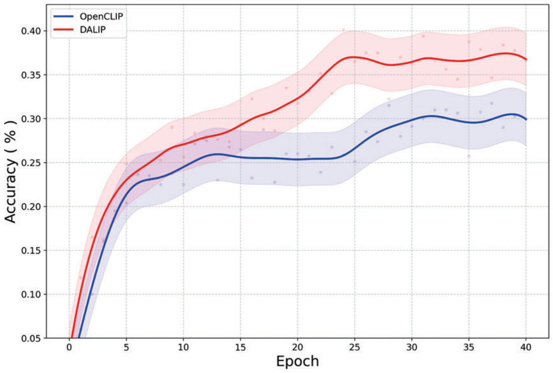

Figure S1. Convergence speed for DALIP and OpenCLIP with tuning on TOL-1M, where accuracies on Fungi are reported. For briefness, we show the results within the first 40 training epochs.

S3. Quality of MLLM-generated Captions

To ensure the quality of generated captions while maintaining computational efficiency, we conducted a systematic evaluation of open-source MLLMs. We compare caption quality and inference efficiency of Qwen2-VL-7B [3], Qwen2-VL-72B [3] and InternVL2-8B [1], aiming to balance practicality and scalability under budget constraints. While GPT-4o achieves SOTA performance, its cost is prohibitive for large-scale datasets: our PlantMix with 13M images would require approximately $150k. To validate caption accuracy and detect hallucinations, we sampled 1K generated captions (on 8 NVIDIA A100 GPUs) and evaluated them using GPT-4o. As shown in Tab S1, Qwen2-VL-7B achieves 91% accuracy and clearly surpasses Qwen2-VL-72B (77%) and InternVL2-8B (87%), Qwen2-vl-7B generates a caption in an average of 0.1s, significantly faster

[*]Corresponding author. †Equal contribution.
E-mail: {wjj_, qlwang}@tju.edu.cn.
Project page: https://github.com/XavierHeart/DALIP.

than Qwen2-vl-72B, which takes 1.0s per caption. It also outperforms InternVL2-8B by 0.1s in generation speed.

S4. Detailed Results of Ablation Study

As shown in Table S3, we give the detailed results on five plant domain tasks (*e.g.*, PlantNet, Fungi, PlantVillage, Medicinal Leaf (Med. Leaf) and PlantDoc, as discussed in Sec. 5.4 of manuscript. From it, we can see that performance on each dataset is consistent with the average.

S5. More Ablation Studies

Here, we further assess the effect of input resolution and visual encoder on our DALIP.

Comparison of Larger Resolution. As shown in the middle of Table S3, we compare CLIP with DALIP models by using two input resolutions. Specifically, we take the models tuned on 224x224 resolution inputs and tune them for 10 epochs using 336x336 resolutions continuously. The results demonstrate that increasing the resolution improves performance for both CLIP and DALIP. CLIP$_{336}$ achieves a 0.7% increase in mean performance over CLIP$_{224}$, while DALIP$_{336}$ shows a more substantial 1.4% improvement over DALIP$_{224}$. Notably, DALIP$_{336}$ outperforms CLIP$_{336}$ by 2.5% on average (51.0% vs. 49.2%).

Visual Encoder of ConvNEXT-base. To further assess the effect of visual encoder, we employ ConvNEXT-base as an alternative to the ViT-B/16 backbone, and compare CLIP with DALIP. As shown in the last two rows of Table S3, DALIP with ConvNEXT-base backbone outperforms CLIP$_{\text{ConvNEXT}}$ across all plant domain tasks. The average performance of DALIP$_{\text{ConvNEXT}}$ is 1.4% higher than that of CLIP$_{\text{ConvNEXT}}$, demonstrating the effectiveness of our DALIP across different visual encoders. Notably, DALIP$_{\text{ConvNEXT}}$ shows significant improvements in tasks such as Fungi and PlantDoc, leading by 1.7% and 1.1% respectively. These results suggest that DALIP is not limited to a specific backbone architecture but can be generalized to other architectures like ConvNEXT.

Sensitivity analysis for λ_1 and λ_2. To investigate how varying λ_1 (with $\lambda_1 + \lambda_2 = 1$) affects model performance on Plant test set, we conduct experiments using multiple different values of λ_1. As shown in Table S4, performance remains stable for λ_1 values between 0.3 and 0.6 (48.1%–49.3%), demonstrating robustness to parameter variations in this range. However, the 8.8% performance gap between optimal (λ_1=0.4) and worst-case (λ_1=1.0) configurations underscores the necessity of avoiding unbalanced weightings.

S6. Comparison of Prediction Examples

Fig. S3 presents a comparative analysis of zero-shot prediction examples between DALIP and BioCLIP across five plant domain tasks, wherein DALIP consistently exhibits superior performance. DALIP's enhanced feature extraction capabilities enable it to discern subtle morphological nuances in fungi, leaf structures, and disease symptoms. This refined ability is particularly evident in its accurate identification of Agaricus augustus Fr. and Alcea rosea, where BioCLIP struggles. Furthermore, DALIP demonstrates a more nuanced contextual understanding, accurately diagnosing plant diseases such as tomato leaf bacterial spot and strawberry leaf scorch. This suggests a heightened capacity to correlate visual cues with specific conditions. Notably, DALIP's consistent accuracy across diverse datasets (Fungi, Med. Leaf, PlantDoc, PlantNet, and PlantVillage) indicates the development of a more comprehensive and transferable knowledge base for plant-related tasks. These results collectively underscore DALIP's advanced proficiency in zero-shot learning scenarios within the plant domain.

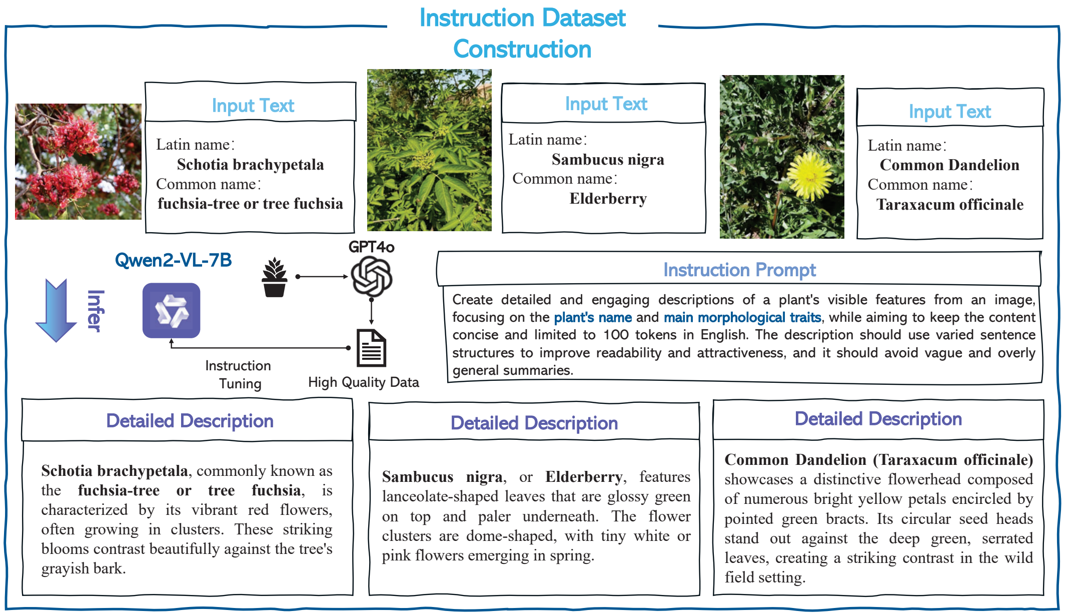
Instruction Dataset Construction
Input Text
Latin name: Schotia brachypetala
Common name: fuchsia-tree or tree fuchsia
Input Text
Latin name: Sambucus nigra
Common name: Elderberry
Input Text
Latin name: Common Dandelion
Common name: Taraxacum officinale
Qwen2-VL-7B
Infer
GPT4o
Instruction Tuning
High Quality Data
Instruction Prompt
Create detailed and engaging descriptions of a plant's visible features from an image, focusing on the plant's name and main morphological traits, while aiming to keep the content concise and limited to 100 tokens in English. The description should use varied sentence structures to improve readability and attractiveness, and it should avoid vague and overly general summaries.
Detailed Description
Schotia brachypetala, commonly known as the fuchsia-tree or tree fuchsia, is characterized by its vibrant red flowers, often growing in clusters. These striking blooms contrast beautifully against the tree's grayish bark.
Detailed Description
Sambucus nigra, or Elderberry, features lanceolate-shaped leaves that are glossy green on top and paler underneath. The flower clusters are dome-shaped, with tiny white or pink flowers emerging in spring.
Detailed Description
Common Dandelion (Taraxacum officinale) showcases a distinctive flowerhead composed of numerous bright yellow petals encircled by pointed green bracts. Its circular seed heads stand out against the deep green, serrated leaves, creating a striking contrast in the wild field setting.

Figure S2. More examples of generating precise and detailed plant descriptions using qwen2VL-7B, based on Latin and Common names, images, and tailored instruction prompts.

Model	Dataset	Imagenet	Plants & Fungi					Plant Mean	Mean
			PlantNet	Fungi	PlantVillage	Med. Leaf	PlantDoc		
Effect of 1st-&2nd-Order Statistics									
DALIP$_{1st}$	PlantMix-1.3M	30.7	79.8	56.3	20.3	34.9	21.4	42.5	36.6
DALIP$_{2nd}$		**32.9**	**81.1**	**58.4**	**24.2**	**36.8**	**23.5**	**44.8**	**38.9**
Comparison of Second-Order Counterparts									
DALIP$_{\text{MP}}$	PlantMix-1.3M	36.0	78.4	54.3	28.7	38.2	18.4	43.6	39.8
DALIP$_{\text{BDC}}$		36.1	82.4	58.0	31.5	41.5	21.6	46.9	41.5
DALIP$_{\text{DeepKSPD}}$		31.9	78.8	55.2	24.7	35.1	20.2	42.8	37.4
DALIP		**36.8**	**85.0**	**61.3**	**33.0**	**43.8**	**23.5**	**49.3**	**43.1**
Comparison of Larger Resolution									
CLIP$_{224}$	PlantMix-13M	46.8	89.9	47.0	32.3	48.9	33.0	50.2	48.5
CLIP$_{336}$		47.2	90.8	48.3	33.6	**49.2**	33.5	51.1	49.2
DALIP$_{224}$		49.2	91.0	52.8	34.5	43.7	**34.3**	51.3	50.3
DALIP$_{336}$		**50.1**	**91.6**	**53.5**	**35.4**	44.6	34.2	**51.9**	**51.0**
Backbone of ConvNEXT-base									
CLIP$_{\text{ConvNEXT}}$	PlantMix-13M	46.5	89.6	45.3	30.1	**47.8**	32.6	49.1	47.8
DALIP$_{\text{ConvNEXT}}$		**48.2**	**90.2**	**47.0**	**32.6**	47.0	**33.7**	**50.1**	**49.2**
Effect of Data Mixing									
DALIP	P: 10M + G: 0M	18.6	**93.0**	**53.7**	**36.4**	40.0	**36.9**	**52.0**	35.3
	P: 10M + G: 1M	43.1	91.8	53.0	34.5	**45.6**	34.6	51.9	47.5
	P: 10M + G: 2M	47.8	91.6	53.2	34.8	44.0	34.4	51.6	49.7
	P: 10M + G: 3M	49.2	91.0	52.8	34.5	43.7	34.3	51.3	**50.3**
	P: 10M + G: 4M	**49.3**	90.1	48.6	30.8	39.3	31.7	48.1	48.7

Table S3. Ablation Study of PlantMix and DALIP

λ_1	0.0	0.2	0.3	0.4	0.5	0.6	0.8	1.0
Acc (%)	44.8	47.5	48.8	49.3	48.6	48.1	46.8	42.5

Table S4. Results of various λ on Plant (Mean). $\lambda_1 + \lambda_2 = 1$.

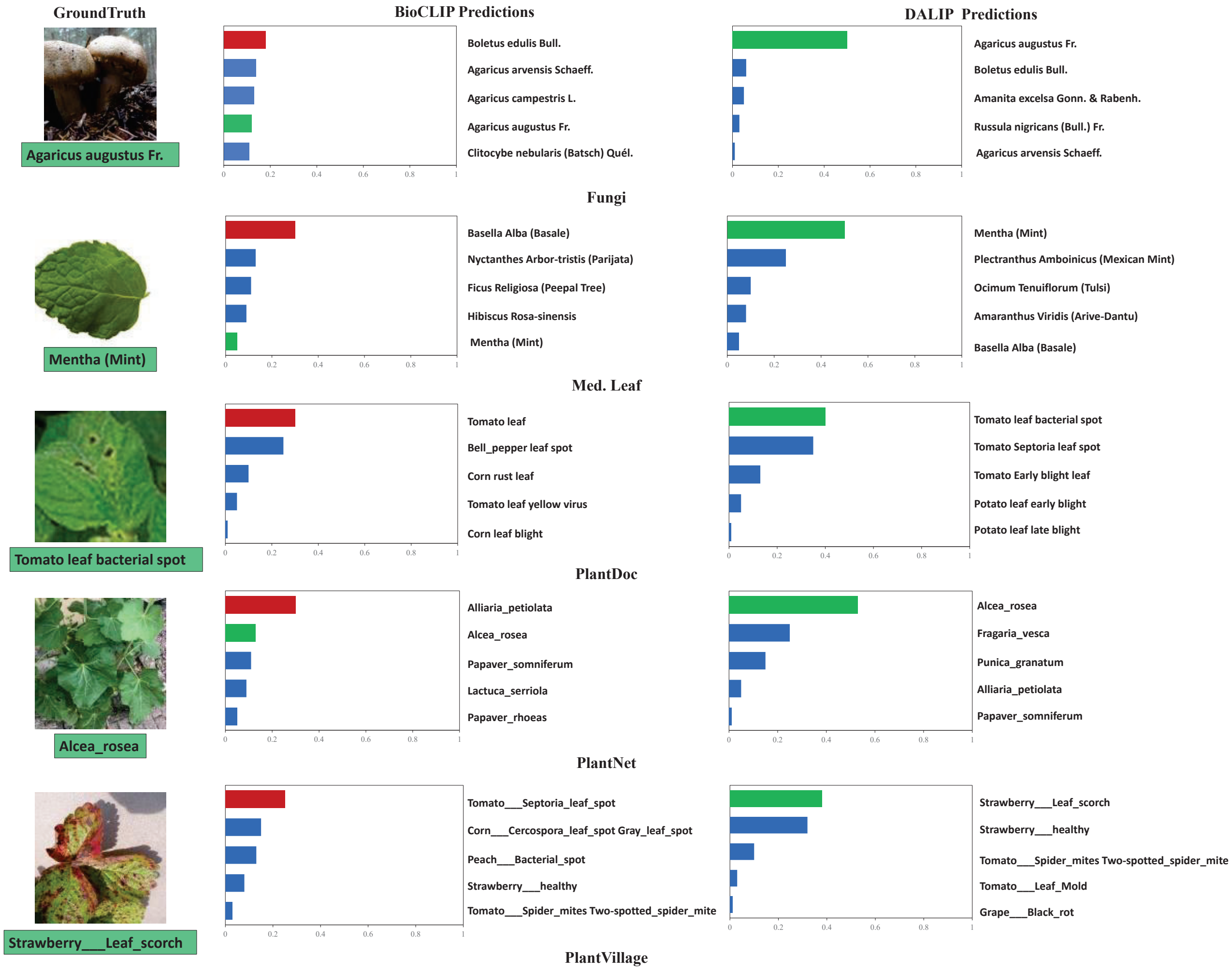

Figure S3. Example predictions for DALIP and BioCLIP on Fungi, Med. Leaf, PlantDoc, PlantNet and PlantVillage. Ground truth labels are green; incorrect predictions are red. Left: Correct DALIP predictions. Center, Right: Images that BioCLIP incorrectly labels, but DALIP correctly labels.

References

[1] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024. 1

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 1

[3] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1