

# LATTECLIP: Unsupervised CLIP Fine-Tuning via LMM-Synthetic Texts

## Supplementary Material

Anh-Quan Cao<sup>2\*</sup> Maximilian Jaritz<sup>1</sup> Matthieu Guillaumin<sup>1</sup> Raoul de Charette<sup>2</sup> Loris Bazzani<sup>1</sup>  
<sup>1</sup>Amazon <sup>2</sup>Inria

In this supplementary material, we first describe the limitation of LATTECLIP in Sec. 1. We then detail the implementation of LATTECLIP and the baselines in Sec. 2, followed by presenting additional ablation studies in Sec. 3. Lastly, we include further qualitative results in Sec. 4.

### 1. Limitations

Despite promising results, LATTECLIP considers a limited number of description types. Expanding description generation to include more contextual levels, such as scenes, objects, and attributes, would provide richer contextual information. Additionally, our performance is constrained by the underlying LMM model, and improvements could be made with better models in the future. Representing a class by a single prototype may limit our ability to capture intra-class variance; exploring multiple prototypes per class could be beneficial. Lastly, it is unclear why the method improves on some datasets but not others. Understanding this discrepancy could lead to better methods.

### 2. Implementation details

We implement LATTECLIP based on the standard fine-tuning pipeline of OpenCLIP [7] using the ViT-B/32 model. The hyperparameters used are the default ones provided in OpenCLIP [7], except for batch size and learning rate. We use a batch size of 512 and a learning rate of  $1e-7$  for the datasets Caltech101 [3], DTD [2], Eurosat [5], FGVC [11], Oxford Pets [13], Cars [8], Flower102 [12], and UCF101 [14]. For the datasets Food101 [1] and SUN397 [15], we use a learning rate of  $1e-6$ . LATTECLIP is trained for  $\min\{2000 \text{ iterations}, 50 \text{ epochs}\}$ .

For FLYP [4], we reimplement it based on its official implementation<sup>1</sup> and OpenCLIP [7], as its idea is intuitive and simple: fine-tuning using contrastive loss with class templates instead of cross-entropy loss. We use the same OpenCLIP-based model and training hyperparameters as LATTECLIP. The pseudo-labels are recalculated after every

weight update, following [9].

For ReCLIP [6], we use the official implementation<sup>2</sup>, but substitute OpenCLIP as the base CLIP model to ensure a fair comparison across all methods. While ReCLIP is designed for transductive learning (train/test on test set), as shown in the paper and by its official implementation, we adapt it to our experimental setup. Specifically, we retrain and evaluate ReCLIP using identical dataset splits as LATTECLIP.

### 3. Additional ablations

**Incorrect images in generating  $T^{\text{group}}$ .** Tab. 1 presents the results across all datasets when varying the number of correct images, which are selected using ground-truth labels, in groups of 4 images used for generating *group-descriptions*. Increasing the number of correct images generally leads to improvements in most datasets. However, the average performance gap remains small, demonstrating the robustness of our method to the presence of incorrect images in the group. This robustness is further evidenced by the performance of LATTECLIP, which remains competitive even when relying on pseudo-labels for image selection instead of ground-truth labels.

**Number of images per group.** Tab. 2 analyzes the performance as the number of images per group used for generating *group-description* increases. Generally, more images per group lead to higher performance on most datasets. This is intuitive, as more images provide richer information and a higher chance of including correct images. Using only two images results in the worst performance because selecting a wrong image would significantly impact the outcome, making 50% or 100% of the selected images incorrect. Consequently, larger groups are more robust to the inclusion of wrong images. As LLAVA [10] has a fixed resolution, adding more images results in lower resolution per image. This could explain the performance plateau on datasets with more image details, such as UCF101 or SUN397.

\*The main work was done while interning at Amazon.

<sup>1</sup><https://github.com/locuslab/FLYP>

<sup>2</sup>[https://github.com/michiganleon/ReCLIP\\_WACV](https://github.com/michiganleon/ReCLIP_WACV)

label type	#correct	Avg.	EuroSAT	Sun397	Food101	Flower102	DTD	FGVC	Oxford Pets	Cars	UCF101	Caltech101
Pseudo (ours)	N/A	72.23	80.27	<b>70.68</b>	<u>79.63</u>	71.94	56.26	<b>22.02</b>	89.21	87.40	<u>70.08</u>	94.77
Ground-truth	1	72.48	80.02	69.19	79.04	<u>72.88</u>	<b>61.11</b>	20.55	<b>89.62</b>	87.35	<b>70.16</b>	94.89
	2	72.61	<b>81.28</b>	69.73	79.13	72.55	<u>60.82</u>	20.76	<u>89.51</u>	<u>87.53</u>	69.65	<u>95.13</u>
	3	<b>72.72</b>	<u>80.81</u>	70.28	<b>79.80</b>	72.72	60.28	<u>21.87</u>	89.48	87.29	69.52	<b>95.17</b>
	4	<u>72.64</u>	80.40	<b>70.54</b>	78.79	<b>72.96</b>	60.17	21.42	<b>89.62</b>	<b>87.96</b>	70.00	94.56

Table 1. Impact of varying the number of correctly chosen images based on ground-truth labels when using 4 images for *group-description* generation. Our approach yields comparable performance despite relying solely on pseudo-labels for image selection.

#Images	Average	EuroSAT	Sun397	Food101	Flower102	DTD	FGVC	Oxford Pets	Cars	UCF101	Caltech101
2	71.55	<b>80.74</b>	69.36	76.03	71.24	56.03	21.12	<u>89.29</u>	87.32	69.88	<u>94.52</u>
4	72.23	80.27	<b>70.68</b>	<b>79.63</b>	71.94	56.26	<u>22.02</u>	89.21	87.40	<b>70.08</b>	<b>94.77</b>
8	<u>72.31</u>	79.90	69.90	<u>79.55</u>	<u>73.04</u>	<u>57.69</u>	22.00	89.18	<u>87.65</u>	69.71	94.44
16	<b>72.49</b>	<u>80.67</u>	<u>70.18</u>	78.24	<b>73.20</b>	<b>58.64</b>	<b>22.28</b>	<b>89.53</b>	<b>87.85</b>	<u>70.05</u>	94.28

Table 2. Impact of increasing the number of images per group for generating group-descriptions. Overall, more images lead to higher performance due to richer information and increased robustness against the inclusion of incorrect images. However, the performance plateaus on some datasets, such as UCF101 or SUN397, could be due to the fixed resolution of LLAVA, resulting in lower resolution per image as the number of images increases.

## 4. Additional results

### Examples of LMM-synthetic texts and pseudo-labels.

Fig. 1 illustrates examples of *image-description*  $T^{\text{image}}$  and *group-description*  $T^{\text{group}}$  generated from individual images  $x$  and image groups  $x^{\text{group}}$ , respectively. The figure also presents ground-truth labels (GT) along with pseudo-labels derived from the frozen CLIP model ( $c_{zs}$ ) and the fine-tuning model ( $c_{ft}$ ). Note that the class-description is generated by substituting the pseudo-label  $c \in \{c_{zs}, c_{ft}\}$  into a pre-defined template: "a photo of a [c].". Combining both types of pseudo-labels increases the chance of capturing the ground-truth label, as each type of pseudo-label is correct for different examples. For instance,  $c_{zs}$  is correct for rows 2, 3, and 4, while  $c_{ft}$  is correct for rows 1 and 4. Regarding the synthetic description,  $T^{\text{group}}$  provides richer contextual information, particularly in rows 1, 2, 4, and 5, and contains less hallucinated information compared to  $T^{\text{image}}$ , as seen in rows 2 and 3, with greater accuracy in rows 1, 4, and 5.

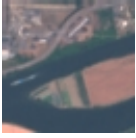


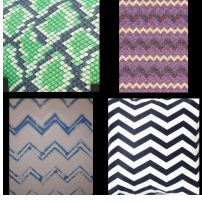

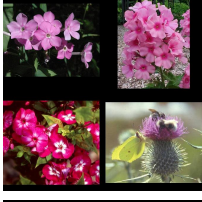

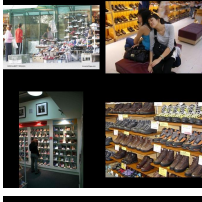

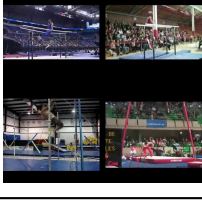
$x$	$T^{\text{image}}$	$x^{\text{group}}$	$T^{\text{group}}$	pseudo-labels	GT
	Buildings and green spaces.		Green, brown, and blue colors, indicating vegetation, soil, and water.	$c_{zs}$ : permanent crop land, $c_{ft}$ : river	river (Eurosat)
	The texture in the photo is a wooden floor with a herringbone pattern.		Zigzag patterns, geometric shapes, and vibrant colors.	$c_{zs}$ : zigzagged, $c_{ft}$ : grooved	zigzagged (DTD)
	The pink primrose flower in the photo is a beautiful and vibrant display of nature's beauty.		Purple and yellow petals, green stems, multiple layers of petals.	$c_{zs}$ : pink primrose, $c_{ft}$ : silverbush	garden phlox (Flower102)
	Woman in white shirt holding blue shoe.		Shoes, women, shopping, retail, store, display, merchandise, fashion, sales, shopping center, mall, department store, commercial, consumer.	$c_{zs}$ : shoe shop, $c_{ft}$ : shoe shop	shoe shop (SUN397)
	Person on trampoline.		Gymnastics, acrobatics, high jumps, flips, and aerial stunts.	$c_{zs}$ : uneven bars, $c_{ft}$ : parallel bars	parallel bars (UCF101)

Figure 1. Examples of *image-description*  $T^{\text{image}}$  generated from image  $x$  and *group-description*  $T^{\text{group}}$  generated from image group  $x^{\text{group}}$ , and two types of pseudo-labels: zero-shot  $c_{zs}$  and fine-tuning  $c_{ft}$ . Note that the class-description is generated by substituting the pseudo-label  $c \in \{c_{zs}, c_{ft}\}$  into a predefined template: "a photo of a [c].".

## References

- [1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101—mining discriminative components with random forests. In *ECCV*, 2014. 1
- [2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 1
- [3] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *CVPRW*, 2004. 1
- [4] Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *CVPR*, 2023. 1
- [5] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *JSTAR*, 2019. 1
- [6] Xuefeng Hu, Ke Zhang, Lu Xia, Albert Chen, Jiajia Luo, Yuyin Sun, Ken Wang, Nan Qiao, Xiao Zeng, Min Sun, et al. Reclip: Refine contrastive language image pre-training with source free domain adaptation. In *WACV*, 2024. 1
- [7] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, jul 2021. 1
- [8] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In

*ICCVW*, 2013. 1

- [9] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICMLW*, 2013. 1
- [10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1
- [11] Subhansu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv*, 2013. 1
- [12] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian conference on computer vision, graphics & image processing*, 2008. 1
- [13] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 1
- [14] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv*, 2012. 1
- [15] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 1