LONGPROLIP: A PROBABILISTIC VISION-LANGUAGE MODEL WITH LONG CONTEXT TEXT

Sanghyuk Chun **Sangdoo Yun**
NAVER AI Lab

ABSTRACT

Recently, Probabilistic Language-Image Pre-Training (ProLIP) has been proposed to tackle the multiplicity issue of vision-language (VL) tasks. Despite their success in probabilistic representation learning at a scale, the ProLIP models cannot handle long context texts longer than 64 context length, which limits their ability to capture rich contextual information from longer text sequences. To address this issue, this paper proposes a fine-tuning strategy for ProLIP to accept longer texts, *e.g.*, 256 text tokens. Experimental results on Urban-1k and the DataComp evaluation suite show that the proposed LongProLIP recipe can improve understanding of long contexts while minimizing the negative effect of fine-tuning. We also observe a trade-off between the long context understanding (measured by Urban-1k) and general zero-shot capability (measured by ImageNet or the average of 38 zero-shot evaluation datasets by DataComp).

1 INTRODUCTION

Probabilistic vision-language models (PrVLMs) (Chun et al., 2021; Ji et al., 2023; Upadhyay et al., 2023; Chun, 2024; Chun et al., 2025) aim to tackle the multiplicity problem of VL tasks, *e.g.*, an image can be described by multiple captions and vice versa. Recently, Chun et al. (2025) proposed ProLIP, the first PrVLM pre-trained on billion-scale image-text pairs (trained on DataComp 1B (Gadre et al., 2024) with 12.8B training seen samples). As shown in Fig. 1, ProLIP estimates both the mean and variance vectors of a Gaussian probabilistic embedding using the `[CLS]` token and the [UNC] token, respectively. By using the uncertainty estimated by the extracted variance, Chun et al. (2025) showed the advantages of uncertainty estimation in VL tasks. For example, ProLIP shows that shorter texts are more uncertain and more general inputs including specific ones which can help the understanding of the given dataset. Also, ProLIP can improve the prompt selection for zero-shot classification tasks and image traversal tasks with uncertainty.

Despite its success on capturing the ambiguity of VL tasks with short captions, ProLIP has a limitation at its text context length; the original ProLIP text encoder only takes at most 64 text tokens. However, in practice, text contexts often exceed 64 tokens. For example, as shown by LongCLIP (Zhang et al., 2024), a text prompt for a text-to-image generation task can be longer than 64 tokens to describe very fine-grained details. As another example, the captions generated by large vision-language models, such as LLaVA (Liu et al., 2024), usually have very long text lengths (*e.g.*, larger than 100 tokens). For practical usability, we argue that 64 token length might not be sufficient.

In this paper, we propose **LongProLIP**, an extension of ProLIP by taking longer text context length, (*i.e.*, 256). Our approach is based on LongCLIP (Zhang et al., 2024), but we provide a more detailed study of the training dataset. Also, we observe that directly applying the LongCLIP recipe to ProLIP often leads to a significant performance drop (*e.g.*, ImageNet zero-shot accuracy becomes 3.7% from 67.8%), especially when the base pre-trained model is not sufficiently strong. More specifically, we observe a trade-off between the long context understanding (measured by Urban-1k) and general zero-shot capability (measured by ImageNet or the average of 38 zero-shot evaluation datasets by DataComp). When we focus on the long context understanding, our LongProLIP model achieved the state-of-the-art performance on the Urban-1k dataset, proposed to measure the long context understanding ability of VLM (Zhang et al., 2024); while LongCLIP ViT-L/14 achieves 82.7% I2T R@1 and 86.1% T2I R@1, our LongProLIP ViT-B/16 achieves 90.8% and 91.8%, respectively. The LongProLIP weights will be available at the HuggingFace hub upon acceptance.

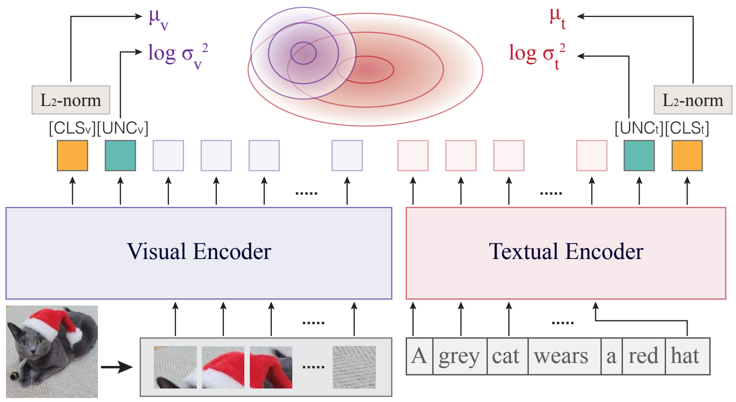

Figure 1: **Overview of ProLIP.** ProLIP consists of an image encoder and a text encoder. The mean and variance are estimated by `[CLS]` token and `[UNC]` token, respectively. Note that the original CLIP text encoder does not use `[CLS]` token, but the ProLIP text encoder uses the last additional two tokens for `[CLS]` and `[UNC]` tokens. ProLIP is trained with probabilistic objective functions, such as probabilistic pairwise contrastive loss, inclusion loss, and variational information bottleneck loss.

2 Results

Pre-trained models. We use the official pre-trained weights provided by Chun et al. (2025), including two ProLIP ViT-B/16 models trained with 12.8B seen samples and 1.28B seen samples, trained on DataComp 1B. To compare LongProLIP and LongCLIP, we compare the ProLIP model trained with 1.28B seen samples to the CLIP model trained on the same number of seen samples.

Fine-tuning dataset. Following LongCLIP (Zhang et al., 2024), we fine-tune the models on the ShareGPT4V dataset (Chen et al., 2024) which contains 1.2M image-text pairs with highly descriptive captions. As shown in Appendix A, these captions contain very long text tokens, such as more than 150. However, our experiments show that using only ShareGPT4V for fine-tuning often leads to a severe performance drop in general zero-shot classification, such as ImageNet. To tackle the issue, we employ a high-quality filtered dataset, the filtered DataComp medium with HYPE (Kim et al., 2024) and Data Filtering Network (DFN) (Fang et al., 2024).

Optimization and hyperparameters. We fine-tune the models with the learning rate of 1e-06, the weight decay of 0.01, and the batch size of 8,192. The other optimization hyperparameters are the same as Chun et al. (2025). Also, following LongCLIP (Zhang et al., 2024), we interpolate the 66-dimensional positional embedding (64 context length + 1 `[CLS]` token + 1 `[UNC]` token), while the first 20 tokens (preserving the pre-trained textual knowledge) and the last 2 tokens (preserving the information of `[CLS]` and `[UNC]` tokens) are fixed during the interpolation.

Evaluation dataset. We mainly compare the models on the Urban-1k dataset proposed by Zhang et al. (2024), which measures the long context text understanding ability. We additionall measure the general zero-shot capability by ImageNet (Russakovsky et al., 2015), ECCC Caption (Chun et al., 2022) (an extended version of COCO Caption (Chen et al., 2015)), and 38 zero-shot evaluation suite by DataComp (Gadre et al., 2024). We list the full dataset of 38 tasks in Appendix B.

Comparison with LongCLIP. Table 1 shows the comparison of LongCLIP and LongProLIP in Urban-1k and ImageNet, where each measures the long context text understanding ability and the general zero-shot classification ability. For a fair comparison, we use both CLIP and ProLIP models pre-trained from DataComp 1B with 1.28B training seen samples.

In the table, we observe two main findings. First, ProLIP and LongProLIP show a better long context text understanding compared to CLIP and LongCLIP. We presume that this is due to uncertainty-aware modeling, which can capture "specificity" or "general" concepts in the given texts. Second, while LongCLIP fine-tuned on ShareGPT4V does not show a significant performance drop in ImageNet zero-shot classification (IN ZSC), ProLIP fine-tuned solely on ShareGPT4V shows a severe

Table 1: **Comparison of LongCLIP and LongProLIP.** We use CLIP and ProLIP models trained on DataComp 1B with 1.28B seen samples. Unlike CLIP, directly applying the LongCLIP recipe to ProLIP leads to a severe performance drop to zero-shot classification (ZSC), such as ImageNet ZSC.

Model	Fine-tuning datasets	Urban-1k I2T	Urban-1k T2I	Urban-1k (Avg)	ImageNet
CLIP	-	41.9	44.8	43.4	67.2
	ShareGPT4V only	83.1 **(+41.2)**	80.1 **(+35.3)**	81.6 **(+38.2)**	63.3 **(-3.9)**
ProLIP	-	56.2	54.7	55.5	67.8
	ShareGPT4V only	84.9 **(+28.7)**	86.6 **(+31.9)**	85.8 **(+30.3)**	3.7 **(-64.1)**
	ShareGPT4V + HYPE + DFN	55.7 **(-0.5)**	62.9 **(+8.2)**	59.3 **(+3.8)**	66.9 **(-0.9)**

Table 2: **LongProLIP results by different fine-tuning datasets.** N denotes the number of the seen samples during fine-tuning. Note that Urban-1k Avg results of LongCLIP are 79.2 for ViT-B/16 and 84.4 for ViT-L/14, respectively. We additionally report ECCV Caption mAP@R and the average score of the 38 tasks of the DataComp evaluation suite. The full results are shown in Table C.1 and Table C.2

Fine-tuning datasets	N	Urban-1k (Avg)	EC mAP (Avg)	ImageNet	38 Tasks (Avg)
-	-	65.4	34.1	74.6	63.3
ShareGPT4V only	24M	87.2 **(+21.8)**	35.7 **(+1.6)**	70.8 **(-3.8)**	60.5 **(-2.8)**
ShareGPT4V only	128M	91.3 **(+25.9)**	33.4 **(-0.7)**	69.5 **(-5.1)**	58.7 **(-4.6)**
ShareGPT4V + HYPE + DFN	128M	77.5 **(+12.1)**	34.6 **(+0.5)**	74.6 **(+0.0)**	63.3 **(+0)**

performance drop in the same setting (67.8 $\rightarrow$ 3.7). Interestingly, as we will show in the latter experiments, if we use a stronger ProLIP backbone, the performance drop is not as significant as Table 1. We presume that this is because the uncertainty modeling of the ProLIP with 1.28B seen sample model is not sufficient, therefore, during the fine-tuning with probabilistic objectives, the learned probabilistic space is largely changed.

We additionally fine-tune ProLIP with high-quality short image-text pairs, *i.e.*, the combination of HYPE medium (Kim et al., 2024) and DFN medium (Fang et al., 2024). In the table, we observe that while the ImageNet performance drop is neglectable (-0.9%), the average long context understanding ability is improved after the fine-tuning (+3.8%). From this observation, we propose to use the LongProLIP model fine-tuned with ShareGPT4V + HYPE + DFN for a general purpose, and use the LongProLIP model fine-tuned solely with ShareGPT4V for a long context-specific purpose.

Main results. In Table 2, we show the LongProLIP results with a stronger ProLIP backbone, *i.e.*, ViT-B/16 with 12.8B seen samples. We observe similar results to Table 1, except the performance drop is not as significant as Table 1. We observe that when we fine-tune ProLIP solely on ShareGPT4V with more iterations (*i.e.*, 24M seen samples vs. 128M seen samples), the Urban-1k result is improved but the other performances are all dropped. On the other hand, when we use ShareGPT4V and HYPE + DFN for fine-tuning, the overall performance has been improved, including retrieval and zero-shot capability. Specifically, in the full results in Table C.1 and Table C.2, we can observe that the retrieval performances (*e.g.*, COCO Caption (Chen et al., 2015), WinoGAViL (Bitton et al., 2022), and ECCV Caption (Chun et al., 2022)) are largely improved with the LongProLIP fine-tuning. While ShareGPT4V only fine-tuning often leads to severe performance drops, the mixed fine-tuning strategy seems to prevent the performance drops (*e.g.*, the KITTI task (Geiger et al., 2012) shows -16.32% and -10.69% drops for 24M and 128M ShareGPT4V fine-tuning, but the mixed strategy shows +3.17% performance improvement compared to the original performance).

Conclusion and future work. This paper proposes LongProLIP, an extension of ProLIP with a longer text context (64 $\rightarrow$ 256). We observe that fine-tuning solely with the ShareGPT4V dataset can lead to a significant performance drop for general zero-shot tasks. The future work would include analyses of why ProLIP is more sensitive than CLIP in a long text fine-tuning setting, and improving our understanding of VL tasks with long captions (*e.g.*, text-to-image generation or highly descriptive captioning) using the newly obtained long context model.

References

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iwildcam 2021 competition dataset. *arXiv preprint arXiv:2105.03494*, 2021.

Yonatan Bitton, Nitzan Bitton Guetta, Ron Yosef, Yuval Elovici, Mohit Bansal, Gabriel Stanovsky, and Roy Schwartz. Winogavil: Gamified association benchmark to challenge vision-and-language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:26549–26564, 2022.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision (ECCV*, pp. 446–461. Springer, 2014.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision (ECCV)*, pp. 370–387. Springer, 2024.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6172–6180, 2018.

Sanghyuk Chun. Improved probabilistic image-text representations. In *International Conference on Learning Representations (ICLR)*, 2024.

Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Sanghyuk Chun, Wonjae Kim, Song Park, Minsuk Chang Chang, and Seong Joon Oh. ECCV Caption: Correcting false negatives by collecting machine-and-human-verified image-caption associations for MS-COCO. In *European Conference on Computer Vision (ECCV)*, 2022.

Sanghyuk Chun, Wonjae Kim, Song Park, and Sangdoo Yun. Probabilistic language-image pre-training. In *International Conference on Learning Representations (ICLR)*, 2025.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)*, 88:303–338, 2010.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. In *International Conference on Learning Representations (ICLR)*, 2024.

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. In *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, 2004.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3354–3361. IEEE, 2012.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *International Conference on Computer Vision (ICCV)*, pp. 8340–8349, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15262–15271, 2021b.

Yatai Ji, Junjie Wang, Yuan Gong, Lin Zhang, Yanru Zhu, Hongfa Wang, Jiaxing Zhang, Tetsuya Sakai, and Yujiu Yang. Map: Multimodal uncertainty-aware vision-language pre-training model. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23262–23271, 2023.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2901–2910, 2017.

Wonjae Kim, Sanghyuk Chun, Taekyung Kim, Dongyoon Han, and Sangdoo Yun. Hyperbolic entailment filtering for underspecified images and texts. In *European Conference on Computer Vision (ECCV)*, 2024.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024.

S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 4. Granada, 2011.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763. PMLR, 2021.

Vikram V Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. Geode: a geographically diverse evaluation dataset for object recognition. *Advances in Neural Information Processing Systems*, 36, 2024.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/recht19a.html`.

William A Gaviria Rojas, Sudnya Diamos, Keertan Ranjan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32:323–332, 2012.

Uddeshya Upadhyay, Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. Probvlm: Probabilistic adapter for frozen vison-language models. In *International Conference on Computer Vision (ICCV)*, pp. 1899–1910, 2023.

Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pp. 210–218. Springer, 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 (NeurIPS)*, pp. 10506–10518, 2019.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3485–3492. IEEE, 2010.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Association for Computational Linguistics (ACL)*, 2:67–78, 2014.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision (ECCV)*, pp. 310–325. Springer, 2024.

APPENDIX

A EXAMPLES OF SHAREGPT4V CAPTION

Here, we provide examples of ShareGPT4V captions (Chen et al., 2024). These captions are highly descriptive and have very long context (*e.g.*, more than 180 tokens).

- This image captures a serene moment on a sandy beach. At the center of the frame, a golden retriever is sitting comfortably inside a heart drawn in the sand. The dog's pink tongue is sticking out, adding a playful touch to the scene. The heart, drawn with a stick or similar object, is large and occupies most of the foreground of the photo. The sandy texture of the beach contrasts with the smooth lines of the heart. The beach itself is empty, creating a sense of tranquility. In the background, you can see the shoreline and the water, which adds depth to the image. Further in the distance, there's a line of trees and houses, providing a hint of civilization. The sky above is overcast, casting a soft light over the entire scene and enhancing the peaceful atmosphere. Despite the absence of bright sunlight, the image exudes warmth, largely due to the presence of the golden retriever at its heart. Overall, this image beautifully combines elements of nature with a symbol of love and companionship.

- In the image, a young man is seen comfortably reclining on a vibrant purple couch. He is wearing a stylish black jacket and a pair of glasses that reflect the light from the two laptops resting on his chest. The laptops, both white in color, are adorned with an assortment of colorful stickers, adding a personal touch to the devices. The man appears to be deeply engrossed in his work or perhaps enjoying some digital entertainment. The couch is positioned on a gray carpeted floor which extends out into the room, providing a soft contrast to the bold color of the couch. The beige wall in the background adds a neutral backdrop to the scene, allowing the focus to remain on the man and his activities. The image captures a moment of modern life, where technology is an integral part of our daily routines. It's a snapshot of digital age comfort and productivity.

- This image captures a dynamic scene on a snowy mountain. At the center of the frame, a skier dressed in a black jacket and pants is in action. The skier, equipped with a helmet and goggles for safety, skillfully maneuvers two ski poles to aid in their descent. The skier is captured mid-turn, their body leaning into the curve as they carve a path down the slope. This action kicks up a spray of snow, creating a dramatic effect against the serene backdrop. The location is a mountain blanketed in snow, its surface untouched except for the trail left by the skier. The mountain's incline suggests a steep descent, adding to the thrill of the scene. In the background, a forest of trees stands tall. Their branches are laden with snow, painting a picturesque winter landscape. The trees appear dense and extend far into the distance, providing a stark contrast to the open space being navigated by the skier. Overall, this image encapsulates the exhilarating sport of skiing against the tranquil beauty of a snowy mountain landscape.

B DATACOMP EVALUATION SUITE DETAILS

The DataComp evaluate suite contains 38 zero-shot tasks: ImageNet (Russakovsky et al., 2015), 6 ImageNet distribution shifts robust benchmarks, including ImageNet-A, ImageNet-O (Hendrycks et al., 2021b), ImageNet-R (Hendrycks et al., 2021a), ImageNet v2 (Recht et al., 2019), ImageNet-Sketch (Wang et al., 2019) and ObjectNet (Barbu et al., 2019), 13 VTAB task (Zhai et al., 2019), including Caltech-101 (Fei-Fei et al., 2004), CIFAR-100 (Krizhevsky et al., 2009), CLEVR Counts, CLEVR Distance (Johnson et al., 2017), Describable Textures (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), KITTI Vehicle Distance (Geiger et al., 2012), Oxford Flowers-102 (Nilsback & Zisserman, 2008), Oxford-IIIT Pet (Parkhi et al., 2012), PatchCamelyon (Veeling et al., 2018), RESISC45 (Cheng et al., 2017), SVHN (Netzer et al., 2011) and SUN397 (Xiao et al., 2010), and 3 retrieval tasks, including Flickr (Young et al., 2014), MS-COCO Caption (Chen et al., 2015) and WinoGAViL (Bitton et al., 2022). There are also 13 additional tasks, such as CIFAR-10, Country211 (Radford et al., 2021), FGVC Aircraft (Maji et al., 2013), Food-101 (Bossard et al., 2014), GTSRB (Stallkamp et al., 2012), MNIST (LeCun et al., 1998), Pascal VOC (Everingham et al., 2010), Rendered SST2 (Radford et al., 2021), STL-10 (Coates et al., 2011), iWildCam (Beery et al., 2021), FMoW (Christie et al., 2018), Dollar Street (Rojas et al., 2022), and GeoDE (Ramaswamy et al., 2024).

Table C.1: **Zero-shot classification full results.** IN dist. denotes the average of 6 ImageNet distribution shift robust benchmarks and VTAB contains 13 zero-shot tasks. Retrieval is the average of three retrieval tasks. "S24M" denotes that the model is fined-tuned on ShareGPT4V with 24M seen samples; S128M is defined similarly (128M seen samples). SHD128M denotes that the model is fine-tuned on ShareGPT4V, HYPE, and DFN with 128M seen samples. Each four row corresponds to each row in Table 2.

	IN dist.	VTAB	Retrieval	Urban I2T R@1	Urban T2I R@1	EC mAP@R I2T	EC mAP@R T2I
ProLIP (pre-trained)	63.0	63.7	59.6	67.5	63.3	28.9	39.2
ProLIP (S24M)	59.1 **(–3.90)**	60.9 **(–2.80)**	61.3 **(+1.70)**	86.9 **(+19.4)**	87.4 **(+24.1)**	30.5 **(+1.60)**	41.0 **(+1.80)**
ProLIP (S128M)	58.1 **(–4.90)**	58.4 **(–5.30)**	59.6 **(+0.00)**	90.8 **(+23.3)**	91.8 **(+28.5)**	29.6 **(+0.70)**	37.2 **(–2.00)**
ProLIP (SHD128M)	62.5 **(–0.50)**	63.0 **(–0.70)**	61.9 **(+2.30)**	75.8 **(+8.3)**	79.1 **(+15.8)**	29.2 **(+0.30)**	40.1 **(+0.90)**

Table C.2: **Zero-shot classification full results.** The detail is same as Table C.1.

	ImageNet 1k	CIFAR-10	CIFAR-100	CLEVR Counts	CLEVR Distance	Country211	Describable Textures	EuroSAT	FGVC Aircraft	Food-101
ProLIP (pre-trained)	74.56	96.42	83.25	29.78	15.13	21.29	66.91	60.89	38.02	91.03
LongProLIP (S24M)	70.76 **(–3.80)**	95.64 **(–0.78)**	81.17 **(–2.08)**	32.18 **(+2.40)**	15.79 **(+0.66)**	18.89 **(–2.40)**	65.37 **(–1.54)**	65.24 **(+4.35)**	34.39 **(–3.63)**	87.71 **(–3.32)**
LongProLIP (S128M)	69.54 **(–5.02)**	91.82 **(–4.60)**	75.10 **(–8.15)**	27.46 **(–2.32)**	15.80 **(+0.67)**	16.95 **(–4.34)**	63.94 **(–2.97)**	53.96 **(–6.93)**	33.69 **(–4.33)**	86.14 **(–4.89)**
LongProLIP (SHD128M)	74.58 **(+0.02)**	96.47 **(+0.05)**	82.94 **(–0.31)**	31.70 **(+1.92)**	15.80 **(+0.67)**	21.29 **(+0.00)**	66.33 **(–0.58)**	57.94 **(–2.95)**	37.57 **(–0.45)**	90.74 **(–0.29)**

	GTSRB	Caltech-101	ImageNet Sketch	ImageNet v2	ImageNet-A	ImageNet-O	ImageNet-R	KITTI Vehicle Distance	MNIST	ObjectNet
ProLIP (pre-trained)	52.79	93.61	63.65	66.66	50.25	45.40	86.00	32.21	84.47	65.80
LongProLIP (S24M)	53.62 **(+0.83)**	92.73 **(–0.88)**	61.12 **(–2.53)**	62.61 **(–4.05)**	44.19 **(–6.06)**	45.75 **(+0.35)**	84.02 **(–1.98)**	15.89 **(–16.32)**	81.60 **(–2.87)**	56.77 **(–9.03)**
LongProLIP (S128M)	48.97 **(–3.82)**	92.88 **(–0.73)**	60.30 **(–3.35)**	61.31 **(–5.35)**	41.91 **(–8.34)**	46.15 **(+0.75)**	83.17 **(–2.83)**	21.52 **(–10.69)**	82.34 **(–2.13)**	55.72 **(–10.08)**
LongProLIP (SHD128M)	51.20 **(–1.59)**	93.52 **(–0.09)**	63.61 **(–0.04)**	66.39 **(–0.27)**	49.71 **(–0.54)**	44.20 **(–1.20)**	85.84 **(–0.16)**	32.63 **(+0.42)**	87.64 **(+3.17)**	65.39 **(–0.41)**

	Oxford Flowers-102	Oxford-IIIT Pet	Pascal VOC 2007	PatchCamelyon	Rendered SST2	RESISC45	Stanford Cars	STL-10	SUN397	SVHN
ProLIP (pre-trained)	78.37	93.45	81.74	61.41	54.31	68.27	91.33	97.91	71.35	72.77
LongProLIP (S24M)	74.30 **(–4.07)**	90.59 **(–2.86)**	77.83 **(–3.91)**	58.08 **(–3.33)**	50.85 **(–3.46)**	62.90 **(–5.37)**	88.02 **(–3.31)**	97.56 **(–0.35)**	70.81 **(–0.54)**	67.17 **(–5.60)**
LongProLIP (S128M)	73.34 **(–5.03)**	89.55 **(–3.90)**	70.11 **(–11.63)**	55.59 **(–5.82)**	52.66 **(–1.65)**	59.86 **(–8.41)**	87.46 **(–3.87)**	97.04 **(–0.87)**	66.08 **(–5.27)**	63.53 **(–9.24)**
LongProLIP (SHD128M)	78.16 **(–0.21)**	92.91 **(–0.54)**	81.55 **(–0.19)**	57.96 **(–3.45)**	54.86 **(+0.55)**	68.48 **(+0.21)**	90.95 **(–0.38)**	98.11 **(+0.20)**	71.33 **(–0.02)**	69.38 **(–3.39)**

	iWildCam	Camelyon17	FMoW	Flickr	MSCOCO	WinoGAViL	Dollar Street	GeoDE	Average
ProLIP (pre-trained)	12.64	57.85	15.12	79.97	53.18	45.56	62.27	90.31	63.31
LongProLIP (S24M)	13.69 **(+1.05)**	50.65 **(–7.20)**	0.00 **(–15.12)**	79.44 **(–0.53)**	55.88 **(+2.70)**	48.61 **(+3.05)**	57.24 **(–5.03)**	89.02 **(–1.29)**	60.48 **(–2.83)**
LongProLIP (S128M)	12.37 **(–0.27)**	51.63 **(–6.22)**	0.00 **(–15.12)**	78.20 **(–1.77)**	55.12 **(+1.94)**	45.59 **(+0.03)**	56.19 **(–6.08)**	86.47 **(–3.84)**	58.67 **(–4.64)**
LongProLIP (SHD128M)	13.54 **(+0.90)**	60.27 **(+2.42)**	14.54 **(–0.58)**	80.13 **(+0.16)**	53.54 **(+0.36)**	51.98 **(+6.42)**	63.20 **(+0.93)**	90.61 **(+0.30)**	63.34 **(+0.03)**

C FULL RESULTS

Table C.1 shows the additional information from Table 2. We can observe that ShareGPT4V fine-tuning for ProLIP may harm the zero-shot ability (*e.g.*, ImageNet, ImageNet robustness benchmark and VTAB), but it can be beneficial to retrieval tasks, such as Flickr, MS-COCO Caption, Wino-GAViL, Urban-1k and ECCV Caption.

Table C.2 shows the full results of the DataComp evaluation suite (38 tasks). Similar to Table C.1, LongProLIP fine-tuning specifically improves retrieval performances.