Unbiased Region-Language Alignment for Open-Vocabulary Dense Prediction

Supplementary Material

We provide an overview of the supplementary materials to ensure a clear and comprehensive understanding.

- In Sec. 1, we detail the limitations and broader impact of DenseVLM.
- In Sec. 2, we present the training details.
- In Sec. 3, we offer supplementary experiments on input image sizes, region proposals, backbones, category sets and threshold of region denoising for training VLMs.
- In Sec. 4, we show visualizations, including confusion matrices and croppping feayures predictions, to demonstrate foreground bias. .
- In Sec. 5, we present the dataset information for training and evaluation.

1. Limitations and broader impact

Limitations: Our aim is to develop a region-language alignment model that effectively integrates local visual and semantic features, thereby improving open-vocabulary dense prediction performance. Compared to previous pre-trained Vision-Language Models (VLMs) [3, 13, 14, 17], our proposed DenseVLM achieves superior results and significantly improves downstream task performance. We believe DenseVLM has even greater potential. 1) Scalability. DenseVLM is designed within an efficient, unsupervised region-language alignment framework, making it adaptable to various datasets. However, computational resource limitations have restricted our ability to scale to larger datasets. 2) Model capacity. We employ the ViT-L/14 model from CLIPSelf [14] as a powerful Pre-trained VLM (P-VLM). Utilizing more robust VLMs can yield better performance, and transferring their rich semantic knowledge to training models is a promising direction. 3) Fine-grained Semantic. We categorize objects into broad *thing* and *stuff* classes. Fine-grained semantic segmentation and decoupled alignment would enhance the model's ability to distinguish between similar categories. We plan to explore these avenues in our future research.

Broader impact: DenseVLM exhibits notable potential for open-vocabulary dense predictions within scenes, which can enhance various applications such as robotics and environmental monitoring. By enabling systems to recognize and interpret a wide range of objects and contexts without prior training on specific categories, DenseVLM facilitates more adaptive and versatile applications. Given its broad applicability and non-specialized nature, our method is designed to support a variety of technical advancements without directly addressing specific societal challenges.

[1] https://developer.nvidia.com/automatic-mixed-precision

item	value
image size	512×512
optimizer	AdamW [12]
learning rate	0.0001
β_1	0.9
β_2	0.98
weight decay	0.1
batch size (per card)	48
warmup steps [6]	1000
epochs	6
learning rate scheduler	cosine decay [11]
number of GPUs	4
automatic mixed precision[1]	True

Table S1. Training details of DenseVLM.

2. Training details

We train all models on NVIDIA A40 GPUs to ensure a fair comparison across experiments. The detailed configuration is provided in Table S1. For the SA-1B dataset [9], we use $8\times$A40 GPUs to ensure efficient and scalable training.

For open-vocabulary segmentation, we train the models such as SAN [15] and CAT-Seg [4] on the COCO-Stuff [1] dataset for $80k$ iterations. For open-vocabulary detection, models are trained for 3 epochs on the OV-COCO [2] benchmark and 48 epochs on OV-LVIS [8] benchmark.

3. Additional experiments

Ablation study on input image sizes. To evaluate the effect of input image size on DenseVLM, we conduct experiments with distinct resolutions: 224, 320, 512, 768 and 1024 pixels, for both training and inference. As shown in Tab. S2, model's performance on the region classification task improves as image resolution increases from 224 to 1024 pixels. This enhancement can be attributed to the greater detail captured at higher resolutions. However, this improvement comes a significant increase in GPU memory usage. Considering the trade-off between computational resources and model performance, we resize the images to 512×512 pixels to achieve an optimal balance.

Ablation study on using region proposals. Following RegionCLIP [17] for fine-tuning VLMs with pseudo-labelled region-text pairs, we compare our approach to CLIPSelf [14] in utilizing these pairs. As shown in Tab. S3, CLIPSelf substitutes random image crops with pseudo

Input Image Size	GPU Memory (per card)	COCO						ADE20K					
		Boxes		**Masks-T**		**Masks-S**		**Boxes**		**Masks-T**		**Masks-S**	
		Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
224	9G	60.1	79.9	49.4	62.4	35.3	64.2	40.0	70.0	36.3	56.4	50.3	77.0
320	11G	66.2	85.4	59.2	73.0	41.0	71.2	45.6	76.0	44.0	67.6	54.3	81.7
512	16G	73.4	90.5	71.0	84.8	45.6	77.8	51.3	82.2	52.1	78.0	57.8	85.5
768	27G	74.4	91.3	75.4	90.1	45.5	79.0	52.7	82.9	55.4	82.6	58.2	86.6
1024	39G	**76.6**	**93.1**	**78.7**	**93.6**	**46.5**	**79.8**	**53.2**	**83.6**	**56.8**	**83.2**	**58.6**	**86.8**

Table S2. Ablation study on input image sizes. We report the Top1 and Top5 mean accuracy on classifying boxes and panoptic masks on COCO panoptic and ADE20K panoptic benchmarks. The GPU memory usage corresponds to a batch size of 12 on A40 GPU.

Method	Region Proposals	COCO						ADE20K					
		Boxes		**Masks-T**		**Masks-S**		**Boxes**		**Masks-T**		**Masks-S**	
		Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
CLIPSelf	✗	69.1	88.2	66.7	83.0	41.7	75.2	48.1	77.7	47.5	74.2	53.7	82.8
CLIPSelf	✓	70.2	89.2	68.1	83.5	35.7	71.8	49.8	79.7	51.5	76.0	50.9	80.7
DenseVLM	✗	73.4	90.5	71.0	84.8	45.6	77.8	51.3	82.2	52.1	78.0	57.8	85.5
DenseVLM	✓	74.4	91.3	75.4	90.1	45.9	79.0	52.7	82.9	55.4	82.6	58.2	86.6

Table S3. Ablation study on using region proposals. We report the Top1 and Top5 mean accuracy on classifying boxes and panoptic masks (thing and stuff) on COCO panoptic and ADE20K panoptic benchmarks.

θ	**Boxes**		**Masks-T**		**Masks-S**	
	Top1	Top5	Top1	Top5	Top1	Top5
0.0	72.1	89.6	68.2	84.3	43.6	76.1
0.1	72.7	90.2	69.1	84.3	44.6	77.2
0.2	73.1	90.4	69.7	84.6	45.1	77.7
0.3	**73.4**	**90.5**	**71.0**	**84.8**	**45.6**	**77.8**
0.4	73.2	90.2	70.2	84.3	45.2	77.5
0.5	73.1	90.0	70.0	84.3	45.0	77.1
0.6	73.1	89.9	69.6	84.0	44.6	76.3

Table S4. Ablation study on threshold of θ in region denoising.

region-text pairs, resulting in an enhanced recognition for foreground objects while concurrently observing a reduction in the accuracy of background identification. In contrast, our proposed DenseVLM achieves a notable improvement in the recognition accuracy of foreground objects while also improving the identification of background stuff.

Ablation study on the threshold θ. We perform an ablation experiment to assess the impact of varying threshold θ values of region denoising. As shown in Tab. S4, the model performs the worst when $\theta = 0$. When θ is set lower, the Top-5 accuracy increases, but results in suboptimal performance. This may be due to low-confidence categories causing alignment confusion for the model. Conversely, setting θ too high filters out too many local images, decreasing performance. By defaul, we select $\theta = 0.3$ for DenseVLM.

Ablation study on various backbones. DenseVLM exhibits adaptability to diverse backbones. As shown in Tab. S5, our models achieve consistent superiority over prior approaches [3, 14] across all dense prediction tasks. Particularly, the ViT-B/16-based DenseVLM performs comparably to the ViT-L/14-based CLIPSelf [14]. Utilizing ViT-L/14 with a large number of parameters as initialization, DenseVLM achieves clearly enhancements across all evaluated metrics, thereby facilitating superior performance in dense prediction tasks.

Ablation study on different category sets. To assess the impact of varying category sets, we conduct experiments using four various sets: 133 (80), 171 (80), 273 (160), and 316 (160), categorized into foreground and background classes. The set of 133 categories exclusively comprises COCO Panoptic [10] class set, while the 171-category set consists solely of COCO-Stuff [1] class set. The 273-category set integrates non-overlapping classes from both COCO-Stuff and the ADE20K Panoptic [18] dataset, which contains 150 categories. The 316-category set encompasses selected background classes from COCO-Stuff and the ADE20K dataset, which includes 847 categories. Our code accurately reflects all specified category sets.

As shown in Tab. S6, with the increase in the number of categories, the performance of our model progressively improves when evaluated on the COCO Panoptic and ADE20K Panoptic benchmarks. This is because the larger category sets provide a richer representation of objects and stuff, enabling the model to capture more fine-grained information, thereby enhancing its overall performance.

backbones	VLMs	COCO						ADE20K					
		Boxes		**Masks-T**		**Masks-S**		**Boxes**		**Masks-T**		**Masks-S**	
		Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
ViT-B/16	OpenCLIP	49.8	74.3	51.9	72.2	29.2	54.9	28.4	54.1	29.6	53.4	37.9	66.6
ViT-B/16	CLIPSelf	67.6	87.8	64.4	81.2	44.5	77.1	43.4	76.0	44.0	71.1	50.7	82.1
ViT-B/16	DenseVLM*	71.9	90.2	70.0	84.3	47.8	79.4	48.5	79.2	49.0	75.2	55.1	85.2
ViT-B/16	DenseVLM	73.4	90.5	71.0	84.8	45.6	77.8	51.3	82.2	52.1	78.0	57.8	85.5
ViT-L/14	OpenCLIP	21.2	45.3	26.6	48.9	11.2	27.2	48.1	11.9	34.1	13.9	11.1	32.4
ViT-L/14	CLIPSelf	68.3	90.1	67.1	84.5	37.7	71.3	47.1	77.5	47.7	74.4	48.9	82.3
ViT-L/14	DenseVLM*	76.2	92.9	73.3	87.3	47.4	79.1	54.0	84.1	54.2	79.9	57.8	85.9
ViT-L/14	DenseVLM	75.2	91.8	73.3	87.1	45.5	78.1	54.5	85.0	55.6	82.1	58.1	86.4

Table S5. Ablation study on various backbones. We report the Top1 and Top5 mean accuracy on classifying boxes and panoptic masks (thing and stuff) on COCO panoptic and ADE20K panoptic benchmarks. * indicates the model initialized by OpenCLIP [3].

Categories	COCO						ADE20K					
	Boxes		**Masks-T**		**Masks-S**		**Boxes**		**Masks-T**		**Masks-S**	
	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
133 (80)	71.1	88.5	68.7	83.0	44.7	75.2	49.4	79.0	48.5	74.2	54.7	82.8
171 (80)	72.3	89.8	69.4	85.8	44.2	76.0	49.8	79.7	48.9	75.1	55.1	82.4
273 (160)	72.3	89.9	70.1	84.4	44.9	76.4	51.0	81.8	49.3	76.5	57.0	84.0
316 (204)	**73.4**	**90.5**	**71.0**	**84.8**	**45.6**	**77.8**	**51.3**	**82.2**	**52.1**	**78.0**	**57.8**	**85.5**

Table S6. Ablation study on different category sets. We report the Top1 and Top5 mean accuracy on classifying boxes and panoptic masks (thing and stuff) on COCO panoptic and ADE20K panoptic benchmarks.

4. Visualizations

Confusion matrix. We compare the region classification results of our method against previous approaches through a confusion matrix visualization for panoptic masks (both thing and stuff categories) on the COCO Panoptic dataset. These confusion matrixes offer a systematic overview of region classification performance, illustrating the incidence of accurate and erroneous classifications, particularly facilitating a precise assessment of models' accuracy in differentiating between thing and stuff categories. As shown in Fig. S1, Fig. S2, and Fig. S3, prior methods, including EVA-CLIP [13], RegionCLIP [17], and CLIPSelf [14], often misclassify background regions as co-occurring foreground classes, such as incorrectly identifying *snow* as *skis* or *sky* as *kite*. In contrast, as demonstrated in Fig. S4, our DenseVLM achieves higher accuracy in recognizing each category, with a notable improvement in the precision of background object identification.

Image grid patches classification. We visualize the classification results of image grid patches using the powerful ViT-L/14 model from CLIPSelf [14]. As shown in Fig. S5, the model focuses heavily on foreground object recognition, but significant portions of background patches are misclassified as foreground objects. The training VLMs are prone to learning these errors. Furthermore, regions with incorrect classifications often have low confidence scores, highlighting the importance of filtering them out.

5. Datasets of training and evaluation

COCO: COCO [10] is a large-scale panoptic segmentation dataset encompassing 80 *Thing* and 53 *Stuff* categories. The dataset comprises 118,000 images designated for the training set and 5,000 images for the validation set.

ADE20k: ADE20k [19] spans a broad spectrum of indoor and outdoor scenes, comprising 2,000 images for the validation set. This dataset includes 100 *Thing* and 50 *Stuff* categories. We evaluate open-vocabulary semantic annotations using both the extensive 847-category version (referred to as A-847) and the more frequently adopted 150-category version (referred to as A-150).

Pascal Context: Pascal-Context [5] constitutes an extensive dataset derived from Pascal-VOC 2010. We evaluate open-vocabulary semantic annotations using the complete set of 459 classes, referred to as PC-459.

OV-COCO: The open-vocabulary detection COCO (OV-COCO) benchmark, introduced in OV-RCNN [16], divides the 65 object categories in the COCO dataset into 48 base categories and 17 novel categories.

OV-LVIS: The open-vocabulary detection LVIS (OV-LVIS) benchmark, introduced in ViLD [7], defines the 337 rare categories from LVIS v1.0 dataset [8] as novel categories.

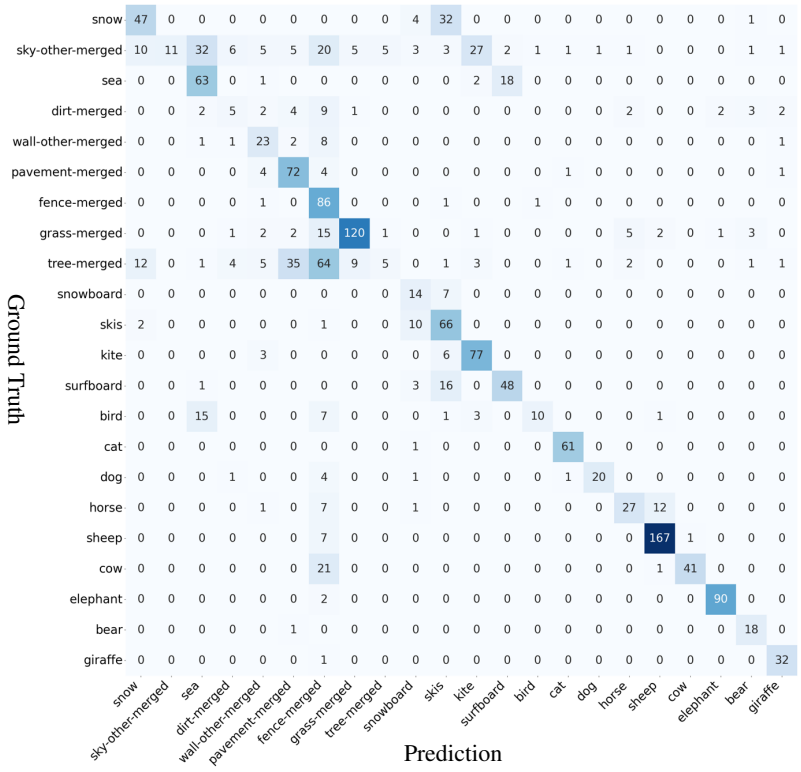

Figure S1. Confusion matrix visualization for region classification results of EVA-CLIP.

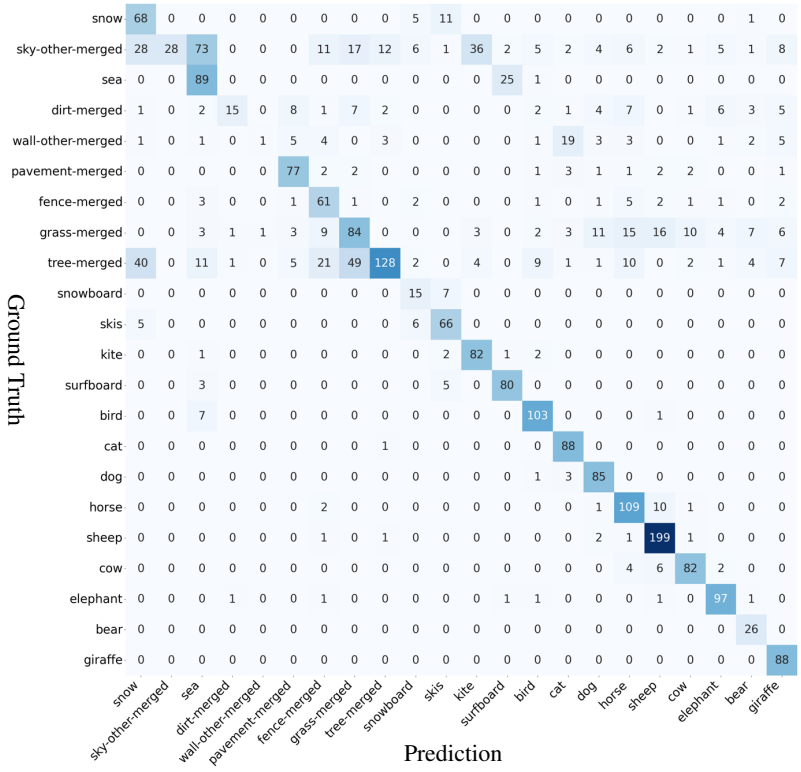

Figure S2. Confusion matrix visualization for region classification results of RegionCLIP.

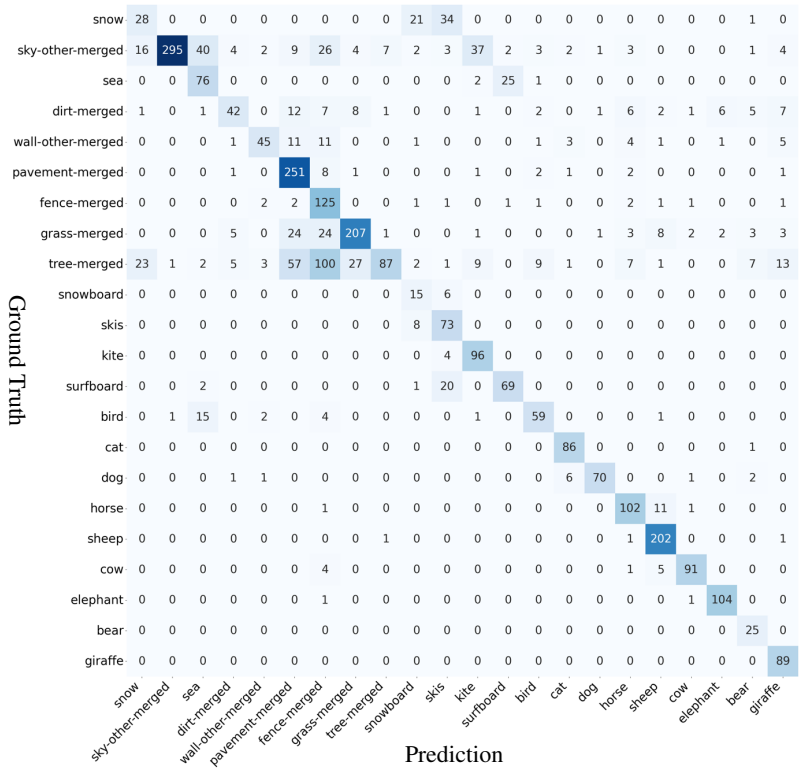

Figure S3. Confusion matrix visualization for region classification results of CLIPSelf.

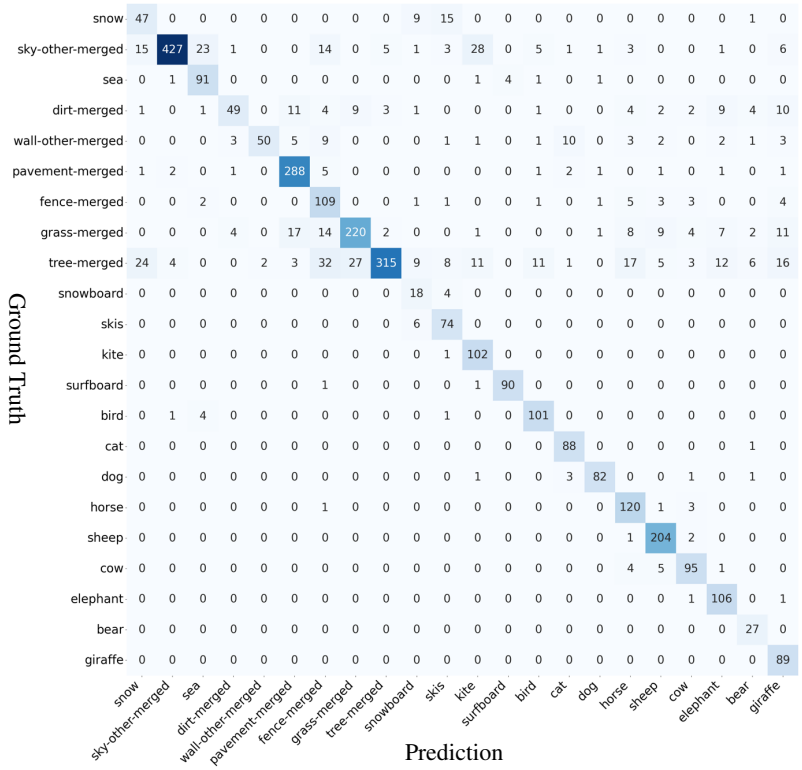

Figure S4. Confusion matrix visualization for region classification results of DenseVLM.

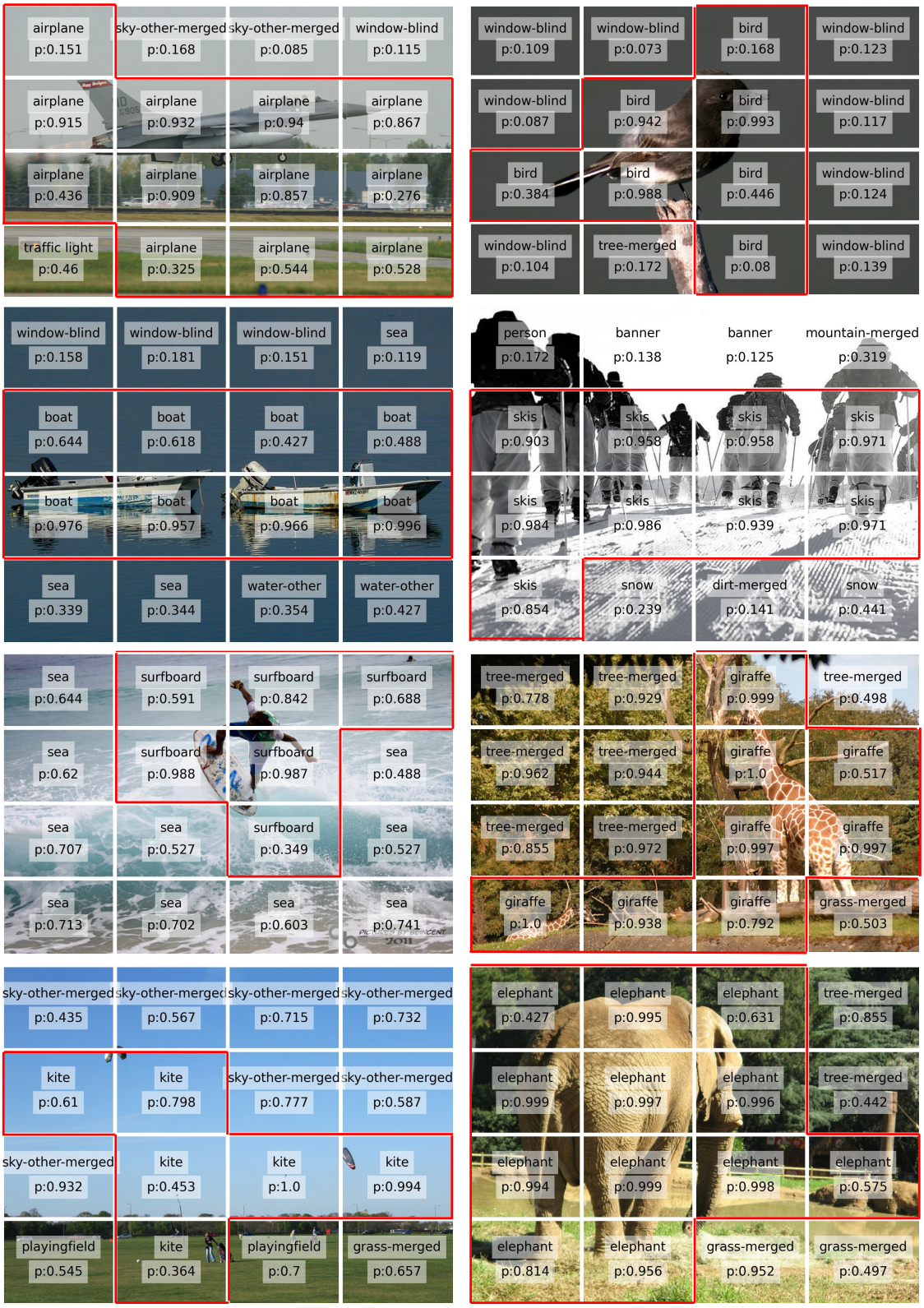

Figure S5. Visualization of image grid patches classification. The powerful ViT-L/14 model exhibits a pronounced focus on foreground object recognition, even when significant portions of background patches are misclassified as foreground objects.

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *CVPR*, pages 1209–1218, 2018. 1, 2

[2] 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. 1

[3] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, pages 2818–2829, 2023. 1, 2, 3

[4] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *CVPR*, 2024. 1

[5] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88:303–338, 2010. 3

[6] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2018. 1

[7] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, pages 1–20, 2022. 3

[8] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 1, 3

[9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 1

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2, 3

[11] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 1

[12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 1

[13] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 1, 3

[14] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. *ICLR*, 2024. 1, 2, 3

[15] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *CVPR*, pages 2945–2954, 2023. 1

[16] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, pages 14393–14402, 2021. 3

[17] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, pages 16793–16803, 2022. 1, 3

[18] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 2

[19] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 127: 302–321, 2019. 3