Learning to Name Classes for Vision and Language Models: Supplementary Materials

Sarah Parisot Yongxin Yang Steven McDonagh
Huawei Noah's Ark Lab
{sarah.parisot, yongxin.yang, steven.mcdonagh}@huawei.com

1. Additional LVIS experiments

We provide additional experiments on the LVIS dataset [S3] to further analyse the behaviour of our method. Firstly, we evaluate the impact of learning more than one word embedding per class; *i.e.* rather than using the query $t_i = [\text{a photo of a }] + [\text{pl}_1^i] + [.]$, we consider multiple placeholders, namely: $t_i = [\text{a photo of a }] + [\text{pl}_1^i] + \cdots + [\text{pl}_m^i] + [.]$, where m is the number of word embeddings to learn. We run experiments for $m = [0, 1, 2, 4, 6, 8]$, with $m = 0$ corresponding to the base model and report results in Figure S1. It can be observed that a large gain in performance is obtained when replacing original words ($m = 0$) with learnable ones ($m = 1$), with overall average precision remaining stable as m increases. Rare classes obtain the largest performance gains, while frequent classes have the smallest increase, as discussed in the main manuscript Sec. 4.2. We observe that rare class performance is more unstable as the number of embeddings increases, which can be attributed to the limited available training data, with increased parameter count increasing related overfitting risks. Based on these observed results, we can see that a single word embedding per class is sufficient to achieve good performance, and additional parameters yield no to very limited improvements.

Secondly, we provide results on the full LVIS validation set, alongside results from the detection specific prompt learning technique DETPRO [S1], in Table S1. We highlight that DETPRO uses VILD [S2] as a base model (pre-trained on base classes only), and is therefore not directly comparable with our results. We observe similar trends as seen on the mini-validation set, with our strategy achieving large performance gains compared to the original model (+2.9 AP ours-base, +4.9 AP Ours-all). We also observe largest gains for common and rare classes (+7.2 AP/c, AP/r, Ours-all). We note that, compared to the model trained on the full dataset, performance gains are more limited than those observed on the mini-validation set. We conjecture that this could be attributed to the larger number of rare classes, with very few training samples available, (the mini-validation set is not comprised of all classes). DETPRO gains on the rare classes additionally highlight the potential for prompt learning to work in conjunction with our approach, by improving performance in the open-vocabulary setting.

Finally, one additional advantage of our method is that we can achieve strong performance using only 10% of the LVIS training data, when learning class names. This notably allows us to boost performance on rare classes, which are typically penalised by the long tail distribution of the training dataset. As further analysis we provide, in Table S2, results using our balanced subset of 10% of the training data, and the entire, imbalanced, training dataset. We can see that overall performance is largely increased using 100% of the training data, and notably better than a model fine-tuned on the whole dataset in all categories (+1.3 AP). However, we observe that our model trained with a balanced subset achieves stronger performance on rare classes (+3.1 AP/r), confirming the supposed advantage of using a balanced dataset. Disentangling whether performance gains, on both frequent and common classes, are predominantly due to the additional data or the positive bias favouring these class groups is a promising avenue for further investigation.

2. Detailed classification results

In this section, we provide detailed, per dataset, results for our classification experiments. Results for model adaptation and open vocabulary experiments are reported in Table S3. In addition to baselines discussed in the main manuscript, we provide, for completeness, results using the CoCoOp method [S6]. As CoCoOp is significantly more computationally expensive, we only provide results reported in [S6] that match our experimental setting (in contrast, all CoOp experiments were reproduced locally using the official code and parameters). CoCoOp is an extension of the CoOp method that introduces an image feedback loop using a learnable so-called meta-network to condition prompt learning. We can see that CoCoOp struggles more at

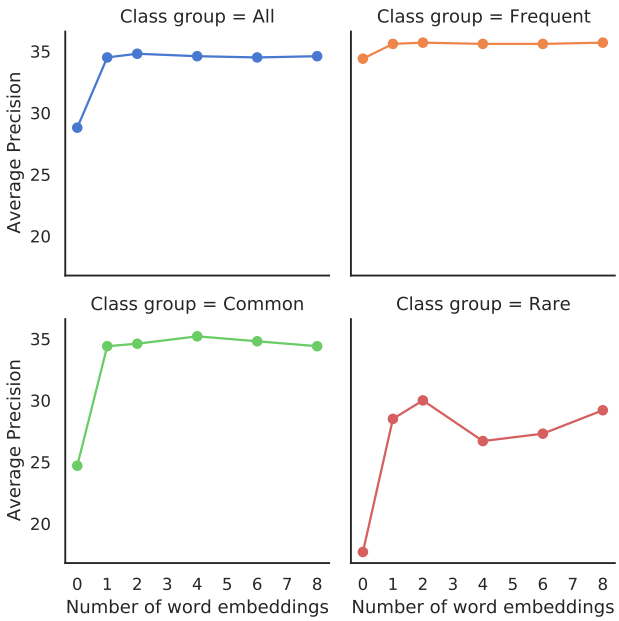


Figure S1. Influence of the number of word embeddings learned, per class, on the average precision for class groups: all classes (top-left); frequent classes (x>100, top-right); common classes (100≥x>10, bottom-left); and rare classes (10≥x≥1, bottom-right), where x pertains to available class samples in each case.

Method	AP	AP/f	AP/c	AP/r
VILD-base	27.5	31.9	27.5	17.4
DETPRO-base	28.4	32.4	27.8	20.8
OWL-vit-base	24.4	30.6	21.4	18.0
Ours-base	27.5	31.6	27.6	18.0
Prompt-all	26.1	31.2	24.5	18.7
Ours-all	29.3	32.0	**28.6**	**25.2**
OWL-vit-all	**30.5**	**35.0**	**28.7**	24.8

Table S1. Average precision detection results on the LVIS full validation set. '-Base' and '-all' indicates the model was trained on base and all classes respectively. 'Ours' and 'prompt' were trained on the OWL-vit-base model and 10% of the training data, OWL-vit-all is trained with 100%.

Method	AP	AP/f	AP/c	AP/r
OWL-vit-base	28.8	34.4	24.7	17.7
Ours-all 10%	34.5	35.6	34.4	**28.5**
Ours-all 100%	**35.8**	**38.5**	**34.8**	25.4
OWL-vit-all	34.5	**38.5**	33.2	19.1

Table S2. Average precision detection results on the LVIS mini-validation set: training class names with 10% of the data *vs*. 100%.

Task 1 (adapting to new datasets) that both CoOp and Ours (-2.1% average accuracy on base classes), achieves stronger open-vocabulary generalisation than CoOp (+7.6% average accuracy, new classes) but worse than ours and CLIP (-2.51%). Interestingly, reported average results on all classes outperform our method (+0.49% average accuracy on all classes), we note the small margin by which our method is outperformed, despite the fact that CoCoOP uses a much more complex and expensive image feedback mechanism. Incorporating a similar feedback loop has the potential to increase the performance of our approach as well.

Detailed results for sequential adaptation are found in Table S4. One notable result from sequential adaptation is that we achieve the most significant gains for datasets with technical class names such as Stanford cars, which can be attributed to the fact that class names are (all) learned independently (*i.e*. there is no mixture of learned and hand-crafted technical names).

3. CLIP engineered templates

We provide the list of manually engineered templates for CLIP zero-shot classification, (mentioned in our main paper, Sec. 4.1), in Table S5. We highlight that four datasets use our standard template ("a photo of a") and three datasets exhibit distinctly disparate sentence templates (namely Eurosat, DTD and UCF 101).

4. Interpretability

We further provide additional results with regards to our interpretability experiments (see main paper Sec. 4.3), for the CODA 2.0 dataset [S5]. In Figures S3 and S4, we illustrate how class names were modified for all 29 classes in the validation set, for the zero-shot and fine-tuned model (LVIS + SODA [S4]), respectively. On the zero-shot models, we can see that self-driving specific terms (*e.g*. pedestrian), map to more common terms with similar meaning (*e.g*. person). We highlight in particular the tricycle class, which is mapped to 'rickshaw', a more appropriate term for the visual content available in the dataset. We additionally note that class names were more substantially modified, as the highest observed self-similarity is 0.74 (motorcycle). In comparison, the fine-tuned model shows much higher class name stability, especially with regards to common classes (first row). This provides insight into what was learned during the fine-tuning process, and the similarity between class names of SODA and CODA 2.0 datasets (*e.g*. tricycle class).

We further note large differences in semantic meaning between the learned representations of the 'misc.' and 'machinery' classes, when comparing both models. This suggests that these classes do not comprise representative components and should be further separated, potentially via a clustering strategy. Finally, we point out how the new em-

Dataset	CLIP*	CoOp	CoCoOp†	Ours	Ours*	CLIP*	CoOp	CoCoOp†	Ours	Ours*	CLIP*	CoOp	CoCoOp†	Ours	Ours*
	Base classes					**New classes**					**All classes**				
Eurosat	56.4	**92.7**	87.9	92.1	91.3	**63.9**	52.7	60.04	59.8	**63.9**	47.7	55.2	-	51.7	**59.5**
Stanford Cars	63.3	77.3	70.49	**80.6**	**80.6**	**75.0**	61.9	73.49	**75.0**	**75.0**	65.3	**65.7**	-	63.1	63.1
Flowers 102	72.2	97.5	94.87	98.0	**98.5**	**77.9**	63.4	71.75	77.1	**77.9**	71.4	71.5	-	81.5	**82.6**
Oxford Pets	91.3	93.7	**95.20**	93.4	93.7	96.9	93.3	**97.69**	96.9	97	**89.1**	88.2	-	84.5	88.0
UCF 101	70.6	84.2	82.33	84.1	84.6	**77.4**	57.3	73.45	73.9	**77.4**	66.7	64.6	-	70.4	**71.4**
Aircraft	27.6	40.5	33.41	**44.0**	43.7	**36.1**	23.5	23.71	33.3	**36.2**	24.6	26.3	-	19.4	**27.7**
DTD	53.4	80.0	77.01	80.3	80.1	**60.3**	41.5	56.00	**60.0**	**60.0**	44.5	49.5	-	54.8	**57.6**
ImageNet	72.4	**76.4**	75.98	74.7	74.7	68.1	68.2	**70.43**	68.1	68.1	66.7	**68.9**	-	67.7	67.7
Caltech 101	97.0	98.1	97.96	97.9	97.9	**94.0**	88.8	93.81	**94.0**	**94.0**	93.0	91.3	-	**93.3**	**93.3**
Food 101	90.1	88.0	**90.70**	86.5	86.7	**91.3**	83.9	**91.29**	91.0	**91.3**	**86.1**	79.9	-	81.2	81.5
Sun 397	69.4	**80.6**	79.74	79.1	79.1	75.5	63.2	**76.86**	75.5	75.5	62.6	62.5	-	**64.3**	**64.3**
Average	69.4	82.6	80.47	82.7	82.8	**74.2**	64.1	71.69	73.1	**74.2**	65.2	65.7	**69.19**	66.5	68.7

Table S3. Detailed results for base to new classification accuracy. * Manually engineered prompt templates. †results copied from [S6].

Dataset	CLIP*	CoOp	Ours	Ours*	CLIP*	CoOp	Ours	Ours*
	New classes				**All classes**			
Eurosat	63.9	92.3	92.9	**93.4**	47.7	65.3	75.1	**76.5**
Stanford Cars	75.0	82.7	**90.8**	**90.8**	65.3	71.8	**80.9**	**80.9**
Flowers 102	77.9	97.0	**98.5**	98.4	71.4	81.3	**96.3**	**96.3**
Oxford Pets	96.9	**97.6**	97.4	97.3	**89.1**	88.5	86.2	87.5
UCF 101	77.4	87.2	**87.7**	87.5	66.7	74.6	**80.3**	80.0
Aircraft	36.1	53.6	**62.4**	**62.4**	24.6	32.1	30.8	**35.6**
DTD	60.3	75.6	**76.0**	75.9	44.5	54.9	**67.1**	66.8
ImageNet	68.1	**72.7**	71.7	71.7	66.7	**70.8**	69.3	69.3
Caltech 101	94.0	**95.9**	95.6	95.6	93.0	**94.2**	93.9	93.9
Food 101	91.3	**91.4**	89.6	89.7	**86.1**	84.9	82.1	81.6
Sun 397	75.5	82.1	**82.4**	**82.4**	62.6	71.0	**71.6**	**71.6**
Average	74.0	84.4	**85.9**	**85.9**	65.2	71.7	75.7	**76.3**

Table S4. Detailed classification accuracy for sequential training. * manually engineered prompt templates.

Dataset	Template
Eurosat	a centered satellite photo of [CLASS].
Stanford Cars	a photo of a [CLASS]. *(default)*
Flowers 102	a photo of a [CLASS], a type of flower.
Oxford Pets	a photo of a [CLASS], a type of pet.
UCF 101	a photo of a person doing [CLASS].
Aircraft	a photo of a [CLASS], a type of aircraft.
DTD	[CLASS] texture.
ImageNet	a photo of a [CLASS]. *(default)*
Caltech 101	a photo of a [CLASS]. *(default)*
Food 101	a photo of [CLASS], a type of food.
Sun 397	a photo of a [CLASS]. *(default)*

Table S5. List of engineered templates used for CLIP zero-shot classification, for each of the eleven datasets.

bedding for the traffic light class is poorly adapted to the class' original meaning. This can be attributed to the fact that the data available for this class only comprises mobile traffic lights, in addition to mislabelled samples (see Figure S2 for examples of training samples). As such, the visual content is highly different from learned mappings between the traffic light term and standard image content. This highlights that we can easily identify failure modes, potentially allowing, in cases of overfitting or poor training examples, to correct new words of poor quality (*e.g.* use the original word embeddings).

Figure S2. Examples of training samples in the traffic light category.

5. Visual examples

In Figure S5, we provide visual examples of improvements to the object detection task for the CODA 2.0 dataset. We compare our fine-tuned model (LVIS + SODA) to performance after learning class names, and provide the class agnostic ground truth as reference. We highlight how our model is able to recognise instances of the misc. class, identifying construction vehicles (*vs.* truck category, improving common class performance), and overall displays a stronger ability to identify corner cases.

6. Implementation

Implementation in MindSpore will be made available at https://gitee.com/mindspore/models/tree/master/research/cv/.

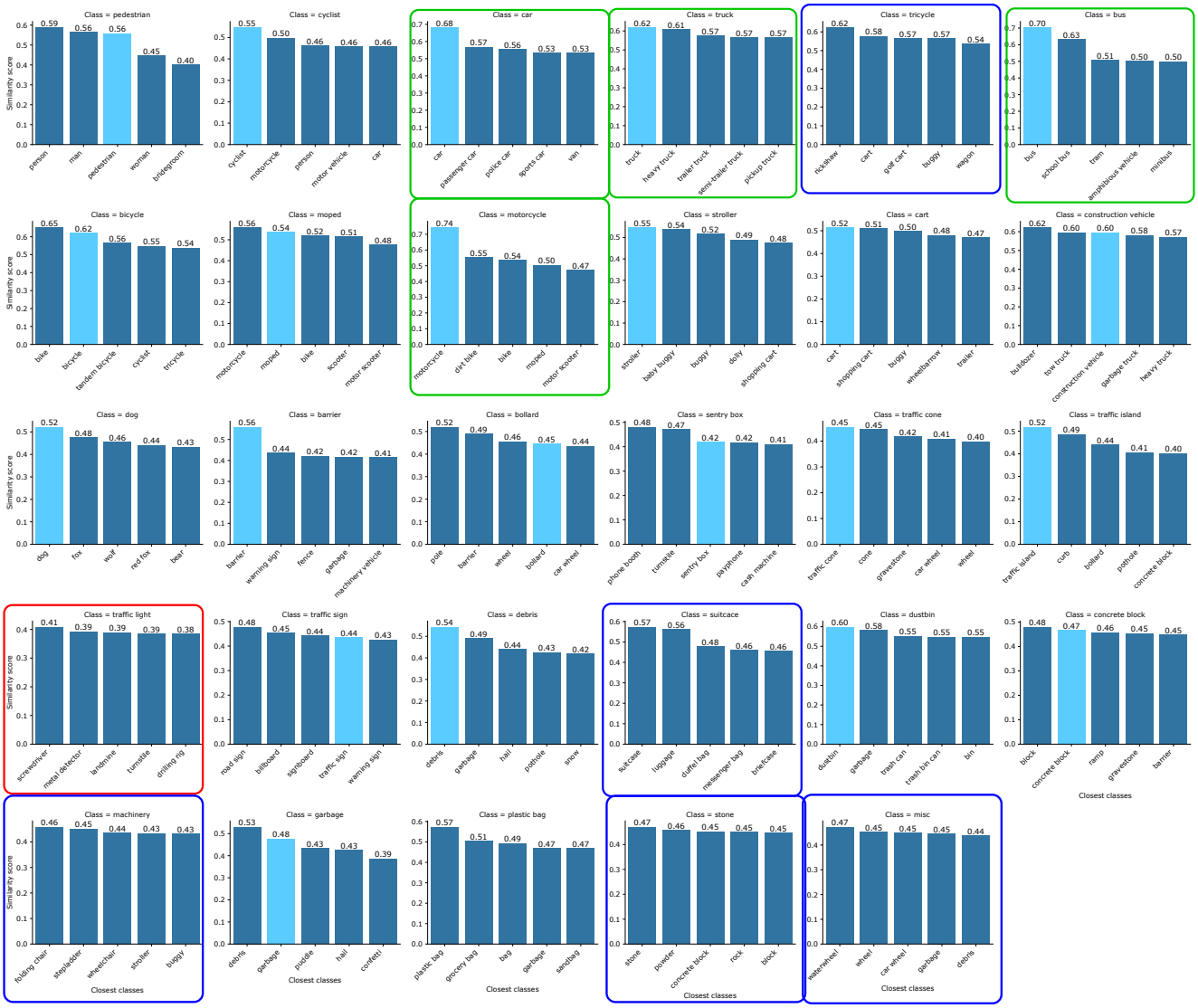


Figure S3. Interpretability results on the CODA 2.0 dataset using the **zero-shot base model** (see main paper, Sec. 4.3). Light blue bars: highlight the similarity between new word embeddings and original class name embeddings; highlighted classes (green boxes): strong similarity with original class name (>0.6, closest word), highlighted classes (blue boxes): original name not included within top 5 most similar classes, highlighted classes (red boxes): modified name semantic meaning markedly different from original.

References

[S1] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 1

[S2] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Zero-shot detection via vision and language knowledge distillation. *CoRR*, abs/2104.13921, 2021. 1

[S3] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1

[S4] Jianhua Han, Xiwen Liang, Hang Xu, Kai Chen, Lanqing Hong, Jiageng Mao, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Xiaodan Liang, et al. Soda10m: A large-scale 2d self/semi-supervised object detection dataset for autonomous driving. *arXiv preprint arXiv:2106.11118*, 2021. 2

[S5] Kaican Li, Kai Chen, Haoyu Wang, Lanqing Hong, Chaoqiang Ye, Jianhua Han, Yukuai Chen, Wei Zhang, Chunjing Xu, Dit-Yan Yeung, et al. Coda: A real-world road corner case dataset for object detection in autonomous driving. *arXiv preprint arXiv:2203.07724*, 2022. 2

[S6] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Com-*

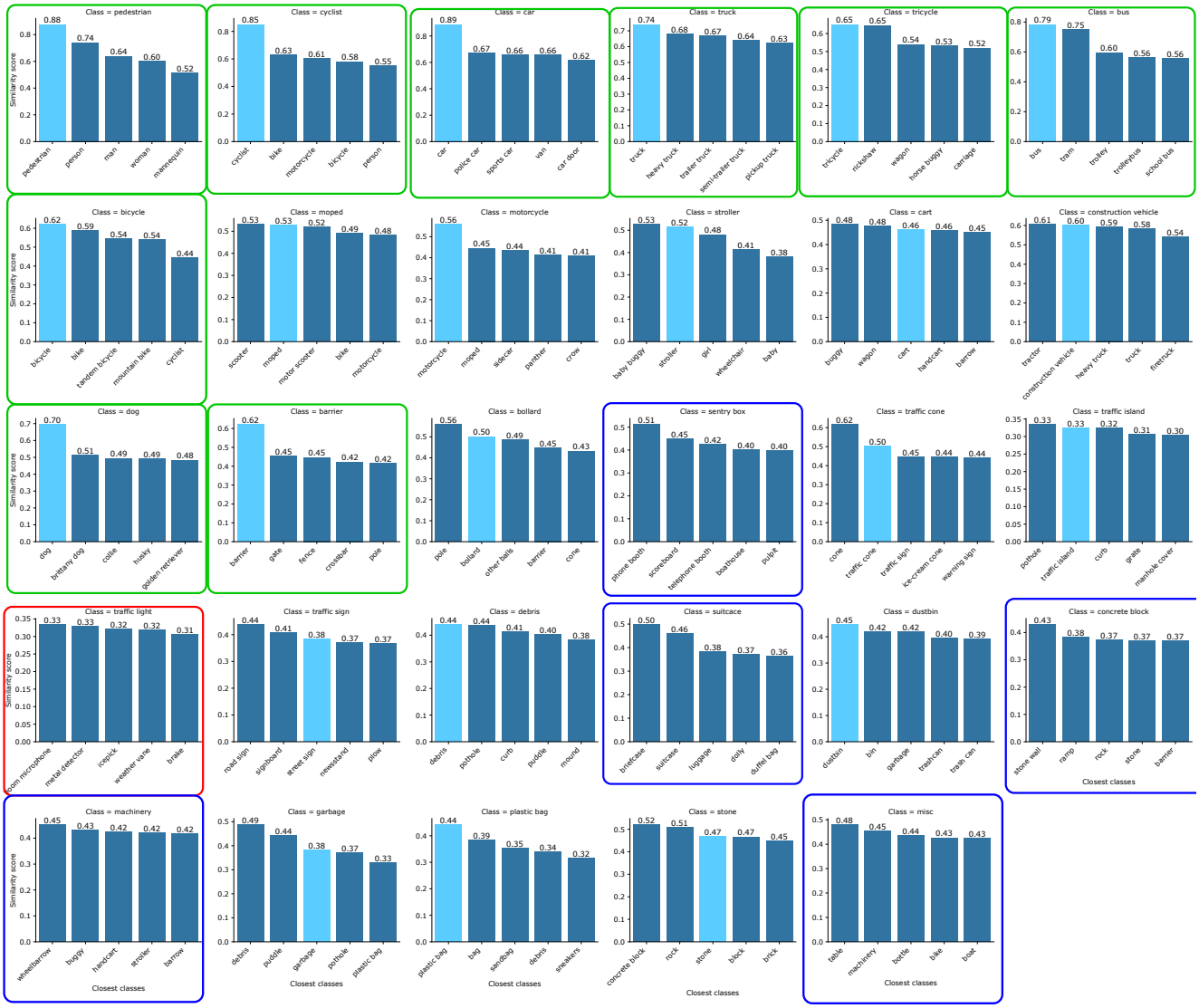
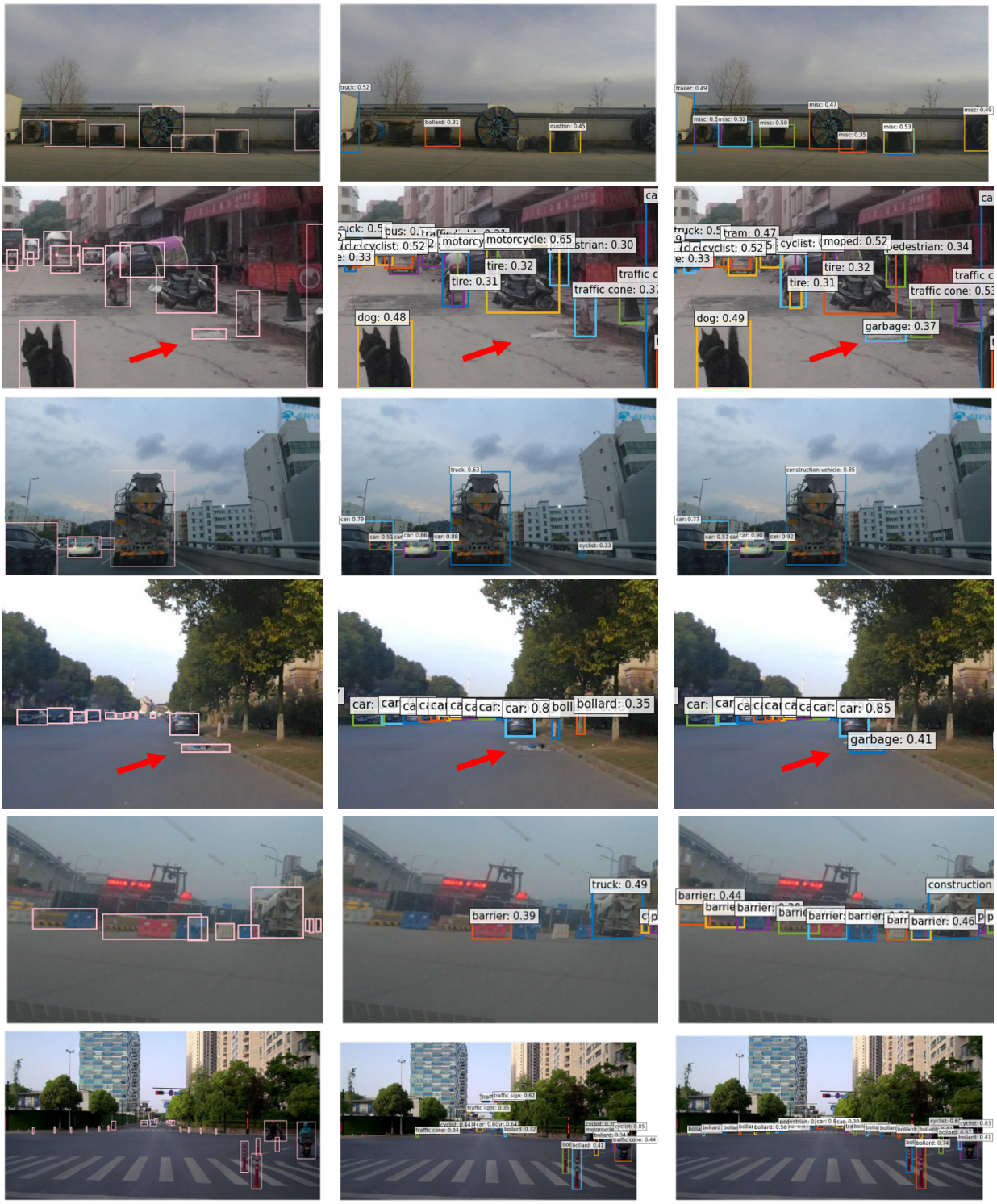


Figure S4. Interpretability results on the CODA 2.0 dataset on the **base model fine-tuned on LVIS and SODA** (see main paper, Sec. 4.3). Light blue bars: highlight the similarity between new word embeddings and original class name embeddings; highlighted classes (green boxes): strong similarity with original class name (> 0.6, closest word), highlighted classes (blue boxes): original name not included within top 5 most similar classes, (red boxes): modified name semantic meaning markedly different from original.

puter Vision and Pattern Recognition, pages 16816–16825,
2022. 1, 3



(a) Ground truth bounding boxes (b) OWL-vit model (c) Ours

Figure S5. Visual results from the CODA 2.0 dataset, comparing our approach to the base model, fine-tuned on LVIS + SODA. Our approach exhibits improved performance on corner-case classes.