Supplementary

A. Base / Novel Attribute Set in VAW

VAW [12] contains a large vocabulary of 620 attributes. In our experiments, considering that VAW attribute vocabulary has certain noise and semantic overlap, instead of taking all 'tail' attributes as the novel set, we sample half of the 'tail' attributes and 15% of the 'medium' attributes as the novel set ($\mathcal{A}_{\text{novel}}$, 79 attributes) and the remaining as the base ($\mathcal{A}_{\text{base}}$, 541 attributes). The novel attributes are: pulled back, smirking, muscular, holed, off white, littered, pepperoni, taupe, tucked in, bell shaped, multicolored, bronze, boiled, caucasian, silk, active, stormy, new, sprinkled, covered in sugar, side view, carried, overgrown, black metal, thatched, dotted, horned, shoeless, stucco, well dressed, barred, half filled, domed, vintage, hiding, gold framed, baked, reddish, rust colored, frizzy, nylon, scruffy, taking photo, opaque, violet, busy, foamy, relaxing, cubed, leaping, moss covered, chocolate, plastic, spreading arms, wispy, arch shaped, bent, bright green, black lettered, patchy, balancing, crocheted, furry, maroon, flat screen, classical, cloudless, partially visible, wearing scarf, orange, slender, eating, doorless, closed, shining, spotted, reflective, barren, wrapped.

B. Summary of Dataset Statistics

In our experiments, we take the standard MS-COCO 2017 [10] and VAW [12] for federated training, with the former for object category classification, and the latter for object attributes classification. In addition, we have harvested external image caption pairs on the COCO and VAW dictionaries from the CC 3M [18] and COCO Captions [4] for training CLIP-Attr. As for evaluation, we also include two additional benchmarks (LSA [13] and OVAD [3]) using official settings in their papers.

LSA [13]. A recent work by Pham *et al.* proposed the Large-Scale object Attribute dataset (LSA). LSA is constructed with all the images and their parsed objects and attributes of the Visual Genome (VG) [7], GQA [5], COCO-Attributes [11], Flickr30K-Entities [14], MS-COCO [10], and a portion of Localized Narratives (LNar) [15]. Here, we evaluate the effectiveness of our proposed method with the same settings proposed in the original paper: LSA common (4921 common attributes for the base set, 605 common attributes for the novel set); LSA common $\rightarrow$ rare (5526 common attributes for the base set, 4012 rare attributes for the novel set).

OVAD [3]. OVAD introduces the open-vocabulary attributes detection task with a clean and densely annotated attribute evaluation benchmark (no training set is provided). The benchmark defines 117 attribute classes for over 14,300 object instances. Tab. 1 contains the detailed statistics for all relevant datasets.

Dataset	Train	Eval.	Description	Images	Categories/Attributes
MS-COCO	-	-	original COCO detection dataset [10]	118K	80
VAW	-	-	original VAW attribute prediction dataset [12]	58K	620
COCO Cap	-	-	COCO Caption dataset [4]	118K	image-text pairs
CC 3M	-	-	Conceptual Captions 3M dataset [18]	3M	image-text pairs
LSA	✓	✓	original LSA dataset for training and evaluating [13]	420K	5526
OVAD	✗	✓	original OVAD benchmark for evaluating [3]	2K	117
COCO-base	✓	✗	base categories on COCO dataset [1]	107K	48
VAW-base	✓	✗	base attributes on VAW dataset	58K	541
CC-3M-sub	✓	✗	available online pairs filtered by the dictionaries	1M	noise
COCO-Cap-sub	✓	✗	image-text pairs filtered by the dictionaries	118K	noise
COCO-novel	✗	✓	65 categories on COCO val dataset setted by [1]	5K	65
VAW-novel	✗	✓	all attributes in VAW test dataset	10K	620

Table 1. A summary of dataset statistics

C. Comparison with the State-of-the-Art

On the OVAD benchmark, training data is not provided, we directly evaluate the OvarNet that is trained with COCO, VAW, and COCO-Cap-sub. On the LSA dataset, we train OvarNet with the base attribute annotations in LSA common and LSA common → rare for evaluation purposes.

Cross-dataset Transfer on OVAD Benchmark. We compare with other state-of-the-art methods on OVAD benchmark [3], following the same evaluation protocol, we conduct zero-shot cross-dataset transfer evaluation with CLIP-Attr and OvarNet trained on COCO Caption dataset. Metric is average precision (AP) over different attribute frequency distributions, 'head', 'medium', and 'tail'. As shown in Tab. 2, our proposed models largely outperform other competitors by a noticeable margin.

Evaluation on LSA Benchmark. We evaluate the proposed OvarNet on the same benchmark proposed by Pham *et al.* [13]. As OpenTAP employs a Transformer-based architecture with object category and object bounding box as the additional prior inputs, we have evaluated two settings. One is the original OvarNet without any additional input information; the other integrates the object category embedding as an extra token into the transformer encoder layer. As shown in Tab. 3, OvarNet outperforms prompt-based CLIP by a large margin and surpasses OpenTAP (proposed in the benchmark paper) under the same scenario, *i.e.*, with additional category embedding introduced. 'Attribute prompt' means the prompt designed with formats similar to "A photo of something that is [attribute]", while 'object-attribute prompt' denotes "A photo of [category] [attribute]". For the 'combined prompt', the outputs of the 'attribute prompt' and the 'object-attribute prompt' are weighted average.

Method	Box Setting	AP_{all}	AP_{head}	AP_{medium}	AP_{tail}
CLIP RN50 [16]	given	15.8	42.5	17.5	4.2
CLIP VIT-B16 [16]	given	16.6	43.9	18.6	4.4
Open CLIP RN50 [6]	given	11.8	41.0	11.7	1.4
Open CLIP ViT-B16 [6]	given	16.0	45.4	17.4	3.8
Open CLIP ViT-B32 [6]	given	17.0	44.3	18.4	5.5
ALBEF [9]	given	21.0	44.2	23.9	9.4
BLIP [8]	given	24.3	51.0	28.5	9.7
X-VLM [20]	given	28.1	49.7	34.2	**12.9**
OVAD [3]	given	21.4	48.0	26.9	5.2
CLIP-Attr RN50 (ours)	given	24.1	54.8	29.3	6.7
CLIP-Attr ViT-B16 (ours)	given	26.1	55.0	31.9	8.5
OvarNet ViT-B16 (ours)	given	**28.6**	**58.6**	**35.5**	9.5
OV-Faster-RCNN [3]	free	14.1	32.6	18.3	2.5
Detic [21]	free	13.3	44.4	13.4	2.3
OVD [17]	free	14.6	33.5	18.7	2.8
LocOv [2]	free	14.9	42.8	17.2	2.2
OVR [19]	free	15.1	46.3	16.7	2.1
OVAD [3]	free	18.8	47.7	22.0	4.6
OvarNet ViT-B16 (ours)	free	**27.2**	**56.8**	**33.6**	**8.9**

Table 2. Cross-dataset transfer evaluation on OVAD benchmark across all, head, medium, and tail attributes. Numbers are copied from [3].

Method	Setting	LSA common			LSA common → rare		
		AP_{base}	AP_{novel}	AP_{all}	AP_{base}	AP_{novel}	AP_{all}
CLIP	attribute prompt	2.53	3.37	2.64	2.62	2.52	2.58
CLIP	object-attribute prompt	0.97	1.56	1.04	1.16	0.73	0.97
CLIP	combined prompt	2.81	3.67	2.92	3.12	2.63	2.91
OpenTAP	w/category prior	14.34	7.62	13.59	15.39	5.37	10.91
OvarNet	wo/category prior	9.15	4.69	8.52	9.46	3.40	6.17
OvarNet	w/category prior	**15.57**	**8.05**	**14.84**	**16.74**	**5.48**	**11.83**

Table 3. Evaluation of LSA common and LSA common → rare. Following the evaluation protocol in the original paper [13], all results are evaluated in a **box-given setting**.

D. Ablation Study

In this section, we provide additional ablation studies that are not included in the main paper, due to space limitations.

Effect of Prompt Vectors. We have conducted experiments by varying numbers of prompt vectors in the CLIP-Attr, all results are obtained from the model after Step-I training. Prompt vectors are split evenly and placed before, between, and after the attribute and parent-class attribute word. As illustrated in Tab. 4, our model is relatively robust to the different number of prompt vectors.

# prompts	VAW		COCO	
	$\mathbf{AP}_{\text{novel}}$	$\mathbf{AP}_{\text{all}}$	$\mathbf{AP}_{\text{novel}}$	$\mathbf{AP}_{\text{all}}$
3	56.94	66.39	45.27	54.45
9	57.13	66.72	45.50	54.86
15	57.24	66.80	45.77	55.05
30	57.39	66.92	45.82	55.21
60	57.41	67.11	45.79	55.32

Table 4. Effect of different numbers of prompt vectors in CLIP-Attr with first step alignment.

Effect of Different Pooling Strategies. We adopt different architectures to extract regional visual features, including CNN and Transformer. The CNN architecture contains three convolution blocks with a stride of 2, followed by average pooling and a 2-layer MLP, while attentional pooling consists of a 4-layer transformer encoder. As illustrated in Tab. 5, we observe that employing the Transformer with attentional pooling to extract regional visual representation significantly outperforms the convolutional blocks w/ or w/o knowledge distillation.

Visual head	Distil.	VAW		COCO	
		$\mathbf{AP}_{\text{novel}}$	$\mathbf{AP}_{\text{all}}$	$\mathbf{AP}_{\text{novel}}$	$\mathbf{AP}_{\text{all}}$
CNN Blocks-AvgPool	none	48.69	60.58	28.63	58.46
Transformer-AttnPool	none	50.53	61.74	30.43	59.83
CNN Blocks-AvgPool	Prob. KL	53.35	65.80	49.40	62.15
Transformer-AttnPool	Prob. KL	56.43	68.52	54.10	67.23

Table 5. Ablation study on different pooling strategy with a box-given setting.

Effect of Transformer Encoder Layers. Here, we also perform an ablation investigation on different numbers of transformer encoder layers in attentional pooling using probability distillation. As indicated in Tab. 6, the number of transformer encoder layers has only a slight influence on performance, and a 4-layer transformer is sufficient to achieve comparable performance.

# layers	VAW		COCO	
	$\mathbf{AP}_{\text{novel}}$	$\mathbf{AP}_{\text{all}}$	$\mathbf{AP}_{\text{novel}}$	$\mathbf{AP}_{\text{all}}$
2	55.02	67.19	52.28	65.33
4	56.43	68.52	54.10	67.23
6	56.71	68.26	53.90	67.17

Table 6. Different number of transformer encoder layers in attentional pooling under a box-given setting.

E. Qualitative Results

In Fig. 1, we show the qualitative results of OvarNet on VAW and MS-COCO benchmarks. OvarNet is capable of accurately localizing, recognizing, and characterizing objects based on a broad variety of novel categories and attributes.

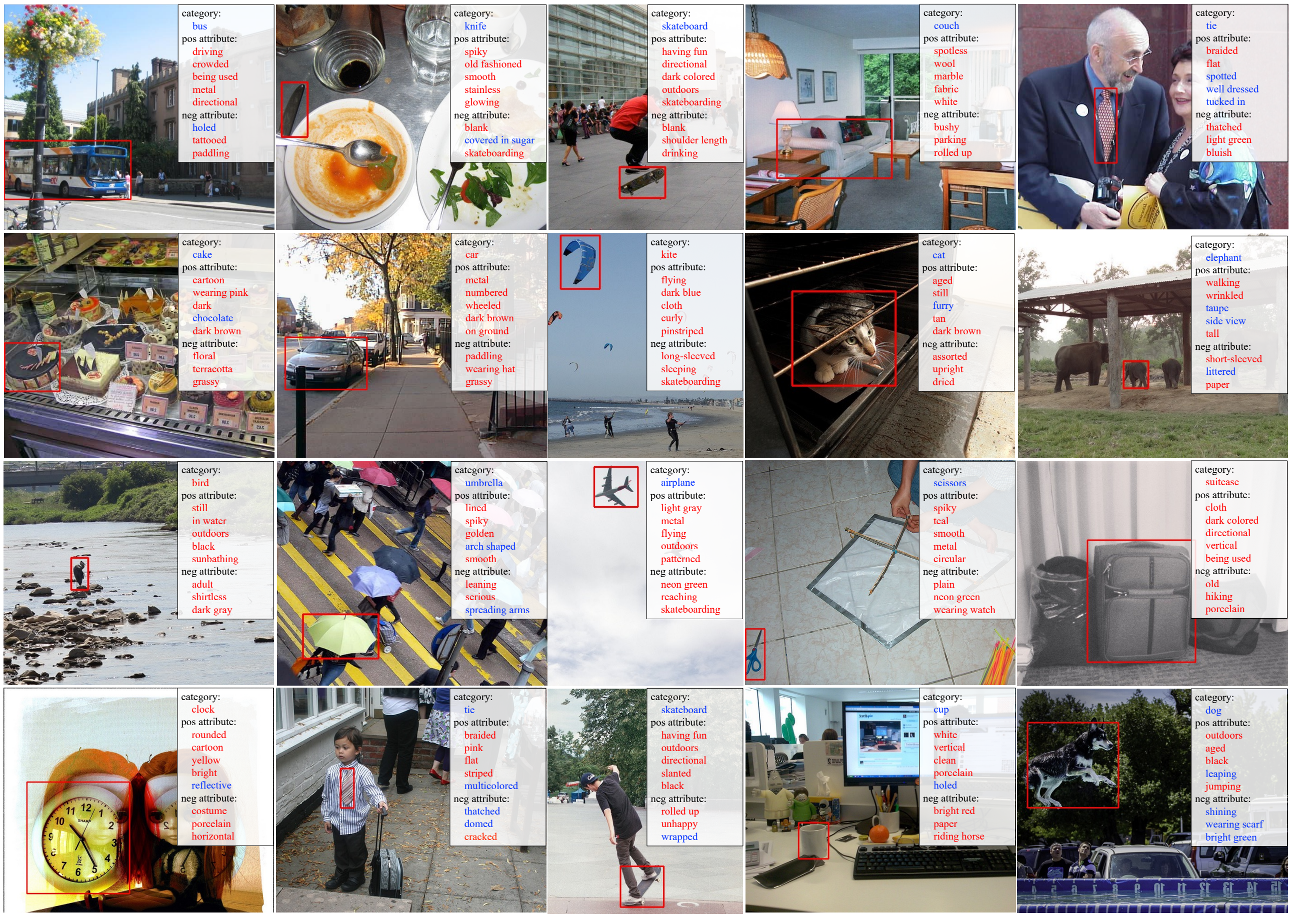

Figure 1. Visualization of prediction results. **Red** denotes the base category/attribute *i.e.*, seen in the training set, while **blue** represents the novel category/attribute unseen in the training set. The first two rows are samples from the VAW test set, while the last two rows are from the COCO val set.

F. Failure Cases & Limitations

In this section, we present some analysis of failure cases, as depicted in Fig. 2, hoping it will inspire future works. Generally speaking, we observe three major failure types: partial localisation, *e.g.*, (a), (b), and (c); misclassification for the semantic category, *e.g.*, (f), (g), and (h); partially inaccurate attribute descriptions, *e.g.*, (d), (e), (i), and (j).

Partial localisation refers to the cases with inaccurate localisation, as shown in Fig. 2 (a), (b), and (c). We discover that a target may be represented by many bounding boxes and that some bounding boxes only encompass a portion of the object, yet they are not removed after non-maximum suppression and have high confidence in the classification score. We believe that partial localisation is mostly caused by the localisation component, and category classification is achieved by following the guidance of response from the partial area in the object.

Misclassification of semantic category denotes that an object is recognized with low confidence, as illustrated in Fig. 2 (f), (g), and (h). Given a box proposal, it is difficult to remove the none-object error boxes, as the classifier may also be able to infer the category with context information. For example, Fig. 2 (h) shows a failure case of a tie.

Partially inaccurate attribute description denotes inaccurate attribute prediction, as illustrated in Fig. 2 (d), (e), (i), and (j). We find the model appears to assign representations of the surrounding environment or background to the object in some circumstances.

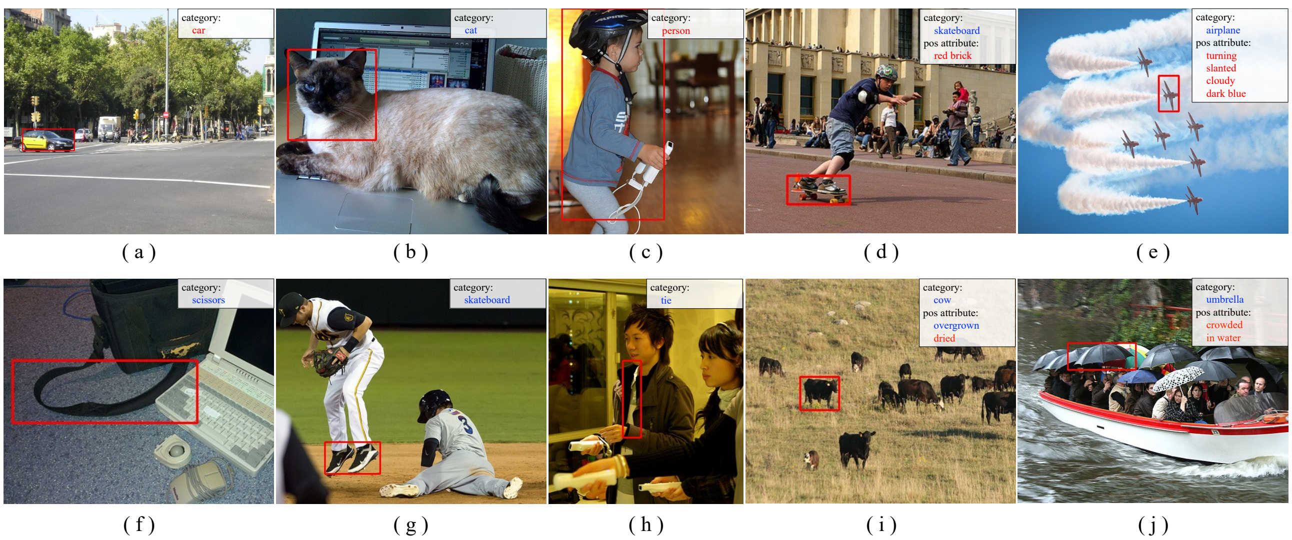

Figure 2. Visualization of failure cases. **Red** denotes the base category/attribute *i.e.*, seen in the training set, while **blue** represents the novel category/attribute unseen in the training set.

References

[1] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 384–400, 2018. 2

[2] Maria A Bravo, Sudhanshu Mittal, and Thomas Brox. Localized vision-language matching for open-vocabulary object detection. *arXiv preprint arXiv:2205.06160*, 2022. 3

[3] María A Bravo, Sudhanshu Mittal, Simon Ging, and Thomas Brox. Open-vocabulary attribute detection. *arXiv preprint arXiv:2211.12914*, 2022. 2, 3

[4] 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. 2

[5] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 2

[6] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. 3

[7] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 2

[8] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 3

[9] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 3

[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 *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[11] Genevieve Patterson and James Hays. Coco attributes: Attributes for people, animals, and objects. In *European Conference on Computer Vision*, pages 85–100. Springer, 2016. 2

[12] Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13018–13028, 2021. 2

[13] Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Improving closed and open-vocabulary attribute prediction using transformers. In *European Conference on Computer Vision*, pages 201–219. Springer, 2022. 2, 3

[14] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 2

[15] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *European conference on computer vision*, pages 647–664. Springer, 2020. 2

[16] 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*, pages 8748–8763. PMLR, 2021. 3

[17] Hanoona Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. *arXiv preprint arXiv:2207.03482*, 2022. 3

[18] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 2

[19] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 3

[20] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021. 3

[21] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Phillip Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. *arXiv preprint arXiv:2201.02605*, 2022. 3