OvarNet: Towards Open-vocabulary Object Attribute Recognition

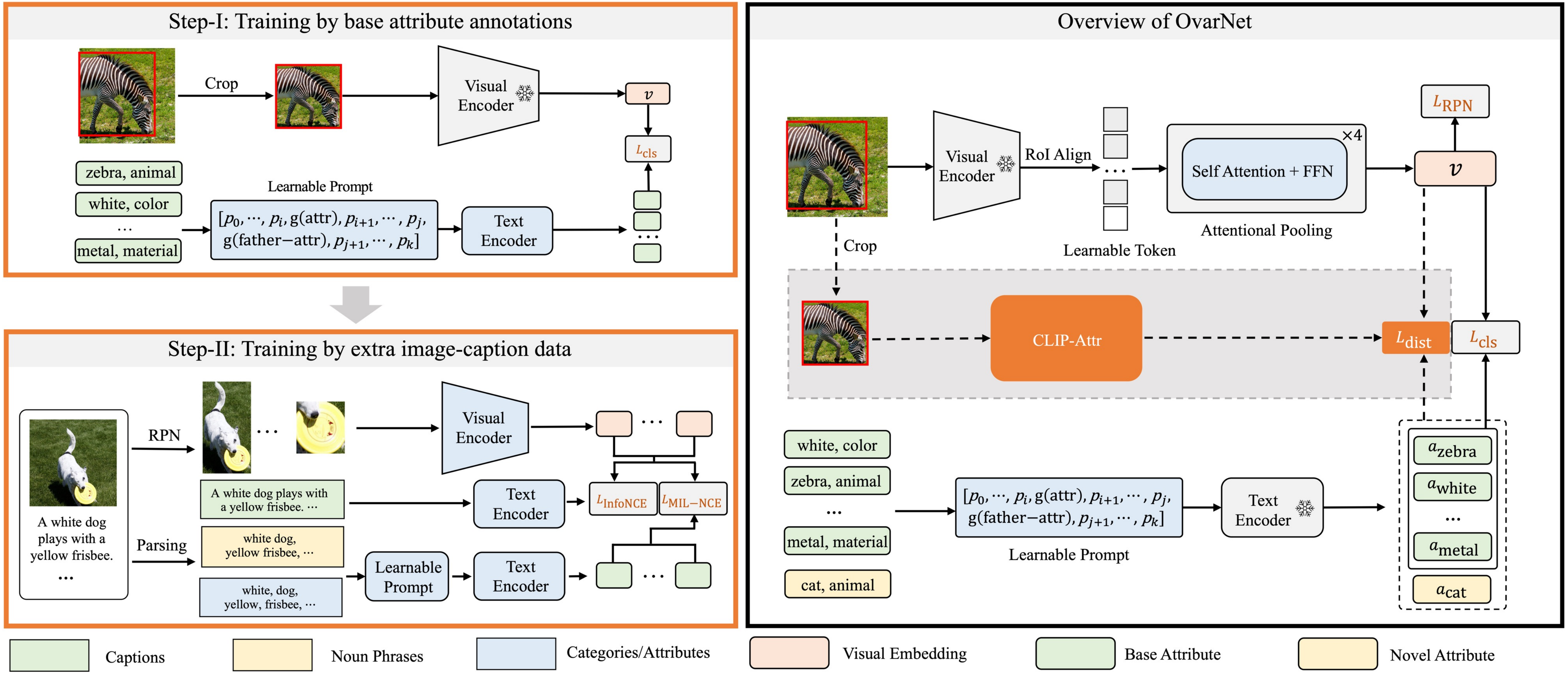
Keyan Chen[1*], Xiaolong Jiang[2*], Yao Hu[2], Xu Tang[2], Yan Gao[2], Jianqi Chen[1], Weidi Xie[3,4✉]

Beihang University[1], Xiaohongshu Inc[2], CMIC Shanghai Jiao Tong University[3], Shanghai AI Laboratory[4]

Paper Tag: THU-PM-278

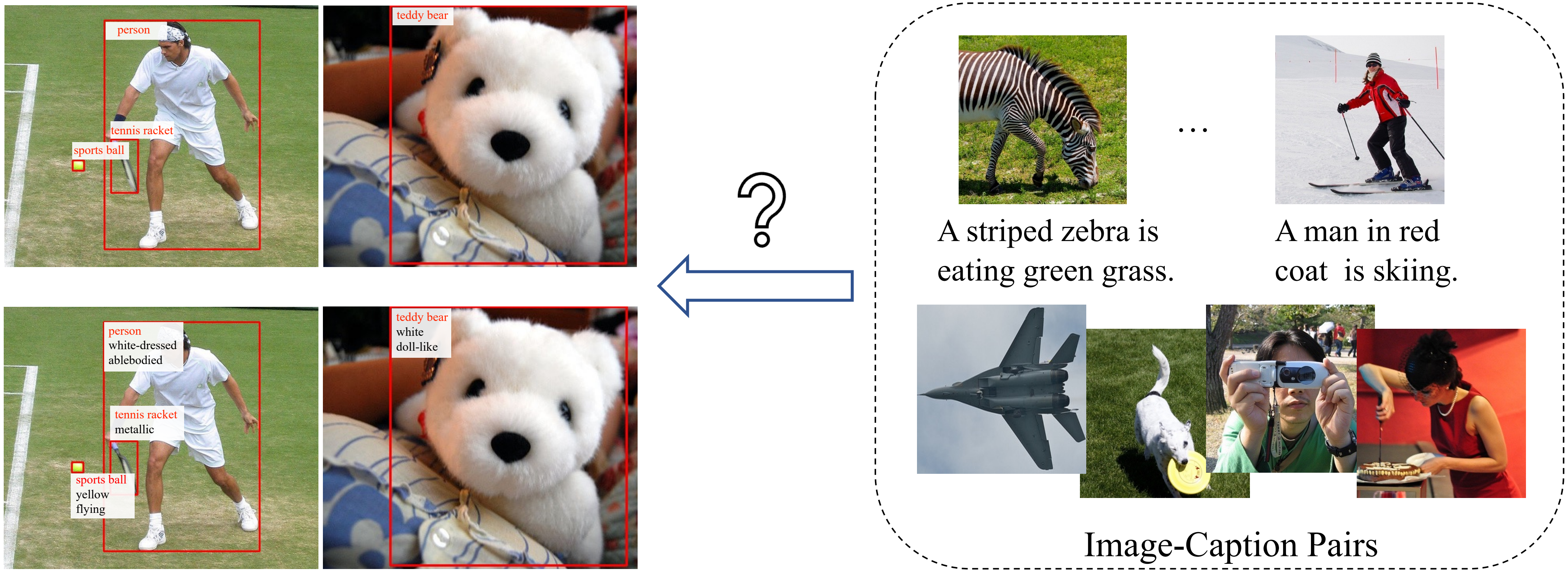
OvarNet : Towards Open-vocabulary Object Attribute Recognition

- Our model can simultaneously localize, categorize, and characterize arbitrary objects in an open-vocabulary scenario.
- In the paper, we leverage Pretrained VL model and freely available image-caption pairs for training and verify that the recognition of semantic category and attributes is complementary for visual scene understanding.



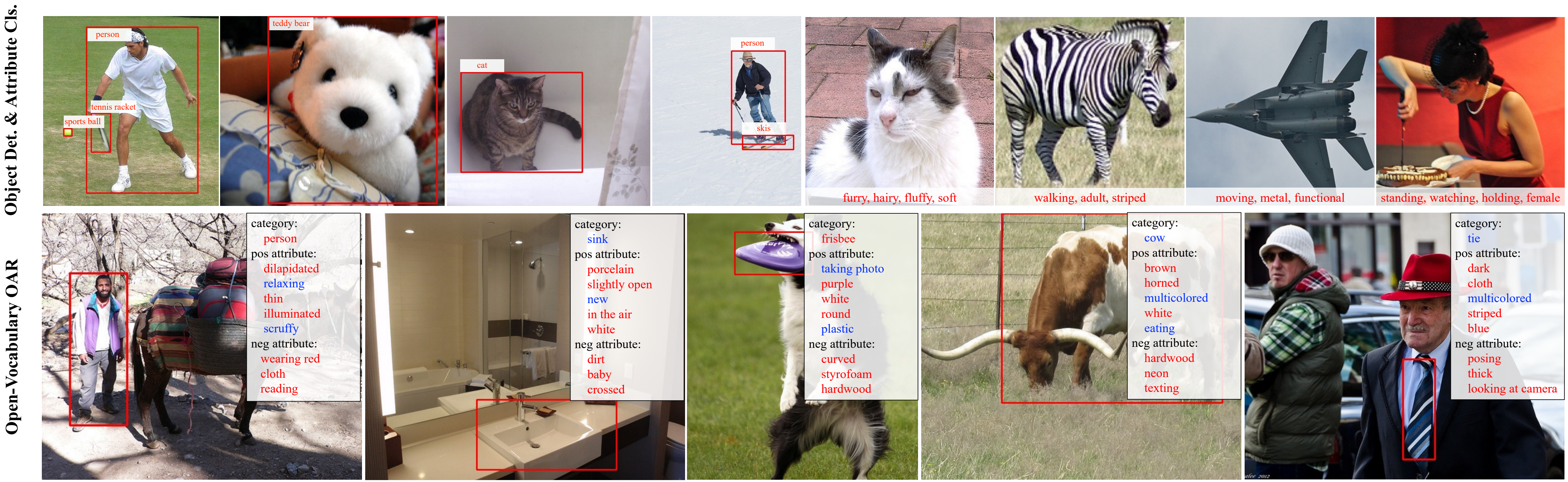
Motivation

- Labelling an object just by category has largely over-simplified our understanding of the visual world.
- When looking around the world, we often understand visual scene by objects via attribute cues.
- Visual language corpora are freely available online, can they be used to aid visual understanding?



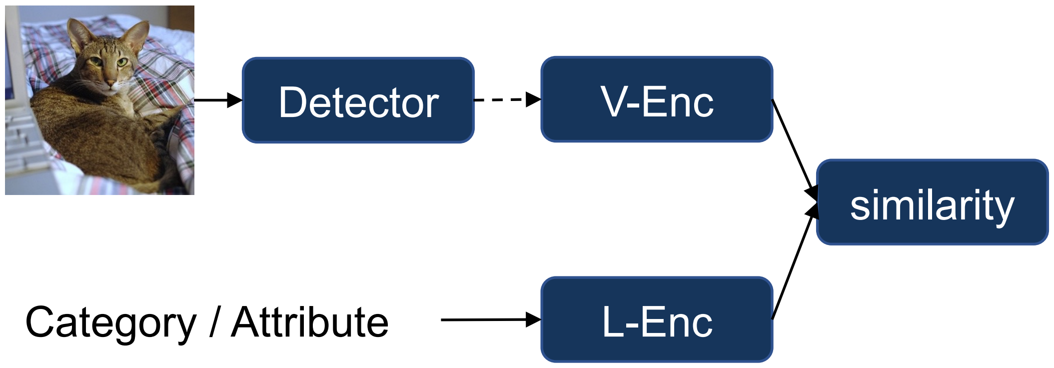
Objective

- Simultaneously localize, categorize, and characterize arbitrary objects in an open-vocabulary scenario.
- Verify that the recognition of semantic category and attributes is complementary for visual scene understanding.

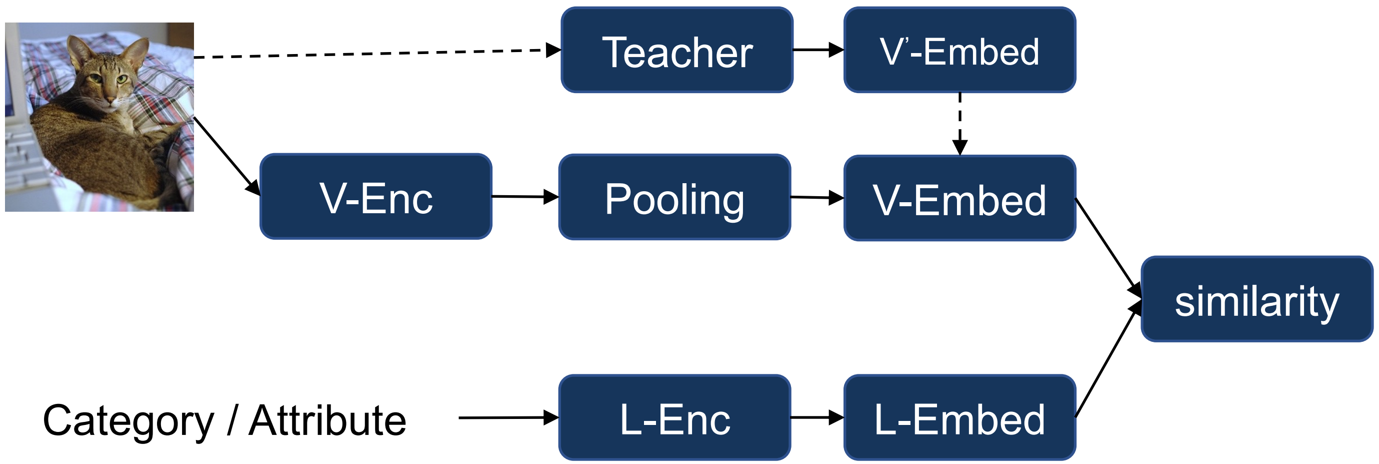


Method

- We start with a naive two-stage approach for open-vocabulary object detection and attribute classification.
- We finetune the VL model by a federated training strategy and investigate the efficacy of leveraging freely available online image-caption pairs.
- We train a Faster-RCNN type model end-to-end with knowledge distillation.



2-Stage Architecture



Distilled 1-Stage Architecture

2-Stage Architecture

- **Problem Setting:**

$$\{\hat{b}_k\} = \Phi_{\text{LOC}} = \Phi_{\text{crpn}}(\mathcal{I})$$
$$\{\hat{c}_k, \hat{a}_k\} = \Phi_{\text{CLS}} = \Phi_{\text{cls}} \circ \Phi_{\text{clip-v}} \circ \Phi_{\text{crop}}(\mathcal{I}, \{\hat{b}_k\})$$

- **Class-agnostic Region Proposal:** propose the candidate regions that potentially have objects situated.
- **Generating Attribute Embedding:** obtain attribute/category embeddings via two variants of prompts.
- **Attribute Classification:** compute the similarity between visual region feature and attribute concept embedding.
- **Training Procedure:** to better align the regional visual feature to the attribute description.

2-Stage Architecture

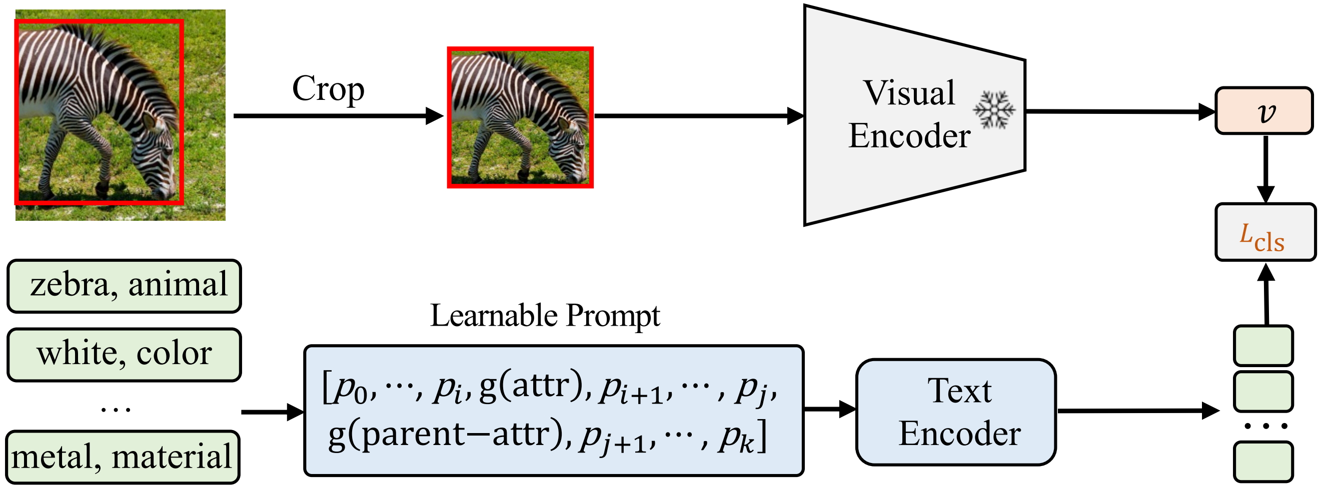
➢ **Generating Attribute Embedding** :

$$\hat{t}_j = \Phi_{\text{clip-t}}([p_0, \cdots, p_i, \mathrm{g}(\text{attribute}), p_{i+1}, \cdots, p_j, \ \mathrm{g}(\text{parent-attribute}), p_{j+1}, \cdots, p_k])$$

- employ prior knowledge of ontologies, and encode their parent-class words along with the attribute.
- augment it with multiple learnable prompt vectors.

➢ **Training Procedure** :

- Step-I: Federated Training. exploit the annotations in existing datasets, i.e., detection and attribute prediction.



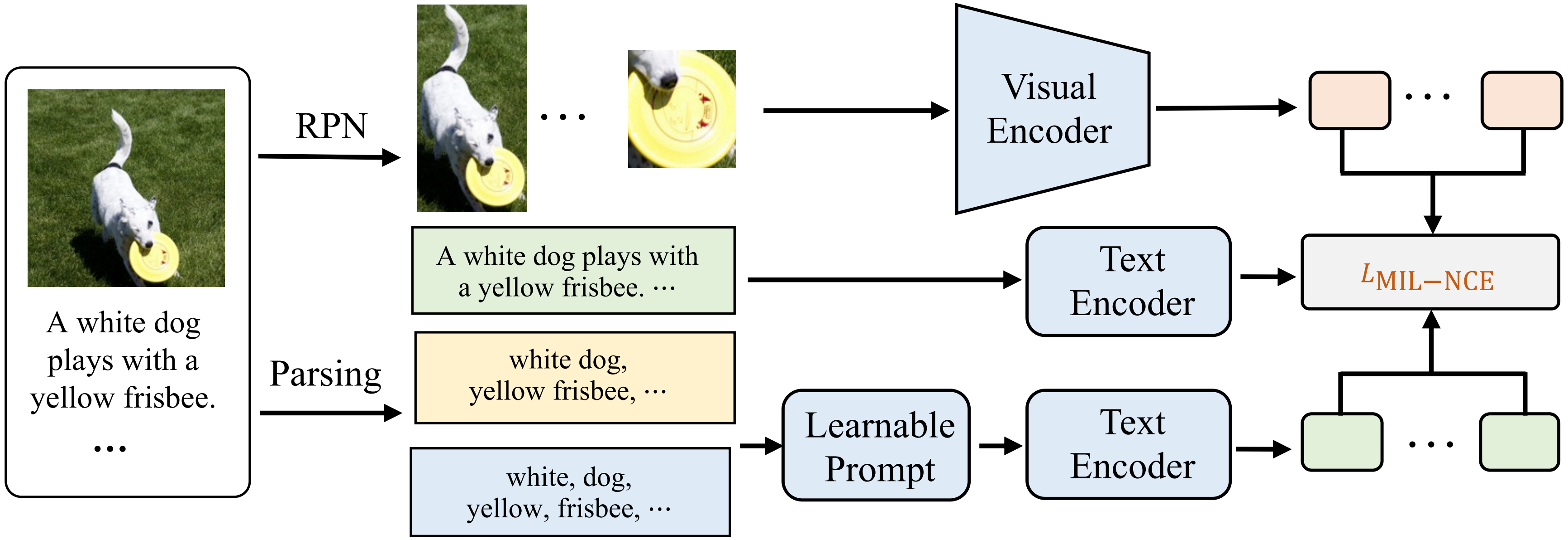
Step-I: Training by base attribute annotations

2-Stage Architecture

➢ **Training Procedure** :

- Step-II: Training with Image-caption Dataset. consider using freely available image-caption datasets to further improve the alignment, especially for novel attributes.

$$\mathcal{L}_{\text{MIL-NCE}} = -\log \frac{\sum_{(v,t)\in\mathcal{P}} \exp(\frac{\langle v^T, t\rangle}{\tau})}{\sum_{(v,t)\in\mathcal{P}} \exp(\frac{\langle v^T, t\rangle}{\tau}) + \sum_{(v',t')\sim\mathcal{N}} \exp(\frac{\langle v'^T, t'\rangle}{\tau})}$$



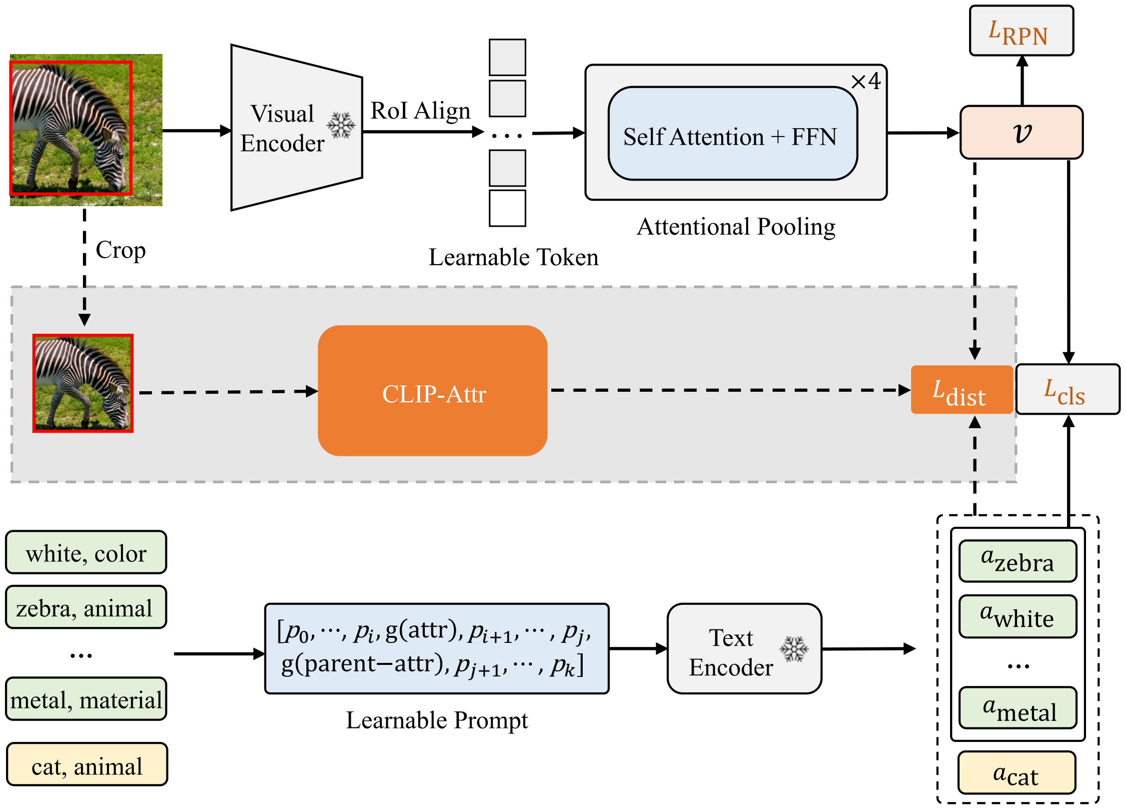
Step-II: Training by extra image-caption data

Distilled Architecture

- **Prediction can be realised with the pre-computed proposals, but the inference is time-consuming.**

- **Problem Setting:**

$$\{\hat{b}_k, \hat{c}_k, \hat{a}_k\} = \Phi_{\text{Ovar}} = \Phi_{\text{cls}} \circ \Phi_{\text{crpn}} \circ \Phi_{\text{v-enc}}(\mathcal{I})$$



- **Visual Encoder:**

  obtain multi-scale feature maps.

- **Class-agnostic Region Encoding:**

  ROI-align with Transformer attentional pooling.

- **Federated Training:**

  jointly supervise localization and classification.

- **Training via Knowledge Distillation:**

  encourage similar prediction between two and one stage model.

Evaluation

- Our considered open-vocabulary object attribute recognition involves two sub-tasks: open-vocabulary object detection and classifying the attributes for all detected objects.
- **Dataset**
  - MS-COCO – 48 classes are selected as base, and 17 classes are used as unseen/novel classes.
  - VAW – half of the 'tail' attributes and 15% of the 'medium' attributes as the novel set.
  - LSA – LSA common and LSA common → rare.
  - OVAD – open-vocabulary attributes detection with a annotated attribute evaluation benchmark.
- **Metrics**
  - both box-given and box-free settings.
  - mAP over base set classes, novel set classes, and all classes.

[1] Bansal, Ankan, et al. "Zero-shot object detection." ECCV. 2018.
[2] Pham, Khoi, et al. "Learning to predict visual attributes in the wild." CVPR. 2021.
[3] Pham, Khoi, et al. "Improving Closed and Open-Vocabulary Attribute Prediction Using Transformers." ECCV. 2022.
[4] Bravo, María A., et al. "Open-vocabulary Attribute Detection." arXiv preprint. 2022.

Results

- Benchmark on COCO and VAW. OvarNet surpasses the recent state-of-the-art ViLD-ens and Detic by a large margin, showing that attributes understanding is beneficial for open-vocabulary object recognition.

Method	Training Data	VAW AP_{base}	VAW AP_{novel}	VAW AP_{all}	COCO $AP50_{base}$	COCO $AP50_{novel}$	COCO $AP50_{all}$
SCoNE [25]	fully supervised	-	-	68.30	-	-	-
TAP [26]	fully supervised	-	-	65.40	-	-	-
OVR-RCNN [38]	COCO Cap	-	-	-	46.00	22.80	39.90
OVR-RCNN [38]	CC 3M	-	-	-	-	-	34.30
ViLD [8]	CLIP400M	-	-	-	59.50	27.60	51.30
Region CLIP [42]	COCO Cap	-	-	-	54.80	26.80	47.50
Region CLIP [42]	CC 3M	-	-	-	57.10	31.40	50.40
PromptDet [6]	Web Images	-	-	-	-	26.60	50.60
Detic [44]	COCO Cap	-	-	-	47.10	27.80	45.00
OvarNet (box-given)	COCO-base + VAW-base	68.27	53.75	66.85	60.94	41.44	55.85
OvarNet (box-given)	+CC 3M-sub	69.30	55.44	67.96	68.35	52.34	64.18
OvarNet (box-given)	+COCO Cap-sub	**69.80**	**56.43**	**68.52**	**71.88**	**54.10**	**67.23**
OvarNet (box-free)	COCO-base + VAW-base	67.71	53.42	66.03	56.20	32.02	49.77
OvarNet (box-free)	+CC 3M-sub	67.32	54.26	66.75	59.50	33.68	52.40
OvarNet (box-free)	+COCO Cap-sub	**68.93**	**55.47**	**67.62**	**60.35**	**35.17**	**54.15**

Table 7. Comparison for open-vocabulary object detection and attribute prediction on the VAW test set and COCO validation.

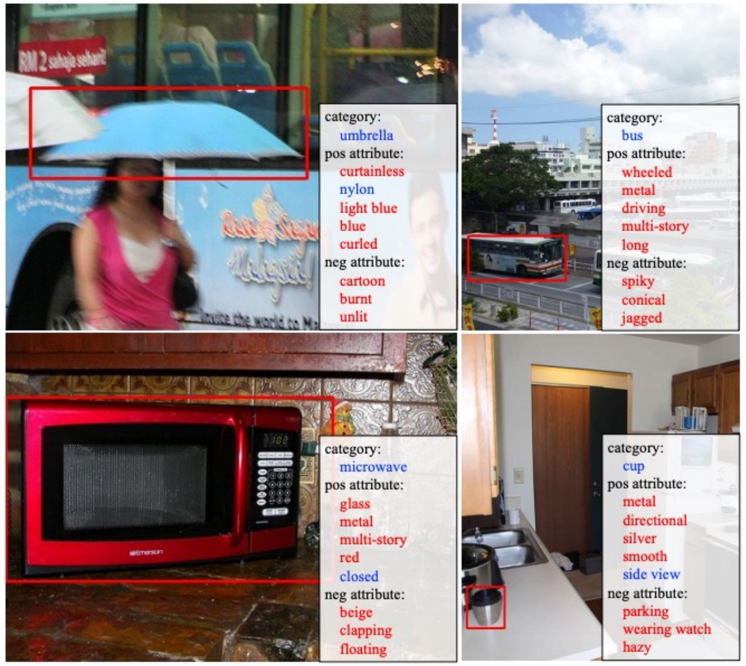


Figure 3. Qualitative visualization from OvarNet. **Red**: base category/attributes. **Blue**: category/attributes.

Results

➢ Cross-dataset Transfer on OVAD Benchmark and Evaluation on LSA Benchmark

- Our proposed models largely outperform other competitors by a noticeable margin.

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**

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**

Qualitative Results

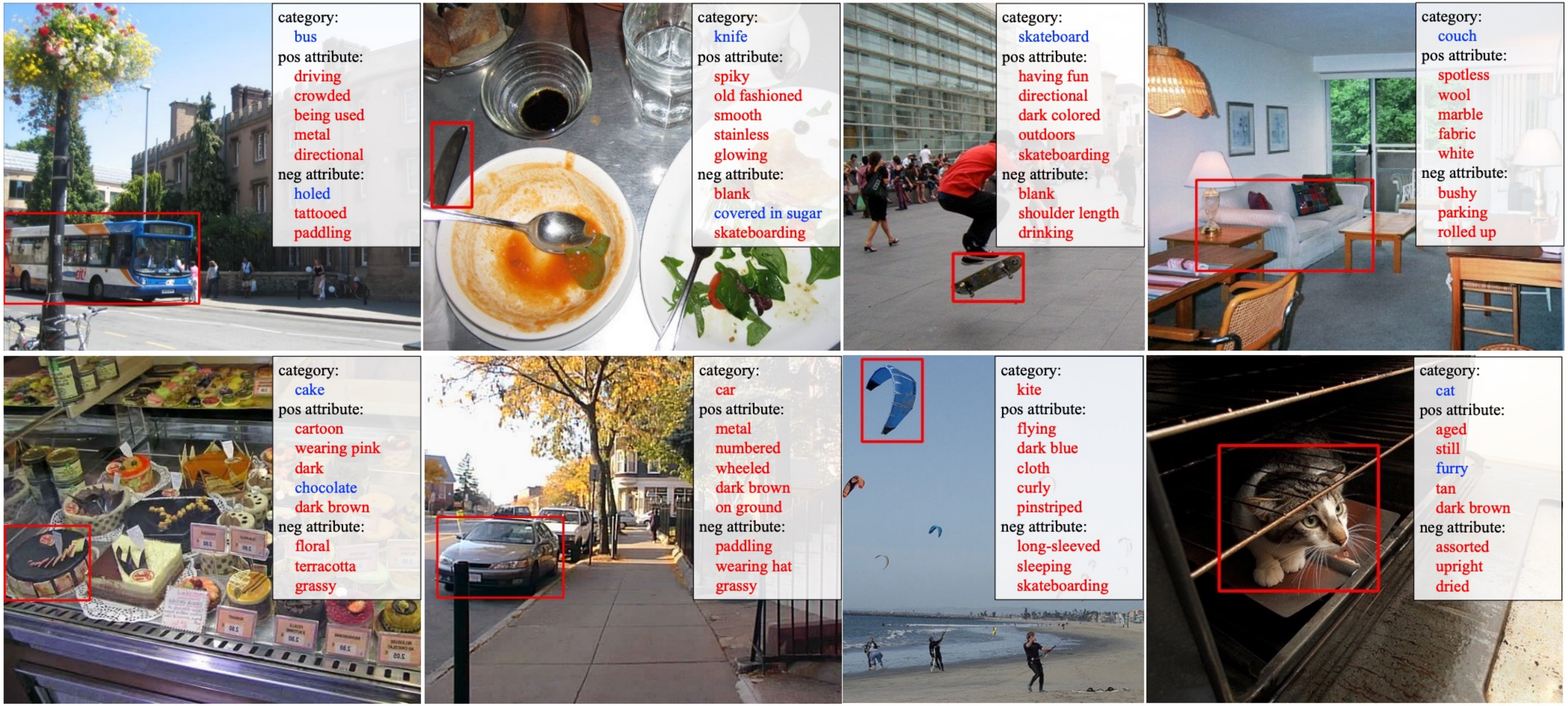


Figure 1. Visualization of prediction results. **Red** denotes the base category/attribute *i.e.*, seen in the training set, while **blue** represents the novel category/attribute unseen in the training set.