OVMR: Open-Vocabulary Recognition with Multi-Modal References

Supplementary Material

A. Ablation Study

A.1. The Number of Visual Tokens

Tab. 10 demonstrates the impact of different numbers of visual tokens on the multi-modal and fused classifiers. We observe the highest average performance for both classifiers across 11 classification datasets when P is set to 2. Consequently, the number of visual tokens is set to 2.

Table 10. Impact of different numbers of visual tokens on average accuracy across 11 classification datasets.

P	OVMR_{VT}	OVMR
1	80.69	82.16
2	**80.99**	**82.34**
4	80.95	82.23
8	80.54	82.17

A.2. Temperature in Preference Generation

Tab. 11 illustrates that setting τ_p to 10 yields the best average performance for the fused classifier. Furthermore, the performance variations under different τ_p values are not significant, suggesting that our method is insensitive to τ_p. This observation validates the robustness of our proposed preference-based fusion approach. Based on these findings, we set τ_p to 10 in our experiments.

Table 11. Average accuracy of the fused classifier with different τ_p across 11 classification datasets.

τ_p	OVMR
1	81.82
10	**82.34**
20	82.07

B. Comparison on Novel Sets

By embedding multi-modal clues of novel categories into vision-language models, the generalization ability of our method is superior to prompt learning methods on novel sets(the last half of categories) of 11 classification datasets. Further comparisons of our method with prompt learning methods across the novel sets of 11 classification datasets are illustrated in Tab. 12. It's important to note that Tab. 1 presents results on the base sets of these 11 classification datasets, conducted in the same 16-shot manner. When evaluating prompt learning methods on novel sets, they do not require additional images after fine-tuning the base categories of each dataset. Our method necessitates a few exemplar images to embed visual clues into the categories of novel sets. Using just one image, our method surpasses current state-of-the-art (SoTA) methods in average performance across 11 datasets. With merely two images, our method achieves an unprecedented average performance exceeding 80.00%. Increasing the number of images to 16 per category boosts average performance to 84.76%, significantly outperforming current SoTA methods by 9.62%. Considering the ease of collecting online data and our method's plug-and-play nature without extra training, it stands as an acceptable competitor to prompt learning methods in low-shot settings. Our method provides a generalizable and efficient approach to embedding multi-modal clues of novel classes into VLMs.

C. Analysis of Preference Weight

In this section, we delve into how the preference-based fusion module mitigates the adverse effects of low-quality text or images. This is achieved by adjusting the preference weights of different classifiers in response to the variable quality of multi-modal references.

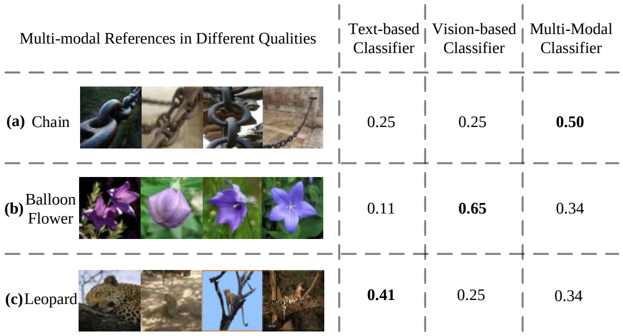

Figure 4. The variation in preference weight for different classifiers corresponding to multi-modal references of various qualities. (a) The category name "Chain" and the exemplar images are both of high quality and effectively complement each other. (b) The category name "Balloon Flower" may not describe the fine-grained flower in detail and is of low quality, whereas the exemplar images accurately represent the category. (c) The exemplar images with various backgrounds, poses, and appearances are of low quality, but the common word "Leopard" clearly defines the animal.

In Fig. 4(a), both the category "Chain" and its corresponding exemplar images are of high quality or complementary to each other. Consequently, this category favors the multi-modal classifier, assigning it the highest preference weight of 0.50. In Fig. 4(b), the category name

Table 12. **Open-vocabulary Classification Results on Novel Sets in Prompt Learning Setup.**

Methods	Shot Number	ImageNet	Caltech101	OxfordPets	Cars	Flowers102	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	Average
CLIP [36]	0	68.14	94.00	97.26	74.89	77.80	91.22	36.29	75.35	59.90	64.05	77.50	74.22
CoOp [59]	0	67.88	89.81	95.29	60.40	59.67	82.26	22.30	65.89	41.18	54.74	56.05	63.22
CoCoOp [58]	0	70.43	93.81	97.69	73.59	71.75	91.29	23.71	76.86	56.00	60.04	73.45	71.69
MaPLe [18]	0	70.54	94.36	97.76	74.00	72.46	92.05	35.61	78.70	59.18	73.23	78.66	75.14
Ours	1	66.67	94.33	95.47	76.60	93.87	89.30	39.10	77.07	61.37	73.60	82.33	77.25
	2	71.83	94.17	97.50	76.90	95.53	91.03	41.40	81.00	67.57	83.83	84.20	80.45
	4	73.03	95.13	97.57	79.90	96.80	91.50	46.17	83.03	69.43	84.30	87.13	82.18
	16	74.87	96.30	97.67	86.23	97.13	91.70	52.03	84.60	74.73	89.57	87.57	84.76

"Balloon Flower" may fail to adequately describe the fine-grained characteristics of the flower in detail, reflecting its low quality, while the exemplar images depict the category more accurately. Thus, the vision-based classifier is assigned the highest preference weight, effectively mitigating the negative impact of the low-quality category name. Conversely, in Fig. 4(c) for the category "Leopard", the exemplar images are of low quality as a result of various backgrounds, poses, and appearances. In contrast, the common word "Leopard" can clearly illustrate the animal. Therefore, the text-based classifier receives the highest preference weight, compensating for the poor quality of the exemplar images.

D. Sources of Exemplar Images

In Tab. 9, we showcase the performance of our method on the base classes of ImageNet using exemplar images sourced from the Internet and ImageNet's training set. When crawling images for a given category, we initiate the process by using the category name as a search query on Google. The first 16 images returned by Google are downloaded as the exemplar images for this category. In Fig. 5, we present a set of examples crawled from the web and examples sampled from ImageNet's training set. It is evident that the images in ImageNet typically exhibit a higher diversity within the same class. The diversity includes differences in the background environment, the number of subjects, their poses, etc. Conversely, images sourced from the Internet often focus on a single subject, featuring simpler poses and backgrounds. Furthermore, images obtained from the Internet can be less reliable due to the noise and ambiguity inherent in text-based queries. For instance, a search for 'Jaguar' may yield images of either the animal or the car, as illustrated in the last row of Fig. 5. As demonstrated in Tab. 9, the use of diverse images from ImageNet's training set as exemplars results in enhanced performance. This improvement is attributed to the closer domain correlation of the test set with the training set in ImageNet, as well as the potential noise and ambiguity of web-crawled images.

E. Visualization of Detection Results

To demonstrate the effectiveness of our method in open-vocabulary detection, we separately showcase the detection results of our OVMR model using the Swin-B backbone, specifically for all categories in Fig. 6 and for the novel categories in Fig. 7, on the LVIS dataset.

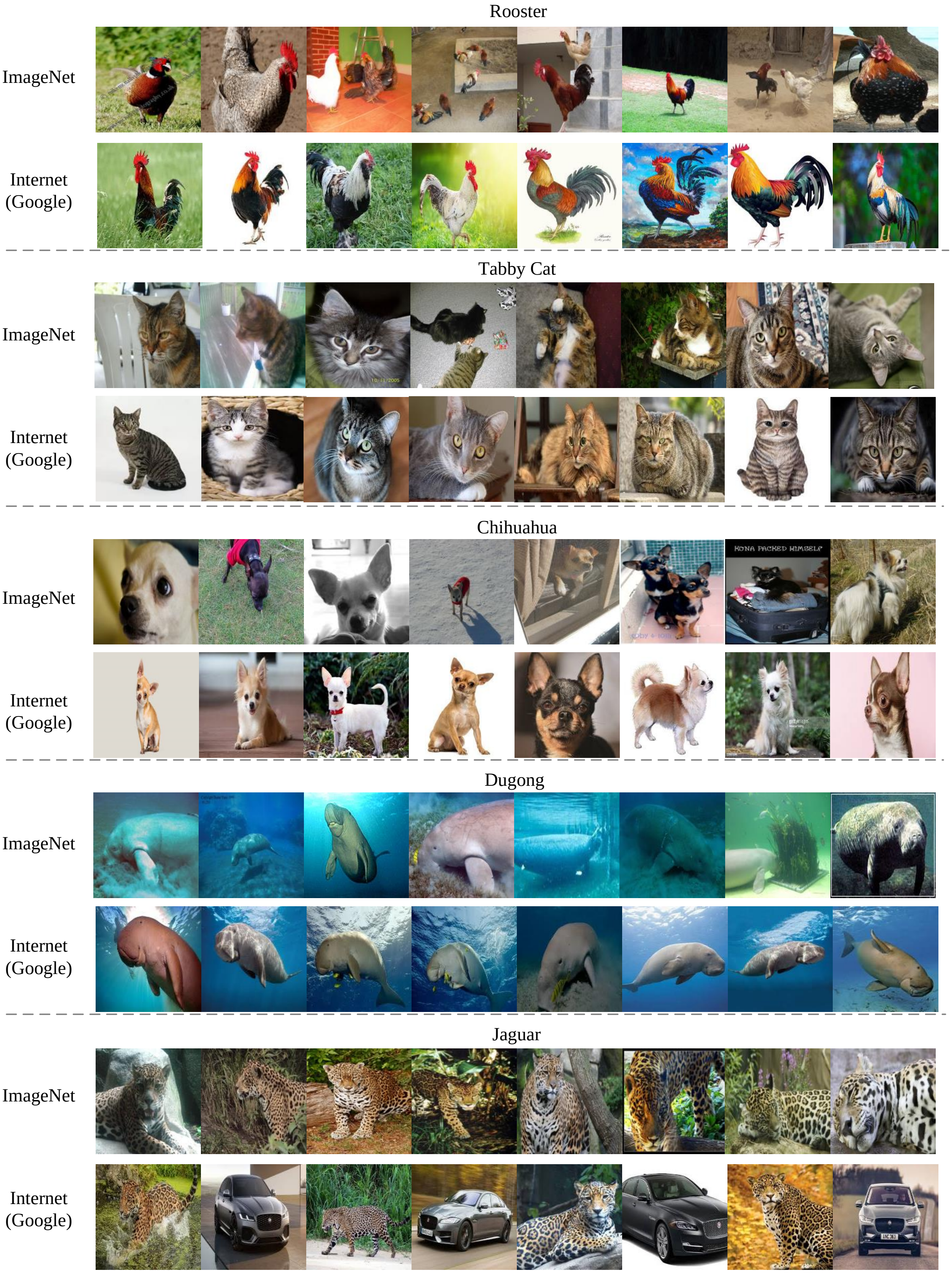

Figure 5. Exemplar images sampled from the training set of ImageNet and the Internet-crawled images.

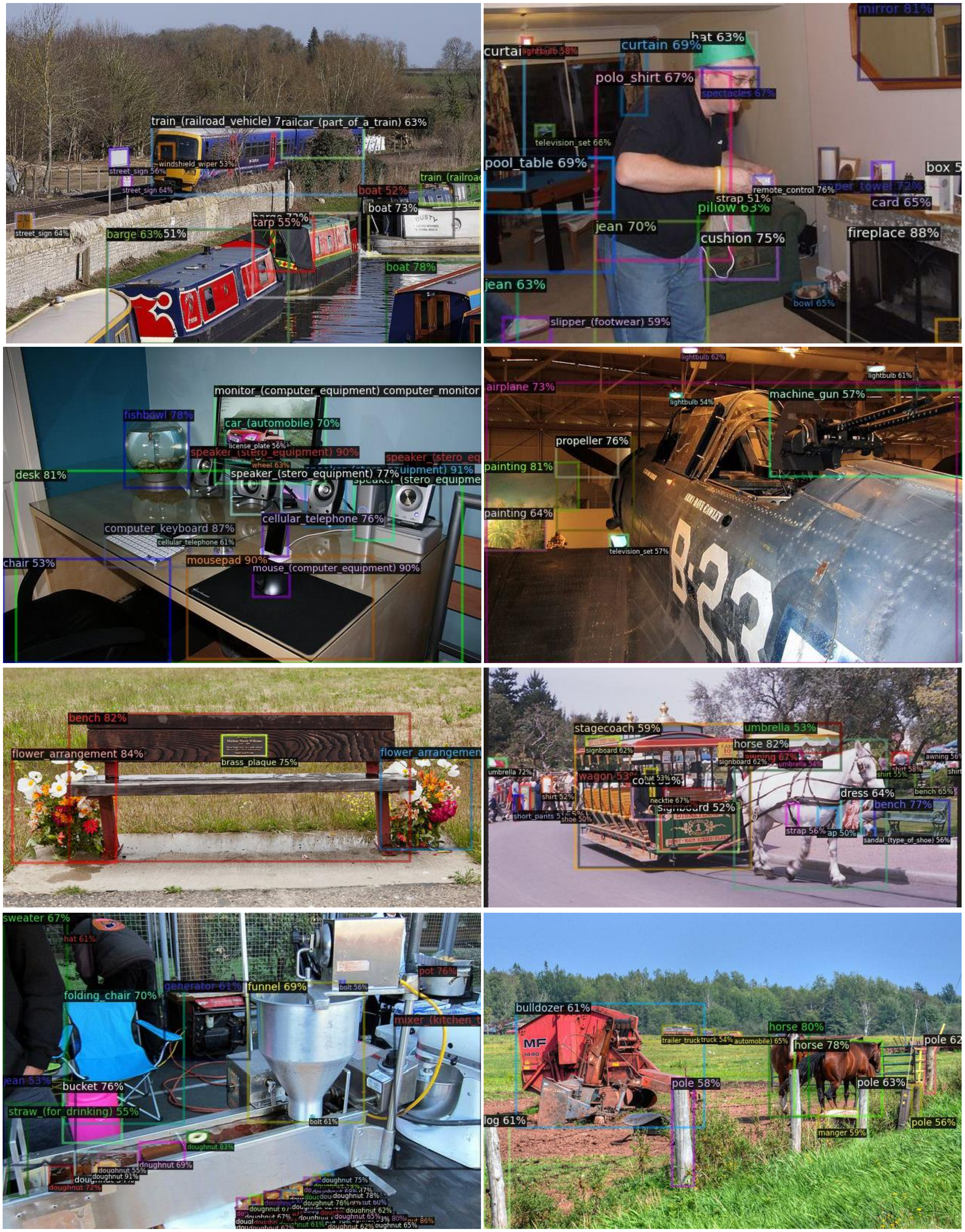

Figure 6. Detection results for all LVIS categories.

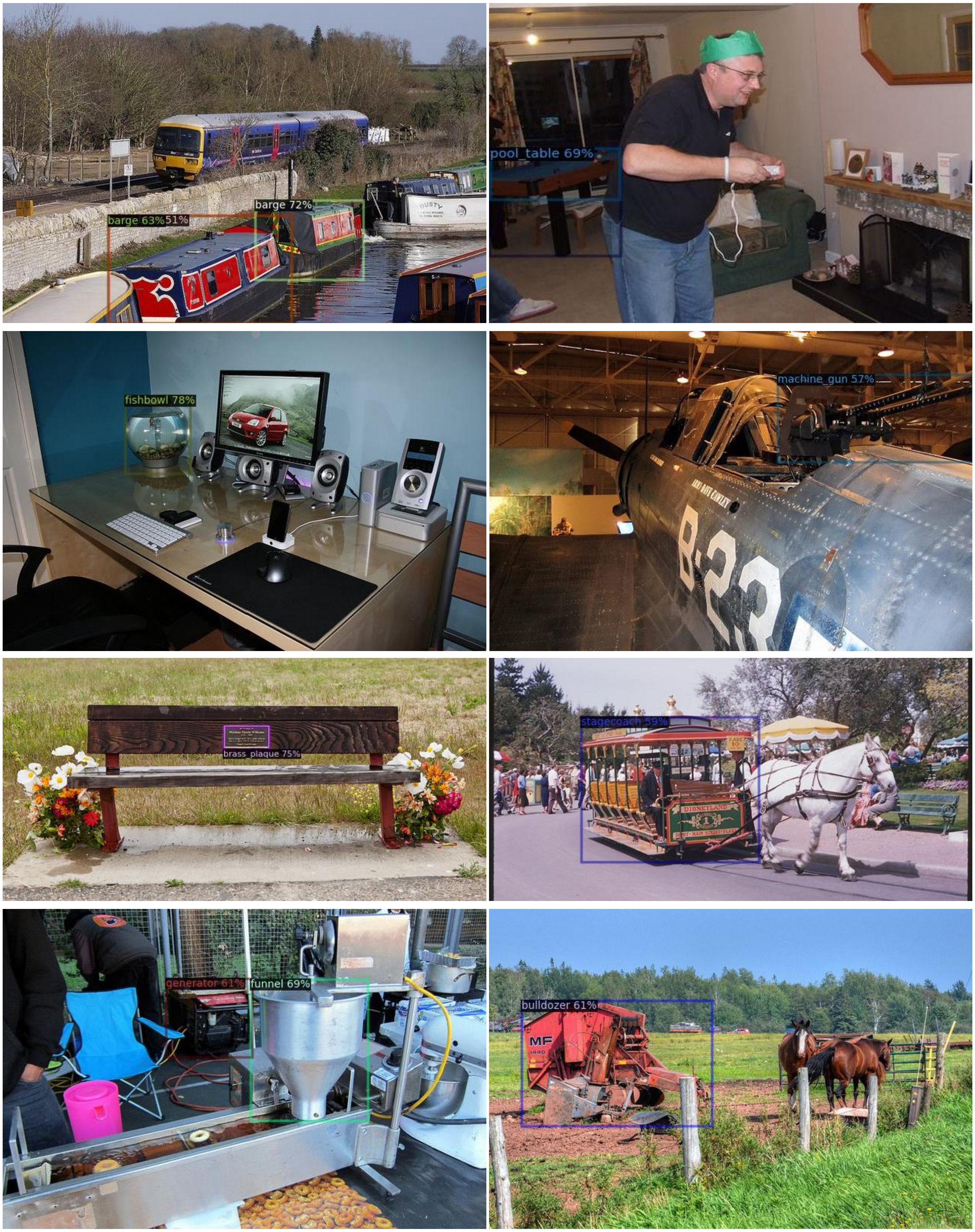

Figure 7. Detection results for novel LVIS categories.