A Appendix

A.1 Constructing dataset variations with small objects

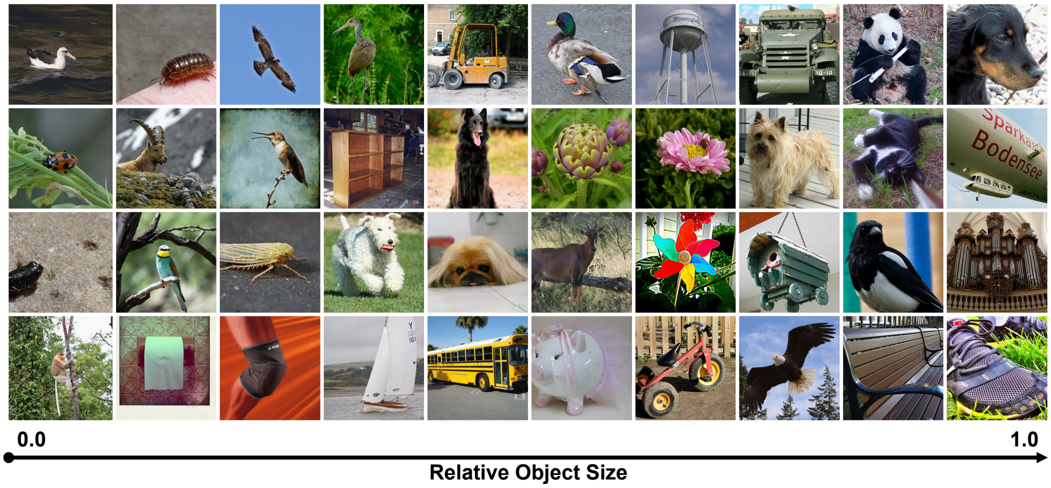


Figure 9: Example images from ImageNetS919 with different relative object sizes.

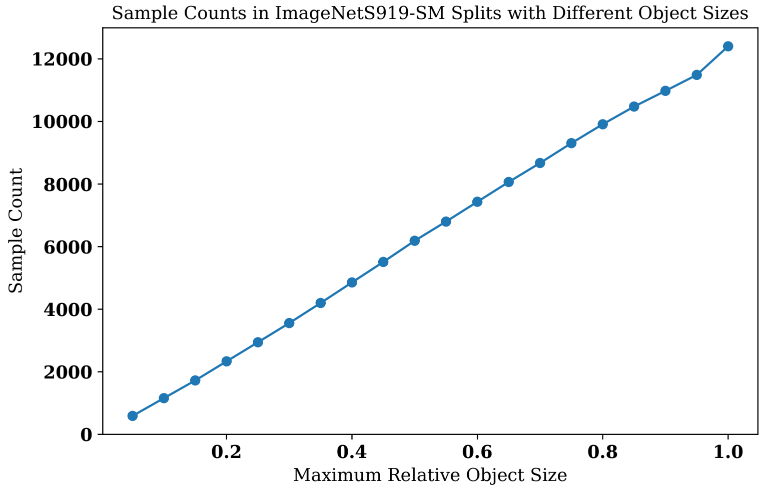


Figure 10: The number of samples in each object size condition of ImageNetS919.

In section 5, we use datasets based on ImageNetS and CUB as well as their small object variations (e.g., ImageNetS-SM and CUB-SM). In this section, we provide more details how those small variations are constructed.

For each image sample, its object size is computed based on object bounding box. In case of CUB, the bounding box is obtained directly from available annotations. However, for ImageNetS, only its pixel-wise segmentation is provided. In this case, object bounding box can be extracted from the segmentation in terms of minimum and maximum coordinates along X and Y axes of object-labelled pixels.

Given an image x_i of size $w \times w$ with the object bounding box represented in terms of minimum/maximum XY coordinates as $(p^X_{min}, p^X_{max}, p^Y_{min}, p^Y_{max})$, relative object size of the image s_{x_i} is the ratio between the area of object bounding box and the total image area which can be computed as follows:

$$s_{x_i} = \frac{(p^X_{max} - p^X_{min})(p^Y_{max} - p^Y_{min})}{w^2}. \tag{4}$$

The value of s_{x_i} will be within the range of $[0, 1]$. Example images with different values of s_{x_i} are shown in Figure 9.

We use s_{x_i} of individual image samples to control object size characteristic of a dataset. In section 5, the datasets with small objects (i.e., ImageNetS919-SM and CUB-SM), are obtained by thresholding s_{x_i} of image samples such that that their values are not larger than 0.2. In section 5.3, multiple thresholds of s_{x_i} are employed on the ImageNetS919 dataset in order to study behavior of our models on different object size conditions. These thresholds are distributed uniformly from 0.05 to 1.0 with the step size of 0.05. The number of samples in each of these object size conditions is presented in Figure 10.

A.2 Additional Zero-Shot Transfer Results

From table 1, we presented zero-shot performance of GC-CLIP variations with different model configurations. In this section, we provide full version of the results including performance of ViT-L/14 and DataComp in Table 3.

A.3 Guided Cropping with Supervised Models

In the main paper, we mainly focus on applying our Guided Cropping to zero-shot models, i.e., CLIP and CALIP. We argue that Guided Cropping can be helpful in this case as image encoders of these models are designed to be generic so that they potentially encode non-discriminative information of input images.

Concerning our Guided Cropping component alone, it is, in fact, orthogonal to supervision strategies. Theoretically, our Guided Cropping can be employed with supervised models as well. In this case, models can be supervisedly trained as normal but, during inference, their input images can be cropped with our Guided Cropping component before forwarding to the models. In this section, we study behaviors of Guided Cropping when it is integrated with few-shot and fully-supervised models.

A.3.1 Few-Shot Models

In this section, we conduct an experiment based on few-shot models, Tip-Adapter and Tip-Adapter-F (Zhang et al., 2021), to learn classification on ImageNetS919-SM and CUB-SM datasets in few-shot (n-shots=16 in our experiment). Its performance without and with Guided Cropping ($\alpha = 0.2$ with no box augmentation) is shown in the table below. According to the table, our Guided Cropping generally improves performance of Tip-Adapter variations. This empirically demonstrates benefits of our Guided Cropping for few-shot models.

A.3.2 Fully-Supervised Models

In this section, we study behaviors of Guided Cropping when it is integrated with pretrained supervised models. In this regard, we utilize ImageNet pretrained models with ViT-B/32, ViT-B/16 and ViT-L/16 backbones from timm (Wightman, 2019), a deep learning library. These models are evaluated on ImageNetS919 and ImageNetS919-SM datsets with/without Guided Cropping. The results are shown in Table 5.

According to the results, optimal performance generally achieves with models without Guided Cropping or with Guided Cropping using large margin ratio, i.e., 0.8, whose crops already cover large context regions. We can observe this behavior even in the case of small objects (ImageNetS919-SM). These results indicate that, for these fully-supervised models, unrelated contexts generally do not degrade classification performance. In contrast, these contexts even improve their performance. This observation is actually not new and has been discussed in shortcut learning literature (Geirhos et al., 2020) that supervisedly trained networks can take unintended visual cues (e.g., background, texture) as shortcuts to gain classification performance on in-distribution samples.

Table 3: Zero-shot classification accuracies from different datasets and model configurations.

Model	Prompt	Guided Cropping	Box Aug.	Dataset			
				ImageNetS919	CUB	ImageNetS919-SM	CUB-SM
CLIP (ViT-B/32)	Category	-	-	63.62	51.83	52.83	49.57
		-	Random Crop	64.42	52.45	53.47	50.79
		✓	-	63.61	52.40	55.18	51.44
		✓	Random Crop	64.46	**53.12**	**56.00**	52.81
		✓	Multi-Margin	**64.66**	**53.12**	**56.00**	**53.09**
	Descriptions	-	-	68.54	53.05	55.70	50.14
		-	Random Crop	69.15	53.62	57.33	50.79
		✓	-	68.59	54.07	58.61	**53.38**
		✓	Random Crop	69.07	54.47	59.08	53.09
		✓	Multi-Margin	**69.62**	**54.56**	**60.07**	52.95
CLIP (ViT-B/16)	Category	-	-	68.60	56.51	57.75	55.54
		-	Random Crop	68.81	56.89	58.05	57.41
		✓	-	68.06	56.09	58.65	55.97
		✓	Random Crop	68.19	56.78	58.35	57.12
		✓	Multi-Margin	**68.94**	**57.30**	**59.81**	**57.63**
	Descriptions	-	-	72.67	57.78	61.61	56.55
		-	Random Crop	73.17	58.87	62.13	57.99
		✓	-	72.61	58.70	63.28	**59.35**
		✓	Random Crop	72.86	58.99	63.32	58.78
		✓	Multi-Margin	**73.49**	**59.34**	**64.05**	59.06
CLIP (ViT-L/14)	Category	-	-	75.15	63.08	64.78	62.16
		-	Random Crop	75.30	63.32	64.70	62.59
		✓	-	75.00	62.96	66.02	62.16
		✓	Random Crop	75.04	63.24	66.54	62.73
		✓	Multi-Margin	**75.71**	**63.63**	**66.92**	**63.17**
	Descriptions	-	-	78.48	64.65	67.78	63.17
		-	Random Crop	78.65	64.60	67.65	**63.96**
		✓	-	78.32	64.67	69.07	63.31
		✓	Random Crop	78.28	**64.88**	69.41	**63.96**
		✓	Multi-Margin	**79.06**	64.76	**69.88**	62.95
DataComp (ViT-L/14)	Category	-	-	82.05	85.57	69.88	85.18
		-	Random Crop	82.10	86.07	69.84	86.04
		✓	-	81.87	85.85	71.04	86.26
		✓	Random Crop	81.75	85.99	71.04	86.04
		✓	Multi-Margin	**82.36**	**86.19**	**71.51**	**86.62**
	Descriptions	-	-	82.66	86.04	70.01	86.12
		-	Random Crop	82.82	86.45	70.48	86.98
		✓	-	82.33	86.57	71.25	87.19
		✓	Random Crop	82.23	86.62	71.25	87.19
		✓	Multi-Margin	**82.93**	**86.83**	**71.68**	**87.41**

Comparing to cases of other supervision strategies, zero-shot and few-shot models are less likely to be affected by shortcut learning since exposing to none (or few) of samples on target datasets make them less likely to learn unintended visual clues from dataset biases.

A.4 Logit refinement on top-k predictions

As per our method mentioned in section 4.1, after computing preliminary logits from conventional CLIP, only top-k predictions are considered and refined with Guided Cropping. We choose $k = 5$ in this work. In this section, we will provide reasons why we adopt this top-k refinement strategy. Two main reasons are given below.

Table 4: Few-shot performance with Tip-Adapter variations. Accuracies gain from Guided Cropping integration are given in parentheses.

Model	Approach	Guided Cropping	Dataset	
			ImageNetS919-SM	CUB-SM
ViT-B/32	Tip-Adapter	-	56.34	53.45
	Tip-Adapter	✓	58.27 (+1.93)	54.53 (+1.08)
	Tip-Adapter-F	-	62.43	60.22
	Tip-Adapter-F	✓	63.15 (+0.72)	60.07 (-0.15)
ViT-B/16	Tip-Adapter	-	62.34	61.44
	Tip-Adapter	✓	64.05 (+1.71)	62.30 (+0.86)
	Tip-Adapter-F	-	68.04	67.12
	Tip-Adapter-F	✓	68.42 (+0.38)	67.05 (-0.07)
ViT-L/14	Tip-Adapter	-	68.77	70.72
	Tip-Adapter	✓	70.44 (+1.67)	71.94 (+1.22)
	Tip-Adapter-F	-	72.24	73.88
	Tip-Adapter-F	✓	72.15 (-0.09)	74.32 (+0.44)

Table 5: Classification accuracies of ImageNet pretrained models with/without Guided Cropping on ImageNet919.

Architecture	Guided Cropping	Margin Ratio	Box Aug.	Dataset	
				ImageNetS919	ImageNetS919-SM
ViT-B/32	-	-	-	76.82	61.53
ViT-B/32	-	-	Random Crop	77.71	62.21
ViT-B/32	✓	0.2	-	77.11	64.05
ViT-B/32	✓	0.2	Random Crop	77.99	**65.04**
ViT-B/32	✓	0.8	-	76.91	62.81
ViT-B/32	✓	0.8	Random Crop	**78.14**	63.84
ViT-B/16	-	-	-	81.72	68.89
ViT-B/16	-	-	Random Crop	**82.11**	**69.37**
ViT-B/16	✓	0.2	-	81.08	68.42
ViT-B/16	✓	0.2	Random Crop	81.16	68.85
ViT-B/16	✓	0.8	-	81.63	68.51
ViT-B/16	✓	0.8	Random Crop	81.94	**69.37**
ViT-L/16	-	-	-	86.09	75.62
ViT-L/16	-	-	Random Crop	86.35	**76.35**
ViT-L/16	✓	0.2	-	85.67	75.92
ViT-L/16	✓	0.2	Random Crop	85.69	75.54
ViT-L/16	✓	0.8	-	86.21	76.26
ViT-L/16	✓	0.8	Random Crop	**86.37**	**76.35**

- Potential Accuracy: We found that there is already high chances that the correct classes are among predicted top-5 classes. To demonstrate this, we analyze top-1, top-5 and top-10 accuracies of conventional CLIP in Table 6. According to the results, large accuracy gaps can be noticed between top-1 and top-5 accuracies (24.53% for ImageNetS919 and 31.79% for CUB). In other words, by considering only 5 classes for refinement with Guided Cropping, upper bounds of final accuracies are already high. It must be noted that, while this upper bound accuracies can be raised further by considering top-10 classes, the gains compared to top-5 classes are relatively small. This may not worth introducing additional computation to the pipeline. Therefore, we decide to perform Guided Cropping based on predicted top-5 classes in this work.

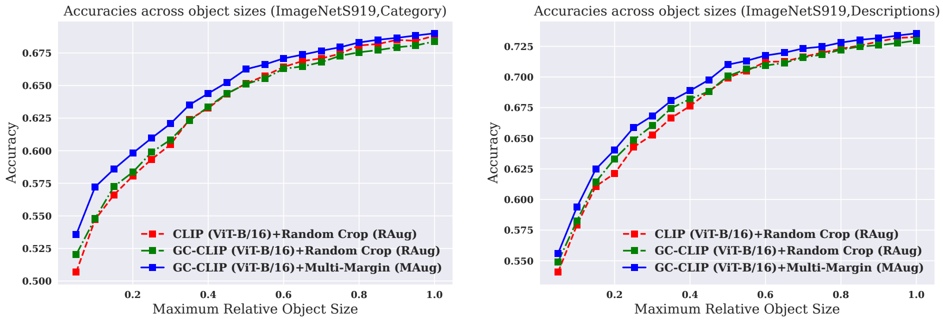
- Common Bounding Boxes: We notice that visual appearances of top-5 classes are relatively similar in most cases. OWL-ViT is also likely to produce similar boxes for these classes. This makes the use of common bounding boxes (e.g., the primary box b_i^0 or the α-margin box b_i^α) among these classes reasonable. To illustrate this, considering each sample in

Table 6: Top-k accuracies from conventional CLIP (ViT-B/32) with category prompts.

Dataset	Accuracy		
	Top-1	Top-5	Top-10
ImageNetS919	63.62	88.15	92.98
CUB	51.83	83.62	90.63

Figure 13 and 14, its primary box generally contains visual features which are (partially) similar to each top class making the box become a decent box candidate for all top classes.

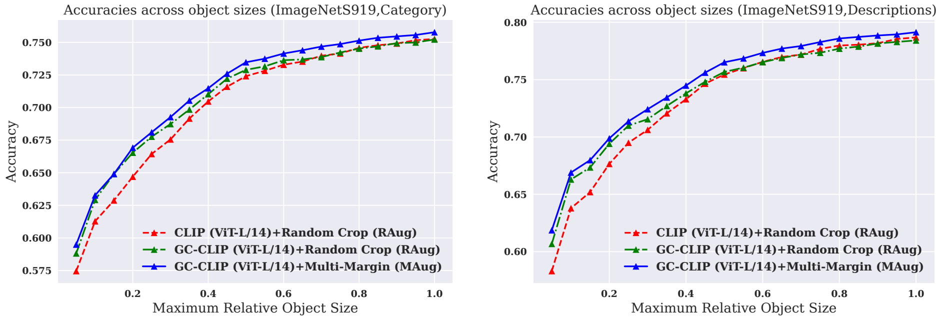
A.5 Accuracies with different object size conditions



(a) Prompt: Category

(b) Prompt: Descriptions

Figure 11: Accuracies (ViT-B/16) on subsets of ImageNetS919 with various object size conditions.



(a) Prompt: Category

(b) Prompt: Descriptions

Figure 12: Accuracies (ViT-L/14) on subsets of ImageNetS919 with various object size conditions.

In section 5.3, we study GC-CLIP performance on various object size conditions and show that GC-CLIP variations outperform baselines especially when target object sizes are small. The plots in Figure 6 are provided for models with ViT-B/32 backbone. In this section, additional evidences with other backbones are provided to support our claim. Figure 11 and 12 show similar plots for models with ViT-B/16 and ViT-L/14 backbones respectively. According to the figures, similar behavior can be observed. There are accuracy gaps between conventional CLIP and GC-CLIP and the gaps are larger on datasets with small objects. This demonstrates that our claim is consistent across different CLIP backbones.

A.6 Inference with OWL-ViT

OWL-ViT performs object detection taking images and text prompts as inputs and producing bounding boxes as well as their scores and class labels as outputs. In this work, for each image sample

Table 7: Accuracies from GC-CLIP (ViT-B/32) with different OWL-ViT inference strategies.

Dataset	Prompt Type	Box Aug.	OWL-ViT Inference	
			Single-Pass	Multi-Pass
ImageNetS919-SM	Category	RAug	54.71	**56.00**
ImageNetS919-SM	Category	MAug	55.61	**56.00**
ImageNetS919-SM	Descriptions	RAug	57.84	**59.08**
ImageNetS919-SM	Descriptions	MAug	59.47	**60.07**
CUB-SM	Category	RAug	50.22	**52.81**
CUB-SM	Category	MAug	**53.09**	**53.09**
CUB-SM	Descriptions	RAug	51.51	**53.09**
CUB-SM	Descriptions	MAug	**53.45**	52.95

Table 8: Average similarity scores between images and their corresponding prompts (i.e., maximum logit values) of correctly classified samples of CLIP (with RAug) and GC-CLIP (with MAug) using ViT-B/32 backbone.

Dataset	Prompt Type	Accuracy with	
		CLIP	GC-CLIP
ImageNetS919-SM	Category	29.39	**29.71**
ImageNetS919-SM	Descriptions	30.17	**30.51**
CUB-SM	Category	33.71	**33.89**
CUB-SM	Descriptions	34.30	**34.55**

x_i, we use OWL-ViT to extract bounding box candidates B_i based on a set of detection prompts of the top-k classes $\{p_j^{det} | j \in J_i^k\}$. Theoretically, there are two possible options to obtain B_i from OWL-ViT.

- Single Forward Pass (Single-Pass): with this option, an input image and all detection prompts are forwarded to OWL-ViT at once. With a single forward pass, OWL-ViT will produce a set of bounding boxes which will be used directly as B_i.
- Multiple Forward Passes (Multi-Pass): with this option, OWL-ViT will perform forward pass with one detection prompt at a time. In other words, there will be k forward passes in total. Each forward pass will produce a set of bounding boxes b_{ij} based on a detection prompt p_j^{det}. Bounding boxes estimated from all forward passes will be merged to get B_i according to equation 2.

As mentioned in section 4.1, we decide to adopt Multi-Pass in our Guided Cropping pipeline as Multi-Pass is more robust to misdetection (if one pass fails, other passes can act as backup passes). In this section, we demonstrate empirically that Multi-Pass can lead to better performance.

In this regard, we conduct an experiment to compare GC-CLIP accuracies when Single-Pass and Multi-Pass are employed. The results are shown in Table 7. According to the results, GC-CLIP with Multi-Pass is consistently better across datasets and model configurations. This confirms our design choice to use Multi-Pass in our Guided Cropping pipeline.

A.7 Similarity between cropped images and their prompts

One motivation of our Guided Cropping is that, by minimizing unrelated information, CLIP image encoder can focus more on target objects leading to better image representations. In section 5.1 better image representations can be indirectly inferred via the improvement of the classification performance. In this section, we would like to analyze image representations in another perspective.

We argue that, if image representations are better, the representations should be not only less similar to prompts of other classes but also more similar to prompts of their own classes. In this regard, we investigate similarities of image embeddings (of the correctly classified samples) to their own prompts. Here, similarity scores are obtained in terms of maximum predicted logit values. Similarity score results of CLIP and GC-CLIP are shown in Table 8. We can notice that similarity scores

Table 9: Performance of GC-CLIP (ViT-B/32) on additional datasets using category-based prompts.

Guided Cropping	Box Aug.	Dataset				
		ImageNet	ImageNetV2	Stanford Dogs	ImageNet-A	ImageNet-R
-	-	58.79	51.88	52.46	29.37	65.26
-	Random Crop	59.31	52.21	53.43	29.28	66.24
✓	-	58.95	52.84	53.92	31.41	65.47
✓	Random Crop	59.46	52.94	**54.73**	31.81	65.99
✓	Multi-Margin	**59.84**	**53.30**	54.12	**31.97**	**66.67**

between images and their corresponding prompts in case of GC-CLIP are consistently higher. This indicates that image representations after Guided Cropping are more similar to their prompts according to our assumption.

A.8 Visualizing example results

In this section, we present top-5 logits estimated from CLIP and GC-CLIP on example samples from ImageNetS919 to demonstrate qualitatively that GC-CLIP can refine logits to make correct predictions. The results are illustrated in Figure 13 and 14.

A.9 Results on additional datasets

In section 5, we aim to study the cases when objects of interest cover small areas of input images. Therefore, image classification datasets with segmentation/bounding box annotations are chosen for evaluation that enable us to quantify the performance on objects covering small areas. Hence, we choose ImageNetS919 and CUB for our evaluation as these datasets provide segmentation/bounding box annotations from which object sizes of image samples can be obtained. These annotations enable more insight studies with different object sizes. These datasets are also commonly used in weakly supervised object localization task (Zhu et al., 2022) as it needs similar annotations during evaluation.

For completeness, we perform evaluation on additional classification datasets without object size annotations as well. However, it must be noted that we may not be able to decouple effects of object size and extraneous image regions in this case. In this section, we present performance of GC-CLIP on ImageNet (Russakovsky et al., 2015), ImageNetV2 (Recht et al., 2019), Stanford Dogs (Khosla et al., 2011), ImageNet-A (Hendrycks et al., 2021b) and ImageNet-R (Hendrycks et al., 2021a) datasets.

The results are shown in Table 9. According to the results, even object sizes of these datasets are not controlled, our GC-CLIP is generally still better than the baselines. The magnitudes of improvement are generally similar to results in Table 1 in the main paper (refering unconstrained variants of ImageNetS919 and CUB).

One interesting observation which must be noted here is GC-CLIP performance on out-of-distribution datasets (i.e., ImageNet-A and ImageNet-R). We can observe that amounts of accuracy gains from GC-CLIP are different depending on out-of-distribution conditions. GC-CLIP benefits better on natural adversarial condition (ImageNet-A) than on rendition condition (ImageNet-R). We attribute this behavior to our dependency of OWL-ViT. In the rendition condition, objects are in unusual contexts such that OWL-ViT performance is not always consistent.

A.10 Comparison with Central Crop

In our work, we demonstrate that image cropping guided by object locations can improve classification performance. To further support this argument, we perform experiments comparing our guided cropping with a deterministic cropping strategy, Central Crop, commonly used for classification (Jia et al., 2021; Zhai et al., 2022; Touvron et al., 2019).

Central Crop benefits under the assumption that target objects likely to locate at the center of input images. During inference, an input image will be cropped around its center according to a predefined

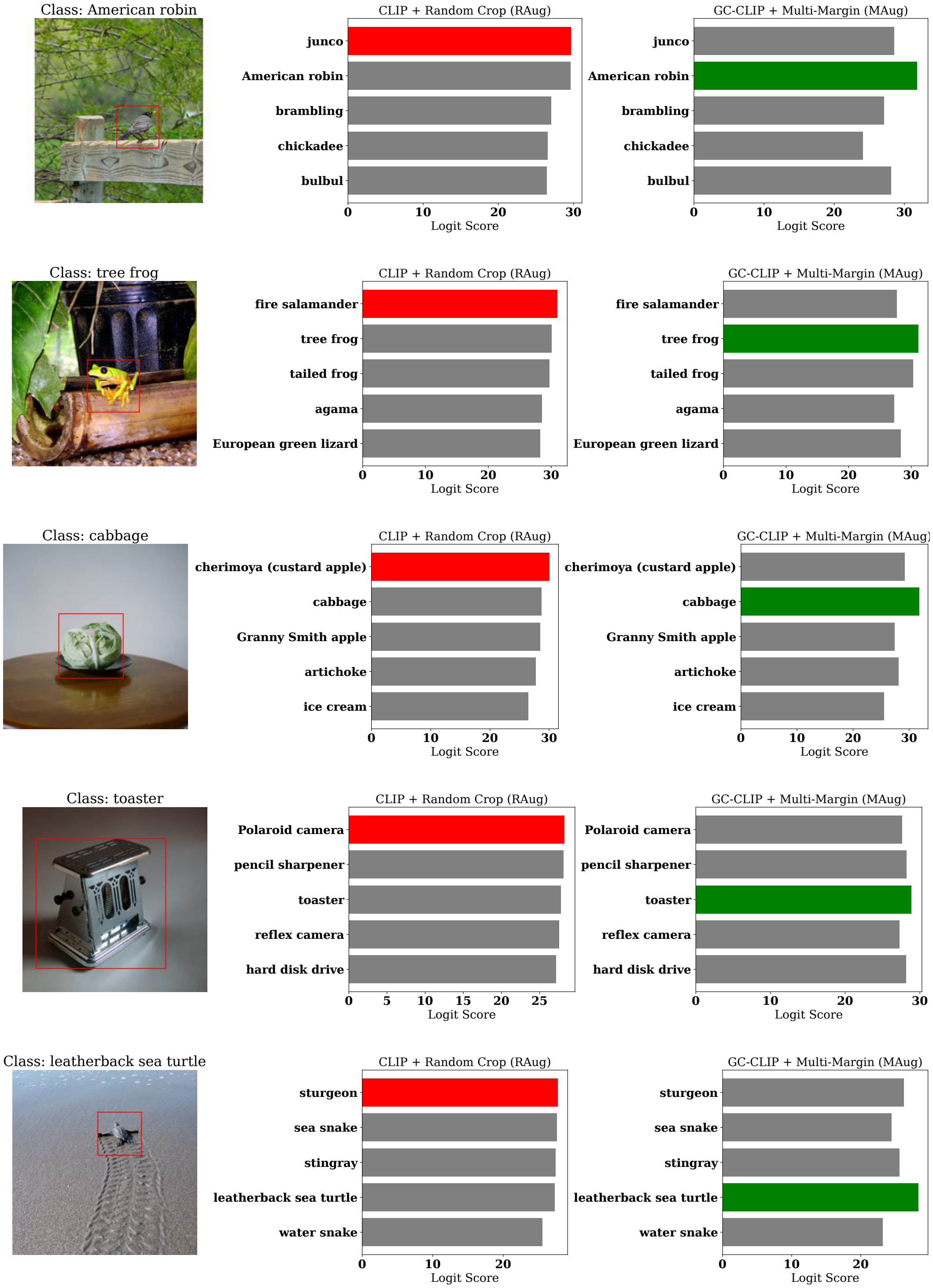


Figure 13: Top-5 logits on example samples improved by Guided Cropping (set 1). Model configurations are CLIP (with RAug) and GC-CLIP (with MAug) using ViT-B/32 backbone and prompt type of descriptions. Red boxes represent primary boxes used in our GC-CLIP pipeline.

cropping ratio from 0.0 to 1.0. The crop ratio of 1.0 refers to the usage of the full images without cropping. Then, the processed image will be resized to a compatible size for employed models be-

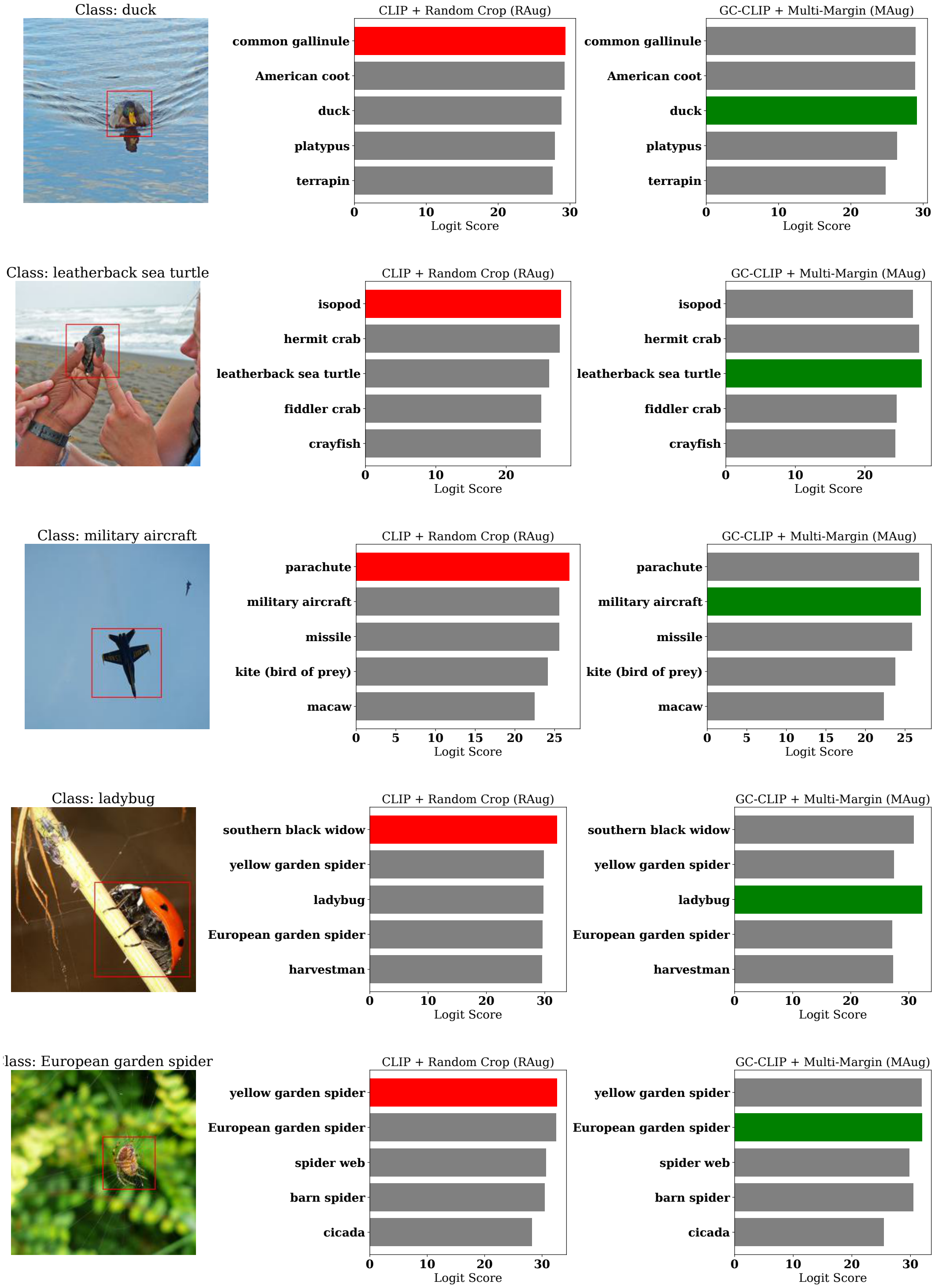
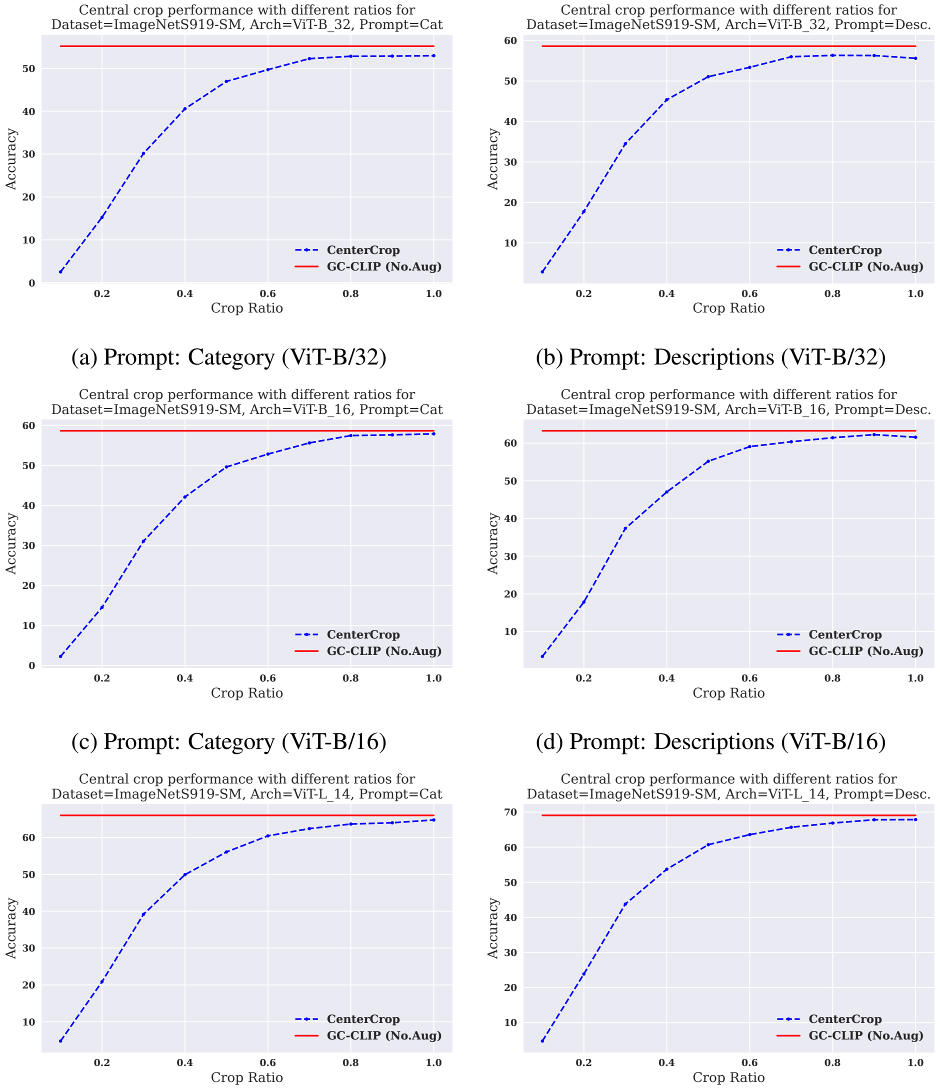


Figure 14: Top-5 logits on example samples improved by Guided Cropping (set 2). Model configurations are CLIP (with RAug) and GC-CLIP (with MAug) using ViT-B/32 backbone and prompt type of descriptions. Red boxes represent primary boxes used in our GC-CLIP pipeline.

fore performing the inference. We conduct experiments with Central Crop using different cropping ratios on ImageNetS919-SM. Its performance can be visualized as in Figure 15.



(a) Prompt: Category (ViT-B/32)

(b) Prompt: Descriptions (ViT-B/32)

(c) Prompt: Category (ViT-B/16)

(d) Prompt: Descriptions (ViT-B/16)

(e) Prompt: Category (ViT-L/14)

(f) Prompt: Descriptions (ViT-L/14)

Figure 15: Central crop performance with different cropping ratios compared to GC-CLIP (without box augmentation) on ImageNetS919-SM.

According to the results, we can see that, models with Central Crop can slightly improve performance compared to vanilla models. For example, according to Figure 15b, the model without Central Crop (ratio=1.0) achieves the accuracy of 55.61 while the model with Central Crop (ratio=0.9) achieves the higher accuracy of 56.30. However, on Figure 15, models with Guided Cropping (without box augmentation) consistently outperform Central Crop. This supports the argument that our cropping approach guided by object locations is preferable over simple cropping at a predefined location.