S1 Training Details and Model Configurations

We set the patch size to be 8. Our model is optimized by AdamW optimizer [3] with a learning rate of 0.0004, 250k training steps, linearly warm-up of 5000 steps and an exponentially weight-decaying schedule. The gradient norm is clipped at 1. We use Pytorch automatic mixed-precision and data paralleling for training acceleration. All models are trained on 4 Nvidia RTX A5000 GPUs with a total batch size of 128. The temperature of cyclic walks is set to 0.1. We use similarity threshold 0.7 and ViT-S8 of DINO [1] for all experiments. We report the mean ± standard deviation of 5 runs with 5 random seeds for all our experiments.

We list the number of slots and image size used for each dataset in Table S1. For Birds, Cars, Dogs and Flowers datasets, we report the performance of Slot-Attention, SLATE and BO-QSA from the work BO-QSA [2]. For other implementation details, we follow all method configurations in the work DINOSAUR [4].

	Birds	Cars	Dogs	Flowers	Pascal VOC 2012	COCO 2017	COCO-Stuff	Movi-C	Movi-E
Slot-Attention	-	-	-	-	6	7	-	11	24
SLATE	-	-	-	-	6	7	-	11	24
DINOSAUR	2	2	2	2	4	7	-	11	24
BO-QSA	-	-	-	-	6	7	-	11	24
Cyclic walks (ours)	2	2	2	2	4	11	11	11	24
Image Size	128	128	128	128	224	224	224	224	224

Figure S1: The choice of the number of slots and image size for all the methods in each dataset. The '-' indicates that the performance results are directly taken from other papers and thus the configuration is not provided here.

S2 Inference Steps of Our Method

S2.1 Unsupervised Foreground Extraction and Unsupervised Object Discovery

During the inference, all the models are asked to predict foreground masks in the unsupervised foreground extraction task and object masks in the unsupervised object discovery task (Section 5.2). We use $M_{x,\hat{s}}$ (Equation 5) as the segmentation masks, where each feature vector at any spatial location of x is softly assigned to a cluster center. The mask with a maximum intersection with the ground truth masks is viewed as the corresponding predicted masks.

S2.2 Unsupervised Semantic Segmentation

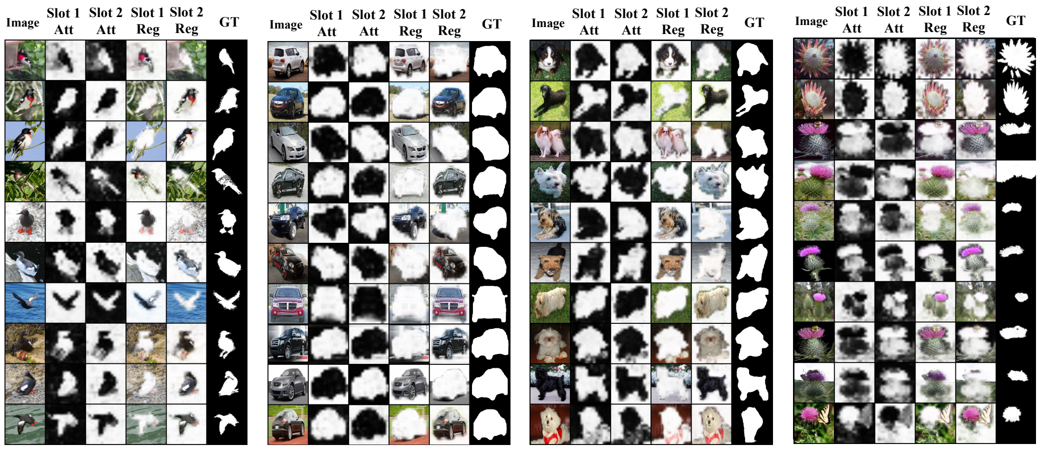
In the task, each pixel of an image has to be classified into one of the pre-defined object categories. To obtain the category labels for each predicted mask, we perform the following inference steps. (a) For each image, we obtain a set of object-centric representations $\hat{s}$. (b) We compute all the object features by taking matrix multiplication between $M_{\hat{s},x}$ and x from all the images and then perform k-means clustering on these feature vectors, in which the number of clusters is the number of semantic categories of the benchmark. (c) A binary matching algorithm is used to match our clustered categories with the ground truth by maximizing mIoU. (d) Each pixel on a given image can be assigned to the predicted class label corresponding to the binded slot basis.

S3 Additional Visualization and Failure Cases

We provide additional visualization results of the unsupervised foreground extraction task on the Birds, Cars, Dogs and Flowers datasets in Figure S2. The same result analysis from Section 5.1 can be applied here.

In the unsupervised object discovery task, we provide additional positive samples of the predicted object masks in Figure S3 and negative samples in Figure S4. As discussed in Section 5.2, our method consistently predicts semantic regions despite zero annotated ground truth masks provided during training.

However, we also notice that our method as well as other methods are not perfect in some cases, especially when the number of slot bases is larger than the number of semantic objects in the scene.



(a) Birds.

(b) Cars.

(c) Dogs.

(d) Flowers

Figure S2: Additional visualization of the predicted foreground masks in the unsupervised foreground extraction task on (a) Birds, (b) Cars, (c) Dogs and (d) Flowers datasets. The design conventions follow Figure 3(a).

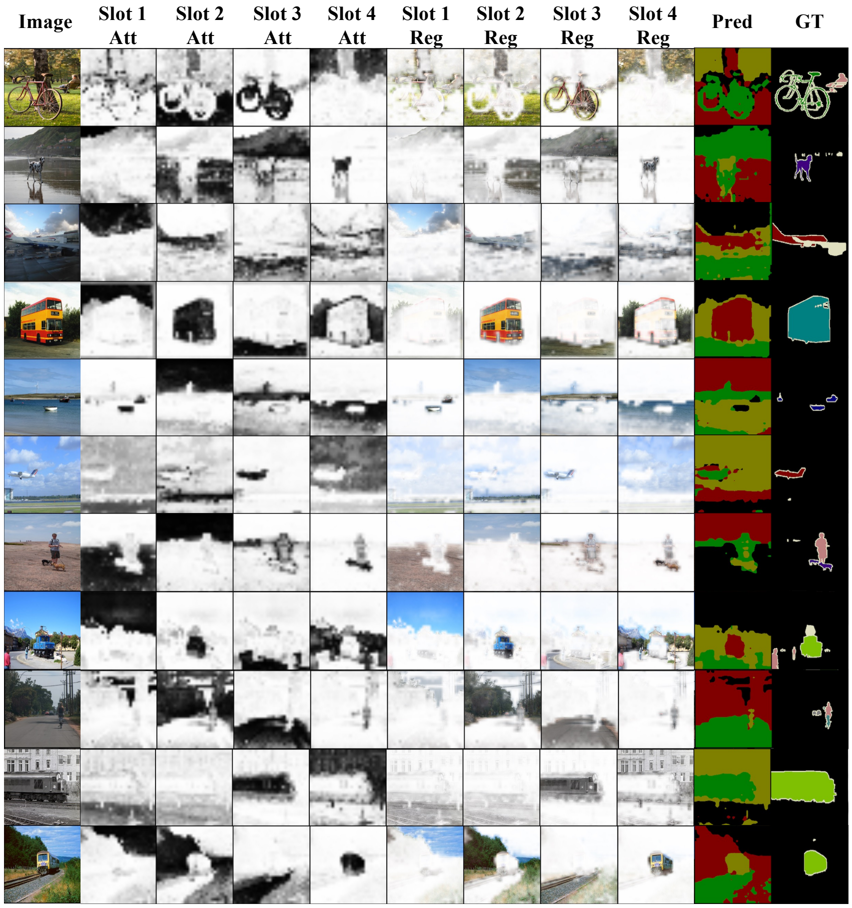


Figure S3: Additional positive samples of the predicted object masks in the unsupervised object discovery task on Pascal VOC 2012. The design conventions follow Figure 3(b).

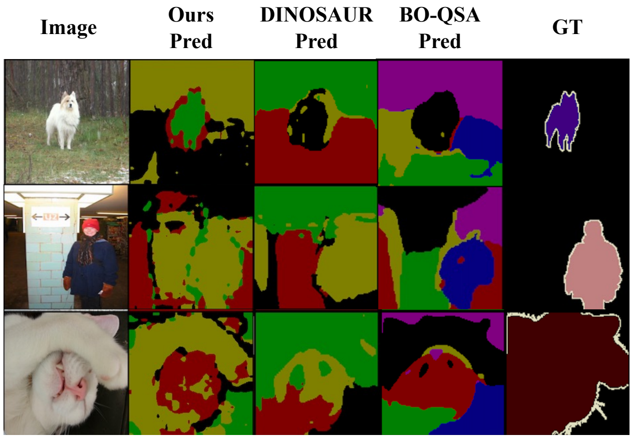


Figure S4: Negative samples of the predicted object masks in the unsupervised object discovery task on Pascal VOC 2012 dataset. From left to right, we show the original image (Col.1), the 4 predicted masks of our cyclic walks corresponding to the 4 slots (Col.2), the 4 masks for DINOSAUR (4 slots, Col.3), the 6 masks for BO-QSA (6 slots, Col.4), and the ground truth object masks (Col.5). Each color indicates a predicted mask from a slot (Col.2-4).

For example, in Row 1 of Figure S4, our method correctly segments trees, dog, and grass, but incorrectly segments the edges of the dogs. In contrast, other methods output completely random segmented masks, carrying no semantic information whatsoever. Given that our method produces more “reasonable" negatively segmented masks compared with other methods, this suggests that the slot bases trained with our contrastive walks are capable of capturing distinct semantic representations and meanwhile, taking into account the holistic view of the scene.

References

[1] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9630–9640. IEEE, 2021.

[2] B. Jia, Y. Liu, and S. Huang. Unsupervised object-centric learning with bi-level optimized query slot attention. *CoRR*, abs/2210.08990, 2022.

[3] I. Loshchilov and F. Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017.

[4] M. Seitzer, M. Horn, A. Zadaianchuk, D. Zietlow, T. Xiao, C. Simon-Gabriel, T. He, Z. Zhang, B. Schölkopf, T. Brox, and F. Locatello. Bridging the gap to real-world object-centric learning. *CoRR*, abs/2209.14860, 2022.