RETHINKING OBJECT-CENTRIC REPRESENTATIONS IN THE ERA OF FOUNDATIONAL SEGMENTATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Object-centric learning (OCL) aims to represent each object's information independently and minimize interference from backgrounds and other objects. OCL is expected to aid model generalization, especially in out-of-distribution (OOD) settings. However, the community's effort has been focused on improving unsupervised entity segmentation performances which is secondary to the main objective. We challenge this. We argue that segmentation is no longer the main barrier: recent class-agnostic segmentation methods reliably localize objects in a zero-shot manner. Instead, we advocate for a renewed emphasis on how decomposed representations can improve OOD generalization. As a first step, we propose Object-Centric Classification with Applied Masks (OCCAM) that exploits discovered objects to extract their representations for downstream classification tasks. Our experiments on datasets with background spurious correlations suggest that even in this task OCL representations do not lead to better generalization than object-centric representations provided by foundational segmentation models. These results showcase the importance of recognizing advances in zero-shot image segmentation when high-performant object-centric representations are the end goal. In addition to that, we suggest exploring new benchmarks for OCL methods evaluation that better reflect the problems these methods are designed to solve and highlight scenarios where OCL methods are more favorable solutions than foundational segmentation models.

1 INTRODUCTION

Object-centric learning (OCL) seeks to learn image representations where each object is encoded independently. The OCL community posits that learning object-centric representations is key for out-of-distribution generalization (Dittadi et al., 2022; Arefin et al., 2024) because it aligns with causal mechanisms (Schölkopf et al., 2021), human cognition (Spelke, 1990; Téglás et al., 2011; Wagemans, 2015) and leverages the compositional nature of scenes (Greff et al., 2020). Some works provided evidence that OCL methods can express compositional generalization (Wiedemer et al., 2024), and help in out-of-distribution (OOD) segmentation (Dittadi et al., 2022) and classification (Arefin et al., 2024). Since direct object-centric representations' quality evaluation is not straightforward, most research has focused on a proxy task - unsupervised image segmentation, where segmentation and matching metrics serve as the primary benchmarks (Locatello et al., 2020; Jiang et al., 2023; Seitzer et al., 2023; Didolkar et al., 2025; Kipf et al., 2022; Elsayed et al., 2022).

In this paper, we question whether the OCL community should continue prioritizing unsupervised segmentation as a core research goal. Recent advances in class-agnostic segmentation models like CropFormer (Lu et al., 2023) have demonstrated remarkable zero-shot object discovery, let alone large-scale foundational models such as Segment Anything (SAM) (Kirillov et al., 2023; Ravi et al., 2025). Identifying objects in an image is not a critical obstacle it once was.

In search of a different task for tracking the progress of OCL methods while having them as competitive baselines, we use OOD image classification as a new benchmark. Conceptually, this task can benefit from OCL methods by capitalizing on the appealing properties of object-centric representations such as disentanglement of different objects' features. To utilize object-centric representations in this task, we introduce **Object-Centric Classification with Applied Masks** (**OCCAM**) that leverages them to boost the performance of zero-shot image classifiers. Our pipeline is described as follows. The first module generates object-centric representations by combining images with objects' masks provided by the underlying mask generator, e.g. an OCL method. The second module selects representations of foreground objects and uses them for image classification, thus making predictions independent of such common sources of spurious correlations as backgrounds.

To measure the performance of the proposed pipeline on OOD test cases we focus on datasets containing **background spurious correlations**, occurring when the central cue for recognition lies not on the foreground object but on the background or other objects. For this purpose, we use spurious background datasets.

To support our claim that unsupervised object discovery might not be a suitable benchmark for OCL methods we show that zero-shot segmentation models significantly outperform the zero-shot object discovery capabilities of existing OCL approaches. Then we leverage these zero-shot capabilities to generate object-centric representations and achieve state-of-the-art results in several robust image classification benchmarks with OCCAM. Our results indicate that foundational segmentation models are more performant alternatives to OCL even for building such robust recognition systems. Such dominance of foundational segmentation models suggests that they should be considered as a strong baseline for building object-centric representations. Our results also point to the need to develop a new benchmark that better reflects the promises and capabilities of OCL, such as the simulation of the development of human cognition (Spelke, 1990; Téglás et al., 2011; Wagemans, 2015) and the causal reasoning capabilities in models (Schölkopf et al., 2021).

2 Related work

Object-centric learning Object-centric learning (OCL) captured widespread attention with the introduction of SlotAttention (Locatello et al., 2020), which enabled iterative learning of separate latent representations for each object in an image. These latent "slots" can then be decoded back to the pixel space. Extensions have included SlotAttention paired with diffusion decoders (Jiang et al., 2023) and SlotAttention architectures built on top of Dino (Seitzer et al., 2023; Didolkar et al., 2025) features. In video contexts, sequential adaptations leverage temporal dependencies (Kipf et al., 2022) and depth information (Elsayed et al., 2022). Some studies also propose theoretical foundations for OCL (Wiedemer et al., 2024; Brady et al., 2023). Although OCL promises a more principled approach of object classification by separating foreground objects from background elements — often the source of spurious correlations (Sagawa et al., 2020; Li et al., 2023; Xiao et al., 2021; Zhang et al., 2024; Wang et al., 2024) — much of the field's progress is tracked by unsupervised object discovery metrics, essentially entity segmentation (Locatello et al., 2020; Jiang et al., 2023; Seitzer et al., 2023; Didolkar et al., 2025; Kipf et al., 2022; Elsayed et al., 2022). Specialized segmentation models already do well on these benchmarks (Table 1 and Figure 2), suggesting that the primary goal of OCL is merely to rediscovering class-agnostic segmentation, but with vastly inferior performance. Furthermore, empirical support for the commonly claimed benefits—such as parameter/learning efficiency (Kipf et al., 2022) and improved generalization (Dittadi et al., 2022; Arefin et al., 2024) or better understanding of representations — remains limited.

For the extended list of related works on foundational segmentation models, spurious correlations, and methods to use object masks, please see § D.

3 Method

This section gives an overview of our proposed method. Subsection §3.1 defines the notation needed for the method description in §3.2.

3.1 Notations

We denote an image as $x \in \mathbb{R}^{[3,\text{ H, W}]}$ and a label as $y \in \mathcal{Y} = \{1, \dots, C\}$, where C is the number of classes. We will write an image encoder, or a feature extractor, as ψ and image embedding, or feature vector, as $\psi(x) \in \mathbb{R}^d$, where $d \geq 1$ is the feature dimensionality. We define the classifier's pre-softmax logits as $f(\psi(x)) \in \mathbb{R}^{|\mathcal{Y}|}$ and softmax probabilities as $p(\psi(x)) = \text{Softmax}(f(\psi(x))) \in$

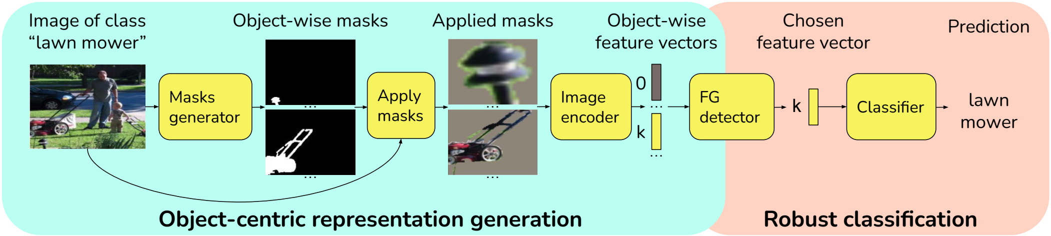

Figure 1: An overview of Object-Centric Classification with Applied Masks (OCCAM). It consists of two main parts. The first part described in § 3.2.1 uses entity segmentation masks for object-centric representation generation. The second part described in § 3.2.2 selects representations corresponding to the foreground object and uses them for classification.

$[0, 1]^{|\mathcal{Y}|}$. We also denote indices for the last two dimensions in tensors as superscripts (e.g. last two dimensions of sizes H, W for x) and all other dimensions as subscripts (e.g. first dimension of size 3 in x). We will use shorthands "FG" and "BG" for foreground and background correspondingly.

3.2 Method

Our Object-Centric Classification with Applied Masks (OCCAM) pipeline is summarized in Figure 1. We use object-centric representations to reduce spurious correlations in image classification. It consists of the two main parts: 1. generate object-centric representations, 2. perform robust classification by classifying an image using only representations of the foreground object. In the following subsections, we will explain these parts in more detail.

3.2.1 Generating object-centric representations

To generate the object-centric representations we first generate masks for all objects and backgrounds in the image using a mask predictor. We then apply generated masks to images by combining masks with images. Each object is then encoded with an image encoder.

Generating masks To produce object representations given an original image $x \in \mathbb{R}^{[3,\text{ H, W}]}$, we generate a set of masks for all the foreground objects and the background. That is done with the help of mask predictor S , which takes x as input and assigns each pixel in x to one of $K_{\max}$ masks. The output of this model is the stack of K binary masks:

$$S_i^{kl}(x) = \mathbf{1}(\text{pixel } x^{kl} \text{ is assigned to mask } i)$$

where $\mathbf{1}(\cdot)$ is the indicator function.

OCL method like FT-Dinosaur (Didolkar et al., 2025) or an external segmentation model like CropFormer (Lu et al., 2023) can be used as masks generator in this pipeline.

Applying masks After producing the binary masks for each object, we segregate the pixel contents for each mask by "applying mask" on the input image. It can be any operation involving an image x and mask m: $a(x, m) \in \mathbb{R}^{[3,\ H,\ W]}, \forall m \in \{S_i, i = 1 \dots K\}, m \in \mathbb{R}^{[H,\ W]}$.

Masks can for example be applied by cropping an image by the mask contours and resizing the result to the full image size. For details of applying masks please see § C.

Encoding applied masks To get the final object-centric representations we encode applied masks by an image encoder ψ and get the following set of object-wise feature vectors:

$$\psi(a(x, S_i(x))) \in \mathbb{R}^d, i = 1 \dots K$$

ViT (Dosovitskiy et al., 2021) can be used as an image encoder for example.

3.2.2 ROBUST CLASSIFIER

We hypothesize that by isolating foreground object representations from the representations of background and other objects we eliminate sources of spurious correlations, hence performing more robust classification. For that reason, we first use the set of object-centric representations obtained in the previous stage to select the single representation that corresponds to the foreground. Then we provide the selected foreground representation to the classifier to make the final prediction.

FG detector After applying masks to the image we select the mask that corresponds to the foreground object by the following process. At first, we compute the *foreground score* $g(x, m) \in [0, 1]$ for all applied masks. This score reflects how likely a given applied mask is to correspond to the foreground object. Then we take the mask with the highest foreground score among all masks for the current image:

$$m^\star = S_{i^\star}(x),$$

where $i^\star = \operatorname{argmax}_i \; g(x, S_i(x))$.

Maximum output probability can be used as a foreground score for example. Details on the foreground scores can be seen in § C.

Image classification using foreground object representations Finally, once we have identified the mask that matches the foreground object, we apply it to the original image and classify the result of this operation. The final output of our method is:

$$\text{OCCAM}(x) = p(\psi(a(x, m^\star))).$$

4 EXPERIMENTS

In this section, we empirically validate the ability of object-centric learning (OCL) approaches to result in robust object classification capabilities by comparing against the viable baseline of using mask predictions from foundational segmentation models, using the OCCAM pipeline in § 3. Throughout the section, we also compare against the state-of-the-art generalization results in each benchmark, noticing that new state-of-the-art results can be achieved by combining CLIP models with OCCAM.

4.1 CASE FOR MOVING BEYOND OBJECT DISCOVERY

In this section, we compare zero-shot (class-agnostic) segmentation methods CropFormer (Lu et al., 2023) and SAM (Kirillov et al., 2023), with the state-of-the-art object-centric learning methods (Jiang et al., 2023; Seitzer et al., 2023; Didolkar et al., 2025) in Table 1 and Figure 2. The goal of these experiments is to see whether, in the presence of foundational segmentation models, OCL methods remain reasonable solutions to the object discovery task (Locatello et al., 2020).

4.1.1 SETUP

Datasets Following the prior work (Kipf et al., 2022; Elsayed et al., 2022; Seitzer et al., 2023; Didolkar et al., 2025) we use two synthetic image datasets Movi-C and Movi-E (Greff et al., 2022). They both contain images of around 1000 realistic 3D-scanned objects placed on HD backgrounds. Movi-C includes 3 to 10 objects per scene. Movi-E includes 11 to 23 objects per scene.

Metrics We evaluate object representation quality by comparing masks predicted for objects with ground truth instance masks. To do so, we use foreground adjusted random index (FG-ARI) (Rand, 1971; Hubert & Arabie, 1985; Kipf et al., 2022) and mean best overlap (mBO) (Pont-Tuset et al., 2015; Seitzer et al., 2023), please see § F.1 for details.

Baselines We compare segmentation models to three recent state-of-the-art methods for object-centric learning in real-world settings: SlotDiffusion (Jiang et al., 2023), Dinosaur (Seitzer et al., 2023), and FT-Dinosaur (Didolkar et al., 2025). Please see § F.1 for details.

	Pre-training Datasets		FT	Movi-C		Movi-E	
	Encoder	Decoder		FG-ARI	mBO	FG-ARI	mBO
Slot Diffusion (Jiang et al., 2023)	OpenImages (1.9M)	COCO (118k)	✗	66.9	43.6	67.6	26.4
Dinosaur (Seitzer et al., 2023)	GLD (1.2M)	COCO (118k)	✗	67.0	34.5	71.1	24.2
FT-Dinosaur (Didolkar et al., 2025)	GLD (1.2M)	COCO (118k)	✓	73.3	44.2	71.1	29.9
CropFormer (Lu et al., 2023) (Ours)	COCO (118k) + EntitySeg (33k)		✗	79.3	65.4	**87.2**	63.8
SAM (Kirillov et al., 2023)	SA-1b (11M)		✗	**79.7**	**73.5**	84.7	**69.7**

Table 1: Quantitative results for object discovery on Movi-C and Movi-E; column “FT” has ✓ if the model was fine-tuned on the training split of the corresponding dataset (Movi-C or Movi-E) and ✗ otherwise.

Method	Masks source	Worst Group Acc
CLIP (Radford et al., 2021)	-	87.2
OCCAM (Ours)	FT-Dinosaur	99.2
OCCAM (Ours)	CropFormer	**100.0**
CoBalT (Arefin et al., 2024)	-	80.0
LLE (Li et al., 2023)	-	90.8*

Table 2: Worst group accuracies on UrbanCars (Li et al., 2023) for ViT-L-14 CLIP model. $\star$ indicates state-of-the-art results.

4.1.2 Results

Factual results As can be seen in Table 1, SAM achieves significantly higher mBO scores than the best object-centric method: 73.5% vs 44.2% on Movi-C and 69.7% vs 29.9% on Movi-E. For FG-ARI, CropFormer-EntitySeg shows large improvement over all methods including SAM, with scores gap of 87.2% vs 71.1% on Movi-E when compared to object-centric baselines. The qualitative results in Figure 2 also confirm the superior performance of segmentation models.

Conclusion. Segmentation models pre-trained with supervision for segmentation tasks and used in a zero-shot manner on new test datasets outperform object-centric methods fine-tuned for these datasets.

4.2 Spurious Correlations Mitigation

The goal of the experiments in this section is to show that object masks help mitigate spurious correlations in the image classification task (Tables 2- 7).

4.2.1 Setup

Datasets We use several popular datasets with spurious backgrounds or co-occuring objects: UrbanCars (Li et al., 2023), ImageNet-D (Zhang et al., 2024), ImageNet-9 (Xiao et al., 2021), Waterbirds (Sagawa et al., 2020) and CounterAnimals (Wang et al., 2024). Please see § B for more details.

Metrics To measure model performance we use CLIP (Radford et al., 2021) zero-shot accuracy.

Baselines For all the experiments in this section we use CLIP model with masked self-attention Transformer (Vaswani et al., 2017) text encoder and ViT-L/14 (Dosovitskiy et al., 2021) image encoder if not stated otherwise.

We use foundational segmentation model CropFormer (Lu et al., 2023) and the state-of-the-art OCL method FT-Dinosaur (Didolkar et al., 2025) for masks prediction.

4.2.2 Results

Factual results Masks usage allows to significantly improve performance on all the considered datasets, sometimes even scoring 100% accuracy, e.g. on UrbanCars (Table 2). However, for

Method	Masks source	Worst Group Acc
CLIP (Radford et al., 2021)	-	83.2
OCCAM (Ours)	FT-Dinosaur	91.6
OCCAM (Ours)	CropFormer	**96.9**
CoBalT (Arefin et al., 2024)	-	90.6
DFR (Kirichenko et al., 2023)	-	91.8⋆

Table 3: Accuracies on Waterbirds (Sagawa et al., 2020) for ViT-L-14 CLIP model. ⋆ indicates state-of-the-art results

Method	Masks source	Accuracy
CLIP (Radford et al., 2021)	-	23.6
OCCAM (Ours)	FT-Dinosaur	56.9
OCCAM (Ours)	CropFormer	**69.7**
LLAVA 1.5 (Liu et al., 2023)	-	73.3⋆

Table 4: Accuracies on ImageNet-D "background" subset (Zhang et al., 2024) for ViT-L-14 CLIP model. ⋆ indicates state-of-the-art results.

max_prob foreground detection method accuracy gains are much smaller, and sometimes performance even drops (Table 8). That highlights a weak spot of the baseline foreground detection method and leaves a vast amount of space for improvement that can fuel future research.

Speaking about masks sources - foundational segmentation model CropFormer provided more useful masks for spurious correlations mitigation in comparison to FT-Dinosaur on all datasets across the board. For example, on ImageNet-D CropFormer allows to achieve near LLAVA 1.5 (Liu et al., 2023) level of accuracy - 69.7% vs 56.9% for FT-Dinosaur masks (Table 4).

Note: The final results are obtained using the best-performing combination of parameters for each test dataset (see detailed analysis in § E).

Conclusion The experiments demonstrate that the quality of mask-based object-centric representations is already high enough to significantly improve performance on all the considered datasets, and even completely solve some of them. However, to be able to use these masks to the full extent one needs a foreground detector that performs on par with g_{oracle} (see § 3.2.2) but does not use ground truth labels.

In addition to that, our proposed framework allowed us to discover that contrary to the authors' hypothesis (Wang et al., 2024) uncommon backgrounds are not the main cause for performance drop between "Common" and "Counter" subsets of CounterAnimals dataset (See Appendix A).

5 Conclusion

Starting from the most popular downstream application for OCL methods, we show that in unsupervised segmentation, these methods are outperformed by foundational segmentation models utilized in zero-shot settings.

With this sober look, we encourage the OCL community to shift focus from unsupervised segmentation to the application that capitalizes on its appealing properties of learning separate representations for separate objects in the image, we propose another downstream task for OCL methods evaluation - spurious background correlations mitigation with the use of object-centric masks. We observe that while OCL representations improve performance in this task, foundational segmentation models provide object-centric representations that lead to even greater performance improvements.

We believe that in the era of foundational segmentation models, works in OCL that aim to improve object-centric representations' quality should either consider this strong baseline in their experiments or justify the assumption that there is no access to the privileged information in the form of foundational segmentation models, which can be downloaded from the internet.

In addition to that, we encourage the community to develop benchmarks that reflect the goals of OCL. These include discovering causal mechanisms (Schölkopf et al., 2021) in model predictions and aligning model predictions with human cognition (Spelke, 1990; Téglás et al., 2011).

References

Md Rifat Arefin, Yan Zhang, Aristide Baratin, Francesco Locatello, Irina Rish, Dianbo Liu, and Kenji Kawaguchi. Unsupervised concept discovery mitigates spurious correlations. In *International Conference on Machine Learning (ICML)*, 2024.

Jack Brady, Roland S. Zimmermann, Yash Sharma, Bernhard Schölkopf, and Wieland and von Kügelgen, Julius Brendel. Provably learning object-centric representations. In *International Conference on Machine Learning (ICML)*, 2023.

Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *International Conference on Computer Vision (ICCV)*, 2021.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022.

Aniket Didolkar, Andrii Zadaianchuk, Anirudh Goyal, Mike Mozer, Yoshua Bengio, Georg Martius, and Maximilian Seitzer. On the transfer of object-centric representation learning. In *International Conference on Learning Representations (ICLR)*, 2025.

Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary universal image segmentation with maskclip. In *International Conference on Machine Learning (ICML)*, 2023.

Andrea Dittadi, Samuele S Papa, Michele De Vita, Bernhard Schölkopf, Ole Winther, and Francesco Locatello. Generalization and robustness implications in object-centric learning. In *International Conference on Machine Learning (ICML)*, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Gamaleldin F. Elsayed, Aravindh Mahendran, Sjoerd van Steenkiste, Klaus Greff, Michael C. Mozer, and Thomas Kipf. SAVi++: Towards end-to-end object-centric learning from real-world videos. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*, 2020.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: a scalable dataset generator. 2022.

Sebastian Gruber and Florian Buettner. Uncertainty estimates of predictions via a general bias-variance decomposition. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Lawrence J. Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 1985.

Jindong Jiang, Fei Deng, Gautam Singh, and Sungjin Ahn. Object-centric slot diffusion. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Thomas Kipf, Gamaleldin F. Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional Object-Centric Learning from Video. In *International Conference on Learning Representations (ICLR)*, 2022.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations (ICLR)*, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *International Conference on Computer Vision (ICCV)*, 2023.

Zhiheng Li, Ivan Evtimov, Albert Gordo, Caner Hazirbas, Tal Hassner, Cristian Canton Ferrer, Chenliang Xu, and Mark Ibrahim. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *European Conference on Computer Vision (ECCV)*, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Qi Lu, Jason Kuen, Shen Tiancheng, Gu Jiuxiang, Guo Weidong, Jia Jiaya, Lin Zhe, and Yang Ming-Hsuan. High-quality entity segmentation. In *International Conference on Computer Vision (ICCV)*, 2023.

Jishnu Mukhoti, Andreas Kirsch, Joost R. van Amersfoort, Philip H. S. Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *International Conference on Learning Representations (ICLR)*, 2024.

Jordi Pont-Tuset, Pablo Arbeláez, Jonathan T. Barron, Ferran Marqués, and Jitendra Malik. Multiscale combinatorial grouping for image segmentation and object proposal generation. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2015.

Lu Qi, Jason Kuen, Yi Wang, Jiuxiang Gu, Hengshuang Zhao, Philip Torr, Zhe Lin, and Jiaya Jia. Open World Entity Segmentation . *Transactions on Pattern Analysis & Machine Intelligence (TPAMI)*, 2023. doi: 10.1109/TPAMI.2022.3227513.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 1971.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations (ICLR)*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and BjÃ¶rn Ommer. High-resolution image synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 2015.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.

B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 2021. doi: 10.1109/JPROC.2021.3058954.

Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, and Francesco Locatello. Bridging the gap to real-world object-centric learning. In *International Conference on Learning Representations (ICLR)*, 2023.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *International Conference on Computer Vision (ICCV)*, 2023.

Elizabeth S. Spelke. Principles of object perception. *Cognitive Science*, 1990. doi: https://doi.org/10.1016/0364-0213(90)90025-R.

Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-clip: A clip model focusing on wherever you want. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Saeid Asgari Taghanaki, Kristy Choi, Amir Hosein Khasahmadi, and Anirudh Goyal. Robust representation learning via perceptual similarity metrics. In *International Conference on Machine Learning (ICML)*, 2021.

Ernő Téglás, Edward Vul, Vittorio Girotto, Michel Gonzalez, Joshua B. Tenenbaum, and Luca L. Bonatti. Pure reasoning in 12-month-old infants as probabilistic inference. *Science*, 2011. doi: 10.1126/science.1196404.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

Johan Wagemans. *The Oxford Handbook of Perceptual Organization*. Oxford University Press, 2015. doi: 10.1093/oxfordhb/9780199686858.001.0001.

Qizhou Wang, Yong Lin, Yongqiang Chen, Ludwig Schmidt, Bo Han, and Tong Zhang. A sober look at the robustness of clips to spurious features. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Thaddäus Wiedemer, Jack Brady, Alexander Panfilov, Attila Juhos, Matthias Bethge, and Wieland Brendel. Provable compositional generalization for object-centric learning. In *International Conference on Learning Representations (ICLR)*, 2024.

Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations(ICLR)*, 2021.

Chenshuang Zhang, Fei Pan, Junmo Kim, In So Kweon, and Chengzhi Mao. Imagenet-d: Benchmarking neural network robustness on diffusion synthetic object. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision (ECCV)*, 2022.

A CounterAnimals: Spurious or simply Hard?

We have demonstrated that our object-centric classification pipeline is able to in-principle isolate the contribution of the object independent of the background, we now investigate the recently proposed CounterAnimal dataset (NeurIPS'24).

Dataset description The goal of the CounterAnimals dataset (Wang et al., 2024) is to highlight models' reliance on spurious backgrounds. This dataset consists of the two splits of realistic images taken from iNaturalist[1] dataset. The first split named "Common" contains images of animals on the most frequently co-occurring backgrounds for example a polar bear on the snow background. The second split named "Counter" contains animals on non-standard backgrounds for example a polar bear on dirt. This dataset includes 45 animal classes. Each of them is part of the ImageNet1k (Russakovsky et al., 2015) class set.

The authors evaluate more than 20 different image classifiers of various architectures and pre-training techniques and observe that they all perform better on the "Common" split than on the "Counter" split. Therefore, the authors reasonably hypothesize that it happens because the Common split in contrast to the Counter split contains backgrounds that often correlate with ground truth labels.

Counterintuitive findings Our results in Table 7 contradict the hypothesis that the "Counter" subset is harder due to uncommon backgrounds. We show that even when backgrounds are removed, and only main objects are classified, still, there is a significant gap between performance on "Common" and "Counter" datasets. For example, the CLIP enhanced with CropFormer masks used with oracle foreground selection and "crop + resize" image combining achieves 66.0% on the "Common" subset but only 55.7% on the "Counter" subset as can be seen in Table 8. This is despite the fact that the background has no influence on predictions as its pixels are replaced by gray pixels. This gap between performance on "Common" and "Counter" subsets remains when classifying only foreground objects for all the CLIP models originally considered by the CounterAnimals authors (Table 5). Even for the best performing ViT-L-14-quickgelu model gap does not change much when classifying only foreground images: 10.2% vs 8.3%.

That suggests that the "Counter" subset is harder because of shifts in the distribution of foreground objects rather than backgrounds. These shifts include occlusions and uncommon camera perspectives.

We believe that the ability of our framework to make such counterintuitive discoveries makes it a valuable tool for new spurious background dataset development.

Note: The fact that absolute accuracies drop when classifying only foreground objects (Cmn/Ctr accuracies are higher than corresponding Cmn-FG/Ctr-FG accuracies in Table 5), e.g. $90.8\%/80.6\%$ vs $78.2\%/69.9\%$ for ViT-L-14-quickgelu model suggests that by isolating foreground objects OC-CAM removes some predictive signal contained in background which can be useful for both "Common" and "Counter" subsets.

[1] https://www.inaturalist.org/observations

Model	Pre-Train-Dataset	Cmn	Ctr	Cmn-FG	Ctr-FG	Gap	Gap-FG
RN50	openai	64.1	40.6	39.8	31.6	23.6	8.3
RN101	openai	65.9	46.5	42.1	33.5	19.3	8.6
RN50x4	openai	71.9	50.8	46.6	38.5	21.1	8.1
RN50x16	openai	78.2	60.4	51.4	41.9	17.7	9.4
RN50x64	openai	80.5	68.5	64.4	52.8	12.0	11.6
ViT-B/32	openai	69.3	45.7	47.8	39.2	23.6	8.6
ViT-B/16	openai	73.4	56.8	53.5	42.6	16.6	10.9
ViT-L/14	openai	85.3	70.4	66.3	56.5	14.9	9.8
ViT-L/14@336px	openai	86.2	73.3	69.2	58.8	12.9	10.4
ViT-B-16	laion400m	73.7	53.6	50.9	41.7	20.2	9.2
ViT-B-16	datacomp	66.1	45.8	46.6	38.0	20.3	8.5
ViT-B-16	laion2b	73.4	53.2	52.3	43.4	20.3	8.9
ViT-B-16	dfn2b	84.7	70.3	67.7	59.0	14.4	8.7
ViT-B-32	laion400m	67.5	38.1	40.9	31.0	29.4	9.9
ViT-B-32	laion2b	73.2	49.1	52.7	40.5	24.1	12.2
ViT-B-32-256	datacomp	80.9	61.6	67.4	55.2	19.4	12.2
ViT-L-14	laion400m	81.7	64.1	61.8	50.1	17.7	11.7
ViT-L-14	datacomp	89.3	79.9	78.3	69.5	9.3	8.8
ViT-L-14	laion2b	82.4	66.6	65.8	54.6	15.8	11.2
ViT-L-14-quickgelu	dfn2b	90.8	80.6	78.2	69.9	10.2	8.3
ViT-H-14	laion2b	85.8	73.5	69.7	61.9	12.3	7.8
ViT-H-14-quickgelu	dfn5b	88.5	79.1	82.7	75.2	9.4	7.5
ViT-H-14-378-quickgelu	dfn5b	90.4	84.1	85.3	78.6	6.3	6.7
ViT-G-14	laion2b	87.5	73.7	71.4	61.6	13.8	9.8
convnext_base	laion400m	74.7	51.8	51.7	40.7	22.9	11.0
convnext_base_w	laion2b	77.3	56.4	58.6	46.9	20.9	11.7

Table 5: Accuracies on CounterAnimals dataset for different CLIP models. Ctr/Cmn stands for "Counter"/"Common" subsets correspondingly; FG stands for accuracy while using OCCAM with CropFormer masks and oracle foreground selection method. Gap(-FG) columns store differences: [Cmn(-FG) - Ctr(-FG)] up to rounding error.

B Extended spurious backgrounds results

In this section, we provide results on ImageNet-9 (Table 6), and CounterAnimals (Table 7) datasets in addition to UrbanCars (Table 2), Waterbirds (Table 3), and ImageNet-D (Table 4) results mentioned in the main paper.

Below we provide details on these datsets (for details on CounterAnimals dataset please see § A):

The core of the datasets consists of the popular choices for robust image classification tasks: UrbanCars (Li et al., 2023), Waterbirds (Sagawa et al., 2020), and ImageNet-9 (Xiao et al., 2021). We also add ImageNet-D (Zhang et al., 2024) dataset to this group as we believe that it contains more realistic images thanks to blending objects with backgrounds using diffusion model (Rombach et al., 2022) instead of cropping foreground objects onto new backgrounds as done in the previous datasets. Finally, we use the CounterAnimals (Wang et al., 2024) dataset, the latest benchmark that contains natural images which are spurious background correlations, specifically designed even for CLIP models.

1. UrbanCars (Li et al., 2023): a dataset for binary classification of cars into “urban” and “country” types. Each image contains one urban or country car paired with either urban or country secondary objects (e.g. fire hydrant for urban or cow for country) on either urban or country backgrounds. It is synthetically generated by placing cut-out cars paired with cut-out secondary objects on urban or rural backgrounds.
2. ImageNet-D (Zhang et al., 2024): a diffusion-model-synthesized dataset for image classification for 113 classes. Its class set is a subset of ImageNet-1k (Russakovsky et al., 2015) class set. We use “background” subset of this dataset which features objects appearing with uncommon backgrounds (e.g. plates in a swimming pool).
3. ImageNet-9 (Xiao et al., 2021): synthetically generated dataset for image classification for 9 classes. It has 9 classes which are supersets of ImageNet classes (e.g. dog, bird etc). We use ”mixed random” subset of this dataset which is generated by placing cut-out objects of one class on backgrounds from images of some other random class.
4. Waterbirds (Sagawa et al., 2020): a dataset for binary classification of birds into “land” and “sea” types. Each image contains land or sea birds on either land or sea backgrounds. It is synthetically generated by placing cut-out birds on land or sea backgrounds.

Method	Masks source	Accuracy
CLIP (Radford et al., 2021)	-	91.9
OCCAM (Ours)	FT-Dinosaur	93.8
OCCAM (Ours)	CropFormer	**96.4**
CoBalT (Arefin et al., 2024)	-	80.3
CIM (Taghanaki et al., 2021)	-	81.1*

Table 6: Accuracies on ImageNet-9 (Xiao et al., 2021) ”mixed rand” subset for ViT-L-14 CLIP model. $\star$ indicates state-of-the-art results

C Implementation details

Foundational segmentation model choice While in general CropFormer and SAM perform similarly on the segmentation task, SAM is much better in the mBO metric. Nevertheless, for the rest of our experiments, we use CropFormer. This choice is justified because we know the exact data it was trained on, and therefore, can confirm it was not trained on images from the spurious correlations datasets on which we test our method in the following section.

Applying masks One way to combine masks and images is to make a crop from the original image that follows the mask contours and then resize it to the original size. This approach follows the steps of masked dataset construction in (Sun et al., 2023). We call such an approach ”crop + resize”.

Method	Masks source	Common	Counter
CLIP (Radford et al., 2021)	-	85.1	70.7
OCCAM (Ours)	FT-Dinosaur	85.8	71.1
OCCAM (Ours)	CropFormer	**87.0**	**72.3**
CLIP-DFN-2B (Wang et al., 2024)	-	90.8*	80.6*

Table 7: Accuracy on CounterAnimals (Wang et al., 2024) for ViT-L-14 CLIP model. $\star$ indicates state-of-the-art results

Another way is to combine masks and images using AlphaClip (Sun et al., 2023), a modified version of CLIP (Radford et al., 2021) model. It adapts the original CLIP model to using masks as a source of additional information. AlphaClip model combines a mask and original image by appending masks as an additional α-channel to the image tensor resulting in RGB-A 4-dimensional tensor. We use checkpoints of this model fine-tuned on mask-grounded data generated from GRIT (Peng et al., 2024) and ImageNet (Russakovsky et al., 2015) datasets combined with proposals from SAM (Kirillov et al., 2023). We call this way of applying masks as "α-channel". Related work for this design choice can be seen in § D.

Foreground scores Currently, we use two types of foreground scores, both computed from the classifier's outputs:

1. **max_prob**: $g_{\text{max\_prob}}(x, m) = \max_c p^c(\psi(a(x, m)))$ - maximum probability across all possible classes (confidence).
2. **oracle**: $g_{\text{oracle}}(x, m) = p^y(\psi(a(x, m)))$ - probability of predicting a ground truth label. We consider this foreground score to measure the efficacy of the object-centric representation rather than to suggest it as a final method to use in practice. Although in reality, we do not have access to ground truth labels, it provides critical signals as to whether the insufficient generalization performance is due to object representation or due to foreground selection and classifier.

D Extended related work

Foundational segmentation models Because OCL methods are often evaluated using segmentation benchmarks, large-scale segmentation models with strong class-agnostic zero-shot performance (Kirillov et al., 2023; Ravi et al., 2025; Cheng et al., 2022; Lu et al., 2023) are particularly relevant. These foundation models were successfully scaled with Segment Anything (SAM) (Kirillov et al., 2023; Ravi et al., 2025), followed by CropFormer (Lu et al., 2023), Mask2Former (Cheng et al., 2022; Qi et al., 2023) In this work, we focus on CropFormer (Lu et al., 2023), which is trained on very limited images from COCO (Lin et al., 2014) and EntitySeg (Lu et al., 2023), yet still achieves competitive segmentation accuracy. CropFormer additionally ensures a large gap (and no overlap) between its training data and our test datasets, ensuring we are indeed measuring OOD generalizable performance.

Using object-centric representations to mitigate spurious correlations. Representations of only Foreground-object masks help improve robustness against backgrounds and co-occurring elements that may introduce spurious signals. This has been explored with various mechanisms, including (Li et al., 2023) uses an augmentation strategy where cropped objects form a background-agnostic branch in a last-layer ensemble (Kirichenko et al., 2023), enhancing generalization. Another strategy (Arefin et al., 2024) employs latent object-centric representations. In our experiments, both methods serve as baselines on relevant datasets. However, because they require fine-tuning, it is not a fair comparison to our zero-shot approach using object-centric representations to enhance CLIP's performance.

Combining masks and images After obtaining masks for foreground objects, we combine them with original images to create object-centric representations suitable for image classification.

To achieve this, masks can be combined with images by directly cropping objects out following the mask contours or highlighting objects corresponding to masks with red circles (Shtedritski et al., 2023).

However, such operations might result in out-of-distribution images for the image classifiers. Therefore, another line of work blends masks with images into a single representation for the CLIP's joint text-image representation space. Namely, AlphaCLIP (Sun et al., 2023) suggests adding projected masks as a fourth α-channel in an extended RGBA image tensor before processing it by image encoder. Another method developed for the attention-based image encoders adds mask's patch tokens to cross-attention layers (Ding et al., 2023) within standard CLIP encoders.

In this project, we focus on the AlphaCLIP approach, as it achieved the best image classification performance among CLIP representations blending approaches (Sun et al., 2023; Ding et al., 2023; Zhou et al., 2022; Shtedritski et al., 2023), and the cropping approach as we find it very intuitive.

Foreground masks selection By itself, having masks for all objects in an image is not enough to solve the classification task. It is also necessary to select the mask that corresponds to the object of interest. This selection depends on the task. The problem can be viewed as a reformulation of OOD detection (Mukhoti et al., 2021; Gruber & Buettner, 2022). Here, masked foreground objects are ID data, and other objects are OOD data. However, to the best of our knowledge, no papers have addressed this problem from this perspective.

Current methods select the foreground masks as the ones that maximize either intersection-over-union with ground truth bounding boxes (Li et al., 2023) or the classifier's ground truth probability (Sun et al., 2023). This project will use the latter as a baseline because while many images have ground truth class labels, fewer have ground truth bounding boxes. We will also use a relaxed version of this baseline that does not require access to the ground truth labels by selecting those masks that maximize maximum class probability (i.e. model's confidence).

E Factors influence on spurious correlations mitigation

E.1 Setup

Here we give a few additional comments to the setup described in § 4.2.1.

Foreground detector To compare different strategies for selecting foreground object masks, we varied the type of foreground detector between max_prob and oracle (see §3.2.2 for details) as indicated by corresponding words in ”FG det” columns. Since AlphaCLIP always expects a mask as part of the input we simulate scenarios when no masks are available by providing a mask covering the whole image as a foreground mask and saying that we use ”whole image” as a foreground detector in that case. However, technically we consider that as mask-free performance as no real masks are used in this case.

Image and mask combining Operation used for combining image and mask is indicated in ”Apply Mask” column. For the cases when we use ”α-channel” image and mask combining we utilize AlphaCLIP version of used CLIP model, described in § 3.2.1. It is fine-tuned for using masks and therefore is not identical to the original CLIP model.

Choosing background. For the ”crop + resize” way of combining masks and images, there is still one more question to answer: what to do with the empty spaces that are left after cropping and resizing the mask? To control for the amount of spurious correlations that come from the background we kept it either as in the original image (we call it “original” in “BG” columns) or substituted all its pixels with gray pixels (we call it “gray” in “BG” columns). By background here we mean all the pixels outside of the foreground mask provided to the model.

E.2 Results

E.2.1 Selecting foreground mask

For AlphaCLIP on most datasets using masks always improves performance with accuracy for max_prob foreground detector being higher than accuracy on clean images and accuracy for oracle foreground detector being much better than they both. For example, on ImageNet-D (Table 8) accuracy steadily grows from 23.6% for clean images to 38.8% for max_prob and 56.0% for oracle.

For ”crop + resize” mask and image combining the overall situation is similar to the ”α-channel” - using masks increases performance by small margin when using max_prob foreground detector and by a large margin when using the oracle foreground detector.

An exception to this can be seen on the ImageNet-9 dataset in Table 8 where the mask-free accuracy of 91.9% drops to 90.9% when using max_prob foreground detector while jumping to 96.4% for oracle.

Therefore, we used oracle foreground detection throughout our experiments.

E.2.2 Combining image and mask

On the majority of datasets using naive ”crop + resize” operation for combining masks and images results in better performance in comparison to ”α-channel” especially when using oracle masks. For instance, it scores 100% accuracy on UrbanCars dataset vs 96.8% for ”α-channel” (Table 8). Performance difference can be explained by the fact that AlphaCLIP fine-tunes original CLIP weights to adapt them for using masks which can harm the overall classification performance. That is why we fixed ”crop + resize” as the main method for combining masks and images throughout our experiments.

Masks source	FG det	Apply Mask	BG	Cmn	Ctr	WB	IN-9	IN-D	UC
-	whole image	α-channel	original	79.0	62.0	79.6	90.2	23.6	87.2
CropFormer	max_prob	α-channel	original	80.1	63.1	82.1	90.7	38.9	87.3
CropFormer	oracle	α-channel	original	87.0	72.3	91.0	93.9	56.0	96.8
-	-	-	original	85.1	70.7	83.2	91.9	17.8	84.0
CropFormer	max_prob	crop + resize	gray	58.9	49.0	74.8	90.9	48.7	94.4
CropFormer	oracle	crop + resize	gray	66.0	55.7	96.9	96.4	69.7	100.0
FT-Dinosaur	max_prob	α-channel	original	79.0	61.4	78.2	90.3	35.6	88.0
FT-Dinosaur	oracle	α-channel	original	85.8	71.1	86.8	93.4	49.3	95.2
FT-Dinosaur	max_prob	crop + resize	gray	62.3	49.2	63.7	85.3	47.1	88.8
FT-Dinosaur	oracle	crop + resize	gray	68.1	54.2	91.6	93.8	56.9	99.2

Table 8: Accuracies on spurious correlations datasets when varying factors for ViT-L-14 CLIP model. Factors columns are described in § E.1. Dataset explanation: Cmn/Ctr - “Common” and “Counter” subsets of CounterAnimal; WB - Waterbirds; IN-9 - ”mixed rand” subset of ImageNet-9; IN-D - “background” subset of ImageNet-D; UC - UrbanCars.

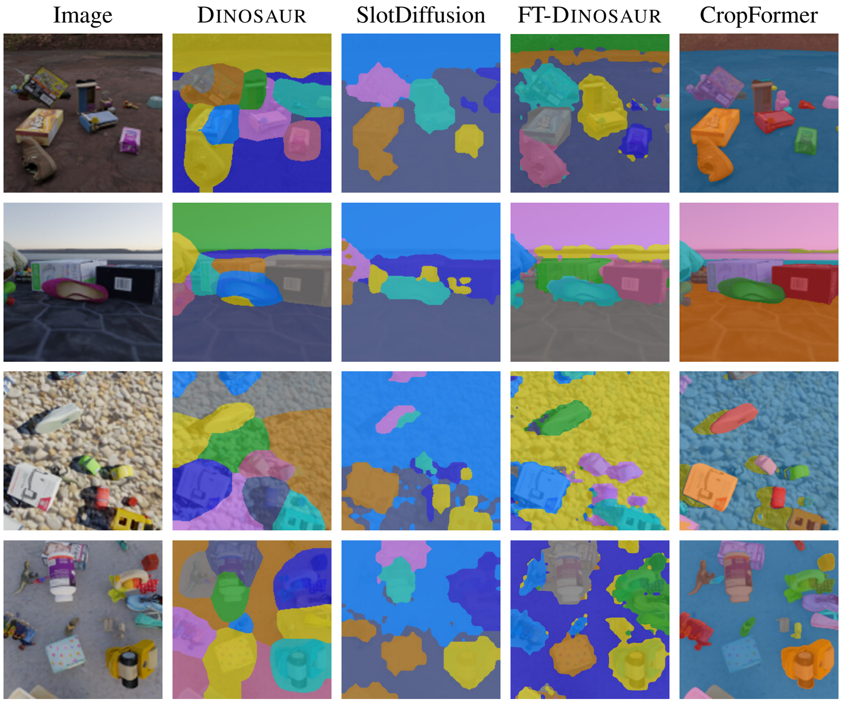

Figure 2: Zero-shot segmentation examples on MOVi-E.

F Extended unsupervised segmentation results

F.1 Quantitative results

Metrics To quantify model performance on unsupervised image segmentation tasks in Table 1, we compute the foreground adjusted random index (FG-ARI) (Rand, 1971; Hubert & Arabie, 1985; Kipf et al., 2022), which is a permutation-invariant clustering metric. It compares pixel clusters formed by predicted segmentation masks with clusters formed by ground-truth masks while ignoring background pixels.

In addition to that, we also compute the mean best overlap (mBO) (Pont-Tuset et al., 2015; Seitzer et al., 2023). This metric assigns each ground-truth mask the predicted mask with the largest overlap. It then averages the intersection-over-union scores of the matched pairs. Unlike FG-ARI, mBO considers background pixels. It also measures how well masks fit objects.

Baselines Dinosaur (Seitzer et al., 2023) uses pre-trained self-supervised Dino (Caron et al., 2021) features as a target for reconstruction loss. This loss is used to train a decoder with Slot Attention (Locatello et al., 2020) on top of the ResNet (He et al., 2016) encoder. FT-Dinosaur (Didolkar et al., 2025) improves Dinosaur by replacing the ResNet encoder with a DINO-ViT (Dosovitskiy et al., 2021) encoder separate from the one used to compute target features. It jointly fine-tunes the encoder with the decoder. SlotDiffusion (Jiang et al., 2023) uses pre-trained features from the Stable Diffusion Encoder (Rombach et al., 2022) and trains a diffusion-based decoder with Slot Attention (Locatello et al., 2020) on top of them.

F.2 Qualitative results

To show that foundational segmentation models outperform state-of-the-art object-centric methods in unsupervised instance segmentation, in addition to the quantitative results in Table 1, we provide qualitative results on the Movi-E dataset in Figure 2. Masks predicted by CropFormer model fits objects much better than masks predicted by OCL methods.