ARE WE DONE WITH OBJECT-CENTRIC LEARNING?

Alexander Rubinstein **Ameya Prabhu** **Matthias Bethge** **Seong Joon Oh**
Tübingen AI Center, University of Tübingen

Project Page OCCAM Codebase

ABSTRACT

Object-centric learning (OCL) seeks to *learn* representations which only encode an object, isolated from other objects or background cues in a scene. This approach underpins various aims, including out-of-distribution (OOD) generalization, sample-efficient composition, and modeling of structured environments. Most research has focused on developing unsupervised mechanisms which separate objects into discrete slots in the representation space, evaluated using unsupervised object discovery. However, with recent sample-efficient segmentation models, we can separate objects in the pixel space and encode them independently. This achieves remarkable zero-shot performance on OOD object discovery benchmarks, is scalable to foundation models, and can handle variable number of slots out-of-the-box. Hence, the goal of OCL methods to obtain object-centric representations has been largely achieved. Despite this progress, a key question remains: How does the ability to separate objects within a scene contribute to broader OCL objectives, such as OOD generalization? We address this by investigating the OOD generalization challenge caused by spurious background cues through the lens of OCL. We propose a novel, training-free probe called **Object-Centric Classification with Applied Masks** (**OCCAM**), demonstrating that segmentation-based encoding of individual objects significantly outperforms slot-based OCL methods. However, challenges in real-world applications remain. We provide the toolbox for the OCL community to use scalable object-centric representations, and focus on practical applications and fundamental questions, such as understanding object perception in human cognition.

1 INTRODUCTION

Object-centric learning (OCL) seeks to develop representations of complex scenes that independently encode each foreground object separately from background cues, ensuring that one object's representation is not influenced by others or the background. This constitutes a foundational element for many objectives: it supports modeling of structured environments (Schölkopf et al., 2021), enables robust out-of-distribution (OOD) generalization (Dittadi et al., 2022; Arefin et al., 2024), facilitates compositional perception of complex scenes (Greff et al., 2020), and deepens our understanding of object perception in human cognition (Spelke, 1990; Téglás et al., 2011; Wagemans, 2015). However, despite these broad goals, most research in OCL has centered on advancing "slot-centric" methods that separate objects and encode them into slots, evaluated using unsupervised object discovery as the primary metric (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 challenge the continued emphasis on developing mechanisms to separate objects in representation space as the main challenge to be addressed in OCL.

We first show that sample-efficient class-agnostic segmentation models, such as High-Quality Entity Segmentation (HQES) (Lu et al., 2023) are far better alternatives to latest slot-centric OCL approaches, already achieve impressive zero-shot object discovery. Moreover, these models are scalable, with foundation models like Segment Anything (SAM) (Kirillov et al., 2023; Ravi et al., 2025) showing remarkable zero-shot segmentation, addressing much of what is usually tackled with slot-centric approaches. Yet, the broader potential of OCL remains largely unexplored. We pose a critical question: How does the ability to separate objects within scenes contribute to other OCL objectives, such as OOD generalization.

Methods	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
HQES (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: **Object Discovery Performance.** Quantitative results for object discovery on Movi-C and Movi-E; column "FT" indicates whether the model was fine-tuned on the training split of the corresponding dataset (Movi-C or Movi-E). HQES outperforms the OCL baselines like Slot Diffusion and Dinosaur, despite being sample-efficient (151k training samples).

We bridge this gap by directly linking OCL to OOD generalization, especially in known hard settings with spurious background cues. We introduce **Object-Centric Classification with Applied Masks** (**OCCAM**), a simple, object-centric probe for robust zero-shot image classification. OCCAM consists of two stages: (1) generating object-centric representations via object-wise mask generation, and (2) applying OCL representations to downstream application by classifying images by selectively focusing on relevant object features while discarding misleading background cues.

Empirically, we find that, on Stage (1), sample-efficient segmentation models outperform current OCL approaches in obtaining object-centric representations without additional training. However, Stage (2)—the task of identifying relevant object cues amidst numerous possible masks—remains a challenge. Nevertheless, when Stage (2) is executed correctly, simple OCL probes such as OCCAM already have the potential for robust OOD generalization.

We recommend more focus by future OCL works on creating benchmarks, methodologies testing real-world applications where object-centric representations offer clear practical benefits, and explore fundamental questions, such as how object perception works in human cognition.

2 Experiments

In this section, we first evaluate slot-centric OCL approaches to foundational segmentation models on unsupervised object discovery tasks. We then evaluate whether OCL methods provide robust object classification by benchmarking them against a strong baseline that uses mask predictions from foundational segmentation models, following the OCCAM pipeline (§A).

2.1 Are we done with object-discovery?

OCL methods are often evaluated by how well they perform on unsupervised object discovery, measured via instance segmentation for every object in the scene. We explore whether the emergence of strong zero-shot segmentation models (class-agnostic) such as HQES (Lu et al., 2023) and SAM (Kirillov et al., 2023) allows reliable decomposition of the scene into objects. We compare these foundational segmenters against state-of-the-art OCL approaches (Jiang et al., 2023; Seitzer et al., 2023; Didolkar et al., 2025).

Setup. We first describe our experimental setup, including datasets, metrics, and compared baselines. Following prior work (Kipf et al., 2022; Elsayed et al., 2022; Seitzer et al., 2023; Didolkar et al., 2025), we use two synthetic image datasets from Greff et al. (2022): Movi-C and Movi-E. We quantify model performance using FG-ARI (Rand, 1971; Hubert & Arabie, 1985; Kipf et al., 2022) and mBO (Pont-Tuset et al., 2015; Seitzer et al., 2023).

Results. Table 1 and Figure 3 show quantitative and qualitative results. Across both metrics FG-ARI and mBO across out-of-distribution benchmarks like Movi-C and Movi-E, HQES far surpasses the OCL baselines. This gap is especially notable in mBO on Movi-E, improving 29.9% to 63.8%. Qualitatively, HQES masks fit objects much better than masks predicted by OCL methods (Figure 3). HQES also shows it is possible to be sample efficient, only being trained on 151k samples in contrast to 11M samples for SAM.

Conclusion. Sample-efficient segmentation models, even in zero-shot setting excel at object discovery, surpassing OCL methods by large margins. This suggests that one key aspect of OCL — decomposing the scene into objects — can be largely solved by powerful pre-trained segmentation models, effectively replacing the slot-based OCL methods. Given the decomposition, we explore in the next section downstream applications where OCL methods can contribute lot of practical value.

(a) ImgNet-D (BG) (Zhang et al., 2024)

Method	Acc. (↑)
CLIP ViT-L	
CLIP (Radford et al., 2021)	23.5
O-D (Ours)	57.7
O-H (Ours)	68.0
CLIP-SigLip (Zhai et al., 2023)	59.4
O-D-SigLip (Ours)	71.5
O-H-SigLip (Ours)	**78.5**
Multi-modal LLMs	
MiniGPT-4 (Zhu et al., 2024)	71.8
LLaVa (Liu et al., 2023a)	52.9
LLaVa-NeXT (Liu et al., 2024b)	68.8
LLaVa-1.5 (Liu et al., 2024a)	73.3⋆

(b) UrbanCars (Li et al., 2023)

Method	WGA (↑)
ViT-L-14 CLIP	
CLIP (Radford et al., 2021)	87.2
O-D (Ours)	98.4
O-H (Ours)	**100.0**
ResNet50 CLIP	
CLIP (Radford et al., 2021)	64.8
O-D (Ours)	98.4
O-H (Ours)	**100.0**
ResNet50	
CoBalT (Arefin et al., 2024)	80.0
LfF (Nam et al., 2020)	34.0
JTT (Liu et al., 2021)	55.8
SPARE (Yang et al., 2023)	76.9
LLE (Li et al., 2023)	90.8⋆

(c) ImgNet-9 (MR) (Xiao et al., 2021)

Method	Acc. (↑)
ViT-L-14 CLIP	
CLIP (Radford et al., 2021)	91.9
O-D (Ours)	93.8
O-H (Ours)	**95.2**
ResNet50 CLIP	
CLIP (Radford et al., 2021)	81.1
O-D (Ours)	80.6
O-H (Ours)	**85.6**
ResNet50	
CoBalT (Arefin et al., 2024)	80.3
SIN (Sauer & Geiger, 2021)	63.7
INSIN (Sauer & Geiger, 2021)	78.5
INCGN (Sauer & Geiger, 2021)	80.1
MaskTune (Asgari et al., 2022)	78.6
CIM (Taghanaki et al., 2021)	81.1⋆

(d) Waterbirds (Sagawa et al., 2020)

Method	WGA (↑)
ViT-L-14 CLIP	
CLIP (Radford et al., 2021)	83.6
O-D (Ours)	92.1
O-H (Ours)	**96.0**
ResNet50 CLIP	
CLIP (Radford et al., 2021)	72.9
O-D (Ours)	83.3
O-H (Ours)	**92.5**
ResNet50	
CoBalT (Arefin et al., 2024)	90.6
GDRO (Sagawa et al., 2020)	89.9
AFR (Qiu et al., 2023)	90.4
SPARE (Yang et al., 2023)	89.8
MaskTune (Asgari et al., 2022)	86.4
CIM (Taghanaki et al., 2021)	77.2
DFR (Kirichenko et al., 2023)	91.8⋆

Table 2: **Object-Centric Learning for Spurious Background OOD Generalization**. We report versions of accuracies in each benchmark. Results are grouped according to backbone architecture. "ImgNet-D (BG)" stands for the ImageNet-D "background" subset. "ImgNet-9 (MR)" stands for the ImageNet-9 "mixed rand" subset. "WGA" stands for the worst group accuracies. O-H/O-D stands for OCCAM with HQES/FT-Dinosaur masks generator correspondingly. For cited methods, we show results reported in the papers (Arefin et al., 2024) and (Zhang et al., 2024). ⋆ indicates the state-of-the-art results in each benchmark.

2.2 Application: Classification with Spurious Background Correlations

As foundational segmentation models outperform OCL methods in decomposing the scene into constituent objects, we take a further step and evaluate OCL methods on a downstream task that leverages the disentangled representations for distinct objects: robust classification under spurious background cues. This subsection demonstrates that object masks are a simple but effective strategy to mitigate the influence of spurious correlations with backgrounds in classification tasks (Table 2).

Setup. We first describe our experimental setup, including datasets, metrics, and compared baselines. We use several standard datasets with spurious backgrounds or co-occurring objects — UrbanCars (Li et al., 2023), ImageNet-D (Zhang et al., 2024) (background subset), ImageNet-9 (Xiao et al., 2021) (mixed rand subset), Waterbirds (Sagawa et al., 2020), and CounterAnimals (Wang et al., 2024). We measure model performance using standard metric used in the respective benchmark: accuracy and worst group accuracy (WGA). We provide per-benchmark comparisons for reference, including results from other relevant methods, citing them alongside their names in the tables. We use the foundational segmentation model HQES (Lu et al., 2023) (O-H) and the state-of-the-art OCL method FT-Dinosaur (Didolkar et al., 2025) (O-D) for mask prediction in our training-free probe, OCCAM. We categorize methods with comparable image encoder backbone for fairness.

Results. Using masks significantly improves performance across all datasets, sometimes reaching 100% accuracy (e.g., on UrbanCars; Table 2(b)) or close to that performance on Waterbirds and ImageNet-9 (mixed rand) subsets. This shows the potential of simple, training-free object-centric methods like OCCAM to address otherwise challenging downstream problems, if we can robustly identify the foreground object of interest. On harder benchmarks like ImageNet-D (background subset), HQES-based masks with SigLip models yield far better performance (78.5%) even compared to recent models like LLAVA 1.5 (Liu et al., 2023a) (73.3%), and outperform their best slot-based counterparts (71.5%) using FT-Dinosaur (Table2(a)). Throughout, HQES consistently provides more effective masks than FT-Dinosaur.

Conclusion. These experiments show that mask-based, training-free object-centric probes can provide practical value on challenging robust classification tasks, if the task of foreground detection is sufficiently addressed (§A.2.2). It provides substantial gains on all tested benchmarks over the state-of-the-art methods for tackling spurious correlations. We hope this encourages the community to develop segmentation-based OCL approaches, and demonstrate practical benefits across a variety of downstream applications. We next perform data-centric analysis with properties of OCCAM.

2.2.1 CounterAnimals: Spurious or Simply Hard?

Our object-centric classification pipeline can isolate an object's influence apart from its background. This property of OCL can be used to analyze the recently proposed CounterAnimals dataset (Wang et al., 2024).

Setup. CounterAnimals (Wang et al., 2024) highlights models' reliance on spurious backgrounds. It consists of two splits from iNaturalist,[1] each containing animals from 45 classes in ImageNet1k (Russakovsky et al., 2015). The Common split features typical backgrounds (e.g., polar bears on snow), while the Counter split features less common ones (e.g., polar bears on dirt). It primarily demonstrates that models consistently perform better on the Common than on the Counter, due to spurious background cues.

What is the Contribution of Spurious Correlations? We perform a simple check using OCCAM – If the drop from Common to Counter is caused by spurious background correlations, then using OCCAM we can ablate the contribution of everything except the foreground object. Ideally, ablating the background should result in roughly equal performance on both Common and Counter sets (the gap should be 0%). However, we see from Table 3, Table 4 that even after ablating the background entirely, there is a substantial gap between the Common and Counter subsets. For example, when using AlphaCLIP the gap reduces from 17.0% to 15.2%. Similarly, using HQES masks and gray background for both sets, we still observe a 8.5% gap. This provides interesting evidence that images in the Common subset might be simply substantially easier than images from the Counter subset by about 8-10%.

CounterAnimals		
Method	Cmn/Cntr (↑)	Cmn-Ctr (↓)
AlphaCLIP ViT-L		
CLIP (Radford et al., 2021)	79.0/62.0	17.0
O-D (Ours)	**85.8/70.5**	15.3
O-H (Ours)	84.4/69.2	15.2

Table 3: **Data-Centric Understanding using OCL.** We report the accuracies on the Common and Counter subset of the Counteranimals dataset. We see that after eliminating the spurious background using OCL methods, the gap (Cmn-Ctr) does not substantially decrease.

Conclusion. OCL methods allow analyzing datasets, and analyse the contribution of individual objects. In the case of CounterAnimals, we find that spurious backgrounds might not be the primary reason models perform worse on the Counter subset than on Common subset, although they are a factor. A significant (10%) gap might be caused by the Counter subset simply being harder to classify than the Common subset due to a wide variety of other factors. Overall, we show the potential for OCL methods to help inform data-centric fields like data attribution.

3 Conclusion and Open Problems

The motivation for object-centric learning (OCL) originates from a variety of goals, including out-of-distribution generalization, sample-efficient composition, and insights into human cognitive object perception. Despite this broad scope, progress has been measured mostly by object-discovery benchmarks only. With the advent of strong segmentation methods such as High-Quality Entity Segmentation (HQES) (Lu et al., 2023), we confirm that class-agnostic segmentation models far surpass slot-based OCL methods in obtaining isolated object representations, effectively meeting OCL's initial goal.

However, its relevance extends beyond object discovery. We advocate for shifting OCL evaluation towards more realistic downstream tasks that leverage object-centric representations, such as mitigating spurious background correlations. We design a simple training-free probe, OCCAM, to show the efficacy of object-centric approaches to help classifiers generalise even in the presence of spurious correlations (§2.2), achieving near-perfect accuracies across many benchmarks (Table 2). By separating object-wise representation (well-addressed by HQES) from object selection (still a key challenge), OCCAM sheds light on where further improvements are needed.

Looking ahead, we hope OCL-based approaches benchmark visual understanding through scene-graph construction, more interpretable intermediate representations, and human-in-the-loop feedback for cue selection. We hope diverse applications and creating corresponding benchmarks will push the field forward. Beyond immediate use cases, OCL may also inform fundamental cognitive questions about how objects and causal structures emerge in the real world and how infants understand objects without explicit supervision (Spelke, 1990; Téglás et al., 2011). Realizing this broader vision will require refining the OCL objective and breaking it down into well-defined subproblems that can further illuminate these deeper inquiries.

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

ACKNOWLEDGMENTS

The authors would like to thank (in alphabetical order): Shyamgopal Karthik, Yash Sharma, Matthias Tangemann, Thaddaeus Wiedemer for insightful feedback and suggestions. This work was supported by the Tübingen AI Center. AP and MB acknowledge financial support by the Federal Ministry of Education and Research (BMBF), FKZ: 011524085B and Open Philanthropy Foundation funded by the Good Ventures Foundation. AR thanks the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for support. This research utilized compute resources at the Tübingen Machine Learning Cloud, DFG FKZ INST 37/1057-1 FUGG.

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. 1, 3, 15, 16

Saeid Asgari, Aliasghar Khani, Fereshte Khani, Ali Gholami, Linh Tran, Ali Mahdavi-Amiri, and Ghassan Hamarneh. Masktune: Mitigating spurious correlations by forcing to explore. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. 3

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Christopher Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub W. Pachocki, Michael Petrov, Henrique Pondé de Oliveira Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. Dota 2 with large scale deep reinforcement learning. *ArXiv*, abs/1912.06680, 2019. 15

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. 16

Christopher P. Burgess, Loïc Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matthew M. Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *ArXiv*, abs/1901.11390, 2019. URL https://api.semanticscholar.org/CorpusID:59523721. 15

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. 15

Stefan Depeweg, José Miguel Hernández-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, 2017. 19

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. 1, 2, 3, 12, 15, 16, 22

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. 1, 15, 16

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. 13, 15, 19

Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham W.Taylor. Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10303–10311, 2018. 15

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. 1, 2, 15, 16

Marco Fumero, Florian Wenzel, Luca Zancato, Alessandro Achille, Emanuele Rodolà, Stefano Soatto, Bernhard Schölkopf, and Francesco Locatello. Leveraging sparse and shared feature activations for disentangled representation learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 27682–27698. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/57fabaa549352c52d5d312171b16970e-Paper-Conference.pdf. 15

Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*, 2020. 1, 15

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. 2

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. 19, 23

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. 15

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. 19

Lawrence J. Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 1985. 2, 15

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip. July 2021. doi: 10.5281/zenodo.5143773. If you use this software, please cite it as below. 23

Jindong Jiang, Fei Deng, Gautam Singh, and Sungjin Ahn. Object-centric slot diffusion. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 1, 2, 15, 16

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. 1, 2, 15

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. 3, 17, 19, 20

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. 1, 2

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. *2013 IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013. URL https://api.semanticscholar.org/CorpusID:14342571. 22

Tejas D. Kulkarni, Ankush Gupta, Catalin Ionescu, Sebastian Borgeaud, Malcolm Reynolds, Andrew Zisserman, and Volodymyr Mnih. Unsupervised learning of object keypoints for perception and control. In *Neural Information Processing Systems*, 2019. 15

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 19

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. 3, 21, 22

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6781–6792. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/liu21f.html. 3

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023a. 3

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, June 2024a. 3

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024b. 3

Yuejiang Liu, Alexandre Alahi, Chris Russell, Max Horn, Dominik Zietlow, Bernhard Schölkopf, and Francesco Locatello. Causal triplet: An open challenge for intervention-centric causal representation learning. In *Conference on Causal Learning and Reasoning*, pp. 553–573. PMLR, 2023b. 15

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. 1, 15

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. 1, 2, 3, 4, 12

Shoya Matsumori, Kosuke Shingyouchi, Yukikoko Abe, Yosuke Fukuchi, Komei Sugiura, and Michita Imai. Unified questioner transformer for descriptive question generation in goal-oriented visual dialogue. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1878–1887, 2021. 15

Toki Migimatsu and Jeannette Bohg. Object-centric task and motion planning in dynamic environments. *IEEE Robotics and Automation Letters*, 5(2):844–851, 2020. doi: 10.1109/LRA.2020.2965875. 15

Bálint Mucsányi, Michael Kirchhof, and Seong Joon Oh. Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 19

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. 19, 23

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546. 3

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019. 19, 23

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. 2, 15

Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023. 3

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. 3, 4, 19, 20, 22

William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 1971. 2, 15

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. 1

Seyed Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian D. Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 658–666, 2019. 23

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. 15, 21

Alexander Rubinstein, Luca Scimeca, Damien Teney, and Seong Joon Oh. Scalable ensemble diversification for ood generalization and detection. *arXiv preprint arXiv:2409.16797*, 2024. 19

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. 4, 19, 21, 23

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations (ICLR)*, 2020. 3, 17, 19, 21, 22

Axel Sauer and Andreas Geiger. Counterfactual generative networks. In *International Conference on Learning Representations*, 2021. 3

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. 1, 15

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. 1, 2, 15, 16

Elizabeth S. Spelke. Principles of object perception. *Cognitive Science*, 1990. doi: https://doi.org/10.1016/0364-0213(90)90025-R. 1, 4, 15

Chen Sun, Per Karlsson, Jiajun Wu, Joshua B Tenenbaum, and Kevin Murphy. Predicting the present and future states of multi-agent systems from partially-observed visual data. In *International Conference on Learning Representations*, 2019. 15

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. 3

Dustin Tran, Jeremiah Zhe Liu, Michael W. Dusenberry, Du Phan, Mark Collier, Jie Jessie Ren, Kehang Han, Z. Wang, Zelda E. Mariet, Huiyi Hu, Neil Band, Tim G. J. Rudner, K. Singhal, Zachary Nado, Joost R. van Amersfoort, Andreas Kirsch, Rodolphe Jenatton, Nithum Thain, Honglin Yuan, E. Kelly Buchanan, Kevin Murphy, D. Sculley, Yarin Gal, Zoubin Ghahramani, Jasper Snoek, and Balaji Lakshminarayanan. Plex: Towards reliability using pretrained large model extensions. *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022. 19, 23

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. 1, 4, 15

Johan Wagemans. *The Oxford Handbook of Perceptual Organization*. Oxford University Press, 2015. doi: 10.1093/oxfordhb/9780199686858.001.0001. 1, 15

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. URL https://api.semanticscholar.org/CorpusID:16119123. 22

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. 3, 4, 21

Nicholas Watters, Loïc Matthey, Matko Bosnjak, Christopher P. Burgess, and Alexander Lerchner. Cobra: Data-efficient model-based rl through unsupervised object discovery and curiosity-driven exploration. *ArXiv*, abs/1905.09275, 2019. 15

Taylor Whittington Webb, Shanka Subhra Mondal, and Jonathan Cohen. Systematic visual reasoning through object-centric relational abstraction. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 15

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. 16

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. 3, 21

Jianwei Yang, Jiayuan Mao, Jiajun Wu, Devi Parikh, David Cox, Joshua B. Tenenbaum, and Chuang Gan. Object-centric diagnosis of visual reasoning. *ArXiv*, abs/2012.11587, 2020. 15

Yu Yang, Eric Gan, Gintare Karolina Dziugaite, and Baharan Mirzasoleiman. Identifying spurious biases early in training through the lens of simplicity bias. *ArXiv*, abs/2305.18761, 2023. 3

Jaesik Yoon, Yi-Fu Wu, Heechul Bae, and Sungjin Ahn. An investigation into pre-training object-centric representations for reinforcement learning. In *International Conference on Machine Learning*, pp. 40147–40174. PMLR, 2023. 15

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952, 2023. 3

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. 3, 21

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2024. 3

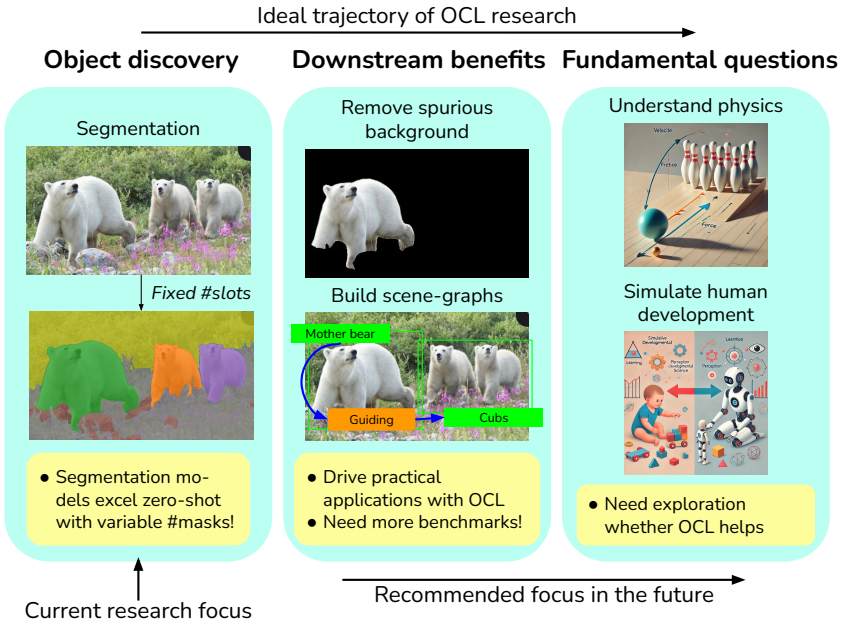

Figure 1: **Where Should We Go?** Object-centric learning (OCL) has focused on developing unsupervised mechanisms to separate the representation space into discrete *slots*. However, the inherent challenges of this task have led to comparatively less emphasis on exploring downstream applications, and exploring fundamental benefits. Here, we introduce simple, effective OCL mechanisms by separating objects in pixel space and encoding them independently. We present a case study that demonstrates the downstream advantages of our approach for mitigating spurious correlations. We outline the need to develop benchmarks aligned with fundamental goals of OCL, and explore the downstream efficacy of OCL representations.

A OCCAM: PROPOSED METHOD

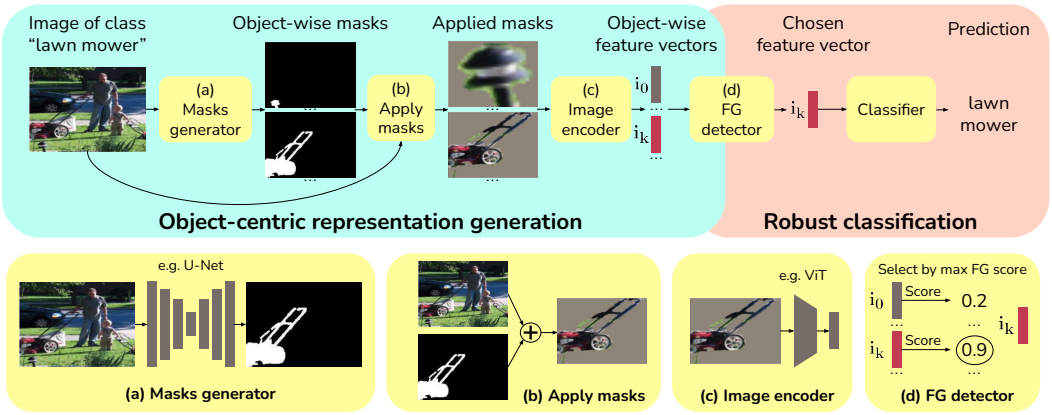

Figure 2: **Overview of Object-Centric Classification with Applied Masks (OCCAM)**. There are two main parts. The first part (§ A.2.1) uses entity segmentation masks for **object-centric representation generation**. The second part (§ A.2.2) performs **robust classification** by selecting representations corresponding to the foreground object and using them for classification. Indices $i_0, \ldots, i_k, \ldots$ correspond to each object in the scene.

This section gives an overview of our proposed method. Subsection §A.1 defines the notation needed for the method description in §A.2.

A.1 NOTATIONS

We denote an image as $x \in \mathbb{R}^{[3,\,\mathrm{H},\,\mathrm{W}]}$ and a label as $y \in \mathcal{Y} = \{1, \ldots, C\}$, 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 $p(\psi(x)) = \mathrm{Softmax}(f(\psi(x))) \in [0,1]^{|\mathcal{Y}|}$. For simplicity, we will use $p(\psi(x))$ and $p(x)$ interchangeably. 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.

A.2 METHOD

Our Object-Centric Classification with Applied Masks (OCCAM) pipeline is summarized in Figure 2. 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.

A.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 $x \in \mathbb{R}^{[3,\,\mathrm{H},\,\mathrm{W}]}$, we generate a set of masks for all the foreground objects and the background. That is done with the help of mask generator 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 each corresponding to a different object. An OCL method like FT-Dinosaur (Didolkar et al., 2025) or an external segmentation model like High-Quality Entity Segmentation (HQES) (Lu et al., 2023) can be used as a mask 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. We will interchangeably call mask applying operation as mask method throughout the paper. One way to apply masks to images is to simply add a gray background to all but selected pixels, cropping the image that follows the mask contours, and resizing the result to the size of the original image. In such case, we call the operation as Gray BG + Crop.

However, a mask method can be any operation involving an image x and mask m: $a(x, m) \in \mathbb{R}^{[3,\ H,\ W]}, \forall m \in \{S_i, i = 1 \dots K\}, m \in \mathbb{R}^{[H,\ W]}$. We additionally show ease-of-use in incorporating latest masking techniques like AlphaCLIP which 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. This allows using masks as a source of focus instead of removing backgrounds entirely, useful for some practical applications. We call such an operation as "α-channel".

Encoding applied masks. To get the final object-centric representations we encode applied masks by an image encoder ψ such as ViT (Dosovitskiy et al., 2021) for example.

A.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* that 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.

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

1. **Ens.** $\mathcal{H}$: $g_{\mathcal{H}}(x, m) = \frac{1}{M} \sum_{k=1}^{M} \mathcal{H}[p_k(x)]$ - ensemble entropy (see details in § C.1). Here, M is the ensemble size, and $\mathcal{H}$ stands for entropy.
2. **Class-Aided**: $g_{\text{class\_aided}}(x, m) = p^y(\psi(a(x, m)))$ - 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.

For the comparison of different foreground scores, see § C.1.

Image classification using FG 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:

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

where $m^\star$ is the FG mask selected by the FG detector.

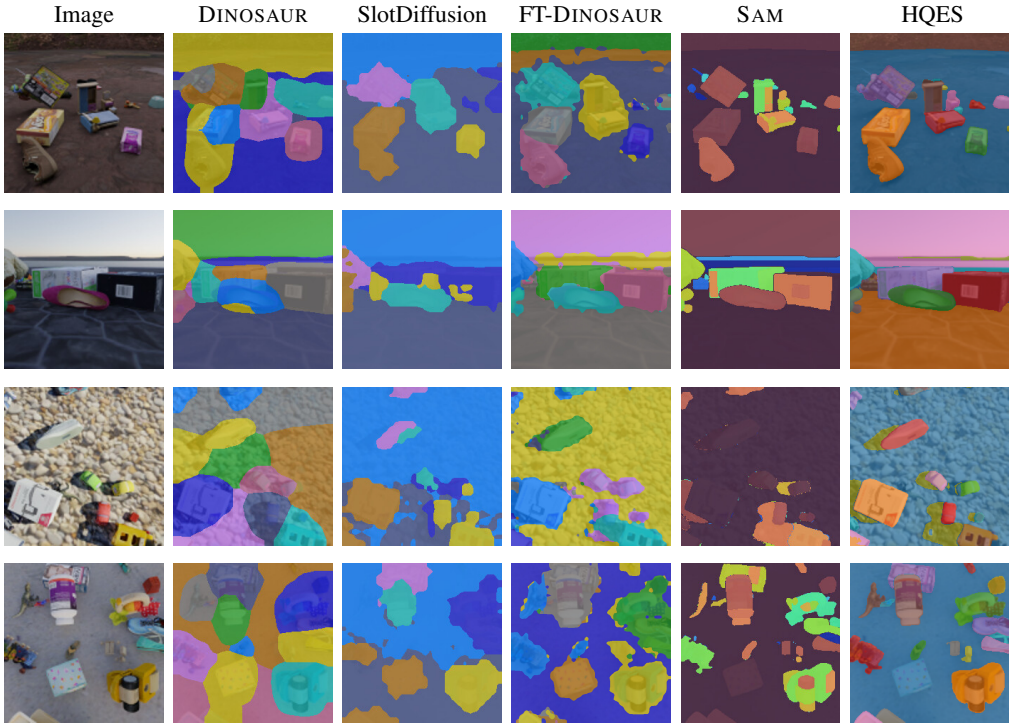

Figure 3: **Qualitative Results on Object Discovery**. DINOSAUR, SlotDiffusion, and FT-DINOSAUR are existing object-centric learning (OCL) approaches. SAM and HQES refer to zero-shot segmentation methods. Images are from MOVi-E. SAM and HQES masks fit objects much better than the masks predicted by OCL methods.

B Related work

We cover prior work in the object-centric learning (OCL) community from three different angles: motivation, evaluation, and methodologies.

Motivation for OCL. The OCL community has inspired research from different perspectives. From one perspective, learning object-centric representations can help discover latent variables of the data-generating process, such as object position and color (Fumero et al., 2023), or even identify its causal mechanisms (Liu et al., 2023b; Schölkopf et al., 2021) by encoding structural knowledge that allows interventions and changes. From another perspective, OCL aims to simulate human cognition (Spelke, 1990; Téglás et al., 2011; Wagemans, 2015) in neural networks. For example, infants intuitively understand physics by tracking objects with consistent behavior over time (Dittadi et al., 2022). They later reuse this knowledge to learn new tasks quickly. Advances in OCL can help neural networks develop this ability as well. In addition to that, some studies focus on understanding the compositional nature of scenes (Greff et al., 2020) by providing separate representations for different elements (e.g. human, hat, bed, table) and their interactions (a cat wearing a hat or a bear guiding cubs). Several papers claim that there is a potential to improve sample efficiency and generalization (Locatello et al., 2020; Kipf et al., 2022; Seitzer et al., 2023) or object-centric methods can be more robust (Seitzer et al., 2023). Others refer to the structure of the world saying that the fundamental structure of the physical world is compositional and modular (Jiang et al., 2023) or that humans understand the world in terms of separate objects (Kipf et al., 2022; Didolkar et al., 2025). However, we have observed a consistent lack of empirical evidence demonstrating that object-centric approaches improve sample efficiency or aid in identifying causal mechanisms. To address this gap, we believe more empirical research is needed. As a first step, we show that robust classification is achievable even in the presence of explicitly distracting backgrounds and other object interference.

OCL evaluation. Measuring progress on the primary motivations of object-centric learning is a hard problem and suffers from chronic lack of scalable benchmarks. Hence, 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. Some papers study the link between object-centric learning and downstream applications. These include reinforcement learning (Watters et al., 2019; Kulkarni et al., 2019; Berner et al., 2019; Sun et al., 2019; Yoon et al., 2023), scene representation and generation (Kulkarni et al., 2019; El-Nouby et al., 2018; Matsumori et al., 2021; Burgess et al., 2019), reasoning (Webb et al., 2023; Yang et al., 2020), and planning (Migimatsu & Bohg, 2020). We highlight that these papers provide valuable contribution to benchmarking progress in OCL field. However, most research does not focus on these tasks. Much of the progress is tracked by unsupervised object discovery benchmarks, 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). Model performance is usually quantified with foreground adjusted random index (FG-ARI) (Rand, 1971; Hubert & Arabie, 1985; Kipf et al., 2022), which is a permutation-invariant clustering metric or mean best overlap (mBO) (Pont-Tuset et al., 2015; Seitzer et al., 2023). These evaluations primarily assess whether slots reliably isolate individual objects—a criterion we argue is overly restrictive in the broader context of object-centric learning. In our paper, we urge more works to additionally evaluate downstream applications, particularly given the emergence of foundational segmentation models that significantly outperform object-centric methods on standard object discovery tasks (see Table 1 and Figure 3).

OCL methodologies. 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. 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. 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). There is also a line of work that studies object-centric representation in the context of out-of-distribution (OOD) generalization in segmentation (Dittadi et al., 2022) and classification, e.g. CoBalT (Arefin et al., 2024) that employs model distillation and slots clustering into concepts to refine features quality. In our experiments, we compare with latest methods – SlotDiffusion (Jiang et al., 2023) and (FT-)Dinosaur (Seitzer et al., 2023; Didolkar et al., 2025) for object discovery and CoBalT (Arefin et al., 2024) across robust classification benchmarks.

C ABLATIONS: IDENTIFYING BOTTLENECKS IN OCCAM

Name	Mask Method	Mask Model	FG Detector	WB↑	IN-9↑	IN-D↑	UC↑	Cmn-Ctr↓
CLIP	-	-	-	83.6	91.9	17.6	87.2	15.0
	Gray BG + Crop	FT-Dinosaur	Ens. $\mathcal{H}$	83.8	84.0	52.4	95.2	13.1
			Class-Aided	92.1	93.8	57.7	98.4	12.7
		HQES	Ens. $\mathcal{H}$	86.8	88.6	60.4	95.2	8.8
			Class-Aided	96.0	95.2	68.0	100.0	8.5
AlphaCLIP	- ($\alpha = 1$)	-	-	79.8	90.2	23.5	87.2	17.0
	α-channel	FT-Dinosaur	Ens. $\mathcal{H}$	81.0	90.3	40.7	92.0	17.2
			Class-Aided	86.9	93.1	49.1	96.0	15.3
		HQES	Ens. $\mathcal{H}$	84.7	91.2	44.7	91.2	16.4
			Class-Aided	89.1	93.1	53.9	97.6	15.2

Table 4: **Factor Analysis for Spurious Background OOD Generalization**. Accuracies on spurious correlations datasets when varying factors for ViT-L-14 CLIP architecture. We use AlphaCLIP for α-channel masking and CLIP for Gray Crop masking. We first report their baseline performances without masking (where mask method and model are both "-") and with 2 different mask models (FT-Dinosaur and HQES) as well as 2 different foreground detectors (Ens. $\mathcal{H}$ and Class-Aided). Results are reported on 5 benchmark datasets, Waterbirds (WB), ImageNet-9 (IN-9), ImageNet-D (IN-D), UrbanCars (UC), and CounterAnimals (Cmn-Ctr). For the CounterAnimals results, we report the gap between the common-split (Cmn) and the counter-split (Ctr) accuracies. Unlike other metrics, a smaller Cmn-Ctr gap is deemed a better generalization.

We now ablate the contributions of different components in the OCCAM pipeline. We first test two CLIP models (CLIP and AlphaCLIP), to see whether our results generalize beyond simply removing backgrounds to recent techniques such as AlphaCLIP which use the α-channel to focus on the mask instead of eliminating the background. Secondly, we study the effect of masking generator, testing HQES along with current SOTA OCL method FT-Dinosaur. Lastly, we study the influence of different FG Detection methods. We showcase our analysis in Table 4.

Effect of mask applying method. Using masks with Class-Aided FG detector improves performance on all the datasets for both Gray BG + Crop and α-channel mask methods, but for the former accuracy is usually higher. For example, on Waterbirds (Table 4), accuracy for Gray BG + Crop mask method and HQES mask generator is 96.0% while for AlphaCLIP it is 89.1%. This indicates the backgrounds have strong spurious correlations still affects α-CLIP to a small extent.

Effect of mask generator. Comparing the rows from mask models to original CLIP model, we see that both FT-Dinosaur and HQES improve performance, across CLIP and AlphaCLIP given that we use Class-Aided FG detector. In this scenario, HQES improves accuracy more than FT-Dinosaur. For example, for the Gray BG + Crop mask method it leads to 68.0% accuracy on ImageNet-D, while FT-Dinosaur reaches only 57.7%. This indicates that the segmentation-based OCL performs better consistently for downstream OCL applications.

Selecting foreground mask. Accuracy gains with Ens. $\mathcal{H}$ are always smaller than for Class-Aided FG detector and sometimes can be negative (Table 4). For example, for Gray BG + Crop mask method and HQES mask generator accuracy on ImageNet-9 drops from 91.9% to 88.6% when using Ens. $\mathcal{H}$ FG detector, while jumping to 95.2% with Class-Aided FG detector. This reveals a weakness in the baseline foreground detection method. It leaves room for improvement and future research.

It is worth mentioning that since Class-Aided foreground detector selects masks based on the highest ground truth probability it can be biased towards non-foreground masks that improve overall accuracy, helping OCCAM achieve the superior performance in image classification. However, on the Waterbirds (Sagawa et al., 2020) dataset for which we have access to ground truth foreground masks (Kirichenko et al., 2023) we observe that it is not usually the case: non-foreground mask was selected by Class-Aided foreground detector in only 5 out of 100 randomly sampled images. At the same time, the accuracy for Class-Aided foreground detector, which is 96.0%, is still lower than the 96.7% achieved using ground truth masks. Based on this, we consider masks selected by Class-Aided foreground detector as the closest approximation of ground truth foreground masks.

Conclusion. The empirical results show that segmentation models outperform current OCL methods in obtaining object-centric representations that result in better classification performance. At the same time, identifying foreground masks among many candidates remains a challenge. The simple Gray BG + Crop mask method generally performs better than the more advanced α-channel mask method.

C.1 FOREGROUND DETECTORS COMPARISON

To justify the choice of $g_{\text{class\_aided}}$ and $g_{\mathcal{H}}$ in § A.2.2, we compare several foreground detection methods. One can notice that foreground detection is an application of an out-of-distribution (OOD) detection, a well-studied problem (Mukhoti et al., 2021; Tran et al., 2022; Gruber & Buettner, 2022) — with foreground objects treated as in-distribution (ID) samples and background objects as OOD samples. Hence, we evaluate OOD detection methods for this task in Figure 4.

Setup. We construct an OOD detection dataset using the ImageNet (Russakovsky et al., 2015) validation set by leveraging ground truth bounding boxes[2] to derive accurate foreground masks (see details in § G). Performance is measured via the area under the ROC curve (AUROC), in line with standard OOD detection frameworks (Mukhoti et al., 2021; Tran et al., 2022; Gruber & Buettner, 2022; Mucsányi et al., 2024; Rubinstein et al., 2024). We use the following strong baselines:

- *Class-Aided (single model)* (Hendrycks & Gimpel, 2017): $p^y(x)$
- *Ensemble entropy* (Ovadia et al., 2019): $\frac{1}{M}\sum_{k=1}^{M}\mathcal{H}[p_k(x)]$
- *Ensemble confidence* (Lakshminarayanan et al., 2017): $\max_c \frac{1}{M}\sum_{k=1}^{M} p_k^c(x)$
- *Confidence (single model)* (Hendrycks & Gimpel, 2017): $\max_c p^c(x)$
- *Entropy (single model)* (Depeweg et al., 2017): $\mathcal{H}[p(x)]$

Here, $p(x)$ denotes the model's probability vector prediction for sample x, y is the corresponding ground truth label, M is the ensemble size, and $\mathcal{H}$ represents entropy. We use ViT-L-14 CLIP model pre-trained by OpenAI (Radford et al., 2021) as the single model, with the ensemble comprising of CLIP (Radford et al., 2021) models with ViT-L-14 (Dosovitskiy et al., 2021) vision encoders pre-trained on different datasets. Note that OpenAI ViT-L-14 was the strongest model by AUROC among the ensemble, hence was used as the single model. Further details are provided in § G.

Results. As shown in Figure G, Class-Aided achieves the highest AUROC of 90.1% whereas the ensemble entropy method yields 89.6%. Other methods perform significantly worse. Nevertheless, all methods scored more than 80% AUROC.

Conclusion. The AUROC performance of Class-Aided and Ens. $\mathcal{H}$ foreground detectors showed only minor differences from each other both scoring around 90% and being the best among the compared methods; however, substantial performance gaps remain when comparing the Class-Aided results with the Ens. $\mathcal{H}$ foreground detector in spurious correlation tasks, possible reason for this is discussed in § D. This disparity highlights two key implications. Current evaluation metrics may have a large research gap to better reflect real-world applications. Conversely, spurious correlation foreground detection might be a promising proxy task for identifying better OOD detection models.

D CLASS-AIDED FOREGROUND DETECTOR YIELDS THE CLOSEST APPROXIMATION TO GROUND TRUTH FOREGROUND MASKS

It is worth mentioning that since Class-Aided foreground detector selects masks based on the highest ground truth probability (§ A.2.2), it can be biased towards non-foreground masks that improve overall accuracy, helping OCCAM achieve the superior performance in image classification. However, on the Waterbirds (Sagawa et al., 2020) dataset for which we have access to ground truth foreground masks (Kirichenko et al., 2023), we observe that it is not usually the case: a non-foreground mask was selected by Class-Aided foreground detector in only 5 out of 100 randomly sampled images. At the same time, the accuracy for Class-Aided foreground detector, which is 96.0%, is still lower than the 96.7% achieved using ground truth masks (see Table 5). We do not see clear evidence that the Class-Aided foreground detector often selects non-foreground masks but we see that it provides masks that perform on par with ground truth masks in image classification with spurious correlations. We therefore treat the masks it selects as the closest approximation to ground truth foreground masks.

[2] https://academictorrents.com/details/dfa9ab25

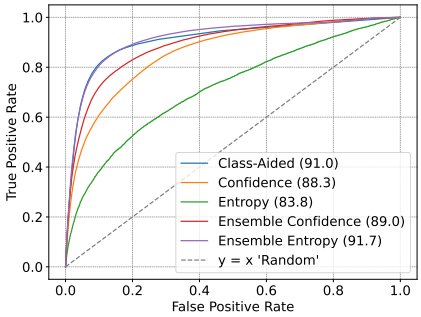

Figure 4: **Foreground Object Detection.** ROC-curves for foreground detection methods. For each scoring scheme, we measure how well the true foreground objects in the ImageNet-validation dataset are detected. More details in § G.

FG Detector	WGA (↑)
-	83.6
Max Prob	78.6
Ens. $\mathcal{H}$	86.8
Class-Aided	96.0
Ground Truth	**96.7**

Table 5: **Different foreground detectors on Waterbirds** We report the worst-group accuracies on the Waterbirds dataset for different foreground detectors. Masks are generated by HQES and applied via "Gray BG + Crop" (see § A.2.1); the classification model is CLIP ViT-L-14 (Radford et al., 2021); "-" stands for classification of original images without using any masks. Max Prob stands for foreground detector that uses the following score (in terms of § A.2.2): $g_{\text{max\_prob}}(x, m) = \max_c p^c(\psi(a(x, m)))$ - maximum probability across all possible classes (its computation is equivalent to confidence in § C.1). Class-Aided and Ens. $\mathcal{H}$ are described in § A.2.2. Ground Truth stands for ground truth foreground masks that are taken from (Kirichenko et al., 2023).

E DETAILS ON SPURIOUS BACKGROUNDS DATASETS

Below we provide details on these datasets:

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.
5. CounterAnimals (Wang et al., 2024): see details in § 2.2.1.

F EXTENDED IMPLEMENTATION DETAILS

Classes for zero-shot classification Following the original CLIP (Radford et al., 2021) work we compute the classifier's pre-softmax logits $f(\psi(x))$ using dot products. These are calculated between image embeddings and text embeddings of class name prompts. Class name prompts follow the format: "A photo of X". Here X takes values from the class names of the corresponding datasets. For Waterbirds (Sagawa et al., 2020) and UrbanCars (Li et al., 2023), we first compute dot products for prompts based on fine-grained classes that come from the CUB (Wah et al., 2011) and StanfordCars (Krause et al., 2013) datasets, respectively. We do that because the foreground objects in Waterbirds and UrbanCars were originally cropped from these datasets. All fine-grained classes are then grouped into two categories. For Waterbirds, they are divided into "land" and "sea" birds. For UrbanCars, they are divided into "urban" and "country" cars. The final prediction is the group that contains the fine-grained class corresponding to the highest dot product.

How resize is done for "Gray BG + Crop" First, find the smallest rectangle that fully contains the foreground object. Next, expand its shortest side to match the longest side while keeping the center of the square the same as the original rectangle's center. Finally, resize the square to the target size.

Fixed number of slots in OCL method When using FT-Dinosaur (Didolkar et al., 2025) as a mask generator we fix the number of slots to 5 as suggested by the original code.

Foundational segmentation model choice While in general HQES 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 HQES. 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.

Mask-free AlphaCLIP AlphaCLIP always requires a mask as input. To simulate cases without masks, we use a mask that covers the entire image as the foreground mask. We call this approach the "- (α = 1)" foreground detector. However, we treat this as mask-free performance since no real masks are used.

Masks filtering Before using masks in our experiments, we filter them using the following rules:

1. **Size:** Remove masks that cover less than 0.001 of the image pixels.
2. **Connected components:** Remove masks with more than 30 connected components.
3. **Background heuristic:** Remove masks that cover at least 6 of the 8 key points (4 corners and 4 side centres of the image).

G FG DETECTION AS A SPECIAL CASE OF OOD DETECTION

In this section, we give details on comparing different candidates for FG detector methods apart from $g_{\text{class\_aided}}$ and $g_{\mathcal{H}}$ (see § A.2.2 for details). We argue that foreground detection can be seen as an out-of-distribution (OOD) detection problem (Mukhoti et al., 2021; Tran et al., 2022; Gruber & Buettner, 2022) with foreground objects being in-distribution (ID) samples while background objects being OOD samples. We then evaluate the OOD detection methods for such an OOD detection task in Figure 4.

Dataset construction details. We generate the following masked dataset as follows. For each image in ImageNet (Russakovsky et al., 2015) validation set that has a corresponding ground truth bounding box for the foreground object (by the foreground object we mean the one for which the ground truth classification label is provided) we predict masks for all the objects that it contains as discussed in "Generating masks" paragraph in §A.2.1. After that, a mask is applied to the image using "Gray BG + Crop" operation as discussed in the "Applying masks" paragraph in §A.2.1. Then such a masked image is assigned label 1 if it corresponds to the foreground object (its corresponding mask has the biggest intersection over union (IoU) (Rezatofighi et al., 2019) with the ground truth bounding box) and 0 otherwise.

Ensemble members.

All models checkpoints are taken from the "openclip" (Ilharco et al., 2021) python library (corresponding pre-training dataset ids: "openai", "datacomp_xl_s13b_b90k", "dfn2b", "laion400m_e31", "laion400m_e32").

We mainly consider ensemble-based baselines for OOD detection as they are the most competitive baseline for this task (Mukhoti et al., 2021; Ovadia et al., 2019).

All these methods are used for the OOD detection in the following manner. Firstly, we use the above-mentioned formulae to compute uncertainty scores from models' outputs for the given sample. Secondly, we use this score as a probability to predict class 1 for this sample in the described binary classification problem setting.

Note: "Ensemble entropy" is equivalent to $g_{\mathcal{H}}$ and "Class-Aided" is equivalent to $g_{\text{class\_aided}}$ described in paragraph "Foreground detector" in § A.2.2.