CTRL-O: Language-Controllable Object-Centric Visual Representation Learning

Aniket Rajiv Didolkar[1,2*] **Andrii Zadaianchuk**[3*] **Rabiul Awal**[1,2*]
Maximilian Seitzer[4] **Efstratios Gavves**[3] **Aishwarya Agrawal**[1,2]
Mila - Quebec AI Institute[1], Université de Montréal[2]
University of Amsterdam[3], MPI for Intelligent Systems[4]

Abstract

Object-centric representation learning aims to decompose visual scenes into fixed-size vectors called "slots" or "object files", where each slot captures a distinct object. Current state-of-the-art models have shown remarkable success in object discovery, particularly in complex real-world scenes, while also generalizing well to unseen domains. However, these models suffer from a key limitation: they lack controllability. Specifically, current object-centric models learn representations based on their preconceived understanding of objects and parts, without allowing user input to guide or modify which objects are represented. Introducing controllability into object-centric models could unlock a range of useful capabilities, such as enabling models to represent scenes at variable levels of granularity based on user specification. In this work, we propose a novel approach that conditions slot representations through guided decomposition, paired with a novel contrastive learning objective, to enable user-directed control over which objects are represented. Our method achieves such controllability without any mask supervision and successfully binds to user-specified objects in complex real-world scenes.

1 Introduction

The goal of object-centric representation learning is to learn strong representations of the objects present in a visual scene. This goal is achieved by decomposing a visual scene into its constituent objects and representing each object as a distinct vector called a *slot*. Slot-based representations are inherently compositional and support many complex downstream tasks such as dynamics modeling [1, 2], control [3–6], and reasoning [7]. While initially these approaches for mainly limited to synthetic domains, recent works [8, 9] have shown that they can be used to extract representation in complex real-world scenes. Studies in cognitive neuroscience [10, 11] have also shown that slot-based representations closely mirror human perception.

One fundamental limitation of existing unsupervised object-centric models [5, 8, 12–14] such as Slot Attention (SA) [14] and DINOSAUR [8] is that they do not offer much control over the object representations. For example, if a user is interested in extracting the representation of a particular object in a scene, there is no way to query a model to do this. These methods allow control only over the number of parts into which the scenes are decomposed but not over the semantic contents of those parts. This lack of control over semantic content can be limiting, as these models always extract a fixed scene decomposition. Such fixed decomposition is not very useful because certain applications might require object representations at different levels of granularity. For instance, a user may be interested in extracting the representation of the wheel of a car rather than the whole car, or the entire car instead of just its parts.

[*] denotes equal contribution

NeurIPS 2024 Workshop on Adaptive Foundation Models.

Moreover, in complex multi-object scenes, one may require representations of a particular subset of objects. Such flexible representation is not achievable with current approaches because: 1) once the slots are extracted from a scene, there is no way to know which objects they represent, and 2) given a fixed scene decomposition, there is no guarantee that the object of interest will be extracted into an independent slot. A fixed scene decomposition may group multiple objects into slots based on its own preconceived understanding of objects and parts (e.g., based on the easiness of reconstructing image regions using one slot).

To address these limitations, we propose introducing controllability into object-centric representation learning. We achieve this by querying the model to represent specific objects in the image. The queries can be in the form of object categories, referring expressions, or center of mass points. In the proposed approach, we condition slots on queries that refer to specific objects in the scene. The main challenge is to ensure that the slots conditioned on a specific query bind to the object referred to by that query. We term the challenge of binding slots to specific objects the *visual grounding problem* [15, 16]. We find that this is not a trivial issue and introduce various architectural modifications along with a contrastive loss to solve it. In our experiments, we demonstrate that the proposed approach can successfully bind to objects referred to by user queries containing center of mass points, object categories or referring expressions in complex real-world scenes with limited supervision.

2 Method

In this section, we describe the proposed approach for injecting controllability into existing object-centric models. We present the visual depiction of our method in Figure 1.

In the setup that we consider here, the input consists of an image X and user-defined queries embedded into vectors $L = \{l_j\}_{j=1}^{M}$, where $l_j \in \mathbb{R}^{D_{\text{emb}}}$. The expected object-centric representation of the image X is a set of slots $S = \{s_i\}_{i=1}^{N}$, where $s_i \in \mathbb{R}^{D_{\text{slot}}}$. The first M slots (we assume that $N \leq M$) represent the object identified by the corresponding queries, while the remaining slots represent the unspecified parts of the scene. This way, the obtained representation is a complete decomposition of the whole image X, while still containing the parts that correspond to the user-specified queries L.

We consider two forms of controllability: *language-based controllability* and *point-based controllability*. For language-based controllability, we rely on the user to provide free-form text specifying object category or object referring expressions whose visual representations are sought after. We encode this text into a fixed-size vector embedding using LLM2Vec [17]. We use LLaMA-3-8B [18] to obtain these embeddings. For point-based controllability, we rely on the user to point to the object whose representations are sought after. We extract the coordinates of the corresponding point and embed them into a fixed-size vector using a learnable linear transformation. These language and point embeddings comprise the queries we feed into the model.

2.1 CTRL-O Architecture

Background We build the proposed approach based on DINOSAUR [8]. DINOSAUR uses Slot Attention module [14] as the object discovery module. Slot Attention is an attention-based differentiable clustering procedure which, given a grid of features $H = \{h_k\}_{k=1}^{K}$ obtained from an image encoder f by processing an image $H = f(X)$ (DINOv2 [19] in our case), outputs a set of slots S such that each slot represents a distinct object in the image. We refer the reader to Appendix B for a detailed description of DINOSAUR.

Query-based Slot Initialization The problem we are trying to solve is essentially a visual grounding problem. Given the query corresponding to an object in the image, we want a slot to bind to exactly that object. The most straightforward way to enforce grounding is to condition the slots directly on the query corresponding to each object. Specifically, we achieve this by adding the object query l_i to one of the slots (see Fig. 1, input to the Slot Attention Module). This approach is similar to SAVi [20], which conditions each slot on the center of mass information for each object. However, Kipf et al. [20] do not evaluate if the slots actually bind to the objects specified by the conditioning. In our experiments, we find that simply conditioning the slots on the queries does not lead to correct grounding; hence, a stronger signal is needed to ensure proper grounding.

Decoder Conditioning Similar to DINOSAUR, we use a broadcast MLP decoder to decode the slots into patch features. The decoder decodes each slot separately into patch features. To further encourage

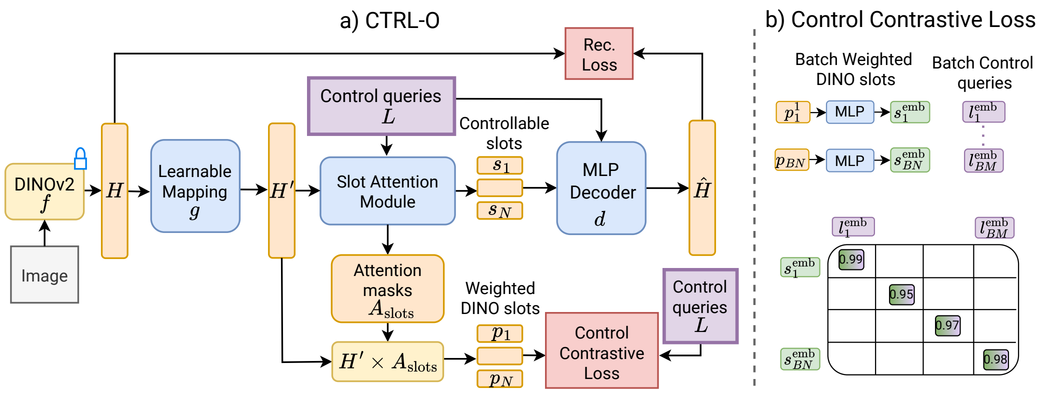

Figure 1: (a) Overview of CTRL-O architecture.

	Slot Init.	GT Masks	Contrastive Loss	Decoder Condn.	binding hits	ARI	mBO
CTRL-O	✓	✓	✗	✗	71.2	69.8	35.4
CTRL-O	✓	✗	✗	✗	8.1	34.52	22.42
CTRL-O	✓	✗	✗	✓	10.11	43.83	25.76
CTRL-O	✓	✗	✓	✗	56.3	44.8	27.3
CTRL-O	✓	✗	✓	✓	61.3	47.5	27.2

Table 1: **Model Component Ablation for Grounding**. This table studies the importance of various components for achieving strong grounding. Here, we use COCO *train* set for training and *val* set for evaluation.

Approach	Model	ARI (%)	mBO (%)
Unsupervised	DINOSAUR (MLP Dec.) [8]	40.5	27.7
	DINOSAUR (TF. Dec.) [8]	34.1	31.6
	Stable-LSD [21]	35.0	30.4
	SlotDiffusion [22]	37.3	31.4
Weakly Supervised	Stable-LSD (Bbox Supervision) [23]	-	30.3
	CTRL-O (Trained on COCO)	47.5	27.2

Table 2: **Segmentation Performance.** Comparison of CTRL-O with unsupervised and weakly supervised object-centric approaches on the COCO dataset.

controllability, we concatenate the slot output from the slot attention with the conditioning query for that slot and then pass the resulting representation through an MLP whose output is then passed to the broadcast decoder as shown in Figure 1 (a). This conditioning helps maintain consistency between the object-specific representation in the slot and the reconstructed output, ensuring that the decoder produces features that are semantically aligned with the intended query.

2.2 Control Contrastive Loss to Enforce Grounding

To enforce grounding, we introduce a novel contrastive loss, as illustrated in Figure 1 (b). The intuition behind this objective as that if a slot s_i is conditioned on a query l_i corresponding to the object o_i, then we want the encoder features corresponding to the slot s_i to be close in embedding space to the query l_i. To obtain the features corresponding to slot s_i, we take the features output of the mapping network (learnable mapping g in Fig. 1 (a)) weighted by the attention scores of slot s_i obtained from the last iteration of slot attention: $p_i = \sum_{k=1}^{K} a_{ik} h_k$, where a_{ik} denotes the attention score of slot s_i on feature h_k. We process p_i using an MLP to output s_i^{emb}, which is used in the contrastive loss. We consider s_i^{emb} and l_i to be a positive pair for the contrastive loss. We compute the similarity between them using cosine similarity. The negatives consist of all (s_i^{emb}, l_t), where $t \neq i$. Note that for the negatives, we consider all the conditioning queries across the entire batch. Considering that there are T conditioning queries in the entire batch, the loss is formalized as follows:

$$\mathcal{L}_{CC} = -\log \frac{\exp(s_i^{\text{emb}} \cdot l_i / \tau)}{\sum_{t=1}^{T} \exp(s_i^{\text{emb}} \cdot l_t / \tau)} \tag{1}$$

Here τ is the temperature, which is set to 0.1. We incorporate this loss in addition to the reconstruction loss from DINOSAUR. For additional implementation details, see Appendix C.

3 Experiments

In this section, we first demonstrate that the grounding problem which we tackle in this paper is not a simple one by showing that even with full supervision i.e. using object mask annotations, models still do not achieve perfect performance. Next, we show that the proposed approach with the contrastive loss achieves decent grounding in complex real-world scenes with limited supervision.

Datasets We use COCO [24] and visual genome [25] as our main datasets to study. COCO consists of natural scenes with multiple objects per scene. Objects in the scene are annotated using category labels and bounding boxes, which provide center of mass coordinates for the queries and the contrastive loss. Visual Genome also consists of equally challenging scenes. Additionally, the objects in the scenes are annotated with referring expressions, which we use as queries for CTRL-O.

Metrics We use the usual metrics such as adjusted rand index (ARI) [26] and mean bounding overlap (mBO) [27] to evaluate the object discovery performance of the model. To measure grounding, we introduce a new metric called *binding hits* which measures accuracy of correct groundings for the conditioned slots. Refer to Appendix D for more details regarding these metrics.

The Grounding Problem Controllability is a new paradigm for object-centric models that have not been explored before. Hence, there are no baselines to which we can compare. Instead, here we try to demonstrate the difficulty of the grounding problem and ablate over the components introduced in Section 2 to understand their importance in achieving good grounding. We use the COCO dataset for this comparison. We use both the center of mass and language categories for the queries. Table 1 presents the results for various ablations. For the model trained with GT masks (row 1 in Table 1), we train it using a reconstruction loss between the predicted and the ground truth masks. Therefore, it is a fully supervised model; hence, it serves as an upper bound for our approach. We can see that it achieves strong segmentation performance (as indicated by ARI and mBO) but still cannot achieve perfect grounding (indicated by Binding Hits), highlighting the difficulty of the grounding problem. Out of the components introduced in Section 2, control contrastive loss is the most crucial for good grounding, followed by decoder conditioning. Without the contrastive loss, the model has no incentive to utilize the queries; hence, it does not achieve good binding.

Comparison to Existing Object-centric Models In Table 2, we compare CTRL-O to existing object-centric models w.r.t segmentation performance. CTRL-O is a weakly supervised approach, as conditioning on language or center-of-mass queries can be considered as a form of weak supervision since we do not require dense labels for every object in an image. We compare to various unsupervised approaches and one weakly supervised approach. We can see that CTRL-O achieves the highest ARI, which means that it can decompose the scenes into objects very well. However, similar to DINOSAUR it achieves a lower mBO as compared to other methods. This means that while it can decompose scenes well, it does not output very sharp masks. We attribute this limitation to the base DINOSAUR model, which also achieves lower mBO. However, CTRL-O is a general approach for producing controllable object-centric representations, and it is not limited to DINOSAUR. Hence, we could apply the proposed approach to other object-centric models like Stable LSD or Slot Diffusion, which produce sharp masks and, thus, stronger mBO.

4 Conclusion

We have introduced CTRL-O, a controllable object-centric model that can be queried to extract representations of specific objects in a scene. Through experimentation, we have shown that representations of specific objects can be extracted in complex real-world scenes based on a range of user queries such as object categories, center of mass points, or referring expressions. This capability expands the applicability of object-centric models to various real-world applications, which we have not explored in this work. For example, one interesting task that CTRL-O enables is instance-based retrieval. Let's say a user is interested in retrieving all images containing a specific object-instance specified by a user-provided image of that object-instance (e.g., more photos of a particular handbag). Since CTRL-O provides a representation specific to that object-instance, this makes it feasible to retrieve all images containing that specific object-instance. This can be very useful in domains such as E-commerce. Furthermore, there can be other applications such as controllable image generation [22, 23] where specific objects could be extracted from multiple images and composed to form a single image. We leave the exploration of these applications for future work.

References

[1] Anirudh Goyal, Aniket Didolkar, Nan Rosemary Ke, Charles Blundell, Philippe Beaudoin, Nicolas Heess, Michael C Mozer, and Yoshua Bengio. Neural production systems. *Advances in Neural Information Processing Systems*, 34:25673–25687, 2021.

[2] Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. SlotFormer: Unsupervised visual dynamics simulation with object-centric models. In *ICLR*, 2023. URL https://openreview.net/forum?id=TFbwV6I0VLg.

[3] Andrii Zadaianchuk, Maximilian Seitzer, and Georg Martius. Self-supervised Visual Reinforcement Learning with Object-centric Representations. In *ICLR*, 2020. URL https://openreview.net/forum?id=xppLmXCbOw1.

[4] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An embodied multimodal language model. In *ICML*, 2023. URL https://arxiv.org/abs/2303.03378.

[5] Aniket Rajiv Didolkar, Anirudh Goyal, and Yoshua Bengio. Cycle consistency driven object discovery. In *ICLR*, 2024. URL https://openreview.net/forum?id=f1xnBr4WD6.

[6] Dan Haramati, Tal Daniel, and Aviv Tamar. Entity-centric reinforcement learning for object manipulation from pixels. In *ICLR*, 2024. URL https://openreview.net/forum?id=uDxeSZ1wdI.

[7] Rim Assouel, Pau Rodriguez, Perouz Taslakian, David Vazquez, and Yoshua Bengio. Object-centric compositional imagination for visual abstract reasoning. In *ICLR2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality*, 2022. URL https://openreview.net/forum?id=rCzfIruU5x5.

[8] 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 *ICLR*, 2023. URL https://openreview.net/forum?id=b9tUk-f_aG.

[9] Aniket Didolkar, Andrii Zadaianchuk, Anirudh Goyal, Mike Mozer, Yoshua Bengio, Georg Martius, and Maximilian Seitzer. Zero-shot object-centric representation learning. *arXiv preprint arXiv:2408.09162*, 2024.

[10] Elizabeth S. Spelke. Core knowledge. *The American psychologist*, 2000. URL https://doi.org/10.1037/0003-066X.55.11.1233.

[11] Steven Pinker. Visual cognition: An introduction. *Cognition*, 1984. URL https://doi.org/10.1016/0010-0277(84)90021-0.

[12] S. M. Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Koray Kavukcuoglu, and Geoffrey E. Hinton. Attend, Infer, Repeat: Fast Scene Understanding with Generative Models. In *NeurIPS*, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/52947e0ade57a09e4a1386d08f17b656-Abstract.html.

[13] Martin Engelcke, Adam R. Kosiorek, Oiwi Parker Jones, and Ingmar Posner. GENESIS: Generative Scene Inference and Sampling with Object-Centric Latent Representations. In *ICLR*, 2020. URL https://openreview.net/forum?id=BkxfaTVFwH.

[14] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-Centric Learning with Slot Attention. In *NeurIPS*, 2020. URL https://proceedings.neurips.cc/paper/2020/file/8511df98c02ab60aea1b2356c013bc0f-Paper.pdf.

[15] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1086. URL https://aclanthology.org/D14-1086.

[16] Kelvin Xu. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2015.

[17] Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*, 2024.

[18] Meta. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

[19] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2023. URL https://arxiv.org/abs/2304.07193.

[20] Thomas Kipf, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional Object-centric Learning from Video. In *ICLR*, 2022. URL https://openreview.net/forum?id=aD7uesX1GF_.

[21] Jindong Jiang, Fei Deng, Gautam Singh, and Sungjin Ahn. Object-centric slot diffusion. In *NeurIPS*, 2023. URL https://arxiv.org/abs/2303.10834.

[22] Ziyi Wu, Jingyu Hu, Wuyue Lu, Igor Gilitschenski, and Animesh Garg. Slotdiffusion: Object-centric generative modeling with diffusion models. In *NeurIPS*, 2023. URL https://arxiv.org/abs/2305.11281.

[23] Krishnakant Singh, Simone Schaub-Meyer, and Stefan Roth. Guided latent slot diffusion for object-centric learning, 2024. URL https://arxiv.org/abs/2407.17929.

[24] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014. URL https://arxiv.org/abs/1405.0312.

[25] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016. URL https://arxiv.org/abs/1602.07332.

[26] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 1985. URL https://link.springer.com/article/10.1007/BF01908075.

[27] Jordi Pont-Tuset, Pablo Arbeláez, Jonathan T. Barron, Ferran Marques, and Jitendra Malik. Multiscale Combinatorial Grouping for Image Segmentation and Object Proposal Generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. URL https://ieeexplore.ieee.org/document/7423791.

[28] Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-Object Representation Learning with Iterative Variational Inference. In *ICML*, 2019. URL https://arxiv.org/abs/1903.00450.

[29] Ondrej Biza, Sjoerd Van Steenkiste, Mehdi SM Sajjadi, Gamaleldin F Elsayed, Aravindh Mahendran, and Thomas Kipf. Invariant slot attention: Object discovery with slot-centric reference frames. *arXiv preprint arXiv:2302.04973*, 2023.

[30] Andrii Zadaianchuk, Maximilian Seitzer, and Georg Martius. Object-centric learning for real-world videos by predicting temporal feature similarities. In *NeurIPS*, 2023. URL https://arxiv.org/abs/2306.04829.

[31] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. *ICCV*, 2021. URL https://arxiv.org/abs/2104.14294.

[32] Avinash Kori, Francesco Locatello, Francesca Toni, and Ben Glocker. Unsupervised conditional slot attention for object centric learning. *arXiv preprint arXiv:2307.09437*, 2023.

[33] Ke Fan, Zechen Bai, Tianjun Xiao, Dominik Zietlow, Max Horn, Zixu Zhao, Carl-Johann Simon-Gabriel, Mike Zheng Shou, Francesco Locatello, Bernt Schiele, Thomas Brox, Zheng Zhang, Yanwei Fu, and Tong He. Unsupervised open-vocabulary object localization in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13747–13755, October 2023.

[34] Dongwon Kim, Namyup Kim, Cuiling Lan, and Suha Kwak. Shatter and gather: Learning referring image segmentation with text supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15547–15557, 2023.

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[36] 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 *ICLR*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[37] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[38] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[39] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021.

[40] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

[41] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1780–1790, 2021.

[42] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[43] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16793–16803, 2022.

[44] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[45] Xiaopeng Zhang, Jiashi Feng, Hongkai Xiong, and Qi Tian. Zigzag learning for weakly supervised object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4262–4270, 2018.

[46] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 350–365. Springer, 2016.

[47] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9215–9223, 2018.

[48] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2601–2610, 2019.

[49] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012), 2012. URL `http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html`.

Appendix

A Related Work

Object-Centric Models Unsupervised object-centric representation learning (OCL) gain a lot of interest in recent years [5, 8, 9, 12–14, 20, 28–30]. OCL aims to extract meaningful entities from unstructured sensory inputs and preserve this separation of information at the representational level. Slot Attention [14] introduces an attention-based mechanism to decompose images into object-centric representations. DINOSAUR [8] builds upon this by utilizing self-supervised DINO features [19, 31] to enhance unsupervised object discovery. While DINOSAUR can effectively identify objects in real-world data [24], it lacks mechanisms for top-down control over the representations. In contrast, the CTRL-O enables controllable OCL by conditioning both the decoder and encoder on control queries using minimal supervision during training for more abstract language-based queries. Some works [20, 32] have explored conditioning in object-centric models. SAVi [20] uses bounding boxes for the initial frame of video data and employs optical flow as a target for conditioning. CoSA [32] conditions on unsupervised vector representations. However, these methods are often limited to specific forms of conditioning and are primarily evaluated on synthetic datasets. Finally, several recent works connect object-centric representations with language. Unsupervised Open-Vocabulary Object Localization [33] and the Shatter and Gather [34] connect object representations with language post-hoc, assigning language labels to discovered slots. By contrast, CTRL-O integrates language and point conditioning directly into the learning process, learning object representations that could be conditioned on user inputs during inference.

Binding in Vision-Language Models Traditional vision-language models rely on holistic feature maps [35] or patch-level representations [36], lacking inherent object-centricity. Object detection models like Mask R-CNN [37] output object-level segmentation and feature vectors but are constrained to predefined categories and lack cross-modal flexibility. Supervised vision-language models (VLMs) such as ViLBERT [38], XVLM [39], Kosmos-2 [40], MDETR [41], UNITER [42], and RegionCLIP [43] integrate region features but depend heavily on dense annotations and treat image regions independently, without maintaining context across the scene. While datasets like COCO [24], RefCOCO [15], and Visual Genome [44] offer rich text-visual alignments for VLMs, weakly supervised binding methods [45–48] reduce the need for labeled data but often operate at a global level or rely on region proposals for a closed set of objects (most commonly PASCAL VOC [49]), limiting flexibility for open-vocabulary tasks. To address these challenges, we introduce a method that decomposes images into object-centric slots conditioned by pre-trained language embeddings and learnable point-based embeddings, enabling robust visual grounding, maintaining context across all objects, and allowing precise manipulation at user-specified granularity through inputs like language descriptions or point prompts.

B DINOSAUR Implementation Details

DINOSAUR uses a DINO [31] encoder to process the image into features. They rely on a feature reconstruction loss to supervise the object discovery process. Throughout the training, the DINO encoder is kept frozen. We adopt a similar approach, however we use a DINOv2 [19] encoder instead of a DINO encoder. Additionally, we have added a learnable mapping network g, which is a 3-layer Transformer after the frozen DINOv2 encoder. SA module is applied on top of the mapping output as shown in Figure 1 (a).

C Implementation Details

Control Contrastive Loss For conditioning, we use either language or point information. However, we assume that each image in our dataset consists of multiple object annotations, each containing a center of mass annotation and a category or referring expression annotation. Therefore, we have two separate contrastive losses - one each for the language information and the point information, as shown in Figure 1 (b).

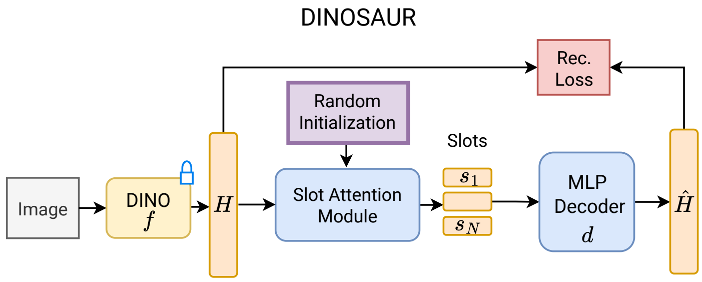

Figure B.1: Overview of DINOSAUR architecture. The image is processed into a set of patch features H by a frozen DINO ViT model. The Slot Attention module groups the encoded features into a set of slots initialized by random queries sampled from the same Gaussian distribution with learnable parameters. By contrast, CTRL-O is initialized by the combination of control queries for conditioned slots and random queries for unconditioned slots. DINOSAUR is trained by reconstructing the DINO features from the slots using MLP decoder [8].

Conditioning We run Slot Attention for a fixed number of slots N. However, in general, we may not have the same number of queries per image. In such cases, we initialize a subset of the slots with the given queries, and the rest are free to bind to any of the other objects in the scene. When computing the contrastive loss, we only consider slots conditioned on some query.

D Metrics

FG-ARI The *adjusted rand index* (ARI) measures the similarity between two clusterings [26]. We use the instance/object masks as the targets. We only compute this metric for pixels in the foreground (hence, FG-ARI). Unlabeled pixels are treated as background.

mBO To compute the mBO [27], each predicted mask is assigned to the ground truth mask with the highest overlap in terms of IoU. The mBO is computed as the average IoU of these mask pairs.

Binding Hits This metric measures controllable grounding. For binding hits, consider that a slot s_i is conditioned on a query L_i identifying an object o_i with ground-truth mask m_i. The broadcast decoder of slot attention outputs a mask per slot. If the overlap between the predicted mask for slot s_i, denoted as $\hat{m}_i$, and the ground truth mask m_i is the highest among all pairs of predicted and ground truth masks, it is considered as a hit (1) else it is considered as a miss (0). binding hits is measured as the average number of hits across the dataset.