OC-CLIP : Object-centric Binding in Contrastive Language-Image Pretraining

Rim Assouel[1,2]*, **Pietro Astolfi**[1], **Florian Bordes**[1], **Michal Drozdzal**[1], **Adriana Romero-Soriano**[1,3,4]
[1]FAIR at Meta, [2]Mila, Université de Montréal, [3]Canada CIFAR AI Chair, [4]McGill University

Abstract

Recent advancements in vision-language models (VLMs) have been driven by contrastive models like CLIP [39], which learn to associate visual information with their corresponding text descriptions. However, these models have limitations in understanding complex compositional scenes involving multiple objects and their spatial relationships. To address these challenges, we propose a novel approach that diverges from traditional data-centric methods of enhancing model performance with hard negatives examples. Our work instead focuses on integrating sufficient inductive biases into pre-trained CLIP-like models to improve their compositional understanding without using additional data annotations. We introduce a binding module that connects a scene graph of the text with an induced graph-like representation of the image, facilitating a structured similarity assessment. We also leverage relationships as text-conditioned visual constraints, thereby capturing the intricate interactions between objects and their contextual relationships more effectively. Our resulting model (OC-CLIP) not only enhances the performance of CLIP in multi-object compositional understanding but also paves the way for more accurate and efficient image-text matching in complex scenes.

1 Introduction

Recent advancements in multi-modal representation learning have primarily been enabled by the introduction of CLIP [39]. CLIP learns aligned image-text representations from Internet-scale data. Despite its success, CLIP exhibits limitations in understanding complex scenes composed of multiple objects [23, 47, 11, 36]. For instance, while capable of recognizing individual objects, CLIP struggles with interpreting spatial relationships among objects in the scene [] (*e.g.*, "the cat is to the left of the mat" *vs.* "the cat is to the right of the mat") and adequately associating objects with their corresponding attributes (*e.g.*, "a red square and a blue circle" *vs.* "a blue square and a red circle"). The process of acquiring this compositional understanding of the world is known as the *binding problem* in the literature, and may be decomposed into *segregation*, *representation*, and *composition* problems [17].

Efforts to improve the compositional understanding of CLIP-like models have largely relied on leveraging *hard negative examples*, either in the text space [22, 48, 54, 11, 36] – to improve sensitivity to the order of words and subtle textual differences – or the image space [3, 25, 53] – to improve sensitivity to subtle visual differences. Although these methods have somewhat improved CLIP-like models' performance on scene compositionality benchmarks [37, 55, 48, 18], they do not explicitly address the binding problem as they focus mainly on enhancing the model's representation capabilities with additional data, hindering their generalization to unseen scene compositions.

Yet, the object-centric representation learning literature [13, 16, 31, 46, 43] has long focused on developing methods to address the segregation and representation problems as a way to facilitate the

*Correspondence to: assouelr@mila.quebec

Submitted to NeurIPS 2024 Workshop on Compositional Learning: Perspectives, Methods, and Paths Forward. Do not distribute.

subsequent compositional processing of images. This has led to the development of inductive biases to segregate different objects in a scene into distinct representational *slots*, which have been shown to naturally scale to an increasing number of visual objects and relations [31, 44, 34, 12].

In this paper, we focus on enhancing the compositional scene understanding of CLIP-like models by leveraging the advances from object-centric representation learning. In particular, we propose to endow CLIP-based vision-language architectures with segregation, representation and composition capabilities. Our core idea is to adapt the slot-centric representation paradigm for CLIP architectures and dynamically align each representational slot with the object entities mentioned in the text. To do so, we design a binding module that connects a scene graph, derived from the textual description, with a slot-structured image representation. We utilize the scene graph's relationships as constraints to effectively capture the complex interactions among the visual entities represented as slots. Our enhanced model, which we refer to as Object-Centric CLIP (OC-CLIP), not only boosts CLIP's performance in understanding multi-object compositional scenes but also improves the accuracy and efficiency of image-text matching in complex and highly compositional visual scenarios.

2 OC-CLIP

Our goal is to enhance CLIP-based architectures with segregation and composition capabilities. Our method starts by extracting representations of distinct objects and relationships in a textual description, as well as representations of patches in an image. Next, a binding module matches the text representation of objects to the relevant image patches, producing a slot-centric representation of the image. Finally, a structured similarity score compares the slot-centric representation with the textual representations of different objects, and leverages the extracted relationships as constraints applied to the visual slots. Our key contributions lie in the design of the binding module and the proposal of the structured similarity score, which we detail below. Figure 1 presents an overview of the proposed approach.

Notation. We denote as $\mathbf{x}$ an image of shape $\mathbb{R}^{h\times w\times 3}$ and as $\bar{\mathbf{x}} = [\bar{\mathbf{x}}^1, ..., \bar{\mathbf{x}}^N] = E_\phi(\mathbf{x}) \in \mathbb{R}^{N\times d}$ its patch-level encoding, where E_ϕ is an image encoder – typically a pre-trained ViT [] – N is the number of patches and d the dimensionality of the patch embeddings. We denote as t the text description, or caption, associated with $\mathbf{x}$. We extract a scene graph, $\mathcal{G}$ from t by leveraging an LLM-based parsing approach. $\mathcal{G}$ is composed of a set of nodes $\mathcal{N} = \{N^1, ..., N^M\}$ representing the M objects in t and of a set of edges $\mathcal{E} = \{(\mathbf{r}^1, s^1, o^1), ..., (\mathbf{r}^P, s^P, o^P)\}$ representing the P relationships in t. Each relationship is represented by a tuple $(\mathbf{r}, s, o)$, where $\mathbf{r}$ is the embedding of the predicate, s the subject and o the object of the relationship. For example, the scene graph of "A red apple to the left of a blue car" will be represented with the set of nodes {"red apple", "blue car"} and the set of edges {("to the left of", "red apple", "blue car")}. In practice, we represent $\mathcal{N}$ as a matrix of node features $\mathbf{N}$, where each row contains the embedding of a node in the graph. Moreover, we represent each s^i and o^i in the relationship tuples as indices referencing the nodes (rows) in $\mathbf{N}$.

Binding Module Our first contribution resides in the binding module. The idea is that when comparing the content of a caption and an image we do not want the features of different objects to interfere with each other but rather keep them separate at a representational level. The role of the binding module is thus to extract a slot-centric representation of an image where the content of the slots are pushed to represent the nodes of the associated scene graph.

To do so, we implement the binding module using a *inverted* cross-attention layer [45], where the queries are the nodes from our scene graph and the keys and values are the image patches. We normalize the attention coefficients over the queries' dimension in order to introduce a competition between queries to explain different parts of the visual input. We follow common practice and set the attention's softmax temperature to $\sqrt{D}$, with D being the dimensionality of the dot-product operation. Applying the softmax along the queries' dimension pushes all the candidate keys to be softly matched to at least one query. However, captions mostly describe specific parts of the image, and rarely capture all the visual information. Since we want only the relevant visual information to be captured by the queries, we add a set of default query tokens, stored in a matrix $\mathbf{Q}_{\text{default}}$, which participate in the competitive attention mechanism – with the goal of absorbing the visual information not captured in the caption. These default query tokens are dropped in the subsequent computation steps of our model (akin to registers in ViT backbones [10]). We find the default query tokens crucial to stabilize the training our model.

The binding module computations are formalized as follows:

$$\mathbf{Q}, \mathbf{K}, \mathbf{V} = \mathbf{W}_q\mathbf{N}, \mathbf{W}_k\mathbf{N}, \mathbf{W}_v\bar{\mathbf{x}} \tag{1}$$

$$\mathbf{Q}' = [\mathbf{Q}; \mathbf{Q}_{\text{default}}],$$

$$\text{Attention}(\mathbf{Q}', \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}' \cdot \mathbf{K}^T}{\sqrt{D}}, \text{dim='queries'}\right) \cdot \mathbf{V},$$

$$\mathbf{S}, \mathbf{S}_{\text{default}} = \text{Attention}(\mathbf{Q}', \mathbf{K}, \mathbf{V}). \tag{2}$$

Here, $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$ are the linear projection weight matrices for the queries, keys, and values, respectively, $\mathbf{S}$ are the visual slots, $\mathbf{S}_{\text{default}}$ are the visual slots from default query tokens, which are discarded for subsequent steps, and [.] denotes the concatenation operation.

Thus, the output of this binding module are the visual slots $\mathbf{S}$. Intuitively, these slots are pushed to represent the visual objects, or entities, that correspond to the nodes of the scene graph. Their object-centric learning is driven by the structured similarity that we detail in the next section.

Structured similarity score Our second contribution resides in the introduction of a structured similarity score, whose goal is to promote the constraints imposed by the scene graph on the learnable visual slots. Our proposed structured similarity score is composed of an *object scoring* function and a *relationship scoring* function. The object scoring function assesses the presence of each node in the scene graph (objects present in the caption). We model this function as the sum of the cosine similarity between each textual node representation $\mathbf{N}^i$ and its assigned visual slot $\mathbf{S}^i$. The relationship scoring function encourages the relational constraints imposed by each edge in the scene graph and is defined as a learnable function f_ϕ of the relationship embedding $\mathbf{r}^i$, and the visual slot representations $\mathbf{S}^{s^i}$ and $\mathbf{S}^{o^i}$ corresponding to the subject and object of the relationship, respectively. We derive the overall structured similarity score over the visual slots $\mathbf{S}$ from an image $\mathbf{x}$ and a graph $\mathcal{G} = (\{N^i\}_{i=1..M}, \{(\mathbf{r}^i, s^i, o^i)\}_{i=1..P})$ such that:

$$S(\mathbf{x}, \mathcal{G}) = \frac{\alpha \sum_{i=1..M} \text{cosine}(\mathbf{N}^i, \mathbf{S}^i) + \beta \sum_{i=1..P} f_\phi(\mathbf{r}^i, \mathbf{S}^{s^i}, \mathbf{S}^{o^i})}{\alpha M + \beta P}, \tag{3}$$

where α and β are parameters controlling the strength of each score. M and P are the number of nodes and relationships in the scene graph $\mathcal{G}$, respectively.

We define f_ϕ as follows:

$$f_\phi(\mathbf{r}, \mathbf{S}^s, \mathbf{S}^o) = \text{cosine}\left(\mathbf{r}, f_s([\mathbf{r}, \mathbf{S}^s]) + f_o([\mathbf{r}, \mathbf{S}^o])\right), \tag{4}$$

where [.] denotes the concatenation of two vectors and f_s and f_o are MLPs that reduce the dimensionality of their inputs. Note that we model the relationship scoring function so that it keeps the same scale as the object scoring function and can take the order of the relationship into account.

Training The model is trained using the following loss:

$$\mathcal{L} = \mathcal{L}_{itc} + \mathcal{L}_{rel}. \tag{5}$$

$\mathcal{L}_{itc}$ is the image-text contrastive loss defined to minimize the distance between image and scene graph representations from paired text-image data while maximizing the distance between image and scene graph representations from unpaired text-image data as:

$$\mathcal{L}_{itc} = -\sum_{i=1}^{B}\left(\log\frac{\exp^{S(\mathbf{x}_i, \mathcal{G}_i)}}{\sum_{j=1}^{B}\exp^{S(\mathbf{x}_j, \mathcal{G}_i)}} + \log\frac{\exp^{S(\mathbf{x}_i, \mathcal{G}_i)}}{\sum_{j=1}^{B}\exp^{S(\mathbf{x}_i, \mathcal{G}_j)}}\right), \tag{6}$$

where B is the number of elements in the batch. Note that the S is the structured similarity score defined in Eq. 3. $\mathcal{L}_{rel}$ is the loss that pushes the model to learn a non-symmetric relationship scores:

$$\mathcal{L}_{rel} = -\sum_{i=1}^{B}\log\frac{\exp^{S(\mathbf{x}_i, \mathcal{G}_i)}}{\exp^{S(\mathbf{x}_i, \mathcal{G}_i)} + \exp^{S(\mathbf{x}_i, \bar{\mathcal{G}}_i)} + \exp^{S(\mathbf{x}_i, \tilde{\mathcal{G}_i})}}, \tag{7}$$

where $\bar{\mathcal{G}}$ and $\tilde{\mathcal{G}}$ are altered scene graphs. In $\bar{\mathcal{G}}$, we swap the order of the subject and the object of a relationship, whereas in $\tilde{\mathcal{G}}$, we randomly chose the relationship's subject and object from the nodes in the scene graph.

3 Results

Setting We train OC-CLIP and finetune OpenCLIP in-domain on a set of datasets relevant for real-world compositional understanding. The training text descriptions representing positive samples are taken from COCO [27], Visual-Genome [24] and GQA [20]. The latter annotates images coming from Visual Genome [24] with objects and both spatial and non-spatial relationships, and thus contains a high representation of spatial prepositions. We evaluate the different models on the most challenging benchmarks representative of compositional understanding, ensuring that we validate both their *attribute binding* and *spatial relationship* understanding capabilities. In particular, we use SugarCrepe [18] and ARO-Attribution (ARO-A) [47] for attribute binding and ARO-Relation (ARO-R) [47], COCO-spatial and GQA-spatial [23] for spatial relationship understanding. The training of the OC-CLIP's binding module is done from scratch along with the finetuning of the text and vision backbones. The text backbone is initialized from OpenCLIP weights [21]. We consider 2 different image base ViT backbones, OpenCLIP (ViT-B-16) [21] and Dinov2 (ViT-B-14) [35], to show the flexibility of our binding module and learned structured similarity score.

Attribute Binding We evaluate the attribute binding capabilities of OC-CLIP and baselines on SugarCrepe [18] and ARO-A [48] benchmarks. We report the results in Table 1. When comparing OpenCLIP-FT to OC-CLIP (both models), we observe notable performance boosts on ARO-A and SugarCrepe's swap-attribute, and swap-object. In particular, OC-CLIP $_{\text{B-14}}$ shows a performance boost of +22.1% on ARO-A, whereas in SugarCrepe, our model reaches improvements of +16.1% on the swap-attribute split, +17.7% on the swap-object split. When comparing with additional contrastive-based models (BLIP and XVLM) finetuned with in-domain data, both OC-CLIP models show notable improvements on SugarCrepe's swap splits – *e.g.*, OC-CLIP $_{\text{B-14}}$ results in +14.6% in object-swap and +12.3% in attribute-swap – despite not relying on additional binding annotations, nor language modeling losses. The results of BLIP and XVLM on ARO-A may be explained by the use of their use of a language modeling prior; It is shown in [18] that language-only models are performing well on this benchmark because the negative caption are often not realistic.

Spatial Relationship Understanding We also evaluate the spatial relationship understanding capabilities of OC-CLIP and baselines on COCO-spatial, GQA-spatial, and ARO-Relation (ARO-R). Note that ARO-Relation contains both spatial and non-spatial relations but about half of the test examples consists of left/right relationships understanding. We report the results in Table 1 and show consistent improvements of both OC-CLIP models over the baseline models and across the 3 datasets. In particular, the best OC-CLIP model outperforms OpenCLIP-FT by +47.9% on COCO-spatial, +46.6% on GQA-spatial, and +34.7% on ARO-R. When compared to contrastive VLMs finetuned with in-domain data (XVLM, BLIP), OC-CLIP models exhibit superior performance, with improvements between +10% and +27% over the strongest contrastive finetuned VLM. Finally, when compared to baselines leveraging hard-negatives (NegCLIP), OC-CLIP remains the highest performer.

	WhatsUp		**SugarCrepe**		**ARO**	
Model	COCO-spatial	GQA-Spatial	swap-obj	swap-att	Att	Rel
OpenCLIP-FT	45.6	49.1	63.1	72.4	59.9	50.1
XVLM [49]	73.6	67	64.9	73.9	86.8	73.4
BLIP 26	56.4	52.6	66.2	76.2	**88.0**	59.0
NegCLIP [47]	46.4	46.7	75.2	75.4	70.5	80.2
OC-CLIP $_{\text{B-16}}$	90.1	93.9	76.3	87.1	80.3	83.7
OC-CLIP $_{\text{B-14}}$	**93.5**	**95.6**	**80.8**	**88.5**	82.0	**84.8**

Table 1: **Compositional Understanding**: Performance on the hardest SugarCrepe, What's Up and ARO Splits. Both OpenCLIP-FT and OC-CLIP are initialized with the same OpenCLIP checkpoints. OC-CLIP is trained with two ViT base backbones with different resolutions: OpenCLIP's backbone (B-16) and Dinov2 (B-14).

4 Conclusion

We propose OC-CLIP, a method that enhances the compositional scene understanding of CLIP-like models by leveraging object-centric representation learning. The results show that OC-CLIP significantly improves performance on challenging real-world compositional image-text matching benchmarks, such as SugarCrepe and Whatsup. Future work could explore ways to improve the scalability of the approach when trained from scratch with noisy alt-text based datasets.

References

[1] Rameen Abdal, Peihao Zhu, John Femiani, Niloy J. Mitra, and Peter Wonka. Clip2stylegan: Unsupervised extraction of stylegan edit directions, 2021.

[2] Rim Assouel, Lluis Castrejon, Aaron Courville, Nicolas Ballas, and Yoshua Bengio. VIM: Variational independent modules for video prediction. In Bernhard Schölkopf, Caroline Uhler, and Kun Zhang, editors, *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 70–89. PMLR, 11–13 Apr 2022.

[3] Rabiul Awal, Saba Ahmadi, Le Zhang, and Aishwarya Agrawal. Vismin: Visual minimal-change understanding, 2024.

[4] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. Paligemma: A versatile 3b vlm for transfer, 2024.

[5] Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Making large multimodal models understand arbitrary visual prompts. In *CVPR 2024*, 2024.

[6] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.

[7] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

[8] Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. Fine-grained image captioning with clip reward, 2023.

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[10] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2024.

[11] Sivan Doveh, Assaf Arbelle, Sivan Harary, Rameswar Panda, Roei Herzig, Eli Schwartz, Donghyun Kim, Raja Giryes, Rogerio Feris, Shimon Ullman, and Leonid Karlinsky. Teaching structured visionlanguage concepts to visionlanguage models, 2023.

[12] Gamaleldin 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 S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 28940–28954. Curran Associates, Inc., 2022.

[13] 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, 2016.

[14] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters, 2021.

[15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017.

[16] Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Chris Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference, 2020.

[17] Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks, 2020.

[18] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality, 2023.

[19] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In *NeurIPS 2023*, 2023.

[20] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[21] 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, 2021.

[22] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning, 2020.

[23] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning, 2023.

[24] 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(1):32–73, 2017.

[25] Tiep Le, Vasudev Lal, and Phillip Howard. Coco-counterfactuals: Automatically constructed counterfactual examples for image-text pairs. In *NeurIPS 2023*, 2023.

[26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

[27] 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 David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014.

[28] Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Revisiting the role of language priors in vision-language models. *arXiv preprint arXiv:2306.01879*, 2024.

[29] 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 2024.

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS 2023*, 2023.

[31] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention, 2020.

[32] Jan Hendrik Metzen, Piyapat Saranrittichai, and Chaithanya Kumar Mummadi. Autoclip: Auto-tuning zero-shot classifiers for vision-language models, 2024.

[33] Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors. *CLIP-Score: A Reference-free Evaluation Metric for Image Captioning*, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[34] Shanka Subhra Mondal, Jonathan D. Cohen, and Taylor W. Webb. Slot abstractors: Toward scalable abstract visual reasoning, 2024.

[35] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024.

[36] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *ICCV 2023*, 2023.

[37] Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 8253–8280. Association for Computational Linguistics, 2022.

[38] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023.

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[40] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[41] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[42] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022.

[43] 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, 2023.

[44] Taylor Webb, Shanka Subhra Mondal, and Jonathan D Cohen. Systematic visual reasoning through object-centric relational abstraction. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 72030–72043. Curran Associates, Inc., 2023.

[45] Yi-Fu Wu, Klaus Greff, Google Deepmind, Gamaleldin F. Elsayed, Michael C. Mozer, Thomas Kipf, and Sjoerd van Steenkiste. Inverted-attention transformers can learn object representations: Insights from slot attention.

[46] Ziyi Wu, Jingyu Hu, Wuyue Lu, Igor Gilitschenski, and Animesh Garg. Slotdiffusion: Object-centric generative modeling with diffusion models, 2023.

[47] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it?, 2023.

[48] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023.

[49] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 25994–26009. PMLR, 2022.

[50] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts, 2022.

[51] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning, 2022.

[52] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022.

[53] Jianrui Zhang, Mu Cai, Tengyang Xie, and Yong Jae Lee. Countercurate: Enhancing physical and semantic visio-linguistic compositional reasoning via counterfactual examples, 2024.

[54] Le Zhang, Rabiul Awal, and Aishwarya Agrawal. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding, 2024.

[55] Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*, 2022.

A Appendix / supplemental material

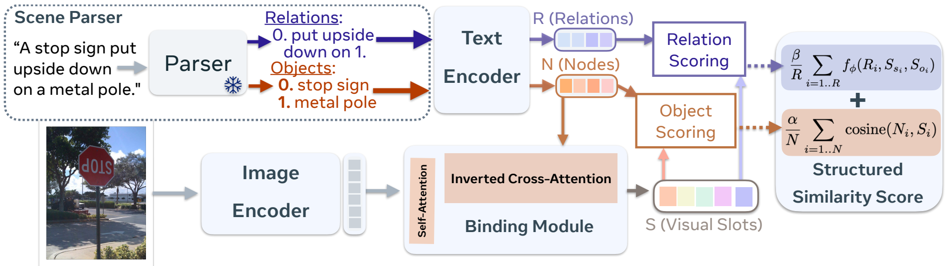

Figure 1: Object-Centric CLIP (OC-CLIP) overview.

A.1 Related Work

Contrastive Pretraining of VLMs. Vision-language models (VLMs) have made substantial strides in both the vision and multi-modal domains. Modern VLMs are pretrained on vast, diverse and oftentimes noisy multi-modal datasets [6, 42, 21, 50] and applied to various zero-shot tasks. CLIP [39] presented a contrastive learning approach used for pretraining, which involves training the model to differentiate between similar and dissimilar image-text pairs. This approach encourages the model to learn a shared representation space for images and text, where semantically similar pairs are close together and dissimilar pairs are far apart. Following CLIP's lead, image-text contrastive learning has become a prevalent strategy for VLM pretraining [30, 5, 29, 9, 51, 7, 4]. Contrastive vision-language pretraining spans numerous downstream applications, including zero-shot image classification [52, 39, 32, 14], text-to-image generation [38, 1, 40, 41], as well as assessing text-image alignment [33, 8]. In this work we are particularly interested the ability of CLIP-based VLMs to evaluate compositional text-image alignment.

Compositional Understanding Benchmarks. Several benchmarks have been developed to assess the compositional understanding of VLMs. In this work, we focus on benchmarks structured as cross-modal retrieval tasks where the model needs to distinguish between correct and incorrect text descriptions given an image, and evaluations are based on accuracy metrics. The majority of these benchmarks [55, 47, 37] rely on the rule-based construction of negative captions and the generation of their associated image counter-factuals [53, 3]. Yet, many of these benchmarks may be solved by leveraging the language prior exclusively [15, 28], hence disregarding the information from the visual input. To address this, benchmarks such as SugarCrepe [19] leverage large language models to generate plausible and linguistically correct hard negatives, and show that previously introduced text-based hard negative strategies are not always effective [48] – *e.g.*, when considering attribute and object swaps between textual descriptions. Other benchmarks focus on assessing the VLMs' spatial understanding [23, 48, 53], and propose to finetune CLIP-based models on data containing a high proportion of spatial relationships since these relationships tend to underrepresented in commonly used pretraining datasets. Interestingly, **(author?)** [23] show that even when finetuning with in-domain data with an overrepresentation of spatial relationships, state-of-the-art models still exhibit a close to random chance performance. In this work, we test the hypothesis that spatial relationship failures are due to the lack composition in the similarity score computation used to train CLIP-like models.

Object-centric Binding Inductive Biases. CLIP has been shown [47] to be pushed to learn disentangled, bag-of-words-style representations from the contrastive loss and the easily distinguishable negatives typically used for pretraining. Although the learned representations might be effective for objects presented in isolation, they struggle with scenes containing multiple objects []. For example, consider a simple scene with a green apple and a yellow banana. In this case, the model must maintain and correctly link the attributes ("green", "yellow") to the objects ("apple", "banana"), without mixing the concepts – *e.g.*, "yellow apple" or 'green banana". This exemplifies the importance of devising robust mechanisms within the CLIP architecture and/or training to accurately handle multiple objects, while preventing feature interferences. In this work, we focus on equipping CLIP with object-centric binding inductive biases and take inspiration from the architectures proposed in the unsupervised object-centric visual representation learning literature [31, 46, 43, 2]. Many recent image-only approaches follow a simple inductive bias introduced by slot Attention [31], where an image – encoded as a set of input tokens – is soft partitioned into K slots. In particular, attention maps are computed via a **inverted cross attention** mechanism [45], where the softmax is applied along the query dimension in order to induce a competition between the slots to explain different groups of input tokens. In this work, we extend these inductive biases to define text-conditioned visual slots from the input image.

A.2 More Compositional Results

We evaluate the attribute binding capabilities of OC-CLIP and baselines on SugarCrepe [18] and ARO-A [47] benchmarks. We report the results in Table 2. When comparing OpenCLIP-FT to OC-CLIP (both models), we observe notable performance boosts on ARO-A and SugarCrepe's swap-attribute, and swap-object. In particular, OC-CLIP B-14 shows a performance boost of +22.1% on ARO-A, whereas in SugarCrepe, our model reaches improvements of +16.1% on the swap-attribute split, +17.7% on the swap-object split, and a smaller +4.7% on the replace-relationship split. Moreover, both OC-CLIP models perform similarly to OpenCLIP-FT on the remaining SugarCrepe splits. This is to be expected since the remaining splits do not require precise binding to distinguish between positive and negative captions and may therefore be solved with a bag-of-words-like representation. When comparing with additional contrastive-based models (BLIP and XVLM) finetuned with in-domain data, both OC-CLIP models show notable improvements on SugarCrepe's swap splits – *e.g.*, OC-CLIP B-14 results in +14.6% in object-swap and +12.3% in attribute-swap – despite not relying on additional binding annotations, nor language modeling losses. The results of BLIP and XVLM on ARO-A may be explained by the use of their use of a language modeling prior; **(author?)** [19] emphasizes that language-only models are performing well on this benchmark because the negative caption are often not realistic. Both OC-CLIP models also improve the results of hard-negative-based methods on SugarCrepe's swap splits as well as ARO-A. In all the remaining splits of SugarCrepe, except add-attribute, OC-CLIP models perform similarly to previous works leveraging hard-negatives. The results achieved by CE-CLIP and CC-CLIP on the add-attribute split could be attributed to an increase of attribute coverage induced by the language model generations.

Model	SugarCrepe – Swap		SugarCrepe – Add		SugarCrepe – Replace			ARO-A
	Object	Attribute	Object	Attribute	Object	Attribute	Relation	
Zero-shot								
OpenCLIP	68.2	66.2	82.7	80.3	93.8	82.8	67.3	58.8
In-domain ft baselines								
BLIP 26†	66.2	76.2	-	-	**96.5**	81.9	68.35	**88.0**
XVLM **(author?)** 49 †	64.9	73.9	-	-	95.2	87.7	77.4	86.8
OpenCLIP-FT	63.1 ±0.6	72.4±1.1	**93.4** ±0.2	83.1 ±0.5	95.4	87.0 ±0.6	75.5 ±0.6	59.9 ±0.2
Hard-Negative based baselines								
NegCLIP [47]†	75.2	75.4	88.8	82.8	92.7	85.9	76.5	70.5
CE-CLIP [54]†	72.8	77	92.4	**93.4**	93.1	88.8	79	76.4
CC-CLIP [53]†	68.6	73.6	86.7	90.3	**95.9**	87.9	76.2	-
Ours								
OC-CLIP B-16	76.3 ±0.7	87.1 ±0.2	91.3	83.8 ±1.0	93.9 ±0.4	88.3 ±0.1	77.0 ±0.2	80.3 ±0.1
OC-CLIP B-14	**80.8** ±0.7	**88.5** ±0.4	93.0±0.3	83.8 ±1.1	95.7 ±0.4	**88.8** ±0.6	**80.2** ±0.2	82.0

Table 2: **Attribute binding: Performance on SugarCrepe and ARO-Attribution (ARO-A).** Both OpenCLIP-FT and OC-CLIP are initialized with the same OpenCLIP checkpoints. OC-CLIP is trained with two ViT base backbones with different resolutions: OpenCLIP's backbone (B-16) and Dinov2 (B-14).

For the parsing of the training and testing data we used a llama-3-70b Instruct model with the following prompt :

Parsing Prompt

Given a caption, your task is to parse it into its constituent noun phrases and relationships. The noun phrases should represent independent visual objects mentioned in the caption without semantic oversimplification. For each caption, output the parsed noun phrases (e.g., entities) and relationships in JSON format, placing the dictionary between `[ANS]` and `[/ANS]` brackets. In the relationships, use indices to specify the subject and object of the relationship mentioned in the caption. The indices of the subject and object should be integers. Here are a few examples:

```
Caption: A large brown box with a green toy in it
Output:
[ANS]
{
  "entities": [
    "large brown box",
    "green toy"
  ],
  "relationships": [
    {
      "relationship": "in",
      "subject": 1,
      "object": 0
    }
  ]
}
[/ANS]

[...]   More examples
```

PAY ATTENTION to the following:
- Relationships MUST relate two different entities in the caption and NOT be unary. For example, in the caption 'red suitcases stacked upon each other', 'stacked upon each other' is not considered a relationship.
- Do not forget any relationships.
- Relationships MUST be directed. 'and' is not a relationship.
- Pay attention to spatial relationships like 'behind', 'left of', 'with', 'below', 'next to', etc. 'and' is not a relationship.
- Check the right dependencies when the relationships are not direct. In the caption template a X with a Y in it, it refers to X.
- Pay attention to co-references.

Now, parse the following caption into its constituting entities and relationships. You MUST place the answer between `[ANS]` and `[/ANS]` delimiters.
Caption: