SPARO: Selective Attention for Robust and Compositional Transformer Encodings for Vision

Ankit Vani[1⋆], Bac Nguyen[2], Samuel Lavoie[1], Ranjay Krishna[3], and Aaron Courville[1,4]

[1] Mila, Université de Montréal
[2] Sony AI
[3] University of Washington, Allen Institute for Artificial Intelligence
[4] CIFAR AI Chair

Abstract. Selective attention helps us focus on task-relevant aspects in the constant flood of our sensory input. This constraint in our perception allows us to robustly generalize under distractions and to new compositions of perceivable concepts. Transformers employ a similar notion of attention in their architecture, but representation learning models with transformer backbones like CLIP and DINO often fail to demonstrate robustness and compositionality. We highlight a missing architectural prior: unlike human perception, transformer encodings do not separately attend over individual concepts. In response, we propose SPARO, a read-out mechanism that partitions encodings into separately-attended slots, each produced by a single attention head. Using SPARO with CLIP imparts an inductive bias that the vision and text modalities are different views of a shared compositional world with the same corresponding concepts. Using SPARO, we demonstrate improvements on downstream recognition, robustness, retrieval, and compositionality benchmarks with CLIP (up to +14% for ImageNet, +4% for SugarCrepe), and on nearest neighbors and linear probe for ImageNet with DINO (+3% each). We also showcase a powerful ability to intervene and select individual SPARO concepts to further improve downstream task performance (up from +4% to +9% for SugarCrepe) and use this ability to study the robustness of SPARO's representation structure. Finally, we provide insights through ablation experiments and visualization of learned concepts.

Keywords: Selective attention · Slot representations · Transformers

1 Introduction

Selective attention is an intrinsic property of human perception [15, 44, 47], enabling people to focus on task-relevant aspects of their surroundings while tuning out the rest [55]. For instance, it allows a driver to focus on traffic signals, road signs, and other vehicles, or an individual searching for their friend

⋆ Correspondence at: `ankit.vani@umontreal.ca`.

in a crowd to attend to details like people's heights and clothing. More broadly, it empowers us to comprehend the vast complexity of our world with limited mental resources [6] by constraining our perception to one concept at a time. When a task requires information for multiple concepts (*e.g.*, looking at traffic signals and pedestrians), this constraint calls for sequentially attending to *separate* salient aspects of our stimuli [59]. The cognitive bottleneck of selective attention, with the ability to separately represent concepts from our environment, enables human perception to remain insensitive to irrelevant distractions and generalize to new compositions of perceivable concepts.

In machine learning, the notion of selective attention is party emulated by the attention mechanism [3], which is a core component of transformers [60] such as the vision transformer (ViT) [18]. However, transformer encodings learnt using approaches like CLIP [50] and DINO [9] still struggle with robustness [30, 31, 62, 70] and compositional generalization [33, 42, 52, 58, 67, 72]. For instance, despite CLIP emerging as the de facto backbone for many vision tasks, it can perform close to chance when evaluated for compositionality [42]. While the attention mechanism equips these models with the ability to emphasize relevant aspects of their intermediate representations, it does not endow them with the means to produce encodings for separately-attended concepts. Consequently, transformer-based representation learning models lack an important prior from human cognition: that perception of a complex input can be broken down into perception of its salient concepts.

This limitation deprives downstream tasks easy access to task-relevant aspects of the data's underlying structure. In a task like ImageNet classification, where most images contain one primary object of interest (*e.g.*, "dog"), information like relationships between objects and attributes of the background are irrelevant and can lead to overfitting if considered. However, these details can become crucial to attain high performance in other tasks like image retrieval (*e.g.*, "dog chasing a frisbee in a park"). Similar to the generalization benefits of selective attention in human perception, encodings that allow easy separation of attended concepts in the data can generalize to a diverse set of downstream tasks. Without additional constraints, however, the emergence of such compositional structure in encodings is unlikely, especially in the common setting of training on noisy internet data.

We propose **Sparo** (**Se**p**a**rate-head **a**ttention **r**ead-**o**ut)[5] as an improved read-out mechanism for transformer encoders in vision inspired by selective attention. Sparo replaces the last transformer block to provide a mechanism for partitioning encodings into *slots* of separately-attended concepts. We design each slot encoding to be a low-dimensional result of a *single-head* attention operation, producing a bottleneck that encourages each slot to "selectively attend." In this mechanism, producing multiple slots can be interpreted as selective attention over different concepts occurring in parallel. When training CLIP, Sparo imparts an inductive bias that both modalities are views onto a shared compositional world with the corresponding slots of both encoders representing its concepts.

[5] Source code: https://github.com/ankitkv/sparo-clip.

In contrast, standard CLIP merely embeds the encodings of both modalities in a shared vector space without imposing any additional structure.

Training SPARO using CLIP and DINO exhibits improved generalization, robustness, and compositionality while retaining the same model size. Our experiments show that using CLIP with SPARO improves generalization for zero-shot recognition, robustness, retrieval, and compositionality (*e.g.*, +14% for ImageNet, +4% for SugarCrepe, when trained with 15M examples; +3% each when trained with 400M examples). We also demonstrate improved linear probe performance of CLIP (+10% trained with 15M examples; +2% with 400M examples), as well as both linear probe and nearest neighbors performance of DINO (+3% each). We showcase how SPARO's representational structure enables the ability to manually intervene and select relevant concepts for downstream tasks, leading to improved compositionality and generalization (up from +4% to +9% for SugarCrepe trained with 15M examples; +3% to +6% with 400M examples). We then study the robustness of this structure, validate the choice of separate-head attention through ablations, and provide insights into the effect of the number and dimensionality of concepts. Finally, we provide visualizations for some of the concepts SPARO learns to attend to.

2 Related work

We situate our work amongst other transformer read-out mechanisms and slot representation learning methods.

Transformer read-outs. For transformer [60] decoders with causal masking, the last hidden state, corresponding to the end of sequence (`EOS`) input token, is used as the output representation. BERT [17] proposed inserting a special `CLS` token before input sequences for transformer encoders where information from all positions is gathered to produce the encoding. Using the `CLS` token output is also popular for Vision transformer (ViT) [9, 12, 18] backbones, along with alternatives such as global average pooling [12, 64]. In the context of language, Multi-CLS BERT [10] argues that an input text sequence can have multiple facets, and that samples can be similar along one facet and dissimilar along another. The authors aim to represent these facets by adding multiple `CLS` tokens to the input sequence. Using separate attention heads for each SPARO slot can also encourage attending to different facets of the inputs, but without requiring expensive accumulations through each layer of the backbone. By keeping SPARO independent of the backbone, we also allow for its easy potential placement on top of frozen pre-trained models without needing to train for added `CLS` tokens.

SPARO is most related to the attentional pooler [39, 66] architecture, which uses multi-head attention with embedded learned queries to produce output encodings. However, SPARO employs a critical structural constraint of not mixing the information from separate attention heads, a design choice we justify empirically in Sec. 5.3. Slot attention [41] also produces slot encodings through the attention mechanism, but rather than having the input positions compete for each slot, the slots compete for each input position. Slot attention also proposes an expensive iterative read-out mechanism where the role of each slot is determined by sampling from a per-sample Gaussian distribution, which can be limiting for large real-world datasets.

Slot representations. Object-centric slot representations have been studied extensively in the literature [7,8,20,21,24–27,41,43,71]. Often, these representations are expected to represent predefined objects and attributes, and evaluated in small-scale or synthetic settings. Recently, the use of slot attention on features produced by a learned encoder [1, 49, 56], and utilizing unsupervised saliency masks for object discovery [69] has enabled unsupervised learning of object-centric representations in larger scale settings. Related to these ideas, Sparo attends to features produced by a transformer using a single attention head per slot, and each head is capable of learning saliency maps without supervision [9]. However, we do not constrain our slots to correspond to predefined types of objects in the data, nor do we impose any independence conditions between the slots. Self-supervised simplicial embeddings (SEM) [37], with only a soft-discrete slot structure constraint, enable improved generalization on downstream tasks. Sparo replaces the separation of softmax application between SEM slots with a separation of underlying attention operations.

3 Method

Having motivated our approach, we now detail the design of Sparo. We start by introducing the relevant notation, then discuss the Sparo module. Finally, we discuss the addition of Sparo in CLIP and DINO, and the applicability of Sparo beyond these settings.

3.1 Notation

We envision Sparo as a special attention-based layer that modifies the top block of a transformer encoder [51,60] backbone (including the ViT [18]). Transformers take as input a sequence $\boldsymbol{x} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ of length n, and produce an output state per position $\{\boldsymbol{h}_1, \dots, \boldsymbol{h}_n\}$, where $\boldsymbol{x}_i \in \mathbb{R}^{\text{input dim}}$, $\boldsymbol{h}_i \in \mathbb{R}^d$, for any $i \in [n]$, with d the model width. Finally, a pooling operation reduces the sequence of output states to output encodings with a fixed size $\boldsymbol{y} \in \mathbb{R}^M$. Examples of pooling operations include attentional pooling [39], global average pooling, and extracting the `CLS` or `EOS` token representations for images or text respectively.

3.2 Separate-head attention read-out (Sparo)

We replace the pooling operation of transformer encoders with a concatenation of outputs of L single-head attention mechanisms. Sparo acts on the transformer outputs $\boldsymbol{H} = \{\boldsymbol{h}_1, \dots, \boldsymbol{h}_n\} \in \mathbb{R}^{n \times d}$ to produce the encoder output

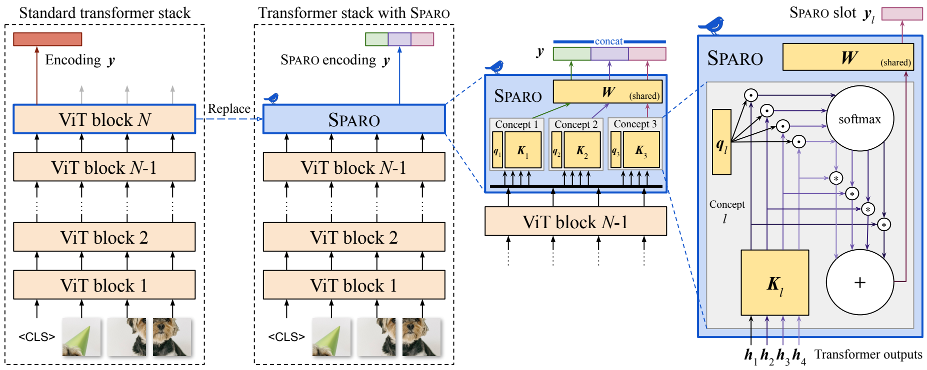


Fig. 1: Illustration of SPARO, a read-out mechanism that structures representations as collections of separately-attended concepts. Take a standard N-block transformer encoder (ViT here as an example), producing an encoding $\boldsymbol{y}$ through extraction of its `CLS` token output. We can replace the Nth transformer block with the SPARO module (typically with equal or fewer parameters) to produce a SPARO encoding $\boldsymbol{y}$, which is a concatenation of L SPARO slots. Each SPARO slot $\boldsymbol{y}_l$ is produced through single-head attention over the backbone outputs using an embedded query $\boldsymbol{q}_l$. The value projection is a composition of slot-specific key projection parameterized by $\boldsymbol{K}_l$ and a slot-wise projection shared between all SPARO concepts parameterized by $\boldsymbol{W}$.

$\boldsymbol{y} = \text{concat}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_L) \in \mathbb{R}^{LV}$, where SPARO slot $\boldsymbol{y}_l \in \mathbb{R}^V$, $l \in [L]$. We illustrate the architecture of SPARO in Fig. 1. Each SPARO slot is produced as:

$$\boldsymbol{y}_l = \boldsymbol{W}\boldsymbol{K}_l\boldsymbol{H}^\top \text{softmax}\left(\frac{\boldsymbol{H}\boldsymbol{K}_l^\top \boldsymbol{q}_l}{\sqrt{D}}\right). \tag{1}$$

Here, $\boldsymbol{K}_l \in \mathbb{R}^{D\times d}$ is a learned key projection weight, and $\boldsymbol{q}_l \in \mathbb{R}^D$ is a learned query embedding for attention. We parameterize the value projection as $\boldsymbol{W}\boldsymbol{K}_l$ where $\boldsymbol{W} \in \mathbb{R}^{V\times D}$ is learned parameter shared between all L slots. Decomposing the value projection in this manner helps reduce the number of model parameters, allowing a larger choice of L with fixed resources.

Generally, we pick V and D to be significantly smaller than the transformer width. In practice, we use the typical value of $D = 64$ used in standard multi-head attention modules, and set V to the same value in our experiments. Therefore, each SPARO slot has limited expressivity and representational capacity. The bottleneck underpinning each SPARO slot is a result of these constraints and the mechanistic bottleneck of a single attention mechanism instantiating competition between its input features.

SPARO can be thought of as multi-head attention [60] where the queries are learned embeddings, key and value projection weights are shared, and the output projection weight is block-diagonal with each block containing the same parameters. Assuming $L = V = D = \sqrt{d}$, SPARO requires a total of $d^2 + 2d$ parameters, compared to multi-head attention's $4d^2$.

3.3 CLIP with Sparo

We add Sparo to both the image and text encoders of CLIP [50], setting the same values of L and V in both encoders. Each Sparo slot is ℓ_2-normalized separately, and we divide their concatenation by $\sqrt{L}$ to produce the global ℓ_2-normalized encoding. With this scheme, the cosine similarity between the image encoding $\boldsymbol{y}^{\mathrm{i}}$ and the text encoding $\boldsymbol{y}^{\mathrm{t}}$ becomes the expected cosine similarity between corresponding slot encodings:

$$\boldsymbol{y}^{\mathrm{i}\top}\boldsymbol{y}^{\mathrm{t}} = \sum_{l=1}^{L}\|\boldsymbol{y}_l^{\mathrm{i}}\|\|\boldsymbol{y}_l^{\mathrm{t}}\|\frac{\boldsymbol{y}_l^{\mathrm{i}\top}\boldsymbol{y}_l^{\mathrm{t}}}{\|\boldsymbol{y}_l^{\mathrm{i}}\|\|\boldsymbol{y}_l^{\mathrm{t}}\|} = \sum_{l=1}^{L}\frac{1}{L}\mathrm{sim}(\boldsymbol{y}_l^{\mathrm{i}},\boldsymbol{y}_l^{\mathrm{t}}) = \mathop{\mathbb{E}}_{l\sim[L]}\left[\mathrm{sim}(\boldsymbol{y}_l^{\mathrm{i}},\boldsymbol{y}_l^{\mathrm{t}})\right]. \quad (2)$$

We train the model using the standard CLIP loss which maximizes $\mathrm{sim}(\boldsymbol{y}^{\mathrm{i}},\boldsymbol{y}^{\mathrm{t}})$ of encodings $\boldsymbol{y}^{\mathrm{i}}$ and $\boldsymbol{y}^{\mathrm{t}}$ for aligned image-text pairs and minimizes it across all other pairings. With Sparo's separate-head attention bottleneck, maximizing $\mathbb{E}_{l\sim[L]}\left[\mathrm{sim}(\boldsymbol{y}_l^{\mathrm{i}},\boldsymbol{y}_l^{\mathrm{t}})\right]$ for aligned pairs across the training dataset encourages both modalities to learn similar attention semantics per slot $l \in [L]$. Consequently, Sparo imparts a prior that both modalities encode a shared world with L selectively attendable concepts, and similarities between inputs can be expressed as their average similarity for these concepts.

3.4 DINO with Sparo

The online and momentum encoders of DINO [9] are each comprised of a ViT backbone which produces the encoding, and a DINO head that transforms it into a categorical distribution for distillation. Unlike CLIP, these encoders have an inherent bias for learning similar encoding functions due to one being an exponentially moving average of the other. However, in standard DINO, there are no constraints for selective attention in the encoding structure. We add Sparo to DINO to impart the prior for representing separately-attendable concepts in its encodings. Concretely, we replace the final transformer block of the ViT backbone with the Sparo, but do not modify the DINO head.

4 Results

In Secs. 4 and 5, we evaluate our method on a variety of downstream tasks, analyze Sparo's representation structure with the ability to intervene on selected slots, perform ablations to support our design, and visualize the learned concepts.

In this section, we validate that partitioning the representation structure as a collection of separately-attended concepts leads to improved generalization for downstream recognition, robustness, compositionality, and retrieval, using Sparo with CLIP [50] in Secs. 4.1 to 4.3 and DINO [9] in Sec. 4.3.

Datasets. We train our CLIP models on one of Conceptual Captions 3M (CC3M) [57], Conceptual Captions 12M (CC12M) [11], a combination of CC3M and CC12M (CC15M), or LAION-400M (L400M) [54]. Appendix C provides the statistics for our training datasets.

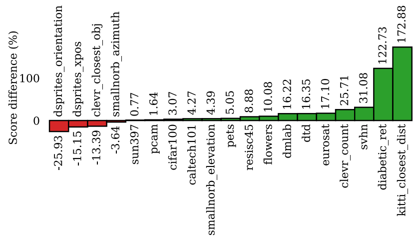
Table 1: Zero-shot recognition, robustness, and SugarCrepe compositionality for CLIP models trained on Conceptual Captions and LAION-400M.

Train	Model	ImageNet- V1	V2	Sketch	A	R	Object Net	Sugar Crepe
CC3M	CLIP^{16} $^{(\mathcal{C})}$	0.141	0.122	0.068	0.033	0.177	0.080	0.611
	$\mathcal{C}$+GAP	0.156	0.134	0.069	0.033	0.187	0.081	0.616
	$\mathcal{C}$+SPARO	**0.170**	**0.140**	**0.088**	**0.035**	**0.221**	**0.098**	**0.625**
CC12M	CLIP^{16} $^{(\mathcal{C})}$	0.361	0.311	0.249	0.091	0.467	0.218	0.697
	$\mathcal{C}$+GAP	0.382	0.330	0.262	0.101	0.501	0.241	0.695
	$\mathcal{C}$+SPARO	**0.406**	**0.350**	**0.298**	**0.113**	**0.559**	**0.268**	**0.723**
CC15M	CLIP^{16} $^{(\mathcal{C})}$	0.384	0.337	0.268	0.105	0.503	0.238	0.699
	$\mathcal{C}$+GAP	0.399	0.343	0.287	0.114	0.531	0.252	0.701
	$\mathcal{C}$+SPARO	**0.437**	**0.378**	**0.317**	**0.145**	**0.579**	**0.279**	**0.730**
L400M	CLIP^{32} $^{(\mathcal{C})}$	0.617	0.531	0.482	0.202	0.719	0.423	0.748
	$\mathcal{C}$+GAP	0.623	0.537	0.492	0.212	0.725	0.440	0.732
	$\mathcal{C}$+SPARO	**0.635**	**0.552**	**0.507**	**0.231**	**0.747**	**0.459**	**0.770**

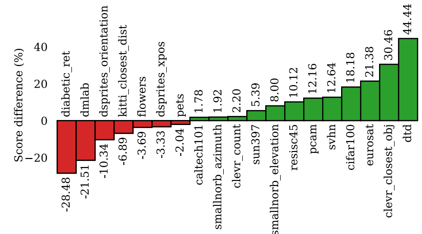
Models. For all experiments in this section, we train CLIP models using the open-source OpenCLIP [34] project. We consider two model sizes: those with a ViT-B/16 visual backbone (CLIP^{16}) and those with a ViT-B/32 backbone (CLIP^{32}). For models trained with SPARO, we replace the last transformer block of the base transformer backbones. For our chosen settings, this ensures that the resulting model size is comparable to the original model size. Further, since only the `CLS` token output from the image transformer and the `EOS` token output from the text transformer are used for the standard CLIP encodings, removing one transformer block ensures that we do not gain an unfair advantage by attending to positions from a layer that standard CLIP discards. Note that standard CLIP also retains the MLP as a part of its final transformer block, which is absent in the SPARO module. We also compare our encodings with those produced by CLIP with global average pooling (CLIP+GAP), which enjoys the advantage of being able to use all of the final layer outputs. When using SPARO, we use $L = V = 64$ when training on Conceptual Captions, and $L = 128, V = 64$ when training on LAION-400M. We also provide results for CLIP with residual network (ResNet) [28] encoders in Appendix B.3.

4.1 Zero-shot recognition, robustness, and compositionality

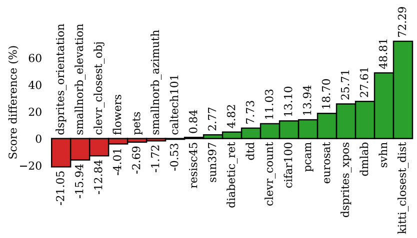
We evaluate the zero-shot classification accuracy of trained CLIP models on ImageNet [16] and a set of robustness benchmarks including ImageNet-V2 [53], ImageNet-Sketch [62], ImageNet-A [31], ImageNet-R [30], and ObjectNet [4]. Additionally, to evaluate the compositionality of our learned encodings in terms of objects, attributes, and relations, we test our trained models on SugarCrepe



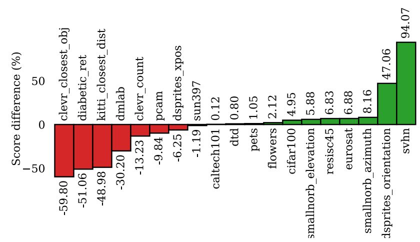
(a) CLIP^{16} models trained on CC3M.



(b) CLIP^{16} models trained on CC12M.



(c) CLIP^{16} models trained on CC15M.



(d) CLIP^{32} models trained on L400M.

Fig. 2: Relative differences of CLIP+Sparo zero-shot accuracies when compared to CLIP+GAP on the VTAB benchmark.

[33]. We report our results in Tab. 1, and provide more fine-grained SugarCrepe numbers in Appendix B.4. We find that using Sparo encodings outperforms CLIP and CLIP+GAP in all settings considered. The improved results on the SugarCrepe benchmark suggests that Sparo encodings exhibit compositionality at a level that enable better binding of object, attribute, and relation properties than baseline encodings on average.

Finally, we also use the VTAB benchmark [70] to evaluate Sparo on a diverse set of datasets [5,13,14,19,22,23,29,35,36,38,45,46,48,61,63,68] that vary more significantly from the training data. We present the relative zero-shot classification accuracy improvements of CLIP+Sparo over CLIP+GAP in Fig. 2 and the corresponding absolute values in Appendix B.6. We see that Sparo outperforms standard CLIP encodings for a majority of VTAB tasks in all four settings.

4.2 Zero-shot image and text retrieval

To separately evaluate the quality of the learned image and text encodings, we consider zero-shot retrieval based on image and text on MS COCO [40], Flickr8k [32], and Flickr30k [65]. Our results are presented in Tab. 2. To compute Recall@5 for 'Text' we find the nearest 5 texts for an image, calculate the fraction of those texts that match the image, and average this metric over all the images (similar for Recall@5 for 'Image'). We find that Sparo encodings improve retrieval across all three datasets for both modalities.

Table 2: Image and text zero-shot Recall@5 retrieval results on MS COCO, Flickr8k, and Flickr30k for CLIP models trained on Conceptual Captions and LAION-400M.

Train	Model	Image R@5			Text R@5		
		COCO	F8k	F30k	COCO	F8k	F30k
CC3M	CLIP^{16} $^{(\mathcal{C})}$	0.237	0.400	0.353	0.294	0.489	0.469
	$\mathcal{C}$+GAP	0.260	0.430	0.400	0.312	0.534	0.497
	$\mathcal{C}$+SPARO	**0.289**	**0.490**	**0.461**	**0.363**	**0.587**	**0.581**
CC12M	CLIP^{16} $^{(\mathcal{C})}$	0.474	0.691	0.715	0.614	0.837	0.817
	$\mathcal{C}$+GAP	0.499	0.714	0.725	0.628	0.819	0.847
	$\mathcal{C}$+SPARO	**0.526**	**0.737**	**0.759**	**0.670**	**0.878**	**0.878**
CC15M	CLIP^{16} $^{(\mathcal{C})}$	0.512	0.738	0.757	0.636	0.844	0.865
	$\mathcal{C}$+GAP	0.535	0.761	0.772	0.659	0.849	0.881
	$\mathcal{C}$+SPARO	**0.557**	**0.778**	**0.793**	**0.696**	**0.898**	**0.905**
L400M	CLIP^{32} $^{(\mathcal{C})}$	0.599	0.822	0.840	0.765	0.923	0.935
	$\mathcal{C}$+GAP	0.610	0.822	0.845	0.768	0.917	0.946
	$\mathcal{C}$+SPARO	**0.616**	**0.836**	**0.854**	**0.774**	**0.935**	**0.950**

Table 3: Left: ImageNet linear probe accuracy for CLIP models trained on Conceptual Captions and LAION-400M. **Right:** ImageNet 20-nearest neighbors and linear probe accuracy for encodings trained with DINO.

Model	ImageNet linear probe			
Train:	CC3M	CC12M	CC15M	L400M
CLIP $^{(\mathcal{C})}$	0.469	0.630	0.646	0.743
$\mathcal{C}$+GAP	0.504	0.649	0.664	0.747
$\mathcal{C}$+SPARO	**0.561**	**0.700**	**0.711**	**0.755**

Model	ImageNet eval	
	20-NN	Linear probe
DINO	0.685	0.735
DINO+SPARO	**0.706**	**0.757**

4.3 Linear probe and DINO nearest neighbors classification

We evaluate the classification accuracy of linear probes trained on the encodings of our trained models. For CLIP models, we follow [50] where we take a subset of the training data to use as a validation set, and sweep for the best weight decay hyperparameter on it for each model setting. Additionally, we evaluate SPARO for self-supervised learning in vision using DINO [9]. We train DINO with a ViT-S/16 backbone on ImageNet [16] without labels with 8 GPUs for 100 epochs without changing any other hyperparameters. Here, we use SPARO with $L = V = 64$, replacing the last transformer block of the ViT encoder. The trained DINO models are evaluated on k-nearest neighbours with $k = 20$ and linear classification on ImageNet with labels. We report our results for CLIP and DINO in Tab. 3. Our results for CLIP align with our previous results, with SPARO outperforming CLIP and CLIP+GAP in all linear probe settings. Furthermore, we see that a straightforward incorporation of SPARO in the DINO encoder

Table 4: Effect of intervening to selecting only the top 32 slots based on zero-shot ImageNet validation accuracy on SugarCrepe compositionality for CC15M-trained CLIP^{16}+Sparo and LAION-400M-trained CLIP^{32}+Sparo.

SugarCrepe		CC15M		L400M	
	Slots:	100%	50%	100%	25%
Replace	Object	0.896	**0.912**	0.927	**0.929**
	Attribute	**0.798**	0.797	0.822	**0.854**
	Relation	0.666	**0.713**	0.678	**0.705**
Swap	Object	0.593	**0.659**	0.585	**0.646**
	Attribute	0.700	**0.703**	0.719	**0.740**
Add	Object	0.779	**0.809**	**0.866**	0.857
	Attribute	0.679	**0.737**	0.794	**0.822**
Average		0.730	**0.761**	0.770	**0.793**

outperforms standard DINO for both nearest neighbors and linear classification. We provide additional experimental details in Appendices C.1 and C.2.

5 Analysis

In this section, we provide insights into encodings learned by Sparo. We demonstrate the ability of using Sparo's slot structure to perform post-hoc selection of slots for improving downstream performance. We then show that without this structure, such selection is more prone to overfitting. Next, we validate the design choice of using single-head attention to model the notion of selective attention through controlled ablations. Finally, we visualize example concepts that CLIP models trained with Sparo learn to attend to.

5.1 Post-hoc concept selection

Sparo enables downstream tasks access to a collection of concepts. However, not all concepts may be important for a specific downstream task. For example, attending to watermarks or caption style can be important when trying to determine the source of a sample, but is unnecessary and often harmful for most downstream tasks that focus on the elements in the scenes. The separation of concepts in the structure of a Sparo encoding makes it possible to manually intervene to select desirable slots.

We demonstrate this ability with a simple heuristic: we perform concept selection by computing the zero-shot ImageNet validation accuracy of each slot separately, and pick only the top-performing 32 slots. In Tab. 4, we show that such an intervention exposes improved compositionality when evaluating CLIP+Sparo models trained on CC15M and LAION-400M on SugarCrepe.

Table 5: Effect of overfitting a partitioned mask for image encodings $\boldsymbol{y}$ of frozen CC15M-trained CLIP^{32}+GAP and CLIP^{32}+SPARO models by training on SugarCrepe. We normalize the relative change on the evaluation benchmarks by the attained relative change on SugarCrepe. The highlighted rows indicate settings where the structure assumed for the mask aligns with SPARO's representation structure.

(# parts, Part size)	Model (CLIP^{32})	Shape of $\boldsymbol{y}$	Training: SugarCrepe Initial	Training: SugarCrepe Masked	VTAB Avg Initial	VTAB Avg Masked	VTAB Avg Relative change	ImageNet 0-shot Acc Initial	ImageNet 0-shot Acc Masked	ImageNet 0-shot Acc Relative change
(64, 64)	+GAP	(4096)	0.692	0.732	0.240	0.217	−1.689	0.343	0.292	−2.598
	+SPARO	(64, 64)	0.710	0.764	0.262	0.272	**+0.497**	0.370	0.364	**−0.190**
(64, 8)	+GAP	(512)	0.678	0.712	0.254	0.219	−2.794	0.345	0.279	−3.924
	+SPARO	(64, 8)	0.696	0.741	0.249	0.260	**+0.688**	0.355	0.347	**−0.347**
(512, 1)	+GAP	(512)	0.678	0.771	0.254	0.215	−1.119	0.345	0.250	**−2.003**
	+SPARO	(64, 8)	0.696	0.809	0.249	0.206	**−1.059**	0.355	0.224	−2.277
(4096, 1)	+GAP	(4096)	0.692	0.815	0.240	0.215	−0.596	0.343	0.286	−0.938
	+SPARO	(64, 64)	0.710	0.881	0.262	0.242	**−0.316**	0.370	0.328	**−0.465**

5.2 Robustness of SPARO's slot structure

We further utilize the notion of post-hoc concept selection to study the robustness of SPARO's representational structure, and compare it with that of standard encodings. We set up the experiment as an inverse of Sec. 5.1 — we perform concept selection using the smaller SugarCrepe benchmark and evaluate the resulting encodings on real-world benchmarks. SugarCrepe contains 7,512 example pairs of positive and negative captions, differing only in compositional interpretation in 7 ways, over a set of 1,561 images. Compositionality is a shared high-level notion present in most real-world data, yet there is much information that exists in data beyond what is necessary to solve the SugarCrepe tasks. The small size of the SugarCrepe as a training dataset, together with the ubiquitous nature of compositionality in machine learning tasks, makes it an ideal candidate to overfit concept selection to for stress-testing the representational structure of SPARO by evaluating on other tasks.

We train a global mask $\boldsymbol{m} \in [0,1]^M$ of the image encodings $\boldsymbol{y} \in \mathbb{R}^M$ to maximize SugarCrepe performance, and evaluate the masked encodings $\boldsymbol{m} \odot \boldsymbol{y}$ on VTAB tasks and ImageNet for zero-shot classification. We train masks for encodings of frozen CLIP^{32}+GAP and CLIP^{32}+SPARO models trained on CC15M with $M = 512$ and $M = 4096$ each. For the SPARO models, we use $L = 64$ and $V = M/L$. For each of these settings, we consider two types of masking: per-dimension and per-slot. To compare the CLIP^{32}+GAP encodings on per-slot masking, we consider their L contiguous equal-sized partitions as the slot encodings. We provide full implementation details of our mask training setup in Appendix C.1.

Table 6: Ablation experiments showing benefits of Sparo's attention bottleneck (Cross-attention is 'Separate-head') and cross-modal attention alignment (Slot-wise LayerNorm and projection are 'Yes' or 'N/A') with CC15M-trained CLIP, CLIP+GAP, CLIP+Sparo, and ablated variants of CLIP+Sparo that include multi-head cross-attention and the attentional pooler (AttPool) [39, 66].

Model	Replace last block	Cross-attention	Slot-wise LayerNorm	Slot-wise projection	Params	FLOPS ($\times 10^9$)	ImageNet zero-shot
CLIP^{32} $(\mathcal{C})$	N/A	N/A	N/A	N/A	151M	7.4	0.329
$\mathcal{C}$+GAP					151M	7.4	0.345
$\mathcal{C}$+AttPool	No	Multi-head	No	No	440M	8.9	0.312
Ablated $\mathcal{C}$+Sparo				Yes	306M	8.8	0.344
			Yes	No	440M	8.9	0.338
				Yes	306M	8.8	0.359
		Separate-head	No	No	306M	8.8	0.344
				Yes	172M	8.7	0.358
			Yes	No	306M	8.8	0.344
				Yes	172M	8.7	**0.374**
$\mathcal{C}$+Sparo			N/A	N/A	161M	8.0	0.372
	Yes				151M	7.4	0.370

We expect the compositionality of the data to be encoded in concepts that are useful for real-world downstream benchmarks. Therefore, the aggressive selection of only the concepts that perform well on a compositionality benchmark should not significantly impact downstream performance. However, since SugarCrepe does not represent all possible compositional manipulations, we expect to see a drop in performance on some tasks due to overfitting. We present our results in Tab. 5. We find that when the mask structure is aligned with the slot structure of Sparo, overfitting concept selection improves the average performance on VTAB tasks, and incurs only a minimal drop in performance on ImageNet. Furthermore, even in settings where the masks are not trained to align with Sparo's slot structure, we see that Sparo remains more resilient to overfitting than standard encodings in a majority of settings.

5.3 Ablating the bottleneck and cross-modal alignment of Sparo

We evaluate the generalization benefits of the separate-head attention bottleneck and of aligning the Sparo attention heads between the modalities in CLIP. We set up Sparo on one end of the ablation spectrum and the cross-attention read-out of attentional pooler (AttPool) [39, 66] to the other. AttPool uses multi-head attention to attend to the backbone encoder's outputs using embedded learned queries whereas Sparo uses separate single-head attention mechanisms. Additionally, AttnPool performs layer normalization [2] followed by a linear projection on the output of attention to produce the output encoding, whereas SPARO directly uses the output of separate-head attention as the encoding.

We train CLIP^{32}, CLIP^{32}+GAP, and variants of CLIP^{32}+SPARO on CC15M with $L = 128, V = 64$ and evaluate for ImageNet zero-shot accuracy and model size. We start by considering the choice of either separate-head or multi-head cross-attention over the backbone outputs. For each cross-attention setting, we apply layer normalization and linear projection operations to the attention module's outputs, similar to AttnPool. However, we additionally also consider slot-wise variants of these operations. Here, a slot-wise operation f_ϕ is one that acts separately on each input slot $\boldsymbol{y}_l$ using the same parameters ϕ, and produces the slot-structured output $\text{concat}(f_\phi(\boldsymbol{y}_1), \ldots, f_\phi(\boldsymbol{y}_L))$. When applied to an input that is not slot-structured, we assume a slot structure by partitioning the input into L contiguous equal-sized partitions. With SPARO, using both slot-wise layer normalization and slot-wise projection enables each read-out attention head from one modality to align with the corresponding head from the other without being influenced by the other heads.

We present our results in Tab. 6, showing a clear advantage of SPARO over AttnPool. Additionally, both the choices of separate-head attention over multi-head and slot-wise operations over the alternatives result in improved generalization, reduced model size, and fewer floating-point operations per second (FLOPS). While the former supports the effectiveness of the separate-head attention bottleneck, the latter validates the benefits of the prior for shared separately-attendable concepts between the modalities enabled by SPARO.

5.4 Visualizations

To conclude our analysis, we qualitatively visualize the concepts represented by SPARO using the CLIP^{16}+SPARO model trained on CC15M from Sec. 4.1. A limitation of visualizing attended positions is that it only provides partial insight into the represented concept — we can tell where the information is taken from, but not how the information is used. Different SPARO slots can have similar attention maps for the same sample, but reveal different semantics when comparing patterns across other images. For instance, a slot that attends to animals (*e.g.* dogs, cats, elephants) and another that attends to transportation (*e.g.* trains, boats, cars) can both have the same attention mask over patches and words for a horse being used for transportation. Furthermore, not all SPARO concepts align between the two modalities for every sample due to information being present in one that is not possible to infer from the other.

To ease our interpretation of attended concepts, we choose concepts to visualize that have sharp attention in the text modality, attend to non-overlapping text tokens, and have cosine similarity more than 0.75 with the corresponding image slots. We present examples of these filtered concepts in Fig. 3. In this example, we see that SPARO is capable of attending to different concepts such as subject of the scene, the activity represented, and the surrounding location, for both vision and text. We provide additional visualizations in Appendix B.7.

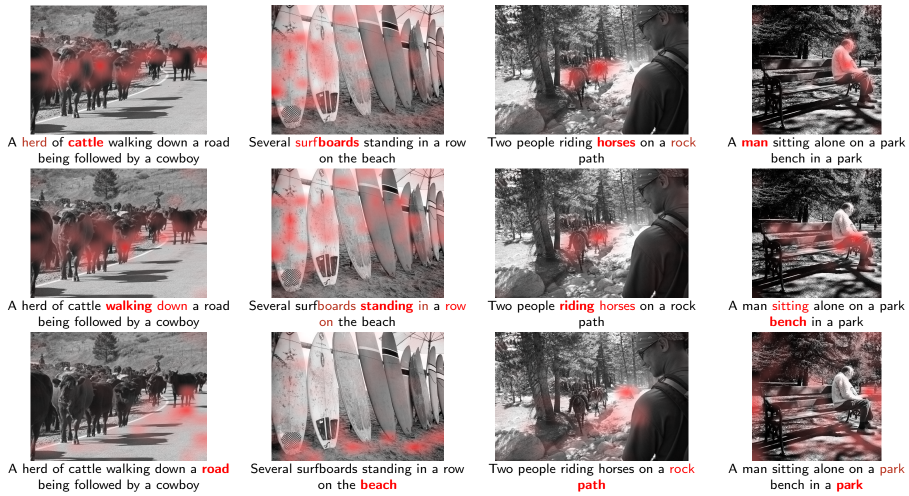


Fig. 3: Visualizing of the attended image and text positions for three SPARO slots (one per row) across four examples (one per column) from MS COCO. We surmise that the SPARO concepts from top to bottom represent the subject, activity, and location.

6 Discussion

Humans utilize a prior for compositionality to coherently represent different subsets of salient aspects of the world. We can enrich our comprehension by selectively attending to new aspects or specialize to a task by filtering out distracting ones. We introduce SPARO with a goal of imparting a similar prior to transformers in representation learning frameworks by partitioning encodings into separately-attended concepts. Although we see evidence of disentanglement through positive results of post-hoc concept selection and visualization of attention maps, we do not impose any explicit independence or disentanglement constraints. In our representation learning frameworks, the pressure for learning distinct concepts arises from the need to explain the factors of variation through the training objective with only the features that can be accumulated by a single head of attention with embedded queries and limited dimensionality. However, auxiliary training objectives to impose stronger distributional conditions on the learned concepts can be explored as promising future directions. Beyond pretraining, we highlight that more sophisticated post-hoc concept selection approaches than explored in our work, like using human interaction or set cover algorithms, can further improve the downstream utility of SPARO.

Acknowledgments

This research was funded by Sony and enabled in part by compute resources provided by the Digital Research Alliance of Canada, Mila, and Sony.

References

1. Aydemir, G., Xie, W., Güney, F.: Self-supervised object-centric learning for videos. arXiv preprint arXiv:2310.06907 (2023)
2. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
4. Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., Katz, B.: Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. Advances in neural information processing systems **32** (2019)
5. Beattie, C., Leibo, J.Z., Teplyashin, D., Ward, T., Wainwright, M., Küttler, H., Lefrancq, A., Green, S., Valdés, V., Sadik, A., et al.: Deepmind lab. arXiv preprint arXiv:1612.03801 (2016)
6. Boff, K.R., Kaufman, L., Thomas, J.P.: Handbook of perception and human performance, vol. 1. Wiley New York (1986)
7. Brady, J., Zimmermann, R.S., Sharma, Y., Schölkopf, B., von Kügelgen, J., Brendel, W.: Provably learning object-centric representations. arXiv preprint arXiv:2305.14229 (2023)
8. Burgess, C.P., Matthey, L., Watters, N., Kabra, R., Higgins, I., Botvinick, M., Lerchner, A.: Monet: Unsupervised scene decomposition and representation. arXiv preprint arXiv:1901.11390 (2019)
9. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
10. Chang, H.S., Sun, R.Y., Ricci, K., McCallum, A.: Multi-cls bert: An efficient alternative to traditional ensembling. arXiv preprint arXiv:2210.05043 (2022)
11. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3558–3568 (2021)
12. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. arXiv preprint arXiv:2104.02057 (2021)
13. Cheng, G., Han, J., Lu, X.: Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE **105**(10), 1865–1883 (2017)
14. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3606–3613 (2014)
15. Colby, C.L., Goldberg, M.E.: Space and attention in parietal cortex. Annual review of neuroscience **22**(1), 319–349 (1999)
16. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. IEEE (2009)
17. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: North American Chapter of the Association for Computational Linguistics (2019)
18. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

19. Dugas, E., Jared, Jorge, Cukierski, W.: Diabetic retinopathy detection (2015), https://kaggle.com/competitions/diabetic-retinopathy-detection
20. Engelcke, M., Kosiorek, A.R., Jones, O.P., Posner, I.: Genesis: Generative scene inference and sampling with object-centric latent representations. arXiv preprint arXiv:1907.13052 (2019)
21. Eslami, S., Heess, N., Weber, T., Tassa, Y., Szepesvari, D., Hinton, G.E., et al.: Attend, infer, repeat: Fast scene understanding with generative models. Advances in neural information processing systems **29** (2016)
22. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE transactions on pattern analysis and machine intelligence **28**(4), 594–611 (2006)
23. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013)
24. Goyal, A., Didolkar, A., Ke, N.R., Blundell, C., Beaudoin, P., Heess, N., Mozer, M.C., Bengio, Y.: Neural production systems. Advances in Neural Information Processing Systems **34**, 25673–25687 (2021)
25. Goyal, A., Lamb, A., Gampa, P., Beaudoin, P., Blundell, C., Levine, S., Bengio, Y., Mozer, M.C.: Factorizing declarative and procedural knowledge in structured, dynamical environments. In: International Conference on Learning Representations (2020)
26. Goyal, A., Lamb, A., Hoffmann, J., Sodhani, S., Levine, S., Bengio, Y., Schölkopf, B.: Recurrent independent mechanisms. arXiv preprint arXiv:1909.10893 (2019)
27. Greff, K., Kaufman, R.L., Kabra, R., Watters, N., Burgess, C., Zoran, D., Matthey, L., Botvinick, M., Lerchner, A.: Multi-object representation learning with iterative variational inference. In: International Conference on Machine Learning. pp. 2424–2433. PMLR (2019)
28. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
29. Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **12**(7), 2217–2226 (2019)
30. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8340–8349 (2021)
31. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15262–15271 (2021)
32. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research **47**, 853–899 (2013)
33. Hsieh, C.Y., Zhang, J., Ma, Z., Kembhavi, A., Krishna, R.: Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. Advances in neural information processing systems (2023)
34. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: OpenCLIP. https://github.com/mlfoundations/open_clip (Jul 2021). https://doi.org/10.5281/zenodo.5143773

35. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2901–2910 (2017)
36. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. Technical report (2009)
37. Lavoie, S., Tsirigotis, C., Schwarzer, M., Vani, A., Noukhovitch, M., Kawaguchi, K., Courville, A.: Simplicial embeddings in self-supervised learning and downstream classification. In: International Conference on Learning Representations (2023), https://openreview.net/forum?id=RWtGreRpovS
38. LeCun, Y., Huang, F.J., Bottou, L.: Learning methods for generic object recognition with invariance to pose and lighting. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004. vol. 2, pp. II–104. IEEE (2004)
39. Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., Teh, Y.W.: Set transformer: A framework for attention-based permutation-invariant neural networks. In: International conference on machine learning. pp. 3744–3753. PMLR (2019)
40. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Fleet, D.J., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V. Lecture Notes in Computer Science, vol. 8693, pp. 740–755. Springer (2014). https://doi.org/10.1007/978-3-319-10602-1_48, https://doi.org/10.1007/978-3-319-10602-1%5F48
41. Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., Kipf, T.: Object-centric learning with slot attention. Advances in Neural Information Processing Systems **33**, 11525–11538 (2020)
42. Ma, Z., Hong, J., Gul, M.O., Gandhi, M., Gao, I., Krishna, R.: Crepe: Can vision-language foundation models reason compositionally? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10910–10921 (2023)
43. Mansouri, A., Hartford, J., Zhang, Y., Bengio, Y.: Object centric architectures enable efficient causal representation learning. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=r9FsiXZxZt
44. Martinez, A.: Putting spatial attention on the map: timing and localization of stimulus selection processes in striate and extrastriate visual areas. Vision Res **41**, 1437–1457 (2001)
45. Matthey, L., Higgins, I., Hassabis, D., Lerchner, A.: dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/ (2017)
46. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian conference on computer vision, graphics & image processing. pp. 722–729. IEEE (2008)
47. O'Connor, D.H., Fukui, M.M., Pinsk, M.A., Kastner, S.: Attention modulates responses in the human lateral geniculate nucleus. Nature neuroscience **5**(11), 1203–1209 (2002)
48. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3498–3505. IEEE (2012)

49. Qian, R., Ding, S., Liu, X., Lin, D.: Semantics meets temporal correspondence: Self-supervised object-centric learning in videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16675–16687 (2023)
50. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
51. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
52. Ray, A., Radenovic, F., Dubey, A., Plummer, B.A., Krishna, R., Saenko, K.: Cola: A benchmark for compositional text-to-image retrieval. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023)
53. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: International conference on machine learning. pp. 5389–5400. PMLR (2019)
54. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)
55. Scott, W.A.: Cognitive complexity and cognitive flexibility. Sociometry pp. 405–414 (1962)
56. Seitzer, M., Horn, M., Zadaianchuk, A., Zietlow, D., Xiao, T., Simon-Gabriel, C.J., He, T., Zhang, Z., Schölkopf, B., Brox, T., et al.: Bridging the gap to real-world object-centric learning. arXiv preprint arXiv:2209.14860 (2022)
57. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers. pp. 2556–2565. Association for Computational Linguistics (2018)
58. Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., Ross, C.: Winoground: Probing vision and language models for visio-linguistic compositionality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5238–5248 (2022)
59. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Cognitive psychology **12**(1), 97–136 (1980)
60. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
61. Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M.: Rotation equivariant cnns for digital pathology. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11. pp. 210–218. Springer (2018)
62. Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. In: Advances in Neural Information Processing Systems. pp. 10506–10518 (2019)
63. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: 2010 IEEE computer society conference on computer vision and pattern recognition. pp. 3485–3492. IEEE (2010)
64. Xu, L., Ouyang, W., Bennamoun, M., Boussaid, F., Xu, D.: Multi-class token transformer for weakly supervised semantic segmentation. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4310–4319 (2022)
65. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2**, 67–78 (2014)
66. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022)
67. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? In: The Eleventh International Conference on Learning Representations (2023)
68. Yuval, N.: Reading digits in natural images with unsupervised feature learning. In: Proceedings of the NIPS Workshop on Deep Learning and Unsupervised Feature Learning (2011)
69. Zadaianchuk, A., Kleindessner, M., Zhu, Y., Locatello, F., Brox, T.: Unsupervised semantic segmentation with self-supervised object-centric representations. arXiv preprint arXiv:2207.05027 (2022)
70. Zhai, X., Puigcerver, J., Kolesnikov, A., Ruyssen, P., Riquelme, C., Lucic, M., Djolonga, J., Pinto, A.S., Neumann, M., Dosovitskiy, A., et al.: A large-scale study of representation learning with the visual task adaptation benchmark. arXiv preprint arXiv:1910.04867 (2019)
71. Zhang, Y., Hare, J., Prugel-Bennett, A.: Deep set prediction networks. Advances in Neural Information Processing Systems **32** (2019)
72. Zhao, T., Zhang, T., Zhu, M., Shen, H., Lee, K., Lu, X., Yin, J.: Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. arXiv preprint arXiv:2207.00221 (2022)