# Your Self-Supervised Projection Head Captures Object Co-Occurrence Statistics

**Arthur Aubret**
Université Clermont Auvergne
Clermont Auvergne INP, CNRS, Institut Pascal
F-63000 Clermont-Ferrand, France
arthur.aubret@irit.fr

**Jochen Triesch**
Frankfurt Institute for Advanced Studies
Frankfurt am Main, 60438, Germany
triesch@fias.uni-frankfurt.de

**Céline Teulière**
Université Clermont Auvergne
Clermont Auvergne INP, CNRS, Institut Pascal
F-63000 Clermont-Ferrand, France
celine.teuliere@uca.fr

## Abstract

Self-Supervised Learning (SSL) has achieved impressive success in learning semantic visual representations, yet the underlying principles driving this success remain underexplored. In this work, we hypothesize that common SSL pretext tasks implicitly model object co-occurrence statistics, a fundamental cue for visual learning. To test this, we curate three datasets of segmented objects from existing vision benchmarks using a state-of-the-art segmentation model. Through experiments across many instance-discrimination SSL models, we reveal a hierarchical encoding of semantic information: while the visual backbone captures object categories, the projection head specializes in encoding co-occurrences between object categories. In addition, the projection head outperforms the backbone in aligning with human judgments of inter-category similarity. Finally, by controlling co-occurrence patterns during pre-training, we demonstrate that encoding object co-occurrences can significantly accelerate the emergence of category-level representations. Our findings uncover a previously hidden learning principle in SSL and suggest a path toward designing more effective pretext tasks by explicitly leveraging object co-occurrence structure.

## 1 Introduction

Recent advances in Self-Supervised Learning (SSL) have made it possible to train vision models without annotation, achieving strong performance on downstream tasks such as categorization. One notable range of SSL methods hinges on a shared pipeline (or pretext task) [Chen et al., 2020, Caron et al., 2021, Grill et al., 2020]: a visual backbone extracts visual features, which feed a multi-layer projection head to learn embeddings invariant to geometric (small random crops, etc.) or color transformations (hue, etc.) of images. Throughout the paper, we refer to this family as instance-discrimination (ID) SSL methods. There is strong empirical evidence that geometric transformations and the projection head are crucial components of SSL [Chen et al., 2020, Caron et al., 2020]. Although these are known to support position and invariance properties in visual embeddings [Tian et al., 2020a, Ericsson et al., 2022, Chen et al., 2020], the underlying principles by which these components support category-wise representation learning remain poorly understood.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

A foundational design principle of large language models is to model the statistical structure of word co-occurrences [Mikolov et al., 2013, Radford et al., 2019]. Combined with large-scale datasets, this principle supports the learning of concept representations that capture the semantic structural aspects of our world. In contrast, the established assumption in ID SSL is that data-augmentations should preserve the task-label information [Tian et al., 2020b]. For example, two random image croppings should show different parts of the same object. However, vision SSL models may also implicitly encode natural object co-occurrence statistics. Pre-training images often reflect real-world object co-occurrence patterns [Turini and Võ, 2022], such as the simultaneous presence of "food" and "forks" in a dining scene. Random cropping frequently simulates a zoom-in on different objects (*e.g.* "food", "fork") within a scene. Consequently, ID SSL models may align embeddings of objects that often co-occur together. This leads to a critical question: Do common SSL methods implicitly capture object co-occurrence statistics, and if so, how does this shape representation learning? To the best of our knowledge, this work is the first to explicitly investigate this question.

To answer it, we construct three datasets of segmented objects by leveraging existing large-scale vision datasets and a state-of-the-art segmentation model. This allows us to remove object co-occurrences from the visual inputs, thereby keeping one category of objects per image. Based on this data, we analyze whether internal layers of dozens of SSL methods capture natural object co-occurrence statistics, *e.g.* whether "food" and "fork" are aligned. To explore the causal impact of such a statistical structure on SSL, we systematically control object co-occurrence statistics during pre-training and quantify their impact on downstream categorization performance.

Our experiments reveal that common ID SSL models hierarchically encode semantic information: the projection head excels at capturing object co-occurrence statistics, while the visual backbone specializes in discriminating object categories. As a consequence, the projection heads significantly outperform the visual backbones for capturing human inter-category similarity judgments. Finally, we provide experimental evidence that encoding object co-occurrences can drastically improve downstream categorization. Together, these findings uncover a hidden learning principle in SSL: the projection head leverages natural co-occurrence statistics to learn strong visual representations. Our contributions are as follows:

- We analyze how object co-occurrences shape self-supervised visual representations learning.
- We show that projection heads contain the most effective layers for encoding object co-occurrences.
- We discover that encoding object co-occurrences allows the projection head to outperform the backbone in capturing inter-category human similarity judgments.
- We provide evidence that encoding object co-occurrences enhances feature extraction for downstream categorization.

More broadly, our work suggests that projection heads should not be overlooked after pre-training when the goal is to study human–machine alignment in higher-order semantic representations and opens new avenues for designing pretext tasks that explicitly leverage object co-occurrence structure. The code is available at `https://github.com/Aubret/object_cooccurrence`.

## 2 State of the art

### 2.1 Self-Supervised Learning

Most ID SSL methods learn by aligning visual embeddings extracted from two views of an image, where each view results from the original image going through two augmentation pipelines [Giakoumoglou and Stathaki, 2024]. Augmentations pipelines typically include cropping, flipping or coloring images. The main difference across SSL models lie in how they prevent representational collapse, *i.e.* learning the exact same embedding for all images. Seminal contrastive learning methods explicitly promote dissimilarity between all embeddings [Chen et al., 2020, He et al., 2020, Wang and Qi, 2022, Hu et al., 2021, Wang et al., 2023a, Dwibedi et al., 2021]. Close to contrastive learning, several approaches strive to explicitly decorrelate features within and across different views [Zbontar et al., 2021, Bardes et al., 2022, Wang et al., 2023b], with approaches bridging contrastive and decorrelation SSL [Zhang et al., 2024, Liu et al., 2022]. Another set of competitive methods consider asymmetric neural network architectures for processing two views [Grill et al., 2020, Chen and He,

2021, Gidaris et al., 2020, Caron et al., 2021, Chen et al., 2021]. More recently, LeJEPA [Balestriero and LeCun, 2025] proposes to enforce an isotropic Gaussian distributed on embedding features, and other methods employ masked modeling to improve upon prior work [Assran et al., 2023, Zhou et al., Oquab et al., 2024, Siméoni et al., 2025]. Rather than building a state-of-the-art SSL model, the objective of this work is to understand the principles underpinning the success of these methods.

### 2.2 Analysis of the SSL projection head

The empirical importance [Chen et al., 2020] of the projection head has attracted attention as 1- the projection head is a key architectural difference compared to supervised learning [Grigg et al.] and 2- the best performance on downstream tasks is obtained when it is fully discarded [Jing et al., Chen et al., 2020, Cosentino et al., 2022, Bardes et al., 2022]. One reason for the latter is that the projection head is too specialized for augmentation invariance, a property misaligned with typical downstream tasks [Grigg et al., Bordes et al., Xue et al.]. In line with this finding, several works have found that the rank of projection embeddings decreases, compared to that of the backbone representations, which is known as dimensional collapse [Gupta et al., 2022, Jing et al., Gui et al., 2023, Garrido et al., 2023, Hua et al., 2021]. A recent analysis further suggests that the projection head filters out information that is irrelevant for the SSL objective [Ouyang et al.]. With spectral analysis, Balestriero and LeCun [2022] found that SSL optimizes embeddings such that their main singular values match those of the inter-sample similarity matrix. This is consistent with a theory that formulates image augmentations as a graph and posits that the overlap between image augmentations guides SSL [Zhang et al., 2025, Wang et al., HaoChen et al., 2021]. Visual learning was later reformulated as an interplay between augmentation overlap and architectural inductive biases [Cabannes et al., 2023, Saunshi et al., 2022]. In contrast to these works, our work analyzes what is concretely encoded in projection heads and how it shapes learning.

### 2.3 Learning Object Co-occurrences

Visual representations that match object co-occurrence statistics are known to emerge in both humans[Turini and Võ, 2022, Bonner and Epstein, 2021] and supervised learning models[Bracci et al., 2023]. In SSL, some studies have assumed the importance of spatial co-occurrences between object features [Aubret et al., 2026, Caron et al., 2020]. However, few studies have investigated the presence of representations capturing co-occurrences between distinct objects in SSL. Aubret et al. [2024] found a hierarchical encoding of semantic information in SSL, but they manually defined object co-occurrences. In contrast, we focus our analysis on natural co-occurrences present in the standard training regime. A more recent study [Schaumlöffel et al., 2026] found that learning temporally consistent representations from human egocentric videos promotes the encoding of objects' context. Yet, they neglect the projection head. To the best of our knowledge, we are the first to show that encoding natural object co-occurrences is a key principle of common visual SSL approaches.

## 3 Method

### 3.1 Motivation

Our working hypothesis presumes that the SSL crop-based augmentation pipeline frequently yields pairs of images depicting different objects. We start by testing this assumption on ImageNet-1k (I1K), the most established dataset for visual SSL. Since I1K lacks semantic segmentation annotations, we use a pretrained ResNet50 as a proxy to estimate the frequency of semantically inconsistent pairs. We sample 1M of crop-augmented image pairs from the train split and compute 1- an upper bound on inconsistency by measuring how often the predicted class labels for a pair differ and 2- A lower bound by counting how frequently each crop's predicted label falls outside the other's top-5 predictions.

As shown in Figure 1A, augmented views are semantically inconsistent approximately 15% of the time under standard hyperparameters (*e.g.*, a minimum crop size of 0.08). Furthermore, Figure 1B reveals that cropping can align co-occurring objects from diverse categories, such as toys, dining objects (chair and glass, mug and food) or meadow-related elements (grass and hedgehog). These observations motivate our investigation into whether SSL models implicitly learn from such co-occurrence patterns.




Figure 1: A) Ratio of semantically inconsistent pair of crops when varying the minimum crop ratio of SSL's augmentation pipeline. B) Examples of semantically inconsistent pairs obtained with random cropping.

### 3.2 Datasets and metrics

**Datasets and object co-occurrence matrices.** We want to assess whether and how SSL methods encode object co-occurrence statistics. We start by deriving a dataset for which each image contains only a single object, devoid of contextual information. This setup ensures that any contextual relationships between objects must emerge from statistical patterns discovered during training. To do so, we use the THINGS dataset [Hebart et al., 2023], which contains 26,107 naturalistic images distributed in 1,854 object concepts, with each image depicting a single, clearly visible object. To further remove residual background information, we preprocess all images with Lang-SAM [Medeiros, 2025], a language-conditioned segmentation pipeline. Lang-SAM combines Grounding-DINO [Liu et al., 2023] to generate concept-specific bounding boxes and SAM-2 [Ravi et al., 2024] to segment the objects within these boxes. We manually filter out poorly segmented images. Additionally, we crop and extract segmented objects larger than $32 \times 32$ pixels from the LVIS validation split [Gupta et al., 2019] as an alternative dataset. Examples of images are provided in Appendix A.

To test whether SSL models encode contextual information in segmented objects, we first estimate natural co-occurrence statistics. We use three large-scale scene datasets that provide semantic annotations of several objects within each image: the COCO training and validation splits, the ADE-20K training split, and the LVIS training split. For each dataset, we compute a co-occurrence matrix by counting the total number of times each pair of object concepts appears together. To align these statistics with our segmented object datasets, we retain only overlapping classes between each scene dataset and its corresponding segmented object dataset. This yields three paired datasets (Co-occurrence dataset-Segmented dataset):

- COCO-THINGS: 736 segmented THINGS images across 53 categories,
- ADE20K-THINGS: 918 segmented THINGS images across 77 categories,
- LVIS-LVIS: 118,068 extracted LVIS object images across 1,002 categories.

**Representational Similarity Analysis.** We perform a Representational Similarity Analysis (RSA) to quantitatively assess whether layer representations reflect natural co-occurrence matrices. For each layer of a model, we average the representations across segmented images that belong to the same category and compute the pair-wise cosine representational similarity matrix (RSM) between category representations. Then, we compute the RSA Kendall's tau between the upper triangle of the RSM and the co-occurrence matrix.

**Context classification accuracy** As an alternative to RSA, we propose a context classification accuracy metric to measure the alignment between object co-occurrences and visual representations. For COCO-THINGS, we manually assign its 53 object categories to 10 spatial contexts in {"Urban street", "Countryside", "Savanna", "Sea", "Snow mountain", "Sport field", "Kitchen", "Bathroom", "Living room", "Bedroom"}. For ADE20k-THINGS and LVIS-LVIS, we run three times a spectral clustering algorithm [Von Luxburg, 2007] on the normalized logarithm of the co-occurrence matrix. Each clustering provides 10 clusters, which we label as co-occurrence contexts. In section 4.4, we experiment by varying the number of clusters.

To compute an accuracy, we first build category representations by averaging the representations across segmented images that belong to the same category, as in section 3.2. Then, we generate a series of tuples composed of one randomly sampled category representation, a random category representation belonging to the same context, and a random category representation from each other context. Then, we compute how often the first representation is closer to that of the same context than to representations from other contexts. For each context assignment, we assess 1,000/10,000 tuples for *-THINGS/LVIS-LVIS and report the average accuracy over tuples and context assignments. In Appendix C.3, we also consider image-level comparison metrics.

### 3.3 Pretrained models

To emphasize the general representational principles of SSL, we evaluate a diverse set of SSL models pretrained on Imagenet-1k, including their projection heads. Our selection spans multiple SSL paradigms (contrastive, decorrelation, distillation) and neural architectures (ViT-B/16, ResNet50), specifically: MoCoV3 with ViT-B/16 and Resnet50, BYOL with Resnet50, VICReg with Resnet50, LeJEPA with resnet50, and DINO with ViT-B/16. In some experiments, we also pretrain SSL models, and use other pretrained models, as detailed in Appendix B.

## 4 Results

### 4.1 The Projection Head Captures Natural Object Co-occurrences

Here, we investigate whether the projection head of classical SSL methods learns representations that capture object co-occurrence statistics. In Table 1, we observe that an internal layer of the projection head consistently outperforms the final layers of the backbone in encoding these statistics across all datasets and SSL methods. Minor exceptions are BYOL-R50 on LVIS-LVIS (RSA Kendall's tau) and MoCoV3-ViT-B on COCO-THINGS (context classification), which remain unexplained. Early and intermediate layers of the projection head are often particularly suited to capture object co-occurrence structure, while the final linear layer tends to perform poorly. One explanation for the poor performance of the final layer is that it memorizes image-level co-occurrence statistics Meehan et al. [2023]. These results are confirmed with a broader set of pretrained models in Appendix C.2 and with more backbone layers in Appendix C.4. When comparing across methods, DINO-ViT-B stands out as the best model for capturing object co-occurrence statistics, with a single exception for Kendall's tau in LVIS-LVIS. As DINO is also the best performing model for ImageNet classification, this suggests that capturing object co-occurrences may correlate with better image representations for downstream classification. We further explore this hypothesis in Section 4.3.

To better understand the semantic aspects captured by the projection head, we visualize in Figure 2 the RSMs of object co-occurrences, representation layers and clusters of contexts. First, we observe that the hand-defined contexts roughly capture the geometry of object co-occurrences. Differences are noticeable, which potentially stem from the difference between COCO and naturalistic environments or imperfect context construction. The backbone of DINO-ViT-B captures only small clusters of object co-occurrences, such as sea-related (kite, surfboard, boat) or dining-related (fork, knife, spoon) objects. In contrast, projection layer P.1 further reflects broader semantic contexts such as in-door household objects or food-related items. The intermediate projection layer P.0 shows in-between behavior.

These results demonstrate that layers of projection heads learn a representational structure aligned with object co-occurrence statistics. However, is the projection head necessary for representing object co-occurrences? To answer this question, we pretrain SSL models while modifying the projection head (cf. Appendix B). In Table 2, we compare co-occurrence encoding metrics with and without a non-linear projection head during pre-training. We observe that the projection head consistently improves co-occurrence encoding across datasets, methods and metrics. There is a single exception for the ADE20k-THINGS context accuracy of SimCLR. Taken together, these findings demonstrate that the projection head is crucial for encoding object co-occurrences.

### 4.2 The Projection Head Better Captures Human Object Similarity Judgments

If the projection head captures semantic information beyond object categorization, it may reflect human inter-category similarity judgments. To test this hypothesis, we leverage the original THINGS

Table 1: Evaluation of the backbone (B) and internal layers of projection heads (P.n) based on whether they capture the co-occurrence statistics of objects. "mP.n" refers to the maximal value within the n-th block "Linear, BatchNorm, ReLU" of the projection head. "P.n" refers to the n-th linear layer of the projection head. "na" occurs when the pretrained model does not contain the n-th linear layer. C-T, A-T and L-L refer to COCO-THINGS, ADE20k-THINGS and LVIS-LVIS, respectively.

| Cooc. | Method | RSA Kendall's tau | | | | Context accuracy | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | B | mP.0 | mP.1 | mP.2 | B | P.0 | P.1 | P.2 |
| C-T | MoCoV3-R50 | 1.5 | **13.7** | 6.4 | na | 39.2 | **47.7** | 25.5 | na |
| | BYOL-R50 | 3.5 | **22.3** | -4.9 | na | 38.6 | **42.0** | 23.4 | na |
| | VICReg-R50 | 2.6 | 14.7 | **16.8** | -3.4 | 35.1 | **40.8** | 39.3 | 22.4 |
| | LeJEPA | 6.5 | 11.4 | 14.8 | **17.8** | 39.5 | 41.0 | 39.7 | **44.1** |
| | MoCoV3-ViT-B | 6.5 | 7.1 | **14.8** | 5.9 | **39.9** | 38.9 | 36.2 | 33.0 |
| | DINO-ViT-B | 9.0 | 21.5 | **33.4** | 13.9 | 43.5 | 45.7 | **50.3** | 40.7 |
| A-T | MoCoV3-R50 | 14.1 | 16.9 | **17.7** | na | 22.0 | **23.6** | 22.6 | na |
| | BYOL-R50 | 9.4 | **11.3** | 8.0 | na | 23.2 | **26.8** | 19.5 | na |
| | VICReg-RN0 | 11.2 | **17.2** | 16.3 | 11.6 | 24.2 | **26.6** | 24.6 | 25.7 |
| | LeJEPA | 15.7 | 14.8 | 17.0 | **17.6** | 21.6 | 21.9 | 23.2 | **24.9** |
| | MoCoV3-ViT-B | 15.8 | **18.6** | 17.0 | 14.8 | 27.0 | 28.1 | **28.2** | 24.9 |
| | DINO-ViT-B | 19.3 | 21.0 | 20.5 | **22.4** | 22.8 | 23.6 | 26.7 | **33.0** |
| L-L | MoCoV3-R50 | 40.7 | **45.7** | 13.2 | na | 19.0 | **20.2** | 14.7 | na |
| | BYOL-R50 | **39.8** | 39.3 | 17.9 | na | 18.5 | **20.6** | 13.3 | na |
| | VICReg-R50 | 37.1 | 40.3 | **49.0** | 16.2 | 19.0 | **19.7** | 18.6 | 16.1 |
| | LeJEPA | 29.4 | 30.9 | **32.8** | 32.6 | 18.6 | 19.2 | 19.6 | **20.3** |
| | MoCoV3-ViT-B | 43.8 | **44.0** | 43.0 | 20.2 | 18.9 | **19.6** | 19.5 | 18.4 |
| | DINO-ViT-B | 39.8 | **43.3** | 40.6 | 27.5 | 21.4 | 21.1 | **23.9** | 22.6 |

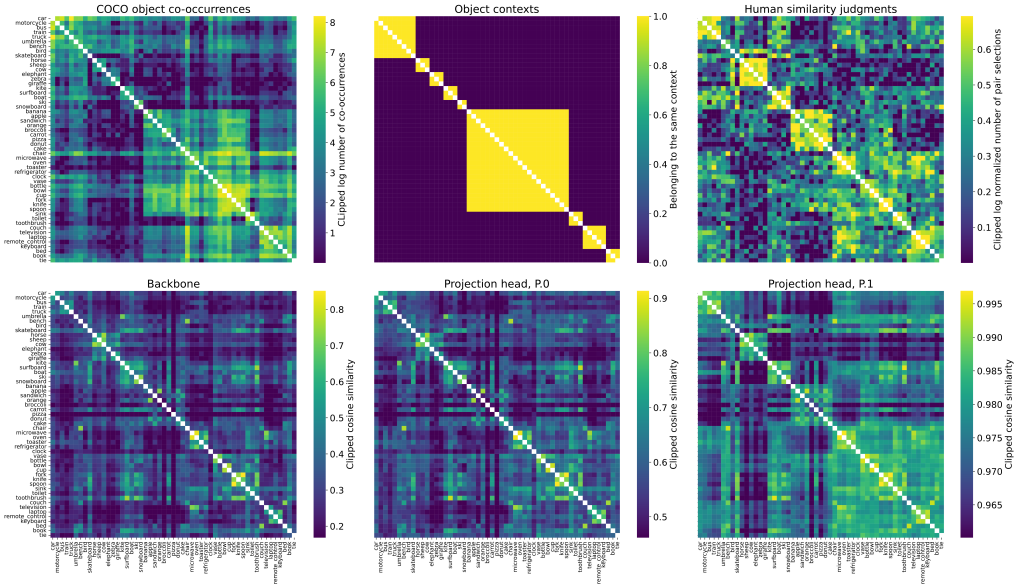


Figure 2: Top row: RSMs for COCO object co-occurrences, context assignments, and human similarity judgments. Bottom row: RSMs for visual representations learnt by DINO-ViT-B after the visual backbone, the first and the second linear layer of the projection head (P.0 and P.1, respectively). To avoid outliers to bias the visualization, we lower bound each matrix by the first decile of the matrix.

Table 2: Comparison of SSL models with and without non linear projection heads, for encoding object co-occurrences. In the "without" condition, the projection head is a linear layer (two for DINO with the final expansion layer). In the "with" condition, we use the setting from the original paper (One hidden layer for SimCLR and BYOL, two for DINO). C-T, A-T and L-L refer to COCO-THINGS, ADE20k-THINGS and LVIS-LVIS, respectively.

| Cooc. | Method | RSA Kendall's tau | | Context accuracy | |
|---|---|---|---|---|---|
| | | without | with | without | with |
| C-T | SimCLR | 4.3 | **14.2** | 36.6 | **42.4** |
| | BYOL | 2.3 | **22.5** | 37.6 | **46.8** |
| | DINO | 12.8 | **21.4** | 36.9 | **41.7** |
| A-T | SimCLR | 12.4 | **19.7** | **24.5** | 22.6 |
| | BYOL | 13.8 | **16.4** | 22.3 | **23.1** |
| | DINO | 16.1 | **19.9** | 26.4 | **26.7** |
| L-L | SimCLR | 26.9 | **38.3** | 18.3 | **20.9** |
| | BYOL | 26.3 | **40.7** | 17.5 | **20.6** |
| | DINO | 39.3 | **42.9** | 18.8 | **20.7** |

dataset, which comprises $\sim$ 4M odd-one-out human judgments over image triplets (train set). We construct a Representational Similarity Matrix (RSM) where each entry reflects the frequency with which a pair of categories was selected by participants as the most similar within the triplet, normalized by the number of times that pair appeared in a triplet. Then, we perform an RSA by measuring the correlation between the human RSM and other RSMs (e.g., encoders' representations or co-occurrence statistics). First, to assess the plausibility of our hypothesis, we compare human semantic similarity and object co-occurrence statistics. We find a positive Kendall's Tau of 0.266 and a Spearman correlation score of 0.374, confirming that object co-occurrences correlate with human similarity judgments. Building on this, we compare the organizational structure of visual representations in Table 3 with that of human RSM. We consistently see that at least one layer of the projection head better encodes inter-category human similarity judgments than the final layer of the backbone.

Next, we investigate whether the projection head's superior performance in human similarity judgments stems from its ability to encode object co-occurrences. We gather 14 publicly pretrained SSL models as well as 12 SSL models pretrained on ImageNet-1k for 100 epochs with different numbers of projection layers (see Appendix B for details). This yields a total of 101 representation layers, covering the final layer of the backbones and the internal layers of the projection heads. As seen in Figure 3, representation layers that excel at encoding object co-occurrences tend to perform well in capturing human similarity judgments. A strong Pearson correlation supports this observation with THINGS-COCO ($r = 0.52$, $p = 1.72 \times 10^{-8}$), ADE20k-THINGS ($r = 0.47$, $p = 6.14 \times 10^{-7}$) and LVIS-LVIS ($r = 0.39$, $p = 6.19 \times 10^{-5}$). Taken together, these results suggest that the projection head's ability to encode object co-occurrences makes it more suitable than the visual backbone for modeling inter-category human similarity ratings.

Table 3: RSA Kendall's tau between human inter-category similarity judgments and the visual representations of layers in a SSL model. "P.n" refers to the maximal value within the n-th block "Linear, BatchNorm, ReLU" of the projection head.

| Method | B | P.0 | P.1 | P.2 |
|---|---|---|---|---|
| MoCoV3-50 | 17.8 | **19.8** | 9.0 | na |
| BYOL-R50 | 18.3 | **22.7** | 8.0 | na |
| VICReg-R50 | 17.7 | 21.0 | 13.4 | **23.5** |
| LeJEPA | 24.9 | 27.5 | 28.9 | **29.5** |
| MoCoV3-ViT-B | 18.1 | 19.0 | **22.9** | 21.6 |
| DINO-ViT-B | 30.5 | 28.5 | **31.1** | 28.0 |

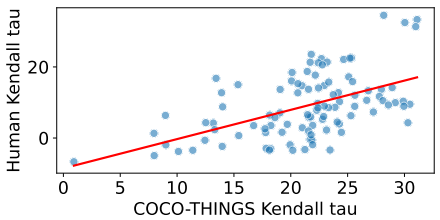


Figure 3: Correlation between encoding COCO-THINGS object co-occurrences and THINGS human similarity judgments, with all pretrained models described in section 3.3.

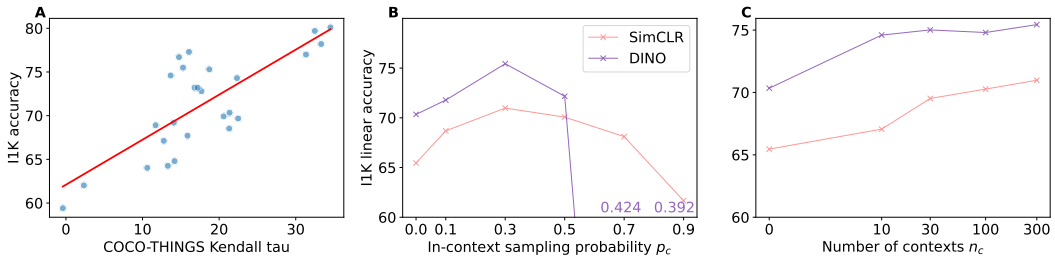


Figure 4: A) Correlation between Imagenet-1k accuracy and the maximum of COCO-THINGS Kendall's tau in the projection head. We plot the Least-Square regression in red. We consider the full set of pretrained models described in section 3.3. B) Imagenet-1k accuracy while varying the probability $p_c$ of sampling two different images from the same context. We use $n_c = 300$. Due to DINO's collapse for $p = 0.7$ and $p = 0.9$, we display the raw accuracies for these experiments. C) Imagenet-1k accuracy against the number of contexts. We use $p_c = 0.3$.

### 4.3 Extracting Object Co-occurrences Boosts Category Learning

Here, we investigate whether learning to encode object co-occurrences enhances category learning in SSL. First, we test whether encoding object co-occurrences in the projection head correlates with improved visual features for downstream categorization. In Figure 4A), we clearly see an increase in category recognition with higher COCO-THINGS Kendall's tau. This is quantitatively confirmed by Pearson scores, with co-occurrences from COCO-THINGS ($r = 0.78$, $p = 7.95 \times 10^{-6}$), ADE20k-THINGS ($r = 0.48$, $p = 1.81 \times 10^{-2}$) and LVIS-LVIS ($r = 0.79$, $p = 4.84 \times 10^{-6}$).

Correlation does not necessarily imply causation. To establish a causal link, we aim to perform controlled experiments that isolate the role of inter-object co-occurrence statistics in shaping category learning. To this end, we define new controlled co-occurrence statistics during training. We leverage images' category-membership information to uniformly and randomly distribute I1K categories into $n_c$ co-occurrence contexts. The higher $n_c$, the lower the number of categories per context. For example, boats and elephants might belong to the same context, while forks and spoons might not. During training, the pipeline remains strictly identical to the original view-invariant SSL method, except for one modification. With probability $p_c$, we sample the second image from a different category within the same context as the first image (e.g. one boat and one elephant in the first and second images, respectively). Therefore, the baseline represented by $p_c = 0$ or $n_c = 0$ is the original SSL method, where no co-occurrence structure is enforced. Figure 6 further illustrates the procedure.

By randomly assigning object categories to contexts, this protocol ensures that any observed performance improvements cannot be attributed to perceptual shortcuts (e.g., images of forks and spoons being similarly represented due to a shared table background). Instead, the model must rely solely on the statistical regularities introduced in our experimental conditions (e.g. boats and elephants are frequently sampled together in a positive pair), which we systematically control by varying $p_c$ and $n_c$. Here, the intuition is analogous to hierarchical feature learning: similar to how supervised category recognition yields early representations of, e.g., edges, learning to align different categories based on co-occurrences may implicitly encourage the model to construct category-level features.

We pretrain two methods (SimCLR, DINO) for 100 epochs under these controlled conditions. In Figure 4B), we see a sweet spot at $p_c = 0.3$, with a drastic accuracy improvement of up to +5.52% compared to the baseline with no hand-crafted contexts ($p_c = 0$). This indicates that co-occurrence statistics can play a substantial role in shaping visual learning. However, the consistent accuracy drop for $p_c > 0.3$ suggests that alignment of same-image augmentations remains a critical learning signal beyond co-occurrence encoding. In Figure 4C), we further find that increasing the number of contexts $n_c$ yields enhanced visual representations, compared to the baseline without hand-crafted contexts ($n_c = 0$). This means that small clusters of repeated co-occurrences, such as one composed of "sink", "toilet" and "toothbrush" (cf. Figure 2), are particularly effective for co-occurrence-based SSL. Due to computational limitations, we compare only $p_c = 0.3$, $n_c = 300$ to $p_c = 0$ for BYOL, VICReg, LeJEPA. Results indicate similar trends (71.60 vs 69.14, 68.57 vs 65.64, 63.74 vs 60.18, respectively). Taken together, these results show that repeated object co-occurrences enhance self-supervised representation learning.

Table 4: In-context one-shot category classification accuracy.

| Method | COCO-THINGS | | | | | LVIS-LVIS | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | B | P.0 | P.1 | P.2 | P.3 | B | P.0 | P.1 | P.2 | P.3 |
| MoCoV3-R50 | **90.0** | 88.7 | 78.6 | na | na | **81.3** | 74.5 | 64.2 | na | na |
| BYOL-R50 | **89.9** | 83.1 | 71.7 | na | na | **80.9** | 76.1 | 65.5 | na | na |
| VICReg-R50 | **87.4** | 83.9 | 81.1 | 72.8 | na | **78.1** | 75.6 | 74.4 | 64.5 | na |
| MoCoV3-ViT-B | **89.7** | 89.6 | 87.1 | 85.0 | na | **80.2** | 79.9 | 76.0 | 72.5 | na |
| DINO-ViT-B | **92.9** | 92.3 | 90.3 | 89.3 | 86.0 | **85.4** | 84.1 | 80.7 | 79.6 | 74.2 |

### 4.4 SSL Hierarchically Extracts Semantic Representations

We found in section 4.1 that the projection head extracts representations that capture object co-occurrence statistics. One possible explanation is that the projection head simply learns better visual representations for capturing diverse semantic information. To verify this, we define the in-context one-shot category classification as follows: we randomly sample an image as well as an image from each category in our context assignments. Then, we assess how frequently an image representation is closer to that of the same category than other categories. We show results only for THINGS-COCO and LVIS-LVIS in Table 4, but consistent results are obtained with THINGS-ADE20k (cf. Appendix C.5). We observe that the performance of the projection consistently decreases with layer depth across all datasets and methods. Together with section 4.1, this demonstrates that SSL models hierarchically organize visual representations, transitioning from categorical to contextual organization in deeper layers.

This result raises the question of whether semantic information is also hierarchically organized within the projection head. Object co-occurrences are naturally hierarchical [Turini and Võ, 2022]. For example, spoons may co-occur very frequently with plates, frequently with bananas, regularly with teapots, occasionally with a television etc. Different layers of the projection head may capture these varying levels of semantic co-occurrences. To investigate this, we define five levels of contextual organization by varying the number of spectral clusters of LVIS categories between {3,10,30,100,300} (cf. section 3.2). Then, we re-train various SSL models for 100 epochs while controlling the number of projection layers (cf. Appendix B). This allows us to compute the context classification accuracy for each layer and contextual level. In Table 5, the results confirm that the projection head consistently outperforms the backbone for encoding object co-occurrences. However, we do not see consistent trends regarding the best layer across levels of contextual organization. This indicates that additional mechanisms might be necessary to enforce a more explicit hierarchical encoding of object co-occurrence statistics within the projection head.

## 5 Conclusion

The learning principles driving visual SSL are poorly understood, the primary assumption being that augmentations should preserve task-label information. Here, we explored the hypothesis that common SSL models also learn by implicitly encoding object co-occurrence statistics. We found that SSL methods trained on Imagenet-1k indeed capture object co-occurrence statistics in their representations. Our analysis further demonstrates that the projection head outperforms the visual backbone in capturing both object co-occurrences and human inter-category similarity judgments. The latter becomes an increasingly important evaluation target in cognitive science Mahner et al. [2025], and our findings suggest that systematically discarding the projection head after pretraining is suboptimal in that case. Furthermore, controlled experiments suggest that encoding object co-occurrences supports the learning of category-wise visual representations in the backbone. Overall, our study reveals the joint role of the projection head and object co-occurrences in visual SSL, paving the way for pre-training strategies that explicitly leverage object co-occurrence statistics.

One limitation of our study is that the controlled co-occurrence statistics introduced in Section4.3 constitute a simplified experimental setting. While they allow us to isolate the role of co-occurrence statistics in category learning, it remains an open question how these statistics relate to the spatial and temporal co-occurrence structure found in heterogeneous datasets such as ImageNet-1K, OpenImages Kuznetsova et al. [2020], or Ego4D Grauman et al. [2022]. Furthermore, we can not rule out that

Table 5: LVIS-LVIS Context classification accuracy for several methods and context definitions.

| Method | # contexts | 3 projection layers | | | | 4 projection layers | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | B | P.0 | P.1 | P.2 | B | P.0 | P.1 | P.2 | P.3 |
| SimCLR-R50 | 3 | 40.4 | 42.9 | **45.4** | 41.5 | 39.9 | 42.2 | 43.2 | **44.6** | 41.7 |
| | 10 | 18.1 | 19.8 | **20.3** | 18.0 | 17.3 | 18.8 | 19.4 | **19.7** | 17.9 |
| | 30 | 12.5 | 14.3 | **15.0** | 13.0 | 13.2 | 13.5 | **14.2** | 13.6 | 12.4 |
| | 100 | 7.7 | 8.7 | **9.4** | 9.2 | 8.1 | 8.5 | 8.1 | 9.1 | **10.0** |
| | 300 | 5.7 | 7.8 | **9.1** | 8.8 | 5.7 | 5.4 | 6.0 | 6.1 | **6.7** |
| BYOL-R50 | 3 | 41.7 | 45.1 | **45.2** | 39.2 | 42.2 | 45.0 | **44.7** | 43.2 | 41.0 |
| | 10 | 17.9 | 20.5 | **20.7** | 16.5 | 17.9 | 20.5 | **20.7** | 16.5 | |
| | 30 | 14.2 | 16.3 | **17.4** | 14.9 | 15.3 | 16.5 | 16.2 | **17.5** | 16.4 |
| | 100 | 8.5 | **9.3** | 8.2 | 8.2 | 8.2 | **9.4** | 8.9 | 8.1 | 8.0 |
| | 300 | 6.5 | **9.3** | 7.3 | 7.8 | 8.0 | 9.4 | 9.4 | 9.6 | **10.8** |
| DINO-ViT-B | 3 | 36.8 | 43.6 | **44.0** | 42.3 | 36.6 | 44.4 | 43.4 | **44.6** | 43.7 |
| | 10 | 18.6 | 19.5 | **20.2** | 18.4 | 19.0 | 19.6 | 19.5 | **20.7** | 18.4 |
| | 30 | 14.6 | 17.3 | **17.7** | 15.5 | 13.6 | 15.1 | 16.4 | 17.7 | **18.4** |
| | 100 | 7.6 | 8.8 | 8.7 | **9.2** | 7.8 | 8.5 | 8.5 | 9.5 | **10.3** |
| | 300 | 5.7 | **10.3** | 7.1 | 5.1 | 5.8 | **9.6** | 8.6 | 9.3 | 6.1 |

some evaluated categories or co-occurrences may be poorly represented in ImageNet-1k, which may harm the absolute performance metric of the models in capturing datasets' object co-occurrence structure.

Our analysis opens several avenues for further investigation. We noticed that multiple factors impact object co-occurrence encoding, such as the neural architecture or the SSL method. Computational constraints prevented the extensive ablation sweeps that would be required to isolate the contribution of individual design choices to the encoding of co-occurrence statistics. Consequently, while our results do not identify which specific design choices are responsible for this effect. Such experiments could nonetheless provide valuable insights for training future self-supervised learning methods. In addition, modern foundational models such as DinoV2 [Oquab et al., 2024] do not release the weights of their projection heads. Future works should confirm our findings on these models. Nevertheless, our study unveils object co-occurrence encoding as a hidden, yet important, learning principle in SSL.

## Acknowledgements

The authors acknowledge the ANR – FRANCE (French National Research Agency) for its financial support of the MeSMERise project ANR-23-CE23-0021-01. This work was funded by the Deutsche Forschungsgemeinschaft: DFG project 5368 ("Abstract REpresentations in Neural Architectures (ARENA)") and supported by the Deutsche Forschungsgemeinschaft (German Research Foundation, DFG) under Germany's Excellence Strategy (EXC 3066/1 "The Adaptive Mind", Project No. 533717223). The authors gratefully acknowledge the computing time provided to them at the NHR Center NHR@SW at Goethe University Frankfurt (project autolearn). This is funded by the Federal Ministry of Education and Research, and the state governments participating on the basis of the resolutions of the GWK for national high performance computing at universities (www.nhr-verein.de/unsere-partner). This work was also supported in part by computational resources provided by the MaSC high-performance computing cluster at Philipps-Universität Marburg, the Hessian Center For Artificial Intelligence and GENCI-IDRIS (Grant 2026-AD011017150).

## References

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020a.

Linus Ericsson, Henry Gouk, and Timothy Hospedales. Why do self-supervised models transfer? on the impact of invariance on downstream tasks. In *The 33rd British Machine Vision Conference, 2022*, page 509. BMVA Press, 2022.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020b.

Jacopo Turini and Melissa Le-Hoa Võ. Hierarchical organization of objects in scenes is reflected in mental representations of objects. *Scientific Reports*, 12(1):20068, 2022.

Nikolaos Giakoumoglou and Tania Stathaki. A review on discriminative self-supervised learning methods. *arXiv preprint arXiv:2405.04969*, 2024.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

Xiao Wang and Guo-Jun Qi. Contrastive learning with stronger augmentations. *IEEE transactions on pattern analysis and machine intelligence*, 45(5):5549–5560, 2022.

Qianjiang Hu, Xiao Wang, Wei Hu, and Guo-Jun Qi. Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1074–1083, 2021.

Xiao Wang, Yuhang Huang, Dan Zeng, and Guo-Jun Qi. Caco: Both positive and negative samples are directly learnable via cooperative-adversarial contrastive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023a.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9588–9597, 2021.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021.

Adrien Bardes, Jean Ponce, and Yann Lecun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR 2022-International Conference on Learning Representations*, 2022.

Feng Wang, Tao Kong, Rufeng Zhang, Huaping Liu, and Hang Li. Self-supervised learning by estimating twin class distribution. *IEEE Transactions on Image Processing*, 2023b.

Yifan Zhang, Zhiquan Tan, Jingqin Yang, Weiran Huang, and Yang Yuan. Matrix information theory for self-supervised learning. In *Proceedings of the 41st International Conference on Machine Learning*, pages 59897–59918, 2024.

Xin Liu, Zhongdao Wang, Ya-Li Li, and Shengjin Wang. Self-supervised learning via maximum entropy coding. *Advances in neural information processing systems*, 35:34091–34105, 2022.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Learning representations by predicting bags of visual words. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6928–6938, 2020.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021.

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15619–15629, 2023.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image bert pre-training with online tokenizer. In *International Conference on Learning Representations*.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, 2024.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Tom George Grigg, Dan Busbridge, Jason Ramapuram, and Russ Webb. Do self-supervised and supervised methods learn similar visual representations?

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*.

Romain Cosentino, Anirvan Sengupta, Salman Avestimehr, Mahdi Soltanolkotabi, Antonio Ortega, Ted Willke, and Mariano Tepper. Toward a geometrical understanding of self-supervised contrastive learning. *arXiv preprint arXiv:2205.06926*, 2022.

Florian Bordes, Randall Balestriero, Quentin Garrido, Adrien Bardes, and Pascal Vincent. Guillotine regularization: Why removing layers is needed to improve generalization in self-supervised learning. *Transactions on Machine Learning Research*.

Yihao Xue, Eric Gan, Jiayi Ni, Siddharth Joshi, and Baharan Mirzasoleiman. Investigating the benefits of projection head for representation learning. In *The Twelfth International Conference on Learning Representations*.

Kartik Gupta, Thalaiyasingam Ajanthan, Anton van den Hengel, and Stephen Gould. Understanding and improving the role of projection head in self-supervised learning. *arXiv preprint arXiv:2212.11491*, 2022.

Yu Gui, Cong Ma, and Yiqiao Zhong. Unraveling projection heads in contrastive learning: Insights from expansion and shrinkage. *arXiv preprint arXiv:2306.03335*, 2023.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International conference on machine learning*, pages 10929–10974. PMLR, 2023.

Tianyu Hua, Wenxiao Wang, Zihui Xue, Sucheng Ren, Yue Wang, and Hang Zhao. On feature decorrelation in self-supervised learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9598–9608, 2021.

Zhuo Ouyang, Kaiwen Hu, Qi Zhang, Yifei Wang, and Yisen Wang. Projection head is secretly an information bottleneck. In *The Thirteenth International Conference on Learning Representations*.

Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. *Advances in Neural Information Processing Systems*, 35:26671–26685, 2022.

Qi Zhang, Yifei Wang, and Yisen Wang. An augmentation overlap theory of contrastive learning. *Journal of Machine Learning Research*, 26(228):1–42, 2025.

Yifei Wang, Qi Zhang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. In *International Conference on Learning Representations*.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in neural information processing systems*, 34:5000–5011, 2021.

Vivien Cabannes, Bobak Kiani, Randall Balestriero, Yann LeCun, and Alberto Bietti. The ssl interplay: Augmentations, inductive bias, and generalization. In *International conference on machine learning*, pages 3252–3298. PMLR, 2023.

Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In *International Conference on Machine Learning*, pages 19250–19286. PMLR, 2022.

Michael F Bonner and Russell A Epstein. Object representations in the human brain reflect the co-occurrence statistics of vision and language. *Nature communications*, 12(1):4081, 2021.

Stefania Bracci, Jakob Mraz, Astrid Zeman, Gaëlle Leys, and Hans Op de Beeck. The representational hierarchy in human and artificial visual systems in the presence of object-scene regularities. *PLoS computational biology*, 19(4):e1011086, 2023.

Arthur Aubret, Céline Teulière, and Jochen Triesch. Seeing the whole in the parts with self-supervised representation learning. *Neurocomputing*, page 132973, 2026.

Arthur Aubret, Timothy Schaumlöffel, Gemma Roig, and Jochen Triesch. Learning object semantic similarity with self-supervision. pages 1–6, 2024.

Timothy Schaumlöffel, Arthur Aubret, Gemma Roig, and Jochen Triesch. Temporal slowness in central vision drives semantic object learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=tEezDE0vWt`.

Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, 12:e82580, 2023.

Luca Medeiros. Language segment-anything. `https://github.com/luca-medeiros/lang-segment-anything`, 2025.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL `https://arxiv.org/abs/2408.00714`.

Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Casey Meehan, Florian Bordes, Pascal Vincent, Kamalika Chaudhuri, and Chuan Guo. Do ssl models have déjà vu? a case of unintended memorization in self-supervised learning. *Advances in Neural Information Processing Systems*, 36:42775–42798, 2023.

Florian P Mahner, Lukas Muttenthaler, Umut Güçlü, and Martin N Hebart. Dimensions underlying the representational alignment of deep neural networks with humans. *Nature Machine Intelligence*, 7(6):848–859, 2025.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022.

Victor Guilherme Turrisi da Costa, Enrico Fini, Moin Nabi, Nicu Sebe, and Elisa Ricci. solo-learn: A library of self-supervised methods for visual representation learning. *Journal of Machine Learning Research*, 23(56):1–6, 2022. URL `http://jmlr.org/papers/v23/21-1155.html`.

Florian Bordes, Randall Balestriero, and Pascal Vincent. Towards democratizing joint-embedding self-supervised learning. *arXiv preprint arXiv:2303.01986*, 2023.

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.

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

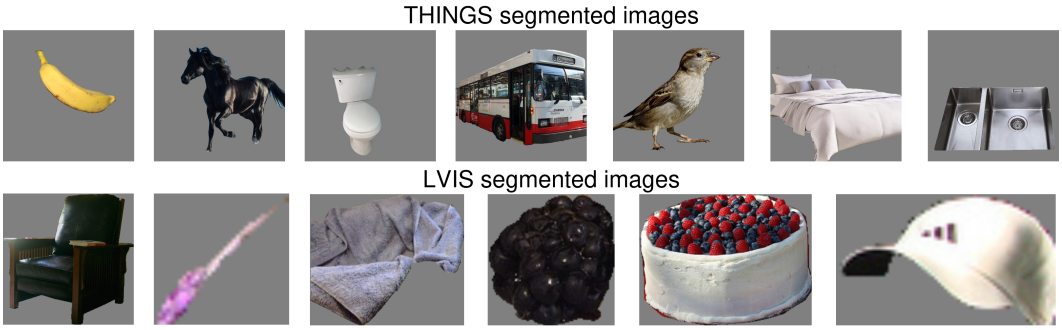


Figure 5: Examples of segmented images from THINGS (top row) and LVIS (bottom row).

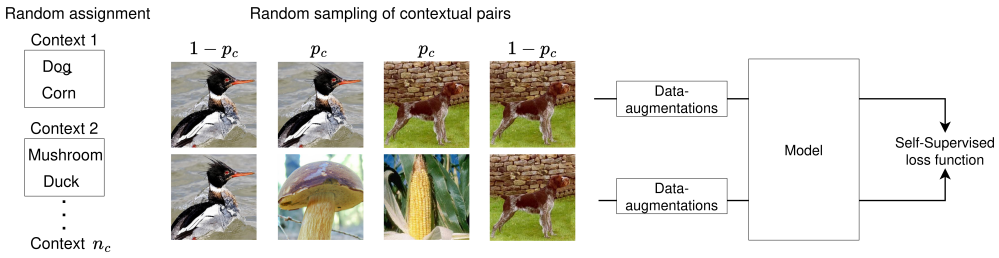


Figure 6: Illustration of the training procedure described in Section C.2.

# A Method visualizations

Figurefig:examples expose examples of images from THINGS and LVIS datasets. Furthermore, Figure 6 further illustrates the training procedure described in Section C.2

# B Pretrained models

Table 6 shows the full set of publicly available pre-trained models used in our experiments. In Appendix C.2 and section 4.4, we pretrain for 100 epochs SSL models with a controlled number of projection heads. We use hyper-parameters provided in the main papers and the solo-learn codebase [da Costa et al., 2022], with few modifications. To adapt to the smaller number of epochs, we increase the base learning rate to 1.6 for Resnet50. We always apply the same augmentation pipeline as in BYOL [Grill et al., 2020]; for DINO, we consider two large crops ([0.4, 1]) and four small crops ([0.05, 0.4]). We always report the final online linear accuracies, following Bordes et al. [2023]. The code and configuration files will be publicly released upon acceptance.

Computational resources required for pre-training the models are described in the original papers. In our experiments, we used 4 NVIDIA A100 for BYOL and 8 AMD Instinct MI210 for SimCLR/DINO. Pretraining SimCLR and BYOL lasted 20 to 30 hours for BYOL/SimCLR, and approximately 72 hours for DINO. Computing the metrics is done in less than 1 hour on four NVIDIA A100.

# C Extended results

## C.1 Standard deviation over context accuracy

Because of space constraints, we could not include the standard deviation of the context accuracy across context definitions in Table 1. Table 7 reports the missing standard deviation. Overall, the accuracy varies only slightly over different context assignments.

Table 6: Full set of publicly available pre-trained models studied in some of our experiments. The top block shows the models used in all our experiments, while others were used for correlation analyses. *We infer that DINO-R50 was pre-trained for 800 or 1000 epochs from the paper.

| Method | Neural network | Pre-training epochs | I1K accuracy |
|---|---|---|---|
| MoCoV3 | Resnet50 | 1000 | 74.6 |
| BYOL | Resnet50 | 800 | 74.4 |
| VICReg | Resnet50 | 1000 | 73.2 |
| MoCoV3 | ViT-B/16 | 300 | 76.7 |
| DINO | ViT-B/16 | 300 | 78.2 |
| MoCoV3 | Resnet50 | 300 | 72.8 |
| MoCoV3 | Resnet50 | 100 | 68.9 |
| MoCoV3 | ViT-S/16 | 300 | 73.2 |
| VICReg | Resnet50x2 | 1000 | 75.5 |
| VICReg | Resnet200x2 | 1000 | 77.3 |
| DINO | Resnet50 | 800-1000* | 75.3 |
| DINO | ViT-S/8 | 300 | 79.7 |
| DINO | ViT-S/16 | 300 | 77 |
| DINO | ViT-B/8 | 300 | 80.1 |

Table 7: Replicate of the context accuracy reported in Table 1, with standard deviation across the three different context assignments.

| Cooc. | Method | Context accuracy | | | |
|---|---|---|---|---|---|
| | | B | P.0 | P.1 | P.2 |
| A-T | MoCoV3-R50 | $22.0 \pm 0.94$ | $\mathbf{23.6 \pm 1.37}$ | $22.6 \pm 1.65$ | na |
| | BYOL-R50 | $23.2 \pm 1.30$ | $\mathbf{26.8 \pm 1.27}$ | $19.5 \pm 0.61$ | na |
| | VICReg-R50 | $24.2 \pm 0.88$ | $\mathbf{26.6 \pm 1.08}$ | $24.6 \pm 0.73$ | $25.7 \pm 1.85$ |
| | MoCoV3-ViT-B | $27.0 \pm 0.86$ | $28.1 \pm 0.95$ | $\mathbf{28.2 \pm 0.84}$ | $24.9 \pm 1.41$ |
| | DINO-ViT-B | $22.8 \pm 1.27$ | $23.6 \pm 1.21$ | $26.7 \pm 1.54$ | $\mathbf{33.0 \pm 0.22}$ |
| L-L | MoCoV3-R50 | $19.0 \pm 0.66$ | $\mathbf{20.2 \pm 0.90}$ | $14.7 \pm 0.41$ | n/a |
| | BYOL-R50 | $18.5 \pm 0.50$ | $\mathbf{20.6 \pm 0.53}$ | $13.3 \pm 0.59$ | n/a |
| | VICReg-R50 | $19.0 \pm 0.51$ | $\mathbf{19.7 \pm 0.51}$ | $18.6 \pm 0.54$ | $16.1 \pm 2.26$ |
| | MoCoV3-ViT-B | $18.9 \pm 0.45$ | $\mathbf{19.6 \pm 0.92}$ | $19.5 \pm 1.14$ | $18.4 \pm 0.77$ |
| | DINO-ViT-B | $21.4 \pm 1.07$ | $21.1 \pm 1.08$ | $\mathbf{23.9 \pm 1.14}$ | $22.6 \pm 1.01$ |

### C.2 Encoding natural co-occurrences with controlled projection heads

Here, we verify whether our findings of section 4.1 generalize to models trained with various projection heads. In Table 8, we see that at least one internal layer in the projection head presents improved object co-occurrence encoding, compared to the final layer of the visual backbone. We also observe that this does not necessarily hold when the projection head is composed of a single linear layer (*-1 methods). This confirms our previous findings. We also see that, the deeper the projection head, the deeper the optimal layer for encoding object co-occurrences. This trend is consistent across all methods, datasets and metrics, with a single exception (BYOL-4 for COCO-THINGS RSA Kendall's tau). Overall, this shows that a representational structure capturing object co-occurrences tend to appear in deeper layers of the projection, though not the final one.

### C.3 Co-occurrence encoding metrics based on image-level comparisons

In our method (section 3.2), we built category representations to derive the representational dissimilarity matrices of models' layers. This tends to suppress intra-category variations between images. Here, we check whether our conclusions hold when we discard this and directly compute similarities between image pairs, which are thereafter averaged across pairs of categories. To derive a context accuracy, we perform one-shot context classification across 2,000, 2,000 and 10,000 tuples of images,

Table 8: Object Co-occurrences encoding metrics for 3 models (method-n) where n represents the number of linear layers in the projection head.

| Method | RSA Kendall's tau | | | | | Context accuracy | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | B | mP.0 | mP.1 | mP.2 | mP.3 | B | P.0 | P.1 | P.2 | P.3 |
| COCO-THINGS | | | | | | | | | | |
| SimCLR-1 | **4.3** | -0.4 | na | na | na | **36.6** | 35.3 | na | na | na |
| SimCLR-2 | 6.3 | **14.2** | -0.6 | na | na | 38.4 | **42.4** | 38.1 | na | na |
| SimCLR-3 | 4.0 | **13.3** | 9.3 | -1.9 | na | 36.5 | **40.7** | 39.6 | 34.8 | na |
| SimCLR-4 | 4.5 | **10.6** | 7.3 | 8.9 | 0.9 | 29.3 | 35.4 | 35.6 | **38.6** | 34.8 |
| BYOL-1 | -1.9 | **2.3** | na | na | na | 34.0 | **37.6** | na | na | na |
| BYOL-2 | 2.2 | **22.5** | -3.4 | na | na | 35.2 | **46.8** | 35.6 | na | na |
| BYOL-3 | 5.3 | 12.0 | **14.1** | -1.9 | na | 36.4 | **42.7** | 41.8 | 33.7 | na |
| BYOL-4 | 6.1 | **15.9** | 15.7 | 13.7 | 4.3 | 39.7 | 42.8 | 40.1 | **43.8** | 35.1 |
| DINO-1 | 7.0 | **12.8** | na | na | na | **36.9** | 30.4 | na | na | na |
| DINO-2 | -2.9 | **21.3** | 8.6 | na | na | 24.9 | **41.0** | 35.3 | na | na |
| DINO-3 | -3.4 | 18.4 | **21.4** | 7.8 | na | 26.8 | **41.7** | 41.4 | 30.3 | na |
| DINO-4 | -3.5 | 12.2 | 12.7 | **20.6** | 9.0 | 23.5 | 37.5 | **38.5** | 36.0 | 33.7 |
| ADE20k-THINGS | | | | | | | | | | |
| SimCLR-1 | **12.4** | 11.5 | na | na | na | 23.7 | **24.5** | na | na | na |
| SimCLR-2 | 11.7 | **19.7** | 12.7 | na | na | 20.6 | **22.6** | 22.2 | na | na |
| SimCLR-3 | 8.8 | 15.6 | **18.6** | 12.2 | na | 20.3 | 23.3 | **24.4** | 22.8 | na |
| SimCLR-4 | 8.2 | 17.0 | **20.5** | 18.2 | 13.4 | 23.0 | 24.4 | **25.9** | 24.6 | 25.4 |
| BYOL-1 | 10.6 | **13.8** | na | na | na | 19.9 | **22.3** | na | na | na |
| BYOL-2 | 9.3 | **16.4** | 11.5 | na | na | 20.5 | **23.1** | 20.7 | na | na |
| BYOL-3 | 9.0 | 17.7 | **24.4** | 10.2 | na | 20.2 | 21.0 | **24.0** | 21.5 | na |
| BYOL-4 | 8.7 | 18.1 | **23.4** | 19.1 | 16.9 | 19.9 | 20.6 | 21.3 | **25.3** | 22.4 |
| DINO-1 | **16.1** | 14.0 | na | na | na | 24.0 | **26.4** | na | na | na |
| DINO-2 | 3.7 | **15.4** | 14.7 | na | na | 12.6 | **27.3** | 25.6 | na | na |
| DINO-3 | 3.2 | 17.0 | **19.9** | 14.2 | na | 13.1 | 24.9 | **26.7** | 19.6 | na |
| DINO-4 | 2.8 | 17.0 | **17.4** | 16.9 | 13.5 | 13.7 | 24.7 | 25.9 | 26.0 | **27.7** |
| LVIS-LVIS | | | | | | | | | | |
| SimCLR-1 | **26.9** | 26.6 | na | na | na | **18.3** | 17.2 | na | na | na |
| SimCLR-2 | 33.5 | **38.3** | 26.0 | na | na | 18.8 | **20.9** | 18.1 | na | na |
| SimCLR-3 | 32.8 | **37.0** | 36.9 | 26.1 | na | 18.1 | 19.8 | **20.3** | 18.0 | na |
| SimCLR-4 | 26.7 | 31.4 | **33.2** | **33.2** | 25.2 | 17.3 | 18.8 | 19.4 | **19.7** | 17.9 |
| BYOL-1 | 23.9 | **26.3** | na | na | na | **17.5** | 16.9 | na | na | na |
| BYOL-2 | 33.5 | **40.7** | 21.7 | na | na | 18.3 | **20.6** | 14.2 | na | na |
| BYOL-3 | 37.1 | 40.9 | **42.0** | 24.5 | na | 17.9 | 20.5 | **20.7** | 16.5 | na |
| BYOL-4 | 38.6 | 41.0 | **42.7** | 35.9 | 26.2 | 18.7 | 20.1 | **20.5** | 20.2 | 18.2 |
| DINO-1 | **39.3** | 27.4 | na | na | na | **18.8** | 17.7 | na | na | na |
| DINO-2 | 41.1 | **43.4** | 25.7 | na | na | 19.1 | **20.6** | 16.8 | na | na |
| DINO-3 | 40.1 | 42.4 | **42.9** | 26.8 | na | 18.6 | 19.5 | **20.2** | 18.4 | na |
| DINO-4 | 39.8 | 41.0 | 40.6 | **42.2** | 26.6 | 19.0 | 19.6 | 19.5 | **20.7** | 18.4 |

Table 9: Image-level object co-occurrences encoding metrics for 3 models (method-n) where n represents the number of linear layers in the projection head.

| Method | RSA Kendall's tau | | | | | Context accuracy | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | B | mP.0 | mP.1 | mP.2 | mP.3 | B | P.0 | P.1 | P.2 | P.3 |
| COCO-THINGS | | | | | | | | | | |
| SimCLR-1 | **0.5** | -4.1 | na | na | na | **27.6** | 26.4 | na | na | na |
| SimCLR-2 | 1.0 | **9.0** | -4.1 | na | na | 29.1 | **31.3** | 28.5 | na | na |
| SimCLR-3 | -1.2 | **6.6** | 4.5 | -5.1 | na | 27.6 | **30.3** | 30.1 | 25.7 | na |
| SimCLR-4 | **9.5** | -1.2 | 6.6 | 4.5 | -5.1 | 25.1 | 27.8 | 27.6 | **29.6** | 27.1 |
| BYOL-1 | -4.1 | **-0.3** | na | na | na | 25.5 | **26.4** | na | na | na |
| BYOL-2 | -2.4 | **20.1** | -5.0 | na | na | 28.9 | **32.9** | 22.2 | na | na |
| BYOL-3 | 1.2 | 6.7 | **11.5** | -3.7 | na | 29.4 | 31.5 | **32.7** | 23.3 | na |
| BYOL-4 | 2.3 | 13.2 | **15.9** | 13.6 | 4.0 | 32.4 | 32.7 | 33.3 | **34.0** | 29.4 |
| DINO-1 | -2.0 | **16.1** | na | na | na | 16.6 | **25.4** | na | na | na |
| DINO-2 | -1.7 | **20.9** | 10.5 | na | na | 17.1 | **29.8** | 24.5 | na | na |
| DINO-3 | -2.7 | 18.6 | **23.4** | 9.7 | na | 17.2 | 30.9 | **33.3** | 26.6 | na |
| DINO-4 | -2.2 | 10.7 | 14.7 | **27.1** | 13.1 | 17.2 | 30.0 | 29.6 | **30.1** | 26.2 |
| ADE20k-THINGS | | | | | | | | | | |
| SimCLR-1 | 5.3 | **5.5** | na | na | na | **20.3** | 19.1 | na | na | na |
| SimCLR-2 | 2.4 | **12.8** | 7.0 | na | na | 17.8 | **20.3** | 18.8 | na | na |
| SimCLR-3 | -0.9 | 7.3 | **12.5** | 6.4 | na | 17.0 | 17.9 | **20.6** | 18.6 | na |
| SimCLR-4 | 0.1 | 8.1 | **12.5** | 11.6 | 7.7 | 16.8 | 17.9 | 19.2 | **20.1** | 18.2 |
| BYOL-1 | 5.5 | **9.9** | na | na | na | **16.6** | 16.1 | na | na | na |
| BYOL-2 | 0.8 | 7.1 | **9.2** | na | na | 17.4 | **18.7** | 17.2 | na | na |
| BYOL-3 | -1.1 | 10.3 | **19.9** | 8.1 | na | 18.0 | 16.9 | **21.0** | 17.7 | na |
| BYOL-4 | -0.9 | 10.7 | **16.7** | 14.5 | 15.9 | 18.2 | 18.2 | **20.3** | 17.7 | 18.9 |
| DINO-1 | 1.5 | **11.7** | na | na | na | 10.3 | **20.5** | na | na | na |
| DINO-2 | 1.6 | **10.9** | 11.5 | na | na | 10.6 | 19.6 | **19.9** | na | na |
| DINO-3 | 1.3 | 5.5 | **11.5** | 10.7 | na | 11.6 | 18.1 | **19.6** | 18.7 | na |
| DINO-4 | 1.0 | 8.1 | 7.6 | 8.2 | **10.4** | 10.8 | 18.6 | 19.1 | 20.9 | **21.5** |
| LVIS-LVIS | | | | | | | | | | |
| SimCLR-1 | 1.4 | **2.2** | na | na | na | **14.3** | 13.1 | na | na | na |
| SimCLR-2 | 3.8 | **11.7** | 1.7 | na | na | 14.1 | **15.2** | 13.6 | na | na |
| SimCLR-3 | -0.9 | 3.8 | **8.1** | 1.8 | na | 13.7 | **14.9** | **14.9** | 13.5 | na |
| SimCLR-4 | 1.1 | 2.8 | 6.1 | **7.9** | 3.6 | 12.9 | 14.2 | 14.2 | **14.6** | 13.2 |
| BYOL-1 | 0.7 | **2.3** | na | na | na | **14.0** | 13.8 | na | na | na |
| BYOL-2 | 0.5 | **11.0** | 0.5 | na | na | 14.0 | **15.7** | 12.3 | na | na |
| BYOL-3 | 4.8 | 7.6 | **11.1** | 0.2 | na | 14.0 | **15.8** | 15.5 | 14.0 | na |
| BYOL-4 | 4.7 | 9.1 | **11.8** | 10.8 | 5.3 | 15.1 | **16.1** | 15.4 | 15.2 | 15.2 |
| DINO-1 | -0.0 | **14.1** | na | na | na | **15.1** | 14.4 | na | na | na |
| DINO-2 | -0.2 | **10.3** | 14.6 | na | na | **15.5** | 15.3 | 14.8 | na | na |
| DINO-3 | -0.6 | 2.3 | 12.2 | **14.9** | na | 14.7 | 15.0 | **15.8** | 14.4 | na |
| DINO-4 | 0.8 | 5.4 | -1.2 | 14.4 | **15.0** | 14.7 | 15.2 | 15.4 | **16.6** | 14.9 |

Table 10: In-context one-shot category classification accuracy. We show average accuracy ± standard deviation, computed across three context assignments.

| Method | B | P.0 | P.1 | P.2 | P.3 |
|---|---|---|---|---|---|
| MoCoV3-R50 | **91.9 ± 0.56** | 84.4 ± 0.87 | 83.7 ± 0.57 | na | na |
| BYOL-R50 | **90.2 ± 0.63** | 86.1 ± 0.32 | 73.5 ± 1.09 | na | na |
| VICReg-R50 | **89.4 ± 0.96** | 85.8 ± 0.69 | 83.8 ± 0.83 | 80.0 ± 1.35 | na |
| MoCoV3-ViT-B | 91.1 ± 1.00 | **91.3 ± 1.25** | 88.3 ± 0.65 | 90.3 ± 1.12 | na |
| DINO-ViT-B | **94.6 ± 0.18** | 93.6 ± 0.23 | 90.7 ± 0.45 | 94.3 ± 0.16 | 89.7 ± 0.71 |

Table 11: ImageNet-1k linear recognition accuracy.

| Method | B | mP.0 | mP.1 | mP.2 |
|---|---|---|---|---|
| MoCoV3-R50 | **73.30** | 72.52 | 64.38 | – |
| BYOL-R50 | **70.76** | 70.43 | 46.07 | – |
| VICReg-R50 | **71.30** | 69.25 | 67.29 | 60.36 |
| MoCoV3-ViT-B | **76.53** | 76.29 | 74.89 | 67.24 |
| DINO-ViT-B | 77.50 | **78.09** | 75.86 | 68.48 |

instead of tuples of category prototypes, for COCO-THINGS, ADE20k-THINGS and LVIS-LVIS, respectively. In Table 9, we see a trend consistent with our previous results: with a non-linear projection head, the projection head consistently outperforms the last layer of the backbone in encoding object co-occurrences. However, the results are generally lower than with our main evaluation protocol (Appendix 8), indicating layers encodes features that vary across images from the same categories, which impedes category co-occurrence encoding.

### C.4 Evaluation of more hidden layers

We found in section 4.1 that the projection head learns representations that better model natural object co-occurrence, compared to the final layer of the visual backbone. Here, we extend our analysis to consider more layers within the neural network model. For internal layers, we take the spatial average of features for the ResNet50, and the register token for Vision Transformers. For the latter, we found it to work better than spatial average in general. In Figure 7A-B), we see that Vision Transformers and ResNets progressively encode inter-object semantic relationships in a different manner. ResNets make little progress through the processing hierarchy for encoding object co-occurrences. In contrast, Vision Transformers exhibit a rapid increase in correlation, before dropping and increasing again with the projection head. However, this specific pattern does not old for ADE20k (Figure 7C), suggesting that this may be due to the specific images of COCO-THINGS. Overall, these results confirm the superiority of layers in the projection head for encoding object co-occurrences.

### C.5 Category classification

In section 4.4, we showed that the in-context category classification decreases along deeper layers of the projection head, for THINGS-COCO and LVIS-LVIS setups. Here, we analyze results for the THINGS-ADE20k setup. In Table 10, we observe that the accuracy generally decreases across layers of the projection head (BYOL-R50, VICReg-R50), confirming our prior findings. However, the accuracy drop is less pronounced than THINGS-COCO and LVIS-LVIS (section 4.4) and remains stable in some cases (MoCoV3-ViT-B).

We further test category recognition abilities in the projection head, by training a linear classifier over each layer on Imagenet-1k. In Table 11, we also generally observe a decrease of category recognition accuracy throughout the projection head.

### C.6 Evaluation of foundational backbones

Despite the absence of publicly released weights in foundational models, we report the performance of CLIP [Radford et al., 2021], DinoV2 [Oquab et al., 2024] and a supervised resnet50 trained on

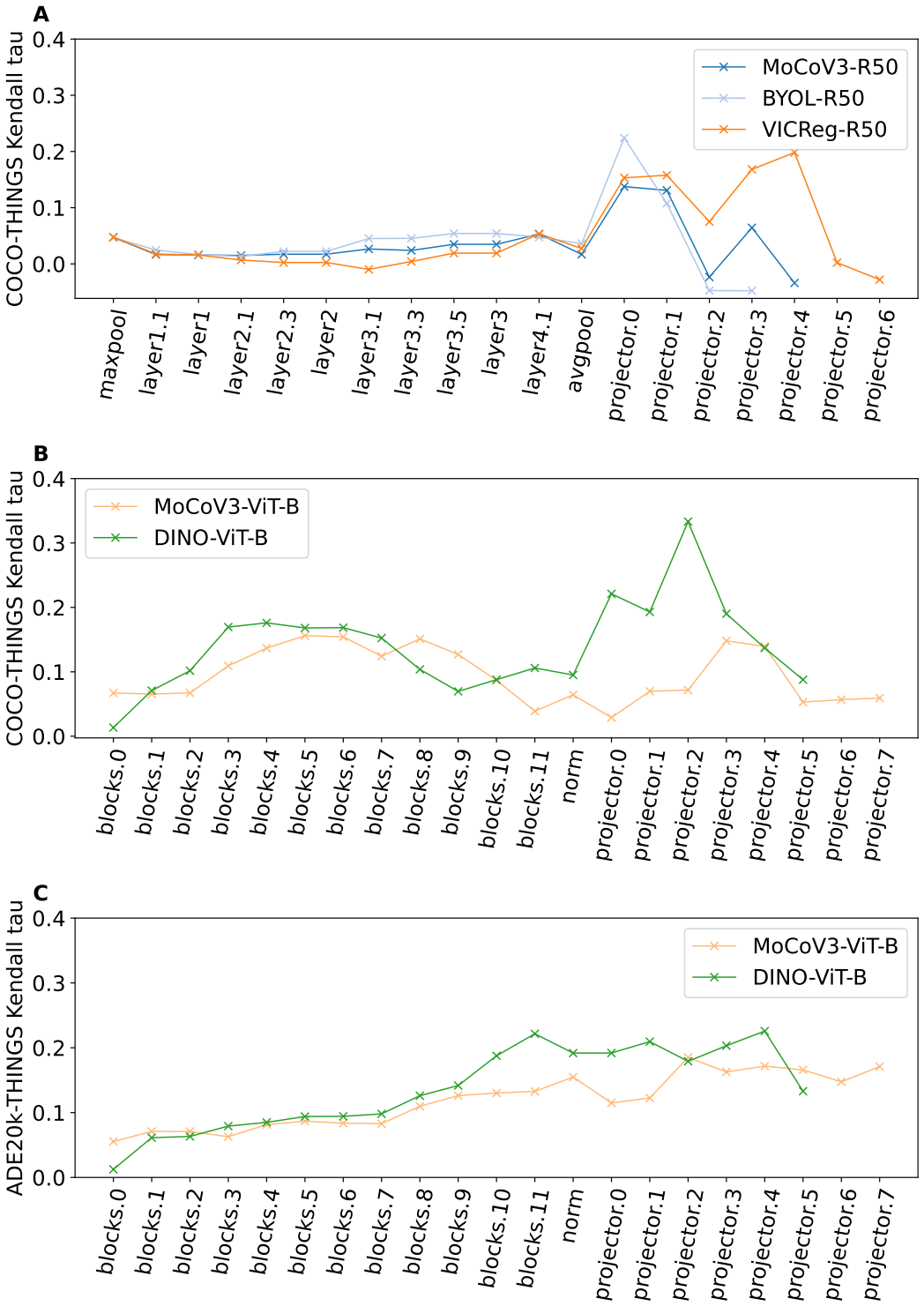


Figure 7: RSA Kendall's tau for 12 layers prior to the projection head, and layers in the projection head for the five main pre-trained models considered in our study. We consider A) COCO-THINGS with ResNet50 models, B) COCO-THINGS with ViT models and C) ADE20k-THINGS with ViT models.

Table 12: Object co-occurrence encoding metrics for each model.

| Method | RSA Kendall's tau | Context accuracy |
|---|---|---|
| COCO-THINGS | | |
| DINO-ViT-B (backbone) | 9.0 | 43.5 |
| DINO-ViT-B (projection head) | **33.4** | 50.3 |
| CLIP-ViT-L (backbone) | 28.1 | **58.2** |
| DINOv2-ViT-L (backbone) | 22.4 | 37.6 |
| Supervised R50 (backbone) | 10.6 | 36.6 |
| ADE20k-THINGS | | |
| DINO-ViT-B (backbone) | 19.3 | 22.4 |
| DINO-ViT-B (projection head) | 22.4 | **33.0** |
| CLIP-ViT-L (backbone) | **24.7** | 28.1 |
| DINOv2-ViT-L (backbone) | 20.4 | 29.9 |
| Supervised R50 (backbone) | 23.5 | 22.6 |
| LVIS-LVIS | | |
| DINO-ViT-B (backbone) | 39.8 | 21.4 |
| DINO-ViT-B (projection head) | 43.3 | 23.9 |
| CLIP-ViT-L (backbone) | **44.4** | 21.9 |
| DINOv2-ViT-L (backbone) | 41.5 | **24.3** |
| Supervised R50 (backbone) | 37.0 | 20.3 |

Imagenet-1k [He et al., 2016] in encoding object co-occurrences. Interestingly, we find that the projection head of DINO rivals with models that benefit supervision or bigger image datasets. The projection head is crucial to bridge the gap between the backbone of DINO and backbones of other methods.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All claims are supported by our experiments.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have included a paragraph about limitations in the Conclusion section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification:

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The method section accurately describes or analysis protocol. Appendix B describes the hyper-parameters used to pretrain our models.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: We plan to release our documented code upon acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix B describes the hyper-parameters used to pre-train our models.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Appendix C.1 reports the missing standard deviation in our experiments. For clarity and compactness, we do not report it for all our results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We give information about our hardware and execution time in Appendix B

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have verified each point of the code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work is an analysis of existing methods. There is no direct societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite each dataset and code that we use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification:

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.