Training Vision Transformers to Focus: Emergence of Spatial Attention Head Specialization

Yuji Nozawa
Kioxia Corporation
yuji1.nozawa@kioxia.com

Yu-Chieh Lin
Kioxia Corporation
yuchieh.lin@kioxia.com

Youyang Ng
Kioxia Corporation
youyang.ng@kioxia.com

Abstract

Self-supervised learning in Vision Transformer models, such as DINO, enables robust visual perception without requiring labeled data by leveraging large-scale image datasets. However, these models lack a focusing mechanism during training, resulting in perceptual representations that are often ambiguous or overly similar. In this work, we introduce two visual focusing loss functions designed to establish correspondence between the model's attention and specific regions within images. Specifically, we leverage the multi-head attention mechanism within the model to selectively steer the model's focus, leading to the emergence of spatial attention head specialization, characterized by diverse and spatially focused perceptual capabilities without requiring annotations. Through both qualitative and quantitative evaluations, we demonstrate that our method substantially increases the spatial selectivity and diversity of attention heads, improving both interpretability and performance on dense prediction tasks such as zero-shot image segmentation.

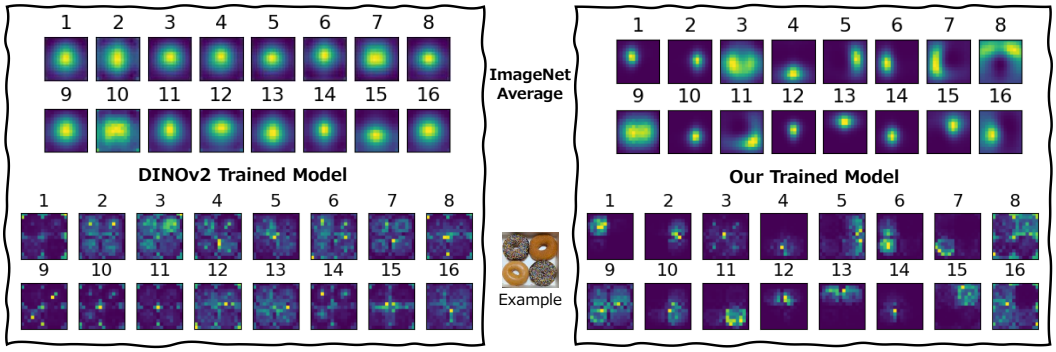

Figure 1: Comparison of attention maps produced by the DINOv2 trained model and our trained model. For each model, we show the attention maps averaged over the validation set of the ImageNet-1k (Russakovsky et al., 2015) and the attention maps obtained from a single example image. Existing learned perceptual representations (left) in individual attention heads often tend to be ambiguous or overly similar within the example image and across images in the dataset, which limits their interpretability. In contrast, our method (right) selectively steers the focus, leading to the emergence of diverse and spatially focused perceptual capabilities. (Images: Visualization of 16 attention heads in the last layer of trained ViT-L/16-Reg model)

1 Introduction

Self-supervised learning has become a powerful paradigm in computer vision, enabling models to acquire rich and meaningful visual representations without the need for costly and time-consuming

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

labeled data. Among various architectures, Vision Transformers (ViTs) (Dosovitskiy et al., 2021) have attracted considerable attention for their remarkable ability to capture long-range dependencies and complex visual patterns within images. Methods such as Self-Distillation with No Labels (DINO) (Caron et al., 2021) have demonstrated that large-scale unlabeled image datasets can effectively train ViTs to produce robust and transferable visual features. Despite these advances, a fundamental challenge remains: unlike supervised approaches that rely on explicit guidance via labels or self-supervised approaches that incorporate prior knowledge to guide attention, self-supervised ViT models without such explicit guidance or prior knowledge typically lack a focusing mechanism during training. Consequently, the learned perceptual representations often tend to be ambiguous or overly similar across attention heads, which limits their discriminative capacity and interpretability.

Addressing this limitation is crucial for extending the applicability of self-supervised ViTs, particularly in tasks that require dense representations such as object segmentation and interpretability analysis. To this end, we propose two visual focusing loss functions that establish correspondence between the model's attention and specific regions within images during training. We refer to this correspondence as *region-to-focus correspondence* (ReFoCo). The first loss, the *inside-focus loss*, is designed to align the model's focus on local regions within images with the cropped images of these corresponding regions. The second loss is inspired by weakly-supervised learning with masked images as proposed in Kumar Singh and Jae Lee (2017), where masking forces the model to focus not only on the most discriminative object parts, but also on other regions. Our proposed second loss, the *outside-focus loss*, is designed to achieve a similar effect in a self-supervised manner by associating partially masked images with the model's focus on regions outside the masked parts.

Our approach leverages the multi-head attention mechanism in ViTs to selectively steer the model's focus, and we find that diverse and spatially focused perceptual capabilities emerge in our trained model, as shown in Fig. 1. Our method also substantially improves the model's interpretability.

We conduct extensive qualitative and quantitative evaluations to validate the effectiveness of our approach. The results demonstrate that incorporating our visual focusing losses substantially improves the model's ability to localize and differentiate between various image regions. These improvements lead to higher zero-shot image segmentation performance, unsupervised object discovery performance, and multi-object image retrieval performance, along with clearer and more interpretable attention maps, highlighting the practical advantages of our method.

We note that while the proposed losses encourage the emergence of diverse and spatially focused perceptual capabilities, we observe a slight reduction in classification performance. Accordingly, our results should not be interpreted as demonstrating universally stronger global or dense representations. Rather, our method aims to introduce focusing mechanisms into self-supervised ViT learning without requiring annotations, encouraging attention heads to capture diverse and spatially focused visual information.

Overall, our contributions are as follows:

- We identify the ambiguity and similarity of perceptual representations present in current self-supervised ViT models, which limits their interpretability.
- We present a new direction for introducing focusing mechanisms into self-supervised ViTs learning without requiring annotations, using visual focusing loss functions that link representation learning more directly to visual understanding.
- We evaluate the performance of models trained using our visual focusing loss functions and demonstrate that they achieve higher zero-shot image segmentation performance, unsupervised object discovery performance, and multi-object image retrieval performance, while also producing more interpretable and meaningful attention maps.

2 Related work

Vision Transformers and self-supervised learning. The advent of ViTs (Dosovitskiy et al., 2021) has ushered in a paradigm shift in self-supervised learning. Leveraging their global receptive fields and adaptive attention mechanisms, ViTs effectively capture long-range dependencies that conventional convolutional neural networks (Simonyan and Zisserman, 2015; He et al., 2016) typically struggle to model. Techniques such as DINO (Caron et al., 2021) and iBOT (Zhou et al., 2022) have demonstrated that ViTs are effective at learning structured, semantically rich representations without

the need for labeled data. Remarkably, both DINO and its successor, DINOv2 (Oquab et al., 2024), shows that ViTs can generate attention maps closely aligned with object boundaries, suggesting an emergent capacity for dense visual recognition.

Visual focusing mechanisms. However, existing self-supervised ViT methods often lack effective focusing mechanisms during training. Consequently, multiple attention heads tend to converge toward similar activation patterns, producing ambiguous or redundant representations (Bhojanapalli et al., 2021). Some approaches have attempted to mitigate this issue by promoting attention map diversification (Ren et al., 2025), employing block dropping (Bhojanapalli et al., 2021), or knowledge distillation (Wu et al., 2022) techniques to reduce redundancy. However, these strategies are typically indirect, relying on architectural heuristics rather than explicitly designed loss functions that encourage interpretability or spatial differentiation. Post-hoc explanation methods, such as attention rollout (Abnar and Zuidema, 2020; Jo et al., 2025) and gradient-based techniques (Selvaraju et al., 2017), can reveal where the model focuses after training but do not influence the training dynamics. Other efforts have introduced architectural modifications (Chen et al., 2023), add-on modules (Lee et al., 2021), or prior knowledge-guidance (Miao et al., 2022) to foster diversity across attention heads or enhance spatial localization. Yet these methods often require supervision or additional annotations, limiting their applicability in fully self-supervised settings.

In summary, although current methods successfully capture global structure and high-level semantics, they often lack mechanisms that promote diverse, meaningful, and discriminative attention across different image regions. The proposed visual focusing losses offer a principled approach to encourage spatial specialization, enhancing interpretability within self-supervised ViT frameworks. Integrating such focusing mechanisms that shape attention during training is a relatively unexplored direction with the potential to link self-supervised representation learning to transparent and spatially grounded visual reasoning.

Visual prompt techniques. Inspired by prompt engineering in natural language processing (Brown et al., 2020), a variety of visual prompts have been studied in computer vision to intuitively modify the behavior of models in a visual manner (Kirillov et al., 2023; Yang et al., 2023; Shtedritski et al., 2023; Zhuang et al., 2024; Jia et al., 2022; Park and Byun, 2024; Yoo et al., 2023; Bahng et al., 2022; Nozawa et al., 2025). In this work, to selectively steer the model's focus, we leverage a visual prompting technique called prompt-guided attention head selection (PHS) (Nozawa et al., 2025). In PHS, the visual prompt is given as a set of binary values R assigned to each pixel in the image, where each value indicates whether it is included in the region of interest (ROI) or not. This has the advantage that the shape of the ROI is not restricted.

3 Proposed method

We consider self-supervised learning settings, such as DINO (Caron et al., 2021), where teacher and student models are used and training is performed by minimizing a loss function between the two models. To enable the model to acquire the ability to focus on the ROI, we propose to introduce two visual focusing loss functions as shown in Fig. 2.

First, we briefly introduce the PHS method (Nozawa et al., 2025) for modifying the model's output using the ROI. Once the ROI R is specified, the attention heads that produce the largest outputs of attention maps within the ROI are selected. By applying the head selection method (Gong et al., 2021; Meng et al., 2022; Ni et al., 2024; Wu et al., 2022; Nicolicioiu et al., 2023; Basile et al., 2025) that emphasizes the outputs of the selected heads, the model's overall output is modified to align with the ROI. Formally, let $H = \{1, 2, \ldots, h\}$ denote the set of the heads, and let $i_{\text{on}} \subseteq H$ be the set of the selected heads. The Multi-Head Attention in the last attention layer, $\text{Linear}\left([\text{head}_1, \text{head}_2, \ldots, \text{head}_h]\right)$ is then modified by replacing head_i with $\text{head}_i^{\text{HS}}$, defined by

$$\text{head}_i^{\text{HS}} = \begin{cases} \text{head}_i \times h/h_{\text{on}} & \text{if } i \in i_{\text{on}}, \\ 0 & \text{if } i \notin i_{\text{on}}, \end{cases} \tag{1}$$

where $h_{\text{on}} = |i_{\text{on}}|$ is the number of the selected heads. As a result, for each input image x, the model's output $P(x)$ changes to $P^{\text{PHS}}(x, R)$.

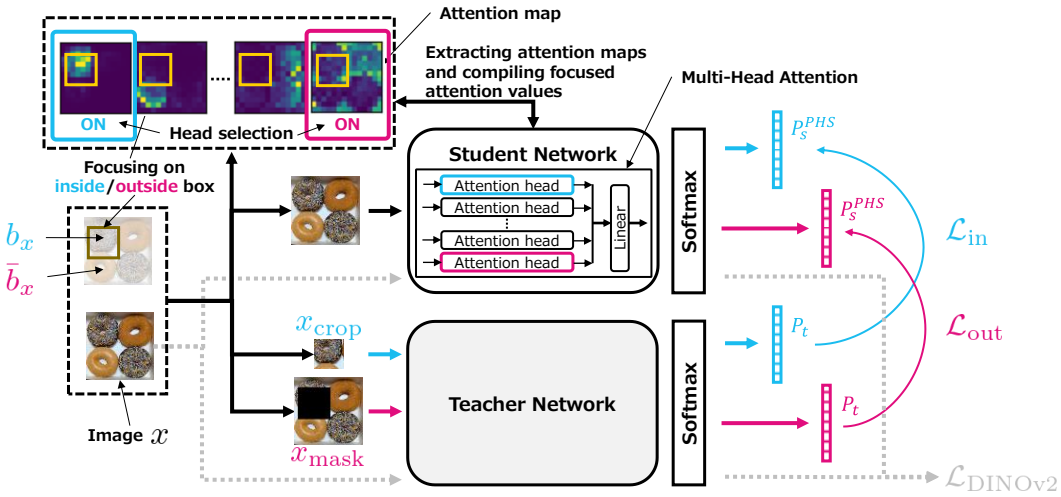

Figure 2: Overview of our proposed visual focusing loss functions $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$.

Inside-focus loss. The first loss function aims to align the model's focus on local areas within images with the cropped images of these corresponding areas. For each training image x, we designate a box b_x within the image as the ROI. Then, the cropped image corresponding to the box, denoted as x_{crop}, is input to the teacher model. Meanwhile, the entire image is input to the student model. However, instead of simply providing the input to the student model, we apply the PHS (Nozawa et al., 2025) method, which enhances the intermediate outputs related to the ROI within the model as described above. The inside-focus loss function $\mathcal{L}_{\text{in}}$ is defined as the cross-entropy loss between the outputs of the teacher and student networks:

$$\mathcal{L}_{\text{in}} = -\frac{1}{B}\sum_{x} P_{\text{t}}(x_{\text{crop}}) \log P_{\text{s}}^{\text{PHS}}(x, R = b_x), \tag{2}$$

where B is the batch size, the summation runs over the batch, P_{t} denotes the softmax output of the teacher network, and $P_{\text{s}}^{\text{PHS}}$ represents the softmax output of the student network with PHS visually prompting the box b_x applied.

Outside-focus loss. To enable the model to focus not only on the most discriminative object or part but also on other regions, we introduce another loss function. Similar to the inside-focus loss function $\mathcal{L}_{\text{in}}$, we designate a box b_x as the ROI for each image x. However, for the teacher model in this case, we input the image in which the region within the box is masked, denoted as x_{mask}. For the student model, the entire image is input, similar to the case of $\mathcal{L}_{\text{in}}$. The difference is the use of PHS, which utilizes the external region of the box, $\bar{b}_x$, as a visual prompt. The outside-focus loss function $\mathcal{L}_{\text{out}}$ is then given by the following cross-entropy loss:

$$\mathcal{L}_{\text{out}} = -\frac{1}{B}\sum_{x} P_{\text{t}}(x_{\text{mask}}) \log P_{\text{s}}^{\text{PHS}}(x, R = \bar{b}_x). \tag{3}$$

We propose that by combining these visual focusing loss functions with conventional loss functions used in self-supervised learning, the model effectively acquire the ability to focus on various objects or regions within the images.

4 Experiments

4.1 Interpretability comparison of attention maps for conventional training methods

Since our training method utilizes attention maps via PHS, interpretability of the maps is crucial. To determine the training method adopted in our method, we compare the attention maps of four different training methods: DeiT III (Touvron et al., 2022), CLIP (Radford et al., 2021), DINOv2 (Oquab

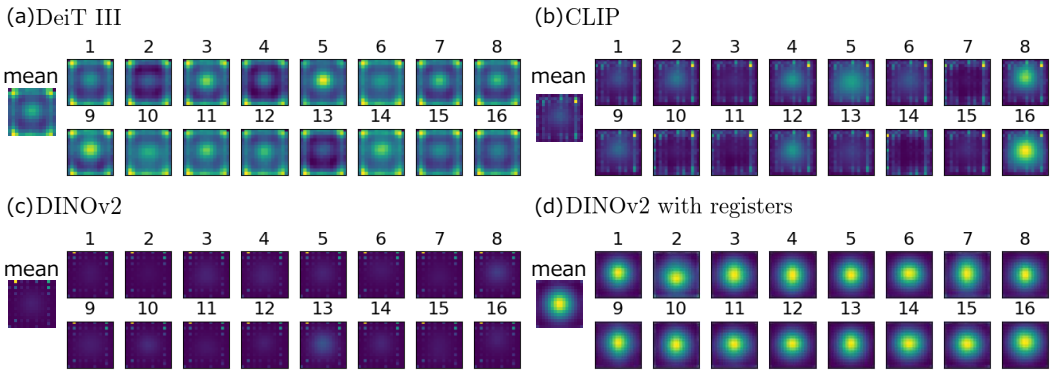

Figure 3: Comparison of the average of attention maps on ImageNet-1k across four training methods using pretrained models (a)-(d).

et al., 2024) and DINOv2 with registers (Darcet et al., 2024), utilizing publicly available pretrained models. DeiT III is a label-supervised training method for ViTs. CLIP is a text-supervised training method using contrastive learning on image-text pairs. DINOv2 with and without registers are self-supervised training methods to learn robust visual representations. DINOv2 with registers extends DINOv2 by introducing additional register tokens to get rid of high-norm tokens in attention maps, which is called artifacts (Darcet et al., 2024). More details about the pretrained models are described in Appendix C.1. Figure 3 shows the average of the attention maps and the head mean of them on the validation set of ImageNet-1k (Russakovsky et al., 2015). Artifacts are observed in the maps produced by DeiT III, CLIP, and DINOv2, which is consistent with Darcet et al. (2024), whereas the maps generated by DINOv2 with registers are clearer. Based on this result and extended analysis presented in Appendix D.1, we adopt DINOv2 with registers as the baseline training method.

4.2 Training

We train ViT models with registers (Darcet et al., 2024) using DINOv2 (Oquab et al., 2024) as the baseline training method to study the effects of incorporating our proposed loss functions on the model's performance. We mainly follow the learning process of DINOv2, but we replace the original DINOv2 loss $\mathcal{L}_{\text{DINOv2}} = \mathcal{L}_{\text{DINO}} + \mathcal{L}_{\text{iBOT}} + 0.1 \cdot \mathcal{L}_{\text{koleo}}$ with the following loss function:

$$\begin{aligned}\mathcal{L}_{\text{focus}} = (1 - w_{\text{in}} - w_{\text{out}}) \cdot \mathcal{L}_{\text{DINO}} + w_{\text{in}} \cdot \mathcal{L}_{\text{in}} + w_{\text{out}} \cdot \mathcal{L}_{\text{out}} \\ + \mathcal{L}_{\text{iBOT}} + 0.1 \cdot \mathcal{L}_{\text{koleo}},\end{aligned} \tag{4}$$

where w_{in} and w_{out} are the weights of the visual focusing loss functions $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$ in Eqs. (2) and (3), respectively. We note that the DINO loss $\mathcal{L}_{\text{DINO}}$ is a cross-entropy loss of the same type as $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$; therefore, its magnitude is comparable to that of $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$. $\mathcal{L}_{\text{iBOT}}$ is the iBOT loss (Zhou et al., 2022), $\mathcal{L}_{\text{koleo}}$ is the KoLeo loss (Sablayrolles et al., 2019), and for $w_{\text{in}} = 0$ and $w_{\text{out}} = 0$, Eq. (4) reduces to the DINOv2 loss. For each training image x, we designate one box b_x, which is specified as a random local region within the image, analogous to the local crops used in DINO (Caron et al., 2021). We share the box b_x for the computation of both $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$. Notably, because the boxes can be designated randomly during training, our approach requires only the images as training data and does not necessitate any ground truth data, such as the positional information of objects within the images. We train the ViT models on the train split of ImageNet-1k dataset (Russakovsky et al., 2015) without labels and use the teacher model in all evaluation experiments. Except for the continual learning discussed in Appendix D.2, the models are trained from scratch. Unless otherwise stated, we use ViT-L/16 with registers (ViT-L/16-Reg) (Darcet et al., 2024) and compare models trained under four different loss weight settings: (1) DINOv2 ($w_{\text{in}} = 0, w_{\text{out}} = 0$), (2) DINOv2 with $\mathcal{L}_{\text{in}}$ ($w_{\text{in}} = 0.1, w_{\text{out}} = 0$), (3) DINOv2 with $\mathcal{L}_{\text{out}}$ ($w_{\text{in}} = 0, w_{\text{out}} = 0.1$), and (4) DINOv2 with $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$ ($w_{\text{in}} = 0.05, w_{\text{out}} = 0.05$). More details about the training conditions are described in Appendix C.1.

Table 1: Zero-shot image segmentation performance. The DINOv2 model that integrates both $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$ outperforms the original DINOv2 model on both ImageNet-Segmentation and PASCAL VOC. (Model: ViT-L/16-Reg)

	ImageNet-Seg.			PASCAL VOC		
Loss	mIoU (%)↑	Acc (%)↑	mAP (%)↑	mIoU (%)↑	Acc (%)↑	mAP (%)↑
DINOv2	63.67	80.69	85.39	58.10	75.28	84.79
w/ $\mathcal{L}_{\text{in}}$	63.93	80.61	86.03	57.82	74.76	85.00
w/ $\mathcal{L}_{\text{out}}$	63.65	80.94	84.69	58.23	75.73	84.97
w/ $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$	**67.29**	**83.20**	**86.43**	**60.68**	**77.36**	**85.27**

4.3 Zero-shot image segmentation

To validate the ability of ViT models to focus on various regions within images, we evaluate their zero-shot image segmentation performance. We follow a standard protocol (Chefer et al., 2021) for evaluating the effectiveness of their heat maps. For each heat map, we compute the head mean of the attention maps, where we use the `[CLS]` tokens from each head in the last attention layer as the attention maps. We then binarize the heat map into a foreground/background segmentation map, where the patch mean of the heat map is used for the threshold value. We utilize two datasets, ImageNet-Segmentation (Guillaumin et al., 2014) and PASCAL VOC (Everingham et al., 2010), for the evaluation following Chefer et al. (2021) and Helbling et al. (2025). ImageNet-Segmentation consists of 4,276 single-label images from the validation set of ImageNet-1k (Russakovsky et al., 2015) with annotated segmentations from 445 categories. For the evaluation on PASCAL VOC, we utilize PASCAL VOC 2012 benchmark (Everingham et al., 2015) and evaluate on a multi-class split (1,449 images). In our multi-class split experiment, all annotated objects are considered as a single foreground.

Quantitative evaluation results. To investigate the impact of changes to the loss function, we quantitatively evaluate zero-shot image segmentation performance using three standard evaluation metrics: mean Intersection over Union (mIoU), pixel-wise accuracy (Acc), and mean Average Precision (mAP). Table 1 provides a comparison of zero-shot image segmentation performance for models trained using four different loss functions. The results indicate that the performance of the DINOv2 models with either $\mathcal{L}_{\text{in}}$ alone or $\mathcal{L}_{\text{out}}$ alone is comparable to that of the original DINOv2 model. In contrast, the model with both $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$ outperforms all the other models across all the evaluation metrics.

4.4 Qualitative analysis and emergent abilities

Dataset-scale attention heads visualization. To investigate which regions of the images the models focus on, we visualize the average of the attention maps across the entire set of evaluation dataset images. Figure 4 shows the average of the attention maps and the head mean of them used as the heat map on the validation set of ImageNet-1k. As a result, we find that the visual loss functions qualitatively change the characteristics of attention maps.

Emergent focusing ability: Spatial Attention Head Specialization. As shown in Fig. 4(a), the attention maps of each head in the original DINOv2 model mainly concentrate their outputs around the center. In contrast, as illustrated in Fig. 4(b), the model with $\mathcal{L}_{\text{in}}$ shows substantial variety in the output locations across different heads. This visualizes the emergence of the ability for each head to focus on distinct spatial positions, and their information is integrated to produce the comprehensive heat map. Notably, the focus area of the heat map in this model is broader compared to that of the original model. In the model with $\mathcal{L}_{\text{out}}$, in addition to heads that primarily focus on the central region as seen in the original model, there also emerge heads that tend to avoid the center and focus on other regions. This behavior is considered to reflect the design of $\mathcal{L}_{\text{out}}$, which encourages attention to objects other than the most discriminative objects. Additionally, a large activation is observed in a single patch at the top-left corner, which is considered to be a type of artifact, as discussed in Darcet et al. (2024). As described above, by using $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$, various attention behaviors emerge across different heads. In the model with both $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$, we observe the coexistence

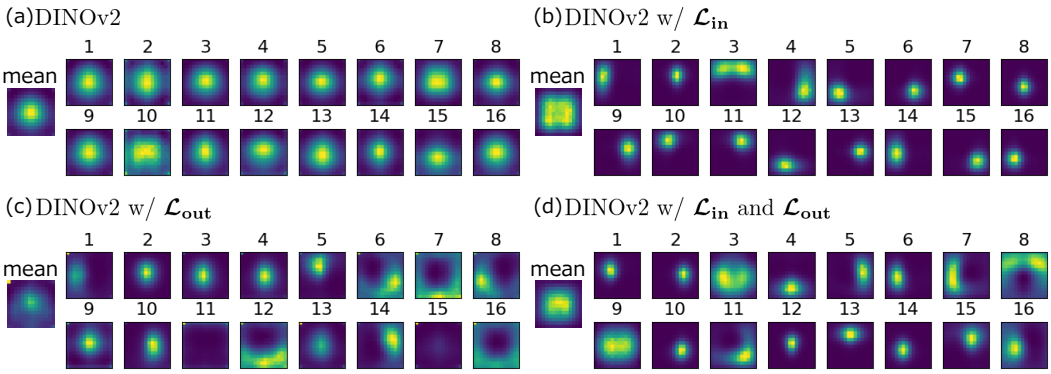

Figure 4: Comparison of the average of attention maps on ImageNet-1k from head1 to head16 and their head mean. Each model is trained using one of four loss functions (a)-(d).

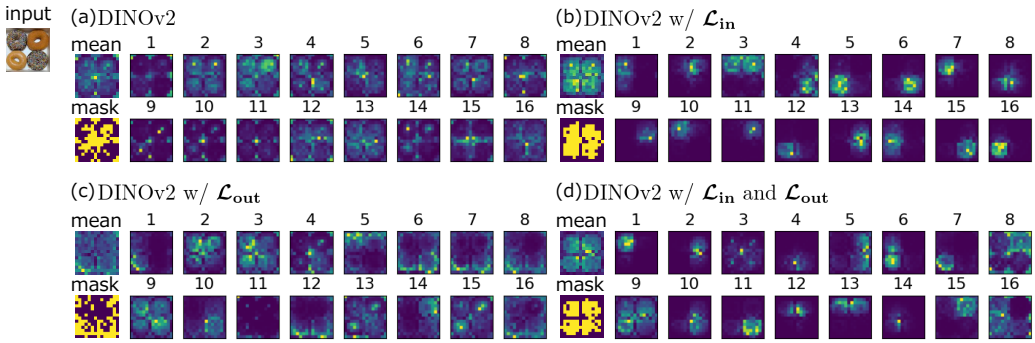

Figure 5: An example of attention maps for each model (a)-(d). The image in the top left is the "input". The images labeled "mean" represent the head means of the attention maps used as heat maps, and the images labeled "mask" represent the segmentation masks.

of the attention maps observed in the model with $\mathcal{L}_{\text{in}}$ alone or $\mathcal{L}_{\text{out}}$ alone, in addition to attention maps that focus on broader regions. We note that the artifacts seen in the model with $\mathcal{L}_{\text{out}}$ alone are absent. While standard DINOv2 already exhibits content-based head specialization, with individual heads capturing different semantic features within an image, our method retains such content-based specialization while introducing an additional, complementary form of specialization across spatial regions. We refer to this phenomenon as the emergence of spatial attention head specialization.

4.5 Empirical analysis of segmentation capability

To understand how segmentation works on individual images, we analyze examples of attention maps for a specific input image, as shown in Fig. 5. In the case of the original DINOv2 model, we observe that all heads focus at least partially on the box behind the donuts, which leads to ambiguous segmentation. We speculate that the original model pays close attention to the box because part of the box is visible in the center of the input image. When $\mathcal{L}_{\text{in}}$ is introduced, we observe that the model focus on object regions corresponding to the parts discussed in the attention map averages in Fig. 4. For instance, head1, 7, 11, and 15 in Fig. 5(d) are able to focus primarily on each donut region in the image. Furthermore, when $\mathcal{L}_{\text{out}}$ is introduced, as seen in head8 of Fig. 5(d), there emerge attention heads that focus on the box regions while avoiding the donuts area. This corresponds to the heads focusing on non-central regions on average, as discussed in Sec. 4.4.

4.6 Segmentation performance by position

As shown in Tab. 1, we find that the model with $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$ exhibits higher segmentation performance compared to the other models evaluated. To investigate the positional dependence of

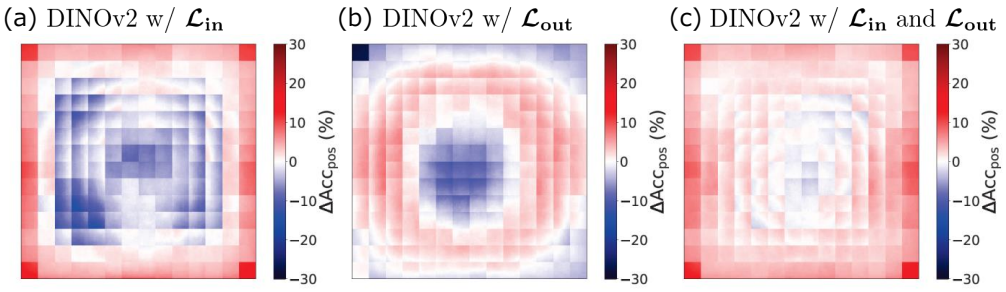

Figure 6: Changes in segmentation accuracy at each position compared to the accuracy of the original DINOv2 model.

Table 2: Unsupervised object discovery performance. (Model: ViT-L/16-Reg)

Loss	VOC07	VOC12	COCO
DINOv2	58.6	63.3	45.4
w/ $\mathcal{L}_{\text{in}}$	56.0	59.5	**45.7**
w/ $\mathcal{L}_{\text{out}}$	42.0	47.3	35.1
w/ $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$	**61.6**	**64.6**	45.5

Table 3: ImageNet-1k classification performance. (Model: ViT-L/16-Reg)

Loss	k-NN (%)	Linear (%)
DINOv2	**81.7**	**82.6**
w/ $\mathcal{L}_{\text{in}}$	79.1	81.4
w/ $\mathcal{L}_{\text{out}}$	79.7	81.7
w/ $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$	80.3	81.8

performance among the models, we calculate the accuracy at each position for each model and examine the changes in accuracy from the original model, referred to as $\Delta\text{Acc}_{\text{pos}}$ as shown in Fig. 6. The figure shows that $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$ play complementary roles in improving segmentation accuracy. First, as shown in Fig. 6(a), the model with $\mathcal{L}_{\text{in}}$ alone decreases accuracy in the central region while improving accuracy in the peripheral regions. This result suggests that whereas the original DINOv2 model tends to focus on objects near the center, $\mathcal{L}_{\text{in}}$ encourages attention to various regions, which enhances segmentation performance in less salient peripheral areas while somewhat reducing focus on the center. In contrast, Fig. 6(b) shows that the model with $\mathcal{L}_{\text{out}}$ alone exhibits improved accuracy in a ring-shaped region surrounding the center. This result can be interpreted as indicating that $\mathcal{L}_{\text{out}}$ primarily enhances segmentation performance in regions surrounding central objects, which are typically located in the center. Although each of these two models shows improved segmentation performance in certain regions, the decline in accuracy around the central area results in overall performance comparable to the original DINOv2 model. Finally, Fig. 6(c) shows that the model with both $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$ successfully integrates the advantages of both loss functions. Furthermore, the accuracy drop observed in the central region in the models (a) and (b) is substantially mitigated in the model (c), resulting in an improvement in Acc.

4.7 Unsupervised object discovery

The ability to focus on a wide range of areas within images is also considered effective for object discovery. Therefore, we evaluate the unsupervised object discovery performance of our trained models. For the evaluation, we use LOST (Siméoni et al., 2021), which infers bounding boxes for objects from the patch tokens of features in the attention layers. Following Darcet et al. (2024) and Jiang et al. (2025), we present the best results of Correct Localization (CorLoc) across the key, query, or value features over the last four attention layers, and details of the evaluation setup are described in Appendix C.2. We utilize three datasets, VOC07 trainval, VOC12 trainval (Everingham et al., 2010), and COCO 20k (Lin et al., 2014; Vo et al., 2020). Table 2 presents the performance of four trained models on the unsupervised object discovery task. We observe that using $\mathcal{L}_{\text{in}}$ or $\mathcal{L}_{\text{out}}$ alone results in lower performance than the original DINOv2 model, except in the case of $\mathcal{L}_{\text{in}}$ on COCO. However, combining $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$ outperforms the original model in all the three datasets. This improvement through combination is consistent with the results in the zero-shot image segmentation task.

Table 4: Focus-Oriented Image Retrieval performance. (Model: ViT-L/16-Reg)

Loss	Retrieval scheme	COCO MP@100 (%) ↑	PASCAL VOC MP@100 (%) ↑
DINOv2	Fixed per image	54.2	72.8
DINOv2	Visual prompt per object	55.2	74.2
w/ $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$	Fixed per image	52.7	72.4
w/ $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$	Visual prompt per object	**55.9**	**75.8**

4.8 Performance on other tasks

In addition to zero-shot image segmentation and unsupervised object discovery, we conduct additional experiments on two complementary tasks: a classification task and an image retrieval task.

Classification performance. We evaluate our models on the ImageNet-1k classification task with k-NN ($k = 20$) and linear probing. Table 3 presents the top-1 accuracy on the validation set of ImageNet-1k. For k-NN classification, we observe that the accuracy drops by approximately $1-3\%$ when visual focusing loss functions are introduced compared to the original DINOv2 model. Our possible explanation for the performance drop is as follows. The classification task is formulated as a single-label problem, where images typically contain a dominant object near the center. This task setting aligns with local-to-global correspondence for single objects in DINO (Caron et al., 2021). In contrast, our proposed loss functions are designed to capture a wide range of regions and focus on various locations, and may therefore be less specialized for this setting. For linear probing classification, the accuracy difference between our visual focusing models and the original DINOv2 model is smaller than that observed with k-NN classification. This result suggests that training a linear classifier for classification purposes enables effective utilization of the frozen features that have been pretrained to focus.

Image retrieval performance. We also evaluate our models using an image retrieval task called focus-oriented image retrieval (FOIR) (Nozawa et al., 2025). FOIR task is a multi-object image retrieval task that evaluates whether models can retrieve appropriate images for each specified object in query images, where the query images may contain multiple objects. We utilize COCO (Lin et al., 2014) and PASCAL VOC (Everingham et al., 2010) datasets for evaluation. We use Mean Precision at k (MP@k) as an evaluation metric to measure the relevance of the top-k retrieved results. Table 4 presents MP@100 for each model and dataset. We evaluate models using two retrieval schemes. The first scheme, referred to as "Fixed per image", utilizes a single retrieval result obtained by extracting one feature vector from each query image for image retrieval. The second scheme, "Visual prompt per object", applies PHS during inference and specifies a bounding box for each object as a visual prompt, thereby obtaining retrieval results that focus on individual objects. First, the consistent performance gain of the visual prompt per object scheme over the fixed per image one across the models indicates that PHS is effective during inference. In a comparison between the models, under the fixed per image scheme, the original DINOv2 model outperforms the model with $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$; however, under the visual prompt per object scheme, the trend reverses, with the model with $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$ outperforming the original DINOv2 model. This result suggests that the feature representations of DINOv2 effectively capture global image characteristics, resulting in averaged retrieval performance over objects, while the model with $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$, consistent with its training objective, exhibits its strengths when focusing on individual objects.

4.9 Additional results on model size variations

To evaluate the robustness of our method, we further examine its effectiveness across different model sizes. Table 5 presents the zero-shot image segmentation performance of the original DINOv2 model and the model with $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$ for ViT-B/16-Reg and ViT-g/16-Reg, evaluated in the same manner as for ViT-L/16-Reg. These results demonstrate that our method consistently improves performance across all evaluated model sizes, confirming its robustness.

Table 5: Zero-shot image segmentation performance. (Model: ViT-{Size}/16-Reg)

Size	Loss	ImageNet-Seg.			PASCAL VOC		
		mIoU (%)↑	Acc (%)↑	mAP (%)↑	mIoU (%)↑	Acc (%)↑	mAP (%)↑
B	DINOv2	68.32	83.81	87.63	62.20	78.14	86.46
	w/ $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$	**69.33**	**84.40**	**87.84**	**62.98**	**78.65**	**86.77**
g	DINOv2	63.37	80.53	85.20	58.34	75.55	84.85
	w/ $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$	**67.61**	**83.31**	**86.86**	**61.31**	**77.65**	**86.09**

5 Discussion

Local-to-global and region-to-focus correspondence. Our results highlight a distinction between conventional local-to-global correspondence (Caron et al., 2021) and region-to-focus correspondence introduced by our visual focusing losses. The main DINOv2 objective is based on local-to-global correspondence and is well suited to images containing a large dominant object, as it encourages correspondence between the global representation of the object and local representations of its parts. In contrast, our method is designed for images containing multiple objects or semantic regions, a common setting in natural images. To address this setting, in addition to establishing local-to-global correspondence, our proposed loss functions encourage correspondence between regional representations obtained through masking or cropping and selective attention toward the corresponding regions. As observed in our experiments, this encourages the emergence of spatial attention head specialization, with performance gains in zero-shot image segmentation, unsupervised object discovery, and multi-object image retrieval.

Practical implications. These properties may also have practical implications for interpretability and customizability. Distinct and spatially specialized attention heads can provide a more structured view of the visual information captured by the model, potentially facilitating the analysis of its behavior and failure cases. Moreover, such specialization may offer a straightforward mechanism for feature manipulation during inference. For example, the output features could be modified by dynamically disabling the signals from individual attention heads during inference, thereby potentially providing a mechanism for on-the-fly model customization. These possibilities could be useful in applications that require selective or dynamically varying visual focus, such as robotic vision and autonomous driving. However, we do not evaluate such applications in this work, and their effectiveness remains an important direction for future investigation.

6 Conclusion

Our work addresses a key limitation of self-supervised ViT models by introducing visual focusing loss functions that establish correspondence between the model's attention and specific regions within images. We find that diverse and spatially focused perceptual capabilities emerge in our trained model, enabling the model to leverage different attention heads to focus on distinct and meaningful image regions. This targeted focusing mechanism effectively mitigates the ambiguity and redundancy often observed in learned representations, thereby improving the interpretability of the model's attention. Through comprehensive experiments, we demonstrate that our approach improves performance in zero-shot image segmentation, unsupervised object discovery, and multi-object image retrieval, while also producing clearer and more informative attention maps that support better visual understanding. Overall, these results highlight the potential of introducing focusing mechanisms into self-supervised representation learning without requiring annotations, providing a new direction for learning diverse and spatially focused visual information.

References

Samira Abnar and Willem Zuidema. Quantifying Attention Flow in Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, 2020.

Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring Visual Prompts for Adapting Large-Scale Models. *arXiv preprint arXiv:2203.17274*, 2022.

Lorenzo Basile, Valentino Maiorca, Diego Doimo, Francesco Locatello, and Alberto Cazzaniga. Head Pursuit: Probing Attention Specialization in Multimodal Transformers. In *NeurIPS*, 2025.

Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding Robustness of Transformers for Image Classification. In *ICCV*, pages 10231–10241, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *NeurIPS*, pages 1877–1901, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *ICCV*, 2021.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer Interpretability Beyond Attention Visualization. In *CVPR*, pages 782–791, 2021.

Xiangyu Chen, Qinghao Hu, Kaidong Li, Cuncong Zhong, and Guanghui Wang. Accumulated Trivial Attention Matters in Vision Transformers on Small Datasets. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3984–3992, 2023.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, pages 2818–2829, 2023.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *CVPR*, pages 3213–3223, 2016.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision Transformers Need Registers. In *ICLR*, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *IJCV*, 88(2):303–338, 2010.

Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes Challenge: A Retrospective. *IJCV*, 111(1):98–136, 2015.

Hongyu Gong, Yun Tang, Juan Pino, and Xian Li. Pay Better Attention to Attention: Head Selection in Multilingual and Multi-Domain Sequence Modeling. In *NeurIPS*, 2021.

Matthieu Guillaumin, Daniel Küttel, and Vittorio Ferrari. ImageNet Auto-Annotation with Segmentation Propagation. *IJCV*, 110(3):328–348, 2014.

Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *ICCV*, pages 991–998, 2011.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.

Alec Helbling, Tuna Han Salih Meral, Benjamin Hoover, Pinar Yanardag, and Duen Horng Chau. ConceptAttention: Diffusion Transformers Learn Highly Interpretable Features. In *International Conference on Machine Learning*, 2025.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual Prompt Tuning. In *ECCV*, pages 709–727, 2022.

Nicholas Jiang, Amil Dravid, Alexei A Efros, and Yossi Gandelsman. Vision Transformers Don't Need Trained Registers. In *NeurIPS*, 2025.

Sehyeong Jo, Gangjae Jang, and Haesol Park. GMAR: Gradient-Driven Multi-Head Attention Rollout for Vision Transformer Interpretability. In *ICIP*, pages 582–587, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment Anything. In *ICCV*, pages 4015–4026, 2023.

Krishna Kumar Singh and Yong Jae Lee. Hide-And-Seek: Forcing a Network to Be Meticulous for Weakly-Supervised Object and Action Localization. In *ICCV*, 2017.

Seung Hoon Lee, Seunghyun Lee, and Byung Cheol Song. Vision Transformer for Small-Size Datasets. *arXiv preprint arXiv:2112.13492*, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014.

Lingchen Meng, Hengduo Li, Bor-Chun Chen, Shiyi Lan, Zuxuan Wu, Yu-Gang Jiang, and Ser-Nam Lim. AdaViT: Adaptive Vision Transformers for Efficient Image Recognition. In *CVPR*, 2022.

Kevin Miao, Akash Gokul, Raghav Singh, Suzanne Petryk, Joseph Gonzalez, Kurt Keutzer, Trevor Darrell, and Colorado Reed. Prior Knowledge-Guided Attention in Self-Supervised Vision Transformers. *arXiv preprint arXiv:2209.03745*, 2022.

Kang Ni, Duo Wang, Zhizhong Zheng, and Peng Wang. MHST: Multiscale Head Selection Transformer for Hyperspectral and LiDAR Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17, 2024.

Armand Mihai Nicolicioiu, Andrei Liviu Nicolicioiu, Bogdan Alexe, and Damien Teney. Learning Diverse Features in Vision Transformers for Improved Generalization. In *ICML Workshop on Spurious Correlations, Invariance and Stability*, 2023.

Yuji Nozawa, Yu-Chieh Lin, Kazumoto Nakamura, and Youyang Ng. Prompt-Guided Attention Head Selection for Focus-Oriented Image Retrieval. In *CVPR Workshops*, pages 4170–4180, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, 2024.

Sungho Park and Hyeran Byun. Fair-VPT: Fair Visual Prompt Tuning for Image Classification. In *CVPR*, pages 12268–12278, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

Yuchen Ren, Zhengyu Zhao, Chenhao Lin, Bo Yang, Lu Zhou, Zhe Liu, and Chao Shen. Improving Adversarial Transferability on Vision Transformers via Forward Propagation Refinement. In *CVPR*, pages 25071–25080, 2025.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Spreading vectors for similarity search. In *ICLR*, 2019.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In *ICCV*, 2017.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does CLIP know about a red circle? Visual prompt engineering for VLMs. In *ICCV*, pages 11987–11997, 2023.

Oriane Siméoni, Gilles Puy, Huy V. Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing Objects with Self-Supervised Transformers and no Labels. In *BMVC*, 2021.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michael Ramamonjisoa, Francisco Massa, Daniel HAZIZA, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Herve Jegou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026.

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556*, 2015.

Thomas Stegmüller, Tim Lebailly, Nikola Đukić, Behzad Bozorgtabar, Tinne Tuytelaars, and Jean-Philippe Thiran. A Simple Framework for Open-Vocabulary Zero-Shot Segmentation. In *ICLR*, 2025.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. DeiT III: Revenge of the ViT. In *ECCV*, 2022.

Huy V. Vo, Patrick Pérez, and Jean Ponce. Toward unsupervised, multi-object discovery in large-scale image collections. In *ECCV*, 2020.

Haiyan Wu, Yuting Gao, Yinqi Zhang, Shaohui Lin, Yuan Xie, Xing Sun, and Ke Li. Self-supervised Models are Good Teaching Assistants for Vision Transformers. In *PMLR*, 2022.

Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. Fine-Grained Visual Prompting. In *NeurIPS*, pages 24993–25006, 2023.

Seungryong Yoo, Eunji Kim, Dahuin Jung, Jungbeom Lee, and Sungroh Yoon. Improving Visual Prompt Tuning for Self-supervised Vision Transformers. In *ICML*, pages 40075–40092. PMLR, 2023.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene Parsing Through ADE20K Dataset. In *CVPR*, pages 5122–5130, 2017.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT Pre-Training with Online Tokenizer. *ICLR*, 2022.

Jiedong Zhuang, Jiaqi Hu, Lianrui Mu, Rui Hu, Xiaoyu Liang, Jiangnan Ye, and Haoji Hu. FALIP: Visual Prompt as Foveal Attention Boosts CLIP Zero-Shot Performance. In *ECCV*, 2024.

A Limitations

While our method can train the model to focus on diverse regions within an image, it does not allow explicit control over where each attention head focuses, as the regions of focus are determined implicitly during training. As a result, it is not possible to know in advance which regions each attention head will focus on until after training. Our method also increases the training time and introduces a trade-off between classification performance and performance improvements in zero-shot segmentation, object discovery, and FOIR tasks. Exploring other effective ways to utilize the visual focusing image representations obtained by our method remains an important direction for future work.

B Societal impacts

In this study, we propose a self-supervised training method, and our trained models demonstrate high-performance segmentation across a wide range of areas within images. Our method can identify both salient and subtle objects in images, and it has potential applications such as medical imaging, autonomous driving, and defect detection. However, incorrect segmentation of subtle structures may lead to erroneous decisions, posing risks especially in scenarios where safety is critical. The ability to extract subtle visible information also has a risk of misuse, such as applying the method to privacy-invasive surveillance. To mitigate these concerns, we emphasize clear communication of limitations and intended use.

C Details of training and evaluation setups

C.1 Training

Most of the training hyperparameters are identical to those used in Oquab et al. (2024) for training ViT-L/16 on ImageNet-1k. Here, we describe the differences from that setting. Due to computational resource constraints, we train the models using 4 NVIDIA H200 GPUs (141GB of memory), instead of 32 NVIDIA A100 GPUs (80GB of memory). Correspondingly, the batch size per GPU is set to 128 for the base and large models, and 64 for the giant models. In addition, for the models with registers, the number of registers is set to 4.

Data augmentation. For each training image, following DINOv2, we apply multi-crop data augmentation, where global and local crops are generated using random resized cropping, together with standard augmentations such as horizontal flipping, color jittering, Gaussian blur, and solarization. The global crops are resized to 224×224 with $\mathrm{scale} = (0.32, 1.0)$ and $\mathrm{ratio} = (3.0/4.0, 4.0/3.0)$, while the local crops are resized to 96×96 with $\mathrm{scale} = (0.05, 0.32)$ and the same aspect-ratio range. When our loss functions $\mathcal{L}_{\mathrm{in}}$ or $\mathcal{L}_{\mathrm{out}}$ are introduced, a box b_x for the PHS method is obtained as a region sampled via random resized cropping from the image, using the same setting as the local crops. We set the number of selected heads h_{on} during training to 5 in our main experiments as discussed in Appendix D.4. For $\mathcal{L}_{\mathrm{out}}$, we also introduce randomness by selecting either the top or bottom edge of the visual-prompt box with equal probability for each batch and aligning the selected edge at the same height across all images in the batch.

Training from scratch. The number of training iterations is set to 500,000. We set the base learning rate, which is the learning rate for a batch size of 1024, to 0.004 for the base and large models, and 0.002 for the giant models for stable training. The training time for ViT-L/16-Reg is approximately 2 days and 22 hours for the DINOv2 model, 3 days and 21 hours for the model with $\mathcal{L}_{\mathrm{in}}$, 4 days and 11 hours for the model with $\mathcal{L}_{\mathrm{out}}$, and 5 days and 10 hours for the model with $\mathcal{L}_{\mathrm{in}}$ and $\mathcal{L}_{\mathrm{out}}$.

Pretrained models and continual learning. For our experiments about pretrained models in Sec. 4.1 and Appendix D.1, we compare DeiT III (Touvron et al., 2022), CLIP (Radford et al., 2021), DINOv2 (Oquab et al., 2024) and DINOv2 with registers (Darcet et al., 2024), utilizing publicly available pretrained models. For DeiT III, we use the ViT-L/16 pretrained on ImageNet-1k at a resolution of 224×224. For CLIP, we use the OpenCLIP implementation (Cherti et al., 2023) and employ the ViT-L/14 image encoder pretrained by OpenAI (Radford et al., 2021). For DINOv2 and

DINOv2 with registers, we use the ViT-L/14 and the ViT-L/14-Reg, respectively, both pretrained on the large-scale LVD-142M dataset. We discuss continual learning using pretrained models in Appendix D.2. To match the patch size of 14, we change the local crops size from 96 to 98. During the initial training stage of the DINO head only, the number of iterations is set to 50,000. In the subsequent stage, where the entire networks are trained, we also set the number of iterations to 50,000. Throughout training, the base learning rate is set to 0.0004, and the number of warmup epochs is set to 1. The training dataset is the train split of ImageNet-1k dataset (Russakovsky et al., 2015) as in training from scratch. The training time for the DINO head of ViT-L/14-Reg is approximately 11 hours, and the training time for the entire networks is approximately 9 hours for the DINOv2 model, 12 hours for the model with $\mathcal{L}_{\text{in}}$, 13 hours for the model with $\mathcal{L}_{\text{out}}$, 16 hours for the model with $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$.

C.2 Evaluation

In all our evaluation experiments except unsupervised object discovery, the input images are resized to 224×224, whereas the unsupervised object discovery task does not involve resizing.

For the unsupervised object discovery task, we utilize three datasets, VOC07 trainval (5,011 images), VOC12 trainval (11,540 images) (Everingham et al., 2010), and COCO 20k (19,817 images) (Lin et al., 2014; Vo et al., 2020). Following Darcet et al. (2024) and Jiang et al. (2025), we present the best results of CorLoc across the key, query, or value features over the last four attention layers, sweeping over a manual bias term to the Gram matrix of features. We normalize the features and sweep the bias term over the range $[-1, 0]$ with a step size of 0.01.

The evaluation procedure for k-NN and linear probing classification follows that of Oquab et al. (2024). For linear probing classification, we report the highest accuracy performing the following grid search of three parameters: learning rate in {0.0001, 0.0002, 0.0005, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.5}, number of output layers in {1, 4}, and concatenate average-pooled tokens in {yes, no}. The linear classifier is trained with SGD for 12,500 iterations.

For the FOIR task, the evaluation procedure follows that of Nozawa et al. (2025). When the retrieval scheme is Visual prompt per object, the number of selected heads at evaluation is set to 2. We use two datasets for this evaluation: COCO (Lin et al., 2014) and PASCAL VOC (Everingham et al., 2010). COCO dataset contains 80 object categories, and we use val2017 for query images (5,000 images) and train2017 for database images (118,287 images). PASCAL VOC dataset contains 20 object categories, and we use test2007 for query images (4,952 images) and trainval2007+2012 for database images (16,551 images).

D Extended analyses

D.1 Extended results on interpretability of attention maps for conventional training methods

In addition to our experiment in Sec. 4.1, we conduct extended analyses on the interpretability of attention maps obtained using four different training methods: DeiT III (Touvron et al., 2022), CLIP (Radford et al., 2021), DINOv2 (Oquab et al., 2024) and DINOv2 with registers (Darcet et al., 2024), utilizing publicly available pretrained models.

Table A1 presents a comparison of zero-shot image segmentation performance of each training method. This result shows that DINOv2 and DINOv2 with registers outperform the other methods. As shown in Fig. A1, the impact of these artifacts is substantial for DeiT III and CLIP, leading to low interpretability of the attention maps. In contrast, DINOv2 produces a single patch that appears to be an artifact, while the remaining patches capture meaningful image information, resulting in a meaningful mask. In DINOv2 with registers, the register tokens are effective in suppressing artifacts, and the object regions are captured more clearly.

D.2 Continual learning in existing models

In this subsection, we consider applying our training method to continual learning in existing models. For the pretrained model, we utilize the publicly available backbone ViT-L/14-Reg weights

Table A1: Zero-shot image segmentation performance comparison of four training methods (a)-(d).

Method	ImageNet-Seg.			PASCAL VOC		
	mIoU (%)↑	Acc (%)↑	mAP (%)↑	mIoU (%)↑	Acc (%)↑	mAP (%)↑
DeiT III (Touvron et al., 2022)	38.21	59.73	74.36	37.98	58.43	77.54
CLIP (Radford et al., 2021)	47.06	72.06	76.41	45.37	69.59	77.66
DINOv2 w/o reg. (Oquab et al., 2024)	52.35	76.14	82.29	52.75	74.40	83.88
DINOv2 w/ reg. (Darcet et al., 2024)	53.71	73.95	81.61	55.50	74.53	84.03

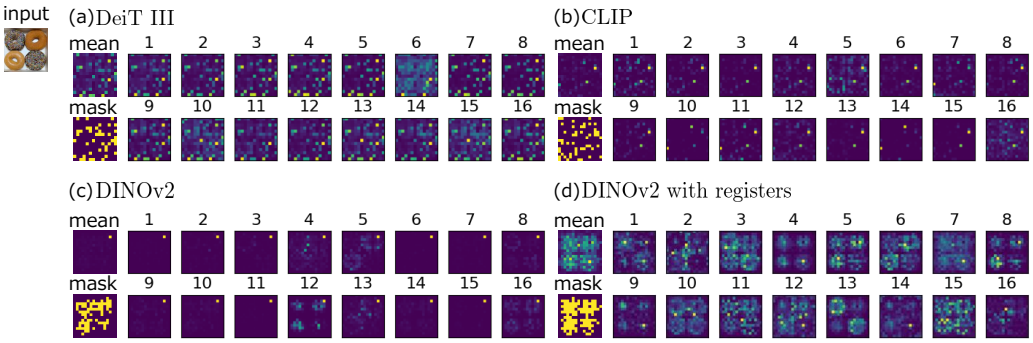

Figure A1: Examples of attention maps for four training methods (a)-(d).

pretrained by DINOv2 with registers (Darcet et al., 2024), which we evaluate in Sec. 4.1 and Appendix D.1. Since pretrained weights of the DINO head is not available, we first train the DINO head using DINOv2 with registers from scratch, while keeping the pretrained backbone frozen. After training the DINO head, we then perform training the entire networks, using each loss function with fewer iterations than training from scratch. We use ImageNet-1k as the training dataset, as in the main experiments. Details of the training setup are described in Appendix C.1.

Table A2 presents a comparison of zero-shot image segmentation performance between the pretrained model and models further trained using four different loss functions. On ImageNet-Segmentation dataset, the performance improves with all loss functions used for continual learning, with the DINOv2 model with $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$ achieving the highest performance. On PASCAL VOC dataset, the DINOv2 loss, DINOv2 loss with $\mathcal{L}_{\text{in}}$, and DINOv2 loss with $\mathcal{L}_{\text{out}}$ all show performance degradation across all metrics. The discrepancy between the results on ImageNet-Segmentation and PASCAL VOC can be attributed to the domain gap between these datasets. While further training on ImageNet-1k improves performance on the in-domain dataset ImageNet-Segmentation, it results in degradation in the out-of-domain dataset PASCAL VOC. Notably, the DINOv2 loss with both $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$ improves performance on two out of three metrics, mIoU and mAP. This indicates that combination of $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$ exhibits stronger robustness to domain shifts and better generalization ability.

For the unsupervised object discovery task, as shown in Tab. A3, we find that the models after continual learning, except for the model with $\mathcal{L}_{\text{out}}$ alone, show improved performance compared to the pretrained model. We also find that the model with $\mathcal{L}_{\text{in}}$ alone has the highest performance in this task, while the original model and the model with $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$ perform at a similar level.

D.3 Loss weights analysis

We investigate the effects of loss weights $(w_{\text{in}}, w_{\text{out}})$ in Eq. (4) that balance the DINO loss $\mathcal{L}_{\text{DINO}}$ and the two visual focusing loss functions $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$ on segmentation performance. In this ex-

Table A2: Zero-shot image segmentation performance in continual learning setting. (Model: ViT-L/14-Reg)

Loss	ImageNet-Seg.			PASCAL VOC		
	mIoU (%)↑	Acc (%)↑	mAP (%)↑	mIoU (%)↑	Acc (%)↑	mAP (%)↑
Pretrained	53.71	73.95	81.61	55.50	**74.53**	84.03
DINOv2	54.90	74.76	82.15	53.07	72.14	83.23
w/ $\mathcal{L}_{\text{in}}$	57.26	75.64	83.87	52.37	70.68	82.70
w/ $\mathcal{L}_{\text{out}}$	60.90	78.79	84.71	55.04	73.32	84.24
w/ $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$	**61.63**	**78.86**	**85.40**	**55.70**	73.30	**84.73**

Table A3: Unsupervised object discovery performance in continual learning setting. (Model: ViT-L/14-Reg)

Loss	VOC07	VOC12	COCO
Pretrained	56.2	60.2	43.0
DINOv2	57.0	61.4	44.6
w/ $\mathcal{L}_{\text{in}}$	**58.1**	**62.9**	**45.1**
w/ $\mathcal{L}_{\text{out}}$	53.5	59.1	42.5
w/ $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$	57.0	61.6	44.6

periment, we train ViT-L/16 with $\mathcal{L}_{\text{in}}$, one with $\mathcal{L}_{\text{out}}$, and one with $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$ with weights set to 0.05 and 0.1, respectively. The results in Tab. A4 show that introducing each weight of 0.05 leads to a substantial improvement in performance compared to the original DINOv2 model, whereas increasing the weights to 0.1 slightly degrades performance compared to 0.05. These findings indicate that simply increasing the weights of $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$ is not necessarily beneficial; rather, balancing between the visual focusing loss and the DINO loss is crucial for achieving optimal performance. We conduct these analyses on training the ViT model without registers and find that the resulting loss weights also perform well with the ViT model incorporating registers. Therefore, we apply these weights in the experiments presented in the main results of our paper.

D.4 Analysis of the number of selected heads during training

The number of selected heads h_{on} during training is a hyperparameter of our training method. The number h_{on} refers to the number of heads selected during head selection in PHS, and it affects training through changes in the output of the student model $P_{\text{s}}^{\text{PHS}}$. To investigate the effect of h_{on} on performance, we train ViT-L/16 by varying h_{on} across three settings: $\{2, 5, 8\}$. Tables A5 and A6 show the effect of h_{on} on the performance for the zero-shot image segmentation task and the FOIR task, respectively. For $h_{\text{on}} = 2$, the models achieve high performance on the segmentation task, whereas their performance on the FOIR task is relatively low. In the case of $h_{\text{on}} = 8$, while the model with $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$ exhibits high performance on the FOIR task, the performance is relatively

Table A4: Zero-shot image segmentation performance across different loss weights. (Model: ViT-L/16)

Loss weights	ImageNet-Seg.			PASCAL VOC		
$(w_{\text{in}}, w_{\text{out}})$	mIoU (%)↑	Acc (%)↑	mAP (%)↑	mIoU (%)↑	Acc (%)↑	mAP (%)↑
$(0, 0)$	51.94	72.04	81.23	49.66	68.81	82.19
$(0.05, 0)$	60.97	78.71	84.76	56.17	73.76	83.90
$(0.1, 0)$	60.79	78.46	83.73	56.04	73.54	83.02
$(0, 0.05)$	64.10	81.26	85.87	60.39	77.35	85.91
$(0, 0.1)$	62.94	80.55	85.73	58.87	76.39	85.35
$(0.05, 0.05)$	65.77	82.21	86.29	59.57	76.53	85.35
$(0.1, 0.1)$	63.65	80.67	85.48	57.97	75.28	84.39

Table A5: Effect of h_{on} on zero-shot image segmentation performance. (Model: ViT-L/16)

		ImageNet-Seg.			PASCAL VOC		
h_{on}	Loss	mIoU (%)↑	Acc (%)↑	mAP (%)↑	mIoU (%)↑	Acc (%)↑	mAP (%)↑
–	DINOv2	51.94	72.04	81.23	49.66	68.81	82.19
2	w/ $\mathcal{L}_{\text{in}}$	64.43	81.12	85.83	58.53	75.51	84.91
5	w/ $\mathcal{L}_{\text{in}}$	60.79	78.46	83.73	56.04	73.54	83.02
8	w/ $\mathcal{L}_{\text{in}}$	58.06	76.50	82.59	53.94	72.01	82.23
2	w/ $\mathcal{L}_{\text{out}}$	62.66	80.35	85.51	56.55	74.72	84.51
5	w/ $\mathcal{L}_{\text{out}}$	62.94	80.55	85.73	58.87	76.39	85.35
8	w/ $\mathcal{L}_{\text{out}}$	57.01	77.45	81.74	55.48	74.74	84.16
2	w/ $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$	66.14	82.13	87.37	59.80	76.38	85.98
5	w/ $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$	65.77	82.21	86.29	59.57	76.53	85.35
8	w/ $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$	63.87	81.01	85.90	59.08	76.28	85.24

Table A6: Focus-Oriented Image Retrieval performance. (Model: ViT-L/16, Retrieval scheme: Visual prompt per object)

		COCO	PASCAL VOC
h_{on}	Loss	MP@100 (%) ↑	MP@100 (%) ↑
–	DINOv2	55.1	73.9
2	w/ $\mathcal{L}_{\text{in}}$	54.9	75.2
5	w/ $\mathcal{L}_{\text{in}}$	57.5	78.1
8	w/ $\mathcal{L}_{\text{in}}$	56.7	77.6
2	w/ $\mathcal{L}_{\text{out}}$	49.4	69.5
5	w/ $\mathcal{L}_{\text{out}}$	55.8	76.0
8	w/ $\mathcal{L}_{\text{out}}$	52.8	72.3
2	w/ $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$	54.2	74.0
5	w/ $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$	55.6	75.4
8	w/ $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$	56.8	77.9

low for other models and tasks. Since $h_{\text{on}} = 5$ yields consistently high performance overall, we adopt $h_{\text{on}} = 5$. We find that the resulting value $h_{\text{on}} = 5$ also performs well with the ViT model incorporating registers. Therefore, we apply $h_{\text{on}} = 5$ in the experiments presented in the main results of our paper.

D.5 Additional qualitative results of attention maps

Figure A2 shows attention maps on additional example input images, similar to those in Fig. 5. It is confirmed that the properties of attention maps discussed in Sec. 4.5 are consistent across different examples.

D.6 Entropy analysis of attention

To complement our qualitative visualizations of attention maps in Fig. 4, we additionally conduct a quantitative analysis of spatial attention head specialization. Specifically, we compute the Shannon entropy of the attention distribution underlying the attention maps. Table A7 shows the following two metrics: (1) Mean Per-Head Entropy: the entropy is first computed separately for each head and then averaged across heads, and (2) Entropy of Head-Averaged Attention: the attention distributions are first averaged across heads, after which the entropy is computed. These entropies are normalized, such that a uniform attention distribution has the maximum value of 1. We use models evaluated in Fig. 4 and report averages over the validation set of ImageNet-1k. Our $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$ model exhibits lower mean per-head entropy than the original DINOv2 model but high entropy after averaging attention across heads, indicating that individual heads produce spatially concentrated attention,

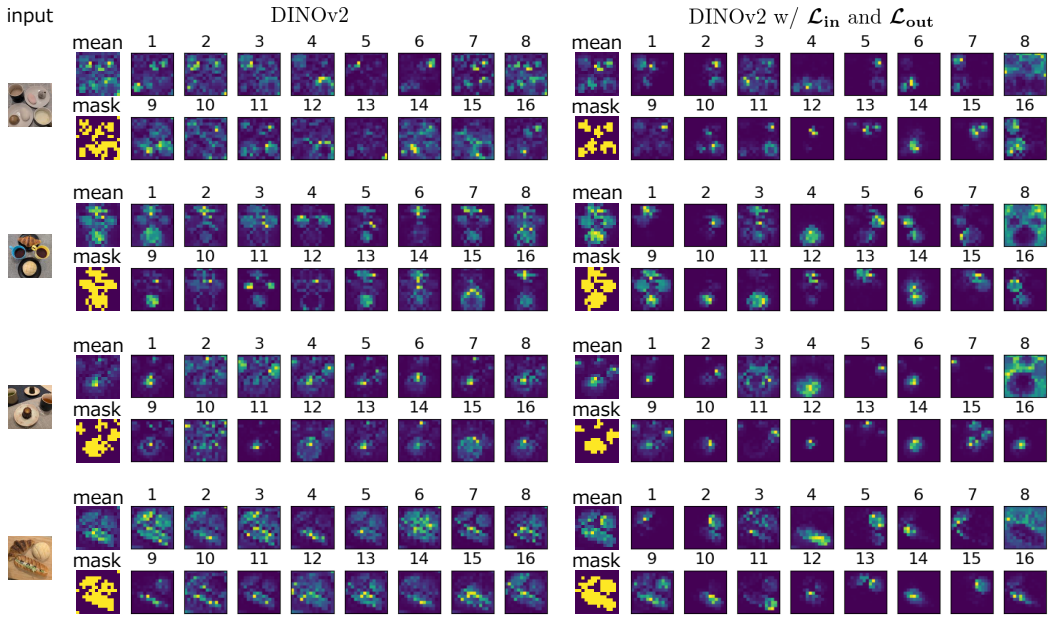

Figure A2: Qualitative comparison of attention maps for the original DINOv2 model and the model with $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$ on example inputs. (Model: ViT-L/16-Reg)

Table A7: Entropy-based evaluation of spatial attention head specialization and diversity

Loss	Mean Per-Head Entropy	Entropy of Head-Averaged Attention
DINOv2	0.8841	0.9265
w/ $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$	0.7492	0.9166

while their aggregation covers a broad range of image regions. This result provides quantitative evidence of spatial specialization and diversity across attention heads.

D.7 Supervised semantic segmentation

We evaluate supervised semantic segmentation using linear probing, following the setting of Siméoni et al. (2026). We conduct the evaluations on ADE20k (Zhou et al., 2017), Cityscapes (Cordts et al., 2016), and PASCAL VOC 2012 (VOC) (Everingham et al., 2015). For VOC, we train the models on the trainaug dataset (Hariharan et al., 2011). We perform a hyperparameter sweep varying the learning rate over $\{1 \times 10^{-4},\ 3 \times 10^{-4},\ 1 \times 10^{-3}\}$ and weight decay over $\{1 \times 10^{-4},\ 1 \times 10^{-3}\}$.

Table A8 compares the supervised semantic segmentation performance of our $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$ model with that of the original DINOv2 model. In terms of model scaling, for both DINOv2 and our model, the Large variant outperforms the Base variant in supervised semantic segmentation. However, in the zero-shot segmentation setting of Tabs. 1 and 5, we observe the opposite trend, with the Base variant outperforming the Large variant. This trend is consistent with both previous work (Stegmüller et al., 2025) and our additional evaluation of the publicly available pretrained DINOv2 models under the same zero-shot segmentation protocol (Tab. A9).

Regarding method comparison in Tab. A8, our $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$ model consistently outperforms the original DINOv2 model across all three datasets for the Base model size. For the Large model size, the $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$ model demonstrates superior performance on Cityscapes, while the original DINOv2 model achieves better results on ADE20k and VOC. Nevertheless, the performance differences between the two models are small in the supervised semantic segmentation setting, with both achieving competitive results and exhibiting no clear performance advantage over one another. These findings

Table A8: Supervised semantic segmentation performance using linear probing. The evaluation metric is mIoU (%). (Model: ViT-{Size}/16-Reg)

Size	Loss	ADE20k	Cityscapes	VOC
B	DINOv2	30.80	54.63	66.82
	w/ $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$	**31.14**	**55.30**	**68.66**
L	DINOv2	**43.15**	64.25	**81.13**
	w/ $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$	42.17	**64.82**	80.89

Table A9: Zero-shot segmentation performance of publicly available DINOv2 pretrained models (Darcet et al., 2024). (Model: ViT-{Size}/16-Reg, Dataset: ImageNet-Seg.)

Size	mIoU (%)↑	Acc (%)↑	mAP (%)↑
B	65.63	82.50	85.69
L	53.71	73.95	81.61
g	55.56	75.47	82.41

Table A10: Comparison of the original model, our model in the main paper, and an additional model with PHS applied to the third-to-last attention block. Zero-shot segmentation is evaluated on ImageNet-Seg. (Model: ViT-L/16-Reg)

Loss		Zero-shot segmentation				
		mIoU (%)↑	Acc (%)↑	mAP (%)↑	k-NN (%)	Linear (%)
DINOv2		63.67	80.69	85.39	**81.7**	**82.6**
w/ $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$,	last block PHS	67.29	83.20	**86.43**	80.3	81.8
w/ $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$,	third-to-last block PHS	**67.85**	**84.09**	86.35	81.0	82.2

clarify that the distinctive property of our model, namely spatial attention head specialization, enables the direct extraction of high-quality features from attention maps in a zero-shot setting while maintaining competitive performance on the supervised downstream task of semantic segmentation.

D.8 Applying head selection beyond the last attention block

While our main experiments apply the head selection method to the last attention block of the student model, we further explore its application to other blocks and discuss the trade-off between the spatial task performance and classification performance. Since attention blocks can be interpreted as integrating information from the `[CLS]` token and patch tokens, we hypothesize that retaining standard attention blocks after the PHS-modified block allows spatially specialized attention information to propagate to the `[CLS]` representation used for classification.

As shown in Tab. A10, applying PHS to the third-to-last attention block with $h_{\text{on}} = 2$ preserves high zero-shot segmentation performance while reducing the accuracy gaps relative to the original DINOv2 model to 0.7% for k-NN classification and 0.4% for linear probing. Although fully closing the classification gap remains an important direction for future work, we believe these additional results highlight the promise of our approach.

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: Yes, the abstract and introduction accurately reflect the paper's contributions and scope.

   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: Limitations of the work is discussed in Limitations section in the Appendix.

   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: The paper does not include theoretical results. However, all the formulas of the proposed method are included in the methodology section.

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: All the information needed to reproduce the main experimental results are explained in details in the experiment section and the Appendix.

   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: [No]

Justification: Open access to the code is not provided at submission time. However, we plan to release our code to the public in future when ready to do so.

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: All the training and test details necessary to understand the results are explained in details in the experiment section and the Appendix.

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: [No]

Justification: The algorithm proposed in the paper requires large-scale vision model training. We do not have the computational resources to perform multiple runs across the experiments.

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: Sufficient information on the computer resources needed to reproduce the experiments is provided in the Appendix.

   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: The research conducted in the paper conform with the NeurIPS 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: [Yes]

    Justification: The paper discuss both potential positive societal impacts and negative societal impacts of the work performed in the societal impacts section in the Appendix.

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: The proposed method does not pose high risk for misuse.

    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: The creators or original owners of assets, used in the paper, are properly credited in the form of citations. The license and terms of use are explicitly mentioned and properly respected where necessary.

    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: The paper does not release new assets.

    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: The paper does not involve crowdsourcing nor research with human subjects.

    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: The paper does not involve crowdsourcing nor research with human subjects.

    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: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    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.