Prompt-Guided Attention Head Selection for Focus-Oriented Image Retrieval (Supplementary Materials)

Yuji Nozawa Yu-Chieh Lin Kazumoto Nakamura Youyang Ng
Kioxia Corporation
{yuji1.nozawa,yuchieh.lin,kazumoto1.nakamura,youyang.ng}@kioxia.com

A1. Extended Analysis

A1.1. Augmenting Current Method with PHS

In this study, we examine the performance of augmenting the *FA* method with our proposed approach. The *FA* method involves an attention additive operation, while our method involves an attention head selection operation. These two methods can be implemented simultaneously without any conflicts. The experimental results, presented in Tab. A1, clearly indicate a complementary relationship between *FA* and our method. Notably, the combined approach, denoted as *FA+Ours*, achieves the highest accuracies across all experimental conditions.

Dataset	Model Size	Method			
		CBIR[1]	FA[9]	Ours	FA+Ours
COCO	small	54.8	55.3	54.9	**55.5**
	base	57.4	57.9	60.6	**61.1**
	large	58.4	58.8	61.3	**61.6**
	giant	58.5	58.8	60.7	**61.0**
PASCAL VOC	small	77.2	77.8	77.4	**77.9**
	base	78.6	79.0	80.9	**81.2**
	large	77.8	78.1	80.3	**80.5**
	giant	78.3	78.5	79.6	**79.8**
Visual Genome	small	30.1	**30.2**	30.1	**30.2**
	base	29.6	29.8	31.4	**31.5**
	large	29.1	29.1	30.2	**30.3**
	giant	29.1	29.2	**29.9**	**29.9**

Table A1. FOIR results when augmenting *FA* with our method (Model: DINOv2, Metric: MP@100 (%)).

A1.2. Method Variations & Parameter Analysis

Our approach offers two distinct modes of operation: (1) Query-Only PHS and (2) Query-DB PHS. The retrieval process of Query-Only PHS mode is compatible with standard prompt-based methods, where PHS is performed solely on the query image. In contrast, Query-DB PHS mode extends the head selection process to the images in the retrieval database, dynamically adapting it for each query. Specifically, this mode modifies each feature in the retrieval database by performing head selection with the same attention heads selected by using the query. By doing so, Query-DB PHS intuitively enhances the feature space of both the query and retrieval database with a query prompt, improving performance in certain scenarios. We mainly report the results of Query-Only PHS as our method in the main paper for its compatible retrieval process. Here, we report additional results for comparing Query-Only PHS and Query-DB PHS.

Dataset	Model Size	DINOv2			CLIP		
		FA[9]	Ours(QO)	Ours(QD)	FA[9]	Ours(QO)	Ours(QD)
COCO	small	**55.3**	54.9	55.2	-	-	-
	base	57.9	**60.6**	60.5	53.3	55.7	**55.8**
	large	58.8	**61.3**	60.8	55.2	58.0	**58.4**
	giant	58.8	60.7	**61.4**	-	-	-
PASCAL VOC	small	**77.8**	77.4	77.3	-	-	-
	base	79.0	**80.9**	80.6	72.2	**73.8**	73.5
	large	78.1	**80.3**	79.8	72.0	**74.2**	73.7
	giant	78.5	79.6	**79.9**	-	-	-
Visual Genome	small	**30.2**	30.1	**30.2**	-	-	-
	base	29.8	**31.4**	31.3	29.5	**30.1**	29.8
	large	29.1	30.2	**30.4**	29.1	**30.2**	29.9
	giant	29.2	29.9	**30.0**	-	-	-

Table A2. FOIR results of method variations. *Ours(QO)* denotes our method with Query-Only PHS. *Ours(QD)* denotes our method with Query-DB PHS (Metric: MP@100 (%)).

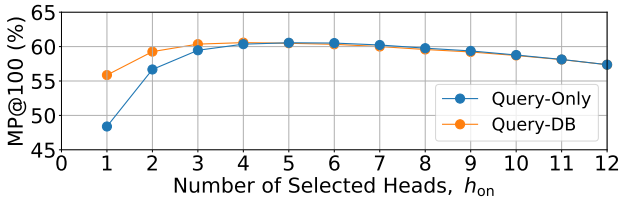


Figure A1. FOIR results with various number of selected heads (Dataset: COCO, Model: DINOv2 *base*).

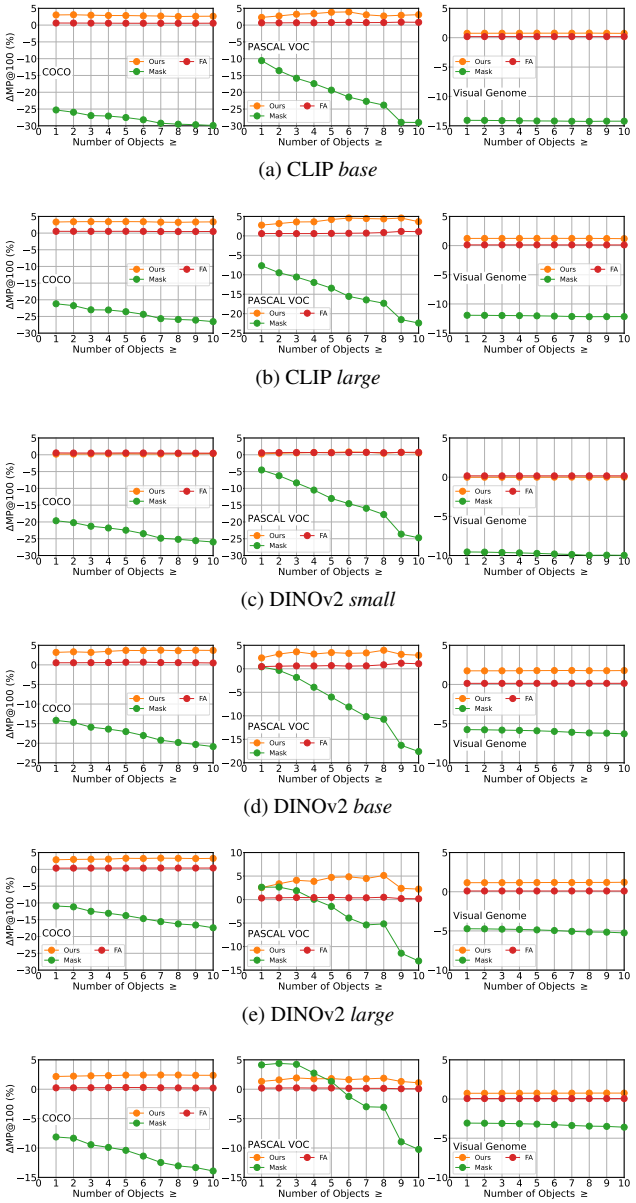
In our method variations, both Query-Only PHS (*QO*) and Query-DB PHS (*QD*) perform similarly and outperform *FA* generally, as shown in Tab. A2. This indicates that ap-

plying our method to query only is sufficient to improve overall performance, while *QD* shows its advantage in certain conditions, highlighting the effectiveness of database-side PHS. The advantage of *QD* can be observed by investigating the parameter of the number of selected heads, h_{on}. We perform a parameter scan for h_{on}. The results in Fig. A1 indicate that while both variations achieve similar performance when h_{on} is set to 5, *QD* demonstrates superior robustness in the selection of h_{on}. This enhancement can be attributed to its retrieval database side PHS component.

Modifying the retrieval database in *QD* incurs higher computational costs. However, the head selection process occurs on the last layer, allowing for caching of query, key, and value features before the attention module. This means that only calculations in half of the last layer are needed, which can be efficiently achieved through GPU parallel processing. Additionally, since LN and FFN operations in Eq. (4) or Eq. (5) of the main paper are applied independently to each token, only the `[CLS]` token needs to be extracted and calculated, further reducing computational requirements.

A1.3. Extended Analysis on Number of Objects

Here, we present an extended analysis on the relationship between the performance of methods and the number of objects in query images. Fig. A2 illustrates the relative performance of the methods with respect to CBIR, where we consider only query images with the number of contained objects equal to or greater than the values on the horizontal axis. This result shows that, except for the DINOv2 *small* model, our method demonstrates substantial enhancements in MP@100 even though the number of objects increases. On the other hand, the *Mask* method consistently exhibits lower performance compared to CBIR as the number of objects increases. In the case of DINOv2 *large* or *giant* for the PASCAL VOC, the *Mask* method outperforms our method when considering all the queries including single-object ones. However, our method outperforms the *Mask* method in both cases of DINOv2 *large* with two or more objects and DINOv2 *giant* with five or more objects, which demonstrates the effectiveness of our method in image retrieval containing many objects.



(a) CLIP *base*

(b) CLIP *large*

(c) DINOv2 *small*

(d) DINOv2 *base*

(e) DINOv2 *large*

(f) DINOv2 *giant*

Figure A2. Relative performance to CBIR. Figures A2a to A2f show the results for different models, respectively. The horizontal axis represents that only query images with the number of contained objects equal to or greater than that value are taken into account.

A1.4. Visual Prompt Noise Analysis: Extended Results

In this study, we examine the influence of noise in visual prompts on the effectiveness of our proposed method when comparing to existing methods. Note that users typically do not generate perfect prompts, necessitating the ability of a prompt-driven method to tolerate some level of noise. To simulate this, we introduce noise into the *Box* prompts by randomly shifting and resizing as described in Sec. A3. The findings, as depicted in Tab. A3, demonstrate that our method's accuracies remain consistently stable even in the presence of prompt noise. This suggests that our method effectively handles imperfect prompts due to its perception matching mechanism. *FA* also performs relatively robust in our experiments due to its attention blending operation, although the accuracies in general are lower than our method. However, *Mask's* accuracies deteriorate when applying prompt noise, indicating the inherent limitation in

image alteration methods.

Dataset	Model Size	Method and Noise						
		CBIR[1]	Mask	Mask-N	FA[9]	FA-N	Ours	Ours-N
COCO	small	54.8	35.1	31.4	55.3	55.2	54.9	54.8
	base	57.4	43.2	38.5	57.9	57.8	60.6	59.6
	large	58.4	47.5	42.1	58.8	58.7	61.3	60.6
	giant	58.5	50.4	44.6	58.8	58.7	60.7	60.4
PASCAL VOC	small	77.2	72.7	65.4	77.8	77.6	77.4	77.1
	base	78.6	79.0	70.5	79.0	78.9	80.9	79.8
	large	77.8	80.4	72.1	78.1	78.0	80.3	79.7
	giant	78.3	82.4	74.1	78.5	78.4	79.6	79.4
Visual Genome	small	30.1	20.5	19.0	30.2	30.2	30.1	30.0
	base	29.6	23.9	21.8	29.8	29.7	31.4	30.9
	large	29.1	24.3	21.9	29.1	29.1	30.2	29.9
	giant	29.1	26.1	22.9	29.2	29.2	29.9	29.7

Table A3. FOIR results with noisy prompts. Method names end with *-N* represent noisy prompts (Model: DINOv2, Metric: MP@100 (%)).

A1.5. ROI Attention Strategy Analysis

In our proposed method, we utilize the *Sum* operation of attention values within the region of interest (ROI) defined by the user-defined prompt to compute the ROI attention for each head. These ROI attentions are then used to determine the selected heads. Alternatively, the *Max* operation can be employed to compute the ROI attention by identifying the patch with the highest value in the ROI. We conducted an ablation study to compare the performance of these two strategies for ROI attention computation. The results presented in Tab. A4 consistently demonstrate that our method, which employs the *Sum* strategy, achieves superior performance across multiple datasets.

Dataset	ROI Attention Strategy		
	CBIR[1]	Max	Sum (ours)
COCO	58.4	60.3	**61.3**
PASCAL VOC	77.8	79.2	**80.3**
Visual Genome	29.1	29.7	**30.2**

Table A4. FOIR results with different ROI attention computation strategies (Model: DINOv2 *large*, Metric: MP@100 (%)).

A1.6. Head Selection Strategy Analysis

In this study, we investigate various strategies for head selection mechanisms. Our method is inspired by Ref. [6], where head selection is performed prior to the output linear projection layer of the MHA module, and the output of the selected heads is multiplied by a scaling factor. It is important to note that alternative head selection strategies exist. For instance, Ref. [2] applies head selection after the output linear projection layer without the use of a scaling factor. Ref. [8] performs head selection before the output linear projection layer, also without using a scaling factor. Additionally, Refs. [3, 5] replaces the attention matrix of selected heads with an identity matrix, which we refer to as the identity type.

In our evaluation, we consider strategies related to the position of head selection and the inclusion of the scaling factor. In Tab. A5, we denote *Before* and *After* to indicate the position of the head selection operation relative to the output linear projection layer. The inclusion of the scaling factor is denoted as *Scale*, and the identity type is denoted as *Identity*. From the results presented in Tab. A5, our approach (*Before+Scale*) demonstrates the highest accuracy among various strategies. This ablation analysis highlights the significance of both the position of head selection and the scaling factor in enhancing the performance of image retrieval.

Dataset	Head Selection Strategy				
	Identity	After	Before	After+Scale	Before+Scale (ours)
COCO	59.5	58.3	59.8	57.9	**61.3**
PASCAL VOC	75.5	77.2	78.7	76.9	**80.3**
Visual Genome	29.0	28.8	29.3	28.8	**30.2**

Table A5. Comparisons of head selection strategies. *Before* and *After* indicate the position of head selection operation in relative to output linear projection layer. *Scale* represents the inclusion of scaling factor. *Identity* denotes the identity matrix replacement method (Model: DINOv2 *large*, Metric: MP@100 (%)).

A1.7. Attention Manipulation Strategy Analysis

In this section, we present an additional study on the attention manipulation strategy in the FOIR task. We create a comparative method called *Attention Mask*, where instead of selecting attention heads, we employ the visual prompt to mask the attentions in the final layer of the ViT model. The results, presented in Tab. A6, demonstrate that the *Attention Mask* approach generally outperforms the previous work of *FA* method. However, our proposed method, PHS, still achieves superior performance compared to *Attention Mask*. Nonetheless, it is worth highlighting that the application of the attention mask in the attention mechanism ensures a more stable performance, avoiding the potential instability that may arise when directly applying the mask to the input image, as shown in the result of *Mask*.

Dataset	ROI Attention Strategy				
	CBIR[1]	Mask	FA[9]	Attention Mask	ours
COCO	58.4	47.5	58.8	59.9	**61.3**
PASCAL VOC	77.8	**80.4**	78.1	78.6	80.3
Visual Genome	29.1	24.3	29.1	29.5	**30.2**

Table A6. FOIR results with different attention manipulation strategies (Model: DINOv2 *large*, Metric: MP@100 (%)).

A1.8. PHS as a Noise Reduction Technique

We conduct an additional study to investigate the potential of our method as a noise reduction technique. In this study, we set up image retrieval by image-region-as-query (IRQ) query format using a crop-based preprocessing technique. Here, we disregard the preprocessing error associated with cropping by utilizing the bounding box labels provided in the datasets as our box prompt. We crop the query images based on the box prompt and resize them to meet the input requirements of the ViT model. We assume that the resulting cropped and resized images contain the necessary information for the retrieval task. In this particular scenario, our method is employed not to select the essential attention, but rather to exclude any undesired noisy attention. To achieve this, we set the value of h_{on} to $h-1$, effectively deactivating a single head corresponding to the undesired noise. The results, as depicted in Tab. A7, indicate that our method achieves superior performance compared to the baseline approach for larger DINOv2 models, although by a slight margin. However, for smaller models, our method performs slightly worse, consistent with our observations in the FOIR results. Nevertheless, it is noteworthy that our method consistently outperforms the baseline approach for all cases involving CLIP models. These outcomes suggest the promising potential of our method as an effective technique for attenuating attention noise in images.

Dataset	Model Size	DINOv2		CLIP	
		CBIR[1]	Ours	CBIR[7]	Ours
COCO	small	**60.0**	58.5	-	-
	base	**66.7**	66.5	45.2	**46.0**
	large	67.3	**67.4**	52.3	**52.5**
	giant	68.7	**68.8**	-	-
PASCAL VOC	small	**86.3**	85.3	-	-
	base	86.8	**86.9**	76.4	**77.0**
	large	83.9	**84.1**	77.8	**77.9**
	giant	84.0	**84.1**	-	-
Visual Genome	small	**34.2**	33.4	-	-
	base	**34.7**	34.6	24.0	**24.4**
	large	**33.4**	**33.4**	25.7	**25.8**
	giant	34.5	**34.6**	-	-

Table A7. PHS as a noise reduction technique (Metric: MP@100 (%)).

A1.9. Visual Analysis with Attention Map: Extended Results

Here, we present the extended results of our visualization analysis on attention maps generated in the final layer of the ViT model, after incorporating our proposed method. As depicted in Fig. A3, our method demonstrates superior intuition in terms of enhanced focus and noise reduction when comparing to *Vanilla* ViT (used in *CBIR*) and *FA*. In contrast, *FA* typically generates attention maps that are comparable to those produced by *Vanilla* ViT, albeit with slightly more concentrated ROI attentions. It is noteworthy that our approach preserves potentially valuable surrounding visual context, which plays a crucial role in reflecting user perception.

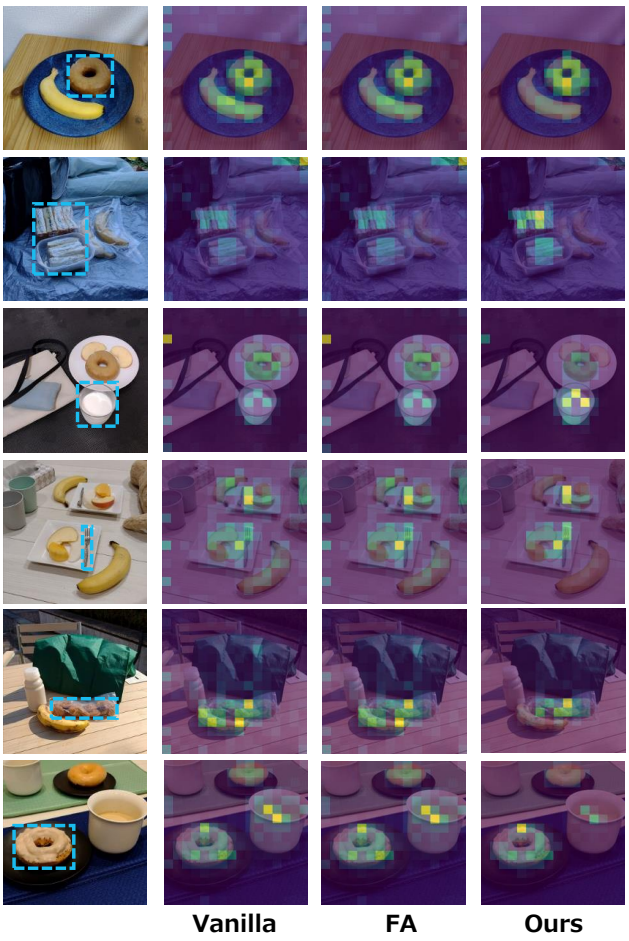


Figure A3. The visualization of attention maps demonstrates that our method performs more intuitively than *Vanilla* and *FA*. Best viewed in color (Model: DINOv2 *giant*).

A1.10. Visual Analysis with Attention Map Across Multiple Model Sizes

In this section, we present a comprehensive visual analysis and investigation of the attention maps generated by *Vanilla* ViT models and our method across different model sizes. Fig. A4 illustrates attention maps of individual heads in the last layer of *Vanilla* ViT for various model sizes. The *base*, *large*, and *giant* models exhibit distinct differentiations across attention heads, indicating the potential of selecting objects based on attention heads. However, the *small* model displays limited differentiation due to its smaller

number of heads. This observation aligns with the overall weaker results obtained with our method on the *small* model in our experiments. When applying our approach, Fig. A5 demonstrates the remarkable alignment between the attention map, visual prompt, and input image with the *giant* and *large* models. Conversely, the *small* model exhibits noisy attention maps even after applying our proposed method. The *base* model's visual quality is somewhere in between. This observation underscores the limitations of our method when dealing with models that have a smaller number of attention heads.

A2. Metrics used in Performance Evaluations

In this section, we describe the details of the performance metrics used in our experiments. To evaluate the performance of our method, we use Mean Precision at k (MP@k) and Mean Average Precision at k (MAP@k), following Ref. [4]. Let $\mathcal{C}$ be the set of categories of objects. We assume that each query image $\mathbf{x}_\mathrm{Q}$ includes objects $o_1, o_2, \ldots, o_{n(\mathbf{x}_\mathrm{Q})}$. We define the category of o_i as $c\left(o_i\right) \in \mathcal{C}$, and the number of objects in category c as $n_c\left(\mathbf{x}_\mathrm{Q}\right)$. In our experiments, it is important to note that the correctness of retrieved images depends on the visual prompt, even if the query image is the same. For a query image $\mathbf{x}_\mathrm{Q}$ with a visual prompt for o_i, we consider the k'th retrieved image $\mathbf{x}_{k'}$ as *correct* if it contains an object in category $c\left(o_i\right)$ and *incorrect* if it does not. We define the score S for $\mathbf{x}_{k'}$ as follows:

$$S\left(\mathbf{x}_{k'}, \mathbf{x}_\mathrm{Q}, o_i\right) = \begin{cases} 1 & \text{if } \mathbf{x}_{k'} \text{ is correct,} \\ 0 & \text{if } \mathbf{x}_{k'} \text{ is incorrect.} \end{cases} \tag{A1}$$

Then, MP@k are calculated by

$$\tilde{\mathrm{P}}@k\left(\mathbf{x}_\mathrm{Q}, o_i\right) = \frac{1}{k} \sum_{1 \leq k' \leq k} S\left(\mathbf{x}_{k'}, \mathbf{x}_\mathrm{Q}, o_i\right), \tag{A2}$$

$$\mathrm{P}@k\left(c\right) = \frac{1}{|\mathcal{I}_{\mathrm{Q},c}|} \sum_{\mathbf{x}_\mathrm{Q} \in \mathcal{I}_{\mathrm{Q},c}} \sum_{i:c(o_i)=c} \frac{\tilde{\mathrm{P}}@k\left(\mathbf{x}_\mathrm{Q}, o_i\right)}{n_c\left(\mathbf{x}_\mathrm{Q}\right)}, \tag{A3}$$

$$\mathrm{MP}@k = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathrm{P}@k\left(c\right), \tag{A4}$$

where $\mathcal{I}_{\mathrm{Q},c}$ is the set of query images that include objects in category c. $\tilde{\mathrm{P}}@k$ is the proportion of correct images in the top-k ones for each visual prompt based query. P@k is the average of $\tilde{\mathrm{P}}@k$ over visual prompt based queries for a fixed c, and MP@k is the average of P@k over $\mathcal{C}$. MAP@k are calculated by

$$\widetilde{\mathrm{AP}}@k\left(\mathbf{x}_\mathrm{Q}, o_i\right) = \frac{1}{|\mathcal{K}|} \sum_{k' \in \mathcal{K}} \tilde{\mathrm{P}}@k'\left(\mathbf{x}_\mathrm{Q}, o_i\right), \tag{A5}$$

$$\mathcal{K} = \left\{k' \in \{1, 2, \ldots, k\} \mid \mathbf{x}_{k'} \text{ is correct}\right\}, \tag{A6}$$

$$\mathrm{AP}@k\left(c\right) = \frac{1}{|\mathcal{I}_{\mathrm{Q},c}|} \sum_{\mathbf{x}_\mathrm{Q} \in \mathcal{I}_{\mathrm{Q},c}} \sum_{i:c(o_i)=c} \frac{\widetilde{\mathrm{AP}}@k\left(\mathbf{x}_\mathrm{Q}, c\right)}{n_c\left(\mathbf{x}_\mathrm{Q}\right)}, \tag{A7}$$

$$\mathrm{MAP}@k = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathrm{AP}@k\left(c\right). \tag{A8}$$

AP@k and MAP@k are metrics that value the higher-ranking images more than P@k and MP@k. In this paper, we employ MP@k and MAP@k as our performance metrics and set k to 100.

A3. Details of Visual Prompt Noise

In our experiments, visual prompt noise is added in the following way. For an object in each original query image, the noiseless box prompt is specified by the positions of the upper left corner (x_0, y_0) and the lower right corner (x_1, y_1) of the box. For the visual prompt with noise, we change (x_0, y_0) and (x_1, y_1) to $(\tilde{x}_0, \tilde{y}_0)$ and $(\tilde{x}_1, \tilde{y}_1)$ randomly as follows:

$$(\tilde{x}_0, \tilde{y}_0) = (x_0, y_0) + (\tilde{c}_x, \tilde{c}_y) - (\tilde{l}_x, \tilde{l}_y), \tag{A9}$$

$$(\tilde{x}_1, \tilde{y}_1) = (x_1, y_1) + (\tilde{c}_x, \tilde{c}_y) + (\tilde{l}_x, \tilde{l}_y), \tag{A10}$$

where $\tilde{c}_x$, $\tilde{c}_y$, $\tilde{l}_x$, and $\tilde{l}_y$ are sampled from the discrete uniform distribution over $[-m, m]$ respectively. The box prompt is shifted by $(\tilde{c}_x, \tilde{c}_y)$ and resized by $(\tilde{l}_x, \tilde{l}_y)$. In all our experiments with noise, we set m to 40 pixels, which is roughly 7.6% of image width and height in average for COCO, 9.4% for PASCAL VOC, and 9.0% for Visual Genome.

A4. Licence info

Table A8 shows the license info of image used in Fig. 7 of the paper.

Image id	563470
URL	http://farm4.staticflickr.com/3370/3518451715_596120fc59_z.jpg
License	CC BY-NC-SA 2.0 DEED http://creativecommons.org/licenses/by-nc-sa/2.0/

Table A8. License info of image in Fig. 7 of the paper.

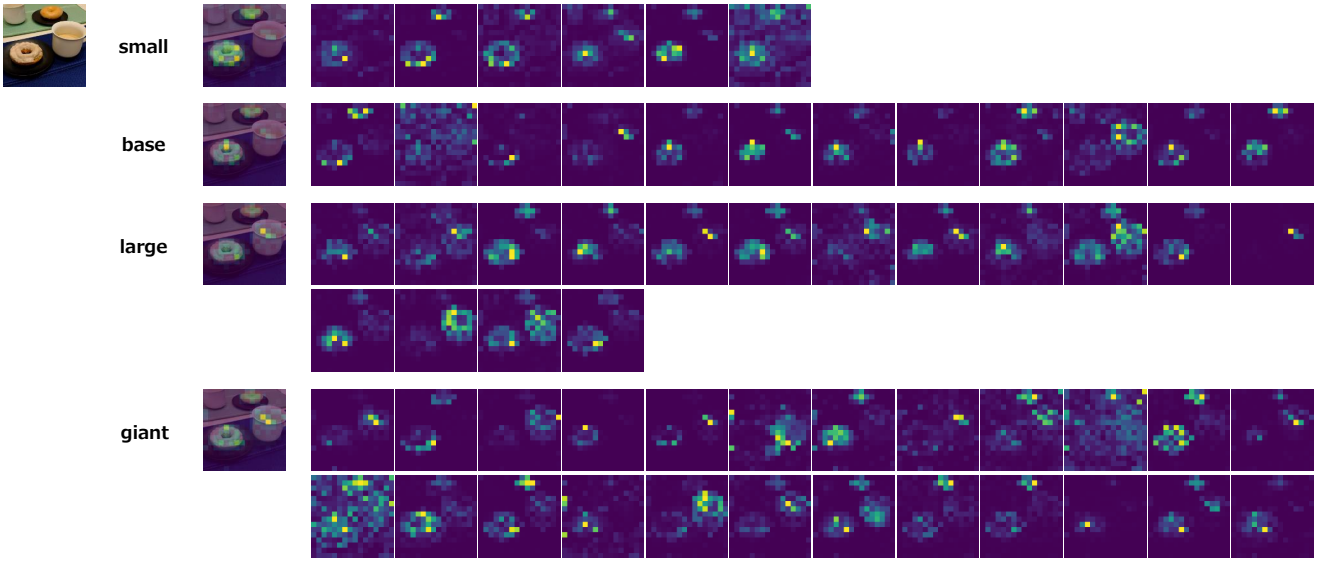


Figure A4. Attention maps visualization across different model sizes for individual attention heads in *Vanilla* ViT. Best viewed in color (Model: DINOv2).

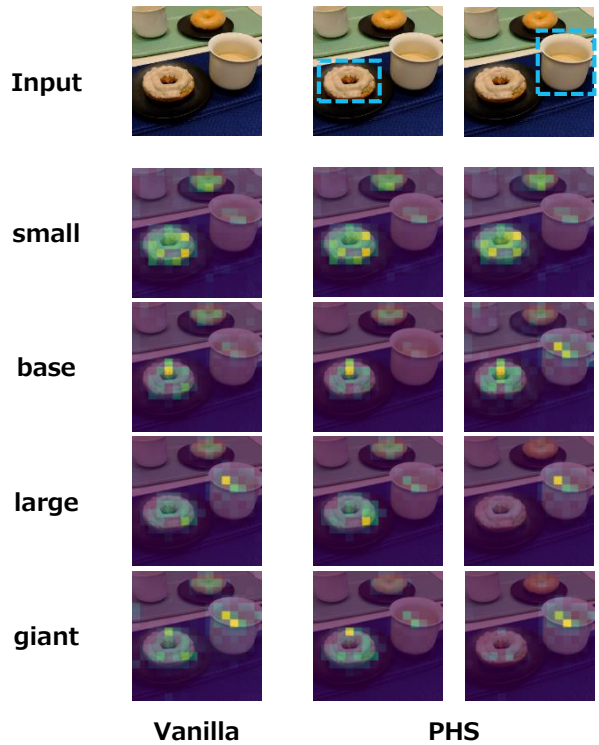


Figure A5. Attention maps visualization across different model sizes when applying our method. Best viewed in color (Model: DINOv2).

References

[1] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision Transformers Need Registers. In *ICLR*, 2024. 1, 3, 4

[2] 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. 3

[3] 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. 3

[4] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A Metric Learning Reality Check. In *ECCV*, 2020. 5

[5] 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. 3

[6] 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. 3

[7] 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. 4

[8] 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. 3

[9] 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. 1, 3