

---

# Learning Visual Prompts for Guiding the Attention of Vision Transformers

---

**Razieh Rezaei\***

School of Computation, Information and Technology  
Technical University of Munich

**Masoud Jalili Sabet\***

AI Innovation, Data:Lab Munich  
Volkswagen AG

**Jindong Gu<sup>†</sup>**

Department of Engineering Science  
University of Oxford

**Daniel Rueckert**

School of Computation, Information and Technology  
Technical University of Munich

**Philip Torr**

Department of Engineering Science  
University of Oxford

**Ashkan Khakzar**

Department of Engineering Science  
University of Oxford

## Abstract

Visual prompting infuses visual information into the input image to adapt models toward specific predictions and tasks. Recently, manually crafted markers such as red circles are shown to guide the model to attend to a target region on the image. However, these markers only work on models trained with data containing those markers. Moreover, finding these prompts requires guesswork or prior knowledge of the domain on which the model is trained. This work circumvents manual design constraints by proposing to learn the visual prompts for guiding the attention of vision transformers. The learned visual prompt, added to any input image would redirect the attention of the pre-trained vision transformer to its spatial location on the image. Specifically, the prompt is learned in a self-supervised manner without requiring annotations and without fine-tuning the vision transformer. Our experiments demonstrate the effectiveness of the proposed optimization-based visual prompting strategy across various pre-trained vision encoders.

## 1 Introduction

The attention mechanism [36] in vision transformers leverages contextually relevant visual information from the entire input image to form embeddings for each token. This process dynamically adjusts the representation of each token, allowing the model to focus selectively on relevant areas, and integrates this context into the embeddings. Such context-aware embeddings lead to enhanced recognition of complex visual data [6]. In our study, we explore whether it is possible to direct the attention mechanism toward a specific input region by introducing and constructing a marker for pre-trained vision transformer models. We pursue this investigation to better *understand what kind of visual information directs the attention mechanism* of various pre-trained vision transformers and to *explore its potential for visual prompting*.

Visual prompting [8] refers to approaches that infuse visual information to the input images to adapt vision foundation models to new tasks. Recent works introduce manual prompting techniques, such as placing colored circle or square markers[32] on the image, or blurring regions surrounding

---

\*Equal Contribution

<sup>†</sup>Corresponding Author

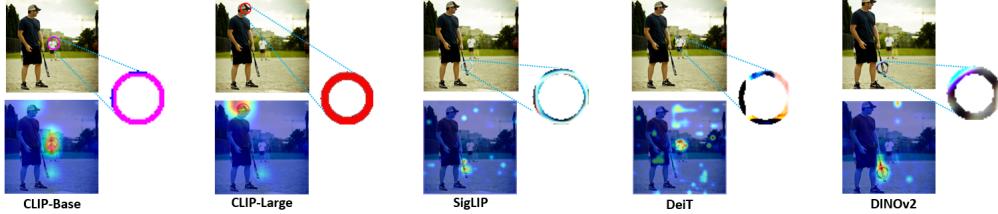


Figure 1: **Learned Prompt for CLIPs [30], SigLIP [43], DeiT [34], DINOv2 [26]**. In our framework, we learn a prompt to draw the attention of the model to a specific point where the prompt is applied. The prompt is optimized for each vision encoder model specifically to generalize across different images. The depicted image is taken from COCO [18].

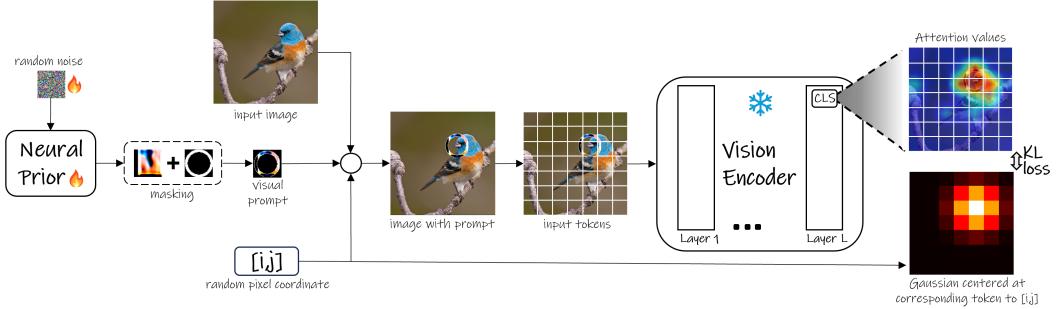
the prompted location [41], and show they are effective in adapting CLIP vision encoder [30] for fine-grained recognition tasks. These works hypothesize that these manual prompting techniques guide the attention of the model and therefore help the model with fine-grained recognition tasks. However, using these manually engineering prompts rely on prior knowledge of biases (or emergent properties) formed in the model during training (It is reported that training data of CLIP contains these markers and blur effects [32, 41]).

In this study, we propose *learning* prompts (without fine-tuning the model) to guide the attention of various large models rather than manually designing them. We explore how we can find visual prompts that attract the attention of a pre-trained (and frozen) vision transformer (ViT [6]) through a self-supervised approach. *Thus, we would not rely on prior knowledge regarding dataset biases, and we can find prompts for any vision transformer trained on any dataset.* The optimization procedure alleviates the need for manual engineering or intuition to find the prompt. Drawing inspiration from universal adversarial patches [22], and employing techniques to improve transferability such as utilizing a network prior [35] to generate adversarial perturbations [29], we propose a self-supervised prompt optimization pipeline. In addition to CLIP, we explore other ViT networks, such as DeiT [34] and DINOv2 [26, 3], each of which follows different training principles than CLIP (e.g. CLIP uses language supervision, while DINOv2 uses self-supervision). Identifying visual prompts for DINOv2 is particularly important because it extracts significantly richer visual features from images and is recently emerging in the new generation of vision-language models [33, 12] due to CLIP’s visual limitations [33]. We observe that manual markers, like a red circle, do not effectively guide DINOv2’s attention, highlighting the need for an optimization-based approach for finding visual prompts.

## 2 Related Work

**Visual Prompting** Prompting has been extensively studied in the NLP community [21]. Recently, researchers have begun exploring the benefits of prompting in image recognition as well. This involves adding a learnable modification to images to guide models towards specific predictions [45, 11, 1, 31, 8]. These prompt-tuning methods [11, 1, 31, 45] optimize a visual prompt that is appended and *fixed* to an input image. For instance, this could involve appending optimizable pixel regions around the image. By doing so, these methods enable the pre-trained model to adapt to new tasks without the need for retraining. This modification is typically applied universally across different images. Follow up approaches have been proposed to improve these visual prompts [44, 42, 16]. More recently, researchers have shown that manually crafted prompts, like a red circle, can effectively guide the attention of models like CLIP when trained on datasets containing similar markers [32]. Moreover, other biases, such as blurring, can be used for prompting [41]. However, such attention-guiding prompts only work effectively on certain models. In this study, we propose learning prompts (without fine-tuning the model) to guide the attention of various models rather than manually designing them.

**Adversarial Patch** The seminal work [27] demonstrates that adversarial examples can be generated by altering just a few pixels in the input. This white-box attack [13, 19, 38, 23] method can deceive models by targeting a very small area. Additionally, certain studies [2, 20] have successfully developed universal, transferable, and targeted adversarial patches. These patches are typically placed



**Figure 2: Overview of Self-supervised Prompt Optimization Framework:** For a given image and a random position for the patch, the patch prior (a random noise) is first passed through a neural prior (an auto-encoder neural network). A desired mask is then applied to the patch to only partially cover the image and avoid losing data. The prompt is positioned on the target location and, after passing through the frozen Vision Encoder, the attention weights are extracted. The desired attention target values are calculated with a Gaussian distribution centered at the token corresponding to the target location. During training, the framework learns a prompt that minimizes the Kullback-Leibler (KL) divergence loss between the attention values of the CLS token and the target distribution.

on the primary object within images. A related study [7] found that model attention can be drawn towards adversarial patches designed to mislead classification results. Our visual prompt can similarly be viewed as an adversarial patch, aiming to divert model attention away from its original focus, but for adapting model behavior towards useful tasks.

**Transferability and Universality of Adversarial Perturbation** The transferability of adversarial perturbation describes the phenomenon that perturbations crafted for one model can deceive another, even with a different architecture [9]. Different from the transferability, the universality of adversarial perturbation is a property that the perturbation is still effective when it is added to different input instances [4]. To improve transferability, we use generative adversarial perturbations [29]. A simple way to increase the universality is to include various input images when optimizing an adversarial perturbation [25], which we also leverage in this work. The transferability and universality of adversarial perturbation have also been studied [2, 20, 40]. These intriguing properties increase the threats in real-world applications. However, how to leverage the two properties for good has not been explored in the community.

### 3 Method

Our goal is to learn a patch such that if it is applied to any part of an input image, the corresponding tokens in the vision encoder would attract more attention, and therefore, the final representation of the image will be manipulated. We propose a self-supervised method to learn such a patch, while the vision-encoder model is utilized in its frozen state and we only require a collection of images without the need for any labels. The learning is about the patch’s RGB color space (three channels) while its shape is predefined, or alternatively, the shape mask can be learned as an additional fourth channel. In the subsequent sections, we provide a detailed explanation of how this prompt is learned through the back-propagation of the self-supervised loss that accounts for the relation between the patch’s location and the attention it receives.

**Input with Prompt.** We aim to discover a visual prompt (patch) that attracts the attention of a transformer-based vision encoder toward a specific location. To achieve this, we work on the RGB image input denoted as  $I \in \mathbb{R}^{n \times n \times 3}$  where  $n$  is the number of pixels in one dimension of the image and 3 indicates the three-dimensional color space. The prompt, denoted as  $\mathcal{P}$ , is also in a similar RGB color space and of size  $m \times m$  (the impact of prompt size on performance is investigated in section 4.1). The designated coordinates  $[i, j]$  within the image  $0 \leq i, j < n$  is then the exact pixel location where the visual prompt will be centrally inserted on. By selecting coordinates only with criteria of  $0 < i \pm \frac{m}{2}, j \pm \frac{m}{2} < n$ , we ensure that the patch falls within the valid range, meaning that no part of the patch extends beyond the boundaries of the image after it is inserted.

The key point here is that by *inserting* we do not simply mean adding the values of the prompt pixel with those of the image pixel. Rather, we aim to identify a general, universal patch that remains effective when applied to different images and locations, regardless of the underlying pixel values. To achieve this, we *replace* the values of the corresponding image pixels with those of the patch pixels  $I[i - \frac{m}{2} : i + \frac{m}{2}, j - \frac{m}{2} : j + \frac{m}{2}] = \mathcal{P}[:, :, :]$ , a process we refer to as *insertion* of patch into the image. To ensure the transferability of the visual prompt, the placement of the patch is randomized across various locations within the image. During the training phase, for an input image, we select  $k$  random locations denoted by coordinates  $[i, j]$  within the validity range criteria mentioned. The patch is then inserted on each coordinate one at a time. This process is repeated for all images  $\mathcal{I}$  in the dataset, resulting in a set of modified input images  $I_{\mathcal{P}}$  for all samples. These modified images, with patches inserted at random locations, are then fed into the vision encoder.

**Transformer-based Vision Encoder.** Within a transformer block with  $L$  layers and  $H$  attention heads, we denote the attention values as  $A_l^i \in \mathbb{R}^{t \times t}$ , where  $A$  is the attention values of the  $i^{th}$  head of the layer  $l$  for an input image with  $t$  tokens. The output of the transformer encoder is typically a contextualized representation of the image tokens, where each token has been influenced by its surrounding tokens through the attention mechanism. This representation, often obtained from a special token (e.g., the CLS token), can then be utilized for downstream vision tasks such as object recognition or image classification. Our method leverages the attention values of the CLS token in the last layer, averaged over the heads of that layer:  $\bar{A}_L[CLS, *] = \sum_{i=1}^H A_L^i[CLS, *]$ .

**Target Gaussian Map.** When the prompt is applied to the input image on the pixel position  $[i, j]$ , it overlays on one or more tiles with the center  $[x, y]$ , refer to Fig. 2. Assuming that the model divides an  $n \times n$  pixel image to  $t \times t$  image tokens with each tile having the pixel size  $n_t$  ( $n/t = n_t$ ), the corresponding patch center  $[x, y]$  can be derived from its pixel position on the image  $[i, j]$  as:  $x, y = (i/n_t, j/n_t)$ .

To construct the target Gaussian map for the patch over the  $t \times t$  token space, a 2D Gaussian distribution is employed. The Gaussian map  $\mathcal{G}(x, y)$  at location  $[x, y]$  is represented by the probability density function of a Gaussian distribution  $N(\mu, \sigma^2)$ , where  $\mu = (x, y)$  denotes the mean (center) of the distribution, and  $\sigma$  is calculated from the Full Width at Half Maximum (FWHM), which in our experiments is set to the patch size in token space  $m/n_t$ :  $\sigma = (m/n_t)/(2\sqrt{2\ln(2)})$ .

**Neural Prior.** To learn the optimized prompt, we aim to avoid the expensive need to fine-tune the large-scale vision encoder by employing it solely in a frozen state. In this way, the only parameters being updated through the training process are the pixel values of the prompt. However, a potential barrier we may encounter in efficiently achieving the optimal patch is the limited learnable variables within the pixel space [35, 15]. Thus we parameterize the patch input space with a neural network as in [35]. Such parameterization is also shown to improve the transferability of optimized perturbations [29]. we use a neural prior  $f$  starting from randomly initialized weights, that receives a random prior input  $\eta \in m \times m$ , and outputs an initial prompt  $\mathcal{P}_{prior} = f(\eta)$  which is then masked by a predefined shape mask  $\mathcal{P} = \mathcal{P}_{prior} * \mathcal{P}_{mask}$  and finally inserted centrally on the  $[i, j]$  pixel of the input image  $I$  to form the input to the vision encoder:  $ViT(I_{\mathcal{P}})$ . Our neural prior  $f$  employs a CNN-based architecture, in particular a U-Net with three layers of downsampling and three upsampling with Sigmoid activation to ensure the values stay between  $[0, 1]$ . Leveraging shared spatial patterns among pixels in the patch enables effective optimization during training which facilitates efficient visual prompt learning [17].

**Objective Function.** Assuming that an image with a patch applied on it  $I_{\mathcal{P}} = Insert(I, \mathcal{P}, [i, j])$  is the input to the  $ViT$ , we define mean of last-layer attention values over attention heads of the vision encoder model as  $\bar{A}_l = Attention(ViT(I_{\mathcal{P}}))$ . Our objective is to train the deep neural prior to output a prompt  $\mathcal{P}$  such that it enhances the attention  $\bar{A}_l$  at corresponding token at  $x, y = (i/n_t, j/n_t)$  with  $n_t \times n_t$  indicating the number of pixels in a token. Having this objective in mind, we calculate the final self-supervised loss as follows:

$$\mathcal{L}(I, [i, j]) = \mathcal{L}_{KL}(\bar{A}_L[CLS, *], \mathcal{G}(x, y)) \quad (1)$$

where  $\mathcal{L}_{KL}$  is the KL-Divergence loss, Throughout the training process, the encoder remains frozen, and only the weights of the neural prior undergo updates as a result of the back-propagating of the defined self-supervised loss.

---

**Algorithm 1** Learning Prompt with Predefined Shape

---

**Require:**  $\mathcal{I}, \eta, n_t$  ▷ Image Collection, Patch Prior Noise,  $n_t^2$  = pixel size of a ViT tile

1: **Initialize:**

    1.  $n = \text{PIXEL\_SIZE}(I)$  ▷ Image size:  $n \times n$

    2: **for**  $I \in \mathcal{I}$  **do**

        3:      $\mathcal{P} \leftarrow f_\theta(\eta)$  ▷ Generate patch from patch prior

        4:      $[i, j] \leftarrow \text{RANDOM}(n)$

        5:      $x, y \leftarrow i/n_t, j/n_t$  ▷ Find corresponding token coordinates

        6:      $I_p \leftarrow \text{INSERT}(I, P, [i, j])$

        7:      $\bar{A}_{CLS} \leftarrow \text{ATTENTION}(\text{ViT}(I_p))$  ▷ Get the averaged attention values for CLS over all heads

        8:      $l \leftarrow \mathcal{L}_{KL}(\bar{A}_{CLS}, \text{GAUSSMAP}(x, y))$

        9:      $\theta \leftarrow \text{ADAMW}(\theta, l)$  ▷ Update the model parameters using AdamW optimizer

    10: **end for**

---

## 4 Experiments

In the following, we first present the learned prompts for CLIP-B/32 and CLIP-L/14, discussing several design considerations and their impacts. We then evaluate the effectiveness of the prompts in the context of Naming Keypoints task. We take it a step further by learning prompts for SigLIP [43], DeiT [34], and DINOv2 [26], each of which either features a slightly different encoder architecture, or has a different training objective function, or varies in pretraining datasets. Finally, we compare how the prompts impact their attention through layers.

**Implementation Details and Settings.** All the vision encoders used are pretrained models available from publicly accessible libraries such as Hugging Face Transformers [39] and PyTorch Hub [28]. Unless specified, we use a learning rate of 1e-3 and batch size of 32 and train our framework for 10 epochs. All experiments were conducted on a machine equipped with a single NVIDIA A100 GPU with 80GB of memory.

**Datasets and Evaluation Metric.** *Training:* To train our neural prior, we use images from ImageNet [5] which consists of millions of labeled images spanning one thousand classes. However, as working with the full ImageNet is resource-demanding, we exploit a subset of it comprising of 10 categories according to [10]. The limited coverage of label classes is not an issue for our work, as our method is not supervised and thus does not rely on specific label types. Furthermore, the over 13k samples in [10] offer a rich and diverse set of images, enhancing the generalizability of our self-supervised training. *Evaluation:* For naming a keypoint task, we utilize CUB-200-2011 test set [37] to evaluate our learned prompt. This dataset contains 5794 bird images of different types, annotated with 15 body-part names and corresponding pixel coordinates as keypoints on the bird image. We evaluate our method by positioning our learned prompt on these keypoints and then calculating the performance based on the percentage of correctly identified body parts. To further evaluate our learned prompt, we test its performance also on the RefCOCO [14], RefCOCO+ [14], and RefCOCOg [24] datasets which consist of images, annotated with bounding boxes around objects in it, each of which is paired with expressions.

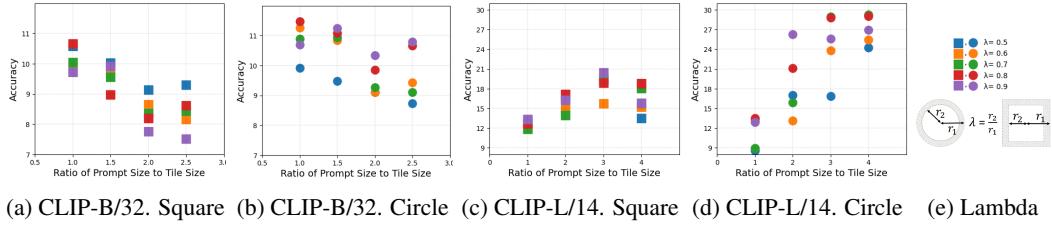
**Models and Baselines.** Specifically, CLIP-B/32 and CLIP-L/14 are employed for their robust zero-shot learning capabilities, leveraging large-scale vision-language pretraining. SigLIP integrates image and language understanding in a synergistic manner. DeiT and DINOv2 are utilized for their state-of-the-art performance in vision transformer architectures, with DeiT focusing on efficient training and DINOv2 providing self-supervised learning benefits. The baseline for comparison involves using the cropped region over the object bounding box in the RefCOCO dataset family. This approach isolates the relevant object from the surrounding context, allowing for a focused evaluation of keypoint identification accuracy. Additionally, random location baselines are employed to assess the robustness of our method against random visual prompts.

### 4.1 Guiding Attention to Name Keypoints

We initiate our experiments with a basic prompt configuration, termed a vanilla patch, which is a simple filled square. This patch undergoes no shape mask filtering during training, and the learned prompt is optimized RGB values. We train patches of varying sizes for the CLIP-B/32 and CLIP-L/14

**Table 1: Vanilla Patches and Scale Constraints.** The tile sizes of the CLIP-B/32 and CLIP-L/14 models are  $32 \times 32$  and  $14 \times 14$ , respectively. We defined the dimensions for the vanilla patch (a simple filled square prompt) accordingly. For CLIP-B/32, the patch sizes are scaled proportionally to the model’s input token at [1, 1.5, 2, 2.5] times larger. For CLIP-L/14, the patch sizes are [1, 2, 3, 4] times larger than the token. These scales are chosen to ensure the patch is sufficiently large for effective learning without excessively covering image information. These prompts are evaluated on the keypoint naming task using the CUB dataset.

Prompt Size	(a) CLIP-B/32				(b) CLIP-L/14			
	32x32	48x48	64x64	80x80	14x14	28x28	42x42	56x56
Visualization								
CUB accuracy	10.97	9.85	9.71	7.35	12.86	16.04	17.41	13.68



**Figure 3: Effect of Patch Parameters:** Performance of CLIP models on CUB dataset for varying ratios of hollow circle and square masks over different patch sizes. The tile size of models ViT-Base and ViT-Large are  $32 \times 32$  and  $14 \times 14$ , respectively. (e) shows the ratio colors and the illustration of  $\lambda$ .

models. Our experiments utilize prompt sizes corresponding to the model token dimensions. As CLIP models split input images into tokens of different pixel sizes (base model:  $32 \times 32$ , large model:  $14 \times 14$ ), our learned prompt of one-token size inherently possesses varying pixel dimensions.

Table 1 presents the visualizations of the prompts learned for both models at different sizes. We select sizes proportional to the tokens: 1, 2, 3, and 4 times larger for CLIP-L/14. For CLIP-B/32, due to the substantial token pixel size, we opt for prompts that are 1, 1.5, 2, and 2.5 times larger to avoid excessive pixel coverage by the patch. We report the accuracy of keypoint naming on the CUB dataset. The results indicate that for CLIP-B/32, the prompt size equal to 1 token yields the best performance, whereas for CLIP-L/14, the 3-times larger prompt achieves the highest accuracy. Looking into this deeper, this suggests that optimal accuracy is achieved when the prompt covers approximately the same area size on the image ( $32 \times 32$  and  $42 \times 42$ ) for both models. This may be because a larger patch, while better at manipulating the image, also covers more of it, leading to information loss and lower performance.

**Beyond vanilla patch towards using shape priors:** Initially, we employed a filled square, but previous studies imply that the geometric shape of the patch influences its effectiveness [32, 1]. To investigate the influence of prompt shape on performance in naming keypoints task for CLIP models, we employ hollow square and circle as predefined shape masks. We define a thickness factor as  $\lambda = \frac{\text{outer Diameter}}{\text{inner Diameter}}$  as shown in Figure 3 and investigate the performance for different thicknesses together with different patch sizes, again proportional to the model’s token size.

What we observe from the performances in Figure 3 is that the circle-shaped prompt yields higher accuracy compared to the square-shaped prompt. For CLIP-B/32, the circle prompt achieves the highest accuracy of 11.51, and the best square prompt yields slightly lower accuracy. For CLIP-L/14, the difference is significant: the circle prompt reaches an accuracy of approximately 30.5, whereas the square prompt only achieves about 20.5. The comparison of patch sizes reveals findings consistent with the vanilla patch results. For CLIP-B/32, the optimal patch size remains the same as the default size for one token. In contrast, for CLIP-L/14, the highest accuracy is obtained with prompt three times larger than one token of the large clip model, the size covering a similar number of pixels. The visualizations of the learned prompts with the highest performance are depicted in Table 2. We observe that the red color for CLIP-L/14 and the pink color for CLIP-B/32 are predominantly present

in the learned prompts. Interestingly, for CLIP-L/14, when the prompt shape is a circle, it learned a solid red color which confirms the findings in [32].

**Table 2: Visualization of optimal prompt for various CLIP encoders.** Here we visualize the best-performing square frame and circle-shape prompts from Figure 3. C stands for Circle and S for Square. Colors of red and pink are predominant for CLIP-B/32 and CLIP-L/14, respectively. It is noteworthy that the optimal prompt for CLIP-L/14 is a red circle. Even when we use a square shape prior, a red circle emerges inside the square. Previously, [32] tried various markers using intuition and identified a red circle as an effective prompt. Interestingly, the marker seems to be the optimal marker for guiding the attention of CLIP-L/14. Though for CLIP-L/14 the red circle is not optimum. Thus it may be an emergent property arising from scale.

Encoder/Shape	CLIP-B/32, C	CLIP-B/32, S	CLIP-L/14, C	CLIP-L/14, S
visualization				

Now that we have developed our prompt and explored the design configurations, we are interested in evaluating its performance compared to baseline methods on different datasets. To this end, we train our method on the ImageNet dataset and test it on the CUB and RefCOCO datasets. These datasets are new to the prompt and have not been seen during the training phase. Additionally, they offer the advantage of annotated image locations with names of specific body parts or objects for CUB and RefCOCO, respectively.

We define the baselines as *Random*: Randomly selected areas. *Crop*: Removing the area around the bounding box for RefCOCO and defining a square similar to the prompt size for CUB, then cutting out that area. *Blur*: Similar to cropping, but with the additional step of blurring the outer area using the averaging method with a kernel size of 5.

As indicated by the results presented in Table 3, the Crop method exhibits superior performance in RefCOCO tasks, but its effectiveness diminishes when applied to the CUB dataset. In contrast, the Blur method shows notable improvements in RefCOCO tasks. Our observations suggest that RefCOCO object descriptions often focus solely on the target object rather than relying heavily on scene context. This tendency explains the success of both the Blur and Crop methods in this dataset.

However, in the CUB dataset, accurately recognizing animals can be challenging for the vision encoders, making it difficult to detect specific body parts. Despite this challenge, our learned prompts generally outperform other baselines in the CUB dataset. In RefCOCO tasks, our method consistently achieves second place after the effective Crop method. These findings underscore the effectiveness of our approach, especially when the context of the image is important.

## 4.2 Investigating Attention Dynamics in Various Vision Encoders

Various vision encoder models are trained on diverse datasets, leading to behaviors that may differ from CLIP models. More importantly, these vision foundations use different training paradigms. CLIP vision encoder is trained with language supervision, while DINOv2 is trained by self-supervision using visual information only. Therefore, during training is no supervision signal for the model to learn that colored markers (such as red circles) signify the importance of the region the marker is placed. Additionally, their architectural details might vary slightly, e.g., SigLIP has no CLS token and instead averages over tokens by an attention pooling mechanism. As a result, the optimal prompt to redirect their attention can also be different. To investigate this, we use a similar self-supervised training approach to learn prompts of SigLIP, DeiT, and DINOv2 as well.

Figure 5 shows the visualization of the learned prompt for each vision encoder. The different appearance of the visual prompts confirms that there is no single prompt that is universally optimal for all encoders. We apply the learned prompts on to three different locations on an image and compare the heatmap visualizations of the prompted image with the original, unprompted image. For almost all encoders, the attention heatmap shifts towards the prompt location, demonstrating the effectiveness of our prompt in manipulating the attention of the corresponding visual encoder models.

**The effect of the optimized prompt on attention across layers:** We examine the attention averaged values of tokens in the area overlaid by the prompt before and after its application for over 1000 images from MS COCO while the prompt location is random. By analyzing attention values across

Table 3: Performance comparison of different methods across various datasets. Bold numbers indicate the best performance, while underlined numbers denote the second-best performance. The Crop method demonstrates superior performance in RefCOCO tasks but exhibits reduced effectiveness in the CUB dataset. Conversely, the Blur method performs significantly better in RefCOCO tasks. Our observations reveal that object descriptions in RefCOCO do not always rely heavily on scene context, explaining the success of both Blur and Crop methods. However, in the CUB dataset, where context is crucial for proper animal detection, losing context leads to decreased model performance.

Encoder	Prompt	CUB			RefCOCO			RefCOCO+			RefCOCOg	
		K2N	TestA	TestB	Val	TestA	TestB	Val	Test	val		
CLIP-B/32	Random	8.2	36.9	39.5	37.8	36.9	39.7	37.9	39.5	39.1		
	Crop	<b>15.6</b>	56.6	47.7	52.6	60.4	51.7	55.9	60.3	59.5		
	Blur	7.4	<u>39.6</u>	<u>41.1</u>	<u>40.0</u>	<u>39.5</u>	<u>41.2</u>	<u>40.0</u>	<u>41.3</u>	<u>41.1</u>		
CLIP-L/14	$\textcircled{3}_3 \times 3_2$	11.39	37.7	40.9	39.3	38.5	40.2	38.5	39.1	39.0		
	Crop	18.5	58.8	47.8	53.5	63.5	51.4	57.5	61.2	60.6		
	Blur	11.0	42.5	41.6	41.2	43.3	41.2	41.9	42.9	43.2		
SigLIP	$\textcircled{4}_2 \times 4_2$	<b>29.7</b>	<u>49.5</u>	<u>43.9</u>	<u>46.7</u>	<u>51.3</u>	<u>46.2</u>	<u>48.1</u>	<u>50.4</u>	<u>49.0</u>		
	Crop	19.6	62.6	51.3	57.9	66.8	56.7	62.3	69.2	68.5		
	Blur	13.1	42.9	41.3	41.8	43.6	43.0	42.8	44.2	44.5		
SigLIP	$\textcircled{4}_8 \times 4_8$	<b>30.8</b>	<u>43.7</u>	<u>43.5</u>	<u>42.9</u>	<u>45.1</u>	<u>43.8</u>	<u>44.0</u>	<u>45.1</u>	<u>45.0</u>		

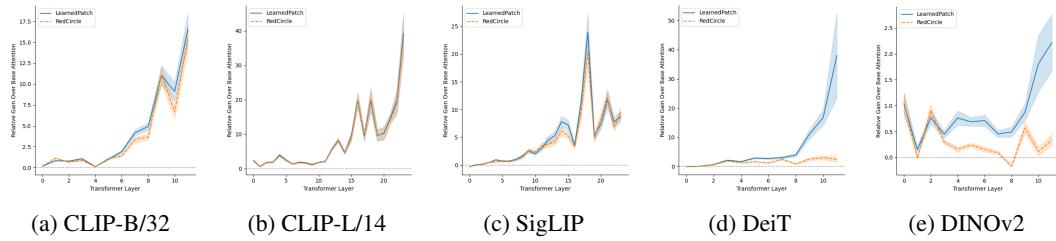
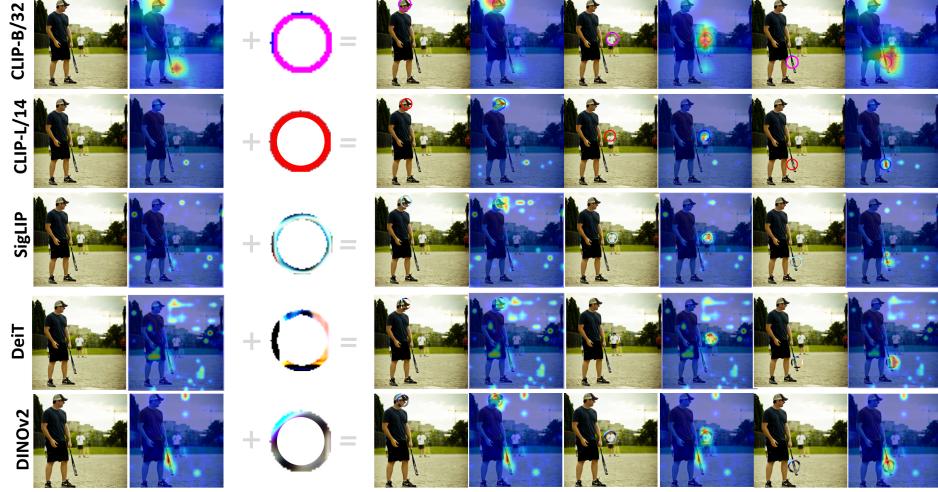


Figure 4: **Attention Gain with Prompt Usage throughout Layers.** We applied our learned prompt to random locations on 1000 samples from the MS COCO dataset (blue line). For comparison, we also used a simple red circle prompt on the same locations (orange line). The Relative Gains are calculated by dividing the difference between attention values before and after applying the prompt on the image by the original attention values of the overlaid tokens without any prompt. A notable observation is that while the simple red circle prompt is as effective for CLIP-L/14 and CLIP-B/32, it is significantly less effective for DeiT and DINOv2 compared to the learned prompt in terms of redirecting attention to a specific location.

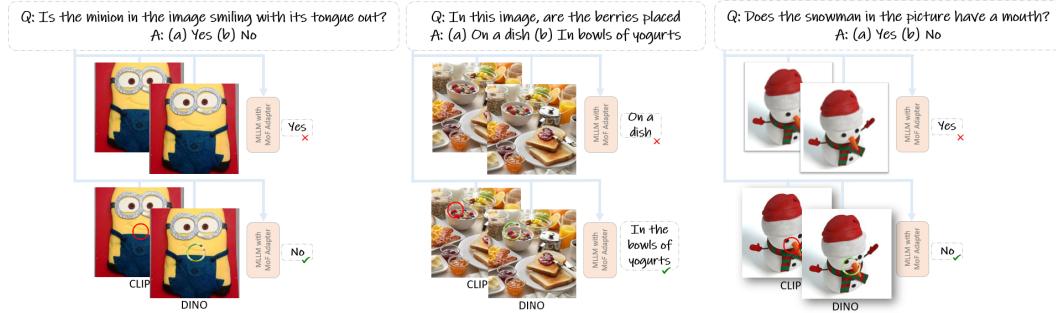
layers, we can assess the prompt’s impact. Since attention values are relative, we define Attention Gain as the ratio of the difference of averaged attention values of tokens influenced by the prompt to their original attention values. Figure 4 illustrates the attention gains across different models. The blue line represents the gains for each model’s optimized prompt, while the orange line shows the gains for a simple red circle marker prompt.

**Red circle does not guide attention of DINOv2 and DeiT:** From Figure 4 we can see that the unique prompts for DeiT and DINOv2 draw significantly larger attention to themselves compared to the red circle prompt, showing the power of our learning framework. This could confirm the suggestion that the presence of red circles in the training data of CLIP models and SigLIP has made them particularly sensitive to this predefined feature [32], which is not the case for other vision encoders such as DeiT and DINOv2.

**The significance of optimizing prompts for self-supervised models such as DINOv2:** To further evaluate the effectiveness of our prompts, we use the MLLM (Multimodal Large Language Model) introduced in [33], which uses an *Interleaved Mixture-of-Features* approach to spatially leverage interleaving CLIP and DINOv2 visual tokens after an adapter. In Figure 6, we compare [33] model’s performance with and without our unique CLIP-L/14 and DINOv2 prompts applied to the image input to see if it improves the question answering responses of their model on proposed MMVP



**Figure 5: Learned Prompt for Different pretrained Vision Transformers.** The optimal visual prompts are not the same and each model has its own unique pattern. The prompt is optimized to generalize across images. Comparing attention heatmaps of the original (on left) and prompted images (on right) reveals how effectively the prompt directs attention to specific locations. The prompt is optimized for each ViT (CLIP-B/32, CLIP-L/14, SigLIP, DeiT, and DINOv2) separately, and is optimized over 20k random samples from ImageNet [5]. The depicted image is taken from MS COCO [18].



**Figure 6: Prompts for the new generation of Vision Language Models:** New vision language models are moving beyond clip vision encoders towards using richer vision encoders such as DINOv2. Therefore it becomes progressively necessary to identify visual prompts for the new generation of models. This figure depicts an example to show the potential of using the optimized prompts in the new generation of vision-language models, in this case, LLaVA+MoF [33] on examples from the MMVP dataset.

(Multimodal Visual Pattern) Benchmark. The examples demonstrate that our prompts enable the model to generate more accurate answers, which means they create improved embeddings. This indicates that our attention-guiding prompts have promising applications in a variety of vision tasks that can be explored in future works.

## 5 Conclusion

In this work, we proposed a self-supervised optimization-based visual prompting technique for guiding the attention of vision transformers, thereby avoiding the limitations of manually crafted prompts. Our method demonstrated the ability to guide the attention of various vision transformer models, such as the CLIP family, SigLIP, DeiT, and DINOv2, without requiring prior knowledge of dataset biases and without fine-tuning the models. The transferability of the prompt across different

images was accomplished by leveraging a deep network prior. Our experiments confirmed that the learned prompts successfully redirect the attention of not only the CLIP family of models which are trained with language supervision, but also the purely self-supervised models such as DINOv2. As new vision language foundation models are leveraging the self-supervised vision encoders for their superior ability to extract visual features, the proposed optimization-based proves to be helpful for prompting upcoming models.

## References

- [1] H. Bahng, A. Jahanian, S. Sankaranarayanan, and P. Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022.
- [2] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer. Adversarial patch. *arXiv:1712.09665v1*, 2017.
- [3] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.
- [4] A. Chaubey, N. Agrawal, K. Barnwal, K. K. Guliani, and P. Mehta. Universal adversarial perturbations: A survey. *arXiv preprint arXiv:2005.08087*, 2020.
- [5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.
- [7] J. Gu, V. Tresp, and Y. Qin. Are vision transformers robust to patch perturbations? In *European Conference on Computer Vision*, pages 404–421. Springer, 2022.
- [8] J. Gu, Z. Han, S. Chen, A. Beirami, B. He, G. Zhang, R. Liao, Y. Qin, V. Tresp, and P. Torr. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*, 2023.
- [9] J. Gu, X. Jia, P. de Jorge, W. Yu, X. Liu, A. Ma, Y. Xun, A. Hu, A. Khakzar, Z. Li, et al. A survey on transferability of adversarial examples across deep neural networks. *arXiv preprint arXiv:2310.17626*, 2023.
- [10] J. Howard. imagenette. <https://github.com/fastai/imagenette>, 2019.
- [11] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022.
- [12] O. F. Kar, A. Tonioni, P. Poklukar, A. Kulshrestha, A. Zamir, and F. Tombari. BRAVE: Broadening the visual encoding of vision-language models. *arXiv preprint arXiv:2404.07204*, 2024.
- [13] D. Karmon, D. Zoran, and Y. Goldberg. Lavan: Localized and visible adversarial noise. In *International Conference on Machine Learning (ICML)*, 2018.
- [14] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.
- [15] A. Khakzar, P. Khorsandi, R. Nobahari, and N. Navab. Do explanations explain? model knows best. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10244–10253, June 2022.
- [16] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan. Maple: Multi-modal prompt learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023.

- [17] A. Krull, T.-O. Buchholz, and F. Jug. Noise2void - learning denoising from single noisy images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.
- [19] A. Liu, X. Liu, J. Fan, Y. Ma, A. Zhang, H. Xie, and D. Tao. Perceptual-sensitive gan for generating adversarial patches. In *AAAI*, 2019.
- [20] A. Liu, J. Wang, X. Liu, B. Cao, C. Zhang, and H. Yu. Bias-based universal adversarial patch attack for automatic check-out. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 395–410. Springer, 2020.
- [21] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.
- [22] H. Luo, J. Gu, F. Liu, and P. Torr. An image is worth 1000 lies: Transferability of adversarial images across prompts on vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=nc5GgFAvtk>.
- [23] J. Luo, T. Bai, and J. Zhao. Generating adversarial yet inconspicuous patches with a single image (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15837–15838, 2021.
- [24] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016.
- [25] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.
- [26] M. Oquab, T. Dariseti, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- [27] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, 2016.
- [28] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.
- [29] O. Poursaeed, I. Katsman, B. Gao, and S. Belongie. Generative adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4422–4431, 2018.
- [30] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/radford21a.html>.
- [31] S. Shen, S. Yang, T. Zhang, B. Zhai, J. E. Gonzalez, K. Keutzer, and T. Darrell. Multitask vision-language prompt tuning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5656–5667, 2024.
- [32] A. Shtedritski, C. Rupprecht, and A. Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11987–11997, October 2023.

- [33] S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024.
- [34] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.
- [35] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018.
- [36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [37] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.
- [38] J. Wang, A. Liu, X. Bai, and X. Liu. Universal adversarial patch attack for automatic checkout using perceptual and attentional bias. *IEEE Transactions on Image Processing*, 31:598–611, 2021.
- [39] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.
- [40] Z. Xiao, X. Gao, C. Fu, Y. Dong, W. Gao, X. Zhang, J. Zhou, and J. Zhu. Improving transferability of adversarial patches on face recognition with generative models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11845–11854, 2021.
- [41] L. Yang, Y. Wang, X. Li, X. Wang, and J. Yang. Fine-grained visual prompting. *Advances in Neural Information Processing Systems*, 36, 2024.
- [42] Y. Zang, W. Li, K. Zhou, C. Huang, and C. C. Loy. Unified Vision and Language Prompt Learning. *arXiv:2210.07225*, 2022.
- [43] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.
- [44] K. Zhou, J. Yang, C. C. Loy, and Z. Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022.
- [45] K. Zhou, J. Yang, C. C. Loy, and Z. Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.