FALIP: Visual Prompt as Foveal Attention Boosts CLIP Zero-Shot Performance

Jiedong Zhuang, Jiaqi Hu, Lianrui Mu, Rui Hu, Xiaoyu Liang, Jiangnan Ye, and Haoji Hu

Zhejiang University
{zhuangjiedong,haoji_hu}@zju.edu.cn

Abstract. CLIP has achieved impressive zero-shot performance after pretraining on a large-scale dataset consisting of paired image-text data. Previous works have utilized CLIP by incorporating manually designed visual prompts like colored circles and blur masks into the images to guide the model's attention, showing enhanced zero-shot performance in downstream tasks. Although these methods have achieved promising results, they inevitably alter the original information of the images, which can lead to failure in specific tasks. We propose a train-free method **F**oveal-**A**ttention C**LIP** (**FALIP**), which adjusts the CLIP's attention by inserting foveal attention masks into the multi-head self-attention module. We demonstrate FALIP effectively boosts CLIP zero-shot performance in tasks such as referring expressions comprehension, image classification, and 3D point cloud recognition. Experimental results further show that FALIP outperforms existing methods on most metrics and can augment current methods to enhance their performance. Our project page is link to https://pumpkin805.github.io/FALIP/

Keywords: zero-shot learning · visual prompt · visual-language model

1 Introduction

Vision-Language Models (VLMs) like CLIP [35] have shown remarkable zero-shot performance in various tasks without further training [10,43,49,62,63]. To further expand CLIP's capability, researchers have explored strategies to manually craft input prompts to CLIP. While previous works mainly focus on *text* prompts inspired by research in large language models (LLMs), *visual* prompts have been recently introduced [24,40,41,55,56], utilizing symbols such as boxes, points, circles, masks, and others to give models additional cues. These techniques achieve various levels of success on tasks including referring expression comprehension [40,41,55,56], part detection [55] and keypoint matching [40].

Despite the promising results on several tasks, we lack a systematic and intuitive understanding of why visual prompts are effective in improving the zero-shot capability of CLIP. To investigate the mechanisms behind the success of manually designed visual prompts, we conduct an in-depth exploration, starting from analyzing the effectiveness of visual prompts on CLIP over the task of

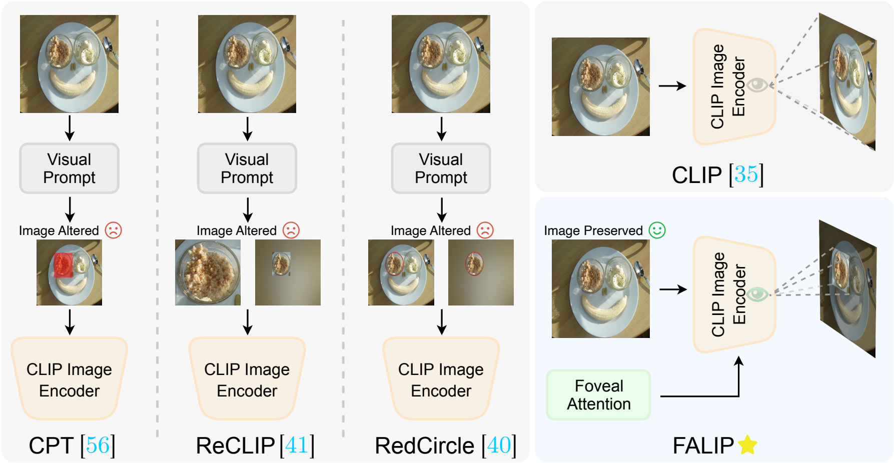


Fig. 1: Overview of visual prompt based methods and FALIP. ***Left*** is the the visual prompt methods [40, 41, 56]. They perform image editing (such as colored boxes, cropping, circles, blur masks, etc.) enabling CLIP to perceive specific regions. ***Bottom right*** is FALIP. It does not alter the content of the original image. The gray dashed line represents the attention of model. Compared to the original CLIP, FALIP aligns more with human visual characteristics.

referring expressions comprehension [33, 58]. Our objectives are twofold: 1) a more principled understanding of the effectiveness of visual prompts, and 2) to design more effective strategies to enhance the zero-shot capability of CLIP.

In our study, we first examine CLIP's attention maps on numerous images. Fig. 2 notes a clear link between visual prompts and model focus: *model attention often zeroes in on areas marked by the visual prompt.* Using a state-of-the-art visual prompt technique (RedCircle [40]) for zero-shot classification with CLIP, we unexpectedly find that its zero-shot efficacy diminishes with the visual prompt in place. This outcome prompts us to re-think the task-specific effectiveness of such visual prompt methods. Crucially, these methods edit the image directly, potentially compromising its integrity by occluding or destroying vital details in the image. For example, RedCircle introduces additional red elements, which could potentially skew fine-grained classification outcomes. Likewise, the "blur mask" obfuscates much of the image, retaining only basic shapes and thus discarding significant detail in certain areas. Consequently, these approaches may be ineffective in scenarios demanding high image fidelity. Our discoveries highlight a paradox in current visual prompt strategies: although aiming to direct CLIP's focus to particular image areas, they inadvertently strip away crucial content, undermining the model's performance. This raises an essential question: is it possible to leverage visual prompts' advantages without sacrificing the integrity of the input image?

In this paper, we introduce **F**oveal-**A**ttention **CLIP**(**FALIP**), a novel approach that aligns regions of attention (ROA) in images with their correspond-

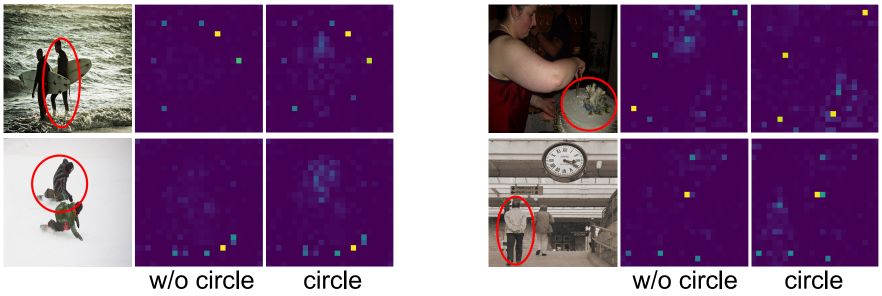


Fig. 2: The shift in the model's attention before and after incorporating visual prompts. It can be observed that visual prompts can guide the model's attention to specific regions.

ing token positions, constructing a foveal attention mechanism into the model's self-attention layer. Drawing inspiration from human visual perception [2, 51] featuring selective focus and specific region processing, FALIP enhances CLIP with similar attentional characteristics. Fig. 1 provides a concise illustration comparing the attention mechanisms of CLIP and human cognition, as well as a comparison between our method and existing techniques. FALIP has been rigorously tested across numerous datasets, demonstrating competitive performance. Remarkably, it is designed to be plug-and-play, adding negligible computational cost, requiring no further training, and complementing existing approaches. In addition, through our experiments, we uncover that the CLIP model attention heads vary in their response to visual prompts, and we find that adjusting these heads may further unleash the effectiveness of visual prompts.

In summary, our main contributions can be outlined as follows: (1) We propose FALIP, a novel method to adaptively guide the attention of CLIP during inference without additional training. (2) We extensively evaluate FALIP on a wide range of tasks and datasets and achieve competitive performance compared to existing methods. (3) We present an in-depth analysis that demystifies the surprising effectiveness of visual prompts and sheds new light on improving the zero-shot inference capability of CLIP. (4) We discover that different attention heads in the CLIP model exhibit varying levels of sensitivity to visual prompts, and they can be adjusted to unleash the full potential of visual prompts.

2 Related Work

Vision-Language Models. CLIP [35] uses massive amounts of image-text paired data for contrastive learning, enabling it to acquire powerful zero-shot image classification and image-text retrieval capabilities. Its introduction also further propelled the subsequent development of models such as Multimodal Large Language Models (MLLMs), BLIP [25, 26], and LLaVA [30, 31, 44, 45] that built upon the foundation established by CLIP. Although recent other vision models based on ViT [11], such as DINO [3, 34], MAE [16], and MoCo [5], have achieved remarkable performance in single-modal visual tasks, they do not possess the cross-modal capabilities of CLIP. This paper focuses on CLIP and proposes a method that can further enhance the zero-shot capability of CLIP.

Prompt. Prompt learning is an emerging topic in computer vision and natural language processing. Previous studies commonly involve inserting learnable tokens into the model input. They add learnable embeddings to the text input [15, 28, 32, 66, 67], or incorporate them into the image input [1, 8, 19]. Other methods incorporate learnable tokens into both the text and image inputs [7,21,38,59]. Most of these methods require retraining because they involve fine-tuning specific parameters of a pre-trained model to adapt to specific downstream task datasets. Some works manually introduce prompts (box, circle, blur mask) within the images to guide the model towards the desired objects or regions [40, 41, 55, 56]. However, the majority of these methods are reliant on the pretraining data of CLIP, and they alter the original information of the images, making it difficult to generalize to certain downstream tasks. Our method can be regarded as a form of *attention prompt*. What sets our method apart from these works is that our method does not need training, introduction of additional models, or altering the content of the original images.

CLIP Region Awareness. To enhance the region awareness of CLIP, several methods have been explored in the field of detection and segmentation. SAN [53] trains a extra transformer network to assist CLIP in recognizing local features. ODISE [52] employ a trainable mask generator to guide CLIP's focus towards specific local regions of interest. RegionCLIP [64], OvarNet [4], Alpha-CLIP [42] and UMG-CLIP [39] use region-level image-text pairs to fine-tune the model. MaskAdaptedCLIP [29] generates mask-text pairs through a pseudo-labeling process to fine-tune CLIP. MaskCLIP [10] and MasQCLIP [54] introduce additional learnable tokens enhancing CLIP's ability to classify objects. Unlike our method, these methods have higher requirements for training data and often require additional training or fine-tuning processes.

3 Method

This section begins with a brief introduction to CLIP and visual prompts. We then proceed to discuss how to apply FALIP to zero-shot tasks such as referring expressions comprehension, image classification and 3D point cloud recognition.

3.1 Preliminary

The core architecture of CLIP encompasses an image encoder $\mathbf{V}$ and a text encoder $\mathbf{T}$. CLIP utilizes contrastive learning to distinguish between matching and mismatched image-text pairs. The text encoder is based on a Transformer [47] architecture, while the image encoder can be either a ViT [11] or a ResNet [18]; our work utilizes the ViT model denoted as $\mathbf{V}$.

When applying different visual prompts to images processed by CLIP's image encoder $\mathbf{V}$, we change the model's attention towards prompted regions, as

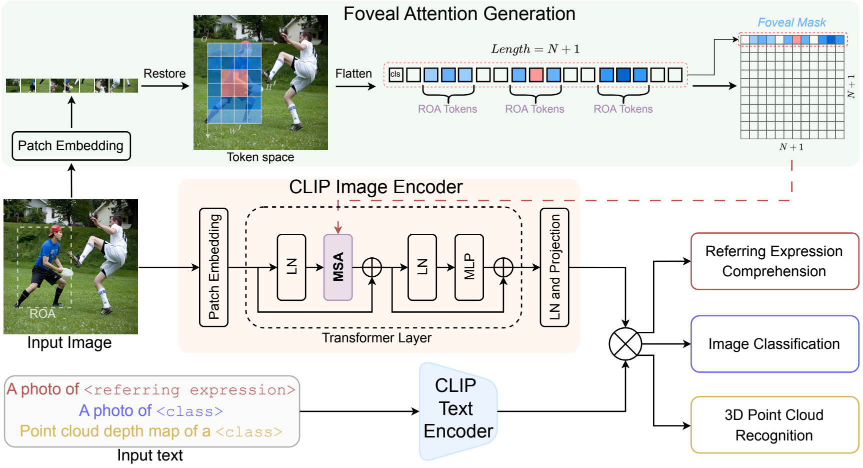


Fig. 3: FALIP Overview. We first input the image into the *foveal attention generation* module to obtain a foveal attention mask. Then, we input original images to the CLIP image encoder, while also providing the foveal attention mask to the Multi-head Self-Attention (MSA) module. With different input images and text prompts, the model can accomplish tasks such as referring expression comprehension, image classification and 3D point cloud recognition.

indicated in Fig. 2 by the brightness of tokens in the multi-head self-attention. This observation suggests that visual prompts significantly influence the model's focus on specific image areas. Based on our finding, we propose a hypothesis that the effectiveness of visual prompts can be fundamentally attributed to their ability to alter the model's attention. Fig. 3 presents the overall framework of our method.

3.2 Foveal Attention

In a typical ViT network, an input image $\mathbf{x} \in \mathbb{R}^{C\times H\times W}$ will be processed as N tokens $x_1, x_2, x_3, \cdots, x_n$. Denoting x_{cls} as the [CLS] token, the input of the transformer layer can be represented as: $X = \{x_{cls}, x_1, x_2, x_3, \cdots, x_n\} \in \mathbb{R}^{(N+1)\times D}$. Our method takes both an image $\mathbf{x} \in \mathbb{R}^{C\times H\times W}$ and its corresponding attention mask $M \in \mathbb{R}^{(N+1)\times(N+1)}$ as inputs. Removing the [CLS] token and restoring the spatial position of X, we can identify the tokens that are originally located on the region of attention (ROA). We represent these tokens as $TOKEN_{roa} = \{x_n | x_n$ located on the ROA$\}$.

To generate foveal attention mask for ROA, we first compute:

$$R_{i,j} = e^{-\frac{[i-(H'-1)/2]^2+[j-(W'-1)/2]^2}{2\sigma^2}} \tag{1}$$

$$R^{norm} = \alpha \times \frac{R - \mathrm{Min}(R) + \epsilon}{\mathrm{Max}(R) - \mathrm{Min}(R) + \epsilon} \tag{2}$$

where σ and α are adjustable parameters, ϵ is a small constant, $H' \in [1, \sqrt{N}]$, $W' \in [1, \sqrt{N}]$ are height and width of ROA in token space. Flattening R^{norm} and aligning its indices with X, the formula for M is given as follows:

$$M_{i,j} = \begin{cases} R_j^{norm} & x_j \in TOKEN_{roa} \quad and \quad i = 0 \\ 0 & x_j \notin TOKEN_{roa} \quad and \quad i = 0 \\ 0 & i > 0 \end{cases} \tag{3}$$

only the first row of M is assigned non-zero value, we will discuss it in Sec. 4.5. The formula for foveal attention is as follows:

$$\text{Foveal-Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}} + M\right)V \tag{4}$$

The design of is inspired by the foveal characteristics in human visual attention. We introduce a gradual blending of the foveal mask to promote smooth transitions between focal regions and surrounding backgrounds. The mask assigns *Gaussian-weighted* coefficients to the attentive tokens, mitigating the interference between background elements and focal regions.

3.3 Applications

Now that we have introduced the main principle of FALIP, we proceed to deploy the augmented model on several zero-shot tasks and discuss detailed considerations specific to each task.

Referring Expression Comprehension. Referring expression comprehension (REC) involves identifying an object in an image based on a textual description that explicitly refers to it. The entire process in a zero-shot manner can be represented as follows. The data for pre-processing includes an image $\mathbf{x} \in \mathbb{R}^{C \times H \times W}$, B boxes and a text t. Based on the aforementioned conclusions, B boxes can be transformed into masks $M^* \in \mathbb{R}^{B \times (N+1)^2} = \{M_1, M_2, M_3, \cdots, M_L\}$. The similarity between a text and a box region in the image and can be represented as follows: $S_i = \mathbf{T}(t) \cdot \mathbf{V}^\top(\mathbf{x}, M_i) \quad i \in [1, B]$, where " $\cdot$ " represents matrix multiplication. Similar to previous work [40], the "subtract" operation is utilized in the post-processing step to weigh down S_i. The best matching mask (box region) M_k to t is given by: $k = \underset{i}{\text{argmax}}\left[S_i - \frac{1}{Q}\sum_{q=1}^{Q}\mathbf{T}(\hat{t}_q) \cdot \mathbf{V}^\top(\mathbf{x}, M_i)\right] \quad \hat{t} \in \hat{T}$, where $\hat{T} = \{\hat{t}_1, \hat{t}_2, \hat{t}_3 \ \cdot \cdot\cdot, \hat{t}_Q\}$. $\hat{T}$ is abtained by randomly sampling Q negative captions related to no instances on the image from the whole dataset.

Image Classification. In this application scenario, the inputs to FALIP include $\mathbf{x} \in \mathbb{R}^{C \times H \times W}$, texts $\widetilde{T} = \{\tilde{t_1}, \tilde{t_2}, \tilde{t_3}, \cdots, \tilde{t_c}\}$, and boxes. Transforming boxes to M the entire classification process formulation is as follows: $Score_i = \mathbf{V}(\mathbf{x}, M) \cdot \mathbf{T}^\top(\tilde{t_i}) \quad i \in [1, c]$, $Pred = \underset{i}{\text{argmax}}\left[\frac{e^{Score_i}}{\sum_{i=1}^{c} e^{Score_i}}\right]$, where $Pred$ is the index of the text corresponding to the image category in $\widetilde{T}$. This task differs from REC, as there is only one mask as input in image classification task.

3D Point Cloud Recognition. CLIP can be deployed for 3D point cloud recognition [62] by projecting a 3D point cloud into six views of 2D depth maps $\overline{\mathbf{x}} \in \mathbb{R}^{6\times C\times H\times W}$. We locate the foreground positions in the depth maps and convert them into $M^* \in \mathbb{R}^{6\times(N+1)^2} = \{M_1, M_2, \cdots, M_6\}$. The texts of the category is $\overline{T} = \{\bar{t}_1, \bar{t}_2, \bar{t}_3, \cdots, \bar{t}_c\}$. The recognition process with FALIP is as follows: $Score_i = \sum_{j=1}^{6} \beta_j \mathbf{V}(\overline{\mathbf{x}}_j, M_j)\cdot\mathbf{T}^\top(\bar{t}_i) \quad i \in [1, c]$, $Pred = \underset{i}{\mathrm{argmax}} \left[\frac{e^{Score_i}}{\sum_{i=1}^{c} e^{Score_i}}\right]$, where β is used to control the weights of views. $Pred$ is the index of the text corresponding to the image category in $\overline{T}$.

4 Experiments

In this section, we begin by comparing our method with existing visual prompt methods on the referring expression comprehension task. Lastly, we extend our method to other tasks like image classification and 3D point cloud recognition, showcasing its superiority. Finally, we present our observations on the visual prompts and introduce the ablation experiments of FALIP. Unless otherwise specified, our experiments are conducted using the OpenAI version of the ViT/B-16 CLIP model. All experiments are performed on two RTX 3090 GPUs. For more experimental details, please refer to the Appendix.

4.1 Referring Expression Comprehension

We conduct the REC task on the RefCOCO [58], RefCOCO+ [58], and RefCOCOg [33] datasets. RefCOCO+ focuses on appearance-based expressions, while RefCOCO and RefCOCOg include relation-based expressions. The test sets are divided into two subsets: TestA (expressions referring to people) and TestB (expressions referring to non-people objects).

In previous works [6,57], some methods first extract proposals from the image using object detectors or instance detectors [17, 36], and then score the matching degree between these bounding boxes and the given text. To simplify this process and alleviate the need for explicit proposals, methods such as ViLT [23] and other methods [12, 20, 27] adopt an end-to-end training approach to predict a bounding box corresponding to the referring expression. We compare our method FALIP with previous zero-shot methods [40, 41, 56] in Tab. 1. Except for CPT [56] using VinVL [60] model, all others use the ViT-B model. We conducte experiments using two bounding box settings "prop" and "gold", which respectively represent the proposals generated by the MAttNet [57] and the annotations from the datasets. The results under the "gold" setting are generally better than those under the "prop" setting, indicating that using a more powerful detector or applying filtering to the proposals can lead to improved zero-shot accuracy. Our method outperforms existing methods in setting "Without E and P". Additionally, it can be combined with existing methods to further enhance their performance. When "subtract" post-progressing is utilized, FALIP shows a significant improvement in accuracy, but it has minimal impact on the performance for the RedCircle. FALIP maintains competitive performance when used

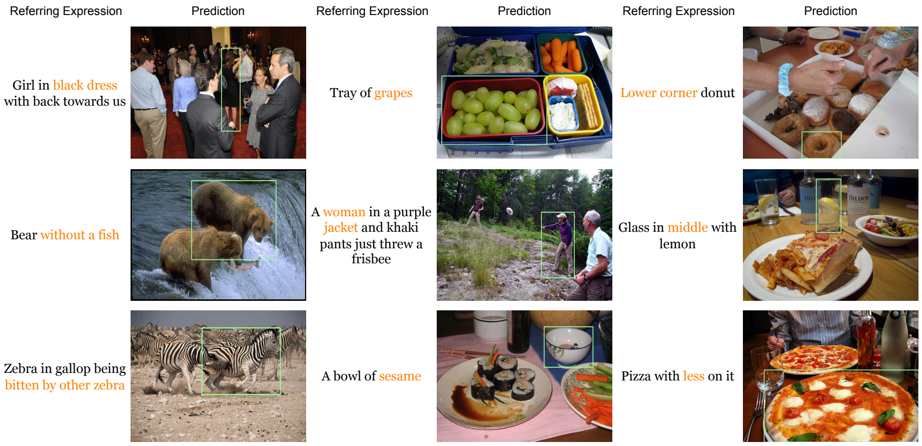


Fig. 4: Visualization of referring expression comprehension. The model predicts the corresponding object in the image based on the given referring expression. The key words in referring expression is colored orange.

in conjunction with ensemble and post-processing, surpassing existing methods by approximately 3%. The results demonstrate that our method is effective in various settings. Visualization results can be seen in Fig. 4.

4.2 Image Classification

For classification, we use the StanfordDogs [22], CUB-200-2011 [48], Waterbirds [37], and ImageNet-S [14] datasets. StanfordDogs and CUB-200-2011 consist of images of 120 different dog breeds and 200 bird species, respectively. Waterbirds contains photographs of waterbirds and landbirds, with bird images from the CUB dataset and backgrounds from the Places dataset [65]. Notably, Waterbirds is a binary classification dataset. ImageNet-S includes 919 classes with semantic segmentation annotations, selected from ImageNet-1k [9].

We compare our method FALIP with previous visual prompt and original CLIP in classification task in Tab. 2. On the four classification datasets, FALIP improves classification accuracy, while the visual prompts RedCircle and Blur respectively leads to slight and significant decreases in accuracy. RedCircle can be seen as contamination to the original fine-grained features. Blur blurs out a significant portion of the background and some foreground subject features. This results in a significant accuracy decrease across the first three datasets. However, in the Waterbirds dataset, where the classification decision relies on the background, Blur eliminates the interfering factor, resulting in only a slight accuracy decrease. We believe that the failure of the visual prompt is due to the contamination introduced by altering the image content, which impacts the fine-grained image classification performance negatively.

Table 1: Results on Referring Expressions Comprehension. "FA": Foveal-attention. "P": Post-progressing. "E": Ensemble-prompt. "‡": Results from the original paper. Our method is effective in this task. The best results are in **bold**.

Method	RefCOCO			RefCOCO+			RefCOCOg	
	TestA	TestB	Val	TestA	TestB	Val	Test	Val
Without E and P								
CPT [56]‡$_{prop}$	36.1	30.3	32.2	35.2	28.8	31.9	36.5	36.7
RedCircle [40]$_{prop}$	38.8	30.5	34.9	41.7	31.9	37.7	39.7	39.7
PASTA [61]$_{prop}$	39.3	32.7	36.3	40.4	36.2	38.4	43.8	43.8
RedCircle+FA$_{prop}$	40.7	31.7	35.9	43.6	34.0	39.2	41.0	41.6
FALIP(Ours)$_{prop}$	41.4	33.2	37.5	44.4	37.6	40.3	45.4	45.6
RedCircle [40]$_{gold}$	41.6	36.2	38.2	44.7	37.7	41.1	45.4	45.7
PASTA [61]$_{gold}$	41.7	37.6	39.5	43.2	40.5	42.4	49.2	49.9
RedCircle+FA$_{gold}$	41.8	36.9	38.2	45.1	38.4	41.6	46.1	46.1
FALIP(Ours)$_{gold}$	**44.2**	**39.4**	**40.8**	**46.8**	**43.1**	**44.5**	**51.5**	**51.3**
With P								
RedCircle$_{prop}$	34.3	30.3	33.8	36.8	31.0	36.3	39.1	39.2
FALIP(Ours)$_{prop}$	49.0	39.1	44.7	52.5	43.0	48.2	51.3	51.6
RedCircle$_{gold}$	39.2	37.7	39.5	42.8	39.7	42.2	44.9	45.3
FALIP(Ours)$_{gold}$	**50.6**	**44.8**	**48.3**	**54.5**	**48.6**	**52.0**	**57.1**	**56.9**
With E and P								
ReCLIP [41]$_{prop}$	42.4	44.4	42.1	42.8	42.3	41.9	56.9	57.1
RedCircle‡$_{prop}$	52.7	36.5	45.3	57.7	40.6	49.4	53.3	53.7
FALIP(Ours)$_{prop}$	51.7	38.3	46.7	57.1	43.0	51.9	54.9	54.2
RedCircle$_{gold}$	**53.5**	41.7	46.9	58.5	45.5	52.3	57.3	56.9
FALIP(Ours)$_{gold}$	53.4	**44.6**	**49.8**	**59.4**	**49.5**	**55.1**	**60.7**	**59.3**

4.3 3D Point Cloud Recognition

For 3D point cloud recognition, we use the ModelNet40 [50] and ScanObjectNN [46] datasets. ModelNet40 has 40 object categories, encompassing common objects like chairs and airplanes. ScanObjectNN comprises real-world 3D scan data with 15 categories of household objects such as tables and lamps.

We extend the proposed FALIP to 3D point cloud recognition. In the experiments, we use PointCLIP [62] as the baseline, which employs the CLIP as the image encoder. The details are shown in Fig. 5. The experimental results are presented in Tab. 3. FALIP outperforms the original CLIP on both datasets, achieving an average accuracy improvement of 1.4%. This encouraging result demonstrates FALIP's potential to extend to 3D data domains.

4.4 Unleash Visual Prompts

To investigate the impact of visual prompts on each attention head, we decouple the output of the image encoder $\mathbf{V}$ into the sum of the individual attention heads. The specific steps are as follows: $X_l' = \text{MSA}[\text{LN}(X_{l-1})] + X_{l-1}$,

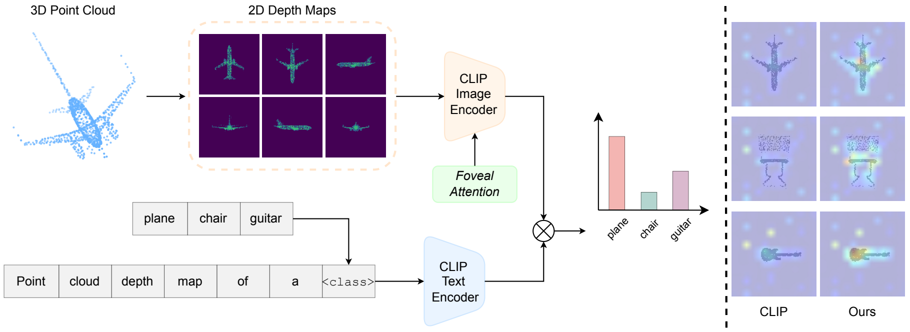


Fig. 5: Pipeline of 3D point cloud recognition. ***Left***: The overall framework remains consistent with PointCLIP [62], with the difference being the insertion of foveal attention in the CLIP image encoder. ***Right***: Attention on the 2D depth maps of original CLIP and our method. It can be observed that our method shows a stronger attention towards the foreground.

Table 2: Results on Image Classification. "Blur" refers to applying blur operation to the areas outside the circle. The superiority of our method lies in preserving the original fine-grained features of the image. The best results are in **bold**, and sub-optimal results are underlined.

Method	StanfordDogs		CUB-200-2011		ImageNet-S		Waterbirds
	Top1	Top5	Top1	Top5	Top1	Top5	Top1
Original CLIP	56.5	85.2	54.2	**83.7**	64.9	88.4	78.2
RedCircle	52.4	82.8	44.2	77.0	62.8	86.5	77.5
Blur	51.9	81.9	39.1	71.0	53.8	77.6	78.1
FALIP(Ours)	**58.3**	**86.0**	**54.3**	83.6	**67.3**	**89.9**	**79.7**

$X_l = \mathrm{MLP}[\mathrm{LN}(X'_l)] + X'_l$, where $l \in [1, L]$, L represents the total layers of model. MSA is multi-head self-attention. LN is LayerNorm operation, X_l is the output of the l-th layer. Denoting $[X_l]_{cls}$ as [CLS] token in X_l, the output of $\mathbf{V}$ can be expressed as:

$$[X_L]_{cls} = [X_0]_{cls} + \sum_{l=1}^{L}\Big[\mathrm{MSA}[\mathrm{LN}(X_{l-1})]\Big]_{cls} + \sum_{l=1}^{L}\Big[\mathrm{MLP}[\mathrm{LN}(X'_l)]\Big]_{cls} \qquad (5)$$

Here we omit the final project layer, and decouple the final output into the sum of MSA and MLP. Further decomposition of the second term in Eq. (5): $\Big[\mathrm{MSA}[\mathrm{LN}(X_{l-1})]\Big]_{cls} = \sum_{h=1}^{H} \sum_{i=1}^{N+1} p_{i,h} \quad p_{i,h} = \gamma_i^h \mathrm{LN}(x_i^{l-1}) W_V^h$, where H is number of attention heads, $N+1$ is the number of input tokens. γ_i^h is the attention weights between [CLS] token and i-th token. x_i^{l-1} is i-th token in X^{l-1}. W_V^h is a mapping matrix of V. Let $G_h = \sum_{i=1}^{N+1} p_{i,h}$, the change of MSA [CLS] token after using visual prompt is: $\Delta = \Big[\mathrm{MSA}[\mathrm{LN}(X_{l-1})]\Big]'_{cls} - \Big[\mathrm{MSA}[\mathrm{LN}(X_{l-1})]\Big]_{cls} =$

Table 3: Results on 3D point cloud recognition. Our method improves the recognition capability of CLIP. The best results are in **bold**.

Method	ModelNet40	ScanObjectNN	Avg
Original CLIP	16.5	14.6	15.6
FALIP(Ours)	**18.6**	**15.3**	**17.0**

Table 4: Ablation on unleashing of viusal prompt on REC task. "R": RedCircle. "B": Blur. "U": Unleash. Adjusting salient attention heads increases the concentration of the model's attention, since unleashed visual prompt outperforms original visual prompt on all metrics. Details are in Sec. 4.4. The best results are in **bold**.

R	B	U	RefCOCO			RefCOCO+			RefCOCOg		Avg
			TestA	TestB	Val	TestA	TestB	Val	Test	Val	
✓			41.6	36.2	38.2	44.7	37.7	41.1	45.4	45.7	41.3
✓		✓	**46.1**	**39.4**	**42.0**	**50.1**	**40.3**	**44.8**	**49.9**	**49.4**	**45.2**
✓	✓		45.4	41.5	43.8	49.3	44.9	46.5	56.6	56.9	48.3
✓	✓	✓	**49.0**	**41.6**	**45.7**	**54.6**	**45.1**	**49.7**	**56.7**	**57.0**	**49.9**

$\sum_{h=1}^{H}(G'_h - G_h)$. Thus, we can easily observe the changes in individual attention heads across different layers before and after using the visual prompt.

We compute the changes in individual attention heads before and after using RedCircle on a large number of images and find that the attention heads in the last 4 layers of the model exhibited significant variations (shown in Fig. 6). Therefore, we propose generating a new [CLS] token by editing these self-attention heads. The formula for the new MSA [CLS] token is as follows:

$$\Big[\text{MSA}[\text{LN}(X_{l-1})]\Big]_{cls} = \sum_{h=1}^{H}[G'_h + (G'_h - G_h)] \qquad l \in [L-3, L] \tag{6}$$

By substituting Eq. (6) into Eq. (5), we can obtain the new output $[X_L]'_{cls}$ of the model. We test the generated new output on the REC task in Tab. 4. After unleashing the potential of RedCircle, the average accuracy has increased by more than 4%. By applying the same method to RedCircle+Blur, an improvement in performance can be seen as well. This phenomenon suggests that the potential of the current visual prompt has not been fully explored. Furthermore, we discover that some attention heads in CLIP are particularly sensitive to RedCircle. Fig. 6 demonstrates some visualizations to explain these.

4.5 Ablation on Foveal Attention

Other implementations. Fig. 7a shows feature mask method, the results can be seem in Tab. 5. The accuracy of feature mask is significantly lower compared

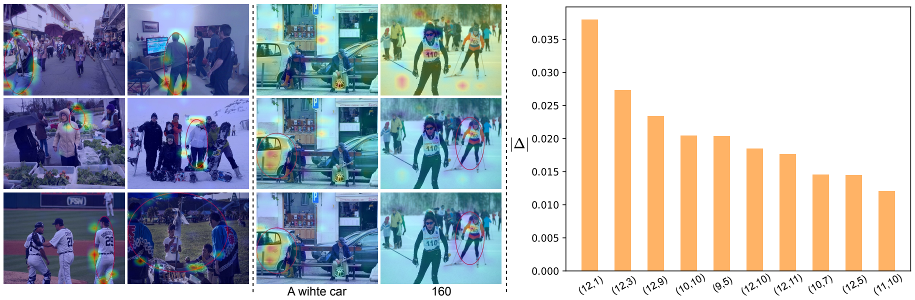


Fig. 6: Visualization of attention. The attention maps are generated by [13]. ***Left***: Attention maps of Layer9-Head5 in CLIP image encoder for various images. This attention head shows sensitivity towards RedCircles of varying positions and sizes. ***Middle***: From top to bottom, are the original image, the image with RedCircle, and the image with unleashed RedCircle. The attention maps for the text input show a gradual decrease in attention on irrelevant backgrounds. ***Right***: Ranking effect on attention head labeled by (Layer, Head).

Table 5: Ablation on different q, k, v in self-attention. The results in the second row and the third row correspond to the methods illustrated in Fig. 7b and Fig. 7c, respectively. The best results are in **bold**.

Method	RefCOCO			RefCOCO+			RefCOCOg		Avg
	TestA	TestB	Val	TestA	TestB	Val	Test	Val	
RedCircle	41.6	36.2	38.2	44.7	37.7	41.1	45.4	45.7	41.3
Replace v	38.5	35.1	36.2	40.6	37.1	38.3	43.4	42.6	39.0
Replace q, k	39.9	32.9	35.9	42.1	35.3	38.3	41.7	42.4	38.6
Feature mask	25.4	26.4	25.6	24.9	27.7	26.2	29.8	30.0	27.0
FALIP(Ours)	**44.2**	**39.4**	**40.8**	**46.8**	**43.1**	**44.5**	**51.5**	**51.3**	**45.2**

to our method. Fig. 7b and Fig. 7c illustrate two methods for performing self-attention by replacing q, k, and v. We observe that regardless of whether q, k, or v is replaced, the accuracy of the RedCircle decreases. This suggests that the q, k, and v generated from the images with trained visual prompts have a strong correlation. The details is shown in Tab. 5.

Forms of masks. As in Fig. 8, we use three different forms of masks. The results in Tab. 6 indicate that *Method a* achieve the highest accuracy. *Method b* causes all other tokens to pay excessive attention to the specified region, disrupting the original information carried by these tokens. On the other hand, *Method c* causes an excessive focus on the tokens within the specified region, neglecting the contextual information. This suggests the row corresponding to [CLS] token plays a crucial role in the model's prediction outcomes.

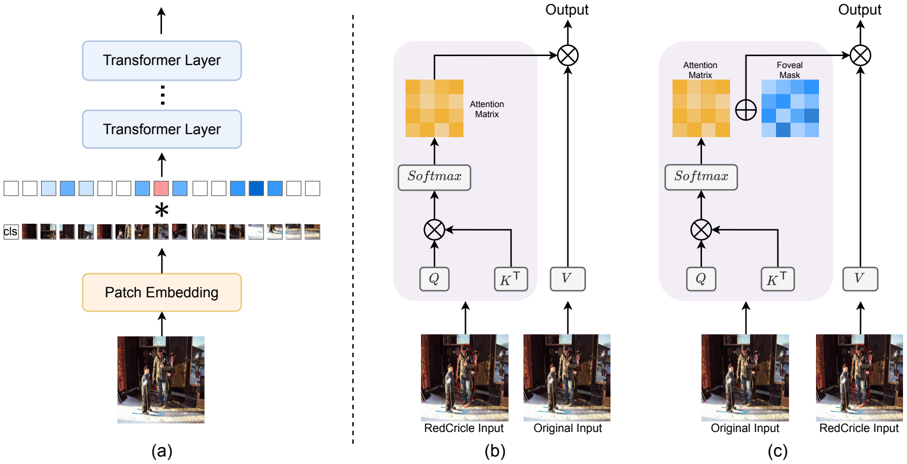


Fig. 7: Illustration of baseline methods in Table 5. (a) Feature mask. "$*$" means element-wise multiplication. (b) Self-attention using the q, k generated from the RedCircle image and the v from the original image. (c) Self-attention using q, k generated from the original image and v from the RedCircle image with foveal attention mask.

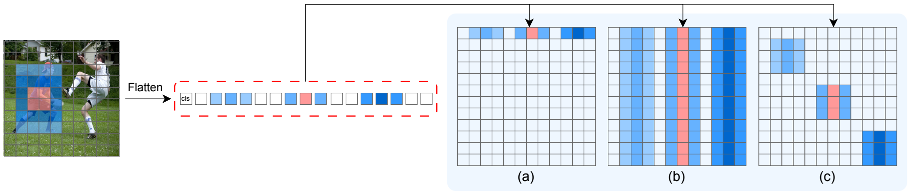


Fig. 8: Several ways to generate foveal attention masks. (a) Only assigning values at the position corresponding to the ROA token in the first row. (b) Assigning values of ROA token position in all rows. (c) Assigning values at the position corresponding to the ROA token on the diagonal line.

Value of α and σ. We conduct ablation experiments on the value of α, σ defined in Sec. 3.2 and present results in Fig. 9.

5 Conclusion

This paper presents an exploration into the surprising effectiveness of visual prompts for CLIP. We discover a close relationship between visual prompts and the attention mechanism within CLIP. Motivated by this finding, we propose FALIP, which enhances the region-awareness capability of CLIP by mimicking human attention characteristics, without incurring additional model fine-tuning or sacrificing its pre-trained knowledge. Our method achieves state-of-the-art performance on the referring expression comprehension task and demonstrates significant improvements on tasks like image classification and 3D point cloud

Table 6: Ablation on the forms of masks. "No mask" means original CLIP. "Method a" is the optimal form of the mask, which preserve the original information carried by tokens, and reassign the weights of all tokens with respect to the [CLS] token. The best results are in **bold**.

Method	RefCOCO			RefCOCO+			RefCOCOg		Avg
	TestA	TestB	Val	TestA	TestB	Val	Test	Val	
No mask	14.8	25.5	19.5	14.7	26.1	19.8	25.5	26.3	21.5
Method a	**44.2**	**39.4**	**40.8**	**46.8**	**43.1**	**44.5**	**51.5**	**51.3**	**45.2**
Method b	36.6	37.2	35.1	39.1	40.3	37.8	43.7	44.2	39.3
Method c	13.3	19.0	15.9	12.5	19.0	15.1	16.1	15.6	15.8

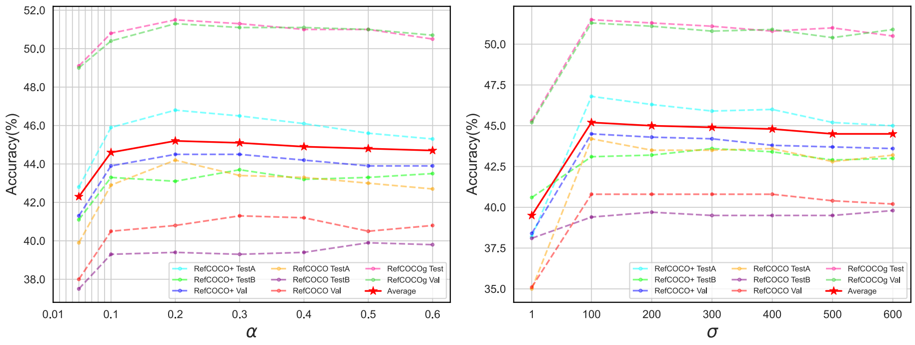


Fig. 9: Effect of α and σ in masks. The accuracy peaks when $\alpha = 0.2$, as it highlights the features of the specific object while preserving the contextual information. When $\sigma = 1$, the concentration of values within a mask can lead to a lack of rich features in specific regions. We consider $\alpha = 0.2$ and $\sigma = 100$ as the optimal value.

recognition. Furthermore, we discover that the full potential of visual prompts can be further unleashed by adjusting the salient attention heads. We hope this work can provide inspiration for the understanding or design of visual prompts and attention mechanisms, sparking greater efforts from the research community dedicated to enhancing our understanding of intriguing phenomena exhibited by AI models.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. U21B2004) and the Zhejiang Provincial Key RD Program of China (Grant No. 2021C01119).

References

1. Bahng, H., Jahanian, A., Sankaranarayanan, S., Isola, P.: Exploring visual prompts for adapting large-scale models. arXiv preprint arXiv:2203.17274 (2022)
2. Burt, R., Thigpen, N.N., Keil, A., Principe, J.C.: Unsupervised foveal vision neural networks with top-down attention. arXiv preprint arXiv:2010.09103 (2020)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
4. Chen, K., Jiang, X., Hu, Y., Tang, X., Gao, Y., Chen, J., Xie, W.: Ovarnet: Towards open-vocabulary object attribute recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23518–23527 (2023)
5. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. in 2021 ieee. In: CVF International Conference on Computer Vision (ICCV). pp. 9620–9629
6. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: European conference on computer vision. pp. 104–120. Springer (2020)
7. Chowdhury, S., Nag, S., Manocha, D.: Apollo: Unified adapter and prompt learning for vision language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 10173–10187 (2023)
8. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. arXiv preprint arXiv:2309.16588 (2023)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
10. Ding, Z., Wang, J., Tu, Z.: Open-vocabulary panoptic segmentation with maskclip. arXiv preprint arXiv:2208.08984 (2022)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
12. Dou, Z.Y., Kamath, A., Gan, Z., Zhang, P., Wang, J., Li, L., Liu, Z., Liu, C., LeCun, Y., Peng, N., et al.: Coarse-to-fine vision-language pre-training with fusion in the backbone. Advances in neural information processing systems **35**, 32942–32956 (2022)
13. Gandelsman, Y., Efros, A.A., Steinhardt, J.: Interpreting clip's image representation via text-based decomposition. arXiv preprint arXiv:2310.05916 (2023)
14. Gao, S., Li, Z.Y., Yang, M.H., Cheng, M.M., Han, J., Torr, P.: Large-scale unsupervised semantic segmentation. IEEE transactions on pattern analysis and machine intelligence (2022)
15. Guo, Z., Dong, B., Ji, Z., Bai, J., Guo, Y., Zuo, W.: Texts as images in prompt tuning for multi-label image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2808–2817 (2023)
16. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
17. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)

18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
19. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European Conference on Computer Vision. pp. 709–727. Springer (2022)
20. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr-modulated detection for end-to-end multi-modal understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1780–1790 (2021)
21. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19113–19122 (2023)
22. Khosla, A., Jayadevaprakash, N., Yao, B., Li, F.F.: Novel dataset for fine-grained image categorization: Stanford dogs. In: Proc. CVPR workshop on fine-grained visual categorization (FGVC). vol. 2. Citeseer (2011)
23. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning. pp. 5583–5594. PMLR (2021)
24. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
25. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
26. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Proceedings of the 39th International Conference on Machine Learning. pp. 12888–12900 (2022)
27. Li, M., Sigal, L.: Referring transformer: A one-step approach to multi-task visual grounding. Advances in neural information processing systems **34**, 19652–19664 (2021)
28. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190 (2021)
29. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7061–7070 (2023)
30. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)
31. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
32. Liu, X., Ji, K., Fu, Y., Tam, W., Du, Z., Yang, Z., Tang, J.: P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 61–68 (2022)
33. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 11–20 (2016)
34. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)

35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
36. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
37. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. arXiv preprint arXiv:1911.08731 (2019)
38. Shen, S., Yang, S., Zhang, T., Zhai, B., Gonzalez, J.E., Keutzer, K., Darrell, T.: Multitask vision-language prompt tuning. arXiv preprint arXiv:2211.11720 (2022)
39. Shi, B., Zhao, P., Wang, Z., Zhang, Y., Wang, Y., Li, J., Dai, W., Zou, J., Xiong, H., Tian, Q., et al.: Umg-clip: A unified multi-granularity vision generalist for open-world understanding. arXiv preprint arXiv:2401.06397 (2024)
40. Shtedritski, A., Rupprecht, C., Vedaldi, A.: What does clip know about a red circle? visual prompt engineering for vlms. arXiv preprint arXiv:2304.06712 (2023)
41. Subramanian, S., Merrill, W., Darrell, T., Gardner, M., Singh, S., Rohrbach, A.: Reclip: A strong zero-shot baseline for referring expression comprehension. arXiv preprint arXiv:2204.05991 (2022)
42. Sun, Z., Fang, Y., Wu, T., Zhang, P., Zang, Y., Kong, S., Xiong, Y., Lin, D., Wang, J.: Alpha-clip: A clip model focusing on wherever you want. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13019–13029 (2024)
43. Tewel, Y., Shalev, Y., Schwartz, I., Wolf, L.: Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17918–17928 (2022)
44. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
45. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
46. Uy, M.A., Pham, Q.H., Hua, B.S., Nguyen, T., Yeung, S.K.: Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1588–1597 (2019)
47. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
48. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
49. Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T.: Cris: Clip-driven referring image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11686–11695 (2022)
50. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1912–1920 (2015)
51. Xia, Y., Kim, J., Canny, J., Zipser, K., Canas-Bajo, T., Whitney, D.: Periphery-fovea multi-resolution driving model guided by human attention. In: Proceedings

of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1767–1775 (2020)

52. Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., De Mello, S.: Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2955–2966 (2023)
53. Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X.: Side adapter network for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2945–2954 (2023)
54. Xu, X., Xiong, T., Ding, Z., Tu, Z.: Masqclip for open-vocabulary universal image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 887–898 (2023)
55. Yang, L., Wang, Y., Li, X., Wang, X., Yang, J.: Fine-grained visual prompting. arXiv preprint arXiv:2306.04356 (2023)
56. Yao, Y., Zhang, A., Zhang, Z., Liu, Z., Chua, T.S., Sun, M.: Cpt: Colorful prompt tuning for pre-trained vision-language models. arXiv preprint arXiv:2109.11797 (2021)
57. Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., Berg, T.L.: Mattnet: Modular attention network for referring expression comprehension. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1307–1315 (2018)
58. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 69–85. Springer (2016)
59. Zang, Y., Li, W., Zhou, K., Huang, C., Loy, C.C.: Unified vision and language prompt learning. arXiv preprint arXiv:2210.07225 (2022)
60. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: Vinvl: Revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5579–5588 (2021)
61. Zhang, Q., Singh, C., Liu, L., Liu, X., Yu, B., Gao, J., Zhao, T.: Tell your model where to attend: Post-hoc attention steering for llms. arXiv preprint arXiv:2311.02262 (2023)
62. Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., Li, H.: Pointclip: Point cloud understanding by clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8552–8562 (2022)
63. Zhao, S., Zhang, Z., Schulter, S., Zhao, L., Vijay Kumar, B., Stathopoulos, A., Chandraker, M., Metaxas, D.N.: Exploiting unlabeled data with vision and language models for object detection. In: European Conference on Computer Vision. pp. 159–175. Springer (2022)
64. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16793–16803 (2022)
65. Zhou, B., Khosla, A., Lapedriza, A., Torralba, A., Oliva, A.: Places: An image database for deep scene understanding. arXiv preprint arXiv:1610.02055 (2016)
66. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16816–16825 (2022)

67. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)

Appendix

Datasets. Tab. 7, Tab. 8 and Tab. 9 provide a brief introduction to the datasets used for tasks referring expression comprehension, image classification and 3D cloud recognition, respectively.

Table 7: Referring expression comprehension datasets. "Refs" means the number of referring expressions.

	RefCOCO			RefCOCO+			RefCOCOg	
	TestA	TestB	Val	TestA	TestB	Val	Test	Val
Images	750	750	1,500	750	750	1,500	2,600	1,300
Refs	1,975	1,810	3,811	1,975	1,798	3,805	5,023	2,573

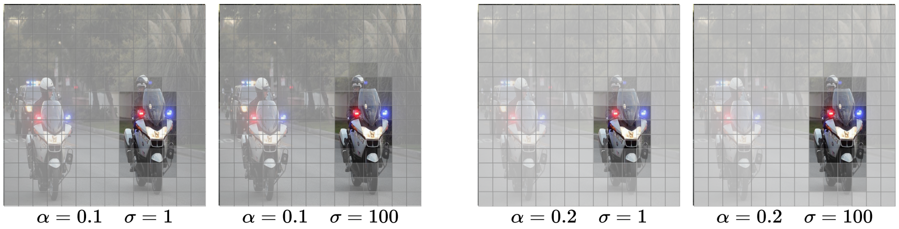
Table 8: Image Classification datasets. "Images used" means the number of images used in our experiments.

	StanfordDogs	CUB-200-2011	ImageNet-S	Waterbirds
Categories	120	200	919	2
Total Images	20,580	11,788	1,223,164	20,580
Images used	20,580	11,788	12,419	5,794

Table 9: 3D cloud recognition datasets. "Clouds used" means the number of clouds used in our experiments.

	ModelNet40	ScanObjectNN
Categories	40	15
Total Clouds	12,311	2,880
Clouds used	2,468	576

Referring Expression Comprehension. Tab. 10 and Tab. 11 present detailed experimental results about α and σ, respectively. We take $\alpha = 0.2$ and $\sigma = 100$ in final result. Fig. 10 illustrates the visual impact of different α and σ on the original image. To investigate the sensitivity of different layers in CLIP to masks, we insert masks at various layers and present results in Tab. 12. We find that inserting masks only in the last 4 layers results in the highest model



$\alpha = 0.1 \quad \sigma = 1$ $\alpha = 0.1 \quad \sigma = 100$ $\alpha = 0.2 \quad \sigma = 1$ $\alpha = 0.2 \quad \sigma = 100$

Fig. 10: Visualizing different values of α and σ on the original image. A large α enhance prominence of the specific region and a large σ preserve more content within the region.

accuracy, which suggests that the attention computations in the later layers play a decisive role in shaping the representation of the model's output, while the initial layers seem to have a minor impact on the results. Fig. 13 depicts the details of the ensemble and Fig. 14 shows the extensive results of referring expression comprehension.

Table 10: Ablation on α. The best results are in **bold**.

α	RefCOCO			RefCOCO+			RefCOCOg		Avg
	TestA	TestB	Val	TestA	TestB	Val	Test	Val	
0.05	39.9	37.5	38.0	42.8	41.2	41.3	49.1	48.8	42.3
0.1	42.9	39.3	40.5	45.9	43.3	43.9	50.8	50.4	44.6
0.2	**44.2**	39.4	40.8	**46.8**	43.1	**44.5**	**51.5**	**51.3**	**45.2**
0.3	43.4	39.3	**41.3**	46.5	**43.7**	44.5	51.3	51.1	45.1
0.4	43.3	39.4	41.2	46.1	43.2	44.2	51.0	51.1	44.9
0.5	43.0	**39.9**	40.5	45.6	43.3	44.0	51.0	51.0	44.8
0.6	42.7	39.8	40.8	45.3	43.5	44.0	50.5	50.7	44.7

Table 11: Ablation on σ. The best results are in **bold**.

σ	RefCOCO			RefCOCO+			RefCOCOg		Avg
	TestA	TestB	Val	TestA	TestB	Val	Test	Val	
1	35.0	38.1	35.1	38.2	40.6	38.4	45.3	45.2	39.5
100	**44.2**	39.4	**40.8**	**46.8**	43.1	**44.5**	**51.5**	**51.3**	**45.2**
200	43.5	39.7	40.8	46.3	43.2	44.3	51.3	51.1	45.0
300	43.5	39.5	40.8	45.9	**43.6**	44.2	51.0	50.8	44.9
400	43.6	39.5	40.8	46.0	43.4	43.8	50.8	50.9	44.9
500	42.8	39.5	40.4	45.2	42.9	43.7	51.0	50.4	44.5
600	43.2	**39.8**	40.2	45.1	43.1	43.6	50.5	50.9	44.5

Image Classification. The image classification experimental results are obtained from testing on the following datasets: entire StanfordDogs, entire CUB-200-2011, test of Waterbirds and validation of ImageNets, which are shown in Tab. 8. Fig. 11 shows the input image of various methods. Tab. 13 demonstrates

Table 12: Effect of which layer to insert masks. “1~4” means layers 1 to 4 are inserted a mask. “9~12” achieves highest performance. The attention in the later layers have a significant impact on shaping the output embedding. The best results are in **bold**.

Layer	RefCOCO			RefCOCO+			RefCOCOg		Avg
	TestA	TestB	Val	TestA	TestB	Val	Test	Val	
1	17.1	25.8	20.6	17.3	26.8	20.6	24.6	26.8	22.4
1~4	20.4	26.1	21.0	21.0	27.1	21.7	27.6	27.3	24.0
1~6	22.3	25.1	22.4	22.1	25.7	23.6	28.6	28.2	24.7
12	39.4	**40.0**	39.7	43.7	**43.8**	42.9	50.9	50.6	43.9
9~12	**44.2**	39.4	40.8	**46.8**	43.1	**44.5**	**51.5**	**51.3**	**45.2**
7~12	43.8	39.4	**41.3**	46.3	42.5	44.2	51.0	51.1	44.9

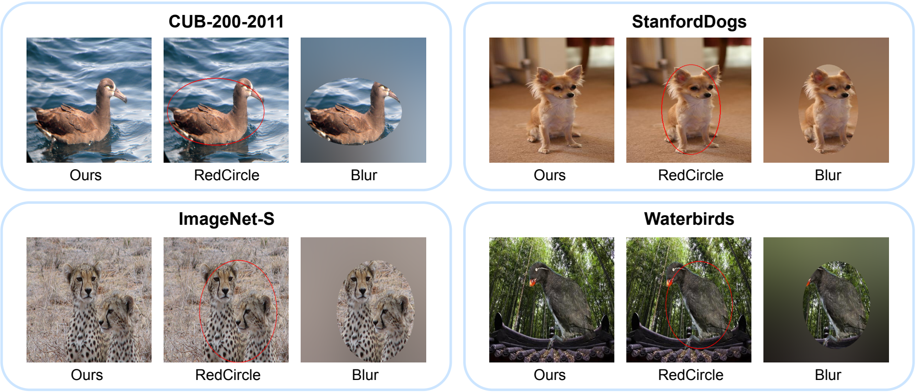


Fig. 11: Examples of input images in each dataset. For each dataset, from the left to right is the input image of model for our method, RedCircle and Blur respectively.

the performance of FALIP on the larger model Vit-L/14, showing an improvement over CLIP in terms of accuracy. Except for the Waterbirds, FALIP achieves the highest accuracy on all other datasets. Tab. 14 illustrates how accuracy is affected by visual prompt of varying sizes. Increasing the range of the RedCircle appropriately can lead to a certain improvement in accuracy. Fig. 13 provides a brief explanation of enlarging size of visual prompt (the maximum size will not exceed the inscribed circle of the image). In Fig. 15 we compare our method with CLIP on the model's attention.

Table 13: Method ablation on Image Classification. The best results are in **bold**, and sub-optimal results are underlined.

Method	Model	StanfordDogs		CUB-200-2011		ImageNet-S		Waterbirds
		Top1	Top5	Top1	Top5	Top1	Top5	Top1
Original CLIP	ViT-B	56.5	85.2	54.2	**83.7**	64.9	88.4	78.2
RedCircle	ViT-B	52.4	82.8	44.2	77.0	62.8	86.5	77.5
Blur	ViT-B	51.9	81.9	39.1	71.0	53.8	77.6	78.1
FALIP(Ours)	ViT-B	**58.3**	**86.0**	**54.3**	83.6	**67.3**	**89.9**	**79.7**
Original CLIP	ViT-L	65.4	89.1	61.4	90.1	72.0	91.1	83.3
RedCircle	ViT-L	63.7	88.6	56.1	87.5	70.9	90.6	80.7
Blur	ViT-L	60.1	85.4	46.1	82.8	63.6	84.2	**85.1**
FALIP(Ours)	ViT-L	**66.6**	**89.8**	**61.7**	**90.7**	**74.8**	**92.7**	84.5

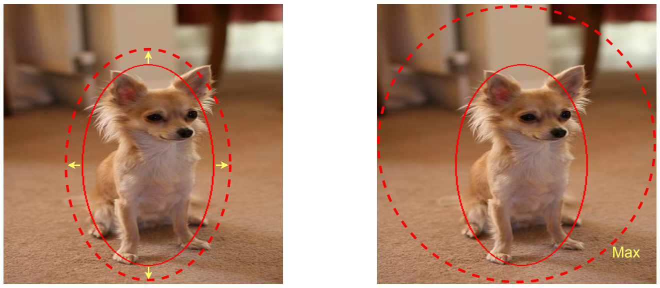


Fig. 12: Enlarge prompts. We increase the pixels in four directions. In this way, the contamination of foreground can be mitigated.

Table 14: Method ablation on size of RedCircle. The best results are in **bold**.

Enlarge Pixels	StanfordDogs		CUB-200-2011		ImageNet-S		Waterbirds
	Top1	Top5	Top1	Top5	Top1	Top5	Top1
0	52.4	82.8	44.2	77.0	62.8	86.5	77.5
5	51.8	81.8	43.2	76.0	63.2	87.2	77.6
10	52.4	82.1	43.8	76.4	63.6	87.3	77.7
20	52.7	82.4	45.6	77.3	**64.3**	87.7	78.0
30	53.1	82.4	46.5	78.0	64.2	**88.1**	78.4
40	**53.2**	82.6	47.1	78.6	64.1	87.9	78.7
50	53.0	**82.7**	46.9	78.8	63.9	87.6	78.7
100	52.9	82.5	47.6	**79.0**	62.6	86.7	78.6
150	52.8	82.4	47.7	78.7	61.8	86.6	78.7
200	52.8	82.4	**47.8**	78.9	61.7	86.2	**78.7**

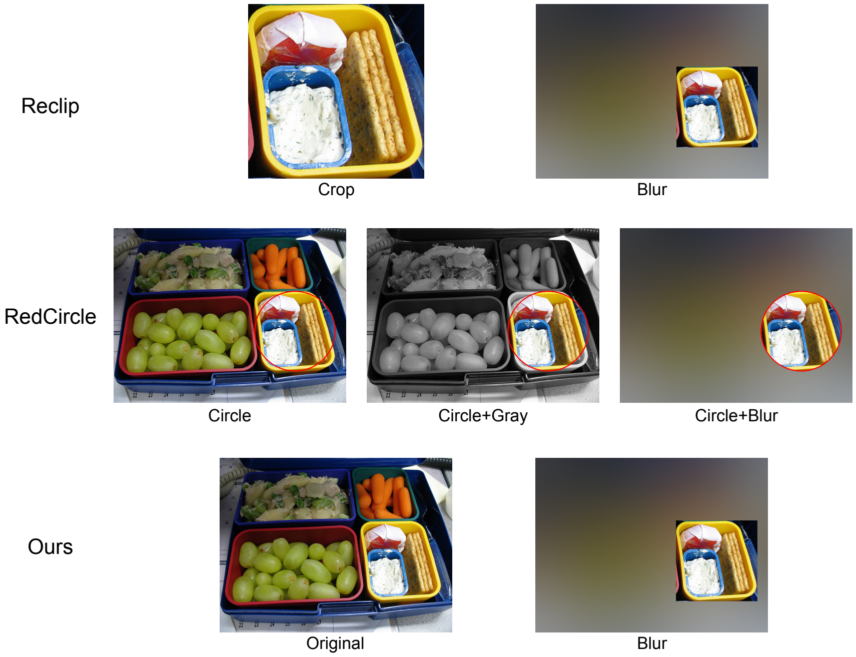


Fig. 13: The specific approaches for ensemble. To ensure a fair comparison, we also adopt the same Blur method used in the previous method.

Pesudo Code. The pesudo code of FALIP is shown in **Algorithm 1**.

Algorithm 1 Image Encoder of Foveal-Attention CLIP

Input: image x, bounding box box

Output: image feature f_v

```
function FALIP($x$, $box$)
    $x^* \leftarrow \text{Preprocess}(x)$
    $X \leftarrow \text{PatchEmbedding}(x^*)$        #Transform image to sequence, $X \in \mathbb{R}^{(N+1)\times D}$
    $T \leftarrow \text{BoxToToken}(x, box)$        #Transform box to token space
    $H, W \leftarrow T.height, T.wdith$
    $R \leftarrow \mathbb{0}^{H\times W}$        #Initialize with 0
    $M \leftarrow \mathbb{0}^{(N+1)\times(N+1)}$        #Initialize with 0, $N+1$ is length of the sequence
    for $i = 0$ to $(H-1)$ do
        for $j = 0$ to $(W-1)$ do
            $R[i][j] \leftarrow e^{-\frac{[i-(H-1)/2]^2+[j-(W-1)/2]^2}{2\sigma^2}}$        #Generate foveal value
        end for
    end for
    $R^{norm} \leftarrow \alpha \times \frac{R-\text{Min}(R)+\epsilon}{\text{Max}(R)-\text{Min}(R)+\epsilon}$        #Normalization
    $R^* \leftarrow \text{Flatten}(R^{norm})$        #Flatten and align indices with $X$
    $M[0] \leftarrow R^*$        #Assgin value to positions in the first row of $M$
    $X^* \leftarrow \text{LayerNorm}(X)$
    $f_v \leftarrow \text{Transformer}(X^*, M)$        #Input sequence and foveal attention mask
end function
```

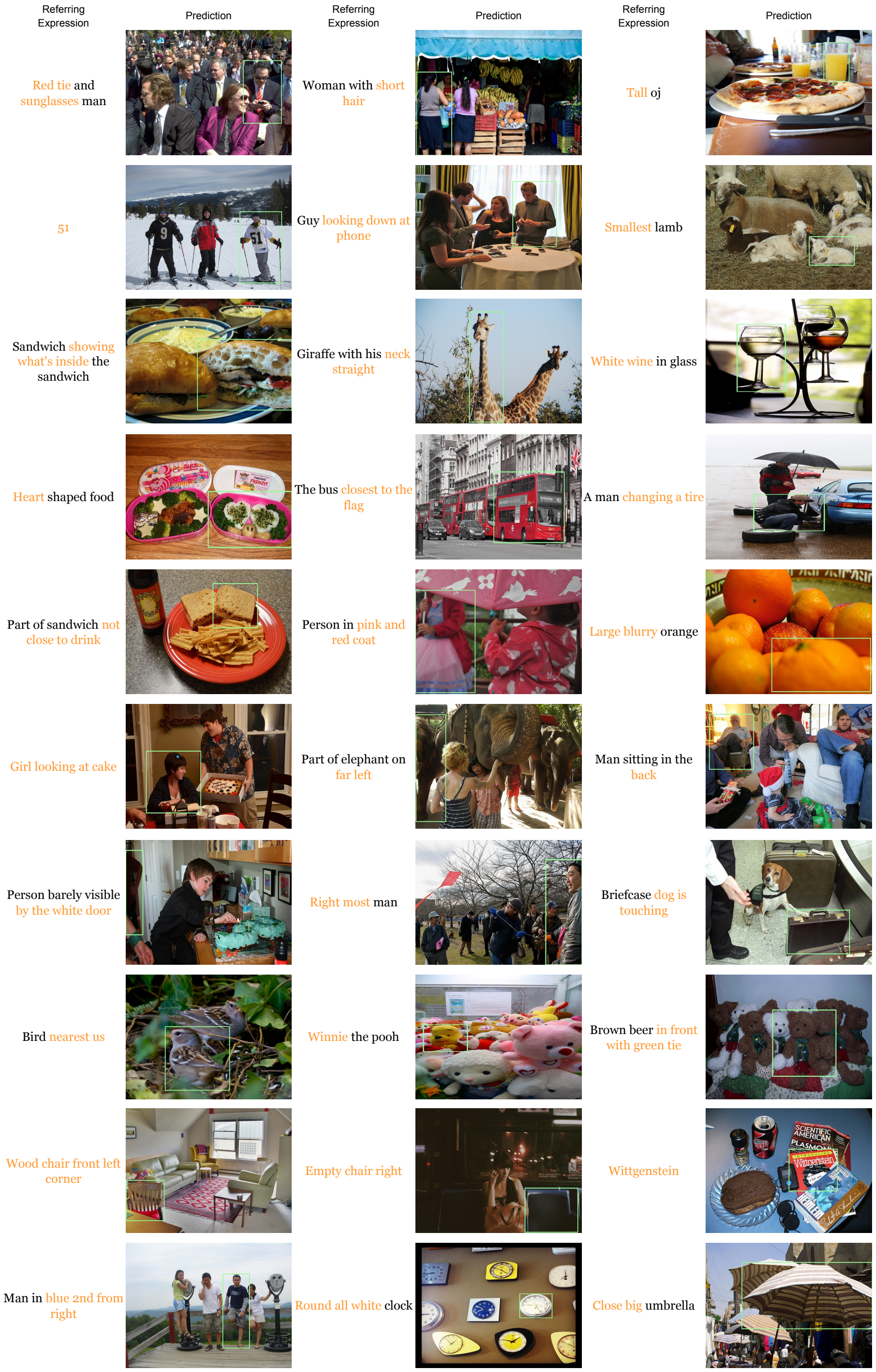

Referring Expression
Prediction
Referring Expression
Prediction
Referring Expression
Prediction
Red tie and sunglasses man
Woman with short hair
Tall oj
51
Guy looking down at phone
Smallest lamb
Sandwich showing what's inside the sandwich
Giraffe with his neck straight
White wine in glass
Heart shaped food
The bus closest to the flag
A man changing a tire
Part of sandwich not close to drink
Person in pink and red coat
Large blurry orange
Girl looking at cake
Part of elephant on far left
Man sitting in the back
Person barely visible by the white door
Right most man
Briefcase dog is touching
Bird nearest us
Winnie the pooh
Brown beer in front with green tie
Wood chair front left corner
Empty chair right
Wittgenstein
Man in blue 2nd from right
Round all white clock
Close big umbrella

Fig. 14: The visualization results of REC. The keywords are highlighted in orange.

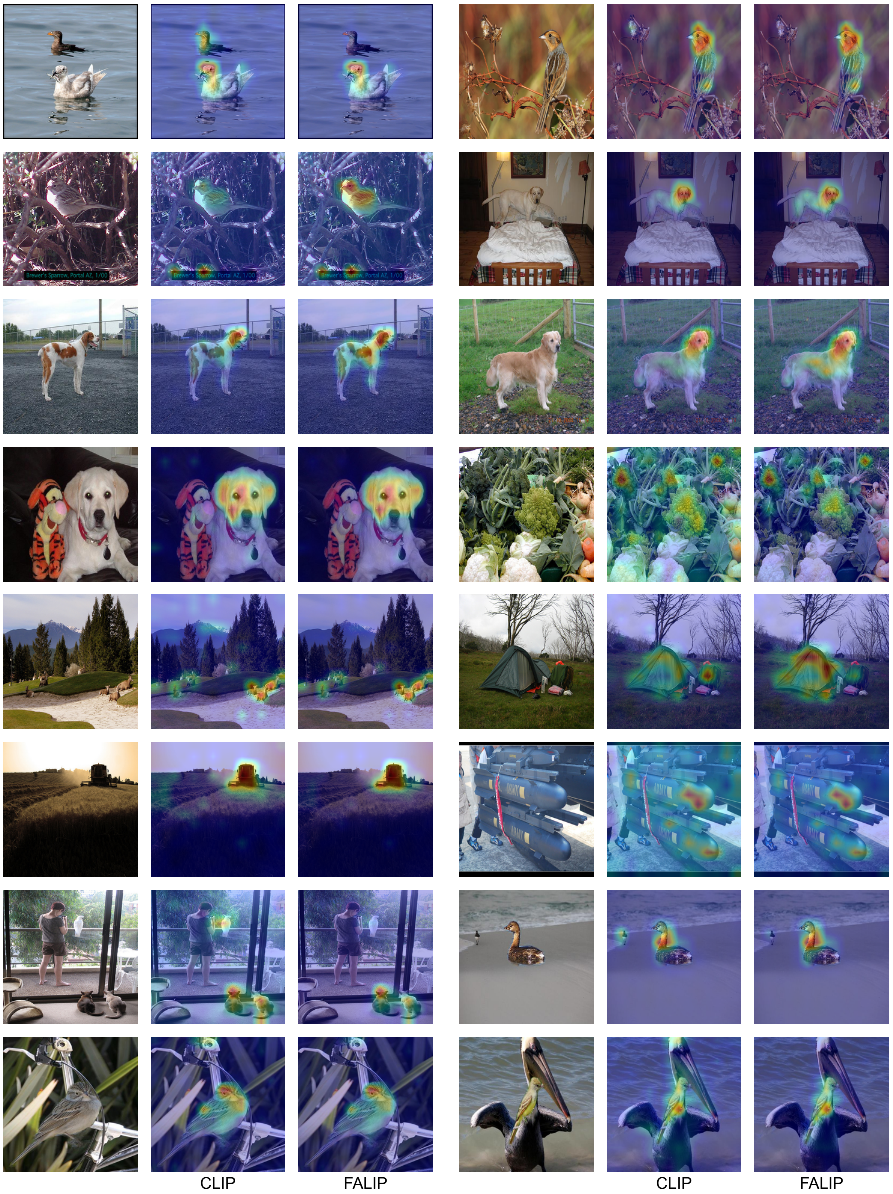


Fig. 15: Attention visualization. Our model demonstrates its ability to better focus on the target objects rather than irrelevant objects in the background.