Appendix

Datasets. Tab. 1, Tab. 2 and Tab. 3 provide a brief introduction to the datasets used for tasks referring expression comprehension, image classification and 3D cloud recognition, respectively.

Table 1: 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 2: 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 3: 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. 4 and Tab. 5 present detailed experimental results about α and σ, respectively. We take $\alpha = 0.2$ and $\sigma = 100$ in final result. Fig. 1 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. 6. We find that inserting masks only in the last 4 layers results in the highest model 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. 4 depicts the details of the ensemble and Fig. 5 shows the extensive results of referring expression comprehension.

Table 4: 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

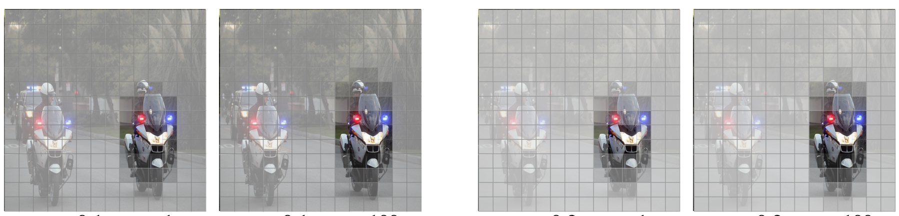
Table 5: Ablation on σ. The best results are in **bold**.

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. 2. Fig. 2 shows the input image of various methods. Tab. 7 demonstrates 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. 8 illustrates how accuracy is affected by visual prompt of varying sizes. Increasing the range of the RedCircle

Table 6: 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



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

Fig. 1: 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.

appropriately can lead to a certain improvement in accuracy. Fig. 4 provides a brief explanation of enlarging size of visual prompt (the maximum size will not exceed the inscribed circle of the image). In Fig. 6 we compare our method with CLIP on the model's attention.

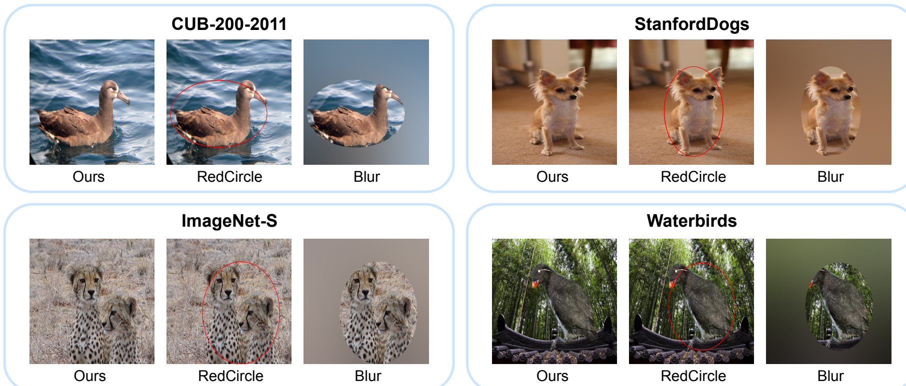


Fig. 2: Examples of input images in each dataset. For each dataset, from the left to right is the input image of model for our method, RedCircle and Blur respectively.

Table 7: Method ablation on Image Classification. The best results are in **bold**, and sub-optimal results are underlined.

Method	Model	StanfordDogs Top1	StanfordDogs Top5	CUB-200-2011 Top1	CUB-200-2011 Top5	ImageNet-S Top1	ImageNet-S Top5	Waterbirds 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

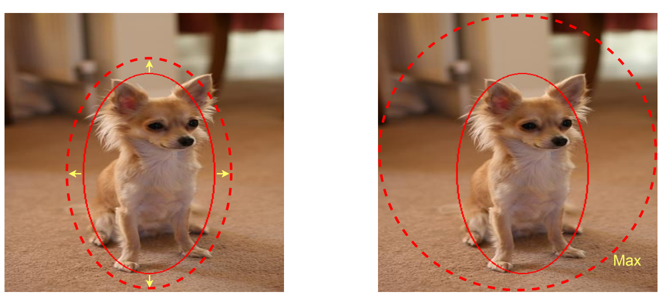


Fig. 3: Enlarge prompts. We increase the pixels in four directions. In this way, the contamination of foreground can be mitigated.

Table 8: 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**

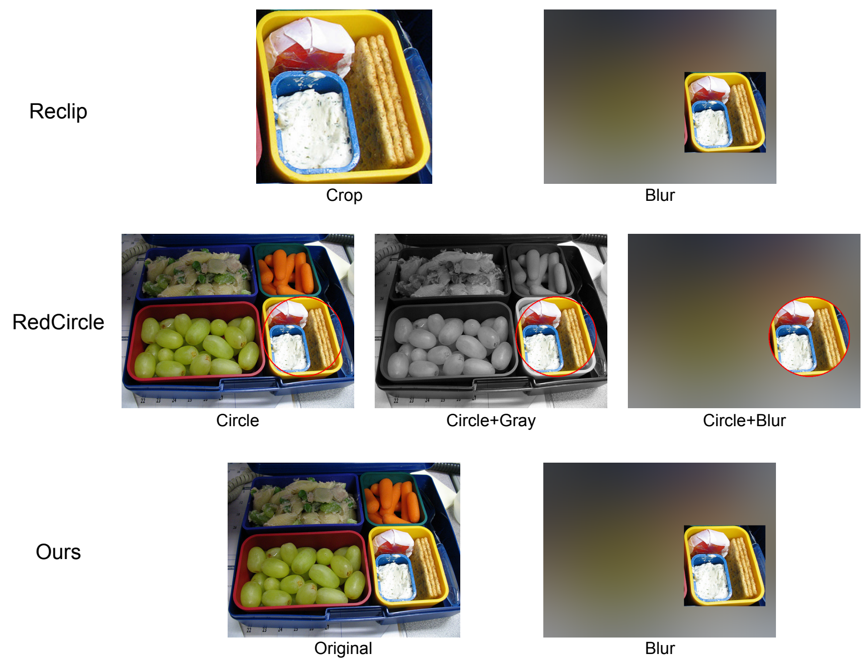


Fig. 4: The specific approaches for ensemble. To ensure a fair comparison, we also adopt the same Blur method used in the previous method.

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

Fig. 5: The visualization results of REC. The keywords are highlighted in orange.

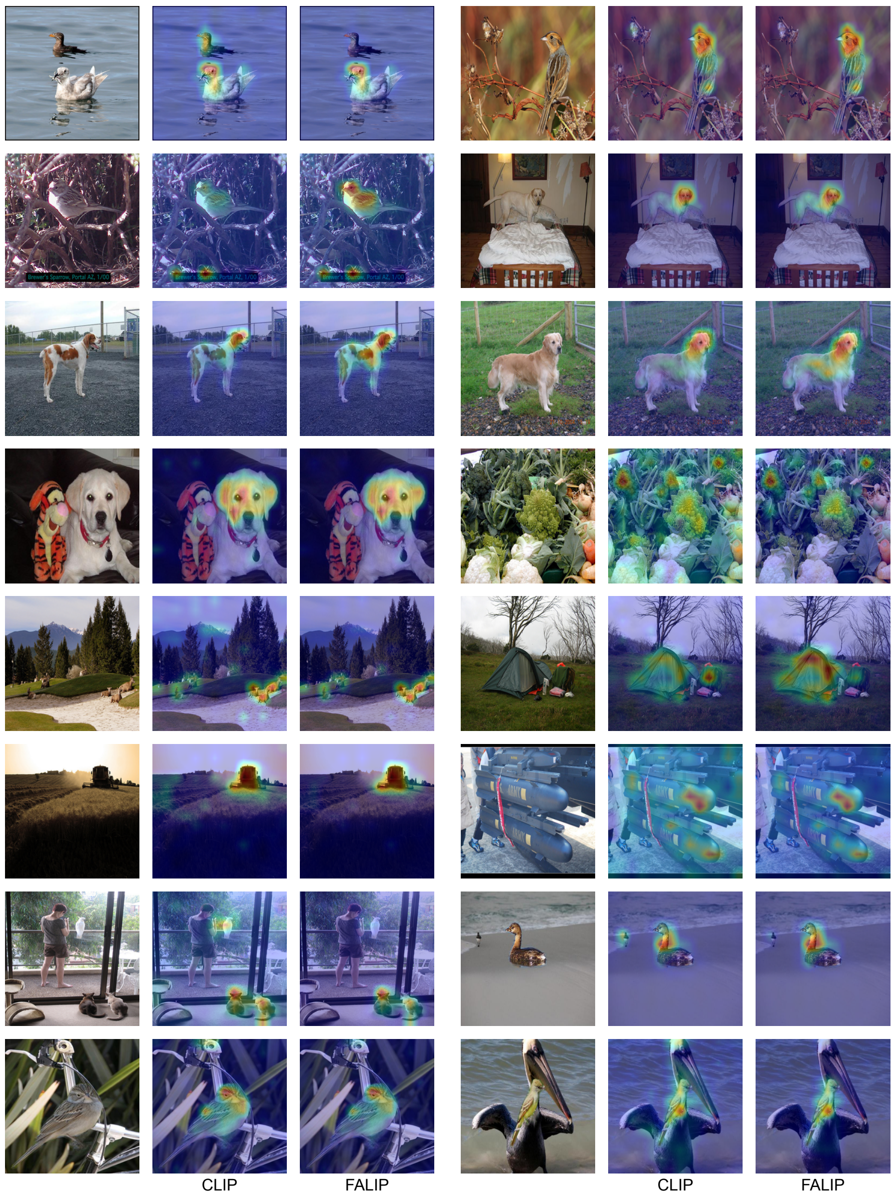


Fig. 6: Attention visualization. Our model demonstrates its ability to better focus on the target objects rather than irrelevant objects in the background.