# EXPLORING VISUAL INTERPRETABILITY FOR CONTRASTIVE LANGUAGE-IMAGE PRE-TRAINING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Contrastive Language-Image Pre-training (CLIP) learns rich representations via readily available supervision of natural language. It improves the performance of downstream vision tasks, including but not limited to the zero-shot, long tail, segmentation, retrieval, caption, and video. However, the visual interpretability of CLIP is rarely studied, especially in the aspect of the raw feature map. To provide visual explanations of its predictions, we propose the Image-Text Similarity Map (ITSM). Based on it, we surprisingly find that CLIP prefers the background regions than the foregrounds, and shows erroneous visualization against human understanding. Experimentally, we find the devil is in the pooling part, where inappropriate pooling methods lead to a phenomenon called semantic shift. To correct and boost the visualization results, we propose the Masked Max Pooling, with attention map from the self-supervised image encoder. Meanwhile, interpretability and recognition require different representations. To address the problem, we propose the dual projections to cater this requirement. We integrate above methods as Interpretable Contrastive Language-Image Pre-training (ICLIP). Our experiments suggest that ICLIP greatly improves the interpretability of CLIP, e.g. nontrivial improvements at 32.85% and 49.10% on VOC 2012 dataset.

## 1 INTRODUCTION

Pre-training is ubiquitously applied in many computer vision tasks such as image classification, object detection and semantic segmentation. To reduce the cost of data acquisition, and broaden the capacity of dataset conveniently, many methods are proposed, such as weakly-supervised pre-training Mahajan et al. (2018) and self-supervised pre-training Doersch et al. (2015); Jaiswal et al. (2020). Compared with the above methods, the Contrastive Language-Image pre-training (CLIP) Radford et al. (2021) learns representations from natural language and leverages a much broader data sources. Then, a series of visual-understanding tasks are improved by it, such as zero-shot and long tail classification Changpinyo et al. (2021), domain generalization Cha et al. (2022), segmentation Xu et al. (2022); Wang et al. (2022), retrieval Luo et al. (2021) and video classification Ni et al. (2022). Follow-up improvements include training scheme Zhai et al. (2022), prompt Zhou et al. (2022) and data Gu et al. (2022), etc.

The applications of CLIP are hot, while to the best of our knowledge, its visual interpretability has not been well explored yet. It means the applications based on CLIP are limited by the inability to explain their decisions to human users. And this lack of explanation punctures the credibility of users, especially in fields like security, clinical decision Tjoa & Guan (2020). For single modality models, prior visual interpretability works Zeiler & Fergus (2014) explain what the convolutional neural network (CNN) learns, and class activation map (CAM) Zhou et al. (2016) reveals the discriminative region via the weights of the classifier. Followed by CAM, many works draw the class activation map by gradient Selvaraju et al. (2017); Chattopadhay et al. (2018) for better performance. Besides CNN, the interpretability of vision transformer (ViT) Dosovitskiy et al. (2020) also have been studied Chefer et al. (2021b). For CLIP, the similarity scores are regarded as "logits" to visualize by gradient in the extended code of Chefer et al. (2021a). However, there is no work to interpret the raw feature map of CLIP, which is direct, basic and simple, without backpropagation.

To visually interpret the predictions of CLIP, we propose a simple and basic concept: Image-Text Similarity Map (ITSM). It is generated by last feature map (image tokens) and the text token with

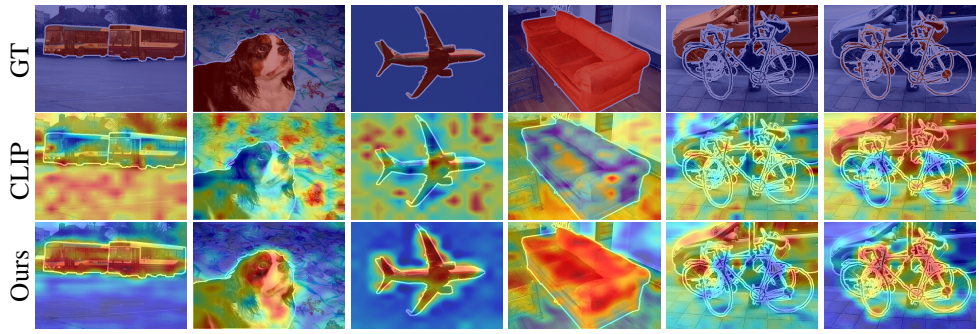



Figure 1: Visualization of CLIP Radford et al. (2021) and ICLIP (ours), based on Image-Text Similarity Map for ViT-B/16 Dosovitskiy et al. (2020). Regions close to red are the target and background is colored in blue. Our ICLIP corrects the erroneous visualization of CLIP to reasonable results.

normalization. It's similar to CAM based on the last feature map, but there are no weights of classifier. Visualized by above ITSM, we surprisingly find that image features from background tokens are more close to the text feature than foregrounds. It means CLIP prefers the background more, which is almost opposite to single modality models and against general understanding.

We experimentally find the devil is in the pooling part. Specifically, we replace the original Attention Pooling (AP, class token) by Global Average pooling (GAP) and Global Max Pooling (GMP), respectively. This problem only disappears in GMP as Fig. 2. And GAP behaves like AP, since it can be regarded as a weighted GAP. We further analyze the reason and find one factor: semantic shift among backgrounds and foregrounds, owing to feature shift by average-like operations. This shift makes text features are matched to background tokens instead of foregrounds, leading to the problem of erroneous visualization.

To further improve the visual interpretability for CLIP, we constrict the max pooling by attention maps from self-supervised image encoder. The attention maps are multiplied to the features of image tokens before computing ITSM. And thus, the features on the foreground are emphasized. We call this pooling as Masked Max Pooling (MMP), which is much better than CLIP visualized by ITSM as Fg. 1. Besides, we find the interpretability task and recognition task require different representations. Thus, we further propose a dual projection architecture to maintain the recognition performance. On PASCAL VOC12 dataset Everingham et al. (2010), two new metrics for interpretability are increased greatly by 32.85 % and 49.10 %, respectively, without obvious loss of recognition accuracy. Our main contributions are summarized as follows:

- This is a prior work to interpret CLIP from raw feature map. Specifically, We use the Image-Text Similarity Map to visualize it, and evaluate the interpretability via two proposed metrics for multi-label images.
- We find CLIP prefers background than foreground, and show erroneous visual results against human understanding. We further locate the problem at the pooling module, and point out one reason is the semantic shift owing to feature shift by average-like operations.
- The Masked Max Pooling is proposed to constrict the semantic shift, and emphasize salient features via attention of self-supervised image encoder. And the architecture of dual projections is deployed to learn different representations for recognition and interpretability.
- These methods are intergraded into Interpretable CLIP (ICLIP), and experiments show nontrivial improvements on the interpretability.

## 2 Task and Problem

### 2.1 Visual Interpretability via Image-Text Similarity Map

In this paper, we explore the visual interpretability by the similarity map between texts and image tokens. We call this map as Image-Text Similarity Map (ITSM). Given an image sample $x$, and the text is regarded as supervision as $y$. After image encoder $f_i$ and the linear projection $\phi_i$ (similar

to fully connected layer), we get $L^p$ normalized features $\boldsymbol{X} \in \mathbb{R}^{1+N_i,C}$ of image tokens as Eq. 1. Here, 1 and $N_i$ mean the class token and image tokens, respectively. $C$ indicates the embedding width, and $\hat{\boldsymbol{X}}$ is the feature matrix before normalization. On the same way, we have the normalized text features $\boldsymbol{Y} \in \mathbb{R}^{N_t,C}$ as Eq. 2, which are the supervision signals during training and weights for ITSM during inference.

$$\hat{\boldsymbol{X}} = f_i(x) \cdot \phi_i, \boldsymbol{X} = \frac{\hat{\boldsymbol{X}}}{||\hat{\boldsymbol{X}}||_p} \tag{1}$$

$$\hat{\boldsymbol{Y}} = f_t(y) \cdot \phi_t, \boldsymbol{Y} = \frac{\hat{\boldsymbol{Y}}}{||\hat{\boldsymbol{Y}}||_p} \tag{2}$$

Then we compute the intermediate similarity matrix $\hat{\boldsymbol{M}} \in \mathbb{R}^{N_i,N_t}$, by inner production between image features $\boldsymbol{X}_{1:,:}$ (class token $\boldsymbol{X}_{:1,:}$ is excluded) and transposed text features $\boldsymbol{Y}^\top$ as Eq. 3.

$$\hat{\boldsymbol{M}} = \boldsymbol{X}_{1:,:} \times \boldsymbol{Y}^\top \tag{3}$$

Then we reconstruct the feature map of **Image-Text Similarity Map** $\mathbf{M} \in \mathbb{R}^{H,W,N_t}$ by reshape, and resize it to the size of the input image via bicubic interpolation, whose width and height are $H$ and $W$, respectively. For larger contrast in visualization, we apply the min-max normalization over dimensions of $H$ and $W$, then we have ITSM as

$$\mathbf{M} = Norm(Resize(Reshape(\hat{\boldsymbol{M}}))). \tag{4}$$

As shown in Fig. 1 and Fig. 6, CLIP shows opposite visualization results. If the performance and real meaning are not cared, the Reversed ITSM (RITSM) $\mathbf{M}_r$ is simple access to visualize without training as Appendix E, where $Abs$ gets the absolute value.

$$\mathbf{M}_r = Abs(1 - \mathbf{M}) \tag{5}$$

For the score vector $\boldsymbol{s}$ for classification, it's the same to the original CLIP as Eq. 6, where the first image token $\boldsymbol{X}_{:1,:}$ is used as the pooled feature.

$$\boldsymbol{s} = \boldsymbol{X}_{:1,:} \times \boldsymbol{Y}^\top \tag{6}$$

### 2.2 Evaluation Metrics and Observed Problem

**mAP for multi-label recognition**. CLIP is designed for image-text pairs, which usually includes multiple objects. For this reason, we evaluate the zero-shot classification on multi-label datasets by mAP, which is the mean area over classes of precision-recall curve.

**mMIoU for prediction map**. For the evaluation of interpretability, previous localization accuracy is not appliable on multi-label datasets. Here, we proposed the **mean Match Inter of Union (mMIoU)** to evaluate interpretability in pixel level, which is similar to mIoU in semantic segmentation task. Compared with mIoU, mMIoU requires the image-level labels to purely evaluate the localization ability without interference of recognition. Besides, we deploy grid search (step 0.01) to find one foreground threshold for each class, to exclude the influence among categories and avoid threshold engineering. We write it as Eq. 7, where $c$, $n$ are numbers of class (without background class) and samples, respectively, and $\boldsymbol{G} \in \mathbb{R}^{H,W}$ is the ground truth matrix, $t$ indicates the searched foreground threshold, $i \in \boldsymbol{G}$ means the usage of image-level labels.

$$mMIoU = \sum_{i=1}^{c} \sum_{j=1}^{n} \frac{(\mathbf{M}^j_{:,:,i} > t) \cap (\boldsymbol{G}^j = i)}{(\mathbf{M}^j_{:,:,i} > t) \cup (\boldsymbol{G}^j = i)}, s.t. i \in \boldsymbol{G} \tag{7}$$

**mFMB for prediction score**. Compared with mMIoU which measures the prediction map, we propose the **mean Foreground Minus Background (mFMB)** to evaluate the prediction score. It directly compared similarity between foreground tokens and backgrounds by minus, and brings new

insights besides the quality of prediction map. It ranges from -1 to 1. When it's lower than 0, the model prefers backgrounds than foregrounds. It's expressed as Eq. 8, where $h$, $w$ are the height and width of maps, and $\odot$ is element-wise multiplication. Note that the left part is the average normalized similarity of foreground tokens, and the right is that of backgrounds.

$$mFMB = \sum_{i=1}^{c}\sum_{j=1}^{n}[\frac{\sum_{k=0}^{h-1}\sum_{l=0}^{w-1} \boldsymbol{M}^{j}_{k,l,i} \odot (\boldsymbol{G}^{j}_{k,l} = i)}{\sum_{k=0}^{h-1}\sum_{l=0}^{w-1}(\boldsymbol{G}^{j}_{k,l} = i)} - \frac{\sum_{k=0}^{h-1}\sum_{l=0}^{w-1} \boldsymbol{M}^{j}_{k,l,i} \odot (\boldsymbol{G}^{j}_{k,l} \neq i)}{\sum_{k=0}^{h-1}\sum_{l=0}^{w-1}(\boldsymbol{G}^{j}_{k,l} \neq i)}], s.t. i \in \boldsymbol{G} \tag{8}$$

**Problem of Interpretability.** As shown in Fig. 1, for the CLIP, image tokens on the backgrounds are surprisingly more close to the text than those on the foregrounds. It presents erroneous visualization results against human understanding. Besides, we measure the quantitative results in Tab. 1. To be specific, the performance of zero-shot multi-label classification is good, but the mMIoU is pretty low, and mFMB indicates the average score on foreground is lower than background by 0.1855. These evidences show that CLIP has the problem of erroneous visualization. For this problem, we locate, explain and solve it in the next section.

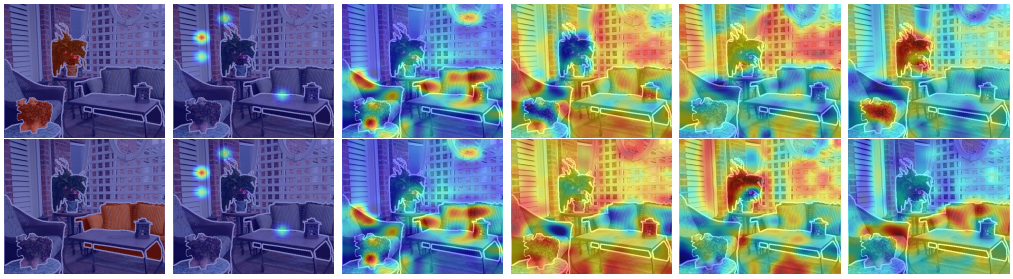
Table 1: Quantitative results of CLIP (ViT-B/16) on VOC12 validation set Everingham et al. (2010), mFMB ranges from -1 to 1.

| mAP (%) ↑ | mMIoU (%) ↑ | mFMB ↑ |
|---|---|---|
| 80.31 | 17.46 | -0.1855 |

## 3 METHODS

### 3.1 THE DEVIL IS IN THE POOLING

**Differences compared with single modality model.** (1) The first influential part is the *image encoder*, which is supervised by text features instead of structure labels. As shown in Fig. 2b, the attention map of the last layer is bad, presenting scatter appearance and focusing on the background more. (2) The second difference is the *pooling method*. In Eq. 6 the score is obtained from class token, which is kind of weighted global average pooling with self-attention. And how to generate the score determines which image token matches best with text. (3) Another difference is the *weight of visualization*. In CAM Zhou et al. (2016), the weights are static parameters of the classifier, and each class has its private weights. While the weights of ITSM are the dynamic feature of text, and both image features and text features require normalization before inner product.



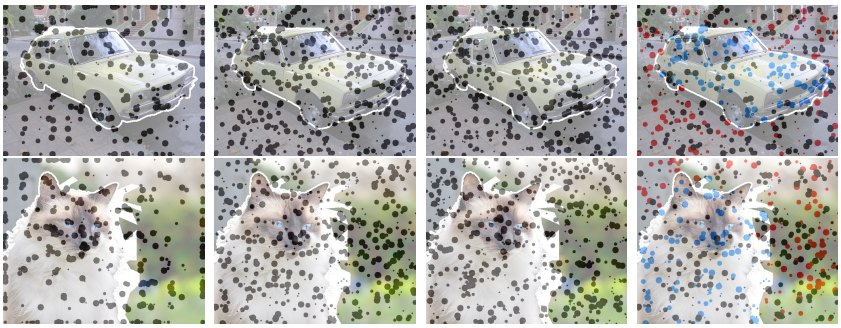
(a) Ground Truth (b) Attention CLIP (c) Attention Ours (d) ITSM AttnPool (e) ITSM AvgPool (f) ITSM MaxPool

Figure 2: Locating the problem. (d) is the ITSM, when the image encoder weights of CLIP (b) are replaced to self-supervised weights (c). Although the attention quality is improved (b vs. c), the problem still exists. When the pooling method is replaced to Max Pooling (f), the problem is solved, and the average pooling (e) behaves like attention pooling (d). Note (c, d, c, e) use the same image encoder and weights. Attention is the class-agnostic self-attention map of the last transformer layer.

**Max pooling solve the problem**. As Fig. 2, we focus on the first two differences and verify their influence by replacing corresponding modules and re-training. For the image encoder, we replace it by the self-supervised image encoder, DINO Caron et al. (2021), and lock the weights of it as LiT Zhai et al. (2022) to speed up the training. As Shown in Fig. 2c, the attention quality is much

better than that of CLIP (Fg. 2b), when the self-supervised image encoder is applied. However, its ITSM based on the original attention pooling still focuses on background. And it shows the image encoder is not the key of erroneous visualization results. For the second difference, we replace the pooling layer to global average pooling and global max pooling, respectively. As Fig. 2e, the average pooling is similar to attention pooling. *While the max pooling solves the problem as Fig. 2f, which suggests the devil is in the pooling module*.

**Reason.** After locate the problem at the pooling module. We analyze the reason by element-wise feature comparison among different pooling methods. Firstly, we match the feature maps before the pooling layer to the values after pooling, and draw the point which is most close to the pooled value for each channel on the image. We can see that many points of max pooling are aggregated and overlapped in Fig. 3a, while average pooling (Fig. 3b) and attention pooling (Fig. 3c) disperse the points. We call this phenomenon as *feature shift*. And feature shift between foregrounds and backgrounds is the *semantic shift* as Fig. 3d. This feature shift leads foregrounds are matched to backgrounds, and background features are turned to foreground regions. Exactly, it behaves as the problem of erroneous visualization results, and explain how it happens. Besides the semantic shift, there may be other reasons, also why CLIP is more sensitive than single modality model about pooling method is waiting to explore.



(a) MaxPool (b) AvgPool (c) AttnPool (d) Avg vs. Max

Figure 3: Illustration of feature shift (a, b, c) and semantic shift (d) on CLIP ViT-B/16. The points indicate the feature scatters before pooling, which is most close to the pooled value for each channel. Larger points mean larger pooled values. Blue points are foreground features of AvgPool, which locates on the background of MaxPool, and red points are from foreground to background. (b) and (c) disperse the overlapped points of (a), leading feature shift. (d) shows this shift leads points to opposite semantic regions, and explain how erroneous visualization happens.

### 3.2 Masked Max Pooling to Boost the Interpretability

As analyzed above, the first principle to improve the interpretability of CLIP is to avoid semantic shift owing to average-like pooling. Another motivation is to emphasize features of foregrounds to boost the interpretability. As shown in Fig. 2b, the attention map of self-supervised model is much better than the original CLIP, and show good interpretability. However, this attention map is class-agnostic. We aim to integrate it into ITSM to get high quality class-aware visualization maps. In this paper, we propose the Masked Mask Pooling (MMP) to constrict the semantic shift, as well as emphasize the discriminative features.

Firstly, we replace the original weights of image encoder to self-supervised weights, and extract the self-attention map $\mathbf{A} \in \mathbb{R}^{N_h,1+N_i,1+N_i}$ from the last transformer layer. Here, $N_h$, $1+N_i$ indicates the number of attention heads and token numbers, respectively. We take the attention map of the first class token, and get mean attention via $Mean$ operation along the first head dimension, with min-max normalization $Norm$. Then we extend the attention to the same embedding channel $C$ as the image feature, and get the expanded mean attention matrix $\boldsymbol{A} \in \mathbb{R}^{N_i,C}$ as Eq. 9.

$$\boldsymbol{A} = Expand(Norm(Mean(\mathbf{A}_{:,0,1:}))) \tag{9}$$

The proposed Masked Max Pooling $Mmp$ is designed to replace the features of class token $\boldsymbol{F}_c \in \mathbb{R}^{1,C}$ from features of image tokens $\boldsymbol{F}_i \in \mathbb{R}^{N_i,C}$ with attention matrix $\boldsymbol{A}$ as Eq.10, where $x$ is the

image sample, and $\hat{f}_s(x)_{1,:}$ indicate the features out of class token before MMP. Note the $max$ operation returns the max value along the first token dimension.

$$\boldsymbol{F}_c = Mmp(\hat{f}_s(x)_{1,:}) = max(\hat{f}_s(x)_{1,:} \odot \boldsymbol{A}) \tag{10}$$

Then we replace the weights of image encoder $f_i$ in Eq. 1 by the self-supervised weights $f_s$ with MMP as Eq. 11, where $\boldsymbol{F}_i = \hat{f}_s(x)_{1,:} \odot \boldsymbol{A}$ is the weighted features of image tokens, and $Cat$ is the concatenate operation on the first dimension, for the same token size 1 + $N_i$ as original CLIP.

$$f_s(x) = Cat(\boldsymbol{F}_c, \boldsymbol{F}_i) \tag{11}$$

### 3.3 Dual Projections and Overall Architecture

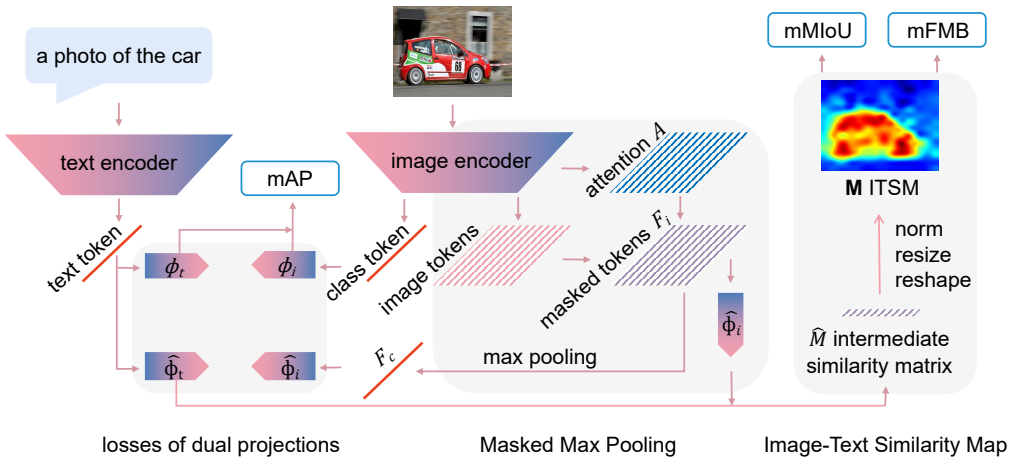



Figure 4: Illustration of ICLIP for single image-text pair. **Middle**: the self-supervised image encoder return features of class token, image tokens, and expanded mean attention map. The feature of image tokens are combined with the attention by element-wise multiplication, and get the pooled features $\boldsymbol{F}_c, (1, C)$ by max pooling among token dimension. **Left**: there are dual projections $\phi_i, \hat{\phi}_i$ with corresponding text projections $\phi_t, \hat{\phi}_t$ to compute contrastive losses. **Right**: for the generation of ITSM, the masked tokens $\boldsymbol{F}_i$ are projected by $\hat{\phi}_i$ and get intermediate similarity matrix $\hat{\boldsymbol{M}}, (N_i, 1)$ with text features from $\hat{\phi}_t$. After reshape, resize and min-max normalization, the ITSM is evaluated by mMIoU and mFMB. The evaluation of mAP uses the outputs of $\phi_i$ and $\phi_t$.

Experimentally, we find the recognition task and interpretability task require different representations. As shown in Tab. 2, although max pooling solves the problem, its performance of zero-shot classification obviously drops. To meet the requirement of two tasks, we propose the dual projections. Specifically, another image linear projection $\hat{\phi}_i$ is applied, with corresponding text linear projection $\hat{\phi}_t$. During inference phase, the features before normalization $\hat{\boldsymbol{X}}$ in Eg. 1 are concatenated $Cat$ from two branches as Eq. 12, where the feature of first class token $\hat{f}_s(x)_{:1,:}$ without attention is inner produced with original projection $\phi_i$, and image tokens are element-wise produced with attention $\boldsymbol{A}$ and inner produced with another projection $\hat{\phi}_i$.

Table 2: Necessity of dual projections. These results are reported on VOC12 validation set, at same training data and architecture. Note, mFMB ∈ [-1, 1].

| Pooling | mAP (%) ↑ | mMIoU (%) ↑ | mFMB ↑ |
|---|---|---|---|
| attention | **76.31** | 19.96 | -0.0877 |
| average | 68.78 | 17.77 | -0.2116 |
| max | 60.13 | **36.31** | **0.1816** |

$$\hat{\boldsymbol{X}} = Cat(\hat{f}_s(x)_{:1,:} \cdot \phi_i, \hat{f}_s(x)_{1,:} \odot \boldsymbol{A} \cdot \hat{\phi}_i) \tag{12}$$

Note that the text features also apply different projections. And the Eq. 3 is modified to $\hat{\boldsymbol{M}} = \boldsymbol{X}_{1:,:} \times \boldsymbol{Y}^{\top}, s.t. \phi_t = \hat{\phi}_t$, where the original text projection $\phi_t$ is changed to $\hat{\phi}_t$.

During training, the contrastive loss of CLIP $\mathcal{L}(f_i(x)_{:1,:} \cdot \phi_i, f_t(y) \cdot \phi_t)$ is expanded to Eq. 13. The first part is for the class token without MMP $\hat{f}_s(x)_{:1,:} \cdot \phi_i$. And the right part is for the pooled token of branch with MMP $f_s(x)_{:1,:} \cdot \hat{\phi}_i$, with extra image projection $\hat{\phi}_i$ and text projection $\hat{\phi}_t$.

$$\mathcal{L}(x, y) = (\mathcal{L}(\hat{f}_s(x)_{:1,:} \cdot \phi_i, f_t(y) \cdot \phi_t) + \mathcal{L}(f_s(x)_{:1,:} \cdot \hat{\phi}_i, f_t(y) \cdot \hat{\phi}_t))/2 \tag{13}$$

For better understanding, we depict the overall framework of ICLIP, including Masked Max Pooling, dual projections, Image-Text Similarity Map and evaluations as in Fig. 4.

# 4 EXPERIMENTS

## 4.1 EXPERIMENTAL SETUP.

**Datasets and evaluation.** The original CLIP Radford et al. (2021) uses 400 millions image-text pairs to train the models. However, this dataset is not available, also too large to reproduce. In this paper, we use the dataset of Google Conceptual Captions 3 millions (GCC3M) Sharma et al. (2018) to train the model, because of the moderate quantity. For the evaluation, we don't report accuracy of zero-shot classification on large scale datasets like ImageNet Lin et al. (2014), because the quantity of training set is not large enough to support it. Another concern is that images in real world usually contains multiple objects, and mAP for multi-label zero-shot classification suits CLIP well. Specifically, we use the Pascal VOC 2012 validation set Everingham et al. (2010) and MS COCO 2017 validation set Lin et al. (2014), which are multi-label datasets with segmentation annotations. Compared with single label dataset like ImageNet in previous interpretability works, the localization metric is changed to the proposed segmentation-like metrics mMIoU and mFMB for finer evaluation. The quantity of VOC 12 validation set is 1449, and COCO 2017 has 5000 validation images. Note that there are 80 foreground categories in COCO, and it's more complex and difficult than VOC whose class number is 20.

**Settings.** For the CLIP models, their ITSM are generated from the official models trained from 400 millions private data, without fine-tuning. And our models are all trained with GCC3M based on ViT-B/16 (patch size 16), which returns 196 image tokens (14×14) at input resolution 224. The text prompt is "a photo of the", and the output similarity is normalized without softmax. For the architecture, the difference compared with CLIP is the image, encoder. In this paper, we use the weights of the self-supervised model, DINO Caron et al. (2021), which is pre-trained from ImageNet without label. During training, the image-encoder is locked without gradient for fast convergence. And the weights of projections and text encoder are updated by AdamW Loshchilov & Hutter (2017), at learning rate 1.25e-4, total batch size 1024, weight decay 0.05 for 30 epochs. Other training settings including augmentation, scheduler are followed by Deit Touvron et al. (2021). Besides, the implementations Xu et al. (2022) of text augmentation and loss are the same to CLIP. Thanks to the locked image encoder, moderate dataset and half precision training, we only need about 33 hours on 4 Nvidia 3090Ti GPUs, or 11 hours for at 10 epochs for comparable results.

## 4.2 RESULTS

Table 3: Accumulative gains of our ICLIP on VOC12 validation set. "+" indicates this module is added based on above improvements. All the methods are trained on GCC3M Sharma et al. (2018) for ViT-B/16. Best results are marked in bold, and second results are noted by underline.

| Improvements of ICLIP | mAP (%) ↑ | mMIoU (%) ↑ | mFMB ([-1,1]) ↑ |
|---|---|---|---|
| basic CLIP | 46.78 | 19.83 | -0.0298 |
| + self-supervised image encoder | <u>76.31</u> | 19.96 | -0.0877 |
| + masked max pooling | 62.80 | <u>49.08</u> | <u>0.2949</u> |
| + dual projections | **78.06** | **50.31** | **0.3055** |

**Ablation study.** We list the accumulative gains of our ICLIP in Tab. 3. Compared with the basic CLIP, the self-supervised image encoder is used as pre-training weights to speed up convergence.

Then the masked max pooling greatly improves the performance of interpretability. While its mAP drops, owing to shared but unsuitable representations. Thus, the dual projections is deployed to maintain the performance of classification, and achieve best results over these three metrics.

**Significant improvements of interpretability.** In Tab. 4, we show the effectiveness of our method by comparing with official CLIP and CLIP with the same dataset. For zero-shot classification, official CLIP works best, since its training data is about 133 time of ours. And our results of mAP are much higher than CLIP in the same dataset, thanks to the self-supervised image encoder and dual projections as Tab. 3. The most significant improvements occur in interpretability metrics. Specifically, compared with official CLIP, mMIoU on VOC and COCO are increased by 32.85% and 16.35%, respectively. From mFMB, the problem of erroneous visualization is corrected, and the improvements are 49.10% on VOC and 43.96% on COCO.

Table 4: Nontrivial improvements on VOC and COCO. Official CLIP uses 400 millions private data, and other experiments are trained with GCC 3 millions. Best results are marked in bold, and second results are noted by underline. Note, ICLIP is our method, and mFMB $\in [-1, 1]$.

| | | VOC 2012 val | | | COCO 2017 val | | |
|---|---|---|---|---|---|---|---|
| Expriments | Data | mAP (%) ↑ | mMIoU (%) ↑ | mFMB ↑ | mAP (%) ↑ | mMIoU (%) ↑ | mFMB ↑ |
| CLIP | 400M | **80.31** | 17.46 | -0.1855 | **53.07** | 9.80 | -0.2230 |
| CLIP | 3M | 46.78 | 19.83 | -0.0298 | 20.93 | 11.33 | -0.0419 |
| **ICLIP** | 3M | <u>78.06</u> | **50.31** | **0.3055** | <u>42.04</u> | **26.15** | **0.2166** |

Besides, we compare ICLIP with the proposed Reversed ITSM (RITSM), and the latest gradient-based method Bi-Model Chefer et al. (2021a) for ViT in Tab. 5. Our RITSM and ICLIP are both beyond the gradient based method at the same backbone, ViT-B/16, in a simpler and direct way. We also extend this part in Appendix E and Appendix F.

Table 5: Comparison with the proposed RITSM and the gradient based method Bi-Model on VOC.

| Method | Type | Data | mAP (%) ↑ | mMIoU (%) ↑ | mFMB ([-1,1]) ↑ |
|---|---|---|---|---|---|
| Bi-Model Chefer et al. (2021a) | Gradient | 400M | **80.31** | 29.84 | 0.0635 |
| RITSM, Eq. 5 | Raw Feature | 400M | **80.31** | <u>36.31</u> | <u>0.1855</u> |
| ICLIP | Raw Feature | 3M | <u>78.06</u> | **50.31** | **0.3055** |

**Visualization.** As shown in Fig.5, we draw the qualitative visualization results. This figure includes VOC12 and COCO17 evaluated above, as well as single label dataset ImageNet. Since there are no pixel-level annotations of ImageNet, and our training dataset is not large enough to support it, we visualize it without evaluation. From these qualitative results, we believe the proposed method is able to explain which parts influence the predictions most, and help us to understand the model for better credibility. Also, these results show the potentiality for tasks like segmentation and localization.

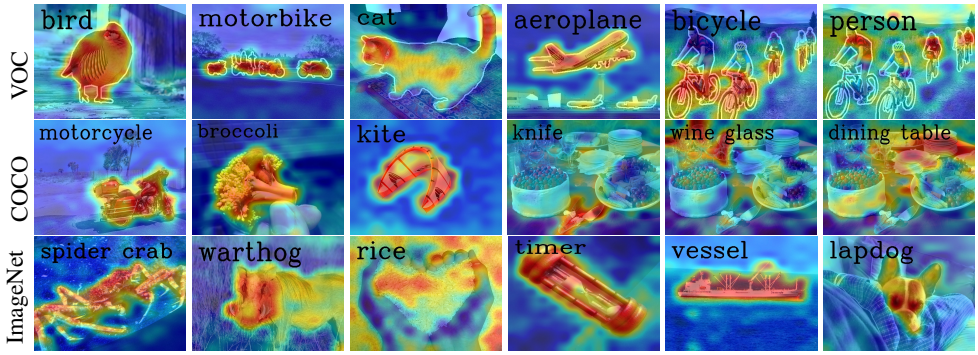



Figure 5: Qualitative visualization.

### 4.3 ANALYSIS

**Details of semantic shift.** We visually analyze the reason of erroneous interpretability in Fig. 3d. And in this part, we count the detailed quantity of semantic shift for varied foreground sizes as Tab. 6. when the size is (0, 1], the overall shift number is 193.5 and takes about 38% of total channels. Especially, when the size is in the middle level (0.25, 0.75], about half channels are shifted in semantics. Except small size, F2B is more than B2F.

Table 6: Quantity statistics of semantic shift on VOC12 validation set of CLIP ViT-B/16 at varied foreground size. B2F means points matched on the background of max pooling are shifted to foreground of average pooling, F2B is opposite to it, and B2B/F2F indicates the number of features without semantic shift.

| Size | B2F | F2B | B2B/F2F |
|---|---|---|---|
| (0, 0.25] | 78.0 | 45.2 | 388.7 |
| (0.25, 0.75] | 111.8 | 136.5 | 263.7 |
| (0.75, 1] | 44.7 | 97.0 | 370.3 |
| (0, 1] | 78.8 | 114.7 | 318.5 |

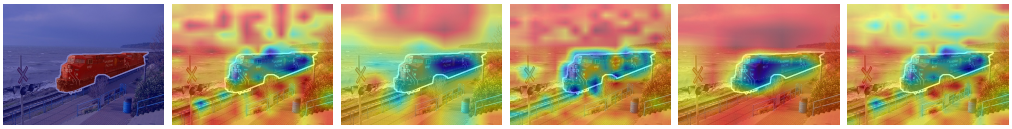
**MMP emphasizes the features of foreground.** Besides the semantic shift to explain why max pooling suits the interpretability of CLIP, we also list the number of channels matched on the foreground as Tab. 7. In this table, attention pooling indicates CLIP uses the same training dataset and weights of image encoder. And our method focus much more on the foregrounds, thanks to the free attention map from self-supervised pre-training. This table quantitatively shows the magnitude of the emphases of the proposed MMP for foreground features.

Table 7: Number of channels matched on the foreground.

| CLIP | attention pooling | Ours |
|---|---|---|
| 160.5 | 169.1 | 222.48 |

**The problem of erroneous visualization is universal for CLIP.**

We visualize the ITSM of CLIP as Fig. 6b, and the problem of erroneous visualization is universal, regardless of the network structure, patch size and visualization method. For the method, we use the gradient based interpretability method Grad_CAM Selvaraju et al. (2017) by replacing original confident to the image-text similarity. But the problem is still existed as Fig. 6f. So we don't extend the gradient based methods in this paper, because the ITSM is more simple and basics.



(a) Ground Truth (b) ITSM ViT-B/16 (c) ITSM ViT-B/32 (d) ITSM ViT-L/14 (e) ITSM RN101 (f) Grad_CAM ViT-B/16

Figure 6: CLIP presents opposite visualization results, regardless of network (d, e), patch size (c, d) and visualization method (f). Regions close to red are the target and background is colored in blue. Besides, we also compare the reversed ITSM in Appendix E, which is worse than our ICLIP too.

## 5 CONCLUSION

In summary, we visually interpret the Contrastive Language-Image pre-training (CLIP) model by the proposed Image-Text Similarity Map (ITSM). Based on it, we observe that CLIP is prone to focus on the background, presenting erroneous visualization results against human understanding. We find one reason is the semantic shift of features in the pooling module. And The problem is solved by removing the average-like pooling. We further propose the Masked Max Pooling (MMP) to avoid the feature shift, and emphasize the features on foreground by attention of self-supervised image encoder. Meanwhile, to cater different representation requirements between interpretability and recognition, we propose the dual projections to maintain the performance of zero-shot classification. The proposed methods are intergraded as Interpretable CLIP (ICLIP), and achieve nontrivial improvements on two new metrics for the interpretability of CLIP.

## REFERENCES


Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. *arXiv preprint arXiv:2204.02491*, 2022.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Junbum Cha, Kyungjae Lee, Sungrae Park, and Sanghyuk Chun. Domain generalization by mutual-information regularization with pre-trained models. *arXiv preprint arXiv:2203.10789*, 2022.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3558–3568, 2021.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 839–847. IEEE, 2018.

Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 397–406, 2021a.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 782–791, 2021b.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pp. 1422–1430, 2015.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021.

Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 317–326, 2016.

Hossein Gholamalinezhad and Hossein Khosravi. Pooling methods in deep neural networks, a review. *arXiv preprint arXiv:2009.07485*, 2020.

Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Minzhe Niu, Hang Xu, Xiaodan Liang, Wei Zhang, Xin Jiang, and Chunjing Xu. Wukong: 100 million large-scale chinese cross-modal pre-training dataset and a foundation framework. *arXiv preprint arXiv:2202.06767*, 2022.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

Suha Kwak, Seunghoon Hong, and Bohyung Han. Weakly supervised semantic segmentation using superpixel pooling network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7331–7341, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021.

Teli Ma, Shijie Geng, Mengmeng Wang, Jing Shao, Jiasen Lu, Hongsheng Li, Peng Gao, and Yu Qiao. A simple long-tailed recognition baseline via vision-language model. *arXiv preprint arXiv:2111.14745*, 2021.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 181–196, 2018.

Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. *arXiv preprint arXiv:2208.02816*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18082–18091, June 2022.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, 32(11):4793–4813, 2020.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 24–25, 2020.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11686–11695, 2022.

Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. Camp: Cross-modal adaptive message passing for text-image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5764–5773, 2019.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18134–18144, 2022.

Nir Zabari and Yedid Hoshen. Semantic segmentation in-the-wild without seeing any segmentation examples. *arXiv preprint arXiv:2112.03185*, 2021.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18123–18133, 2022.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, pp. 1–12, 2022.

## A Related Work

Extensive works have been developed on the interaction of computer vision and natural language processing, such as text-to-image retrieval Wang et al. (2019), visual question answering Antol et al. (2015) and referring segmentation Wang et al. (2022). Recently, contrastive language-image pre-training (CLIP) Radford et al. (2021) has gained substantial attention, with its impressive performance, and superior transfer ability over diverse classification datasets. CLIP takes natural language as efficient supervision to learn rice representations. Along this direction, some works (e.g., CoOp Zhou et al. (2022) and CLIP-Adapter Gao et al. (2021)) dedicate to improving the performance by fine-tuning CLIP with adapter or prompt optimization, using either abundant or limited training data. LiT Zhai et al. (2022) gives a different pre-training strategy which only tunes the text model in the CLIP using image-text pairs, and locks its image model. In addition, some followers focus on developing CLIP to a variety of practical downstream tasks Xu et al. (2022); Ni et al. (2022); Cha et al. (2022); Changpinyo et al. (2021). For example, Lei et al. (2021); Luo et al. (2021) presents that the model pre-trained by huge amount of image-text pairs can contribute to retrieval task. Ma et al. (2021) leverages language modality via CLIP backbone to facilitate long-tailed recognition. Rao et al. (2022) extend image-text relationships to pixel-text relationships to guide the training of dense prediction models. Xu et al. (2022) explores zero-shot transfer learning to semantic segmentation tasks with only language supervision.

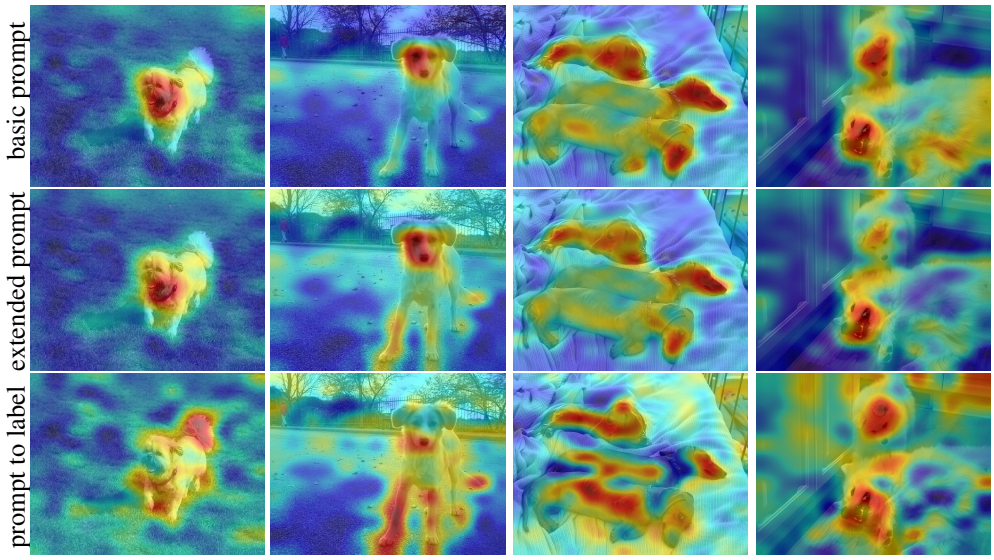
Although we have witnessed many CLIP-related works achieving high performance in a variety of visual tasks, to our best acknowledge, the visual interpretability of CLIP rarely studied. Most previous works about interpretability Zeiler & Fergus (2014); Zhou et al. (2016) are based on a single modality. The prior work, CAM Zhou et al. (2016), locates the discriminative regions of CNN via drawing the class activation mapping. Followed by it, some works (e.g., Score CAM Wang et al. (2020), and Grad CAM Selvaraju et al. (2017)) explore different methods to generate the CAM, to reveal visual cues distributed on images more precisely and effectively. Besides, the interpretability of the vision transformer Dosovitskiy et al. (2020) has also been investigated. Chefer et al. (2021b) designs a class-specific visualization method for self-attention models of vision transformer. For CLIP, Chefer et al. (2021a;b) uses the similarity scores as "logits" and interpret with gradient, and some works Bar-Tal et al. (2022); Zabari & Hoshen (2021) use this relevance map in downstream

tasks. However, explanation from raw feature has not been studied yet, which is more basic and direct. Most importantly, our results in Tab. 5 and Fig.10 suggest our raw feature based method performs better, even the simplest version, reversed ITSM, performs better than the previous work.

Besides the interpretability methods, we also investigate the pooling methods, as the Masked Max Pooling is one key module of our ICLIP. In the survey Gholamalinezhad & Khosravi (2020) about pooling layer, most pooling methods are design to improve the recognition task, like Compact Bilinear Pooling Gao et al. (2016) for fine-grained classification. And some methods serve for down stream tasks, such as Super-pixel Pooling for segmentation Kwak et al. (2017), Region of Interest Pooling for detection Girshick (2015). Our Masked Max Pooling (MMP) is also a kind of pooling method for down stream task. While MMP is design for the interpretability of CLIP. Besides the varied task, the mask from self-supervised attention map is newly applied as weight of pooling to emphasize the foreground regions.

## B VISUALIZATION FOR LOCAL PARTS VIA EXTENDED PROMPTS

For CLIP, the prompt influences the final results a lot. In this part, we explore the potentiality of prompt on localization of local parts of object. As the models are prone to learn the most discriminative parts, our idea is to highlight other regions by extended prompts. Taking the text of dog as an instance, we provide four extended keywords as extra prompts as Fig. 7. The basic prompt with text is "a photo of the dog", and the extended prompt is "tail of the dog" for Fig. 7a. Note that "prompt to label" indicates the text is replaced by the prompt keyword. We can see from the figures, the extended prompts work occasionally (e.g. legs of Fig. 7b). And "prompt to label" focuses more on the target parts, but the visualization results are not good as basic prompt. From these visualizations, we believe extended prompts are able to influence the ITSM, while how to make good use of it is waiting to explore.




(a) tail (b) legs (c) body (d) head

Figure 7: Visualization with extended prompts. For (a), the extended prompt is "tail of the dog", and the prompt to label is "a photo of the tail", by replacing "dog" of the basic prompt to the keyword "tail". Extended prompt emphasizes the target region occasionally (b, d), and prompt to label focus more on the targets with worse visualization results.

## C FAILURE CASES AND EXISTING PROBLEMS

Besides the qualitative results in Fig. 5, we report the failure cases in Fig. 8, and summarize existing problems for reference. The first problem is about complex scenarios. When the image contains

many categories or the targets are not obvious, the performance is worse than anticipation. This problem is highly related to the performance of zero-shot classification. The second problem is similar to the difficulty of fine-grained classification. For example, it's hard to distinguish truck, car and bus, presenting similar ITSM. One possible reason is about the attention. Since it's class-agnostic, and comes from mean operation, fine-grained classes or related categories are very likely to be confused by the shared attention map. Besides, some common categories behave bad, such as person. It's very common, and many image-text pairs do not mention it, which lead to noisy supervision towards certain categories. Besides, the self-supervised image encoder learns a lot about animals and plants, with little representations about human. Furthermore, the interpretability of CLIP is newly studied, and exists many problems waiting to solve.

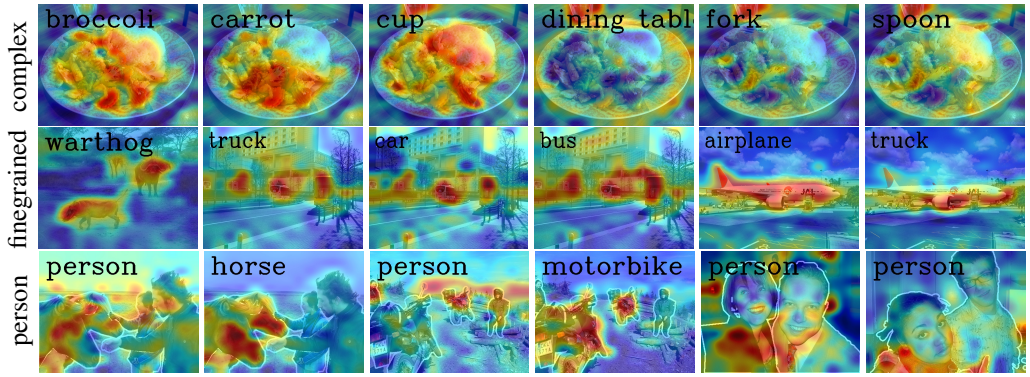
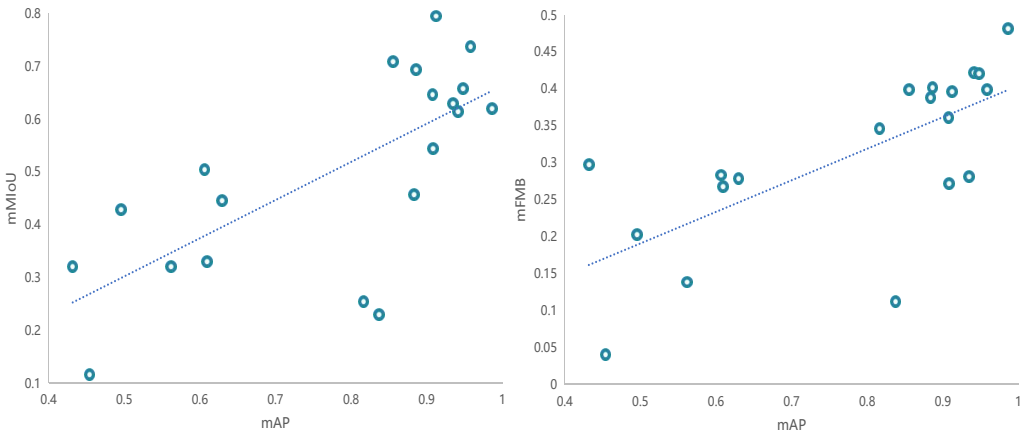



Figure 8: Failure cases about complex scenarios, fine-grained categories and certain categories.

## D RELATION BETWEEN INTERPRETABILITY AND RECOGNITION.

We analyze the relation between classification and interpretability as Fig. 9. The X-axis is the mAP of zero-shot classification on VOC 2012 validation set, and the Y-axis is mMIoU of Fig. 9a, and mFMB of Fig. 9b. These two metrics are both positively correlated with mAP, which suggests these two tasks are mutual benefit. It means we are able to pick high quality visualization results, by its zero-shot classification score.




(a) mMIoU-mAP

(b) mFMB-mAP

Figure 9: Relation between zero-shot classification and visual interpretability. The points are results of classes on PASCAL VOC 2012 validation set.

Table 8: . Performance of reversed ITSM on VOC 2012 validation set.

| Experiments | Data | mAP (%) ↑ | mMIoU (%) ↑ | mFMB ([-1,1]) ↑ |
|---|---|---|---|---|
| CLIP | 400M | **80.31** | 17.46 | -0.1855 |
| Reversed CLIP | 400M | **80.31** | <u>36.31</u> | <u>0.1855</u> |
| Reversed CLIP | 3M | 46.78 | 23.22 | 0.0298 |
| ICLIP | 3M | <u>78.06</u> | **50.31** | **0.3055** |

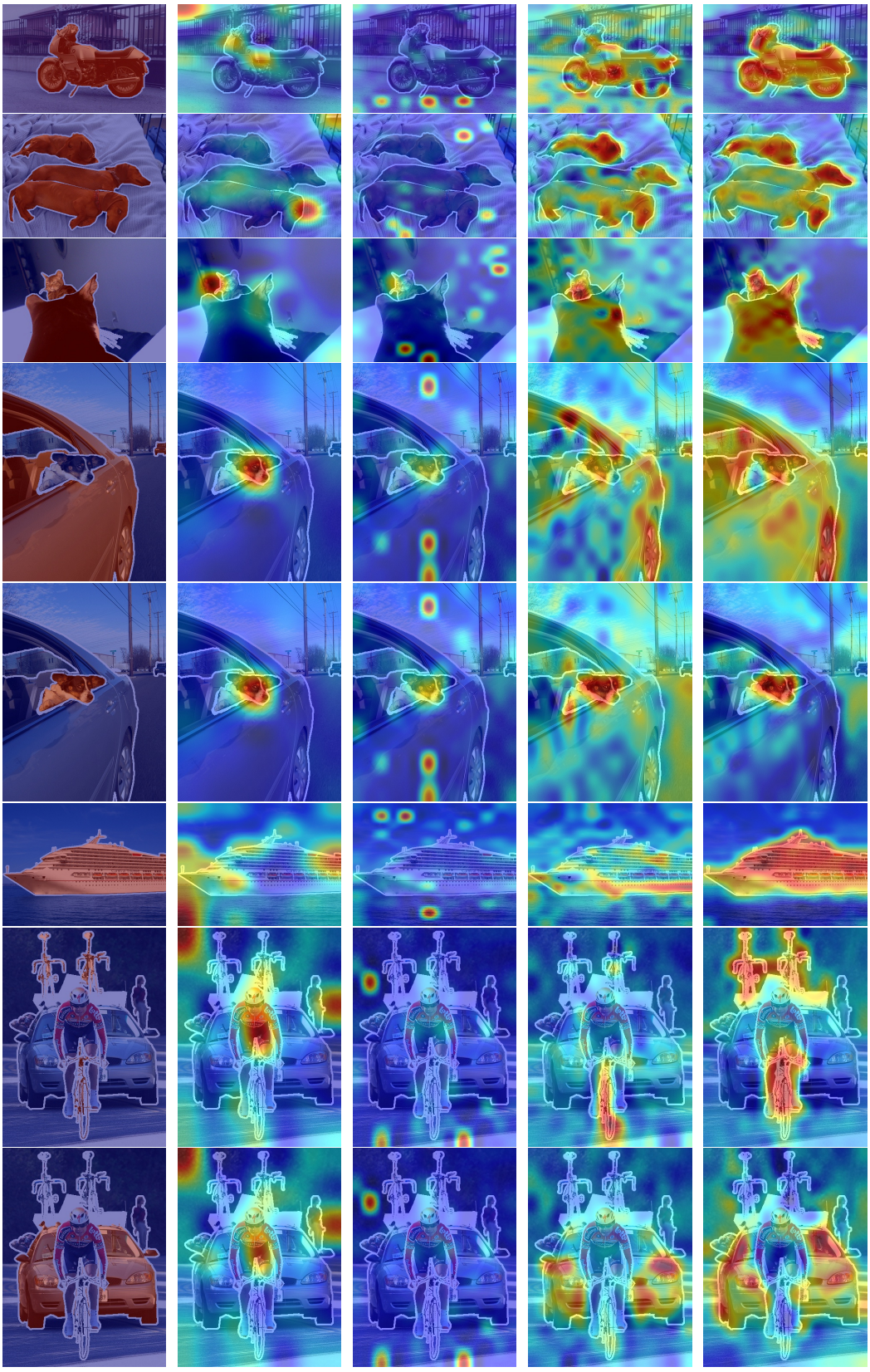
# E MAKE THE BEST OF A MISTAKE: REVERSE THE ITSM OF ORIGINAL CLIP

As shown in Fig. 6 and Tab. 4, CLIP shows opposite visualization results and prefers the background more. Analyzed in Fig. 3d, this problem is owing to semantic shift. So, one simple idea is to make the best of a mistake by reversing the ITSM.

Then we evaluate the reversed ITSM in equation 5 as Tab. 8. And we find the reversed ITSM is more reasonable than the original ITSM, but it's worse than the proposed ICLIP. And the reproduced CLIP in the same data performs much worse, because the dataset size is much smaller. Compared with reversed CLIP (3M), our ICLIP is 27.09% higher at mMIoU, and it's 14% higher than official CLIP trained with 400 millions image-text pairs. Although, the official CLIP is obviously worse than our ICLIP, but it's also able to use directly without retraining, if you do not care too much about the performance and real meaning.

# F VISUAL COMPARISON WITH GRADIENT BASED INTERPRETABILITY METHOD

We draw visualization results in Fig. 10 to compare our Reversed ITSM and ICLIP with the gradient based interpretability method Bi-Model. Note that, the interpretability of CLIP is not involved in the paper, while its official code provides the visualization method for CLIP. This method requires back-propagation to generate the heatmap, while our methods obtain it from the raw feature map, without complex operations. Besides the simplicity, our methods perform better. Specifically, the visualization results are not in scatter-like appearance, instead, the foregrounds are clearly distinguished from backgrounds, especially for ICLIP. We also notice that, Bi-Model performs worse when the patch size is larger, and patch size 32 is better. Since its scatter is larger (1/7 vs. 1/14 of patch 16), and interpolation operation expands it to wider region in smoothing appearance.



(a) Groud Truth (b) Bi-Model (32) (c) Bi-Model (d) Reversed ITSM (e) ICLIP

Figure 10: Qualitative visualization of our ICLIP and RITSM, compared with and gradient based method, Bi-Model Chefer et al. (2021a). All the method use the same backbone: ViT-B/16, but Bi-Module (32) uses ViT-B/32.