Gradient-based Visual Explanation for Transformer-based CLIP

Chenyang Zhao [1 2] **Kun Wang** [2] **Xingyu Zeng** [2] **Rui Zhao** [2] **Antoni B. Chan** [1]

Abstract

Significant progress has been achieved on the improvement and downstream usages of the Contrastive Language-Image Pre-training (CLIP) vision-language model, while less attention is paid to the interpretation of CLIP. We propose a Gradient-based visual Explanation method for CLIP (Grad-ECLIP), which interprets the matching result of CLIP for specific input image-text pair. By decomposing the architecture of the encoder and discovering the relationship between the matching similarity and intermediate spatial features, Grad-ECLIP produces effective heat maps that show the influence of image regions or words on the CLIP results. Different from the previous Transformer interpretation methods that focus on the utilization of self-attention maps, which are typically extremely sparse in CLIP, we produce high-quality visual explanations by applying channel and spatial weights on token features. Qualitative and quantitative evaluations verify the superiority of Grad-ECLIP compared with the state-of-the-art methods. A series of analysis are conducted based on our visual explanation results, from which we explore the working mechanism of image-text matching, and the strengths and limitations in attribution identification of CLIP. Codes are available here: https://github.com/Cyang-Zhao/Grad-Eclip.

1. Introduction

Recently, by learning the representations for matching caption text and its corresponding image, the Contrastive Language-Image Pre-training (CLIP) model (Radford et al., 2021) has introduced a simple and effective dual-encoder pre-training paradigm for the interaction between natural language processing and computer vision. CLIP significantly improves the performance on various downstream tasks, such as classification (Changpinyo et al., 2021; Cha et al., 2022), retrieval (Luo et al., 2022) and segmentation (Wang et al., 2022; Xu et al., 2022), with zero-shot and finetuning methodologies. Inspired from CLIP, multi-modal pre-training has been further developed by exploring different perspectives, including unifying vision-language understanding and generation (Yu et al., 2022; Li et al., 2022a), prompt design (Zhou et al., 2022b; Chen et al., 2022), and region-aware enhancement (Li et al., 2020; Wang et al., 2023; Zhong et al., 2022). Although researchers devote many efforts into improving multi-modal pre-training or exploring the usages in downstream tasks, less attention has been focused on the interpretation or explanation of CLIP.

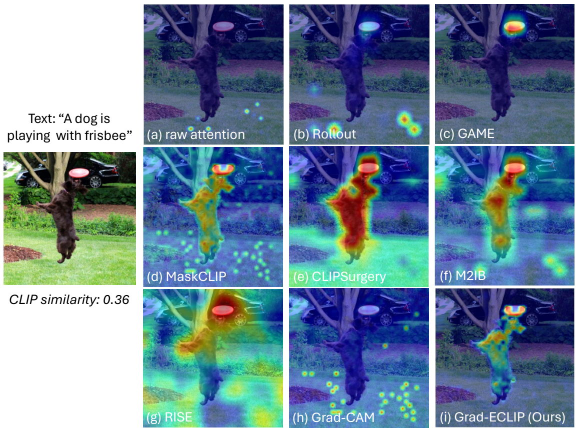


Figure 1: A visual explanation of CLIP for the image with the text "A dog is playing with frisbee" using (a) raw attention in the last layer; (b) Rollout; (c) GAME; (d) MaskCLIP; (e) CLIPSurgery; (f) M2IB; (g) RISE; (h) Grad-CAM; and (i) Our method Grad-ECLIP.

Previous visual explanation works have considered interpreting the transformer architecture used by CLIP. Attention Rollout (Abnar & Zuidema, 2020) generates explanations by aggregating attention maps computed along the forward pass of the model. Relevance-based methods (Chefer et al., 2021b;a) apply Layer-wise Relevance Propagation (LRP) (Bach et al., 2015) and also rely on the attention mechanism in the model architecture. Since Rollout and many LRP variants are class-agnostic, Transformer interpretability (Chefer et al., 2021b) and Generic Attention-Model Explainability (GAME) (Chefer et al., 2021a) build class-specific relevance-based explanations using the self-attention or co-

[1] Department of Computer Science, City University of Hong Kong, HongKong [2] SenseTime Group Ltd. Correspondence to: Chenyang Zhao <chenyzhao9-c@my.cityu.edu.hk>, Antoni B. Chan <abchan@cityu.edu.hk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

attention. However, simply treating CLIP as a vision transformer (ViT) and generating visual explanations based on self-attention sometimes leads to confusing results because of the sparse attention map (see Fig. 1a-c).

ECLIP (Li et al., 2022b) and CLIPSurgery (Li et al., 2023) (see Fig. 1e) explore explanations for CLIP by computing an image-text similarity map, and solve the counter-intuitive problem that background patch features have higher similarity with the text feature than the foreground. However, to obtain reasonable similarity maps, new additional projection layers or structure changes of the original CLIP are required. Although the parameters of CLIP encoders are frozen, learning more black-box parameters with extra data or modifying original model architecture makes the explanation less interpretable. MaskCLIP (Zhou et al., 2022a) also provides a technique to calculate class-specific image-text similarity map. By passing the value features of the last attention layer through later linear layers as image patch features, the similarity map is able to localize the concept in the text (see Fig. 1d), but has noisy backgrounds and confusingly highlights points on the locations unrelated to the explained target. The disadvantage of these similarity-map methods is that they are only forward processing, and the attended features are not necessarily used in the final prediction.

To better focus on the discriminative features used in the prediction, gradient-based methods with class-activation maps (CAM) (Zeiler & Fergus, 2014), such as Grad-CAM (Selvaraju et al., 2017), Layer-CAM (Jiang et al., 2021) and FullGrad (Srinivas & Fleuret, 2019), consider the gradient of the prediction with respect to features from a CNN layer as weights, and locates the class-specific discriminative regions by weighted aggregation of the features maps. Fig. 1h shows the visualization when adapting Grad-CAM on CLIP, where the cosine similarity of image-text pair is adopted as the prediction and the gradients are calculated w.r.t. the patch tokens from the ViT layers. Since there are no gradients w.r.t. the patch tokens in final layer because they are not involved in the calculation of the matching score, feature outputs from the penultimate layer of ViT are adopted. However, the results of Grad-CAM are do not well explain CLIP and suffer from the same problem of highlighting unrelated points as MaskCLIP, which suggests that the layer features of ViT are not suitable for CAM methods.

In this paper, we explore a more effective way to interpret CLIP, by analyzing how CLIP obtains the final feature embedding, and deriving the relationship between the embedding and intermediate features. Based on the CAM principle, we propose a novel gradient-based visual explanation method for CLIP (Grad-ECLIP), which generates the importance heat map by aggregating the intermediate features with result-related *channel* and *spatial* importance weights. Our proposed method uses the gradients of the image-text matching score w.r.t. the attention layer as the importance for feature channels. For the spatial importance, because the softmax attention typically yields sparse attention maps, we propose a loosened attention map for computing the spatial importance, which can better reflect the importance of more regions, as compared to directly using the strict softmax attention. Then our Grad-ECLIP explanation map is calculated with the *values* in the attention layer as the feature map, weighted by the channel and spatial importances. Note that Grad-ECLIP is result-specific and is suitable for both the image and text encoders, i.e., the visual explanation on image is text-specific and the word attention degrees in a sentence is image-specific. The evaluations in experiments show the superiority of our proposed Grad-ECLIP compared with other explanation methods. Finally, using Grad-ECLIP, we further conduct a visualization-based analysis on CLIP, and reveal working mechanisms and advantages/limitations of the CLIP model. We hope our proposed method can be helpful for researchers to explore more properties of vision-language models.

In summary, the contributions of this paper are:

1. We propose Grad-ECLIP, a gradient-based visual explanation approach for CLIP to produce high-quality result-specific heat maps for explaining the matching of image-text pairs.
2. We demonstrate the superiority of the proposed Grad-ECLIP with comprehensive evaluations comparing with the state-of-the-art explanation methods for Transformers and CLIP.
3. By using Grad-ECLIP, we explore the properties of CLIP, and reveal the model's ability of concept decomposition and addibility, as well as strengths and weaknesses in attribution identification.

2. Related Work

Contrastive language-image pre-training. Many multimodal works have been developed and focus on the interaction of computer vision and natural language processing, such as text-image retrieval (Wang et al., 2019b), image captioning (Xu et al., 2015), visual question answering (Antol et al., 2015), and visual grounding (Plummer et al., 2015). Contrastive language-image pre-training (CLIP) performs contrastive learning on very large-scale web-curated image-text pairs. It shows promising pre-trained representations with superior zero-shot transfer ability on diverse datasets and impressive fine-tuning performance on various downstream tasks. Subsequent works extend and improve CLIP from different aspects: Zhou et al. (2022b); Chen et al. (2022) improve the aspects of prompt design and optimization; Yu et al. (2022); Li et al. (2022a) unifies the vision-language understanding and generation by adding text decoders with image-text cross-attention during pre-training; Li et al. (2020); Wang et al. (2023); Zhong et al. (2022)

builds an alignment between region feature or position information with fine-grained object descriptions. Although significant results have been achieved with CLIP and its development, less effort and exploration is focused on its interpretability through visual explanations. In this paper, we propose a novel visual explanation method, which generates high-quality heat maps that reveal the important regions or words used for CLIP's scoring of an image-text pair.

Explainability in computer vision. Since visualizing the importance of input features is a straightforward approach to interpret a model, many works visualize the internal representations of CNNs or Transformers with heat maps. Most explanation methods can be categorized as: CAM methods, perturbation methods, Shapley-value methods, or attribution propagation (relevance-based) methods.

CAM methods, such as CAM (Zeiler & Fergus, 2014), Grad-CAM (Selvaraju et al., 2017), and Grad-CAM++ (Chattopadhay et al., 2018), generate the explanation heat map from a selected feature layer by linearly aggregating the activation maps with weights that indicates each feature's importance. Grad-CAM computes the weights with global average pooling on the gradients of the model's prediction w.r.t. the feature layer. Gradient-free CAMs (Ramaswamy et al., 2020; Wang et al., 2020b;a) generate weights from the prediction score changes when perturbing features.

Perturbation-based methods (Ribeiro et al., 2016; Petsiuk et al., 2018; Fong & Vedaldi, 2017; Lundberg & Lee, 2017; Wagner et al., 2019; Lee et al., 2021; Petsiuk et al., 2021) perturb the input and observe the changes in output scores to determine the importance of input regions. Such black-box methods are intuitive and highly generalizable, but computationally intensive. The quality of these methods are often greatly influenced by the type or resolution of the perturbations used. While having solid theoretical justification, Shapley-value methods (Lundberg & Lee, 2017) also suffer from large computational complexity.

The attribution propagation methods recursively decompose the network output into the contribution of early layers, based on the Deep Taylor Decomposition (DTD) (Montavon et al., 2017). LRP (Bach et al., 2015) and its variants (Lundberg & Lee, 2017; Nam et al., 2020; Shrikumar et al., 2017) propagate relevance from the prediction to the input image based on DTD and generate class-agnostic explanations, while Contrastive-LRP (Gu et al., 2019) and SG-LRP (Iwana et al., 2019) generate class-specific explanations. Some works (Qiang et al., 2022; Xie et al., 2022; Yu & Xiang, 2023) are proposed to interpret Transformers. Abnar & Zuidema (2020) proposed an Attention flow and Rollout method, which is based on all attention maps in the forward process of model. Since Rollout is class-agnostic, Transformer interpretability (Chefer et al., 2021b) and GAME (Chefer et al., 2021a) build class-specific relevance-based method for explaining transformer with the internal attention mechanism. However, we found that the explanation methods relying on attention maps in Transformer cannot generate satisfactory results with CLIP, possibly because the sparse attention patterns on the $\mathrm{softmax}$ map. The recent M2IB (Wang et al., 2024) applies information bottleneck principle to CLIP, which develop an optimization objective to find the compressed representations for both image features and text features. However, a series of hyperparameters are adopted during the optimization, which limits the generalization in practical application.

The similarity based methods (Li et al., 2022b; 2023; Zhou et al., 2022a), which use the cosine similarity map between the image local features and the text features as the explanation map, have the disadvantage that they are only based on the forward (bottom-up) process and thus the attended features are not necessarily used in the final prediction. In contrast, we propose Grad-ECLIP as an effective approach to interpret CLIP, which highlight features that have largest influence on the prediction as measured by the gradient, which is a top-down process.

3. Method

Our method serves as a gradient-based visual explanation for interpreting the CLIP matching performed on image-text pairs. We start with a brief introduction of CLIP. Then, by decomposing the layers of the transformer and exploring the relationship between the final output and intermediate features, we give the formulation of our gradient-based visual explanation for CLIP (Grad-ECLIP).

3.1. Preliminary on CLIP

CLIP learns both visual and language representations from large-scale raw web-curated image-text pairs. It consists of an image encoder $\mathcal{I}(\cdot)$ and a text encoder $\mathcal{T}(\cdot)$, which are jointly trained to respectively extract image and text feature embedding in a unified representation space. Given image-text pair (I, T), the matching score between their extracted image feature $f_I \in \mathbb{R}^D$ and text feature $f_T \in \mathbb{R}^D$ is:

$$S(f_I, f_T) = \cos(f_I, f_T) = \frac{f_I f_T^\mathsf{T}}{\|f_I\|\|f_T\|}. \qquad (1)$$

Contrastive learning is used on the matching scores, regarding ground-truth image-text pairs as positive samples and other mismatched pairs as negatives. Here we focus on the CLIP model where both encoders are transformers. Our method is derived based from the transformer architecture, and thus is suitable for interpreting both image and text encoders. In the followed section we present our Grad-ECLIP from the image viewpoint, where the visualization is generated on the input image I and shows important regions related to producing the matching score $S_T(f_I) \triangleq S(f_I, f_T)$, with the given specific text prompt T. The application of Grad-ECLIP from the text viewpoint, where the visualiza-

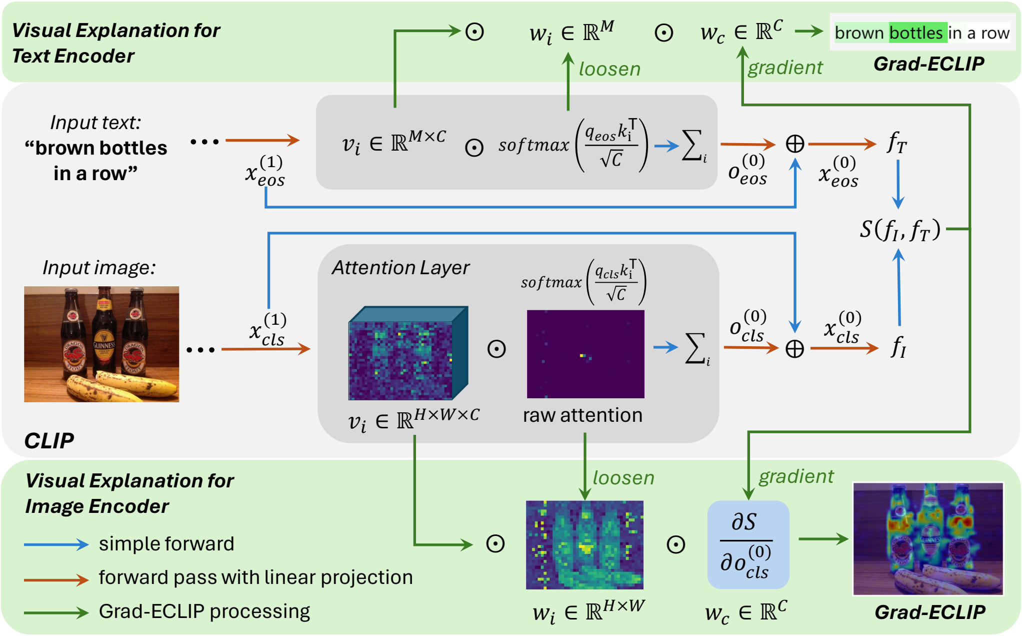

Visual Explanation for Text Encoder
$w_i \in \mathbb{R}^M$
$w_c \in \mathbb{R}^C$
brown bottles in a row
loosen
gradient
Grad-ECLIP
Input text: "brown bottles in a row"
$x_{eos}^{(1)}$
$v_i \in \mathbb{R}^{M\times C}$
$softmax\left(\frac{q_{eos}k_i^\top}{\sqrt{C}}\right)$
$o_{eos}^{(0)}$
$x_{eos}^{(0)}$
f_T
$S(f_I, f_T)$
Input image:
$x_{cls}^{(1)}$
Attention Layer
$softmax\left(\frac{q_{cls}k_i^\top}{\sqrt{C}}\right)$
$v_i \in \mathbb{R}^{H\times W\times C}$
raw attention
$o_{cls}^{(0)}$
$x_{cls}^{(0)}$
f_I
CLIP
Visual Explanation for Image Encoder
simple forward
forward pass with linear projection
Grad-ECLIP processing
loosen
gradient
$\frac{\partial S}{\partial o_{cls}^{(0)}}$
$w_i \in \mathbb{R}^{H\times W}$
$w_c \in \mathbb{R}^C$
Grad-ECLIP

Figure 2: Illustration of our method. Image-text pair specific visual explanation is generated by weighting and aggregating the *values* as feature map in attention layer with spatial importance w_i and channel importance w_c. Gradients are propagated to the attention layer output to produce w_c, and the loosened attention map is applied as w_i.

tion is generated for the text prompt T given the input image I, can be obtained analogously by considering the $[eos]$ token (end of sentence token) from the text encoder, which is analogous to the $[cls]$ token in the image encoder.

3.2. Grad-ECLIP

Following convention, we denote $x^{(n)}$ as the input of layer $L^{(n)}$ and output of layer $L^{(n+1)}$, where $n \in [0...N]$ is the index in a Transformer that consists of N layers. $x^{(N)}$ is the input of the network, $x^{(1)}$ is the input of the last layer and $x^{(0)} = \mathcal{I}(x^{(N)})$ is the output of the network. The image feature is $f_I = \mathcal{LP}(x^{(0)}[cls])$, where $\mathcal{LP}$ denotes linear projections, and $[cls]$ represents the operation to get the feature vector on the class token. Thus, except for the class token, all the final layer features of the other tokens (image patch tokens) are not used during contrastive learning of CLIP. Therefore, to interpret the $S_T(f_I)$ w.r.t. image feature map, we explore the relationship between the last layer class token feature $x_{cls}^{(0)}$ and the intermediate spatial feature maps.

As shown in the illustration of Fig. 2, looking closely into the last layer of the network, the image embedding from visual encoder can be formulated as:

$$f_I = \mathcal{LP}(x_{cls}^{(0)}) = \mathcal{LP}(\mathcal{A}(x^{(1)}) + x^{(1)})[cls] \quad (2)$$

$$= \mathcal{LP}(o_{cls}^{(0)}) + \mathcal{LP}(x_{cls}^{(1)}) \quad (3)$$

$$= \mathcal{LP}\big(\sum_i \text{softmax}(\tfrac{q_{cls}k_i^\top}{\sqrt{C}})v_i\big) + \mathcal{LP}(x_{cls}^{(1)}), \quad (4)$$

where the output of attention layer on class token is

$$o_{cls}^{(0)} = \mathcal{A}(x^{(1)})[cls] = \sum_i \text{softmax}(\tfrac{q_{cls}k_i^\top}{\sqrt{C}})v_i, \quad (5)$$

and $\mathcal{A}$ represents the attention layer in the Transformer, q_{cls} is the *query* embedding for the class token, while k_i and v_i represent the *key* and *value* embeddings at spatial location i, with C as their channel dimension. The softmax operation inside the attention layer measures the weight of the value on each location. Multi-heads are usually used in the attention layer to group the channel of $\{q, k, v\}$ into several heads, and (5) is operated inside each head with the softmax calculated over subsets of the channels. Then the final attention layer output is obtained by concatenating the results of each head together. In practice, we formulate the o_{cls} with one attention head in the forward pass and operate the softmax over all channels as (5). We discuss the influence multi-heads to visual explanation in the Appendix.

Then, with **only considering the last layer**, the relationship between the matching score and the spatial feature map can be approximately formed as:

$$S_T\left(f_I\right) \approx S_T\left(w_c o_{cls}\right) \approx S_T(w_c \sum_i w_i v_i), \quad (6)$$

where w_c and w_i are the linear weights on channel and spatial dimensions. Thus, the target-specific heat map is

$$H_i = \text{ReLU}(\sum_c w_c w_i v_i), \quad (7)$$

which is produced by summarizing the feature maps with weight w_c and w_i that capture the importance of the c-th channel and i-th location, respectively,

$$w_c = \frac{\partial S_T(f_I)}{\partial o_{cls}}, \quad w_i = \Phi(q_{cls}k_i^{\mathsf{T}}), \tag{8}$$

where Φ is a normalization function discussed next.

The channel importance w_c: The previous gradient-based works (Ribeiro et al., 2016; Selvaraju et al., 2017; Chattopadhay et al., 2018) have shown that the partial derivative w.r.t. the intermediate feature can reflect the influence of feature elements on the final output. Thus, based on (6), we set the weight of channels according to the gradient of the explanation target, which is the similarity of the image-text pair, with respect to the output class token feature of attention layer, as in (8).

The spatial importance w_i: Eq. 5 shows the softmax attention represents the importance of the values at each location. However, from the visualization, we discover that the $\mathrm{softmax}$ self attention function is extremely sparse. Important information may be encoded in different locations, but the $\mathrm{softmax}$ only selects the largest activation, which is not appropriate as a spatial weight. Thus, when calculating the spatial weight, we replace the $\mathrm{softmax}$ with a looser similarity function , which is the inner product between q_{cls} and k_i, normalized by Φ to $[0, 1]$. In the experiments, we compare using the loosened w_i and without w_i to show the effect of spatial weights, qualitatively and quantitatively.

Grad-ECLIP generates visual explanation for the CLIP encoder by (7) with weights in (8) using the last layer values v as the feature map. Finally, based on (4), the explanation can be **aggregated over all the layers** by recursively processing each layer. In the experiments, we use the last layer to explain the image encoder, and the last eight layers for interpreting the text encoder. The ablation study for the influence of different number of layers involved in image and text explanation is shown in Appendix.

4. Experiments

In this section we conduct experiments on Grad-ECLIP to: 1) evaluate its visual explanation qualitatively and quantitatively, and compare with the current SOTA methods; 2) evaluate the processing time; 3) gain insight about CLIP by analyzing the visual explanations.

We conducted the experiments with the ViT-B/16 architecture. We considered two versions of our approach: the full version of Grad-ECLIP using w_i defined in (8), and a version without w_i (denoted as "w/o w_i") that replaces the proposed spatial weights with $w_i = 1$. We compared with representative baseline methods from the four categories: 1) attention map-based *Rollout* (Abnar & Zuidema, 2020), which takes into account all the attention maps computed along the forward pass, and *raw attention* in the last visual encoder layer, both of which are not result-specific explanation; 2) classical gradient-based method *Grad-CAM* (Selvaraju et al., 2017), which takes the image-text similarity as target and calculate the gradients w.r.t. the ViT layer output; 3) relevance-based *GAME* (Chefer et al., 2021a), which integrates the relevancies and gradients propagated through the network; 4) cosine-based *MaskCLIP* (Zhou et al., 2022a) and *CLIPSurgery* (Li et al., 2023), which generates a similarity value on each location by the cosine between text feature and processed values as local image features. 5) *M2IB* (Wang et al., 2024), which applies an information bottleneck principle to generate explanation maps for CLIP. Each baseline is built with different properties and assumptions over the architecture. We also show visualization comparisons with the typical black-box perturbation method *RISE* (Petsiuk et al., 2018), but did not conduct quantitative comparisons with black-box perturbation and Shapley methods, due to their computational complexity and inherent differences with our proposed approach.

4.1. Qualitative evaluation

Comparison of visual explanations. We compare the visualizations of raw attention, Rollout, Grad-CAM, GAME, MaskCLIP, CLIPSurgery, M2IB, RISE and our Grad-ECLIP (w/ or w/o w_i in Eq. 7) in Fig. 3 with the images from ImageNet (Russakovsky et al., 2015) and MS COCO (Lin et al., 2014). Except raw attention and Rollout, which are defined to be text-agnostic, the others are all text-specific, so we test the same image with two different text inputs on MS COCO. Our Grad-ECLIP demonstrates a strong ability of generating clear and distinct text-specific heat maps, and gives reasonable explanation of verbs for interpreting CLIP. For example, the highlights for "holding" focus around the person's hands (the 5th row of Fig. 3k), while "standing" highlights the person's legs (the 6th row of Fig. 3k). We also notice that the sticks in the background are highlighted, which is probably because the sticks are regarded as "standing" on snow.

Compared with the full Grad-ECLIP, the variant w/o w_i contains more noise near object boundaries and on the background (Fig. 3j), but are otherwise consistent with full Grad-ECLIP. The result of using $w_i = \mathrm{softmax}$ (Fig. 3i) is equivalent to raw attention (Fig. 3a). In contrast to our methods, Grad-CAM and MaskCLIP can produce highlights on the explained object, but both also generate significant noise. CLIPSurgery tends to put high and coarse attention on the text target region, but also contains background noises. M2IB and MaskCLIP fails when the texts are verbs, while RISE performs the worst at interpreting CLIP. The results of GAME and Rollout, which are both based on self-attentions of the model, generate confusing heat maps due to the sparse attention between tokens in some layers.

Explanations on image/sentence pairs. The explanation map from Grad-ECLIP can also be generated from text encoder viewpoint. Using the gradient of matching score and

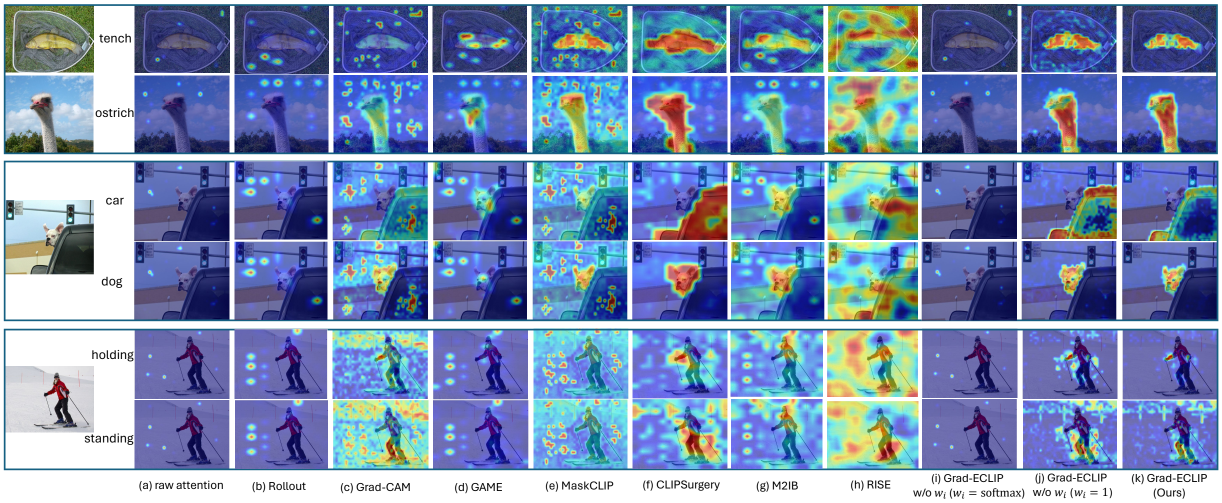


Figure 3: Comparison of heat maps from: (a) raw attention map in the last ViT; (b) Rollout (Abnar & Zuidema, 2020); (c) Grad-CAM (Selvaraju et al., 2017); (d) GAME (Chefer et al., 2021a); (e) MaskCLIP (Zhou et al., 2022a); (f) CLIPSurgery (Li et al., 2023); (g) M2IB (Wang et al., 2024); (h) RISE (Petsiuk et al., 2018); (i) variants of our proposed Grad-ECLIP using softmax attention as w_i, or (j) using $w_i = 1$ (w/o w_i); (k) our proposed Grad-ECLIP. Visual explanations are provided for the matching score between the image and the specific text prompts, which can be nouns (*e.g.*car, dog) or verbs (*e.g.*holding, standing). From the visualization comparison, Grad-ECLIP shows the superior explanation ability on different types of text prompts.

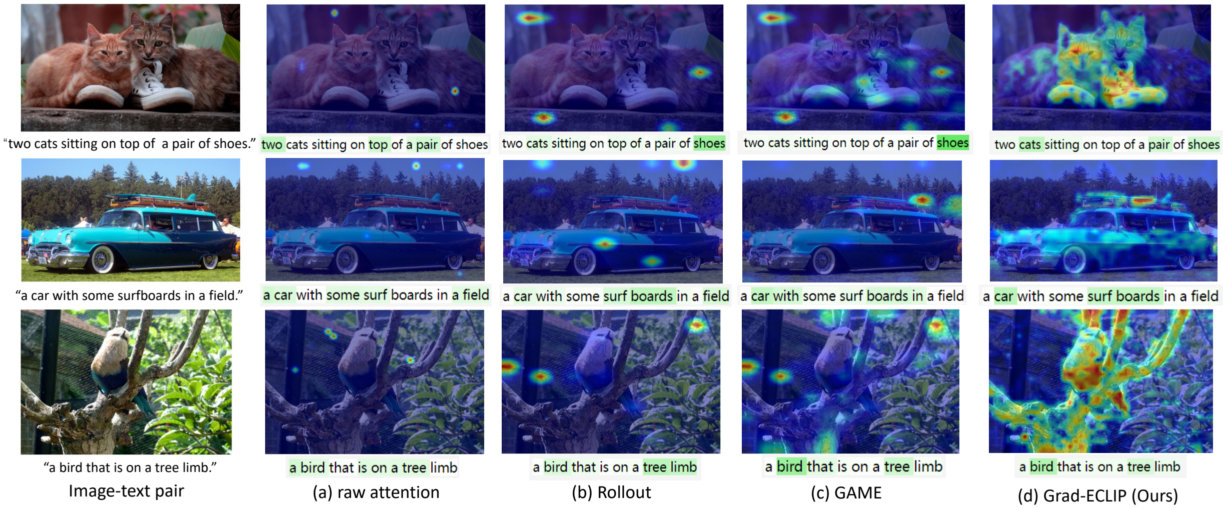

"two cats sitting on top of a pair of shoes."
two cats sitting on top of a pair of shoes
two cats sitting on top of a pair of shoes
two cats sitting on top of a pair of shoes
two cats sitting on top of a pair of shoes
"a car with some surfboards in a field."
a car with some surf boards in a field
a car with some surf boards in a field
a car with some surf boards in a field
a car with some surf boards in a field
"a bird that is on a tree limb."
a bird that is on a tree limb
a bird that is on a tree limb
a bird that is on a tree limb
a bird that is on a tree limb
Image-text pair
(a) raw attention
(b) Rollout
(c) GAME
(d) Grad-ECLIP (Ours)

Figure 4: Explanation for image-text pairs from MS COCO using: by (a) raw attention, Transformer interpretation methods (b) Rollout, (c) GAME, and our method (d) Grad-ECLIP. The importances of words are visualized by the degree of green color.

Table 1: Faithfulness evaluation of **image** explanation on the *ImageNet* validation dataset: AUC for Deletion and Insertion curves, based on Top-1 (@1) or Top-5 (@5) classification accuracy. Either the ground-truth or the prediction are used as the text input into CLIP. The second best is shown with underline.

	Deletion↓				Insertion↑			
	Ground-truth		Prediction		Ground-truth		Prediction	
Method	@1	@5	@1	@5	@1	@5	@1	@5
raw attention	0.3831	0.6239	-	-	0.2492	0.4195	-	-
Rollout	0.4082	0.6556	-	-	0.2803	0.4665	-	-
Grad-CAM	0.3417	0.5628	0.3518	0.5817	0.2682	0.4454	0.2526	0.4206
GAME	0.3356	0.5734	0.3497	0.5938	0.3611	0.5636	0.3425	0.5384
MaskCLIP	0.2848	0.4885	0.2886	0.4957	0.3335	0.5351	0.3275	0.5267
CLIPSurgery	0.3115	0.5235	0.3217	0.5412	0.3832	**0.6021**	**0.3727**	0.5719
M2IB	0.3630	0.5953	0.3633	0.5951	0.3351	0.5411	0.3347	0.5410
Ours w/o w_i	0.2535	0.4379	0.2634	0.4568	0.3715	0.5831	0.3528	0.5556
Ours	**0.2464**	**0.4272**	**0.2543**	**0.4420**	**0.3838**	0.5993	0.3672	**0.5749**

Table 2: Evaluation of **text** explanation faithfulness on *MS COCO image-text retrieval (Karpathy's split)* validation dataset: AUC for Deletion and Insertion curves with reporting image retrieval (IR) and text retrieval (TR) performance.

	Deletion↓		Insertion↑	
Method	IR	TR	IR	TR
raw attention	0.2843	0.4917	0.0065	0.0328
Rollout	0.1221	0.2389	0.1052	0.2070
GAME	0.1083	0.2084	0.1146	0.2301
M2IB	0.2139	0.4256	0.0063	0.0375
Ours w/o w_i	0.1116	0.2113	0.1123	0.2361
Ours	**0.0996**	**0.1770**	**0.1292**	**0.2536**

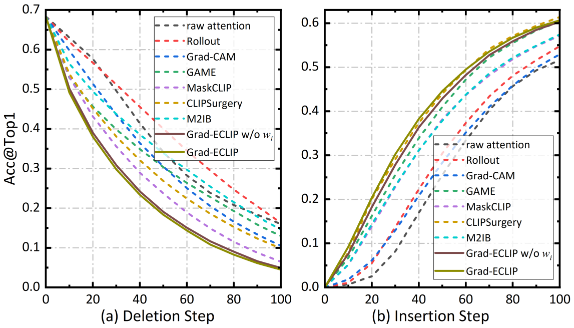


Figure 5: Classification accuracy at Top-1 vs. (a) Deletion steps and (b) Insertion steps, on the *ImageNet* validation dataset with explanation heat maps from our Grad-ECLIP (solid) and other methods (dash).

Table 3: Evaluation of localization ability using the Point Game (PG and PG-energy) and Segmentation test (Pixel Acc., AP and MaskIoU) on the *ImageNet-S* validation dataset.

Method	PG	PG-energy	Pixel Acc.	AP	maskIoU
raw attention	0.1219	0.1321	0.0278	0.2877	0.0013
Rollout	0.1375	0.2835	0.2524	0.3345	0.011
Grad-CAM	0.1845	0.3154	0.5457	0.4050	0.1251
GAME	0.4706	0.4438	0.4765	0.4072	0.089
MaskCLIP	0.4041	0.1408	0.718	0.4557	0.2481
CLIPSurgery	0.5759	0.3983	**0.7546**	0.4608	**0.3471**
M2IB	0.264	0.3557	0.6194	0.4003	0.1474
Ours w/o w_i	0.8356	0.4409	0.7365	0.5163	0.3314
Ours	**0.8899**	**0.5997**	0.7056	**0.5662**	0.2869

the feature embeddings of word tokens, Grad-ECLIP can show the importance of each word in the given sentence when matching with an image. Fig. 4 shows example explanations for image-text pairs from MS COCO. Although Rollout and GAME can highlight important words in the sentence, Grad-ECLIP is the only one showing good correspondence between image attention regions and important words. From the explanation of the sentence, we can identify which words are more important for CLIP when matching with the specific image, and conversely the text-specific important regions on the image are shown with image explanation. This word importance visualization of the input text can be helpful when designing text prompts for image-text dual-encoders in practical applications.

4.2. Quantitative evaluation

We next perform quantitative evaluations of Grad-ECLIP comparing with baselines in this section. The explanation faithfulness is evaluated by the Deletion and Insertion metrics (Samek et al., 2016; Chattopadhay et al., 2018; Wang et al., 2020b;a; Petsiuk et al., 2021), which is also called perturbation tests (Chefer et al., 2021b;a). Moreover, we evaluate localization ability, when considering each visualization as a soft-segmentation of the image, using PointGame (Zhang et al., 2018; Zhao & Chan, 2022) and segmentation tests (Chefer et al., 2021b).

Deletion and Insertion. A faithful explanation method should produce heat maps highlighting the important content in the image that has greatest impact on the model prediction. Deletion (negative perturbation) replaces input image pixels by random values step-by-step with the important pixels removed first based on the ordering of the heat map values, while recording the drop in prediction performance. Insertion adds image pixels to an empty image step-by-step based on the heat map importance, and records the performance increase. We consider each step as 0.5% of number of image pixels, and record results for 100 steps. The model performance is measured using top-1 or top-5 zero-shot classification accuracy on the validation set of ImageNet (Russakovsky et al., 2015) (ILSVRC) 2012, consisting of 50K images from 1000 classes.

The insertion/deletion curves for top-1 accuracy are presented in Fig. 5, and the corresponding area under the curve (AUC) with top-1 and top-5 accuracy are presented in Tab. 1. Steeper drop of performance with deletion steps corresponds to a lower deletion AUC, while quicker increase of performance with insert steps outputs a higher insertion AUC. Our method obtains the fastest performance drop for Deletion and largest performance increase for Insertion compared with most related works, showing that regions highlighted in our heat maps better represent explanations of CLIP. CLIPSurgery has comparable results on the Insertion metric to ours, while performs poorly when evaluated with Deletion. The reason is that the CLIPSurgery shows similar and high heat map values on the explained target region, so the deletion operation fails to delete the most important pixels on the image at the beginning steps, which causes the deletion curve to decrease gradually, producing the high deletion value. Since CLIPSurgery can successfully locate the explanation target with high values on the map, it performs well in the Insertion test. Our method's variant without using the loosened attention (w/o w_i) has slightly worse performance, but is still better than other baselines. As with Chefer et al. (2021b;a), we also use both the ground-truth class and the predicted class as the text prompt to generate heat maps, and our method is consistent with them, showing gains when using ground-truth text prompts.

We further evaluate the faithfulness of our text explanations using the *text version* of Deletion and Insertion metric, where words are deleted or inserted based on the order of importance in the text heat map. Using images and caption annotations in MS COCO Karpathy's split, we record the image-text retrieval performance changing with total 5 steps, with one word deleted/inserted at a time. The results in Tab. 2 show that Grad-ECLIP has the highest faithfulness, i.e., the best deletion and insertion test performance, compared with the other Transformer explanation methods. This demonstrates that Grad-ECLIP also has the excellent ability for image-specific text explanation.

Table 4: Comparison of the averaged processing time per image for generating the explanation map.

Method	raw attention	Rollout	Grad-CAM	GAME	MaskCLIP	CLIPSurgery	M2IB	RISE	Grad-ECLIP(Ours)
time (s/img)	0.0117	0.0298	0.0114	0.0228	0.0117	2.9423	0.5781	6.2376	0.0165

Point Game and Segmentation Test. We next evaluate the localization ability of the visual explanations. We adopt the ImageNet-Segmentation (ImageNet-S) (Gao et al., 2022) validation set, which provides segmentation annotations on 12,419 images of 919 categories from ImageNet. Point Game (PG) is a commonly used metric to evaluate the localization correctness of visual explanation. PG counts a hit score if the location with the largest value in the text-specific heat map lies within the object region, which can be defined by the class segmentation mask. Then the PG accuracy is measured by averaging all samples. Since PG only considers the maximum point, but not the spread of the heat map, we also conduct the energy-PG (Wang et al., 2020b), which calculate the proportion of heat map energy within the ground-truth mask versus the whole map. Similar to the evaluation by Chefer et al. (2021b;a), regarding the heat maps as soft-segmentation results, we adopt pixel accuracy (Pixel Acc.), average precision (AP), and averaged mask intersection over union (maskIoU).

The evaluation results for localization are shown in Tab. 3. Both versions of Grad-ECLIP significantly outperform other explanation methods, especially on PG, which demonstrates that Grad-ECLIP can well show the attention of CLIP on the object with the correct category as the text prompt. Comparing Grad-ECLIP with and without w_i, Grad-ECLIP without w_i obtains relatively higher performance on pixel accuracy and maskIoU, since heat maps that contain more high-value pixels within the ground-truth mask have advantage on these two metrics. In Fig. 3(j,k), using w_i reduces the values on the mask while removing the surrounding noise. Due to a similar reason, CLIPSurgery obtains higher pixel accuracy and maskIoU, since it tends to put high heat map values on all the pixels of object region, and gets higher score when aggregating the heatmaps inside the object mask in these two evaluations. However, its lower PG, PG-energy and AP demonstrate that there are more high values generated outside of the object boundary. Better segmentation does not necessarily result in faithful explanations, in terms of both insertion and deletion metrics, as indicated in Table 1.

4.3. Processing time comparison

In Tab. 4, we show the average processing time per image, which counts the total duration from inputting the image and text into CLIP to obtaining the explanation map. Since the gradient can be easily and quickly obtained through the autograd function of Pytorch, both our method and Grad-CAM take similar processing time as raw attention and MaskCLIP, which obtain the map from the forward calculation and require other minor operations. Note that for gradient-based methods, the backpropagation does not need to go all the way to the input layer, but stops at an intermediate upper layer, and thus the extra computation required is not much. RISE needs the longest processing time, which is the common drawback of perturbation-based methods.

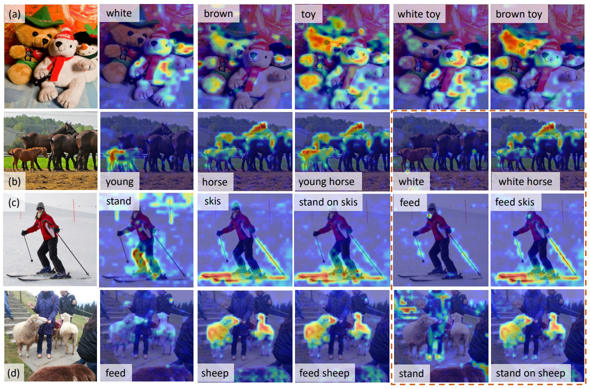


Figure 6: Visual explanation heat maps generated for single words and word phrases using Grad-ECLIP on CLIP. The dashed box contains examples where the text does not match the image.

4.4. Analysis of CLIP based on Grad-ECLIP

Useful explanation methods can be used to identify failure modes, establish appropriate users' confidence and give insight to developers to improve models. Therefore, in this section, we use the text-specific visual explanation maps generated by Grad-ECLIP and give examples of exploring the mechanism in text and image matching, and analyze the strengths and weaknesses of CLIP. We hope that our explanation tool can help researchers discover more interesting properties of language-image pre-training models, and inspire further development of these models.

Concept decomposition and addibility in image-text matching. Examining the visualizations shown in Fig. 3h, CLIP can well recognize the single concepts (nouns) and has good attention about actions (verbs). An interesting question is how does it process the combination of words, *e.g.*adjective and noun, verb and noun? To examine the working function of phrase matching, we conducted experiments comparing the explanation heat maps for single words and combined phrases using Grad-ECLIP.

The results are shown in Fig. 6. Considering adjective-noun combinations in (a), the highlights are put on all three toys when matching with "toy", and CLIP can successfully highlight the correct toy when the color adjective is included in the text. In the case of "young horse" in (b), the other horses are still highlighted, while the highlights on the young one is strengthened by adding the attribute "young". The examples of verb-noun cases in (c) and (d) also show similar addibility pattern on the heat maps: (c) with the verb "stand", the re-

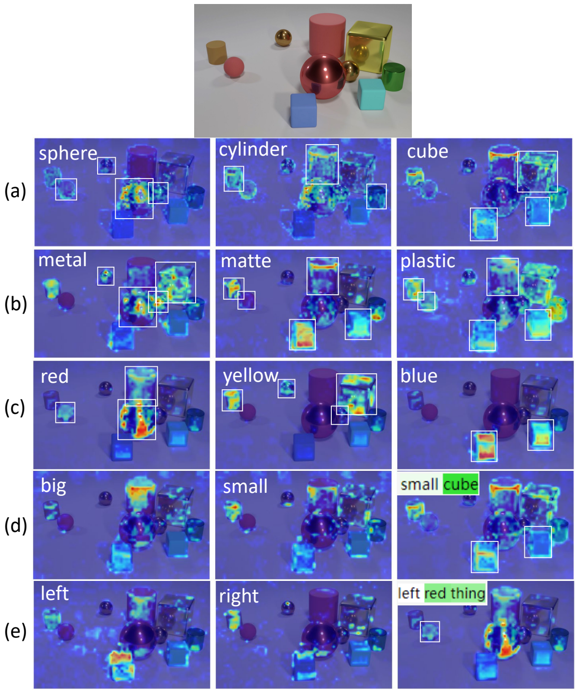


Figure 7: Visual explanations on image matching with different kinds of attributions: (a) shape; (b) material; (c) color; (d) size; (e) position. For visualization, the ground-truth corresponding to the text prompt are outlined with white boxes, except for cases involving relative adjectives, *e.g.*“big”, “small”, “left”, “right”. The text explanation maps are also shown for “small cube” and “left red thing” combinations.

gion of person’s leg is highlighted along with the“skis”; (d) with the verb “feed”, the people’s hands are also highlighted together with sheep. We also show some non-existent concepts or strange word combinations in the dashed box of Fig. 6, e.g., “white horse” in (b), “feed skis” in (c), “stand on sheep” in (d). In these cases, the visualization shows that CLIP will mainly focus on the reasonable part of the concept, such as “horse”, “skis” and “sheep”. For the non-existent “white” concept in image (b), the visual explanation does not highlight anything.

Therefore, we infer that when processing the matching of image and phrases, the model has the ability of decomposition and addibility of different concepts. This can help the model to generalize to different scenarios and could be the source of the strong zero-shot ability of CLIP.

Diagnostics on attribution identification. In Fig. 6(a), we see that CLIP has an ability to distinguish color attributes, and mark out the corresponding regions on image. To explore further, we conduct an experiment to test CLIP’s ability to identify different types of object attributes. We adopt an example image from CLEVR (Johnson et al., 2017), a diagnostic dataset for visual reasoning, and visualize image/text matching with various attributes: shape (sphere, cylinder, cube), material (metal, matte, plastic), color (red, yellow, blue), size (big, small), position (left, right).

Fig. 7 shows the visual explanation heat maps generated with each image-attribution pair. We have the following findings: 1) for shape and material, the heat maps can show partial correct attention with some obvious objects, such as the metal sphere for “sphere” and the highlighted cylinder and cube for “matte”. However, there are also false positive and false negative errors in (a) and (b). Thus, CLIP possess a certain but limited knowledge about object shapes and materials. 2) For the color attribute in row (c), the results further verify that the model can have good ability to distinguish different colors. 3) As for comparative attributes, size (big or small) in (d) and position (left or right) in (e), the visual explanations also show that CLIP produces some erroneous results. Comparing the heat map of “small cube” and “cube” in (a), “left red thing” and “red” in (c), there are little difference between them, which demonstrates that the word “cube” and “red” take the major role in the matching. This is also confirmed by the text heat maps in the figure.

Overall, from the above analysis, we can infer that CLIP has advantages with common perceptual attributes like color, but cannot well handle physical attributes like shape and material, and is weak at grounding objects with comparative attributes, like size and position relationships. Related to the addibility of concepts in the previous section, it is reasonable to expect that attributes that have concrete visual appearance, such as color, will contribute more to the matching score, compared with the abstract comparative attributes.

5. Conclusion

In this paper, we propose Grad-ECLIP, a novel white-box gradient-based visual explanation method for CLIP, the dual-encoder pre-trained model for image-text matching. Grad-ECLIP can be applied on both the image and text encoder to produce heat maps, which indicate the importance of image regions or words for the image-text matching score. Qualitative and quantitative evaluations demonstrate the advantages of our method compared with existing explanation methods designed for transformers or CLIP. We also adopt Grad-ECLIP to analyze the properties of CLIP model, where we discover its ability of concept decomposition and addibility, and advantages/limitations on different attribute identification. By introducing these analyses as examples, we hope the proposed interpretation method can be used to help with both development and understanding of language-image models. In future work, we will also explore more usages of the visual explanation for improving the contrastive pretraining model and scheme. We will also consider how to associate individual words from the sentence to regions in the image, and vice versa.

Acknowledgements

This work was supported by Strategic Research Grants from City University of Hong Kong (Project. Nos. 7005840 and 7005995) and Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone (HZQB-KCZYZ-2021045).

Impact Statement

This paper presents work whose goal is to advance the field of explainable AI. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abnar, S. and Zuidema, W. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, 2020.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Boecking, B., Usuyama, N., Bannur, S., Castro, D. C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pp. 1–21. Springer, 2022.

Cha, J., Lee, K., Park, S., and Chun, S. Domain generalization by mutual-information regularization with pretrained models. In *European Conference on Computer Vision*, pp. 440–457. Springer, 2022.

Changpinyo, S., Sharma, P., Ding, N., and Soricut, R. 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.

Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. 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.

Chefer, H., Gur, S., and Wolf, L. 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.

Chefer, H., Gur, S., and Wolf, L. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 782–791, 2021b.

Chen, G., Yao, W., Song, X., Li, X., Rao, Y., and Zhang, K. Prompt learning with optimal transport for vision-language models. *arXiv preprint arXiv:2210.01253*, 2022.

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.

Fong, R. C. and Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pp. 3429–3437, 2017.

Gao, S., Li, Z.-Y., Yang, M.-H., Cheng, M.-M., Han, J., and Torr, P. Large-scale unsupervised semantic segmentation. 2022.

Gu, J., Yang, Y., and Tresp, V. Understanding individual decisions of cnns via contrastive backpropagation. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pp. 119–134. Springer, 2019.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8340–8349, 2021a.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15262–15271, 2021b.

Iwana, B. K., Kuroki, R., and Uchida, S. Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 4176–4185. IEEE, 2019.

Jiang, P.-T., Zhang, C.-B., Hou, Q., Cheng, M.-M., and Wei, Y. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021.

Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Lee, J., Yi, J., Shin, C., and Yoon, S. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2643–2652, 2021.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., and Hoi, S. C. H. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705, 2021.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022a.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, pp. 121–137. Springer, 2020.

Li, Y., Wang, H., Duan, Y., Xu, H., and Li, X. Exploring visual interpretability for contrastive language-image pre-training. *arXiv preprint arXiv:2209.07046*, 2022b.

Li, Y., Wang, H., Duan, Y., and Li, X. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*, 2023.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., and Li, T. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K.-R. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition*, 65:211–222, 2017.

Nam, W.-J., Gur, S., Choi, J., Wolf, L., and Lee, S.-W. Relative attributing propagation: Interpreting the comparative contributions of individual units in deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 2501–2508, 2020.

Petsiuk, V., Das, A., and Saenko, K. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

Petsiuk, V., Jain, R., Manjunatha, V., Morariu, V. I., Mehra, A., Ordonez, V., and Saenko, K. Black-box explanation of object detectors via saliency maps. In *CVPR*, pp. 11443–11452, 2021.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

Qiang, Y., Pan, D., Li, C., Li, X., Jang, R., and Zhu, D. Attcat: Explaining transformers via attentive class activation tokens. *Advances in neural information processing systems*, 35:5052–5064, 2022.

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.

Ramaswamy, H. G. et al. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 983–991, 2020.

Ribeiro, M. T., Singh, S., and Guestrin, C. ” why should i trust you?” explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252, 2015.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., and Müller, K.-R. Evaluating the visualization of what a deep neural network has learned. *IEEE Trans. Neural Netw Learn Syst*, 28(11):2660–2673, 2016.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. 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.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. 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.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In *International conference on machine learning*, pp. 3145–3153. PMLR, 2017.

Srinivas, S. and Fleuret, F. Full-gradient representation for neural network visualization. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Wagner, J., Kohler, J. M., Gindele, T., Hetzel, L., Wiedemer, J. T., and Behnke, S. Interpretable and fine-grained visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9097–9107, 2019.

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019a.

Wang, H., Naidu, R., Michael, J., and Kundu, S. S. Ss-cam: Smoothed score-cam for sharper visual feature localization. *arXiv preprint arXiv:2006.14255*, 2020a.

Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., and Hu, X. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *CVPR Workshops*, pp. 24–25, 2020b.

Wang, J., Zhou, P., Shou, M. Z., and Yan, S. Position-guided text prompt for vision-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23242–23251, 2023.

Wang, Y., Rudner, T. G., and Wilson, A. G. Visual explanations of image-text representations via multi-modal information bottleneck attribution. *Advances in Neural Information Processing Systems*, 36, 2024.

Wang, Z., Liu, X., Li, H., Sheng, L., Yan, J., Wang, X., and Shao, J. Camp: Cross-modal adaptive message passing for text-image retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5764–5773, 2019b.

Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., and 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.

Xie, W., Li, X.-H., Cao, C. C., and Zhang, N. L. Vit-cx: Causal explanation of vision transformers. *arXiv preprint arXiv:2211.03064*, 2022.

Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., and Wang, X. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18134–18144, 2022.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pp. 2048–2057. PMLR, 2015.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Yu, L. and Xiang, W. X-pruner: explainable pruning for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24355–24363, 2023.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pp. 818–833. Springer, 2014.

Zhang, J., Bargal, S. A., Lin, Z., Brandt, J., Shen, X., and Sclaroff, S. Top-down neural attention by excitation backprop. *IJCV*, 126(10):1084–1102, 2018.

Zhao, C. and Chan, A. B. Odam: Gradient-based instance-specific visual explanation for object detection. In *The Eleventh International Conference on Learning Representations*, 2022.

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.

Zhou, C., Loy, C. C., and Dai, B. Extract free dense labels from clip. In *European Conference on Computer Vision*, pp. 696–712. Springer, 2022a.

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

A. Deletion and Insertion with image visual explanation on MS COCO

The results of Deletion and Insertion with *image-specific text explanation* on MS COCO Karpathy's split validation set are reported in Tab. 2. In this section, we also conduct the Deletion and Insertion experiments with *caption-specific image explanations*, which record the image and text retrieval performance changing when deleting and inserting pixels of the input image with 100 steps. From the results shown in Tab. 5, it can be seen that Grad-ECLIP surpasses the other methods on most metrics, which further demonstrates that our method produce high-quality visual explanation on both the text and image encoder, for the specific image and text pair, regardless if the text is the class categories as in ImageNet or long captions as in MS COCO.

Table 5: Evaluation of image explanation faithfulness on *MS COCO image-text retrieval (Karpathy's split)* validation dataset: AUC for Deletion and Insertion curves with reporting image retrieval (IR) and text retrieval (TR) performance.

Method	Deletion↓						Insertion↑					
	IR			TR			IR			TR		
Method	@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10
raw attention	0.1708	0.3554	0.4558	0.1923	0.3720	0.4654	0.1247	0.2552	0.3292	0.1544	0.2969	0.3477
Rollout	0.1948	0.3946	0.4977	0.2268	0.4238	0.5240	0.1294	0.2932	0.3423	0.1753	0.3503	0.3841
Grad-CAM	0.1717	0.3502	0.4462	0.2161	0.4008	0.4927	0.1027	0.2216	0.2903	0.1152	0.2327	0.2947
GAME	0.1706	0.3552	0.4560	0.1982	0.3800	0.4736	0.1537	0.3083	0.3885	**0.2097**	0.3735	0.4186
MaskCLIP	0.1321	0.2841	0.3722	**0.1516**	0.2949	0.3755	0.1423	0.2953	0.3785	0.1891	0.3514	0.4056
Ours w/o w_i	0.1390	0.2940	0.3805	0.1827	0.3386	0.4200	0.1403	0.2895	0.3729	0.1735	0.3279	0.3894
Ours	**0.1246**	**0.2670**	**0.3480**	0.1550	**0.2933**	**0.3701**	**0.1576**	**0.3203**	**0.4065**	0.2056	**0.3761**	**0.4321**

B. The influence of the number of layers used to compute the visual explanation

As introduced in Section 3.2, the explanation can be aggregated over all the layers in the Transformer by recursively processing each layer with Eq. 7. In this section, we conduct experiments to study the influence of using different number of layers to generate heat maps for image and text.

With different layer number N, the visualizations for images with specific texts are shown in Fig. 8, and the corresponding caption explanations are shown in Fig. 9. $N = 1$ means the visualization is generated only with the final Transformer layer, while $N = 12$ means all the layers are involved. The image explanations become worse when increasing the number of layers involved, since the features in lower layer may introduce more noise to the heat map. Therefore, it is the best to just use the last layer in the calculation of image visual explanation, where this conclusion is consistent with the classical gradient-based CAM methods.

As for the text explanation, there is no obvious difference of the visualization quality, since the highlights are basically focusing on "dog", "car" and "traffic lights" with some minor variations. Therefore, we perform the Deletion and Insertion experiments as in Sec. 4.2 on text explanation map with different number of layers N. The results are shown in the following Table 6. The explanation faithfulness has the trend that it first increases with more layers used and then goes down with the lower-layer features involved ($N > 8$). Therefore, we aggregate the last eight layers maps for interpreting the text encoder in our experiments.

Table 6: The **text** explanation faithfulness vs. the involved layer numbers. Evaluating on *MS COCO image-text retrieval (Karpathy's split)* validation dataset: AUC for Deletion and Insertion curves with reporting image retrieval (IR) and text retrieval (TR) performance.

	Deletion↓		Insertion↑	
N	IR	TR	IR	TR
1	0.1118	0.2087	0.1059	0.2196
2	0.1021	0.1826	0.1186	0.2351
4	**0.0995**	0.1786	0.1242	0.2428
6	0.0989	**0.1761**	0.1273	0.2490
8	0.0996	0.1770	**0.1292**	**0.2536**
10	0.1008	0.1843	0.1288	0.2472
12	0.1095	0.2087	0.1219	0.2364

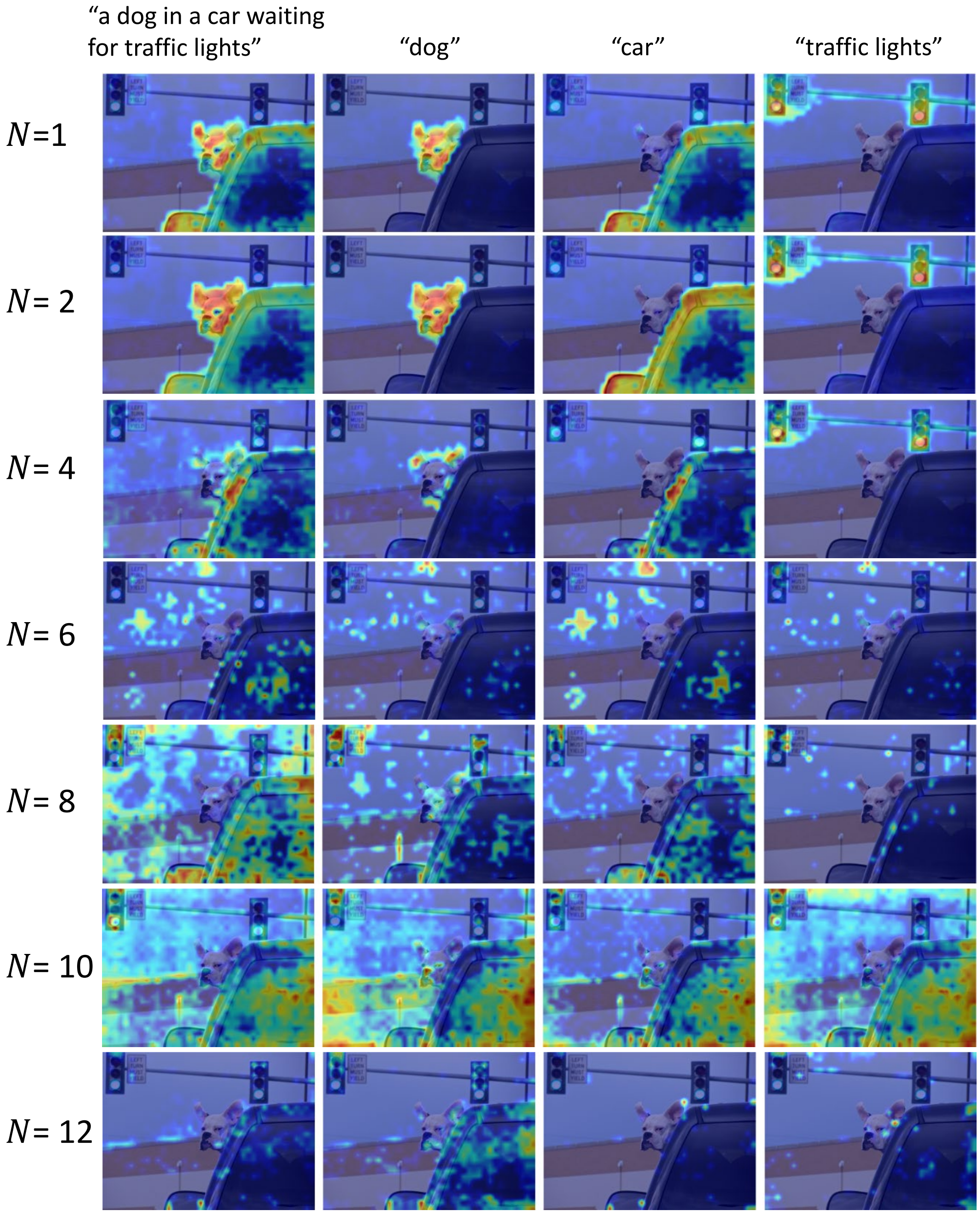


Figure 8: The *image* visual explanations generated with different N layers.

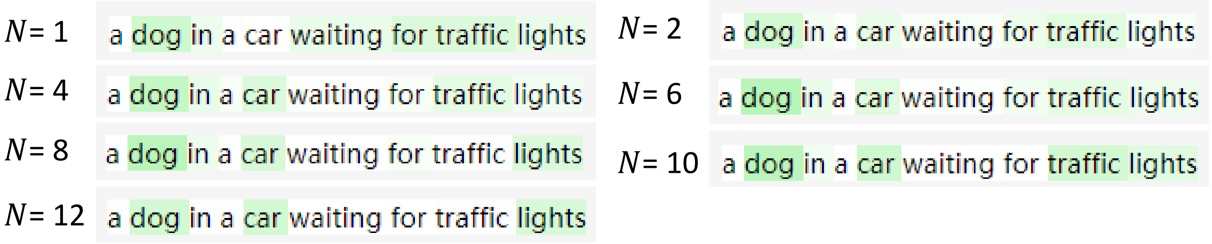
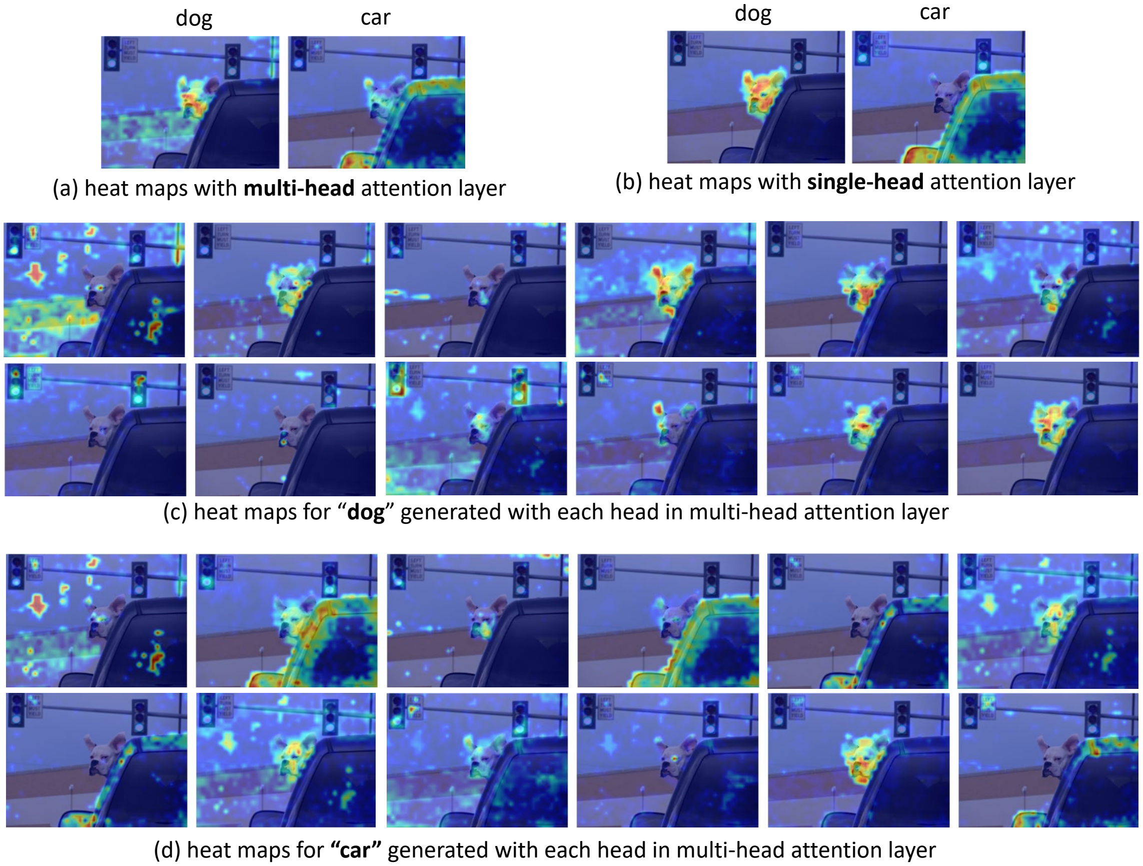

N= 1
a dog in a car waiting for traffic lights
N= 2
a dog in a car waiting for traffic lights
N= 4
a dog in a car waiting for traffic lights
N= 6
a dog in a car waiting for traffic lights
N= 8
a dog in a car waiting for traffic lights
N= 10
a dog in a car waiting for traffic lights
N= 12
a dog in a car waiting for traffic lights

Figure 9: The *text* visual explanations generated with different N layers.

C. The influence of multi attention heads on visual explanation

As mentioned in Section 3.2 (5), CLIP performs the forward pass with a single head in the attention layer instead of the original multi-head attention layer when producing the Grad-ECLIP visual explanation. In Fig. 10, we show the visualization of explanation maps when using multi-head attention layers, compared with passing forward through a single head. Comparing Fig. 10 (a) and (b), it can be seen that some surrounding context information is also highlighted with the explained object. To further investigate, we further produce the heat maps for *each* attention head, using the $q \in \mathbb{R}^{D/12}$, $k \in \mathbb{R}^{D/12}$, $v \in \mathbb{R}^{D/12}$, and attention output $o_{cls} \in \mathbb{R}^{D/12}$, where D is channel number before going into multi heads, and visualize them in Fig. 10 (c) for the target "dog", and (d) for the "car". The visual explanation in each head highlights different regions, in addition to the target object. We can infer that the channels assigned to each heads can preserve different information, and the $\mathrm{softmax}$ inside each head helps the model to encode more context information. In contrast, with the single head setting, the $\mathrm{softmax}$ is performed over all channels, which selects out the most important information, and our explanation method can show the model's attention on the specific explained target, as shown in Fig. 10 (b).



(a) heat maps with **multi-head** attention layer

(b) heat maps with **single-head** attention layer

(c) heat maps for "**dog**" generated with each head in multi-head attention layer

(d) heat maps for **"car"** generated with each head in multi-head attention layer

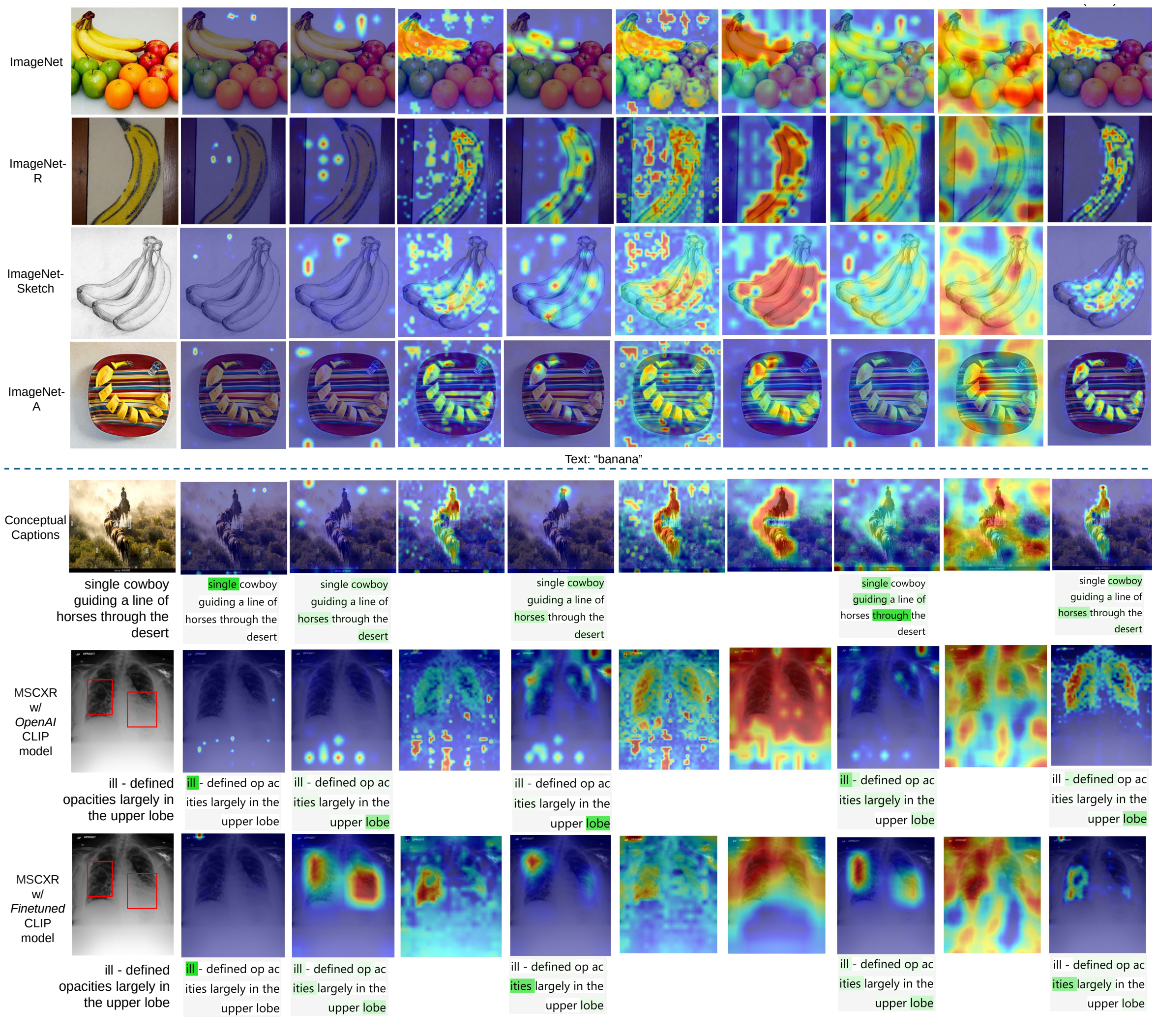
Figure 10: The visual explanation maps when using (a) multi-head attention layer; (b) single-head attention layer; (c) each head in multi-head attention layer for text "dog"; (d) each head in multi-head attention layer for text "car"

D. More visualization examples on diverse datasets of different domain

We show the visualization comparison of different methods on the samples from different image domains, including the original ImageNet and ImageNet in different domains: rendition (ImageNet-R (Hendrycks et al., 2021a)), pencil sketch (ImageNet-Sketch (Wang et al., 2019a)), natural adversarial example (ImageNet-A (Hendrycks et al., 2021b)), web images with captions (Conceptual Captions (CC) (Sharma et al., 2018)), and chest x-ray with text (MSCXR (Boecking et al., 2022)) in Fig. 11. For the image-caption pairs from the web-collected CC and chest X-ray data MSCXR, we generate explanations for both image and text encoder, and compare with the other methods that also provide text encoder explanations, including the raw attention, Rollout, GAME, M2IB.

Our Grad-ECLIP explanations provide interesting insights into how CLIP handles different image domains. In Fig. 11, given a normal banana image and text "banana" Grad-ECLIP reveals that the yellow color is dominant to CLIP. However, when given a pencil sketch without color (ImageNet-Sketch), Grad-ECLIP reveals that CLIP looks at the curvature of the banana. For the color sketch of the banana (ImageNet-R), Grad-ECLIP shows that the color of the banana is mainly used, and not the black curved lines. Thus, from these examples, we may infer that CLIP prefers using the yellow color over the curved shape for matching with the "banana" text.

Grad-ECLIP also provides interesting insights on how the original CLIP fails on novel domains. The last two rows of Fig. 11 show the explanations for chest x-ray images and text for the OpenAI CLIP model and a fine-tuned CLIP model (on MSCXR). The Grad-ECLIP explanation shows that the original CLIP uses the whole lobe to match with the words "defined" and "lobe". In contrast, the fine-tuned CLIP locates the actual anomaly and matches it with the text "defined opacities largely". The reason is that the finetuned model is trained to the specific domain that matches the X-ray and the illness location descriptions, while the original OpenAI CLIP model is more general and apparently "lobe" is the key word and the main object in the image-text pair.


ImageNet
ImageNet-R
ImageNet-Sketch
ImageNet-A
Text: "banana"
Conceptual Captions
single cowboy guiding a line of horses through the desert
MSCXR w/ OpenAI CLIP model
ill - defined opacities largely in the upper lobe
MSCXR w/ Finetuned CLIP model
ill - defined opacities largely in the upper lobe

(a) raw attention (b) Rollout (c) Grad-CAM (d) GAME (e) MaskCLIP (f) CLIPSurgery (g) M2IB (h) RISE (i) Grad-ECLIP (Ours)

Figure 11: The visual explanation comparison of different methods on samples from different image domains.

E. Adapt Grad-ECLIP to other VLMs

Since our method is designed based on CLIP encoders, which are Transformer-based, our method can be easily adopted to generate visual explanations for the models that are also Transformer-based. Here we extend Grad-ECLIP to ViT-based classifier (Dosovitskiy et al., 2020) and BLIP (Li et al., 2022a) to show the generalized applicability of our method in Fig. 12. When matching the same image-text pair, different models put attention on different regions, shown by the visual explanations. For example, BLIP notes the fins, while CLIP notes the fish body to match the image to "tench". When matching with the sentence "a dog is playing with a frisbee", BLIP puts attention on the dog on the image, while CLIP shows more attention on the frisbee.

Other VLMs like CoCa (Yu et al., 2022) and ALBEF (Li et al., 2021) add additional attention layers after the encoder, and thus our current method is not directly applicable since our method assumes that the last layer attention output has linear relationship with the final feature embedding. Our future work will investigate adapting our method to these modified ViT frameworks, e.g., to handle the attention pooling after the last layer in CoCa, or using cross attention to fuse image and text in ALBEF. Nonetheless, our ability to explain CLIP and other VLMs with similar architecture is significant considering that CLIP is by far the most widely used VLM.

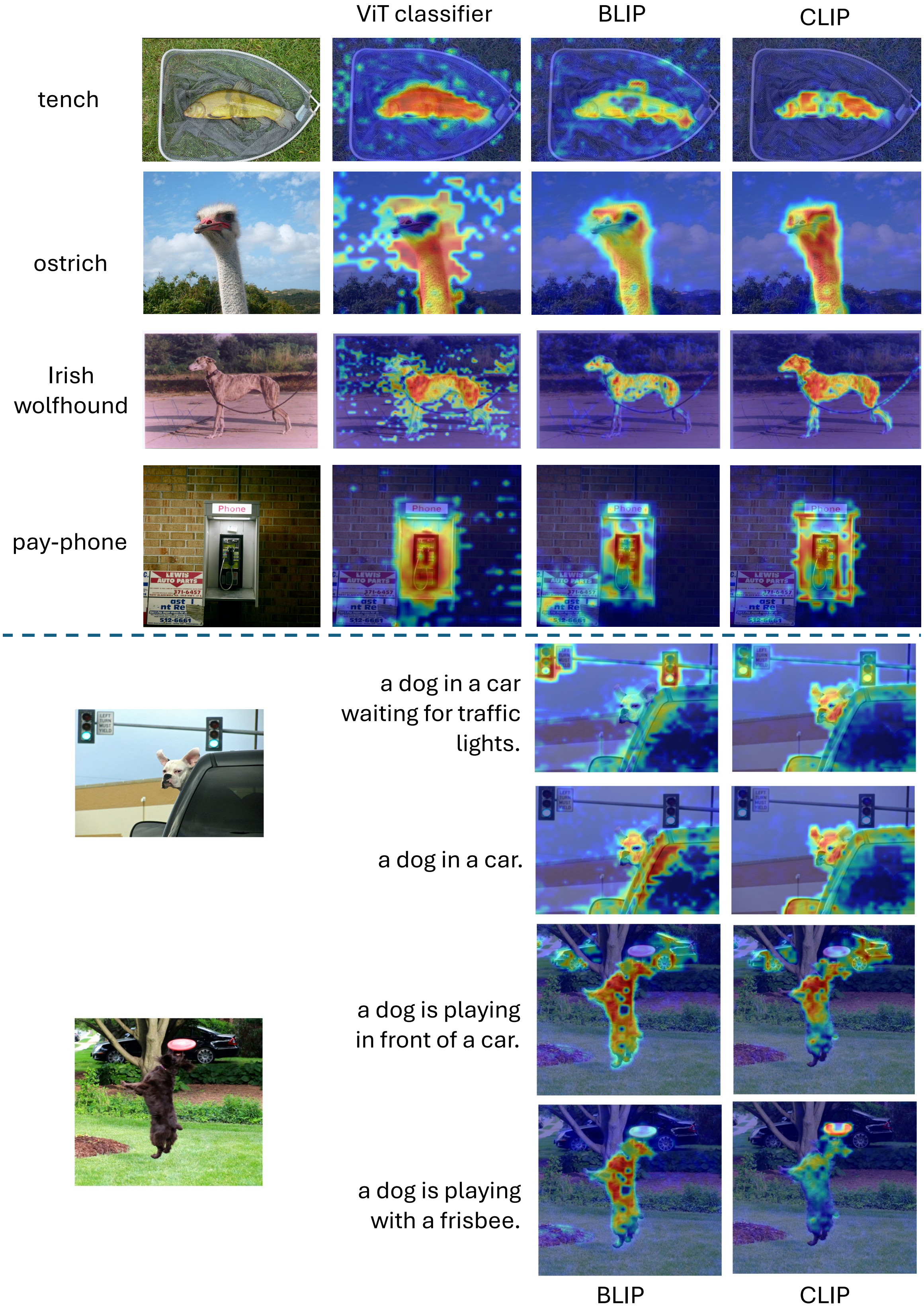

ViT classifier
BLIP
CLIP
tench
ostrich
Irish wolfhound
pay-phone
a dog in a car waiting for traffic lights.
a dog in a car.
a dog is playing in front of a car.
a dog is playing with a frisbee.
BLIP
CLIP

Figure 12: Adapt Grad-ECLIP to ViT-based classifier and BLIP .

F. More visual explanations by Grad-ECLIP

Here we present more visualization of explanation results by Grad-ECLIP on CLIP. Figure 13 shows some examples with images in *ImageNet* val dataset with the category as the text prompt. Then, Fig. 14 visualizes the explanation maps for image-text pair with samples of *MS COCO Karpathy's split* validation set. The image explanation is text-specific, while the text explanation is image-specific.

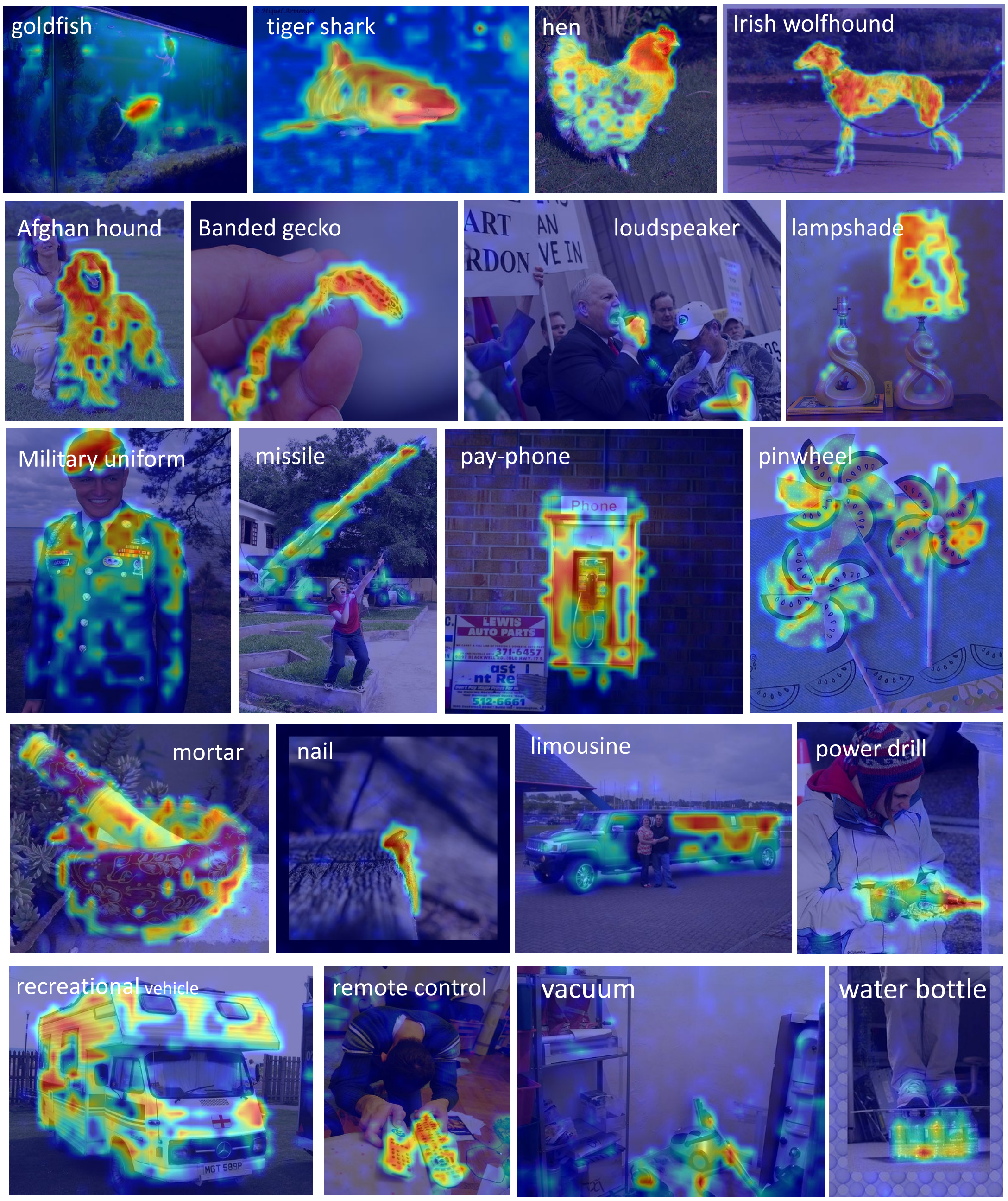
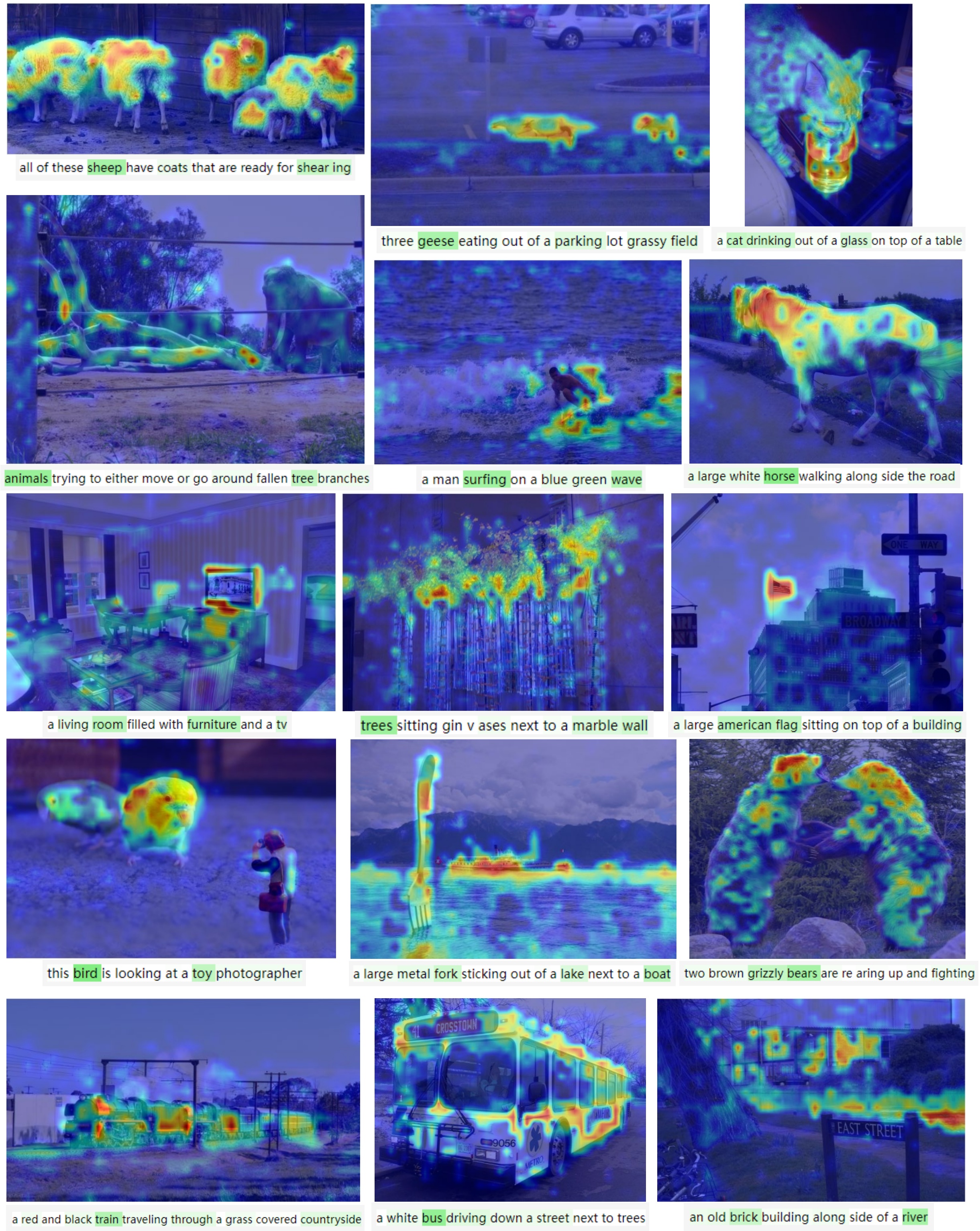


Figure 13: More visual explanation examples in *ImageNet* validation set.



all of these sheep have coats that are ready for shear ing

three geese eating out of a parking lot grassy field

a cat drinking out of a glass on top of a table

animals trying to either move or go around fallen tree branches

a man surfing on a blue green wave

a large white horse walking along side the road

a living room filled with furniture and a tv

trees sitting gin v ases next to a marble wall

a large american flag sitting on top of a building

this bird is looking at a toy photographer

a large metal fork sticking out of a lake next to a boat

two brown grizzly bears are re aring up and fighting

a red and black train traveling through a grass covered countryside

a white bus driving down a street next to trees

an old brick building along side of a river

Figure 14: More visual explanation examples for image-text pair in *MS COCO Karpathy's split* validation set.