Grad-ECLIP: Gradient-based Visual and Textual Explanations for CLIP

Chenyang Zhao, Kun Wang, Janet H. Hsiao and Antoni B. Chan,

***Abstract*—Significant progress has been made in the improvement and downstream applications of the Contrastive Language-Image Pre-training (CLIP) vision-language model, while less attention has been paid to the interpretation of CLIP. We propose a Gradient-based visual and textual Explanation method for CLIP (Grad-ECLIP), which interprets the matching result of CLIP for a specific input image-text pair. By decomposing the encoder's architecture and identifying the relationship between matching similarity and intermediate spatial features, Grad-ECLIP generates effective heat maps that reveal the impact of image regions or words on the CLIP results. Unlike previous Transformer interpretation methods that focus on utilizing self-attention maps, which are typically extremely sparse in CLIP, we produce high-quality visual explanations by applying channel and spatial weights to token features. Qualitative and quantitative evaluations verify the effectiveness and superiority of Grad-ECLIP compared with the state-of-the-art methods. Finally, a series of analyses are conducted based on our visual and textual explanation results, from which we explore the working mechanism of image-text matching, the strengths and limitations in attribution identification of CLIP, and the relationship between the concreteness/abstractness of a word and its usage in CLIP. The code of Grad-ECLIP is available here: https://github.com/Cyang-Zhao/Grad-Eclip.**

***Index Terms*—gradient-based explanation, visual and textual explanation, explainable AI, contrastive language-image pre-training**

I. Introduction

Recently, Vision-Language Pre-training (VLP) models have bridged the gap between the low-level visual information in pixels and high-level conceptual knowledge embedded in language. By leveraging self-supervised objectives on web-scale data, VLP models learn an aligned understanding of both vision and language, and can flexibly perform applications on various tasks, such as classification, image-text retrieval, visual questioning answering (VQA), image captioning, and serve as foundational technology for the visual inputs to generative AI systems. The evolution of VLP has been marked by several influential architectures. The Contrastive Language-Image Pre-training (CLIP) [1] model, which employs a dual-encoder architecture trained with a contrastive objective to facilitate image-text matching, is a notable milestone. CLIP significantly improves performance on various downstream tasks [2, 3, 4, 5, 6], utilizing both zero-shot and fine-tuning methodologies. Inspired by CLIP, VLP has been further developed by exploring different perspectives, including unifying vision-language understanding and generation [7, 8], prompt design [9, 10], and region-aware enhancement [11, 12, 13]. 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 the learned vision and language matching process.

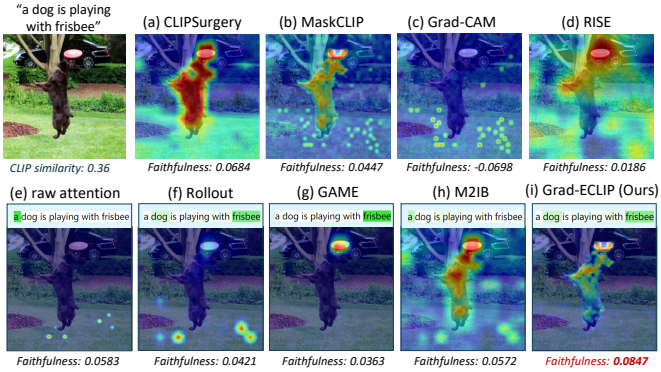

Fig. 1: Visual and textual explanations of CLIP for the image with the text "A dog is playing with frisbee" using (a) CLIPSurgery [14]; (b) MaskCLIP [15]; (c) Grad-CAM [16]; (d) RISE [17]; (e) raw attention in the last layer; (f) Rollout [18]; (g) GAME [19]; (h) M2IB [20]; and (i) Our Grad-ECLIP. For (e) to (i), textual explanations on the sentence are shown, where the degree of green color represents the word importance. Other methods (a-d) are not applicable to text. The faithfulness (IMD) of each method on the image is shown under the heat map of each method (higher is better).

Previous visual explanation works have considered interpreting the transformer architecture used by VLP models. Attention Rollout [18] generates explanations by aggregating attention maps computed along the forward pass of the model. Relevance-based methods [19, 21] apply Layer-wise Relevance Propagation (LRP) [22] and also rely on the attention mechanism in the model architecture. Since Rollout and many LRP variants are class-agnostic, Transformer interpretability [21] and Generic Attention-Model Explainability (GAME) [19] build class-specific relevance-based explanations using the self-attention or co-attention. However, just treating VLP models as a vision transformer and generating explanations based on self-attention sometimes leads to low-faithfulness and confusing visualization results because of sparse attention maps (see Fig. 1e-g). Figure 1 shows the visual explanation heat maps on CLIP with the evaluated IMD faithfulness, where the value represents an average influence on the image-text similarity score within 100 steps when deleting and inserting image pixels based on the heat maps (see §IV-B1 for more details about the IMD metric). Higher IMD values are better,

Chenyang Zhao, and Antoni B. Chan (corresponding author) are with the Department of Computer Science, City University of Hong Kong. Janet H. Hsiao is with the Division of Social Science and Department of Computer Science & Engineering, Hong Kong University of Science & Technology, and Kun Wang is with the SenseTime Group Ltd. E-mail: zhaocy2333@gmail.com, abchan@cityu.edu.hk.

indicating better fidelity to the actual important regions.

The vision-language matching interpretation is explored by ECLIP [23] and CLIPSurgery [14] with CLIP by computing an image-text similarity map. They attempt to solve the counter-intuitive problem that background patch features get higher similarity with the text feature than the foreground. However, to obtain reasonable similarity maps, new additional projection layers or changing the structure of the original CLIP are required. Although the parameters of CLIP encoders are frozen, learning more black-box parameters with extra data or modifying the original model architecture makes the explanation less interpretable. MaskCLIP [15] 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. Although the similarity maps from CLIPSurgery and MaskCLIP are able to localize the concept in the text (see Fig. 1a-b), lower faithfulness results come from a noisy background and confusingly highlight points on the location unrelated to the explained target matching text. 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) [24], such as Grad-CAM [16], Layer-CAM [25], and FullGrad [26], 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 feature maps. Fig.1c shows the faithfulness and visualization when adapting Grad-CAM on CLIP, where the cosine similarity of the 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 the final layer because they are not involved in the calculation of the matching score, feature outputs from the penultimate layer of ViT are adopted. The results of Grad-CAM do not well explain CLIP, and suffer from the same problem of highlighting unrelated points as MaskCLIP, suggesting that the layer features of ViT are unsuitable for CAM methods.

In this paper, we explore a more effective way to interpret the image-text matching in VLP models. Using the representative CLIP dual-encoders as the paradigm architecture, we analyze how CLIP obtains the final feature embedding and derive the relationship between the image and text embedding similarity score and intermediate features via a series of approximations. We first make the derivation based on the more widely used ViT-based CLIP, which is suitable for both the transformer-based image and text encoders. Then, we give the adaptation of the CNN-based architecture. Based on the CAM principle, we then propose a novel gradient-based visual explanation method for CLIP (Grad-ECLIP), which generates the importance heat map by aggregating the intermediate image 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. The same method used to generate the explanation of the image encoder can also be applied to the text encoder to obtain a textual explanation for CLIP. 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 textual explanation of a sentence is image-specific. As the example shows in Fig. 1i, Grad-ECLIP produces the best faithfulness result on this example image-text pair. The heat map on the image shows the important region when matching the image with the specific text "a dog is playing with frisbee", while the degree of green color on the sentence represents the important words, where the most important words "dog" and "frisbee" correspond to the highlighted regions in the image heat map.

In the experiments, we conduct both qualitative evaluation by visualization of explanation maps and quantitative evaluation compared with other types of explanation methods, and show the superiority of our proposed Grad-ECLIP. Moreover, in the qualitative evaluation, we demonstrate the generalizability of Grad-ECLIP by applying our method on diverse datasets of different domains and showing it is applicable to both ViT and CNN-based CLIP, as well as the ViT classifier and other transformer-based vision language models like BLIP [8]. Then, using Grad-ECLIP, we further conduct a visualization-based analysis on CLIP, and reveal working mechanisms and advantages/limitations of the CLIP model, including the type of attributes and the concreteness/abstractness of words used by CLIP. We hope our proposed method can be helpful for researchers to explore more properties of vision-language models like CLIP, as well as understand their current limitations and how this may affect downstream tasks.

Finally, we also present an application of Grad-ECLIP to fine-tuning the CLIP model to boost the fine-grained understanding. Since the ViT-based CLIP model has been shown to have limitations in producing dense representations, due to the pretraining focusing on the whole image-text matching [12, 13, 27, 28], we propose to generate detailed region-phrase corresponding pairs via Grad-ECLIP so as to enhance the fine-grained understanding of the CLIP model during fine-tuning. Experiments with zero-shot region classification and downstream open-vocabulary detection application show that the Grad-ECLIP-enabled fine-tuning is effective.

In summary, the contributions of this paper are fivefold:

1) We explore the explanation of image-text matching in VLP models by investigating Grad-ECLIP, a gradient-based visual and textual explanation approach for CLIP dual-encoders 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) We show the generalizability of Grad-ECLIP by verifying the applicability on both ViT and CNN-based CLIP, other transformer-based VLP models like BLIP, as well as ViT classifier, and presenting explanations on datasets of different domains.
4) By using Grad-ECLIP, we explore the properties of CLIP, and reveal the model's ability of concept decomposition and addability, strengths and weaknesses in attribution identification, as well as the relationship between word usage and concreteness in the image and text matching.

A preliminary version of our work appears in [29]. The extensions over the conference version are as follows. First, we provide a more detailed derivation of Grad-ECLIP, based on the transformer model. Second, we enrich the qualitative evaluation and verify the generalizability of Grad-ECLIP by adding: (1) the visualization on diverse datasets of different domains; (2) application on ViT-based classifier and other vision language models; (3) adaptation with CNN-based CLIP. Third, we include new ablation studies for the Grad-ECLIP design, including: (1) the effect of the proposed loosen spatial weight; (2) the influence of the number of layers involved in the calculation; (3) the influence of multi-attention heads on the visual explanation results. Fourth, we add quantitative experiments for the analysis of CLIP and a new exploration about the relationship between word concreteness and word usage in image-text matching with CLIP. Moreover, we conduct a user study to evaluate the trustability of the visual explanations.

The remainder of this paper is organized as follows. The related works about CLIP and explainability in computer vision are briefly reviewed in §II. Grad-ECLIP is introduced in §III, and the experiment results are presented in §IV, including qualitative evaluations, quantitative evaluations, and ablation studies. Finally, we perform analysis of CLIP based on the proposed Grad-ECLIP in §V.

II. Related Works

We first briefly review the CLIP model, and then discus different types of visual explanation methods in computer vision.

A. Contrastive language-image pre-training

Many multi-modal works have been developed and focus on the interaction of computer vision and natural language processing, such as text-image retrieval [30], image captioning [31], visual question answering [32], and visual grounding [33]. 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: [9, 10] improve the aspects of prompt design and optimization; [7, 8] unify the vision-language understanding and generation by adding text decoders with image-text cross-attention during pre-training; [11, 12, 13, 28] build 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 are 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.

B. 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 [24], Grad-CAM [16], and Grad-CAM++ [34], generate the explanation heat map from a selected feature layer by linearly aggregating the activation maps with weights that indicate 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 [35, 36, 37] generate weights from the prediction score changes when perturbing features.

Perturbation-based methods [17, 38, 39, 40, 41, 42, 43] 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 to different architectures and tasks, but computationally intensive. The quality of these methods is often greatly influenced by the type or resolution of the perturbations used. While having solid theoretical justification, Shapley-value methods [40] also suffer from large computational complexity.

Attribution propagation methods recursively decompose the network output into the contribution of early layers, based on the Deep Taylor Decomposition (DTD) [44]. LRP [22] and its variants [40, 45, 46] propagate relevance from the prediction to the input image based on DTD and generate class-agnostic explanations, while Contrastive-LRP [47] and SG-LRP [48] generate class-specific explanations.

These previous works are mainly proposed for interpreting CNN-based models. Due to the introduction of self-attention mechanisms in Transformers, recent works [49, 50, 51] have also looked at visual explanations for the Transformer architecture. [18] proposed an Attention flow and Rollout method, which is based on all attention maps in the forward process of the model. Since Rollout is class-agnostic, Transformer interpretability [21] and GAME [19] build class-specific relevance-based methods for explaining the 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 of the sparse attention patterns on the $\mathrm{softmax}$ map. The recent M2IB [20] applies the information bottleneck principle to CLIP, which develops 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 applications.

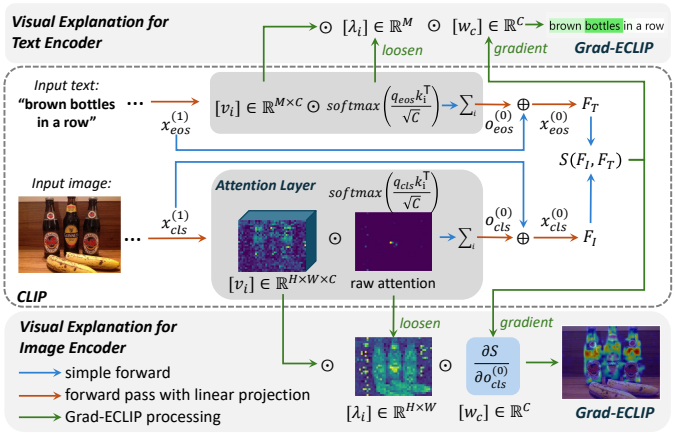

Fig. 2: Illustration of Grad-ECLIP. An image-text pair specific visual explanation is generated by weighting and aggregating the *values* as feature map in the attention layer with spatial importance λ_i and channel importance w_c. Gradients are propagated to the attention layer output to produce w_c, and the loosened attention map is applied as λ_i.

Finally, existing approaches for explaining CLIP [14, 15, 23], 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 highlights features that have the largest influence on the prediction as measured by the gradient, which is a top-down process.

III. Grad-ECLIP: Gradient-based Visual Explanation for CLIP

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

A. Preliminaries

CLIP learns both visual and language representations from large-scale 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 $F_T \in \mathbb{R}^D$ (both row vectors) is:

$$S(F_I, F_T) = \cos(F_I, F_T) = \frac{F_I F_T^\top}{\|F_I\| \|F_T\|}. \tag{1}$$

The model is trained using contrastive learning on the matching scores, regarding the ground-truth image-text pairs as positive samples and other mismatched pairs as negatives. In practice, both encoders can be implemented as transformers. In §III-B, our method is derived based on the transformer architecture, and thus is suitable for interpreting both ViT-based image and transformer-based text encoders. Alternatively, the image encoder could be a CNN-based ResNet followed by an attention pooling layer, which is basically the same as an attention layer in the Transformer. Thus, the proposed Grad-ECLIP is also applicable to the CNN-based CLIP, for which we display the visualization results for CLIP with ResNet [52] backbone in §IV-A5.

For a Transformer that consists of N layers, 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 layer index. $x^{(N)}$ 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 $F_I = \mathcal{LP}(x_{cls}^{(0)})$, where $\mathcal{LP}$ denotes linear projections, and x_{cls} is the feature vector from the $[cls]$ 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. an image feature map, we explore the relationship between the last layer class token feature $x_{cls}^{(0)} \in \mathbb{R}^C$ and the intermediate spatial feature maps.

B. Methodology of Grad-ECLIP

Here we present our derivation of 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 visualization 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.

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

$$F_I = \mathcal{LP}(x_{cls}^{(0)}) = \mathcal{LP}(o_{cls}^{(0)} + x_{cls}^{(1)}) \tag{2}$$

$$\approx \mathcal{LP}(o_{cls}^{(0)}) + \mathcal{LP}(x_{cls}^{(1)}), \tag{3}$$

where the approximation is based on assuming linearity of $\mathcal{LP}$[53, 54]. Noting that $\mathcal{LP}(x_{cls}^{(t)}) = \mathcal{LP}(o_{cls}^{(t)} + x_{cls}^{(t+1)})$, we substitute recursively to obtain the approximation:

$$F_I \approx \mathcal{LP}(o_{cls}^{(0)}) + \cdots + \mathcal{LP}(o_{cls}^{(N-1)}) + \mathcal{LP}(x_{cls}^{(N)}) \tag{4}$$

$$\triangleq \sum_{t=0}^{N} F_I^{(t)}, \tag{5}$$

where we define $F_I^{(t)} = \mathcal{LP}(o_{cls}^{(t)})$ for $t < N$, and $F_I^{(N)} = \mathcal{LP}(x_{cls}^{(N)})$. Thus from (5), the image feature F_I is approximately an aggregation of features from each layer $F_I^{(t)}$.

The feature vector from each layer is computed from a self-attention operation. For example, for the last layer ($t = 0$),

$$F_I^{(0)} = \mathcal{LP}(o_{cls}^{(0)}) = \mathcal{LP}\big(\sum_i \text{softmax}(\tfrac{q_{cls} k_i^\top}{\sqrt{C}}) v_i\big), \tag{6}$$

where the output of attention layer ($\mathcal{A}$) on the class token is

$$o_{cls}^{(0)} = \mathcal{A}(x^{(1)})[cls] = \sum_i \text{softmax}(\tfrac{q_{cls} k_i^\top}{\sqrt{C}}) v_i, \tag{7}$$

and $\mathcal{A}$ represents the attention layer in the Transformer, q_{cls} is the *query* embedding for the class token, while k_i and $v_i \in \mathbb{R}^C$ represent the *key* and *value* embeddings at spatial location i, with C as their channel dimension.[1] The $\mathrm{softmax}$ 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 channels of $\{q, k, v\}$ into several heads, and (7) is operated inside each head with the $\mathrm{softmax}$ calculated over subsets of the channels. Then, the final attention layer output is obtained by concatenating the results of each head together. In practice for visualization, we formulate the $o_{cls}^{(0)}$ with one attention head in the forward pass and operate the $\mathrm{softmax}$ over all channels as in (7). We discuss the influence of multi-heads on visual explanation in the §IV-C3.

Assuming that the feature vectors (F_T, F_I) are normalized and using (5), the matching score can be approximated as an aggregation of partial scores for feature vectors from each layer,

$$S_T(F_I) = \sum_c F_T[c]F_I[c] \approx \sum_c F_T[c] \sum_t F_I^{(t)}[c] \tag{8}$$

$$= \sum_t \Big(\sum_c F_T[c]F_I^{(t)}[c] \Big) \triangleq \sum_t S_T(F_I^{(t)}), \tag{9}$$

where $[c]$ selects the c-th channel, and $S_T(F_I^{(t)})$ denotes the score from layer t. Next, to calculate the heat map for the contribution of o_{cls} on the partial matching score, we write the partial matching score as a function of its o_{cls}. Specifically, looking at the last layer ($t = 0$) as an example,

$$S_T(F_I^{(0)}) = \sum_c F_T[c]F_I^{(0)}[c] \tag{10}$$

$$= \sum_c F_T[c]\mathcal{LP}(o_{cls})[c]^{(0)} \triangleq f(o_{cls}), \tag{11}$$

where we have defined the matching score as a function of o_{cls}, i.e., $f(o_{cls})$. We define the approximation of the matching score as a weighted combination of the channel features in o_{cls}, we have

$$f(o_{cls}) \approx \tilde{f}(o_{cls}) \triangleq \sum_c w_c o_{cls}[c] = w o_{cls}^{\mathsf{T}} \tag{12}$$

where w_c is the weight for the c-th channel, and $w = [w_c]_c \in \mathbb{R}^C$ the corresponding weight vector for all channels. To obtain the channel weights w, we aim to match the first derivatives (gradients) of the original matching score f and its approximation $\tilde{f}$, leading to the optimization problem:

$$w = \underset{w}{\mathrm{argmin}} \left\| f'(o_{cls}) - \tilde{f}'(o_{cls}) \right\|^2 \tag{13}$$

$$= \mathrm{argmin} \left\| \frac{\partial f}{\partial o_{cls}} - w \right\|^2, \tag{14}$$

which has the solution

$$w = \frac{\partial f}{\partial o_{cls}} = \frac{\partial S_T(F_I)}{\partial o_{cls}}. \tag{15}$$

[1]We skip the superscript (0) on $\{q, k, v\}$ identifying the layer for brevity.

Therefore, substituting (7) and (15) into (12),

$$S_T(F_I) \approx \sum_c w_c o_{cls}[c] \tag{16}$$

$$= \sum_c \frac{\partial S_T(F_I)}{\partial o_{cls}[c]} \sum_i \mathrm{softmax}(\tfrac{q_{cls}k_i^{\mathsf{T}}}{\sqrt{C}}) v_{ic} \tag{17}$$

$$= \sum_i \Big[\sum_c \frac{\partial S_T(F_I)}{\partial o_{cls}[c]} \mathrm{softmax}(\tfrac{q_{cls}k_i^{\mathsf{T}}}{\sqrt{C}}) v_{ic} \Big], \tag{18}$$

where v_{ic} is the c-th channel of v_i. Regarding $[v_{ic}]$ as the intermediate feature map, and $\lambda_i = \mathrm{softmax}(\frac{q_{cls}k_i^{\mathsf{T}}}{\sqrt{C}})$, then the importance of the i-th spatial location is defined as:

$$H_i = \mathrm{ReLU}\big(\sum_c w_c \lambda_i v_{ic} \big) \tag{19}$$

where ReLU means that we only focus on positions that have a positive effect on the final score.

The weight w_c represents the importance of each feature channel, and λ_i represents the importance of the values at each location via the softmax attention. However, from the visualization, we discover that the output of 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. Therefore, we replace the $\mathrm{softmax}$ with a "loosened" correlation by applying 0-1 normalization on the similarities $[q_{cls}k_i^{\mathsf{T}}]_i$, i.e., $\lambda_i \approx \Phi(q_{cls}k_i^{\mathsf{T}})$, where Φ is the 0-1 normalization function applied over the set of similarities. In the experiments and §IV-C1, we compare using the loosened λ_i and without λ_i to show the effect of spatial weights, qualitatively and quantitatively.

Therefore, with *the channel importance* w_c and *the spatial importance* λ_i, where

$$w_c = \frac{\partial S_T(F_I)}{\partial o_{cls}^{(0)}[c]}, \quad \lambda_i = \Phi(q_{cls}k_i^{\mathsf{T}}), \tag{20}$$

we obtain the proposed Grad-ECLIP explanation map $H = [H_i]_i$ for the last layer, where H_i is defined in (19) using the last layer values v as the feature map.

Visual and textual explanations: For the image encoder, the flattened heat map H_i is reshaped and interpolated to the original image's height and width, while for the text encoder, the heat (importance) value on the i_{th} tokens is remapped to the original word position in the sentence. Finally, based on the approximation in (9), the final explanation can be *aggregated over all the layers* by recursively processing each layer to obtain its heat map from (19). 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 numbers of layers involved in image and text explanation is shown in §IV-C2.

CNN-based CLIP: Our proposed Grad-ECLIP can also be applied to CNN-based CLIP. The CNN-based CLIP model is composed of a ResNet backbone followed by an attention pooling. Thus, we can use the final attention layer in the pooling to conduct our explanation, which uses the same implementation as for ViT-based CLIP. The visualizations shown in §IV-A5 verify the effectiveness of Grad-ECLIP on CNN-based CLIP model.

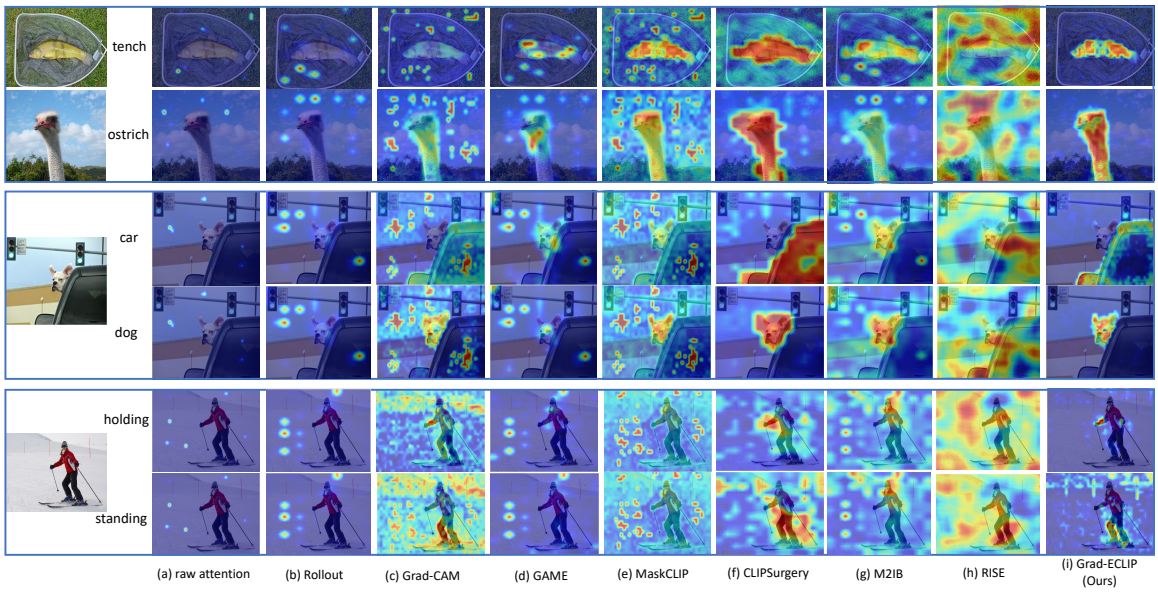

Fig. 3: Comparison of heat maps from: (a) the raw self-attention map in the last ViT layer; (b) Rollout [18]; (c) Grad-CAM [16]; (d) GAME [19]; (e) MaskCLIP [15]; (f) CLIPSurgery [14]; (g) M2IB [20]; (h) RISE [17]; (i) our proposed Grad-ECLIP. Visual explanations are provided for the matching score between the image and the specific text prompts, which can be nouns (*e.g.*, car, dog) or verbs (*e.g.*, holding, standing). From the comparison of visualizations, Grad-ECLIP exhibits superior explanation ability on different types of text prompts.

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

IV. Experiments with Grad-ECLIP

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) conduct ablation studies, including the effect of spatial weight, the involved layers, and attention heads.

Unless otherwise specified, we conducted the experiments with the ViT-B/16 architecture. We compared with representative baseline XAI methods from each category: 1) attention map-based *Rollout* [18], 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 explanations; 2) classical gradient-based method *Grad-CAM* [16], which takes the image-text similarity as target and calculates the gradients w.r.t. the ViT layer output; 3) relevance-based *GAME* [19], which integrates the relevancies and gradients propagated through the network; 4) cosine-based *MaskCLIP* [15] and *CLIPSurgery* [14], which generate similarity value on each location by the cosine between text feature and processed values as local image features; 5) M2IB [20], which applies the information bottleneck principle to generate an explanation map for CLIP. Each baseline is built with different properties and assumptions over the architecture. For evaluation of text encoder explanations, we compare Rollout, GAME, and M2IB, and our Grad-ECLIP by generating word importance on the input sentences – the other methods (Grad-CAM, MaskCLIP, and CLIPSurgery) are excluded since they are only applicable to the image encoder. We also show visualization comparisons with the typical perturbation method RISE [17] (and IG [55] in the Appendix), but did not conduct quantitative comparisons with perturbation

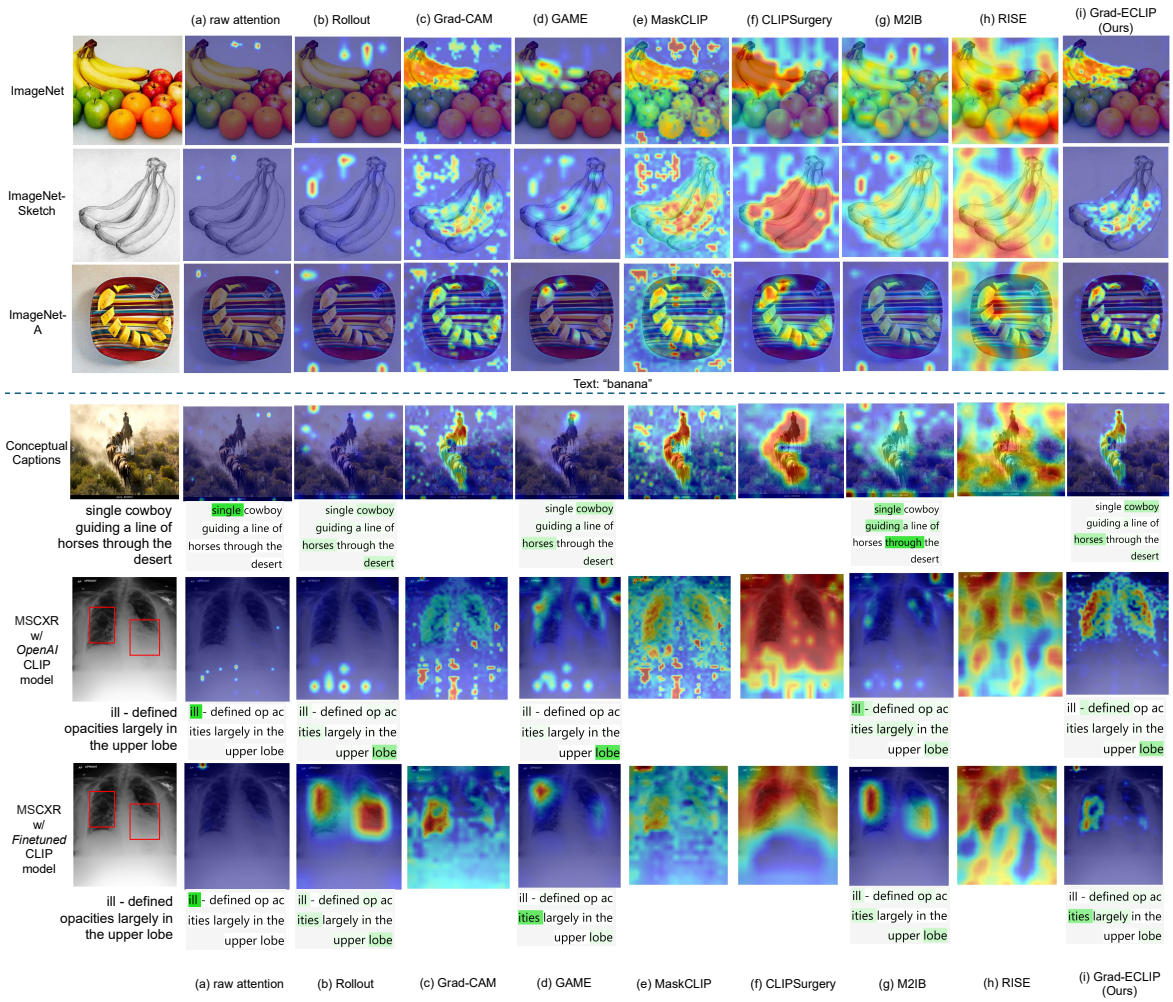

Fig. 5: Comparison of visual explanations for different methods on image samples from different image domains: natural images (ImageNet), renditions (ImageNet-R), pencil sketch (ImageNet-Sketch), natural adversarial examples (ImageNet-A), web images with captions (Conceptual Captions), and chest X-ray (MSCXR). On MSCXR, explanations are provided for both the OpenAI CLIP model and a fine-tuned version.

and Shapley methods, due to their computational complexity and black-box property. Since the proposed Grad-ECLIP is a white-box method designed for CLIP, we prioritize comparison with other XAI methods designed for CLIP (CLIPSurgery, MaskCLIP, M2IB) or similar transformer architectures (Rollout, GAME), and the classical CAM-based method Grad-CAM, both qualitatively and quantitatively.

A. *Qualitative evaluation*

In this section, we evaluate the proposed Grad-ECLIP qualitatively. The comparisons of visual explanations from different types of methods are presented in §IV-A1, and the visual explanations on both image and text encoders with image-sentence pair examples are presented in §IV-A2. Then, we conduct the visualizations of explanations on diverse image domains in §IV-A3 and adapt the Grad-ECLIP to other Vision Language Models (VLMs) in §IV-A4. Finally, we show that the Grad-ECLIP is applicable to CNN-based CLIP with ResNet backbone in §IV-A5.

1) Comparison of visual explanations: We compare the visualizations of raw self-attention, Rollout, Grad-CAM, GAME, MaskCLIP, CLIPSurgery, M2IB, RISE, and our Grad-ECLIP in Fig. 3 with images from ImageNet [56] and MS COCO [57]. Except for 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 distinct text-specific heat maps, and gives corresponding explanation of verbs for interpreting CLIP. For example, the highlights for "holding" focus around the person's hands (the 5th row of Fig. 3i), while "standing" highlights the person's legs (the 6th row of Fig. 3i). We also notice that the sticks in the background are highlighted, which is probably because the sticks are regarded as "standing" in the snow.

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 object region, but also contains background noise. M2IB and MaskCLIP fail when the texts are verbs ("holding" and "standing"), while RISE performs the worst when interpreting the CLIP model. The results of GAME and Rollout, which are both based on the self-attention of the model, generate confusing heat maps due to the sparse attention between tokens in some layers.

2) Explanations on image-sentence pairs: The explanation map from Grad-ECLIP can also be generated from the text encoder viewpoint. Using the gradient of the matching score and 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 visual and textual 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 correspndingly the text-specific important regions on the image are shown in the visual explanation. This word importance visualization of the input text can be helpful when designing text prompts for image-text dual-encoders in practical applications.

3) Visualization examples on diverse image domains: We show the visualization comparison of different methods on the samples from different image domains, including the original ImageNet, ImageNet in different domains: rendition (pencil sketch (ImageNet-Sketch [58]), natural adversarial example (ImageNet-A [59]), web images with captions (Conceptual Captions (CC) [60]), and chest x-ray with text (MSCXR [61]) in Fig. 5. For the image-caption pairs from the web-collected CC and chest X-ray data MSCXR, we generate explanations for both the image and text encoders, and compare with the other methods that also provide text encoder explanations, including the raw attention, Rollout, GAME, and M2IB.

Our Grad-ECLIP explanations provide interesting insights into how CLIP handles different image domains. In Fig. 5 (top), 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. With the sample of a web image and caption in the CC dataset, Grad-ECLIP generates visual highlights and the corresponding textual explanation map, showing that "cowboy", "horses" and "desert" are the main concepts used in the matching (from high to low importance).

Grad-ECLIP also provides interesting insights into how the original CLIP fails on novel domains. The last 2 rows of Fig. 5 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 fine-tuned 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 keyword and the main object in the image-text pair.

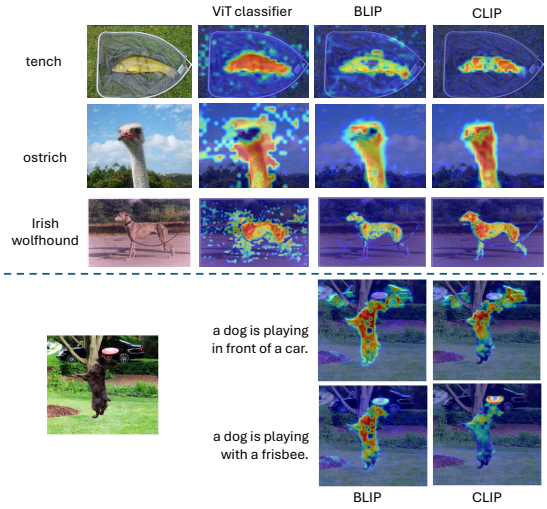

Fig. 6: Grad-ECLIP visual explanations of the ViT classifier and BLIP.

4) Explaining ViT classifier and BLIP with Grad-ECLIP: Since our explanation method is designed for CLIP encoders, which are Transformer-based, our method can be easily adapted to generate visual explanations for other Transformer-based models. Here we adapt Grad-ECLIP to a ViT-based classifier [62] and BLIP [8] to show the generality of our method.

For explaining the ViT-based classifier, the classification score on the corresponding category is used to calculate the gradient and generate the heat map. From the examples for the ViT classifier in Fig. 6, we can see that although there is some noise on the background, Grad-ECLIP can well mark out the important region on the image for the specific class. As for the VLMs, shown by the visual explanation results presented in Fig. 6, when matching the same image-text pair, different models put attention on different regions. 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 more importance on the dog in the image, while CLIP places more importance on the frisbee.

Other VLMs like ALBEF [63] add additional attention layers after the encoders to fuse the image and text features, and thus our current method is not directly applicable since our method assumes that the last layer attention output has a linear relationship with the final feature embedding. Our future work will investigate adapting our method to these modified ViT frameworks, e.g., the ALBEF model that uses cross attention to fuse image and text. 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.

5) Applying Grad-ECLIP to explain CNN-based CLIP: Although the methodology in §III-B for Grad-ECLIP is derived from CLIP with Transformer architecture, here we show that our method is also applicable to CNN-based CLIP by using the attention layer in the final attention pooling. Figure 7 shows the visual explanation results for ResNet50-CLIP using Grad-ECLIP, and compared to other explanation methods that are compatible with CNN-based CLIP, including Grad-CAM,

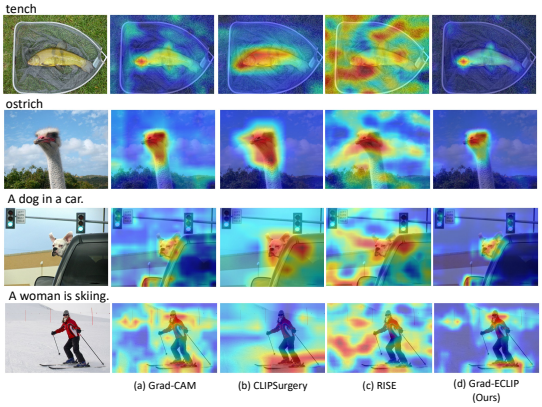

Fig. 7: Visual explanations of the CNN-based CLIP with ResNet50 architecture using Grad-ECLIP and other visual explanation methods.

CLIPSurgery, and RISE. From the examples, our Grad-ECLIP is able to generate text-specific heat maps with less noise for interpreting CNN-based CLIP. In contrast to our method, Grad-CAM can produce similar highlights on the explained objects, but its heat map is noisier in the background. Similar to its performance on ViT-based CLIP, CLIPSurgery generates rougher heat maps, which tend to put high values on a coarse region of the object. Meanwhile, RISE fails to explain the CLIP model.

B. Quantitative evaluation

In this section, we perform quantitative evaluations of Grad-ECLIP compared with baselines. In §IV-B1, the explanation faithfulness is evaluated by the Deletion and Insertion metrics [34, 36, 37, 43, 64], which are also called perturbation tests [19, 21]. Moreover, in §IV-B2, we evaluate the localization ability of CLIP using our Grad-ECLIP and the compared methods, when considering each visualization as a soft-segmentation of the image, using PointGame [65, 66] and segmentation tests [21]. Then, we conduct a user trust study to evaluate the trustability of the visual explanations and report the results in §IV-B3. Finally, in §IV-B4, we evaluate the processing time of Grad-ECLIP compared with other visual explanation methods. In this section, in order to understand the effect of the spatial importance term in (19), we also present Grad-ECLIP without using the spatial weight λ_i, i.e., setting $\lambda_i = 1$, which is denoted as "w/o λ_i".

1) Deletion and Insertion: The insertion/deletion process for an example image over the first 50 steps is displayed in Fig. 8 left. To summarize the insertion and deletion curves, we compute the area under the curve (AUC),

$$\text{Deletion} = \frac{1}{N_{step}} \sum_{n=1}^{N_{step}} Acc_{del}^{(n)}, \tag{21}$$

$$\text{Insertion} = \frac{1}{N_{step}} \sum_{n=1}^{N_{step}} Acc_{ins}^{(n)}, \tag{22}$$

where $Acc_{ins}^{(n)}$ and $Acc_{del}^{(n)}$ are the model performance (e.g., accuracy) with inserted and deleted images at step n. For the deletion curve, a more faithful model will have larger drops in model performance (e.g., accuracy) as pixels are removed, and thus lower AUC for deletion curves is better. On the opposite, for the insertion curve, a more faithful model will have a larger increase in model performance as pixels are added, and thus a higher AUC for insertion curves is better. We also define an overall metric IMD (Insertion Minus Deletion), which summarizes the two AUCs:

$$\text{IMD} = \frac{1}{N_{step}} \sum_{n=1}^{N_{step}} Acc_{ins}^{(n)} - Acc_{del}^{(n)}. \tag{23}$$

Higher IMD indicates better faithfulness, corresponding to higher insertion AUC and lower deletion AUC. We consider each step as 0.5% of the 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 [56] (ILSVRC) 2012, consisting of 50K images from 1000 classes. In particular, the perturbed image and each of the class names is fed into CLIP, and then classes with the highest scores are selected.

The insertion/deletion curves for top-1 accuracy are presented in Fig. 8 right, and AUC values for top-1 and top-5 accuracy are presented in Tab. I. Our method obtains the fastest performance drop for Deletion and the largest performance increase for Insertion compared with most related works ($p<.001$), showing that regions highlighted in our heat maps better represent explanations of CLIP. CLIPSurgery has comparable results to ours for Insertion ($p=0.374$, 0.413, 0.074, 0.425), while performing poorly when evaluated with Deletion. The reason is that CLIPSurgery exhibits heat maps with nearly the same high values on the explained target region, so that 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 AUC. Since CLIPSurgery can locate the explanation target with high values on the heat map, it performs well in the Insertion test. Overall, combining the deletion and insertion, our method has achieved the highest results in terms of IMD ($p<.001$), which shows that changing image inputs based on Grad-ECLIP maps produces the greatest impact on the model prediction. Our method without using the spatial importance (w/o λ_i) has slightly worse performance, but is still better than other baselines. As with [19, 21], 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 next evaluate the Deletion and Insertion performance for image and text retrieval tasks on Karpathy's validation split of MS COCO. To evaluate the image explanations, image pixels are removed or added step-by-step based on the text-specific explanation heat map. With the modified image replacing the original image, we record the image and text retrieval performance (recall @ top-1 and top-5 matching) to draw the deletion and insertion curve. Then, the AUC results of Deletion and Insertion with text-specific image explanations are reported in Tab. II. Grad-ECLIP surpasses the other methods on most metrics ($p<.001$), which further demonstrates that our method produces high-quality visual explanation, regardless of whether the text is the class categories as in ImageNet or long captions as in MS COCO.

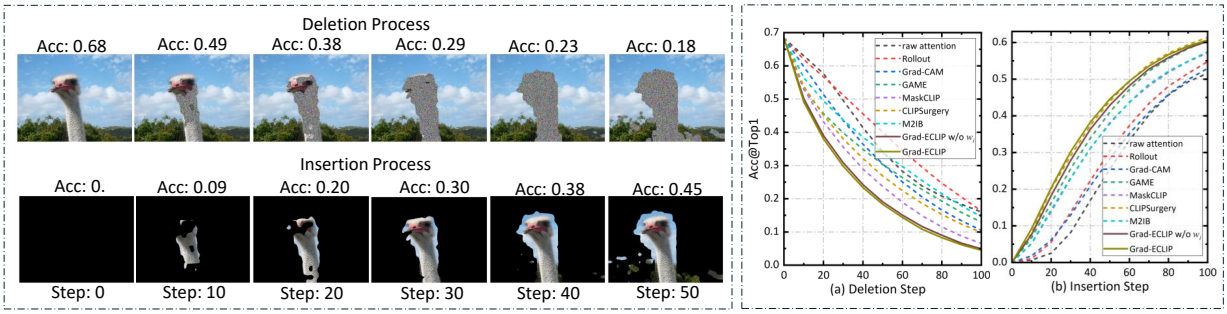

Fig. 8: (left) An example showing the deletion and insertion process according to the Grad-ECLIP heat map on an image for 0 to 50 steps. The classification accuracy on each step is shown on the top, which forms the points of the deletion or insertion curves. (right) Classification accuracy at Top-1 vs. (a) Deletion steps and (b) Insertion steps, on the ImageNet validation dataset with visual explanation heat maps from our Grad-ECLIP (solid) and other methods (dashed).

TABLE I: Faithfulness evaluation of **image** explanation on the *ImageNet* validation set: Deletion, Insertion and IMD (Insertion minus Deletion), based on Top-1 (@1) or Top-5 (@5) classification accuracy. Either the ground-truth or the prediction is used as the text input into CLIP. The arrows indicate lower/higher value is better. The last row shows the p-value for a paired sample t-test between the top-2 methods (bolded and underlined), excluding Ours w/o λ_i. Significant values (p<.05) are highlighted in blue.

	Deletion↓				Insertion↑				IMD ↑			
	Ground-truth		Prediction		Ground-truth		Prediction		Ground-truth		Prediction	
Method	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5
raw attention	0.383	0.624	-	-	0.249	0.419	-	-	-0.134	-0.204	-	-
Rollout	0.408	0.656	-	-	0.280	0.467	-	-	-0.128	-0.189	-	-
Grad-CAM	0.342	0.563	0.352	0.582	0.268	0.445	0.253	0.421	-0.074	-0.117	-0.099	-0.161
GAME	0.336	0.573	0.350	0.594	0.361	0.564	0.343	0.538	0.025	-0.010	-0.007	-0.055
MaskCLIP	0.285	0.489	0.289	0.496	0.334	0.535	0.328	0.527	0.049	0.047	0.039	0.031
CLIPSurgery	0.312	0.524	0.322	0.541	0.383	**0.602**	**0.373**	0.572	0.070	0.079	0.051	0.030
M2IB	0.363	0.595	0.363	0.595	0.335	0.541	0.335	0.541	-0.028	-0.054	-0.029	-0.054
Ours	**0.246**	**0.427**	**0.254**	**0.442**	**0.384**	0.599	0.367	**0.575**	**0.137**	**0.172**	**0.113**	**0.133**
- w/o λ_i	0.254	0.438	0.263	0.457	0.372	0.583	0.353	0.556	0.118	0.145	0.089	0.099
p-value	<.001	<.001	<.001	<.001	0.374	0.413	0.074	0.425	<.001	<.001	<.001	<.001

TABLE II: Evaluation of **image** explanation faithfulness on *MS COCO image-text retrieval (Karpathy's split)* validation dataset: Deletion, Insertion and IMD with performance on image retrieval (IR) and text retrieval (TR) tasks. The last row shows the p-value for a paired sample t-test between the top-2 methods (bolded and underlined), excluding Ours w/o λ_i. Significant values (p<.05) are highlighted in blue.

	Deletion↓				Insertion↑				IMD↑			
	IR		TR		IR		TR		IR		TR	
Method	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5
raw attention	0.171	0.355	0.192	0.372	0.125	0.255	0.154	0.297	-0.046	-0.038	-0.100	-0.075
Rollout	0.195	0.395	0.227	0.424	0.129	0.293	0.175	0.350	-0.065	-0.052	-0.128	-0.105
Grad-CAM	0.172	0.350	0.216	0.401	0.103	0.222	0.115	0.233	-0.069	-0.101	-0.129	-0.168
GAME	0.171	0.355	0.198	0.380	0.154	0.308	**0.210**	0.374	-0.017	0.012	-0.047	-0.006
MaskCLIP	0.132	0.284	**0.152**	0.295	0.142	0.295	0.189	0.351	0.010	0.038	0.011	0.057
CLIPSurgery	0.179	0.365	0.238	0.429	0.142	0.294	0.177	0.338	-0.037	-0.061	-0.071	-0.091
M2IB	0.179	0.367	0.206	0.391	0.147	0.300	0.206	0.369	-0.033	0.000	-0.067	-0.021
Ours	**0.125**	**0.267**	0.155	**0.293**	**0.158**	**0.320**	0.206	**0.376**	**0.033**	**0.051**	**0.053**	**0.083**
- w/o λ_i	0.139	0.294	0.183	0.339	0.140	0.289	0.174	0.328	0.001	-0.009	-0.005	-0.011
p-value	<.001	<.001	0.106	0.705	0.011	0.001	0.297	0.645	<.001	<.001	0.001	0.001

Finally, we evaluate *the faithfulness of our text explanations* using the *text version* of the 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 for the modified caption, changing with a total of 5 steps, with one word deleted/inserted at a time. The results in Tab. III show that Grad-ECLIP has the highest faithfulness (best deletion, insertion and IMD scores with p=0.019) compared with the other Transformer explanation methods.

The last row of Tables I, II, and III displays the paired sample t-test results conducted between the AUCs of the top-2 methods in the column, where pairs correspond to the same step on the two curves.

2) Point Game and Segmentation Test: Following the pointing game and segmentation test metric adopted in related works [17, 21, 25, 36, 65, 67], we evaluate the localization ability of CLIP using our Grad-ECLIP and the compared methods. It should be noted that these localization tests are mainly a rationalization of the CLIP prediction, rather than a faithfulness metric. Indeed, localization or segmentation

TABLE III: Evaluation of **text** explanation faithfulness on *MS COCO image-text retrieval (Karpathy's split)* validation dataset: Deletion, Insertion and IMD with performance on image retrieval (IR) and text retrieval (TR) tasks. The last row shows the p-value for a paired sample t-test between the top-2 methods (bolded and underlined), excluding Ours w/o λ_i. Significant values (p<.05) are highlighted in blue.

	Deletion↓		Insertion↑		IMD↑	
Method	IR	TR	IR	TR	IR	TR
raw attention	0.284	0.492	0.007	0.033	-0.278	-0.459
Rollout	0.122	0.239	0.105	0.207	-0.017	-0.032
GAME	<u>0.108</u>	<u>0.208</u>	<u>0.115</u>	<u>0.230</u>	<u>0.006</u>	<u>0.022</u>
M2IB	0.214	0.426	0.005	0.024	-0.233	-0.419
Ours	**0.100**	**0.177**	**0.129**	**0.254**	**0.076**	**0.030**
- w/o λ_i	0.112	0.211	0.112	0.236	0.001	0.025
p-value	0.019	0.017	0.025	0.026	0.019	0.019

performance is not necessarily indicative of the faithfulness, in terms of the insertion and deletion metrics, as indicated in Table I.

We adopt the ImageNet-Segmentation (ImageNet-S) [68] validation set, which has segment annotations on 12,419 images of 919 categories from ImageNet. Point Game (PG) is a commonly used metric to evaluate the localization correctness of a visual explanation. PG counts a hit score if the location with the largest value in the visual explanation heat map lies within the object region, which can be defined by the class segmentation mask. Then the PG accuracy is measured by averaging over all samples. Since PG only considers the maximum point, but not the spread of the heat map, we also conduct energy-PG [36], which calculates the proportion of heat map energy within the ground-truth mask versus the whole map. Similar to the evaluation by [19, 21], regarding the heat maps as soft-segmentation results, we adopt pixel accuracy (Pixel Acc.), average precision (AP), and averaged mask intersection over union (maskIoU) as additional metrics.

The results for localization are shown in Tab. IV. Both versions of Grad-ECLIP significantly outperform other explanation methods on PG and energy-PG (χ^2 test on PG hit numbers between Grad-ECLIP and the second best method CLIPSurgery, χ^2=593.3, df=1, p<.001), which demonstrates that Grad-ECLIP can well reveal that the important pixels for CLIP are inside the object region. Comparing Grad-ECLIP with and without λ_i, Grad-ECLIP without λ_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 an advantage on these two metrics. In Fig. 9(b,c) of the main paper, using λ_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 the object region, and gets a higher score when aggregating the heatmaps inside the object mask in these two evaluations. However, the lower PG, energy-PG, and AP demonstrate that there are more high values generated outside of the object boundary.

3) User Study on Improving Understanding: In this section, we evaluate whether the Grad-ECLIP explanations can improve a user's understanding of CLIP's predictions. We adopt a *forward simulation* task [69], where the user is given an image and two captions and is asked to judge which caption will receive a higher matching score from the model under two conditions: with or without corresponding visual explanations provided. If the users' judgment accuracy improves when provided with the explanations, then it suggests that the explanations allow the users' to better understand the model's predictions.

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

Method	PG	energy-PG	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	0.2640	0.3557	0.6194	0.4003	0.1474
Ours	**0.8899**	**0.5997**	0.7056	**0.5662**	0.2869
- w/o λ_i	0.8356	0.4409	0.7365	0.5163	0.3314

Ten images and their corresponding captions were randomly selected from the dataset. A questionnaire was created, where each image was displayed with two text caption options, and the participant was asked to select the caption that they think the CLIP model will give a higher matching score. The images were grouped into two sets of five, denoted as Set I and Set II. In Set I, the image and captions were displayed without explanations, and in Set II they were displayed with the corresponding Grad-ECLIP explanations for each image-caption pair. Before each Set, a few examples of image-caption pairs, their CLIP scores, and explanations (Set II only) were presented to the participant in a training phase. Set I was presented before Set II so that the participant's initial understanding of the model was not biased by the explanations. The assignment of images to Set I and Set II was counterbalanced so that a given image appeared the same number of times in Set I and Set II over all participants. We collected responses from a total of 34 participants. More details about the questionnaires, including sample selection and examples, are presented in the Appendix.

The results of the user study indicate that the participants' correct judgment of CLIP's higher-matched sentence significantly improved when the Grad-ECLIP explanations were provided, from 67.6% to 80.0% accuracy (paired t-test, t(33)=1.904, p=0.032). Thus, the user study demonstrates that the Grad-ECLIP explanations can effectively assist users in understanding the prediction-making mechanism of the CLIP model, highlighting its superior image-text matching interpretability.

4) Processing time comparison: In Tab. V, 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 the raw attention and MaskCLIP, which obtain their heat maps from the forward pass of the model and some other

TABLE V: Comparison of the average processing time (on RTX3090 GPU) 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

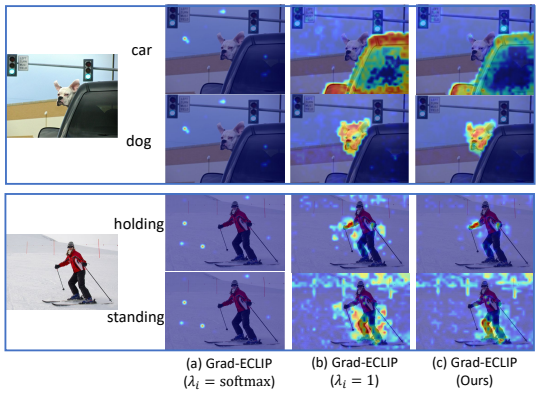

Fig. 9: The effect of spatial weight λ_i. We compare the explanation maps of Grad-ECLIP with a version without the proposed λ_i (denoted as "w/o λ_i", which replaces the proposed spatial weights with (a) $\lambda_i = \mathrm{softmax}$, (b) $\lambda_i = 1$), and (c) the full Grad-ECLIP.

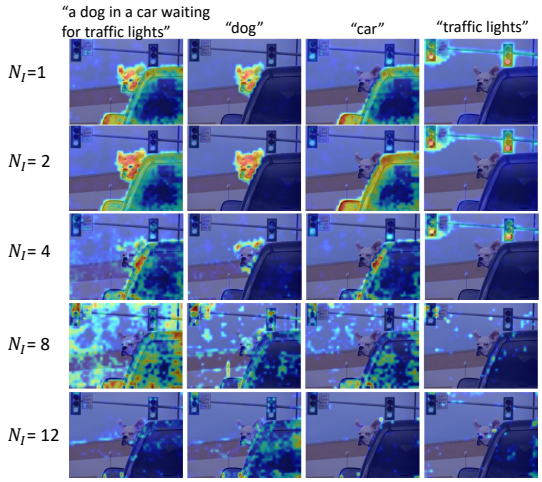

Fig. 10: The **image** visual explanations generated when aggregating over N_I layers of the image transformer encoder.

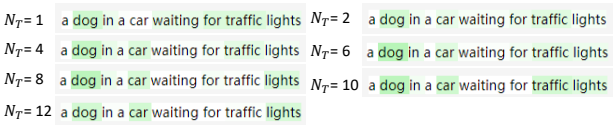

Fig. 11: The **textual** explanations generated when aggregating over N_T layers of the text transformer encoder.

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 a common drawback of perturbation-based methods.

C. Ablation study

In this section, we conduct ablation studies to illustrate the influence of the proposed loosened spatial weight (§IV-C1), the number of involved layers (§IV-C2), and multi attention heads (§IV-C3) in the calculation of Grad-ECLIP.

1) Effect of the loosened spatial weight: Here we conduct an ablation study on the λ_i in Eq. 20 to show the effect of the proposed spatial weight. We consider two versions of Grad-ECLIP with modified spatial importance λ_i: the first version uses the softmax attention rather than 0-1 normalization (denoted as $\lambda_i = \mathrm{softmax}$); the second version removes the spatial importance completely by setting $\lambda_i = 1$.

A comparison of visualizations is presented in Fig. 9. Compared with the heat maps generated by the full Grad-ECLIP in Fig. 9c, the version that removes spatial importance altogether ($\lambda_i = 1$) contains more noise near object boundaries and on the background (Fig. 9b), but is otherwise consistent with full Grad-ECLIP. The result of using $\lambda_i = \mathrm{softmax}$ (Fig. 9a) is equivalent to raw attention (Fig. 3a) due to the output of the $\mathrm{softmax}$ being extremely sparse. We also provide the quantitative comparisons of Grad-ECLIP using $\lambda_i = 1$ in the §IV-B1 and Appendix C, denoted as "w/o λ_i".

2) Effect of number of layers on visual explanation: As introduced in §III-B, the explanation can be aggregated over all the layers in Transformer by recursively processing each layer with Eq. 19. In this section, we conduct the experiments to discuss the influence of using different numbers of layers to generate heat maps for images and text.

Let N_I and N_T be the number of layers used for explaining the image and text encoders, respectively, where $N_I = 1$ means the visualization was generated with only the final Transformer layer (for image encoder), and $N_I = 12$ means that the visualizations are aggregated over all Transformer layers. The visualizations for different N_I for image explanations are shown in Fig. 10, and the corresponding caption explanations are shown in Fig. 11 for different N_T. Table VI evaluates the image explanation faithfulness vs. the number of transformer layers (N_I) aggregated in the image encoder. The setting of $N_I = 1$ gives the highest IMD. Figure 10 shows the corresponding example for different N_I – the lower layers of the image encoder include less semantic information and may introduce more noise to the heat maps, which is why using only the final layer ($N_I = 1$) in the image encoder has higher faithfulness. Therefore, it is best to just use the last layer in the calculation of image visual explanation, and this conclusion is consistent with the classical gradient-based CAM methods for CNNs.

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

3) Effect of multi attention heads on visual explanation: As mentioned in (7) of §III-B, for producing the Grad-

TABLE VI: The **image** explanation faithfulness vs. the number of transformer layers aggregated for the explanation. Evaluating on *MS COCO image-text retrieval (Karpathy's split)* validation dataset: Deletion, Insertion, and IMD with reporting image retrieval (IR) and text retrieval (TR) performance.

	Deletion↓		Insertion↑		IMD↑	
N_I	IR	TR	IR	TR	IR	TR
1	**0.1246**	**0.1551**	**0.1576**	**0.2056**	**0.0330**	**0.0505**
2	0.1279	0.1657	0.1547	0.2010	0.0269	0.0253
4	0.1622	0.2051	0.1182	0.1399	-0.0439	-0.0652
8	0.2316	0.2879	0.0484	0.0440	-0.1832	-0.2439
10	0.2121	0.2643	0.0715	0.0774	-0.1407	-0.1868
12	0.2163	0.2642	0.0699	0.0773	-0.1465	-0.1869

TABLE VII: The **text** explanation faithfulness vs. the number of transformer layers aggregated for the explanation. Evaluating on *MS COCO image-text retrieval (Karpathy's split)* validation dataset: Deletion, Insertion and IMD with reporting image retrieval (IR) and text retrieval (TR) performance.

	Deletion↓		Insertion↑		IMD↑	
N_T	IR	TR	IR	TR	IR	TR
1	0.1118	0.2087	0.1059	0.2196	-0.0059	0.0108
2	0.1021	0.1826	0.1186	0.2351	0.0166	0.0525
4	**0.0995**	0.1786	0.1242	0.2428	0.0247	0.0643
6	0.0989	**0.1761**	0.1273	0.2490	0.0284	0.0729
8	0.0996	0.1770	**0.1292**	**0.2536**	**0.0296**	**0.0766**
10	0.1008	0.1843	0.1288	0.2472	0.0279	0.0629
12	0.1095	0.2087	0.1219	0.2364	0.0125	0.0277

ECLIP visual explanation, we set CLIP to perform the forward pass with a single head in the attention layer instead of the original multi-head attention layer. In Fig. 12, we show the visualization of explanation maps when using multi-head attention layers, compared to using a single head. Comparing Fig. 12 (a) and (b), using multi-head attention results in some surrounding context information being highlighted with the explained object.

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. 12 (c) for the target "dog", and (d) for the "car". The visual explanation in each head highlights different regions, not only seeing the target object. We can infer that the channels assigned to each head 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. 12 (b).

V. Analysis of CLIP using 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 visual explanation maps and textual explanations generated by Grad-ECLIP to give examples of how to explore the mechanism in text and image matching, and analyze the strengths, weaknesses, and preferences of

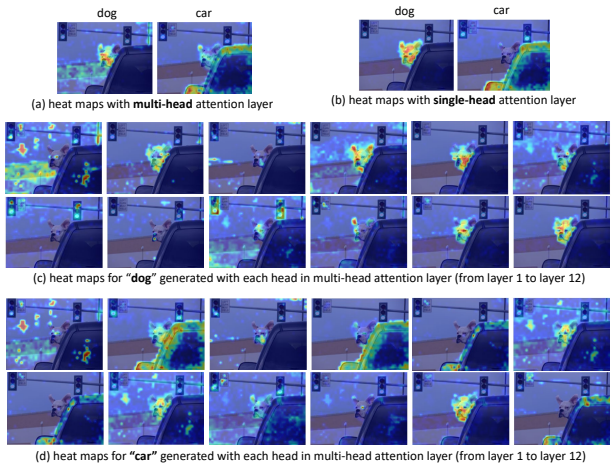

Fig. 12: The visual explanation maps using (a) multi-head attention layer; (b) single-head attention layer; (c) each head in the multi-head attention layer for text "dog"; (d) each head in the multi-head attention layer for text "car".

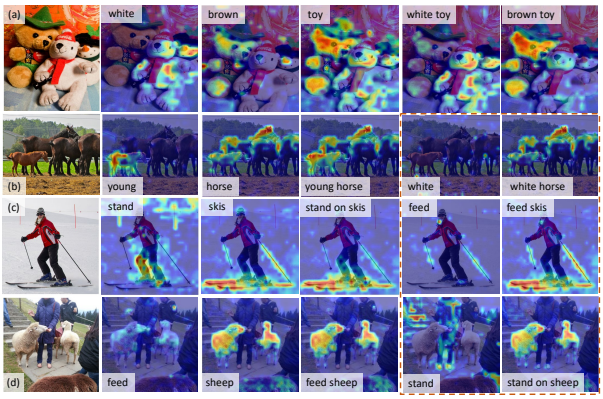

Fig. 13: Visual explanation heat maps generated for single words and word phrases using Grad-ECLIP on CLIP. The dashed box contains examples where the text does not match the image.

the CLIP model. We hope that our explanation tool can help researchers discover more interesting properties of pre-trained image-language models and inspire further development of these models. Here we analyze three aspects of CLIP: 1) the decomposition and addibility in image-text matching (§V-A); 2) types of attributes that can be identified by CLIP (§V-B); 3) the concreteness/abstractness of words learned by CLIP (§V-C). We note that there is a risk that the heat map explanations may not be accurate when explaining the model, and we can temper this risk by selecting the XAI that has higher faithfulness to the model. We have shown in Section IV that our method Grad-ECLIP has the highest faithfulness among the related methods, and thus, this provides a justification for using it to perform our interpretation/analysis of CLIP here. Future XAI methods may improve on faithfulness and may update and refine these interpretations. Nonetheless, this is the nature of scientific progress.

A. Concept decomposition and addibility

Examining the visualizations shown in Fig. 3h, CLIP can well recognize the single concepts (nouns) and has good attention to actions (verbs). An interesting question is how it processes 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. 13. 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 are strengthened by adding the attribute "young". The examples of verb-noun cases in (c) and (d) also show a similar addibility pattern on the heat maps: (c) with the verb "stand", the region of the 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. 13, 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.

To verify the above observation quantitatively, we conduct an experiment on 60 randomly selected images from the MS COCO validation set. Concretely, for each image, we provide three words based on the content of the image, including one *base word*, one *mismatched word*, and one *matched word*. For example, in Figure. 13(b), the base word is "horse", and the matched and mismatched words are "young" and "white", respectively. Combining the base word with the matched word or mismatched word produces the *matched phrase*, e.g. "young horse", and the *mismatched phrase* "white horse", respectively. We randomly select 30 images for the combination of adjective and noun, and another 30 images for verb and noun combinations.

To measure concept addibility, we calculate the cosine similarity between the explanation heat map for the whole phrase, and the combination of corresponding word-based heat maps, e.g., the heat map for the matched phrase "young horse" and the combination of maps for "young" and for "horse". The averaged similarities over all images are shown in Tab. VIII , where we list the results using different combination methods. Combining word-based maps using addition leads to the highest similarity with the phrase-based heat map (p<.001, paired t-test between addition and other methods), regardless of whether it is a matched or mismatched phrase. Meanwhile, combinations using multiplication have significantly lower similarity compared to addition, which supports the above addibility hypothesis observed from Fig. 13.

We conducted a further analysis by plotting the heat map similarity of the matched/mismatched word or phrase compared to the base word, versus the CLIP image-text matching score, relative to the base word, in Fig. 14(left). The mismatched phrase (red "+") usually obtains a lower CLIP matching score than the base word (the yellow star) but a higher score than the mismatched word (purple circle), while the explanation heat maps of mismatched phrases have high similarity with the base word (near 1). Therefore, combining

TABLE VIII: Average cosine similarity between the heat map of a phrase and the combined heat maps of individual words using different combination methods. The last column shows the p-value for a paired sample t-test between "add maps" and other methods.

	Matched Phrase		Mismatched Phrase		p-value
Combination method	adj.+noun	verb+noun	adj.+noun	verb+noun	vs. add maps
add maps	0.9750	0.9814	0.9480	0.9755	-
multiply maps	0.7871	0.7231	0.5810	0.5400	<.001
maximum maps	0.9619	0.9741	0.9319	0.9711	<.001

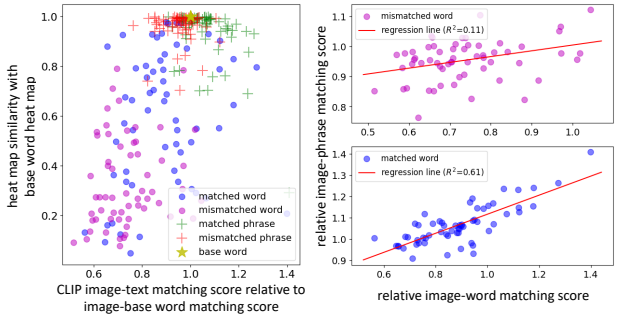

Fig. 14: (Left) Scatter plot of explanation heat map similarity vs. CLIP image-text matching score for matched words, mismatched words, matched phrases, and mismatched phrases. The heat map similarity is with the base word, and the CLIP score is relative to the CLIP score using the base word. Thus, all base word points are at coordinates (1,1). (Right) Scatter plot and linear regression of the relative CLIP image-phrase matching score vs. the relative image-word matching score for mismatched words (top-right) and matched words (bottom-right).

with the verified addibility pattern, we infer that when CLIP matches the image with a phrase, it separates the phrase into words and makes the matching result based on putting attention to each word. Thus, for the case of a mismatched phrase (e.g., the white horse in Fig. 13(b)) where the object does not exist in the image, the model basically looks at the image region corresponding to the base word "horse" and makes the prediction. The scatter plot in Fig. 14 (top-right) also verifies that the CLIP matching score for mismatched words is usually lower than the base words (lower than 1), while the scores for the corresponding mismatched phrases are close to 1, i.e., similar to just base-word matching score. While there is a linear trend between the relative image-mismatched-word score and the image-phrase score ($R^2 = 0.11$, $p = 0.007$ in linear regression), the R^2 value is extremely low, indicating significant unexplained variance. On the other hand, in Fig. 14 (bottom-right), most matched phrases will have increased score compared to the baseline word, and there is a strong linear trend ($R^2 = 0.61$, $p < .001$), where matched words with higher scores will also have higher scores in the matched phrase. A matched word can provide further positive effect to using just the base word in the matching process, so that most matched phrases (green "+") obtain higher CLIP scores than the base word (the yellow star).

Therefore, we infer that when processing the matching of the 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.

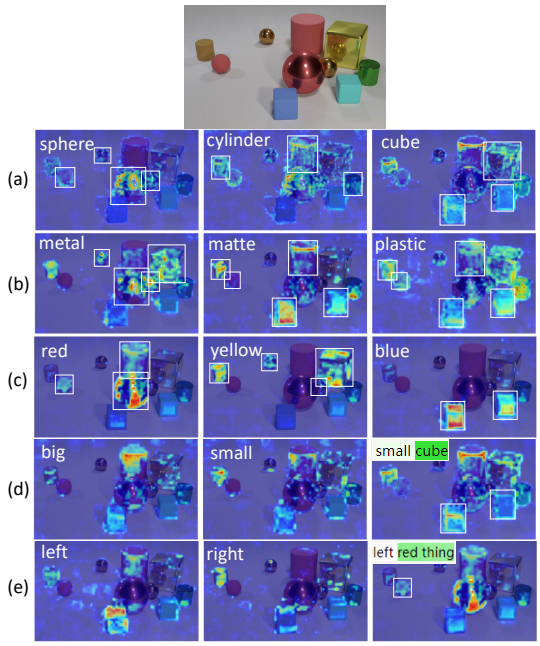

Fig. 15: Visual explanations on image matching with different kinds of attributions: (a) shape; (b) material; (c) color; (d) size; (e) position. For visualization, the ground-truth corresponding to the text prompt is outlined with white boxes, except for cases involving relative adjectives, *e.g.*"big", "small", "left", "right". The text explanation maps are also shown for "small cube" and "left red thing" combinations.

B. Diagnostics on attribution identification

In Fig. 13(a), we see that CLIP has an ability to distinguish color attributes and mark out the corresponding regions on the 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 [70], 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. 15 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 possesses 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 a good ability to distinguish different colors. 3) 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. For example, there are a few differences between the heat maps of "small cube" and "cube" in (a), or "left red thing" and "red" in (c), which demonstrates that the words "cube" and "red" take the major role in the matching. This is also confirmed by the text heat maps in the figure.

We next conduct a quantitative experiment to explore CLIP's ability to encode different object attributes. We utilize the annotations of the CLEVR dataset, which provides the center position of each object with its properties, including color, shape, material, size, and location relationship. With the object coordinates, we use SAM [71] to generate segmentation masks for the objects, and then perform manual checking to obtain the ground-truth masks. We randomly select 50 images from the dataset. An image example with the ground truth masks and explanation heat maps for specific attribute text is shown in Fig. 16. Using different attribute words, we measure similarity or distance between the CLIP explanation maps and the corresponding ground-truth, with higher similarity (lower distance) indicating that CLIP encodes relevant features for that attribute by focusing on the corresponding object.

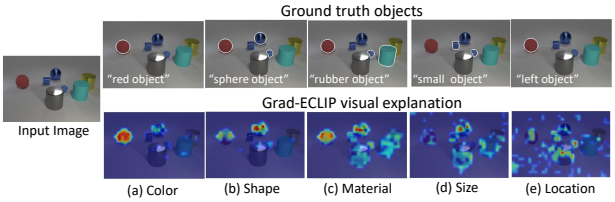

Fig. 16: The ground truth objects (outlined in white) and the Grad-ECLIP explanation heat maps corresponding to image-text match for specific attribute words: (a) Color "red object"; (b) Shape "sphere object"; (c) Material "rubber object"; (d) Size "small object"; (e) Location relationship "left object".

TABLE IX: Averaged similarity/distance between the ground-truth mask and Grad-ECLIP heat maps generated when CLIP matches attribute words in different categories.

	Color	Shape	Material	Size	Location
Cosine similarity ↑	**0.6654**	0.5959	0.5772	0.4645	0.2758
Correlation ↑	**0.0419**	0.0218	0.0150	0.0100	0.0107
OTD ↓	0.0085	**0.0084**	0.0090	0.0093	0.0356
KLD ↓	**2.0237**	2.0968	3.1444	4.3012	5.0392
JSD ↓	**0.2881**	0.3222	0.3072	0.3856	0.5327

The results averaged over attribute categories are presented in Tab. IX. The higher average similarities (or lower distances) for color words demonstrate that the attention of CLIP can better locate the correct object when matching with the given object colors, which means CLIP has a better ability to distinguish or encode color attributes. Following color, the shape and material attributes are the 2nd and 3rd best, while size and location are the worst attributes. To further verify the bad matching with location relationship words, we further conduct an ablation study by comparing using only shape, e.g., "sphere", the combination of location and shape, e.g., "left sphere", and only location, "e.g., left", to describe the left object. The results show that the location relationship words bring no influence to the CLIP matching process, with the heatmap similarity unchanged when adding the location attribute word. See Appendix B for details.

Overall, from the above analysis, we 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 §V-A, it is reasonable to expect that attributes that have a concrete visual appearance, e.g., color, will contribute more to the matching score, compared with the abstract comparative attributes.

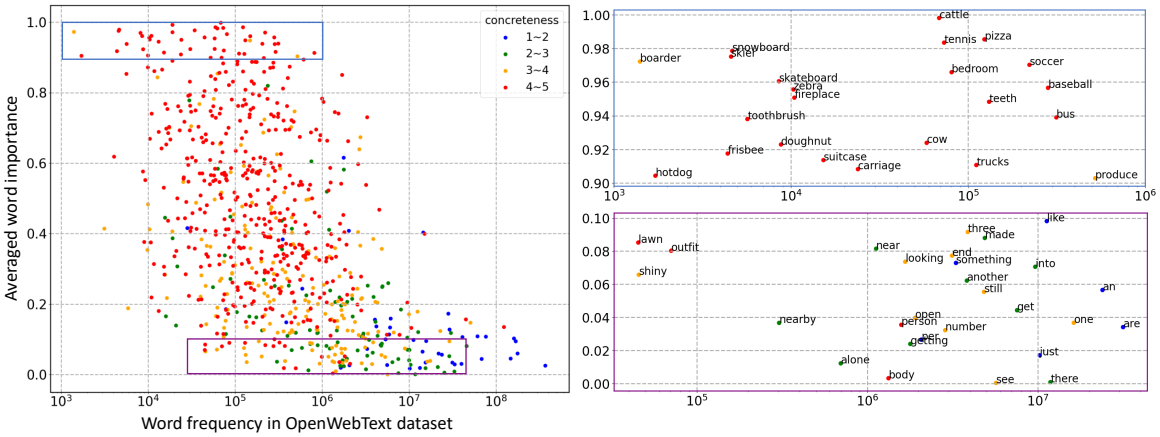

Fig. 17: The average word importance (via Grad-ECLIP) vs. the word frequency in the OpenWordText dataset for the top-1000 most frequent words in the MS COCO Karpathy's validation split. The colors represent the concreteness level of each word according to [72]. (right) zoom-in of the blue and purple boxes to show the word examples with high word importance (blue box) and low word importance (purple box).

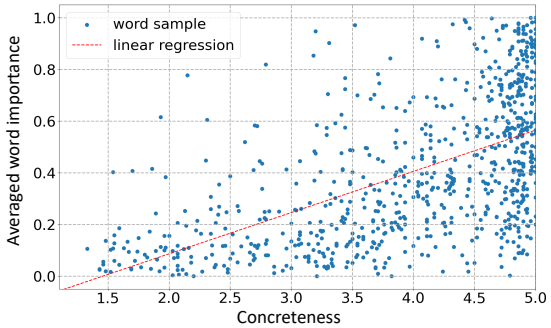

Fig. 18: The average word importance (via Grad-ECLIP) vs. word concreteness for the top-1000 most frequent words in the MS COCO Karpathy's validation split. The red dashed line shows the linear regression result over the word samples; $r^2 = 0.32, p < 0.001$.

C. Relationship between word importance and concreteness

From §V-B, we have inferred that the concrete visual concepts (e.g., "red" and "blue") contribute more to the image and text matching than the abstract attributes (e.g., "left" and "right"). Therefore, based on the text explanation obtained from Grad-ECLIP, we further explore the relationship between word importance in matching and word concreteness, and analyze which type of concepts and words (concrete vs. abstract) the CLIP model has actually learned and uses most often for matching. For an image-text pair, we calculate the textual explanation from Grad-ECLIP, and then obtain the importance value of each word by normalizing such that the maximum word importance value is 1 in the sentence. We then calculate the average word importance on the top-1000 most frequent words in the MS COCO caption (Karpathy's split) validation set. For the word concreteness, we adopt the open-sourced database from [72], which provides the concreteness for 40,000 common English words measured by human rating. The concreteness is a value from 1 to 5, where 5 means the most concrete and 1 means the most abstract.

Based on the 1000 selected words, Fig. 18 presents a scatter plot of average word importance versus concreteness value. We perform linear regression analysis on this data (red dashed line), and the regression result was statistically significant ($r^2 = 0.32$, $p < 0.001$). The scatter distribution and linear regression result reveal that CLIP places higher word importance on more concrete words, and vice versa, less word importance on more abstract words. *Therefore, the words that CLIP has learned for matching are biased towards concrete words.*

CLIP's learned bias towards concrete words could be due to frequency bias in the training set, i.e., concrete words could appear more frequently during training. To investigate this possibility, we attempt to count the number of occurrences of the selected words in CLIP's training set WebImageText [1]. However, since WebImageText is not publicly available, we instead compute these statistics from the OpenWebText [73] dataset, which follows the same methodology to reproduce the data characteristics and structure of the WebImageText corpus.

The relationship between average word importance and the word frequency in the training corpus is plotted in Fig. 17, with different sample colors representing the levels of concreteness. There is no obvious relationship between the word attention and the frequency. Many high-frequency words obtain a low word importance when the concreteness value is very low, and on the opposite, low-frequency words of concrete concepts may obtain a high word importance. Fig. 17 (right) shows the zoom-in on the samples in the blue box, which have high average word importance, and in the purple box with low average importance. Comparing these two boxes, the words with concrete visual meaning, especially nouns, are indeed more likely to obtain higher importance than the abstract words. Therefore, the analysis using Grad-ECLIP text explanation further supports the conclusion from §V-B – *concrete visual concepts are learned better and contribute more to the CLIP image-text matching score than abstract concepts, and this phenomenon is not due to word frequency bias.*

D. Limitations and future work

In this paper, we have analyzed CLIP via XAI methods by observing and interpreting the explanation heat maps on different samples, such as in Sections V-I and V-II. However, it should be noted that there are limitations of heat map-based analysis, which stem from two aspects. Firstly, the observation results are illustrated through a small set of examples (compared to the size of the dataset), where limited samples will reduce the credibility of conclusions. Secondly,

the faithfulness of any XAI method cannot reach a hundred percent, and the importance ranks obtained by certain methods cannot perfectly display both the necessity and sufficiency [74]. Therefore, conclusions drawn from XAI results from certain methods are subject to some extent. In future work, we will consider when given image-sentence pair (with a sentence containing multiple words), how to associate each important word from the sentence to each important region in the image, and vice versa. Future work can also consider how to extend Grad-ECLIP to other VLMs with modified dual-encoder architectures, e.g ALBEF with additional cross-attention layers to fuse image and text features.

VI. Conclusion

In this paper, we propose Grad-ECLIP, a novel white-box gradient-based visual and textual explanation method for CLIP, the dual-encoder pre-trained model for image-text matching. Grad-ECLIP is applicable to both the image and text encoders, producing heat maps that indicate the importance of image regions or words for the image-text matching score. Qualitative and quantitative evaluations demonstrate the advantages of Grad-ECLIP compared with existing explanation methods designed for transformers/CLIP, and the adaptation experiments exhibit the generalizability of our method. Finally, we also adopt Grad-ECLIP to analyze the properties of the pre-trained CLIP model, where we discover its ability of concept decomposition and addibility, advantages/limitations on different attribute identification, and its bias towards learning concrete words over abstract words. By introducing these analyses as examples, we hope the proposed interpretation method can be used to help with both development and understanding of VLMs.

Acknowledgments

This work was supported by Strategic Research Grants from City University of Hong Kong (Project. Nos. 7030010 and 7005995).

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[2] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *CVPR*, 2021, pp. 3558–3568.

[3] J. Cha, K. Lee, S. Park, and S. Chun, "Domain generalization by mutual-information regularization with pre-trained models," in *ECCV*, 2022.

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

[5] Z. Wang, Y. Lu, Q. Li, X. Tao, Y. Guo, M. Gong, and T. Liu, "Cris: Clip-driven referring image segmentation," in *CVPR*, 2022.

[6] J. Xu, S. De Mello, S. Liu, W. Byeon, T. Breuel, J. Kautz, and X. Wang, "Groupvit: Semantic segmentation emerges from text supervision," in *CVPR*, 2022.

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

[8] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, 2022, pp. 12 888–12 900.

[9] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *IJCV*, vol. 130, no. 9, pp. 2337–2348, 2022.

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

[11] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *ECCV*, 2020.

[12] J. Wang, P. Zhou, M. Z. Shou, and S. Yan, "Position-guided text prompt for vision-language pre-training," in *CVPR*, 2023, pp. 23 242–23 251.

[13] Y. Zhong, J. Yang, P. Zhang, C. Li, N. Codella, L. H. Li, L. Zhou, X. Dai, L. Yuan, Y. Li *et al.*, "Regionclip: Region-based language-image pretraining," in *CVPR*, 2022, pp. 16 793–16 803.

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

[15] C. Zhou, C. C. Loy, and B. Dai, "Extract free dense labels from clip," in *ECCV*. Springer, 2022, pp. 696–712.

[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017, pp. 618–626.

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

[18] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in *ACL*, 2020, pp. 4190–4197.

[19] H. Chefer, S. Gur, and L. Wolf, "Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers," in *ICCV*, 2021, pp. 397–406.

[20] Y. Wang, T. G. Rudner, and A. G. Wilson, "Visual explanations of image-text representations via multi-modal information bottleneck attribution," *NeurIPS*, 2024.

[21] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *CVPR*, 2021, pp. 782–791.

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

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

[24] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, 2014, pp. 818–833.

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

[26] S. Srinivas and F. Fleuret, "Full-gradient representation for neural network visualization," in *NeurIPS*, vol. 32, 2019.

[27] D. Kim, A. Angelova, and W. Kuo, "Region-aware pretraining for open-vocabulary object detection with vision transformers," in *CVPR*, 2023.

[28] S. Wu, W. Zhang, L. Xu, S. Jin, X. Li, W. Liu, and C. C. Loy, "Clipself: Vision transformer distills itself for open-vocabulary dense prediction," *arXiv preprint arXiv:2310.01403*, 2023.

[29] C. Zhao, K. Wang, X. Zeng, R. Zhao, and A. B. Chan, "Gradient-based visual explanation for transformer-based clip," in *ICML*, 2024.

[30] Z. Wang, X. Liu, H. Li, L. Sheng, J. Yan, X. Wang, and J. Shao, "Camp: Cross-modal adaptive message passing for text-image retrieval," in *ICCV*, 2019.

[31] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015, pp. 2048–2057.

[32] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *ICCV*, 2015.

[33] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *ICCV*, 2015.

[34] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *WACV*, 2018, pp. 839–847.

[35] H. G. Ramaswamy *et al.*, "Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization," in *WACV*, 2020.

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

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

[38] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *ACM SIGKDD*, 2016.

[39] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *ICCV*, 2017, pp. 3429–3437.

[40] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *NeurIPS*, vol. 30, 2017.

[41] J. Wagner, J. M. Kohler, T. Gindele, L. Hetzel, J. T. Wiedemer, and S. Behnke, "Interpretable and fine-grained visual explanations for convolutional neural networks," in *CVPR*, 2019, pp. 9097–9107.
[42] J. Lee, J. Yi, C. Shin, and S. Yoon, "Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation," in *CVPR*, 2021.
[43] V. Petsiuk, R. Jain, V. Manjunatha, V. I. Morariu, A. Mehra, V. Ordonez, and K. Saenko, "Black-box explanation of object detectors via saliency maps," in *CVPR*, 2021, pp. 11 443–11 452.
[44] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern recognition*, vol. 65, pp. 211–222, 2017.
[45] W.-J. Nam, S. Gur, J. Choi, L. Wolf, and S.-W. Lee, "Relative attributing propagation: Interpreting the comparative contributions of individual units in deep neural networks," in *AAAI*, 2020.
[46] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *ICML*, 2017.
[47] J. Gu, Y. Yang, and V. Tresp, "Understanding individual decisions of cnns via contrastive backpropagation," in *ACCV*, 2019, pp. 119–134.
[48] B. K. Iwana, R. Kuroki, and S. Uchida, "Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation," in *ICCVW*, 2019.
[49] Y. Qiang, D. Pan, C. Li, X. Li, R. Jang, and D. Zhu, "Attcat: Explaining transformers via attentive class activation tokens," *NeurIPS*, 2022.
[50] W. Xie, X.-H. Li, C. C. Cao, and N. L. Zhang, "Vit-cx: Causal explanation of vision transformers," in *IJCAI*, 2023, pp. 1569–1577.
[51] L. Yu and W. Xiang, "X-pruner: explainable pruning for vision transformers," in *CVPR*, 2023, pp. 24 355–24 363.
[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
[53] T. Sarlos, "Improved approximation algorithms for large matrices via random projections," in *2006 47th annual IEEE symposium on foundations of computer science (FOCS'06)*. IEEE, 2006, pp. 143–152.
[54] K. Vu, P.-L. Poirion, and L. Liberti, "Random projections for linear programming," *Mathematics of Operations Research*, vol. 43, no. 4, pp. 1051–1071, 2018.
[55] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.
[56] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *IJCV*, vol. 115, pp. 211–252, 2015.
[57] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.
[58] H. Wang, S. Ge, Z. Lipton, and E. P. Xing, "Learning robust global representations by penalizing local predictive power," *NeurIPS*, 2019.
[59] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," in *CVPR*, 2021, pp. 15 262–15 271.
[60] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *ACL*, 2018, pp. 2556–2565.
[61] B. Boecking, N. Usuyama, S. Bannur, D. C. Castro, A. Schwaighofer, S. Hyland, M. Wetscherek, T. Naumann, A. Nori, J. Alvarez-Valle *et al.*, "Making the most of text semantics to improve biomedical vision–language processing," in *ECCV*, 2022.
[62] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
[63] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *NeurIPS*, vol. 34, pp. 9694–9705, 2021.
[64] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Trans. Neural Netw Learn Syst*, vol. 28(11), pp. 2660–73, 2016.
[65] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," *IJCV*, vol. 126(10), pp. 1084–102, 2018.
[66] C. Zhao and A. B. Chan, "Odam: Gradient-based instance-specific visual explanation for object detection," in *ICLR*, 2022.
[67] Y. Wang, T. G. Rudner, and A. G. Wilson, "Visual explanations of image-text representations via multi-modal information bottleneck attribution," *Advances in Neural Information Processing Systems*, vol. 36, pp. 16 009–16 027, 2023.
[68] S. Gao, Z.-Y. Li, M.-H. Yang, M.-M. Cheng, J. Han, and P. Torr, "Large-scale unsupervised semantic segmentation," *TPAMI*, 2022.
[69] P. Hase and M. Bansal, "Evaluating explainable AI: Which algorithmic explanations help users predict model behavior?" in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 5540–5552. [Online]. Available: https://aclanthology.org/2020.acl-main.491/
[70] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in *CVPR*, 2017.
[71] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," *arXiv:2304.02643*, 2023.
[72] M. Brysbaert, A. B. Warriner, and V. Kuperman, "Concreteness ratings for 40 thousand generally known english word lemmas," *Behavior research methods*, vol. 46, pp. 904–911, 2014.
[73] A. Gokaslan and V. Cohen, "Openwebtext corpus," http://Skylion007.github.io/OpenWebTextCorpus, 2019.
[74] Z. Wang, P. Mardziel, A. Datta, and M. Fredrikson, "Interpreting interpretations: Organizing attribution methods by criteria," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 10–11.

Chenyang Zhao received the B.Eng. degree in Electrical Engineering from Xiamen University, Xiamen, China, and M.S. degree in Computer Science from School of Electronic and Computer Engineering, Peking University, Shenzhen, China, in 2016 and 2019, respectively. She is currently working towards the Ph.D. degree in Computer Science at the City University of Hong Kong. Her research interests include explainable AI and object detection.

Kun Wang is a senior researcher at SenseTime Group Limited. He holds an MPhil degree from the Department of Electronic Engineering at the Chinese University of Hong Kong. His research interests focus on computer vision and representation learning. Presently, he is engaged in developing applications utilizing large language models and large multimodal models within the industry.

Janet H. Hsiao received the B.S. degree in Computer Science & Information Engineering from National Taiwan University, the M.S. degree in Computing Science from Simon Fraser University, and the Ph.D. degree in Informatics from University of Edinburgh. She is currently a Professor in the Division of Social Science and Department of Computer Science & Engineering at Hong Kong University of Science and Technology. She is also a Fellow of the Cognitive Science Society and serves on the Governing Board. Her research interests include cognitive science, computational modelling, learning and visual cognition, and explainable AI.

Antoni B. Chan received the B.S. and M.Eng. degrees in electrical engineering from Cornell University, Ithaca, NY, in 2000 and 2001, and the Ph.D. degree in electrical and computer engineering from the University of California, San Diego (UCSD), San Diego, in 2008. He is currently a Professor in the Department of Computer Science, City University of Hong Kong. His research interests include computer vision, machine learning, pattern recognition, and music analysis.