Enhancing Trust in AI-Driven Dermatology: CLIP for Explainable Skin Lesion Diagnosis

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Skin carcinoma is the most common cancer worldwide and costs over $8 billion annually. Early diagnosis is vital for improving melanoma survival rates from 23% to 99%. Deep neural networks show promising results in classifying skin lesions as benign or malignant, but black-box methods are typically not trusted by doctors. In this paper we use the CLIP (Contrastive Language-Image Pretraining) model, trained on various skin lesion datasets, to capture meaningful relationships between visual features and related diagnostic terms in an effort to increase explainability. We also use a gradient-based visual explanation method for CLIP, known as Grad-ECLIP, which highlights the critical regions in images linked to specific diagnostic descriptions. This pipeline not only classifies skin lesions and generates corresponding descriptions but also adds a layer of visual explanations.

1 Introduction

Cancer is characterized by the uncontrolled growth of body cells and is a major global health concern. Among its various forms, skin cancer is the most common, primarily affecting areas of the body frequently exposed to the sun, such as the face, lips, back, head, and legs. The primary cause of skin cancer is excessive exposure to ultraviolet (UV) radiation, which can lead to life-threatening conditions in as little as six weeks. Early identification of skin diseases is critical as it can significantly improve outcomes and reduce healthcare costs.

Various methods have been developed to detect and differentiate skin lesion types [8] [1] [10] [2] [27]. Melanomas, the most serious form of skin cancer, exhibit a range of characteristics, such as the presence or absence of pigmentation and diagnostic features like the whitish veil. Clinicians have established different guidelines such as the ABCDE rule—Asymmetry, Border irregularity, Color variation, Diameter, and Evolution—to track changes in lesions [22]. However, variations in image resolution can complicate the extraction of lesion diameters, and these features alone are often insufficient for accurately diagnosing different types of melanomas. Consequently, the Menzies method was developed as a simplified dermoscopy technique for diagnosing melanomas, focusing on the presence or absence of "negative" and "positive" features [31]. Despite its improved accuracy over the ABCDE rule, the Menzies method has high sensitivity [19] [4], which can lead to false-positives, especially when used by less experienced clinicians. To overcome the limitations of the Menzies method, the 7-point checklist was introduced [33]. However, this method also presents challenges for non-experts, as accurate diagnosis without specialized tools is difficult.

The complexity of diagnosing skin lesions highlights the need for manual evaluation by clinicians. Nonetheless, automated techniques using deep neural networks offer promising solutions by improving the precision and reliability of skin lesion detection and classification [17] [36]. Despite their potential, these methods are often perceived as "black boxes", making it challenging for clinicians to trust their outputs. While some studies have focused on enhancing the explainability of medical data to build transparency and trust [9] [23], they have not addressed the importance of highlighting specific regions in relation to their corresponding textual descriptions, which would further enhance explainability and interpretability. Furthermore, no existing classification method fully integrates all diagnostic techniques.

To address this research gap we developed a pipeline that fine-tunes the Contrastive Language-Image Pretraining (CLIP) model [25] on skin lesion datasets including images along with their descriptions using features from all diagnostic techniques. By employing the gradient-based method Grad E-CLIP [38] we enhance explainability by visually and textually highlighting the features in an image that are most relevant to the diagnosis. This method also illustrates how specific textual descriptions correspond to these highlighted regions, thereby bridging the gap between visual data and diagnostic terms. This enhanced transparency promotes trust among clinicians, enabling them to understand and verify the AI's decision-making process.

2 Related Work

CLIP has generated significant interest in a number of medical domains. Med-CLIP [34] uses a semantic matching loss based on medical knowledge to improve zero-shot prediction, supervised classification, and image-text retrieval. eCLIP [14] incorporates radiologist eye-gaze heatmaps to address data scarcity. Mammo-CLIP [6] processes multi-view mammograms and corresponding text using early feature fusion, and ConVIRT [37] is an unsupervised strategy for pretraining medical image encoders using paired descriptive text through a bidirectional contrastive objective. MITER (Medical Image–TExt joint adaptive pRetraining) [30] proposes a joint adaptive pretraining framework that combines multi-level contrastive learning with dynamic hard negative sample selection to enhance medical image and text models. pathCLIP [11] uses image-text contrastive learning to create embeddings of image snippets and text descriptions for better identification of genes and gene relations. CLIPath [15] uses Residual Feature Connections to fine-tune CLIP with few trainable parameters by fusing task-specific and pre-trained knowledge, enhancing performance on pathology classification tasks with limited annotated samples. PubMedCLIP [7], a fine-tuned version of CLIP for the medical domain using PubMed articles, outperforms state-of-the-art MAML networks on MedVQA benchmarks by up to 3% in overall accuracy. Despite these advancements, several research gaps remain in using CLIP models for medical tasks, particularly concerning generalizability across diverse medical domains and explainability.

Numerous studies have focused on enhancing the explainability of models applied to medical data [21]. In the realm of image-based explanations, the primary objective is to identify the specific parts of an image (such as pixels or segments) that most significantly influence a model's prediction. Prominent techniques for this purpose include gradient-based methods for convolutional neural networks (CNNs), such as Guided Backpropagation, CAM, Grad-CAM [28], GradCAM++ [5], Guided GradCAM [29], SmoothGrad [32] and DeepLIFT [16]. These were developed to further enhance the interpretability of model predictions by offering more refined visual explanations that highlight the regions of the input most responsible for the model's decisions in medical data.

In addition to gradient-based methods, other approaches like SHAP (SHapley Additive exPlanations) [18], LIME (Local Interpretable Model-agnostic Explanations) [26] and Layer-wise Relevance Propagation (LRP) [3] have been developed to provide more generalizable explanations across different types of data. These techniques offer insights into the contribution of individual features to the model's predictions, thereby enhancing the interpretability of AI systems in healthcare.

For a broader understanding of how a model operates across different data points, global explanation methods like SP-LIME offer a comprehensive view of a model's behavior by selecting diverse, representative explanations. These techniques help clinicians understand model predictions more thoroughly, building trust and ensuring safe AI deployment in healthcare. However, some medical data often rely on multiple data sources such as images, EHR and clinical notes. Grad E-Clip's ability to generate comprehensive explanations across different modalities is a largely unexplored area.

3 Method Overview

3.1 Dataset

Because our work requires annotated images with specific dermoscopic structure criteria, we used the PH² and Derm7pt datasets. The PH² [20] image database contains a total of around 200 dermoscopic images of melanocytic lesions, including common nevi, atypical nevi, and melanomas. The PH² database includes clinical and histological diagnoses and the identification of several dermoscopic structure criteria (colors, pigment network, dots/globules, streaks, regression areas, blue-whitish veil).

Similarly, Derm7pt [13] is a dermoscopic image dataset that contains over 2000 clinical and dermoscopy images along with corresponding structured metadata tailored for training and evaluating computer aided diagnosis (CAD) systems. This dataset includes the 7-point checklist for assessing the malignancy of skin lesions, making it a valuable resource for our study.

3.2 Data Prepration

Once the data are collected, each image is paired with its corresponding text description. The dataset is organized so that each row represents a single image-text pair, with duplicates removed to avoid overfitting and redundancy.

For text preprocessing, special characters and unnecessary punctuation are removed. At the same time, images are resized to 224x224 pixels to meet the input requirements of the image encoder. To increase the number of image-text pairs, augmentations are applied: images are augmented through flipping and rotating, while text descriptions are reordered to create variations for the same image. These augmented text descriptions are then tokenized to create a format compatible with the CLIP model's text encoder, splitting the text into tokens (words or subwords) that can be converted into embeddings. The images are then fed into the image encoder, and the text is fed into the text encoder. These augmented image-text pairs are then subsequently split into training and testing datasets. This careful pairing and preprocessing of images and text is crucial, as CLIP relies on learning the relationships between image-text pairs to function effectively.

3.3 Contrastive Learning Image Pretained - CLIP

Contrastive Language-Image Pre-training (CLIP) has shown its capability to learn distinctive visual representations and generalize across a wide range of downstream vision tasks. Trained on a dataset of 400 million image-text pairs sourced from the web, CLIP effectively aligns image and text features, allowing for rich incorporation of diverse visual concepts. This extensive pre-training enhances the transferability of the learned features to various applications.

As shown in Figure 1, CLIP consists of two key components: an image encoder and a text encoder, both of which are jointly trained to extract feature embeddings from images and text into a shared representation space. In this study, a pre-trained model with a vision transformer (ViT) is used as an image encoder, while a transformer-based encoder is used for text. Given an image-text pair (I, T), the matching score between their extracted image features $f_I \in \mathbb{R}^D$ and text features $f_T \in \mathbb{R}^D$ is:

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

CLIP maximizes the cosine similarity between embeddings of positive pairs, while minimizing it for negative pairs using a contrastive loss.

3.3.1 Fine-Tuning CLIP

We conducted our experiments on Google Colab with a TESLA T4 GPU. Fine-tuning the CLIP model, the most computationally intensive task, took less than an hour for each 30-epoch run. CLIP was fine-tuned on colored dermoscopic images collected from the PH² and Derm7pt datasets. These images were paired with dermoscopic structure criteria, which served as descriptive annotations. Since CLIP is trained to align images with their corresponding text features, we utilized these descriptive annotations during training, resulting in updated weights that were subsequently saved. These fine-tuned weights were then employed for the classification of new image-text pairs.

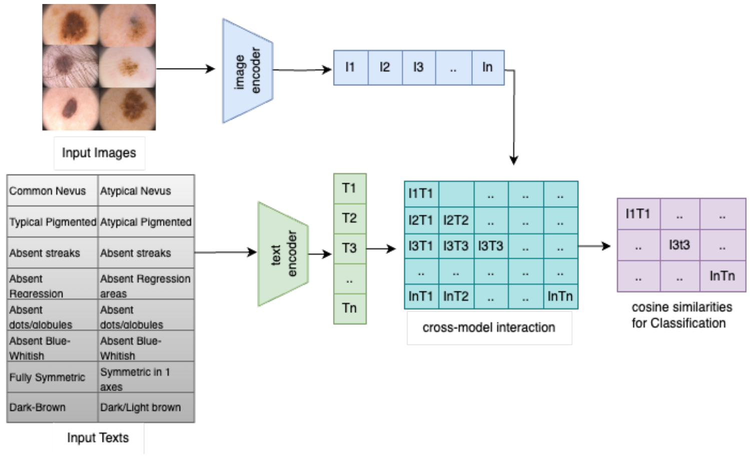


image encoder
I1
I2
I3
..
In
Input Images
Common Nevus | Atypical Nevus
Typical Pigmented | Atypical Pigmented
Absent streaks | Absent streaks
Absent Regression | Absent Regression areas
Absent dots/globules | Absent dots/globules
Absent Blue-Whitish | Absent Blue-Whitish
Fully Symmetric | Symmetric in 1 axes
Dark-Brown | Dark/Light brown
Input Texts
text encoder
T1
T2
T3
..
Tn
I1T1 | | .. | .. | ..
I2T1 | I2T2 | .. | .. | ..
I3T1 | I3T3 | I3T3 | .. | ..
.. | .. | .. | .. | ..
InT1 | InT2 | .. | .. | InTn
cross-model interaction
I1T1 | .. | ..
.. | I3t3 | ..
.. | .. | InTn
cosine similarities for Classification

Figure 1: **CLIP Overview for Custom Dataset:** Given skin lesion images, we encode them alongside their descriptive criteria to generate image and text embeddings. These embeddings are then combined in a cross-modal interaction module, where cosine similarities between the image-text pairs are calculated to assess the alignment between lesions and their diagnoses. The final classification output is determined by the degree of alignment, ensuring accurate diagnosis.

Formally, as illustrated in Figure 1, let f_x represent the image features extracted by CLIP's image encoder for lesion image x. The text features, which include descriptive criteria such as "melanoma", "symmetry", and "presence of blue-whitish veil", are extracted using CLIP's text encoder, resulting in a set of w_i features $W = \{w_i\}_{i=1}^{K}$, where K represents the number of classes, such as Melanoma, Atypical Nevus, Common Nevus, Seborrheic Keratosis, etc., along with their descriptions. The probability of predicting class i (e.g., melanoma) given input image x is computed as:

$$p(y = i|x) = \frac{\exp(\cos(w_i, f_x)/\tau)}{\sum_{j=1}^{K} \exp(\cos(w_j, f_x)/\tau)}, \tag{2}$$

where $\cos(\cdot, \cdot)$ denotes the cosine similarity between two vectors, and τ is a scaling factor learned by CLIP [35]. During fine-tuning, the model learns to maximize the cosine similarity between the image features f_x and the correct text features w_i for the true class. Simultaneously, it minimizes the cosine similarity between f_x and text features w_j for all incorrect classes j $\neq$ i. This fine-tuning aligns the image and text embeddings in the feature space, enhancing the model's ability to accurately match images with their corresponding diagnoses.

3.4 Explainable AI (XAI)

With the increasing use of deep learning for detection, classification, and segmentation of medical images, it has become challenging for clinicians to trust these models due to their black box nature. Therefore, building trust and transparency in their output is crucial for user acceptance.

Various XAI methods have been developed for different tasks [21]. SHAP (SHapley Additive Explanations) represents a game theoretic approach by computing the importance of input features (image pixels) with respect to model output [18]. LIME (Local Interpretable Model-agnostic Explanations) [26] is a model-agnostic algorithm that generates interpretable, locally faithful explanations for the predictions of any classifier. Layer-wise relevance propagation (LRP) is another XAI explanation technique applicable to models structured as neural networks. It assigns relevance scores to each neuron in the model and shows the importance of different neurons by propagating the prediction backwards in the neural network by means of purposely designed local propagation rules for the decision of the model [3].

Saliency maps are a popular technique used to highlight the key regions in input data that significantly contribute to a given prediction. In the domain of dermatology, a series of Class Activation Mapping (CAM) techniques (CAM [12], Grad-CAM [28], and Grad-CAM++ [5]) have been employed to explain CNN models for image analysis. Each method has its limitations, prompting further development. CAM, for example, is limited to CNNs with a Global Average Pooling (GAP) layer before the fully connected layer and requires retraining multiple linear classifiers after training the base model. Grad-CAM addresses this by introducing a backpropagation concept that considers partial derivatives to solve for the weight independent from the position of a particular activation map. However, Grad-CAM's heatmaps may fail to localize the entire region of the object, which led to the development of Grad-CAM++. It uses a weighted combination of the positive partial derivatives of the last convolutional layer feature maps to produce more detailed heatmaps, even though related features might be confined to a limited pixel area.

These explainability methods were designed to focus on image-only or text-only data, and are not suited for models that handle both image and text inputs simultaneously, explaining how text relates to the image. To bridge this gap, we applied Grad E-CLIP [38], a technique specifically developed for the CLIP model, which effectively addresses the challenges of image-text explainability. This method provides valuable insights into how the CLIP model makes its predictions by highlighting the connections between visual features and their corresponding textual descriptions. By doing so, Grad E-CLIP enhances our understanding of the model's decision-making process.

3.4.1 Gradient-based Explanation for CLIP (Grad E-CLIP)

Grad-ECLIP is a method designed to provide visual explanations for the CLIP model by analyzing the output of attention layers, particularly focusing on the final layer. The method works by examining the interaction between the class token and spatial feature maps within the model, and by calculating the importance of each channel and spatial location using a modified attention mechanism.

The math behind Grad-ECLIP is in section A.2. For the purposes of this paper, Grad-ECLIP improves interpretability of the CLIP model by aggregating explanations across all layers, capturing the contributions of features throughout the model. By applying this method to both the image and text encoders, Grad-ECLIP effectively addresses the black-box nature of CLIP. In this study, we demonstrate the utility of Grad-ECLIP in explaining different description criteria associated with each type of skin lesion.

4 Our Approach

In this study, we developed a fully connected pipeline for the classification and differentiation of various skin lesions. As illustrated in Figure 2, data were collected from two different databases, including images and their corresponding text descriptions. The collected images and text were then pre-processed, involving resizing, organization, and data augmentation. The dataset was split into 75% training data and 25% testing data. The training data were used to fine-tune the CLIP model, which was trained for 30 epochs with a batch size of 64, using the Adam optimizer with a learning rate of 1e-5. The loss use was the mean of image and text cross-entropy (see section A.1 for details).

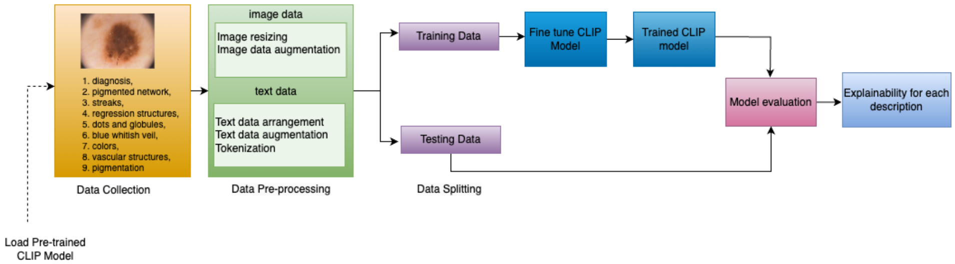


Figure 2: Proposed pipeline

After training, the new weights were used to evaluate the model on the test data. To enhance the interpretability of the newly trained CLIP model, explainability was applied using Grad E-CLIP, which provides visual and textual insights into the model's decision-making process.

We used the state-of-the-art CLIP model that has not been extensively explored for skin lesion classification. Our approach extends beyond just applying existing explainability techniques like Grad E-CLIP to a pre-trained CLIP model by developing a comprehensive classification framework that effectively integrates image and text pairs. This dual-modality strategy not only enhances the model's ability to differentiate between various skin lesions but also deepens our understanding of the relationship between visual and textual data.

A critical aspect of our methodology is the analysis of visual and textual features highlighted by Grad E-CLIP, which helps us identify and address potential biases in the model's predictions during training. By fine-tuning the CLIP model with our dataset, we aim to improve both its accuracy and relevance for the skin lesion classification task. This approach not only enhances classification accuracy but also emphasizes the importance of interpretability and transparency, making a significant contribution to AI-driven medical diagnostics.

5 Experiments

We conducted experiments using the ViT-B/16 architecture, which is based on a transformer model with a 16x16 patch size. The experiments were conducted in two main parts.

1) Performance Evaluation of Fine-tuning the CLIP Model on a Custom Dataset:

In this part we evaluate the performance of the pre-trained CLIP model on a custom skin lesion dataset, followed by the performance of the model after fine-tuning it on the same dataset. The dataset was split into 75% for training and 25% for testing. The testing dataset was used to evaluate both the pre-trained and the fine-tuned CLIP models, allowing for a direct comparison of their performance. We found that the performance of the CLIP model improved significantly after fine-tuning on the custom dataset. The evaluation metrics for the training data are presented in Table 1, while Table 2 shows the performance on the test data before and after fine-tuning. The loss reported in the tables is cross-entropy.

Evaluation Metrics	**Value**
Accuracy	81.80%
Loss	0.4771
Precision	0.8195
Recall	0.818
F1-score	0.8179
Sensitivity	0.818
Specificity	0.9971

Table 1: Model metrics on training data after fine-tuning

	Before fine-tuning	**After fine-tuning**
Number of test samples	1215	1215
Batch size	64	64
Accuracy	2.06%	80.08%
F1-Score	0.0153	0.8011
Average Loss	4.1579	0.4954
Average CLIP Score	0.3081	0.9655

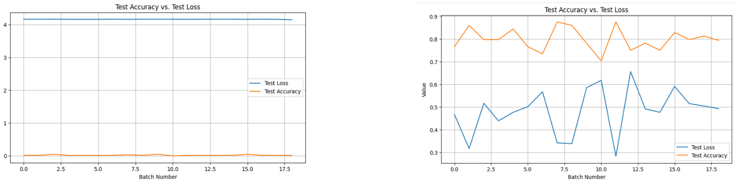
Table 2: Model metrics on test data before and after fine-tuning

Accuracy, being the most commonly used metric, evaluates the overall performance of deep learning models by measuring the proportion of correct predictions out of all predictions made. In addition to accuracy, other evaluation metrics such as Sensitivity, Specificity, Precision, and F1-score are also assessed for the CLIP model. These provide a more comprehensive evaluation of the model's performance by offering insights into its ability to correctly identify true positives, avoid false negatives, and maintain a balance between precision and recall. Furthermore, the CLIP Score (S_{CLIP}) is calculated as the cosine similarity between the image and text embeddings.

$$S_{\text{CLIP}} = \frac{f_{\text{img}}(I) \cdot f_{\text{text}}(T)}{\|f_{\text{img}}(I)\| \|f_{\text{text}}(T)\|} \tag{3}$$

These metrics indicate the effectiveness of the learning algorithm, as the training curves reach a point of stability. In contrast, Figure 3 shows the learning curves for test accuracy and loss before and after fine-tuning on the custom dataset, respectively. It is clear that the performance of the CLIP model improves significantly after fine-tuning on the custom dataset, leading to enhanced classification performance. Similar plots for training accuracy and training loss can be found in section A.3.

2) Performance Evaluation of CLIP's Explainability on a Custom Dataset:



(a) Test accuracy and loss on pre-trained model

(b) Test accuracy and loss on fine-tuned model

Figure 3: Comparison of test accuracy and loss for pre-trained and fine-tuned models

The second part of the experiments focused on evaluating the explainability of the CLIP models both before and after fine-tuning. Gradient-based explainability methods, such as Grad-CAM and Grad E-CLIP, were employed to analyze the image-text pair understanding of model's decision-making process on the skin lesion dataset. The results indicate that fine-tuning not only improved the model's accuracy on the skin lesion dataset but also influenced the explainability of the model's outputs.

Figures 4, 6 and 8 present the explainability results from the pre-trained CLIP model, comparing Grad E-CLIP and Grad-CAM. For the CLIP model, Grad-CAM was evaluated based on the cosine similarity of the image-text pair, using the gradients calculated with respect to the patch tokens from the ViT layers. In contrast, Figures 5, 7 and 9 show the explainability results after fine-tuning. Each column in these figures represents a different skin lesion condition, with characteristics such as "common nevus", "typical pigmented", or "absent streaks". The top row in these figures shows the Grad E-CLIP visualizations, highlighting areas of the image that contribute to the model's predictions. The bottom row shows the Grad-CAM visualizations, with heatmaps indicating regions of importance in the image for the model's output. These figures suggest that Grad E-CLIP provides superior explainability in relation to each input text compared to Grad-CAM. While in few cases Grad E-CLIP's performance on the fine-tuned model is not as strong as its original performance on the pre-trained CLIP model, it produces clearer and more focused visualizations that avoid highlighting irrelevant areas, resulting in better alignment with the corresponding texts.

As discussed in [38], the CLIP model excels at identifying common perceptual attributes such as color, but it struggles with physical attributes like shape and material, and is less effective at grounding objects with comparative attributes, like size and positional relationships. The explainability visualizations, as shown in Figure 9, clearly highlight these strengths and weaknesses of the CLIP model. For instance, in the explanations of Common Nevus, it is evident that CLIP performs better when color is provided as a text input, compared to other attributes like absent streaks or full asymmetry.

3) Performance Evaluation of Insertion and Deletion for Grad E-CLIP Explainability:

To evaluate the effectiveness of explanations provided by machine learning models, several metrics have been developed, including the area focus score, border focus score, and insertion and deletion metrics. The insertion and deletion metrics, introduced by [24], are widely used to assess the faithfulness of explanations.

The insertion metric measures the improvement in the model's performance as pixels, ranked by their importance, are gradually added to an empty image. A higher insertion score suggests that the heatmap has correctly identified the most important pixels, resulting in a rapid increase in model performance as these pixels are reintroduced. Conversely, the deletion metric evaluates how much the model's prediction degrades as important pixels are sequentially removed from the image, based on their importance as indicated by the heatmap. A lower deletion score indicates that the heatmap has effectively identified the crucial pixels, leading to a swift decline in model performance when these pixels are removed.

	Melanoma	**Atypical pigmented**	**Present streaks**	**Regression areas**	**Atypical dots/globules**
Insertion ↑	0.2928	0.2913	0.2757	0.2808	0.2881
Deletion ↓	0.2801	0.2809	0.2955	0.2852	0.2743
	Blue-whitish veil	**Fully asymmetric**	**White/dark-brown/blue-gray/black**	**Missing vascular structures**	**Missing pigmentation**
Insertion ↑	0.2968	0.2868	0.2718	0.2980	0.2976
Deletion ↓	0.2846	0.2864	0.2964	0.2781	0.2785

Table 3: Comparison of Insertion and Deletion Metrics from Pre-trained Grad E-CLIP on Various Diagnostic Features. (Visualization shown in Figure 6)

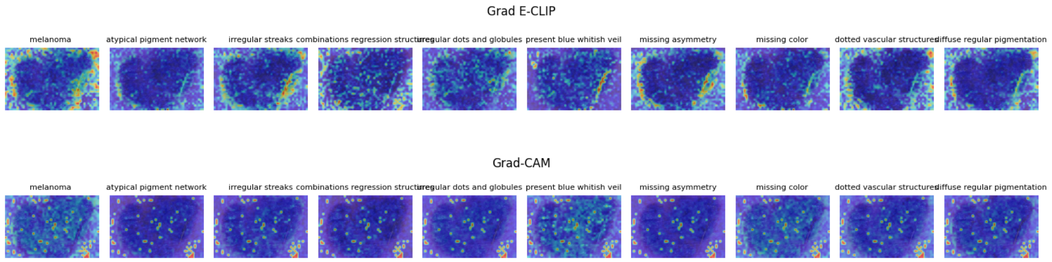

Grad E-CLIP
melanoma
atypical pigment network
irregular streaks
combinations regression structures
irregular dots and globules
present blue whitish veil
missing asymmetry
missing color
dotted vascular structures
diffuse regular pigmentation
Grad-CAM
melanoma
atypical pigment network
irregular streaks
combinations regression structures
irregular dots and globules
present blue whitish veil
missing asymmetry
missing color
dotted vascular structures
diffuse regular pigmentation

Figure 4: Comparison of Grad E-CLIP and Grad-CAM Visualizations on **Melanoma** using a Pre-trained CLIP Model.

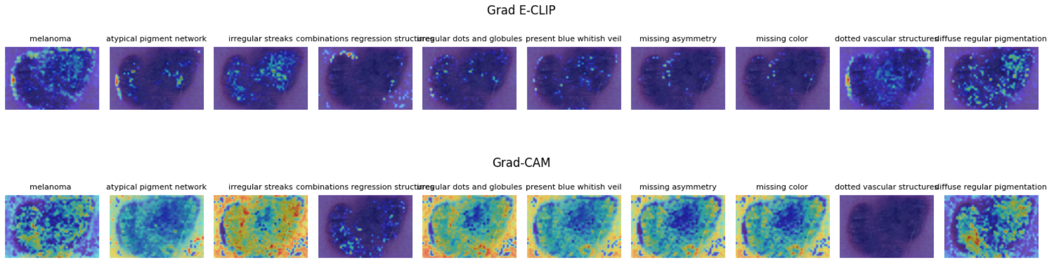

Grad E-CLIP
melanoma
atypical pigment network
irregular streaks
combinations regression structures
irregular dots and globules
present blue whitish veil
missing asymmetry
missing color
dotted vascular structures
diffuse regular pigmentation
Grad-CAM
melanoma
atypical pigment network
irregular streaks
combinations regression structures
irregular dots and globules
present blue whitish veil
missing asymmetry
missing color
dotted vascular structures
diffuse regular pigmentation

Figure 5: Comparison of Grad E-CLIP and Grad-CAM Visualizations on **Melanoma** using a Fine-tuned CLIP Model.

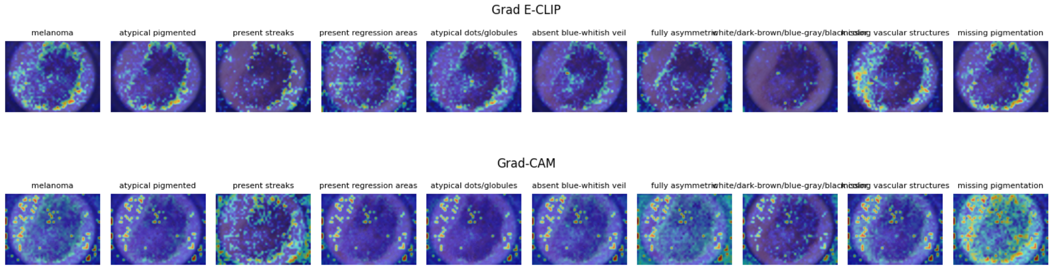


Figure 6: Comparison of Grad E-CLIP and Grad-CAM Visualizations on **Melanoma** Using a Pre-Trained CLIP Model.

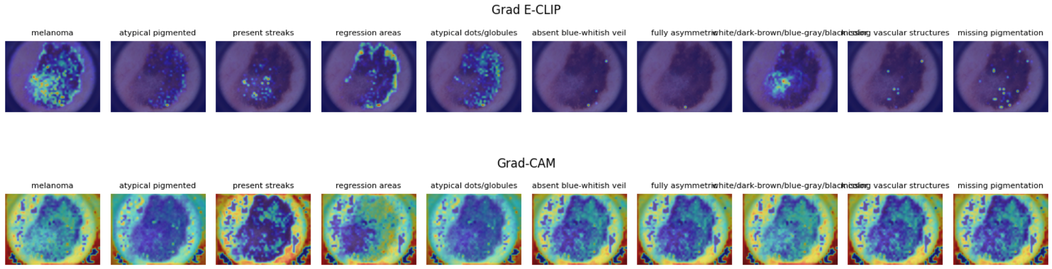


Figure 7: Comparison of Grad E-CLIP and Grad-CAM Visualizations on **Melanoma** using a Fine-tuned CLIP Model.

	Melanoma	**Atypical pigmented**	**Present streaks**	**Regression areas**	**Atypical dots/globules**
Insertion ↑	0.2294	0.2085	0.2088	0.2164	0.2136
Deletion ↓	0.2059	0.2149	0.2097	0.2160	0.2078
	Blue-whitish veil	**Fully asymmetric**	**White/dark-brown/blue-gray/black**	**Missing vascular structures**	**Missing pigmentation**
Insertion ↑	0.1980	0.1961	0.2243	0.1982	0.2049
Deletion ↓	0.2108	0.2070	0.2106	0.2087	0.2143

Table 4: Comparison of Insertion and Deletion Metrics from Fine-Tuned Grad E-CLIP on Various Diagnostic Features (Visualization shown in Figure 7)

In our study, we compared the evaluation metrics of Grad E-CLIP applied to both pre-trained and fine-tuned CLIP models as shown in Table 3 and Table 4, specifically for the Melanoma class (visually

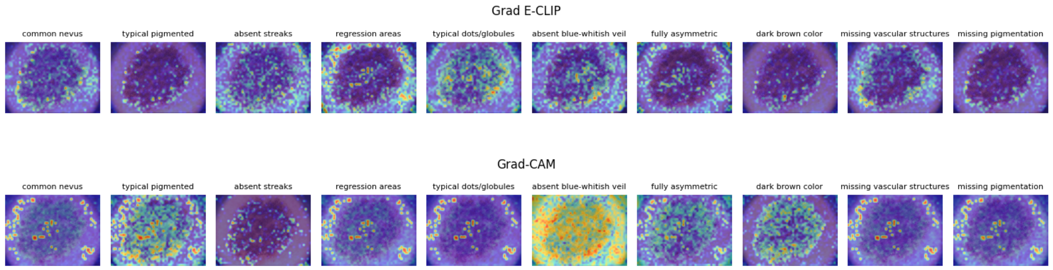


Figure 8: Comparison of Grad E-CLIP and Grad-CAM Visualizations on **Common Nevus** using a pre-trained CLIP Model.

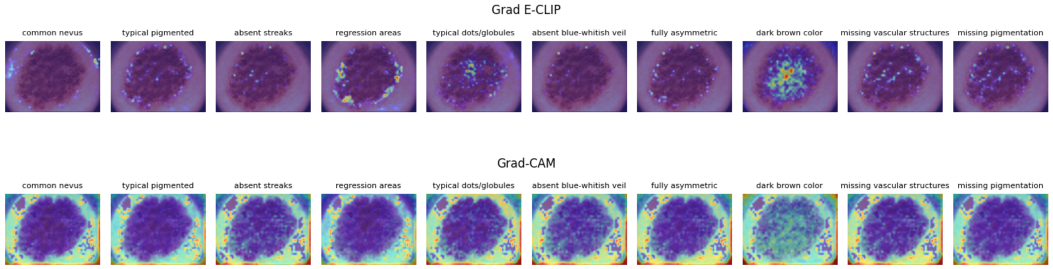


Figure 9: Comparison of Grad E-CLIP and Grad-CAM Visualizations on **Common Nevus** using a Fine-tuned CLIP Model.

illustrated in Figures 6 and 7) across different dermoscopic criteria. The results reveal that the Grad E-CLIP pre-trained model achieves higher insertion scores than the fine-tuned model, while the fine-tuned Grad E-CLIP model exhibits lower deletion scores.

The pre-trained Grad E-CLIP has higher insertion scores which suggest that these features might not be informative or relevant enough to significantly increase confidence when they are the only features present. However, this tends to focus on areas unrelated to melanoma, producing noisier heatmaps. This scatter of attention reduces its interpretability and precision, making it less reliable for identifying the critical regions relevant to melanoma diagnosis.

Conversely, the fine-tuned Grad E-CLIP model, despite showing a slightly lower Area Under Curve (AUC) for insertion, exhibits a clear advantage with its lower deletion score. These lower AUC for deletion, highlights the model's high sensitivity to the removal of key features, indicating that the model is identifying features that it heavily relies on crucial melanoma-related areas. This focused attention, with less noise, enhances the model's reliability and accuracy in pinpointing the most relevant regions, making the fine-tuned model more robust and trustworthy for clinical applications. The fine-tuned model effectively excludes non-essential regions, leading to improved precision in its output. While it sacrifices some of its region-retention capability, this trade-off results in better specificity and precision, which is particularly beneficial in medical data analysis where accurately highlighting only the most relevant regions is critical (AUC plots can be found in A.4).

6 Conclusion

This paper showed that fine-tuning the CLIP model on a custom skin lesion dataset significantly enhances both classification accuracy and explainability. The fine-tuned model not only achieves improved accuracy but also generates more precise and relevant visualizations when using gradient-based explainability methods (Grad E-CLIP). To the best of our knowledge, this is the first work that comprehensively uses CLIP and evaluates image-text pair explanations for skin lesions. There are limitations to this work, particularly in the explainability of image-text pair relevance for certain cases, such as common nevus and atypical nevus, where the alignment is less clear. In future work, we plan to enhance the dataset with more detailed descriptions of each skin lesion and improve explainability, focusing on better aligning image-text pairs in the skin lesion dataset to ensure stronger correlations.

References

[1] Adekanmi Adegun and Serestina Viriri. Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of state-of-the-art. *Artificial Intelligence Review*, 54(2):811–841, 2021.

[2] Md Shahin Ali, Md Sipon Miah, Jahurul Haque, Md Mahbubur Rahman, and Md Khairul Islam. An enhanced technique of skin cancer classification using deep convolutional neural network with transfer learning models. *Machine Learning with Applications*, 5:100036, 2021.

[3] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[4] Cristina Carrera, Michael A Marchetti, Stephen W Dusza, Giuseppe Argenziano, Ralph P Braun, Allan C Halpern, Natalia Jaimes, Harald J Kittler, Josep Malvehy, Scott W Menzies, et al. Validity and reliability of dermoscopic criteria used to differentiate nevi from melanoma: a web-based international dermoscopy society study. *JAMA dermatology*, 152(7):798–806, 2016.

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

[6] Xuxin Chen, Yuheng Li, Mingzhe Hu, Ella Salari, Xiaoqian Chen, Richard LJ Qiu, Bin Zheng, and Xiaofeng Yang. Mammo-clip: Leveraging contrastive language-image pre-training (clip) for enhanced breast cancer diagnosis with multi-view mammography. *arXiv preprint arXiv:2404.15946*, 2024.

[7] Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1181–1193, 2023.

[8] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[9] Zongyuan Ge, Sergey Demyanov, Rajib Chakravorty, Adrian Bowling, and Rahil Garnavi. Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*, pages 250–258. Springer, 2017.

[10] Nils Gessert, Maximilian Nielsen, Mohsin Shaikh, René Werner, and Alexander Schlaefer. Skin lesion classification using ensembles of multi-resolution efficientnets with meta data. *MethodsX*, 7:100864, 2020.

[11] Fei He, Kai Liu, Zhiyuan Yang, Yibo Chen, Richard D Hammer, Dong Xu, and Mihail Popescu. pathclip: Detection of genes and gene relations from biological pathway figures through image-text contrastive learning. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[12] Hyungsik Jung and Youngrock Oh. Towards better explanations of class activation mapping. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1336–1344, 2021.

[13] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, Ghassan Hamarneh, and Seven-Point Checklist. Skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, 23:538–546, 2019.

[14] Yogesh Kumar and Pekka Marttinen. Improving medical multi-modal contrastive learning with expert annotations. *arXiv preprint arXiv:2403.10153*, 2024.

[15] Zhengfeng Lai, Zhuoheng Li, Luca Cerny Oliveira, Joohi Chauhan, Brittany N Dugger, and Chen-Nee Chuah. Clipath: Fine-tune clip with visual feature fusion for pathology image analysis towards minimizing data collection efforts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2374–2380, 2023.

[16] Junbing Li, Changqing Zhang, Joey Tianyi Zhou, Huazhu Fu, Shuyin Xia, and Qinghua Hu. Deep-lift: Deep label-specific feature learning for image annotation. *IEEE transactions on Cybernetics*, 52(8):7732–7741, 2021.

[17] Adria Romero Lopez, Xavier Giro-i Nieto, Jack Burdick, and Oge Marques. Skin lesion classification from dermoscopic images using deep learning techniques. In *2017 13th IASTED international conference on biomedical engineering (BioMed)*, pages 49–54. IEEE, 2017.

[18] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[19] Ammara Masood and Adel Ali Al-Jumaily. Computer aided diagnostic support system for skin cancer: a review of techniques and algorithms. *International journal of biomedical imaging*, 2013(1):323268, 2013.

[20] Teresa Mendonça, Pedro M Ferreira, Jorge S Marques, André RS Marcal, and Jorge Rozeira. Ph 2-a dermoscopic image database for research and benchmarking. In *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 5437–5440. IEEE, 2013.

[21] Michael Munn and David Pitman. *Explainable AI for practitioners*. " O'Reilly Media, Inc.", 2022.

[22] Franz Nachbar, Wilhelm Stolz, Tanja Merkle, Armand B Cognetta, Thomas Vogt, Michael Landthaler, Peter Bilek, Otto Braun-Falco, and Gerd Plewig. The abcd rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions. *Journal of the American Academy of Dermatology*, 30(4):551–559, 1994.

[23] Natasha Nigar, Muhammad Umar, Muhammad Kashif Shahzad, Shahid Islam, and Douhadji Abalo. A deep learning approach based on explainable artificial intelligence for skin lesion classification. *IEEE Access*, 10:113715–113725, 2022.

[24] V Petsiuk, A Das, and K Saenko. Rise: Randomized input sampling for explanation of black-box models. arxiv 2018. *arXiv preprint arXiv:1806.07421*, 1806.

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

[26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[27] Thomas G Salopek, Alfred W Kopf, Catherine M Stefanato, Katrien Vossaert, Mark Silverman, and Sandhya Yadav. Differentiation of atypical moles (dysplastic nevi) from early melanomas by dermoscopy. *Dermatologic clinics*, 19(2):337–345, 2001.

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

[29] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016.

[30] Chang Shu, Yi Zhu, Xiaochu Tang, Jing Xiao, Youxin Chen, Xiu Li, Qian Zhang, and Zheng Lu. Miter: Medical image–text joint adaptive pretraining with multi-level contrastive learning. *Expert Systems with Applications*, 238:121526, 2024.

[31] Deepika Singh, Diwakar Gautam, and Mushtaq Ahmed. Detection techniques for melanoma diagnosis: A performance evaluation. In *2014 International Conference on Signal Propagation and Computer Technology (ICSPCT 2014)*, pages 567–572. IEEE, 2014.

[32] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[33] Fiona M Walter, A Toby Prevost, Joana Vasconcelos, Per N Hall, Nigel P Burrows, Helen C Morris, Ann Louise Kinmonth, and Jon D Emery. Using the 7-point checklist as a diagnostic aid for pigmented skin lesions in general practice: a diagnostic validation study. *British Journal of General Practice*, 63(610):e345–e353, 2013.

[34] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022.

[35] Ziyi Wu, Xudong Liu, and Igor Gilitschenski. Eventclip: Adapting clip for event-based object recognition. *arXiv preprint arXiv:2306.06354*, 2023.

[36] Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen. Attention residual learning for skin lesion classification. *IEEE transactions on medical imaging*, 38(9):2092–2103, 2019.

[37] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022.

[38] Chenyang Zhao, Kun Wang, Xingyu Zeng, Rui Zhao, and Antoni B Chan. Gradient-based visual explanation for transformer-based clip. In *International Conference on Machine Learning*, pages 61072–61091. PMLR, 2024.

A Appendix

A.1 Loss Functions Used to Fine-Tune CLIP

During training, the loss was calculated as follows:

$$\text{Loss}_{img} = \frac{1}{N}\sum_{i=1}^{N} \text{CrossEntropyLoss}(\mathbf{L}_{img}^{(i)}, \mathbf{y}^{(i)}) \tag{4}$$

$$\text{Loss}_{txt} = \frac{1}{N}\sum_{i=1}^{N} \text{CrossEntropyLoss}(\mathbf{L}_{txt}^{(i)}, \mathbf{y}^{(i)}) \tag{5}$$

$$\text{Total Loss} = \frac{\text{Loss}_{img} + \text{Loss}_{txt}}{2} \tag{6}$$

N is the batch size, $\mathbf{L}_{img}^{(i)}$ and $\mathbf{L}_{txt}^{(i)}$ are the logits for the i-th image and text, respectively, and $\mathbf{y}^{(i)}$ is the ground truth label for the i-th sample.

A.2 Grad-ECLIP

The process starts by extracting the image embedding f_I from the class token $x_{\text{cls}}^{(0)}$ in the final layer of the network, where $x^{(1)}$ is the input to the last layer and $x^{(0)}$ is the output of the network. The class token from the penultimate layer, $x_{\text{cls}}^{(1)}$, after applying attention $\mathcal{A}$ is $\mathcal{A}(x_{\text{cls}}^{(1)})$. The image embedding is then computed by applying a linear projection (LP) to the sum of $\mathcal{A}(x^{(1)})[\text{cls}]$ and $x^{(1)}[\text{cls}]$, where [cls] denotes getting the feature vector from the class token.

$$f_I = \text{LP}(x_{\text{cls}}^{(0)}) = \text{LP}(\mathcal{A}(x^{(1)})[\text{cls}] + x^{(1)}[\text{cls}]), \tag{7}$$

The attention layer $\mathcal{A}$ calculates the contribution of each feature by aggregating the weighted outputs, determined by the softmax function applied to the scaled dot product of the query (q_{cls}), key (k_i), and value v_i embeddings as shown int the equation below, where C is the channel dimension:

$$x_{\text{cls}}^{(0)} = \mathcal{A}(x^{(1)})[\text{cls}] = \sum_i \text{softmax}\left(\frac{q_{\text{cls}}k_i^T}{\sqrt{C}}\right) v_i, \tag{8}$$

To generate a target-specific heatmap that highlights the significant regions influencing the model's prediction, Grad-ECLIP computes heatmap H_i using the following equation:

$$H_i = \text{ReLU}\left(\sum_c w_c w_i v_i\right). \tag{9}$$

w_c represents the channel importance, which is derived from the gradient of the similarity between the image-text pair with respect to the output class token:

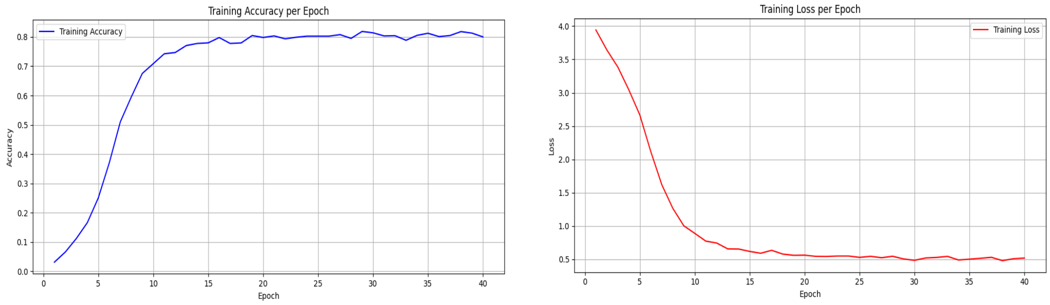
$$w_c = \frac{\partial S_T(f_I)}{\partial o_{\text{cls}}}, \tag{10}$$

w_i denotes the spatial importance, which is computed by normalizing the inner product of the query and key embeddings to the range [0, 1]:

$$w_i = \Phi(q_{\text{cls}}k_i^T). \tag{11}$$

Φ is a normalization function used to scale the importance values appropriately.

A.3 Training Accuracy and Loss for CLIP

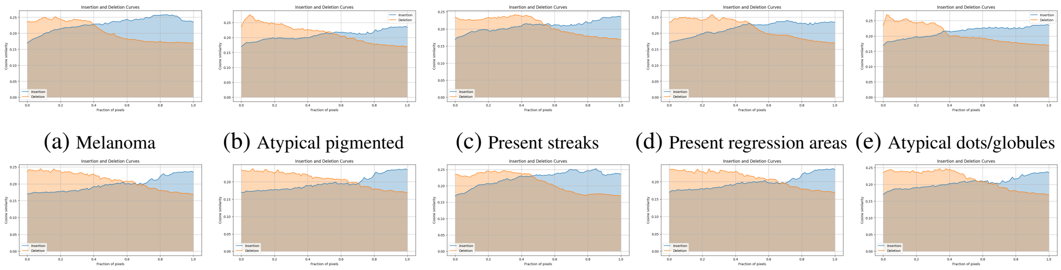


(a) Learning Curve for Training Accuracy

(b) Learning Curve for Training Loss

Figure 10: Comparison of Training Accuracy and Training Loss

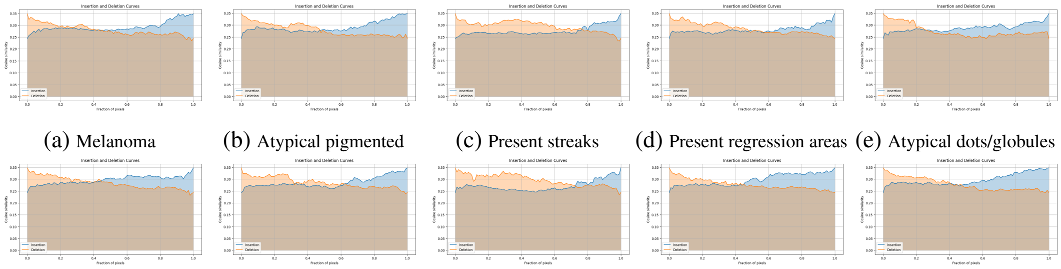
A.4 Area Under the Curve (AUC) for Insertion and Deletion



(a) Melanoma (b) Atypical pigmented (c) Present streaks (d) Present regression areas (e) Atypical dots/globules

(f) Absent blue-whitish veil (g) Fully asymmetric (h) White/dark-brown/blue-gray/black color (i) Missing vascular structures (j) Missing pigmentation

Figure 11: AUC for Deletion and Insertion curves of Fine-Tuned Grad E-CLIP Across Various Diagnostic Features (Visualization shown in 6



(a) Melanoma (b) Atypical pigmented (c) Present streaks (d) Present regression areas (e) Atypical dots/globules

(f) Absent blue-whitish veil (g) Fully asymmetric (h) White/dark-brown/blue-gray/black color (i) Missing vascular structures (j) Missing pigmentation

Figure 12: AUC for Deletion and Insertion curves for Pre-Trained Grad E-CLIP on Various Diagnostic Features (Visualization shown in Figure 7).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our claims about improving accuracy and explainability for medical imagery are clearly stated in both the abstracyt and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: There is a discussion of limitations at the end of the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: There are no theoretical results in the paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Datasets, models, and hyperparameter settings for all experiments are clearly stated.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The datasets are all publicly available, as are the pre-trained models used. If the paper is accepted we will make the code publicly available.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The section on empirical results is very clear about all of these items.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We use a single train/test split for all experiments so there is no way to compute statistical significance. Indeed, the differences between the pre-trained and fine-tuned models is so large that it seems unlikely that additional random splits would make any difference.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Those details are given in the section on empirical results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Based on our reading of the NeurIPS Code of Ethics the answer here is clearly yes.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: There are clear potential positive impacts as discussed in the introduction and abstract. There are no obvious negative impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We use public datasets and public models that are not generative.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Appropriate citations are given for all publicly available resources that were used in this paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: There are no new assets introuced in this paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper used existing datasets and did no work with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: See answer to previous question.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.