INTRIGUING PROPERTIES OF VISUAL-LANGUAGE MODEL EXPLANATIONS

Chirag Agarwal
chiragagarwall12@gmail.com

ABSTRACT

The growing popularity of large-scale visual-language models (VLMs) has led to their employment in various downstream applications as they provide a rich source of image and text representations. However, these representations are highly entangled and complex to interpret by machine learning developers and practitioners. Recent works have shown visualizations of image regions that VLMs focus on but fail to describe the change in explanations generated for visual-language classifiers in zero-shot (using image and text representations) vs. fine-tuned settings (using image representations). In this work, we perform the first empirical study to establish the trustworthy properties of explanations generated for VLMs used in zero-shot vs. fine-tune settings. We show that explanations for zero-shot visual-language classifiers are more faithful than their fine-tuned counterpart. Further, we demonstrate that VLMs tend to attribute high importance to gender, despite being non-indicative of the downstream task. Our experiments on multiple real-world datasets show interesting VLM behavior in zero-shot vs. fine-tuned settings, opening up new frontiers in understanding the trustworthiness of large-scale visual-language models.

1 INTRODUCTION

Explaining complex machine learning models remains a challenge despite several efforts to develop new explanation methods (Fong & Vedaldi, 2017; Agarwal & Nguyen, 2020; Smilkov et al., 2017), benchmark state-of-the-art explainers (Agarwal et al., 2022a), and explore their interplay with trustworthy properties like fairness and robustness (Pawelczyk et al., 2022; Agarwal et al., 2022b). Th problem is exacerbated by the arrival of large-scale visual-language models (VLMs) (Radford et al., 2021; Li et al., 2022; 2021; Kim et al., 2021; Singh et al., 2022) pre-trained with millions of image and text caption data, providing dense image and text representations to perform accurate predictions in both zero-shot or fine-tuned settings.

To this end, most works explaining image classification models focus on generating an attribution map that finds image regions that contribute more to the final classification (Smilkov et al., 2017; Fong & Vedaldi, 2017) by assigning an importance score to each pixel in the image. Recent works for explaining large-scale VLMs employ a similar technique using a fine-tuned VLM for specific image classification datasets (Chen et al., 2022). However, there is little to no work on explaining how differently VLMs behave when used as a zero-shot vs. fine-tuned image classifier. In particular, *how faithful and reliable are explanations generated for zero-shot and fine-tuned VLMs?*

Present work. In this work, we perform the first empirical exploration to understand the performance of explanations generated for zero-shot vs. fine-tuned visual-language image classifiers. The core idea is to analyze key properties of fidelity and reliability of explanations generated for VLMs in different classification settings. Our empirical results using the real-world occupation dataset shows that explanations generated using zero-shot visual-language classifiers are less biased than their fine-tuned counterpart. We also observe that the explanations generated using zero-shot VLMs are more localized and successfully leverage textual information in the image to perform classification. Further, our results establish that VLMs perform differently in zero-shot vs. fine-tuned settings, highlighting an inherent trade-off between accuracy, explainability, and fairness.

2 METHOD

Next, we define the problem of generating attribution maps and feature attribution methods for generating explanations of large-scale zero-shot vs. fine-tuned VLM classifiers.

2.1 PRELIMINARIES

We formally describe visual-language models and the problem of generating attribution maps.

Vision-Language Models. Generally, a visual-language model comprises of i) an image encoder $\mathcal{I} : \mathbb{R}^{m\times n\times 3} \rightarrow \mathbb{R}^d$ that maps a three-channel image $\mathbf{X}$ to a d-dimensional representation, and ii) a text encoder $\mathcal{T} : \mathbb{R}^t \rightarrow \mathbb{R}^d$ that maps a given set of text tokens $\mathbf{T}$ to a similar d-dimensional latent space. Recently developed VLMs (Radford et al., 2018; Li et al., 2022; Singh et al., 2022) learn image $\mathcal{I}(\mathbf{X})$ and text $\mathcal{T}(\mathbf{T})$ representations in a similar latent space using self-supervised learning. The similarity between the image and text representations is obtained using cosine similarity $\text{sim}(\mathbf{X}, \mathbf{T}) = \frac{\mathcal{I}(\mathbf{X})\cdot\mathcal{T}(\mathbf{T})^\top}{||\mathcal{I}(\mathbf{X})||\ ||\mathcal{T}(\mathbf{T})||}$. The complex representation of VLMs can be used for several downstream tasks, and we utilize them for performing image classification in this work.

Zero-shot vs. Fine-tuned Setting. In a zero-shot image classification setting, a VLM is provided with an image to be classified and a range of text inputs with the prompt "a photo of a $\{\texttt{class}\}$" where $\{\texttt{class}\}$ is substituted with all possible categories in the given classification dataset. We classify the image into the category which obtains the maximum cosine similarity (defined above). However, for a fine-tuned image classification setting, we use the image encoder $\mathcal{I}$ of the VLM followed by a new classification layer on top to produce logits with output sizes similar to the number of classes in the given image classification dataset.

Problem Formulation (Attribution Map). Let $f : (\mathbb{R}^{m\times n\times 3},\ \mathbb{R}^{d\times t}) \rightarrow \mathbb{R}$ be a visual-language model employed for image classification task in zero-shot and fine-tuned settings. We aim to generate an attribution map $\mathbf{A} \in \mathbb{R}^{m\times n}$, where each input pixel has a corresponding attribution score $\mathbf{A}_{i,j} \in [0, 1]$ indicating the contribution of the pixel to the final model prediction score.

2.2 ATTRIBUTION MAPS

We leverage the meaningful perturbation (MP) (Fong & Vedaldi, 2017) explanation algorithm to generate explanations for a zero-shot and fine-tuned visual-language model in image classification tasks. The MP method learns a minimal and continuous attribution map $\mathbf{A}$ that identifies the smallest image region which blurs the input image, leading to the minimization of the target class probability. In particular, the MP method solves the following optimization problem:

$$\mathbf{A}^* = \arg\min_{\mathbf{A}} \lambda_1 \|\mathbf{A}\|_1 + \lambda_2 \text{TV}(\mathbf{A}) + \hat{y}, \tag{1}$$

where $\text{TV}(\cdot)$ is the total-variation loss that acts as a smoothness prior over the attribution map, (λ_1, λ_2) are the regularization coefficient for the ℓ_1- and TV-loss, and $\hat{y}$ is the predicted probability for the blurred image obtained using a Gaussian blurring kernel over the input image.

Note that the first two terms are similar for both zero-shot and fine-tuned VLM settings. Next, we discuss how to compute the third term (prediction probability).

Zero-Shot Setting. In the zero-shot setting, we first convert all the dataset's classes into captions such as *"a photo of a chef"*, *"a photo of a doctor"*, *"a photo of a police"*, etc. We get the text representation of each of these captions using the pre-trained text encoder $\mathcal{T}$, *i.e.,* $\mathbf{z}_{t1}$, $\mathbf{z}_{t2}$, and $\mathbf{z}_{t3}$. Given an image of a *doctor*, we encode the image using the encoder $\mathcal{I}$ to a representation $\mathbf{z}_i$. We then calculate the similarity of these text and image representations using cosine similarity and return the similarity score of the target class as the final predicted score, *i.e.,* $\hat{y} = \text{sim}(\mathbf{z}_i, \mathbf{z}_{t2})$.

Fine-tuned Setting. For a fine-tuned VLM, we simply train an additional classification head using CrossEntropy loss and Adam optimizer (please see Appendix A.1 for more training details). For the third term in Eqn. 1, we apply the softmax operator on the output logits and utilize the softmax probability of the target class for generating attribution maps.

3 EXPERIMENTS

Next, we present results from our empirical study and address three key questions: Q1) Are explanations more faithful to zero-shot or fine-tuned visual-language classifiers? Q2) How does the interplay between vision and language affect the generated explanations? Q3) Do VLMs rely on protected attributes to achieve high predictive performance in zero-shot and fine-tuned settings?

3.1 DATASET AND EXPERIMENTAL SETUP

Datasets. We experiment with two datasets. 1) The *StanfordCars* (Krause et al., 2013) dataset contains 16,185 images of 196 car classes split into 8,144 training images and 8,041 testing images. 2) The *Profession* (Ola) dataset contains images of identifiable professionals and comprises 11,000 images that span over ten profession categories. Further, we crawled an additional set of 200 profession-based images for studying the fairness properties of explanations, where we compute the segmentation mask for the faces using MTCNN Face identification network (Zhang et al., 2016) and mask out the face from each image. In particular, we added a black mask on the face of humans from each profession so that we can determine whether the VLM focuses on the face to classify occupations. See Figure 6 in the appendix for some example images.

Performance evaluation. We use two evaluation metrics to quantify the fidelity and reliability of output explanations. For fidelity, we use the *deletion* metric (Petsiuk et al., 2018) that measures the area under the curve of the predicted class probability as we zero out input pixels of the highest attribution as identified by an output explanation, where lower deletion scores are considered more accurate. For fairness, we extend the pointing game (Zhang et al., 2018) metric to evaluate if the most salient pixels identified by an explanation lie within the annotated face mask of the image. The final fairness accuracy is computed as: Fairness Accuracy $= \frac{\text{\#Images with high attribution inside face mask}}{\text{Total number of images}}$.

Baseline Model and Explainer. To investigate the explanation behavior of VLMs in zero-shot and fine-tuned settings, we consider the CLIP (Radford et al., 2021) model and use the meaningful perturbation (Fong & Vedaldi, 2017) to generate explanations for model predictions. We set all model and explainer hyperparameters following the authors' guidelines. For a fair comparison, we tested the explanations for images correctly classified by zero-shot and fine-tuned CLIP models.

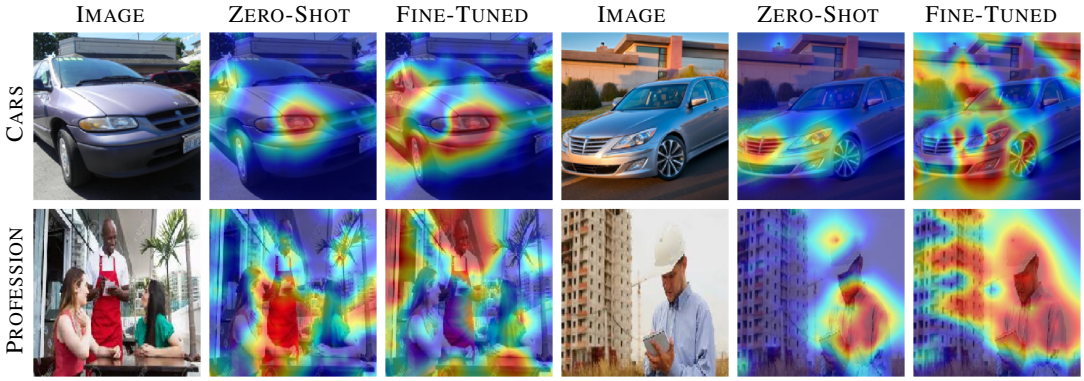

Figure 1: Attribution maps for zero-shot visual-language classifiers are more faithful (accurate localization) than their fine-tuned counterpart. In the occupation examples, we observe that explanations in the zero-shot column focus only on profession-related regions (*e.g.,* bib apron for *Chef* and helmet and notes for *Engineer*).

3.2 RESULTS

Q1) Zero-shot VLM explanations are more reliable. Across both datasets, we observe that explanations generated for zero-shot CLIP are more reliable than their fine-tuned counterparts. On average, explanations for zero-shot CLIP improve the deletion score by 18.52% and 5.44% for StanfordCars and Occupation datasets, respectively (Figure 4). Further, in Figure 1, we find that attribution maps for zero-shot CLIP are more localized to the objects in the image and identify regions that are more related to the context of the classification task, *i.e.,* car model and profession.

Figure 2: Attribution maps for zero-shot visual-language StanfordCar classifier show that CLIP utilizes the text in images to classify specific car categories by detecting relevant text like car *manufacturer* and *model*.

Q2) Zero-shot VLM explanations identify relevant text. One key difference between the zero-shot and fine-tuned CLIP settings is that zero-shot CLIP explicitly uses the representations from the text encoder to make classification, whereas, in the fine-tuned setting, we only use CLIPs' image encoder to perform classification. Thus, we hypothesize that explanations from zero-shot CLIP should identify more textual content for the StanfordCars dataset as there are several images with information about the car model and make in the text form. In Figure 2, we show that explanations generated for zero-shot CLIP identify the text in the image, like the car manufacturer name or model, better than explanations from fine-tuned CLIP. Please see Figure 8 for more such examples.

Q3) Zero-shot VLM explanations are less biased. Here, we demonstrate the fairness properties of explanations generated using zero-shot vs. fine-tuned CLIP by generating explanations for human professional images with masked faces (see Figure 6 for examples) so that an explanation identifies the importance using image context (*e.g.,* Bib Aprons for *Chef*, a lab coat and stethoscope for *Doctor*, etc.) Across different genders, we find that explanations for zero-shot and fine-tuned CLIP are more faithful and achieve 6.16% lower deletion scores when generated using face-masked images. Further, we observe that explanations generated for zero-shot CLIP do not highlight regions inside the face masks as the *most* salient pixels (Figure 3). We show that despite occluding the face, fine-tuned CLIP models attribute high face importance as if it has learned the face location as a proxy of the gender bias for the classification task. In Table 1, we calculate the fairness accuracy and find that the explanations of the zero-shot model achieve lower accuracy than their fine-tuned counterparts, *i.e.,* the number of images for which we get high attribution on the masked faces are more for explanations of fine-tuned CLIP and male professions, highlighting gender disparity.

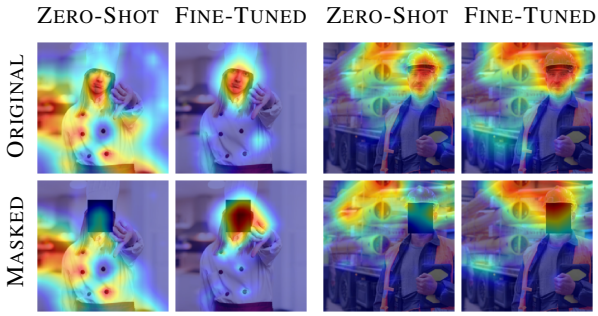

Figure 3: Masking faces show contrasting behavior between zero-shot vs. fine-tuned visual-language classifiers. Explanations for zero-shot classifiers attribute low importance to masked faces than their original and fine-tuned counterparts. See Figure 9 for more examples.

Gender	Zero-Shot	Fine-Tuned
Female	**12.50%**	20.45%
Male	**21.17%**	29.41%

Table 1: Fairness Accuracy of explanations generated for zero-shot vs. fine-tuned VLM. Explanations for zero-shot classifiers achieve lower fairness accuracy, *i.e.,* they attribute lower importance to faces.

4 CONCLUSION

In this work, we perform the first empirical study to evaluate the trustworthiness of explanations generated for zero-shot vs. fine-tuned VLMs. Our results on real-world datasets show that explanations generated for zero-shot CLIP classifiers are more faithful (lower deletion scores) and reliable (lower gender bias) than their fine-tuned counterpart. We observe that explanations for zero-shot VLMs achieve better localization and utilize multi-modal features to perform classification. Our preliminary exploration paves the way for several exciting future directions in understanding and developing trustworthy large-scale ML models.

REFERENCES

Idenprof. https://github.com/OlafenwaMoses/IdenProf. (Accessed on 03/31/2023).

Chirag Agarwal and Anh Nguyen. Explaining image classifiers by removing input features using generative models. In *ACCV*, 2020.

Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. Openxai: Towards a transparent evaluation of model explanations. *NeurIPS*, 2022a.

Chirag Agarwal, Marinka Zitnik, and Himabindu Lakkaraju. Probing gnn explainers: A rigorous theoretical and empirical analysis of gnn explanation methods. In *AISTATS*, 2022b.

Peijie Chen, Qi Li, Saad Biaz, Trung Bui, and Anh Nguyen. gscorecam: What objects is clip looking at? In *ACCV*, 2022.

Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *ICCV*, 2017.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV workshops*, 2013.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS*, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

Martin Pawelczyk, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. Exploring counterfactual explanations through the lens of adversarial examples: A theoretical and empirical analysis. In *AISTATS*, 2022.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *BMVC*, 2018.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

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 *ICML*, 2021.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *CVPR*, 2022.

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

Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *IJCV*, 2018.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 2016.

A APPENDIX

Here, we discuss the datasets, training details, and additional results from our experiments.

A.1 TRAINING DETAILS

Datasets. We experiment with two datasets. 1) The *StanfordCars* (Krause et al., 2013) dataset contains 16,185 images of 196 car classes. The dataset is split into 8,144 training images and 8,041 testing images, where each class has been split roughly in a 50-50 split. Different car classes are typically at the level of make, model, and year of a car. 2) The *Occupation* (Ola) dataset contains images of identifiable professionals collected in order to ensure that ML models can recognize professionals using image context (mode of dressing). The dataset comprises 11,000 images that span over ten profession categories, where there are 1100 images in each class, with 900 images for training and 200 images for testing.

Implementation details. For training the fine-tuned visual-language model, we add a linear layer to produce logits of size 196 (for StanfordCars dataset) and 10 (for Occupation dataset), where the logit sizes were determined using the number of classes in the respective datasets. We freeze the image and text encoder of the VLMs and fine-tune the linear layer using an Adam optimizer with a learning rate of 0.01 and CrossEntropy loss. We follow Fong & Vedaldi (2017) and optimize a coarse 28×28 mask and finally upsample the mask to the full 224×224 image using bilinear interpolation.

A.2 RESULTS

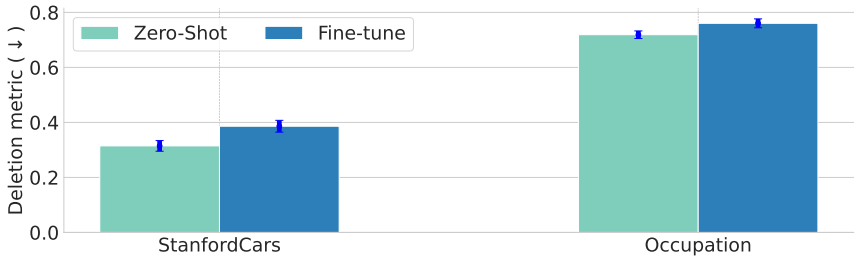

Figure 4: Deletion metric (lower is better) values for zero-shot and fine-tuned CLIP model on StanfordCars and Occupation dataset. Zero-shot CLIP classifier achieves more faithful explanations than their fine-tuned counterparts.

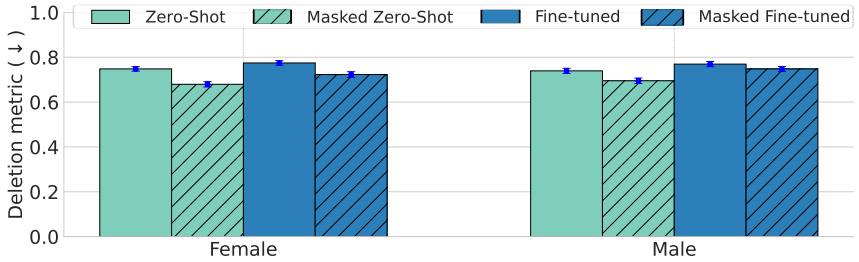

Figure 5: Deletion metric (lower is better) values for explaining zero-shot and fine-tuned CLIP predictions of 200 original and face-masked images of the Occupation dataset. Explanations generated for face-masked occupation images achieve better deletion scores than their original counterparts.

StanfordCars		Occupation	
Zero-Shot	Fine-Tuned	Zero-Shot	Fine-Tuned
58.67%	**78.26%**	92.70%	**95.80%**

Table 2: Evaluation of StanfordCars and Occupation dataset. We show the testing accuracy obtained from the zero-shot and fine-tuned CLIP model. As expected, the CLIP model fine-tuned using an additional linear layer achieves higher predictive performance than its zero-shot counterpart.

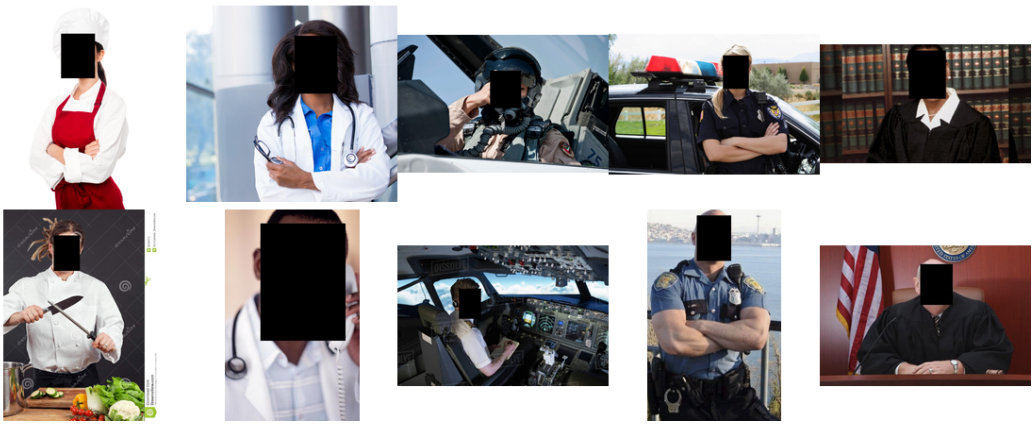

Figure 6: Random images from the masked profession dataset. We utilize these images to quantify the fairness accuracy of zero-shot and fine-tuned VLMs. Here, we mask the face by inserting a black patch on top so that machine learning models can learn to classify using profession-based features from the image and not bias on the gender of the human.

Figure 7: Attribution maps for zero-shot visual-language classifiers are more faithful (accurate localization) than their fine-tuned counterpart.

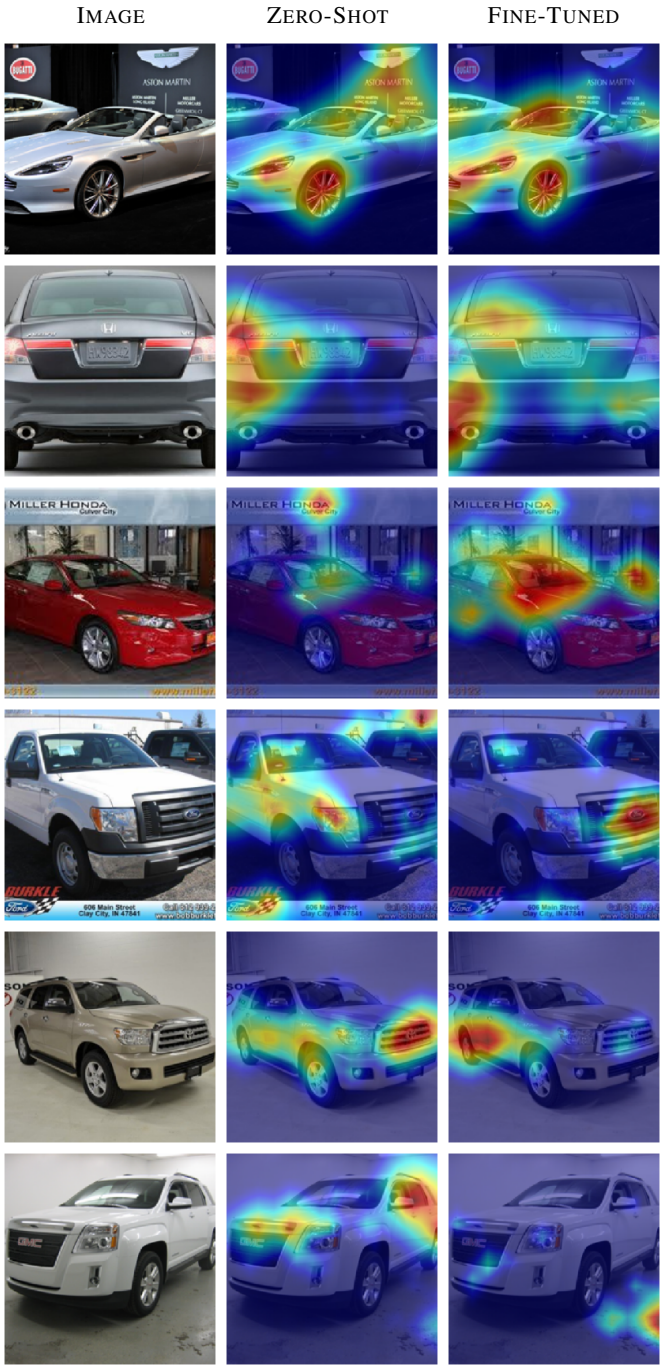

Figure 8: Attribution maps for zero-shot visual-language StanfordCar classifier show that CLIP utilizes the text in images to classify specific car categories by detecting relevant text like car *manufacturer*, *model*, or *logo*.

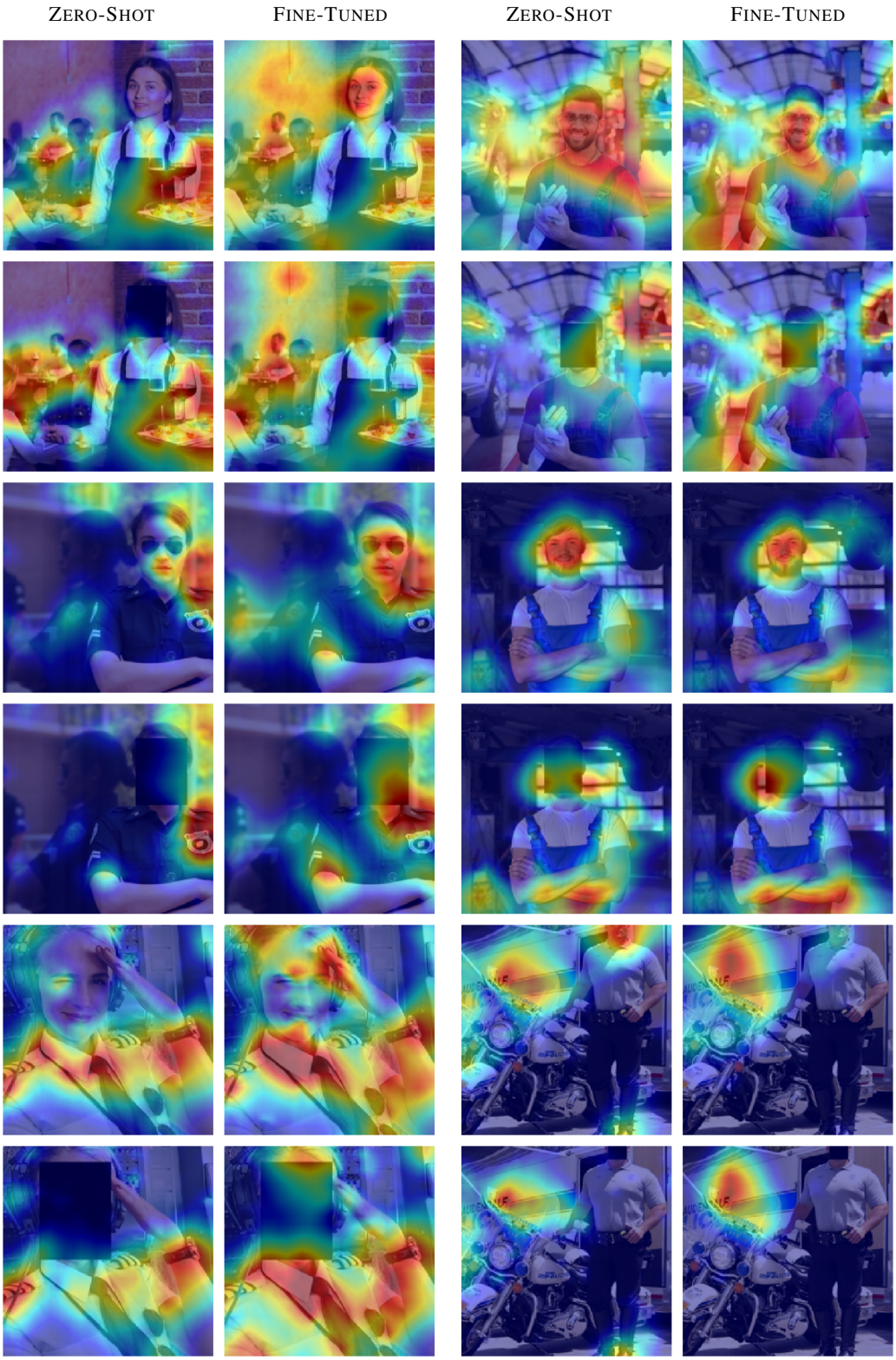

Figure 9: Masking faces show contrasting behavior between zero-shot vs. fine-tuned visual-language classifiers. Explanations for zero-shot classifiers attribute low importance to masked faces than their original and fine-tuned counterparts.