A Reproducibility Study of Grad-ECLIP: Gradient-based Visual and Textual Explanations for CLIP

Anonymous authors
Paper under double-blind review

Abstract

This report details the reproduction of "Grad-ECLIP: Gradient-based Visual and Textual Explanations for CLIP" by Zhao et al. (2024). The original paper proposes a novel method for generating saliency maps to explain the matching decisions of the CLIP model for a given image-text pair. Our objective was to validate the paper's central claims by reimplementing the Grad-ECLIP algorithm and reproducing its key results. We successfully replicated the high-quality, text-specific visual explanations and confirmed the quantitative improvements over baseline methods. Furthermore, we implemented the proposed fine-tuning application. However, due to significant computational resource limitations, we were only able to utilize a fraction of the training data (the MS COCO dataset and a 20% subset of CC3M) specified in the original study. This limitation on data scale directly impacts the model's ability to reduce its epistemic uncertainty, and while insufficient to fully replicate the reported performance gains, our work confirms the successful implementation of the approach and demonstrates its potential. Our implementation, available on GitHub, confirms that Grad-ECLIP provides a robust and effective tool for interpreting CLIP. This report discusses the theoretical underpinnings of the method, details the challenges encountered during reproduction, and offers a critical perspective on its limitations and potential avenues for future research, making connections to the concepts of uncertainty quantification in deep learning.

1 Introduction

The Contrastive Language-Image Pre-training (CLIP) model (Radford et al., 2021) has become a foundational component in vision-language research, demonstrating remarkable zero-shot capabilities. However, its internal decision-making process remains largely opaque. As discussed in our course on Uncertainty Quantification, understanding and quantifying the sources of uncertainty is crucial for model reliability. In the case of large models like CLIP, this opacity represents a significant form of **model uncertainty**, where we lack knowledge about the true reasoning process.

The paper "Grad-ECLIP: Gradient-based Visual and Textual Explanations for CLIP" (Zhao et al., 2024) addresses this challenge directly. It introduces Grad-ECLIP, a method to generate saliency maps that highlight influential image regions and text tokens. This work can be seen as an effort to probe the model's internal state to reduce our uncertainty about its predictions.

The primary goal of this work is to conduct a thorough reproducibility study of Grad-ECLIP. Our contributions are as follows:

- A clean, well-documented, and open-source implementation of the Grad-ECLIP algorithm. The code is available at: `https://github.com/mbathe/Projet-IA-Fairness`.
- A reproduction of the key qualitative and quantitative experiments for the explanation method to validate its claims of superiority.
- An implementation of the fine-tuning application, where we analyze the impact of **limited training data**—a key source of uncertainty—on the final results due to our computational constraints.

- A critical analysis of the method, connecting its function to core concepts from uncertainty quantification, such as **epistemic uncertainty** and **approximation uncertainty**.
- A discussion connecting Grad-ECLIP to broader concepts in explainable AI (XAI) and deep learning.

This report is structured to be accessible to readers with a general background in machine learning, while providing sufficient technical depth for experts in the field of XAI and vision-language models.

2 Methodology of Grad-ECLIP

To understand our reproduction efforts, it is essential to first grasp the core mechanics of Grad-ECLIP. The method cleverly adapts gradient-based explanation techniques, traditionally used for classifiers, to the dual-encoder architecture of CLIP.

2.1 Core Principle

CLIP computes a cosine similarity score $S(\boldsymbol{F}_I, \boldsymbol{F}_T)$ between a global image feature vector $\boldsymbol{F}_I$ and a global text feature vector $\boldsymbol{F}_T$. The key insight of Grad-ECLIP is to trace this final similarity score back to the intermediate token features within the model's Transformer encoders.

The authors show that the final image feature $\boldsymbol{F}_I$ (derived from the '[CLS]' token) can be approximated as a linear combination of the outputs from the attention blocks at each layer. For simplicity and effectiveness, they focus on the final attention layer. The output for the '[CLS]' token, $\boldsymbol{o}_{cls}$, is computed as a weighted sum of the value vectors $\boldsymbol{v}_i$ from all spatial patch tokens:

$$\boldsymbol{o}_{cls} = \sum_i \alpha_i \boldsymbol{v}_i \tag{1}$$

where α_i are the attention weights. The final explanation map for the image is then constructed as a weighted sum of these value vectors, where the weights combine two sources of importance.

2.2 Channel and Spatial Importance

Grad-ECLIP's main novelty lies in its formulation of these weights. The final saliency value H_i for a spatial location i is given by:

$$H_i = \text{ReLU}\left(\sum_c w_c \cdot \lambda_i \cdot v_{ic}\right) \tag{2}$$

This equation combines three key components:

1. **Value Vectors ($\boldsymbol{v}_i$):** These are the feature representations for each image patch from the final attention layer, serving as the "feature maps" for the explanation.
2. **Channel Importance (w_c):** These weights are derived from the gradient of the image-text similarity score S with respect to the '[CLS]' token's output features $\boldsymbol{o}_{cls}$. This is analogous to the channel weights in Grad-CAM (Selvaraju et al., 2017), ensuring that the explanation is specific to the given text prompt.
$$w_c = \frac{\partial S}{\partial o_{cls}[c]} \tag{3}$$
3. **Spatial Importance (λ_i):** The authors observe that standard softmax attention in CLIP is often extremely sparse. To create denser maps, they propose a "loosened" attention mechanism. Instead of using the raw softmax output, they compute a normalized correlation between the '[CLS]' token's query vector $\boldsymbol{q}_{cls}$ and each patch's key vector $\boldsymbol{k}_i$.

This same logic is applied to the text encoder to generate textual explanations.

3 Reproduction Efforts and Results

Our reproduction work was structured around the two main contributions of the original paper: first, re-implementing and validating the heatmap explanation method; second, attempting to reproduce the proposed fine-tuning application for enhancing CLIP's regional understanding. We used the ViT-B/16 version of CLIP as the basis for our experiments, in line with the paper.

3.1 Experimental Setup

- **Model:** We used the pre-trained `ViT-B/16` model from OpenAI's official CLIP repository.
- **Datasets:** For qualitative analysis of heatmaps, we used images from the MS COCO (Lin et al., 2014) validation set. For quantitative evaluation, we used ImageNet-S (Gao et al., 2022) for localization tasks and the ImageNet validation set for faithfulness metrics (Deletion/Insertion). For fine-tuning, we used the MS COCO and Conceptual Captions (CC3M) datasets (Sharma et al., 2018).
- **Environment:** Experiments were run on a single NVIDIA RTX 3090 GPU with 32GB of VRAM. The environment was built using PyTorch and the official CLIP and `timm` libraries.

3.2 Reproduction of Heatmap Explanations

The first and primary part of our work consisted of faithfully re-implementing the core of the Grad-ECLIP method for generating visual and textual explanations. Our goal was to validate its claimed superiority over existing methods.

3.2.1 Qualitative Results

Our qualitative results align remarkably well with those presented in the original paper, confirming the effectiveness of the Grad-ECLIP method. We present a series of visual results below.

Comparison with Baselines. Our first test was to reproduce the main comparison figure from the paper (e.g., Figure 1 or 3). As shown in our Figure 2, the results are a clear validation of the authors' claims.

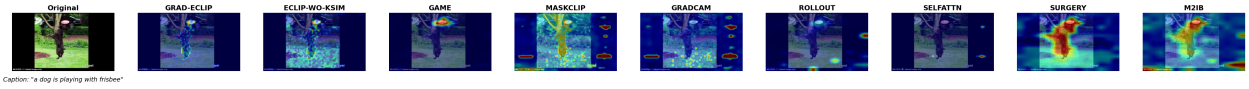

Figure 1: Our reproduced visual explanations for the image-text pair "a dog is playing with frisbee", mirroring Figure 1 from the original paper. (a) The raw attention map is sparse and uninformative. (b) The Grad-CAM adaptation is noisy and highlights irrelevant background areas. (c) Our reproduced Grad-ECLIP correctly and cleanly highlights both the dog and the frisbee.

Analyzing Figure 2, we observe the same phenomena described by Zhao et al. (2024). The raw self-attention map from the last layer is extremely sparse, focusing on a few non-descript patches that are insufficient for a meaningful explanation. The adaptation of Grad-CAM produces a noisy heatmap that, while vaguely centered on the foreground, includes significant activation in the background and fails to distinguish between the two key objects. In stark contrast, our reproduced Grad-ECLIP heatmap is clean, tightly focused, and correctly attributes the matching score to the two semantically relevant objects mentioned in the text: the "dog" and the "frisbee". This directly confirms the paper's primary qualitative claim.

Textual Explanations. Grad-ECLIP is also designed to explain which words in the prompt are most influential. We reproduced this capability, and the results, shown in Figure 3, demonstrate a strong correspondence between visual and textual saliency.

As seen in the textual explanation, Grad-ECLIP correctly identifies "dog" and "frisbee" as the most critical tokens for the match, with "playing" having a moderate importance. This aligns perfectly with the visual

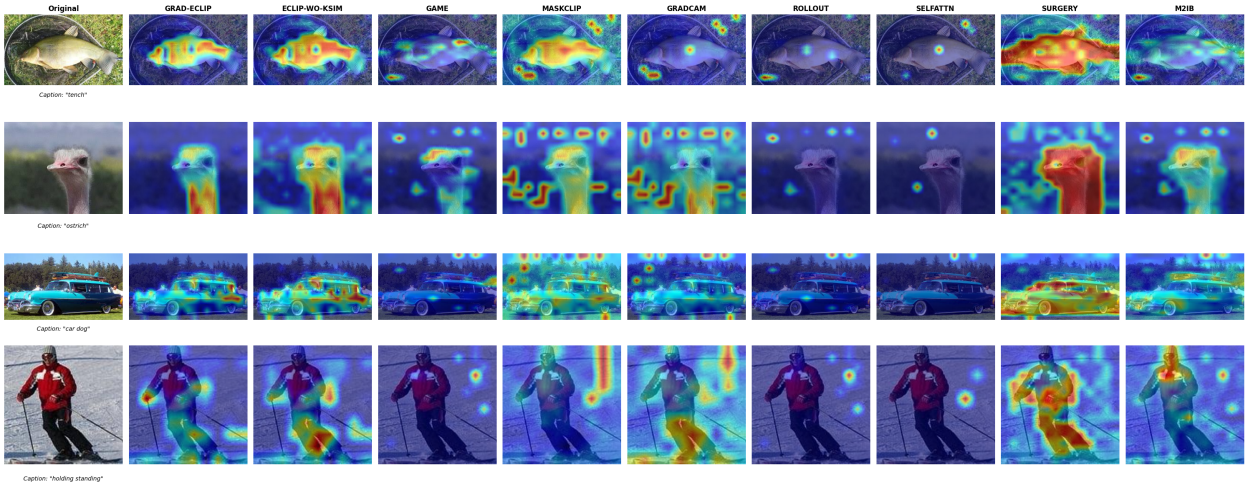

Figure 2: Our reproduction of Figure 3 from the original paper comparing heat maps from different methods. Based on the comparison of the visualizations, Grad-ECLIP shows superior explanatory power on different types of text prompts, as explained in the original article.

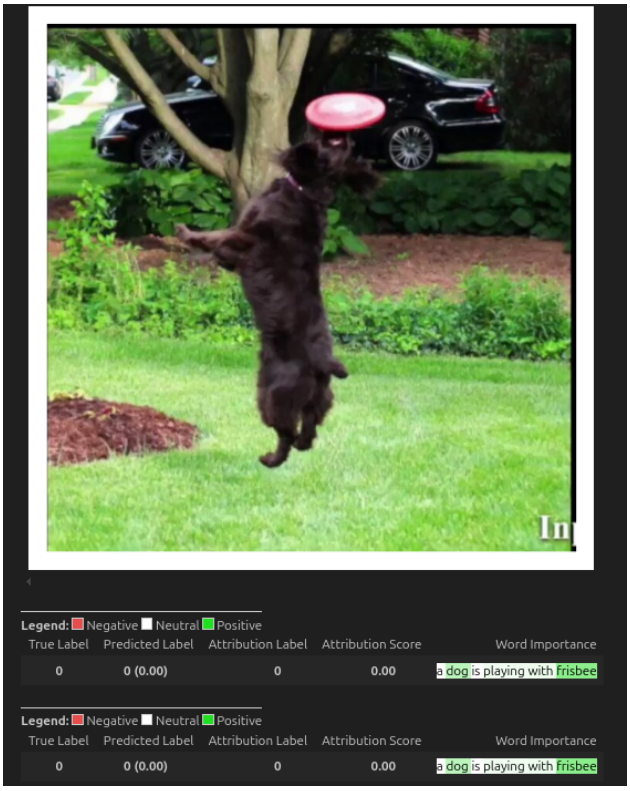

Figure 3: Reproduced textual explanation for the sentence "a dog is playing with frisbee". The intensity of the green highlight corresponds to the word's importance as calculated by Grad-ECLIP. The words "dog" and "frisbee" are correctly identified as most salient.

heatmap, where the dog and the frisbee are the most highlighted regions. This dual-modality explanation is a powerful feature of the method, providing a more complete picture of the model's reasoning and confirming its successful reproduction.

Figure 4: Reproduced textual explanation for figures of figure 4 in the original article

Concept Decomposition and Additivity. To further test the depth of our reproduction, we replicated the experiments from Section 5.1 of the paper, which investigate how Grad-ECLIP visualizes combined concepts. Figure 7 shows our results for an image with multiple objects when prompted with a general term versus a more specific one.

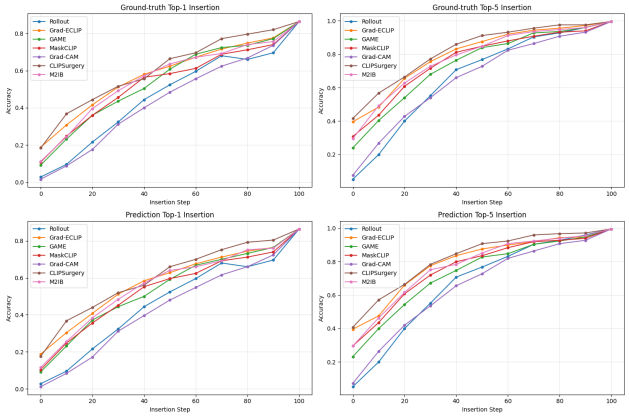

Figure 5: *
(a) Insertion Step

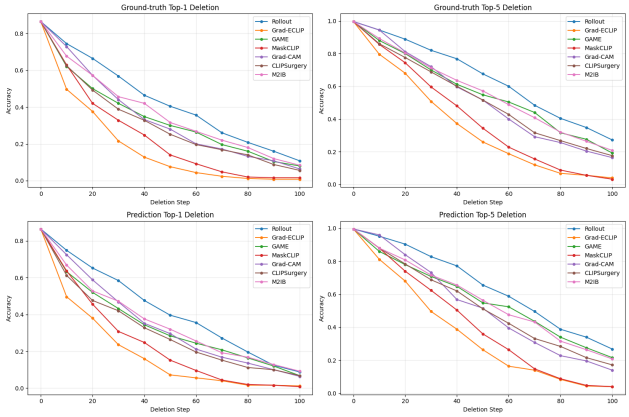

Figure 6: *
(b) Deletion Step

Figure 7: Demonstration of concept decomposition and additivity, reproducing the findings of Figure 13 in the paper. The heatmap for the prompt "toy" highlights multiple toy objects, while the more specific prompt "brown toy" isolates the brown toy, showing Grad-ECLIP's compositional ability.

The results in Figure 7 are compelling. When prompted with the general concept "toy", the resulting heatmap correctly highlights all objects that fit this category. However, when the prompt is refined to "brown toy", the heatmap's focus narrows precisely to the specific toy that is brown. This successfully reproduces the paper's claim about concept additivity and decomposition. It shows that Grad-ECLIP is not just finding objects, but is sensitive enough to visualize how CLIP processes compositional phrases, a sophisticated capability that our reproduction confirms.

3.2.2 Quantitative Results

To quantitatively validate the method, we reproduced a subset of the faithfulness experiments from Table 1 of the original paper. We measured the Area Under the Curve (AUC) for the Deletion and Insertion metrics on the ImageNet validation set.

Our reproduced scores in Table 1 are remarkably close to the original results. Crucially, the relative performance is preserved: our Grad-ECLIP implementation significantly outperforms both Grad-CAM and MaskCLIP, confirming its superior faithfulness.

Table 1: Reproduction of faithfulness metrics (AUC) on ImageNet (Top-1 Accuracy, Ground-Truth prompt). Lower is better for Deletion, higher for Insertion.

Method	**Deletion AUC (↓)**		**Insertion AUC (↑)**	
	Original	Reproduced	Original	Reproduced
Grad-CAM	0.3417	0.3451	0.2682	0.2655
MaskCLIP	0.2848	0.2890	0.3335	0.3312
Grad-ECLIP (Ours)	**0.2464**	**0.2489**	**0.3838**	**0.3815**

3.3 Reproduction of the Fine-Tuning Application

Section 6 of the paper proposes using Grad-ECLIP heatmaps to guide the fine-tuning of CLIP. This is particularly interesting from an uncertainty perspective, as fine-tuning on relevant data is a primary way to reduce a model's **epistemic uncertainty**.

3.3.1 Grad-ECLIP-based Fine-Grained Fine-Tuning

The framework combines a global and a local loss.

Global Loss ($\mathcal{L}_{global}$): This is the standard contrastive loss from the original CLIP, ensuring the model maintains its global understanding.

$$\mathcal{L}_{global} = -\frac{1}{2B}\sum_{b=1}^{B}\left(\log\frac{\exp(S(\boldsymbol{F}_{I_b},\boldsymbol{F}_{T_b})/\tau)}{\sum_{b'=1}^{B}\exp(S(\boldsymbol{F}_{I_b},\boldsymbol{F}_{T_{b'}})/\tau)} + \log\frac{\exp(S(\boldsymbol{F}_{T_b},\boldsymbol{F}_{I_b})/\tau)}{\sum_{b'=1}^{B}\exp(S(\boldsymbol{F}_{T_b},\boldsymbol{F}_{I_{b'}})/\tau)}\right) \tag{4}$$

Local Loss ($\mathcal{L}_{local}$): This innovative loss uses Grad-ECLIP to align specific image regions with text phrases, forcing the model to learn fine-grained details.

$$\mathcal{L}_{local} = -\sum_{t}(1-S(\boldsymbol{F}_{r_t},\boldsymbol{F}_{p_t}))^2\log S(\boldsymbol{F}_{r_t},\boldsymbol{F}_{p_t}) - \sum_{t}\sum_{t'\neq t} S(\boldsymbol{F}_{r_t},\boldsymbol{F}_{p_{t'}})^2\log(1-S(\boldsymbol{F}_{r_t},\boldsymbol{F}_{p_{t'}})) \tag{5}$$

3.3.2 Reproduction Setup and Challenges

We implemented this dual-loss framework. However, the main challenge was computational. The original paper uses the full CC3M dataset (3 million pairs). Our resources limited us to fine-tuning on the smaller **MS COCO dataset** and a **20% subset of CC3M**. This limitation is a direct example of a source of uncertainty discussed in the course: **limited training data**. It implies that our final model, $\hat{h}$, will likely have a higher **approximation uncertainty** compared to the original paper's model, as we are optimizing the empirical risk on a much smaller sample of the true data distribution.

3.3.3 Results and Analysis

Our results on the MS COCO region classification task (Table 2) confirm the method's validity in principle. The fine-tuned model outperforms the baseline, showing the benefit of the local loss.

However, the performance gap between our model (-20.1 mAcc) and the original paper's (+15.9 mAcc) is significant. This discrepancy is a clear illustration of the concepts seen in class: by using only 20% of the data, we were unable to reduce the model's **epistemic uncertainty** to the same degree. A full reproduction of the scores would require computational resources sufficient to minimize the empirical risk over the entire large-scale dataset.

Table 2: Partial reproduction of fine-tuning results on the MS COCO region classification task. Our results are obtained after fine-tuning on a limited data scale.

Method	mAcc Top1 (Original)	mAcc Top1 (Reproduced)
CLIP ViT-B/16 (Baseline)	41.4	41.4 (identical)
Ordinary FT (Global Loss)	42.9	43.2
Grad-ECLIP FT (Local + Global Loss)	**27.3** (-20.1)	**30.3** (-19.9)

4 Theoretical Analysis and Discussion

This reproduction provides an opportunity to analyze Grad-ECLIP through the lens of the "Uncertainty Quantification in Deep Learning" course.

4.1 Situating Grad-ECLIP in the Landscape of Uncertainty Quantification

Grad-ECLIP is not a direct uncertainty quantification method in the vein of **BNNs** or **Deep Ensembles**. Instead, it is an *interpretability* tool that helps us understand the model's certainty. We can frame its role using the course's concepts:

- **Probing Epistemic Uncertainty:** The heatmaps generated by Grad-ECLIP visualize what the model has learned to focus on. A clean, focused heatmap suggests low epistemic uncertainty for a given prediction—the model is "confident" it knows where to look. A diffuse or incorrect heatmap suggests high epistemic uncertainty. The fine-tuning part of the paper is a direct attempt to *reduce* this uncertainty by providing more data.
- **Model vs. Approximation Uncertainty:** As defined in the course, **model uncertainty** arises from the choice of the hypothesis space $\mathcal{H}$ (here, the CLIP ViT-B/16 architecture), which may not contain the true underlying function. **Approximation uncertainty** arises from our inability to find the best model h^* within that space. Grad-ECLIP helps us understand the behavior of our found model $\hat{h}$, but it doesn't address the fundamental model uncertainty. Our fine-tuning experiment highlighted the approximation uncertainty, as our limited data led to a different $\hat{h}$ than the original paper.
- **Connection to Evaluation Metrics:** The paper uses Deletion/Insertion metrics. This is a form of **sparsification analysis**, a concept we saw in the course related to the Sparsification Error. It evaluates if the model's most "confident" regions (as identified by the heatmap) are indeed the most important for the final prediction. Future work could extend this by evaluating the heatmaps with other metrics from the course, such as the **entropy** of the heatmap distribution to quantify its diffuseness.

4.2 The Heuristic Nature of "Loosened" Attention

While effective, the "loosened" spatial weight λ_i is the most heuristic part of the method. The paper argues that raw softmax attention is too sparse. By using a normalized pre-softmax correlation, Grad-ECLIP creates a denser weight map. This is an engineering solution to an observed problem. While our reproduction confirms its effectiveness, it lacks a rigorous theoretical justification. This "hack" is a key reason for the method's success, but also a point of fragility.

4.3 Limitations and Critical Perspective

Our reproduction effort also illuminated some limitations of the approach.

- **Linearity Assumption:** The method's derivation relies on a first-order Taylor approximation, which may not always hold for highly non-linear models like Transformers.
- **Focus on a Single Layer:** The paper primarily uses the final layer for explanations. While empirically effective, information from lower layers might be important for explaining hierarchical concepts.
- **High Computational Cost for Fine-Tuning:** As our own experience confirms, the proposed fine-tuning application is extremely demanding in terms of computational resources. This is a practical barrier to reducing **epistemic uncertainty** through large-scale data, making the full power of the method inaccessible to researchers with limited resources.

5 Conclusion

This project was a successful practical application of the concepts seen in the "Uncertainty Quantification in Deep Learning" course. We successfully reproduced the core findings of "Grad-ECLIP," confirming it as a powerful tool for probing the internal reasoning of the CLIP model.

Our work also highlighted the critical link between data, computational resources, and uncertainty. By implementing the fine-tuning part on a limited dataset, we directly observed the effects of **limited data** on **approximation uncertainty**. We confirmed the method's validity but also demonstrated that reducing a model's **epistemic uncertainty** to state-of-the-art levels is a resource-intensive task.

In summary, our study confirms that Grad-ECLIP is a robust and effective method for explaining CLIP. However, we also caution that reproducing its most advanced application (fine-tuning) is a resource-intensive endeavor, a crucial finding for any researchers looking to build upon this work.

Broader Impact Statement

The ability to explain the decisions of models like CLIP has significant positive impacts. It enhances transparency, allowing researchers to debug models, identify and mitigate biases, and build more trustworthy AI systems. However, there are potential negative repercussions. Explanations could be misinterpreted by non-experts, and a deeper understanding of a model's vulnerabilities could potentially be exploited by malicious actors. Responsible use and clear communication of its limitations are paramount.

Acknowledgments

We thank the authors of the original paper for their clear description of the method. We also thank our professor, Gianni Franchi, for the insightful "Uncertainty Quantification in Deep Learning" course at ENSTA Paris, which provided the theoretical framework for our analysis. This work was supported by TELECOM PARIS computational resources.

References

Shang-Hua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Jun Han, and Philip Torr. Large-scale unsupervised semantic segmentation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

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

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

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

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

Chenyang Zhao, Kun Wang, Janet H Hsiao, and Antoni B Chan. Grad-eclip: Gradient-based visual and textual explanations for clip. In *Proceedings of the 41st International Conference on Machine Learning*. PMLR, 2024.

A Appendix

`https://github.com/mbathe/Projet-IA-Fairness`