What does CLIP see per token?*

Wajahat Ali Khan
Kyung Hee University
South Korea
wajahat@khu.ac.kr

Seungkyu Lee
Kyung Hee University
South Korea
seungkyu@khu.ac.kr

Abstract

*While Vision Language Models (VLMs) like Contrastive Language Image Pre training (CLIP) demonstrate remarkable zero-shot generalization, their internal decision-making processes remain highly opaque. Current post-hoc interpretability methods typically provide holistic explanations by correlating the entire image with the full text prompt. This approach tends to highlight only the most dominant object, failing to answer a fundamental question: what exactly does CLIP see for each individual word? To bridge this gap, we propose Token CLIP (**TCLIP**), a training free, gradient free interpretability method that establishes a direct, dense relevance mapping between individual text tokens and image patches. By intercepting intermediate representations and leveraging multi-scale translation, our approach suppresses background noise to yield sharp, focused spatial heatmaps for any specific text token. Extensive quantitative and qualitative experiments demonstrate that TCLIP achieves SOTA scores on standard benchmarks. TCLIP is orders of magnitude more computationally efficient than previous baselines and offers a deeper, transparent understanding of CLIP's conceptual grounding.*

1. Introduction

Vision-Language Models (VLMs) like CLIP [17] demonstrate remarkable zero-shot performance by aligning images and text in a shared latent space. However, this alignment relies on a single, global cosine similarity score computed from the final `[CLS]` and `[EOS]` tokens [8]. This holistic metric reduces the model to a black box, concealing a fundamental question: *When given a complex prompt, what exactly does CLIP see for each individual word?* Prior research shows CLIP struggles with compositionality and spatial reasoning [10, 11], often operating less like a true language system and more like a sophisticated "bag-of-words" [21].

Uncovering this token-level visual grounding is challenging. Classic interpretability methods adapted for VLMs (e.g., GradCAM [18], gScoreCAM [15], HilaCAM [3]) typically produce a single, aggregated heatmap for the entire text prompt, failing to spatially differentiate the visual grounding of an adjective from its corresponding noun. Conversely, recent attempts to extract finer-grained alignments, such as CLIPSurgery [12] or Grad-ECLIP [23], require complex architectural modifications or memory-intensive backpropagation, limiting their practical utility.

To answer the question of what CLIP sees per token without the overhead of gradients or retraining, we propose **Token CLIP (TCLIP)**. Instead of relying on global summary tokens, TCLIP intercepts the intermediate hidden states to compute a fine-grained relevance matrix directly between individual text tokens and image patches. By leveraging a multi-scale translation technique, we generate high-fidelity spatial heatmaps for any specific word in a prompt, requiring zero backward passes and remaining computationally efficient.

The main contributions of this work are as follows:

1. We introduce TCLIP, a novel, gradient free method that establishes a direct, dense relevance mapping between individual text tokens and image patches.
2. We generate high fidelity spatial heatmaps that successfully isolate the specific visual grounding of individual words within a complex prompt.
3. We achieve state-of-the-art scores on zero shot object detection, localization and segmentation tests on standard benchmarks providing transparent, token-level insights into CLIP's conceptual grounding capabilities.

2. Methodology

In a ViT-based CLIP model [5], an input image is divided into N_{patches} patches, yielding intermediate hidden states $\mathbf{H}_{\text{img}} \in \mathbb{R}^{(N_{\text{patches}}) \times D_i}$ at the final transformer layer. Concurrently, the tokenized text prompt yields hidden states $\mathbf{H}_{\text{text}} \in \mathbb{R}^{L \times D_t}$. To ensure our local embeddings inhabit the identical D-dimensional multimodal space optimized by CLIP, we project the full hidden state sequences using

*Accepted at The 5th Explainable AI for Computer Vision (XAI4CV) Workshop at CVPR 2026.

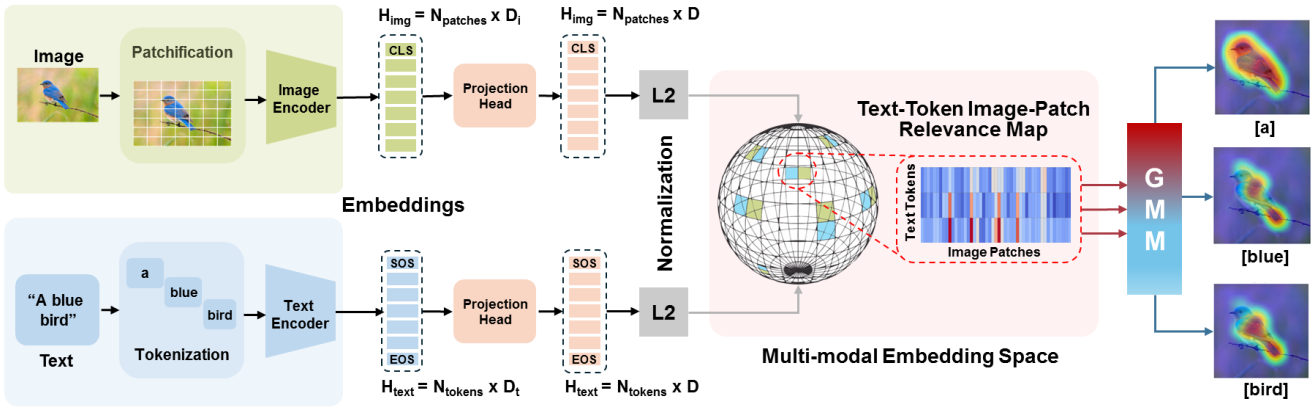

Figure 1. TCLIP works on the latent space to enable dense, fine-grained text-token to image-patch analysis. By intercepting the full sequences of patch and token hidden states and mapping them into the shared multi-modal embedding space, we compute a direct relevance matrix. Gaussian Mixture Model (GMM) thresholding suppresses background noise, yielding high fidelity spatial heatmaps for any individual word without requiring gradients or retraining.

the model's pre-trained linear projection heads ($\mathbf{W}_{\text{img}}$ and $\mathbf{W}_{\text{text}}$):

$$\mathbf{I}' = \mathbf{H}_{\text{img}} \cdot \mathbf{W}_{\text{img}} \quad , \quad \mathbf{T}' = \mathbf{H}_{\text{text}} \cdot \mathbf{W}_{\text{text}} \tag{1}$$

This yields the complete projected image tensor $\mathbf{I}'$ and text tensor $\mathbf{T}'$. While standard CLIP computes similarity exclusively between the global `[CLS]` and `[EOS]` representations, TCLIP intercepts these entire tensors to formulate a dense, fine-grained relevance map. To compute this map, we filter $\mathbf{I}'$ and $\mathbf{T}'$ to isolate content-bearing embeddings by discarding the `[CLS]`, `[SOS]`, `[EOS]`, and padding tokens. This yields the final patch matrix $\mathbf{I} \in \mathbb{R}^{N_{\text{patches}} \times D}$ and token matrix $\mathbf{T} \in \mathbb{R}^{N_{\text{tokens}} \times D}$. Following CLIP's objective, we L2-normalize all row vectors to produce $\hat{\mathbf{I}}$ and $\hat{\mathbf{T}}$. The dense relevance matrix $\mathbf{S}$ is generated via a single matrix multiplication:

$$\mathbf{S} = \hat{\mathbf{T}} \cdot \hat{\mathbf{I}}^T \tag{2}$$

Each element $\mathbf{S}_{ij}$ quantifies the relevance between the i-th text token and the j-th image patch, decomposing the global CLIP score into a transparent grid of visual-linguistic alignments. To visualize this alignment, the vector $\mathbf{S}_{i,:}$ for a token i is reshaped into its spatial grid and upsampled. To overcome the coarse ViT patch resolution and suppress background noise, we employ a multi-scale translation aggregation technique. We generate K (25) augmented images by applying translations at multiple pixel scales across eight directions. A relevance map is computed for each, geometrically realigned, and aggregated using a pixel-wise trimmed mean.

To distinguish salient foreground activations, we fit a two-component Gaussian Mixture Model (GMM) to the aggregated similarity scores $\{\mathbf{S}_{t,p}\}_{p=1}^{N_{\text{patches}}}$. The intersection τ of the two learned probability density functions serves as a data-driven threshold to create a focused relevance map $\mathbf{M}_t$:

$$\mathbf{M}_{t,p} = \begin{cases} \mathbf{S}_{t,p} & \text{if } \mathbf{S}_{t,p} \geq \tau \\ \min(\mathbf{S}_{t,:}) & \text{if } \mathbf{S}_{t,p} < \tau \end{cases} \tag{3}$$

Suppressing values to $\min(\mathbf{S}_{t,:})$ pushes background patches to the baseline minimum rather than zero. Finally, a Gaussian filter is applied to reduce minor blocky artifacts. Note that this GMM thresholding is utilized strictly for generating these continuous soft heatmaps, whereas Otsu's method is applied exclusively for generating discrete hard bounding boxes during our zero-shot evaluation.

3. Experimental Evaluation

To validate TCLIP, we conduct evaluations using the pre-trained `ViT-B/32` model to demonstrate its fine-grained interpretability. We perform our analyses on standard benchmarks, including the MS COCO 2017 [13] and ImageNet-S [7] validation sets, which provide challenging samples across diverse semantic, primitive, and abstract concepts.

3.1. Qualitative Evaluation

Comparison with Interpretability Methods: TCLIP is compared with prior interpretability methods like GradCAM [18], gScoreCAM [15], and HilaCAM [3]. As demonstrated in Figure 2, TCLIP consistently produces more precise heatmaps. For prompts like "cat" and "laptop", TCLIP's activation is tightly focused on the object of interest, while baselines produce diffuse or misaligned results.

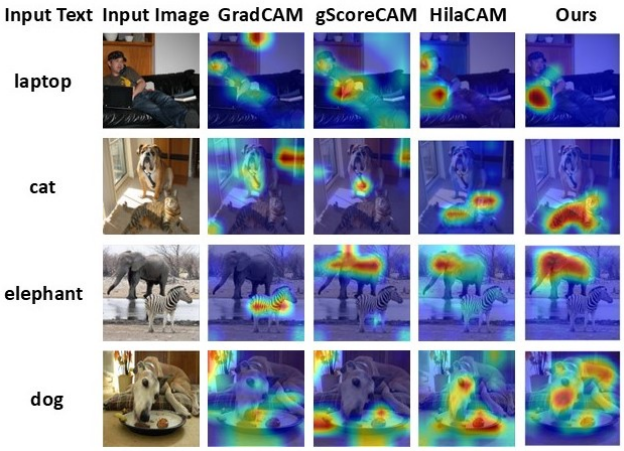

Figure 2. Comparison of TCLIP with established interpretability methods: GradCam [18], gScoreCAM[15], and HilaCAM[3].

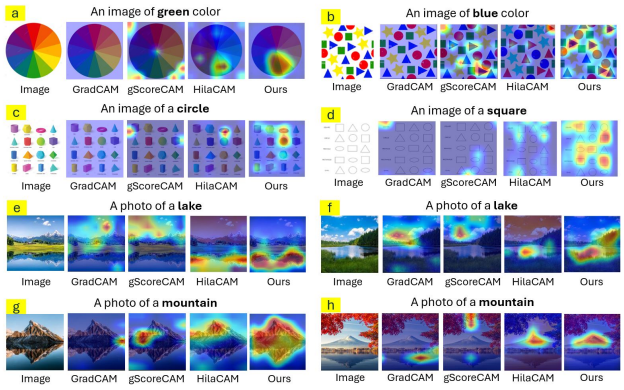

Figure 3. TCLIP accurately grounds primitive concepts like shape and color and demonstrates semantic consistency, localizing "mountain" and "lake" across diverse images, often with better object coverage than other baselines.

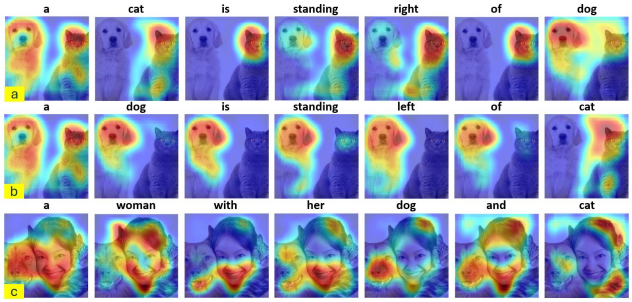

Figure 4. This figure demonstrates TCLIP's ability to deconstruct complex prompts showing Spatial Relations, Order Invariance and Human Object Interaction in CLIP.

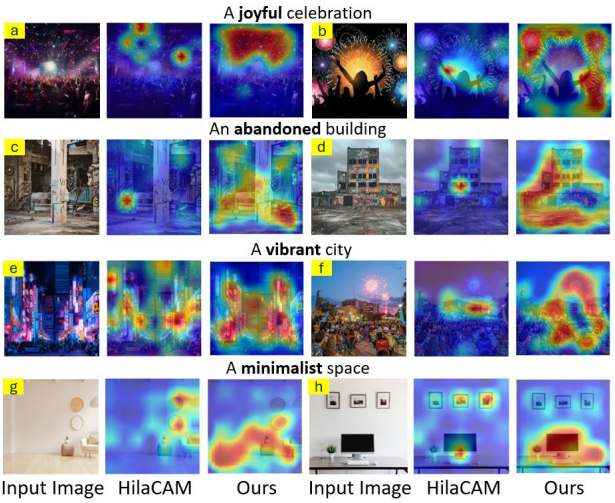

Figure 5. This figure shows a key novelty of TCLIP, that is we can deconstruct even the abstract concepts in a text prompt.

Grounding of Fundamental Visual Attributes: TCLIP demonstrates a strong capacity for grounding primitive concepts beyond high-level objects. As shown in Figure 3 (c,d), the heatmaps precisely localize "circular" and "square" shapes while ignoring other geometric forms. Similarly, color tokens like "green" and "blue" (Figure 3 a,b) show unambiguous correspondence, avoiding the incorrect activations common in prior methods. TCLIP also maintains semantic consistency, accurately localizing "mountain" or "lake" across diverse images despite significant variations in scale or viewing angle (Figure 3 e,f,g,h).

Disentangling Compositional Prompts: A key strength of TCLIP is deconstructing multi-entity prompts. For spatial relations like (Figure 4 a,b), TCLIP correctly localizes the noun tokens to their respective entities while producing distinct heatmaps for spatial tokens like "left" and "right". Reversing the prompt preserves the correct noun-entity localization, demonstrating order invariance. Furthermore, it disambiguates human-object interactions; for "a woman with her dog and cat" (Figure 4 c), the model isolates all three entities. The abstract relational token "with" generates a focused heatmap bridging the woman and her pets.

Visualizing Abstract and Subjective Concepts: TCLIP successfully interprets abstract concepts lacking concrete visual referents. For "a minimalist space" (Figure 5 g,h), TCLIP highlights regions of negative space, indicating CLIP associates minimalism with visual sparsity, whereas HilaCAM primarily activates salient objects. For concepts like "joyful" or "vibrant" (Figure 5 a,b,e,f), TCLIP identifies correlated visual elements (e.g., fireworks, city lights). This demonstrates that TCLIP extends interpretability beyond object localization, offering a unique window into CLIP's internal representation of abstract ideas.

Table 1. Zero-Shot Object Detection on MS COCO. Accuracy is Box Accuracy (BoxAcc) at IoU $>$ 0.5. FP/BP = Forward/Backward passes per heatmap.

Method	BoxAcc (↑)	Backbone	FP	BP
GradCAM [18]	11.59	RN50x16	1	1
xGradCAM [6]	5.60	RN50x16	1	1
GradCAM++ [2]	9.68	RN50x16	1	1
LayerCAM [9]	9.19	RN50x16	1	1
RISE [16]	7.26	RN50x16	8001	0
GroupCAM [22]	13.06	RN50x16	96	1
scoreCAM [19]	20.43	RN50x16	3073	0
gScoreCAM [15]	12.73	ViT-B/32	301	1
HilaCAM [3]	12.82	ViT-B/32	1	1
TCLIP [Ours]	**19.07**	ViT-B/32	**25**	**0**

3.2. Quantitative Evaluation

While TCLIP is primarily designed to reveal the specific visual grounding of individual text tokens, we must evaluate it against prior interpretability baselines to rigorously validate its spatial accuracy. To achieve this, we conduct comprehensive quantitative evaluations across two distinct downstream tasks. First, we perform a zero-shot object detection test on MS COCO, measuring how well our generated token-level heatmaps align with ground-truth bounding boxes via Intersection over Union (IoU). Second, we evaluate fine-grained localization and pixel-wise semantic segmentation on the ImageNet-S dataset, utilizing Pointing Game metrics and maskIoU to demonstrate TCLIP's precise spatial alignment and superior shape adherence.

Zero Shot Object Detection on MS COCO: We evaluate Zero Shot Object Detection on the MS COCO validation set (5,000 images, 80 categories). Crucially, the majority of COCO categories consist of a single word, which perfectly aligns with our token-level methodology. For the few categories that span multiple tokens, their corresponding heatmaps are simply averaged. To convert our continuous heatmaps into discrete bounding boxes, we employ a multi-step protocol: (1) we aggregate the 25 translated heatmap instances via a trimmed mean, (2) we binarize the final heatmap using Otsu's method [14] to create a foreground mask, and (3) we apply connected-component analysis to extract a tight, axis-aligned bounding box enclosing the largest component.

As detailed in Table 1, TCLIP achieves a state-of-the-art Box Accuracy among ViT-based interpretability methods, significantly outperforming HilaCAM and gScoreCAM. Notably, TCLIP is highly competitive with the top-performing CNN-based method ScoreCAM while being orders of magnitude more computationally efficient. Generating a TCLIP heatmap requires only one forward pass per image translation (totaling 25 forward passes) and zero memory intensive backward passes.

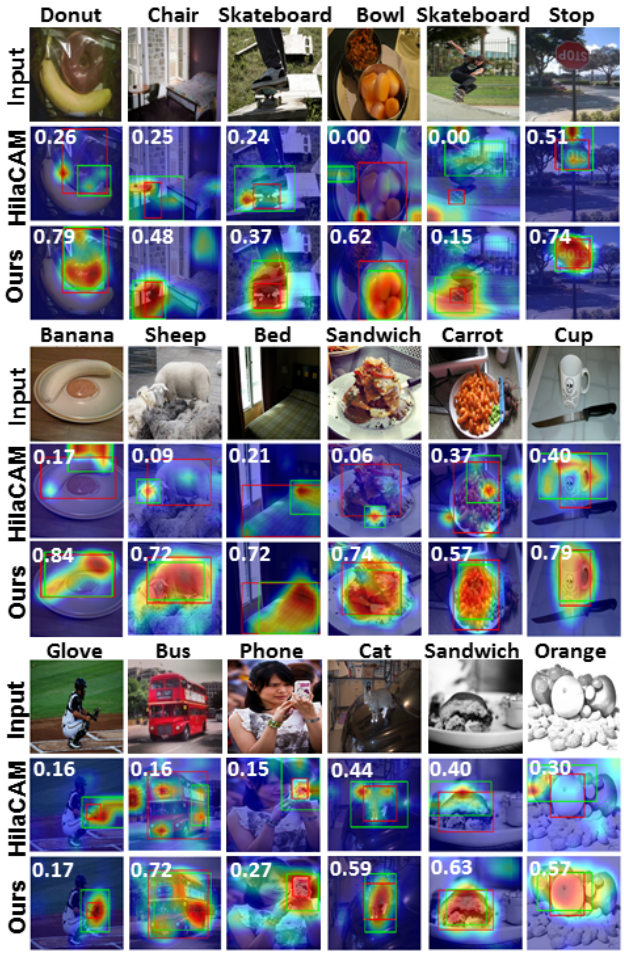

Figure 6. Comparison of bounding box results: Red Box is Ground-truth, Green Box is Predicted. IoU is shown on top-left of every image. All samples are from the COCO validation dataset.

These strong quantitative findings are mirrored in our qualitative bounding box visualizations (Figure 6). TCLIP's heatmaps remain visibly more focused, leading to superior IoU scores and tighter boundary adherence compared to baseline methods.

Localization and Segmentation on ImageNet-S: We do further quantitative analysis by comparing TCLIP with state-of-the-art CLIP interpretability methods on the challenging tasks of fine-grained localization and segmentation on the ImageNet-S validation dataset [7].

For the localization test, we employ the Pointing Game (PG) metric (which counts a hit if the maximum heatmap intensity falls within the ground-truth mask, otherwise a miss) and energy-PG (which measures the overall heatmap energy concentrated within the mask). As shown in Table 2, TCLIP achieves a new state-of-the-art Pointing Game score

Table 2. Comparison of different methods on localization test using Pointing Game and Energy Pointing Game Evaluation Metric on the ImageNet-S validation dataset.

Method	Pointing Game (↑)	energy-PG (↑)
Raw Attention	0.121	0.132
Rollout [1]	0.137	0.283
GradCAM [18]	0.184	0.315
M2IB [20]	0.264	0.355
MaskCLIP [4]	0.404	0.140
CLIPSurgery [12]	0.575	0.398
HilaCAM [3]	0.470	0.443
TCLIP [Ours]	**0.775**	**0.445**

Table 3. Comparison of different methods on Segmentation test using Pixel Accuracy, Average Precision and mask IoU on the ImageNet-S validation dataset.

Method	Pixel Acc.	Avg. Precision	maskIoU
Raw Attention	0.0278	0.2877	0.0013
Rollout [1]	0.2524	0.3345	0.0110
GradCAM [18]	0.5457	0.4050	0.1251
GradECLIP [23]	0.7056	0.5662	0.2869
MaskCLIP [4]	0.7180	0.4557	0.2481
CLIPSurgery [12]	**0.7546**	0.4608	0.3471
M2IB [20]	0.6194	0.4003	0.1474
HilaCAM [3]	0.4765	0.4072	0.0890
TCLIP [Ours]	0.6925	**0.6295**	**0.3817**

of 0.775, significantly surpassing the previous best method. Our leading energy-PG score of 0.445 further demonstrates that our heatmaps not only point to the correct location but remain highly concentrated within the object's boundaries.

Furthermore, we conduct a full pixel wise segmentation test evaluating Pixel Accuracy, Average Precision (AP), and mask IoU (Table 3). While CLIPSurgery achieves a higher Pixel Accuracy, TCLIP is decisively state-of-the-art in the metrics that evaluate the structural quality of the segmentation shape. TCLIP establishes a new SOTA in both Average Precision and maskIoU. This substantial improvement in maskIoU over the next best method strongly indicates that TCLIP's token level relevance maps capture the true shape and spatial extent of objects with much higher fidelity than existing baselines.

4. Conclusion and Limitations

In this work, we introduced TCLIP, a training-free and computationally efficient interpretability method that provides a direct text-token-to-image-patch relevance mapping for CLIP. Our quantitative evaluations demonstrate that TCLIP achieves state-of-the-art zero-shot localization and segmentation performance compared to existing baselines. Qualitatively, TCLIP moves beyond simple object localization to provide unique insights into CLIP's compositional understanding and its grounding of abstract concepts.

While highly effective, our approach has certain limitations that present clear avenues for future work. First, TCLIP is specifically designed for Vision Transformer (ViT) architectures to leverage their direct patch-token alignment. Extending this precise mapping to CNN backbones would require adapting our framework to compute relevance over the final convolutional feature maps instead of discrete ViT patch embeddings. Finally, TCLIP currently computes relevance based on the final projected embeddings, effectively assuming an equal initial contribution from all text tokens. To address this, future iterations will explore weighting the initial token embeddings by the magnitude of the model's internal self-attention weights. This will account for unequal initial token contributions and further refine the interpretability of complex compositional prompts.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers, 2020. 5

[2] 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, 2018. 4

[3] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers, 2021. 1, 2, 3, 4, 5

[4] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Maskclip: Masked self-distillation advances contrastive language-image pretraining, 2023. 5

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 1

[6] Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns, 2020. 4

[7] Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. 2022. 2, 4

[8] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022. 1

[9] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical

class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021. 4

[10] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning, 2023. 1

[11] Martha Lewis, Nihal V. Nayak, Peilin Yu, Qinan Yu, Jack Merullo, Stephen H. Bach, and Ellie Pavlick. Does clip bind concepts? probing compositionality in large image models, 2024. 1

[12] Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A closer look at the explainability of contrastive language-image pre-training, 2024. 1, 5

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 2

[14] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. 4

[15] Saad Biaz Trung Bui Peijie Chen, Qi Li and Anh Nguyen. gscorecam: What is clip looking at? In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2022. 1, 2, 3, 4

[16] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models, 2018. 4

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1

[18] 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. *International Journal of Computer Vision*, 128(2):336–359, 2019. 1, 2, 3, 4, 5

[19] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks, 2020. 4

[20] Ying Wang, Tim G. J. Rudner, and Andrew Gordon Wilson. Visual explanations of image-text representations via multimodal information bottleneck attribution, 2024. 5

[21] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it?, 2023. 1

[22] Qinglong Zhang, Lu Rao, and Yubin Yang. Group-cam: Group score-weighted visual explanations for deep convolutional networks, 2021. 4

[23] Chenyang Zhao, Kun Wang, Janet H. Hsiao, and Antoni B. Chan. Grad-eclip: Gradient-based visual and textual explanations for clip, 2025. 1, 5