What does CLIP see per token?

Wajahat Ali Khan, Seungkyu Lee
Kyung Hee University, South Korea

PerCV: Perception & Computer Vision Lab. — KYUNG HEE UNIVERSITY

CVPR JUNE 3-7, 2026 DENVER COLORADO

Abstract

While CLIP [1] demonstrate remarkable zero-shot generalization, its internal decision making processes remains highly opaque. Current explainability methods provide holistic explanations for the entire text prompt on the image, failing to answer a fundamental question: what exactly does CLIP see for each individual word in the image? **TCLIP** is a training-free, gradient-free interpretability method that establishes a direct, dense relevance mapping between individual text tokens and image patches.

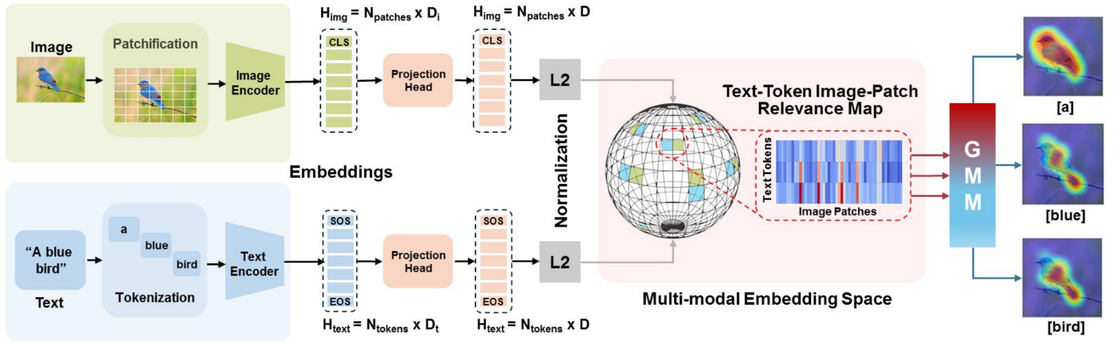
Problem

- CLIP relies on a single, global cosine similarity score computed from the final [CLS] and [EOS] tokens [2].
- Standard XAI methods generate aggregated heatmaps for the entire text prompt, failing to spatially differentiate the visual grounding of specific words.
- Gradient based alternatives require complex architectural modifications or memory intensive backpropagation.

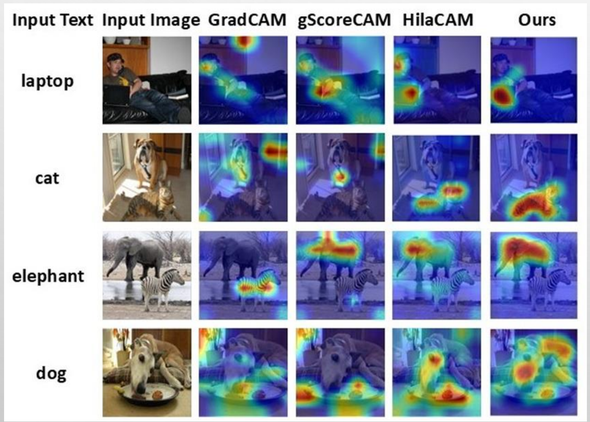
Results

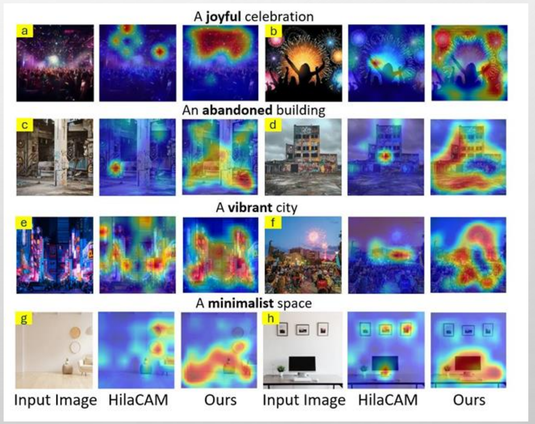
- **High Fidelity Token Grounding:** TCLIP's activation is tightly focused on the object of interest (laptop, cat), avoiding the diffuse results of baseline methods [3], [4], [5].
- **Abstract Reasoning:** TCLIP extends interpretability beyond objects, isolating visual stimuli for subjective concepts like "minimalist" (negative space) or "joyful", "vibrant" (lights).
- **Zero Shot Object Detection (MS COCO):** Achieved 19.07% Box Accuracy at IoU > 0.5, vastly outperforming ViT-based methods.
- **Computational Efficiency:** Requires 0 backward passes and only 25 forward passes, remaining highly computationally efficient compared to CNN based baselines.
- **Latent Segmentation (ImageNet-S):** Achieved SOTA Average Precision and maskIoU scores, proving TCLIP captures true shape geometry [6], [7], [8], [9].

Methodology



TCLIP operates directly 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, we map them into the shared multimodal embedding space. After L2 normalization, we compute a direct relevance matrix ($S = \hat{T} \cdot \hat{I}^T$). Finally, Gaussian Mixture Model (GMM) thresholding dynamically suppresses background noise, yielding high fidelity spatial heatmaps for any individual word without requiring gradients or retraining.





COCO

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

ImageNet

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

Conclusions

- TCLIP successfully exposes a transparent grid of visual-linguistic alignments native to CLIP.
- It provides unique insights into CLIP's compositional understanding and its grounding of abstract concepts without the overhead of retraining.
- Crucially, TCLIP deconstructs complex compositional mechanisms and reveals the emergent mathematical structures driving abstract visual associations.

Contact

Wajahat Ali Khan
Kyung Hee University, South Korea
Email: wajahat@khu.ac.kr
Website: wajahat-alikhan.github.io

References

1) Alec Radford, et al. Learning transferable visual models from natural language supervision, 2021
2) Jack Hessel, et al. Clipscore: A reference-free evaluation metric for image captioning, 2022
3) Ramprasaath R. Selvaraju, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization.
4) Saad Biaz, et al. gscorecam: What is clip looking at?
5) Hila Chefer, et al. Generic attention model explainability for interpreting bi-modal and encoder-decoder transformers, 2021.
6) Chenyang Zhao, et al. Grad-eclip: Gradient-based visual and textual explanations for clip, 2025.
7) Xiaoyi Dong, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining, 2023.
8) Yi Li, et al. A closer look at the explainability of contrastive language-image pre-training, 2024
9) Ying Wang, et al. Visual explanations of image-text representations via multi-modal information bottleneck attribution, 2024.