
Sparsification of CLIP using conceptual decomposition of images

Mansour Osman

`mansouro@andrew.cmu.edu`

Mona Aman

`amona@andrew.cmu.edu`

Ahmed Alhassan

`aalhassa@andrew.cmu.edu`

Azza Osman

`aiosman@andrew.cmu.edu`

Abstract

Recent advances in deep multimodal networks, such as CLIP models, have achieved impressive performance on various tasks but remain largely opaque, limiting their interpretability and diagnostic utility. Many prior research has attempted to address this issue. In this paper, we propose a unified interpretability framework that combines sparse semantic decomposition, gradient-based visualization, and model sparsification to enhance the interpretability of deep multimodal networks. Our approach first uses SpLiCE to transform dense CLIP image embeddings into sparse human-interpretable representations. These sparse concept activations are then spatially localized using GRAD-ECLIP, a gradient-based method to generate heat maps for each term, revealing the spatial contribution of individual concepts. Then, we apply SPADE, a sparsity-guided debugging technique, to prune the network in a sample-specific manner. Extensive experiments demonstrate that our multistage pipeline not only preserves downstream performance but also significantly improves the clarity, fidelity, and diagnostic value of visual explanations, as demonstrated by GRAD-ECLIP.

[Github Repo](#)

1 introduction:

Deep multimodal models, such as OpenAI’s CLIP [1], have revolutionized vision-language learning by enabling zero-shot transfer across a wide range of tasks. By training on large-scale image-text pairs, CLIP learns a shared embedding space where both modalities are aligned. Despite their impressive capabilities, these models remain largely black-box systems, making it difficult to understand what semantic concepts they rely on when making predictions. This lack of interpretability hinders trust, accountability, and downstream debugging, particularly when such models are deployed in real-world applications.

SpLiCE[2] and GRAD-ECLIP[3] form a complementary pair of interpretability methods designed specifically for the multimodal architecture of CLIP. SpLiCE transforms dense CLIP embeddings into a sparse, human-interpretable representation by decomposing them into non-negative linear combinations of semantic concepts. This approach highlights the key factors driving the model’s latent space without requiring additional supervision or training labels. In parallel, GRAD-ECLIP leverages the gradient information within the Transformer-based CLIP model to produce high-resolution heat maps for each identified concept, thus localizing which pixels contribute the most to a particular

semantic component[3]. By coupling SpLiCE’s sparse concept decomposition with GRAD-ECLIP’s spatial attributions, researchers can diagnose how CLIP balances multiple overlapping themes, such as objects and context, while offering clearer and more granular insights into the decision-making processes of the model.

The ability to explain and interpret deep neural network decisions has become increasingly critical, particularly in high-stakes applications like medical diagnosis. Although modern image classification models achieve impressive accuracy, their black-box nature raises concerns about reliability, bias, and trustworthiness. Methods that provide visual explanations of model decisions, such as SPADE[4], have emerged as powerful tools to understand how these models arrive at their classifications. However, as of today, SPADE is only implemented for RESNET[5] and GPT[6] models.

In this work, we present a unified interpretability framework that enhances transparency in CLIP-like models by decomposing their dense representations into sparse, human-interpretable concepts, localizing these concepts spatially within images, and refining model explanations through structured sparsification. Our approach integrates three complementary techniques:

- **Sparse linear concept embeddings (SpLiCE)** - We first employ SpLiCE[2] to transform dense CLIP embeddings into sparse, interpretable representations. By expressing an image as a non-negative linear combination of a predefined dictionary of concepts, SpLiCE enables us to understand the high-level features CLIP is paying attention to in an image.
- **Gradient-Based Explanation for Transformer-Based CLIP (GRAD-ECLIP)** To further refine these interpretations, we utilize GRAD-ECLIP, a gradient-based attribution method tailored for the CLIP Transformer architecture. By generating heat maps corresponding to each sparse concept, GRAD-ECLIP provides spatial visualizations of how specific concepts influence the model predictions.[3]
- **Sparsity-Guided Debugging (SPADE)**. We apply SPADE[4] to sparsify the model execution trace by pruning extraneous neurons and connections in a sample-specific manner. This disentangling step removes polysemantic activations, making model decisions more interpretable and debuggable.

By combining these techniques, our framework overcomes several limitations of previous interpretability methods. Unlike standard saliency maps, which often suffer from visual clutter and fail to distinguish between conceptually meaningful features, our sparse decomposition ensures that explanations are both concise and semantically meaningful. Furthermore, SPADE improves interpretability by isolating critical pathways within the network, addressing polysemanticity, a key challenge in mechanistic interpretability.

We evaluated our approach in multiple datasets, demonstrating that it maintains CLIP’s zero-shot classification performance while significantly improving interpretability. Our experiments show that our method enables more effective model debugging, facilitates human alignment with model decisions, and uncovers spurious correlations within datasets.

In summary, this work contributes a novel integrated pipeline that enhances the interpretability of deep multimodal models by combining sparse semantic decomposition, gradient-based attribution, and structured pruning. By bridging the gap between high-dimensional feature representations and human-understandable concepts, our framework paves the way for more transparent and trustworthy AI systems.

2 Related work:

Recent advances in multimodal models have transformed how we leverage large-scale language-vision representations for downstream tasks. In particular, the CLIP model demonstrated that training in image-text pairs can produce highly performant yet opaque dense embeddings[1]. Although these representations enable zero-shot classification, image retrieval, and even image captioning [4][7], their lack of interpretability has motivated a surge of research on understanding and debugging such models.

A substantial body of work has focused on post hoc explanation techniques to interpret deep network predictions. Gradient-based methods, including saliency maps [8], Integrated Gradients [9], and Grad-CAM[10], compute the influence of input features on the output of the model. Although widely

adopted, these approaches often produce noisy or overly diffuse explanations. Perturbation-based techniques such as LIME[11] and occlusion methods[12] offer alternative perspectives but are limited by issues related to input distribution shifts when masking.

Beyond pixel-level attributions, concept-based methods aim to bridge the semantic gap between model representations and human-understandable concepts. TCAV[13] introduced the idea of concept activation vectors to quantify the sensitivity of model predictions to high-level concepts. Similarly, concept bottleneck models (CBMs) enforce an intermediate representation based on interpretable concepts, albeit often requiring expert-defined concept labels or task-specific datasets. In contrast, our approach leverages a large and overcomplete dictionary to extract semantic decompositions without the need for labeled concepts, a strategy aligned with recent efforts in dictionary learning for interpretability[14].

Mechanistic interpretability has also focused on understanding the inner workings of deep neural networks by visualizing the activation patterns of individual neurons[15][16]. These studies have revealed that neurons in large networks are often polysemantic, that is, they respond to multiple distinct features, complicating direct interpretation. Disentangling these mixed signals is crucial for developing clearer explanations, and several works have attempted to achieve this through specialized regularization during training[16] or through post hoc methods.

Sparsity has emerged as a promising tool for improving interpretability. It has been shown that enforcing sparsity in the final layers can improve the attribution of predictions by reducing the number of contributing neurons[17]. Building on this idea, SPADE (Sparsity-Guided Debugging for Deep Neural Networks)[4] introduces a sample-specific pruning strategy that extracts a “trace” of the network execution, effectively disentangling polysemantic neurons without requiring retraining. This approach not only yields more faithful saliency maps and neuron visualizations but also provides actionable insights for model debugging.

Our work is also closely related to recent efforts that aim to decompose dense representations into interpretable components. In “Interpreting CLIP with Sparse Linear Concept Embeddings (SpLiCE)”[2], the authors take advantage of the structured latent space of CLIP to express image embeddings as sparse, non-negative combinations of human-interpretable concepts. SpLiCE is task-agnostic and can be applied post hoc, aligning with our goal of enhancing transparency without modifying the original model.

It is well known that sparse models are easier to interoperate[18][19]. Previous work attempted to achieve sparsity by altering the model itself, either by pre-trained with sparsity constraints or apply sparsity after training, which can disrupt the network’s original performance and behavior[20][21][22]. recent study has introduced SPADE a method which introduces sample-specific pruning to improve explanation fidelity. Outperform previous methods in deletion/insertion metrics.

In summary, our framework builds on the foundations of multimodal learning and post hoc interpretability by integrating three complementary techniques. First, we use SpLiCE to obtain sparse semantic decompositions that reveal the underlying concepts in CLIP embeddings. Next, GRAD-ECLIP employs gradient-based methods tailored for Transformer-based CLIP models to generate spatial heat maps of these concepts. Finally, SPADE further refines these explanations through sample-specific network pruning, addressing the challenges of neuron polysemanticity. Together, these advances address the limitations of prior methods by offering a unified, efficient, and interpretable pipeline for debugging and understanding deep multimodal networks.

3 Methodology:

We introduce a comprehensive interpretability framework that unifies three novel techniques, SpLiCE[2], GRAD-ECLIP[3], and SPADE[4], to enhance the transparency and diagnostic power of deep multimodal models such as CLIP. Our approach first employs Sparse Linear Concept Embeddings (SpLiCE) to transform high-dimensional, dense CLIP image embeddings into sparse, human-interpretable representations. By expressing each image as a sparse, nonnegative linear combination on a large, overcomplete dictionary of semantic concepts (for example, ‘coffee’, ‘birthday party’), SpLiCE isolates the underlying semantic factors embedded in the latent space, revealing which concepts the model relies on during inference.

To create this vocabulary using SPLICE methodology[2], we analyzed the most common one- and two-word bigrams in the text captions of the LAION-400m dataset, which is commonly used for training CLIP versions. We filter the captions to remove inappropriate content, and we trim our concept set so that no two terms have a cosine similarity of more than 0.9. Then, we delete bigrams with a cosine similarity of more than 0.9 to the average of individual words. We selected the top 10,000 most common single-word and 5000 most common two-word concepts for our concept vocabulary.[2]

Building on this semantic decomposition, we then introduce GRAD-ECLIP[3], a gradient-based visual explanation method specifically tailored for transformer-based CLIP models. GRAD-ECLIP leverages the sparse concept activations from SpLiCE and generates high-resolution heat maps that spatially localize the contribution of each identified concept within the input image. These heat maps allow for a more intuitive visualization of the decision process of the model by clearly indicating the regions that activate particular semantic components.

Then, we integrate SPADE[4], a sparsity-guided debugging technique, to further refine the interpretability of the model. SPADE operates as a sample-specific preprocessing step that prunes the network weights in a targeted manner while preserving its original predictions. This selective pruning disentangles polysemantic neurons by eliminating extraneous connections and highlighting the most critical pathways used by the model for a given sample. The resulting trace of the execution of the network improves the fidelity of both the visualizations of neurons and the saliency maps, offering clearer insights into the internal workings of the network.

The faithfulness of the visual explanations produced by Grad-ECLIP is statistically assessed using insertion and deletion criteria. These metrics evaluate the degree to which Grad-ECLIP’s heat maps correctly pinpoint words or areas that are essential to CLIP’s image-text matching predictions. begins with a baseline image that is empty or neutral, then gradually adds pixels according to their significance as shown by the Grad-ECLIP heat map. At every stage, the model’s prediction performance such as its accuracy is noted. This is the same metric used in SPADE[4].

In summary, our unified framework, combining SpLiCE’s semantic decomposition, GRAD-ECLIP’s spatial attribution, and SPADE’s sample-specific network pruning, provides a robust, efficient, and interpretable solution for demystifying decision-making processes of complex multimodal neural networks. This work paves the way for more transparent AI systems that can be audited, debugged, and trusted more easily in practical applications, as shown in Figure 1.

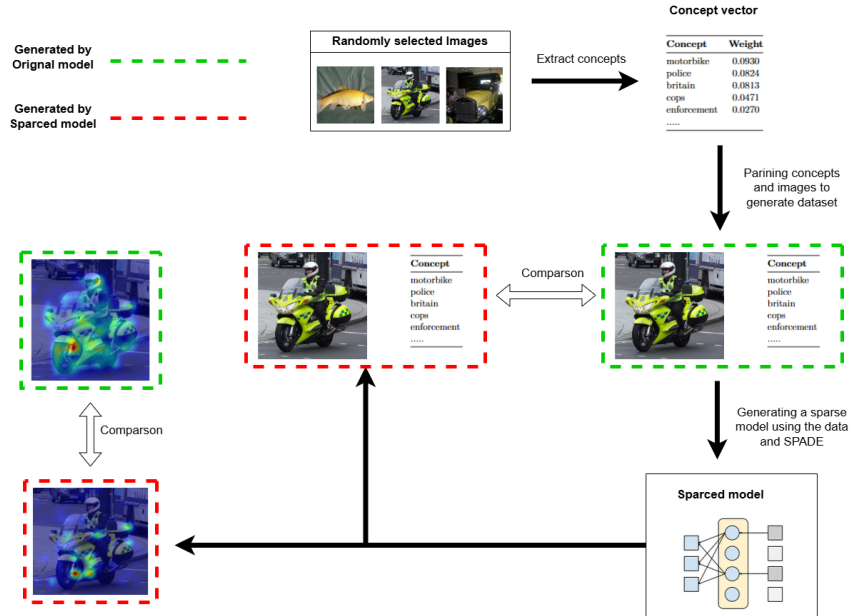


Figure 1: Methodology Diagram

4 Baseline and Extensions

4.1 Baseline

Our baseline approach employs the original dense CLIP embeddings combined with standard interpretability techniques (e.g., Grad-CAM, Integrated Gradients, and occlusion methods). Although these methods provide a general overview of the behavior of the model, they often yield noisy polysemantic explanations that obscure the underlying decision process.

4.2 Extensions

To overcome these limitations, we extend the baseline with three key components.

- **SpLiCE:** Transforms dense CLIP representations into sparse, non-negative linear combinations of semantic concepts. This transformation isolates distinct, human-understandable features, thus improving interpretability.
- **GRAD-ECLIP:** Leverages gradient information from transformer-based CLIP models to produce spatially resolved heat maps for each sparse concept, localizing the contributions of individual features within the input image.
- **SPADE:** Applies sample-specific network pruning to remove extraneous connections, effectively disentangling polysemantic neurons and highlighting the critical pathways that drive model predictions.

Together, these extensions refine the interpretability of CLIP predictions without compromising performance, offering clearer and more actionable insights for model debugging and analysis.

5 Results and analysis:

Our goal is to implement SPADE on CLIP architecture method while maintaining reasonable accuracy levels and maintaining close SPLICE interpretation between the sparse model and the full model and comparable GRAD-ECLIP heat maps from the concepts extracted by SPLICE.

We tested our pipeline on an image of a police motorcycle in traffic - used to demonstrate SPLICE[2] - and have show the heatmaps of the concept motorbike for different sparsification percentages (1%, 10%, 25%, and 50%)

Figure 3 displays GRAD-ECLIP heat maps corresponding to CLIP models with different sparsity levels for the concept of “motorbike.” The activation of the concept “motorbike” is most intense around the body and wheels of the vehicle in all models except 50%, with it being more pronounced in the model with 25% sparsity, indicating the model’s focus on those distinctive elements. Together, these visualizations underscore how SpLiCE and GRAD-ECLIP can unravel themes (e.g., motorbike) within a single scene and assign interpretable heat maps that localize the contribution of each concept in the image.

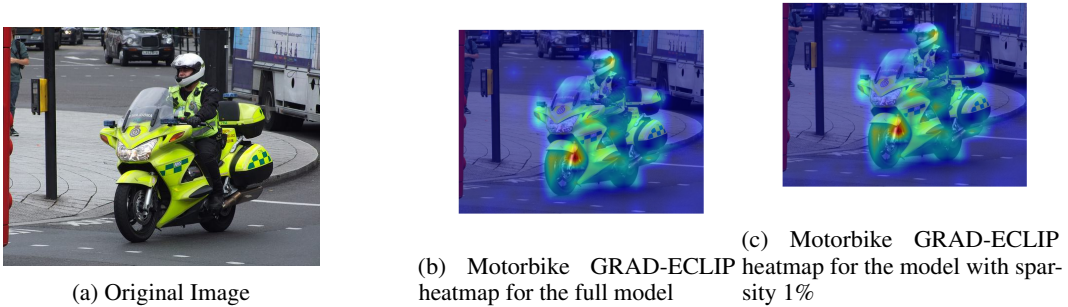


Figure 2: GRAD-ECLIP heatmap for top Motorbike for the full and 99% of the model



(a) Motorbike GRAD-ECLIP heatmap for the model with sparsity 10%



(b) Motorbike GRAD-ECLIP heatmap for the model with sparsity 50%



(c) Motorbike GRAD-ECLIP heatmap for the model with sparsity 50%

Figure 3: GRAD-ECLIP heatmap for top Motorbike for the 90%, 75%, and 50% of the model

Then an image from the dataset was selected to extract the concepts from it using SpLiCE with the generated models. For most of the models, all the top 5 concepts are relevant to the image, except for the 50% sparsified model, with all concepts not being remotely close to the image.



Figure 4: Image used to test SPLICE

6 Proposed Improvements and Future Steps

Our current work has successfully executed the three individual methods, and we have established a meaningful link between SpLiCE and GRAD-ECLIP. These components work together to transform dense CLIP embeddings into sparse, interpretable semantic representations and to generate spatially resolved gradient-based explanations, respectively.

However, our initial implementation of SPADE for CLIP remains in its early stages. At present, SPADE has not been extensively tested, and we have not yet integrated it with the SpLiCE and GRAD-ECLIP pipeline. Moreover, we have not yet selected a representative set of images on which to apply SPADE for sparsifying the CLIP model.

In future work, we plan to:

Table 1: Top-weighted concepts at different sampling fractions

Level of sparsification									
Full		0.01		0.1		0.25		0.5	
Concept	wt.	Concept	wt.	Concept	wt.	Concept	wt.	Concept	wt.
roadster	0.1279	ford	0.0516	ford	0.0546	opel	0.0661	keychain	0.0550
olds	0.0469	automobile	0.0465	auto	0.0546	diecast	0.0607	bird	0.0514
autos	0.0464	auto	0.0425	preserved	0.0451	roadster	0.0434	pest	0.0466
chevrolet	0.0441	antique	0.0423	automobile	0.0438	car	0.0420	insect	0.0459
ford	0.0431	garage	0.0373	opel	0.0380	acquisition	0.0375	perch	0.0420

- Integrate SPADE with SpLiCE and GRAD-ECLIP to create a unified, end-to-end interpretability framework.
- Experimenting thoroughly to evaluate how SPADE-based model sparsification improves the interpretability and diagnostic value of CLIP.
- Analyze the impact of SPADE on downstream tasks such as model debugging and editing, particularly in terms of disentangling polysemantic neurons.

These steps will help us further refine our approach and enhance the overall transparency of multi-modal neural networks.

References

- [1] A. Radford, J. W. Kim, C. Hallacy, *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*, PmLR, 2021, pp. 8748–8763.
- [2] U. Bhalla, A. Oesterling, S. Srinivas, F. P. Calmon, and H. Lakkaraju, “Interpreting clip with sparse linear concept embeddings (splice),” in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, *et al.*, Eds., vol. 37, Curran Associates, Inc., 2024, pp. 84 298–84 328. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/996bef37d8a638f37bdfcac2789e835d-Paper-Conference.pdf.
- [3] C. Zhao, K. Wang, J. H. Hsiao, and A. B. Chan, “Grad-eclip: Gradient-based visual and textual explanations for clip,” *arXiv preprint arXiv:2502.18816*, 2025.
- [4] A. S. Moakhar, E. Iofinova, E. Frantar, and D. Alistarh, “Spade: Sparsity-guided debugging for deep neural networks,” *arXiv preprint arXiv:2310.04519*, 2023.
- [5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [6] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, “Improving language understanding by generative pre-training,” 2018.
- [7] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical text-conditional image generation with clip latents,” *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.
- [8] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” *arXiv preprint arXiv:1312.6034*, 2013.
- [9] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in *International conference on machine learning*, PMLR, 2017, pp. 3319–3328.
- [10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
- [11] M. T. Ribeiro, S. Singh, and C. 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*, 2016, pp. 1135–1144.
- [12] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, Springer, 2014, pp. 818–833.
- [13] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, *et al.*, “Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav),” in *International conference on machine learning*, PMLR, 2018, pp. 2668–2677.
- [14] K. P. Panousis, D. Ienco, and D. Marcos, “Sparse linear concept discovery models,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 2767–2771.

- [15] C. Olah, A. Mordvintsev, and L. Schubert, “Feature visualization,” *Distill*, vol. 2, no. 11, e7, 2017.
- [16] A. Nguyen, J. Yosinski, and J. Clune, “Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks,” *arXiv preprint arXiv:1602.03616*, 2016.
- [17] E. Wong, S. Santurkar, and A. Madry, “Leveraging sparse linear layers for debuggable deep networks,” in *International Conference on Machine Learning*, PMLR, 2021, pp. 11 205–11 216.
- [18] H. Cunningham, A. Ewart, L. Riggs, R. Huben, and L. Sharkey, “Sparse autoencoders find highly interpretable features in language models,” *arXiv preprint arXiv:2309.08600*, 2023.
- [19] A. Tamkin, M. Taufeeque, and N. D. Goodman, “Codebook features: Sparse and discrete interpretability for neural networks,” *arXiv preprint arXiv:2310.17230*, 2023.
- [20] S. Srinivas, A. Subramanya, and R. Venkatesh Babu, “Training sparse neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 138–145.
- [21] M. M. Junaid Baig and A. Stepanyants, “Sparse learning enabled by constraints on connectivity and function,” *arXiv e-prints*, arXiv–2409, 2024.
- [22] R. Gong, Y. Yong, Z. Wang, *et al.*, “Fast and controllable post-training sparsity: Learning optimal sparsity allocation with global constraint in minutes,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 12 190–12 198.

Contributions

- Brainstroming: mansouro, amona, aalhassa, aiosman.
- Code: aalhassa.
- Report: mansouro, aiosman.
- Review: amona.