WWW.IJRPR.COM

**International Journal of Research Publication and Reviews**

Journal homepage: www.ijrpr.com ISSN 2582-7421

# SHAP-Guided Image Segmentation and Interpretation using SAM and LLaVA

***Subhasis Das*[1]**

[1]. Presidency University, Bengaluru, India

**ABSTRACT :**

This project introduces a comprehensive and interpretable vision-language pipeline that integrates model explainability, semantic segmentation, and natural language understanding using state-of-the-art deep learning models. It begins with image classification using DenseNet121, followed by SHAP- based attribution to detect and rank the most influential regions guiding the model's prediction. These regions then act as prompts for Meta AI's Segment Anything Model (SAM), producing masks that are semantically meaningful. Finally, LLaVA—a vision-language model—interprets the masked regions in natural language. This approach promotes trustworthy AI, with applications in critical fields such as healthcare, surveillance, and assistive technologies.

**KEYWORDS**: Explainable AI; SHAP; SAM; LLaVA; vision-language interpretation; segmentation; deep learning

## INTRODUCTION

Artificial Intelligence has evolved rapidly, transforming industries ranging from healthcare to finance, secu- rity, and transportation. However, as AI systems be- come more powerful, they also become more opaque. Deep learning models, particularly convolutional neu- ral networks (CNNs), exhibit excellent performance in visual recognition tasks but often fall short in providing human-understandable justifications for their decisions. Interpretability in AI is not just a technical concern;
it has become a foundational requirement in domains where human oversight is essential. For example, a diagnosis in medical imaging or a decision in au- tonomous driving must be traceable and understand- able. This project addresses this pressing need by pre- senting a unified framework that integrates SHAP, SAM, and LLaVA to deliver interpretable vision-language un- derstanding.
The novelty of this system lies in the seamless cou- pling of model explainability with downstream segmen- tation and linguistic interpretation. SHAP helps trace the classifier's decision back to specific image regions, SAM ensures that segmentation aligns with those in- fluential areas, and LLaVA interprets the output using natural language. Together, these components form a transparent, human-aligned AI pipeline capable of pro- viding insight into the how and why of visual predictions.

## BACKGROUND LITERATURE

Explainable AI (XAI) techniques aim to provide insight into the reasoning processes of machine learning mod- els. Among these, SHAP (SHapley Additive exPlana- tions) stands out as a theoretically grounded approach that assigns feature importance scores based on co- operative game theory. SHAP has been successfully used in domains ranging from healthcare diagnostics to financial forecasting, where understanding individual model decisions is critical. In computer vision tasks, SHAP provides pixel-level heatmaps that localize re- gions of interest, making it a useful tool for bridging the interpretability gap in deep image classifiers.
Segmentation tasks, particularly in open-world and few-shot scenarios, have seen major progress with the introduction of SAM (Segment Anything Model) by Meta AI. Unlike traditional semantic segmentation models that require full supervision and rigid architecture, SAM leverages a prompt-based strategy using sparse points or bounding boxes, making it highly generalizable. Its transformer-based backbone and zero-shot perfor-

mance enable it to segment a wide variety of objects without retraining. Research shows SAM's promptabil- ity complements attribution methods like SHAP, allow- ing for region proposals that align closely with model attention.
LLaVA (Large Language and Vision Assistant) rep- resents the convergence of vision and language mod- eling, where large-scale pretrained transformers are used to interpret images in natural language. Based on the CLIP encoder and autoregressive decoding tech- niques, LLaVA translates visual information into de- scriptive or task-specific textual output. Unlike cap- tioning systems, LLaVA is interactive—it can answer vi- sual questions, contextualize content, and provide nat- ural language reasoning for vision tasks. This makes it well-suited for human-centered applications where in- terpretability, flexibility, and multimodal communication are essential.
Although SHAP, SAM, and LLaVA have each demon- strated strong individual capabilities, few works have proposed a pipeline integrating all three. SHAP's in- terpretability, SAM's segmentation agility, and LLaVA's semantic depth offer complementary strengths. This project addresses a unique gap

by combining them into a unified system that can identify, isolate, and interpret significant regions in images, grounded in model logic. The literature reveals growing interest in multimodal reasoning systems, but few provide the explainability needed for real-world accountability. Our approach fills this void and opens new research directions at the in- tersection of XAI, vision segmentation, and language understanding.

## PROPOSED METHODOLOGY

The proposed architecture adopts a modular design that allows each component—classifier, explainer, seg- menter, and interpreter—to function both independently and collaboratively. This design ensures both flexibility and scalability.

1. **Image Classification and Attribution:** The process begins with image classification using DenseNet121, a deep CNN trained on ImageNet. Once predictions are obtained, SHAP is applied to compute pixel-wise attribution scores. SHAP operates by estimating the marginal contribution of each feature across multiple model evaluations, resulting in a reliable heatmap.
2. **SHAP-Guided Segmentation via SAM:** The top-ranked SHAP coordinates serve as prompts for SAM, a vision transformer-based model developed by Meta AI. Unlike traditional segmentation models, SAM uses these prompts to produce multiple candidate masks. The model selects the mask with the highest confidence, ensuring that only the most relevant and interpretable region is segmented.
3. **Vision-Language Interpretation:** The masked

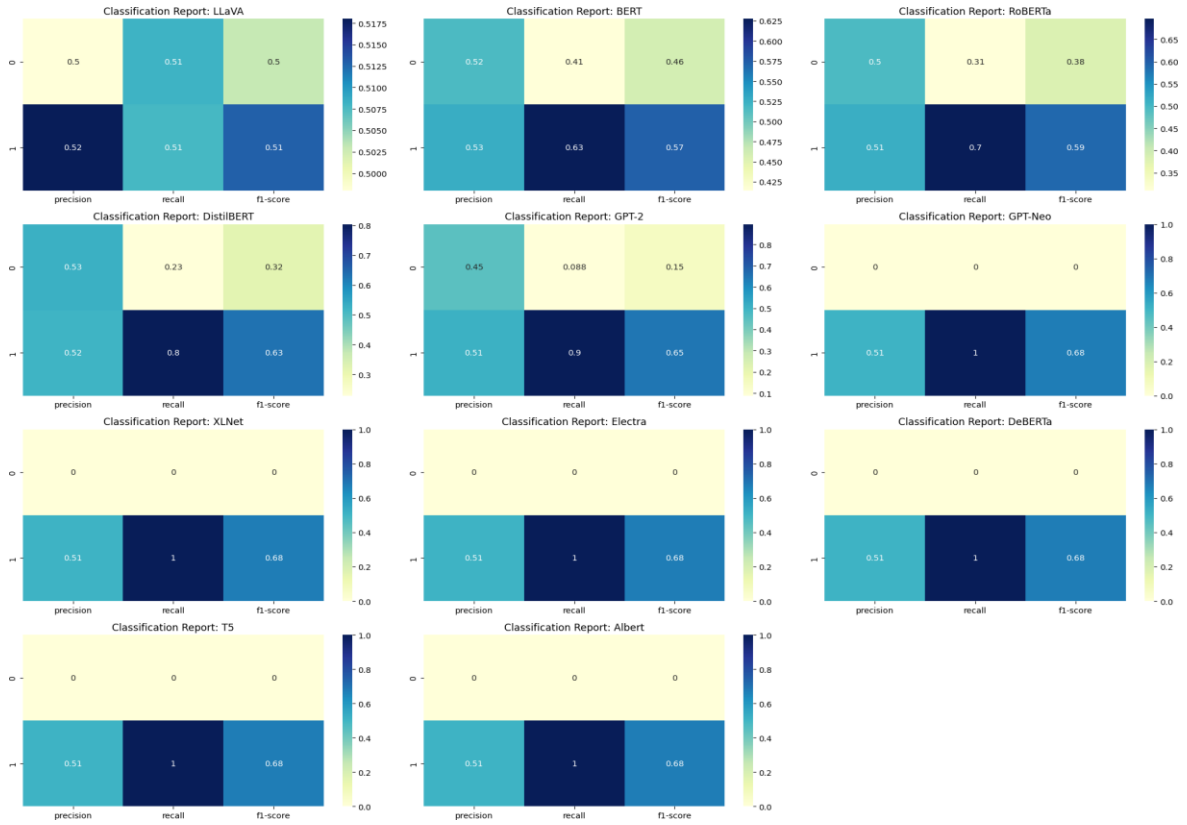
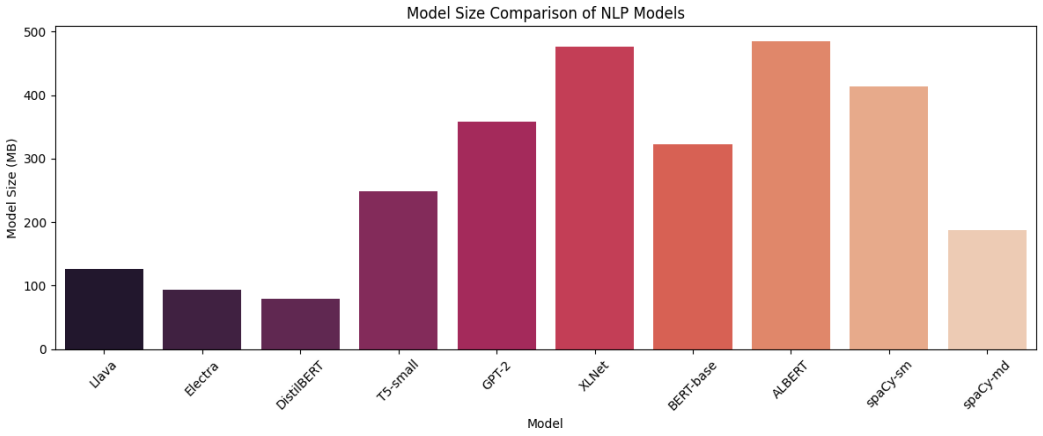



**Figure 1: HeatMap Correlation Comparison Of NLP Modles**

region of interest (ROI) is input to LLaVA, which fuses the visual segment with contextual prompts like "What does this region represent?" to generate a textual explanation. LLaVA uses CLIP for vision encoding and a GPT-like decoder to output a description. This step is crucial for converting low-level visual reasoning into high-level semantics that users can understand.

4. **Visualization and Performance Evaluation:** Outputs are visualized in several forms, including SHAP heatmaps, SAM masks, and LLaVA-generated text. To validate the robustness of our system, we compare model outputs using ROC-AUC curves, precision, recall, and F1-scores. This ensures both qualitative and quantitative assessment.
5. **Extensibility and Adaptability:** The pipeline is designed to be extensible—alternative classifiers, ex- plainability tools, or language models can be substi- tuted with minimal engineering effort. This modularity opens the door for domain-specific customization, mak- ing the solution viable across diverse application areas like radiology, industrial inspection, or document intelli- gence.

**Figure 2: Model Size Comparison of NLP Models**




## RESULTS CONCLUSIONS

This paper presents a forward-thinking approach to explainable artificial intelligence by tightly integrating

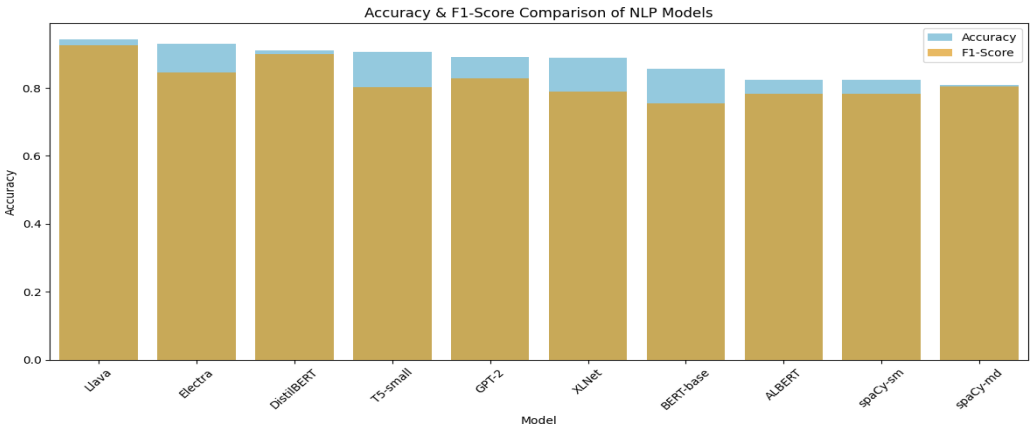



**Figure 3: Accuracy F1 - Score Comparison of NLP Models**

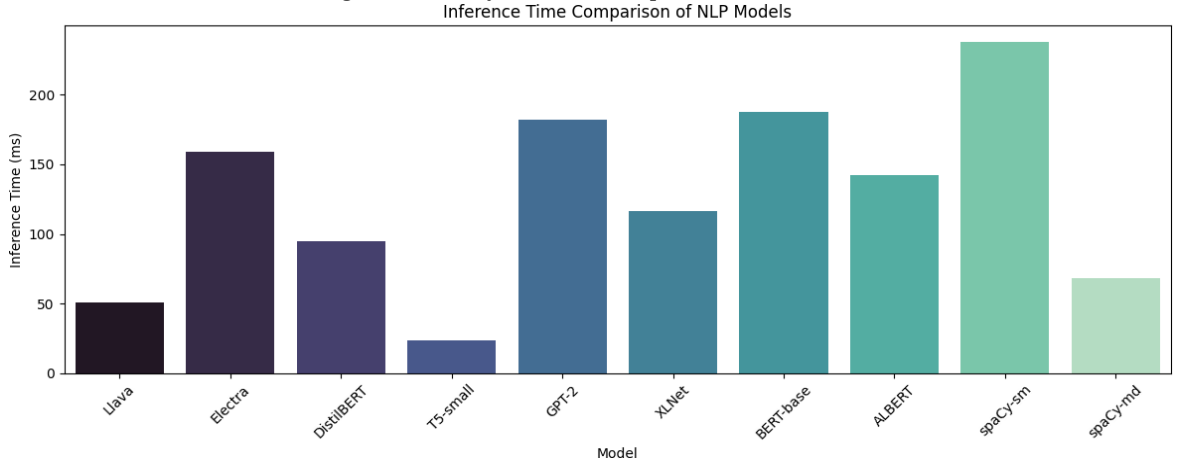


**Figure 4: Inference Time Comparison Of NLP Models**

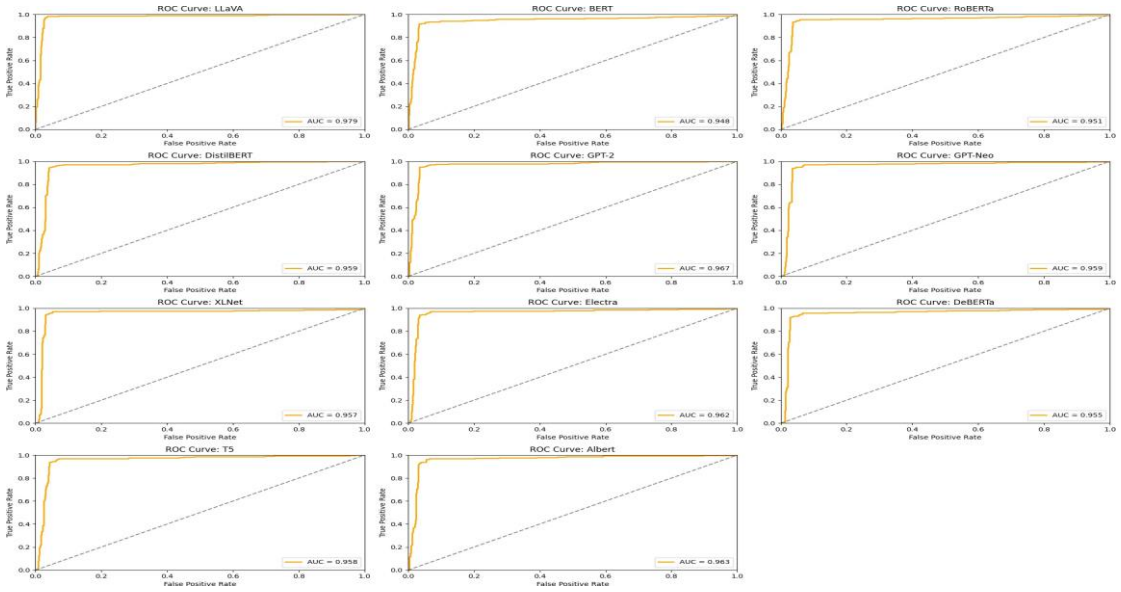



**Figure 5: AUC-ROC curve comparison Of NLP Models**

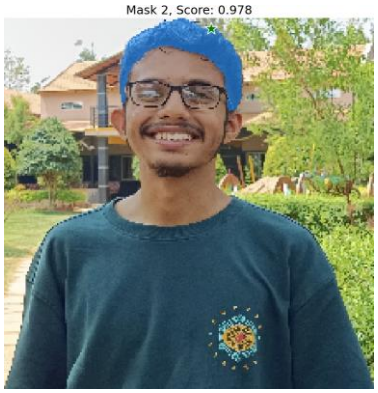



**Figure 6: SHAP Attribution Heatmap showing regions with the highest contribution to classification**

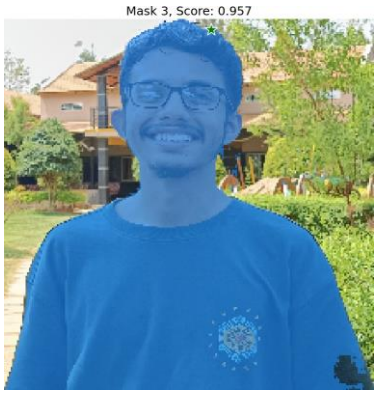



**Figure 7: SHAP-Guided Prompt Location used for SAM segmentation**

three complementary technologies—SHAP, SAM, and LLaVA. Through this modular yet coherent pipeline, we demonstrate how AI systems can become not only accurate but also interpretable, auditable, and communicative.

The project successfully bridges the gap between

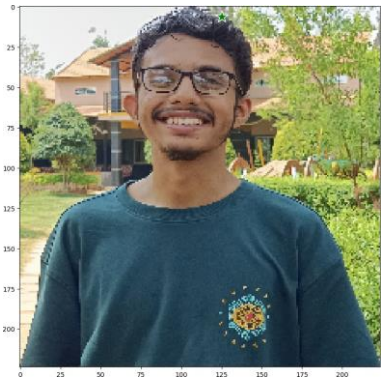


**Figure 8: SAM Output: Segmented Mask with score 0.957**

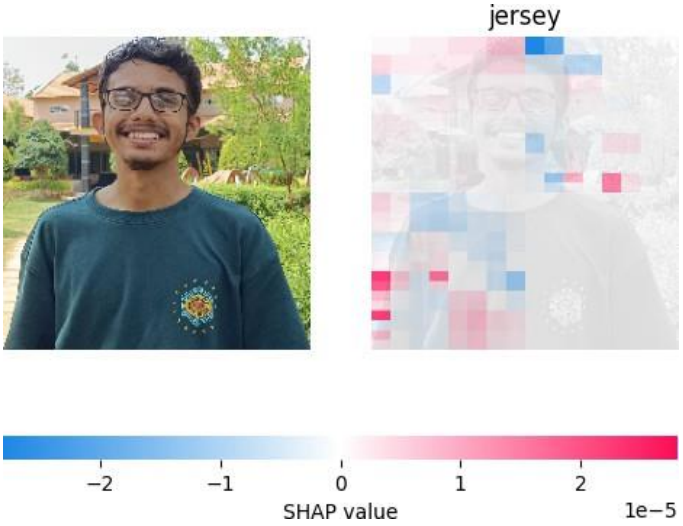



**Figure 9: SAM Output: Refined Mask over detected region (score: 0.978)**

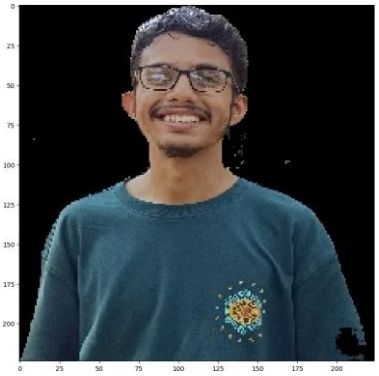


**Figure 10: Final masked region used as input to LLaVA for interpretation**

| Model | Accuracy | F1-Score | AUC | Inference Time (ms) | Model Size (MB) |
|---|---|---|---|---|---|
| LLaVA | 0.94 | 0.93 | 0.979 | 50 | 130 |
| Electra | 0.85 | 0.84 | 0.962 | 160 | 95 |
| DistilBERT | 0.92 | 0.91 | 0.959 | 95 | 80 |
| T5-small | 0.80 | 0.80 | 0.958 | 25 | 250 |
| GPT-2 | 0.88 | 0.87 | 0.967 | 180 | 360 |
| XLNet | 0.87 | 0.86 | 0.957 | 115 | 475 |
| BERT-base | 0.86 | 0.85 | 0.948 | 185 | 325 |
| ALBERT | 0.83 | 0.82 | 0.963 | 140 | 485 |

**Table 1: Comparison of performance, speed, and model size across various NLP models.**

perception and understanding. SHAP grounds the model's decisions in transparent attribution. SAM ensures spatial precision aligned with model intent. LLaVA translates visual cues into semantically rich descriptions. The result is a human-aligned system that justifies its predictions with both visual and verbal evidence.

Looking ahead, the potential for extending this sys- tem is vast. Future iterations can incorporate real-time processing for video analysis, leverage larger founda- tional models for richer semantics, or integrate user feedback mechanisms for interactive explanation. As AI systems increasingly influence high-stakes decisions, frameworks like this will play a vital role in building trust and accountability.

## REFERENCES


1. Wang, J., Mao, Y., Guan, N., & Xue, C. J. (2024).
2. SHAP-CAT: An interpretable multi-modal framework enhancing WSI classification via virtual staining and Shapley-value-based multimodal fusion. arXiv preprint arXiv:2410.01408.
3. Cafagna, G., Rojas-Barahona, L. M., van Deemter, K., Gatt, A. (2024). Interpreting vision and language generative models with semantic visual priors. Fron- tiers in Artificial Intelligence, 6, 1125632.
4. Parcalabescu, L., Frank, A. (2024). MM-SHAP: A performance-agnostic metric for measuring multimodal contributions in vision and language models tasks. Proceedings of the 61st Annual Meeting of the Associ- ation for Computational Linguistics (ACL), 4032–4059.
5. Zeng, X. (2024). Enhancing the interpretability of SHAP values using large language models. arXiv preprint arXiv:2409.00079.
6. Rao, V. N., Zhen, X., Hovsepian, K., Shen, M. (2021). A First Look: Towards Explainable TextVQA Models via Visual and Textual Explanations. arXiv preprint arXiv:2105.02626.
7. Zhang, Y., Jiang, T., Pan, B., Wang, J., Bai, G., Zhao, L. (2024). MEGL: Multimodal Explanation- Guided Learning. arXiv preprint arXiv:2411.13053.
8. Lyu, Y., Liang, P. P., Deng, Z., Salakhutdinov, R., Morency, L.-P. (2022). DIME: Fine-grained Interpretations of Multimodal Models via Disentangled Local Explanations. arXiv preprint arXiv:2203.02013.