UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO

DAVI MONTEIRO PAIVA

HERE COMES THE SAM: BRINGING LIGHT TO BLACK BOX MODELS APPLIED TO VIDEO CONTENT

Recife
2025

DAVI MONTEIRO PAIVA

HERE COMES THE SAM: BRINGING LIGHT TO BLACK BOX MODELS APPLIED TO VIDEO CONTENT

Tese/Dissertação apresentada ao Programa de Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de bacharel em Ciência da Computação.

Área de Concentração: Ciências da Computação.

Orientador: Francisco Paulo Magalhaes Simoes
Corientadora: Veronica Teichrieb

Recife

2025

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

PAIVA, Davi Monteiro.
Here comes the sam: bringing light to black box models applied to video content / Davi Monteiro PAIVA. - Recife, 2025.
28 : il., tab.

Orientador(a): Francisco Paulo Magalhães Simões
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Informática, Ciências da Computação - Bacharelado, 2025.
Inclui referências, apêndices.

1. IA explicável. 2. Segmentação de vídeo. 3. Explicações agnósticas de modelo. I. Paulo Magalhães Simões, Francisco . (Orientação). II. Título.

000 CDD (22.ed.)

DAVI MONTEIRO PAIVA

HERE COMES THE SAM: BRINGING LIGHT TO BLACK BOX MODELS APPLIED TO VIDEO CONTENT

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de bacharel em Ciência da Computação.

Aprovado em: 31/03/2025

BANCA EXAMINADORA

_____

Prof. Dr. Francisco Paulo Magalhaes Simoes (Orientador)

Universidade Federal de Pernambuco

_____

Prof. Dr. Cleber Zanchettin (Examinador Interno)

Universidade Federal de Pernambuco

Dedico este trabalho especialmente à minha orientadora, Professora Drª. Veronica Teichrieb, cuja força e perseverança diante dos desafios enfrentados têm sido um exemplo inspirador.

AGRADECIMENTOS

Gostaria de expressar minha profunda gratidão ao professor Dr. Ricardo Bastos Cavalcante Prudêncio, por ter me apresentado a área de Inteligência Artificial Explicável e por ter proporcionado a oportunidade inicial de desenvolver o trabalho que mais tarde viria a ser este TCC.

Agradeço imensamente ao professor Dr. João Marcelo Xavier Natário Teixeira, que me orientou com dedicação, oferecendo suporte fundamental na pesquisa, além de incentivar e dar liberdade às minhas ideias, mesmo as mais ousadas.

Agradeço também ao professor Dr. Francisco Paulo Magalhães Simões, por ter aceitado gentilmente representar minha orientadora em sua ausência, garantindo o andamento deste trabalho.

Também deixo meu agradecimento especial ao meu pai, pela orientação constante e pelos conselhos valiosos ao longo da minha trajetória acadêmica. Ao meu irmão, agradeço pela preciosa ajuda na escrita e revisão deste trabalho.

ABSTRACT

This study presents a novel model-agnostic framework aimed at enhancing the explainability of black-box video models by integrating advanced video segmentation techniques. We propose utilizing the Segment Anything Model 2 (SAM) to generate semantically meaningful and spatio-temporally coherent segments, which we subsequently employ within a Local Interpretable Model-agnostic Explanations (LIME)-inspired approach. Our method addresses the inherent limitations of traditional image-based explainability techniques, such as temporal inconsistency and semantic incoherence when applied to video content. By systematically perturbing these meaningful video segments, we develop intuitive and faithful local surrogate explanations that highlight the model's decision-making process clearly and effectively. Experimental evaluations using the Kinetics-400 action recognition dataset demonstrate that our approach produces superior explanations compared to baseline methods, significantly improving interpretability and temporal coherence. The insights provided by this enhanced explainability framework hold particular relevance for critical domains like surveillance, medical diagnostics, and autonomous systems, where understanding model decisions is essential for reliability and user trust.

Keywords: Explainable AI; Video segmentation; Model-Agnostic explanations.

RESUMO

Este trabalho apresenta um método agnóstico ao modelo que melhora a explicabilidade de modelos caixa-preta aplicados a vídeos, integrando técnicas avançadas de segmentação de vídeo. Propomos utilizar o *Segment Anything Model* 2 (SAM) para gerar segmentos espacial e temporalmente coerentes e semanticamente significativos, que são posteriormente utilizados em explicações locais baseadas no método *Local Interpretable Model-agnostic Explanations* (LIME). Ao empregar a segmentação proporcionada pelo SAM, preservamos limites importantes dos objetos e a consistência temporal, proporcionando explicações mais intuitivas e confiáveis. Os resultados experimentais obtidos com o conjunto de dados *Kinetics-400*, voltados para o reconhecimento de ações, mostram que nossa abordagem gera explicações superiores em comparação com métodos tradicionais, aumentando significativamente a interpretabilidade e a coerência temporal. As melhorias na clareza das explicações proporcionadas por esse método são especialmente importantes em áreas críticas, como vigilância, diagnósticos médicos e sistemas autônomos, onde compreender as decisões tomadas pelos modelos é essencial para garantir confiabilidade e confiança por parte dos usuários.

Palavras-chave: IA explicável; Segmentação de vídeo; Explicações agnósticas de modelo.

SUMÁRIO

1 INTRODUCTION

Deep learning models have achieved remarkable success in analyzing video data, excelling in tasks like action recognition, object detection in dynamic scenes, and event classification. However, many of these high-performing models operate as black-boxes, offering little insight into their internal decision making processes. As the reliance on such models grows in critical domains—such as surveillance, medical diagnostics, and autonomous systems—ensuring that their outputs are explainable and trustworthy becomes increasingly important.

Explainable AI (XAI) efforts in vision often focus on static images, utilizing techniques like saliency maps, class activation mappings, or perturbation-based methods to identify image regions that strongly influence a model's prediction. Adapting these methods directly to video is nontrivial. Videos add a temporal dimension and often involve complex, evolving scenes. Naive extension of image-based techniques can yield noisy or temporally inconsistent explanations, ultimately reducing their utility and comprehensibility.

In this study, we propose a novel framework that enhances the explainability of black-box video models by leveraging advanced segmentation techniques. Our approach builds upon Local Interpretable Model-agnostic Explanations (LIME) (RIBEIRO; SINGH; GUESTRIN, 2016) but applies a Segment Anything Model (SAM) (RAVI et al., 2024) to generate coherent spatio-temporal segments that serve as meaningful units of explanation. By integrating SAM-based segmentation, we preserve important object boundaries and temporal consistency, providing explanations that are more intuitive and faithful.

As our main contribution, we: introduce a SAM-based segmentation procedure tailored to generate coherent spatio-temporal segments in video data; adapt a LIME-inspired local surrogate explanation method for videos, utilizing these segments to enhance temporal consistency and interpretability; and demonstrate that improved segmentation results in more faithful and comprehensible explanations.

The remainder of the study is structured as follows. Section 2 reviews related work on explainable AI in images and video. Section 3 details our methodology, describing the SAM-based segmentation and the adaptation of LIME to video. Section 4 outlines the experimental setup, datasets. In Section 5, we present qualitative results. Section 6 discusses insights, limitations, and implications for future research. Finally, Section 7 concludes and highlights potential extensions of our approach.

2 RELATED WORK

Explainable AI methods aim to provide insights into why a model produces certain predictions. By “explaining a prediction,” we mean presenting textual or visual artifacts that provide qualitative understanding of the relationship between the instance’s components (e.g., words in text, patches in an image) and the model’s prediction, as proposed in the LIME paper. Trust in AI can be understood in two dimensions: trusting a prediction—whether a user finds an individual prediction reliable enough to act on it, and trusting a model—whether the user believes the model will behave reasonably when deployed (RIBEIRO; SINGH; GUESTRIN, 2016). We argue that explaining predictions is an important aspect of enabling humans to trust and effectively use machine learning, provided the explanations are faithful and intelligible. In image analysis, popular techniques include saliency maps (SIMONYAN; VEDALDI; ZISSERMAN, 2013), Grad-CAM (SELVARAJU, et al., 2020), and perturbationbased methods like LIME. These methods highlight pixels or regions of interest but were initially designed for static images, where temporal coherence is not a concern.

When extending explanations to video, methods must consider both spatial and temporal dimensions. Prior work often adapts image-based techniques frame-by-frame, potentially leading to inconsistent explanations over time (RAVI et al., 2024).

Although video.explainability is a highly relevant area, it remains significantly underexplored. Some approaches focus on spatio-temporal saliency estimation (WANG et al., 2016), while others attempt to adapt perturbation strategies by sampling multiple frames (ROY et al., 2023). However, these methods often struggle to maintain coherent units of explanation that map onto meaningful objects or events. Our work builds upon REVEX: A Unified Framework for Removal-Based Explainable Artificial Intelligence in Video (GAYA-MOREY; RUBIO; MANRESA-YEE, 2016), which provides a robust foundation for architecture independent video explanations. By leveraging the concepts introduced in REVEX, we extend the framework to develop model-agnostic explanations that address the limitations of prior methods and ensure coherent, meaningful interpretations across spatial and temporal dimensions.

In image contexts, segment-based perturbation methods—such as LIME—rely on superpixels or segmented regions to generate local surrogates. Good segmentation is crucial for coherent explanations. Extending this concept to video requires the use of supervoxels, which represent spatio-temporal regions that group pixels across both spatial

and temporal dimensions. Supervoxels provide a natural way to create stable and interpretable explanation units that can track objects or actions over time.

High-quality supervoxels play a critical role in simplifying the task for the surrogate model by enabling it to more accurately predict the importance of each supervoxel cluster. Moreover, meaningful supervoxel segmentation helps users clearly identify which regions of the video are most relevant to the model's predictions, improving the interpretability and usability of explanations. By creating temporally consistent and spatially meaningful clusters, good supervoxels enhance both the computational and visual clarity of video-based explanations.

The Simple Linear Iterative Clustering (SLIC) algorithm (ACHANTA et al., 2012) has been a popular choice for generating superpixels in image analysis due to its computational efficiency and ability to maintain temporal consistency. SLIC operates by adapting the k-means clustering algorithm to work in a combined space of color and spatial coordinates, creating compact and nearly uniform supervoxels. For video applications, SLIC extends this approach to include the temporal dimension, clustering pixels based on their color similarity and spatio-temporal proximity.

While SLIC provides a reasonable baseline for video segmentation, it has limitations. The algorithm relies heavily on low-level features (color and position) and can sometimes fail to capture semantic object boundaries, especially in complex scenes with varying lighting conditions or motion.

Our approach utilizes SAM 2, the successor to SAM specifically designed for video segmentation, to generate what are referred to as masklets. A masklet, a concept introduced in the SAM 2 paper (RAVI et al., 2024), represents a spatio-temporal mask that tracks an object or region of interest across multiple frames in a video. For our purposes, a masklet serves a similar function to a supervoxel, we will use these terms interchangeably throughout the text.

Unlike SLIC, which relies on a geometric clustering technique, SAM 2 employs a deep learning-based method that comprehends semantic content and object relationships within the video. This enables SAM 2 to produce segmentations that more closely align with human perception and accurately capture object boundaries. By leveraging SAM 2, we achieve robust and consistent segmentation across frames, facilitating the creation of meaningful, temporally coherent explanation units that reflect the dynamic nature of video data. Moreover, the semantic understanding embedded in SAM 2 allows for more intuitive and interpretable explanations compared to the purely geometric approach used by SLIC.

LIME introduced the concept of generating local surrogate models around a given instance to explain the predictions of any black-box model. It works by perturbing the input data and observing how the model's predictions change. For each instance to be explained, it creates a set of perturbed samples by randomly removing or modifying features, then trains an interpretable linear model on this data, the linear model will try to learn the behavior of the black-box model only in this instance. The weights of this linear model reveal which features were most important for the original prediction.

While its application to video presents unique challenges, LIME remains a versatile tool for model-agnostic explanations, allowing compatibility with any classification or detection algorithm and maximizing its applicability. The removal based approach in LIME is particularly suitable for video analysis as it allows us to understand which spatial-temporal regions most influence the model's decision by systematically removing them and measuring the impact on the prediction.

LIME has already demonstrated superior performance compared to other methods in removal-based explanation tasks, making it the preferred algorithm for this approach (GAYA-MOREY; RUBIO; MANRESA-YEE, 2016). The removal-based methodology provides intuitive explanations by identifying which parts of the video, when removed, most significantly affect the model's output. This approach is more interpretable than attribution-based methods as it directly shows the causal relationship between video regions and predictions.

Building on REVEX, we adopt LIME as our core removal based explanation algorithm. By integrating LIME with SAMbased segmentation, we extend its functionality to address the added spatial and temporal complexities of video data, enabling robust and interpretable explanations. This combination allows for more precise and semantically meaningful perturbations, as SAM provides high-quality segmentation masks that can be used to remove coherent objects or regions.

3 METHODOLOGY

We consider a black-box video model f that takes as input a video V = {frame $_1$, frame $_2$, … ,frame $_T$} consisting of T frames. The model produces a prediction $y = f(V)$, where y can represent a class label (e.g., for action recognition) or a set of bounding boxes and classes (e.g., for object detection).

To analyze f, we leverage SAM 2 to generate a segmentation map S with the same shape as V. Each pixel in V is assigned a number in S, representing the supervoxel to which the pixel belongs. The segmentation map S consists of N supervoxels s, where each supervoxel groups together spatiotemporally coherent regions of the video.

A perturbation set is a copy of S and we create N perturbation sets by randomly perturbing the supervoxels. For each perturbation set, each supervoxel s has a 50% probability of being "removed." This process generates diverse perturbations of V, allowing us to train an explainable model that can better infer the importance of different supervoxels.

The objective is to explain the prediction y by identifying the supervoxels $s \in S$ that most significantly influence the decision of f.

The key idea is to generate spatio-temporally coherent segments that capture semantically meaningful entities (objects, actions, or events) throughout the video. SAM (RAVI et al., 2024) provides a robust segmentation backbone designed primarily for image data. In our framework, we leverage SAM 2, which supports video processing by utilizing SAM's image segmentation capabilities to create automatic mask encodings for individual video frames and aggregating and aligning these frame-level encodings to generate temporally consistent spatio-temporal segments.

While SAM 2 effectively segments many regions, it does not automatically assign every pixel to a cluster. To address this, we include all unassigned pixels in an additional cluster, ensuring complete coverage of the video frames.

To enhance the quality of explanations, we carefully tune SAM 2's parameters, optimizing them to produce segments that better align with meaningful objects, actions, or events in the video. The specific parameter values and tuning process are provided in the appendix for reproducibility and further exploration. This process yields a set of N+1 spatio-temporal segments $S = \{s_1, s_2, \ldots, s_N\}$, with an extra segment s_{N+1} containing the unassigned pixels. Each segment ideally represents a coherent object or activity,

maintaining spatio-temporal consistency across frames while ensuring no pixel is excluded from analysis.

We adapt the LIME framework for explaining video model predictions by employing a perturbation approach similar to that introduced by REVEX. Specifically, we simulate the removal of selected regions by masking their corresponding pixels in video frames with black color. This perturbation strategy maintains methodological consistency with REVEX. Alternative masking techniques exist but may unintentionally introduce out-of-domain data due to changes in pixel color distributions. REVEX identifies additional masking strategies, including median colors, grayscale, blurring, and up-scaled black. Future research could explore the comparative effects of these alternative perturbation techniques on the interpretability and accuracy of our video model.

To estimate the importance of each region, we employ a linear model to fit the black-box model's predictions using the perturbed data. The linear model assigns weights to regions, representing their contributions to the prediction.

To account for the temporal aspect of videos, we adapt the segmentation process to generate spatiotemporal superpixels. For this, we use both SAM 2 and SLIC as segmentation algorithms, enabling a comparative evaluation of their impact on the explanation results. SAM 2 generates supervoxels by grouping spatiotemporally coherent regions across frames, while SLIC creates spatially contiguous regions within individual frames. By considering both methods, we aim to provide insights into the effectiveness of these approaches for videospecific explainability. Our approach does not require access to model internals (weights, activations) and can be applied to any type of video model—ranging from CNN-based classifiers to transformer-based detectors. As long as we can query the model with perturbed video inputs and obtain predictions, we can produce explanations.

4 EXPERIMENTAL SETUP

We evaluate our approach using the Kinetics-400 dataset (KAY et al., 2017), a benchmark for action recognition tasks. This dataset consists of diverse videos spanning 400 human activity classes, offering a comprehensive testbed for assessing our method's performance.

We conduct our experiments using the Swing3D T model (LIU et al., 2022), a transformer-based architecture designed for spatiotemporal action recognition. The model leverages attention mechanisms to capture both spatial and temporal features, making it particularly suited for tasks requiring fine-grained motion analysis.

SAM 2 is a computationally heavy model, which posed challenges during our experiments. Despite utilizing a powerful consumer GPU (NVIDIA RTX 4090), processing a single video containing 300 frames often takes hours. This extended runtime highlights the demanding nature of generating spatiotemporally coherent segments with SAM 2.

For the initialization of spatio-temporal segments, we also employed SAM 2 in its image mode to generate mask auto-encodings for individual frames. While this process is relatively faster compared to video processing, it required careful tuning of parameters. Finding the optimal parameters was challenging, as they needed to maximize the utility of generated masks—capturing meaningful regions—without consuming excessive storage or computational resources.

These computational considerations and parameter tuning processes were critical to achieving high-quality results, and we provide detailed configurations and guidelines in the appendix to facilitate reproducibility and github repo.

At the time these experiments were conducted, there was no official implementation of LIME for video data. Consequently, we developed our own implementation tailored for video analysis. To streamline the segmentation process, we utilized the slic implementation from the skimage library, which supports 3D data and is optimized for efficiency. This allowed us to generate super-voxels effectively, ensuring compatibility with our LIME-based framework.

For our implementation, we utilized the SAM 2.1 tiny version for both image and video mask generation. The automatic image mask generator was configured with the following parameters:

```
“mask_generator = SAM2AutomaticMaskGenerator( model=sam2,
pred_iou_thresh=0.7,
stability_score_thresh=0.5,
stability_score_offset=0.5,
crop_n_layers=1,
box_nms_thresh=0.7,
crop_n_points_downscale_factor=2,
min_mask_region_area=25.0,
use_m2m=True,
)”
```

This configuration was chosen by trying different combinations to try and balance segmentation quality and cover all possible areas of image, without leaving a gap of background. The SAM 2.1 tiny model was selected for its reduced computational requirements while still providing sufficient segmentation quality for our explanation framework.

5 RESULTS AND ANALYSIS

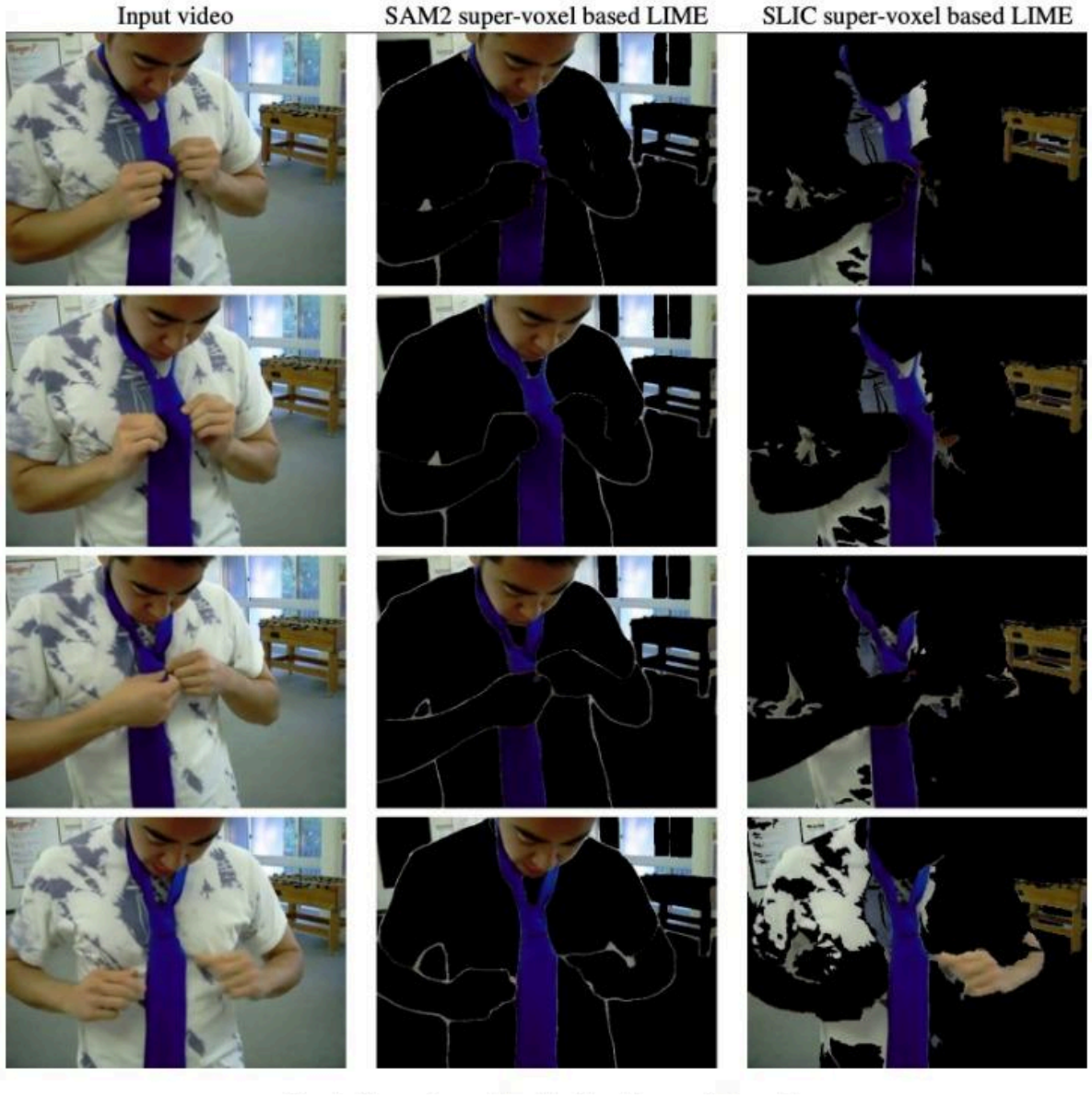
To provide interpretable visual feedback of our model's decision-making process, we implemented a visualization scheme that highlights the most significant regions identified by our adapted LIME algorithm. Following the established visualization approach used in image-based LIME explanations, we represent the importance of different video segments through a selective masking process.

Our visualization method ranks all spatio-temporal segments according to their importance scores derived from the LIME surrogate model. We establish a threshold at the 80th percentile of these scores, then mask (set to black) all pixels belonging to segments with importance scores below this threshold.

This approach creates a clear visual distinction between regions the model considers crucial for its prediction (which remain visible) and less important regions (which are masked in black). By maintaining the original appearance of the most influential segments while obscuring the less relevant ones, we provide an intuitive visualization that allows users to directly observe which parts of the video most strongly influenced the model's decision.

This visualization technique effectively communicates the model's focus areas while maintaining temporal consistency across frames, as segments are evaluated and masked based on their importance across the entire temporal sequence rather than frame by frame.

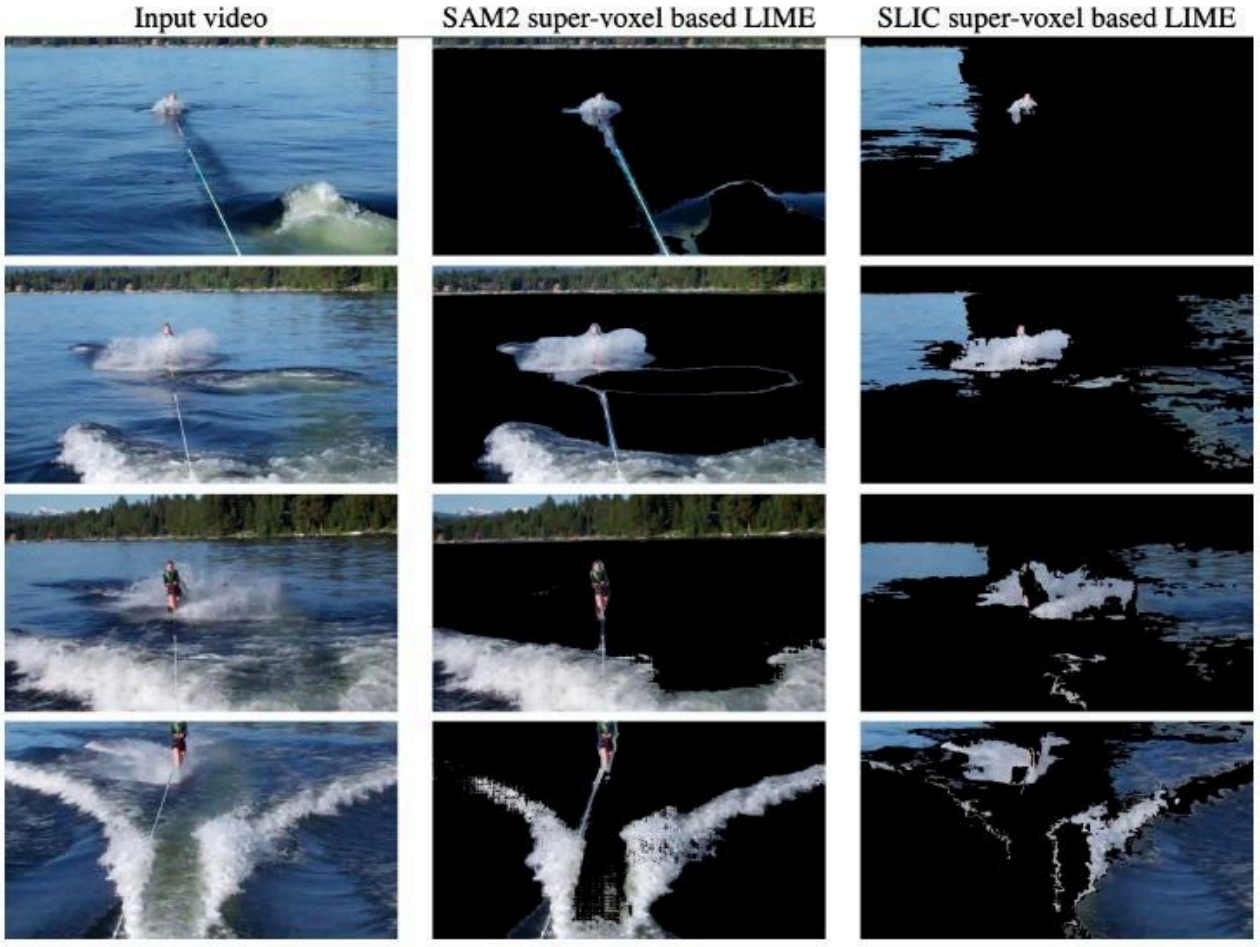
Figure 1 – The predicted action in this video is "tying tie." Notice how SAM2 effectively segments the shirt into a single coherent cluster, and how accurately the tie is segmented compared to the results obtained by SLIC.



Source: Author (2025).

Figure 1 presents a sample video frame from the KinectTest dataset, comparing the original video, SAM 2-based segmentation, and SLIC-based segmentation methods. It is evident that SAM 2 excels at creating semantically meaningful clusters, whereas SLIC struggles to segment entire objects consistently. SLIC often leaves parts of objects unsegmented or fails to form coherent superclusters, particularly during camera movement. We refer to this issue as “cluster collapse”, a problem observed in both techniques but significantly more pronounced in SLIC. In contrast, SAM 2 successfully highlights important elements with minimal interference from surrounding regions. For instance, the tie in the image is clearly identified as a key region by SAM 2, demonstrating its ability to isolate significant features while maintaining cluster integrity.

Figure 2 – The predicted action in this video is "water skiing." SAM2 effectively segmented the water, accurately distinguishing even small objects such as the rope pulling the skier.



Source: Author (2025).

Figure 2 demonstrates SAM's superior ability to segment challenging objects, such as separating water from the waves, delineating the horizon line, and isolating the person in the frame. For color-based algorithms like SLIC, dynamic objects such as water pose significant challenges due to continuous changes in color and texture. SAM 2, leveraging its semantic understanding, handles such scenarios more effectively, maintaining consistent object boundaries even in visually complex scenes. Both images are displayed in large formats in the appendix.

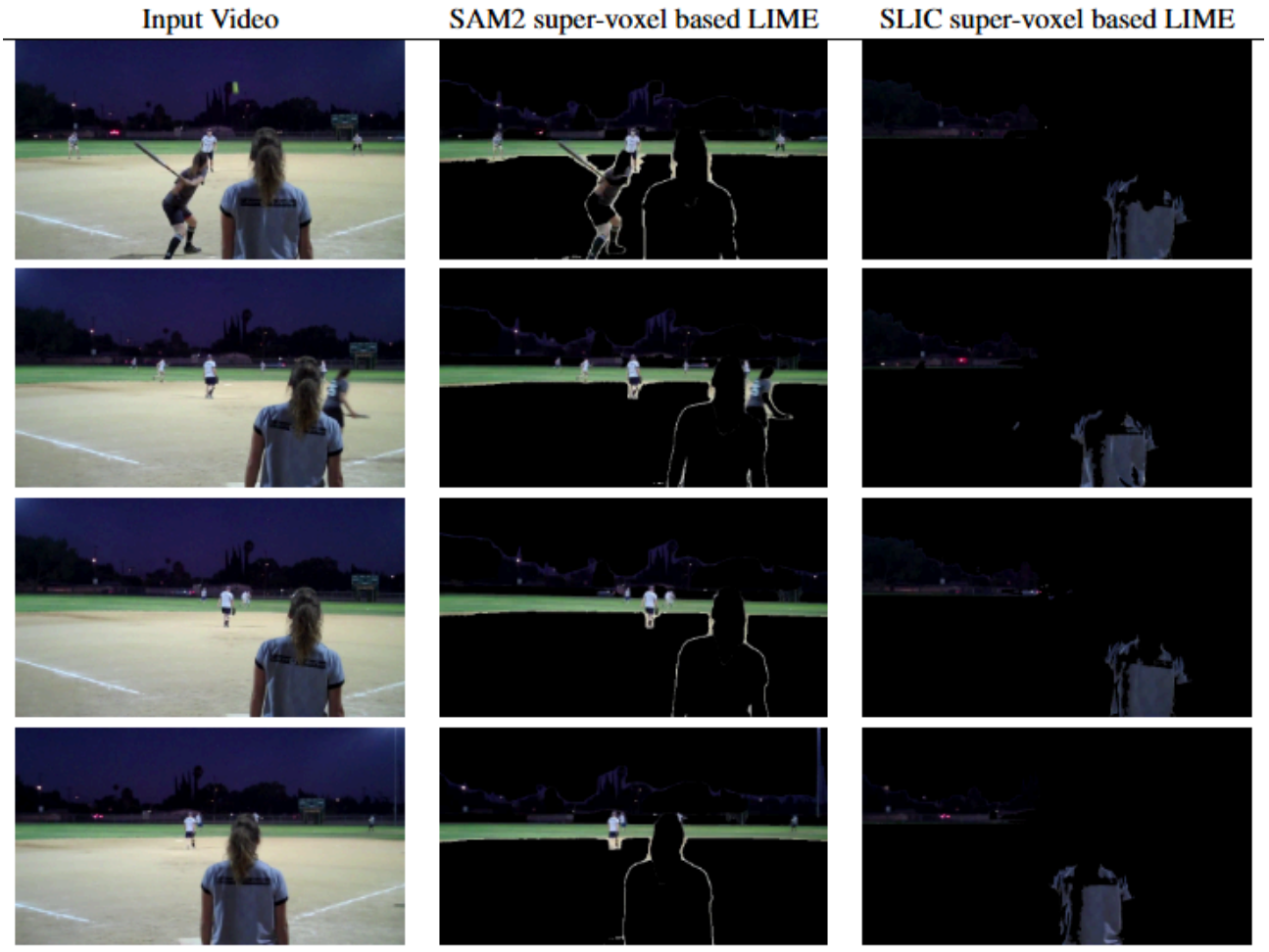
By comparing these examples, it becomes clear that SAM's ability to account for both semantic and spatial coherence offers a significant advantage in scenarios where SLIC struggles to adapt, particularly in environments with rapid changes or complex textures.

Two parameters in SAM 2 can be adjusted, which directly influence the resulting cluster size. Ideally, clusters should not be unnecessarily large; however, they should

correspond appropriately to the size of objects in the scene. Often, an object of interest may be an entire person, though in some cases, better interpretability is achieved by subdividing a person into separate components, as illustrated in Figure 1. Excessively large clusters may disproportionately influence predictions due to their size rather than their actual relevance to the object being predicted. It is important to note that clusters cannot be removed entirely from an image—only their color can be altered, for instance, to black—thus, their presence remains relevant to the overall interpretation.

Figure 3 – The predicted action in this video is "riding scooter." In this comparison, neither method shows a clear advantage; results appear subjective. However, SAM2 generates well-defined boundaries around clusters, whereas the segmentation produced by SLIC appears more concentrated.

Figure 4 – This figure demonstrates the superior segmentation performance of SAM2, which correctly identifies the entire field as a single coherent object rather than dividing it into multiple segments, leading to improved interpretability.



6 DISCUSSION

Our results show that integrating a robust segmentation model like SAM into the explanation pipeline significantly improves the quality and coherence of local surrogate explanations. By treating stable spatio-temporal segments as perturbation units, we produce explanations that better reflect object boundaries and maintain consistency over time.

This model-agnostic approach is particularly valuable in complex video analysis scenarios. As new architectures and tasks emerge, having a flexible, generalizable explanation method ensures that insights into model behavior remain accessible. Nonetheless, this flexibility comes with added computational overhead and complexity, especially for long, high-resolution videos.

While we focused on improving the structural coherence of explanations, there remains room for incorporating semantic priors, leveraging motion cues more explicitly, or integrating audio-visual modalities. Addressing these challenges can further enhance explainability in future work.

Understanding how deep learning models make decisions in video analysis is crucial for both technical advancement and responsible AI deployment. Our approach to explainable video analysis has several key implications for the field.

From a development perspective, the ability to visualize and understand model decisions enables researchers and engineers to identify potential weaknesses or biases in their models. This insight is invaluable for iterative improvement of video analysis systems, helping create more robust and reliable models. When models make incorrect predictions, our explanation method can reveal whether the error stems from focusing on irrelevant features or missing crucial information in the video sequence.

Fairness in AI systems is becoming increasingly critical as these technologies impact more aspects of society. Video analysis systems can inadvertently perpetuate or amplify existing societal biases, particularly in applications like security surveillance, job interview analysis, or behavior monitoring. Our explanation method provides a crucial tool for fairness auditing by revealing whether models disproportionately focus on sensitive attributes like skin color, gender-specific features, or cultural elements. This transparency enables developers and stakeholders to identify and address potential discriminatory patterns in model behavior before deployment.

In terms of accountability, explainability becomes particularly crucial in sensitive applications such as surveillance, medical diagnosis, or autonomous vehicle systems. When these systems make critical decisions, stakeholders need to understand the reasoning behind these choices. Our method provides a transparent way to audit model decisions, helping identify potential biases or systematic errors that could lead to unfair treatment of certain groups or dangerous failures in critical situations.

This work contributes to the broader goal of creating more transparent, fair, and accountable AI systems, particularly in the complex domain of video analysis where traditional explanation methods may fall short.

7 CONCLUSION

We presented a novel, model-agnostic approach to explaining black-box video models by combining advanced segmentation methods with a LIME-inspired framework. By leveraging the Segment Anything Model, we enhanced spatio-temporal coherence, resulting in explanations that align more closely with the behavior of the underlying model while maintaining human interpretability.

For future work, we recognize several avenues to further improve and expand this approach. One direction involves experimenting with explanations beyond LIME 3D. While LIME 3D provides a straightforward framework for generating local surrogate explanations, its reliance on linear approximation limits its capacity to capture complex decision boundaries. Exploring other paradigms, such as RISE (PETSIUK; DAS; SAENKO, 2018) or SHAP (KAY et al., 2017), could yield more nuanced insights, particularly for intricate video models. These methods offer the potential to better capture relationships within the data and reveal alternative perspectives on model behavior that extend beyond local fidelity.

Another promising avenue lies in post-processing existing techniques to enhance the quality of segmentation and clustering outputs. Addressing artifacts or inconsistencies in these methods can significantly improve the interpretability and coherence of the generated explanations. For instance, morphological operations or smoothing filters could refine segment boundaries, while semantic-based strategies for merging or splitting clusters could produce explanation units that are both computationally efficient and visually meaningful. Additionally, we propose enabling user-defined segmentation for the first frame of the video as a way to guide the segmentation process throughout the sequence. By allowing users to highlight what they consider the most important regions or objects in the initial frame, the method can propagate this guidance across the video, aligning the segmentation with human-defined priorities and improving the relevance of the resulting explanations.

We also aim to test our approach on a broader range of models and datasets to evaluate its robustness and generalizability. This includes applying the method to diverse video models, such as transformer-based architectures, convolutional networks, vision-language models (VLM), and hybrid systems. Expanding our evaluation across datasets from various domains—including medical imaging, surveillance, sports analysis, and autonomous driving—will help us identify domainspecific challenges and validate the

applicability of our method in real-world scenarios. By addressing these dimensions, we seek to push the boundaries of explainable AI in video analysis, ensuring its relevance and utility across a variety of contexts and applications.

REFERENCES

ACHANTA, R. et al. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 34, n. 11, p. 2274–2282, nov. 2012.

GAYA-MOREY, F. Xavier; BUADES-RUBIO, Jose M.; MANRESA-YEE, Cristina. Local Agnostic Video Explanations: a Study on the Applicability of Removal-Based Explanations to Video. **arXiv preprint arXiv:2401.11796**, 2024. Available: https://arxiv.org/abs/2401. 11796. Acess: 13 mar. 2025

KAY, Will et al. The kinetics human action video dataset. **arXiv preprint arXiv:1705.06950**, 2017. Available: https://arxiv.org/abs/1705. 06950. Acess: 13 mar. 2025

LIU, Ze et al. Video swin transformer. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. 2022. p. 3202-3211. Available:https://arxiv.org/abs/2106.13230. Acess: 13 mar. 2025

LUNDBERG, Scott M.; LEE, Su-In. A unified approach to interpreting model predictions. **Advances in neural information processing systems**, v. 30, 2017. Available:https://proceedings.neurips.cc/paperfiles/paper/2017/file8a20a8621978632d76c43dfd28b67767-Paper.pdf. Acess: 13 mar. 2025

PETSIUK, Vitali; DAS, Abir; SAENKO, Kate. Rise: Randomized input sampling for explanation of black-box models. **arXiv preprint arXiv:1806.07421**, 2018. Available:https://arxiv.org/abs/1806.07421. Acess: 13 mar. 2025

RAVI, N. et al. SAM 2: Segment Anything in Images and Videos. **arXiv (Cornell University)**, 1 ago. 2024. Available: https://arxiv.org/abs/2408. 00714. Acess: 13 mar. 2025

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16**, p. 1135–1144, 2016. Available: https://arxiv.org/abs/1602. 04938. Acess: 13 mar. 2025

ROY, C. et al. Explainable Activity Recognition in Videos using Deep Learning and Tractable Probabilistic Models. **ACM Transactions on Interactive Intelligent Systems**, v. 13, n. 4, p. 1–32, 8 dez. 2023.

SELVARAJU, R. R. et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. **International Journal of Computer Vision**, v. 128, n. 2, p. 336–359, 1 fev. 2020. Available: http://dx. doi.org/10.1007/s11263-019-01228-7. Acess: 13 mar. 2025

SIMONYAN, Karen; VEDALDI, Andrea; ZISSERMAN, Andrew. Deep inside convolutional networks: Visualising image classification models and saliency maps. **arXiv preprint arXiv:1312.6034**, 2013. Available: https://arxiv.org/abs/1312.6034 Acess: 13 mar. 2025

WANG, L. et al. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. **Computer Vision – ECCV 2016**, p. 20–36, 2016. Available:https://arxiv.org/abs/1608.00859. Acess: 13 mar. 2025