Here comes the SAM: bringing light to black box models applied to video content

Davi Monteiro Paiva
UFPE
Recife, Brasil
vico.paiva@gmail.com

Joao Marcelo Xavier Natario Teixeira
UFPE
Recife, Brasil
jmxnt@cin.ufpe.br

Veronica Teichrieb
UFPE
Recife, Brasil
vt@cin.ufpe.br

***Abstract*—This paper introduces a model-agnostic approach to improving explainability in black-box video models by integrating advanced segmentation techniques. Leveraging the Segment Anything Model 2 (SAM) to create coherent spatio-temporal segments, we adapt a LIME-inspired framework to generate more intuitive local surrogate explanations. Our method allows for the extraction of meaningful regions within video frames, providing clearer insights into the model's decision-making process. Experimental results demonstrate that employing better segmentation leads to more faithful and interpretable explanations, highlighting the benefits of this generalizable strategy for a wide range of video-based classification and detection tasks.**

***Keywords*—Explainable Artificial Intelligence; Video Segmentation; Model-Agnostic Explanations**

I. 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. Naïve extension of image-based techniques can yield noisy or temporally inconsistent explanations, ultimately reducing their utility and comprehensibility.

In this paper, 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) [1] but applies a Segment Anything Model (SAM) [2] 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.

Our main contributions are:

- We introduce a SAM-based segmentation procedure tailored to generating coherent spatio-temporal segments in video data.
- We adapt a LIME-inspired local surrogate explanation method to video and by using these segments, thereby improving temporal consistency and interpretability.
- We demonstrate that better segmentation leads to more faithful and comprehensible explanations.

II. Related Work

A. Video-Based Explainability

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 [2].

Although video explainability is a highly relevant area, it remains significantly underexplored. Some approaches focus on spatio-temporal saliency estimation [3], while others attempt to adapt perturbation strategies by sampling multiple frames [4]. 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 [5], 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.

B. Segmentation and Superpixels in Explanations

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 [6] 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 [2], 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.

C. LIME and Model-Agnostic Approaches

LIME introduced the concept of generating local surrogate models around a given instance to explain the predictions of any black-box model. LIME 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 [5]. 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 SAM-based 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.

III. Methodology

A. Problem Formulation

We consider a black-box video model f that takes as input a video $V = \{\text{frame}_1, \text{frame}_2, \ldots, \text{frame}_T\}$ 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.

B. Segment Anything Model 2

The key idea is to generate spatio-temporally coherent segments that capture semantically meaningful entities—such as objects, actions, or events—throughout the video. SAM [2] provides a robust segmentation backbone originally designed for image data. In our framework, we extend this by using SAM 2, which supports video processing by first applying SAM's image segmentation capabilities to generate automatic mask encodings for each video frame. These frame-level encodings are then aggregated and temporally aligned to produce consistent spatio-temporal segments across the video sequence.

While SAM 2 effectively segments many regions, it does not guarantee that every pixel is assigned to a cluster. To ensure complete spatial coverage, we assign all unsegmented pixels to an additional cluster, preserving the integrity of the entire frame for downstream analysis.

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 $S = s_1, s_2, ..., s_N$, 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.

C. Adapting LIME to Video Segments

We adapt the LIME framework to explain video model predictions by following a similar approach to REVEX. Specifically, we simulate the removal of selected regions by making the corresponding pixels in the video black, effectively masking out the content. This ensures consistency with REVEX's methodology for perturbation.

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 video-specific 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.

IV. Experimental Setup

Quantifying the notion of explanation remains a complex challenge, as the very act of explaining is inherently abstract and subjective. Nevertheless, to obtain a measurable sense of how informative a model's predictions are, we adopt the Deletion and Preservation games introduced in the REVEX framework. These games offer a structured way to evaluate the contribution of individual spatiotemporal regions to the final prediction.

In our experiments, we apply this evaluation strategy to three distinct action recognition models, providing a diverse benchmark for explanation quality.

To perform the segmentation needed for these games, we integrate the SAM2 segmenter into the REVEX test pipeline and compare its results against the LIME segmenter originally used in the REVEX paper. All evaluations are conducted using the Kinetics-400 dataset.

A. Datasets

We evaluate our approach on the Kinetics-400 dataset [7], a widely-used benchmark for human action recognition. This dataset includes a broad range of video clips covering 400 distinct human activity classes, offering a diverse and challenging testbed for assessing the robustness and generalizability of both the models and their corresponding explanations. For our evaluation, we randomly selected 30 videos from the test set that were correctly classified by all three models. This

filtering ensures that our analysis of explanation quality is not confounded by model misclassifications, and focuses solely on the interpretability of accurate predictions.

B. Models Under Test

To assess the quality of explanations across different model architectures, we evaluate three state-of-the-art networks:

TimeSformer [8], a space-time attention-based transformer that factors spatial and temporal attention separately for scalable video understanding.

TPN (Temporal Pyramid Network) [9], which leverages multi-scale temporal features through a top-down pathway and lateral connections for robust temporal modeling.

TANet (Temporal Aggregation Network) [10], a model that introduces hierarchical temporal aggregation blocks to capture both short- and long-term dependencies.

These models were selected from the MMAction2 library, following the same protocol used in the original REVEX paper. This choice ensures consistency with prior work while introducing architectural variability across the evaluated models.

C. Implementation Details

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 took several hours. To mitigate this and accelerate testing, we applied a temporal stride of 2, effectively dropping every other frame during segmentation. This reduction in temporal resolution helped strike a balance between processing time and explanation quality, while preserving the overall semantic structure of the video.

For the initialization of spatio-temporal segments, we also employed SAM 2 in its image mode to generate mask autoencodings 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.

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(
    points_per_side = 64,
    crop_n_layers   = 1,
    crop_n_points_downscale_factor = 2,
    pred_iou_thresh        = 0.85,
    stability_score_thresh = 0.85,
    min_mask_region_area   = 0,
    box_nms_thresh  = 0.65,
    crop_nms_thresh = 0.65,
    points_per_batch = 128,
)
```

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.

V. Results and Analysis

All experiments were implemented using the MMAction2 framework, with models initialized from pre-trained weights. We adopted the REVEX evaluation procedure, applying the *Deletion* and *Insertion* metrics to quantify the effectiveness of the generated explanations. The SAM2 and LIME segmenters were used to mask regions of interest and measure the impact of feature perturbation on model predictions.

Table I
Average Insertion and Deletion scores for each evaluated model using SAM2 and SLIC.

Model	Insertion Score (↑)		Deletion Score (↓)	
	SAM2	SLIC	SAM2	SLIC
TimeSformer	0.75	0.76	0.54	0.48
TPN	0.59	0.59	0.39	0.32
TANet	0.67	0.70	0.46	0.40

While Table I provides a quantitative summary of model performance in terms of Insertion and Deletion scores, it is important to note that these metrics do not fully capture the qualitative differences in explanation strategies. Specifically, our approach—based on object-level segmentations—tends to preserve and highlight entire objects that contribute to the model's prediction. In contrast, methods like the original LIME segmenter often focus on selecting disconnected parts or patches of an object, which may not align with semantically meaningful units.

This distinction means that our explanations may appear less optimized according to the Insertion/Deletion curves, which favor fine-grained perturbations over holistic object-level understanding. Consequently, the numerical scores in the table should be interpreted with caution: they are useful for comparing general trends, but they may undervalue explanation

methods that prioritize coherence and semantic completeness over granular feature attribution.

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 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.

This approach creates a clear visual distinction between regions the model considers crucial for its prediction (which remain visible) and less important regions. 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. Left is using SAM segmentation and right is using SLIC. The action predicted in this video was tap dancing

Figure 1 shows a frame from the Kinetics-400 test set comparing the explanations produced using SAM and SLIC segmentations for a correctly classified tap dancing video. The difference in segmentation granularity is clearly illustrated: while SLIC focuses on a smaller, more localized region—highlighting only the dancer's shoes—SAM successfully segments the entire dancer. This demonstrates a key strength of SAM-based explanations: they preserve the semantic unity of the object responsible for the action, rather than isolating disjointed or overly specific parts. Although SLIC's segmentation may appear more focused, it fails to capture the holistic context of the action, which in this case involves coordinated full-body movement. By segmenting the complete subject, SAM offers explanations that are not only more interpretable to humans but also more faithful to the actual decision-making process of the model.

Figure 2. Left is using SAM segmentation and right is using SLIC. The action predicted in this video was making sushi

Figure 2 presents another example from the Kinetics-400 dataset, illustrating the segmentation-based explanations for a video correctly classified as making sushi. In this case, the SAM-based explanation highlights the entire plate of food—the central object associated with the predicted action—while the SLIC-based explanation selects only small, scattered regions of the plate. Although SLIC produces a more narrowly focused mask, it fails to capture the full extent of the relevant object, which limits the semantic coherence of the explanation. SAM, on the other hand, succeeds in isolating the complete plate, offering a more intuitive and human-understandable rationale for the model's decision. This again underscores SAM's strength in generating holistic, object-level explanations that align better with how humans perceive meaningful visual elements.

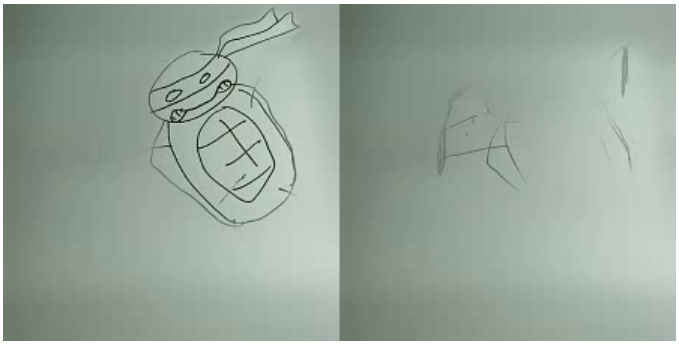

Figure 3. Left is using SAM segmentation and right is using SLIC. The action predicted in this video was drawing

Figure 3 illustrates a case where the SAM-based segmentation successfully identifies the entire children's drawing as a coherent, unified object, forming a single cluster that aligns with the semantic focus of the action. In contrast, the SLIC-based

explanation fragments the drawing into multiple disjointed regions, failing to preserve its visual and conceptual integrity. This example highlights SAM's superior ability to capture complex, irregularly shaped objects as meaningful wholes—an essential trait for generating human-interpretable explanations in tasks involving fine-grained visual context.

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.

VI. 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.

A notable limitation we observed during experimentation is the occurrence of what we term cluster collapse. This phenomenon typically arises when there is rapid camera movement or abrupt scene transitions, causing segmentation quality to degrade across frames. In such cases, the segments lose temporal consistency and may fragment or fail to track meaningful objects, thereby weakening the explanation's interpretability. A promising direction for future work is the introduction of dynamic cluster tracking mechanisms. For instance, maintaining a per-frame cluster count and triggering a re-segmentation process when a collapse is detected could help restore semantic coherence in challenging video conditions.

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.

VII. 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 [11] or SHAP [12], 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. 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 will help us identify domain-specific 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

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," 2016. [Online]. Available: https://arxiv.org/abs/1602.04938

[2] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "Sam 2: Segment anything in images and videos," 2024. [Online]. Available: https://arxiv.org/abs/2408.00714

[3] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool, "Temporal segment networks: Towards good practices for deep action recognition," 2016. [Online]. Available: https://arxiv.org/abs/1608.00859

[4] C. Roy, M. Nourani, S. Arya, M. Shanbhag, T. Rahman, E. D. Ragan, N. Ruozzi, and V. Gogate, "Explainable activity recognition in videos using deep learning and tractable probabilistic models," *ACM Transactions on Interactive Intelligent Systems*, vol. 13, no. 4, pp. 1–32, 2023.

[5] F. X. Gaya-Morey, J. M. Buades-Rubio, I. S. MacKenzie, and C. Manresa-Yee, "Revex: A unified framework for removal-based explainable artificial intelligence in video," 2024. [Online]. Available: https://arxiv.org/abs/2401.11796

[6] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels," *Technical report, EPFL*, 06 2010.

[7] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017. [Online]. Available: https://arxiv.org/abs/1705.06950

[8] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" 2021. [Online]. Available: https://arxiv.org/abs/2102.05095

[9] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou, "Temporal pyramid network for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[10] Z. Liu, L. Wang, W. Wu, C. Qian, and T. Lu, "Tam: Temporal adaptive module for video recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 13 708–13 718.

[11] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," 2018. [Online]. Available: https://arxiv.org/abs/1806.07421

[12] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf