It makes sense! Exploring user preferences for AI explanations on video

Francesc Xavier Gaya-Morey[1,*], Jose Maria Buades-Rubio[1], Ian Scott MacKenzie[2] and Cristina Manresa-Yee[1]

[1]Universitat de les Illes Balears, Carretera de Valldemossa, km 7.5, 07122, Palma, Illes Balears, Spain
[2]York University, 4700 Keele St, North York, ON M3J 1P3, Canada

Abstract

Many explainable artificial intelligence methods exist; however, there is a lack of user evaluations on explainability or trustworthiness. Consequently, it remains unclear which XAI methods are appropriate based on users and their preferences. We present an evaluation with expert users of six video removal-based XAI methods applied across three networks and two datasets. Experts consistently preferred the video-adapted RISE method, while identifying the video-adapted univariate predictors method as the least preferred. These findings provide insight for researchers and practitioners on the preferred XAI methods to use with videos, while also expanding the understanding of XAI methods from a human perspective.

Keywords

explainable artificial intelligence, evaluation, human-centered XAI, video-based XAI methods

1. Introduction

The increasing presence of AI-driven systems across many domains [1] highlights the need for explainability. Since the appearance of deep learning, numerous explainable AI (XAI) methods have emerged with the goal of explaining the AI rationale to a human user [2, 3] .

While developing novel XAI methods is important, research is needed on human-centered XAI (HCXAI) since explanations must align with specific users and in different contexts. That is, integrating the human factor into the research and development of XAI [4] is paramount.

Although social sciences routinely explore human explanation processes, research in XAI often relies on researchers' intuition for a "good" explanation [5] and overlooks human understanding, recipient profiles, and contextual factors around the explanation. On the one hand, recent studies highlight a gap, revealing under-utilization of human-centered methods from human-computer interaction (HCI) in the design of XAI systems [6, 7, 8]. Research shows that human-centered approaches in XAI are relevant for user-driven technical choices, identifying pitfalls in XAI methods, and providing conceptual frameworks for human-compatible XAI [9]. On the other hand, the literature shows a lack of frameworks, methods, and metrics to evaluate whether XAI methods provide adequate explainability to humans [10, 11, 12]. The few XAI user evaluations that exist often lack insight from cognitive or social sciences [6] and do not follow a standard procedure for measuring, quantifying, and comparing the explainability of AI systems [13]. Further supporting this observation, Wells and Bednarz [14] conducted a systematic review examining XAI studies through a user-focused lens. Their findings revealed that many studies did not involve users. Even when user testing was conducted, key details were often omitted, such as the number of participants, recruitment methods, or participants' level of expertise in machine learning. This limits the transparency and reproducibility of evaluations.

Interacción '25: XXV International Conference on Human-Computer Interaction, September 03–05, 2025, Valladolid, Spain
*Corresponding author.
francesc-xavier.gaya@uib.es (F. X. Gaya-Morey); josemaria.buades@uib.es (J. M. Buades-Rubio); mack@yorku.ca (I. S. MacKenzie); cristina.manresa@uib.es (C. Manresa-Yee)
0000-0003-1231-7235 (F. X. Gaya-Morey); 0000-0002-6137-9558 (J. M. Buades-Rubio); 0000-0003-1731-9651 (I. S. MacKenzie); 0000-0002-8482-7552 (C. Manresa-Yee)

Studies empirically evaluating XAI methods with specific tasks, users, and contexts show different needs and preferences of users [15, 16, 17]. When addressing visual data (viz., images or videos), evaluations of XAI methods with users are found for images [18, 19, 20, 21], but to the best of our knowledge, there is no work addressing videos. Therefore, research is needed to fully understand the impact of XAI methods for video and the effectiveness of the explanations.

In this work, we describe a quantitative study with experts assessing six video removal-based XAI methods applied across three networks and two datasets. To achieve this, we adapted six widely used XAI methods, originally for image-based local explanations, to the video domain. To reveal differences, we generated explanations for three networks representing varied approaches: transformers and convolutional models. Additionally, we utilized two publicly available human-action datasets: one recorded in a controlled environment and another comprising videos from YouTube.

We quantify user preferences of the six types of XAI methods. Our findings show an agreement both with the preferred and the least preferred methods. In light of the findings, researchers and practitioners have concrete design implications for user-driven choices for XAI methods.

The paper is organized as follows: Section 2 provides a review of the key related concepts. Section 3 details the AI-driven system, including the datasets, neural networks, and XAI methods used. Section 4 outlines the methodology, covering the participants, apparatus, procedure, and design of the study. The results are presented in Section 5, followed by a discussion in Section 6. Finally, Section 7 concludes and highlights potential directions for future research.

2. Related work

2.1. Human-centered explainable artificial intelligence

XAI refers to methods and techniques that make the decisions and inner workings of AI models more transparent and understandable to humans. As AI systems are increasingly complex, particularly in deep learning, understanding how they arrive at outcomes is more difficult. XAI addresses this by providing insight into model behavior, thus allowing users to interpret, trust, and validate AI-driven decisions [2, 3].

Explainability techniques are generally categorized along several key dimensions: data versus model focus, direct versus post-hoc explanations, global versus local scope, and static versus interactive presentation [22]. First, explanations may aim to clarify either the properties of the input data or the behavior of the model itself. When explaining the model, the distinction is between directly interpretable models (e.g., linear regression, tree-based methods) and post-hoc explainability, which is applied after the model is trained. Additionally, explanations may target individual predictions (local) or the model's behavior as a whole (global). Finally, explanations can be static or, as recommended by Miller [5], designed to support interactive user engagement for deeper understanding.

Human-centered XAI builds on the foundation of traditional XAI by focusing not just on technical explainability, but also on aligning explanations with human needs, cognitive processes, and context [23, 24]. Rather than assuming that an explanation is sufficient, human-centered XAI emphasizes usability, interpretability, and relevance for diverse users, including non-experts [25]. The goal is to create explanations that are intuitive and context-aware and thereby support decision-making to improve the collaboration between humans and AI systems [9, 26]. This perspective recognizes that the effectiveness of an explanation depends as much on the user as on the model itself.

2.2. XAI applied to video data

While image-based XAI methods are extensively studied (e.g., [27, 28, 29]), video-based XAI methods, particularly model-agnostic ones, remain relatively underexplored due to the unique challenges posed by video. However, model-agnostic methods are valuable because they offer flexibility and broad applicability for real-world scenarios. For an up-to-date and comprehensive overview of XAI methods designed for video data, we refer the reader to the review by Gaya-Morey et al. [30].

2.3. User studies for visual data

Most existing work on XAI focuses on algorithmic metrics, often overlooking how actual users interpret, trust, or benefit from these explanations [5]. Consequently, relatively few user studies evaluate image and video explanations with human participants. This gap is especially pronounced in the video domain, where the temporal dimension adds complexity to human interpretation. Evaluating explanations with real users is crucial for understanding their practical utility, improving design choices, and ensuring that such systems align with human reasoning and decision-making.

Regarding user studies on XAI images, Aechtner et al. [21] studied users' perception on local vs. global explanations, showing the preference for AI novices toward local explanations. Manresa et al. [19] also studied local vs. global explanations, engaging 104 users on aspects such as perceived trust or understanding. Higher scores were observed for combinations of both explanations. Alqaraawi et al. [18] also studied the performance of saliency maps in a user study. They reported a preference for LRP and noted the limited help of explanations in predicting the network's output for new images or in identifying image features the system is sensitive to. Selvaraju et al. [31] explored whether Grad-CAM explanations helped users establish appropriate trust in predictions. Their results showed that Grad-CAM enabled untrained users to successfully differentiate a "stronger" deep network from a "weaker" one, even when they produced identical predictions.

In our review of the literature for XAI applied to video, we did not find any work evaluating or comparing different explanations on videos from a human perspective.

3. AI-driven system and XAI Methods

To evaluate user preferences for video XAI methods, we created a sample set combining three networks, two datasets, and six XAI methods. This allowed us to introduce variation and thereby identify influences across the three networks.

3.1. Datasets

We selected two datasets with distinct characteristics to train the models and evaluate the XAI methods: Kinetics 400 [32] and EtriActivity3D [33].

The Kinetics 400 dataset is a large-scale collection of YouTube videos covering 400 human action categories, with at least 400 video clips per class. The dataset focuses on diverse human activities, including both interactions between people and interactions with objects. It features a wide variety of participants, environments, and objects, alongside challenges such as camera motion and video edits within the same clip, contributing to its complexity.

In contrast, EtriActivity3D is a more specialized dataset containing 112,620 video samples across 55 activity classes. It focuses on everyday tasks performed by 100 individuals, half of whom are over 64 years old, providing insight into older demographics. The videos were captured in home environments across multiple rooms and from eight fixed cameras, ensuring a stable, unedited recording for each clip. This controlled setup allows for a consistent evaluation, free from the variations introduced by camera movements or edits.

3.2. Neural networks

We used three networks: TimeSformer [34], TANet [35], and TPN [36]. These networks represent different architectural approaches—transformers and convolutional models—allowing us to explore both similarities and differences in user evaluations. The choice of networks is justified by their performance in action classification tasks and their public availability within the MMAction2 PyTorch-based open-source toolbox for video analysis [37].

TimeSformer, a variant of the Vision Transformer, captures spatio-temporal features by processing frame-level patches. TANet incorporates a Temporal Adaptive Module (TAM) within its 2D CNN

framework, enabling the capture of both short-term and long-term temporal dynamics using a two-level adaptive mechanism. The Temporal Pyramid Network (TPN), on the other hand, extracts and integrates spatial, temporal, and semantic information using hierarchical rescaling; that is, enhancing performance for tasks with temporal variability. For both TANet and TPN, we used the ResNet50 architecture as the backbone.

For the Kinetics 400 dataset, we utilized pre-trained weights available in the MMAction2 framework. For the EtriActivity3D dataset, we fine-tuned the networks using the pre-trained Kinetics 400 weights, training them for 10 epochs with 5-fold cross-validation.

3.3. XAI methods

Since we adopted networks with varying architectures, we opted for model-agnostic XAI methods, which generate explanations independent of the underlying model. Specifically, we employed publicly available[1] video adaptations of widely used model-agnostic XAI methods, including LIME (Video LIME) [27], Kernel-SHAP (Video Kernel-SHAP) [29], RISE (Video RISE) [28], occlusion sensitivity (Video SOS) [38], LOCO (Video LOCO) [39], and univariate predictors (Video UP) [40]. Aligned with the previously mentioned dimensions of XAI, these methods aim to explain the model and are characterized by being post-hoc, local in scope, and static in presentation [30].

The operation of these methods involves four main steps: (1) segmenting the input video into regions consisting of pixels from different frames, (2) occluding these regions and passing the modified video through the model, where predictions change based on the occluded regions, (3) summarizing the relevance of each region to the target prediction, and (4) visualizing the explanations. The exact parameters used in each step depend on the method. The application of these XAI methods to explain a model's prediction for a given video produces an explanation in the form of a video, within which each pixel represents the relevance of the corresponding pixel in the original video. Figure 1 displays an explanation example using each method.

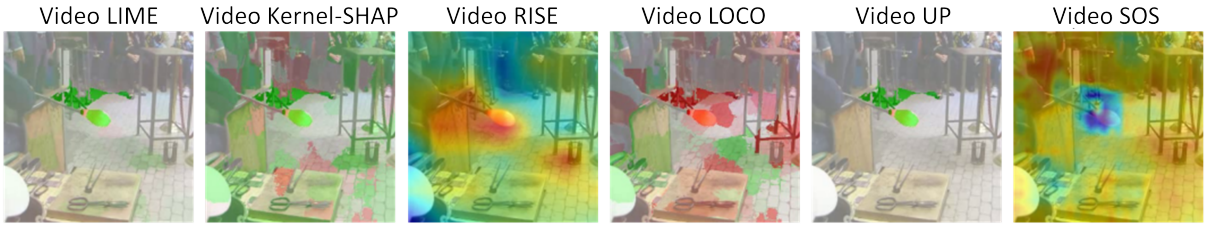

Figure 1: Example of computed explanations on the TimeSformer model and Kinetics 400 dataset, using the different methods. Only one frame of the moment of maximum relevance is shown per method.

3.4. Video explanations

For the evaluation, we selected a random sample of 30 videos from each dataset—Kinetics 400 and EtriActivity3D. For consistency, we included only videos that were correctly classified by all three networks used in the study. If a video was misclassified by any network, it was replaced with another randomly selected video. To ensure a fair comparison, we enforced equal conditions for the methods, such as the number of features, samples, and occlusion type.

Each of the 30 videos from both datasets was processed through the three networks—TimeSformer, TPN, and TANet. For every prediction, explanations were generated using all six XAI methods described earlier. This resulted in $30 \times 6 \times 3 \times 2 = 1,080$ explanations across the experiment.

To enhance the clarity and interpretability of the explanations, only the top 30% most relevant regions were retained, filtering out less significant areas. Additionally, we applied histogram stretching to ensure the explanations utilized the full range of the color spectrum, making the visualizations more

[1] https://github.com/xavi3398/revex_framework

distinct. Furthermore, negative relevance values were removed for two main reasons: to simplify the information presented to users assessing the explanations and to standardize outputs across all XAI methods, as not all methods provide both positive and negative relevance scores.

4. Method

Explanations were presented to the users to assess their preferences.

4.1. Participants

Six volunteer participants (three female) from the local university were recruited. Ages ranged from 24 to 47 years (mean = 34.7, *SD* = 10.4). The experts have extensive experience in both AI and XAI, with their expertise grounded in years of specialized research and practical applications. Two of the experts, the younger ones, worked in AI for at least three to four years and have spent the past two years working in XAI. The more experienced experts have an extensive background both in AI and XAI, having worked in the latter area for a minimum of five years. Further, three of the experts focus their research in HCI. Their research spans a broad spectrum, including computer vision and deep learning applied to HCI problems. Although all participants had experience with AI and XAI, their familiarity did not extend to all XAI methods.

4.2. Apparatus

A user interface was developed to display the video, its associated class, the explanation, and a corresponding color map to aid users in their evaluation. See Figure 2. The interface included the 1,080 explanations, with each screen displaying only one type of explanation.

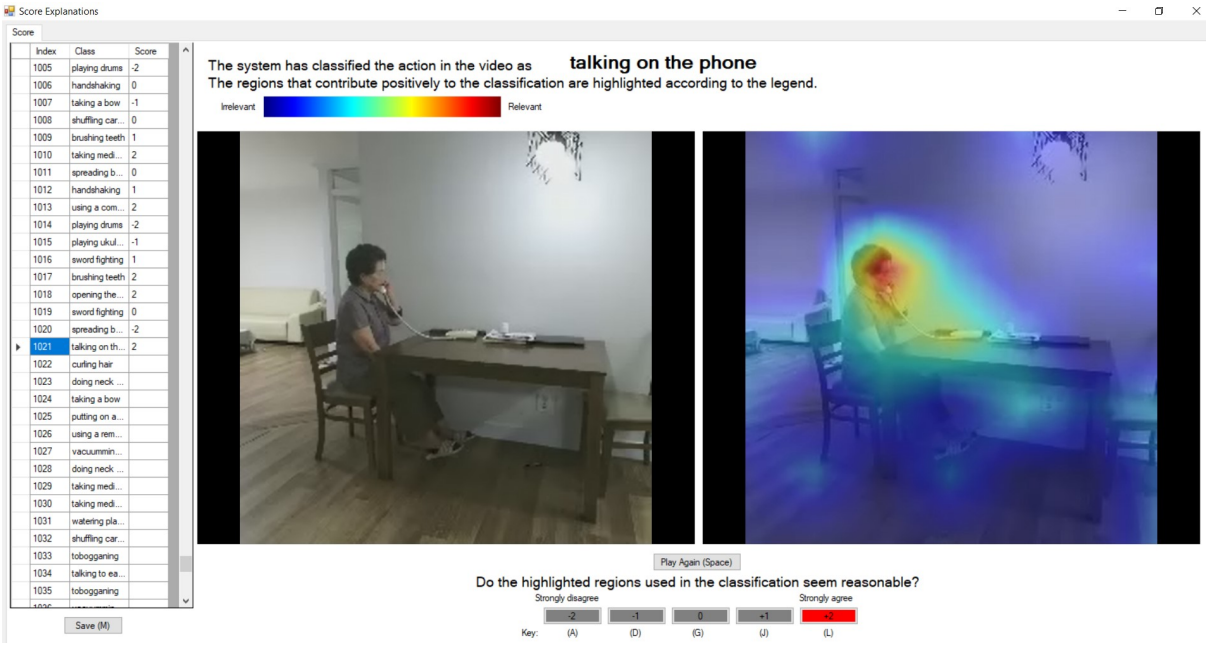

Figure 2: User interface to score the explanations generated by the XAI method. On the left, a scrollable list of videos with their corresponding class and the user-assigned scores. At the top, pertinent information about the current explanation is displayed, including the class and the color map representing the explanation. In the center, the video being explained is presented (left) alongside its corresponding explanation (right). At the bottom, the user is presented with a question and response options, with the selected response in red.

The question posed to the users during evaluation was: "Do the highlighted regions used in the classification seem reasonable?" The response options ranged from -2 (*strongly disagree*) to +2 (*strongly*

agree). Thus, negative scores highlight the deviation of participants from the explanation. The query sought to determine whether the highlighted regions align with users' perceptions when identifying specific actions in the video. To mitigate any potential bias, explanations are presented in random order and without information regarding the network, dataset, or XAI method. This ensures a "blind" evaluation.

4.3. Procedure

The study was conducted using a laptop with the program installed locally, which participants were allowed to take home. Each participant was tasked with evaluating 1,080 explanations, a process that required approximately 3 to 4 hours. To accommodate this, participants were given the flexibility to pause and resume the evaluation at their convenience.

The explanations were presented to all participants in the same order, with the next explanation automatically displayed after one was assessed. However, participants had the flexibility to navigate freely between explanations, enabling them to revisit, reassess, and update their scores as needed.

For each method, the mean score was computed for all participants for each XAI method. In addition, we created aggregated barplots of the participant scores by method, dataset, and network and analyzed statistical significance of the different factors.

4.4. Design

The study was a $6 \times 3 \times 2$ within-subjects design with the following independent variables and levels:

- XAI method (Video RISE, Video Kernel-SHAP, Video LOCO, Video LIME, Video SOS, Video UP)
- Network (TimeSformer, TANet, TPN)
- Dataset (EtriActivity3D, Kinetics400)

The dependent variable was the score for reasonableness on a 5-point Likert scale from -2 (*strongly disagree*) to 2 (*strongly agree*).

The total number of trials was 6,480 (= 6 participants $\times$ 6 XAI methods $\times$ 3 networks $\times$ 2 datasets $\times$ 30 videos per condition).

5. Results

We now present the results for the participant assessments by the conditions tested. The grand mean over all 6,480 explanations was 0.292. To the question of interest, this represents an overall participant response between 0 (*neutral*) and 1 (*mildly agree*). Thus, there was a general tendency for participants to feel the explanations from the XAI methods leaned toward "reasonable." By XAI method, the means were -0.806 (Video UP), -0.093 (Video LOCO), 0.368 (Video SOS), 0.356 (Video Kernel-SHAP), 0.540 (Video LIME), and 1.390 (Video RISE). By network, the means were 0.200 (TANet), 0.389 (TimeSformer), and 0.288 (TPN). By dataset, the means were 0.096 (EtriActivity3D) and 0.489 (Kinetics 400). Further analyses by combinations of these conditions are now described. The user-based evaluation results are presented in Figures 3 and 4. Figure 3 illustrates the average scores per method, while Figure 4 aggregates the scores by dataset, network, and method.

A three-way ANOVA was conducted to evaluate the effects of Dataset, Neural Network architecture, and XAI Method on user ratings. Significant main effects were found for Dataset ($F_{1,6444}$ = 162.83), Network ($F_{2,6444}$ = 12.49), and XAI Method ($F_{5,6444}$ = 369.43). Additionally, significant interactions were observed between Dataset $\times$ Network ($F_{2,6444}$ = 10.98), Dataset $\times$ XAI Method ($F_{5,6444}$ = 43.17), Network $\times$ XAI Method ($F_{10,6444}$ = 9.06), and the three-way interaction Dataset $\times$ Network $\times$ XAI Method ($F_{10,6444}$ = 5.61). In all cases, $p < .001$. These results indicate that user perception depends not only on individual factors but also on their combinations, with the XAI Method showing the strongest effect on ratings.

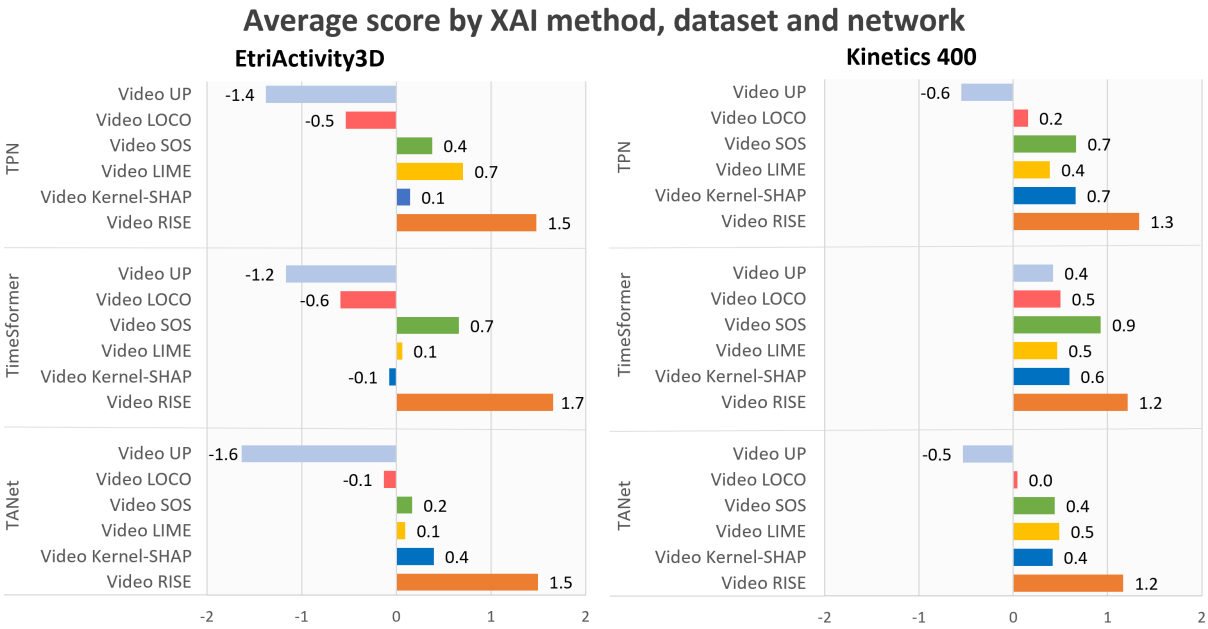

Figure 3: User scores by XAI method, dataset, and network.

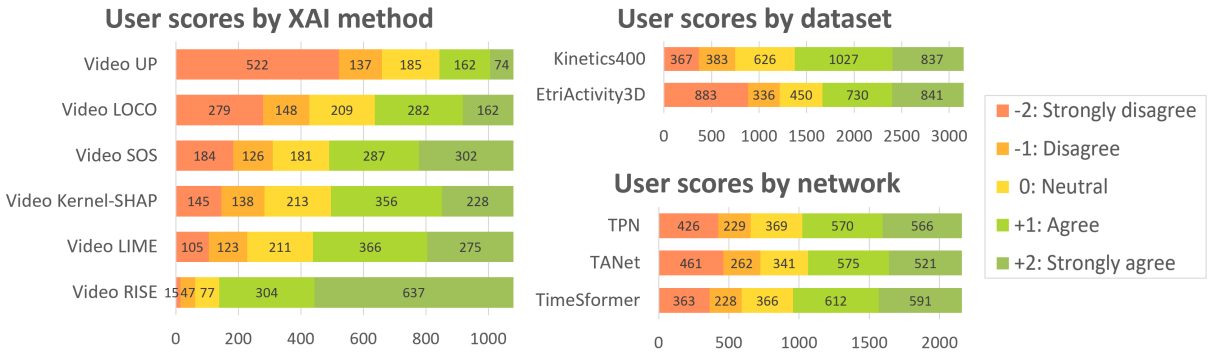

Figure 4: User scores separately by XAI method, dataset, and network.

To assess the explanatory power of different factors, we computed R^2 for three models: one including all factors (Dataset, Network, and XAI Method), one considering only the XAI Method, and one including only Dataset and Network. The full model achieved R^2 = .273, indicating that the factors together explain 27.3% of the variance in user ratings. The model considering only the XAI Method yielded R^2 = .208, confirming that XAI Method is the most influential factor. Conversely, the model including only Dataset and Network yielded only R^2 = .024, suggesting that these factors alone contribute minimally to explaining user ratings. Additionally, the full model had the lowest AIC (21,216) and BIC (21,460), indicating the best balance between goodness-of-fit and model complexity.

A post hoc Tukey HSD test was conducted to analyze pairwise differences between XAI Methods. The results revealed significant differences in most comparisons (p < .05), except between Video Kernel-SHAP and Video SOS (p = .999), where no significant difference was found. The Video RISE method consistently obtained significantly higher ratings compared to other methods, with the largest differences observed against Video UP (mean difference = 2.20, p < .001). Conversely, Video UP received significantly lower ratings than all other methods. These findings confirm that the choice of XAI Method strongly influences user ratings. Video RISE exhibits the most favorable results, attaining an average score of 1.39 within the range [-2, 2]. In the second position, is Video LIME (0.54), closely followed by Video SOS (0.37) and Video Kernel SHAP (0.36). Conversely, Video LOCO scores poorly (-0.09), and Video UP receives a score of -0.81, the lowest score.

A post hoc Tukey HSD test was also performed to analyze pairwise differences between the three

neural network architectures. The results revealed that TimeSformer received significantly higher ratings than TANet (mean difference = 0.19, $p < .001$). However, the differences between TPN and TANet (mean difference = 0.09, $p = .119$) and between TimeSformer and TPN (mean difference = 0.10, $p = .056$) were not statistically significant.

6. Discussion

The preference for Video RISE by experts suggests that placing important regions over the image made sense to the users (see Figure 5, first row, third column, explanation for the "brushing hair" class). Also, the smooth explanations displayed by Video RISE, without hard edges, were favored over other methods. This observation prompts the question of whether introducing smoothness in other methods, such as through a Gaussian filter, would positively influence the quality of the explanation according to users. While Video RISE consistently achieved superior results across datasets and networks, the performance of other methods varied depending on these factors. For instance, Video UP scored approximately one point higher on Kinetics 400 than on EtriActivity3D, and Video SOS performed better on TimeSformer than on other networks. This suggests that certain XAI methods may be better suited to specific neural networks or data characteristics.

The dataset also influenced user ratings. On average, scores for Kinetics 400 were 0.39 points higher than those for EtriActivity3D, with the ANOVA confirming this difference as significant ($F_{1,6444} = 162.8$, $p < .001$). We attribute the difference to dataset complexity: Kinetics 400 features more challenging videos with camera movements, cuts, and a broader range of action classes, making explanation generation more difficult. In contrast, EtriActivity3D offers a simpler context for identifying important regions for classification, which likely influenced user scores.

Regarding network selection, we observed significant differences in average user scores between two models: TimeSformer (average score = 0.39) and TANet (average score = 0.20). However, no significant difference was found between TPN (average score = 0.29) and either of the other two. This demonstrates that, even when trained under identical conditions, architectural differences between models impacts user evaluations. For example, the explanations with Video UP and Video SOS consistently received higher scores when generated for TimeSformer compared to the other two networks, as shown in Figure 3. Consequently, to ensure a fair assessment of XAI methods, experiments should include multiple networks representing diverse architectural designs.

A limitation of this study is the participant sample size. However, the unanimous agreement among participants on both the best and worst explanations strengthens our confidence in the findings. Figure 5 presents examples of explanations that unanimously received the highest and lowest scores.

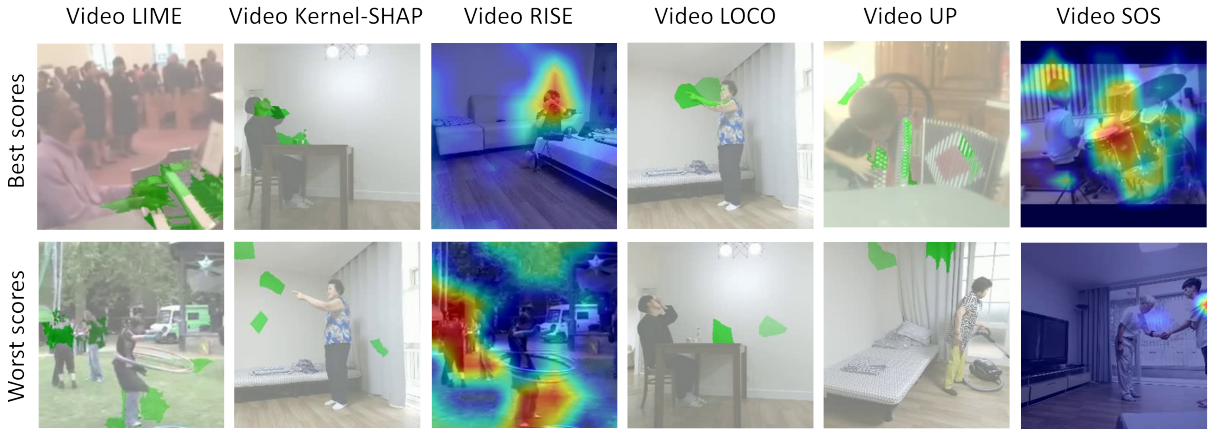

Figure 5: Examples images receiving the highest and lowest scores from the six experts. Columns represent different XAI methods, while rows show the highest-scored images (top) and lowest-scored images (bottom).

7. Conclusion

Although numerous XAI methods exist for generating explanations, selecting the most suitable method remains a challenge for both researchers and practitioners. This study marks a step forward in understanding how users perceive six well-known local XAI methods (including LIME, SHAP, or RISE) when adapted to the video domain. By applying explanations on diverse datasets and networks, we explored the influence of those factors. Remarkably, and although the sample of experts is small, there was consensus: Video RISE was preferred by participants while Video UP received the lowest scores.

User studies to evaluate XAI methods are essential for gaining insights into how users interact with and interpret explanations from AI systems. This knowledge can guide technical decisions based on users' explainability preferences, and help in choosing an appropriate XAI method for real world applications. However, the studies are time-consuming and costly, requiring significant resources to gather meaningful data. To speedup the evaluation process, automatic metrics such as area under the curve (AUC) can offer more efficient ways to assess XAI methods. However, an ongoing debate persists on whether the performance of XAI methods through objective metrics should take precedence over user preferences in determining their effectiveness or application. Nevertheless, we believe it is fundamental to test how well automatic metrics align with the user perspective. Bridging this gap will ensure the evaluation process remains both effective and representative of real-world user experiences.

Future work will involve evaluations with a larger sample of users to further validate and test our findings. Additionally, incorporating greater participant diversity—such as variations in AI knowledge, age, and other demographics—will provide deeper understanding of XAI methods from a human perspective.

Acknowledgments

This work is part of the Project PID2023-149079OB-I00 (EXPLAINME) funded by MICIU/AEI/10.13039/501100011033/ and ERDF, EU and of Project PID2022-136779OB-C32 (PLEISAR) funded by MICIU/ AEI /10.13039/501100011033/ and FEDER, EU. F. X. Gaya-Morey was supported by an FPU scholarship from the Ministry of European Funds, University and Culture of the Government of the Balearic Islands.

Declaration on generative AI

The authors have not employed any Generative AI tools.

References

[1] S. Dong, P. Wang, K. Abbas, A survey on deep learning and its applications, Computer Science Review 40 (2021) 100379. doi:10.1016/j.cosrev.2021.100379.

[2] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion 58 (2020) 82–115. doi:/10.1016/j.inffus.2019.12.012.

[3] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160. doi:10.1109/ACCESS.2018.2870052.

[4] T. A. Schoonderwoerd, W. Jorritsma, M. A. Neerincx, K. van den Bosch, Human-centered XAI: Developing design patterns for explanations of clinical decision support systems, International Journal of Human-Computer Studies 154 (2021) 102684. doi:10.1016/j.ijhcs.2021.102684.

[5] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38. doi:10.1016/j.artint.2018.07.007.

[6] Y. Rong, T. Leemann, T.-T. Nguyen, L. Fiedler, P. Qian, V. Unhelkar, T. Seidel, G. Kasneci, E. Kasneci, Towards human-centered explainable AI: A survey of user studies for model explanations , IEEE

Transactions on Pattern Analysis & Machine Intelligence 46 (2024) 2104–2122. doi:10.1109/TPAMI.2023.3331846.

[7] S. Kaplan, H. Uusitalo, L. Lensu, A unified and practical user-centric framework for explainable artificial intelligence, Knowledge-Based Systems 283 (2024) 111107. doi:10.1016/j.knosys.2023.111107.

[8] S. Mohseni, N. Zarei, E. D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable AI systems, ACM Transactions on Interactive Intelligent Systems 11 (2018) 24:1–24:45. doi:10.1145/3387166.

[9] Q. V. Liao, K. R. Varshney, Human-centered explainable AI (XAI): From algorithms to user experiences, 2022. doi:10.48550/arXiv.2110.10790.

[10] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. Vayena, AI4People: An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations, Minds and Machines 28 (2018) 689–707. doi:10.1007/s11023-018-9482-5.

[11] R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Metrics for explainable AI: Challenges and prospects, 2019. doi:10.48550/arXiv.1812.04608.

[12] M. Miró-Nicolau, A. J. i Capó, G. Moyà-Alcover, Assessing fidelity in XAI post-hoc techniques: A comparative study with ground truth explanations datasets, Artificial Intelligence 335 (2024) 104179. doi:10.1016/j.artint.2024.104179.

[13] N. Burkart, M. F. Huber, A survey on the explainability of supervised machine learning, Journal of Artificial Intelligence Research 70 (2021) 245–317. doi:10.1613/jair.1.12228.

[14] L. Wells, T. Bednarz, Explainable AI and reinforcement learning: Asystematic review of current approaches and trends, Frontiers in Artificial Intelligance 4 (2021) 1–15. doi:10.3389/frai.2021.550030.

[15] J. Dodge, Q. V. Liao, Y. Zhang, R. K. E. Bellamy, C. Dugan, Explaining models: An empirical study of how explanations impact fairness judgment, in: Proceedings of the 24th International Conference on Intelligent User Interfaces – IUI '19, ACM, New York, 2019, pp. 275–285. doi:10.1145/3301275.3302310.

[16] U. Ehsan, S. Passi, Q. V. Liao, L. Chan, I.-H. Lee, M. Muller, M. O. Riedl, The who in XAI: How AI background shapes perceptions of AI explanations, in: Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems – CHI '24, ACM, New York, 2024, pp. 316.1–316.32. doi:10.1145/3613904.3642474.

[17] M. Szymanski, M. Millecamp, K. Verbert, Visual, textual or hybrid: The effect of user expertise on different explanations, in: Proceedings of the 26th International Conference on Intelligent User Interfaces – IUI '21, ACM, New York, 2021, pp. 109–119. doi:10.1145/3397481.3450662.

[18] A. Alqaraawi, M. Schuessler, P. Weiß, E. Costanza, N. Berthouze, Evaluating saliency map explanations for convolutional neural networks: A user study, in: Proceedings of the 25th International Conference on Intelligent User Interfaces – IUI '20, ACM, New York, 2020, p. 275–285. doi:10.1145/3377325.3377519.

[19] C. Manresa-Yee, S. Ramis, F. X. Gaya-Morey, J. M. Buades, Impact of explanations for trustworthy and transparent artificial intelligence, in: Proceedings of the XXIII International Conference on Human Computer Interaction– Interacción '23, ACM, New York, 2024. doi:10.1145/3612783.3612798.

[20] A. Heimerl, K. Weitz, T. Baur, E. Andre, Unraveling ML models of emotion with NOVA: Multi-level explainable AI for non-experts, IEEE Transactions on Affective Computing 1 (2020) 1–13. doi:10.1109/TAFFC.2020.3043603.

[21] J. Aechtner, L. Cabrera, D. Katwal, P. Onghena, D. P. Valenzuela, A. Wilbik, Comparing user perception of explanations developed with XAI methods, in: Proceedings of the IEEE International Conference on Fuzzy Systems – FUZZ-IEEE '22, IEEE, New York, 2022, pp. 1–7. doi:10.1109/FUZZ-IEEE55066.2022.9882743.

[22] V. Arya, R. K. E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović, S. Mourad, P. Pedemonte, R. Raghavendra, J. T. Richards, P. Sattigeri,

K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, Y. Zhang, AI explainability 360: An extensible toolkit for understanding data and machine learning models, Journal of Machine Learning Research 21 (2020) 1–6. URL: http://jmlr.org/papers/v21/19-1035.html.

[23] Q. V. Liao, D. Gruen, S. Miller, Questioning the AI: Informing design practices for explainable AI user experiences, in: Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems – CHI '20, ACM, New York, 2020, pp. 1–15. doi:10.1145/3313831.3376590.

[24] A. J. Barda, C. M. Horvat, H. Hochheiser, A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare, BMC Medical Informatics and Decision Making 20 (2020). doi:10.1186/s12911-020-01276-x.

[25] P. Lopes, E. Silva, C. Braga, T. Oliveira, L. Rosado, XAI systems evaluation: A review of human and computer-centred methods, Applied Sciences 12 (2022). doi:10.3390/app12199423.

[26] U. Ehsan, P. Wintersberger, Q. V. Liao, E. A. Watkins, C. Manger, H. Daumé III, A. Riener, M. O. Riedl, Human-centered explainable AI (HCXAI): Beyond opening the black-box of AI, in: Extended Abstracts of the ACM SIGCHI Conference on Human Factors in Computing Systems – CHI EA '22, ACM, New York, 2022. doi:10.1145/3491101.3503727.

[27] M. T. Ribeiro, S. Singh, 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 – KDD '16, ACM, New York, 2016, pp. 1135–1144. doi:10.1145/2939672.2939778.

[28] V. Petsiuk, A. Das, K. Saenko, RISE: Randomized input sampling for explanation of black-box models, in: Proceedings of the British Machine Vision Conference – BMVC '18, Newcastle, UK, 2018, pp. 1–151. doi:10.48550/arXiv.1806.07421.

[29] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems – NIPE '17, 2017, pp. 4768–4777. URL: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[30] F. X. Gaya-Morey, J. M. Buades-Rubio, I. S. MacKenzie, C. Manresa-Yee, Revex: A unified framework for removal-based explainable artificial intelligence in video, 2024. doi:https://doi.org/10.48550/arXiv.2401.11796, submitted for publication.

[31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the International Conference on Computer Vision – ICCV '17, IEEE, New York, 2017, pp. 618–626. doi:10.1109/ICCV.2017.74.

[32] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, A. Zisserman, The kinetics human action video dataset, 2017. doi:10.48550/arXiv.1705.06950.

[33] J. Jang, D. Kim, C. Park, M. Jang, J. Lee, J. Kim, ETRI-Activity3D: A large-scale RGB-D dataset for robots to recognize daily activities of the elderly, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems – IROS '20, IEEE, New York, 2020, pp. 10990–10997. doi:10.1109/IROS45743.2020.9341160.

[34] G. Bertasius, H. Wang, L. Torresani, Is space-time attention all you need for video understanding?, in: Proceedings of the 38th International Conference on Machine Learning Research – PMLR '21, volume 139, ML Research Press, Maastricht University, Netherlands, 2021, pp. 813–824. URL: https://proceedings.mlr.press/v139/bertasius21a/bertasius21a-supp.pdf.

[35] Z. Liu, L. Wang, W. Wu, C. Qian, T. Lu, TAM: Temporal Adaptive Module for video recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision – ICCV '21, IEEE, New York, 2021, pp. 13688–13698. doi:10.1109/ICCV48922.2021.01345.

[36] C. Yang, Y. Xu, J. Shi, B. Dai, B. Zhou, Temporal pyramid network for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition – CVPR '20, IEEE, New York, 2020, pp. 591–600. doi:10.1109/CVPR42600.2020.00067.

[37] OpenMMLab, OpenMMLab's next generation video understanding toolbox and benchmark, https://github.com/open-mmlab/mmaction2, 2020.

[38] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Proceedings of the 13th European Conference on Computer Vision – ECCV '14 (LNCS 8689), Springer, Cham, 2014, pp. 818–833. doi:10.1007/978-3-319-10590-1_53.

[39] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, L. Wasserman, Distribution-free predictive inference for regression, Journal of the American Statistical Association 113 (2018) 1094–1111. doi:10.1080/01621459.2017.1307116.

[40] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, Journal of Machine Learning Research 3 (2003) 1157–1182. URL: https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf.