Learn, Generate, Rank, Explain: A Case Study of Visual Explanation by Generative Machine Learning

CHRIS KIM, Ontario Tech University, Canada
XIAO LIN, SRI International, USA
CHRISTOPHER COLLINS, Ontario Tech University, Canada
GRAHAM W. TAYLOR, University of Guelph and Vector Institute for AI, Canada
MOHAMED R. AMER, SRI International, USA

While the computer vision problem of searching for activities in videos is usually addressed by using discriminative models, their decisions tend to be opaque and difficult for people to understand. We propose a case study of a novel machine learning approach for generative searching and ranking of motion capture activities with visual explanation. Instead of directly ranking videos in the database given a text query, our approach uses a variant of Generative Adversarial Networks (GANs) to generate exemplars based on the query and uses them to search for the activity of interest in a large database. Our model is able to achieve comparable results to its discriminative counterpart, while being able to dynamically generate visual explanations. In addition to our searching and ranking method, we present an explanation interface that enables the user to successfully explore the model's explanations and its confidence by revealing query-based, model-generated motion capture clips that contributed to the model's decision. Finally, we conducted a user study with 44 participants to show that by using our model and interface, participants benefit from a deeper understanding of the model's conceptualization of the search query. We discovered that the XAI system yielded a comparable level of efficiency, accuracy, and user-machine synchronization as its black-box counterpart, if the user exhibited a high level of trust for AI explanation.

CCS Concepts: • **Human-centered computing** → *User studies*; *Information visualization*; • **Computing methodologies** → *Machine learning;*

Additional Key Words and Phrases: Explainable artificial intelligence, model-generated explanation, trust and reliance, user study

The reviewing of this article was managed by special issue associate editors Shixia Liu, Daniel Archambault, Tatiana von Landesberger, Remco ChangCagatay Turkay.
Chris Kim and Xiao Lin contributed equally to this research.
This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions, and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC).
Authors' addresses: C. Kim and C. Collins, Ontario Tech University, 2000 Simcoe St. N, Oshawa, ON L1G 0C5, Canada; email: chris.kim@ontariotechu.net; X. Lin and M. R. Amer, SRI International, 201 Washington Rd, Princeton, NJ 08540, United States; G. W. Taylor, University of Guelph, 50 Stone Rd E, Guelph, ON N1G 2W1, Canada and Vector Institute for AI, Toronto, Ontario, Canada.

2160-6455/2021/08-ART23 $15.00
https://doi.org/10.1145/3465407

ACM Reference format:
Chris Kim, Xiao Lin, Christopher Collins, Graham W. Taylor, and Mohamed R. Amer. 2021. Learn, Generate, Rank, Explain: A Case Study of Visual Explanation by Generative Machine Learning. *ACM Trans. Interact. Intell. Syst.* 11, 3-4, Article 23 (August 2021), 34 pages.
https://doi.org/10.1145/3465407

1 INTRODUCTION

Explainable Artificial Intelligence (XAI) has recently emerged due to increased interest for **Artificial Intelligence (AI)** systems. XAI enables new machine learning techniques, specifically deep learning, to yield *explainable models*. These explanations can be developer-focused (to help in understanding, designing, and improving models) or user-centric (to help in knowing how and when to trust the outputs of AI tools). From the user-centric point of view it is of crucial importance to explain the decisions of an AI system with effective explanation techniques to enable end users to understand, appropriately trust, and effectively manage the decisions made by AI. An effective explainable AI system assists in the human decision-making supported by the system, in particular, whether to accept the recommendations or classifications suggested by the model. In modern AI systems, the most critical and most opaque components are based on machine learning. There is an inherent tension between machine learning performance (predictive accuracy) and explainability; often the highest performing methods, such as deep learning, are the least explainable, while the most explainable, such as decision trees, are the least accurate.

From a decision making point of view, the goal of XAI systems is to maintain performance while being explainable. The target of XAI is an end user who depends on decisions, recommendations, or actions produced by an AI and therefore needs to understand the rationale for the system's decisions. For example, a test operator of a newly developed autonomous system will need to understand why the system makes its decisions so that they can decide how to use it in the future. A successful XAI system should provide end users with an explanation of individual decisions, enable users to understand the system's overall strengths and weaknesses, convey an understanding of how the system will behave in the future, and in some cases even suggest how to correct the system's mistakes. Explanations can be global, explaining how the model works and the ways in which it encodes knowledge, or local, explaining the provenance and confidence in individual decisions or recommendations. Our work focuses on the local level, using visual explanations to help users decide whether to trust an individual output of an AI system.

Explainable models might be created by learning to associate explanatory semantic information with features of the model; by learning simpler models that are easier to explain; by learning richer models that contain more explanatory content; or by inferring approximate models solely for the purpose of explanation. Another critical component of XAI systems are *explanation interfaces* that enable explainable models [24]. For example, Reference [35] provide an example of the development and evaluation of a basic explanation interface. The work followed a complete development strategy that included identifying principles of explainability, developing an interface from those principles, and evaluating the effectiveness of the explanations provided by the interface. The system explained a very simple naive Bayesian text classifier.

In this article, we introduce the paradigm of *explanation by generation*. We propose a novel generative XAI system for human activity search and ranking in motion capture data.

Visual search and ranking is a prominent problem in the computer vision community. Applications of visual search and ranking range from commerce, to surveillance, to robotics. Recent work on ranking focused on discriminative methods [4]: given a query, the goal is to retrieve videos containing the query, along with the query's location visually highlighted in the video. We depart from the discriminative ranking paradigm, and propose a generative ranking framework based on the **Dense Validation Generative Adversarial Networks (DVGANs)** approach [37]. We defined the problem as follows: given a text query, generate multiple video hypotheses representing the query, then search for the query using the model-generated videos. In this case, by having the model generate the visual information, presented in an analytics dashboard, we can give the user an insight on what the model "thinks" the query looks like, hence, it becomes more explainable. The underlying model is a **Generative Adversarial Network (GAN)** [22] for human motion generation from text.

Innovations in interface design enable the presentation of detailed results, beyond simple yes-or-no answers. Our interactive explanation interface acts as the mediator between the user and the model, permitting the model's rationale for decisions to be explained in a variety of ways. The explanation interface combines visualization of exemplars generated by the generative model and a confidence score associated with each search result. This further supports the user's understanding about why a specific instance was returned and how confident the system was in its decision. Our approach explores the nuanced confidence and sensitivity in the decision, thereby helping a user set an appropriate level of trust in decisions made by the system. The interface is designed such that it enables visualization of explanations generated by the model and allow the user to drill down on decisions. As users determine and establish trustworthiness with provenance information [26], we see an opportunity for bringing a new type of provenance information model to the forefront.

Finally, we evaluate our XAI system in the context of surveillance, where the user is querying a database of videos and searching for a specific activity in various scenarios. We used the CMU motion capture database [1] to devise an extensive user study that evaluates the quality of explanations, gauges the users' satisfaction with the explanations, and assesses the users' mental model as well as trust and reliance in the system. Our XAI generative ranking system improves explainability while maintaining a high level of performance, comparable to a black-box AI discriminative ranking system. We found that the XAI system yielded a comparable level of efficiency, accuracy, and user-machine synchronization as its black-box counterpart, if the user exhibited a high level of trust for AI explanation. *Our contributions are threefold:*

(1) A GAN-based explainable AI system, based on a generative model with performance comparable to its discriminative counterpart, for human activity search and ranking.
(2) A visual interface for traversing exemplars created by the generative model and exploring confidence and sensitivity in model decisions.
(3) The findings of a user study evaluating the GAN-based explainable AI system in comparison to black-box AI, based on various criteria including accuracy, speed, and user satisfaction.

2 RELATED WORK

In this section, we review the relevant literature on recent XAI approaches, literature related to our specific explainable model approach, and finally, literature related to the explanation interface.

2.1 Related Work on Explainable AI

Our approach is inspired by recent work produced in the XAI domain, as well as new opportunities that emerged therein. Traditionally, as various inference systems extended their capabilities, there has been a need to trace and represent each system's decision making process to justify its conclusions, identify any contradictions, and further improve the corresponding operations [12]. Extending on the need to understand both automatically and manually coded rules, there has been a demand for more transparent, explainable AI systems. We divide the related work into two different groups: global explanations and local explanations. While an exhaustive review of XAI is beyond the scope of this article, Molnar provides one [40].

Global explanations focus on analyzing overall learned representations, for example, understanding and visualizing representations in deep learning (e.g., convolutional neural networks) [30, 42, 55], analyzing representations learned by deep reinforcement learning agents (e.g., deep Q-networks) [54] or learning disentangled representations [28]. In the global explanation case, after the model is learned, the explanation is then extracted from the representation learned by the model itself.

Local explanations focus more on grounding the explanations on specific data, for example, finding influential features [45] and grounding them on the input image. Other methods focused on finding influential data points [33] and parameterizing training batches. Recent work focused on generating textual explanations by training a second deep network to generate explanations without explicitly identifying the semantic features of the original network [27]. Finally, attention-based methods for explanation, such as show and tell networks, couple captioning with attention on images [51], using attributes for attention [9] or using guided attention [36].

Beyond the above developer-centric explanations that focus on AI models and corresponding data points, however, there is also a call for more intuitive and interpretable explanations for human users. Some XAI projects pursue "human-in-the-loop," user-centric systems that produce trustworthy answers that without significantly compromising the system performance [45]; other applications seek different ways to ensure fairness and accountability by providing users alternative outcomes using counterfactual statements ("had a number of conditions been different, the outcome would change") via intuitive voice assistants [48]; finally, ideal XAI projects also provide contextually relevant recommendations and explanations to their end users who may have little to no technical knowledge in AI systems, but are experts in their own domains [24].

2.2 Related Work on Explanation by Generation

Since our explainable model is based on GANs, we briefly review the relevant literature. GANs [22] are a class of implicit generative models that learn directly from examples. They have been employed successfully in many problems, mostly in the area of computer vision where GANs are trained directly on pixels. There are multiple variations of GANs, many of which propose a variation of the objective function to address different needs. Starting from the original formulation [22], the extension to **Conditional GANs (CGAN)** [20] was introduced to enable conditioning on a class label, **Wasserstein GANs (WGAN)** [3] was introduced to improve the stability of GANs, and finally **WGANs with Gradient Penalty (WGAN-GP)** [23] improved WGAN's stability even further by replacing weight clipping with a gradient penalty in the loss function.

We specifically focus on approaches for human motion generation, since it is most related to our work. There are two types of synthesis: (1) Motion completion, starting from a short clip and extrapolating to a longer clip; (2) Motion generation, starting with a label and generating full clips. Recent work on human motion modeling for motion completion successfully used RNNs [19, 21, 29, 39]. However, they did not do human motion generation from scratch. Recently, GAN-based approaches have been applied successfully to synthesize human motion from text [2, 5, 37] by formulating a sequence-to-sequence model using a GAN framework [39].

2.3 Related Work on Explanation Interfaces

Our research into explainable interfaces for artificial intelligence is inspired by recent advances in interpretability, trust, and explainability in the information visualization and human-computer interaction fields [10, 11, 38]. Information visualizations are often populated with the outputs of machine learning techniques, however, simply visualizing the outputs of an ML system is insufficient as an explanation. Where visualizations such as a topic model plot [16], rendering of features in CNNs [38], or a t-SNE model visualization [50] may be useful for those familiar with the workings of the algorithm, they are inappropriate for data domain experts.

Provenance is a key consideration for supporting decision-making in data analytics, and providing traces of both data provenance and analytic provenance has been used to enhance the trustworthiness of analytic outcomes using visual analytics [46, 52]. Analytic provenance tools have recently been the focus of much visual analytics research, and are often a variation of an automatically populated storyboard showing the history of interaction [17, 56].

2.4 Related Works on Explanation Assessment

Designed to closely evaluate each participant's interaction with explanations, our user study is informed by a wide range of work surrounding user trust and reliance in AI explanations, as well as data-driven recommendation systems.

Example-based explanations are generally considered an acceptable way to rationalize algorithmic behaviour [7, 34], while participant reactions to these explanations vary greatly and are subject to individual differences, including self-confidence and prior experience with explanations [13, 47]. Users also generally prefer to retain control over the application, leaving recommendation and AI-driven behaviour only accessible by request [15]. Users are also most likely to opt into utilizing AI assistance to quickly identify the answer, with little interest to learn how or why AI arrived at its solution [14, 43].

When users do accept automatic assistance, however, they do so at arm's length: distancing their own decisions from algorithmic behaviour and adjusting trust levels according to the accuracy of systems [53]. Finally, past work has observed general indifference to how these explanations were presented, as there is no strong preference for the level of detail or the type of visualization presented in the explanations [14, 47].

3 HUMAN ACTIVITY GENERATION FOR SEARCH AND RANKING

In this section, we specify the model used for human activity search and retrieval. We also specify the black-box discriminative AI model used to contrast our generative XAI.

3.1 Problem Statement

To study the effect of global explanations, we use human activity search and ranking (querying a video database for certain activities of interest) as a use case. Human activity understanding is a rich area of research in robotics, computer vision and machine learning, due to the challenges it offers. In this work, we focus on the surveillance use-case [4].

Due to the large size and number of video frames, instead of searching the pixels directly, the video is processed by extracting visual abstractions such as objects, parts and their spatial configurations. This is usually done using detectors [8, 44, 49]. For studying the effect of global explanations, in this work, we normalize the abstraction process by using a database of captured motions where human body joints are accurately localized in 3D using motion capture devices. In this way, we can focus on the second stage where the spatial motion of human body parts are connected to the query of interest. Specifically, we perform experiments on the **CMU Motion Capture database (CMU Mocap)** [1].

3.2 CMU Motion Capture Database

CMU Mocap is a large-scale motion capture dataset of open-ended activities. It contains 2,548 motion capture videos from 113 actors performing 1,095 unique activities captured at 120 frames-per-second with descriptions in text. There are activities with different styles and transitions such

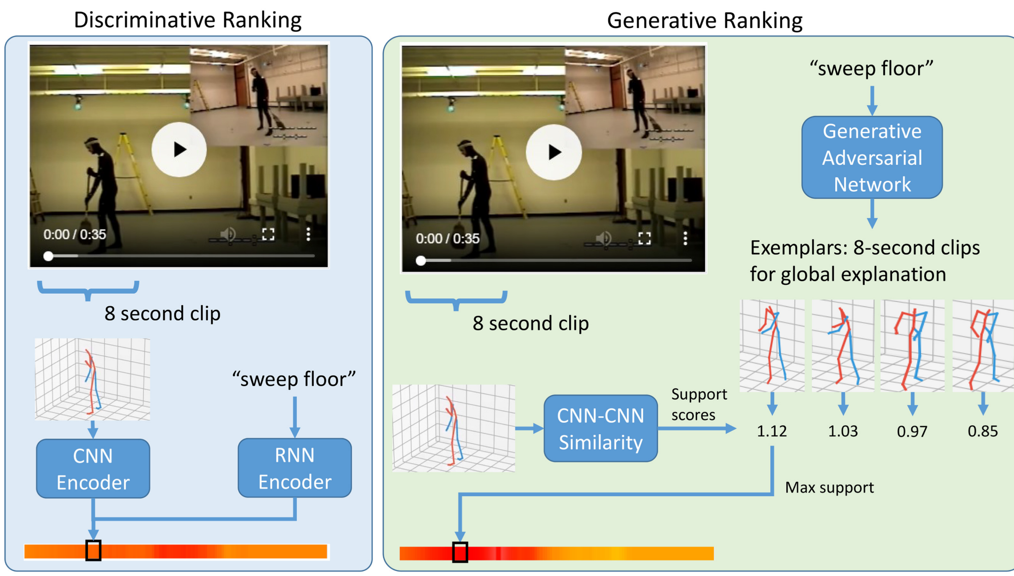


Fig. 1. Illustration of discriminative ranking vs. generative ranking. (left) Given a fixed-length video clip in the video database and a query "sweep floor," discriminative ranking uses a CNN-RNN model to score the clip. (right) Generative ranking first generates exemplar clips of what the model thinks is "sweep floor" then uses the clips to score the video database through a CNN-CNN similarity function. The score is indicated by the confidence bar, where red indicates a higher level of confidence.

as "walk on uneven terrain," "dance—expressive arms, pirouette," "punch and kick," and "run to sneak." Having such fine-grained activities brings us close to human activity understanding in the wild, and CMU Mocap is the largest dataset of its kind. We use data in the BVH format provided by Reference [25], where the human body skeletons are represented by 31 joints, closely following Reference [37]. The joint angles are pre-processed into the exponential map representation and activities spanning less than 8 s are filtered out. The filtered dataset contains 573 actions across 1,125 videos totaling 8 h. We use 757 videos for training the AI retrieval system and 368 videos for evaluation.

Our AI retrieval systems spot query activity using frame-by-frame sliding windows of 8-second clips. We compare two state-of-the-art deep models: (1) a discriminative ranking model that does not provide global explanation and (2) a generative ranking model that provides exemplar-based global explanations about the retrieval decisions. The discriminative ranking and generative ranking systems are illustrated in Figure 1.

3.3 Discriminative Ranking Implementation

The discriminative ranking model is inspired by the state-of-the-art model for ranking image captions [18]. Given a query and a video clip, discriminative ranking computes a ranking score of how well the clip matches the query. The input action text is encoded into a query vector using the skip-thought vectors model [32], and fed to a GRU **recurrent neural network (RNN)** language model. The input video clip is encoded into a video clip vector using a 1D residual **convolutional neural network (CNN)**. The matching score between the query and video clip is computed as the dot product between the query vector and the video clip vector. The AI retrieval system selects videos with highest average score over all its sliding window clips as the output. The ranking scores for

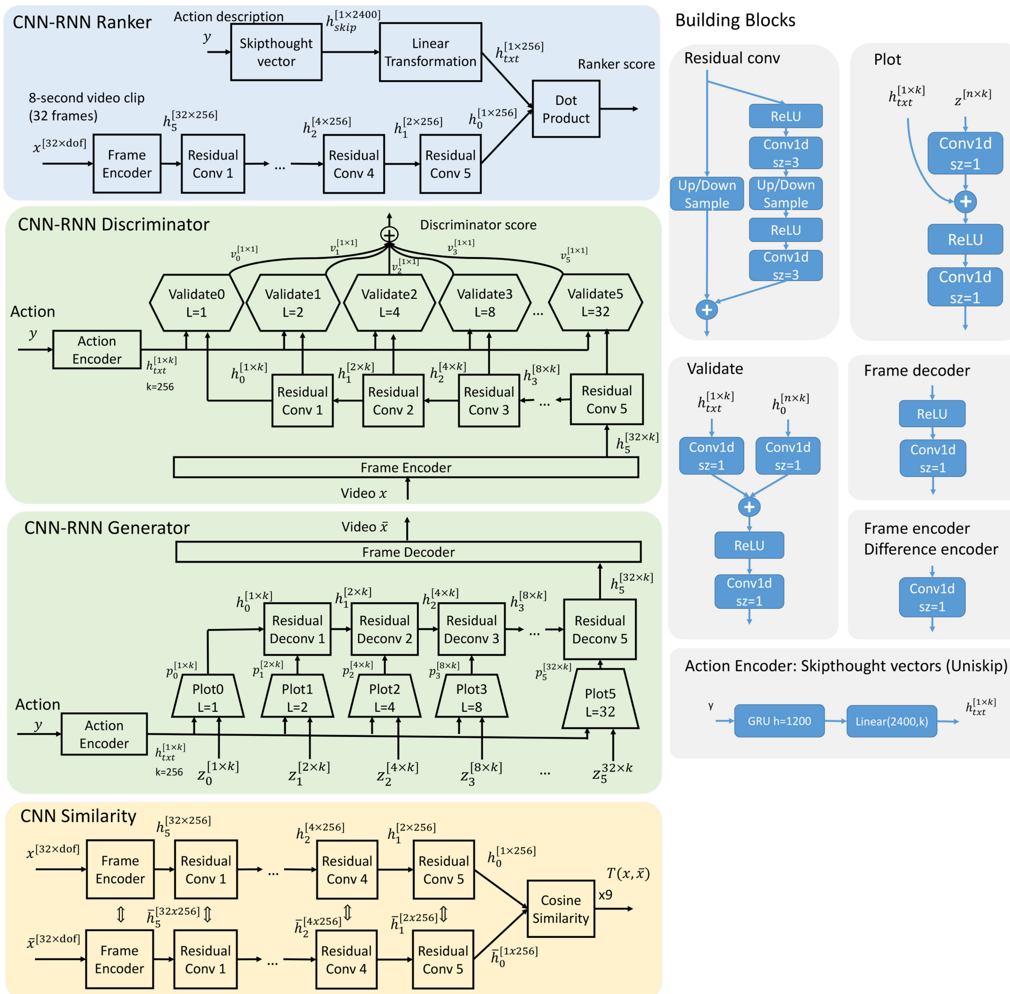


Fig. 2. Model architecture diagrams of (1) Top: CNN-RNN for discriminative ranking. (2) Middle: CNN-RNN generators and discriminator following Reference [37] for the GAN component in generative ranking. (3) Bottom: Siamese CNN similarity function for the similarity component in generative ranking. The basic building blocks for the model architectures are shown on the right. All models are implemented in the PyTorch framework.

sliding window clips are visualized to the user to explain the retrieval decisions. Figure 2 (top) illustrates the architecture of the discriminative ranking model.

The discriminative ranking model is trained jointly for action classification and retrieval of human activity videos. The action classification task is given a fixed-length video clip, retrieve its original description from a pool of $K = 250$ descriptions. Similarly the action retrieval task is given an action description, retrieve the video clip that corresponds to the action description from a pool of $K = 250$ fixed-length video clips. We optimize the negative log-likelihood action retrieval and action classification losses to learn parameters of the CNN. We use the Adam optimizer with learning rate 1×10^{-4} over 100 epochs.

3.4 Generative Ranking Implementation

Given a query, the generative ranking model first generates exemplar clips using a text-conditioned GAN [37]. For example, for a query "walking," the model will generate a set of exemplar clips that it "thinks" represent the "walking" action. Given a video clip, the generative ranking model computes the ranking score by comparing the video clip with the exemplar clips using a learned similarity metric. Similar to discriminative ranking, videos with the highest average score over all sliding window clips are retrieved as the output. This particular method is ideal for our interface that allows the users to query by natural language, as this is the only GAN-based approach that requires a text query to generate a video clip without example frames as an additional input. Additionally, our generator performs well against an existing approach that requires such frames [37].

For generative ranking, the generated exemplar clips are global explanations of the retrieval decisions. The generated exemplar clips are presented to the user along with the ranking scores to explain the retrieval decisions. While we acknowledge that the diversity of the generated examples is also important for GAN-based approaches, we also recognize that some users may want a more "deterministic" explanation. The quality of these generated clips has been assessed as part of the user study by external data annotators.

3.4.1 Generative Adversarial Network (GAN) for Text-conditioned Activity Generation. GANs [22] consist of two components that learn as adversaries: a generator and an discriminator. Let video clip be x and query be y. In the context of video generation, given a query "sweep floor," a discriminator $D(x, y)$ is a video appraiser that tries to tell if a video is an authentic "sweep floor" video in the training set, or a generated counterfeit made by the generator. A video generator $\bar{x} = G(y, z)$, however, starts from a random Gaussian noise vector z and transforms this using a feedforward neural network to generate realistic enough "sweep floor" videos that fool the discriminator. $\hat{x} = \alpha x + (1 - \alpha)\bar{x}$, where $\hat{x}$ is derived from $\bar{x}$ and x with α uniformly sampled between 0 and 1. $\hat{x}$ takes an interpolation between a random real sample and a random generated sample.

Our activity generation GAN is trained uses the Wasserstein GAN [3] objective function (Equation (1)), which is the average score assigned to real videos of the target action, minus the average score for generated videos, plus a gradient penalty term for stabilizing optimization. The discriminator maximizes while the generator minimizes the objective. Parameters of the discriminator and generator are learned through alternating gradient descent, in which the discriminator learns to improve its classification performance, followed by the generator learning to improve the quality of the videos it creates. Reference [3] proves that when Equation (1) reaches equilibrium, the generator will generate the real video distribution $P(x|y)$ and the discriminator will not be able to tell generated videos from real videos:

$$\begin{aligned} J_{GP}(D, G) \quad &= \quad \underbrace{\mathbb{E}_{\mathbf{x}\sim p_x, \mathbf{y}\sim p_y} D(\mathbf{x}, \mathbf{y})}_{\text{Real}} - \underbrace{\mathbb{E}_{\bar{\mathbf{x}}\sim p_G, \mathbf{y}\sim p_y} D(\bar{\mathbf{x}}, \mathbf{y})}_{\text{Generated}} \\ &+ \quad \underbrace{\lambda \mathbb{E}_{\hat{\mathbf{x}}\sim p_{\hat{x}}, \mathbf{y}\sim p_y} \left[(\|\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}}, \mathbf{y})\|_2 - 1)^2 \right]}_{\text{Gradient Penalty}}. \end{aligned} \tag{1}$$

Following Reference [37], the GAN discriminator is a 1D residual CNN with dense validation blocks, and the GAN generator is a 1D residual deconvolution CNN. The architectures are shown in Figure 2 (middle). We train the model on the CMU Mocap training set, using the Adam optimizer [31] with learning rate 1×10^{-4} for 1,000 epochs.

3.4.2 Generative Ranking Using Generated Exemplars. The second stage of the generative ranking system computes matching scores between the set of exemplar video clips $\{\bar{x}\}$ generated for query y and a target video clip x.

Our generative ranking approach computes as matching score the **point-wise mutual information (PMI)** $\mathrm{PMI}(x, y) = \frac{P(x,y)}{P(x)P(y)}$ between query y and video clip x it captures how often the video and the query are seen together. In addition, our generative ranking approach computes $\mathrm{PMI}(x, y)$ using only $\{\bar{x}\}$ and x and without using y, to guarantee that the set of exemplars $\{\bar{x}\}$ faithfully explain the decisions. Notice that if we can learn $T(x, \bar{x}) = \log \frac{P(x,\bar{x})}{P(x)P(\bar{x})}$, we will have:

$$\begin{aligned}
\log \frac{P(x,y)}{P(x)P(y)} &= \log \sum_{\bar{x}} \frac{P(x|\bar{x}, y)P(\bar{x}|y)}{P(x)} \\
&= \log \sum_{\bar{x}} \frac{P(x|\bar{x})P(\bar{x}|y)}{P(x)} \qquad \text{(Assuming } x \perp\!\!\!\perp y | \bar{x}\text{)} \\
&= \log \mathbb{E}_{\bar{x} \sim P(\bar{x}|y)} \frac{P(x|\bar{x})}{P(x)} \\
&= \log \mathbb{E}_{\bar{x} \sim P(\bar{x}|y)} e^{T(x,\bar{x})}, \qquad (2)
\end{aligned}$$

which shows that the matching score $\mathrm{PMI}(x, y)$ can be computed as the log-mean-exp (a soft version of max function) of $T(x, \bar{x})$ over target video clip x and exemplars $\{\bar{x}\}$.

The key function, $T(x, \bar{x})$ can be approximated using a neural network learned by maximizing **mutual information (MI)** lower-bound Equation (3) discovered by References [6, 41]:

$$\mathbb{I}(X; \bar{X}) \geq \max_T \mathbb{E}_{(x,\bar{x}) \sim P(x,\bar{x})} T(x, \bar{x}) - \mathbb{E}_{x \sim P(x)} \mathbb{E}_{\bar{x} \sim P(\bar{x})} e^{T(x,\bar{x})} + 1. \qquad (3)$$

Equation (3) equality is reached when $T(x, \bar{x}) = \log \frac{P(x,y)}{P(x)P(y)}$.

For the model architecture of $T(x, \bar{x})$, we use a residual 1D CNN with shared parameters to encode a pair, x and $\bar{x}$ (i.e., a Siamese-CNN) into vectors. $T(x, \bar{x})$ is computed as the cosine similarity scaled by a constant factor of 9.[1] The CNN-similarity network architecture is shown in Figure 2 (bottom). We optimize Equation (3) on the CMU Mocap training set using the Adam optimizer [31] with a learning rate 1×10^{-4} for 100 epochs and apply $T(x, \bar{x})$ on the test set for predicting $\log \frac{P(x,y)}{P(x)P(y)}$.

In summary, for generative ranking, given query y, we first use the GAN generator $\bar{x} = G(y, z)$ under different noise vectors z to generate $N = 30$ exemplars $\{\bar{x}\}$. The decision to use 30 exemplars is a balanced one between two considerations: diversity and retrieval performance for machines and usability for human users. While it is important to have more exemplars for diversity and subsequent machine performance, we also recognize that human users cannot digest an overwhelming number of exemplars.

Thereafter, we compute log-mean-exp over the $N = 30$ $T(x, \bar{x})$ matching scores between video clip x and every exemplar $\bar{x}$ to estimate $\mathrm{PMI}(x, y)$ as the ranking score.

3.5 Machine Learning Performance

We benchmark both approaches, based on the same CNN model, by their top-1 accuracy when ranking 248 8-s video clips with unique action descriptions in the CMU Mocap test set. Random ranking is $1/248 \approx 0.4\%$. The discriminative ranking approach achieves 35% top-1 accuracy, while

[1]The output range of $T(x, \bar{x})$ affects confidence of MI estimation [6]. Empirically, this controls the output range of $T(x, \bar{x}) \in [-9, 9]$ and makes optimization more stable.

the generative ranking approach achieves 33% top-1 accuracy. Therefore their performance is comparable. For global explanation, discriminative ranking provides only the matching scores, while generative ranking, in addition to scores, naturally generates exemplar-clips for the query that can be shown to the user. Finally, generating 5, 10, and 20 exemplars resulted in top-1 retrieval performance to drop by 1.6%, 1%, and 0.5%, respectively, compared to 30 exemplars.

We found that using the same CNN model to generate vectors of real and generated structural motion data gives 3% better top-1 retrieval performance than using two separate CNNs. Our hypothesis is that the generated clips are qualitatively very similar to real structural motion data, and therefore using a single CNN model is sufficient. Less parameters empirically reduces the gap between performance on training/testing data in machine learning, and hence, we predict that generalization performance is improved, because a single CNN model has less parameters than two separate CNN models.

On our evaluation dataset with 368 videos and 248 unique action keywords, we computed the precision-recall curves for each keyword and the resultant mean average precision (mAP)—area under the precision-recall curve averaged across all keywords—for both approaches. The discriminative approach reaches 0.457 mAP and the generative approach using 30 exemplar clips achieves 0.425 mAP. The generative approach using 1, 2, 5, 10, 20 exemplar clips achieve 0.404, 0.413, 0.421, 0.422, 0.423 mAP, respectively. Observations are consistent with our top-1 accuracy metric, that discriminative ranking performs slightly better than the generative counterpart. We conclude that the performance gap is small, and using more exemplar clips improve performance.

4 EXPLANATION INTERFACE

The interface acts as a mediator between the human and the AI, to help understand the AI's rationale for decisions through a variety of explanation approaches. The explanation interface combines visualization of generator instances from the AI generator, as well as uncertainty in the outcomes from the ranker. We set out to bridge the gap and make the explanation interface appropriate for people who are not AI experts. To achieve this, the explanations consist primarily of relatable constructs such as animated motion sequences, rather than abstract visualizations of hidden model states or other low level features. We adopt a surveillance use case where our target user is an analyst searching for an activity in a large video database.

The explainable interface assumes a dashboard design geared towards domain experts who may not have deep knowledge of artificial intelligence, but are well-versed in the data upon which the system operates. Purpose-built to visualize AI rationale for individual video clips in the dataset, the interface is also designed to handle similar application areas and accompanying datasets, enabling users to answer textual queries with visual rationales and confidence scores for the answers. The interface also allows a data domain expert to observe the evidence and make an informed decision whether to trust the XAI system, by supporting "drill-down" into deeper evidence for the provided answers and the level of confidence the model is reporting. A detailed illustration of the interface is available in Figure 3.

The interface enables the user to input a search query and select a ranking algorithm (*Generative* or *Discriminative*). Once the search button is clicked, all the videos in the database are ranked in a new list with the most relevant video on top. Once one of the videos is selected, the video becomes available for the user to view, along with a *confidence bar* visualization located below the navigation bar, showing areas of interest where the query term is most likely to occur. In the case of generative ranking, in addition to the confidence bar, the user is also presented with an unsorted set of *generated evidence* that the algorithm used to rank the list of videos. Upon clicking on the confidence bar in a certain segment, the score for that segment is presented to the user, and the list of

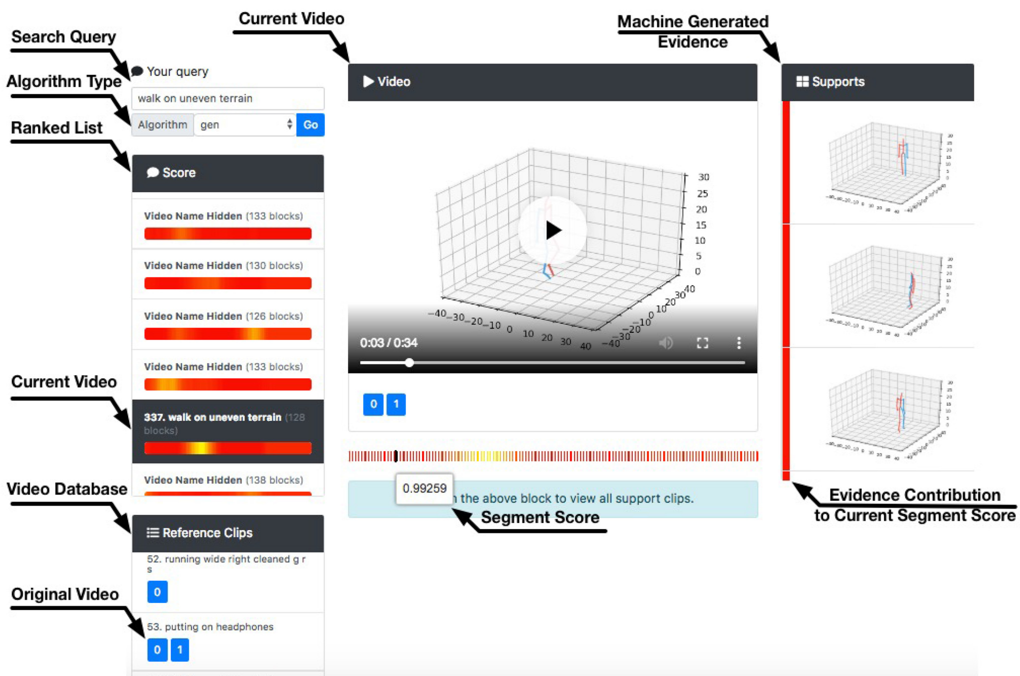


Fig. 3. Annotated view of the explanation interface. The flow of the interface starts from left to right. In the left column, the user starts by inputting a query and selecting the ranking algorithm of interest. Once the search button is clicked, all the videos in the database are ranked in a new list with the most relevant video at the top. In the middle column, once one of the videos is selected from the ranked list, the video becomes available for the user to view, along with a segmented confidence bar (red indicating higher confidence) located below the navigation bar. In the right column, which is available only in the generative ranking case, the user is also presented with an unsorted set of generated evidence that the algorithm used rank the list of videos. Upon clicking on the confidence bar in a certain segment, the score for that segment is presented to the user, and the list of evidence gets sorted showing the most important evidence first. Upon clicking on one of the evidence clips, a pop-up video player is presented to the user to view it.

evidence is sorted to show the most important evidence first. Upon clicking on one of the evidence clips, a pop-up video player is presented to the user to review the generated evidence in detail.

The generative ranking interface enables the user to drill down into deeper evidence for the provided answers and see the level of confidence the model is reporting. This allows a data domain expert to make an informed decision whether to trust the system. In the discriminative ranking case, the user would have little evidence to support the provided decision. The ability to drill down on evidence ends at the ranked list and confidence bar. One of the key research questions of our user study, presented in the next section, is the investigation of the role of trust and whether generated evidence engenders appropriate trust in AI systems: how will an explanation interface lead a user to accept or reject the answer provided by the AI model?

5 USER STUDY

The user study consists of various components to assess different factors from a mental model to trust and reliance. Designed as a more linear, guided variation of the explanation interface, the study presents numerous instances of three main tasks:

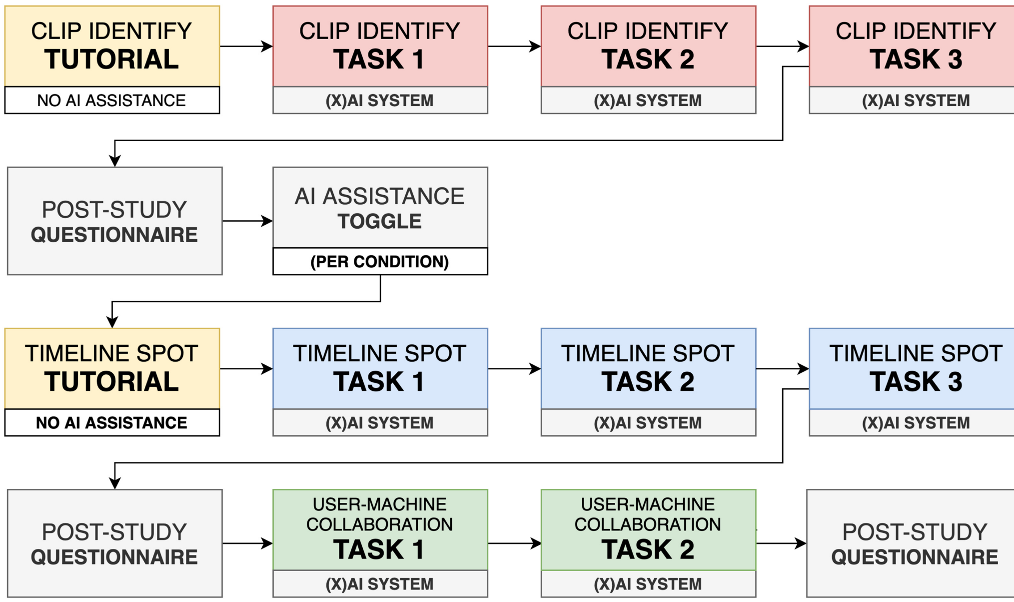


Fig. 4. Illustration of the user study session flow. Throughout each session, the user performs numerous trials of three distinct tasks, each accompanied by a post hoc questionnaire specific to the task and the AI system's performance. Depending on the allocated experimental condition, the session may switch to the alternate AI system without warning.

(1) Identifying one or more video clips that best illustrate the displayed query.
(2) Spotting one or more segments in a single video clip that best illustrate the keyword.
(3) Collaborating with AI to solve a more complex challenge of identifying a longer video clip that best illustrates a complex query with multiple actions.

Through comparing two conditions—the XAI system or a black-box AI system, powered by generative and discriminative models, respectively—we aim to assess the benefit, if any, of XAI. The study also rigorously records the user's subjective experience with questionnaire components after each task. The web-based study interface is available for public access at http://gr.ckprototype.com/.

5.1 Objectives

Our three-stage user study sets out to evaluate the model, the interface and the benefits of using the XAI system. Featuring a variety of interactive modules, this web-based interface was refined through an internal pilot and was deployed as part of a randomized controlled study, the results of which we report in this article.

Hypothesis. We hypothesize that the explanation interface will facilitate the user's understanding of the XAI system's behavior, while improving the user's task performance by building a correct mental model of the AI and establishing appropriate trust and reliance on the system.

Mental Model. A successful XAI system should allow users to gain a better understanding of the system's behavior, thus building a correct mental model of its operations. In this study, we use a series of prediction tasks and questionnaires to better understand the benefits of using an XAI system over a black-box one for mental model formation. Prior to gaining access to the XAI

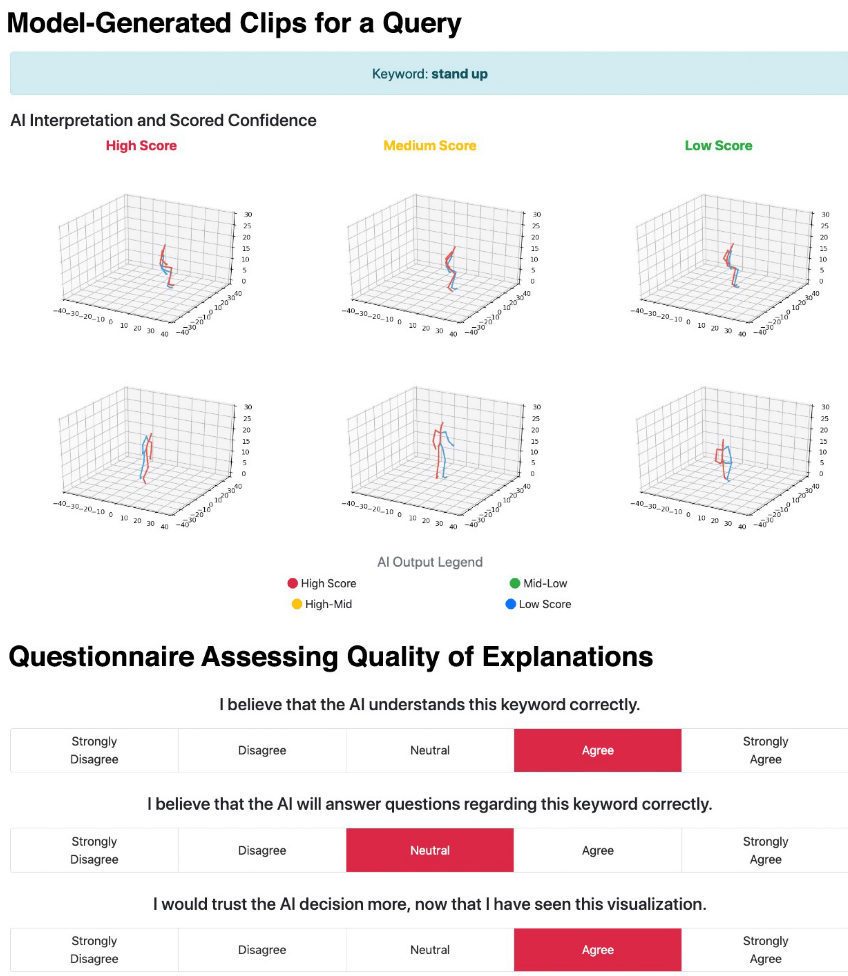


Fig. 5. The XAI evidence screen. This is the first step towards assessing the mental model of the user. Before each task, the participant is presented with a query and a sorted list of generated clips depicting what the XAI thinks the query visually looks like. The participant is then prompted to answer a brief survey about their expectations of the XAI's ability to answer correctly and its understanding of the query.

system's assistance, the user is presented with a sorted list of XAI-generated clips that illustrate how the system interprets the query. They are then asked to fill a short questionnaire to assess their expectations of the system as shown in Figure 5. Given the presented clips, the user is asked to predict the decision of the AI on a specific task. The user's work is then compared to that of the XAI system to gauge whether the user was able to predict the system's behaviour. In addition to prompting the user with prediction tasks, the study also presents a number of assertions about the XAI system and asks the user to agree or disagree with the assertions. This way, their mental model is compared with an ideal model of the XAI system.

Task Performance. The user's task performance alongside the AI system is measured by comparing the resultant output with ground truth and observing the user's acceptance of the system.

Given a task, the user has an option to view and use the AI system's output as the user's own. The study monitors the user's decision to accept the system's assistance, and examines the similarity between ground truth, the system output, and the user's answer. Other miscellaneous parameters, including the user's task completion time and interaction with different interface modules, are recorded for further analysis.

Appropriate Trust and Reliance. Explanations should help users to develop more appropriate trust and reliance toward an XAI system and enable users to better achieve their goals. Maintaining close ties with the user's mental model of the AI system and resultant task performance, the study measures the user's trust and reliance by asking the user to assess the level of confidence for the XAI system's output. The user can select an answer from a 5-level Likert scale, ranging from "Strongly Disagree" to "Strongly Agree," to indicate the user's confidence in the XAI system. The study also measures the user's reliance on the system by examining whether the user solicits the XAI assistance (through interaction log files) and continues to use it for subsequent tasks.

5.2 User Study Design

The study deconstructs the explanation interface into modular, guided user experiences to evaluate the benefits of using the XAI system over the traditional AI counterpart. Featuring numerous instances of three distinct tasks—*Clip Identify*, *Timeline Spot*, and *User-Machine Collaboration Task*—the study offers either the AI or XAI system to assist each participant along the way.

Overview. The study is a three-part experience featuring 2 different modes of AI assistance and 3 task types (3+3+2 repetitions) for a total of 8 AI-assisted tasks based on more than 40 different, randomly sampled configurations. The study is designed to take a maximum of 50 min to complete, and each prompted task is accompanied by Likert-scale questionnaires designed to record the user's subjective experience with the task. The study design and instructions were pilot tested with colleagues and students who were not part of the participant pool.

Participants. A total of 44 undergraduate students from computer science and information technology disciplines were recruited to participate in the between-groups study. The participants had no prior experience with XAI systems but had used commercial video search tools (e.g., YouTube). As there is no evidence that gender or age would be relevant factors, this information was not collected. Participants were compensated $20 for 1 h of their time at the end of each session.

Experimental Design. There were two conditions and three tasks. To alleviate order effects due to participant fatigue and practice with AI assistance, the following conditions were established:

(1) AI system only
(2) XAI system only
(3) AI system, then switch to XAI system halfway
(4) XAI system, then switch to AI system halfway

Each study session, dedicated to a single condition, was initiated with a brief introduction to the procedure and a tutorial about the study interface. Twelve participants were invited to each session, with at least ten participants successfully completing each condition, and no participants engaging in more than one condition. Participants worked individually on computers within a computing lab. The experimenter was available to answer participant questions throughout the session. Each participant received detailed training prior to beginning the study session, including completing the aforementioned sample tutorial tasks and watching video recordings that illustrate ideal interaction scenarios.

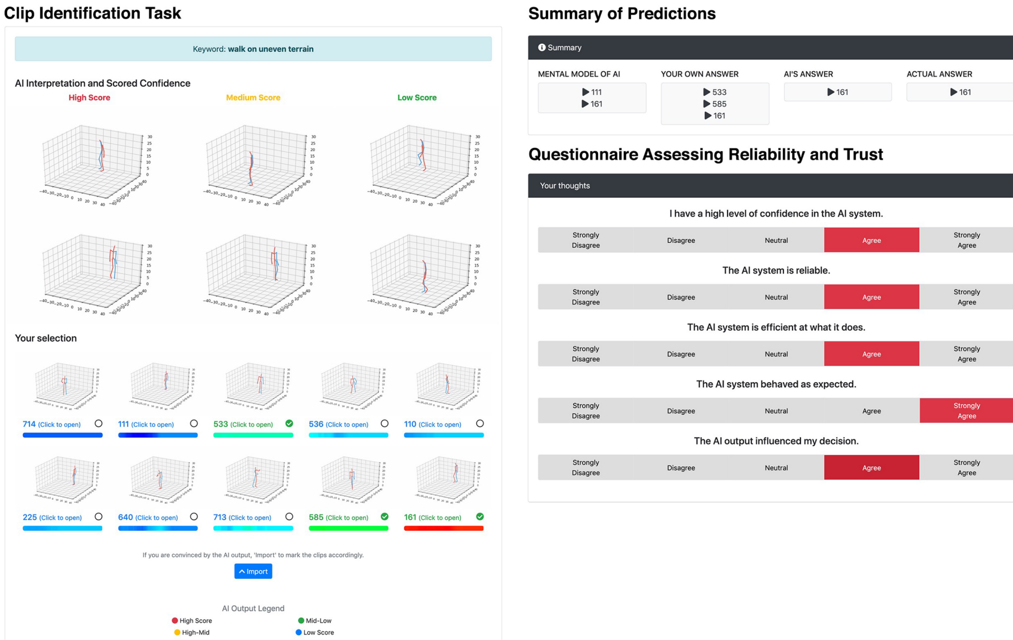


Fig. 6. (left) The *Clip Identify* task with the XAI system. Given a set of ten video clips, the user is asked to pick the top three most relevant clips to the displayed query. The XAI system (shown), unlike the AI system, provides assistance with explanations using model-generated clips. (top right) A summary showing the correct answer, the user's answer, the user's prediction of the system's answer (mental model of AI), and the system's answer. (bottom right) A questionnaire per trial assessing the performance of the system.

General Structure. The study presents a number of text queries, accompanied by user tasks specific to each part, as well as applicable (X)AI assistance and questionnaire components. At the end of each task, the study also displays the summary that compares ground truth to user- and AI-provided answers. A detailed outline of the study is shown in Figure 4.

Part 1: Clip Identify. In this first part of the study, the participant is prompted to investigate a given set of ten video clips and pick up to three clips most relevant to the displayed keyword using a drag-and-drop sort interface as shown in Figure 6 (left). During this stage, the participant is presented a total of four trials with randomly selected keywords and associated identification tasks. The first trial serves as a tutorial and is not included in the results.

For each trial, the participant is first presented with the mental model questionnaire and asked to predict the system's answer. In the XAI condition, the mental model questions are accompanied by the sorted list of generated clips of what the system "thinks" the query looks like, as shown in Figure 5. After the mental model questions, the participant is presented with the task along with the results from assigned (X)AI assistance. The system's assistance is provided through sorting the list of clips, along with a confidence bar below each clip, showing the system's score over the length of the clip. The participant may view any clip during each task, and optionally import the system's suggestion as the solution. In the case of XAI, the user is also presented with AI-generated clips as evidence supporting the XAI system's interpretation of the text query. The evidence clips are rearranged automatically once a video is selected, according to which generated clips contributed

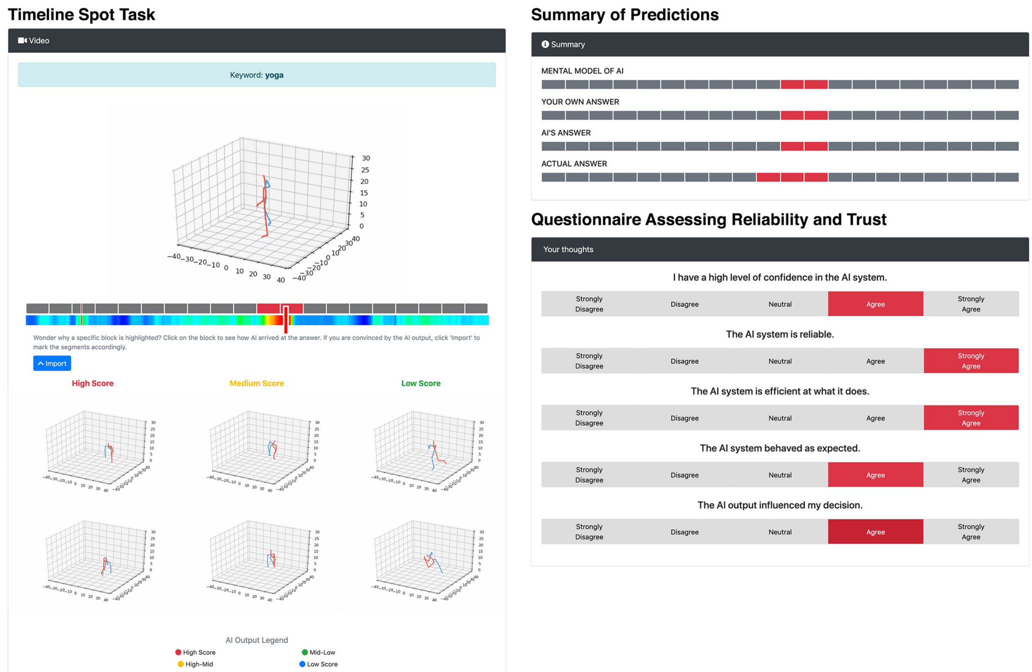


Fig. 7. (left) The *Timeline Spot* task with the XAI system. Given a long video, the user's task is to highlight segments where the activity described by the given keyword query exists. The XAI system, unlike the AI system, provides assistance with explanations using model-generated clips. (top right) A summary of timelines with the location of the correct answer, the user's answer, the user's prediction of the system's answer, and the system's answer. (bottom right) A questionnaire per trial assessing the performance of the system.

the most to the system's decision. At the end of the trial, the participant is presented a summary showing the correct answer, their answer, their prediction of the system's answer based on their mental model, and the system's answer, as shown in Figure 6 (top right). Finally, the participant completes a questionnaire for this specific trial, assessing the performance of the system as shown in Figure 6 (bottom right).

Part 2: Timeline Spot. The second part of the study provides a single but lengthier video clip, consisting of multiple individual clips as shown in Part 1, to localize a specific activity. Presenting a single keyword in the same fashion as the first part, the study prompts the participant to search for different parts of the video that best illustrate the keyword. The user can play or scrub the video to locate the parts that match the keyword, and mark them using the timeline interface as shown in Figure 7 (left). AI assistance is once again available for the user to consult, complete with AI-generated clips exclusive to the XAI system. Once the user clicks on the confidence bar below the clip, the supporting evidence consisting of generated clips is sorted automatically to show the most contributing evidence to a specific time segment. The trial structure mirrors the Clip Identify task, with a summary of results as shown in Figure 7 (top right), and a questionnaire as shown in Figure 7 (bottom right).

Part 3: User-Machine Collaboration Task. Combining interface elements and challenges of parts 1 and 2, the third and final part of the study provides a large set of lengthier video clips and

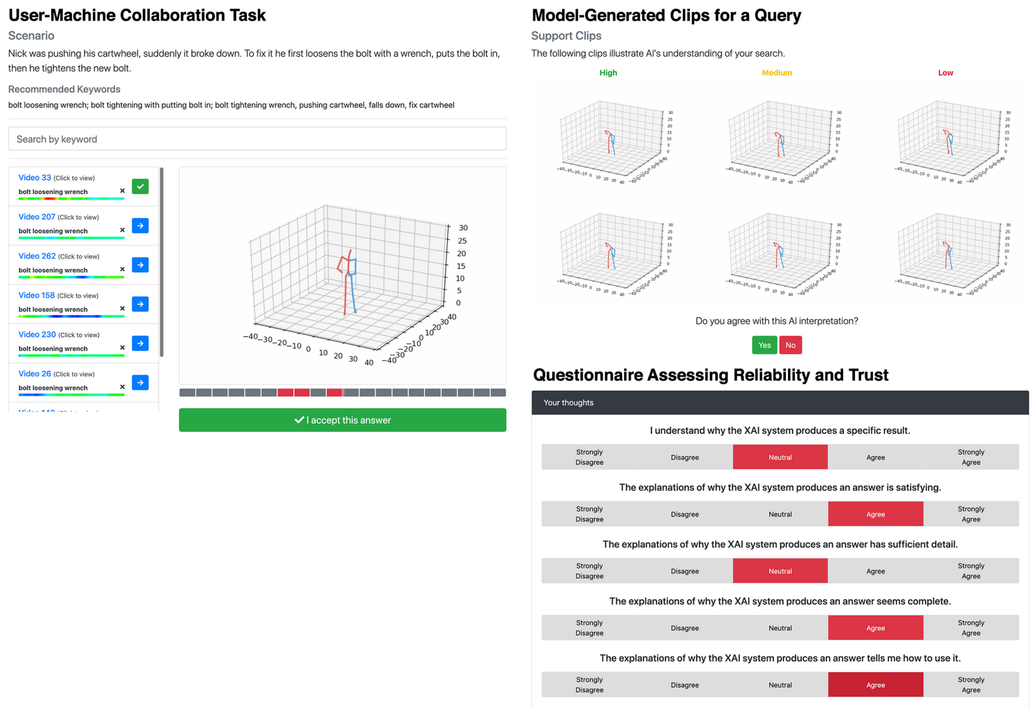


Fig. 8. (left) The *User-Machine Collaboration* task with the (X)AI system. Given a series of long video clips and a more complex scenario, the user's task is to select a video that best represents the text description. The interface features a search box that allows the user to consult the (X)AI system to facilitate the investigation. (top right) The XAI system, unlike the AI system, provides assistance with explanations using model-generated support clips. (bottom right) A questionnaire per trial assessing the performance of the system.

prompts a randomly selected scenario. Each of the seven scenarios was manually constructed by concatenating previously available motion-capture clips. The user is encouraged to deconstruct the provided description and search for the video clip that best illustrates the scenario, but we suspect the task will be overwhelming enough for the user to request AI assistance as required. Illustrated in Figure 8, the interface provides a total set of three main elements: the search box, the clip list, and the video player. The user can independently browse and investigate the individual video clips to complete the task, but is encouraged to use the AI system to facilitate the investigation. Upon submitting one or more text queries, the AI system will highlight the clips that are most likely to illustrate the user's query. The XAI system, in alignment with its behavior in parts 1 and 2, presents its interpretation of the query through generated supporting evidence before the user accepts AI assistance in sorting the video clips. Finally, the user must continue the investigation until the correct video clip is selected, and then is presented a summary of results and a questionnaire.

5.3 External Data Annotation

In addition to collecting participant reactions to model-generated video clips as part of the study, we recruited five external data annotators and launched a post hoc analysis of AI explanation quality. Each annotator was presented a series of keyword queries and corresponding AI explanations, and asked to rate how well the model-generated video clips represent each query, on a

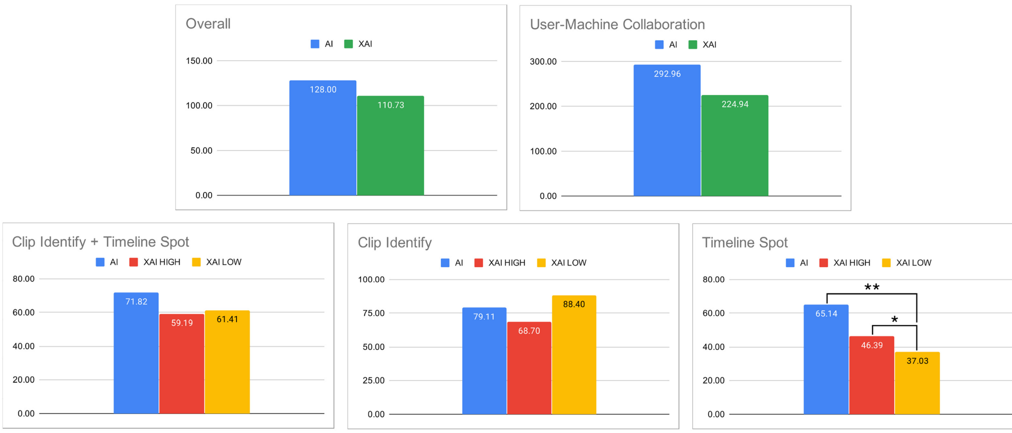


Fig. 9. Task completion times between the two AI systems, and across the three task clusters. Segmentation between XAI HIGH and XAI LOW was determined by the user's express level of trust in AI explanation. UMC completion times using the XAI system remain unsegmented, as users did not provide express level of trust for corresponding AI explanation.

5 point Likert scale. These ratings, collected from annotators with no prior experience with the study, were used as a proxy for the quality of AI explanations as well as an indicator of participant attentiveness throughout the study.

6 OUTCOMES

Participant activities recorded during each session have been collated and thoroughly analyzed to test the original hypothesis that the XAI system will facilitate the user's understanding of the AI system and in turn improve the user's task performance. Any other notable insights that arose during this process have been also been collected for discussion below.

Measures. With a total of 44 participants engaging in more than 350 distinct AI-assisted tasks, the collected data features completion time, user and AI accuracy, and user reaction to the AI system or AI-provided explanations as applicable per task. In the following discussion, we note statistical tests with * at $p < 0.05$ and ** at $p < 0.005$.

Task Clusters. Upon observing divergence in participant performance and reaction between those who explicitly stated low levels of trust in AI explanation and those who did not, the study results were further segmented into three separate groups: AI tasks (51.7%), XAI tasks completed by users with low levels of trust (XAI LOW, 16.5%), and finally, the remainder of XAI tasks where the user did not express explicit distrust or instead expressed trust (XAI HIGH, 31.8%). Segmentation between XAI LOW and XAI HIGH clusters was determined by user response to the question "I would trust the AI decision more, now that I have seen this visualization," where the "Disagree" or "Strongly Disagree" response serving as a qualifier for XAI LOW. Some task trials were discarded due to user or system error, resulting in a slight imbalance between AI and XAI task numbers.

6.1 Speed

Overview. Speed is defined as the elapsed time in completing a single task trial, illustrated in Figure 9. Speed determines the efficiency advantage of using the AI or XAI system. Adjusting for variance in internal loading and computation time for both AI and XAI systems, a typical task was completed on average in 119 s, although it is important to note that **User-Machine**

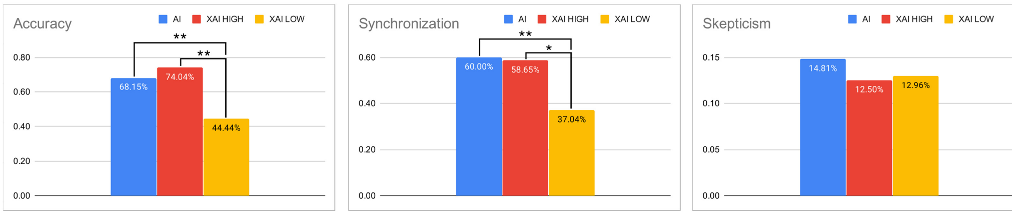


Fig. 10. Results of accuracy (left), user-machine synchronization (middle), and user skepticism (right) across the three task clusters. Segmentation between XAI HIGH and XAI LOW was determined by the users' expressed level of trust in each AI explanation.

Collaboration (UMC) tasks are more complex and hence more time-consuming for users. Excluding these collaborative tasks that require about 257 s to complete on average, the average completion time hovered around 66 s. Computation time was offset to allow for direct comparison between AI and XAI systems after the study, although we recognize that the users may have deemed computation time excessive and influential to user satisfaction with the system.

Results. Without task segmentation, the XAI system (111 s) presented negligible advantage over the AI counterpart (128 s), but more significant divergence emerged upon segmenting the XAI results by trust and task types. An ANOVA revealed no significant effect of task cluster on speed for Clip Identify (AI: M = 79 s, SD = 78 s, XAI HIGH: M = 69 s, SD = 58 s, XAI LOW: M = 88 s, SD = 70 s). Similarly, no significant effect of condition was found for the UMC tasks (AI: M = 293 s, SD = 206 s, XAI: M = 225 s, SD = 164 s). However, the Timeline Spot tasks varied significantly (**, $F(3,126) = 5.21$, $p = 0.007$) with XAI LOW tasks being completed most quickly (M = 37 s, SD = 39 s), followed by XAI HIGH Tasks (M = 46 s, SD = 36 s) and AI tasks (M = 65 s, SD = 46 s). Post hoc pairwise t-tests with Bonferroni correction for repeated measures revealed significant differences in completion time between AI and XAI HIGH (*, $p = 0.03$) and between AI and XAI LOW (**, $p = 0.002$). There was no significant difference between XAI HIGH and XAI LOW.

Discussion. The provision of XAI support did not aid in the speed of task completion for the Identify task, as participants generally viewed multiple clips in detail, irrespective of XAI support. In the Timeline Spot task, overall completion times were shorter than either UMC or Identify counterparts, indicating a simpler task overall: participants could use the XAI support to know quickly whether to accept the AI answer or at least seek the playback to the highest rated positions to check them. The UMC task was designed as a complex challenge that would maximize the support provided to participants, the results were not significantly different between the two systems, likely due to the very high variance between participants on the time to complete this task. This points to the individualized nature of the provision of evidence, and that it may be important to provide support on demand, while putting potentially distracting explanations out of the way when they are not requested or required.

6.2 Accuracy

Overview. Accuracy, depicted in Figure 10, is the portion of instances where the user, assisted by the AI system, was able to identify the correct answer in a single task trial. Accuracy determines whether the system is able to produce more correct answers than others, resulting in a less error-prone experience.

Results. The accuracy was highest for the XAI HIGH cluster (74.0%), followed by AI (68.2%) and XAI LOW (44.4%). Pairwise chi-square tests with Bonferroni correction revealed significant

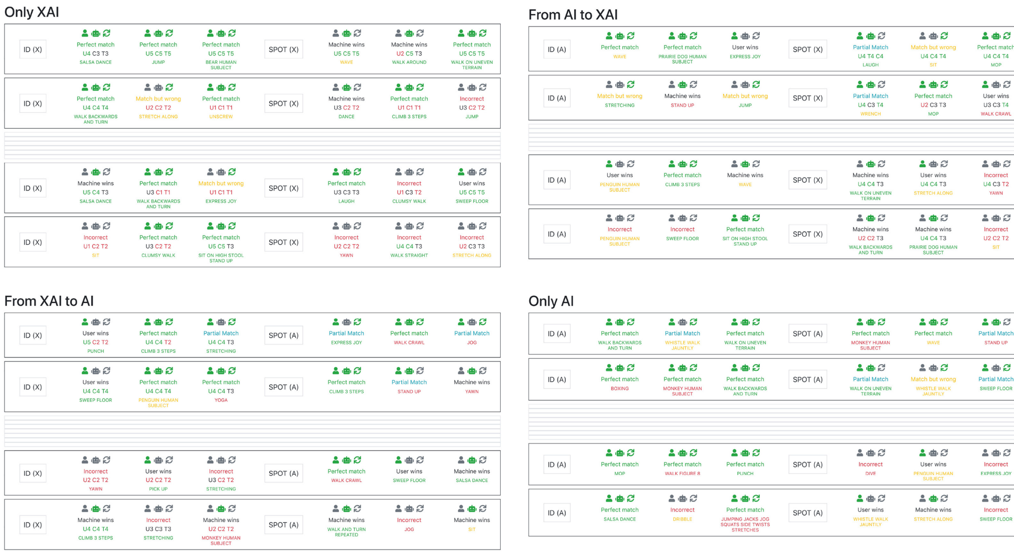


Fig. 11. Truncated summary of individual participant "journeys" through the study across different experimental conditions. Each section displays two of the most successful journeys (where both the user and the AI system were able to identify the correct answer), as well as two of the least successful per corresponding condition. The icons indicate the user's accuracy and synchronization with the AI system's interpretation, and the numbers below indicate the user's understanding, confidence, and trust pertaining to the XAI system. Queries used in each trial are also displayed, with their colors indicating the ratings by external data annotators.

differences between AI and XAI LOW (**, $\chi^2(1{,}189) = 9.14$, $p = 0.002$) and between XAI HIGH and XAI LOW (**, $\chi^2(1{,}158) = 13.50$, $p = 0.0002$). The difference between AI and XAI HIGH was not significant.

Discussion. The accuracy results were lowest when users indicated low trust for AI explanations. This may indicate that when trust is low, the user may assume the generated evidence is unreliable, and proceeded to submit their own (often incorrect) answer. This conjecture is reinforced by the fact that when trust is low, affecting accuracy, synchronization is usually also low. These results, for clarity, entirely depend on each user's interaction with the system, independent of the underlying algorithm: one can choose to accept or ignore AI assistance, regardless of the system type in use.

6.3 User-Machine Synchronization

Overview. Determined as the instance where both the user and the AI systems select the same answer regardless of its accuracy, this measure represented in Figure 10 defines the level of synchronization between the user and the AI system. As a whole, about 55% of all user and AI answers were synchronized. A sample of these journeys is illustrated in Figure 11.

Results. In strong alignment with the previous accuracy results, XAI LOW tasks resulted in a significantly lower synchronization rate of 37.40% in comparison to AI (60.0%) and XAI HIGH (58.65%) tasks. Post hoc chi-square tests with Bonferroni correction revealed significant differences between AI and XAI LOW (**, $\chi^2(1{,}189) = 8.17$, $p = 0.004$) and between XAI HIGH and XAI LOW (*, $\chi^2(1{,}158) = 6.65$, $p = 0.0099$). The difference between AI and XAI HIGH was not significant.

Discussion. AI and XAI HIGH results indicate higher user-machine synchronization than XAI LOW. This may indicate that the provision of trustworthy evidence (XAI HIGH) does not help any more than no evidence (AI), but the provision of untrustworthy evidence, such as poorly generated clips (XAI LOW) can actually drive participants away from AI suggestions. This is in fact the desired result, as we hope that users will appropriately choose to find their own answers when they do not trust the AI system to do the job.

6.4 User Skepticism

Overview. Whenever the user decides that the AI system's assistance is unhelpful and even incorrect, the user may explicitly exhibit a level of skepticism, illustrated in Figure 10, by choosing a correct answer despite the AI system's invalid suggestion. About 14% of all tasks reflected this rare but consistent behaviour.

Results. There was no significant deviation to trend across the three clusters, with AI, XAI HIGH, and XAI LOW tasks exhibiting evidence of skepticism 14.8%, 12.5%, and 13.0% at a time, respectively.

Discussion. User skepticism serves as a proxy measure of user attention to the task, indicating that the participants sometimes went against AI suggestions and did not blindly accept them. This phenomenon was consistent across all task clusters, and there was no correlation between this behaviour and expert ratings per clip.

6.5 Questionnaire Responses

Observation. While the AI-only system originally seemed to yield higher overall satisfaction amongst the participants, there was a sharp divide in satisfaction between the participants with a high level of trust and reliance for the XAI system compared to those without. Upon segmenting the responses from the XAI system as illustrated in Figure 12, it was evident that the XAI system resulted in a more positive experience overall compared to the AI system, should the users have a high level of trust and reliance for the system.

6.6 Additional Findings

Overview. Below are some of the secondary findings that do not directly correspond our hypothesis, but are notable and warrant further investigation in future work.

Distribution of User Reactions to AI Explanation. The user study collected using the three distinct questions, to individual AI explanations: "I believe that the AI understands this keyword correctly" (UNDERSTAND), "I have a high level of confidence in the AI system" (CONFIDENCE), and "I would trust the AI decision more, now that I have seen this visualization" (TRUST). Upon visualizing these reactions, there was apparent bimodal behaviour, as illustrated in Figure 13, across all three categories, indicating that users often exhibit less ambiguous reactions to presented AI explanations. We recognize that these reactions are biased to each participant's subjective experience.

Correlation Between User Reactions to AI Explanation. Beyond the anecdotal tendency where individual users who exhibit trust in the AI system may also indicate confidence in the AI system, as illustrated in Figure 14, there was significant correlation between the user's three responses to a specific AI explanation. Post hoc multiple correlation tests revealed significant positive correlation across the board: UNDERSTAND and CONFIDENCE (**, $r(121) = 0.7979$, $p < 0.00001$), UNDERSTAND and TRUST (**, $r(121) = 0.7777$, $p < 0.00001$), and CONFIDENCE and TRUST (**, $r(121) = 0.7777$, $p < 0.00001$). This, along with the shifting user reactions in Figure 14, suggest that users do actively respond to presented AI explanation and change their opinions accordingly, and

Question	AI System	XAI System High trust	XAI System Low trust
MAIN_Q1 I have a high level of confidence in the AI system.	3.37	3.95	3.02
MAIN_Q2 The AI system is reliable.	3.42	4.23	3.00
MAIN_Q3 The AI system is efficient at what it does	3.46	4.25	3.14
MAIN_Q4 The AI system behaved as expected.	3.51	4.27	3.23
MAIN_Q5 The AI output influenced my decision.	2.93	3.94	2.69
MENTAL_Q1 I have a clear understanding of how AI would answer this question.	3.71	4.09	3.09
MENTAL_Q2 I am confident that my answer would match the AI interpretation.	3.71	4.28	3.03
END_Q1 I understand why the AI system produces a specific result.	3.86	4.10	3.26
END_Q2 The explanations of why the AI system produces an answer is satisfying.	3.77	4.40	3.15
END_Q3 The explanations of why the AI system produces an answer has sufficient detail.	3.59	4.26	2.91
END_Q4 The explanations of why the AI system produces an answer seems complete.	3.63	4.15	3.06
END_Q5 The explanations of why the AI system produces an answer tells me how to use it.	3.80	4.15	3.03

Fig. 12. Summary of questionnaire responses about the (X)AI system, listing individual questions featured in the original study session. Responses pertaining to the XAI system are split into two parts, based on the overall level of trust and reliance indicated by each study participant, represented by individual responses to questions pertaining to model-generated clips. If the user expressed general lack of confidence in AI explanation, then the user was classified as "low trust." The XAI system performed better than the AI system without explanation, if the user had a higher level of trust in the system.

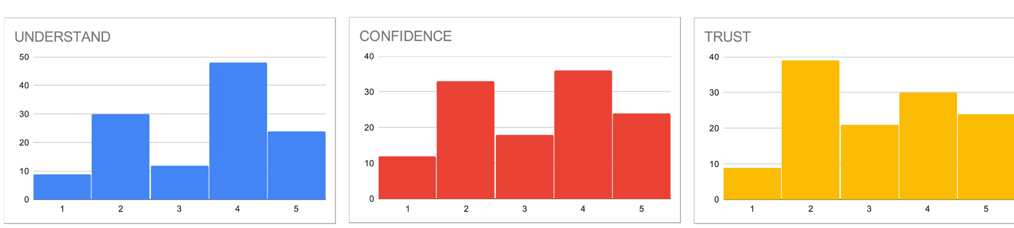


Fig. 13. Histograms illustrating user reactions to AI explanation, represented by responses to three distinct questions about AI assistance. The pattern shows a bimodal distribution showing that participants formed clear opinions for most task trials, especially on the understand and confidence questions.

that not all three questions may be necessary in future studies to measure the level of trust and reliance on the system.

Alignment with External Ratings. There was no apparent correlation between user reactions to individual clips and externally annotated ratings, as indicated in Figure 15. This may indicate that clip quality assessment criteria differed between our experts and participants, or that overall clip quality did not strongly influence the user's trust or confidence in the AI agent. It was notable, however, that positive user reactions were clustered around clips that feature exaggerated motions and cartoon-like premises, such as "bear (human subject)," "salsa dance," and "express joy," while

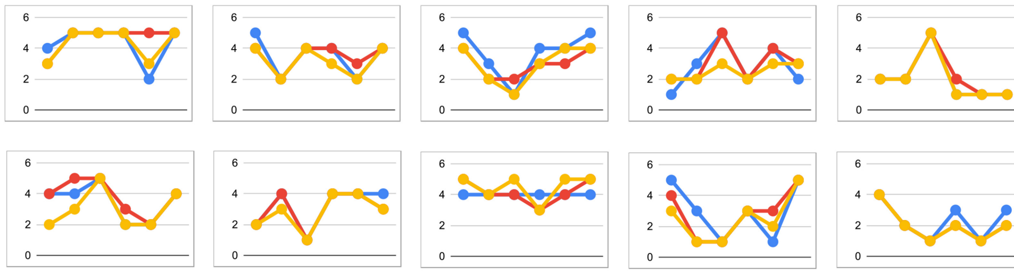


Fig. 14. Small multiples illustrating example reactions to AI explanations when using the XAI system, represented by three distinct questions: UNDERSTAND (blue), CONFIDENCE (red), and TRUST (yellow). Each plot represents a single participant across the tasks completed in the XAI experimental condition.

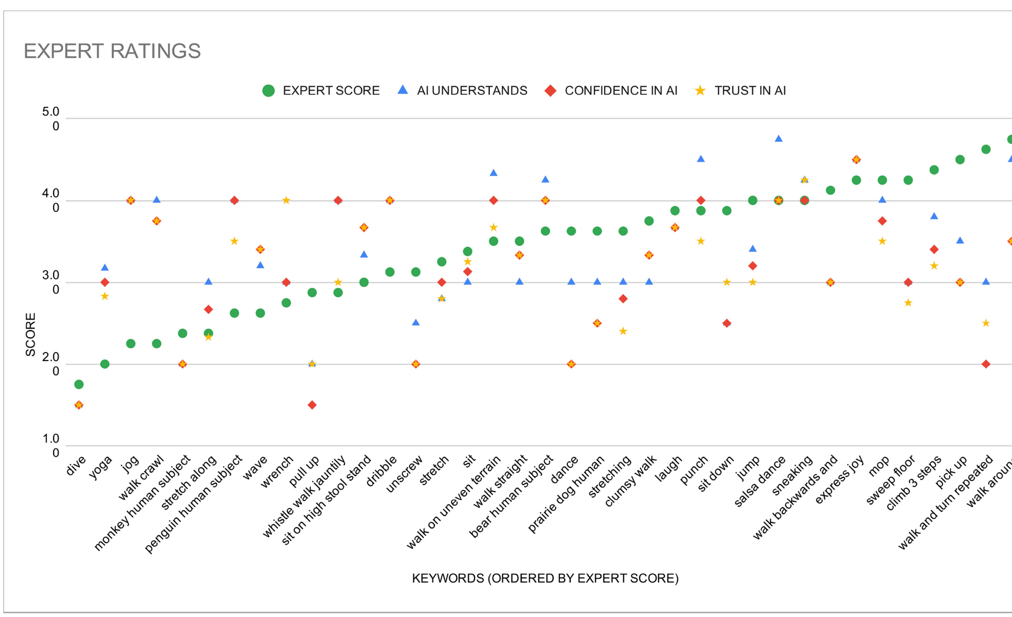


Fig. 15. Alignment between external ratings and individual user reactions to AI explanations. X axis is anchored by keywords ordered by external rating scores.

more generic and muted clips such as “pull up” and “walk and turn repeated” received negative reactions.

Participant Comments. There was divergence between clusters of participants who found the AI system to be reliable and influential to their decision-making processes, and those who deemed the system to be counter-intuitive and underwhelming. One participant wrote “(AI explanation) is a good basis of determining the reliability of AI in terms of (whether) the AI is able to detect the proper animations,” and another expressed satisfaction, stating “(I am) impressed of what the AI system outputs.” However, some expressed caution and distrust, with one writing “the AI system often interpreted small portions of movements as if they met the definition of the keyword although it was a mere segment of the movement,” and another writing “I didn’t trust it completely as it directed similar movements and categorized it as the real one.” Two participants plainly wrote

"I did not see the explanation," alluding to the possibility that the definition or qualifications of what constitutes an AI *explanation* may vary between individuals or may require additional training or clearer messaging to help people interpret generated clips as explanations.

6.7 Summary

The following insights were observed based on the above findings:

(1) The XAI system yielded a comparable level of efficiency, accuracy, and user-machine synchronization as the AI system, but only if the user exhibited a high level of trust.
(2) The XAI system yielded a significantly lower level of efficiency, accuracy, and user-machine synchronization if the user exhibited a lower level of trust.
(3) The XAI system yielded higher overall user satisfaction, but only if the user exhibited a higher level of trust.

7 DISCUSSION

Outcomes. The user study outcomes present significant evidence that the XAI system and its generative examples can facilitate task performance consistent with the AI system, offer improved performance in select task types, and provide a more satisfying overall user experience. However, this is only applicable if the users decide to trust the provided AI explanations.

We claim that the presence of AI explanations, characterized by exemplar clips and the corresponding interactive visualization, does not improve the user's performance in search tasks, but helps one to know when to trust or reject AI assistance, thus indirectly influencing performance. Additionally, the presence of such visualization helps to identify the user as belonging in one of the two groups: those who exhibit a higher level of trust and satisfaction for the system, and those exhibit skepticism and yield a lower level of efficiency and accuracy.

We observed a significant divide in behavior and performance between users who chose to trust the AI explanations and those who did not, and this divide impacted all performance-related measures including speed, accuracy, and user-machine synchronization. While there was no significant indication that the user was able to correctly accept or reject the XAI system's assistance, the results were largely comparable with the AI counterpart. These results suggest that users form trust and affinity for the XAI or AI system more or less based on instinct, and the system may produce video clips that ultimately result in correct answers, but not necessarily seem logical or comprehensible to human users. This disparity contributes to lack of perceived performance improvement.

Future Improvements. The XAI system could be further improved in numerous areas to gain a more significant advantage over the AI-only counterpart. The generative model could be improved to produce more exemplars that achieve the same or higher level of accuracy as the AI-only system. Also, the XAI system could produce more high-quality model-generated clips that best represent individual queries and result in higher user satisfaction and user-machine synchronization. Furthermore, XAI explanations must be short and require little effort to interpret, or the advantage they offer will be outweighed by the extra time and effort they require.

We also hypothesize that the evaluation dataset may contain biases that may contribute to the system behaving in a way unexpected and even jarring to the users. For users to make more accurate, informed decisions, the system will need to transparently communicate what the potential biases are, and why its decision, although less intuitive, can result in the correct answer.

There are other implications pertaining to the experiment design as well. Users may exhibit a higher level of trust and reliance for the XAI system, should the individual tasks present a higher stake and a more captivating incentive. Task formulation is an important consideration as well:

all the presented tasks in the study are simple permutations of the same dataset and the interface components, yet user performance and satisfaction noticeably differ across the tasks. Experimenting with different configurations may be useful in identifying user biases and designing tasks with more balanced challenges.

8 CONCLUSION AND FUTURE WORK

We presented a novel explainable approach for searching and ranking videos using textual queries and visual exemplars. We argue that the decisions of our generative ranking approach are more explainable than its discriminative counterpart, as it is able to display supporting evidence to reveal its understanding of the concept space.

We also discussed our findings from the user study, facilitated by the explanation interface for exploring modal-generated decisions and viewing visual exemplars as applicable. In our study, we discovered that the XAI system yielded a comparable level of efficiency, accuracy, and user-machine synchronization as the AI system, but only if the user exhibited a high level of trust for AI explanation. However, the XAI system yielded a significantly lower level of efficiency, accuracy, and user-machine synchronization if the user instead maintained a lower level of trust for AI explanation. It is also notable that the XAI system, in addition, presented a noticeable advantage in overall user satisfaction should the user exhibit a high level of trust.

While the study outcomes do not offer concrete support for the XAI system in realms of overall explainability, trustworthiness, and accuracy, there is significant evidence that the XAI system does provide a more satisfying experience for the users who expressed a higher level of trust for the AI system's explanation. With these results, we believe that this work is one of the early steps in examining, measuring, and dissolving levels of tension and distrust between human user and the AI system, and pave way to future research opportunities pertaining to human-in-the-loop AI systems. As a follow-up to this case study, we plan to further assess the level of trust the user has in the decisions from a generative system versus a discriminative one.

In future work, we plan to extend the explainable interface to support machine learning practitioners. We believe that this interface will enable strong debugging tools for the developer to understand more about the models, ranging from their learned representations to decision boundaries, and improve the machine learning model accordingly.

SUPPLEMENTARY MATERIALS

Clip name	Study Rating Q1 AI understands	Study Rating Q2 Confidence in AI	Study Rating Q3 Trust in AI	External Rating Accurate Representation
express joy	4.50	4.50	4.50	4.25
salsa dance	4.75	4.00	4.00	4.00
sneaking	4.25	4.00	4.25	4.00
bear human subject	4.25	4.00	4.00	3.63
dribble	4.00	4.00	4.00	3.13
jog	4.00	4.00	4.00	2.25
punch	4.50	4.00	3.50	3.88
walk on uneven terrain	4.33	4.00	3.67	3.50
penguin human subject	4.00	4.00	3.50	2.63
walk around	4.50	3.50	3.50	4.75
walk crawl	4.00	3.75	3.75	2.25
mop	4.00	3.75	3.50	4.25
laugh	3.67	3.67	3.67	3.88
whistle walk jauntily	4.00	4.00	3.00	2.88
sit on high stool stand up	3.33	3.67	3.67	3.88
climb 3 steps	3.80	3.40	3.20	4.38
wave	3.20	3.40	3.40	2.63
wrench	3.00	3.00	4.00	2.75
clumsy walk	3.00	3.33	3.33	3.75

(a)

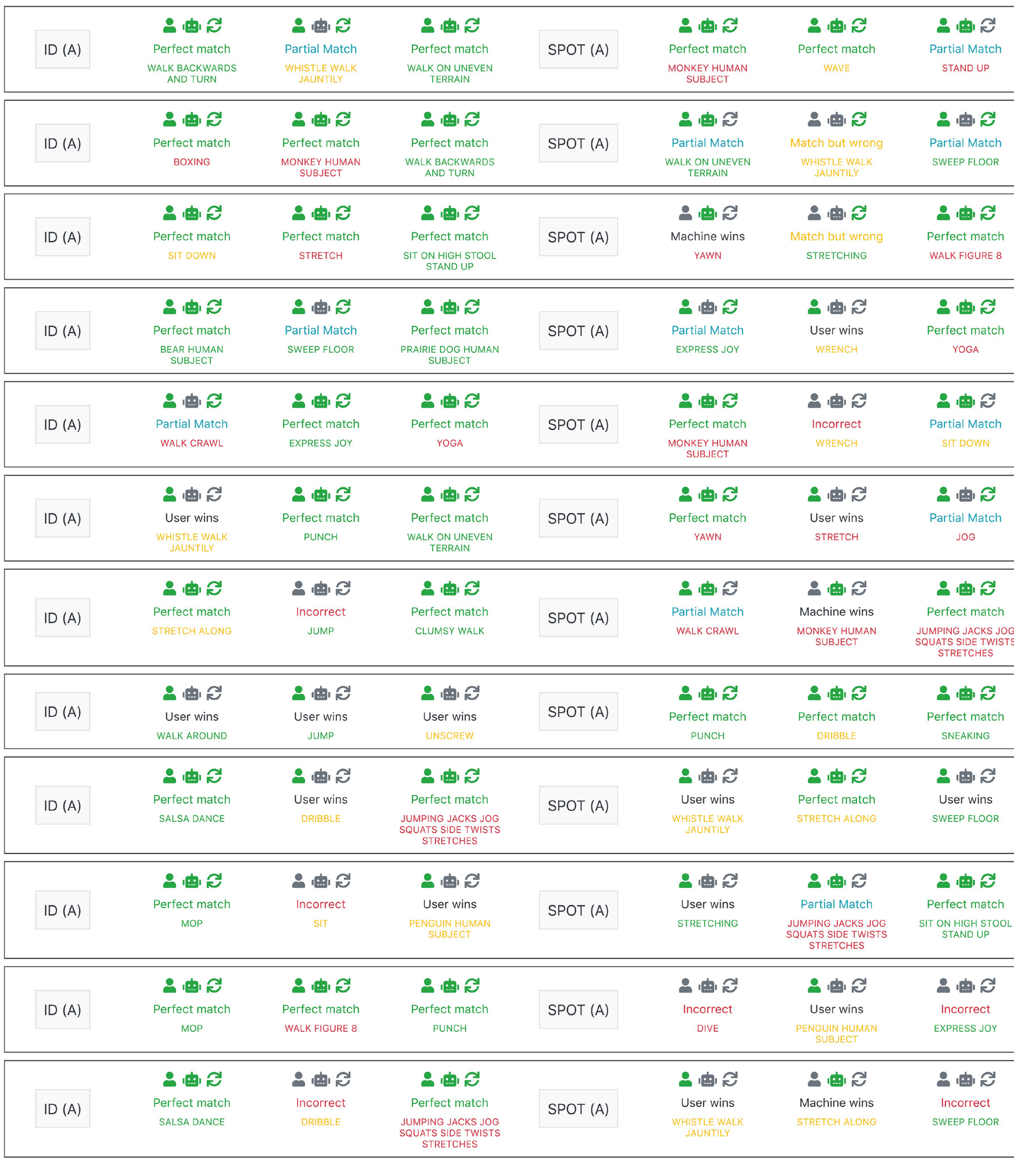
Fig. S1. Panels (a) and (b) provide a summary of questionnaire responses pertaining to AI explanations presented to the participant throughout the study session. Each participant was asked to evaluate whether the XAI system understands the query, and to rate their confidence and trust in the system. These responses are accompanied by explanation quality ratings collected from an external group of annotators and then sorted and clustered by the overall level of participant satisfaction per keyword. Poorly received keywords are marked by particularly lower ratings from study participants and external annotators.

Clip name	Study Rating Q1 AI understands	Study Rating Q2 Confidence in AI	Study Rating Q3 Trust in AI	External Rating Accurate Representation
walk straight	3.00	3.33	3.33	3.50
jump	3.40	3.20	3.00	4.00
pick up	3.50	3.00	3.00	4.50
sit	3.00	3.13	3.25	3.00
yawn	3.22	3.00	2.89	1.50
walk backwards and turn	3.00	3.00	3.00	4.13
yoga	3.17	3.00	2.83	2.00
sweep floor	3.00	3.00	2.75	4.25
stretch	2.80	3.00	2.80	2.38
stretching	3.00	2.80	2.40	3.63
prairie dog human subject	3.00	2.50	2.50	3.63
sit down	2.50	2.50	3.00	3.38
stretch along	3.00	2.67	2.33	3.25
walk and turn repeated	3.00	2.00	2.50	4.63
dance	3.00	2.00	2.00	3.63
unscrew	2.50	2.00	2.00	3.13
monkey human subject	2.00	2.00	2.00	2.38
pull up	2.00	1.50	2.00	2.88
dive	1.50	1.50	1.50	1.75

(b)

Fig. S1. Continued

Only AI



(a) Participant journeys using an AI-only system.

Fig. S2. Diagrams (a) through (d) illustrate the level of synchronization between user and AI responses as well as each user.s trust and reliance in AI explanation across different experimental conditions. Icons indicate whether the user and/or the AI.s answers match the ground truth for each trial, and whether the two answers overlap, indicating user-machine synchronization. Questionnaire responses pertaining to the user.s experience with the XAI system (available in Figure S1) are also indicated below the icons, along with the query used in each trial. Color of the query indicates the quality of generated videos based on external ratings.

(b) Participant journeys using an XAI-only system.

Fig. S2. Continued

(c) Participant journeys using an AI-only system initially then moving to an XAI system.

Fig. S2. Continued

(d) Participant journeys using an XAI system initially then moving to an AI-only system.

Fig. S2. Continued

REFERENCES

[1] CMU. 2018. CMU Graphics Lab Motion Capture Database. Retrieved from http://mocap.cs.cmu.edu/.

[2] Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. 2018. Text2Action: Generative adversarial synthesis from language to action. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'18)*. 1–5. https://doi.org/10.1109/ICRA.2018.8460608

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein GAN. Retrieved from https://arXiv:1701.07875.

[4] George Awad, Asad Butt, Jonathan Fiscus, David Joy, Andrew Delgado, Martial Michel, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, Georges Quenot, Maria Eskevich, Roeland Ordelman, Gareth J. F. Jones, and Benoit Huet. 2017. TRECVID 2017: Evaluating ad-hoc and instance video search, events detection, video captioning and hyperlinking. In *Proceedings of the Annual TREC Video Retrieval Evaluation (TRECVID'17)*. NIST.

[5] Emad Barsoum, John Kender, and Zicheng Liu. 2017. HP-GAN: Probabilistic 3D human motion prediction via GAN. Retrieved from https://abs/1711.09561.

[6] Ishmael Belghazi, Sai Rajeswar, Aristide Baratin, R. Devon Hjelm, and Aaron Courville. 2018. MINE: Mutual information neural estimation. Retrieved from https://arXiv:1801.04062.

[7] Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, 258–262.

[8] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. 7291–7299. https://doi.org/10.1109/CVPR.2017.143

[9] Hui Chen, Guiguang Ding, Zijia Lin, Sicheng Zhao, and Jungong Han. 2018. Show, observe and tell: Attribute-driven attention model for image captioning. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

[10] Jaegul Choo and Shixia Liu. 2018. Visual analytics for explainable deep learning. *IEEE Comput. Graph. Appl.* 38, 4 (2018), 84–92.

[11] J. Chuang, D. Ramage, C. Manning, and J. Heer. 2012. Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

[12] Johan de Kleer and Raymond Reiter. 1987. Foundations for assumption-based truth maintenance systems: Preliminary report. In *Proceedings of the American Association for Artificial Intelligence National Conf* erence. 183–188.

[13] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining models: An empirical study of how explanations impact fairness judgment. Retrieved from https://arXiv:1901.07694.

[14] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark Riedl. 2019. Automated rationale generation: A technique for explainable AI and its effects on human perceptions. Retrieved from https://arXiv:1901.03729.

[15] Malin Eiband, Sarah Theres Völkel, Daniel Buschek, Sophia Cook, and Heinrich Hussmann. 2019. When people and algorithms meet: User-reported problems in intelligent everyday applications. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, 96–106.

[16] M. El-Assady, V. Gold, C. Acevedo, C. Collins, and D. Keim. 2016. ConToVi: Multi-party conversation exploration using topic-space views. In *Proceedings of the Computer Graphics Forum*, Vol. 35. 431–440.

[17] D. Gotz et al. 2010. HARVEST: An intelligent visual analytic tool for the masses. In *Proceedings of the International Workshop on Intelligent Visual Interfaces for Text Analysis*.

[18] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. VSE++: Improved visual-semantic embeddings. Retrieved from https://arXiv:1707.05612.

[19] Katerina Fragkiadaki, Sergey Levine, and Jitendra Malik. 2015. Recurrent network models for kinematic tracking. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*.

[20] John Gauthier. 2014. *Conditional Generative Adversarial Nets for Face Generation*. Technical Report. Stanford. CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester.

[21] Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges. 2017. Learning human motion models for long-term predictions. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*.

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. 2672–2680.

[23] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. 2017. Improved training of Wasserstein GANS. In *Advances in Neural Information Processing Systems*. 5767–5777.

[24] David Gunning. 2016. *Explainable Artificial Intelligence*. Technical Report DARPA-BAA-16-53. DARPA. Retrieved from https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf.

[25] Bruce Hahn. Accessed 2018. CMU Graphics Lab Motion Capture Database Motionbuilder-friendly BVH conversion. Retrieved from https://sites.google.com/a/cgspeed.com/cgspeed/motion-capture/cmu-bvh-conversion.

[26] Jim Hendler and Tim Berners-Lee. 2010. From the semantic web to social machines: A research challenge for AI on the world wide web. *Artific. Intell.* 174, 2 (2010), 156–161.

[27] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations. In *Proceedings of the European Conference on Computer Vision*. Springer, 3–19.

[28] Irina Higgins, Arka Pal Loic Matthey, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. Beta-VAE: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the International Conference on Learning Representations*.

[29] Ashesh Jain, Amir R. Zamir, Silvio Savarese, and Ashutosh Saxena. 2016. Structural-RNN: Deep learning on spatio-temporal graphs. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*.

[30] A. Karpathy, J. Johnson, and L. Fei-Fei. 2016. Visualizing and understanding recurrent networks. In *International Conference on Learning Representations*.

[31] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

[32] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*. 3294–3302.

[33] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the International Conference on Machine Learning*.

[34] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2019. Personalized explanations for hybrid recommender systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, 379–390.

[35] T. Kulesza, M. Burnett, W. K. Wong, and S. Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the International Conference on Intelligent User Interfaces*.

[36] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. 2018. Tell me where to look: Guided attention inference network. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*.

[37] Xiao Lin and Mohamed R. Amer. 2018. Human motion modeling using DVGANs. Retrieved from https://arXiv:1804.10652.

[38] M. Liu, J. Shi, Z. Li, C. Li, J. Zhu, and S. Liu. 2016. Towards better analysis of deep convolutional neural networks. In *Trans. Visual. Comput. Graph.* 23 (2016) 91–100.

[39] Julieta Martinez, Michael J. Black, and Javier Romero. 2017. On human motion prediction using recurrent neural networks. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*.

[40] Christoph Molnar. 2019. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Retrieved from https://christophm.github.io/interpretable-ml-book/.

[41] XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. 2010. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Trans. Info. Theory* 56, 11 (2010), 5847–5861.

[42] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. 2018. The building blocks of interpretability. In *Distill Publication*. Retrieved from https://distill.pub/2018/building-blocks/.

[43] Sean Penney, Jonathan Dodge, Claudia Hilderbrand, Andrew Anderson, Logan Simpson, and Margaret Burnett. 2018. Toward foraging for understanding of StarCraft agents: An empirical study. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*. ACM, 225–237.

[44] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: Better, faster, stronger. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*.

[45] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*.

[46] Dominik Sacha, Hansi Senaratne, Bum Chul Kwon, Geoffrey Ellis, and Daniel A. Keim. 2015. The role of uncertainty, awareness, and trust in visual analytics. *IEEE Trans. Visual. Comput. Graph.* 22, 1 (2015), 240–249.

[47] James Schaffer, C. A. Playa Vista, John O'Donovan, James Michaelis, M. D. Adelphi, Adrienne Raglin, and Tobias Höllerer. 2019. I can do better than your AI: Expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, 240–251.

[48] Kacper Sokol and Peter A. Flach. 2018. Glass-Box: Explaining AI decisions with counterfactual statements through conversation with a voice-enabled virtual assistant. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 5868–5870.

[49] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. DeepFace: Closing the gap to human-level performance in face verification. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*.

[50] L. van der Maaten and G. Hinton. 2008. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9 (2008), 2579–2605.

[51] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and tell: A neural image caption generator. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*.

[52] Ruoyu Wang, Daniel Sun, Guoqiang Li, Muhammad Atif, and Surya Nepal. 2016. Logprov: Logging events as provenance of big data analytics pipelines with trustworthiness. In *Proceedings of the IEEE International Conference on Big*

Data. 1402–1411.

[53] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2019. Do I trust my machine teammate?: An investigation from perception to decision. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, 460–468.

[54] Tom Zahavy, Nir Ben Zrihem, and Shie Mannor. 2016. Graying the black box: Understanding DQNs. In *Proceedings of the International Conference on Machine Learning*.

[55] M. D. Zeiler and R. Fergus. 2014. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision*.

[56] J. Zhao, C. Collins, F. Chevalier, and R. Balakrishnan. 2013. Interactive exploration of implicit and explicit relations in faceted datasets. *Trans. Visual. Comput. Graph.* 19 (2013), 2080–2089.

Received November 2019; revised May 2021; accepted May 2021