Explainable Activity Recognition in Videos: Lessons Learned

Chiradeep Roy[1] | Mahsan Nourani[2] | Donald R. Honeycutt[2] | Jeremy E. Block[2] | Tahrima Rahman[1] | Eric D. Ragan[2] | Nicholas Ruozzi[1] | Vibhav Gogate[1]

[1]Department of Computer Science, The University of Texas at Dallas, Texas, USA
[2]Department of Computer & Information Science & Engineering, University of Florida, Florida, USA

Correspondence
Chiradeep Roy, The University of Texas at Dallas. Email: chiradeep.roy@utdallas.edu

Present Address
The University of Texas at Dallas

Summary

We consider the following activity recognition task: given a video, infer the set of activities being performed in the video and assign each frame to an activity. This task can be solved using modern deep learning architectures based on neural networks or conventional classifiers such as linear models and decision trees. While neural networks exhibit superior predictive performance as compared with decision trees and linear models, they are also uninterpretable and less explainable. We address this *accuracy-explanability gap* using a novel framework that feeds the output of a deep neural network to an interpretable, tractable probabilistic model called dynamic cutset networks, and performs joint reasoning over the two to answer questions. The neural network helps achieve high accuracy while dynamic cutset networks because of their polytime probabilistic reasoning capabilities make the system more explainable. We demonstrate the efficacy of our approach by using it to build three prototype systems that solve human-machine tasks having varying levels of difficulty using cooking videos as an accessible domain. We describe high-level technical details and key lessons learned in our human subjects evaluations of these systems.

KEYWORDS:
video activity recognition, cutset networks, temporal and time-series models, tractable probabilistic models, debugging machine learning models, human–computer interaction

1 | INTRODUCTION

Activity recognition in video, the task of inferring and assigning a predefined set of activities to frames or segments of a given video, is an important sub-task in *video content analysis* with a myriad of applications including video surveillance, video summarization, improving situational awareness (detecting dangerous situations) and home security. While this task is hard in general, today, with advances in deep learning, especially given the remarkable predictive performance of deep models, the task can be solved accurately when ample labeled (videos) training data is available. Unfortunately, despite high accuracy, deep neural models are black-boxes: it is difficult to explain the reasoning behind their decisions and answers. This lack of explanability is often undesirable, especially for solving interactive human-machine tasks[25] where the user makes decisions with the aid of a machine learning (ML) system. In such cases, the users need to trust the predictions of the system and be able to easily determine when the ML system is (typically) correct and when it is not. To facilitate this, the system should be such that both its functioning and the reasoning behind its actions are clear to the users, i.e., the system should be explainable.

The purpose of this paper is to describe a probabilistic modeling framework for activity recognition that is accurate yet interpretable and explainable, and to show that it helps improve the users' trust and understanding of the system. We define

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/ail2.59

"explainable" here as the ability of our system to be able to compute the answers to probabilistic queries posed over videos and provide justifications to its answers in the form of explanations (similar to the notion of explainability in the probabilistic programming literature[44,23,53]). Since the explanations are generated directly from our joint model, they have high fidelity and can be used by end-users to build a mental model of the system's functioning by posing multiple queries to it.

At a high level, our model (referred to as "the model" in the remainder of the paper) has two layers. The top layer, with which a user interacts, is a tractable, interpretable probabilistic model, specifically a cutset network[48]. Unlike conventional graphical models such as Bayesian and Markov networks in which probabilistic inference is NP-hard in general and inaccurate in practice, cutset networks have two desirable properties. First, they are tractable in that they provide linear-time reasoning capabilities. Second, their predictive performance is substantially superior to Bayesian and Markov networks[48,47,52,27]. While all probabilistic models that represent a joint probability distribution over the variables of interest are capable of generating explanations, under the assumption that an explanation is a query over the model, the use of tractable cutset networks ensures that explanations can be computed both accurately and efficiently. The bottom layer of our model is a deep neural network that yields accurate estimates, which are fed into the cutset network layer. A possible interpretation of this model is that the deep learning layer provides noisy sensory inputs to the cutset network layer, which in turn removes the noise, makes decisions and generates explanations (for the decisions) by performing fast, accurate inference. To model temporal aspects in video, we further refine this model, propose a novel temporal probabilistic modeling framework called dynamic cutset networks, and show that it improves the estimation accuracy.

We demonstrate the efficacy of our two-layer model by using it to build an *explainable activity recognition system* for an accessible domain—*cooking videos* given in the Textually Annotated Cooking Scenes (TACoS) dataset[50]. Our system is set up as a visual question-answering system where the user can ask the system a series of questions *(e.g. "Did the person cut a carrot?")* and the system provides answers to questions posed as probabilistic queries *(e.g. "Yes")* as well as explanations for these answers *(e.g. "the activity took place between 00:12 to 00:45 with a probability of 0.72")*. The system provides three types of explanations that are relevant to the question and its answer: (1) visual explanations which show video segments; (2) ranked explanations which show top-3 combinations of objects, locations and actions and (3) marginal explanations which show the confidence in the detected objects, locations and actions. Note that the goal of this framework is *not* to provide low-level pixel explanations for the deep learning model; rather, it is to build an activity recognition system that is explainable *as a whole* by generating high-level explanations for each probabilistic query at the level of label probabilities. To this end, we hypothesize that the three kinds of explanations that can be generated from our two-layer model for each probabilistic query should give end users a good understanding of the strengths and the weaknesses of the *entire* system. For example, end users might notice that the system is able to accurately predict an orange only when a knife and cutting board are present in the given frame. These kinds of insights are either absent or very difficult to extract from pixel-level explanations.

We evaluated our system by using it to solve three human-machine tasks having varying levels of difficulty. The three tasks were: (1) answering whether an activity was performed in a video by watching the video (Yes/No questions); (2) answering whether a set of policies were followed or not in a collection of videos; and (3) debugging the model to find its error patterns. We built novel interactive visual interfaces for each task and conducted human subjects studies. Our studies showed that when the underlying (human-machine) task is relatively easy (yes/no questions) and the model is accurate, explanations helped in improving time to completion, accuracy and trust in the system. However, on more involved tasks such as verifying policies in a set of videos, we found that the effects of explanations are not universally beneficial. Specifically, if first impressions of system outputs are skewed towards either system strengths or system weaknesses it can significantly affect users' trust, reliance and (mis-)understanding of the ML model, regardless of the presence of explanations. Additionally, we show how the explanations generated by our system can be used in a visual analytics tool to support model debugging through identification of both instance-level problems as well as global error patterns.

2 | RELATED WORK

2.1 | HMMs and DBNs

There has been significant work in the past that have used state-space models such as Hidden Markov Models (HMMs)[58,4,34,40] and Dynamic Bayesian Networks (DBNs)[6,18,1,41] on top of low-level feature extractors to model persistence and dynamics. However, HMMs typically make a lot of restrictive assumptions and inference complexity increases exponentially in the size of the state space. The expressive power of DBNs, on the other hand, are offset by exact inference being NP-hard[8] in these

networks. The model we use in this paper aims to bridge this expressiveness-tractability gap by extending cutset networks[48] into the temporal domain (much like DBNs extend Bayesian Networks into the temporal domain). This formulation will help us efficiently compute posterior probabilities which can be used as explanations for queries (which we explain in the later sections).

2.2 | Models that use Black-boxes as Sensors

There has been research in the past that has attempted to combine black-box models with probabilistic models in various ways. Some of the earlier attempts were the work of Pei et al.[43], Morariu et al.[29] and Brendel et al.[5]. The work by Pei et al. is a grammar-based approach[21,55,22,60] that uses an AND/OR graph to model the semantics of the video[56]. While this approach has the advantage of providing a global overview of the functioning of the model (i.e. interpretability), it is unable to efficiently compute action-level conditional probabilities over time such as the probability that a cutting activity over a carrot would take place in the sixth interval given that a picking up knife activity took place in the first. It can only do this after enumerating all possible combination of missing activities (i.e. inference is NP-hard). The latter works by Morariu et al. and Brendel et al. use probabilistic relational logic to refine the probabilities of candidate event hypotheses. Although probabilistic relational models work very well in situations where prior knowledge about the domain is known a-priori, learning these relations from data is typically NP-hard[16]. Further, inference is also typically intractable without making some very restrictive assumptions (for example, the work by Morariu restricts the treewidth of the MLN to make exact inference tractable). The model we propose in this paper addresses all of these problems by formulating as a temporal multi-label classification (TMLC) problem where the relations between labels are encoded into a compact representation of the joint probability distribution. Further, using tractable prior and transition distributions in our Dynamic Conditional Cutset Network (DCCN) framework allows for both efficient and accurate posterior estimates using particle filtering.

2.3 | Explainable Systems and Trust

There have been a number of studies on how trust influences interactions between humans and automated systems, e.g., Muir et al. (1994),[31], Muir et al. (1996),[30], Lee et al.[26], and Hoffman et al.[19]. These studies examine factors that might affect the trust of the user in the system, such as showing the past performance of the system and making the working of the system more understandable[26]. The work by Hoffman et al.[20] provides a more detailed taxonomy of such factors and explains how trust is context-specific and dynamic. In other words, trust might vary with respect to specific contexts of automation and must also be maintained over time. Our aim is to be able to measure and control user trust with respect to these systems in order to better understand what kind of explanations influence the trust variable.

2.4 | Activity Recognition and NLP

This work is closely related to the work of Rohrbach[51] and Donahue[13] on generating a semantic representation from videos at an activity level using deep learning architectures. Instead of generating sentences in natural language however, we assign a number of pre-defined labels divided into categories. Related efforts have considered the task of dense captioning[24], i.e., generating summaries of texts from particular segments. Song et al.[57] attempted to create captioning methods that require minimum supervision on the TACoS dataset. Duan et al.[15] attempted to combine caption generation and sentence localization to feed off of each other to create a weakly supervised training model. While these works focus on constructing text summaries, our work is different in that it aims to create a semantic representation for activities in each frame that can be used to both answer queries easily as well as generate explanations (via probabilistic inference) that justify these answers.

3 | XAI SYSTEM DESCRIPTION

3.1 | The Explainable Activity Recognition Task for Cooking Videos

We evaluated and tailored our system to the Textually Annotated Cooking Scenes (TACoS) dataset[51]. Each frame in each video in this dataset is labeled with an (*action, object, location*) triple; this triple defines an *activity*. The *action* component forms the core part of the activity. These are usually verbs like wash, cut, slice, open, etc. The *object* component denotes the entities over which the activity is performed. These are generally nouns such as apples, refrigerator, cutting board, knife, etc. Finally, the

location component tells us *where* the activity is taking place. These are generally location nouns such as kitchen, counter top, sink, etc. but can also overlap with the nouns we use as objects. The dataset has 28 labels (our vocabulary) which includes 12 actions, 7 objects, 8 locations and a special label called 'Nothing' or 'None'.

Users interact with our system by posing so-called *selection questions*: "Did a particular activity defined by the triple (*action*, *object*, *location*) happen in the video?" where object and location can be "None," but action is not allowed to be "None." Examples of selection questions include: (1) "Did the person slice an orange on the counter?" where slice, orange, and counter denote the action, object, and location respectively; and (2) "Did the person take out grapes from the refrigerator?" where take out, grapes, and refrigerator denote the action, object, and location respectively.

Our goal is to build an explainable system that provides three types of explanations following an answer to a selection question:

1. **Video Explanations:** When the system answers "yes," we want the system to highlight (possibly more than one) segments of the video where the activity happened. For "no" answers, we want the system to highlight segments where a related activity happened, e.g., carrots were cut in the video but not oranges. If no related activity is found in case of a "no answer," we want the system to output the most likely activity in the video.

2. **Ranked (action, object, location) Triples:** We want the system to display the top-k predicted activity triples in the video that are relevant to the query.

3. **Most Probable Entities:** We want the system to display the most probable actions, objects and locations, along with their likelihood that are relevant to the query.

3.2 | System Architecture

Fig. 1 shows a high-level overview of the components of the system and the processing pipeline. Roughly speaking, the system comprises the following two layers of models: (a) *video classification layer* which takes as input video frames and a vocabulary file and assigns a set of labels from the vocabulary to each frame; and (b) *explanation layer* which takes the predicted labels from the video classification layer as input, corrects them using a probabilistic model, and outputs (potentially more accurate) labels and explanations.

3.3 | Video Classification Layer

For this layer, we used GoogLeNet[59], a 22-layer neural network that is pre-trained on the ImageNet dataset[54]. To tailor GoogLeNet to our video dataset which has 28 labels, we replaced the topmost softmax layer in GoogLeNet with a fully-connected layer with 28 nodes that uses sigmoid cross-entropy loss. The latter is used because it is a standard loss function for solving multi-label classification problems; note that we are solving a multi-label classification task since each frame can have multiple objects and locations. We used the backpropagation algorithm with stochastic gradient descent (SGD) and Adaptive Moment Estimation (Adam) optimizers to further train the pre-trained model on the TaCOS dataset and found that Adam yields the best performance.

3.4 | Explanation Layer

In this section, we present dynamic conditional cutset networks (DCCNs), a new tractable temporal probabilistic representation. We will use DCCNs in the explanation layer to: (a) correct errors in the labels predicted by the GoogLeNet at each frame; (b) model the dynamics as well as persistence (activities do not change rapidly between frames) in the video; and (c) provide explanations via poly-time probabilistic inference.

3.4.1 | Conditional Cutset Networks

Tractable probabilistic models (TPMs)[2,28,9] are probabilistic models which admit poly-time posterior marginal inference (MAR)—the task of computing marginal probability distribution over each variable given evidence which is defined as an assignment of values to a subset of variables—and maximum-a-posteriori (MAP) inference—the task of computing the most likely assignment to all non-evidence variables given evidence. Examples of popular TPMs include cutset networks[48], arithmetic circuits[9], sum-product networks[45] and probabilistic sentential decision diagrams[3]. Although, TPMs are less expressive

than intractable (latent) probabilistic models and as a result have slightly poor generalization performance as compared to the latter, their accuracy at test time is often much higher than intractable models. This is because tractable models use exact inference at prediction time while one has to use inaccurate approximate inference algorithms in intractable models.

Cutset networks[48] are TPMs which represent multidimensional joint probability distributions using a rooted (directed) OR tree[10] with tree Bayesian networks at each leaf node of the OR tree (see Fig. 2). Each OR node in the OR tree is labeled with a variable and just like in decision trees represents conditioning over the variable. Unlike decision trees however, the arcs in the OR tree are labeled with conditional probability of the variable taking the corresponding value given an assignment of values from the root node to the OR node. Tree Bayesian networks at each leaf l represent the conditional distribution $P(\boldsymbol{X}|path(l))$[1] where $path(l)$ denotes the assignment from the root to l and $\boldsymbol{X}$ is the subset of variables not assigned in $path(l)$. MAR and MAP inference over cutset networks can be performed in linear time in the size of the network using cutset conditioning[42,10]. However, unlike conventional cutset conditioning algorithms, cutset networks take advantage of context-specific independence, dynamic variable orders and determinism. As a result, cutset networks can compactly represent and perform tractable inference over probability distributions that admit high treewidth probabilistic graphical models[11,12,47].

Recently, Rahman et al.[49] proposed a new framework called *conditional cutset networks* (CCNs) that extends the cutset networks framework to compactly represent and perform efficient reasoning over high-dimensional conditional probability distributions, namely distributions of the form $P(\boldsymbol{Y}|\boldsymbol{X})$ where both $\boldsymbol{X}$ and $\boldsymbol{Y}$ are sets of random variables. To compactly represent the (exponentially many) conditional joint distributions over the variables $\boldsymbol{Y}$ given each assignment $\boldsymbol{X} = \boldsymbol{x}$, CCNs use a cutset network structure over $\boldsymbol{Y}$ whose conditional probability distributions $P(Y|path(n))$ at each OR node n as well as those attached to each variable in the Bayesian networks is replaced by calibrated probabilistic classifiers[35]. The latter takes an assignment $\boldsymbol{x}$ to $\boldsymbol{X}$ and $path(n)$ as input and outputs a conditional probability distribution over Y, namely $P(Y|\boldsymbol{X} = \boldsymbol{x}, path(n))$ by using only polynomial (in $|\boldsymbol{X}|$) number of parameters. For example, when we use logistic regression, we have $P(Y = 1|\boldsymbol{X} = \boldsymbol{x}, path(n)) = \sigma(w_0 + \sum_{x_i \in \boldsymbol{x}} w_i x_i)$ where w_i's are the weights (parameters) and σ denotes the *sigmoid* function. We learn the parameters of the calibrated classifier (e.g., logistic regression) using a subset of the data that is consistent with $path(n)$. CCNs are conditionally tractable in that given an assignment $\boldsymbol{X} = \boldsymbol{x}$, each (probabilistic) classifier yields a probability distribution over the (class) variable Y and thus given $\boldsymbol{X} = \boldsymbol{x}$, a CCN yields a (tractable) cutset network. (see Fig. 3 for an example).

The structure of CCNs is learnt using the top-down induction algorithm detailed in Rahman et al.'s[49] paper. The base case occurs when the dataset contains either (a) a very small number of examples/records or (b) a very small number of variables. In such a situation, a simple tree-structured Bayesian network is powerful enough to represent the data distribution and can be learned using the Chow-Liu algorithm[7]. If the base condition is not satisfied, then the algorithm will recursively and heuristically select a single variable from the set of all currently available variables and then condition on it. For example, in a dataset defined over variables $\boldsymbol{Y} = \{Y_1, Y_2, Y_3, Y_4, Y_5\}$, the algorithm might find that Y_1 leads to the largest information gain in the data and choose to condition over it (such as in the CCN in Fig. **??**). It will keep recursively doing this until it reaches the base case. After the structure is learned, the branch probability functions at each OR node in the CCN as well as each tree-structured Bayesian Network at the leaves is learned using calibrated classifiers such as logistic regression and neural networks with a sigmoid layer on top (see Fig. 3). The best calibrated classifier is chosen via cross-validation.

To use CCNs in our framework, we feed the output of GoogLeNet to the CCN. More formally, let $\boldsymbol{Y}$ denote the set of output nodes of GoogLeNet and $\boldsymbol{X}$ denote the set of true labels at a frame. We use the CCN to model $P(\boldsymbol{X}|\boldsymbol{Y})$ and learn its structure and parameters using a dataset constructed as follows. Each frame in each video is a training example and is composed of true labels ($\boldsymbol{X}$) and labels predicted by GoogleNet ($\boldsymbol{Y}$) with the pixels in the frame as input. At test time, at each frame, we instantiate all the classifiers in the CCN using the predicted labels to yield a cutset network and then perform inference over the cutset network to yield the final set of labels. In other words, the CCN treats the output of the neural network as a noisy sensor (see Fig. 1) and computes a conditional joint probability distribution over the true labels given the predicted (noisy) labels.

We now describe how the compactness and tractability of CCNs can be leveraged to model the temporal dynamics in videos.

3.4.2 | Dynamic Conditional Cutset Networks

An issue with CCNs is that they are static and do not explicitly model temporal aspects of video. For instance, we can use persistence, namely objects do not change their position rapidly between subsequent frames to correct prediction errors at a frame by using data from neighboring frames. To address this issue, we propose a novel framework called dynamic conditional

[1] We denote random variables as uppercase letters e.g. X, Y etc., sets of random variables as bold uppercase letters e.g. $\boldsymbol{X}, \boldsymbol{Y}$ etc. and assignment of values to all the variables in a set as bold lowercase letters, e.g. $\boldsymbol{x}, \boldsymbol{y}$ etc.

cutset networks (DCCNs). Formally, let a video consist of n frames, let $\boldsymbol{Y}_t$ and $\boldsymbol{X}_t$ be the set of true labels and predicted labels (evidence) at frame t. Then, the DCCN represents the following probability distribution:

$$P(\boldsymbol{y}_{1:n}|\boldsymbol{x}_{1:n}) = P(\boldsymbol{y}_1|\boldsymbol{x}_1)\prod_{i=2}^{n} P(\boldsymbol{y}_i|\boldsymbol{y}_{1:i}, \boldsymbol{x}_{1:i-1}), \tag{1}$$

where the notation $\boldsymbol{x}_{1:n}$ (similarly $\boldsymbol{y}_{1:n}$) denotes an assignment of values to all predicted (true) labels in frames 1 to n. We will use the notation $\boldsymbol{X}_{1:n}$ to denote the set $\bigcup_{i=1}^{n} \boldsymbol{X}_i$.

The representation given in Eq. (1) is not compact as n increases. To circumvent this issue, we use two standard assumptions widely used in temporal or dynamic probabilistic models—the 1-Markov and stationarity assumptions[46]. Specifically, we assume that each frame is conditionally independent of all frames before it given the previous frame (1-Markov) and all conditional distributions are identical (stationarity). With these assumptions, we can represent $P(\boldsymbol{y}_{1:n}|\boldsymbol{x}_{1:n})$ using

$$P(\boldsymbol{y}_{1:n}|\boldsymbol{x}_{1:n}) = P(\boldsymbol{y}_1|\boldsymbol{x}_1)\prod_{i=2}^{n} P(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{y}_{i-1}), \tag{2}$$

where $P(\boldsymbol{Y}_1|\boldsymbol{X}_1)$ and $P(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{Y}_{i-1})$ are conditional cutset networks and $P(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{Y}_{i-1})$ is the same for all i.

We learn DCCNs using the following approach. The prior model $P(\boldsymbol{Y}_1|\boldsymbol{X}_1)$ is the same as the CCN described in the previous section. To learn the structure and parameters of $P(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{Y}_{i-1})$, we construct the dataset as follows. Each frame in each video is a training example and is composed of true labels at frame i ($\boldsymbol{Y}_i$), true labels at frame $i-1$ ($\boldsymbol{Y}_{i-1}$) and labels predicted by GoogleNet at frame i ($\boldsymbol{X}_i$) using the pixels in the frame as input.

Inference over DCCNs can be performed using sequential sampling approaches such as particle filtering and smoothing[14]. Here, we generate k assignments $(\boldsymbol{y}_1^{(1)}, \ldots, \boldsymbol{y}_1^{(k)})$ uniformly at random from $P(\boldsymbol{Y}_1|\boldsymbol{x}_1)$, then for each assignment $\boldsymbol{y}_1^{(i)}$ we sample one assignment from $P(\boldsymbol{Y}_2|\boldsymbol{x}_2, \boldsymbol{y}_1^{(i)})$, and so on. At the end of the sampling process, we will have k particles from $P(\boldsymbol{Y}_{1:n}|\boldsymbol{x}_{1:n})$. The main virtue of DCCNs is that unlike widely used temporal models such as dynamic Bayesian networks[32], the particles in DCCNs are generated from the posterior distribution $P(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{y}_{i-1})$ at each frame. As a result, issues such as particle degeneracy—particles vanish because their weights become too low as i increases—that typical sequential sampling algorithms suffer from will be less severe in DCCNs.

The three explanation types (see section 3.1) can be computed by performing MAP and MAR inference in CCNs and DCCNs. To compute video explanations, we use an ontology that models relationships between activities and objects (e.g., 'chef's knife' is related to 'kitchen knife', 'slice' is related to 'cut', etc.) and display video segments in which the marginal probability of the queried activity or activities related to it (recall that activity is a (*action*, *object*, *location*) triple) is larger than a threshold. Ranked triples over each segment in video explanations are computed by performing k-best MAP inference. Again, the key advantage of CCNs and DCCNs is that k-best MAP inference is linear in k and the number of parameters at each frame. Most probable entities in each highlighted segment are derived as follows. We compute the marginal probability of each entity in each highlighted segment given evidence (via MAR inference) and display the top k most likely ones according to the marginal probability. Note that these inferences are not possible on GoogLeNet or recurrent deep architectures[13] unless we treat the labels as independent entities.

3.5 | Video Compilation and Query Processing

In this section, we explain how each video is individually processed and compiled when it becomes available to the system and how this compiled knowledge is later used for answering queries and providing explanations. We call this the compilation and query processing pipeline (for an example, see Fig. 4). The pipeline has the following two phases:

- **Compilation Phase:** When a new video becomes available to the system, relevant information about the video is compiled and stored in a database. The database is then used to answer queries in real-time. More specifically:
  1. Each frame t of the video is passed through GoogleNet which infers a set of noisy labels $\boldsymbol{x}_t$.
  2. All the frames are then passed to the dynamic cutset network which finds the top-k activities in each frame t by constructing and inferring over the posterior distribution $P(\boldsymbol{y}_t|\boldsymbol{x}_{1:t})$ where $\boldsymbol{y}_t$ denotes the true labels associated with frame t and $\boldsymbol{x}_{1:t}$ denotes the noisy labels from frame 1 to frame t.
  3. All consecutive frames with the *same* top explanation are then grouped into a *segment*. In addition, the dynamic cutset network also computes the component-wise marginal probabilities and these are averaged over each segment.

4. Finally, the explanations and marginal probabilities are stored in a database for fast retrieval.

- **Query Phase:** In this phase, the user poses a selection query to the system. These queries are in the form of whether a given combination of *action*, *object* and *location* exists in the video. The system then searches over all the compiled segments (stored in a database) for matches on the most probable activity for each segment. If at least one *complete* match is found, then the system answers "Yes" and returns *all* the segments that match the query, along with their explanations. Otherwise, the system answers "No" and as explanations displays all segments with partial matches. For example, if the query is asking if the person cuts a carrot on a cutting board (*cut*, *carrot*, *cutting board*) and there are no exact matches, then the system might return segments where the person is cutting a carrot, but on a plate (*cut*, *carrot*, *plate*) or washing a carrot in the sink (*wash*, *carrot*, *sink*).

3.6 | Interactive Explanatory Interface

We originally sought after an interactive interface that allowed users to load videos, ask queries, and review the model output along with the explanations. The goal for the interface design was to limit the amount of model information presented to the users in order to avoid overwhelming them with information. To this end, we first designed the interface with a predefined list of queries which users were able to select and explore the model outputs for, and later, sought after a new design where users were allowed to build their own queries. Fig. 5 and Fig. 6 show the resulting interfaces respectively.

Initially, the queries were in form of yes/no questions that consisted at least two of the three combinations of particular *actions*, *objects*, and *locations* that defined an activity. An example query (as seen in Fig. 5) is "Does the person *peel* an *onion*?" However, later-on, we opted for a query building tool where users could choose an *action*, an *object*, and a *location* from a list of potential vocabulary. Unlike the first approach, this method allowed users to search by selecting as little as one activity component. With this approach, people could form they query implicitly in their minds and by looking at the selected activity components. For example, one possible query is (*peel*, *any object*, *cutting board*), which can be mapped to the natural-language question: "Does the person *peel anything* on the *cutting board*?"

The model then attempted to answer the query. For the initial interface, it was a simple response of yes or no. This was under the assumption that the video was already loaded, i.e., the person could see and select a query from the list of predefined queries for the loaded video. However, with the query building tool, we wanted to allow a user to show the model response to all the videos before deciding which video to inspect. Because of this, we designed the interface to organize the videos into two groups, based on whether model found a video was a match for the query or not.

With both of the designs, we showed the same interface elements for the explanations. These explanations comprised the video segments that were most relevant to finding the answer to the query, and for each of these segments, the top three activities found in the query as well as the model's confidence in observing individual activity components (i.e., *actions*, *objects*, or *locations*). This information was always visible per video, i.e., they were local explanations. Thus, for the initial interface (as seen in Fig.5), users always accessed the information while for the second interface, this information would pop-up when a video was selected (in interface shown in Fig. 6).

Upon selecting a query (and choosing a video in the second interface), the video player highlights the most relevant segments of the video through visual annotations added under the video play bar (shown as light and dark blue under the video in Fig. 5). The video player will also automatically jump to the appropriate segment to help users see the video frames most important for determining the output. For each selected video segment, the detected combination of components showed the top three detected activities within this segment (in form of a table), as well as the individually-detected activity components. To help users to quickly judge these component scores, graphical bars are shown underneath detected components to visually represent the values of the component scores. Users can select different video segments to view the corresponding component scores and combinations from different portions of the video. The selected segment was represented with a darker color.

One of our main goals when designing these interfaces (particularly, second design) was supporting users to build better mental models of the model strengths and weaknesses (i.e., errors). However, both of the interfaces we described gave greater attention to false positives while giving less attention to false negative errors. In other words, it was easier/faster to discover and pay attention to false positive errors as opposed to their counterparts. Since it is important to understand both these types of errors to 1) build a more accurate mental model of the system and 2) be able to fix the errors with the model when designing an algorithm, we aimed to design yet another exploratory interface to provide more support for both of the tasks.

We hypothesized that the addition of holistic explanations to provide broader model-level information will increase user's ability to equally pay attention to and identify false negative and false positive errors, as well as higher-level model-based

problems across multiple videos. We refer to these more holistic representations as "global explanations" as they provide insights about the model's logic by displaying multiple possible outputs at once with overall system performance indicators (i.e. accuracy, precision, recall, etc.). This view was designed to aid user understanding of higher-level model-based problems across multiple videos rather than focusing on problem identification on a per-video basis.

Figure 7 shows the visual design for the global view of the interface. In the video selection panel, each video shows a list of the top five components the system detected alongside an estimated rate of false positives and false negatives for the video. With each video, a set of temporal heat maps are provided, representing the model's detection confidence for each component in the system's vocabulary over the video's duration. Additionally, a global information panel is included to provide general system performance information. This includes an estimation of the overall system detection accuracy, false positive rate, and false negative rate. Finally, this panel contains a bar chart for both false positive and false negative rates per each object in the system.

4 | EVALUATION

4.1 | Machine Learning Evaluation

We selected 60313 frames for training and 9355 frames for testing distributed over 17 videos in the TACoS dataset. For each set, we selected a set of ground labels and used the video classification layer to generate the predicted labels. We performed exact inference over CCNs and used particle filtering with 100 particles for inference in DCCNs. We performed the following ablation study: (1) Our system in which the explanation layer is removed (GoogLeNet); (2) Our system which uses (static) conditional subset networks in the explanation layer (CCNs); and (3) the full system (dynamic CCNs).

Table 1 outlines the accuracy scores for correct activity recognition according to various evaluation metrics. Since predicting each activity correctly is a multilabel classification task, we use K-Group measures to calculate the overall percentage of instances where K labels out of the total number of labels were predicted correctly. We report K-1, K-2, and K-3 since each activity comprises of *action*, *object* and *location*. In addition, we also use standard measures such as the Hamming Loss and the Jaccard index. We notice that the full system that uses the dynamic CCNs yields the best scores. This is expected, since the dynamic CCN encodes both the error distribution of the labels predicted by the neural network at each given frame as well as the transition distribution of how labels evolve over time.

4.2 | Human Evaluations

To evaluate the effectiveness of the explanations with user interactions, we conducted a series of studies that focused on human performance. These experiments aimed to study various human factors with explainable AI systems, specifically, in the context of decision-support systems, human-AI collaboration, and video activity recognition. In this section, we rely on a brief description for these studies and refer the audience to the full manuscripts to learn more about the details of each study[38,36,37].

4.2.1 | Evaluating Human-Machine Query Verification

The primary goal in this study was to measure the degree to which the explanations generated by our system would benefit the end users with little to no understanding of how ML systems work. We hypothesized that the presence of explanations would improve both the speed and the accuracy of decision-making. We also hypothesized that user agreement with the system's outputs would significantly increase with explanations. A user's answer is said to 'agree' with the system's when they are the same.

To test our hypotheses, we designed a user study where participants were set to review the model's responses and explanations to a set of queries about videos, using a similar design from interface shown in Fig. 5. More detailed study description, other findings and results, and discussions from this study is available in our previous paper[37]. The query-review task was designed to assess the participants' ability to accurately determine the correct answer to queries with the aid of the system. To clarify, we define a "correct" answer here as the actual answer (or ground truth) of a given query. This is different from "agreement" (defined above) where the user's answer matches the one from the system. For example, there could be cases where both the user and the system agree on the same (incorrect) answer; however, this might be different from the actual, "correct" answer. Through a between-subjects user study, we divided the participants into two groups: *with* and *without* explanations. Participants from both groups had access to the video player and the system's answer to each question. However, while those *with explanations* were provided with the interface seen in Fig. 5, their counterparts did not see any explanations (i.e. they were only able to view

the system's answers and not the video segments, the detected combination of components,and the component scores). Overall, each participant reviewed 20 unique queries with a ratio of 16–4 correct–incorrect system answers.

The experiment was completed online by 80 AMT workers. Of these participants, 40 of them were shown explanations while the other 40 were not. After pre-processing the data and removing outliers that did not fall within 1.5 × IQR, we analyzed results from 38 participants for the *with explanations* category and 40 for the *without explanations* category. We analyzed the results using the Kruskal-Wallis non-parametric test to measure the difference between the two groups. We observed a significant difference on error per trial $\left(\chi^2(1,76) = 5.63, p < 0.05\right)$, showing that the participants *with explanations* had significantly less error than those *without explanations*. Our experiment also detected a significant difference on average time per trial $\left(\chi^2(1,76) = 28.1, p < 0.001\right)$. Participants *with explanations* were significantly faster. Additionally, the results from the user agreement with the system show that participants *with explanations* significantly agreed with the system more than their counterparts $\left(\chi^2(1,76) = 8.00, p < 0.01\right)$. Fig. 8a shows average participant error per trial.

Overall, these results support our hypothesis that the addition of these explanations significantly improves user task-performance in our system. Through this study, we learned that providing more information (through post-hoc explanations) can support user understanding of the system and judge when it is correct. It would seem that the explanations encouraged the users to *correctly* trust its output. Since the *with explanations* category also had significantly better performance results, this suggests that the higher rate of agreement was not simply blind trust or *automation bias*[17], where humans tend to trust an intelligent system by virtue of its 'intelligence' alone. However, it is to be noted that our study was not designed to specifically focus on the potential effects of explanations on automation bias.

8.2.2 | Evaluating Open-ended Human-Machine Video Review

In the first evaluation, we designed a task where users reviewed and answered queries about activities and objects in designated videos. Building on these results, we used the same underlying algorithm, to explore how users attempt to understand model competencies and weaknesses when given the freedom to explore. Additionally, we also wanted to understand the role of first impressions on users' mental model formation. We performed a user study based on the interface described in Section 3.6 and seen in Fig. 6, and the results of this work were recently published[38,36]. We will therefore briefly describe the study and key findings here and advice those who seek to learn about the study in more detail to refer to our prior work.

We designed a policy-verification task, where participants were asked to verify whether a set of kitchen guidelines and policies are being followed by the people performing cooking activities by utilizing the query-building system. The policies were designed such that half were in reference to known system weaknesses while the others exposed system competencies. Assuming that most users began from the top of the list and worked their way down, we changed the order of the policy sets and compared how participants experienced the system. As an additional variable, half of the participants were provided explanations while others were not. After a brief video tutorial, users freely explored how the tool classified different combination of activated components as they tried to verify system policies before being asked about their impressions of the system. We ultimately asked 110 participants to review the system, 54 observed explanations: 28 of whom saw policies that exposed model capabilities first and another 26 observed those that exposed model weaknesses early-on. Of those provided no explanations, the number of participants were 29 and 27, with the respective order.

First, we measured the proportion of policies that were answered correctly (user-task error). Our results indicate that participants who saw system strengths early-on made significantly more errors in the policy-verification task compared to those who encountered weaknesses first, with $F(1,106) = 6.55$, $p < 0.05$, $\eta_p^2 = 0.058$. This indicates that encountering strengths earlier can lead to over-reliance on the model outcomes (i.e., automation bias), while seeing weaknesses in the beginning can prevent this problem. Fig. 8b shows the distribution of the task-error results across the conditions.

After completing the policy-review task, participants were asked to estimate the model's detection accuracy (percentage) for several activity components in the system's vocabulary that corresponded to both system strengths and weaknesses. For each participant, we calculated the error in their estimated accuracy for that component. For components that corresponded to system strengths, participants who observed weaknesses first significantly underestimated the model's detection accuracy compared to their counterparts, with $F(1,106) = 6.24$, $p < 0.05$, $\eta_p^2 = 0.056$. Additionally, those who observed weaknesses early-on were significantly less confident about their estimations, with $F(1,106) = 3.94$, $p < 0.05$, $\eta_p^2 = 0.036$.

This shows that participants who observed system weaknesses first had problems forming their mental models of the system competencies and strengths. They significantly underestimated the system capabilities while also having less confidence in their estimations. These users were skeptical of system strengths yet showed hesitancy in their skepticism because their earlier negative observations obscured their judgment of the system capabilities. With a negative first impressions of the system capabilities,

users tended to rely more on their own abilities and focused less on how the model performed. Since they were more focused on completing the task themselves, questions about their impressions of the system capabilities may have been unexpected—leading to the confusion reported in our results. On the other hand, users who experienced system strengths first tended to exhibit behaviors related to overconfidence. Their performance on the policy review task was significantly worse than their weakness first counterparts. They appear to develop a false sense of security (i.e. automation bias) early on, showing that when they looked at the later policies, they generally continued to rely on the system, even when it made mistakes. Having overconfidence in system capabilities had no influence on completion time. Yet, if users took about the same amount of time to complete the policy review task, it appears that being overconfident had the undesired effect of decreasing user's critical review of system outputs.

Overall, these results suggest an additional nuance to the findings from the previous user study described in Section 4.2.1. While the prior study found the addition of explanations significantly improved human verification performance, the second study demonstrates that explanations cannot be assumed to be universally beneficial. Even more notable is the knowledge that explanations might increase the likelihood that users will develop unfounded knowledge of the model (i.e., they think they understand how the model works when, in fact, they do not). In safety critical applications, these impression effects can have disastrous consequences. In this study, we found evidence to support that early impressions of a machine learning model is fundamental for users to establish their assessment of model capabilities. When introducing end-users to XAI systems, attention should be given to how to expose users to both system weaknesses and strengths in a balanced way early on to help ensure accurate mental models are developed and maintained by users.

4.2.3 | Using Global Explanations to Debug the Model

While the previously described user studies focused on human-machine performance for video review and inspection user tasks, the third evaluation—using the interface described in Section 3.6 and seen in Fig. 7—addresses a different use case involving developers debugging a model. To demonstrate the utility of the tool for deeper, more detailed model inspection, we present case studies using two datasets in which an experienced machine learning designer from the research team used the global explanation view to identify various types of errors and problem patterns in the underlying model for debugging purposes. The first case study was based on the activity recognition model created with the previously described TACoS video corpus[50] of kitchen activities. For the second case study, we chose a set of six videos from the Wet Lab dataset[33], which consists of videos of laboratory experiments involving chemical and biological testing. While we performed and included a comprehensive case study in a prior work[39], we will only briefly touch upon the case studies in the current paper and refer those interested to learn more about the details on the system design goals and choices, usability user evaluation, and the two case studies to read our past paper[39].

Our interface from Fig. 7 provides possibilities to pay attention to both false positive and false negative errors, equally, as demonstrated by both of our case studies. To achieve this, the subject examined which video to explore by referring to the objects that have the highest False Positive (FP) and False Negative (FN) rates using the bar charts in the global information panel. Once he decided to explore a specific object, e.g., object X, he could use the top-five components to identify which video highly includes object X. By comparing the heat maps based on their colors and (i.e., the model's detection confidence per object, per second) using the blue bar, he could explore many scenarios where X is a FP or FN *individually*, or may compare various heat maps against X to find cases where X (or other components) are FP or FN in the same *activity*. This information can then be used to penalize the model parameters proportional to the observed probability of the error to ensure the final model learnt will make fewer errors of each type. This is one approach to use this exploratory tool, and allows a model designer to navigate and find various types of errors (with or without combinations) or even find the sources of a problem by comparing similar combinations or individual components with errors across multiple videos to find similar global error patterns. For example, a debugger might notice that a given video has a high number of false positives. After clicking on the video, the global information panel might show that the false positive rate for "plate" is 20%. The debugger can then browse through the sections in the video that have high confidence scores for plate by looking at the heat map (see Fig. 7) and notice that the probability for plate goes up whenever there is a frying pan in the frame. More technical details of these findings and a more detailed walk-through of the case studies are available in our prior paper[39].

Through our case studies, we learnt that providing higher-level explanations that are not limited to a query or specific activity can support identification of various specific types of model errors and allows more experienced model architects to explore the model to find high-level, global error patterns as well as instance-level problems. This can be used to alleviate the bias towards focusing primarily on false positive errors that we observed with AI-assisted querying.

5 | CONCLUSION

In this paper, we presented the technical details of our proposed approach that seeks to address the following accuracy-explainability gap in existing machine learning technology: deep neural networks yield accurate predictions but are not explainable while conventional classifiers such as linear models and decision trees are explainable but substantially less accurate. The key idea in our approach is to *compose* neural networks with explainable, tractable probabilistic logic representations. We presented a version of this general approach in which we first used a neural network to yield accurate estimates over the labels (or classes), then used these estimates as *soft evidence* in a probabilistic model called *dynamic cutset networks*, and finally derived decisions and explanations by performing reasoning over the latter. The main virtue of dynamic cutset networks is that they are tractable. Therefore, they can answer reasoning queries—both decision and explanation queries—accurately and often in linear time in the size of the model. We showed that for the task of activity recognition in cooking videos derived from the TACoS dataset, an XAI system based on our approach not only has better predictive performance than the one based on neural networks alone but also has superior explanatory power. The two main lessons learned from building this explainable system were:

- **Lesson 1:** To build models that are both accurate and explainable, *compose* a neural network defined over low-level features with probabilistic models that offer superior reasoning capabilities (and thus better explanations via reasoning).
- **Lesson 2:** In order to ensure that explanations are accurate, generated in real-time and faithful to the model used to make decisions, use *tractable* models.

We evaluated our explainable video activity recognition system by using it to solve three human-machine tasks: (1) answering yes/no questions posed over a given video; (2) answering whether a set of policies were followed or not in a given collection of videos; and (3) debugging the XAI model. We built novel interactive visual interfaces for each task and conducted human subjects studies. The lessons learned from these studies were:

- **Lesson 3:** If the XAI system for solving a human-machine task is highly accurate and the task is relatively easy, provide detailed explanations. Explanations significantly and justifiably improve the task completion time and the user's trust and reliance in the system, and reduce the gap between the user's perceived accuracy and the system's actual accuracy. On the other hand, if the XAI system for solving a human-machine task is not accurate, explanations wrongly amplify the automation bias in that they increase the gap between the user's perceived accuracy and the system's actual accuracy by increasing the former.
- **Lesson 4:** When the XAI system is used for solving relatively harder human-machine tasks (e.g., in safety-critical applications) and has known weaknesses, provide a balanced, detailed view of the system's strengths and weaknesses when the user begins to use the system. These early impressions play a critical role in establishing the user's assessment of model capabilities and developing an accurate mental model.
- **Lesson 5:** For expert users of the XAI system, develop novel visual exploratory tools that clearly depict not only local explanations specific to a query but also more global, higher-level explanations that can support identification of various types of model errors. These global views can help alleviate various biases such as automation bias, anchoring bias and bias towards focusing on false-positive errors, and help the expert user find high-level, global error patterns as well as instance-level problems.

6 | ACKNOWLEDGEMENTS

This work was supported by the DARPA Explainable Artificial Intelligence (XAI) Program under contract number N66001-17-2-4032.

References

1. M. Al-Hames and G. Rigoll. A multi-modal mixed-state dynamic bayesian network for robust meeting event recognition from disturbed data. In IEEE International Conference on Multimedia and Expo (ICME), pages 45–48, 2005.

2. F. R. Bach and M. I. Jordan. Thin junction trees. In Advances in Neural Information Processing Systems (NeurIPS), pages 569–576, 2002.

3. J. Bekker, J. Davis, A. Choi, A. Darwiche, and G. Van den Broeck. Tractable learning for complex probability queries. In Advances in Neural Information Processing Systems (NeurIPS), pages 2242–2250, 2015.

4. A. F. Bobick and A. D. Wilson. A state-based approach to the representation and recognition of gesture. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 19(12):1325–1337, 1997.

5. W. Brendel, A. Fern, and S. Todorovic. Probabilistic event logic for interval-based event recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3329–3336, 2011.

6. H. Buxton and S. Gong. Visual surveillance in a dynamic and uncertain world. Artificial Intelligence, 78(1-2):431–459, 1995.

7. C. K. Chow and C. N Liu. Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory, 14:462–467, 1968.

8. Gregory F Cooper. The computational complexity of probabilistic inference using bayesian belief networks. Artificial intelligence, 42(2-3):393–405, 1990.

9. A. Darwiche. A differential approach to inference in Bayesian networks. In Proceedings of the Sixteenth Conference on Uncertainty in Artifical Intelligence (UAI), pages 123–132, 2000.

10. R. Dechter and R. Mateescu. AND/OR search spaces for graphical models. Artificial Intelligence, 171:73–106, 2007.

11. N. Di Mauro, A. Vergari, and T. M. A. Basile. Learning Bayesian random cutset forests. In Foundations of Intelligent Systems - 22nd International Symposium, pages 122–132, 2015.

12. N. Di Mauro, A. Vergari, and F. Esposito. Learning accurate cutset networks by exploiting decomposability. In The Fourteenth International Conference of the Italian Association for Artificial Intelligence, pages 221–232, 2015.

13. J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2625–2634, 2015.

14. A. Doucet, N. de Freitas, K. P. Murphy, and S. J. Russell. Rao-Blackwellised particle filtering for dynamic Bayesian networks. In Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI), pages 176–183, 2000.

15. X. Duan, W. Huang, C. Gan, J. Wang, W. Zhu, and J. Huang. Weakly supervised dense event captioning in videos. In Advances in Neural Information Processing Systems (NeurIPS), pages 3063–3073, 2018.

16. N. Friedman, L. Getoor, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI), pages 1300–1309, 1999.

17. K. Goddard, A. Roudsari, and J. C. Wyatt. Automation bias: a systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association, 19(1):121–127, 2012.

18. S. Gong and T. Xiang. Recognition of group activities using dynamic probabilistic networks. In IEEE International Conference on Computer Vision (ICCV), pages 742–749, 2003.

19. K. A. Hoff and M. Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors, 57(3), 2015.

20. R. R. Hoffman. A taxonomy of emergent trusting in the human–machine relationship. Cognitive Systems Engineering: The Future for a Changing World, pages 137–164, 2017.

21. Y. A. Ivanov and A. F. Bobick. Recognition of visual activities and interactions by stochastic parsing. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 22(8):852–872, 2000.

22. S.W. Joo and R. Chellappa. Recognition of multi-object events using attribute grammars. In IEEE International Conference on Image Processing (ICIP), pages 2897–2900, 2006.

23. A. Kimmig, B. Demoen, L. De Raedt, V. S. Costa, and R. Rocha. On the implementation of the probabilistic logic programming language problog. Theory and Practice of Logic Programming, 11(2-3):235–262, 2011.

24. R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. Dense-captioning events in videos. In IEEE International Conference on Computer Vision (ICCV), pages 706–715, 2017.

25. T. Kulesza, M. Burnett, W. Wong, and S. Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In Proceedings of the Twentieth International Conference on Intelligent User Interfaces (IUI), 2015.

26. J. D. Lee and K. A. See. Trust in automation: Designing for appropriate reliance. Human factors, 46(1):50–80, 2004.

27. Y. Liang, J. Bekker, and G. Van den Broeck. Learning the structure of probabilistic sentential decision diagrams. In Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence (UAI), 2017.

28. D. Lowd and P. M. Domingos. Learning arithmetic circuits. In Proceedings of the Twenty-Fourth Conference in Uncertainty in Artificial Intelligence (UAI), pages 383–392, 2008.

29. V.I. Morariu and L.S. Davis. Multi-agent event recognition in structured scenarios. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3289–3296, 2011.

30. B. M. Muir and N. Moray. Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation. Ergonomics, 39(3):429–460, 1996.

31. B. M. Muir. Trust in automation: Part i. theoretical issues in the study of trust and human intervention in automated systems. Ergonomics, 37(11):1905–1922, 1994.

32. K. P. Murphy. Dynamic Bayesian networks: representation, inference and learning. PhD thesis, University of California, Berkeley, 2002.

33. I. Naim, Y. Song, Q. Liu, H. Kautz, J. Luo, and D. Gildea. Unsupervised alignment of natural language instructions with video segments. Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, pages 1558–1564, 2014.

34. P. Natarajan and R. Nevatia. Coupled hidden semi markov models for activity recognition. In IEEE Workshop on Motion and Video Computing (WMVC), pages 10–10, 2007.

35. A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In Proceedings of the Twenty-Second International Conference on Machine Learning (ICML), pages 625–632, 2005.

36. M. Nourani, D.R. Honeycutt, J.E Block, C. Roy, T. Rahman, E.D. Ragan, and V. Gogate. Investigating the importance of first impressions and explainable ai with interactive video analysis. In Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems Extended Abstracts, pages 1–8, 2020.

37. M. Nourani, C. Roy, T. Rahman, E. D. Ragan, N. Ruozzi, and V. Gogate. Don't explain without verifying veracity: An evaluation of explainable ai with video activity recognition. arXiv preprint arXiv:2005.02335, 2020.

38. M. Nourani, C. Roy, J.E. Block, D. R. Honeycutt, T. Rahman, E. D. Ragan, and V. Gogate. Anchoring bias affects mental models and user reliance in explainable ai systems. In Proceedings of the Twenty-Sixth International Conference on Intelligent User Interfaces (IUI), 2021.

39. M. Nourani, C. Roy, D. R. Honeycutt, E. D. Ragan, and V. Gogate. Detoxer: A visual debugging tool with multi-scope explanations for temporal multi-label classification models. In Manuscript submitted and under-review, 2021.

40. N. M. Oliver, B. Rosario, and A. P. Pentland. A bayesian computer vision system for modeling human interactions. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 22(8):831–843, 2000.

41. K. Otsuka, J. Yamato, Y. Takemae, and H. Murase. Conversation scene analysis with dynamic bayesian network basedon visual head tracking. In IEEE International Conference on Multimedia and Expo (ICME), pages 949–952, 2006.

42. J. Pearl. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, 1988.

43. M. Pei, Y. Jia, and S.C. Zhu. Parsing video events with goal inference and intent prediction. In IEEE International Conference on Computer Vision (ICCV), pages 487–494, 2011.

44. D. Poole. Probabilistic horn abduction and bayesian networks. Artificial intelligence, 64(1):81–129, 1993.

45. H. Poon and P. Domingos. Sum-Product Networks: A New Deep Architecture. In Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI), pages 337–346, 2011.

46. L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2):257–286, 1989.

47. T. Rahman and V. Gogate. Learning ensembles of cutset networks. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, pages 3301–3307, 2016.

48. T. Rahman, P. Kothalkar, and V. Gogate. Cutset networks: A simple, tractable, and scalable approach for improving the accuracy of chow-liu trees. In Proceedings of the 2014 Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD), pages 630–645, 2014.

49. T. Rahman, S. Jin, and V. Gogate. Cutset Bayesian Networks: A New Representation for Learning Rao-Blackwellised Graphical Models. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI), pages 5751–5757, 2019.

50. M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal. Grounding action descriptions in videos. Transactions of the Association for Computational Linguistics (TACL), 1:25–36, 2013.

51. A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, and B. Schiele. Coherent multi-sentence video description with variable level of detail. In The Thirty-Sixth German conference on Pattern Recognition, pages 184–195, 2014.

52. A. Rooshenas and D. Lowd. Learning sum-product networks with direct and indirect variable interactions. In Proceedings of the Thirty-First International Conference on Machine Learning (ICML), pages 710–718, 2014.

53. S. Roy and D. Suciu. A formal approach to finding explanations for database queries. In ACM SIGMOD International Conference on Management of Data (MOD), pages 1579–1590, 2014.

54. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV), 115(3), 2015.

55. M. S. Ryoo and J. K. Aggarwal. Recognition of composite human activities through context-free grammar based representation. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1709–1718, 2006.

56. Z. Si, M. Pei, B. Yao, and S.C. Zhu. Unsupervised learning of event and-or grammar and semantics from video. In IEEE International Conference on Computer Vision (ICCV), pages 41–48, 2011.

57. Y. C. Song, I. Naim, A. Al Mamun, K. Kulkarni, P. Singla, J. Luo, D. Gildea, and H. A. Kautz. Unsupervised alignment of actions in video with text descriptions. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI), pages 2025–2031, 2016.

58. T. Starner and A. Pentland. Real-time american sign language recognition from video using hidden markov models. In Motion-based recognition, pages 227–243. Springer, 1997.

59. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S.E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1–9, 2015.

60. Z. Zhang, T. Tan, and K. Huang. An extended grammar system for learning and recognizing complex visual events. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 33(2):240–255, 2010.

Metric	GoogLeNet	CCNs	Dynamic CCNs
K-1	0.9335	0.9677	**0.9687**
K-2	0.8557	0.8998	**0.9197**
K-3	0.7918	0.7962	**0.8168**
Jaccard Index	0.8608	0.8559	**0.8674**
Hamming Loss	0.1392	0.1286	**0.1160**

TABLE 1 Accuracy for Activity Recognition on Test Videos. Bold results indicate the best performing model.

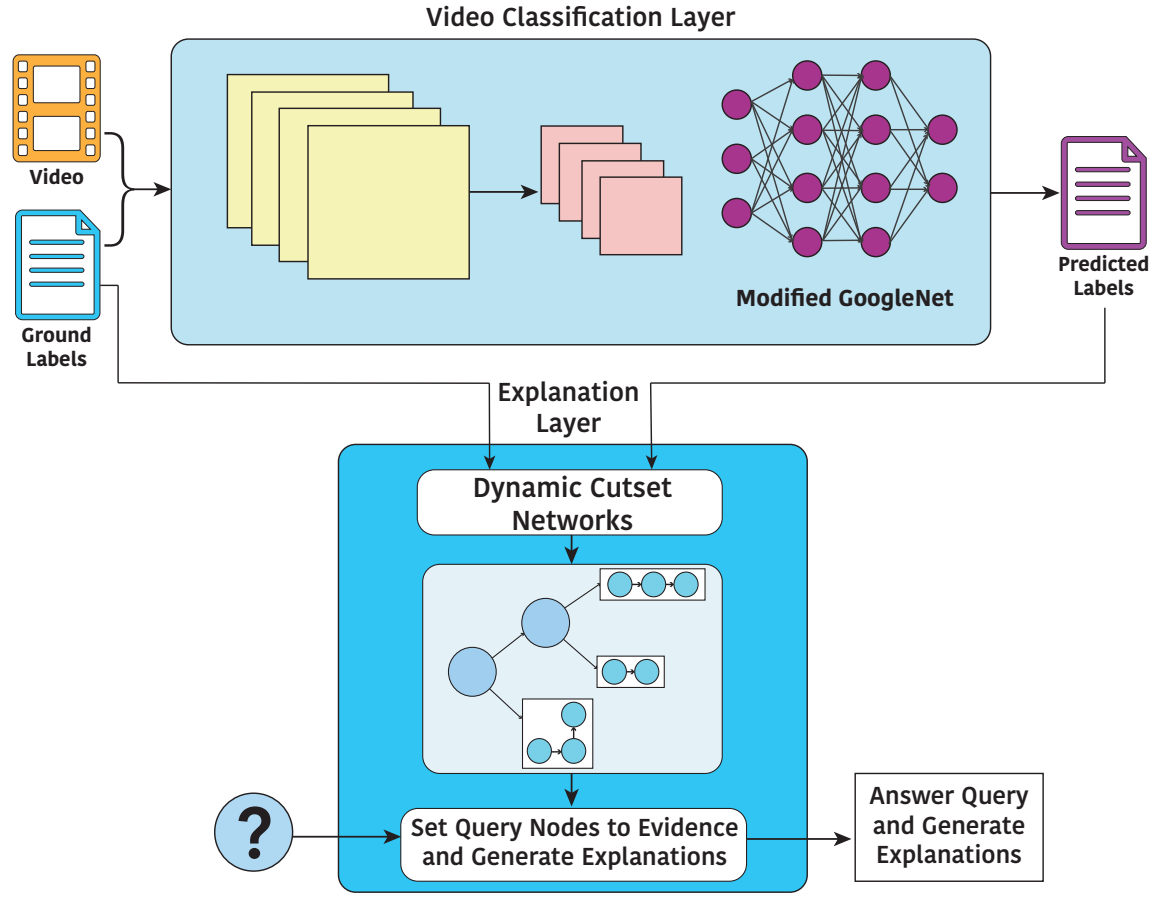
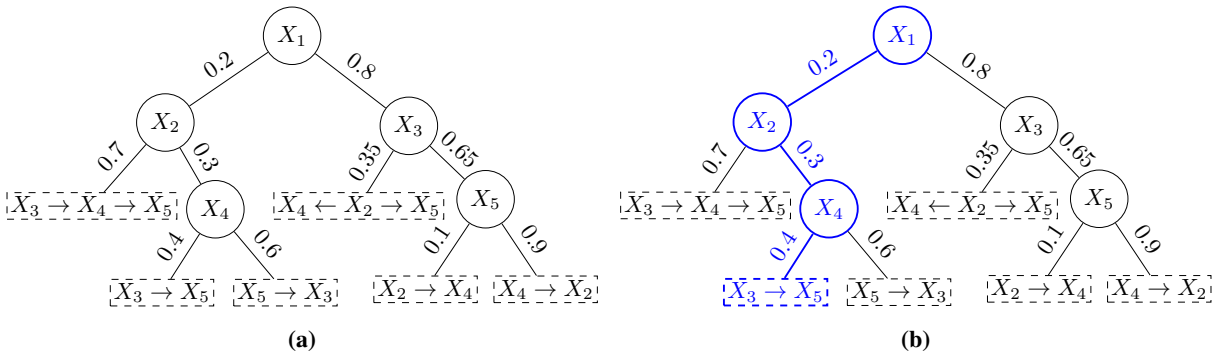


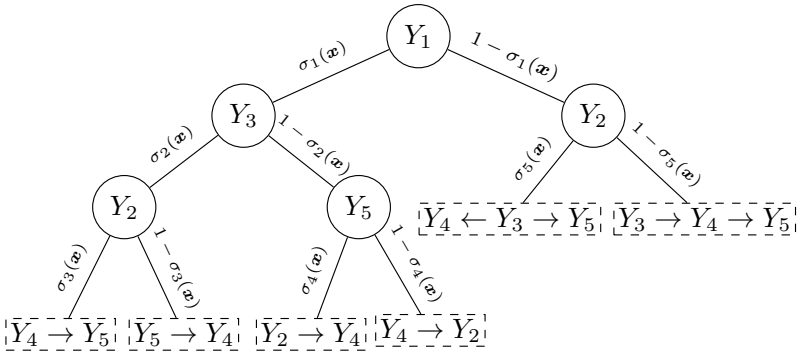
FIGURE 1 High-level Architecture and Data Processing Pipeline. Our system has two layers: a *video classification layer* based on a deep learning model whose output is fed to an *explanation layer* which is based on cutset networks[48], an interpretable, tractable probabilistic model. During the learning phase, the classification layer uses the video frames and the ground truth activities (labels) as input and learns a mapping from frames to object, action and location. On the other hand, during the learning phase, the explanation layer uses the labels predicted by the classification layer and ground truth as input and learns a mapping from predicted labels to the ground truth. During the query phase, the system answers questions by performing inference over the cutset network (in the explanation layer).

How to cite this article: C. Roy, M. Nourani, D. Honeycutt, J. Block, T. Rahman, E. Ragan, N. Ruozzi, and V. Gogate (2021), Explainable Activity Recognition in Videos: Lessons Learned, *Applied AI Journal*, *2021;00:1–10*.



(a) **(b)**

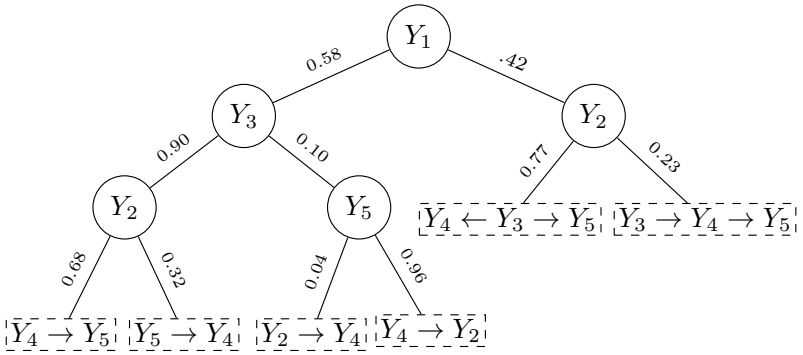
FIGURE 2 (a) A cutset network over 5 variables $\{X_1, \ldots, X_5\}$. Each variable takes a value from the binary domain $\{true, false\}$. OR nodes are denoted by circles. X_1 is the root node of the OR tree. Left and right arcs emanating from an OR node labeled by X_i indicate conditioning over *true* and *false* values of X_i respectively. Arcs emanating from OR nodes are labeled with conditional probabilities. For example, the arc labeled with 0.4 denotes the conditional probability $P(X_4 = true | X_1 = true, X_2 = false)$. The leaf nodes are denoted by dashed rectangles and contain tree Bayesian networks over the remaining variables not appearing on the path to the leaf. (b) The nodes, arcs and leaves activated during the computation of the query $P(X_1 = true, X_2 = false, X_3 = false, X_4 = true, X_5 = false)$.



(a)

$$\boldsymbol{W}_1 = [0.22, 0.3, -0.2, 0.1]$$
$$\boldsymbol{W}_2 = [1, 0.7, 0.5, 0.1]$$
$$\boldsymbol{W}_3 = [2.0, -0.33, -0.89, 0.9]$$
$$\boldsymbol{W}_4 = [-0.3, -2.0, -1.0, 1.0]$$
$$\boldsymbol{W}_5 = [0.01, 0.4, 0.8, 0.1]$$

(b)



(c)

FIGURE 3 (a) A conditional cutset network (CCN) representing $P(Y_1, \ldots, Y_5 | X_1, X_2, X_3)$. Arcs emanating from OR nodes are labeled with sigmoid functions σ_1 through σ_5. For brevity, we omit showing sigmoid functions for the conditional probability distributions in the tree Bayesian networks at the leaves. (b) Each $\boldsymbol{W}_i$ is the parameter vector of the corresponding sigmoid function $\sigma_i(\boldsymbol{x})$. (c) Given the assignment ($X_1 = true$,$X_2 = true$,$X_3 = false$), the CCN yields a cutset network having the same structure as the one given in (a) except that the parameters will be computed using the classifiers denoted by $\sigma_i(\boldsymbol{x})$.

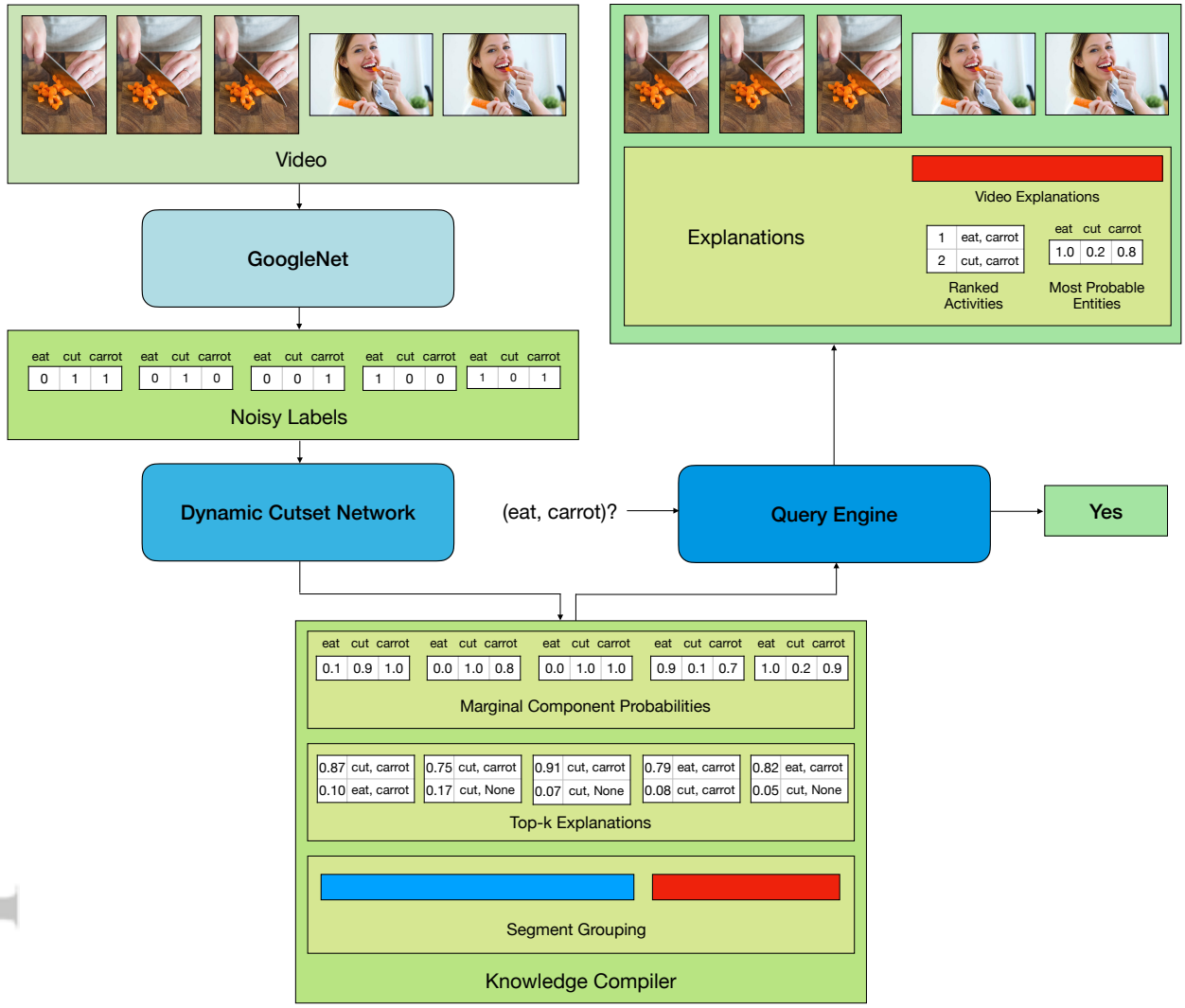


FIGURE 4 An example of a dummy video with five frames being passed through the compilation and query pipeline over three labels – *eat*, *cut* and *carrot*. For brevity, an activity is treated as a *pair* instead of a *triple* and comprises of only *action* and *object*. The frames are first individually passed through GoogleNet which assigns noisy labels to them. They are then passed through the Dynamic Cutset Network that groups the video into segments based on the top most likely activity given the noisy labels. In addition, it also computes the marginal component probabilities and other top-k explanations and stores them. Finally, when a query is posed to the system, the query engine searches for a segment that *completely* matches the query on *all* components. If such a segment exists, then it answers "Yes" (otherwise it answers "No" and the partial matches are shown as explanations). It then fetches the explanations and shows them to the user by highlighting the segment of the video that is matched and showing the ranked activities and most probable entities averaged over the entire segment.

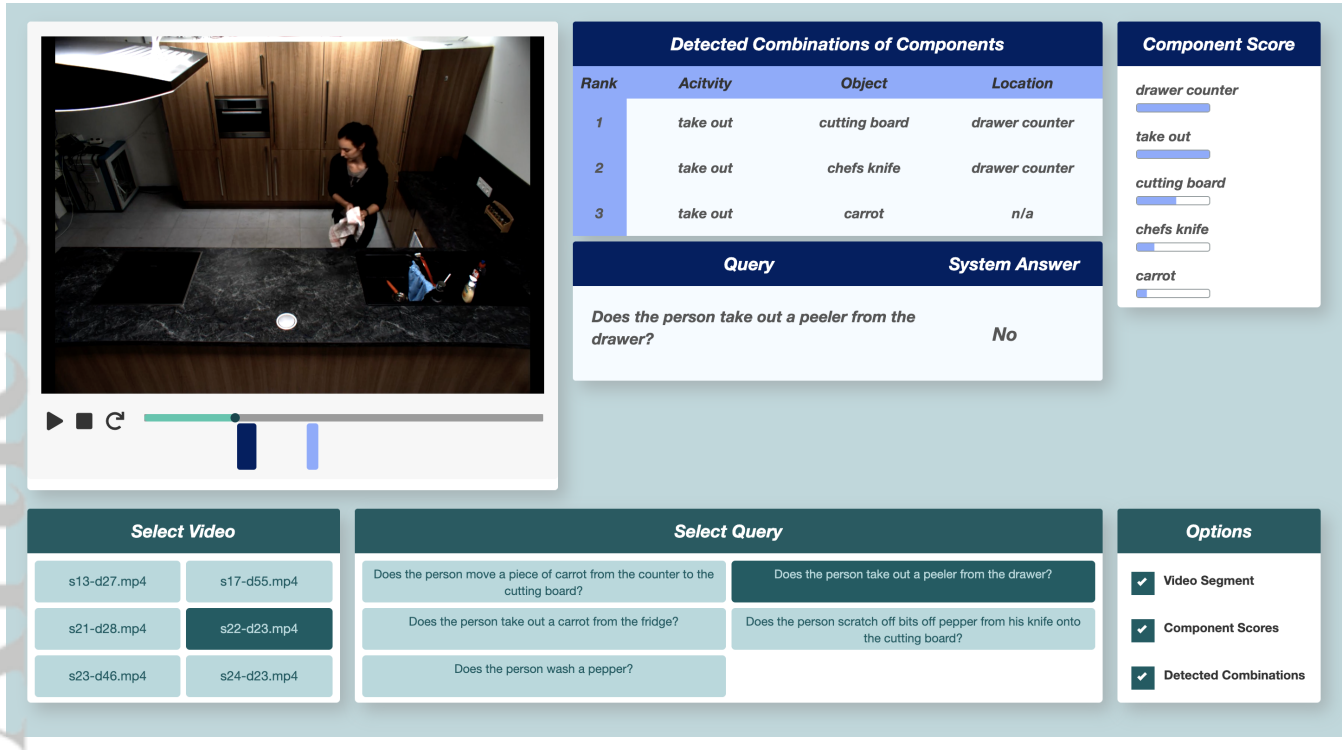


FIGURE 5 The interactive visual interface allows users to load videos and ask queries. The interface shows the ML system's answer along with explanatory elements for the output. The most relevant portions of the video play time are shown by colored bars beneath the video, and the right side shows detected video components and combinations of components relevant to the video and query. A similar version of this system was used in one of our previous user studies[37].

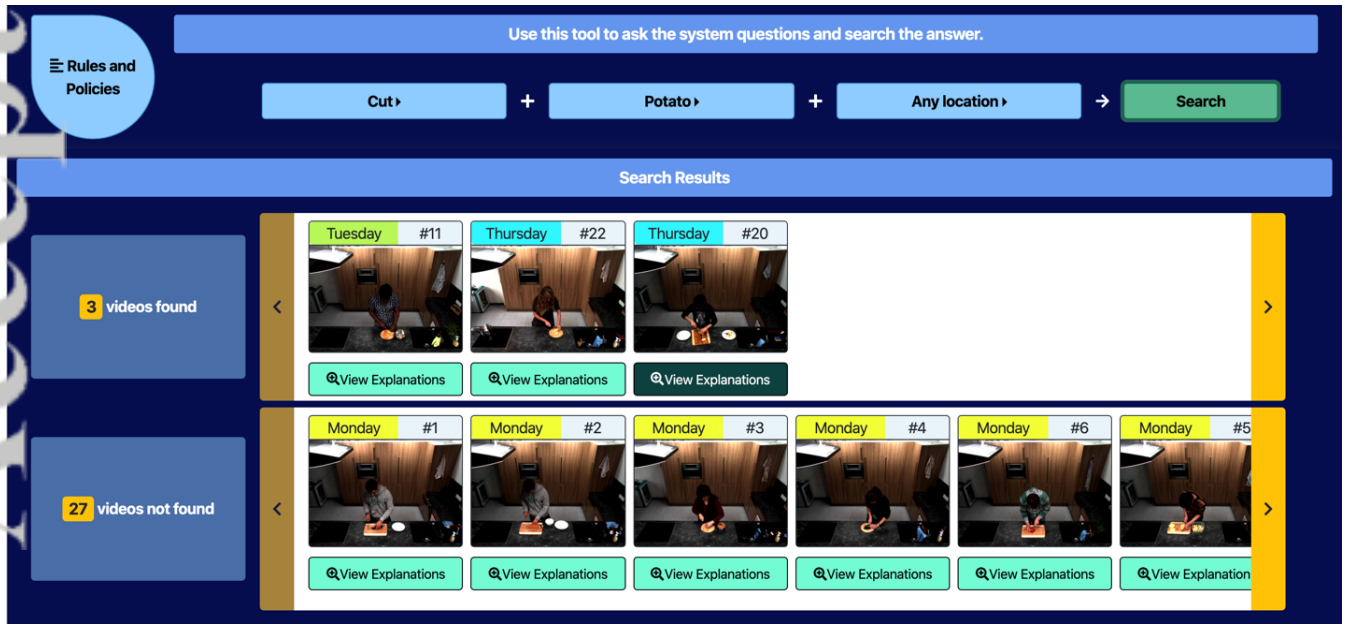
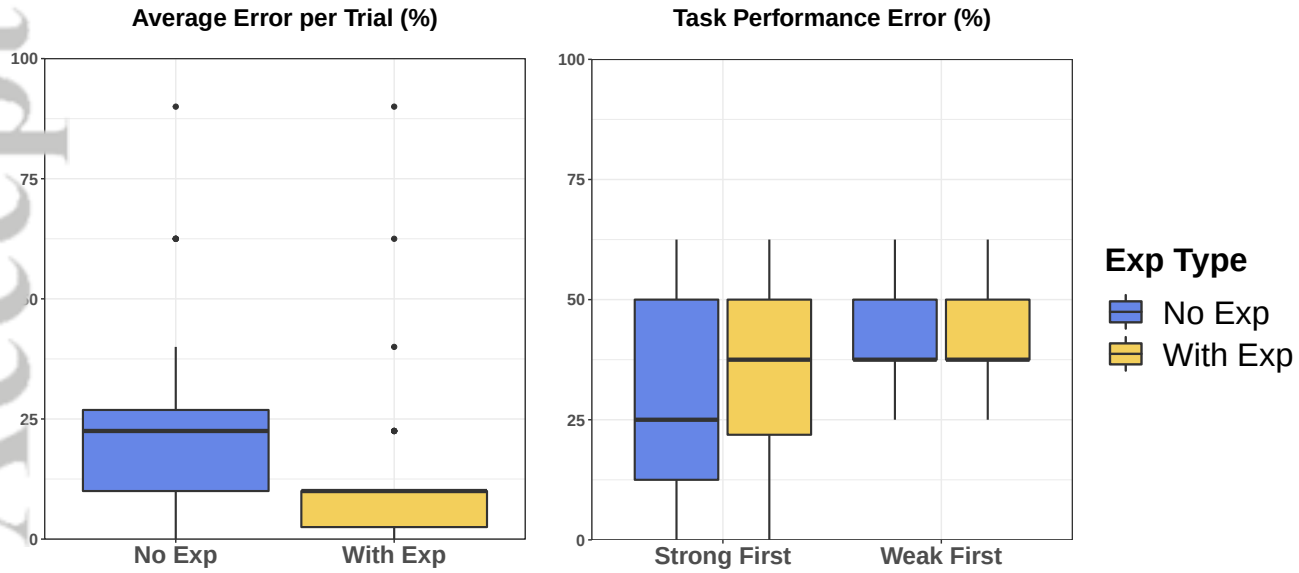


FIGURE 6 The overview of the query selection interface—as discussed in more detail in our previous work[38]—used in the policy-verification task, where users were able to query the model and find all the relevant videos. Upon clicking each video (or the View Explanations Bottom), a similar interface to Fig. 5 would pop-up.

FIGURE 7 The overview of the interface that included the global representations of the model as well as instance-level outputs. This interactive interface is not dependent on specific queries like those demonstrated in Fig. 5 and Fig. 6.



(a) Experiment 1: Human-Machine Query Verification **(b) Experiment 2**: Open-ended Video Review

FIGURE 8 (a) Participant error on the policy task (Percentage) from experiment 1 (See Section 4.2.1). **(b)** Average participant error per trial (Percentage) from experiment 2 (See Section 4.2.2).