2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Relevance for Human Robot Collaboration

Xiaotong Zhang, Dingcheng Huang and Kamal Youcef-Toumi

Abstract— **Inspired by the human ability to selectively focus on relevant information, this paper introduces relevance, a novel dimensionality reduction process that enables robots to identify relevant scene elements in a scene and generate responses that are seamless, fast, and accurate. To accurately and efficiently quantify relevance, we developed an event-based framework that maintains a continuous perception of the scene, evaluates cue sufficiency within the scene, and selectively triggers relevance determination. Within this framework, we developed a probabilistic methodology that considers various factors and is built on a novel structured scene representation. Both simulations and experimental results demonstrate the effectiveness of our relevance concept, as well as the proposed framework and methods for relevance quantification. Simulation results demonstrate that the relevance framework and methodology accurately predict the relevance of a general Human Robot Collaboration (HRC) setup, achieving a precision of 0.99, a recall of 0.94, an F1 score of 0.96, and an object ratio of 0.94. Relevance demonstrates broad benefits across multiple aspects of HRC, yielding a 79.56% reduction in task planning time compared with a state-of-the-art (SOTA) task planner for a cereal task, a 26.53% decrease in perception latency for object detection, an improvement of up to 13.50% in HRC safety, and an 80.84% reduction in the number of inquiries required during collaboration. A real-world demonstration highlights the effectiveness of the relevance framework, together with its modules, in providing intelligent and seamless assistance to humans during everyday tasks.**

I. Introduction

The growing integration of robots and automated systems into modern society underscores the necessity of advancing human-like intelligence and cognitive functions [1], [2] . A key example is the human capability to selectively focus on relevant physical elements in the environment, as well as on relevant abstract concepts or information in the input, guided by context, objectives, prior experiences, and/or reasoning. This cognitive function is closely associated with the reticular activating system (RAS), a network of neurons in the brain responsible for filtering sensory information, capturing relevant details, and suppressing irrelevant stimuli to optimize cognitive processing and decision-making [3]. Various studies have demonstrated the aforementioned human ability. [4].

Empowering robots with this human-like cognitive capability offers two significant benefits. First, robots engaged in close interaction with humans can achieve a level of scene understanding and reasoning similar to that of humans, leading to more natural and effective behavior and interaction [5], [6]. Second, by selectively focusing on the most relevant elements within a scene, robots can allocate their limited computation resources more effectively, reducing computational requirements, increasing processing speed, and even enhancing the precision and safety of robots in HRC tasks [7].

In this paper, we define and introduce a novel concept and approach, emulating the human RAS, for dimensionality reduction, termed 'relevance.' Relevance is defined as a dimensionality reduction process that continuously perceives the scene, identifies and detects its elements, and organizes them into a hierarchical representation, such as attribute classes. This process utilizes contextual cues to selectively reduce input dimensionality. When cues are insufficient, relevance-driven processing iteratively gathers additional cues and reapplies dimensionality reduction until sufficient information is obtained.

Previous works in this area to identify prominent features have primarily focused on visual saliency and attention mechanisms. Visual saliency aims to identify the most distinctive regions within an image based on the image features, such as color, intensity, and orientation [8]. As a result, most saliency-based methods detect visually conspicuous regions [9], [10], yet they usually overlook the rich contextual information that is critical for HRC applications. The attention mechanism constructs a mapping from the embedded input data to importance weights through a structured learnable function [11]. It acts as an auxiliary module within neural network models, enabling the system to weigh different components of the input in order to enhance performance. Attention has been widely adopted and implemented in various applications, ranging from end-to-end frameworks that generate actions directly from sensory inputs [12] to specialized sub-modules within robotic pipelines [13], [14].

Compared with saliency and attention, relevance is a comprehensive framework and concept that dramatically extends the scope and functionalities of prior approaches. Our relevance, in its current form, already considers contextual information, such as human objectives, tasks, environmental information, human preferences, etc. These factors are essential to enable proactive robotics assistance. Moreover, beyond assigning importance to input elements, relevance is also a novel framework that integrates a continuously running multi-modality perception module, a novel hierarchical scene representation, consideration of cue sufficiency in the input, etc. Those functions of relevance, which are not considered in saliency and attention, are essential for accurate, reliable,

All authors are with the Mechatronics Research Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA. Xiaotong Zhang and Dingcheng Huang contributed equally. {kevxt, dean1231, youcef}@mit.edu

This research was made possible by the support and partnership of King Abdulaziz City for Science and Technology (KACST) through the Center for Complex Engineering Systems at Massachusetts Institute of Technology (MIT) and KACST.

and seamless HRC. Last but not least, our relevance features a flexible formulation and computational principle not limited to the structure of neural networks. In this paper, as an illustration. We develop a probabilistic model that leverages a large language model (LLM) within an AI toolkit, human preferences, low-level attributes, and constraints to generate the necessary information. This formulation supports anticipatory capabilities, facilitates interactions across diverse information sources, and enables complex reasoning processes, ultimately pointing toward the long-term goal of artificial general intelligence (AGI). The demonstration in this paper shows, with relevance, that the robot can generate optimal reactions and seamless assistance to humans without transfer learning.

To quantify relevance, we propose and develop a novel event-based framework containing four modules: perception, triggers, relevance determination, and decision-making. The perception module operates continuously and acquires cues from the environment through multi-modality sensors and processing. The event-based mechanism selectively triggers relevance determination only under specific circumstances, resulting in a significant increase in computational efficiency. In the relevance determination module, flexible formulations and computation principles can be applied. As an illustration, we developed a unified probabilistic methodology based on a novel hierarchical representation of scenes, enabling accurate quantification of relevance. Finally, the decision making module generates appropriate actions based on these outcomes, leading to substantial improvements in HRC performance in terms of safety, efficiency, and seamlessness.

To summarize, the contributions of this paper are fourfold: 1. We introduced relevance, a novel concept and dimensionality reduction approach that enables robots to filter out redundant input and focus on relevant elements, thereby supporting fast, accurate, seamless, and proactive responses in HRC. 2. We developed an event-based framework uniquely for relevance that integrates a continuously operating perception module, selectively triggers downstream relevance determination and decision-making under specific circumstances, and incorporates cue sufficiency evaluation to enhance efficiency and accuracy of processing. 3. We developed a probabilistic methodology for relevance quantification and determination based on a novel hierarchical scene representation and designed with flexibility to accommodate diverse contextual factors. 4. We validated our framework and methodology using both simulations and real-world demonstrations, showing great improvements in processing efficiency, proactive assistance, and safety in HRC.

II. Related Works

Two areas of research, i.e., visual saliency and attention, share similar goals to our work but with much narrower scopes and functionalities. We also include a review of related HRC literature. To the best of our knowledge, this work is unique in HRC in that it empowers the robots with the cognitive capability to focus on selective components in the input by developing a novel framework and methodology with unique advantages, which are essential for HRC.

A. Visual Saliency

Visual saliency seeks to identify and localize the most distinctive regions or objects within an image based on visual conspicuity. Traditional methods are developed based on the features at the pixel level, including colors, intensity, orientation, etc [8], [9], [10]. Deep learning-based methods can also be developed and trained leveraging visual saliency datasets that are annotated either manually with a mouse click or automatically with an eye gaze tracker [15]. However, visual saliency mainly considers visual conspicuity, neglecting important considerations, such as context, for proactive assistance in HRC.

B. Attention

Attention is another mechanism that selectively focuses on relevant parts of the input data, but with limited scope and functionalities compared to relevance. In a nutshell, attention is structured with self-attention, cross-attention, multi-layer perception (MLP), etc., to formulate and optimize weights for the input features [11]. However, attention in robotics is mainly applied to end-to-end algorithms for better completion of the short-term action, such as pick and place [12]. Moreover, attention has several challenging limitations, such as limited applicability to end-to-end methods only, limited interpretability, and myopia. Relevance not only resembles the cognitive capability to reduce the dimension of original input but also alleviates the limitations of attention by developing a novel event-based framework and distinctive methodologies, such as a flexible formulation, continuously running perception, hierarchical scene representation, multi-level relevance determination, etc. Thus, our relevance can integrate the information and cues from an LLM in an AI toolkit, human preferences, human objectives, environmental constraints, etc., to enable accurate, seamless, and proactive assistance in HRC. Finally, relevance extends beyond current and past information to include future factors, such as those associated with a defined objective.

C. Human Robot Collaboration

There have been significant advancements in the fields related to HRC. We classify those works into two categories. The first category of works focuses on one specific important function for HRC, such as object detection [7], semantic segmentation [16], human intention prediction [17], [18], human motion prediction [19], human-aware task planning [20], etc. Those works, though important, have fundamentally different purposes from relevance. They are not for dimensionality reduction. The results of those modules can be applied for relevance quantification, and relevance can, vice versa, benefit these modules. The second category of work aims to develop a framework that directly generates actions from sensory inputs. Those works include learning from demonstration [21], reinforcement learning [22], etc. To incorporate the selective processing idea into these functions,

some work leverages attention mechanisms if the model is neural network-based [23]. However, because of the narrower scope of attention, as described in Section II.B, our relevance possesses more functions and introduces more benefits than attention. As shown in Sections V and VI, with relevance, robots can generate proactive, seamless, accurate, and structured reactions autonomously and efficiently.

III. Framework for Relevance

In this section, we introduce our novel event-based framework for relevance determination and robotic sensorimotor action generation, as shown in Fig. 1. The framework contains a perception module, triggers T check, relevance determination, and decision making.

A. Perception module

The perception module continuously processes the sensor inputs (camera, microphone, etc.) with an AI toolkit and obtains the features $\mathcal{F}$ of the scene. Unlike other pipeline frameworks for robots that process the perception and decision-making in a serial manner [19], our developed and proposed framework dynamically allocates the limited computation resources with an event-based mechanism as introduced in the following subsections. This enables our framework to perceive the scene in an uninterrupted manner to capture the fast and dynamic changes in the scene. Potential algorithms in the AI toolkit include You Only Look Once (YOLO), MMPose, CLIP, Visual Language Models (VLM), etc.

B. Triggers Check

The triggers T are to determine if a subprocess should be initialized for relevance determination based on the features $\mathcal{F}$ derived from the perception module. Potential criteria for trigger activation include changes in scene elements, updated objectives, or the introduction of new tasks for the human and/or the robot.

C. Relevance Determination

One unique advantage of relevance is that it enables flexible formulations and computation principles. In this paper, as an illustration, our relevance determination methodology is a probability framework based on a new type of scene representation comprising classes and elements, as described in Section IV.A. The relevance determination methodology proceeds in two steps that determine the relevance of classes and elements in a sequential manner. At the same time, the methodology evaluates cue sufficiency. With sufficient information, the framework completes the methodology and determines the relevance of the scene. With insufficient information, the framework requests or waits for the perception module to provide additional cues. The waiting period may apply to either a complete task or a subtask. During this time, the robot can continue performing other activities and will engage in the subtask once sufficient conditions are met. A comprehensive description of this methodology is provided in Section IV.

D. Decision Making

The output of the relevance determination module is subsequently leveraged in the decision-making to enable natural, faster, precise, and safer actions for proactive HRC.

IV. Methodology for Relevance

In this section, we introduce the methodology of the relevance determination module in our framework. An example instantiation of the derivation of each term's value is provided in Sections V and VI.

A. Class and Element Scene Representation

To quantify relevance efficiently and systematically, we propose a new representation $\mathcal{S}$. In $\mathcal{S}$, the scene is represented as a set of classes of elements:

$$\mathcal{S} = \{C_1, C_2 \ldots, C_m\} \tag{1}$$

where C_i represents the ith class of the classification results of all elements in $\mathcal{E}$. $\mathcal{E}$ is the set of all elements in the scene. m is the number of classes.

Each C_i contains a set of elements classified into the class C_i according to the criterion $k_i \in K$, where $K = \{k_1, k_2, \ldots, k_m\}$ is the set of classification criteria for each class. Thus, the set of elements belonging to the class C_i based on criterion k_i can be mathematically represented as:

$$C_i = \{e_j \mid e_j \in \mathcal{E}, e_j \text{ satisfies criterion } k_i\}. \tag{2}$$

Each element e_j is included in the class C_i if it satisfies the criteria k_i.

In organizing elements, we consider two primary types of criteria in this work: attribute criteria and functional criteria. Attribute criteria group elements based on shared characteristics or properties, such as similar semantics. For example, attribute-based classes include general groupings like a class of fruits or books. Other examples within attribute criteria are hierarchical criteria, temporal criteria, spatial criteria, quantitative criteria, etc. Functional criteria, on the other hand, group elements based on their roles, actions, or contributions to a common objective. Examples include behavioral classes, conceptual classes, and thematic classes.

B. Relevance Determination Mechanism

We developed and proposed a multi-level relevance determination mechanism that first operates on the class level and subsequently on the element level within the relevant classes. The relevance mechanism in both class and element level is defined as $\mathcal{M}$. The mechanism $\mathcal{M}$ takes the set of classes or elements and the set of features $\mathcal{F}$ as inputs. $\mathcal{F}$ can be represented as $\mathcal{F} = \{\mathcal{F}_v, \mathcal{F}_a, \mathcal{F}_c, \ldots\}$, where $\mathcal{F}_v$, $\mathcal{F}_a$, and $\mathcal{F}_c$ represent visual features (e.g., elements detected, classes classified, motion information), auditory features (e.g. conversational cues in the scene), and contextual features (e.g., objective, tasks, or environmental conditions), respectively.

There are two types of outcomes from the mechanism $\mathcal{M}$. Suppose the information is sufficient for the relevance quantification and determination. In this case, the outcomes

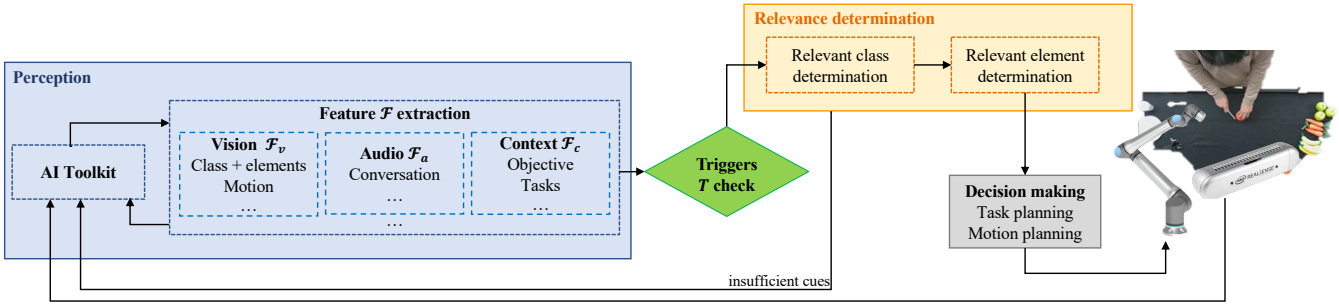

Fig. 1. Overview of the framework for relevance quantification and application of relevance for proactive human-robot collaboration. The framework consists of a continuously running perception module, a trigger check module to selectively initialize relevance determination, a two-level relevance determination methodology, and a decision-making module to generate natural and efficient human-robot interaction.

are the relevance scores for each class or element, and, at the same time, the set of relevant classes $\mathcal{C}_r$ or the set of relevant elements $\mathcal{E}_r$. If the information in $\mathcal{F}$ is not sufficient for relevance quantification and determination, then the algorithm will refer to the perception module to request or obtain new cues $\mathcal{F}_{new}$ from the scene before continuing the processing. Thus, the mechanism $\mathcal{M}$ can be mathematically represented as:

$$\mathcal{M}: \mathcal{F}, \mathcal{X} \rightarrow \begin{cases} \mathcal{R}(x)\ \forall x \in \mathcal{X}, \text{ and } \mathcal{X}_r & \text{if sufficient } \mathcal{F} \\ \text{ask or wait for } \mathcal{F}_{new}, & \text{else} \end{cases} \tag{3}$$

where $\mathcal{X}$ represents a set of classes or elements, x represents a specific class or element, and $\mathcal{X}_r$ represents a set of relevant classes or elements.

Both on the class level and the element level, the relevance $\mathcal{R}(x)$ is defined as the probability of x being relevant given the objective of the human $\mathcal{O}$:

$$\mathcal{R}(x) = P(x \mid \mathcal{O}) \tag{4}$$

(4) can also be conditioned on other factors according to the requirements of the applications. After $\mathcal{R}(x)$ is quantified, there are two possibilities: if $\mathcal{R}(x) \geq \tau_x$, then x is relevant to the objective in the scene and is added to the relevant set $\mathcal{X}_r$:

$$R(x) \geq \tau_{\mathrm{x}} \rightarrow x \text{ is relevant} \rightarrow \mathcal{X}_r = \mathcal{X}_r \cup x \tag{5}$$

If $\mathcal{R}(x)$ is smaller than τ_x, then x is deemed as irrelevant:

$$R(x) < \tau_{\mathrm{x}} \rightarrow x \text{ is irrelevant} \tag{6}$$

(4) can be influenced by a comprehensive set of possible factors. In this paper, we compute (4) based on four factors: the human's objective, tasks, preferences, and spatial placement.

C. Class Relevance Modeling

In this part, we model the class relevance based on (4), which is defined as:

$$\mathcal{R}(C_i) = P(C_i \mid \mathcal{O}) \tag{7}$$

There could be several possible tasks $\mathcal{T}$ associated with the objective $\mathcal{O}$. Thus, we use the total probability to decompose (7) and obtain:

$$\mathcal{R}(C_i) = \sum_{\mathcal{T}} P(C_i \mid \mathcal{T}, \mathcal{O}) P(\mathcal{T} \mid \mathcal{O}) \tag{8}$$

In (8), the first term $P(\mathcal{C}_i \mid \mathcal{T}, \mathcal{O})$ represents the probability that class C_i is relevant given the task is $\mathcal{T}$, and the objective is $\mathcal{O}$. The second term $P(\mathcal{T} \mid \mathcal{O})$ represents the probability that the current task is $\mathcal{T}$ given the objective is $\mathcal{O}$. If the human's historical preference data about the specific class C_i for the task $\mathcal{T}$ and objective $\mathcal{O}$ is available, then the first term can be derived as the probability of C_i is relevant based on the preference. If the historical preference data is not available for class C_i for task $\mathcal{T}$, the probability $P(C_i \mid \mathcal{T}, \mathcal{O})$ can be derived from predictions based on the tools in the AI toolkit, such as Large Language Models (LLM), other datasets that incorporate action sequences, other human cues, etc. The second term in (8) can be derived from prediction models for human tasks.

If any term in (8) is unavailable, or available but associated with high uncertainty, the information to determine the relevant class is deemed as not sufficient, and $\mathcal{F}_{new}$ is required for accurate quantification of the class relevance. The uncertainty of the term $P(\mathcal{T}|\mathcal{O})$ can be estimated using entropy as:

$$H(P(\mathcal{T}|\mathcal{O})) = -\sum_{\mathcal{T}} P(\mathcal{T}|\mathcal{O}) \ln P(\mathcal{T}|\mathcal{O}) \tag{9}$$

where H represents the entropy of a probability.

D. Element Relevance Modeling

With a similar methodology to class relevance modeling, the relevance for an element can be defined and modeled as:

$$\begin{aligned} \mathcal{R}(e_j) &= P(e_j|\mathcal{O}) \\ &= P(e_j|C_i, \mathcal{O}) P(C_i|\mathcal{O}) \\ &\quad + P(e_j|\neg C_i, \mathcal{O}) P(\neg C_i|\mathcal{O}) \end{aligned} \tag{10}$$

where $e_j \in C_i \in C_r$, and the symbol $\neg$ denotes logical negation. $P(e_j|\neg C_i, \mathcal{O})$ represents the probability that e_j is

relevant given the class that e_j belongs to is not relevant and the objective is $\mathcal{O}$, which equals to 0. Thus, (10) becomes:

$$\mathcal{R}(e_j) = P(e_j|C_i, \mathcal{O})\mathcal{R}(C_i) \tag{11}$$

An important observation is that, given the hierarchical representation and probabilistic formulation, (11) can be further extended to additional layers. For example, when modeling element relevance, we can model our element relevance based on a two-hierarchy approach in which each element is characterized by middle-level attributes, such as type, and low-level attributes, such as spatial or temporal factors. With this consideration, (11) can be computed as:

$$\begin{aligned}\mathcal{R}(e_j) &= \mathcal{R}(C_i)P(e_j|C_i, \mathcal{O}) \\ &= \mathcal{R}(C_i)(P(e_j|h_j, C_i, \mathcal{O})P(h_j|C_i, \mathcal{O}) \\ &\quad + P(e_j|\neg h_j, C_i, \mathcal{O})P(\neg h_j|C_i, \mathcal{O})) \\ &= R(C_i)P(e_j|h_j, C_i, \mathcal{O})P(h_j|C_i, \mathcal{O})\end{aligned} \tag{12}$$

where h_j represents the event that the middle-level attributes of e_j are relevant. The value $P(e_j|\neg h_j, C_i, \mathcal{O}) = 0$ in the above equation.

When determining if the information is sufficient for relevant element determination, the availability and/or uncertainty of the final three terms in (12) should be carefully examined, similarly to relevant class determination.

E. Constraints and Dependencies

Constraints and dependencies among elements, such as temporal and spatial constraints, can be considered in a post hoc manner for the downstream decision-making modules after relevance is determined. If a relevant element depends on or is constrained by an element, this element should also be considered in relevance and decision-making. It is critical to consider these constraints so that the downstream decision-making module can be informed comprehensively and produce better actions.

V. Simulation Evaluation

In this paper, we first conduct a comprehensive simulation-based evaluation to assess both the effectiveness of relevance quantification and its applicability to HRC.

A. Simulation Domain Setup

Since the concept of relevance is newly proposed in this paper, we introduce two distinct and representative HRC testing domains for evaluation. However, our relevance framework and methodology are sufficiently general to be extended to non-tabletop applications.

Coffee: The Coffee domain involves making coffee with creamer and passing it to humans. The original simple problem setup includes 19 objects with no task dependencies. In randomly generated cases, 10 to 30 elements from 200 irrelevant kitchenware items are added, along with three random spatial constraints introducing task dependencies. The hard problem setup starts with 36 elements, and randomly generated cases add 20 to 50 more elements, with eight spatial constraints. In both cases, the constraints require moving an object on top of another to access the lower one.

Cereal: The Cereal domain focuses on a robot preparing cereal, serving it to a person, and returning the items to their original places. In the original simple problem, there are 18 elements and 20 natural task dependencies, such as needing to open a cabinet to access the cereal. In random problem instances, 10 to 45 irrelevant kitchenware items are added. The hard version starts with 26 objects and 20 task dependencies, with an additional 30 to 60 irrelevant objects added in random cases without new task dependencies.

B. Relevance Quantification Evaluation

For the relevance quantification evaluation, we adopted and developed several metrics. First, we define precision $\mathfrak{P}$, recall $\mathfrak{R}$, and F_1 score $\mathfrak{F}$, which evaluate the prediction of the relevant element set based on the ground truth of the relevant element set. Moreover, we introduced a metric, $\mathfrak{N}$, representing the number ratio of predicted relevant elements to actual relevant elements for assessing speed improvement in downstream tasks. A higher $\mathfrak{N}$ will probably slow down the downstream tasks. Those metrics systematically evaluate the relevance quantification from aspects of effectiveness, completeness, and conciseness. Different applications will require different combinations of optimal metrics.

In this paper, we emphasize the simulation for the coffee domain for the brevity and clarity of our analysis. In the simulation, the objective is predefined as "get something to drink at the break of a conference", and the task is defined as "drink cold brew coffee". A large language model (LLM) GPT-4o is adopted to classify all the elements into the scene representation $\mathcal{S}$ using a criterion of semantics and provide the probabilities necessary for the relevance quantification, as illustrated in Section IV. We tested 25 combinations of the class thresholds τ_c and element thresholds τ_e, with 30 cases randomly generated for each combination.

The simulation results are shown in Fig. 2, which agree well with common sense. As τ_c and τ_e increase, more classes and elements will be pruned. Thus, the recall decreases, the precision increases, and the relevant object ratio decreases as these two thresholds increase. The F1 scores demonstrate a trend that increases first and then decreases with the threshold increase.

The trend of $\mathfrak{R}$ in Fig. 2(a) demonstrates that when τ_c and τ_e are equal to or smaller than 0.2, most relevant elements can be retained. At the same range of the thresholds, according to Fig. 2(d), the number of relevant elements predicted is very close to the number of relevant elements in the ground truth, which is very small when compared with the total number of elements in the scene. This demonstrates that our relevance quantification methodology can successfully remove most irrelevant elements while preserving relevant elements.

We select the thresholds τ_c and τ_e to be 0.2 to ensure better inclusion of relevant objects. At this threshold combination, our methodology achieves a recall $\mathfrak{R}$ of 0.94, a precision $\mathfrak{P}$ of 0.99, an F1 score $\mathfrak{F}$ of 0.96, and an object ratio $\mathfrak{N}$ of 0.94.

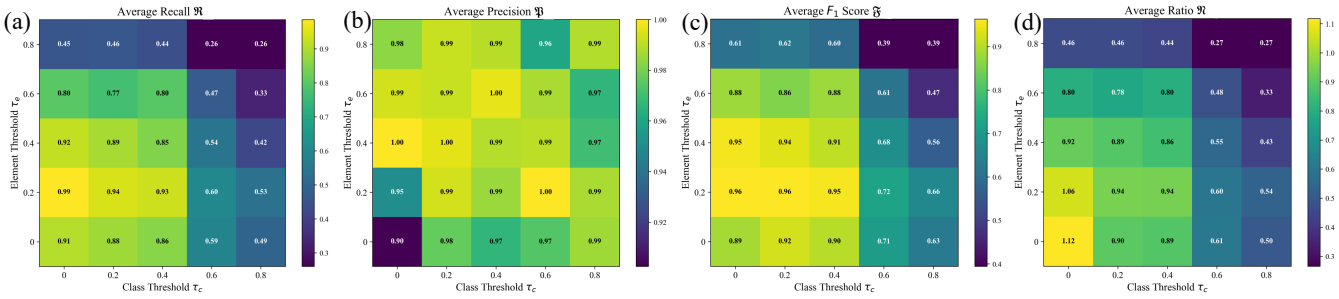

Fig. 2. Simulation results for the coffee domain. The values in the figures are averaged across 30 cases for each threshold combination. When τ_c and τ_e equal to 0.2, our methodology archives a recall $\mathfrak{R}$ of 0.94, a precision $\mathfrak{P}$ of 0.99, an F1 score $\mathfrak{F}$ of 0.96, and an object ratio $\mathfrak{N}$ of 0.94.

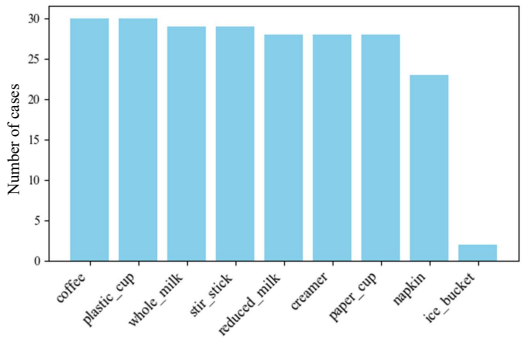

Fig. 3. Number of cases each element is predicted to be relevant out of 30 cases without considering constraints. τ_c and τ_e are 0.2. Our relevance quantification method predicts relevance accurately and reliably.

These results demonstrate that our methodology is effective and accurate in dimension reduction. The number of cases in which each type of element is relevant is shown in Fig. 3 without considering the constraints. Of the 30 cases, a couple of them randomly add an ice bucket to the setup and predict that the ice bucket is relevant, which agrees with common sense in the context of drinking cold brew coffee. All 30 cases predict that coffee and plastic cups are relevant. These results further demonstrate that relevance prediction is very accurate.

C. Relevance in HRC

One benefit of relevance in HRC is that relevance can help the robot better reason about the scene and the human's requirements in a well-structured manner, which resembles how humans interact with other humans. This benefit can be demonstrated with a simple robot inquiry example in a coffee serving setup. The human-like robot will first detect the relevant classes and types of elements. If the relevance of any classes or types of elements is high, those classes or types are deemed necessary (coffee in this domain). Other relevant classes or types of elements are optional, and the robot asks the human if they need those relevant elements. The number of inquiries required using different methods is shown in Table I. Note that the number of inquiries required without relevance equals the average number of objects in the testing setup. It is shown that with accurate and proper reasoning

TABLE I

NUMBER OF INQUIRIES REQUIRED FOR APPROPRIATE ASSISTANCE. RELEVANCE DRAMATICALLY REDUCES THE NUMBER OF INQUIRIES.

method	inquiry elements	count
relevance with necessity	**creamer, plastic cup, paper cup, whole milk, reduced milk, stick, napkin**	**7**
relevance	coffee, creamer, plastic cup, paper cup, whole milk, reduced milk, stick, napkin	8
no relevance	Everything in the scene	36.53

about the relevance of an element to a task or objective, the number of inquiries required for appropriate assistance is reduced by 80.84% compared with the number without relevance. Thus, human-robot interactions can be much more natural and fluent. We use objective metrics to evaluate effectiveness, as they provide rigorous, consistent, and comparable results, while subjective evaluations are left for future work. A full demonstration and additional discussions are provided in Section VI, along with the accompanying video.

D. Relevance in Task Planning

This section demonstrates the effectiveness of relevance in task planning. A Fast Downward planner is employed. We compare three methods: our relevance with 0.2 as thresholds, random relevance with 0.2 as thresholds, and pure planning without relevance. For each method, we evaluate the processing time (relevance computation + planning), the timeout rate (120s), and the failure rate. Results are shown in Table II. Note that the failure rate is not applicable to pure planning, as solutions are guaranteed to be found if they exist.

Relevance dramatically improves task planning performance. Compared with Pure Planning, relevance significantly reduces the time cost by up to 79.56% and the timeout rate across all problem formulations. As the problem formulation becomes more complex, we observe a notable increase in the timeout rate and planning time for Pure Plan, highlighting its inefficiency in solving complex task-planning problems. When compared with random relevance, we observe a substantial reduction in failure rates, demonstrating that our relevance successfully and accurately determines the relevant elements of the scene. Moreover, the highest failure rate after adopting relevance in the four test cases is 0.04, which demonstrates that our relevance can effectively, accurately,

TABLE II

TASK PLANNING PERFORMANCE COMPARISON. OUR METHODOLOGY OF RELEVANCE DETERMINATION ACCURATELY PREDICTS RELEVANCE AND DRAMATICALLY REDUCES THE PLANNING TIME.

Domains	Metrics	Relevance	Pure Planning	Random Relevance
Coffee simple	Time (s)	**16.76**	45.84	21.61
	Timeout rate	**0.00**	0.22	0.05
	Failure rate	0.04	-	0.62
Coffee hard	Time (s)	36.91	-	**21.38**
	Timeout rate	**0.00**	1.00	0.02
	Failure rate	0.01	**0.00**	0.88
Cereal simple	Time (s)	**13.37**	63.84	39.61
	Timeout rate	**0.00**	0.38	0.05
	Failure rate	0.02	-	0.58
Cereal hard	Time (s)	**30.70**	-	53.82
	Timeout rate	**0.00**	1.00	0.12
	Failure rate	0.04	-	0.60

and robustly reduce the task planning time.

E. Relevance in Perception

One previous work demonstrated that by selectively processing the regions containing relevant elements, a notable maximum reduction of 30.09% in inference time and 26.53% in total time per frame of an object detector is achieved. Additionally, this processing strategy improves two safety metrics by 11.25% and 13.50%, respectively [7].

VI. EXPERIMENTAL DEMONSTRATION

In this section, we present a real-world demonstration to verify the effectiveness of our proposed framework and relevance determination methodology.

A. Experimental Setup

The experimental setup is shown in Fig. 4(a). On one end of the table, a UR5 robot arm with a robotiq gripper is mounted to reason about the relevance and generate actions to assist humans. The table is a snack table for a conference with desserts, drinks, and utensils. A microphone is placed on the table to pick up the audio information and cues in the scene. The HRC task is for the robot to assist two humans into the scene one by one. The robot utilized relevance for optimal decision-making and HRC.

The demonstration is implemented in Python using the `threading` package for multi-threading and asynchronous computation. The communication between threads is achieved using `Event` and `Queue`. At the beginning of the code, we initialized two threads, as shown in the perception module of Fig. 1, one for the visual information processing using the OpenAI API and another one for picking up the microphone and processing the audio information using the Assembly AI API. In the visual information processing thread, we currently implement a VLM model for visual information processing, extracting the semantic information of all the objects on the table, classifying the objects into class and element representation, and detecting the human motion and objectives. For each iteration in the visual threads, the trigger criteria are checked. We implement the trigger criterion, the changes in human numbers in the scene, which is sufficient for our current demonstration. Once a trigger criterion is met, a new thread for relevance determination is initialized, and relevance is determined based on the methodologies in Section IV. In the setup without preference information, the probabilities required for the computation are derived from an LLM model in the AI toolkit. We currently leverage the OpenAI API, and the model we used is GPT-4o. Within the same thread, the action is generated based on the relevance identified to best assist humans.

B. Demonstration Results

The demonstration results are shown in Fig. 4(b-h). In Fig. 4 (b), in the absence of a human in the scene, due to the lack of triggers, the perception module is continuously running, and no relevance determination is initialized. In Fig. 4(c-d), a human agent with recorded preferences for coffee enters the scene and grabs a coffee. The trigger is activated, and the human's objective and relevance are determined based on the methodologies we described in Section IV. Based on the recorded preferences, the robot actions are automatically generated to serve the human agent whole milk and a stir stick. In Fig. 4(e), a second human enters without indicative actions. The trigger is activated, but the relevance determination module lacks sufficient cues. The system continues looping through the perception module to collect more information. In Fig. 4(f-g), a conversation between two human agents about coffee provides more cues to determine relevance. Without preference for the second human, the robot first generates actions related to the necessary elements (i.e., the coffee) and then inquires about the second human's preferences for making the coffee. Based on the human response, the system generates a complete action sequence to serve the second human agent. In Fig. 4(h), two humans successfully get their preferred coffee with the robot's assistance, demonstrating the effectiveness of relevance.

Without the unique components in relevance, this type of HRC assistance in a general setup will be challenging. Cue sufficiency determination prevents the robot from providing false assistance, such as when no human is present or when a second person enters without indicative cues. Our flexible probabilistic framework accommodates diverse information sources, including typical human behavior, individual preferences, proximity, etc. The effectiveness of our hierarchical scene representation is further validated through this demonstration. In the video, the robot first generates its inquiries at the class level and subsequently reasons using proximity information to refine its decisions at the element level. This hierarchical reasoning process enables the robot to act in a manner that is both more natural and intuitive.

VII. CONCLUSION

This paper introduced relevance, a novel concept and framework for dimensionality reduction in human–robot

Fig. 4. The illustration of (a) experimental setup and (b-h) demonstration results. With relevance, the robot successfully and seamlessly assists two human agents with cold-brew coffee drinking.

collaboration. The proposed framework integrates a continuously operating perception module, an event-based triggering mechanism, cue sufficiency consideration, and a probabilistic methodology formulated on a hierarchical scene representation to enable accurate and efficient identification of relevant elements. Comprehensive simulation studies validated the effectiveness of the framework, achieving a precision of 0.99, recall of 0.94, F1 score of 0.96, and object ratio of 0.94, while reducing task planning time by 79.56%, perception latency by 26.53%, and the number of inquiries by 80.84%, with a 13.50% improvement in safety. A real-world demonstration further confirmed the capability of relevance to facilitate proactive, seamless, and structured human–robot interaction. These findings establish relevance as a substantive advancement toward the development of versatile robotic assistants capable of providing intelligent, context-aware, and human-centered support in everyday tasks.

REFERENCES

[1] X. Zhang, A. Al Alsheikh, and K. Youcef-Toumi, "Systematic evaluation and analysis on hybrid strategies of automatic agent last-mile delivery," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 2733–2740.

[2] X. Zhang and K. Youcef-Toumi, "Magnetohydrodynamic energy harvester for low-power pipe instrumentation," *IEEE/ASME Transactions on Mechatronics*, vol. 27, no. 6, pp. 4718–4728, 2022.

[3] G. Moruzzi and H. W. Magoun, "Brain stem reticular formation and activation of the eeg," *Electroencephalography and clinical neurophysiology*, vol. 1, no. 1-4, pp. 455–473, 1949.

[4] A. Mack, "Inattentional blindness: Looking without seeing," *Current directions in psychological science*, vol. 12, no. 5, pp. 180–184, 2003.

[5] X. Zhang, "thesis in the field of mechanical engineering: Relevance for human-robot collaboration: Definitions, systems, algorithms, and applications," Ph.D. dissertation, Massachusetts Institute of Technology, 2025.

[6] X. Zhang, D. Huang, and K. Youcef-Toumi, "Relevance-driven decision making for safer and more efficient human robot collaboration," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 5899–5905.

[7] X. Zhang, J. Chong, and K. Youcef-Toumi, "How does perception affect safety: New metrics and strategy," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 13 411–13 417.

[8] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[9] D. Meger, P.-E. Forssén, K. Lai, S. Helmer, S. McCann, T. Southey, M. Baumann, J. J. Little, and D. G. Lowe, "Curious george: An attentive semantic robot," *Robotics and Autonomous Systems*, vol. 56, no. 6, pp. 503–511, 2008.

[10] D. Chung, R. Hirata, T. N. Mundhenk, J. Ng, R. J. Peters, E. Pichon, A. Tsui, T. Ventrice, D. Walther, P. Williams, *et al.*, "A new robotics platform for neuromorphic vision: Beobots," in *Biologically Motivated Computer Vision: Second International Workshop, BMCV 2002 Tübingen, Germany, November 22–24, 2002 Proceedings 2*. Springer, 2002, pp. 558–566.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[12] Y. Zhu, A. Joshi, P. Stone, and Y. Zhu, "Viola: Imitation learning for vision-based manipulation with object proposal priors," in *Conference on Robot Learning*. PMLR, 2023, pp. 1199–1210.

[13] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[14] K. Kedia, A. Bhardwaj, P. Dan, and S. Choudhury, "Interact: Transformer models for human intent prediction conditioned on robot actions," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 621–628.

[15] J. Pan, C. C. Ferrer, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. Giro-i Nieto, "Salgan: Visual saliency prediction with generative adversarial networks," *arXiv preprint arXiv:1701.01081*, 2017.

[16] C. Munasinghe, F. M. Amin, D. Scaramuzza, and H. W. van de Venn, "Covered, collaborative robot environment dataset for 3d semantic segmentation," in *2022 IEEE 27th International Conference on Emerging Technologies and Factory Automation (ETFA)*. IEEE, 2022, pp. 1–4.

[17] V. Hernandez-Cruz, X. Zhang, and K. Youcef-Toumi, "Bayesian intention for enhanced human robot collaboration," in *2025 IEEE 21st International Conference on Automation Science and Engineering (CASE)*, 2025, pp. 1024–1030.

[18] L. Hu, X. Zhang, and K. Youcef-Toumi, "Eye movement feature-guided signal de-drifting in electrooculography systems," in *2025 IEEE 21st International Conference on Automation Science and Engineering (CASE)*, 2025, pp. 519–526.

[19] A. Kothari, T. Tohme, X. Zhang, and K. Youcef-Toumi, "Enhanced human-robot collaboration using constrained probabilistic human-motion prediction," in *2025 IEEE 21st International Conference on Automation Science and Engineering (CASE)*, 2025, pp. 541–547.

[20] Y. Cheng, L. Sun, and M. Tomizuka, "Human-aware robot task planning based on a hierarchical task model," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1136–1143, 2021.

[21] W. Wang, R. Li, Y. Chen, Z. M. Diekel, and Y. Jia, "Facilitating human–robot collaborative tasks by teaching-learning-collaboration from human demonstrations," *IEEE Transactions on Automation Science and Engineering*, vol. 16, no. 2, pp. 640–653, 2019.

[22] M. El-Shamouty, X. Wu, S. Yang, M. Albus, and M. F. Huber, "Towards safe human-robot collaboration using deep reinforcement learning," in *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2020, pp. 4899–4905.

[23] R. Zhang, J. Lv, J. Li, J. Bao, P. Zheng, and T. Peng, "A graph-based reinforcement learning-enabled approach for adaptive human-robot collaborative assembly operations," *Journal of Manufacturing Systems*, vol. 63, pp. 491–503, 2022.