A Dynamic Bayesian Network Based Framework for Multimodal Context-Aware Interactions

Violet Yinuo Han
Meta Reality Labs
Redmond, WA, USA
Carnegie Mellon University
Pittsburgh, PA, USA
yinuoh@andrew.cmu.edu

Tianyi Wang
Meta Reality Labs
Redmond, WA, USA
tianyiwang@meta.com

Hyunsung Cho
Meta Reality Labs
Redmond, WA, USA
Carnegie Mellon University
Pittsburgh, PA, USA
hyunsung@cs.cmu.edu

Kashyap Todi
Meta Reality Labs
Redmond, WA, USA
kashyap.todi@gmail.com

Ajoy Savio Fernandes
Meta Reality Labs
Redmond, WA, USA
ajoyferns@meta.com

Andre Levi
Meta Reality Labs
Redmond, WA, USA
andrelevi@meta.com

Zheng Zhang
Meta Reality Labs
Redmond, WA, USA
University of Notre Dame
Notre Dame, IN, USA
zzhang37@nd.edu

Tovi Grossman
University of Toronto
Toronto, ON, Canada
tovi@dgp.toronto.edu

Alexandra Ion
Carnegie Mellon University
Pittsburgh, PA, USA
alexandraion@cmu.edu

Tanya R. Jonker
Meta Reality Labs
Redmond, WA, USA
tanya.jonker@meta.com

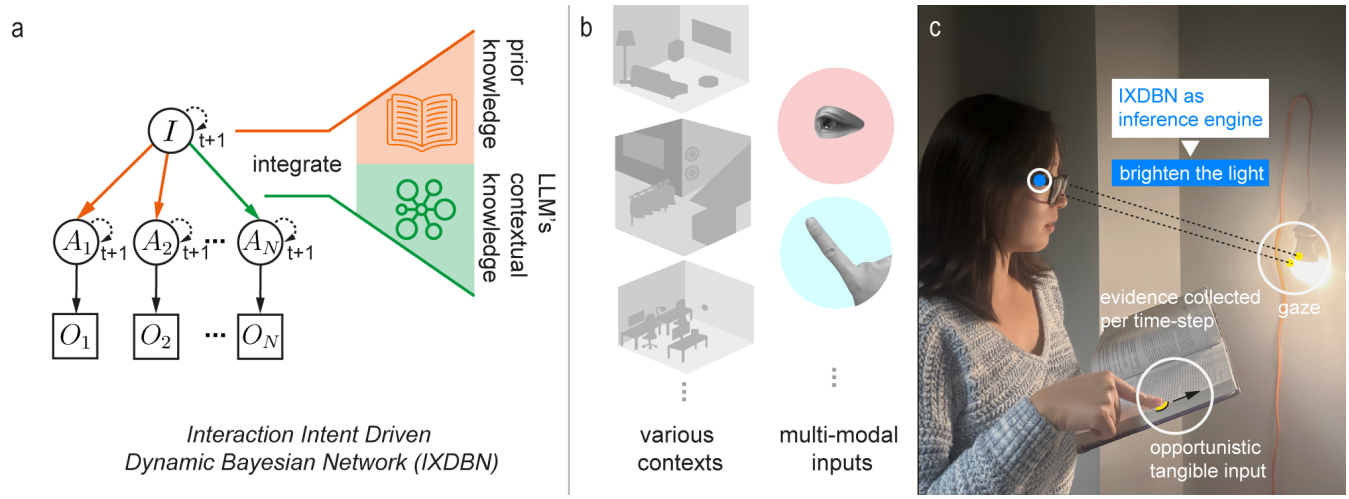

Figure 1: (a) We propose a dynamic Bayesian network that integrates prior knowledge and a large language model (LLM)'s contextual knowledge to support multimodal interactions. An example of our approach, Interaction Intent-driven Dynamic Bayesian Network (IXDBN), models the user's intent (I)-to-action (A) process based on observations (O). (b) Our approach enables adaptation to various contexts and scalability to different multimodal inputs. (c) During interaction, the IXDBN acts as an inference engine on a wearable device, inferring a user's interaction intent dynamically on-the-go. In this example, the user's gaze is directed at a light fixture while performing an opportunistic tangible input–pressing and sliding upward on a nearby surface. The IXDBN collects these multimodal signals as evidence over time, interpreting them as an intent to adjust the lighting, and executes the inferred interaction intent of brightening the light.

IUI '25, Cagliari, Italy

ACM ISBN 979-8-4007-1306-4/25/03
https://doi.org/10.1145/3708359.3712070.

Abstract

Multimodal context-aware interactions integrate multiple sensory inputs, such as gaze, gestures, speech, and environmental signals, to provide adaptive support across diverse user contexts. Building such systems is challenging due to the complexity of sensor fusion, real-time decision-making, and managing uncertainties from noisy inputs. To address these challenges, we propose a hybrid approach combining a dynamic Bayesian network (DBN) with a large language model (LLM). The DBN offers a probabilistic framework for modeling variables, relationships, and temporal dependencies, enabling robust, real-time inference of user intent, while the LLM incorporates world knowledge for contextual reasoning. We demonstrate our approach with a tri-level DBN implementation for tangible interactions, integrating gaze and hand actions to infer user intent in real time. A user evaluation with 10 participants in an everyday office scenario showed that our system can accurately and efficiently infer user intentions, achieving 0.83 *per frame* accuracy, even in complex environments. These results validate the effectiveness of the DBN+LLM framework for multimodal context-aware interactions.

CCS Concepts

• **Human-centered computing** → **HCI theory, concepts and models**; • **Computing methodologies** → *Modeling methodologies.*

Keywords

Computational Interaction, Dynamic Bayesian Networks, Multimodal Interaction, Context-Aware Interaction, Bayesian Inference, Large Language Models, User Modeling

ACM Reference Format:

Violet Yinuo Han, Tianyi Wang, Hyunsung Cho, Kashyap Todi, Ajoy Savio Fernandes, Andre Levi, Zheng Zhang, Tovi Grossman, Alexandra Ion, and Tanya R. Jonker. 2025. A Dynamic Bayesian Network Based Framework for Multimodal Context-Aware Interactions. In *30th International Conference on Intelligent User Interfaces (IUI '25), March 24–27, 2025, Cagliari, Italy.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3708359.3712070.

1 Introduction

Multimodal context-aware interactions refer to interactions that integrate multiple sensory inputs, such as gaze, gestures, speech, and environmental signals, to provide adaptive real-time support across a wide range of user contexts [52, 53]. Such interactions aim to respond dynamically to user intent and environmental changes, using data from various sources to infer meaning and deliver services or control devices seamlessly. For instance, a smart home system may combine gaze tracking, voice commands, and hand gestures to control lights or appliances, while an augmented reality (AR) interface could use gaze data, accelerometer input, and geometry of objects in the environment to adjust their visibility, placement, or level of detail of user interface elements.

The idea of multimodal interactions was pioneered by "Put-That-There" in 1980, which demonstrated intuitive interactions with a computing system using speech and hand gestures [7]. Since then, research has expanded to explore combinations of different types of user input. For instance, gesture input has commonly been used with eye-based inputs, such as gaze [11] and blinking [69] to disambiguate otherwise error-prone user inputs [33], such as gaze input alone. Many input types have been explored for multimodal interactions, such as tongue movement [20] or facial expressions [73].

The integration of new input modalities for existing applications is difficult to achieve [39]. One common approach is to develop tools that enable end users to define mappings, such as mapping rules between body gestures and an interface output [18]. These tools take different mechanisms, for example, the trigger-action paradigm [21] or programming by demonstration [41].

Beyond supporting diverse user input modalities, multimodal interactions must also adapt to contextual changes that users experience (e.g., transitioning from a private to a public space) [52]. Prior work has explored context-dependent multimodal interaction in various settings, including smart homes systems [36], design studios [1], vehicles [19], and accessible spaces [67]. These studies demonstrate the interaction benefits of pre-designed context-aware interactions tailored to specific environments.

A key characteristic of an ideal system would be to continuously adapt to user input flexibly, spanning different input modalities and evolving contexts to provide interaction support that feels natural and fluid across different scenarios. Building such multimodal, context-aware interactions presents several challenges. Sensor fusion, or combining and interpreting data from multiple, often noisy input sources, is complex and error-prone. Multimodal context-aware interaction systems must not only process data in real time but also need to understand the user's intent across varied contexts, which can change dynamically. Furthermore, handling uncertainty and errors from sensor inputs, such as false gesture detection or gaze drift, adds another layer of complexity.

In this paper, we introduce a novel approach to address these challenges and support multimodal, context-aware interaction (Figure 1). The main components of our computational framework are a *dynamic Bayesian network (DBN)* to model users' real-time intent-to-action process, and elicitation from a *large language model (LLM)* to integrate contextual and domain-specific reasoning (Figure 1a).

The DBN provides a probabilistic framework for integrating real-time user input, allowing us to handle the uncertainties inherent in these systems. By updating its beliefs dynamically as new data become available over time, the DBN can *infer user intent even when some inputs are noisy or incomplete.* The DBN models variables, their relationships, and temporal dependencies, enabling continuous tracking of context and more accurate, real-time decision-making while making the model *more transparent and explainable.* Additionally, the DBN's structure *supports the fusion of diverse data streams*, offering a robust mechanism for managing multimodal interaction.

We model relationships between inputs and outcomes using DBNs and extend the flexibility of our framework by incorporating a large language model (LLM) to introduce world knowledge and generalize beyond explicit DBN models. The LLM dynamically applies contextual and domain-specific reasoning, allowing the system to infer relationships and outcomes that the DBN may not explicitly model. This hybrid approach leverages the LLM's ability to process unstructured inputs like natural language and abstract concepts, enhancing the DBN's inference capabilities and enabling

more adaptive, intelligent interactions across varied, evolving contexts.

To demonstrate our computational framework, we model gaze and touch interactions on opportunistically available surfaces using our *Interaction Intent-driven Dynamic Bayesian Network* (IXDBN), as illustrated in Figure 1. Our system processes a stream of gaze data from an AR headset and touch input from a ring-based wearable device to infer which objects the user intends to interact with and how. In Figure 1c, the user looks at a light fixture and slides their finger upward to brighten the light. In this example, our DBN processes data from these two input streams, but it is designed to be easily extended to additional user input streams as needed.

We evaluated our implementation with 10 participants in an everyday office setting. Participants used gaze and surface-based touch gestures to interact with various Internet of Things (IoT) devices and AR elements. Results show that our system accurately infers user interaction intentions with low latency, even in challenging scenarios involving overlapping objects and closely placed items.

The main contributions of this work are:

(1) A hybrid DBN+LLM framework that integrates probabilistic modeling and world knowledge to enhance inference capabilities in multimodal, context-aware interactions.
(2) A system implementation to demonstrate the effectiveness of our approach, which processes streams of gaze and touch input, models the user's intent-to-action process from observations, and infers the interaction intentions on-the-go.
(3) A user evaluation demonstrating accurate and low-latency online inference of user intentions.

2 Related Work

We discuss related work on multimodal context-aware interactions, uncertainty modeling in human-computer interaction (HCI), the use of natural inputs for embedded interactions, and applications of dynamic Bayesian networks (DBNs).

2.1 Multimodal Context-Aware Interactions

Previous works on multimodal interactions explored a combined use of various input modalities, including gaze [11, 20, 40], gesture [11, 46], pen, and speech [7], utilizing information in multiple channels to enable rich and efficient interaction. Many approaches use separate modules to process individual input modalities into individual decision outputs, and later combine multiple decision outputs together to make an interaction decision. However, in such approaches, the combined decision heavily relies on individual decisions, making it prone to error stemming from uncertainty in individual modules.

Incorporating contextual information further allowed context-dependent interactions in various contexts, such as smart homes [36], design studios [1], vehicles [19], public spaces [65], and accessible spaces [67]. Authoring these interactions could be enabled through end-user authoring tools [18, 21, 41] for end users to author multimodal interactions themselves, providing flexibility in tailoring interactions to their everyday use. However, they require predefined interaction mappings, making this approach difficult to scale to many contexts. In this work, we address the reliability and scalability challenges of multimodal context-aware interaction systems by integrating both input modality diversity and contextual adaptation through a DBN-based framework that incorporates elicited knowledge from an LLM.

2.2 Probabilistic Modeling in HCI

HCI modeling aims to represent, explain, and reason about interactions [51]. Uncertainty often arises in interactions, making them hard to model with deterministic language. Uncertainty can occur in various phases of interacting with an interface. Schwarz synthesized that within an interaction, uncertainty may stem from sensor noise, input interpretation, or application actions [61]. Probabilistic modeling allows for reasoning under uncertainty and can be integrated with established models and theories.

Sensing input presents many uncertainties. Touch-based input is commonly used for today's flat-screen devices. Unfortunately, problems such as the "Fat Finger" problem and palm rejection often trigger false input sensing. Probabilistic modeling approaches, such as the utilization of Bayes' rule [5] and probabilistic inference using a particle filter [59], can mitigate uncertainty and improve input accuracy. In input interpretation, probabilistic modeling can extend classic HCI models such as Fitts' Law, allowing for its extensions to 2D pointing [24], pointing at targets of arbitrary shapes [25], and integration of prior knowledge [76]. At the application interface level, Bayesian Information Gain (BIG) has been introduced as a probabilistic framework that can generalize to various applications, such as navigation [43], file retrieval [44], and information exploration [64]. Besides providing reasoning under uncertainty, BIG also offer the benefit of efficient interactions.

Moving toward new computing interfaces, spatial computing interfaces bring additional uncertainties due to the complexity of spatial environments [12, 14] and ambiguity in natural input modalities [61]. We address these uncertain interactions in this new computing paradigm, with a DBN that can incorporate previous HCI knowledge, and propagate belief with continuously updated evidence.

2.3 Applications of Dynamic Bayesian Networks

Dynamic Bayesian Networks (DBNs) are probabilistic graphical models that are powerful tools for reasoning under uncertainty in problems involving the temporal dimension [49]. Applications of DBNs span diverse disciplines, such as medical diagnosis [10], gesture recognition [66], operation risk assessment [4], , and identification of gene networks [77]. This is largely due to the benefits of DBNs that can integrate dispersed knowledge from various sources, such as domain knowledge from experts [54], contextual knowledge [70], and data-driven knowledge [23]. In general, they are capable of information fusion [74], including fusing multimodal sensor readings [63]. At the same time, they are also inference engines, making them capable of inferring based on integrated knowledge.

To enable multimodal inputs in HCI, Pavlovic discussed ideas regarding the use of DBNs for information fusion for human-computer interfaces [55]. This approach has since been adapted for adjacent disciplines such as gesture recognition [66], tracking [30], social network modeling [58], and decision making [75].

We introduce a DBN-based computational framework for human-computer interactions that appear as stochastic processes with large amounts of uncertainty. Using a DBN, prior knowledge from previous HCI research and theories can be integrated, as well as causal relationships, contextual information, and other relevant factors. We further introduce integrating LLM-elicited knowledge for the DBN to adapt to various environments. We directly use this DBN as an inference engine, enabling real-time intent recognition in user interactions.

2.4 Using Natural Inputs for Embedded Interactions

In this paper, we showcase an implementation of our DBN-based computational framework with an example application in embedded interactions. Embedded interactions [37, 60] are digital interactions embedded within everyday physical environments. IoT devices, environmentally integrated AR information, and tangible interface could all form embedded interactions. By extending interactions beyond flat-screen devices into the physical world, users have opportunities to leverage opportunistically available physical affordances for input [28, 29]. Wearable devices can detect users' hand inputs and allow users to control IoT devices through gesture-based interactions [2, 31, 62]. In addition to hand input, gaze is a natural input modality that effectively communicates interest and attention [3, 32, 42, 56]. In embedded interaction scenarios, gaze has been used for target selection [48], interaction with public displays [34], and direct control of ambient devices [68]. Rather than relying on a single natural input modality for embedded interactions, multimodal inputs can be combined [36] to enhance interaction efficiency.

Although using natural inputs for embedded interactions can enable intuitive user experiences, they also introduce uncertainties to the interaction system. These can arise during sensing, input interpretation level, and application action processes [61]. Current embedded interactions utilizing natural inputs are mostly deterministic implementations, failing to account for these uncertainties. Moreover, as the number and variability of environments and embedded interactions grow, scalability poses a challenge in embedded interactions [36].

To address this, we introduce a probabilistic modeling approach to this interaction scenario, accounting for emerging uncertainties. Our probabilistic model further integrates LLM's world knowledge, to enable scalability.

3 A Dynamic Bayesian Network-Based Framework For Multimodal Context-Aware Interactions

We introduce a DBN-based computational framework for multimodal context-aware interactions (Figure 2). Our approach combines (1) the DBN's ability to integrate knowledge and dynamically perform probabilistic reasoning for inference with (2) the LLM's world knowledge to fill in knowledge gaps in the DBN, for scalability to various environments.

3.1 Walkthrough: Interacting with a Multimodal Context-Aware System Based on Our Framework

A walkthrough of a multimodal context-aware system based on our computational framework is illustrated in Figure 4.

A user's smart glasses and wristband are integrated with a multimodal context-aware interaction system built on our computational framework. The system runs on a version that supports gaze and touch input.

When a user enters a new environment, their wearable device(s) (e.g., smart glasses) will automatically connect to embedded interactions in the environment (e.g., IoT devices, AR elements), and scan the room to acquire their positions. The system's prompting engine uses these embedded interactions to prompt an LLM to fill in missing knowledge specific for this context, storing it in the DBN (Figure 4a). As the user interacts in this environment, the DBN infers the user's interaction intention by dynamically updating probability distributions based on evidences collected (Figure 4b).

3.2 Problem Formulation: Improvised Interactions as a Stochastic Process

Our goal is to enable users to interact effortlessly *without predefined interaction mappings* in *various contexts* using *multimodal inputs* such as gaze and touch. This scenario can be conceptualized as users improvising embedded interactions [13], enabled by the system inferring their intentions in real time. This presents a complex, stochastic problem domain due to the many uncertainties and temporal dynamics involved.

A stochastic process describes phenomena that evolve over time and involve uncertainties [16]. Our interaction problem can be understood as such a process, characterized by a sequence of random variables $\{X_t^1, X_t^2, ..., X_t^n\}$ where each variable X_t^i represents a component of the interaction–either observable or latent–at discrete time steps indexed by t. For example, a latent random variable could be a user's interaction intent (e.g., brightening the light), and an observable random variable could be a user's gaze position or hand movement. The outcome of each random variable is uncertain. Its possible values are defined over a probability space. The domain of each variable, therefore, includes all potential outcomes it can assume, with probabilities assigned to each possible outcome.

3.3 Bayesian Inference: Reasoning for Interacting Under Uncertainty

Bayesian inference provides a robust framework to reason under uncertainty. It integrates prior knowledge and updates beliefs as new evidence is observed. Bayesian updating can be described by:

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

Bayesian inference is effective for interaction problems with uncertainty [71]. For our interaction problem, which we have formulated as a stochastic process, Bayesian inference allows the system to dynamically update its beliefs about latent variables based on observable evidence from sensors [51].

Consider a latent variable L_t representing the user's interest at time step t. The domain of L_t is the set O, which encompasses all

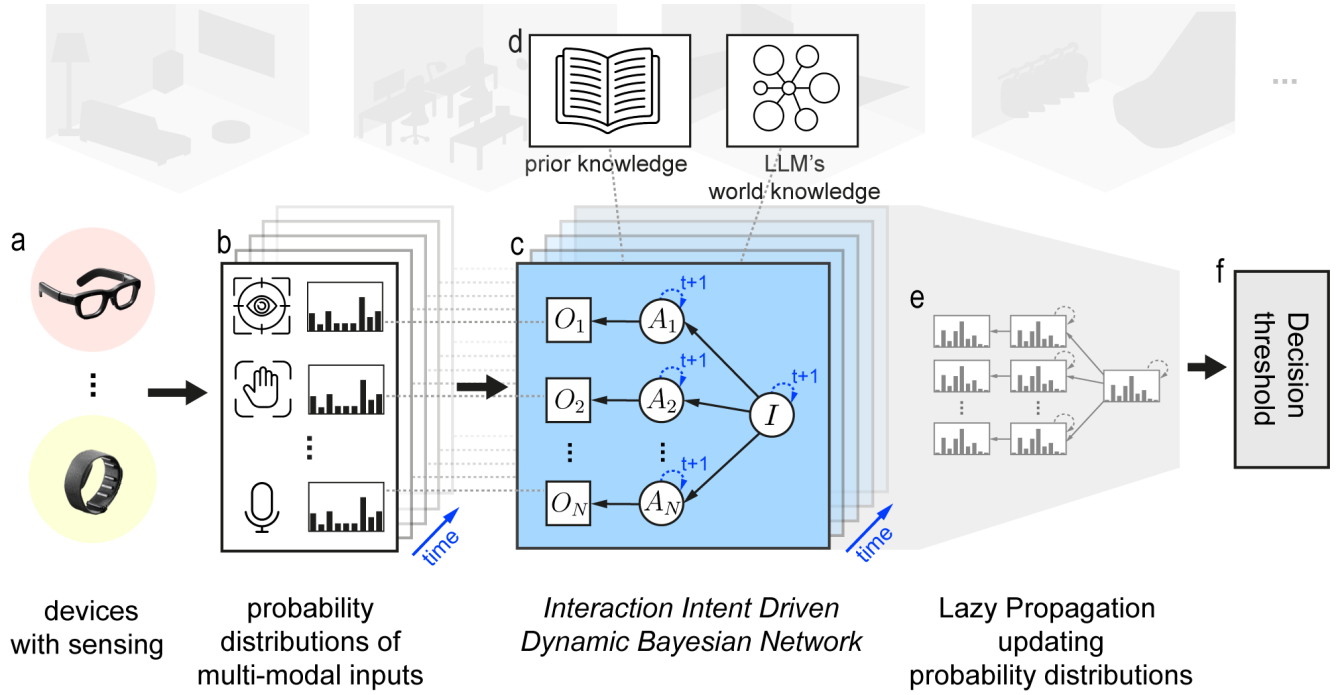

Figure 2: Overview of our dynamic Bayesian network (DBN)-based computational framework for multimodal context-aware interactions. (a) Devices with multimodal sensing, e.g., gaze and gesture sensors, collect data. (b) Each multimodal input makes predictions represented as probability distributions over time. (c) The interaction intent-driven dynamic Bayesian network (Figure 3) models relationships between observations (O), actions (A), and inferred interaction intent (I), integrating world knowledge from the LLM and prior knowledge at each time step. (d) Prior and world knowledge are used to enable scalability to diverse environments. (e) Lazy propagation updates probability distributions over time, (f) resulting in decision-making based on a decision threshold.

objects in the environment that offer embedded interactions. Initially, at time step $t = 0$, user interest L_0 is assumed to be uniformly distributed across all objects in O. As the system begins to observe user behavior at time step $t = 1$, specifically where the user's gaze G_1 is directed, it collects this evidence to refine its beliefs. For example, if the user's gaze at time $t = 1$ focuses on object O_1, the likelihood $P(G_1|L_1 = O_1)$ that O_1 is of particular interest increases. This observation updates the posterior probability $P(L_1 = O_1|G_1)$, proportionally to the product of the likelihood of observing G_1 given $L_1 = O_1$ and the initial uniform belief. This posterior at $t = 1$ becomes the new prior for the next observation at $t = 2$. If, at $t = 2$, the system again observes the user's gaze dwelling on O_1, it will further increase the probability for O_1 being the point of interest. Conversely, if the gaze shifts to another object O_2 at $t = 2$, the evidence from this time step will modify the posterior to reflect increased interest in O_2, while reducing interest for O_1. This iterative process allows the system to continuously collect evidence and update its belief with refined understanding and dynamic adjustments over time.

3.4 Dynamic Bayesian Network: Inference with Integrated Knowledge

Bayesian networks (BN) are shown to be effective structured knowledge representations and inference engines [45, 49, 63]. They encode prior beliefs about elements in a system and their relationships, and allow integration of different kinds of knowledge from various sources [9, 57], such as domain knowledge [54], data-driven knowledge, and causal relationships. DBNs extend BNs to the temporal domain, and are used for modeling stochastic processes, such as our interaction problem, as formulated above in the previous Section 3.2. Similar to BNs, DBNs allows integrating dispersed knowledge in one graphical model.

The integrated knowledge in a DBN can be conceptualized as priors encoded in the model. Some of these priors may be invariant across time steps, representing stable knowledge or assumptions about the interaction system. Others may be updated as new evidence is collected, reflecting the dynamic nature of the interaction process.

Building upon the general concept of Bayesian updating described in Section 3.3, we use a DBN as both a structured integration of knowledge and an inference engine during interaction. This dual role allows us to: (1) encode prior knowledge about the interaction

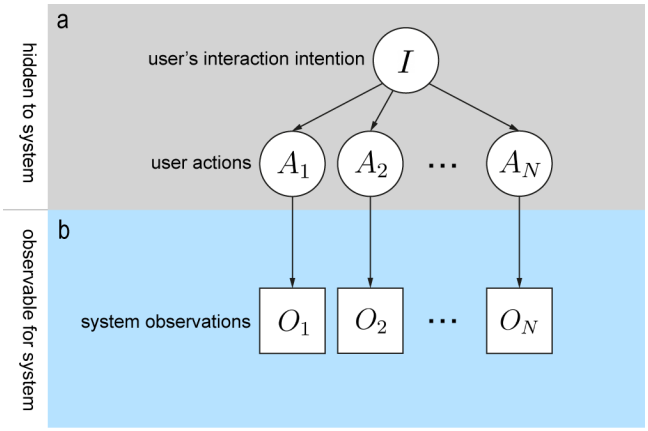

Figure 3: Hierarchical structure of the interaction intent-driven dynamic Bayesian network (IXDBN). (a) The DBN models user interaction intentions (I), which are inferred from user actions (A_1, A_2, ..., A_n). (b) These actions are linked to system observations (O_1, O_2, ..., O_n), which are derived from multimodal sensing inputs. The upper layer (a) represents the hidden aspects of the system (user intentions and actions), while the lower layer represents observable data that informs inference.

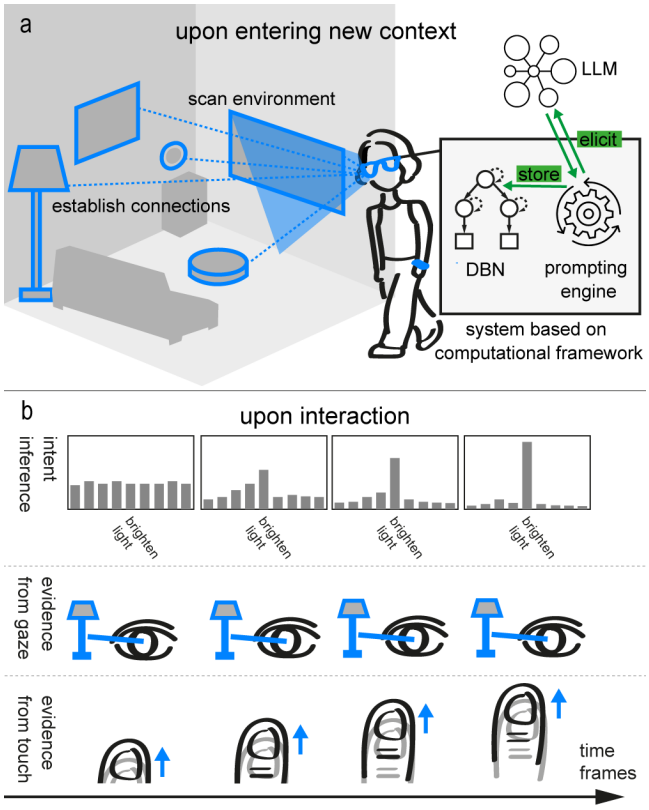

Figure 4: A walkthrough of our framework for interacting with a multimodal context-aware system. (a) Upon entering a new context, the system scans the environment, establishes connections with nearby devices, and integrates information using the computational framework that includes the DBN and LLM. The LLM supports context-specific reasoning, while the DBN manages inference and decision making. (b) During interaction, evidences from gaze and touch are collected at every time step, and the framework dynamically updates intent inference based on these evidences across time frames.

system, including user behavior, environmental factors, and device capabilities; (2) continuously collect evidence during the interaction process; and (3) update beliefs regarding key hypotheses, such as the user's interaction intentions, in real time.

We chose DBNs over other temporal probability models, such as Hidden Markov Models (HMMs) and Partially Observable Markov Decision Processes (POMDPs), due to their advantages in user modeling. DBNs provide modularity, a natural way to encode causality, and an effective means of integrating prior knowledge with data, making them ideal for our interaction framework. Compared to POMDPs, DBNs offer a more computationally efficient way to model uncertainty without requiring extensive policy optimization, making them more suitable for real-time inference in interactive systems. For a detailed comparison, we refer readers to Oliver et al. [50].

We describe our high-level approach to designing our DBN below and provide implementation details in Section 4.

3.4.1 User Intention-Action Model. The stochastic process in our interaction scenario is fundamentally driven by user intentions. We adopt the causal theory of action [6, 15], which posits that a user only performs an action if they have an internal state of intention leading to that action. From our system's perspective, both the user's intention and actions are latent variables, as they cannot be directly observed through sensor readings.

Building on intention recognition models for intelligent agents [27], we introduce an additional causal relationship: user actions result in observable sensor readings. This intention-to-action model governs the overall structure of our DBN design (Figure 3), comprising three levels: the user's intention (latent), the user's actions as caused by their intention (latent), and sensor readings resulting from their actions (observable). This structure allows our system to infer user intentions from observable data, while accounting for the causal relationship and inherent uncertainty in the process.

3.4.2 Affordance Theory. We incorporate the theory of affordance into our DBN to reason about interactions and inputs in digital environments. Affordances are perceivable actionable possibilities determined by both object properties in a given environment and an individual's action capabilities [22]. In our interaction scenario, environmental objects (e.g., IoT devices, embedded AR interfaces) afford interaction possibilities, while users can act on these possibilities using natural input modalities enabled by our sensing and inference system.

Our DBN design reflects this relationship probabilistically. At each discrete time step, the probability of a user intending an interaction L with an object O is governed by whether O affords

L:

$$P(L \mid O) = \begin{cases} P_{\text{low}}, & \text{if Afford}(O, L) = 0 \\ P_{\text{high}}, & \text{if Afford}(O, L) = 1 \end{cases}$$

This probabilistic formulation allows our DBN to infer likely user intentions based on the affordances present in the environment, guiding the system towards more accurate interpretations of user actions.

3.4.3 Multimodal Fusion. The DBN fuses multimodal sensing data together [26, 38]. Instead of sequentially making decisions based on separate rule-based decisions (e.g., if gaze lingers on the light beyond a threshold, and a hand touches a surface and slides up...), the DBN allows for joint reasoning across multiple sensor inputs together. For example, at every time step, gaze input and hand input are treated as probability distributions, which are jointly considered to infer other hidden states).

For this, we model nodes at the observation level to represent sensor inputs from various sensors, while nodes at the action level, which connect to them, correspond to the names of human actions as communicated by sensors. Our DBN can horizontally expand to integrate additional sensing capabilities as they become available.

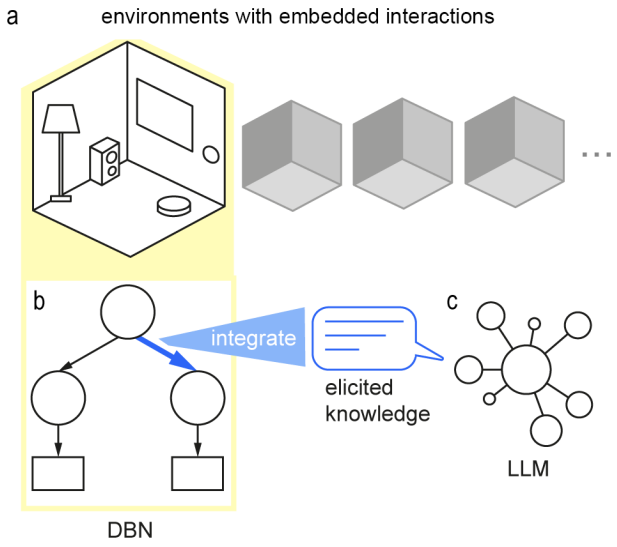

Figure 5: Integration of embedded interactions using DBN and LLM. (a) To adapt to diverse environments that feature embedded interactions with different sets of devices and actions, (b) the DBN integrates data from multimodal devices to model user interactions and infer intent, based on input from (c) the large language model (LLM) that provides elicited knowledge which supports contextual reasoning. Integrating the DBN and LLM enhances the system's inference capabilities by adding relevant world knowledge across diverse environments.

3.4.4 Domain Knowledge from an LLM. Traditionally, DBN construction relies on domain knowledge elicited from human experts. However, to make our system adaptable to various environments, we propose incorporating an LLM's world knowledge into the DBN (Figure 5).

For each new environment a user enters, we prompt an LLM to act as an interaction design expert. The LLM generates a confidence matrix that maps embedded interactions in that environment to potential user hand inputs. We incorporate this dynamically generated knowledge into our DBN through an "adaptable edge" that updates its value based on the current environment. This approach offers several advantages. It provides *scalability*, allowing the system to adapt to new environments without requiring manual expert input for each scenario. It also offers *flexibility*, as the LLM can generate nuanced mappings that account for the specific characteristics of each environment. By integrating LLM-derived knowledge, our DBN can make *more informed inferences* about user intentions across a wide range of environments and interaction scenarios.

Moreover, we design our framework such that the LLM is *only* prompted upon *entering unseen environments*. The retrieved knowledge is then *stored* in the DBN, which alone runs as the inference engine for interactions in the environment. This avoids prompting latency that occurs in end-to-end LLM systems, where the LLM would otherwise be prompted *per potential interaction* during interactions, rather than *once per unseen environment* before interactions in our framework.

4 System Implementation

To demonstrate our framework, we implement a system for multimodal interactions with gaze and touch.

We detail our implementation of the DBN, mathematical formulations of variables and their relationships, inference algorithm, and LLM prompting methods. The problem formulation for our interaction scenario and the modeling approach for our framework, which guide this implementation, are described in the previous Section 3. We use the PyAgrum package [17] for our implementation. .

4.1 Background: Dynamic Bayesian Networks

A DBN is a probabilistic graphical model used to represent temporal sequences of random variables. It models the dependencies between hidden states and observations across time. At each time step t, the system includes a set of hidden variables X_t and observable variables O_t. The hidden state at time t is influenced by the previous state, as defined by the transition model $P(X_t \mid X_{t-1})$, while the observation model $P(O_t \mid X_t)$ captures the relationship between the hidden state and the observed data. The DBN relies on the first-order Markov assumption, which means that the state at time t depends only on the state at time $t-1$. The joint probability of a sequence from time $t = 1$ to T is given by:

$$P(X_1, X_2, \ldots, X_T, O_1, O_2, \ldots, O_T) = P(X_1) \prod_{t=2}^{T} P(X_t \mid X_{t-1}) \prod_{t=1}^{T} P(O_t \mid X_t)$$

This formulation allows the DBN to infer and update its inference of hidden states from sequences of observations, making it suitable for time-dependent inference tasks, such as our interaction scenario.

4.2 DBN Structure Design

4.2.1 Observation Model. Our observation model is structured to represent the causal relationships in recognition of a user's intention-to-action process. We follow a tri-level intending agent behavior model [27] as seen in Figure 6. We horizontally expand the action and observation layers to account for both gaze and hand inputs. Our observation model spans from the intention node I on the first intention level with edges connecting it to the action nodes GA (gaze action node) and HA (hand action node).

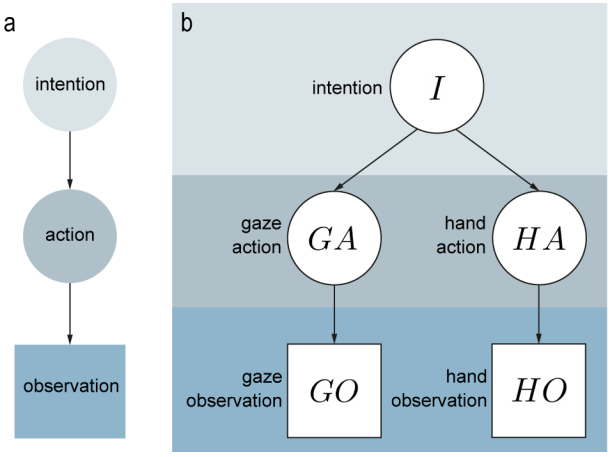

Figure 6: We structure our observation model based on (a) a tri-level behavior model for an intending agent, (b) horizontally expanding it to account for actions and observations from both gaze and hand.

GA and HA are then connected to GO (gaze observation node) and HO (hand observation node), respectively. This represents that users' acting with their gaze and hands produce observable states that can be directly observed by the system. Mathematically, this can be described as:

$$P(I, GA, HA, GO, HO) = P(I) \cdot P(GA \mid I) \cdot P(HA \mid I) \cdot P(GO \mid GA) \cdot P(HO \mid HA)$$

This structure encodes the conditional dependencies among these variables, making our tri-level observation model both efficient and generalizable.

4.2.2 Transition Model. To account for how the model should dynamically propagate probability beliefs across time, we design a transition model that encodes the evolution of hidden states across time slices. This is done by assuming a first-order Markov assumption, where the state at time t depends only on the state at time $t-1$.

As shown in Fig. 7, we model temporal transitions by establishing conditional dependencies between each time slice t and its predecessor $t-1$ for all three hidden variables I_t, GA_t, and HA_t, while preserving the observation model's dependencies.

$$P(I_t, GA_t, HA_t \mid I_{t-1}, GA_{t-1}, HA_{t-1}) = \begin{cases} P(I_t \mid I_{t-1}) \\ P(GA_t \mid GA_{t-1}, I_t) \\ P(HA_t \mid HA_{t-1}, I_t) \end{cases}$$

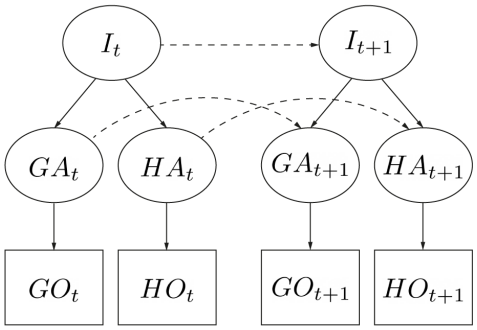

Figure 7: The transition model defines how hidden states in the current time slice are influenced by those in the previous time slice.

4.3 Random Variables and Their Conditional Probability Distributions

Each node in our DBN is a discrete random variable. It can take a set of distinct values, known as its domain, with each value having a specific probability:

$$P(X = x_i) = p_i, \quad \text{where} \quad \sum_i P(X = x_i) = 1$$

Table 1 provides an overview of all the random variables and their domains. Although pairs HA_t and HO_t, as well as GA_t and GO_t, both have the same domains, they are distinct variables. Action variables represent how the user acts with their gaze and hand, as caused by having an intention to interact. Action variables are latent to the inference system. Observation variables represent sensor readings that the inference system can directly access. They are governed by conditional probability distributions and do not assume the same values.

Edges in a Dynamic Bayesian Network represent conditional dependencies between nodes, defined by Conditional Probability Distributions (CPDs). A CPD $P(X \mid \text{Parents}(X))$ specifies how the value of a node depends on its parent nodes, capturing the probabilistic relationships between variables. In the following, we detail each random variable and CPD for the reproducibility of this work.

4.3.1 Intention Prior. Intention I is positioned at the top level of the network in each time slice. At the initial time step $t = 0$, the prior distribution for I_0 is drawn from a Dirichlet distribution:

$$P(I_0) \sim \text{Dirichlet}(\alpha)$$

where the concentration parameter α is evenly distributed across all possible actions, and weight controls the concentration of the distribution:

$$\alpha = \text{weight} \cdot \mathbf{1}_k$$

This Dirichlet prior serves as a conjugate prior to the multinomial distribution, which governs I_t. The sampled values introduce

Table 1: Overview of random variables in our DBN and their domains.

Random Variable	Description	Domain	Domain Description
Latent Intention and Action Parameters			
I_t	Interaction intention at time t	$\mathcal{E} = \{E_1, E_2, \ldots, E_k\}$	Embedded interactions in environment (e.g., turn off light, turn up volume)
GA_t	Gaze targeting action at time t	$\mathcal{O} = \{O_1, O_2, \ldots, O_n\}$	Objects with interaction affordances in environment (e.g., IoT lamp, IoT speaker)
HA_t	Hand input action at time t	$\mathcal{A} = \{A_1, A_2, \ldots, A_m\}$	Hand inputs (e.g., 0D touch, slide right, slide left...)
Observation Parameters			
GO_t	Gaze position at time t	$\mathcal{O} = \{O_1, O_2, \ldots, O_n\}$	Objects with interaction affordances in environment (e.g., IoT lamp, IoT speaker)
HO_t	Hand observation at time t	$\mathcal{A} = \{A_1, A_2, \ldots, A_m\}$	Hand inputs (e.g., 0D touch, slide right, slide left...)

variability in the initial conditions, allowing the model to incorporate uncertainty and reflect the diversity of possible outcomes.

4.3.2 Intention at Time Slice t. After the initial time slice, the conditional probability distribution (CPD) for the intention variable I_t, given the intention I_{t-1}, is modeled as follows:

The Dirichlet parameters for I_t, denoted as α_{It}, are organized into a matrix of shape $k \times k$, where k is the number of embedded interactions. Initially, each element in α_{It} is set to a base value (e.g., 3) to ensure a moderate degree of variability across transitions. The diagonal elements of α_{It} are then set to a higher value $w_{\text{given } I_0}$ (e.g., 10) to bias the distribution towards maintaining the same intention across consecutive time slices.

The CPD for I_t is sampled from a Dirichlet distribution for each row i of the matrix, reflecting the probability distribution over possible intentions I_t given $I_{t-1} = i$:

$$P(I_t \mid I_{t-1} = i) \sim \text{Dirichlet}(\alpha_{It}[i, :])$$

This introduces a higher probability for intentions to remain consistent across time slices, while still allowing for transitions to other intentions.

4.3.3 Gaze Action Prior. At the initial time step $t = 0$, the prior distribution for Gaze Action GA_0 is drawn from a Dirichlet distribution. The parameter vector α_{GA0} is of length $m + 1$, where m is the number of objects affording interactions, and the extra component represents a non-action or null state. Each element of α_{GA0} is set to a uniform base value controlled by the `weight` parameter. The prior is sampled to model uncertainty:

$$P(GA_0) \sim \text{Dirichlet}(\alpha_{GA0})$$

where:

$$\alpha_{GA0} = \texttt{weight} \cdot \mathbf{1}_{m+1}$$

4.3.4 Gaze Action at Time Slice t. At any time slice $t > 0$, the CPD for GA_t is modeled using a Dirichlet distribution that incorporates dependencies on both the previous gaze action GA_{t-1} and the current intention I_t. The conditional probability is expressed as:

$$P(GA_t \mid GA_{t-1}, I_t) \sim \text{Dirichlet}(\alpha_{GAt\_GA_{t-1}\_It})$$

where the parameter matrix $\alpha_{GAt\_GA_{t-1}\_It}$ is influenced by both GA_{t-1} and I_t.

Since I_t's domain is all the embedded interactions in the environment $\mathcal{E}$, and GA_t's domain is all the objects with interaction affordances $\mathcal{O}$, we express these domains as:

$$I_t \in \mathcal{E} = \{E_1, E_2, \ldots, E_m\}, \quad GA_t \in \mathcal{O} = \{O_1, O_2, \ldots, O_n\}$$

We draw on affordance theory to motivate the prior:

$$P(E_i \mid O_j) = \begin{cases} P_{\text{low}}, & \text{if Afford}(O_j, E_i) = 0 \\ P_{\text{high}}, & \text{if Afford}(O_j, E_i) = 1 \end{cases}$$

We utilize affordance relationships between embedded interactions and objects affording interactions to prioritize gaze actions. Specifically, when $I_t = E_i$ (where E_i is an embedded interaction afforded by an object O_j), the Dirichlet parameters are adjusted to increase the likelihood that $GA_t = O_j$:

$$\alpha_{GAt\_GA_{t-1}\_It}[i, j] \propto w_{\text{given } I_t} \cdot \text{Affords}(O_j, E_i)$$

To promote consistency in gaze actions over time, a higher weight $w_{\text{given } GA_{t-1}}$ is applied to the probability of continuing the previous gaze action GA_{t-1} into the current time step. This creates a bias toward maintaining the same gaze action unless influenced by other factors. The model combines the effects of both the previous gaze

action GA_{t-1} and the current intention I_t by adjusting the Dirichlet parameters accordingly. This approach balances gaze consistency with responsiveness to current intentions, allowing the model to predict gaze actions that are both consistent with past behaviors and adaptive to the user's current intentions.

4.3.5 Hand Action Prior. At the initial time slice $t = 0$, the prior distribution for hand action HA_0 is drawn from a Dirichlet distribution. The parameter vector α_{HA0} corresponds to the number of possible gestures, each representing a different hand action. Each element of α_{HA0} is set to a uniform value controlled by the `weight` parameter, modeling uncertainty:

$$P(HA_0) \sim \text{Dirichlet}(\alpha_{HA0})$$

where:

$$\alpha_{HA0} = \texttt{weight} \cdot \mathbf{1}_{\text{gestures}}$$

4.3.6 Hand Action at Time Slice t. At any time slice $t > 0$, the CPD for HA_t is modeled using a Dirichlet distribution that depends on both the previous hand action HA_{t-1} and the current intention I_t:

$$P(HA_t \mid HA_{t-1}, I_t) \sim \text{Dirichlet}(\alpha_{HAt\_HA_{t-1}\_It})$$

where the parameter matrix $\alpha_{HAt\_HA_{t-1}\_It}$ is shaped by both HA_{t-1} and I_t.

Given that I_t spans all embedded interactions in the environment $\mathcal{E}$, and HA_t spans all hand inputs $\mathcal{A}$:

$$I_t \in \mathcal{E} = \{E_1, E_2, \ldots, E_m\}, \quad HA_t \in \mathcal{A} = \{A_1, A_2, \ldots, A_n\}$$

We construct the Dirichlet parameters to prioritize hand inputs that align with interaction intentions, using an expert elicitation matrix L elicited from an LLM . Specifically, when $I_t = E_i$ (where E_i is associated with hand input A_j), the Dirichlet parameters are adjusted to increase the likelihood that $HA_t = A_j$:

$$\alpha_{HAt\_HA_{t-1}\_It}[i, j] \propto w_{\text{given } I_t} \cdot L[i, j]$$

where $L[i, j]$ indicates the relevance of hand input A_j for embedded interaction E_i.

To maintain consistency in hand actions over time, a higher weight $w_{\text{given } HA_{t-1}}$ is applied to the probability of continuing the previous hand action HA_{t-1} into the current time slice, favoring continuity unless other factors intervene. By combining the effects of both the previous hand action HA_{t-1} and the current intention I_t, the model achieves a balance between maintaining hand action consistency and adapting to the user's current intentions.

4.3.7 Gaze Observation (GO) and Hand Observation (HO). Gaze observation GO_t and hand observation HO_t serve as observable evidence within the model, reflecting the user's gaze and hand actions at any time slice $t > 0$. Unlike the intention variables, GO_t and HO_t do not have priors at $t = 0$; instead, they are conditionally dependent on the corresponding gaze action GA_t and hand action HA_t, respectively.

The CPD for GO_t is modeled as:

$$P(GO_t \mid GA_t) \sim \text{Dirichlet}(\alpha_{GO})$$

where the Dirichlet parameter matrix α_{GO} is designed to align GO_t closely with GA_t, applying a higher weight $w_{\text{GA\_GO\_match}}$ when the gaze observation matches the gaze action.

Similarly, the CPD for HO_t is modeled as:

$$P(HO_t \mid HA_t) \sim \text{Dirichlet}(\alpha_{HO})$$

where the Dirichlet parameter matrix α_{HO} applies a higher weight $w_{\text{HA\_HO\_match}}$ when the hand observation matches the hand action.

These CPDs ensure that the observable evidence GO_t and HO_t reflect the most likely gaze and hand observations given the corresponding actions, thereby reinforcing the alignment between observed and intended actions.

4.4 Inference

We use Lazy Propagation [47] as our inference algorithm, which provides a way of fast inference updating the entire body of random variables in the DBN.

4.5 LLM Elicitation

To provide necessary knowledge to the DBN in different contexts, we elicit an LLM for this knowledge and integrate it into the DBN's CPD(s). For the multimodal interaction system with gaze and touch input, we require prior conditional probabilities mapping embedded interactions in different contexts to system-supported touch inputs.

This elicitation happens upon a user entering an unseen environment. The elicited information is processed and stored in the DBN for the user to conveniently interact in this environment. No other reconfiguration or training is needed.

We implement our prompting engine with a Python interface connecting to OpenAI API. We prompt gpt-4o-mini and ask the model to adopt a persona of an interaction designer, as human expert knowledge elicitation is an established way of constructing DBNs [72]. We provide an output structure[1] to the LLM to ensure that the output is a confidence matrix establishing an expert's prior belief on conditional probabilities mapping context embedded interactions and system-supported touch inputs. This output is then processed into a Dirichlet distribution to form the CPD for node HA_t.

Our elicited information consists of simple input-output likelihoods for IoT devices and AR interfaces, which are typically easy for human designers and users to agree upon. We did not experience any prompting inconsistencies in our experiments. See Appendix A for the prompts used.

5 Evaluation

To evaluate our DBN's performance during online interaction intention inference, we conducted a study with 10 participants (F = 4, M = 5, Non-binary = 1; Average age = 28.3). We simulate an everyday office scenario with embedded interactions and implement natural input modalities of gaze and surface-based hand input. Participants were asked to act on embedded interactions using natural inputs in ways most intuitive to them. During this process, our DBN performs online inference and outputs inferred interactions.

[1] https://platform.openai.com/docs/guides/structured-outputs

5.1 Apparatus and Study Setup

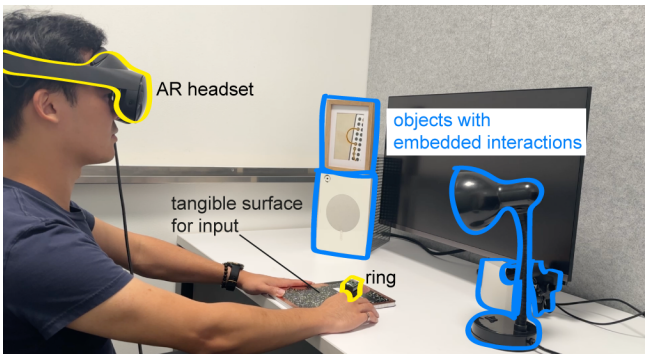

Figure 8: Study setup mimicking an office environment. Participants wore an AR headset and a ring (highlighted in yellow) that tracked gaze and hand positions and touch input on passive surfaces. Five physical objects (highlighted in blue) simulated objects that afford embedded interactions: an AR photo album, an IoT speaker, an IoT desk lamp, an IoT temperature monitor, and an IoT web camera, arranged from left to right.

We used a commercial AR headset (Meta Quest Pro) for tracking gaze and hand positions and a ring-based input device [35] for sensing touch on passive surfaces. For hand input, we simulated opportunistic tangible interactions that are generalizable to act on many embedded interactions. For this reason, we detected surface-based inputs on passive surfaces, and specifically implemented per-frame detection of a 0D press and 1D slides in the four cardinal directions [8]. We used Unity to implement tracking input and built a socket connection to run our DBN in real time with Python 3.11.7 for online inference.

The physical study environment was set up to mimic an office environment (Figure 8). We arranged five physical objects to simulate objects affording embedded interactions, including an IoT desk lamp, an AR photo album, an IoT web camera, an IoT speaker, and an IoT temperature monitor. To test our DBN's ability in mitigating uncertainty, we deliberately design the study setup to include a small object (IoT web camera), objects placed closely together (AR photo album and IoT speaker), and overlapping objects (IoT web camera and IoT temperature monitor). Each object had at least two interaction affordances, with a total of 12 embedded interactions in the study scenario.

We further set up a notebook which was taped with copper tape decorations on its cover. This was because the ring-based input device detected passive touch on conductive surfaces. We used this notebook to simulate opportunistically available surfaces that users could leverage in any environment for natural input.

5.2 Procedure

A study conductor first introduced participants to objects affording embedded interactions in their environment. Participants were asked to look at each object as it was introduced. The notebook was introduced as an opportunistically available surface for tangible input. Participants were onboarded with the five types of surface-based inputs implemented, including 0D touch and 1D slides in the four cardinal directions. They tried performing these five surface-based inputs using their index finger on the notebook's cover.

Before data collection began for each participant, the study conductor ran a script that generated a random order for performing the 12 embedded interactions five times. Participants were informed that they would be given an interaction intention and should use their gaze and tangible input on the notebook to perform the interaction. The study conductor then provided interaction intentions in a random order, as determined by the script. After reading an interaction intention, the study conductor logged it as ground truth by pressing a key on the keyboard for the duration of the interaction.

As participants performed natural inputs, the DBN observed their gaze and hand tracking data at each time step and dynamically updated its belief about their interaction intention. The system outputted a decision when the skewness value of the interaction intention variable exceeded a threshold.

An interaction was concluded once the system outputted a decision matches the ground truth or when the participant decided to move on to the next interaction. The study conductor then provided the participant the next interaction intention. On average, the data collection process took approximately 10 minutes per participant.

5.3 Logged Data

For each interaction, the study system logged data at every frame. The logged data include user inputs (gaze position, hand touch state, and hand movement), probability distributions for all system variables, ground truth interaction intention (as given by the study conductor), and system-inferred interaction intention. The logged data was used for further analysis. The study conductor also noted anomaly behaviors from participants during the study.

After the study, the logged data was cleaned to remove human errors that influence the analysis of the results. This included occasions of misinterpretations in which participants missed, misheard, or asked follow-up questions regarding the prompted interaction intention. We used a 2 s reaction time threshold to filter such data that added noise to our system evaluation. In these occasions, there exist relative large time discrepancies between a ground truth interaction intention being logged, and for it to actually become the participant's internal intention. Figures 9 and 10 show logged inference and ground truth data from two study participants.

5.4 Study Results

Online inference results (Figure 11) show that our system achieves overall high *per-frame* accuracy with 0.83 per-frame accuracy, 0.87 per-frame precision, 0.83 per-frame recall, and 0.83 per-frame F1 score on average across interaction intentions. Figure 12 shows accuracy metrics for each interaction intention: "Increase volume" and "Decrease volume" have low recall scores. Figure 11 shows that the two were often mistaken with "Play/pause". This is because "Increase volume," "Decrease volume," and "Play/pause" all belong to the same object, the IoT music player. Vertical sliding inputs, which participants commonly used for increasing and decreasing volume, require touching the input surface first, which is often taken as a "press" input for the first few frames. However, our system was

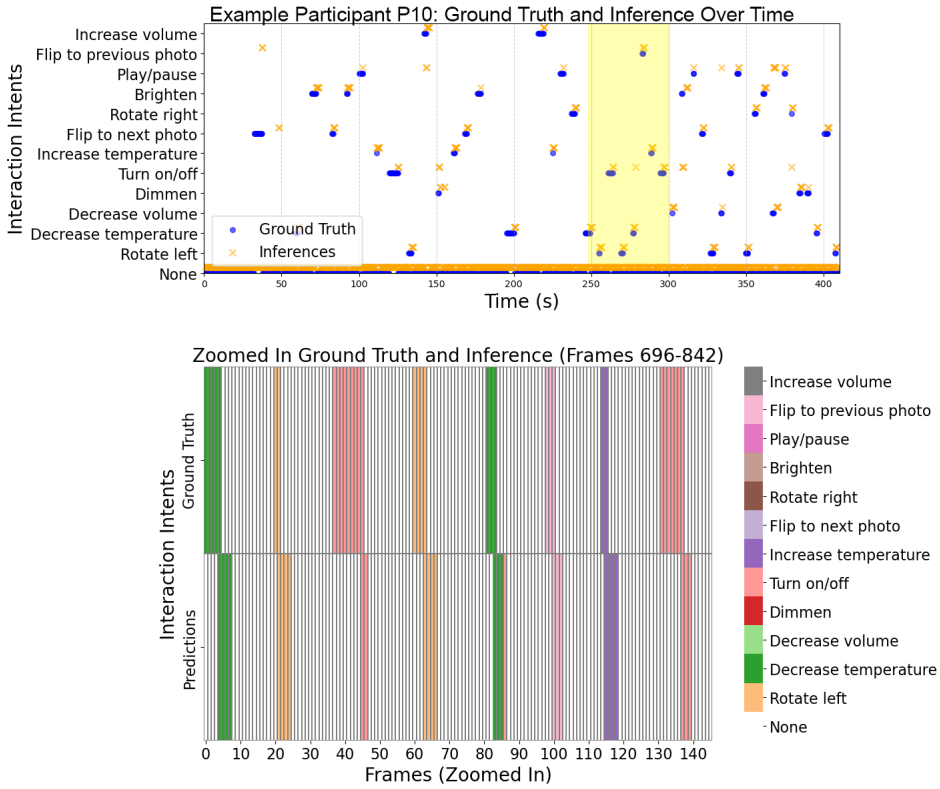

Figure 9: Logged interaction data for P10. The top figure plots the ground truth interaction intents (blue dots) and our system's inferences (orange crosses) over time. The bottom figure provides a zoomed-in view of a time segment (yellow-highlighted region) from the top figure, showing a frame-by-frame comparison between the ground truth interaction intents (top row) and the system's predicted intents (bottom row). The color-coded legend corresponds to different interaction intents. This illustrates that the system's inferences closely align with the ground truth and that the system quickly infers the correct interaction intent after the ground truth action occurs.

able to dynamically update and change its inference rather quickly. Figure 10 shows an example of this dynamic change around time 50 s.

5.5 Discussion

Discussion. The results show that our system employing IXDBN handles uncertanties in noisy multimodal inputs of gaze and ring-based gestures effectively to distinguish users' intent to interact, even in a complex setup where the embedded devices with similar affordances are placed in proximity to each other.

As shown in the confusion matrix in Figure 11, the current per-frame results involve misclassifications caused by fast decision made by the system (e.g., the first tap of a slide as a tapping gesture). While this highlights the fast inferencing capability of the system using our DBN-based approach, it degrades the per-frame classification accuracy for the embedded interaction. The classification accuracy using our model can be improved by improving the decision threshold for the final classification decision (Figure 2f) based on different applications.

We also observed that during the study, some participants demonstrated contrasting habits toward multimodal interaction. For example, for the interactions "Flip to next photo" and "Flip to previous photo" embedded in an AR photo album, some participants swiped in the opposite direction to the majority. Future work could address such individual preferences through personalization of the DBN's CPD to adapt to user preferences over time.

Limitations. While our initial evaluation demonstrated the promise of our approach, several limitations warrant discussion. Our current study is focused primarily on evaluating the real-time system performance with participants, but it lacks comparative analysis against alternative approaches such as end-to-end LLM prompting systems, data-driven models, or rule-based systems. Such comparisons would better illuminate the relative strengths and weaknesses of our approach. Additionally, our evaluation would benefit from greater contextual and demographic diversity—testing across various environments (e.g., living rooms, kitchens, museums) and with a more diverse participant pool (varying in age, physical capabilities, and interaction habits), providing more comprehensive insights into the system's consistency and adaptability (including potential biases in LLM-elicitation).

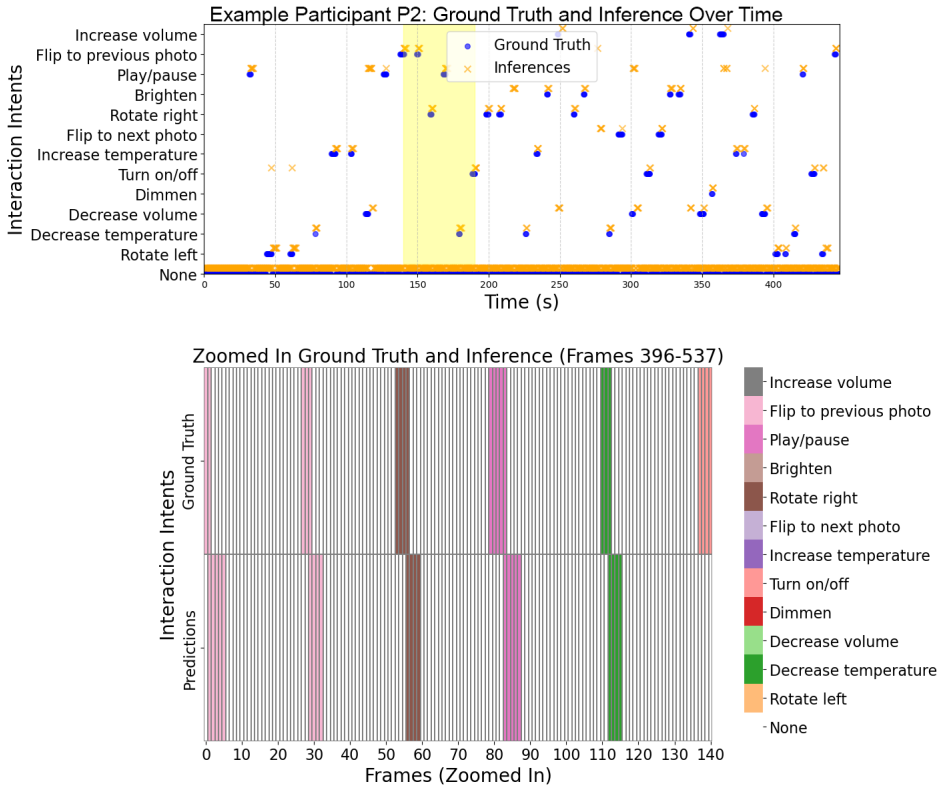

Figure 10: Logged interaction data for P2.

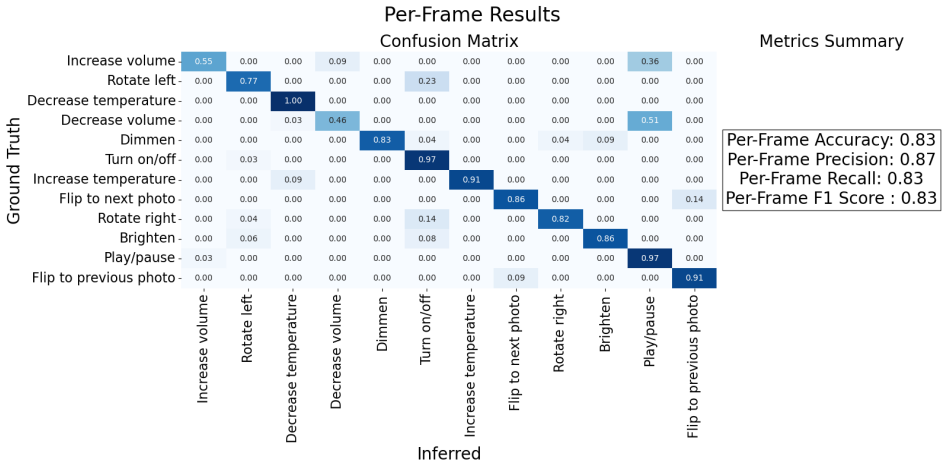

Figure 11: Confusion matrix of per-frame results. The results highlight occurrences of misclassifications where the first tapping gesture when sliding (e.g., increasing/decreasing volume) is misclassified into the tap gesture (e.g., turn on/off), due to the fast inference made by our system.

Future Work. Our framework opens several promising directions for future research, spanning both systems development and practical applications. From a systems perspective, longitudinal studies could explore how the DBN's parameters adapt to accumulated user data over time, potentially improving personalization and inference accuracy gradually. Additionally, the hybrid nature of our approach enables the integration of various pre-trained models into the DBN, expanding its capabilities across diverse use cases. From an applications standpoint, our multimodal context-aware interaction system

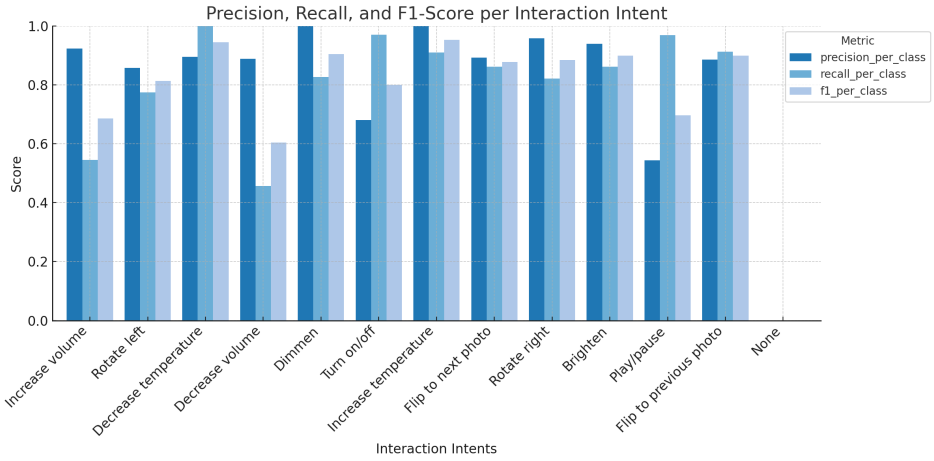

Figure 12: Accuracy metrics for individual interaction intentions shows overall good performance. Recall scores are low for "Increase Volume" and "Decrease Volume" due to misdetected touch input in the first few frames, which our system dynamically addressed in following frames.

shows significant potential for users with physical limitations, enabling more convenient and accessible environmental interactions across various contexts. However, further research is needed to assess the system's specific benefits and limitations for different user populations. Future studies could also explore how our framework adapts to emerging interaction contexts, such as smart homes, healthcare environments, and educational settings, where multimodal context-aware interactions could provide substantial value.

6 Conclusion

In this work, we proposed a novel dynamic Bayesian network-based computational framework for multimodal context-aware interactions. We implemented a system utilizing this framework that allows online interaction intent inference without pre-mappings of multimodal interactions using gaze and touch inputs. We performed an evaluation study to assess our system's performance.

The key concept of our approach is to use a DBN to structure and integrate various prior knowledge, while leveraging an LLM's world knowledge to fill in the unknowns in the DBN, for it to become an effective inference engine for interactions. This approach poses several benefits of: maintaining a small network while being scalable to various contexts; temporally aligning sensor inputs; and allowing for scalability to more sensor inputs. Our approach also accounts for the uncertainties that exist in interaction, while performing per-frame inference with low latency. Future works can adapt our core concept to various intent-driven interactions.

Our tri-level DBN is a generalizable structure for different intent-driven interactions. However, it has not yet understood the more expressive and less common inputs, such as rotating a coffee mug to play music volume. Explorations of using other Artificial Intelligence systems and external knowledge bases are needed for this understanding and integration into our DBN.

Furthermore, the explainability of our approach (outputting probability distributions for every system variable at every time step) offers opportunities to use this for interaction, such as feedforward and feedback visualizations [61]. Integrations of such mechanisms may account for the interaction context, as our current relatively simple multi-modal interactions are intuitive to control and happen very fast. Feedback and feedforward mechanisms may be added to more elaborate interactions.

Acknowledgments

We thank Yiming Zhang, Xingyu Bruce Liu, Romain Nith, Jacqui Fashimpaur, and Anna Yu for insightful discussions, and Jom Preechayasomboon, Sebastian Freitag, and David Moncada for device testing. The first author is grateful to Yujie Hui, Tim Lebailly, Wenxuan Guo, Yingsi Qin, Zixiong Su, Jingxuan Fan, Dorian Chan, Shwetha Rajaram, Luis Hernan Cubillos, and fellow interns for their support and encouragement throughout this work.

References

[1] Aaron Adler, Jacob Eisenstein, Michael Oltmans, Lisa Guttentag, and Randall Davis. 2004. Building the Design Studio of the Future.. In *AAAI Technical Report (6)*. 1–7.

[2] Amr Alanwar, Moustafa Alzantot, Bo-Jhang Ho, Paul Martin, and Mani Srivastava. 2017. SeleCon: Scalable IoT Device Selection and Control Using Hand Gestures. In *Proceedings of the Second International Conference on Internet-of-Things Design and Implementation* (Pittsburgh, PA, USA) *(IoTDI '17)*. Association for Computing Machinery, New York, NY, USA, 47–58. https://doi.org/10.1145/3054977.3054981

[3] Stelios Asteriadis, Paraskevi Tzouveli, Kostas Karpouzis, and Stefanos Kollias. 2007. Non-Verbal Feedback on User Interest Based on Gaze Direction and Head Pose. In *Second International Workshop on Semantic Media Adaptation and Personalization (SMAP 2007)*. 171–178. https://doi.org/10.1109/SMAP.2007.50

[4] Shubharthi Barua, Xiaodan Gao, Hans Pasman, and M Sam Mannan. 2016. Bayesian network based dynamic operational risk assessment. *Journal of Loss Prevention in the Process Industries* 41 (2016), 399–410.

[5] Xiaojun Bi and Shumin Zhai. 2013. Bayesian touch: a statistical criterion of target selection with finger touch. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. 51–60.

[6] John Christopher Bishop. 1989. *Natural agency: An essay on the causal theory of action*. Cambridge University Press.

[7] Richard A Bolt. 1980. "Put-that-there" Voice and gesture at the graphics interface. In *Proceedings of the 7th annual conference on Computer graphics and interactive techniques*. 262–270.

[8] William Buxton. 1983. Lexical and pragmatic considerations of input structures. *ACM SIGGRAPH Computer Graphics* 17, 1 (1983), 31–37.

[9] Andrés Cano, Andrés R. Masegosa, and Serafín Moral. 2011. A Method for Integrating Expert Knowledge When Learning Bayesian Networks From Data. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 41, 5 (2011), 1382–1394. https://doi.org/10.1109/TSMCB.2011.2148197

[10] Theodore Charitos, Linda C Van Der Gaag, Stefan Visscher, Karin AM Schurink, and Peter JF Lucas. 2009. A dynamic Bayesian network for diagnosing ventilator-associated pneumonia in ICU patients. *Expert Systems with Applications* 36, 2 (2009), 1249–1258.

[11] Ishan Chatterjee, Robert Xiao, and Chris Harrison. 2015. Gaze+Gesture: Expressive, Precise and Targeted Free-Space Interactions. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (Seattle, Washington, USA) *(ICMI '15)*. Association for Computing Machinery, New York, NY, USA, 131–138. https://doi.org/10.1145/2818346.2820752

[12] Jiajian Chen and Blair MacIntyre. 2008. Uncertainty Boundaries for Complex Objects in Augmented Reality. In *2008 IEEE Virtual Reality Conference*. 247–248. https://doi.org/10.1109/VR.2008.4480784

[13] Xiang 'Anthony' Chen and Yang Li. 2017. Improv: An Input Framework for Improvising Cross-Device Interaction by Demonstration. *ACM Trans. Comput.-Hum. Interact.* 24, 2, Article 15 (April 2017), 21 pages. https://doi.org/10.1145/3057862

[14] Enylton Machado Coelho, Blair MacIntyre, and Simon J. Julier. 2005. Supporting interaction in augmented reality in the presence of uncertain spatial knowledge. In *Proceedings of the 18th Annual ACM Symposium on User Interface Software and Technology* (Seattle, WA, USA) *(UIST '05)*. Association for Computing Machinery, New York, NY, USA, 111–114. https://doi.org/10.1145/1095034.1095052

[15] Wayne A Davis. 2010. The causal theory of action. *A Companion to the Philosophy of Action* 1 (2010), 32–39.

[16] Joseph L Doob. 1942. What is a stochastic process? *The American Mathematical Monthly* 49, 10 (1942), 648–653.

[17] Gaspard Ducamp, Christophe Gonzales, and Pierre-Henri Wuillemin. 2020. aGrUM/pyAgrum : a Toolbox to Build Models and Algorithms for Probabilistic Graphical Models in Python. In *10th International Conference on Probabilistic Graphical Models (Proceedings of Machine Learning Research, Vol. 138)*. Skørping, Denmark, 609–612. https://hal.archives-ouvertes.fr/hal-03135721

[18] Martina Eckert, Marcos López, Carlos Lázaro, and Juan Meneses. 2019. MoKey: a versatile exergame creator for everyday usage. *Assistive Technology* (2019).

[19] Philipp Fischer and Andreas Nurnberger. 2008. Adaptive and multimodal interaction in the vehicle. In *2008 IEEE International Conference on Systems, Man and Cybernetics*. IEEE, 1512–1516.

[20] Tan Gemicioglu, R. Michael Winters, Yu-Te Wang, Thomas M. Gable, Ann Paradiso, and Ivan J. Tashev. 2023. Gaze & Tongue: A Subtle, Hands-Free Interaction for Head-Worn Devices. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 456, 4 pages. https://doi.org/10.1145/3544549.3583930

[21] Giuseppe Ghiani, Marco Manca, Fabio Paternò, and Carmen Santoro. 2017. Personalization of context-dependent applications through trigger-action rules. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 2 (2017), 1–33.

[22] JJ Gibson. 1977. The theory of affordances. *Perceiving, acting and knowing: Towards an ecological psychology/Erlbaum* (1977).

[23] Alex Greenfield, Christoph Hafemeister, and Richard Bonneau. 2013. Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. *Bioinformatics* 29, 8 (2013), 1060–1067.

[24] Tovi Grossman and Ravin Balakrishnan. 2005. A probabilistic approach to modeling two-dimensional pointing. *ACM Trans. Comput.-Hum. Interact.* 12, 3 (Sept. 2005), 435–459. https://doi.org/10.1145/1096737.1096741

[25] Tovi Grossman, Nicholas Kong, and Ravin Balakrishnan. 2007. Modeling pointing at targets of arbitrary shapes. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 463–472.

[26] David L Hall and James Llinas. 1997. An introduction to multisensor data fusion. *Proc. IEEE* 85, 1 (1997), 6–23.

[27] Clint Heinze. 2004. Modelling intention recognition for intelligent agent systems. (2004).

[28] Steven Henderson and Steven Feiner. 2009. Opportunistic tangible user interfaces for augmented reality. *IEEE Transactions on Visualization and Computer Graphics* 16, 1 (2009), 4–16.

[29] Steven J. Henderson and Steven Feiner. 2008. Opportunistic controls: leveraging natural affordances as tangible user interfaces for augmented reality. In *Proceedings of the 2008 ACM Symposium on Virtual Reality Software and Technology* (Bordeaux, France) *(VRST '08)*. Association for Computing Machinery, New York, NY, USA, 211–218. https://doi.org/10.1145/1450579.1450625

[30] Gang Hua and Ying Wu. 2006. Measurement integration under inconsistency for robust tracking. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 1. IEEE, 650–657.

[31] Darshan Iyer, Fahim Mohammad, Yuan Guo, Ebrahim Al Safadi, Benjamin J. Smiley, Zhiqiang Liang, and Nilesh K. Jain. 2016. Generalized Hand Gesture Recognition for Wearable Devices in IoT: Application and Implementation Challenges. In *Machine Learning and Data Mining in Pattern Recognition*, Petra Perner (Ed.). Springer International Publishing, Cham, 346–355.

[32] Qiang Ji and Zhiwei Zhu. 2003. Non-intrusive eye and gaze tracking for natural human computer interaction. *MMI-Interaktiv* 6 (2003), 1439–7854.

[33] Ed Kaiser, Alex Olwal, David McGee, Hrvoje Benko, Andrea Corradini, Xiaoguang Li, Phil Cohen, and Steven Feiner. 2003. Mutual disambiguation of 3D multimodal interaction in augmented and virtual reality. In *Proceedings of the 5th international conference on Multimodal interfaces*. 12–19.

[34] Mohamed Khamis, Axel Hoesl, Alexander Klimczak, Martin Reiss, Florian Alt, and Andreas Bulling. 2017. EyeScout: Active Eye Tracking for Position and Movement Independent Gaze Interaction with Large Public Displays. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (Québec City, QC, Canada) *(UIST '17)*. Association for Computing Machinery, New York, NY, USA, 155–166. https://doi.org/10.1145/3126594.3126630

[35] Wolf Kienzle, Eric Whitmire, Chris Rittaler, and Hrvoje Benko. 2021. ElectroRing: Subtle Pinch and Touch Detection with a Ring. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 3, 12 pages. https://doi.org/10.1145/3411764.3445094

[36] Jung-Hwa Kim, Seung-June Choi, and Jin-Woo Jeong. 2019. Watch & Do: A smart iot interaction system with object detection and gaze estimation. *IEEE Transactions on Consumer Electronics* 65, 2 (2019), 195–204.

[37] Matthias Kranz, Paul Holleis, and Albrecht Schmidt. 2009. Embedded interaction: Interacting with the internet of things. *IEEE internet computing* 14, 2 (2009), 46–53.

[38] Dana Lahat, Tülay Adali, and Christian Jutten. 2015. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proc. IEEE* 103, 9 (2015), 1449–1477.

[39] Fabrizio Lamberti, Andrea Sanna, Gilles Carlevaris, and Claudio Demartini. 2015. Adding pluggable and personalized natural control capabilities to existing applications. *Sensors* 15, 2 (2015), 2832–2859.

[40] Jaewook Lee, Jun Wang, Elizabeth Brown, Liam Chu, Sebastian S. Rodriguez, and Jon E. Froehlich. 2024. GazePointAR: A Context-Aware Multimodal Voice Assistant for Pronoun Disambiguation in Wearable Augmented Reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 408, 20 pages. https://doi.org/10.1145/3613904.3642230

[41] Toby Jia-Jun Li, Amos Azaria, and Brad A Myers. 2017. SUGILITE: creating multimodal smartphone automation by demonstration. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 6038–6049.

[42] Yixuan Li, Pingmei Xu, Dmitry Lagun, and Vidhya Navalpakkam. 2017. Towards Measuring and Inferring User Interest from Gaze. In *Proceedings of the 26th International Conference on World Wide Web Companion* (Perth, Australia) *(WWW '17 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 525–533. https://doi.org/10.1145/3041021.3054182

[43] Wanyu Liu, Rafael Lucas D'Oliveira, Michel Beaudouin-Lafon, and Olivier Rioul. 2017. BIGnav: Bayesian Information Gain for Guiding Multiscale Navigation. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5869–5880. https://doi.org/10.1145/3025453.3025524

[44] Wanyu Liu, Olivier Rioul, Joanna Mcgrenere, Wendy E Mackay, and Michel Beaudouin-Lafon. 2018. BIGFile: Bayesian information gain for fast file retrieval. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[45] Jiebo Luo, Andreas E Savakis, and Amit Singhal. 2005. A Bayesian network-based framework for semantic image understanding. *Pattern recognition* 38, 6 (2005), 919–934.

[46] Mathias N. Lystbæk, Peter Rosenberg, Ken Pfeuffer, Jens Emil Grønbæk, and Hans Gellersen. 2022. Gaze-Hand Alignment: Combining Eye Gaze and Mid-Air Pointing for Interacting with Menus in Augmented Reality. *Proc. ACM Hum.-Comput. Interact.* 6, ETRA, Article 145 (May 2022), 18 pages. https://doi.org/10.1145/3530886

[47] Anders L Madsen and Finn V Jensen. 1999. Lazy propagation: a junction tree inference algorithm based on lazy evaluation. *Artificial Intelligence* 113, 1-2 (1999), 203–245.

[48] Simon Mayer and Gábor Sörös. 2014. User Interface Beaming – Seamless Interaction with Smart Things Using Personal Wearable Computers. In *2014 11th International Conference on Wearable and Implantable Body Sensor Networks Workshops*. 46–49. https://doi.org/10.1109/BSN.Workshops.2014.17

[49] V Mihajlovic and Milan Petkovic. 2001. Dynamic bayesian networks: A state of the art. (2001).

[50] Nuria Oliver and Eric Horvitz. 2005. A comparison of hmms and dynamic bayesian networks for recognizing office activities. In *International conference on user modeling*. Springer, 199–209.

[51] Antti Oulasvirta, Per Ola Kristensson, Xiaojun Bi, and Andrew Howes. 2018. *Computational interaction*. Oxford University Press.

[52] Sharon Oviatt. 2022. Multimodal interaction, interfaces, and analytics. In *Handbook of Human Computer Interaction*. Springer, 1–29.

[53] Sharon Oviatt, Björn Schuller, Philip Cohen, Daniel Sonntag, and Gerasimos Potamianos. 2017. *The handbook of multimodal-multisensor interfaces, volume 1: Foundations, user modeling, and common modality combinations*. Morgan & Claypool.
[54] Diane Oyen and Terran Lane. 2012. Leveraging domain knowledge in multitask Bayesian network structure learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 26. 1091–1097.
[55] Vladimir Ivan Pavlovic. 1999. *Dynamic Bayesian networks for information fusion with applications to human-computer interfaces*. University of Illinois at Urbana-Champaign.
[56] Christopher Peters, Catherine Pelachaud, Elisabetta Bevacqua, Maurizio Mancini, and Isabella Poggi. 2005. A model of attention and interest using gaze behavior. In *International Workshop on Intelligent Virtual Agents*. Springer, 229–240.
[57] K.W. Przytula and D. Thompson. 2000. Construction of Bayesian networks for diagnostics. In *2000 IEEE Aerospace Conference. Proceedings (Cat. No.00TH8484)*, Vol. 5. 193–200 vol.5. https://doi.org/10.1109/AERO.2000.878490
[58] Vasanthan Raghavan, Greg Ver Steeg, Aram Galstyan, and Alexander G Tartakovsky. 2014. Modeling temporal activity patterns in dynamic social networks. *IEEE Transactions on Computational Social Systems* 1, 1 (2014), 89–107.
[59] Simon Rogers, John Williamson, Craig Stewart, and Roderick Murray-Smith. 2011. AnglePose: robust, precise capacitive touch tracking via 3d orientation estimation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 2575–2584. https://doi.org/10.1145/1978942.1979318
[60] Albrecht Schmidt, Matthias Kranz, and Paul Holleis. 2005. Interacting with the ubiquitous computer: towards embedding interaction. In *Proceedings of the 2005 Joint Conference on Smart Objects and Ambient Intelligence: Innovative Context-Aware Services: Usages and Technologies* (Grenoble, France) *(sOc-EUSAI '05)*. Association for Computing Machinery, New York, NY, USA, 147–152. https://doi.org/10.1145/1107548.1107588
[61] Julia Schwarz. 2014. *Monte Carlo Methods for Managing Uncertain User Interfaces*. Ph. D. Dissertation. PhD thesis, Carnegie Mellon University.
[62] Vasileios Sideridis, Andrew Zacharakis, George Tzagkarakis, and Maria Papadopouli. 2019. GestureKeeper: Gesture Recognition for Controlling Devices in IoT Environments. In *2019 27th European Signal Processing Conference (EUSIPCO)*. 1–5. https://doi.org/10.23919/EUSIPCO.2019.8903044
[63] Amit Singhal and Christopher R Brown. 1997. Dynamic bayes net approach to multimodal sensor fusion. In *Sensor Fusion and Decentralized Control in Autonomous Robotic Systems*, Vol. 3209. SPIE, 2–10.
[64] Kihoon Son, Kyungmin Kim, and Kyung Hoon Hyun. 2022. BIGexplore: Bayesian Information Gain Framework for Information Exploration. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 37, 16 pages. https://doi.org/10.1145/3491102.3517729
[65] Tanya Stivers and Jack Sidnell. 2005. Introduction: multimodal interaction. (2005).
[66] Heung-Il Suk, Bong-Kee Sin, and Seong-Whan Lee. 2010. Hand gesture recognition based on dynamic Bayesian network framework. *Pattern recognition* 43, 9 (2010), 3059–3072.
[67] Mohan M Trivedi, Shinko Yuanhsien Cheng, Edwin MC Childers, and Stephen J Krotosky. 2004. Occupant posture analysis with stereo and thermal infrared video: Algorithms and experimental evaluation. *IEEE transactions on vehicular technology* 53, 6 (2004), 1698–1712.
[68] Eduardo Velloso, Markus Wirth, Christian Weichel, Augusto Esteves, and Hans Gellersen. 2016. AmbiGaze: Direct Control of Ambient Devices by Gaze. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems* (Brisbane, QLD, Australia) *(DIS '16)*. Association for Computing Machinery, New York, NY, USA, 812–817. https://doi.org/10.1145/2901790.2901867
[69] Bryan Wang and Tovi Grossman. 2020. BlyncSync: enabling multimodal smartwatch gestures with synchronous touch and blink. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
[70] Xiaoyang Wang and Qiang Ji. 2012. Incorporating contextual knowledge to dynamic bayesian networks for event recognition. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE, 3378–3381.
[71] John H Williamson, Antti Oulasvirta, Per Ola Kristensson, and Nikola Banovic. 2022. *Bayesian Methods for Interaction and Design*. Cambridge University Press.
[72] Cao Xiao, Yan Jin, Ji Liu, Bo Zeng, and Shuai Huang. 2018. Optimal expert knowledge elicitation for Bayesian network structure identification. *IEEE Transactions on Automation Science and Engineering* 15, 3 (2018), 1163–1177.
[73] Yukang Yan, Chun Yu, Wengrui Zheng, Ruining Tang, Xuhai Xu, and Yuanchun Shi. 2020. FrownOnError: Interrupting Responses from Smart Speakers by Facial Expressions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376810
[74] Yongmian Zhang and Qiang Ji. 2006. Active and dynamic information fusion for multisensor systems with dynamic Bayesian networks. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 36, 2 (2006), 467–472.
[75] Yongmian Zhang, Qiang Ji, and C.G. Looney. 2002. Active information fusion for decision making under uncertainty. In *Proceedings of the Fifth International Conference on Information Fusion. FUSION 2002. (IEEE Cat.No.02EX5997)*, Vol. 1. 643–650 vol.1. https://doi.org/10.1109/ICIF.2002.1021215
[76] Hang Zhao, Sophia Gu, Chun Yu, and Xiaojun Bi. 2022. Bayesian Hierarchical Pointing Models. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 87, 13 pages. https://doi.org/10.1145/3526113.3545708
[77] Min Zou and Suzanne D Conzen. 2005. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics* 21, 1 (2005), 71–79.

A LLM Elicitation

This section contains prompts for LLM elicitation at each new context.

A.1 System Prompt

You are an experienced interaction designer with in-depth knowledge of the gestures commonly used for various embedded interactions involving IoT devices and AR interfaces in user environments. I am designing a probabilistic system to infer users' interaction intentions by observing their gestures. You will provide expert advice on constructing conditional probability distributions for this purpose.

A.2 User Prompt

I have a list of gestures and a set of embedded interactions. Each interaction specifies an action and its associated object. Here are the gestures: {gestures}. The embedded interactions are: {embedded_interactions_strs}. For each interaction, could you assign a most probable gesture, and a confidence score on a scale from 1 to 7 (1 = least confident, 7 = most confident) indicating how likely it is to be used for that interaction?