

Beyond Failure Recovery: An Engagement-Aware Human-in-the-loop Framework for Robotic Systems

Jiaying Fang, Joyce Yang, Zhanxin Wu, Bohan Yang, Tapomayukh Bhattacharjee
Cornell University

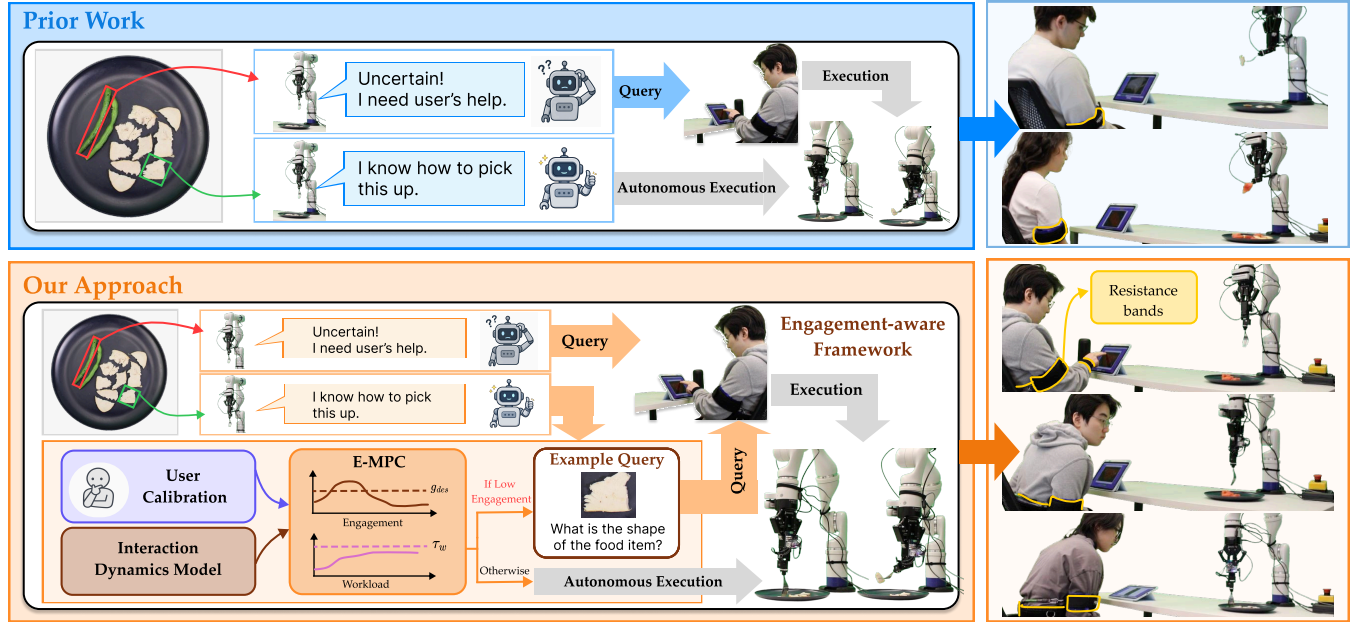


Fig. 1: We propose **Engagement-aware MPC (E-MPC)**, a human-in-the-loop framework for robotic systems that explicitly models user engagement. E-MPC involves users in the robot’s decision-making not only when user’s assistance is needed, but also when interaction is desired to sustain user’s preferred engagement level throughout the task.

Abstract—Conventional human-in-the-loop approaches typically involve users only when a robot encounters failure or uncertainty, treating humans primarily as tools to improve robot performance. However, in many human-centered robotics settings, interaction should support user engagement, keeping users meaningfully involved in decision-making rather than limiting them to failure-driven interventions. For many users, this cannot be achieved through limited, failure-driven interaction alone. This is particularly compelling in physical caregiving, where mobility limitations can reduce users’ ability to intervene or modulate the robot’s behavior in the moment. As a result, interaction policies that engage users only upon failure may further reduce engagement by relegating users to passive observers for long stretches of the task. For example, a user with mobility limitations may experience reduced engagement when being continuously and passively fed by a robot. At the same time, overly frequent interaction can be tiring and increase the user’s workload.

To address this trade-off, we propose Engagement-aware MPC (E-MPC), a user-engagement-aware method that plans interaction to maintain engagement while respecting a workload constraint. E-MPC leverages a user interaction dynamics model that captures how user engagement evolves as a function of both the frequency and type of interaction. Rather than requesting input only when difficulties arise during task execution, the robot proactively considers the user’s preferred level of engagement throughout the task, balancing autonomy and interaction while ensuring task success. We evaluate E-MPC in simulation with several ablations and baseline comparisons. Baselines optimize for task

success alone or jointly for user workload and task success. Results demonstrate the effectiveness of our approach across diverse user personas. In addition, we conduct a real-world user study with participants with emulated mobility limitations on a robot-assisted bite acquisition system, showing that E-MPC improves user experience while maintaining task success. Website: <https://emprise.cs.cornell.edu/empc>

I. INTRODUCTION

Human-in-the-loop robotic policies has traditionally been motivated by the goal of improving the robustness and reliability of autonomous robot policies. In such settings, human guidance can help provide feedback [36, 35, 41] or assist with difficult perception and control challenges [32, 19, 23]. As robots become more capable, these human-in-the-loop mechanisms are increasingly being deployed in human-centered applications like caregiving and household robots, where failures can be costly and reliable execution is essential.

Despite these advances, current interaction strategies are often designed from a robot-centric perspective: the human is consulted mainly when the robot requires additional information to complete an action successfully. In many household or assistive tasks where the human is involved, the quality of robot’s assistance is shaped not only by task success and failure recovery, but also by how the person experiences the process

[3, 6]. This is particularly critical in caregiving applications, where maintaining engagement helps preserve a sense of agency and long-term acceptance of robotic assistance. For example, during robot-assisted feeding, users may prefer to shape how the robot picks up and delivers food, rather than passively receiving continuous robot autonomy [3]. However, meaningful interaction must be balanced carefully: interaction that is too sparse may leave users disengaged, while interaction that is too frequent can increase user workload [1]. This tension is especially pronounced in long-horizon tasks, where the robot repeatedly makes decisions over time and the user's preferences for engagement vary across individuals and circumstances.

In this work, we propose **Engagement-aware Model Predictive Control (E-MPC)**, a framework that explicitly treats user engagement as an objective during task execution. Rather than viewing interaction solely as a mechanism to resolve uncertainty, *E-MPC* enables the robot to proactively decide when and how to involve the user in order to maintain a desired level of engagement throughout the task. Our approach leverages a predictive model of interaction dynamics, capturing how engagement evolves based on both the frequency and type of interaction, while also enforcing constraints on user workload and task success. *E-MPC* provides a principled way to balance autonomy and interaction in long-horizon tasks. In general, our contributions are:

- A human-in-the-loop framework for robotic systems that jointly reasons about user engagement, user workload and task success to improve user experience.
- An interaction dynamics model that models user engagement as a function of interaction frequency and type.
- Extensive evaluations of the framework both in simulation with different personas and on a real robot-assisted bite acquisition system through an in-lab user study with participants with emulated mobility limitations.

II. RELATED WORKS

A. Engagement in Human-Robot Interaction

Prior research in Human-Robot Interaction has highlighted the importance of sustaining user engagement in long-horizon interactions [29, 21, 9, 37]. Studies suggest that engagement can diminish over time without deliberate interaction design, even when robotic task execution remains successful [9, 14, 24]. Engagement has also been linked to a user's sense of agency [6, 39, 12] - the perceived ability to influence ongoing outcomes - which has been shown to support psychological well-being [33] and acceptance of assistive technologies [7, 27]. In addition, in assistive robotics, users can prefer meaningful involvement in robot decision-making over full autonomy [26], valuing a sense of control even when it may reduce task efficiency [3].

A related line of work focuses on recognizing and detecting engagement during interaction [31, 25, 34]. However, these approaches do not model engagement as an objective that can be actively shaped in long-horizon tasks. More closely related to our work, Del Duchetto and Hanheide [8] consider engagement in interaction planning, but aim to maximize

engagement inferred from sensory input. In contrast, we regulate engagement toward a user-specified target and model it as a continuous state governed by engagement dynamics. Our work thus introduces a principled engagement-aware interaction planning framework that treats engagement as a controllable objective.

B. Human-in-the-loop Robotic Policies

Most human-in-the-loop robotic policies seek human assistance when the robot is uncertain about how to complete the task or encounters execution failure [36, 19, 32, 23, 35, 41, 20, 30, 22, 40]. In the DAgger line of work [32, 19, 23], the human expert is iteratively queried for good actions on challenging or out-of-distribution observations. In [35, 20, 30], users provide corrective feedback to the robot policy to help it complete the task. Recently, some work extends the human-in-the-loop framework to jointly optimize task success and human workload, recognizing that excessive querying can increase users' workload [1, 2]. However, these methods do not explicitly incorporate user engagement objectives into policy optimization. In contrast, our method models and optimizes for user engagement in an MPC framework.

Another line of human-in-the-loop robotic policies is shared autonomy, which blends continuous human inputs with autonomous robot control [15, 11]. While these approaches operate at the control level, our setting instead considers a higher-level interaction planner that selects from a discrete set of query actions.

III. PROBLEM FORMULATION

We investigate the problem of achieving successful task completion with the robotic systems while ensuring bounded user workload and optimizing user engagement toward a desired engagement level. The robotic systems considered in this work are designed to accomplish specific tasks with a human user involved. At each time step, the robot may either query the user for input or proceed autonomously.

We assume access to a skill library $\mathcal{S} = \{\phi^1, \phi^2, \dots, \phi^N\}$ consisting of N skills that the robot can use to complete the task. Task execution involves a planner that produces a sequence of skills over discrete time steps $t = 1, 2, \dots, T$ to achieve the goal. Each skill ϕ_t^i executed at time t is associated with a confidence value $c_t = \mathcal{C}(\phi_t^i, \mathbf{o}_t) \in [0, 1]$, where $\mathbf{o}_t$ denotes the observation at time t and $\mathcal{C}$ is the confidence estimation function. Larger values of c_t indicate a higher likelihood of successful skill execution. If a skill execution fails, the system may retry up to τ_r times.

In our human-in-the-loop robotic system, the robot may interact with the user by selecting an interaction action $\mathbf{q}_t \in \mathcal{Q}$, where $\mathcal{Q}$ denotes the finite set of available query types. Because the robot may also choose to proceed autonomously without querying the user, this set includes a no-query action represented by the zero vector $\mathbf{0} = [0, 0] \in \mathcal{Q}$. Each query action is parameterized by a two-dimensional vector $\mathbf{q}_t = [q_{d,t}, q_{r,t}]$, where $q_{d,t}$ denotes the query difficulty and $q_{r,t}$ denotes the query response difficulty.

Each skill ϕ^i is associated with a set of *task-assistive* query types, $\mathcal{Q}^{\text{assist},\phi^i} \subseteq \mathcal{Q}$, which provides functional user input that can directly help the robot successfully executing the skill under uncertainty. For example, in a bite-acquisition system, when the autonomous policy is highly uncertain about where to skewer a food item, a drawing-based query that asks the user to indicate the skewering point on an image can directly support the successful execution of the skewering skill. In addition, each skill ϕ^i is associated with a set of *feasible* query types, $\mathcal{Q}^{\text{feas},\phi^i} \subseteq \mathcal{Q}$, representing all task-relevant queries that the robot may ask during execution. This feasible set includes the *task-assistive* queries, $\mathcal{Q}^{\text{assist},\phi^i} \subseteq \mathcal{Q}^{\text{feas},\phi^i}$, as well as queries that may increase user engagement but do not really help the robot complete the task. For instance, the robot may ask “fake” queries, such as asking the user about the shape of a food item. These interactions can enhance engagement but do not affect execution of the skewering skill. At each time step t , before executing skill ϕ_t^i , the robot can either ask a query by selecting $\mathbf{q}_t \in \mathcal{Q}^{\text{feas},\phi_t^i}$ or execute the skill autonomously ($\mathbf{q}_t = \mathbf{0}$). If a query is asked, the user provides a response, and the robot may incorporate this information to carry out the skill. We denote the time steps at which the robot issues a query ($\mathbf{q}_t \neq \mathbf{0}$) by $\{t_k\}_{k=1}^K$, where $t_k \in \{1, \dots, T\}$, $K \leq T$.

Finally, we represent user workload at time t as $w_t \in [0, 1]$, and user engagement as $g_t \in [0, 1]$. The user’s desired engagement level is denoted by $g_{\text{des}} \in [0, 1]$.

IV. APPROACH

A. Interaction Dynamics Modeling

1) *Engagement-Related Interaction Variables*: We hypothesize that user engagement is related to:

- **Query response difficulty** $q_{rd,t}$: This represents the different response difficulties of a robot query. Different query response difficulties can lead to different engagement outcomes. We consider two types of query response difficulties: (a) the response difficulty of a multiple-choice question, for example, asking the user to choose one skill from multiple skills (b) the response difficulty of a drawing-based query, for example, asking user to draw action points or a bounding box on an image.
- **Query frequency** δ_t : This represents the time gap between queries. Specifically, $\delta_t = \text{seconds_per_timestep} \times (t - t_k)$, represents the time gap in seconds between the current time step t and the time step t_k which the previous query happened.

2) *Interaction Dynamics Model*: Inspired by the above two variables, we treat engagement g_t as a latent internal user state that is not directly observable, but estimated online through interaction outcomes and calibrated query impacts. We model the interaction dynamics as:

$$g_t = e^{-\lambda_g \delta_t} g_{t_k} + (1 - e^{-\lambda_g \delta_t}) k_{q_{rd}} q_{rd,t}, \quad (1)$$

where

$$q_{rd,t} = \mathbf{w}_{\text{resp}}^T \mathbf{q}_t, \quad (2)$$

and $\lambda_g > 0$ is the decay rate. g_{t_k} denotes the engagement level at the most recent timestep t_k when a query action was issued. The operator $\mathbf{w}_{\text{resp}} = [0, 1]$ projects the interaction action $\mathbf{q}_t = [q_{d,t}, q_{rd,t}] \in \mathcal{Q}$ onto the query response difficulty component $q_{rd,t}$. $k_{q_{rd}}$ denotes an one-to-one mapping from query response difficulty $q_{rd,t}$ to query impact values in $[0, 1]$, which capture the immediate engagement boost induced at the moment of interaction. In real-robot experiments, this mapping is obtained through a user-specific calibration process conducted before the user study. When no query is issued at time t , i.e., $\mathbf{q}_t = \mathbf{0}$, the immediate engagement boost $k_{q_{rd}} q_{rd,t}$ evaluates to zero, and engagement evolves purely via decay.

The proposed interaction dynamics model exhibits several key behavioral characteristics:

- Engagement evolves as a combination of discounted prior engagement level and instantaneous engagement impulse. The effect of query frequency δ_t is captured through the discounting of prior engagement. The effect of query response difficulty $q_{rd,t}$ is encoded in the instantaneous engagement impulse.
- Older interactions gradually fade, while recent ones exert greater influence, reflecting the temporally evolving nature of engagement [38, 28].
- Different query response difficulties lead to different query impacts, which aligns with the finding in [4, 10]. The engagement level is bounded by the range of the query impacts.

We can summarize the parameter λ_g has the following roles: It controls the relative weighting between previous engagement level and the instantaneous engagement impulse. A larger λ_g means less persistent engagement memory and more sensitive to current engagement impulse. A smaller λ_g means more persistent engagement memory and less sensitive to current engagement impulse.

B. User Workload Modeling

Following [1], we directly adopt a linear discrete-time user workload model:

$$w_t = \gamma w_{\text{init}} + \sum_{j=0}^{H-1} \mathbf{w}_j^T \mathbf{q}_{t-j} + w_0 \quad (3)$$

where γw_{init} is the weighted initial workload, $\mathbf{q}_{t-j} = [q_{d,t-j}, q_{rd,t-j}]$ is the interaction action happened j time steps in the past, $\mathbf{w}_j$ represents the effect of the interaction action $\mathbf{q}_{t-j} \in \mathcal{Q}$ on the workload at the current time t , H is the length of the interaction history the workload model considers, and w_0 is a bias term.

C. Engagement-aware MPC

Given a robotic system, we aim to ensure task success while bounding user workload and optimizing user engagement toward a user-specified desired engagement level. This problem is challenging due to the subjectiveness in engagement and workload modeling, as well as the noise in the user’s stated engagement preference. We address these challenges by formulating the problem as a model predictive control

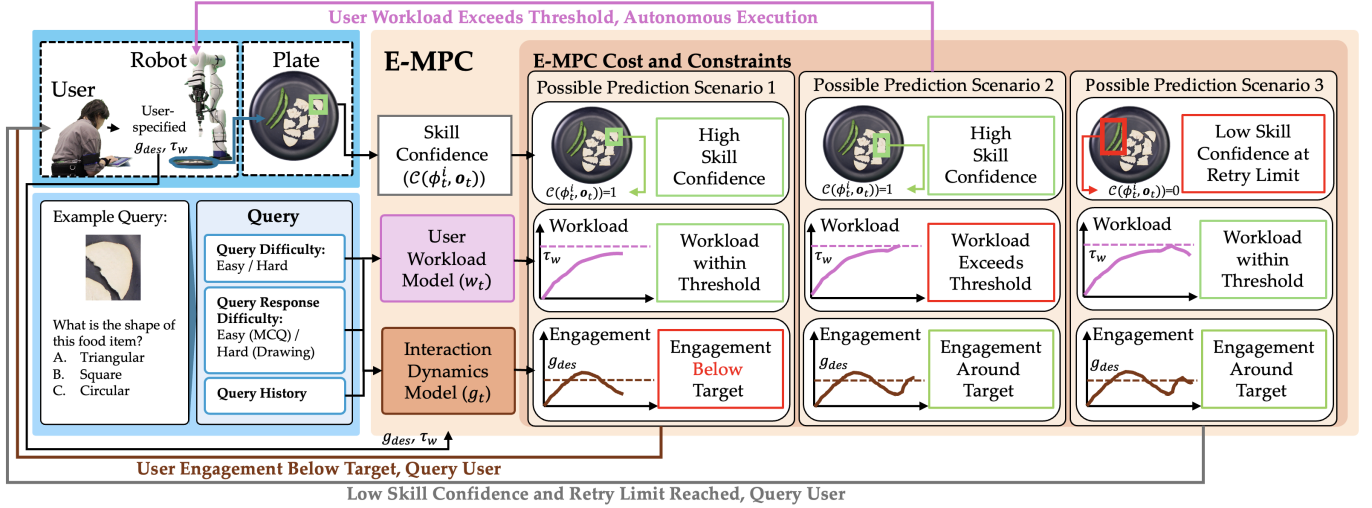


Fig. 2: **E-MPC framework.** *E-MPC* proactively issues user queries to regulate engagement g_t toward a user-specified desired level g_{des} throughout task execution. By predicting the engagement effect of different query types, *E-MPC* selects interactions that provide engagement boosts while keeping user workload w_t bounded. When task success is at risk (low confidence $\mathcal{C}(\phi_t^i, \mathbf{o}_t)$ and retries exhausted), the controller requires *task-assistive* queries, and replans in a receding-horizon manner.

(MPC) framework that jointly reasons about task success, user workload, and engagement level, while allowing the user to correct their engagement preference g_{des} online. The MPC optimizes over interaction actions $\mathbf{q}_t \in \mathcal{Q}$ only; task-level skills are provided by an external planner.

1) *State Variables and Control Input:* To reason about user engagement and user workload, the controller maintains three state variables:

$$g_t \in [0, 1], \quad g_{des} \in [0, 1], \quad w_t \in [0, 1].$$

g_t captures the current estimated user engagement level at time t , g_{des} is the desired user engagement level specified by the user. We assume that g_{des} remains fixed over each MPC rollout, but may be updated between rollouts based on user feedback. w_t is the estimated user workload. The dynamics model of g_t and w_t are explained in Sec. IV-A and Sec. IV-B. The control of the MPC controller is the interaction action $\mathbf{q}_t \in \mathcal{Q}$.

2) *User Desired Engagement Level:* The user desired engagement level g_{des} is a self-reported preference, which specifies how actively the user wishes to participate in decision-making during the interaction and serves as the target engagement level g_{des} for the system. Users can specify g_{des} prior to the robot system rollout and update it during the rollout.

3) *Engagement Tracking Cost:* At prediction step l , let g_{t+l} be the predicted engagement state. The MPC cost is

$$\mathcal{J}_g = (g_{des} - g_{t+l})^2 \quad (4)$$

This term encourages the estimated engagement level g_{t+l} to track the desired engagement level g_{des} .

4) *Task Success Constraint:* A task-related constraint is set to ensure the successful completion of the task. At prediction step l , the confidence of the skill to be executed is $\mathcal{C}(\phi_{t+l}^i, \mathbf{o}_{t+l})$. When this confidence falls below a threshold τ_c , and the current number of retries $r_t^{\phi_{t+l}^i}$ of the skill reaches the maximum allowed number of retries τ_r , the robot is required to seek user assistance. Otherwise, the robot may either select any query

from the feasible set $\mathcal{Q}^{feas, \phi_t^i}$ or proceed autonomously without querying the user. Accordingly, we impose the following constraint on the interaction action:

$$\mathbf{q}_{t+l} \in \begin{cases} \mathcal{Q}^{assist, \phi_{t+l}^i}, & \text{if } \mathcal{C}(\phi_{t+l}^i, \mathbf{o}_{t+l}) \leq \tau_c \\ & \text{and } r_t^{\phi_{t+l}^i} = \tau_r, \\ \mathcal{Q}^{feas, \phi_{t+l}^i} \cup \{\mathbf{0}\} & \text{otherwise.} \end{cases} \quad (5)$$

5) *User Workload Constraint:* To avoid making the user feel overwhelmed by the robot queries, we also set a safe workload threshold τ_w . At prediction step l , let w_{t+l} be the predicted workload, the workload constraint is formulated as a cost term:

$$\mathcal{J}_w = \lambda_w [\max(0, w_{t+l} - \tau_w)]^2, \quad (6)$$

where λ_w is a large constant. This term allows the robot to exceed the workload threshold τ_w if necessary, but any violation is heavily penalized by the quadratic cost.

6) *Finite-Horizon MPC Optimization:* At time t , with prediction horizon L , the controller solves:

$$\begin{aligned} \min_{\{\mathbf{q}_{t+l}\}} \quad & \sum_{l=0}^{L-1} \left((g_{des} - g_{t+l})^2 + \lambda_w [\max(0, w_{t+l} - \tau_w)]^2 \right) \\ \text{s.t.} \quad & \text{for } l = 0, \dots, L-1, \\ & \mathbf{q}_{t+l} \in \begin{cases} \mathcal{Q}^{assist, \phi_{t+l}^i}, & \text{if } \mathcal{C}(\phi_{t+l}^i, \mathbf{o}_{t+l}) \leq \tau_c \\ & \text{and } r_t^{\phi_{t+l}^i} = \tau_r, \\ \mathcal{Q}^{feas, \phi_{t+l}^i} \cup \{\mathbf{0}\} & \text{otherwise.} \end{cases} \end{aligned} \quad (7)$$

Only the first control input is applied, after which the horizon recedes and the optimization is repeated.

V. EVALUATION SCENARIO

We now instantiate the proposed *E-MPC* framework on a robot-assisted bite acquisition system. Robot-assisted feeding enhances the autonomy of individuals with mobility limitations [18, 13, 16, 3]. Respecting and supporting user engagement in

long-horizon tasks like bite acquisition in feeding can further empower users and promote a sense of agency [17, 18].

A. System Overview

The goal of our robot-assisted bite acquisition system is to acquire food items from a plate using a utensil. At each step of the task plan, the robot locates a food item from an image of the plate and executes one food acquisition skill. If the skill execution fails, the system can retry two times ($\tau_r = 2$) on the same food item after the first failure. Following [2], we estimate the acquisition 6D pose for different food items using a skill parameter estimation model, which is a transformer-based vision-language-action model. The model takes as input a textual skill description, along with the current bounding box image as visual context, and outputs the corresponding acquisition pose. The primary uncertainty in this system comes from this skill parameter estimation model. To get human assistance, the robot can ask the user to draw the acquisition position and angle on the bounding box image.

B. Skill Library

The skill library $\mathcal{S}_{\text{food}}$ of the robot-assisted bite acquisition system consists of three acquisition skills:

- *Skewering* (ϕ^1): Skewers the target food item. The system estimates the 3D interaction point between the utensil and the food, along with the utensil roll angle.
- *Scooping* (ϕ^2): Scoops the food item. The system estimates the 3D start and end points of the scooping trajectory.
- *Twirling* (ϕ^3): Twirls the food item. The system estimates the 3D interaction point and the corresponding utensil roll angle required for twirling.

C. Skill Confidences and Query Types

The confidence score for each acquisition skill is modeled as a binary value, derived from two sources: (1) the uncertainty of the transformer-based skill parameter estimation model, and (2) a failure identification module. The uncertainty of the skill parameter estimation model comes from the variance of the its predictions using Monte Carlo Dropout [2]. The failure detection module detects whether the predicted interaction point (e.g., a skewering location) is within the segmentation mask of the food item (see Appendix for details). Since food acquisition failures often stem from inaccurate parameter estimation, we design the following *task-assistive queries* to obtain user assistance:

- *Skewering* query: The user is shown a cropped bounding-box image of the target food item and asked to draw the skewer point and angle.
- *Scooping* query: The user is shown a cropped bounding-box image and asked to draw the start and end points of the scooping trajectory.
- *Twirling* query: The user is shown a cropped bounding-box image and asked to draw the twirling point and angle.

In addition to these required queries, we introduce an auxiliary query type intended solely to promote user engagement:

- *Food item shape* query: The user is shown a cropped image of the food item and asked to select its shape from multiple-choice options.

The query type of *task-assistive* queries is denoted as $\mathbf{q}_t^{\text{food,points}} = [q_{d,t} = 2(\text{Hard}), q_{rd,t} = 2(\text{Hard})]$, while the engagement-only query type is denoted as $\mathbf{q}_t^{\text{food,shape}} = [q_{d,t} = 1(\text{Easy}), q_{rd,t} = 1(\text{Easy})]$.

VI. SIMULATION EXPERIMENTS

In simulation experiments, we aim to answer the following question: compared to baseline policies, does *E-MPC* optimize user engagement towards a desired target level while achieving task success and maintaining bounded user workload?

A. Baselines

We compare *E-MPC* against the following four baselines:

- *AlwaysQuery*: asks the *task-assistive* query at every step.
- *NeverQuery*: executes all skills autonomously without querying the user.
- *TaskConfidenceAware*: asks the *task-assistive* query when the predicted skill success confidence $\mathcal{C}(\phi_t^i, \mathbf{o}_t)$ falls below a threshold. This baseline optimizes solely for task success.
- *WorkloadAware*: asks the *task-assistive* query when the skill uncertainty, $1 - \mathcal{C}(\phi_t^i, \mathbf{o}_t)$, exceeds a scaled version of the user workload $\gamma_{\text{scale}} w_t$, thereby jointly accounting for task success and user workload. This baseline is the method proposed in [1], which is the state-of-the-art method that considers both task success and user workload.

B. Metrics

We evaluate the performance of different methods using three metrics: *Skill Success*, *Engagement Tracking Accuracy*, and *Satisfaction*.

- 1) *Skill Success* measures the fraction of skill executions in which the skill succeeds within a maximum of τ_r retries.
- 2) *Engagement Tracking Accuracy* evaluates how well a method maintains the desired engagement level g_{des} . We define it as one minus the Root-Mean-Square Error (RMSE) between the tracked engagement g_t and g_{des} over a horizon of length T : $\text{RMSE} = \sqrt{\frac{1}{T} \sum_{t=1}^T (g_t - g_{des})^2}$, $\text{EngagementAcc} = 1 - \text{RMSE}$.
- 3) *Satisfaction* captures the overall user experience by jointly accounting for engagement tracking, workload compliance, and task completion. The satisfaction score is defined as

$$\text{Satisfaction} = \begin{cases} 1 - |g_t - g_{des}|, & \text{if } w_t \leq \tau_w \text{ and skill} \\ & \text{succeeds in } \tau_r \text{ retries,} \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

This metric reflects that accurate engagement tracking is only meaningful when achieved under bounded workload and reliable task success.

Note that the metrics used in simulation experiments here are not intended to measure whether the engagement model aligns with true user engagement, instead they are used to evaluate whether *E-MPC* can effectively optimize engagement

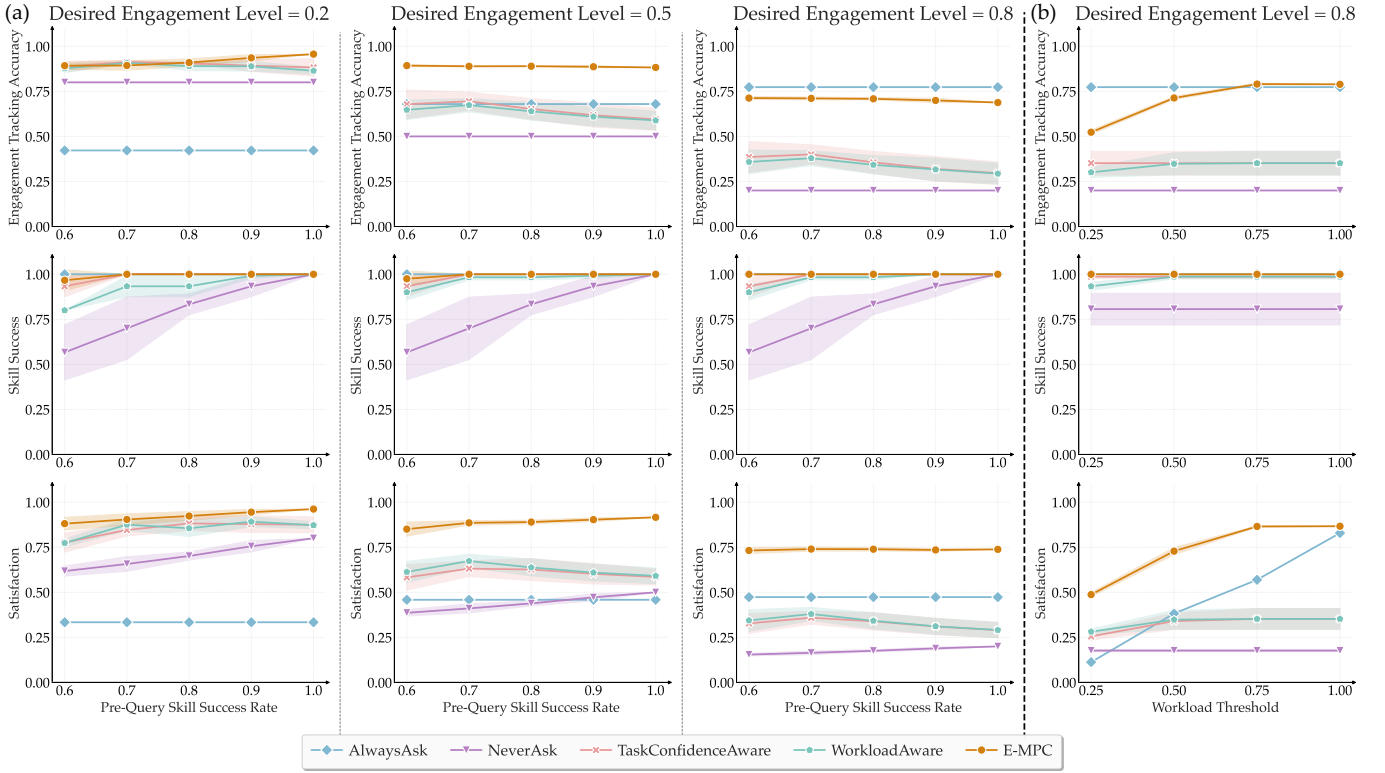


Fig. 3: Simulation Results. (a) Performance of methods across varying pre-query skill success rates and user personas. Results are averaged over workload thresholds $\tau_w \in \{0.25, 0.5, 0.75, 1.0\}$. For *WorkloadAware* baseline, we set the value of the hyperparameter γ_{scale} as the value that maximizes the *Satisfaction* metrics in each setting. (b) Performance of methods across varying workload thresholds τ_w with high-engagement user persona. Results are averaged over pre-query skill success rates $p_{\text{success}} \in \{0.6, 0.7, 0.8, 0.9, 1.0\}$.

and outperform baseline methods, while assuming our model reflects true user engagement. In Sec. VII, the user study metrics then evaluate whether our framework with the proposed engagement model aligns with the user experience.

C. Experiment Settings

In simulation, the robot can issue two types of queries that are mentioned in Sec. V-C: easy MCQs and hard drawing-based questions. We assign the first query type a query impact k_{qrd} of 0.5; the second query type a query impact k_{qrd} of 1.0. Thus, the easy MCQs provide a modest engagement boost with low workload cost, whereas drawing-based queries yield stronger engagement gains but impose substantially higher workload.

The skill confidence $\mathcal{C}(\phi_t^i, \mathbf{o}_t)$ is modeled as a noisy binary estimate, reflecting real-world prediction uncertainty with imperfect true-positive and false-positive rates (TPR= 0.9, FPR= 0.1). Each rollout consists of $T = 10$ task-planning steps, with a maximum of $\tau_r = 2$ retries allowed after a skill execution failure. The MPC prediction horizon is set to $L = 5$.

To capture diverse user engagement preferences, we consider three personas with low, medium, and high desired engagement levels: $g_{\text{des}} \in \{0.2, 0.5, 0.8\}$. We evaluate each method under varying pre-query skill success rate $p_{\text{success}} = P(\phi_t^i \text{ succeeds without querying}) \in \{0.6, 0.7, 0.8, 0.9, 1.0\}$ and workload thresholds $\tau_w \in \{0.25, 0.5, 0.75, 1.0\}$. Each condition $(p_{\text{success}}, \tau_w)$ is evaluated over three rollouts.

D. Results

As shown in Fig. 3(a), *E-MPC* consistently achieves high *Engagement Tracking Accuracy* across all three personas. The *TaskConfidenceAware* and *WorkloadAware* baselines attain comparable tracking performance when the desired engagement level g_{des} is relatively low. However, as g_{des} increases, the advantage of *E-MPC* becomes more pronounced when compared to these two baselines. When g_{des} becomes very high, *AlwaysQuery* can also track engagement well, as it maximizes interaction regardless of workload or task conditions.

In addition, *E-MPC* achieves higher *Skill Success* than all other baselines except *AlwaysQuery*. This improvement arises because *E-MPC* queries the user not only when the predicted skill confidence $\mathcal{C}(\phi_t^i, \mathbf{o}_t)$ is low, but also when interaction is beneficial for maintaining the target engagement level g_t . By doing so, *E-MPC* can mitigate the impact of occasional false-positive confidence estimates caused by noisy skill confidences, leading to higher overall success rates.

Finally, *E-MPC* outperforms all baselines on the *Satisfaction* metric. This indicates that *E-MPC* not only tracks engagement effectively, but also ensures task success while keeping user workload bounded. Among the baselines, *WorkloadAware* is the strongest overall; however, because it does not explicitly account for engagement, its satisfaction scores drop substantially for medium- and high-engagement personas. *AlwaysQuery* achieves strong engagement tracking and high task success for highly engaged users, but its disregard for workload leads to

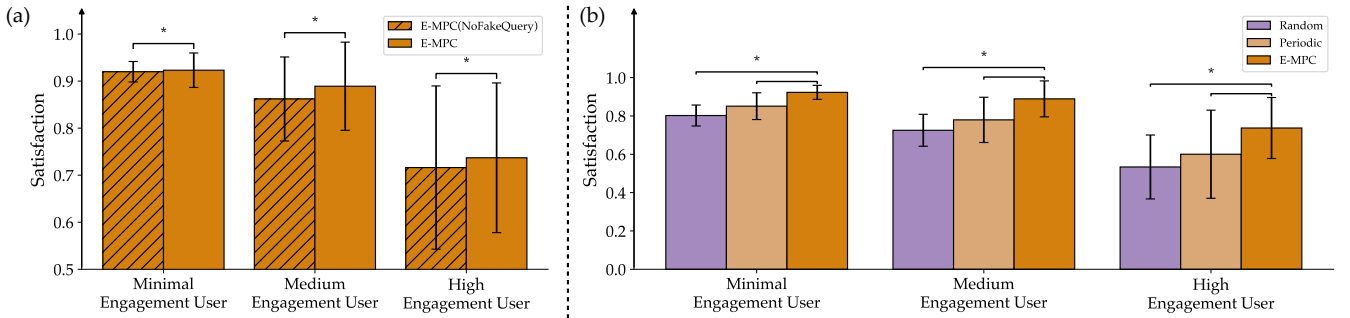


Fig. 4: Effect of “Fake” Queries and Heuristic Baselines. (a) *Satisfaction* for three personas under *E-MPC* and its No-Fake-Queries ablation. (b) *Satisfaction* for three personas under *E-MPC* and two heuristic baselines: *Random* and *Periodic*. Results are averaged over pre-query skill success rates $p_{\text{success}} \in \{0.6, 0.7, 0.8, 0.9, 1.0\}$ and workload thresholds $\tau_w \in \{0.25, 0.5, 0.75, 1.0\}$. * indicates statistical significance $p < 0.05$, determined via a Mann-Whitney U test.

TABLE I: Performance of *E-MPC* and its ablations across three desired engagement levels (pre-query skill success rate $p_{\text{success}} = 80\%$ and workload threshold $\tau_w = 0.5$)

g_{des}	Mode	Engagement Acc. $\uparrow$	Workload Compliance (%) $\uparrow$	Satisfaction $\uparrow$
0.2	E-MPC	0.910 ± 0.021	100.0 ± 0.000	0.933 ± 0.012
	NoWorkloadCst	0.910 ± 0.021	100.0 ± 0.000	0.933 ± 0.012
	NoEngagementTrack	0.862 ± 0.013	100.0 ± 0.000	0.880 ± 0.013
0.5	E-MPC	0.917 ± 0.003	100.0 ± 0.000	0.942 ± 0.007
	NoWorkloadCst	0.919 ± 0.001	100.0 ± 0.000	0.933 ± 0.012
	NoEngagementTrack	0.594 ± 0.025	100.0 ± 0.000	0.607 ± 0.024
0.8	E-MPC	0.720 ± 0.013	97.22 ± 4.811	0.730 ± 0.028
	NoWorkloadCst	0.786 ± 0.000	50.00 ± 0.000	0.382 ± 0.000
	NoEngagementTrack	0.300 ± 0.027	100.0 ± 0.000	0.307 ± 0.024

reduced satisfaction.

E. Workload Thresholds

Fig. 3(b) compares baseline methods and *E-MPC* under a high desired engagement level ($g_{\text{des}} = 0.8$). When the maximum allowable workload threshold τ_w is low, *E-MPC* tracks engagement less accurately than *AlwaysAsk*, since it prioritizes maintaining workload compliance over engagement interaction. Importantly, this prioritization leads to a better overall user experience: this is reflected in the *Satisfaction* metric, where *E-MPC* achieves substantially higher satisfaction for low to medium workload thresholds.

F. Ablations

In addition to the baselines above, we also compare *E-MPC* against the following ablations: (1) *NoWorkloadCst*: our method without the user workload constraint; (2) *NoEngagementTrack*: our method without the engagement tracking cost.

In Table I, beyond *Engagement Tracking Accuracy* and *Satisfaction*, we further report the *Workload Compliance Rate*, which measures the fraction of timesteps during which the user workload remains below the maximum allowable threshold τ_w . When g_{des} is low or moderate, *NoWorkloadCst* performs comparably to *E-MPC*. However, as g_{des} increases to 0.8, *NoWorkloadCst* frequently violates the workload constraint, exceeding τ_w in roughly 50% of timesteps. Since *NoEngagementTrack* does not account for engagement, it consistently underperforms *E-MPC* across all settings.

G. Effect of “Fake” Queries

We also explore the effect of engagement-oriented (“fake”) queries in simulation. In our setting, these queries take the form of easy MCQs that are intended solely to increase user engagement and do not help the robot complete the task. We compare *E-MPC* against an ablation that excludes these engagement-oriented queries and permits only *task-assistive* queries. As shown in Fig. 4(a), incorporating engagement-oriented queries yields a statistically significant improvement in the *Satisfaction* metric across all three personas. This improvement arises because these “fake” queries provide engagement boosts while imposing relatively low additional workload, enabling the robot to better align with user engagement preferences without exceeding maximum allowed workload.

H. Comparison with Heuristic Baselines

We compare *E-MPC* against two heuristic baselines that approximate engagement via simple query scheduling. The *Random* baseline queries with probability $P(\mathbf{q}_t \neq \mathbf{0}) = g_{\text{des}}$, while the *Periodic* baseline queries at fixed intervals determined by g_{des} . As shown in Fig. 4(b), *E-MPC* consistently achieves higher satisfaction across all user types compared to heuristic baselines. While *Random* and *Periodic* regulate query frequency, they fail to adapt interaction to task uncertainty or user state, leading to suboptimal behavior. In contrast, *E-MPC* explicitly models engagement and selects both when and how to query, leading to better engagement tracking and higher satisfaction.

VII. REAL-ROBOT USER STUDY

We conducted a Cornell University IRB-approved real-robot user study on a robot-assisted bite acquisition system to evaluate the effectiveness of *E-MPC*. The study involved 10 participants (4 female, 6 male; ages 19-29). Participants provided informed consent and were compensated for their time.

A Kinova Gen3 7-DoF arm with a Robotiq 2F-85 gripper is used in the study (Fig. 5(left)). The tool attached to the end-effector is adapted from pre-existing tools [17]. As described in Sec. V-C, the robot-assisted bite acquisition system supports two types of user queries. Among them, only the drawing-based query is needed for successful task completion when the robot becomes uncertain. The multiple-choice question about food shape is a “fake” query that is related to the task but does not

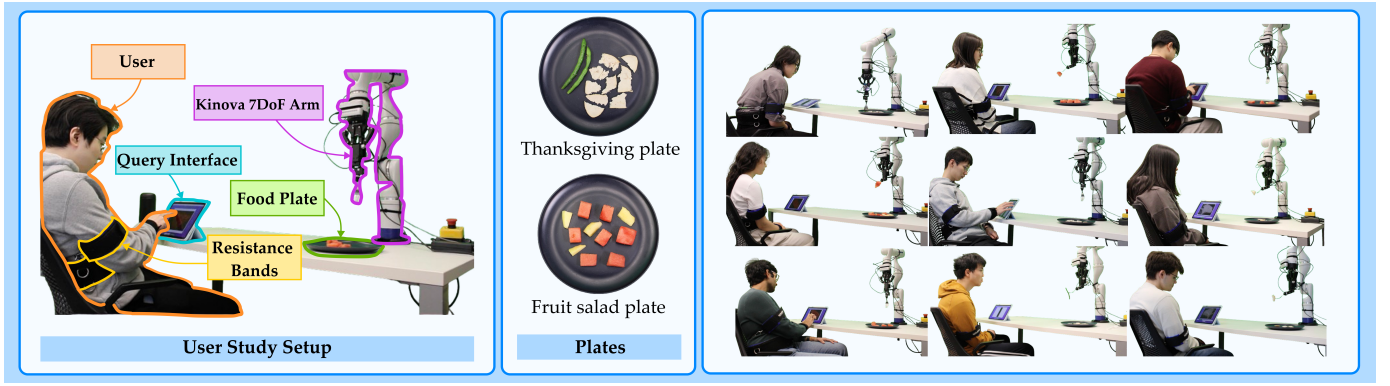


Fig. 5: **Study Setup.** (left) User study setup. (middle) Plates used in the study: a Thanksgiving plate and a fruit salad plate. (right) Participants in the user study with emulated mobility limitations using resistance bands (faces shown with full permission).

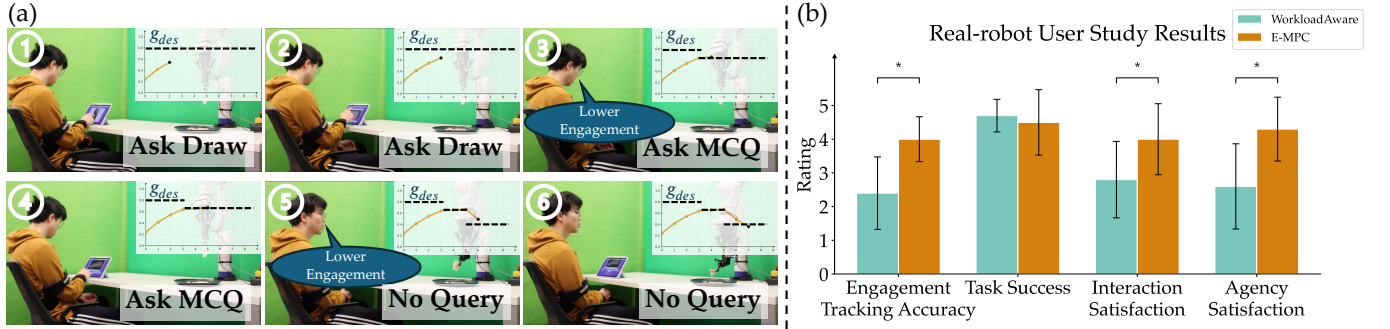


Fig. 6: **User Study Results.** (a) Visualization of a rollout from the user study. The user initially set the $g_{des} = 0.8$. During the process, the user lowered g_{des} twice. *E-MPC* adapts to the updated g_{des} . (b) Participants ratings on four subjective metrics. * indicates statistical significance ($p < 0.05$), determined via a Mann-Whitney U test.

help the execution of the skill. During each study session, the robot attempts to acquire 10 food items from the plate, and up to $\tau_r = 2$ retries are permitted.

A. Baseline and Metrics

Since *WorkloadAware* is the strongest baseline overall based on our simulation results and is the state-of-the-art method proposed in [1], we therefore select *WorkloadAware* as the primary comparison method alongside *E-MPC* in the real-robot user study.

The user study metrics are designed to evaluate whether our framework with the proposed engagement model aligns with what the users care about. Therefore, these metrics are directly self-reported by participants based on their responses to the following five questions, which we use to evaluate user experience and system performance. All user study metrics except *Workload Compliance* are reported on a 5-point Likert scale, while *Workload Compliance* is reported as a binary (yes/no) outcome:

- 1) *Engagement Tracking Accuracy*: "How successful was it for keeping you engaged at your desired engagement level?"
- 2) *Workload Compliance*: "Is the workload you felt during the task acceptable?"
- 3) *Task Success*: "How satisfied are you with how successfully the robot completed the task?"
- 4) *Interaction Satisfaction*: "How satisfied are

you with how the robot interacted with you?"

- 5) *Agency Satisfaction*: "How satisfied are you with how much control/agency you felt during the task?"

B. Study Procedure

Before the user study, we perform several user-specific calibration steps. First, we calibrate the query impact parameter k_{qrd} . Specifically, we present each of the two feasible query types to the participant and ask them to rate the maximum engagement level they would expect if that query were issued at every timestep over a prolonged interaction. Next, we collect the participant's desired engagement level g_{des} using a continuous slider, where 0 indicates no engagement and 1 indicates full engagement. Finally, the participant specifies the maximum workload they are willing to tolerate during the task via a slider, which we set as the workload threshold τ_w . After calibration, each subject evaluates the baseline method and *E-MPC* on the same plate. Across the study, we use two plates, a Thanksgiving plate and a fruit salad plate, with plate assignment counterbalanced between participants (Fig. 5 (middle)).

Following each method, participants completed a post-method questionnaire with five questions corresponding to our five evaluation metrics. After completing the full study, participants completed a post-study questionnaire in which they selected their preferred method and provided additional qualitative feedback.

The calibration details and post-method questionnaire are described in the Appendix.

C. Results

From user-reported ratings shown in Fig. 6(b) (see Appendix for detailed statistics), we find that *E-MPC* achieves significantly better performance than the baseline on *Engagement Tracking Accuracy*, *Interaction Satisfaction*, and *Agency Satisfaction*. Both the *WorkloadAware* [1] and *E-MPC* perform comparably on the *Task Success* metric. In addition, all 10 participants reported that *Workload Compliance* is achieved, meaning that their workload remained acceptable under both methods throughout the study. In the post-study questionnaire, 9 out of 10 participants preferred *E-MPC* over the state-of-the-art baseline [1] for providing a more aligned engagement experience. One participant noted, “*I do think more control is better for me before it gets tiring*,” and another commented, “*The balance of engagement and successful runs brought satisfaction*.” This feedback reinforces our quantitative findings, suggesting that participants valued the increased engagement enabled by *E-MPC* while maintaining task success and a comfortable interaction experience.

Notably, among all queries that the robot asked users during the user study, 51.65% were “fake” queries, while 48.35% were *task-assistive* queries. These results suggest that *E-MPC* effectively maintains user engagement near their desired level with strategic use of “fake” queries.

To complement self-reports, we also include an additional eye-gaze analysis from study videos. We observe that when user engagement g_t deviates from the desired level g_{des} , users’ gaze behavior becomes less task-focused, suggesting a measurable behavioral difference associated with engagement misalignment, which is consistent with prior work [5, 31]. Detailed analysis and results are provided in the Appendix.

VIII. DISCUSSIONS AND LIMITATIONS

We present *E-MPC*, a human-in-the-loop framework for robotic systems that explicitly models and optimizes user engagement. *E-MPC* enables robots to reason about when and how to involve users. Through simulation across diverse user personas and a real-robot bite-acquisition user study, we demonstrated that *E-MPC* improves interaction satisfaction and perceived agency without sacrificing task success or exceeding user workload limits. These results highlight the importance of user-engagement-aware interaction planning for human-in-the-loop policies, and open new opportunities for designing robotic systems that better support user-centered assistance.

An important direction for future work is to make engagement regulation more transparent and negotiable between the user and the robot. In the current formulation, *E-MPC* can use “fake” queries that do not directly affect task execution to increase engagement with relatively low additional workload. This design helps separate engagement regulation from task assistance, but it may also raise concerns about transparency and deception: users may interpret these queries as task-relevant even when they are not. An alternative approach would

be to increase engagement through explicit and transparent interaction strategies.

Our findings also motivate a deeper investigation into the relationship between user engagement, workload, task success, and perceived agency. In our current framework, engagement and workload are modeled explicitly, whereas perceived agency is evaluated only through self-report. Although engagement may contribute to agency, agency is a distinct construct that cannot be fully explained by engagement alone. For example, maintaining engagement near a desired level may help users feel that they can meaningfully influence the robot’s behavior, but this effect may depend on whether the interaction also feels manageable and whether the robot succeeds at the task. Excessive workload may make users feel burdened rather than agentic, while too little involvement may make them feel passive or disconnected from the task outcome. Moreover, perceived agency may depend on other factors beyond engagement, workload, and task success, such as interface design, transparency, and user expectations. Therefore, modeling perceived agency would require a dedicated study, which is beyond the scope of this project.

One limitation of our work is the reliance on an explicit calibration procedure. In the current system, the pre-task calibration process is used both to set the user’s desired engagement level and workload threshold, and to calibrate the engagement model using hypothetical interaction scenarios, which can be noisy. Future work could improve this process by estimating both the user’s current engagement level and desired engagement level online from behavioral signals, interaction history, and direct user feedback.

A second limitation is that our current formulation models interaction primarily through query actions. Queries provide a convenient and interpretable mechanism for eliciting user input. However, real-world human-robot interaction supports a broader range of interaction modalities, including shared control, conversational feedback, implicit corrections, physical guidance, explanations, and socially grounded behaviors, etc. Extending *E-MPC* beyond explicit queries would allow the framework to capture richer and more natural forms of engagement.

Finally, while our user study evaluates *E-MPC* on a real robot in an assistive bite-acquisition task, it does not include participants with mobility limitations, who are a key target population for assistive robotics. Evaluating the framework with users who have lived experience with mobility constraints is an important next step. Such studies may reveal different engagement preferences, workload sensitivities, and would provide a stronger evaluation of whether *E-MPC* can improve user engagement during real-world assistive robot deployments.

ACKNOWLEDGMENTS

This work was partly funded by National Science Foundation IIS #2132846, and CAREER #2238792. The authors would like to additionally thank Rohan Banerjee for helping with the workload model, Yunting Yan for helping with the user study design, and all of the participants in our user study.

REFERENCES

- [1] Rohan Banerjee, Rajat Kumar Jenamani, Sidharth Vasudev, Amal Nanavati, Katherine Dimitropoulou, Sarah Dean, and Tapomayukh Bhattacharjee. To ask or not to ask: Human-in-the-loop contextual bandits with applications in robot-assisted feeding. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1378–1384. IEEE, 2025.
- [2] Rohan Banerjee, Krishna Palempalli, Bohan Yang, Jiaying Fang, Alif Abdullah, Tom Silver, Sarah Dean, and Tapomayukh Bhattacharjee. A human-in-the-loop confidence-aware failure recovery framework for modular robot policies. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2026.
- [3] Tapomayukh Bhattacharjee, Ethan K Gordon, Rosario Scalise, Maria E Cabrera, Anat Caspi, Maya Cakmak, and Siddhartha S Srinivasa. Is more autonomy always better? exploring preferences of users with mobility impairments in robot-assisted feeding. In *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, pages 181–190, 2020.
- [4] Laura Esther Bijkerk, Anke Oenema, Nicole Geschwind, and Mark Spigt. Measuring engagement with mental health and behavior change interventions: an integrative review of methods and instruments. *International journal of behavioral medicine*, 30(2):155–166, 2023.
- [5] Paolo Buono, Berardina De Carolis, Francesca D’Errico, Nicola Macchiarulo, and Giuseppe Palestra. Assessing student engagement from facial behavior in on-line learning. *Multimedia Tools and Applications*, 82(9):12859–12877, 2023.
- [6] Maggie A Collier, Rithika Narayan, and Henny Admoni. The sense of agency in assistive robotics using shared autonomy. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 880–888. IEEE, 2025.
- [7] Patricia Cornelio, Patrick Haggard, Kasper Hornbaek, Orestis Georgiou, Joanna Bergström, Sriram Subramanian, and Marianna Obrist. The sense of agency in emerging technologies for human–computer integration: A review. *Frontiers in Neuroscience*, 16:949138, 2022.
- [8] Francesco Del Duchetto and Marc Hanheide. Learning on the job: Long-term behavioural adaptation in human-robot interactions. *IEEE Robotics and Automation Letters*, 7(3):6934–6941, 2022.
- [9] Francesco Del Duchetto, Paul Baxter, and Marc Hanheide. Are you still with me? continuous engagement assessment from a robot’s point of view. *Frontiers in Robotics and AI*, 7:116, 2020.
- [10] Kevin Doherty and Gavin Doherty. Engagement in hci: conception, theory and measurement. *ACM computing surveys (CSUR)*, 51(5):1–39, 2018.
- [11] Anca D Dragan and Siddhartha S Srinivasa. A policy-blending formalism for shared control. *The International Journal of Robotics Research*, 32(7):790–805, 2013.
- [12] Felix Glawe, Tim Schmeckel, Philipp Brauner, and Martina Zieffle. Human autonomy and sense of agency in human-robot interaction: A systematic literature review. *arXiv preprint arXiv:2509.22271*, 2025.
- [13] Ethan K Gordon, Rajat Kumar Jenamani, Amal Nanavati, Ziang Liu, Daniel Stabile, Xilai Dai, Tapomayukh Bhattacharjee, Tyler Schrenk, Jonathan Ko, Haya Bolotski, et al. An adaptable, safe, and portable robot-assisted feeding system. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 74–76, 2024.
- [14] Bahar Irfan, Aditi Ramachandran, Samuel Spaulding, Dylan F. Glas, Iolanda Leite, and Kheng Lee Koay. Personalization in long-term human-robot interaction. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 685–686, 2019. doi: 10.1109/HRI.2019.8673076.
- [15] Shervin Javdani, Siddhartha Srinivasa, and Andrew Bagnell. Shared autonomy via hindsight optimization. In *Proceedings of Robotics: Science and Systems*, Rome, Italy, July 2015. doi: 10.15607/RSS.2015.XI.032.
- [16] Rajat Kumar Jenamani, Daniel Stabile, Ziang Liu, Abrar Anwar, Katherine Dimitropoulou, and Tapomayukh Bhattacharjee. Feel the bite: Robot-assisted inside-mouth bite transfer using robust mouth perception and physical interaction-aware control. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 313–322, 2024.
- [17] Rajat Kumar Jenamani, Priya Sundaresan, Maram Sakr, Tapomayukh Bhattacharjee, and Dorsa Sadigh. Flair: Feeding via long-horizon acquisition of realistic dishes. *arXiv preprint arXiv:2407.07561*, 2024.
- [18] Rajat Kumar Jenamani, Tom Silver, Ben Dodson, Shiqin Tong, Anthony Song, Yuting Yang, Ziang Liu, Benjamin Howe, Aimee Whitneck, and Tapomayukh Bhattacharjee. Feast: A flexible mealtime-assistance system towards in-the-wild personalization. *arXiv preprint arXiv:2506.14968*, 2025.
- [19] Michael Kelly, Chelsea Sidrane, Katherine Driggs-Campbell, and Mykel J Kochenderfer. Hg-dagger: Interactive imitation learning with human experts. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8077–8083. IEEE, 2019.
- [20] Ji Woong Kim, Juo-Tung Chen, Pascal Hansen, Lucy Xiaoyang Shi, Antony Goldenberg, Samuel Schmidgall, Paul Maria Scheikl, Anton Deguet, Brandon M White, De Ru Tsai, et al. Srt-h: A hierarchical framework for autonomous surgery via language-conditioned imitation learning. *Science robotics*, 10(104):eadt5254, 2025.
- [21] Guy Laban, Arvid Kappas, Val Morrison, and Emily S Cross. User experience of human-robot long-term interactions. In *Proceedings of the 10th international conference on human-agent interaction*, pages 287–289, 2022.
- [22] Boyi Li, Philipp Wu, Pieter Abbeel, and Jitendra Malik. Interactive task planning with language models. *arXiv preprint arXiv:2310.10645*, 2023.

- [23] Huihan Liu, Soroush Nasiriany, Lance Zhang, Zhiyao Bao, and Yuke Zhu. Robot learning on the job: Human-in-the-loop autonomy and learning during deployment. *The International Journal of Robotics Research*, 44(10-11): 1727–1742, 2025.
- [24] Kayla Matheus, Rebecca Ramnauth, Brian Scassellati, and Nicole Salomons. Long-term interactions with social robots: Trends, insights, and recommendations. *ACM Transactions on Human-Robot Interaction*, 14(3):1–42, 2025.
- [25] AJung Moon, Daniel M Troniak, Brian Gleeson, Matthew KXJ Pan, Minhua Zheng, Benjamin A Blumer, Karon MacLean, and Elizabeth A Croft. Meet me where i’m gazing: how shared attention gaze affects human-robot handover timing. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 334–341, 2014.
- [26] Jauwairia Nasir, Barbara Bruno, Mohamed Chetouani, and Pierre Dillenbourg. What if social robots look for productive engagement? automated assessment of goal-centric engagement in learning applications. *International Journal of Social Robotics*, 14(1):55–71, 2022.
- [27] Simone Nertinger, Robin Jeanne Kirschner, Abdeldjalil Naceri, and Sami Haddadin. Acceptance of remote assistive robots with and without human-in-the-loop for healthcare applications. *International Journal of Social Robotics*, 16(6):1131–1150, 2024.
- [28] Heather L O’Brien and Elaine G Toms. What is user engagement? a conceptual framework for defining user engagement with technology. *Journal of the American society for Information Science and Technology*, 59(6): 938–955, 2008.
- [29] Catharine Oertel, Ginevra Castellano, Mohamed Chetouani, Jauwairia Nasir, Mohammad Obaid, Catherine Pelachaud, and Christopher Peters. Engagement in human-agent interaction: An overview. *Frontiers in Robotics and AI*, 7:92, 2020.
- [30] Allen Z Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, et al. Robots that ask for help: Uncertainty alignment for large language model planners. *arXiv preprint arXiv:2307.01928*, 2023.
- [31] Charles Rich, Brett Ponsler, Aaron Holroyd, and Candace L Sidner. Recognizing engagement in human-robot interaction. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 375–382. IEEE, 2010.
- [32] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.
- [33] Richard M Ryan and Edward L Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist*, 55(1):68, 2000.
- [34] Jyotirmay Sanghvi, Ginevra Castellano, Iolanda Leite, André Pereira, Peter W McOwan, and Ana Paiva. Automatic analysis of affective postures and body motion to detect engagement with a game companion. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 305–312, 2011.
- [35] Pratyusha Sharma, Balakumar Sundaralingam, Valts Blukis, Chris Paxton, Tucker Hermans, Antonio Torralba, Jacob Andreas, and Dieter Fox. Correcting robot plans with natural language feedback. *arXiv preprint arXiv:2204.05186*, 2022.
- [36] Lucy Xiaoyang Shi, Zheyuan Hu, Tony Z Zhao, Archit Sharma, Karl Pertsch, Jianlan Luo, Sergey Levine, and Chelsea Finn. Yell at your robot: Improving on-the-fly from language corrections. In *Robotics: Science and Systems*, 2024.
- [37] Alessandra Sorrentino, Laura Fiorini, and Filippo Cavallo. From the definition to the automatic assessment of engagement in human–robot interaction: A systematic review. *International Journal of Social Robotics*, 16(7): 1641–1663, 2024.
- [38] Mingfei Sun, Zhenjie Zhao, and Xiaojuan Ma. Sensing and handling engagement dynamics in human-robot interaction involving peripheral computing devices. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 556–567, 2017.
- [39] Claire Yang, Heer Patel, Max Kleiman-Weiner, and Maya Cakmak. Preserving sense of agency: User preferences for robot autonomy and user control across household tasks. *arXiv preprint arXiv:2506.19202*, 2025.
- [40] Ruolin Ye, Yifei Hu, Yuhan Anjelica Bian, Luke Kulm, and Tapomayukh Bhattacharjee. Morpheus: a multimodal one-armed robot-assisted peeling system with human users in-the-loop. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9540–9547. IEEE, 2024.
- [41] Lihan Zha, Yuchen Cui, Li-Heng Lin, Minae Kwon, Montserrat Gonzalez Arenas, Andy Zeng, Fei Xia, and Dorsa Sadigh. Distilling and retrieving generalizable knowledge for robot manipulation via language corrections. In *2024 IEEE international conference on robotics and automation (ICRA)*, pages 15172–15179. IEEE, 2024.