Human-Robot Commensality: Bite Timing Prediction for Robot-Assisted Feeding in Groups

Anonymous authors

Abstract: We develop data-driven models to predict when a robot should feed during social dining scenarios. Being able to eat independently with friends and family is considered one of the most memorable and important activities for people with mobility limitations. Robots can potentially help with this activity but robot-assisted feeding is a multi-faceted problem with challenges in bite acquisition, bite timing, and bite transfer. Bite timing in particular becomes uniquely challenging in social dining scenarios due to the possibility of interrupting a social human-robot group interaction during commensality. Our key insight is that bite timing strategies that take into account the delicate balance of social cues can lead to seamless interactions during robot-assisted feeding in a social dining scenario. We approach this problem by collecting a multimodal Human-Human Commensality Dataset (HHCD) containing 30 groups of three people eating together. We use this dataset to analyze human-human commensality behaviors and develop bite timing prediction models in social dining scenarios. We also transfer these models to human-robot commensality scenarios. Our user studies show that prediction improves when our algorithm uses multimodal social signaling cues between diners to model bite timing. The HHCD dataset, videos of user studies, and code will be publicly released after acceptance.

Keywords: Multimodal Learning, HRI, Assistive Robotics, Group Dynamics

1 Introduction

Nearly 27% of people living in the United States have a disability, and close to 24 million people aged 18 years or older need assistance with activities of daily living (ADL) [1]. Key among these activities is *feeding*, which is both time-consuming for the caregiver, and challenging for the care recipient (patient) to accept socially [2]. Indeed, needing help with one or more ADLs is the most cited reason for moving to assisted or institutionalized living [3, 4]. Although there are several automated feeding systems on the market [5–13], they have lacked widespread acceptance. One of the key reasons is that all of them require manual triggering of bite timing by the user, which is challenging for users with cognitive disabilities and inconvenient in social settings. A key challenge for the realization of autonomous robotic feeding systems is therefore to infer proper bite timing [14].

While existing systems focus on solitary dining (e.g. [15–32]), **commensality**, the act of eating together, is often the practice of choice. People like to share meals with others. The social experience of a shared meal is an important part of the overall eating experience and current robot feeding systems are not designed with that experience in mind. Transferring the challenge of inferring appropriate bite timing to a social dining setting requires not only attuning to the user's eating behavior but also to the complex social dynamics of the group. For example, a robot should not attempt to feed a user who is actively engaged in conversation. Here we ask the seemingly simple question: *How should an assistive feeding robot decide the right timing for feeding a user in ever-changing and dynamic social dining scenarios?*

We developed an intelligent autonomous robot-assisted feeding system that uses multimodal sensing to feed people in dynamic social dining scenarios. We collected a novel audio-visual Human-Human Commensality Dataset (HHCD) capturing human social eating behaviors. Using this data we then trained multimodal machine learning models to predict bite timing in human-human commensality. We explored how our models trained on human-human commensality scenarios performed in a

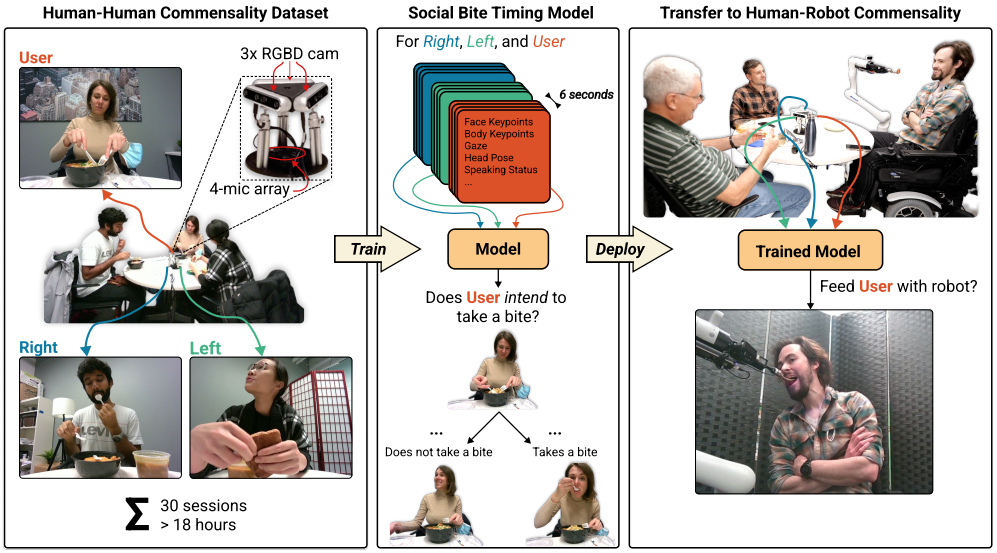

Figure 1: Our bite timing prediction workflow: **(Left)** Human-Human Commensality Dataset collection: We record audio and video of participants eating food in triads. **(Middle)** Our Social Nibbling NETwork (SoNNET) learns to predict whether a user intends to take a bite based on a 6-second window of social signals. **(Right)** We conduct a social robot-assisted feeding user study by deploying a variation of SoNNET on a robot. We refer to the **User** also as a **Target user**.

human-robot commensality setting and evaluated them in a user study. The overall workflow is shown in Fig. 1. Our results indicate that bite timing prediction improves when our model accounts for social signaling among diners, and such a model is preferred over a manual trigger and a fixed-interval trigger. Our main contributions include:

- A SOcial Nibbling NETwork (SoNNET) which captures the subtle inter-personal social dynamics in human-human and human-robot groups for predicting bite timing in social-dining scenarios.
- Methods that can successfully transfer bite timing strategies learned from human-human commensality cues to human-robot commensality situations, which we evaluate in a user study with a robot in 10 triadic human groups.
- A socially-aware robot-assisted feeding system that extends our capacity to feed people in solitary settings to groups of people sharing a meal.
- An analysis of various social and functional factors that affect human feeding behaviors during human-human commensality.
- A novel *Human-Human Commensality Dataset (HHCD)* containing multi-view RGBD video and directional audio recordings capturing 30 groups of three people sharing a meal.

2 Human-Robot Commensality

Eating is a complex process that requires the sensitive coordination of a number of motor and sensory functions. Anyone who has fed another knows that feeding, particularly *social feeding* where a person is eating or being fed in a social setting, is a delicate *dance* of multimodal signaling (via gaze, facial expressions, gestures, and speech, to name a few). Research on *commensality*, the practice of eating together, has highlighted the importance of the social nature of eating for social communion, order, health, and well-being [33]. As a consequence, digital commensality has focused on understanding the role of technology in facilitating or inhibiting the more pleasurable social aspects of dining [34].

When a person relies on assisted feeding, meals require that patient and caregiver coordinate their behavior [35]. To achieve this subtle cooperation, the people involved must be able to initiate, perceive, and interpret each other's verbal and non-verbal behavior. The main responsibility for this cooperation lies with caregivers, whose experiences, educational background, and personal beliefs may influence the course of the mealtime [36]. Our goal in this work is to understand the rhythm and timing of this *dance* to enable an automated feeding assistant to be thoughtful of when it should

feed the user in social dining settings. We introduce the concept of **Human-Robot Commensality** at the intersection of commensality and robot-assisted feeding in social group settings.

Our research is fueled by the key insight that bite timing strategies that take into account ever-changing social signals and group dynamics can lead to a seamless human-robot collaboration in social dining scenarios. Fueled by this insight, we believe a feeding device that takes the initiative and offers bites proactively during the meal at times when a bite is likely to be desired will create a more seamless dining experience than a device that requires the user to initiate bites. Herlant [37] designed an HMM to predict bite timing in dyadic robot-assisted feeding. However, her model only considered the social cues of the user. Bhattacharjee et al. [38] found users preferred less intrusive interfaces in a social dining scenario, specifically a web interface over a voice interface. Our work aims to build non-intrusive bite timing strategies by focusing on learning when to feed a user in triadic scenarios while using implicit social features from all diners.

Particularly, bite timing is important because the consequences of presenting a bite to the diner earlier than expected is poorly tolerated. This can include an interruption to conversation or to finishing chewing the prior bite. The consequences of presenting a bite later than desired can include frustration towards the robot and disruption of the natural flow of conversation during the meal. Parallels can be drawn to interruptibility research on finding the most appropriate timing to probe a user. Researchers have found that people performed best on a task if interruptions were mediated rather than timed immediately or on scheduled intervals [39, 40], often mediated based on modelling contextual and social factors [41–44].

A socially-aware robot-assisted feeding system should be designed such that if needed, the user should be able to communicate these intentions via multiple different modalities such as body language, gaze, or speech. These various modalities have been found to be effective in modelling social interactions [45–48]. Capturing these natural social interactions in computational models are likely crucial to provide accurate bite timing without distracting users from the social ambiance.

3 Problem Formulation

The objective of the bite timing prediction problem in robot-assisted feeding with a single diner is to predict the timing of *when* this user will take a bite of food by capturing their signals $\mathbf{U}$ such as voice, body gestures, head movements or speaking status. We define the proper timing for when a robot should feed as when the user *intends* to take a bite of food. It takes input signals $\mathbf{U}(t_0 : t)$ from time t_0 to time t and learns a function $\mathcal{F}(\mathbf{U})$ to predict a Boolean $y(t+h) = \mathcal{F}\left(\mathbf{U}\left(t_0 : t\right)\right)$, which indicates whether the user intends to take a bite in the time horizon h and trigger a bite transfer at time $t+1$. When a person lifts their fork off the plate to eat, they intend to take a bite of food, where this time horizon h is the time it takes to transfer the food to their mouth from their plate.

In this paper, we consider a social variant of the bite timing prediction problem where a user is interacting with two co-diners. Our goal is to predict the timing of a user to take a bite of food based on the social cues within the interaction. From an initial time t_0 to time t, the user receives social signals $\mathbf{L}(t_0 : t)$ and $\mathbf{R}(t_0 : t)$ from their left and right conversational co-diners, respectively. Given these external social signals and the target user's own history of signals $\mathbf{U}(t_0 : t)$, we aim to predict y. We note that it may not always be possible to track the same set of features for a user and their co-diners. Therefore, for some time range $k = t - t_0$ and feature dimensions n, m for the user and co-diners respectively, $\mathbf{U} \in \mathbb{R}^{k \times n}$ while $\mathbf{L}, \mathbf{R} \in \mathbb{R}^{k \times m}$, where n does not necessarily equal m. The function to learn is:

$$y(t+1) = \mathcal{F}\left(\mathbf{U}\left(t_0 : t\right), \mathbf{L}\left(t_0 : t\right), \mathbf{R}\left(t_0 : t\right)\right)$$

4 Model: SOcial Nibbling NETwork (SoNNET)

We present the SOcial Nibbling NETwork (SoNNET) that predicts when a user has the intention to eat based on various social signals. We selected features to represent both human eating and social behavior: bite features, which include the number of bites taken so far and the time since the last bite of food $b \in \mathbb{R}^2$, a diner's gaze and head pose direction $d \in \mathbb{R}^4$, binary speaking status $s \in \{0, 1\}$, and face and body keypoints $o \in \mathbb{R}^{168}$ from OpenPose [49]. We note that, in our case, the bite

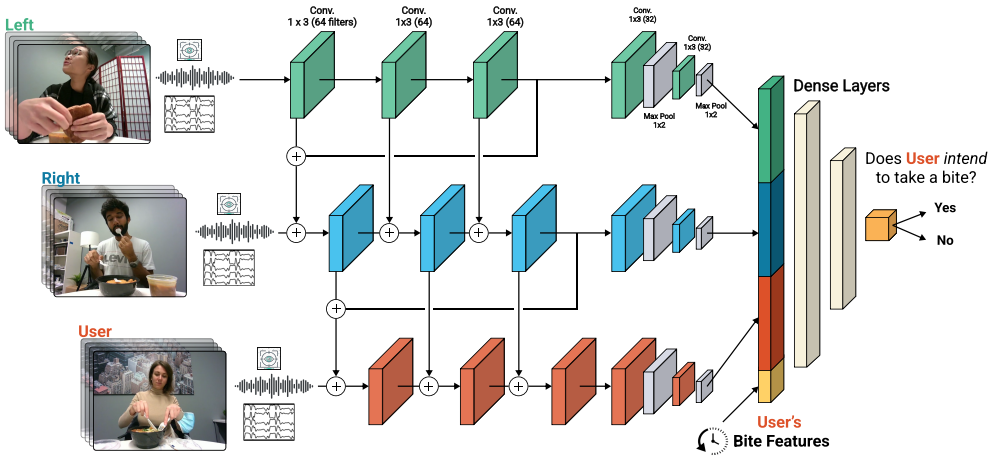

Figure 2: **Triplet-SoNNET** contains three interacting channels for features of a **target user** and two co-diners. Each channel concatenates the input of time, gaze, speaking and skeleton features from each single diner. **Couplet-SoNNET** eliminates all features from the **target user** by dropping the last channel; however, it continues to use the user's bite features. Batch normalization layers are not shown in the figure.

features b are computed only for the user and not the co-diners, since we do not estimate in real-time whether a co-diner is taking a bite of food. Thus, for a time interval $k = t - t_0$, these features are temporally stacked to construct the input signals $\mathbf{U} \in \mathbb{R}^{k \times 175}, \mathbf{L} \in \mathbb{R}^{k \times 173}, \mathbf{R} \in \mathbb{R}^{k \times 173}$ for the user, left co-diner, and right co-diner, respectively.

Recently, convolutional neural networks (CNNs) have demonstrated significant success for multi-channel time series classification from various kinds of signals [50–52]. Wu et al. [44] proposed PazNet: a multi-channel deep convolutional neural network which is able to handle inputs of different dimensions. PazNet is designed to predict the interruptibility of individual drivers. However, the information of different channels are not shared, and it lacks ability to capture social interactions among multiple people.

We design the Social Nibbling NETwork (SoNNET), a new model architecture which follows a multi-channel pattern allowing multiple interconnected branches to interleave and fuse at different stages. We create input processing channels for each diner, then add interleaving tunnels between each convolutional module and adjacent branches. The information capturing visually-observable behaviors between the diners is allowed to flow between the frames and channels. We conjecture that our model will learn a socially-coherent structure, allowing the model to implicitly represent the diners in an embedding space. Therefore, each channel has the same structure but does not share the same weight parameters. To help capture informative features, we performed dimension-reduction after the interleaving components using max pooling layers and 1×1 convolutional layers. These per-diner channels are concatenated and then followed by 2 dense layers for classification, which decides whether the user intends to feed or not. For SoNNET, the range between t and t_0 is six seconds. The social signals in this range are used to predict a user's bite intentions.

Triplet-SoNNET. For modelling the bite-timing prediction of three users with no mobility limitations, we propose Triplet-SoNNET which uses social signals from the left and right co-diners $\mathbf{L}, \mathbf{R}$ and signals from the user $\mathbf{U}$. Depicted in Fig. 2, Triplet-SoNNET ensures that the features from other co-diners $\mathbf{L}, \mathbf{R}$ interleave into the target user's features $\mathbf{U}$.

Couplet-SoNNET. To run Triplet-SoNNet in a robot-assisted feeding setting, there would be a distribution shift in the kinds of signals a target user outputs. Our goal is to feed people with mobility limitations while they are engaged in social conversations. The features from someone self-feeding are inherently different from someone using a robot-assisted feeding system. In the case of body pose, a target user with mobility limitations would be largely still, which is different from the training data. Our Human-Human Commensality Dataset consists of healthy adult diners, thus applying a trained Triplet-SoNNET model to robot-assisted feeding of a user with disabilities would be out-of-distribution. We create Couplet-SoNNET, where we ignore most social signals

from the target user by removing the last channel in Triplet-SoNNET. Therefore, the intention to feed $y(t+1) = \mathcal{F}\left(\mathbf{U}_b\left(t_0:t\right), \mathbf{L}\left(t_0:t\right), \mathbf{R}\left(t_0:t\right)\right)$, where $\mathbf{U}_b \in \mathbb{R}^{k\times 2}$ are the user's bite features for $k = t - t_0$. The user's bite features, such as the time since the last bite and the number of bites since the onset of the feeding activity, are the only social signals from the target user.

5 Human-Human Commensality Dataset (HHCD)

We introduce a novel Human-Human Commensality Dataset (HHCD) of three healthy adult participants eating in a social scenario. We used this dataset to develop models that predict a diner's intention to take a bite of food while taking into account subtle social cues. We deployed the trained models in a social robot-assisted feeding setting where one diner is fed by a robot.

Data Collection Setup. We recruited 90 people among our Institution-affiliated fully-vaccinated students, faculty, and staff to eat a meal in a triadic dining scenario. Each participant was 18+ years old and took part in the study only once. The study setup is illustrated in Fig. 1 (left). There are three cameras (mutually at 120°) in the middle of the table, each capturing one participant, and a fourth camera capturing the whole scene. All four cameras are Intel RealSense Depth Cameras D455 [53]. The scene audio was captured by a microphone array ReSpeaker Mic Array v2.0 [54] placed in the middle of the table. The ReSpeaker microphone array has four microphones arranged at the corners of a square and estimates the direction of sound arrival.

Participants were free to bring any kind of food and any utensil with them. They could also bring a drink (some drank from a cup, others from a bottle or both, with or without a straw) and were provided with napkins. Before the study, each participant was asked to fill in a pre-study questionnaire about their demographic background, relationship to other participants, and social dining habits. The experimenter then asked them to eat their meals and have natural conversations. At this point, the experimenter started the recording and left the room. When *all* three participants finished eating or after 60 minutes have passed, whichever was earlier, the experimenter stopped the recording and asked participants to fill in a post-study questionnaire about their dining experience. The specific questions asked in both pre/post-study questionnaires can be found in App. 8.1.2. The study was approved by our Institution's IRB.

Data Annotation. We annotated each participant's video based on their interactions with food, drink, and napkins. In particular, we annotated *food_entered*, *food_lifted*, *food_to_mouth*, *drink_entered*, *drink_lifted*, *drink_to_mouth*, *napkin_entered*, *napkin_lifted*, *napkin_to_mouth*, and *mouth_open* events. We chose these events as they are key transition points during feeding. We spent 151 hours annotating and used the ELAN annotation tool [55]. We assigned the annotation value $\in$ {*fork*, *knife*, *spoon*, *chopsticks*, *hand*} based on the utensil performing the food-to-mouth handover. While annotating, we also noted down per-participant food types and observations of interesting behaviors. All annotation types with detailed rules are provided in App. 8.1.3.

Data Statistics. There were 56 female and 34 male participants, and their ages ranged 18-38 ($\mu = 22$, $\sigma = 3$) years. Session durations ranged 21-55 ($\mu = 37$, $\sigma = 9$) minutes and 1 session was at breakfast, 10 at lunch, and 19 at dinner time. For additional dataset statistics, see App. 8.1.4.

For a summary of all available data in the dataset and its detailed analysis, see App. 8.1. For the purposes of this work, we only consider bite features, speaking status, gaze and head pose, and body and face keypoints.

6 Model Evaluation on Human-Human Commensality Dataset

In this section, we evaluate Triplet- and Couplet-SoNNET against other models on the HHCD. In particular, we compare against a regularized linear SVM trained with SGD to evaluate performance of linear classifiers. We also consider a Temporal Convolution Network (TCN) [56, 57], which uses causal convolutions and dilations to represent temporal data. TCNs have been found to perform better than LSTMs and GRUs on temporal anomaly detection [58] and robot food manipulation tasks [20], therefore they would provide a strong baseline to compare our models to. We also perform an ablation study to investigate the importance of various modalities. Implementation details about baseline models, SoNNET, and training procedure can be found in App. 8.2.

For training, we use 6811 *food_lifted* annotations as positive training labels since they precede an actual bite of food and indicate an intention to eat. We use a time interval of $k = t - t_0 = 6$ seconds because it takes roughly 6 seconds for the robot to move from its wait position to feeding the user. Since bite actions are sparsely distributed over time, we select 2486 6-second clips as negative samples that are in the middle of two *food_lifted* annotations. All reported models are trained with leave-one-session-out (LOSO) cross-validation to evaluate generalizability to new groups of people. Due to an issue with recording, we train over 29 sessions if speaking status features are used.

The user's bite features $b \in \mathbb{R}^2$ (time since last bite and the number of bites eaten since the start) are indicators of hunger. To ensure this feature is not dominated by higher dimensional features, we scale the size of the input by γ. This hyperparameter γ scales $b \in \mathbb{R}^2 \rightarrow b \in \mathbb{R}^{2\gamma}$. We selected $\gamma = 100$ after a grid search over the training set on the TCN and SoNNET models.

Evaluation Metrics. A high recall indicates that our model can reliably feed when it should. In contrast, a high precision indicates that a model tends to be stricter in deciding when to feed. Due to our dataset imbalance, the average accuracy across 29 sessions for a model that predicts it should always feed is 71.56%. This classifier achieves perfect recall, and relatively high precision, causing the model to have a high F1 score. It is clear that given this class imbalance, a high F1 score poorly represents the capabilities of this model. To evaluate our model effectively, we consider the normalized Matthews Correlation Coefficient (nMCC) in addition to F1 score, precision, recall, and accuracy. Unlike F1 score, nMCC considers the size of the majority and minority classes, and can only produce high scores if a classifier is able to make correct predictions for a majority of both the negative and positive classes [59]. A value of 0.5 indicates random prediction, while 0 is inverse prediction and 1 is perfect prediction.

Effects of Modality. We are interested in investigating features that are the most informative for designing a good bite timing predictor in social dining. We perform a feature ablation study on the Triplet-SoNNET model, as shown in Table 1. We selectively remove feature streams, such as body and face data from OpenPose, gaze and head features from RT-GENE, speaking status signals, and the user's bite features. We find that users' bite features such as the time since last bite and the number of bites are important, as accuracy drops drastically without them. Intuitively, we believe this feature is important because a user's bite features are a proxy for their level of hunger. We also see that without body and face features, F1 and recall slightly increase. This could be due to the fact these data streams are noisy; however, as indicated by the lower accuracy and nMCC when removing OpenPose features, these features are useful.

Effects of Model Type. Table 2 shows the outcomes of various model comparisons when trained using LOSO. We compare performance of Triplet-SoNNET against a linear SVM and TCN trained on all three diners. We call this TCN a Triplet-TCN. Triplet-TCN has all the diners' features concatenated per-timestep, and we compare this result to Triplet-SoNNET. We find that Triplet-SoNNet achieves higher accuracy and nMCC compared to all other models; however, it performs worse on recall and F1 score compared to Triplet-TCN. In our scenario, we want to ensure that the robot feeds when it should and does not feed when it should not. A bite prediction model that overfeeds or underfeeds is not ideal. A high nMCC balances the roles of recall and precision and better represents whether a classifier should or should not feed. Therefore, for our scenarios, Triplet-SoNNET is a more effective predictor of bite timing than other models trained on all three diners.

Effects of Social Scenario. We are interested in comparing the ability of models to learn social behaviors using only two co-diners' features rather than having full observability. We compare Couplet-SoNNET to a similarly-named Couplet-TCN trained on two co-diners' features and a user's bite features. As expected, Couplet-TCN and Couplet-SoNNET perform worse than their Triplet-counterparts, with Couplet-TCN being close to random prediction with an nMCC of 0.5539 while Couplet-SoNNET has an nMCC of 0.6648. We find that Couplet-SoNNET performs better than Couplet-TCN. This result reveals Couplet-SoNNET is able to understand social signals better than a predictor that always feeds. This implies that it is possible to predict the behavior of a user using only their co-diner information, which indicates that there is social coordination in human-human commensality. These findings also suggest that social signals were captured by the interleaving structure of the SoNNET models.

Table 1: Ablation study on different modalities from various data sources. We use average over LOSO cross-validation.

Method	Acc.	Prec.	Rec.	F1	nMCC
Triplet-SoNNET	**0.820**	0.861	0.871	0.862	**0.772**
- Speaking Status	0.816	**0.864**	0.863	0.856	0.771
- Gaze & Head Pose	0.815	0.863	0.863	0.856	0.769
- Bite Features	0.781	0.832	0.855	0.834	0.727
- Body & Face	0.820	0.854	**0.886**	**0.865**	0.771

Table 2: Model comparison on LOSO cross-validation over 29 sessions.

Method	Acc.	Prec.	Rec.	F1	nMCC
Always Feed	0.72	0.72	1	0.83	0.5
Linear SVM (SGD)	0.68	0.82	0.77	0.74	0.64
Triplet-TCN	0.82	0.82	**0.96**	**0.88**	0.72
Triplet-SoNNET	**0.82**	**0.86**	0.87	0.86	**0.77**
Couplet-TCN	0.73	0.73	**0.98**	0.83	0.55
Couplet-SoNNET	**0.76**	**0.78**	0.96	**0.85**	**0.66**

7 Transferring from Human-Human to Human-Robot Commensality

Our objective is to develop a bite timing strategy for a robot that feeds a user in a social dining setting. We design a study where users evaluate the effect of different bite timing strategies on their overall social dining experience. To simulate robotic caregiving scenarios for people with upper-extremity mobility limitations, we instructed users to not move their upper body. This study was approved by our Institution's IRB.

Experimental Setup. We evaluate three bite timing strategies for triggering the robot to feed a user, also depicted in Fig. 3 (middle):

1. **Learned Timing.** This social, fully autonomous bite timing strategy feeds based on our Couplet-SoNNET model's output. We sample this model every three seconds with the last six seconds of preprocessed features at a rate of 15 frames per second. This approach takes into account the social context. Since we want to evaluate the generalization performance, we train Couplet-SoNNET on 80% of the HHCD data and use the remaining 20% of HHCD data to select early-stopping criteria.
2. **Fixed-Interval Timing.** This fully autonomous bite timing strategy feeds every 44.5 seconds, which is a scaled average time a robot should take to feed a user after it has picked up a food item. To derive this value, we first find the appropriate scaling factor between human motion from the HHCD and robot motion. We note the average time for a human from the *food_entered* transition to *food_lifted* transition is 9.9 seconds. The robot end-effector motion is not designed to match the human speed but rather to be perceived as safe and comfortable to a user being fed. We find the equivalent key transitions for the robot to be $5x$ slower than a human. Since we define the intention to take a bite as when the food is lifted, the robot should take 49.5 seconds to feed a user after picking up a food item. Given the robot takes roughly 5 seconds to move to its wait position after picking the food, the robot waits 44.5 seconds.
3. **Mouth-Open Timing.** This partially autonomous bite timing strategy feeds only when the user prompts the robot by opening their mouth. The target user is prompted each time by the robot saying "When ready, look at me and open your mouth". This approach gives the user explicit control of when the robot should feed [38].

The robot user is seated on a wheelchair mounted with a Kinova Gen3 6-DoF arm [60], which is used to feed the participant (Fig. 3, left). For implementation details of the robot study, see App. 8.3.

Experimental Procedure. In this study, participants are seated in a similar setup as that used for HHCD data collection in Sec. 5. All participants were asked to bring their own food, and each group chose who would be fed by the robot. We recruited 30 participants over 10 sessions. There were 16 female and 14 male participants, and their ages ranged from 19-70 ($\mu = 27$, $\sigma = 9$) years.

A single trial consists of bite acquisition, followed by one of the three bite timing strategies, then bite transfer. For bite acquisition, the robot alternates feeding the user cantaloupes and strawberries. We chose these fruits due to their high acquisition success rates [24]. We used the bite acquisition strategies and bite transfer strategies from [25, 26]. All participants take a survey after each trial, which administers a forced-choice question on the participants' preferences between the previous and current conditions. Each pair of comparisons between any two conditions occurs three times, leading to ten trials. The condition orderings are counterbalanced over ten trials. Additionally,

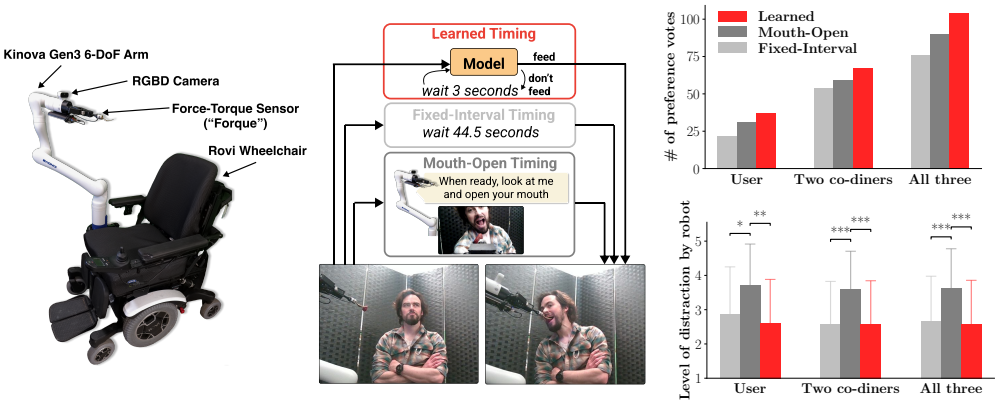

Figure 3: **Left:** We use a 6-DoF Kinova Gen3 robotic arm [60] on Rovi wheelchair [61]. **Middle:** User study conditions/bite timing strategies: Learned, Fixed-Interval, and Mouth-Open Timings. **Top right:** Preferences for bite timing strategies rated by users, two co-diners, and all three diners. **Bottom right:** Level of distraction by the robot perceived by users, two co-diners, and all three diners on a Likert scale 1-5 (agreement with "I felt distracted by the robot"), for each bite timing strategy. $*, **, ***$ denote statistically significant differences with $p_{0.05}, p_{0.005}, p_{0.0005}$ respectively.

we ask participants whether they felt the robot fed them too early, on-time, or too late. To avoid interruptions in social conversations due to the presence of a robot in human groups, we provide the participants with a list of questions (see App. 8.3.1), which they could optionally use to help get the conversation started at each trial, similarly to previous work [37]. The post-study questionnaire included questions after each trial or condition about bite timing appropriateness, distractions due to the robot, ability to have natural conversations, ability to feel comfortable around the robot, as well as system reliability and trust in the robot [62].

Results and Discussion. As shown in Fig. 3 (top right), users and co-diners preferred the Learned strategy for bite timing as compared to Fixed-Interval or Mouth-Open Timing. This confirms that our insight to incorporate social signals in model structure (SoNNET) improves bite timing prediction. These results using Couplet-SoNNET also imply that it is possible to predict the behavior of a user using only their co-diner information, which indicates that there is social coordination in human groups even in the presence of a robot. In Fig. 3 (bottom right) we further compared the level of distraction by the robot as perceived by participants. We performed Kruskal-Wallis H-tests and Tukey HSD post-hoc tests and found that Mouth-Open Timing distracts dining participants significantly more than Learned or Fixed-Interval Timing. We believe this is because the Mouth-Open strategy prompts the user using a voice interface, which can disrupt the rhythm of conversation. Even though the participants had a clear preference for the Learned strategy when given a forced-choice, when asked to individually rate the conditions using a 5-point Likert scale, interestingly we could not find any statistically significant differences between Mouth-Open Timing and Learned Timing. This is probably because the Mouth-Open Timing strategy provides full control of bite timing to the users themselves. Note, regardless of the conditions, the users found the system comfortable, reliable, and trustworthy. Detailed analysis is given in App. 8.3.

Limitations. There is a risk that our results from human-robot user studies with healthy adults may not generalize to those with people with mobility limitations. People with mobility limitations may have different preferences and cognitive workload associated with a robotic intervention. It remains to be investigated how our models perform with our target population. We also made multiple assumptions when transferring our results from human-human to human-robot commensality scenarios. During human-human commensality, the user was self-feeding whereas in human-robot commensality the user was being fed. We also assumed that the addition of a robot into a human-human commensality scenario does not change the social dynamics of the diners significantly. Given these assumptions, it would be an interesting exploration to see how our models perform when trained on similar human-robot commensality scenarios. Finally, it is an open question as to how these models would perform with groups of different cultures. This motivates further investigation into human-robot commensality, both from technical and societal perspectives.

References

[1] U. S. C. Bureau. Americans with disabilities: 2014. 2014.

[2] L. Perry. Assisted feeding. *Journal of advanced nursing*, 62(5):511–511, 2008.

[3] M. E. Mlinac and M. C. Feng. Assessment of activities of daily living, self-care, and independence. *Archives of Clinical Neuropsychology*, 31(6):506–516, 2016.

[4] Agis living. http://www.agis.com/Document/484/assisted-living-care-with-an-independent-flavor.

[5] Obi, 2018. https://meetobi.com/,[Online; Retrieved on 25th January, 2018].

[6] My spoon, 2018. https://www.secom.co.jp/english/myspoon/food.html,[Online; Retrieved on 25th January, 2018].

[7] Meal-mate, 2018. https://www.made2aid.co.uk/productprofile?productId=8&company=RBF%20Healthcare&product=Meal-Mate,[Online; Retrieved on 25th January, 2018].

[8] Meal buddy, 2018. https://www.performancehealth.com/meal-buddy-system,[Online; Retrieved on 25th January, 2018].

[9] Winsford feeder, 2018. https://www.youtube.com/watch?v=KZRFj1UZl-c,[Online; Retrieved on 15th February, 2018].

[10] Bestic, 2019. https://www.camanio.com/us/products/bestic/,[Online; Retrieved on 18th April, 2019].

[11] The mealtime partner dining system, 2018. http://mealtimepartners.com/,[Online; Retrieved on 15th February, 2018].

[12] Neater eater robot, 2019. http://www.neater.co.uk/neater-eater-2-2/,[Online; Retrieved on 18th April, 2019].

[13] Beeson automaddak feeder, 2019. https://abledata.acl.gov/product/beeson-automaddak-feeder-model-h74501,[Online; Retrieved on 18th April, 2019].

[14] T. Bhattacharjee, G. Lee, H. Song, and S. S. Srinivasa. Towards robotic feeding: Role of haptics in fork-based food manipulation. *IEEE Robotics and Automation Letters*, 2019.

[15] A. Candeias, T. Rhodes, M. Marques, M. Veloso, et al. Vision augmented robot feeding. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018.

[16] E. K. Gordon, S. Roychowdhury, T. Bhattacharjee, K. Jamieson, and S. S. Srinivasa. Leveraging post hoc context for faster learning in bandit settings with applications in robot-assisted feeding. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10528–10535. IEEE, 2021.

[17] H. Higa, K. Kurisu, and H. Uehara. A vision-based assistive robotic arm for people with severe disabilities. *Transactions on Machine Learning and Artificial Intelligence*, 2(4):12–23, 2014.

[18] T. Bhattacharjee, M. E. Cabrera, A. Caspi, M. Cakmak, and S. S. Srinivasa. A community-centered design framework for robot-assisted feeding systems. In *The 21st international ACM SIGACCESS conference on computers and accessibility*, pages 482–494, 2019.

[19] A. Jardón, C. A. Monje, and C. Balaguer. Functional evaluation of asibot: A new approach on portable robotic system for disabled people. *Applied Bionics and Biomechanics*, 9(1):85–97, 2012.

[20] T. Bhattacharjee, G. Lee, H. Song, and S. S. Srinivasa. Towards robotic feeding: Role of haptics in fork-based food manipulation. *IEEE Robotics and Automation Letters*, 4(2): 1485–1492, 2019.

[21] R. M. Alqasemi, E. J. McCaffrey, K. D. Edwards, and R. V. Dubey. Wheelchair-mounted robotic arms: Analysis, evaluation and development. In *Proceedings, 2005 IEEE/ASME International Conference on Advanced Intelligent Mechatronics.*, pages 1164–1169. IEEE, 2005.

[22] Z. Bien, M.-J. Chung, P.-H. Chang, D.-S. Kwon, D.-J. Kim, J.-S. Han, J.-H. Kim, D.-H. Kim, H.-S. Park, S.-H. Kang, et al. Integration of a rehabilitation robotic system (kares ii) with human-friendly man-machine interaction units. *Autonomous robots*, 16(2):165–191, 2004.

[23] D. Park, Y. K. Kim, Z. M. Erickson, and C. C. Kemp. Towards assistive feeding with a general-purpose mobile manipulator. *arXiv preprint arXiv:1605.07996*, 2016.

[24] R. Feng, Y. Kim, G. Lee, E. K. Gordon, M. Schmittle, S. Kumar, T. Bhattacharjee, and S. S. Srinivasa. Robot-assisted feeding: Generalizing skewering strategies across food items on a plate. In *The International Symposium of Robotics Research*, pages 427–442. Springer, 2019.

[25] D. Gallenberger, T. Bhattacharjee, Y. Kim, and S. S. Srinivasa. Transfer depends on acquisition: Analyzing manipulation strategies for robotic feeding. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 267–276. IEEE, 2019.

[26] S. Belkhale, E. K. Gordon, Y. Chen, S. Srinivasa, T. Bhattacharjee, and D. Sadigh. Balancing efficiency and comfort in robot-assisted bite transfer. *arXiv preprint arXiv:2111.11401*, 2021.

[27] G. Canal, G. Alenyà, and C. Torras. Personalization framework for adaptive robotic feeding assistance. In *International conference on social robotics*, pages 22–31. Springer, 2016.

[28] T. Rhodes and M. Veloso. Robot-driven trajectory improvement for feeding tasks. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2991–2996. IEEE, 2018.

[29] I. Naotunna, C. J. Perera, C. Sandaruwan, R. Gopura, and T. D. Lalitharatne. Meal assistance robots: A review on current status, challenges and future directions. In *2015 IEEE/SICE International Symposium on System Integration (SII)*, pages 211–216. IEEE, 2015.

[30] E. K. Gordon, X. Meng, M. Barnes, T. Bhattacharjee, and S. S. Srinivasa. Learning from failures in robot-assisted feeding: Using online learning to develop manipulation strategies for bite acquisition. 2019.

[31] D. Park, Y. Hoshi, and C. C. Kemp. A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder. *IEEE Robotics and Automation Letters*, 3(3):1544–1551, 2018.

[32] D. Park, H. Kim, Y. Hoshi, Z. Erickson, A. Kapusta, and C. C. Kemp. A multimodal execution monitor with anomaly classification for robot-assisted feeding. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5406–5413. IEEE, 2017.

[33] H. Jönsson, M. Michaud, and N. Neuman. What is commensality? a critical discussion of an expanding research field. *International Journal of Environmental Research and Public Health*, 18(12):6235, 2021.

[34] C. Spence, M. Mancini, and G. Huisman. Digital commensality: Eating and drinking in the company of technology. *Frontiers in psychology*, 10:2252, 2019.

[35] E. Athlin, A. Norberg, and K. Asplund. Caregivers' perceptions and interpretations of severely demented patients during feeding in a task assignment system. *Scandinavian Journal of Caring Sciences*, 4(4):147–156, 1990.

[36] E. Athlin and A. Norberg. Interaction between the severely demented patient and his caregiver during feeding. *Scandinavian Journal of Caring Sciences*, 1(3-4):117–123, 1987.

[37] L. V. Herlant. Algorithms, implementation, and studies on eating with a shared control robot arm. *PhD Dissertation*, 2018. URL http://www.cs.cmu.edu/afs/cs/user/lcv/www/herlant-thesis.pdf.

[38] T. Bhattacharjee, E. K. Gordon, R. Scalise, M. E. Cabrera, A. Caspi, M. Cakmak, and S. S. Srinivasa. Is more autonomy always better? exploring preferences of users with mobility impairments in robot-assisted feeding. In *2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 181–190. IEEE, 2020.

[39] D. C. McFarlane. Comparison of four primary methods for coordinating the interruption of people in human-computer interaction. *Human-Computer Interaction*, 17(1):63–139, 2002.

[40] M. Czerwinski, E. Cutrell, and E. Horvitz. Instant messaging: Effects of relevance and timing. In *People and computers XIV: Proceedings of HCI*, volume 2, pages 71–76, 2000.

[41] M. Pielot, B. Cardoso, K. Katevas, J. Serrà, A. Matic, and N. Oliver. Beyond interruptibility: Predicting opportune moments to engage mobile phone users. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3), Sept. 2017. doi:10.1145/3130956. URL https://doi.org/10.1145/3130956.

[42] Y. Takemae, T. Ohno, I. Yoda, and S. Ozawa. Estimating interruptibility in the home for remote communication based on audio-visual tracking. *IPSJ Digital Courier*, 3:125–133, 2007.

[43] S. Banerjee, A. Silva, K. Feigh, and S. Chernova. Effects of interruptibility-aware robot behavior. *arXiv preprint arXiv:1804.06383*, 2018.

[44] T. Wu, N. Martelaro, S. Stent, J. Ortiz, and W. Ju. Learning when agents can talk to drivers using the inagt dataset and multisensor fusion. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(3):1–28, 2021.

[45] J. Wang, H. Xu, M. Narasimhan, and X. Wang. Multi-person 3d motion prediction with multi-range transformers. *Advances in Neural Information Processing Systems*, 34, 2021.

[46] P. Müller, M. X. Huang, X. Zhang, and A. Bulling. Robust eye contact detection in natural multi-person interactions using gaze and speaking behaviour. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pages 1–10, 2018.

[47] H. Park, E. Jain, and Y. Sheikh. 3d social saliency from head-mounted cameras. *Advances in Neural Information Processing Systems*, 25, 2012.

[48] H. Joo, T. Simon, M. Cikara, and Y. Sheikh. Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10873–10883, 2019.

[49] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018.

[50] B. Zhao, H. Lu, S. Chen, J. Liu, and D. Wu. Convolutional neural networks for time series classification. *Journal of Systems Engineering and Electronics*, 28(1):162–169, 2017.

[51] J. Yang, M. N. Nguyen, P. P. San, X. Li, and S. Krishnaswamy. Deep convolutional neural networks on multichannel time series for human activity recognition. In *Ijcai*, volume 15, pages 3995–4001. Buenos Aires, Argentina, 2015.

[52] C.-L. Liu, W.-H. Hsaio, and Y.-C. Tu. Time series classification with multivariate convolutional neural network. *IEEE Transactions on Industrial Electronics*, 66(6): 4788–4797, 2018.

[53] RealSense. Introducing the Intel® realsense™ depth camera D455, 2020. URL https://www.intelrealsense.com/depth-camera-d455/.

[54] B. Zuo. ReSpeaker Mic Array v2.0, 2018. URL https://wiki.seeedstudio.com/ReSpeaker_Mic_Array_v2.0/.

[55] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. Elan: A professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1556–1559, 2006.

[56] C. Lea, R. Vidal, A. Reiter, and G. D. Hager. Temporal convolutional networks: A unified approach to action segmentation. In *European conference on computer vision*, pages 47–54. Springer, 2016.

[57] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017.

[58] Y. He and J. Zhao. Temporal convolutional networks for anomaly detection in time series. In *Journal of Physics: Conference Series*, volume 1213, page 042050. IOP Publishing, 2019.

[59] D. Chicco and G. Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13, 2020.

[60] Kinova gen3. https://www.kinovarobotics.com/product/gen3-robots, 2022. Accessed: 2022-06-14.

[61] Rovi wheelchair. https://www.rovimobility.com/, 2022. Accessed: 2022-06-14.

[62] J.-Y. Jian, A. M. Bisantz, and C. G. Drury. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics*, 4(1):53–71, 2000.

[63] T. Fischer, H. J. Chang, and Y. Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 334–352, 2018.

[64] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[65] T. Fischer and Y. Demiris. Markerless perspective taking for humanoid robots in unconstrained environments. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3309–3316. IEEE, 2016.

[66] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[67] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.