8 Appendix

For a video of our work, see `https://www.youtube.com/watch?v=5KLGrvrPjMc`

8.1 Human-Human Commensality Dataset (HHCD) Details

8.1.1 Summary of Available Data

Overall, the Human-Human Commensality Dataset (HHCD) contains 30 sessions, totalling over 18 hours of multistream, multimodal recordings of 90 people, and provides the following data.

- ROS bags with topics: 4x mic audio, mixed audio, sound direction, per-participant RGBD, and scene RGBD
- Raw data (extracted from ROS bags): scene audio, sound direction, per-participant videos, and scene videos
- Processed data (extracted from raw data): per-participant speaking status, per-participant face and body keypoints from OpenPose [51], per-participant gaze and head pose from RT-GENE [69], per-participant bite count, and per-participant times since last bite lifted and since last bite delivered to mouth
- Annotations: per-participant interactions with food, drink, and napkins (all entered, lifted, delivered to mouth, and mouth open events), per-participant food type labels and observations of interesting behaviors

The HHCD dataset is available at `https://emprise.cs.cornell.edu/hrcom/`

8.1.2 Data Collection Setup Measures

We set up the data collection study with the following measures, depicted in Fig. 4:

- Table diameter: 105 cm
- Distance between the ground and the top of the table: 72.5 cm
- Distance between the table center and a participant camera center: 6 cm horizontally
- Camera triangle side: 11.6 cm horizontally
- Distance between the top of the table and the center of a participant camera lens: 14 cm vertically
- Participant camera tilt: 12° above the horizontal plane
- Distance between the table center and the scene camera: 170 cm horizontally
- Distance between the ground and the scene camera: 119.5 cm
- Angle between the camera of the participant at position 1 and the scene camera: 41° in the clockwise direction (toward the participant at position 3) in horizontal plane around the table center
- Microphone array square side: 4.5 cm
- Angle between the zero degree sound direction of the microphone array and the camera of the participant at position 3: 49° in the counter-clockwise direction (toward the participant at position 1) in horizontal plane around the table center

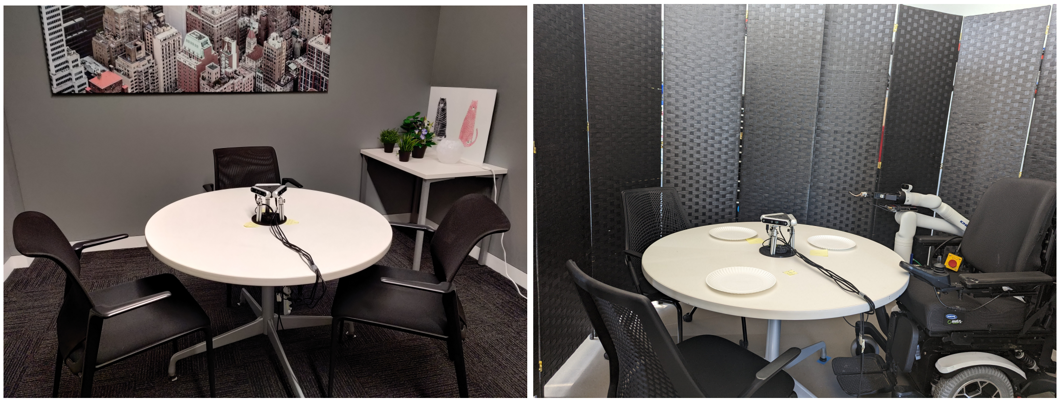


Figure 4: **Left:** Human-human commensality experimental setup, described in App. 8.1.2. **Right:** Human-robot commensality experimental setup, described in App. 8.3.

8.1.3 Questionnaires

The questions we asked the participants in the pre-study and post-study questionnaires are shown in Fig. 14 and Fig. 15 respectively.

8.1.4 Data Annotation Details

Using the ELAN annotation tool [57], we annotated each participant's video (excluding the scene videos) based on participant's interactions with food, drink, and napkins. We defined the following annotation types and associated sets of annotation values. The annotation value was assigned based on the type of utensil involved.

- ***mouth_open*** $\in \{\emptyset\}$: From the time the mouth opened due to an immediately following food-to-mouth handover until it closed. The frames where the mouth was open for other reasons were ignored. If the mouth was open even when not eating, the mouth_open annotation began when the mouth started opening more due to an incoming food item and similarly, the mouth_open annotation ended when the mouth closed the most the first time after eating the bite.
- ***food_to_mouth*** $\in$ {fork, knife, spoon, chopsticks, hand, $\emptyset$}: From the time the food item entered the mouth (i.e., got above teeth) until the given utensil/hand first lost contact with the mouth (or started moving away from mouth in case the utensil/hand did not touch the mouth). Subsequent actions (if any) to correct/fix an unsuccessful feeding attempt were ignored unless they involved a proper food item pick up. There was exactly one food_to_mouth annotation for each mouth_open annotation such that the mouth_open annotation always started before the food_to_mouth annotation but they could have ended in any order. If the food was consumed without the use of utensil/hand and the person just moved head towards the table to eat a bite, an empty annotation value was assigned.
- ***food_entered*** $\in$ {fork, knife, spoon, chopsticks, hand}: First 400 ms after the person touched/entered the food with a utensil/hand. If there were multiple such events before the next food_to_mouth annotation (e.g., the person first entered the food, then rested, and later entered the food again), only the first such event was annotated. The reason was to record the first intention to eat. Events when the utensil touched/entered the food just because it was put on top of the food to free up hands were ignored, and the food_entered annotation started once they touched/entered the food again. So there was exactly one food_entered annotation prior to each food_to_mouth annotation. However, when the person used two/more kinds of utensils/hands at the same time, the food_entered annotation was made for each utensil/hand independently and not each of them was followed by the food_to_mouth annotation of the same utensil/hand type (e.g., food entered by fork, food entered by knife, food lifted by fork, food delivered to mouth using fork but without knife). Also, when the food was grabbed by hand, there might not have been a food_entered annotation prior to each food_to_mouth annotation (e.g., when the person kept holding their food, such as a sandwich, in their hand between bites). If two/more food_entered annotations with different values overlapped, some annotations were shortened below 400 ms, as ELAN does not allow overlapping annotations within one tier.
- ***food_lifted*** $\in$ {fork, knife, spoon, chopsticks, hand, $\emptyset$}: First 400 ms after the utensil performing the food-to-mouth handover lost contact with the rest of the food or with another utensil/hand involved in food manipulation, whichever occurred later. In case the food was grabbed by hand, the first 400 ms after the food started moving towards the mouth. If there were multiple such events before the food_to_mouth annotation (e.g., the person first lifted the food item a bit, then returned it back to the rest of the food to dip it in a sauce, and later lifted it again), only the last lift off event was annotated. The reason was to record only such food lift off events that immediately led to feeding. So there was exactly one food_lifted annotation prior to each food_to_mouth annotation. However, when the person used two/more kinds of utensils/hands at the same time, the food_lifted annotation was made only for the last lift off before the food_to_mouth annotation of the same utensil/hand type (e.g., if the food was entered by fork, lifted by fork, handed over to spoon, lifted by spoon, and finally, delivered to mouth using spoon, then the fork lift off was not annotated). If the food was consumed without the use of utensil/hand and the person just moved head towards the table to eat a bite, the annotation was made when the head started moving towards the food item and the empty annotation value was used. When candies/chocolates were consumed, the food_lifted annotation was made only after the candy/chocolate was unwrapped.

- ***drink_to_mouth*** $\in$ {cup, bottle}: From the time the cup/bottle/straw touched the mouth until it left the mouth.
- ***drink_entered*** $\in$ {cup, bottle}: First 400 ms after the person grabbed the drink with their hand. If there were multiple such events before the drink_to_mouth annotation (e.g., the person first grabbed the drink, then dropped it, and later grabbed the drink again), only the first such event was annotated. The reason was to record the first intention to drink. So there was exactly one drink_entered annotation prior to each drink_to_mouth annotation unless they kept holding the drink between two drink_to_mouth annotations. Also, if the person used a bottle to pour drink into a cup, the drink_entered annotation was made for both: when they grabbed the bottle and when they grabbed the cup.
- ***drink_lifted*** $\in$ {cup, bottle}: First 400 ms after the drink lost contact with the table and started moving towards the mouth (or just started moving towards the mouth in case they kept the drink in hand after the last drink_to_mouth annotation). If there were multiple such events before the drink_to_mouth annotation (e.g., the person first moved the drink towards the mouth, then stopped a bit, and later completed the move) only the last move towards the mouth was annotated. The reason was to record only such drink lift off events that immediately led to drinking. So there was exactly one drink_lifted annotation prior to each drink_to_mouth annotation.
- ***napkin_to_mouth*** $\in \{\emptyset\}$: From the time the napkin touched the mouth until it left the mouth.
- ***napkin_entered*** $\in \{\emptyset\}$: First 400 ms after the person grabbed the napkin with their hand. If there were multiple such events before the napkin_to_mouth annotation (e.g., the person first grabbed the napkin, then dropped it, and later grabbed the napkin again), only the first such event was annotated. The reason was to record the first intention to use the napkin. So there was exactly one napkin_entered annotation prior to each napkin_to_mouth annotation unless they kept holding the napkin between two napkin_to_mouth annotations.
- ***napkin_lifted*** $\in \{\emptyset\}$: First 400 ms after the napkin lost contact with the table and started moving towards the mouth (or just started moving towards the mouth in case they kept the napkin in hand after the last napkin_to_mouth annotation). If there were multiple such events before the napkin_to_mouth annotation (e.g., the person first moved the napkin towards the mouth, then stopped a bit, and later completed the move) only the last move towards the mouth was annotated. The reason was to record only such napkin lift off events that immediately led to its use. So there was exactly one napkin_lifted annotation prior to each napkin_to_mouth annotation.
- ***disruption*** $\in$ {light_off, participant_left}: From the time the recording became disrupted due to the light turning off or due to a participant leaving the room until the normal conditions were restored.

We further defined the following additional annotation rules:

- When people were just unpacking their food/drink or loading their plates from shared bowls/containers
  - No food/drink_entered and food/drink_lifted associated annotations
  - Reason: we are not researching the preparation phase prior to eating
- When people tore their food (e.g., a piece of bread)
  - No additional food_entered annotations when the other hand touches the food
  - Reason: we consider tearing the food as a part of the food manipulation that follows the most recent food_entered annotation and precedes the food_lifted annotation
- When people licked their empty utensil/fingers/hands or foils (e.g., yogurt lid)
  - No food/drink_entered, food/drink_lifted, food/drink_to_mouth, and mouth_open associated annotations
  - Reason: there is no food/drink consumed
- When people smelled their food/drink
  - No food/drink_entered, food/drink_lifted, food/drink_to_mouth, and mouth_open associated annotations
  - Reason: there is no food/drink consumed

- When people used a napkin for anything else than cleaning their mouth (e.g., blowing/swiping their nose, cleaning their hands/eyes/utensil/table)
  - No napkin_entered, napkin_lifted, napkin_to_mouth associated annotations, but if the person cleaned their hands and then suddenly decided to clean their mouth, then the napkin_lifted annotation was made when the napkin started to move towards mouth and also the napkin_to_mouth annotation was made. If the initial intention to pick up the napkin seemed to be to eventually clean the mouth, then also the napkin_entered annotation was made.
  - Reason: blowing/swiping nose and cleaning hands/etc. is not directly related to eating/drinking
- When people picked up a napkin from their lap
  - No napkin_entered associated annotation
  - Reason: the napkin was most likely entered earlier and just put on their lap
- When people grabbed the bottle only to close it or read its label
  - No drink_entered, nor drink_lifted associated annotations
  - Reason: there is no drink consumed
- When the food/drink/napkin_to_mouth or mouth_open event was already in progress at the beginning of the video
  - No food/drink/napkin_to_mouth, and mouth_open associated annotation
  - Reason: we are not able to determine the beginning of such an event
- When there was a disruption (light went off or participant left)
  - No other annotations (besides the disruption annotation) during the disruption interval. New *_entered and *_lifted annotations had to be made after the disruption (i.e., any *_entered and *_lifted annotations from before the disruption occurred were forgotten).
  - Reason: the data from the disrupted interval are not used and the disruption is considered as a reset
- When people used coffee stirrer sticks to put spread/jam on a piece of bread
  - All the associated events were annotated with the ”knife” annotation value
- When people drank soup (e.g., from a cup)
  - The food_entered/lifted/to_mouth and mouth_open annotations were used with the ”hand” annotation value.
- When people drank from the bottle cap
  - All the associated events were annotated with the ”cup” annotation value
- When people grabbed or lifted the food/drink/napkin outside of the camera view
  - The start of the associated annotation was estimated but the annotation was not skipped
- When people picked up and ate small food items such as crumbs
  - The food_entered/lifted/to_mouth and mouth_open annotations were not skipped
- When people ate a sandwich/wrap and decided to pick a small piece with fingers from the rest of the sandwich
  - The food_entered/lifted/to_mouth and mouth_open annotations were not skipped
- When the food entered the mouth but the person did not take a bite
  - The food_entered/lifted/to_mouth and mouth_open annotations were not skipped
- When there was an incomplete (*_entered, *_lifted, *_to_mouth) sequence at the beginning or end of the video
  - For example, the first annotation could be food_lifted, mouth open or food_to_mouth without prior food_entered. Similarly, the last annotation could be food_entered or food_lifted.
- When the feeding failed at the mouth (e.g., even if the whole food item falls down during the food-to-mouth handover)
  - The food_entered/lifted/to_mouth and mouth_open annotations were not skipped

Table 3: HHCD: Annotation counts by annotation type.

Annotation type	Count
mouth_open	6,834
food_entered	6,000
food_lifted	6,830
food_to_mouth	6,834
drink_entered	755
drink_lifted	981
drink_to_mouth	978
napkin_entered	380
napkin_lifted	600
napkin_to_mouth	598
disruption	16
Total	30,806

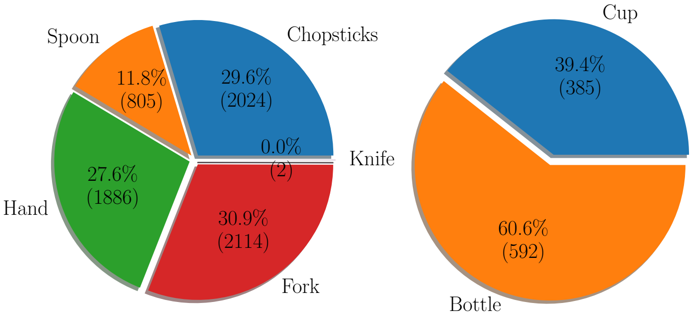


Figure 5: HHCD: Distribution of annotations by annotation value. **Left:** Distribution of *food_to_mouth* annotations. **Right:** Distribution of *drink_to_mouth* annotations. All annotations of interactions with napkin have an empty annotation value. The *disruption* annotations include one annotation of participant leaving the room for a while and 15 annotations of light turning off for a bit.

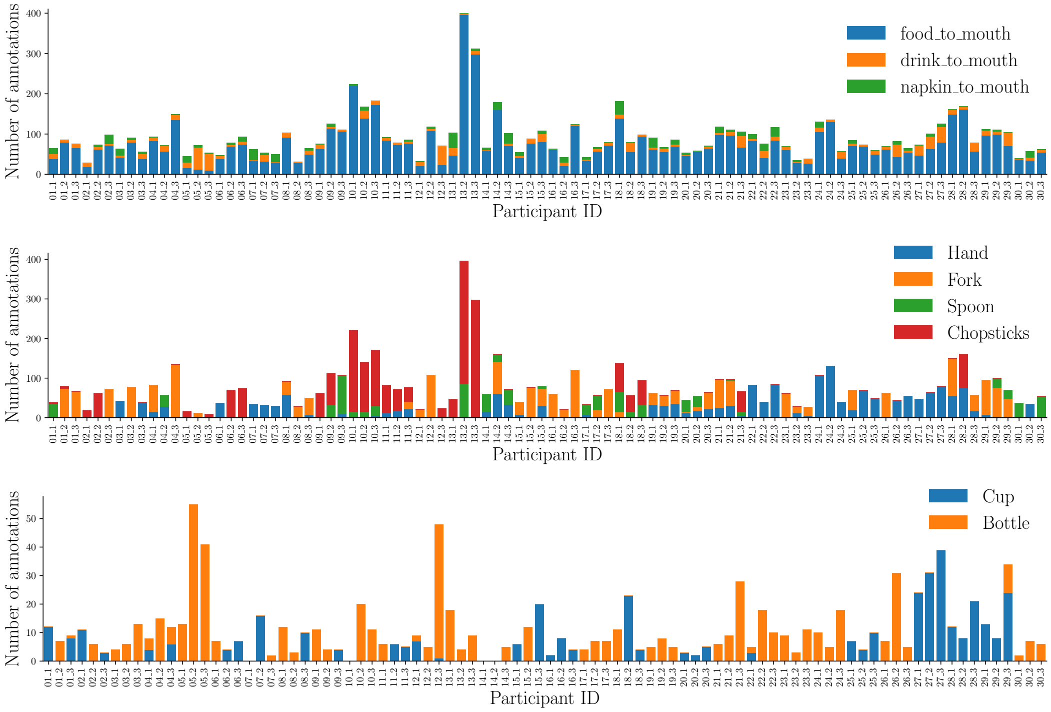


Figure 6: HHCD: Distribution of annotations across participants/videos. **Top:** Distribution of annotations by annotation type. **Middle** Distribution of *food_to_mouth* annotations by annotation value. **Bottom:** Distribution of *drink_to_mouth* annotations by annotation value. Participant ID is encoded as {session-number}_{participant-position}.

8.1.5 Additional Data Statistics

Annotation counts. The summary of all annotation counts by annotation type is provided in Tab. 3 and the distribution of annotations by annotation value is shown in Fig. 5. Figure 6 further shows the distribution of annotations by types and values across participants/videos.

Annotation durations. Means and standard deviations of annotation durations by annotation type and annotation value are shown in Tab. 4.

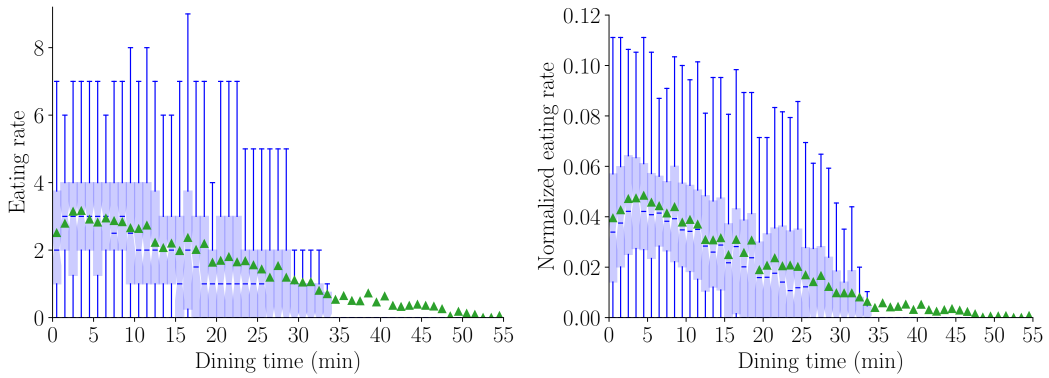


Figure 7: HHCD: Eating rate during dining. **Left:** Eating rate: number of eating actions per minute. **Right:** Normalized eating rate: number of eating actions per minute normalized by the total number of eating actions the diner made. One eating action corresponds to one *food_to_mouth* annotation.

Time gaps between annotations. In Tab. 5, we report mean and standard deviation of duration (time gap) between two consequent annotations of both the same annotation type (e.g., from *food_lifted* to *food_lifted*) as well as different annotation type (e.g., from *food_lifted* to *food_to_mouth*). We aggregate the times by annotation type and annotation value.

Eating rate during dining. Figure 7 (left) shows the eating rate (number of eating actions per minute) where one eating action corresponds to one *food_to_mouth* annotation. Since the number of eating actions might vary based on the total amount of food the diner had (and hence total number of eating actions they made), in Fig. 7 (right) we also normalize the eating rate by the total number of eating actions the diner made. As we can see in both cases the eating rate increases from the start till around the 5th minute of dining time and decreases thereafter. This confirms the eating is a non-stationary activity and needs to be accounted for when designing models of commensality.

Food types. The distribution of types of food the participants ate can be found in Fig. 8.

Demographic background. 82 participants were right-handed and 8 left-handed. The distribution of participants' race is shown in Fig. 9 (left).

Relationship between diners. The distributions of co-diner relationship types, durations, and frequency of eating together are provided in Fig. 10 (left top-bottom) respectively.

Social dining habits. The distributions of participants' typical co-diner type, social dining frequency, and dining location are shown in Fig. 10 (right top-bottom) respectively.

Dining experience. Participants' ratings of their overall meal experience, social interactions with other participants, and food are presented in Fig. 9 (right).

Table 4: HHCD: Annotation durations (mean ± std) by annotation type and annotation value. We report only variable-length annotation types and exclude *disruption*. Cell colors (yellow–red) correspond to mean annotation duration on a scale 0.7–3.0 seconds.

Annotation type	Duration (s)
mouth_open	1.2 ± 0.9
food_to_mouth	0.9 ± 0.8
drink_to_mouth	2.9 ± 1.8
napkin_to_mouth	1.6 ± 1.8

Annotation value	Duration (s)
Chopsticks	0.8 ± 0.6
Spoon	1.0 ± 0.4
Hand	1.9 ± 1.2
Fork	0.9 ± 0.6
Chopsticks	0.7 ± 0.4
Spoon	0.9 ± 0.5
Hand	1.5 ± 1.3
Fork	0.7 ± 0.4
Bottle	3.0 ± 1.6
Cup	2.7 ± 2.0

Table 5: HHCD: Time gaps (mean ± std) between two consequent annotations of the same annotation type **(left)** and of a different annotation type **(right)**. Aggregated by annotation type and annotation value. The *disruption* annotation type is excluded. Cell colors (yellow–red) correspond to mean time gap between annotations on a scale 18.9–196.3 seconds (left) and 0.3–11.3 seconds (right).

Annotation type	Ann. value	Time gap (s)
mouth_open	*All*	23.5 ± 39.8
	Chopsticks	18.9 ± 34.6
	Spoon	27.9 ± 48.6
	Hand	26.6 ± 41.5
	Fork	23.6 ± 38.7
food_entered	*All*	26.5 ± 47.2
	Chopsticks	19.2 ± 34.0
	Spoon	27.4 ± 51.8
	Hand	47.1 ± 71.8
	Fork	24.7 ± 40.2
food_lifted	*All*	23.6 ± 39.8
	Chopsticks	18.9 ± 34.8
	Spoon	28.0 ± 48.6
	Hand	26.6 ± 41.4
	Fork	23.6 ± 38.7
food_to_mouth	*All*	23.5 ± 39.8
	Chopsticks	18.9 ± 34.6
	Spoon	27.9 ± 48.6
	Hand	26.6 ± 41.5
	Fork	23.6 ± 38.7
drink_entered	*All*	192.4 ± 222.1
	Bottle	196.3 ± 206.2
	Cup	187.2 ± 241.5
drink_lifted	*All*	144.3 ± 204.3
	Bottle	138.0 ± 188.9
	Cup	154.1 ± 225.8
drink_to_mouth	*All*	143.8 ± 204.8
	Bottle	137.1 ± 189.4
	Cup	154.2 ± 226.3
napkin_entered	∅	184.0 ± 253.5
napkin_lifted	∅	134.1 ± 209.5
napkin_to_mouth	∅	132.6 ± 206.2

Annotation sequence	Ann. value	Time gap (s)
food_entered ↓ food_lifted	*All*	9.9 ± 27.3
	Chopsticks	8.9 ± 24.0
	Spoon	10.0 ± 19.5
	Hand	9.9 ± 28.5
	Fork	10.8 ± 31.9
food_lifted ↓ food_to_mouth	*All*	1.8 ± 4.0
	Chopsticks	1.3 ± 2.1
	Spoon	1.5 ± 2.6
	Hand	1.9 ± 4.3
	Fork	2.3 ± 5.2
mouth_open ↓ food_to_mouth	*All*	0.3 ± 0.2
	Chopsticks	0.3 ± 0.1
	Spoon	0.3 ± 0.1
	Hand	0.3 ± 0.2
	Fork	0.3 ± 0.2
drink_entered ↓ drink_lifted	*All*	9.1 ± 37.1
	Bottle	7.3 ± 32.7
	Cup	11.3 ± 41.8
drink_lifted ↓ drink_to_mouth	*All*	4.2 ± 8.9
	Bottle	5.1 ± 10.5
	Cup	2.9 ± 5.2
napkin_entered ↓ napkin_lifted	∅	3.0 ± 24.0
napkin_lifted ↓ napkin_to_mouth	∅	1.5 ± 2.0

Replies to open-ended post-study questions. We also analyze the study participants' answers to open-ended questions in the post-study questionnaire (Fig. 15 (right)). We observe the following patterns.

∗ *When participants think it is appropriate to take a bite of food when they are eating with others*

- **Talking-related rules:** *”When I am not speaking”, ”When listening to others”, ”When others are talking or if there is a pause in the conversation”, ”After sharing a long piece of speech and expecting a lot of response”, ”When someone else is talking and I don't think they're going to ask me anything”, ”It is appropriate when someone is not talking about a very serious topic you need to give your full attention to.”*
- **Eye gaze-related rules:** *”When the person talking is not making eye contact”, ”. . . when i'm not making direct eye contact with someone, . . . ”*
- **Diner physical state-related rules:** *”when you are hungry, it should be ok to take a bite of food.”*
- **Social interaction-related rules:** *”. . . when other people are taking a bite too”, ”It is appropiate when my bite it is at the same time when the others are putting food in their mouth. . . . ”, ”. . . when two other people are having a subconversation that I am less engaged in”*
- **Time-related rules:** *”every 10 seconds or so, . . . ”, ”. . . when a lot of time has passed between your previous bite”*
- Several participants also replied with *”whenever i want”* or similarly.

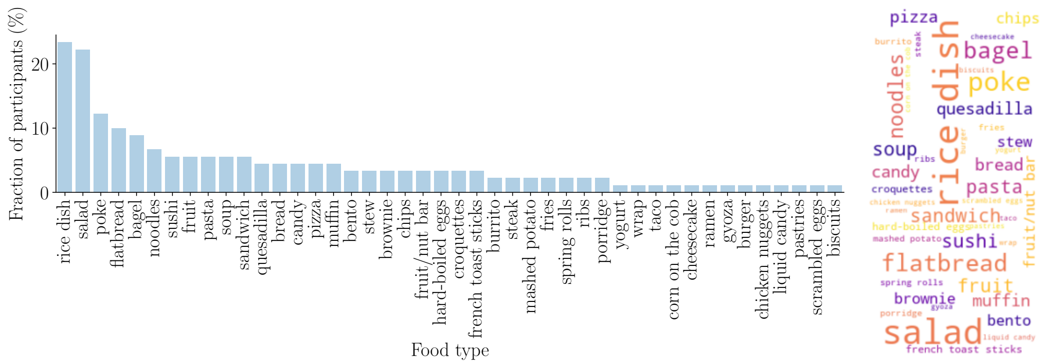


Figure 8: HHCD: Distribution of types of food the participants ate. Some participants ate multiple types of food.

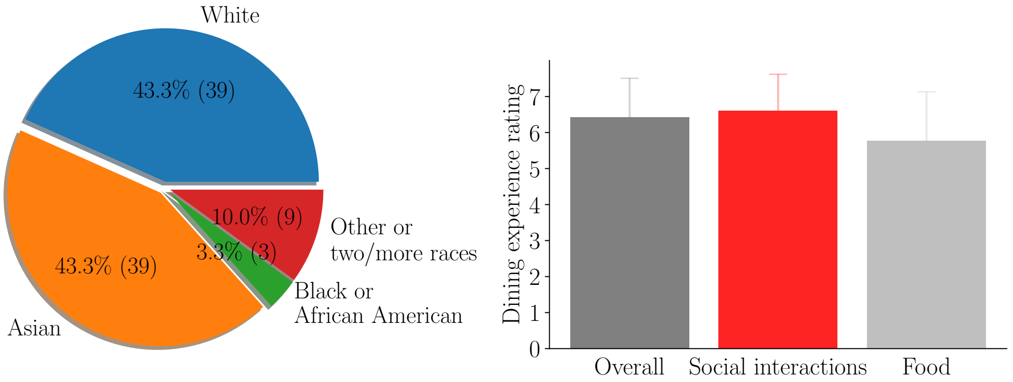


Figure 9: HHCD: **Left:** Distribution of participants' race. *"Other or two/more races"* includes three White-Asians, three White-Hispanics/Latinos, two Latinos, and one Asian-Hispanic. **Right:** Participants' ratings of their overall meal experience, social interactions with other participants, and food on a Likert scale 1-7 (Strongly disagree - Strongly agree with positive experience).

Note, the replies to the bite timing questions align with choices of modalities and features we use for bite timing prediction.

∗ *What participants liked about the meal experience*

- Most participants liked **food, conversation, and time spent with friends**. For example, *"It was super interactive and I got to know my friends better"*
- **Research contribution:** *"Time with friends, spicy foods, contributing to research", "fun experience to help the robots take over the world"*
- **The study environment:** *"It felt comfortable and natural and the food was yummy.", "Felt like a natural interaction", "I enjoyed the food, being able to soley talk to my friends without distraction", "food was good, after getting used to cameras conversation felt pretty natural"*

∗ *What participants did not like about the meal experience*

- 18 participants (20%) replied *"nothing"* or similarly.
- **Complaints about food they brought:** *"We didn't buy enough food.", "i ate too much and my stomach hurts", "one friend talked too much, it was a bit long, I ordered too much food and did not eat all of it.", "The pizza was slightly cold and i ordered the wrong pizza from domino pizza company."*
- **The study room and the recording setup:** *"I think I would rather be in a more comfortable chair and have lower lighting", "The room was too quiet for my comfort", "It was in an enclosed room. The physical setting didn't feel natural.", "A little conscious of the camera", "It was a little odd to be monitored the whole time", "I was nervous speaking about somethings because it was recorded", "The camera directly in our faces", "Not much!*

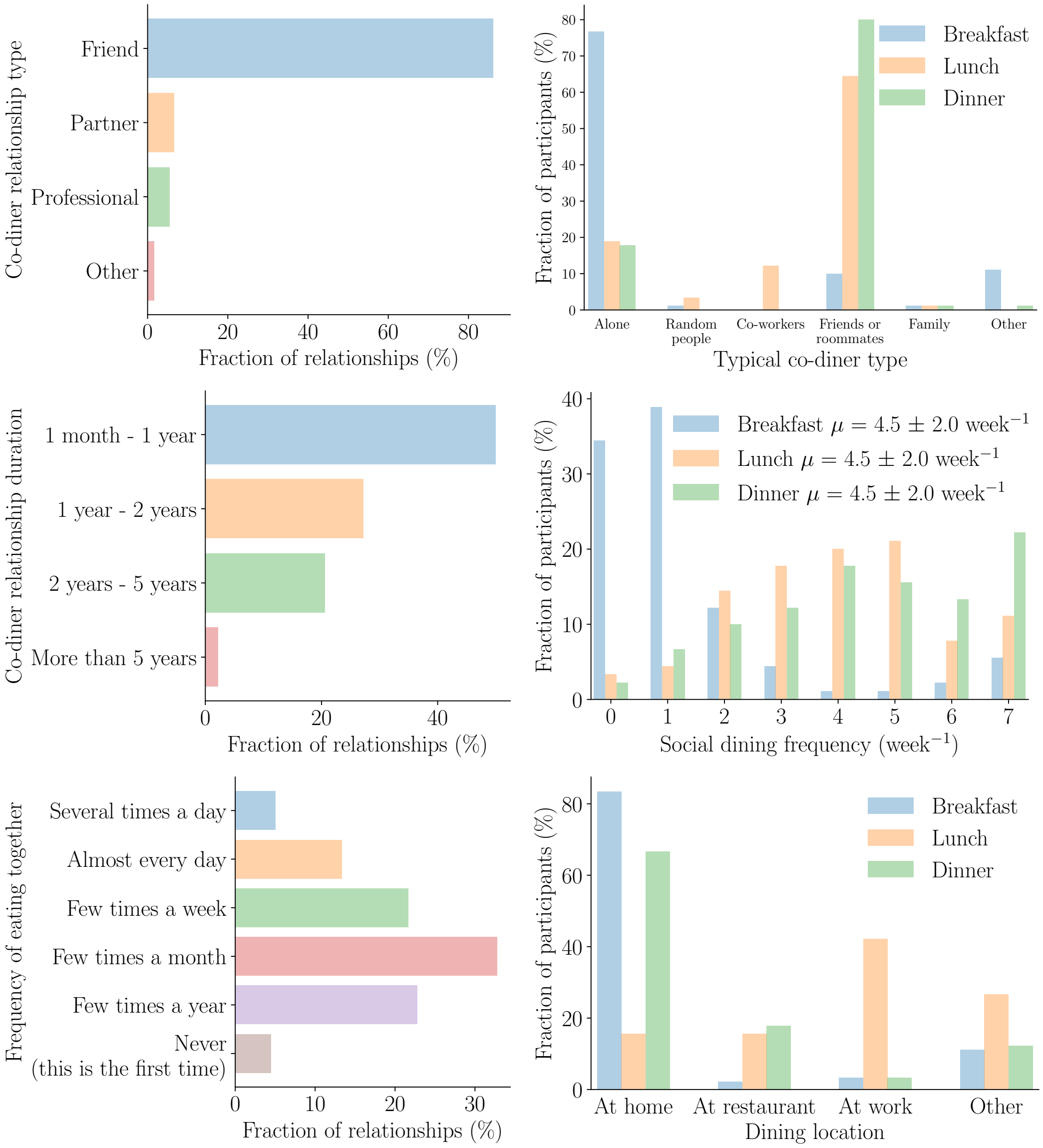


Figure 10: HHCD: **Left top-bottom:** Distributions of co-diner relationship types, durations, and frequency of eating together. *"Other"* co-diner relationship type includes two acquaintances and one boyfriend. **Right top-bottom:** Distributions of participants' typical co-diner type, social dining frequency, and dining location. *"Other"* typical co-diner type includes partner or n/a for skipped breakfasts. *"Other"* typical dining locations include dining hall, campus, n/a for skipped breakfast, or a friends' place for dinner.

Cameras in the middle of the table made it slightly more awkward to pass food, I guess.", "Maybe the fact that we were being watched, recorded; felt a like bit performative", "I did not like that I felt that I had to lean backwards to fit in the frame", "We definitely knew and acted like we were being recorded at times", "I think just because we were participating in the study but I didn't feel uncomfortable with the cameras or anything. So, I feel like our dinner was still authentic."

- **Conversation topics:** *"Participant 2 was talking too much about politics that were boring."*
- **Dining duration:** *"The amount of time I spent could have been longer to have more of a conversation"*
- **Eating with others:** *"I did not choose the food we ordered and I didn't enjoy the food very much, and I get embarrassed eating around others"*

- **Use of mobile devices while eating:** *"some things that i dont like about the meal is sometimes people tend to still like using their devices, which makes it feel like they dont want to be there eating a meal with you"*

8.2 Human-Human Commensality Model Experiments

8.2.1 Feature Extraction

We utilize several feature extraction techniques to obtain various high-level features that might indicate semantic visual and audio cues. We combine these features from each person and align target user with two co-diners for each sample event.

1. **Visual features:** Video clips from cameras facing diners explicitly capture dining behaviours and social interactions. We estimate people's body, hand and face skeletons using OpenPose [51] across consecutive frames. Each frame at time t contains body gesture and face representation as a 168-dimensional vector $o \in \mathbb{R}^{168}$. Gaze plays a crucial role during communications and interactions. It is a predictor of participants' interests in human-robot interactions [70]. We extract participants' gaze and head pose directions using Real-Time Eye Gaze and Blink Estimation in Natural Environments (RT-GENE) [69]. Gaze and head pose direction data points are represented by Euler angles θ and ϕ, and together form the feature $d \in \mathbb{R}^4$.
2. **Audio features:** Using ReSpeaker Mic Array v2.0 [56] we extract raw audio (a mixture of three diners' voices) and a sound direction channel from ROS messages. We use ROS messages as they can be naturally transferred to a robot. We align these ROS messages to video frames using nearest neighbor based on the video frame and audio message timestamps. There can be repeated audio frames aligned to video frames due to audio messages varying in speed. For each aligned audio message, we apply voice activity detection using WebRTC VAD [71] and use k-means clustering on the sound direction information to localize speakers in the scene. We combine the directional clusters and detected voice activity to create a binary vector that indicates whether a diner is speaking or not at each video frame. We also refer to this binary feature as speaking status.
3. **Temporal features:** Upon analyzing eating rate in HHCD (App. 8.1.5), we notice that the participant's eating rate increases a bit at the beginning and then decreases as the dining comes to the end. Therefore, we believe that explicitly providing the model with time and bite count information can better capture the non-stationary nature of commensality. We thus generate two bite features $b \in \mathbb{R}^2$, which indicate the time since the last bite of food was taken and the number of bites a person has consumed during the eating session.

8.2.2 Design Choices for Couplet-SoNNET

We chose to restrict the features of the user in Couplet-SoNNET due to a distribution shift between human-human commensality and human-robot commensality. We chose to remove most social signals from the user because it would be more generalizable across our target population. Whether a user is talking could be relevant to predicting bite timing. During preliminary testing however, we found that modeling the user's speaking status led to the model never feeding at all if they kept talking. This makes sense as talking is highly correlated with not-feeding in HHCD. Since the user is not self-feeding, they are not incentivized to stop talking. Therefore, we believe some level of coercion is required to ensure the user is fed, which we realize is a common subtle practice when we spoke with the caregivers who feed care recipients. By removing the user's speaking status, we can ensure that feeding does occur.

8.2.3 Implementation Details of the SoNNET

Both Triplet-SoNNET and Couplet-SoNNEt are trained using an Adam optimizer with a learning rate of 0.0001 and a batch size of 128. To prevent overfitting, we early stop if the validation loss does not increase after 10 epochs. The number of filters at each convolutional layer can be seen in Fig. 2. We use batch normalization layers after each convolutional layer. All experiments are performed on a 64-core cluster with five NVIDIA RTX 3090s.

8.2.4 Implementation Details of the Baseline Models

We use the Keras TCN implementation [72] and train the Triplet-TCN and Couplet-TCN using the same hyperparameters as the SoNNET models. We set the filter size to 50, which ensures a similar number of learnable parameters as the SoNNET models. In the case of the Triplet-TCN, we simply concatenate all the features of all three participants, while for the Couplet-TCN, we use features of the co-diners and only the bite features of the User.

8.3 Human-Robot Commensality (HRCom) Experiments

8.3.1 Study Design Rationale

We considered various experimental designs for our user study. Our experiment design is a within-subjects repeated-measures design where the conditions are counterbalanced such that A→B and B→A occur a total of 3 times and there is only 1 bite per condition at one time. This helps mitigate the recency tendency and guarantees that within each session, there is a tie-breaker. Within one session, there are 9 forced-choice comparisons from 10 trials. Across 10 sessions we ensure that each ordered pair occurs an equal number of times. This gives us a total of 30 comparisons, 15 A→B and 15 B→A (and similarly for BC, AC). With these forced-choice questions, this study design is generally better for eliciting preference data [73] and has less recency bias [74] than alternative study designs where conditions are presented repeatedly at a time. For example, an alternative study design could have X bites of condition A, followed by X bites of condition B, and X bites of condition C with a preference question at the end. This alternative study design would exacerbate the recency effect, as the participants would have to remember what they felt several bites ago. We also considered other similar alternative study designs but settled with the one we presented as we believe this design would represent people's preferences in a sample-efficient manner with less recency biases.

8.3.2 Bite Timing Strategy Details

We designed our three bite timing strategies (Learned Timing, Fixed-Interval Timing, and Mouth-Open Timing) based on discussions with care recipients, occupational therapists, and caregivers.

While consulting people with mobility limitations, caregivers, and occupational therapists on what features we should look at based on what movements are consistent across people with mobility limitations. Our target users (with C3-C5 SCI) cannot move their arms to feed themselves. Also, there is a huge spectrum of severity of mobility limitations depending on the users' conditions, and their movements are not consistent across these users. Therefore, the Learned Timing strategy uses Couplet-SoNNET which does not use arm gesture features but uses only the features of the co-diners to make it more generalizable across this target population.

For the Mouth-Open Timing, caregivers said that they estimate the appropriate time to feed when care recipients open their mouth and provide an explicit cue. This directly inspired the design of our Mouth-Open Timing. As described in App. 8.3.1, our study design mitigates the recency effect and ensures useful comparisons between conditions. Since a new condition is presented after each bite, we use a speaker to prompt the user to open their mouth when they are ready.

To decide on the wait time for the Fixed-Interval Timing, we used data from HHCD to find a user-inspired wait-time. We found that a human on average takes 1.8s from lifting a food item off a plate / bowl (food_lifted) to bringing it to the mouth (food_to_mouth). The robot's equivalent approach duration is on average around 9 seconds (taking into account the variable motion planning time). Though the robot could mechanically move at a faster speed, we chose the speed that would feel safe and comfortable to a user when they are being approached (fed) by a robot with a fork. We determined this velocity of our robot to be perceived as safe and comfortable based on [40] which explicitly did a study on what approach speeds are preferred by people with mobility limitations. This scaling factor (9s / 1.8s = 5) between robot speeds and human speeds is thus user-inspired. Once we determined this scaling factor, we use this same scaling factor to scale up the bite timing from HHCD (9.9s) to human-robot commensality (9.9s*5 - 5s [for robot bite acquisition to bite-timing waiting position] = 44.5s) to make sure that the proportion of time for different phases of feeding (bite acquisition - bite timing - bite transfer) are all proportional and balanced. We would also like to note that although the average time for "food_entered → food_lifted" was 9.9s in HHCD, the standard deviation was 27.3s. So a wait-time of 44.5s is roughly 1 standard deviation away from the equivalent annotation in the HHCD data.

8.3.3 Experimental Setup Details

The experiment was set up similarly the human-human commensality dataset collection described App. 8.1.2 and is depicted in Fig. 4. For the Learned timing, RT-GENE [69] and OpenPose [51] need to process video streams from all three cameras in real-time, in addition to the robot's planning and perception stack. We thus distribute compute over two machines: a 24-core PC with an NVIDIA

RTX 3060 and a 32-core PC with an NVIDIA RTX 3090. We downsample the 30 FPS video streams to 15 FPS to ensure real-time performance.

As noted in the formulation of the Fixed-Interval timing strategy (Sec. 7), the robot is $5x$ slower during feeding as compared to a human. This means there is a distribution shift in the time since the last bite was taken on the robot compared to the training data for Couplet-SoNNET. To mitigate this distribution shift, we scale down the computed time since the last bite during the user study by a factor of 5.

The robot used joint space velocity control. The robot's motion was generated from different calls to a library of planners available to our platform:

- `planToConfiguration(goal_config)`: plans from current configuration to a joint space goal configuration (6 degrees-of-freedom)
- `planToTaskSpaceRegion(ee_goal_pose, constraints)`: plans from current configuration to a task space end-effector (EE) goal pose with some given constraints [1]
- `planToEEOffset(ee_offset)`: plans such that the end-effector moves in the direction of a certain vector.

It is paramount for the our robot platform to ensure safety to the user, so we familiarized participants to the four levels of safety we designed:

1. We placed a conservative collision model around the user's head. The users in our study were familiar with the general workspace of the robot (we moved the robot while they were seated on the chair as a part of the pre-study familiarization procedure).
2. The fork has a Force/Torque sensor attached to it, where if a certain threshold of force is reached (beyond acceptable safety / comfortable force thresholds), the arm stops immediately.
3. We had an observer watch the experiment while the emergency stop was ready to press in the case of unexpected behaviors. Additionally, an experimenter watched the system and was ready to stop it for safety.
4. The compliant robot arm is also set up so that the user can stop it if absolutely necessary. We also designed the speed of the robot to be at comfortable levels.

8.3.4 Experimental Procedure Details

Each participant was compensated for each hour of their time and for their food expenses. All participants were instructed to bring their own food. The user who was fed by the robot only ate fruits. Each of the other two participants could choose if they also want to eat fruits during the study or the food they brought.

Before starting the study, we familiarized the participants with robot-assisted feeding by showing them a trial of the Mouth-Open condition and shortened Fixed-Interval condition, along with explaining safety measures. The procedure than continues as described in Sec. 7.

8.3.5 Conversation Starters

List of questions that the user study participants could optionally use to help get the conversation started at each trial, similar to the past work [39].

- What are you studying?
- Who is your favorite singer and why?
- What is your favorite food and why?
- What is your favorite color and why?
- Do you give back or volunteer with any organizations?
- What are your favorite writers and books?
- Do you have any pets and if so, what are they?
- What sports do you play or watch and why?
- What is your favorite movie and why?
- Who is your favorite actor and why?
- Which languages do you speak and which ones do you want to learn?
- What was your favorite vacation?
- What are your hobbies?

8.3.6 Questionnaires

The questions we asked the participants in the pre-study questionnaire included all the questions asked during data collection (Fig. 14) and an additional question about the participant's level of hunger (Fig. 16 (a)). The questions in the experiment questionnaire we asked after each trial are shown in Fig. 16 (b) and the final post-study questionnaire at the very end of the study is shown in Fig. 16 (c).

8.3.7 Additional Results

Bite timing. Besides the forced-choice assessment of bite timing strategies in terms of bite timing in Sec. 7, we also evaluate absolute ratings of "how timely" each trial was. In fact, one robot user reported that *"Slight timing changes seemed more noticeable than I expected."* As we can see from Fig. 11 (left), the only statistically significant differences are with respect to Fixed-Interval timing suggesting that the user as well as all three diners found this strategy feeds rather late compared to other strategy/ies. It might be interesting to further evaluate whether diners would prefer Fixed-Interval timing with a higher feeding frequency.

We also investigate whether the robot users' pre-study hunger level affected their bite timing ratings. As shown in Fig. 11 (right), we do not find any statistically significant differences with $p_{0.05}$ between the three hunger levels users reported. This could suggest that their bite timing ratings were not biased by their hunger level. However, we cannot draw any strong conclusions as the hunger level self-assessment is a very subjective metric.

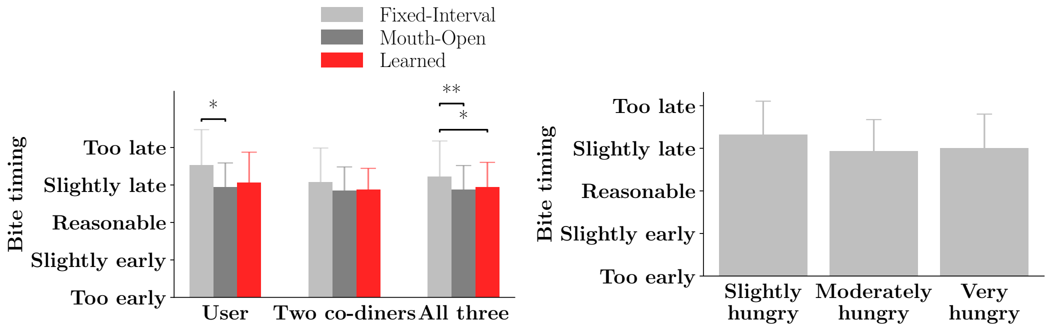


Figure 11: HRCom: **Left:** Bite timing perceived by users, two co-diners, and all three diners on a Likert scale 1-5 (Too early - Too late), for each bite timing strategy (Fixed-Interval, Mouth-Open, and Learned timing). $*, **$ denote statistically significant differences with $p_{0.05}, p_{0.005}$ respectively. **Right:** Effect of robot users' hunger level on their bite timing ratings. Our study did not find any statistically significant differences with $p_{0.05}$.

Other factors. Besides bite timing itself, we evaluate differences between bite timing strategies for other factors: distraction by the robot (already discussed in Sec. 7), ability to have natural conversations (Fig. 12 (top left)), ability to feel comfortable around the robot (Fig. 12 (top right)), system reliability (Fig. 12 (bottom left)), system trustworthiness (Fig. 12 (bottom right)), overall experience of the meal (Fig. 13 (left)), social interactions with other participants (Fig. 13 (right)). We can see that the ability to have natural conversation and feel comfortable around robot is significantly lower for Mouth-Open timing than for Learned or Fixed-Interval timing. This aligns with the finding in Sec. 7 that Mouth-Open timing distracts dining participants significantly more than Learned or Fixed-Interval timing. It is however interesting to note that co-diners, not users of the robot, felt less comfortable around the robot during Mouth-Open Timing. We speculate this is because co-diners perceive the prompt from a voice interface as an external disruption factor not related to their own eating whereas for robot users it is what makes them feed so it does not set robot users into discomfort as much. For users, co-diners as well as all three diners, we do not find any statistically significant differences between bite timing strategies in system reliability, trustworthiness, overall experience nor social interactions they had.

Replies to open-ended post-study questions. We also analyze the study participants' answers to open-ended questions from Fig. 16 (c). We observe the following patterns.

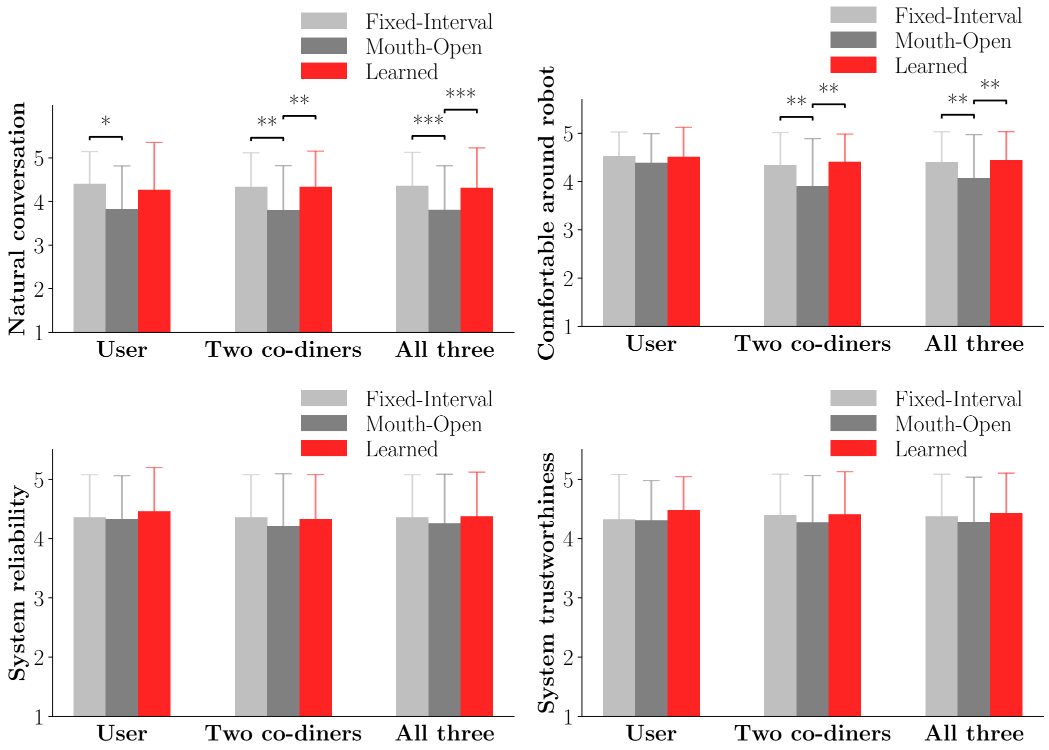


Figure 12: HRCom: **Top left:** Ability to have natural conversation. **Top right:** Ability to feel comfortable around the robot. **Bottom left:** System reliability. **Bottom right:** System trustworthiness. Perceived by users, two co-diners, and all three diners on a Likert scale 1-5 (Strongly disagree - Strongly agree), for each bite timing strategy (Fixed-Interval, Mouth-Open, and Learned timing). $*, **, ***$ denote statistically significant differences with $p_{0.05}, p_{0.005}, p_{0.0005}$ respectively.

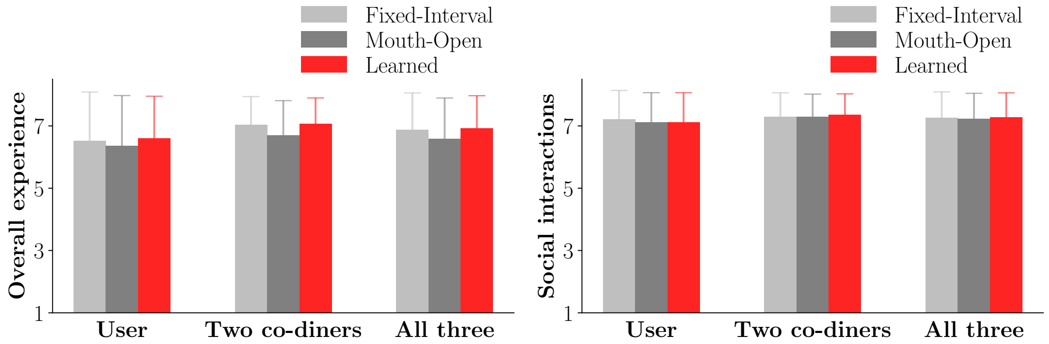


Figure 13: HRCom: **Left:** Overall meal experience **Right:** Social interactions perceived by users, two co-diners, and all three diners on a Likert scale 1-7 (Strongly disagree - Strongly agree), for each bite timing strategy (Fixed-Interval, Mouth-Open, and Learned timing). Our study did not find any statistically significant differences with $p_{0.05}$.

$*$ *Whether participants felt safe around the robot*

- Robot users:
  - 7 users (70%) replied with *"Yes, . . . "*
  - Their main concern was **safety when robot was moving around with the fork**: *"Yes for the most part. Sometimes it felt surprising how close it got to my face when it went to pick up food.", "Yes. At first I thought it was going to stab me in the face but it moved slow and never hurt me.", "Sort of. My primary concern was the robot's resting*

state. When the neutral position has the fork poised at eye level it is very concerning. Simply aiming the fork down and away from the table would make a huge difference.”

  - Many noted that the **initial familiarization with the robot helped**: *”Not very safe at beginning , but with trials go on, I feel more safe. If I have e stop myself I’ll feel more safe.”, ”I was a bit nervous, but after the first few trials I felt more comfortable around the robot”, ”Yes! Very quickly got used to it’s actions, which were very regular so easy to get used to safety-wise”*

- Other co-diners:
  - **No major concerns:** *”Yes. It seemed to be under control nicely.”, ”Yes: it helped that the robot moved pretty slowly and along familiar ”tracks” through the air. The e-stop was nice to have too!”, ”Yes. It avoided my friends and I well. It didnt seem overpowering.”, ”Moderately. The robot was cutting it close to the person’s face while going to grab the food”*
  - Similarly to robot users, the **initial familiarization with the robot helped**: *”Yes, after a few trials I felt safe around the robot. But might be because it’s far away from me as well”, ”I was a little uncomfortable initially but I started feeling safe after a few trials.”, ”I was little uneasy at first but then I quickly forgot about it and was comfortable”, ”Yes. I was a little worried at first, but wound up feeling very comfortable around the robot.”*

These replies show that familiarizing users as well as other participants with the robot helps them to feel safer around the robot.

∗ *When participants think it is appropriate to take a bite of food when they are eating with others*

- **Talking-related rules:** *”Usually when someone else is speaking and you are not expecting a question to be asked to you”, ”When I am not talking, or being directly talked to”*
- **Eye gaze-related rules:** *”After a few second pause in speech combined with a stationary eye position.”, ”. . . if it is a very serious topic, or they are making eye contact, I would probaly wait.”*
- **Diner physical state-related rules:** *”When you are not speaking and have the desire to take a bit.”*
- **Social interaction-related rules:** *”. . . when the present speaker is not saying anything very emotional, energetic, or charged. For example I would not like to take a bite when consoling a crying friend.”, ”. . . if it is a very serious topic, or they are making eye contact, I would probaly wait.”*
- **Time-related rules:** *”When there is a stop(all people stop talking) longer than 1.5s, I feel it’s right time . . . ”*
- **Robot-related:** *”The robot shouldn’t wait too long after the food is on the fork.”, ”Almost always. I would say worst case the biter can wait to make the move towards the robot, but it seems very appropriate for the robot to “always” be feeding and take a bite almost immediately when it’s ready.”*

These replies match the same kinds of rules we find in replies to the same question asked during human-human commensality (App. 8.1.5).

∗ *What participants liked about the meal experience*

- Robot users:
  - **Food and conversation:** *”Conversation with people, fruit”, ”It was still easy to have a natural conversation, . . . ”*
  - **Robot and its behavior:** *”I liked that the robot did the same thing over and over, making it easy to ignore”, ”. . . the robot was relatively quiet. It’s kind of nice to be fed and I like fruit.”, ”The food item is placed in a proper position, not too far or close, I have the choice to eat or not.”, ”The robot was generally out of the way. Once we went through a few trials, the robot was less distracting”, ”It was very nice not having to think about bite acquisition and delivery”, ”. . . enjoyed the novelty of the robot”*
- Other co-diners:
  - **Food and conversation:** *”My food was great. People were too.”, ”I was able to have a natural conversation”, ”. . . The robot was not too intrusive and was almost a cool fourth diner.”*

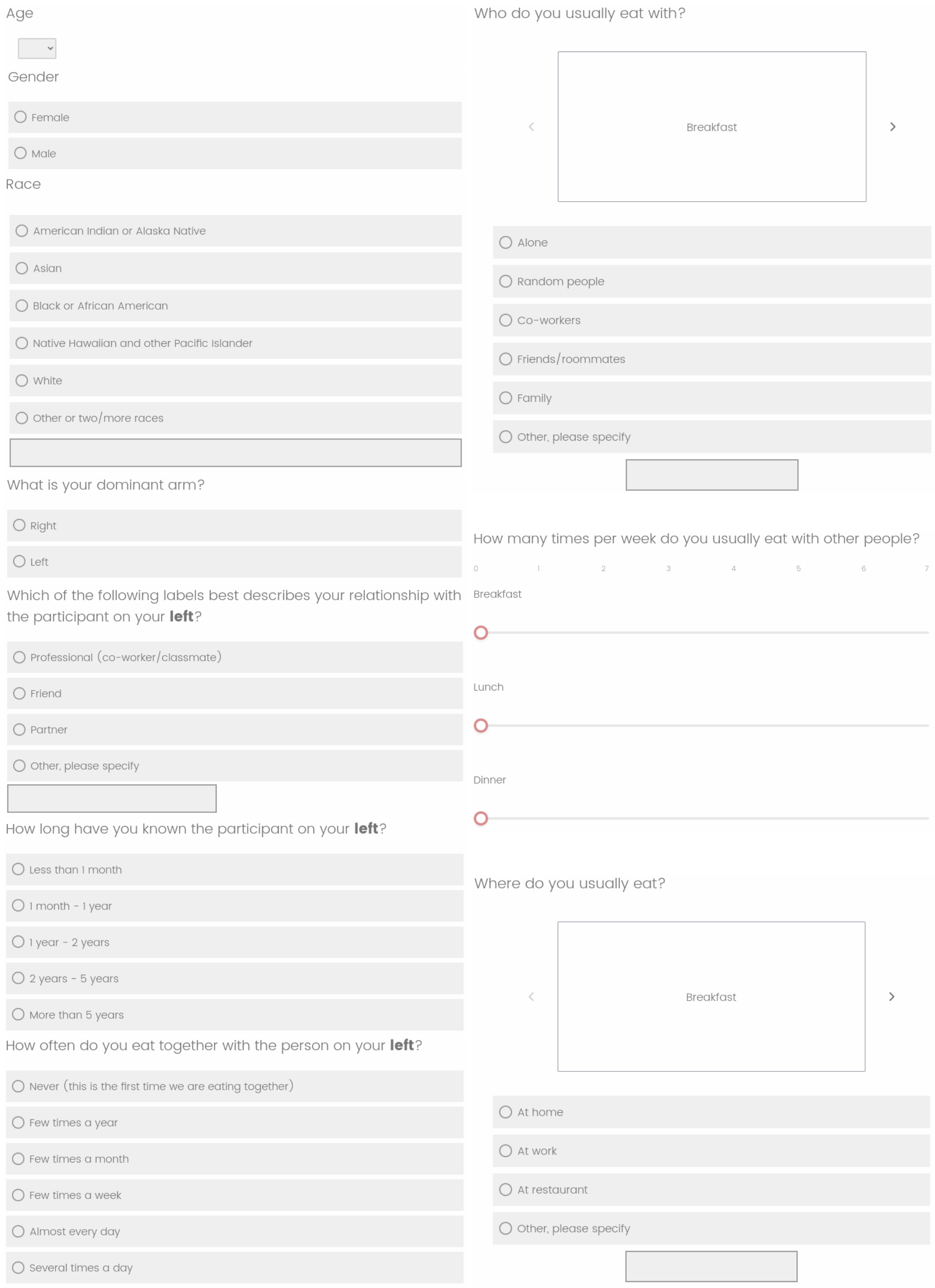
- **Robot and its behavior:** *"It did not take too long to become accustomed to the robot.", "... the robot was not as much of a disruption as I imagined.", "I feel the pace was nice, and I felt more normal than I expected. The conversation flow was good and not interrupted by the robot.", "... that the robot waited till was a natural pause in conversation from participant 1 before "speaking" or coming forward with the food so the experience was pretty smooth.", "interesting to watch the robot moves, and I felt the robot wasn't that distracting when it didn't make any sound", "There were many times when the robot was very much in the background of the conversation and the conversational flow was uninterrupted"*
- **Dining setting:** *"Circular table made for nice discussion atmosphere. 3 people is also nice so we can talk while the other is being fed."*

Both users and co-diners liked food and conversation which matches what participants during human-human commensality experienced (in replies to open-ended questions in App. 8.1.5). This suggests that the addition of the assistive feeding robot does not remove these particular factors of commensality that people like. Also, participants seemed to like the robot behavior and its presence as a new element in commensality.

∗ *What participants did not like about the meal experience*

- Robot users:
  - 7 robot users (70%) found the **voice prompts during Mouth-Open Timing distracting**: *"Robot was very distracting especially when it spoke commands", "When the robot talks, it breaks the flow of the conversation.", "The voice that told me to look at the robot was sometimes distracting.", "I didn't like the trials when the robot prompted to eat", "I don't like voice interruption by robot ... ", "When the robot spoke it would cut off the conversation."*
  - **Robot position, speed, and noise:** *"I didn't like how it became harder to make eye contact why talking because sometimes the robot would block out eye level ... ", "... robot was a bit slow so I didn't get to eat much ... ", "... the noise make by robot in operation makes others voice hard to heard clearly."*
  - **Questionnaires after each trial:** *"... taking survey in between bites also was challenging as it interrupted the flow"*
  - **Bite timing:** *"Sometimes the robot was distracting when I was in the middle of a story", "... it was weird because I felt like I couldn't signal when I wanted the food and had to wait.", "The robot took too long to feed me. It would take several hours to eat enough food with its speed. There is a tradeoff between timely feeding and fast enough feeding to finish a meal in a proper amount of time."*
- Other co-diners:
  - 9 co-diners (45%) found the **voice prompts during Mouth-Open Timing distracting**: *"I didn't love the trials when the robot spoke ... ", "Robot was sometimes speaking in the middle of the conversation"*
  - **Robot position:** *"robot blocked my sight when I talked to the person on the left side"*
  - **Time to get used to the robot:** *"Not much. It took a while to get used to the robot.", "When the machine spoke over us it was hard to keep the conversatiom going, although this became easier over time."*
  - **Questionnaires after each trial:** *"Interruptions for the survey broke up the conversation"*
  - **Bite timing:** *"A couple of trials, the robot came in slightly early or waited for a while."*
  - **Conversation content:** *"We all consciously or unconsciously had to structure our conversation around what the robot was doing at a particular point of time."*

These replies clearly show that both robot users as well as co-diners find the Mouth-Open bite timing strategy disrupts the flow of the conversation. As several participants reported that the robot movements interrupted their mutual eye contacts, it would be interesting to explore robot bite transfer trajectories that minimize eye gaze blockage.



Age

Gender

○ Female

○ Male

Race

○ American Indian or Alaska Native

○ Asian

○ Black or African American

○ Native Hawaiian and other Pacific Islander

○ White

○ Other or two/more races

What is your dominant arm?

○ Right

○ Left

Which of the following labels best describes your relationship with the participant on your **left**?

○ Professional (co-worker/classmate)

○ Friend

○ Partner

○ Other, please specify

How long have you known the participant on your **left**?

○ Less than 1 month

○ 1 month - 1 year

○ 1 year - 2 years

○ 2 years - 5 years

○ More than 5 years

How often do you eat together with the person on your **left**?

○ Never (this is the first time we are eating together)

○ Few times a year

○ Few times a month

○ Few times a week

○ Almost every day

○ Several times a day

Who do you usually eat with?

‹ Breakfast ›

○ Alone

○ Random people

○ Co-workers

○ Friends/roommates

○ Family

○ Other, please specify

How many times per week do you usually eat with other people?

0 1 2 3 4 5 6 7

Breakfast

Lunch

Dinner

Where do you usually eat?

‹ Breakfast ›

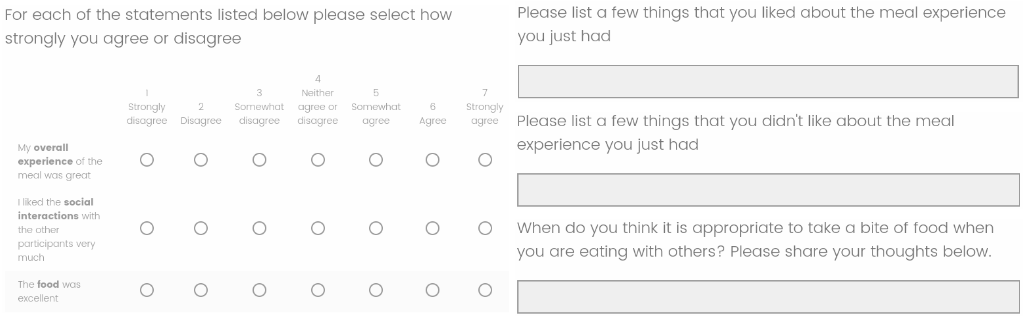
○ At home

○ At work

○ At restaurant

○ Other, please specify

Figure 14: HHCD and HRCom: Pre-study questionnaire: questions about demographic background, relationship to other participants (the same questions were asked in relation to the participant on the right), and social dining habits.



For each of the statements listed below please select how strongly you agree or disagree

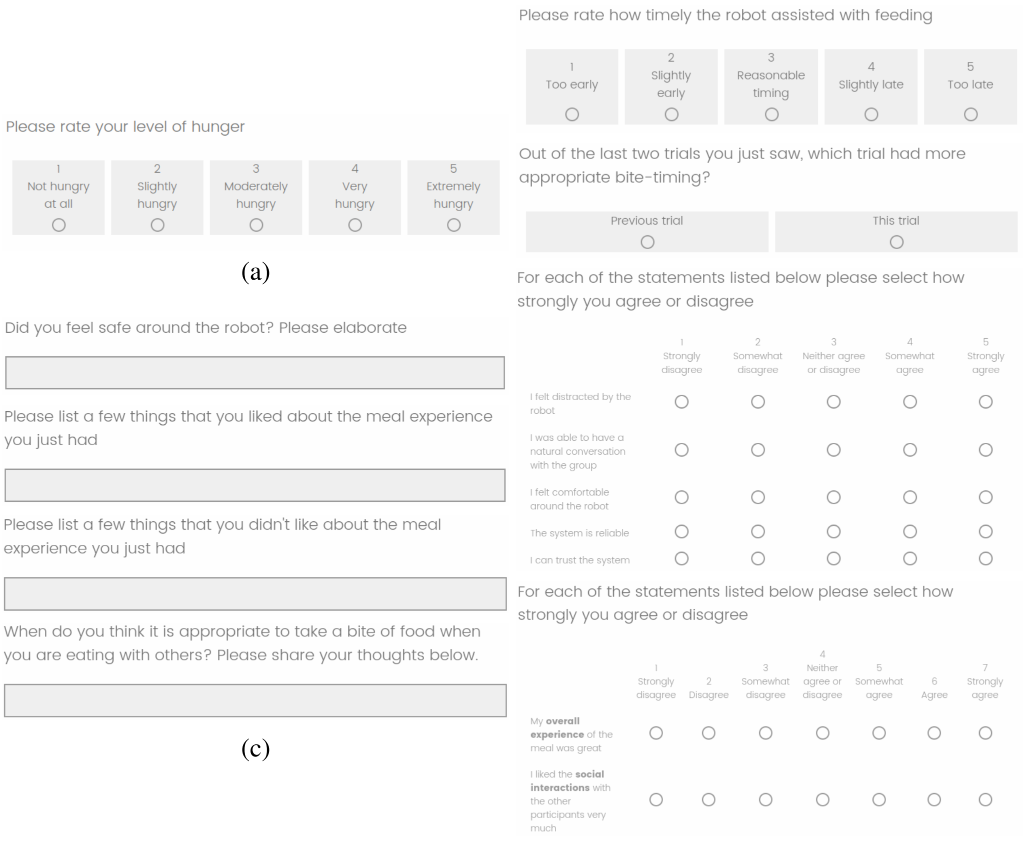
	1 Strongly disagree	2 Disagree	3 Somewhat disagree	4 Neither agree or disagree	5 Somewhat agree	6 Agree	7 Strongly agree
My **overall experience** of the meal was great	○	○	○	○	○	○	○
I liked the **social interactions** with the other participants very much	○	○	○	○	○	○	○
The **food** was excellent	○	○	○	○	○	○	○

Please list a few things that you liked about the meal experience you just had

Please list a few things that you didn't like about the meal experience you just had

When do you think it is appropriate to take a bite of food when you are eating with others? Please share your thoughts below.

Figure 15: HHCD and HRCom: Post-study questionnaire: questions about dining experience.



Please rate your level of hunger

1 Not hungry at all	2 Slightly hungry	3 Moderately hungry	4 Very hungry	5 Extremely hungry
○	○	○	○	○

(a)

Did you feel safe around the robot? Please elaborate

Please list a few things that you liked about the meal experience you just had

Please list a few things that you didn't like about the meal experience you just had

When do you think it is appropriate to take a bite of food when you are eating with others? Please share your thoughts below.

(c)

Please rate how timely the robot assisted with feeding

1 Too early	2 Slightly early	3 Reasonable timing	4 Slightly late	5 Too late
○	○	○	○	○

Out of the last two trials you just saw, which trial had more appropriate bite-timing?

Previous trial	This trial
○	○

For each of the statements listed below please select how strongly you agree or disagree

	1 Strongly disagree	2 Somewhat disagree	3 Neither agree or disagree	4 Somewhat agree	5 Strongly agree
I felt distracted by the robot	○	○	○	○	○
I was able to have a natural conversation with the group	○	○	○	○	○
I felt comfortable around the robot	○	○	○	○	○
The system is reliable	○	○	○	○	○
I can trust the system	○	○	○	○	○

For each of the statements listed below please select how strongly you agree or disagree

	1 Strongly disagree	2 Disagree	3 Somewhat disagree	4 Neither agree or disagree	5 Somewhat agree	6 Agree	7 Strongly agree
My **overall experience** of the meal was great	○	○	○	○	○	○	○
I liked the **social interactions** with the other participants very much	○	○	○	○	○	○	○

(b)

Figure 16: HRCom: Questionnaires: (a) Additional question asked in pre-study questionnaire in addition to questions in Fig. 14. (b) Experiment questionnaire asked after each trial. Note, we did not ask the second forced-choice question after the first trial. (c) Post-study questionnaire.