Templates and Graph Neural Networks for Social Robots Interacting in Small Groups of Varying Sizes

Sarah Gillet
KTH Royal Institute of Technology
Stockholm, Sweden
sgillet@kth.se

Sydney Thompson
Yale University
New Haven, USA
sydney.thompson@yale.edu

Iolanda Leite
KTH Royal Institute of Technology
Stockholm, Sweden
iolanda@kth.se

Marynel Vázquez
Yale University
New Haven, USA
marynel.vazquez@yale.edu

***Abstract*—Social robots need to be able to interact effectively with small groups. While there is a significant interest in human-robot interaction in groups, little focus has been placed on developing autonomous social robot decision-making methods that operate smoothly with small groups of any size (e.g. 2, 3, or 4 interactants). In this work, we propose a Template- and Graph-based Modeling approach for robots interacting in small groups (TGM), enabling them to interact with groups in a way that is group-size agnostic. Critically, we separate the decision about the target of their communication, or "whom to address?" from the decision of "what to communicate?", which allows us to use template-based actions. We further use Graph Neural Networks (GNNs) to efficiently decide on "whom" and "what". We evaluated TGM using imitation learning and compared the structured reasoning achieved through GNNs to unstructured approaches for this two-part decision-making problem. On two different datasets, we show that TGM outperforms the baselines encouraging future work to invest in collecting larger datasets.**

***Index Terms*—Human-Robot Interaction, Groups, Social behavior generation**

I. Introduction

People interact in a variety of small groups in different everyday situations: at work in teams, at home with their family, or at a park with friends. The size of these groups differs based on the need or arrangements in the context they are in, for instance, a family of three and a pair of work colleagues. Social groups can also change dynamically – the family is sometimes joined by the grandmother. Hence, socially assistive robots that interact with groups of people must be able to interact with small groups of varying sizes.

There has been a growing interest in the Human-Robot Interaction (HRI) community to explore how a robot can support interactions in small groups of people [1], [2]. Researchers have used robots to act as a facilitator in educational groups [3], to support conflict resolution [4], [5], and encourage participation [6]–[9]. These prior works focus on revealing underlying phenomena of group human-robot interaction – often in groups of fixed sizes. Other works pioneered group-specific computational approaches [8]–[10] focusing on either dyads or triads of people in HRI settings. However, to the best of our knowledge, no prior work has explored robot decision-making, i.e., selecting an action based on the perceived state, for small groups with methods agnostic to the group size.

We propose a group-size agnostic Template- and Graph-based Modeling approach for robots interacting in small groups (TGM) which we evaluate in an imitation learning setting. The approach uses template-based actions to enable robots to make decisions in groups of different sizes and Graph Neural Networks (GNNs) to select these actions efficiently.

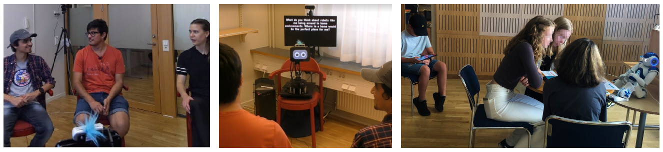

Fig. 1. We evaluate our approach, TGM, on two datasets: (1) an interaction between the robot Shutter and two to four human group members (*left* and *middle* images), (2) the robot Nao, controlled by the teenager on the left, interacting with the other three teenagers at the table (*right*).

A key insight is that robots must always choose "what" to do to interact with people and "whom" from the group to address, the latter being unique to small group settings. For example, imagine a robot that is brainstorming with three people about its use in a home environment, as in Fig. 1 (left). The robot might want to encourage balanced participation to learn about everyone's opinion. Thus, it might first need to choose an addressee and then decide what to do to achieve this goal. It could select a quiet group member (whom) to just look at (what) for attention or ask the loudest contributor (whom) to summarize their thoughts (what). Template-based actions allow for this sequential decision-making, resulting in a constant template-based action space for "what" with an auxiliary choice for "whom" instead of a combinatorial action space with every combination of "whom" and "what".

Our approach leverages GNNs to model the robot's behavior policy. GNNs can capture the underlying structure of the human-robot interaction by modeling it as a graph. This way, each group member (node) and the interaction between group members (edges) can be modeled explicitly.

We evaluated TGM on two datasets with distinct interactions illustrated in Fig. 1: 1) a brainstorming discussion among adults and 2) a discussion-based design and decision-making task among teenagers. Our goal in evaluating TGM on these two datasets is twofold. First, we show that TGM enables a robot to clone demonstrations across groups of varying sizes in a realistic dataset with heuristic demonstrations collected in brainstorming sessions with two to four people. This dataset setup can be considered idealistic because the heuristic demonstrations are the most consistent possible. Second, we provide

Accepted to the 20th ACM/IEEE International Conference on Human-Robot Interaction, 2025, Melbourne, Australia

insight into how TGM performs in a dataset with complex human demonstrations collected from teenagers. In an extensive ablation study, we explore the diversity of group sizes needed for generalizing to new groups of seen sizes and new groups of unseen sizes. We demonstrate that TGM outperforms both Multi-Layer Perceptrons (MLPs) and Random Forest (RF) baselines in choosing "whom" to address and "what" to do relative to the addressee. In sum, our main contributions are:

1) A novel approach (TGM) to model a robot's decision-making enabling interactions in groups of varying sizes;

2) An evaluation of the efficacy of TGM on two datasets (one with simple, ideal demonstrations and one with complex human demonstrations). Our evaluation shows that:

- TGM outperforms linear input modeling baselines when trained on multiple group sizes and subsets of the data.
- TGM generalizes to group sizes unseen during training.

II. Background and related work

Robots in Groups: The study of robots and groups has gained importance in HRI [2], including how people perceive robots in groups and how they influence and facilitate group dynamics [1], [11], [12]. In particular, prior works explored robots in a variety of group constellations, facilitating interactions among adults [9], [13], [14], children [15]–[17], mixed groups of toddlers and parents [18], and students [3], [19]. These groups are often studied in fixed size and constellation when interacting with a robot as peripheral companion [8], [20], through conversational engagements [7], [21]–[24], in multi-party games [4], [25]–[27], or collaborative tasks [5], [28]–[30]. Most related to our work are investigations in which people might join or leave a small group [31], [32]. In both works, the robot's behavior is group size agnostic, but the robot cannot choose to direct its actions to a specific individual.

Imitation Learning in HRI: Imitation learning (also learning from demonstration in robotics [33], [34]) has been demonstrated to generate robot policies for a variety of human-robot interaction scenarios. A common approach is to learn robot behaviors from expert human demonstrations, such as manipulation skills [35], [36]. Similarly, we use expert demonstrations provided by a person that directly controlled a robot but focus on modeling and learning for social interactions. Prior work on generating social behaviors explored, e.g., backchanneling [37], [38] or responding to ambiguous questions [39]. These works typically use a human-human interaction dataset for one-on-one interaction instead of investigating datasets of human-robot interactions and groups of people.

Template-based action spaces: Template-based action spaces have previously been used in natural language generation for interactive fiction or adventure games [40], [41]. In these games, the action space consists of templates such as *put* __ *in* __. The underlined portions of the template are filled with objects from the environment. Template-based actions were proposed in this problem space to overcome the challenges of combinatorial action spaces. Given their straightforward nature, combinatorial action spaces that comprise all combinations of "whom" and "what" have previously been used for learning robot decision-making policies in small group settings [10]. We utilize a template-based action space and avoid a combinatorial action space by viewing the target, "whom", as an additional context to the action. In the field of HRI, templates have been used for explanation generation [42] or dialog generation [43]. Our work expands the use of template-based action spaces to multi-modal actions in the context of group human-robot interaction.

GNNs for HRI: Prior work used GNNs to model groups in HRI settings to predict group positioning behavior [44] or detect backchanneling behavior [45] through graph convolutional networks. The latter work compares modeling individuals vs. the whole group through GNNs with a beneficial outcome for GNNs. Another work compares temporal graph models to a rule-based history baseline to model social dynamics and predict the next gaze and speakers [46]. Our approach uses a Message-Passing Graph Neural Network (MPGNN) [47] because they are well-suited for problems that require node, edge and/or whole graph predictions. MPGNNs are composed of one or more *Graph Network layer* (GN layer). A GN layer takes as input a directed graph $\mathcal{G}$ and produces an updated graph $\mathcal{G}'$. Let $\mathcal{G} = (u, V, E)$, where u is a global attribute (or feature) for the graph, $V = \{v_i\}_{i=1:n}$ are attributes of the graph's nodes, and $E = \{(e_k, r_k, s_k)\}_{k=1:m}$ corresponds to the edges. Each e_k in E is an edge attribute with $(r_k,\ s_k)$ being the corresponding indices of the receiver and sender nodes. Then, a GN layer operates in three steps. First, the edge features are updated using an edge update function ϕ^e such that $e'_k = \phi^e(e_k, v_{r_k}, v_{s_k}, u)$ for a given edge e_k. Second, the node features are updated in a similar fashion but aggregate information from edges. For example, for node i, $v'_i = \phi^v(\bar{e}'_i, v_i, u)$ with $\bar{e}'_i = \rho^{e\rightarrow v}(\{(e'_k, r_k, s_k)\}_{r_k=i, k=1:m})$ being aggregate information from all edges that have the node i as receiver. Third, the global feature u for the graph is updated as $u' = \phi^u(\bar{e}', \bar{v}', u)$ using all the edges, $\bar{e}' = \rho^{e\rightarrow u}(\{(e'_k, r_k, s_k)\}_{k=1:m})$, and node information, $\bar{v}' = \rho^{v\rightarrow u}(\{v'_i\}_{i=1:n})$. The update functions $\phi^e(\cdot), \phi^v(\cdot), \phi^u(\cdot)$ and the aggregate functions $\rho^{e\rightarrow v}(\cdot), \rho^{e\rightarrow u}$, $\rho^{v\rightarrow u}(\cdot)$ are differentiable functions, which allows training the GNN via gradient descent. Importantly, the aggregate functions are often implemented via symmetric mathematical functions, like element-wise averaging, because nodes and edges in a graph typically lack a natural order.

Prior work in HRI using MPGNNs generated robot poses joining a group [48] or detected the presence of groups [49]. Whereas these works are more concerned with physical movement and positions, our work considers social signals and behaviors. Different from prior work on social tasks, we compare GNNs to other group-size agnostic baselines. Additionally, no prior work explored imitating social robot behaviors from demonstrations through the combination of MPGNNs and template-based actions for group HRI.

III. Method

We study the problem of generating social robot behavior for interacting in small groups of varying sizes in an imitation

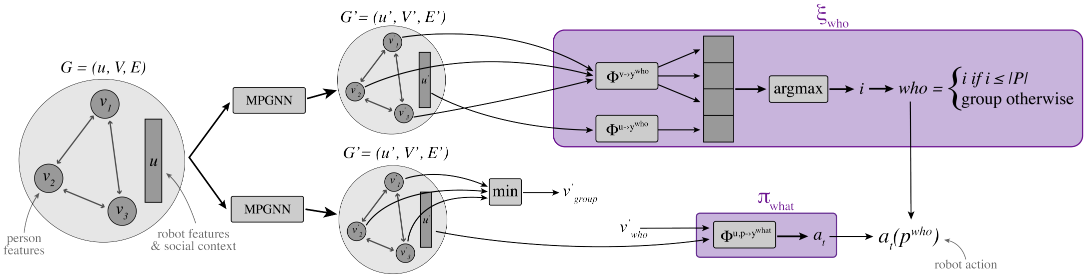

Fig. 2. The architecture of TGM. Each node, v_i, in the graph, represents a person, and the global graph attribute, u, contains information about the robot and interaction context. The original graph G is passed through an MPGNN to create an updated graph, G'. The policy for choosing the robot's addressee, ξ_{who}, consists of two steps. The updated node values are passed through $\phi^{p \to y^{\text{who}}}$ and the updated global graph attribute is passed through $\phi^{u \to y^{\text{who}}}$ to generate a score for each person. The person with the maximum score becomes p^{who} - the robot's addressee. The policy for choosing what to do relative to p^{who} is selected using π_{what}. Using the updated global graph attribute, u', and the features of the addressee v'_{who}, $\phi^{u,p \to y^{\text{what}}}$ chooses a single action template a_t.

learning setting. We view the problem as a decision-making problem and model it through a Markov Decision Process (MDP). The MDP is defined by a set of states $s \in \mathcal{S}$ describing the group interaction and a set of multi-modal actions $a \in \mathcal{A}$ suitable for the robot to interact with the group. We generally assume that the transition function $\mathrm{T}(s, a, s')$ of the MDP describing how state s transitions into state s' as a result of action a is unknown. The reason is that transition functions in group HRI are hard to model due to the unpredictability of human-human interactions, unknown effects of robots on humans, and the limited availability of HRI data. The robot's goal is to then learn a policy $\pi : \mathcal{S} \to \mathcal{A}$ that indicates which action a to take in a given state s. We train the policy π through behavioral cloning [50]. That is, we use supervised learning to map observed states s_t to actions a_t given paired input-target data from a dataset with t indicating the time step.

A. *Template- and Graph-Based Modeling Approach for Interactions in Groups (TGM)*

We propose TGM, a novel approach for decision-making in small group human-robot interactions that is group size agnostic and can be used across varied interaction settings. We model groups as graphs captured in state s_t, which allows us to reason upon this group structure through GNNs. For modeling the action a_t, we propose to use multi-modal template-based action spaces. The template-based action space serves to avoid a combinatorial increase in the action space due to having to address both "whom" and "what" decisions.

a) Modeling groups as graphs: We model the state s of a small group human-robot interaction as a graph $\mathcal{G} = (u, V, E)$, where each node $v \in V$ represents features of one human group member. We denote human group members with $p_i \in P$. Edges $e_{ij} \in E$ encode the relation between group members p_i and p_j through, for example, the physical distance between them. The robot's and group's overall state is captured in the global graph attribute u of $\mathcal{G}$. This formulation is applicable to a robot that interacts with 2 or more people, but our primary focus is on small groups of 2 to 4 people.[1]

b) Template-based action space and policies: We represent the actions that the robot takes with action templates [40]. The action templates structure the robot's actions by using information about "whom" the robot targets with the action - a given person or the whole group. For instance, at time t, if person $p_t^{\text{who}} \in P$ is selected to be addressed and, the selected action template $a_t \in \mathcal{A}$ is "Ask __ to elaborate", then the filled action template $a_t(p_t^{\text{who}})$ is "Ask $\underline{\;p_t^{\text{who}}\;}$ to elaborate". The filled template $a_t(p_t^{\text{who}})$ represents an action that can be taken by the robot and specifies *what* the robot should do when addressing person p_t^{who}. Note that this need not be an utterance but could indicate the behavior more abstractly.

We propose to reason about the template-based actions in two steps due to a natural dependency between the "whom" and "what" decisions. First, the robot chooses "whom" to address for its next action. Second, the robot decides which action to take based on the world state s_t and the result of the "whom" decision, p_t^{who}. Because of this natural split, we implement the auxiliary choice that represents the addressee separate from the behavior policy choosing the robot's action. The first choice that selects "whom" to address is defined as follows:

$$\xi_{\text{who}} : f(s_t) \to p_t^{\text{who}} \tag{1}$$

where p_t^{who} can be an individual person or the whole group. The first choice ξ_{who} then informs which action template is chosen to decide "what" to do when addressing p_t^{who}:

$$\pi_{\text{what}} : f(s_t, p_t^{\text{who}}) \to a_t \tag{2}$$

where a_t is a template which can then be executed as $a_t(p_t^{\text{who}})$.

Template-based actions are beneficial for this problem because of their flexibility while providing structure. Templates also help to avoid a combinatorial action space. For instance, consider a robot in a group with three people that can choose to tell a joke, make a request, or ask a question. The robot can address any individual or the entire group. A combinatorial action space has 12 possible outcomes. Using templates allows us to reduce the complexity from a single decision with 12 options to two decisions with 4 options for "whom" and 3 options for the robot's action template. As illustrated, using templates and thereby separating the decision into two distinct

[1] Our formulation generally assumes that there is a single robot in the interaction – but if there were more, they could be added as nodes to the graph. Evaluating the latter setup is out of the scope of this paper.

decisions – “whom” and “what”– allows the action space to remain unchanged even with varying group sizes.

c) Decision-making with GNNs: Both decisions, ξ_{who} and π_{what}, are separately modeled through a MPGNN. In our experimental evaluation, we explore the use of one or two computational graph layers. The l^{th} layer of the MPGNN transforms the input graph $\mathcal{G}$ to a new graph $\mathcal{G}'$. This transformation takes place in three steps, as explained in Sec. II. Assuming the input graph in the first layer of the MPGNN is $\mathcal{G} = (u, V, E)$, the MPGNN first updates the edges through ϕ^e resulting in e'_k. Second, features from incoming edges for each node are aggregated $\rho^{e \rightarrow v}$ and then used to update the node features so that each node gets updated through ϕ^v resulting in $v'_i \in V'$. Third, the global feature u for the graph is updated through ϕ^u resulting in u'. The latter function ϕ^u uses aggregated edge $\rho^{e \rightarrow u}$, and aggregated node features $\rho^{v \rightarrow u}$.

For ξ_{who}, we use final-layer node embeddings v'_i and u' if the whole group can be addressed and transform them through $\phi^{v \rightarrow y^{\text{who}}}$ and $\phi^{u \rightarrow y^{\text{who}}}$ to a prediction $y_i^{\text{who}} \in \mathbb{R}$ with $i \in [0, |V|]$, where $i < |V|$ corresponds to the likelihood of selecting an individual and $i = |V|$ is the whole group. We choose p^{who} through $argmax_i(y_i^{\text{who}})$. For implementing π_{what}, we use the final-layer global output u' and the v'_{who}, concatenate them, and transform them through a dense layer, $\phi^{u,v \rightarrow y^{\text{what}}}$, to $y^{\text{what}} \in \mathbb{R}^{|\mathcal{A}|}$. The respective functions ϕ are modelled with MLPs for non-sequential data or Long-short Term Memory (LSTM) for sequential data.

B. Baselines - Linear inputs and models

Traditionally, linear models have been used to reason about group interactions [51], [52]. We compare TGM against two linear modeling approaches: one that handles arbitrary group sizes and one that does not. In both cases, the inputs to the linear model are constructed from the graph. For the MPGNN, the state s is the entire graph, but for linear models, s must be represented as a vector. We construct input vectors $s_{\text{dependent}}$, which is dependent on the group’s size, and $s_{\text{independent}}$, which is independent of group size. For $s_{\text{dependent}}$, we concatenate the features for all nodes and the global graph attribute to one input vector, $s_{\text{dependent}} = [\, \|_{j=0}^{|V|} v_j, u \,]$. We use the same action space as for TGM, i.e., we directly output probabilities from π^{who} for each $p_i \in P$ and from π^{what} for each $a_t(p_i) \in \mathcal{A}$ with p_i being the output of π^{who}. For an arbitrarily-sized group, we construct the model input, $s_{\text{independent}}$, for a single individual, p_i, by concatenating the node features v_i with a group context feature formed by min-pooling[2] the node values of other group members, $v_i^{\text{min}} = [\, \min\{(v_j)_k \text{ for } j \in [0, |V|),\ j \neq i\}, k \in [0, len(v))\,]$, and the global graph attribute u to construct $s_{\text{independent}} = [v_i, v_i^{\text{min}}, u]$. This way $s_{\text{independent}}$ is a fixed-length input vector regardless of the group size. We use a different model output when using $s_{\text{independent}}$ to accommodate different group sizes, output $y_{p_i}^{\text{address}} \in [\text{address}, \text{not address}], p_i \in P$. To compute the final *who*, we select the person with the highest certainty for “address”, $p_i^{\text{who}=argmax[y_{p_i}]}$. When the robot can address the whole group, we choose the whole group if “not address” was chosen for all individuals.

For training π^{what}, we follow a similar approach, using our constructed states - $s_{\text{independent}}$ or $s_{\text{dependent}}$ - *and* the p_i chosen from p_i^{who}. We use the same input vector but build on the dependency of first deciding “whom” and then “what” and only keep the features of the person who was addressed.

IV. Datasets

For the datasets, we chose group interactions in which the robot supports the group by shaping interactions among people [11] as our application context. We explore the performance of TGM in a purposefully collected dataset of two to four people interacting with the robot – the brainstorming dataset. This dataset allows us to explore the performance in a realistic dataset with ideal, post-annotated demonstrations. We also use a dataset with human demonstrations – the teenager dataset – in which three teenagers interact with the robot [15].

A. The brainstorming dataset

For collecting a dataset of two to four people interacting with the robot with the potential to shape interactions, we designed a brainstorming task in which the robot can make autonomous decisions at every conversational turn. We used an Azure Kinect Camera and individual close-talk microphones to collect audio and posture information from participants.

During the data collection, the robot acted autonomously but followed a random policy within the given turn-based system. This was to collect a realistic interaction but make the dataset available to a larger variety of learning algorithms, e.g., offline RL or causal learning. For this work, we needed demonstrations for behavioral cloning. We designed a literature-informed heuristic that provided idealized and simple demonstrations through post annotations as outlined below. We release the dataset for benchmarking and reproducibility efforts[3].

The brainstorming task: The task for the group was to brainstorm about robots in home environments. The robot, thereby, moderated the discussion, i.e., it asked general questions to discuss specific topics and asked follow-up questions or encouraged more ideas (see Action space).

Participants: 78 participants total across 31 groups with two (18 groups), three (12 groups), or four (1 group) human group members joined the data collection. 15 participants were between 18 and 24 years old, 40 were 25-34 years old, 9 were 35-44 years old, 9 were 45-54 years old, 2 were 55-64 years old, and 3 were 65+ years old. 43 identified as male, 34 as female, and 1 as non-binary.

Size of the dataset: The 31 brainstorming sessions provided 3424 decisions taken when the robot could take the turn or when one participant took the turn, i.e., started speaking.

Action space: With a previously chosen p^{who}, the following multi-modal action templates were available: (1) Look more

[2] We chose min-pooling in any place where we use pooling per expert advice on graph neural networks. A pre-study showed no notable differences across pooling functions.

[3] We can only release the processed and not the raw data. Other work published a different extract from the same raw data to study topic changes [53].

at $\underline{\;p^{who}\;}$, (2) Ask opinion question to $\underline{\;p^{who}\;}$, (3) Ask more ideas question to $\underline{\;p^{who}\;}$, (4) Ask concerns question to $\underline{\;p^{who}\;}$.

Feature extraction - Graph state space: We collected features that are relevant to perceiving and shaping group dynamics based on prior work [10], including prosody (energy, pitch), speaking duration, and additional features detailed below. We also define derived features that indicate how balanced a person's participation is. Specifically, we defined a person's speech share, sp_r^i, as the relative participation with respect to all other participants. We also define participation unevenness [8] as $\text{uneven}^i = \text{sp}_r^i - \frac{1}{|P|}$ and compute the unevenness of the group in general as $\text{uneven}_g = \sum_i |\text{uneven}^i|$. We normalize and reverse the measure to compute a metric of group balance $\text{bal}_g = 1 - \frac{\text{uneven}_g}{\max(\text{uneven}_g)}$. Balance from the perspective of p_i is then defined as $\text{bal}^i = \frac{1}{|P|} - \frac{|\text{uneven}^i|}{\max(\text{uneven}_g)}$.

Features of each turn were collected as the graph $\mathcal{G} = (u, V, E)$. A node contained a summary of features from a human group member over the past turn, namely the following thirteen features: balance from an individual participants' perspective, bal^i; relative amount of speaking for the current topic, sp_r^i; percent of time the participant has been participating in the last turn; percent of time the robot looked at the participant in the last turn; whether the robot is currently looking at the person; mean, standard deviation, minimum, and maximum of pitch and energy.

Edges contained five features: mean, standard deviation, minimum, and maximum head angle between participants; euclidean distance between the participants. The global graph attribute, u, summarized features of the robot not specific to individual participants: the percentage of time the robot changed gaze over the last turn; overall group balance, bal_g.

Post-annotations: We post-annotated the dataset with an ideal and simple rule set that was inspired by heuristics from the interaction-shaping robotics community. This heuristic chooses p^{who} as the person with the lowest speech share, sp_r^i, that did not speak in the last turn. "What" the robot should do when addressing p^{who} is chosen based on the relative speaking amounts of p^{who} and the pitch of the last speaker in the following manner: If the speech share of p^{who} was at least 90% of a balanced interaction, that is $\text{sp}_r^i \geq 0.9 * 1/|P|$, the robot should choose gaze, Action (1), if the pitch of the last speaker was high[4] and ask an opinion question, Action (2), if the pitch was low. If the speech share, sp_r^i, of the target person was low and the pitch of the last speaker was low, the robot should ask a concerns question, Action (4), and ask an ideas question, Action (3), if the pitch was high.

B. The teenager dataset

The goal of this dataset collection was to obtain demonstrations on how a robot could enable "better" group interactions among a group of teenagers. The dataset was kindly provided as an anonymized comma-separated file by Gillet et al. [15], who conducted the original study and collected the dataset.

The task: Three teenagers worked on a discussion-based task with the robot acting as a facilitator. The robot was controlled through a tablet by a fourth teenager who decided whom to address and which robot action to take to achieve the goal. The interaction is visualized in Figure 1 (right).

Participants: 8 boys and 8 girls of age 12-15 (M = 12.8).

Size of the dataset: The dataset contains 48 interactions each 15 minutes long. Teenagers split into three groups and rotated to either work on the group task or control the robot. Within these 48 interactions, teenagers made 2654 decisions.

Actions space: To control the robot, teenagers used a tablet interface displaying people and the available "what" actions (see [15], Fig. 3 for an illustration). They chose an addressee and how the addressee should be addressed before clicking a 'Send' button. For the training process, we joined two actions, i.e., the prompt discussion and ask opinion actions due to their similar nature which resulted in the following action space: (1) Prompt discussion/ask opinion to $\underline{\;p^{who}\;}$, (2) Agree with $\underline{\;p^{who}\;}$, (3) Disagree with $\underline{\;p^{who}\;}$, (4) Ask $\underline{\;p^{who}\;}$ to elaborate, (5) Nod looking at $\underline{\;p^{who}\;}$, (6) Look at $\underline{\;p^{who}\;}$, (7) Praise $\underline{\;p^{who}\;}$, (8) Tell joke toward $\underline{\;p^{who}\;}$, (9) Ask $\underline{\;p^{who}\;}$ to focus, (10) Remind $\underline{\;p^{who}\;}$ to cooperate.

In addition to addressing individuals, teenagers could address the whole group by not choosing a specific group member.

Feature extraction - Graph state space: The dataset is comprised of continuous audio data from each individual captured at 2Hz through individual lapel microphones. The audio stream has 37 features[5] for each human group member and two general features describing the group interaction. We represent the provided data as a graph, $\mathcal{G} = (u, V, E)$.

We ran a backward feature selection and selected the features used in 75% of the top 5% models: mean and standard deviation of energy and pitch, the first derivative of the pitch. A sequence of these five features is used for the nodes, V.

In this dataset, edges had no features and were solely used to connect the individual nodes. The global graph feature, u was unevenness in participation uneven_g and the time since the the robot took the last action. The concrete utterances the robot used upon selecting an action are listed in Appendix A2.

V. Experiments

To evaluate TGM in groups of varying sizes and with varying complexity in demonstrations, we use the two different datasets outlined in Section IV. With these datasets, we explore the following research questions:

RQ1. *Can we use TGM to model the decision-making process in groups of varying sizes and outperform the linear baselines on a realistic dataset with ideal demonstrations?*

[4] We used the mean of the dataset to determine high and low pitch values.

[5] Loudness of the speech in dB, accumulated relative amount of speech compared to all speech in the group, and whether the human group member is currently speaking, mean and standard deviation over the past second (34 features) of: 13-dimensional mel-frequency cepstrum coefficients (MFCC, every 25ms, sliding hamming window: 40ms), 4-dimensional prosody features (speech intensity through yin-energy, pitch through the fundamental frequency as well as the first derivative of these features).

Human groups naturally occur in varying sizes. For instance, small groups of 2 or 3 people might be common in the home [54] while groups are larger in public environments [55] or educational settings [56]. Further, people might dynamically join and leave human-robot groups [32], [57]. Therefore, TGM needs to handle different group sizes and generalize to group sizes unseen during training. This prompted us to ask:

RQ2. *Can TGM enable the learned behaviors to generalize to group sizes unseen during training?*

In a data ablation study, we further explore how the availability of different group sizes in the training data influences the ability of the model to generalize. Therefore, we also investigate:

RQ3. *How do the size of the idealized dataset and diversity of group sizes impact generalization to unseen group sizes?*

We use the teenager dataset in which group sizes are fixed to three but which offers complex human demonstrations to extend on ***RQ1*** and explore:

RQ4. *Is TGM beneficial when cloning complex human demonstrations?*

A. Training on the Brainstorming dataset

To answer *RQ1* and *RQ2*, we split the brainstorming data by full session/groups into train, validation and two test sets. We first separated the single group of four as a test set for *RQ2*, generalization to unseen group sizes. 10% of dyads and triads then formed the second test set and 16% of the remaining data the validation set. We normalized based on the training data and applied the normalization to validation and test data.

We trained all models on a CPU and seeded the process to reduce the variability between runs. We performed hyperparameter searches for our proposed models and our baselines. We first evaluated each hyperparameter combination on the average macro F1 over three seeds and then verified the stability of the hyperparameter combination by training the 10 best models on 7 additional seeds for a total of ten different seeds. We report results from the model with the highest average macro F1 over the ten seeds. We used mean-squared error loss and tuned the learning rate in the first step of training where appropriate. To answer *RQ3*, we constructed three different subsets of training data as depicted in Table I. Since the size of this dataset is smaller, we ran six-fold cross-validation on each hyperparameter set. The code and exemplary commands are available on https://github.com/sarahgillet/TGM-SmallGroups.

1) Training TGM: To train TGM, we used Stochastic Gradient Descent (SGD) [58] to train one- or two-layer GNNs. The hyperparameter search explored different MLP architectures for the respective message-updating ϕ^e and node-updating neural networks ϕ^v and other parameters detailed below. $\phi^{v\rightarrow y^{\text{who}}}$ and $\phi^{u,v\rightarrow y^{\text{what}}}$ were implemented through one dense layer and we used min-pooling for the aggregation functions. We applied a softmax before computing the loss.

TABLE I
OVERVIEW ON THE DIFFERENT SETS OF DATA AND THEIR SPLIT USED TO EVALUATE *RQ3* EXPLORING THE IMPORTANCE OF THE DIVERSITY OF THE DATASET. DYADS REFER TO GROUPS OF 2 AND TRIADS TO GROUPS OF 3.

	Train	Validation	Test
Set Dyad	10 dyads	2 dyads	12 triads, 1 group of 4
Set Triad	10 triads	2 triads	18 dyads, 1 group of 4
Set Mix	5 dyads, 5 triads	1 dyad, 1 triad	6 triads, 12 dyads, 1 group of 4

We explored a different number of hidden layers (1,2), batch sizes (32,64,128), dropout in the MLPs (0.2,0.5), and the number of weights per hidden layer (8-16-4, 4-8, 8-16 for node- and message update on the first layer, 8-4, 4-8, 8-16 for message update on the second layer, 2-4, 4-8, 4-2 for node update on the second layer, all but the last numbers indicate hidden layer sizes, the last number indicates the number of outputs of the specific ϕ). Further, we use dropout between layers and ReLUs as activations. The hyperparameters were chosen after initial sparse exploration of the hyperparameters.

2) Training the baselines: The brainstorming dataset consists of groups of different sizes which limited us to explore the group size-independent linear input modeling baseline. We explored two machine learning techniques that use vectors as inputs instead of a graph: MLPs and RFs.

We explored different numbers of hidden layers and sizes (8-16, 4-8, 16-4, 8-4, 4-16, 16-8, 32-4, 8-2-4, 8, 4, 4-8-4), different percentages of dropout between layers (0.2, 0.5) and batch sizes (32, 64). For RFs[6], we used different numbers of estimators (100, 200, 300, 400, 500), allowed maximum depths (automatic, 10, 20, 30, 40, 50), minimum numbers of samples for a split (2, 5, 10) and for a leaf (1, 2, 4), maximum number of features used (sqrt, log2), if bootstraping is used (True, False) and the splitting criterion (gini impurity, entropy).

B. Training on the teenager dataset

To answer *RQ4*, we trained TGM on the teenager dataset as well as two baselines, the group-size independent and the group-size dependent linear input modeling baseline.

We split the data for training into the train, validation, and test data. Given the complexity and small size of the dataset, we reduced the size of validation and test data compared to the brainstorming dataset and use 6% of random instances from the sessions as test data and 10% as validation data. Note that all sets contained instances from all groups and sessions.

We trained all models on a CPU and seeded the process to reduce the variability between runs. Though we trained models on this dataset with multiple seeds, we observed high variability. This variability might be a result of the small size of the dataset and the complexity of demonstrations. Therefore, we only report the results from a single seed chosen randomly before any training was performed. We report results from the model with the highest macro average F1 score on the validation set. The code is available on https://github.com/sarahgillet/TGM-SmallGroups.

1) Training TGM: To train TGM, we used Decoupled Weight Decay Regularization (AdamW) [59] to train one- or two-layer GNNs. We explore different LSTM architectures for

[6] We use the implementation in scikit-learn: https://scikit-learn.org/

TABLE II

Results for *RQ1* and *RQ2* evaluating TGM in comparison to the linear baslines for group size seen under training, dyads and triads, and group sizes unseen under training, groups of four. All values are mean ± std of the **macro average F1 score** over 10 different seeds.

	Test - dyads, triads (RQ1)		Test - group of four (RQ2)	
	ξ_{who}	π_{what}	ξ_{who}	π_{what}
Chance	$.438 \pm .049$	$.25$	$.25$	$.25$
RF	$.780 \pm .128$	$.185 \pm .143$	$.701 \pm .040$	$.311 \pm .04$
MLP	$.582 \pm .135$	$.181 \pm .083$	$.661 \pm .099$	$.526 \pm .066$
TGM	$\mathbf{.900 \pm .046}$	$\mathbf{.890 \pm .057}$	$\mathbf{.764 \pm .049}$	$\mathbf{.726 \pm .066}$

the respective message-updating ϕ^e and node-updating neural networks ϕ^v to accommodate the sequential nature of the data. We use min-pooling for the aggregation functions.

As hyperparameters, we explored the number of features in the hidden state of the LSTM (2, 4, 8, 16 for node update on the first layer and message update on the second layer; 4, 8, 16, 32 for message update on the first layer; 2, 4, 8 for node update on the second layer), batch sizes (32,64,128), $\phi^{v \rightarrow y^{\text{who}}}$ and $\phi^{u,v \rightarrow y^{\text{what}}}$ were implemented through one dense layer. We applied a softmax before computing the loss.

2) Training the baselines for the teenager dataset: We used AdamW to train one or two layers of LSTMs followed by an MLP. For both baselines, we explored one or two layers of LSTMs as well as the number of features in the hidden state of these LSTMs (4, 8, 16, 32, 64 for the first layer; 2, 4, 8, 16, 32 for the second layer). We also explored different architectures of the last layer MLP with one or two hidden layers (2, 4, 8, 16, 4-8, 4-16, 8-16, 8-2, 16-4, 16-8, 32-4). We further explored different batch sizes (32, 64, 128).

C. Results

To answer ***RQ1*** and ***RQ2***, we first explore the performance of TGM and the baselines on the full brainstorming dataset trained as outlined in V-A. The results with respect to the best model's macro average F1 score are presented in Table II with full details (validation set, hyperparameters) in Appendix B1.

The first two columns in Table II present the results for ***RQ1*** and show that TGM outperforms the baselines when generalizing to new groups of seen group sizes (groups of 2 and 3 people) on both tasks, i.e., for deciding "whom" and "what". Results relevant for ***RQ2*** are presented in the latter two columns of Table II. These results further demonstrate that TGM outperforms the baselines when generalizing to group sizes unseen during training for deciding "whom" and "what".

Toward understanding ***RQ3***, we also trained on subsets of the data as detailed in Table I. We kept the dataset size constant between group sizes to be able to differentiate between the diversity of the data, i.e., only groups of 2, only groups of 3, and groups of 2 and 3, without conflating it with the importance of the size of the dataset. We ran cross-validation to counteract the small size of the dataset. The last row of Table III shows that when training on a smaller dataset that has the same diversity as the full dataset, i.e., groups of 2 *and* 3, the general trend of performance between approaches holds.

TABLE III

Results for *RQ2* training π_{who} and π_{what} on limited data. Models were trained by limiting the dataset size and/or group size, and tested on two test sets with groups of sizes unseen under training. We report the mean ± std of the macro average F1 score over 10 seeds and 6 folds.

Train set	Model	Test 'Dyads, Triads'		Test - Group of 4	
		ξ_{who}	π_{what}	ξ_{who}	π_{what}
		'Triads'			
'Dyad'	Chance	$.33$	$.25$	$.25$	$.25$
	RF	$\mathbf{.50 \pm .06}$	$.17 \pm .02$	$\mathbf{.33 \pm .05}$	$.30 \pm .02$
	MLP	$.35 \pm .03$	$.15 \pm .07$	$.27 \pm .04$	$\mathbf{.36 \pm .10}$
	TGM	$.44 \pm .04$	$\mathbf{.41 \pm .05}$	$.23 \pm .05$	$\mathbf{.36 \pm .09}$
		'Dyads'			
'Triad'	Chance	$.5$	$.25$	$.25$	$.25$
	RF	$\mathbf{.84 \pm .02}$	$.08 \pm 0$	$\mathbf{.72 \pm .05}$	$.29 \pm .01$
	MLP	$.65 \pm .04$	$.14 \pm .03$	$.69 \pm .05$	$.44 \pm .06$
	TGM	$\mathbf{.84 \pm .03}$	$\mathbf{.45 \pm .11}$	$\mathbf{.72 \pm .05}$	$\mathbf{.67 \pm .12}$
		'Mixed Dyads, Triads'			
'Mix'	Chance	$.415$	$.25$	$.25$	$.25$
	RF	$.70 \pm .04$	$.22 \pm .03$	$.61 \pm .08$	$.29 \pm .02$
	MLP	$.46 \pm .06$	$.19 \pm .04$	$.54 \pm .07$	$.45 \pm .06$
	TGM	$\mathbf{.83 \pm .02}$	$\mathbf{.60 \pm .09}$	$\mathbf{.67 \pm .09}$	$\mathbf{.57 \pm .12}$

TABLE IV

Results for *RQ4* from training the full hyperparameter search for the teenager dataset on seed 0. We report the macro average F1 score on the validation and test set.

	Validation		Test	
	ξ_{who}	π_{what}	ξ_{who}	π_{what}
Chance	.25	.1	.25	.1
Linear - $s_{\text{dependent}}$	.274	.109	.282	.084
Linear - $s_{\text{independent}}$	.283	.108	.211	.113
TGM	**.351**	**.243**	**.306**	**.175**

That is, TGM outperforms the linear baseline on both tasks when generalizing to seen and unseen group sizes.

Training on the less diverse datasets, i.e., Dyad and Triad, shows that overall generalization is less successful when only seeing dyads, i.e., groups of two. When choosing "whom", RFs seem advantageous for capturing the simple rules used for post-annotations in this dataset. However, the more complex rule set used for deciding "what" the robot should do is captured more successfully by TGM and outperforms the baselines in all cases but one exception for MLPs (generalizing to groups of 4 after training on the Dyad set). Overall, we can conclude that the set 'Dyad' as the data subset with the smallest group size does not allow for generalization in any of the explored models, as indicated by the low performance.

For exploring ***RQ4***, we trained TGM and the two linear baselines using a group size-independent or a group size-dependent linear input on the teenager dataset as described in Section V-B. Table IV shows that TGM outperforms both baselines on the teenager dataset (more details in Appendix B2).

VI. Discussion

This paper proposes TGM to enable robots to learn effectively to interact with small groups of varying sizes. Answering ***RQ1***, our results demonstrate that TGM can serve for learning from demonstrations and outperform the baselines. All methods can generalize to group sizes unseen during training (***RQ2***) but TGM outperforms the baselines.

RQ3 explored how diversity and dataset size would influence TGM and the baselines. In general, less diversity and the smaller size of the dataset reduce the overall performance with the exception of deciding "whom" trained on only triads. The simple rule-based demonstrations seem to be well captured through groups of three leading to higher performance specifically for condition-based methods such as the RF. At the same time, the rule set used to decide "what" is poorly captured by the baselines with lower-than-chance performance. The rules used for "what" use the pitch of the previous speaker whose data the baselines can only access in the group context, i.e., the aggregated features of all other group members. This aggregation was necessary to allow for a group size agnostic baseline, however, it leads to the data of the previous speaker being entangled with the data of other group members. TGM is group size agnostic without the need to aggregate the data and can, therefore, better capture the decision of "what".

To answer ***RQ4***, our results indicate rather low performance. We note, though, that the teenager dataset is small given the expected complexity of demonstrated behaviors by the fifteen different wizards (teenagers formed groups of five rotating the robot controller). We hypothesize that TGM would be advantageous if the dataset was larger. We base this hypothesis on the exploration for ***RQ3***. For ***RQ3***, we found that limiting the dataset to smaller subsets of the brainstorming dataset resulted, as expected, in lower performance than using the larger full dataset. However, the general trend of performance between TGM and the baselines remained the same, TGM outperforms the baselines also on smaller subsets of the dataset especially for the more complex "what" decision. Therefore, the trend between methods for the teenager dataset is likely to scale, encouraging future work to collect larger datasets with varying groups of people and group sizes beyond four people.

We used an expert heuristic to post-annotate the brainstorming dataset for sensitive facilitation of group interactions inspired by prior work [6], [8]. The post-annotations provided consistent and simple human demonstrations that match the complexity of the dataset which allowed us to explore behavior cloning. Using simple heuristics to approximate human demonstrations might be an interesting method for technical HRI research in general. These simple demonstrations allowed us to explore the promise and limitations of TGM and the baselines without the need to collect a large dataset. It enables future research to justify the collection of larger datasets given our insights on which approach might be promising.

Template-based action spaces show promise for group human-robot interaction and might also be useful for other areas of HRI. Templates have been a common approach for generating dialogue and have been used to solve interactive fiction games [40], [41]. Nonetheless, template-based actions have not been commonly used in HRI. The concept of templates allows the embedding of a variety of contexts, e.g., the context of a person in a group, in our case, into the decision-making. If the interaction setting was to require first choosing the template, e.g., as in interactive fiction games, the templates can also provide the context of the available actions to choosing the filler, e.g., objects in interactive fiction games or people in other group HRI settings. The versatility of template-based action spaces can enable future HRI research to reason more efficiently over actions taken toward people or toward objects or possibly even a combination of them.

Limitations: HRI datasets are often small, which is also a limitation of this work. This might be an additional reason for methods performing lower than chance in the brainstorming dataset. The low performance of all approaches on the teenager dataset, which is small but features complex human demonstrations, indicates that more data would be needed to achieve acceptable performance. Our results show promise and justify that future work invests resources to collect larger datasets.

Another limitation of this work is the lack of standardized benchmarks. Unfortunately, there are no benchmarks nor adopted common baselines for evaluating robot policies in group HRI. We further did not evaluate our policy in interaction with people. To approximate how the robot would act in new groups of people, we extracted features the same way as during a deployment and constructed our test sets to contain groups unseen under training. A real-world deployment in a user study would further face the challenge of measuring performance. Measuring performance through the effectiveness of the robot's behavior would likely be confounded with the design of the heuristic that provided the post-annotation. Future work should therefore explore which measures other than accuracies or F1 scores could be used.

VII. Conclusion

In this work, we propose a Template- and Graph-Based Modeling approach (TGM) for social robot decision-making in groups enabling robots to interact in various real-world groups of varying sizes. TGM allows a trained model to be agnostic to the group size by using a fixed-size template-based action space which separates the decision into "what" to do and "whom" to address. In addition, we model the group as a graph. We evaluated TGM by comparing structured reasoning through GNNs with linear models. Given consistent and simple human demonstrations, TGM outperformed the baselines and led to results that appeared sufficient for deploying the learned policies in new group interactions, even of different sizes. Future work will need to collect a larger dataset of human demonstrations and evaluate if the robot's policy can follow the demonstrations close enough in real-world interactions. Our results show promise to allow for robots to support people's complex everyday interactions.

Acknowledgment

This work was supported by the Swedish Foundation for Strategic Research (SSF FFL18-0199), the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, and the National Science Foundation (Grant No. IIS-2143109).

References

[1] S. Sebo, B. Stoll, B. Scassellati, and M. F. Jung, "Robots in Groups and Teams: A Literature Review," *Proc. ACM Hum.-Comput*, vol. 4, no. October, p. 37, 2020.

[2] E. Schneiders, E. Cheon, J. Kjeldskov, M. Rehm, and M. B. Skov, "Non-dyadic interaction: A literature review of 15 years of human-robot interaction conference publications," *J. Hum.-Robot Interact.*, vol. 11, no. 2, 2022.

[3] E. Mizrahi, N. Danzig, and G. Gordon, "vrobotator: A virtual robot facilitator of small group discussions for k-12," *Proc. ACM Hum.-Comput. Interact.*, vol. 6, no. CSCW2, Nov. 2022.

[4] E. S. Short, K. Swift-Spong, H. Shim, K. M. Wisniewski, D. K. Zak, S. Wu, E. Zelinski, and M. J. Mataric, "Understanding social interactions with socially assistive robotics in intergenerational family groups," *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 236–241, 2017.

[5] M. F. Jung, N. Martelaro, and P. J. Hinds, "Using Robots to Moderate Team Conflict: The Case of Repairing Violations," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. Portland, Oregon, USA: Association for Computing Machinery, 2015, p. 229–236.

[6] S. Gillet, W. van den Bos, and I. Leite, "A social robot mediator to foster collaboration and inclusion among children," in *Proceedings of Robotics: Science and Systems*, Corvalis, Oregon, USA, July 2020.

[7] S. Gillet, R. Cumbal, A. Pereira, J. Lopes, O. Engwall, and I. Leite, "Robot Gaze Can Mediate Participation Imbalance in Groups with Different Skill Levels," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. New York, NY, USA: ACM, 3 2021.

[8] H. Tennent, S. Shen, and M. Jung, "Micbot: a peripheral robotic object to shape conversational dynamics and team performance," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2019, pp. 133–142.

[9] A. Shamekhi and T. W. Bickmore, "A multimodal robot-driven meeting facilitation system for group decision-making sessions," *ICMI 2019 - Proceedings of the 2019 International Conference on Multimodal Interaction*, pp. 279–290, 2019.

[10] S. Gillet, M. T. Parreira, M. Vázquez, and I. Leite, "Learning gaze behaviors for balancing participation in group human-robot interactions," in *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '22. IEEE Press, 2022, p. 265–274.

[11] S. Gillet, M. Vázquez, S. Andrist, I. Leite, and S. Sebo, "Interaction-shaping robotics: Robots that influence interactions between other agents," *J. Hum.-Robot Interact.*, vol. 13, no. 1, mar 2024.

[12] H. Javed, W. Wang, A. B. Usman, and N. Jamali, "Modeling interpersonal perception in dyadic interactions: towards robot-assisted social mediation in the real world," *Frontiers in Robotics and AI*, vol. 11, 2024. [Online]. Available: https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2024.1410957

[13] D. Utami, T. W. Bickmore, and L. J. Kruger, "A robotic couples counselor for promoting positive communication," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2017, pp. 248–255.

[14] T. V. Pham, T. H. Weisswange, and M. Hassenzahl, "Embodied mediation in group ideation – a gestural robot can facilitate consensus-building," in *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, ser. DIS '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 2611–2632. [Online]. Available: https://doi.org/10.1145/3643834.3660696

[15] S. Gillet, K. Winkle, G. Belgiovine, and I. Leite, "Ice-Breakers, Turn-Takers and Fun-Makers: Exploring Robots for Groups with Teenagers," in *31st IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 8 2022.

[16] I. Neto, F. Correia, F. Rocha, P. Piedade, A. Paiva, and H. Nicolau, "The robot made us hear each other: Fostering inclusive conversations among mixed-visual ability children," in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 2023, pp. 13–23.

[17] R. Gomez, D. Szapiro, S. Cooper, N. Bougria, G. Pérez, E. Nichols, J. Giménez-Figueroa, J. M. Perez-Moleron, M. Peavy, D. Serrano, and L. Merino, "Design of embodied mediator haru for remote cross cultural communication," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 5505–5511.

[18] O. Gvirsman and G. Gordon, "Effect of social robot's role and behavior on parent-toddler interaction," in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 222–230.

[19] C. Birmingham, Z. Hu, K. Mahajan, E. Reber, and M. J. Matarić, "Can i trust you? a user study of robot mediation of a support group," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 8019–8026.

[20] G. Hoffman, O. Zuckerman, G. Hirschberger, M. Luria, and T. Shani Sherman, "Design and evaluation of a peripheral robotic conversation companion," in *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, 2015, pp. 3–10.

[21] B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, and N. Hagita, "Footing in human-robot conversations: how robots might shape participant roles using gaze cues," in *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. ACM, 2009, pp. 61–68.

[22] N. Mirnig, A. Weiss, G. Skantze, S. Al Moubayed, J. Gustafson, J. Beskow, B. Granström, and M. Tscheligi, "Face-to-face with a robot: What do we actually talk about?" *International Journal of Humanoid Robotics*, vol. 10, no. 01, p. 1350011, 2013.

[23] M. L. Traeger, S. Strohkorb Sebo, M. Jung, B. Scassellati, and N. A. Christakis, "Vulnerable robots positively shape human conversational dynamics in a human–robot team," *Proceedings of the National Academy of Sciences*, vol. 117, no. 12, pp. 6370–6375, 2020.

[24] S. Reig, M. Luria, J. Z. Wang, D. Oltman, E. J. Carter, A. Steinfeld, J. Forlizzi, and J. Zimmerman, "Not some random agent: Multi-person interaction with a personalizing service robot," in *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, 2020, pp. 289–297.

[25] M. Vázquez, A. May, A. Steinfeld, and W.-H. Chen, "A deceptive robot referee in a multiplayer gaming environment," in *2011 international conference on collaboration technologies and systems (CTS)*. IEEE, 2011, pp. 204–211.

[26] M. Vázquez, E. J. Carter, J. A. Vaz, J. Forlizzi, A. Steinfeld, and S. E. Hudson, "Social group interactions in a role-playing game," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, 2015, pp. 9–10.

[27] F. Correia, S. F. Mascarenhas, S. Gomes, P. Arriaga, I. Leite, R. Prada, F. S. Melo, and A. Paiva, "Exploring prosociality in human-robot teams," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2019, pp. 143–151.

[28] J. Connolly, V. Mocz, N. Salomons, J. Valdez, N. Tsoi, B. Scassellati, and M. Vázquez, "Prompting prosocial human interventions in response to robot mistreatment," in *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, 2020, pp. 211–220.

[29] M. F. Jung, D. Difranzo, S. Shen, B. Stoll, H. Claure, and A. Lawrence, "Robot-assisted tower construction—a method to study the impact of a robot's allocation behavior on interpersonal dynamics and collaboration in groups," vol. 10, no. 1, oct 2020.

[30] J. Nasir, B. Bruno, and P. Dillenbourg, "Social robots as skilled ignorant peers for supporting learning," *Frontiers in Robotics and AI*, vol. 11, 2024. [Online]. Available: https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2024.1385780

[31] D. Bohus, S. Andrist, and E. Horvitz, "A study in scene shaping: Adjusting f-formations in the wild," in *2017 AAAI Fall Symposium: Natural Communication for Human-Robot Collaboration*, September 2017.

[32] M. Vázquez, E. J. Carter, B. McDorman, J. Forlizzi, A. Steinfeld, and S. E. Hudson, "Towards robot autonomy in group conversations: Understanding the effects of body orientation and gaze," in *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*. IEEE, 2017, pp. 42–52.

[33] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and autonomous systems*, vol. 57, no. 5, pp. 469–483, 2009.

[34] A. Billard and D. Grollman, "Robot learning by demonstration," *Scholarpedia*, vol. 8, no. 12, p. 3824, 2013.

[35] B. Akgun, M. Cakmak, J. W. Yoo, and A. L. Thomaz, "Trajectories and keyframes for kinesthetic teaching: A human-robot interaction perspective," in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 2012, pp. 391–398.

[36] C. L. Mueller and B. Hayes, "Safe and robust robot learning from demonstration through conceptual constraints," in *Companion of the*

2020 ACM/IEEE International Conference on Human-Robot Interaction, 2020, pp. 588–590.

[37] V. Jain, M. Leekha, R. R. Shah, and J. Shukla, "Exploring semi-supervised learning for predicting listener backchannels," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–12.

[38] M. T. Parreira, S. Gillet, and I. Leite, "Robot duck debugging: Can attentive listening improve problem solving?" in *Proceedings of the 25th International Conference on Multimodal Interaction*, ser. ICMI '23. New York, NY, USA: Association for Computing Machinery, Sep. 2023, p. 527–536.

[39] M. Doering, D. F. Glas, and H. Ishiguro, "Modeling interaction structure for robot imitation learning of human social behavior," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 3, pp. 219–231, 2019.

[40] M. Hausknecht, P. Ammanabrolu, M.-A. Côté, and X. Yuan, "Interactive fiction games: A colossal adventure," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 7903–7910.

[41] X. Peng, M. Riedl, and P. Ammanabrolu, "Inherently explainable reinforcement learning in natural language," *Advances in Neural Information Processing Systems*, vol. 35, pp. 16 178–16 190, 2022.

[42] G. LeMasurier, A. Gautam, Z. Han, J. W. Crandall, and H. A. Yanco, "Reactive or proactive? how robots should explain failures," in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 413–422.

[43] M. Mitchell, D. Bohus, and E. Kamar, "Crowdsourcing language generation templates for dialogue systems," in *Proceedings of the INLG and SIGDIAL 2014 Joint Session*, 2014, pp. 172–180.

[44] F. Yang, W. Yin, T. Inamura, M. Björkman, and C. Peters, "Group behavior recognition using attention-and graph-based neural networks," in *ECAI 2020*. IOS Press, 2020, pp. 1626–1633.

[45] G. Sharma, K. Stefanov, A. Dhall, and J. Cai, "Graph-based group modelling for backchannel detection," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 7190–7194.

[46] J. Taery Kim, A. Naik, I. Jayarathne, S. Ha, and J. Y. Chew, "Modeling social interaction dynamics using temporal graph networks," *arXiv e-prints*, p. arXiv:2404.06611, Apr. 2024.

[47] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner *et al.*, "Relational inductive biases, deep learning, and graph networks," *arXiv preprint arXiv:1806.01261*, 2018.

[48] M. Vázquez, A. Lew, E. Gorevoy, and J. Connolly, "Pose generation for social robots in conversational group formations," *Frontiers in Robotics and AI*, p. 341, 2022.

[49] S. Thompson, A. Gupta, A. W. Gupta, A. Chen, and M. Vázquez, "Conversational group detection with graph neural networks," in *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 248–252.

[50] D. Pomerleau, "An autonomous land vehicle in a neural network," *Advances in Neural Information Processing Systems; Morgan Kaufmann Publishers Inc.: Burlington, MA, USA*, 1998.

[51] M. Swofford, J. Peruzzi, N. Tsoi, S. Thompson, R. Martín-Martín, S. Savarese, and M. Vázquez, "Improving social awareness through dante: Deep affinity network for clustering conversational interactants," *Proc. ACM Hum.-Comput. Interact.*, vol. 4, no. CSCW1, may 2020.

[52] I. Leite, M. McCoy, D. Ullman, N. Salomons, and B. Scassellati, "Comparing models of disengagement in individual and group interactions," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 99–105.

[53] G. Hadjiantonis, S. Gillet, M. Vázquez, I. Leite, and F. I. Dogan, "Let's move on: Topic change in robot-facilitated group discussions," in *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, 2024, pp. 2087–2094.

[54] B. Scassellati, L. Boccanfuso, C.-M. Huang, M. Mademtzi, M. Qin, N. Salomons, P. Ventola, and F. Shic, "Improving social skills in children with asd using a long-term, in-home social robot," *Science Robotics*, vol. 3, no. 21, p. eaat7544, 2018.

[55] L. Moshkina, S. Trickett, and J. G. Trafton, "Social engagement in public places: a tale of one robot," in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, 2014, pp. 382–389.

[56] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro, "Interactive robots as social partners and peer tutors for children: A field trial," *Human–Computer Interaction*, vol. 19, no. 1-2, pp. 61–84, 2004.

[57] D. Bohus and E. Horvitz, "Managing human-robot engagement with forecasts and... um... hesitations," in *Proceedings of the 16th international conference on multimodal interaction*, 2014, pp. 2–9.

[58] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, Y. Lechevallier and G. Saporta, Eds. Paris, France: Springer, August 2010, pp. 177–187.

[59] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

Appendix

A. Details on the data collection

In the following we provide the concrete utterances that the robot could choose from when selecting one of the actions. We repeat the action space for convenience. The full data collection protocol can be found at https://github.com/sarahgillet/TGM-SmallGroups.

1) The brainstorming dataset:

Action space: With a previously chosen "who" denoted with p_{who}, the following multimodal action templates describe "how" the chosen person should be addressed throughout one turn:

1) Look more at $\underline{\quad p_{who} \quad}$
   No speech, Gaze between $p \in P$ with focus on p_{who}
2) Ask opinion question to $\underline{\quad p_{who} \quad}$
   Ask question, gaze at p_{who}
3) Ask more ideas question to $\underline{\quad p_{who} \quad}$
   Ask question, gaze at p_{who}
4) Ask concerns question to $\underline{\quad p_{who} \quad}$
   Ask question, gaze at p_{who}

Implementation: The system was realized in the Robot Operating System (ROS) and the behavior of the robot was controlled by a behavior tree. When the behavior tree reached the decision-making node, the robot would first choose the action and then select the respective concrete sentence the robot would say from a list of available options. To avoid repetitiveness, the robot would select each option once before shuffling the available options to select from them again.

Options for action (2) Ask opinion question to $\underline{\quad p_{who} \quad}$

- What do you think?
- How do you feel about that?
- Do you like that idea?
- Do you like that idea?
- Does that sound good to you?
- Do you agree with that?
- Does that make sense to you?
- What's your take on this?
- Do you feel the same way?
- Do you have the same opinion?
- Whats your opinion on this?
- Do you think this is a good idea?
- Do you feel this is a good idea?
- How do you like this?

Options for action (3) Ask more ideas question to $\underline{\quad p_{who} \quad}$

- Do you have other ideas to share?
- Can you think of something more?
- Could you share something more that excites you when thinking about this question?

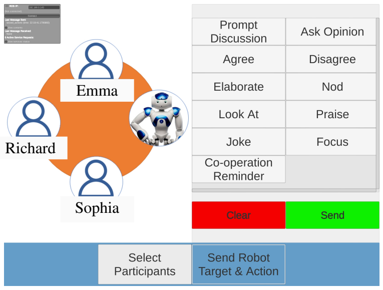

Fig. 3. The tablet interface that teenagers used to control the robot and provide demonstrations for "whom" and "what".

- Is there any other point you would like to make?
- Anything else you can think off?
- Do you have a different idea?
- Do you have something else on your mind?
- Anything else that comes to your mind?
- Any other point that comes to your mind?

Options for action (4) Ask concerns question to $\underline{\quad p_{who} \quad}$

- Is there anything that you would recommend to avoid?
- Do you feel there could be problems with that?
- Do you think there could be any issues with that?
- I wonder if there are any cases where this is unsuitable. What do you think?
- Is there anything that someone could find concerning?

2) The teenager dataset:

Actions space: To control the robot, teenagers used a tablet interface illustrated in Figure 3 which also displays the available "what" actions. They had to chose an addressee from the table on the left side of the screen and how the addressee should be addressed from the table on the right side of the screen before clicking the 'Send' button. We joined the prompt discussion and ask opinion actions due to their similar nature which resulted in the following action space: (1) Prompt discussion/ ask opinion to $\underline{\quad p_{who} \quad}$, (2) Agree with $\underline{\quad p_{who} \quad}$, (3) Disagree with $\underline{\quad p_{who} \quad}$, (4) Ask $\underline{\quad p_{who} \quad}$ to elaborate, (5) Nod looking at $\underline{\quad p_{who} \quad}$, (6) Look at $\underline{\quad p_{who} \quad}$, (7) Praise $\underline{\quad p_{who} \quad}$, (8) Tell joke toward $\underline{\quad p_{who} \quad}$, (9) Ask $\underline{\quad p_{who} \quad}$ to focus, (10) Remind $\underline{\quad p_{who} \quad}$ to cooperate.

Implementation: Options for action (1) Prompt discussion/ ask opinion to $\underline{\quad p_{who} \quad}$:

- Do you have an opinion on this matter?
- What about you?
- What do you think?
- How do you feel about this?
- You haven't spoken so much yet, do you have a different opinion?
- Great question! Let's talk about this.

- Let’s start the discussion.
- Can you explain more?
- Let’s discuss our opinion about this.
- I think you’ll do great! Go team!
- You are such a smart and creative group, I think you will do great!

Options for action (2) Agree with $\underline{p_{who}}$:

- I think so, too.
- I agree.
- I agree with you.
- I like this idea.
- I like this idea (Nodding).

Options for action (3) Disagree with $\underline{p_{who}}$: :

- I disagree.
- I don’t agree with this.
- I don’t agree with you.

Options for action (4) Ask $\underline{p_{who}}$ to elaborate: :

- Really? Why do you think so.
- Can you maybe explain more what you mean.
- Sounds awesome! Can you elaborate a bit more.
- I didn’t quite understand, could you elaborate a bit more.
- Great, but could you elaborate a bit more.
- Thank you for bringing that up, can you elaborate a bit more.
- You seem to be on a great track, can you explain more.
- Seems like you know a lot about this. Anything else about this that you would like to share.

Action (5) Nod looking at $\underline{p_{who}}$ does not have any variants.
Action (6) Look at $\underline{p_{who}}$ does not have any variants.
Options for action (7) Praise $\underline{p_{who}}$:

- Interesting, good job.
- Good work! Keep it up!
- Great job.
- Nice work.
- Good thinking.
- That was a really good point.
- Great work! Let’s keep going.
- Wow! Good thing you brought that up.
- That sounds awesome!
- Yes! Such a good answer.
- This is going really well! I might as well leave! Too lazy though!
- You are working so good!

Options for action (8) Tell joke toward $\underline{p_{who}}$:

- I wrote a song of a tortilla. Actually, it’s more of a wrap.
- How do you tell if a vampire is sick? (Pause) By how much is coffin!
- What do you call a pig that can do karate? (Pause) A pork chop.
- My wife left a note on the fridge saying this is not working, but when I opened the fridge it was working fine!
- Why did the robot get upset? (Pause) Because everyone was pushing its button!
- What happens to robot after they go defunct? (Pause) They rust in peace!
- Why did the robot fall in love with the magnet? (Pause) It could not resist!
- Which band do robot listen to? (Pause) Metallica!!
- You don’t need a parchude to go skydiving, you need one to go twice!

Options for action (9) Ask $\underline{p_{who}}$ to focus:

- Please can we stay focused on the activity?
- Focus people!
- Remember, focus is important!
- Okay guys, remember to focus!
- Let’s read the question again, shall we?
- You can do this! Just a bit more focus.
- Don’t forget the task!
- Get back to work!
- Stay focused!

Options for action (10) Remind $\underline{p_{who}}$ to cooperate:

- We have different opinions here, lets try and find some common ground.
- We have different opinions here, lets try and find some common ground.
- Remember that we are a team, we have to cooperate.
- Let everyone talk.
- Remember: teamwork divides the task and multiplies the success.
- Let’s listen to each other and stop interrupting.
- This is a group experience, let’s act more like it.
- Let’s compromise because you need to work together.

TABLE V

Top model hyperparameters for training TGM and the baselines on the full dataset. We repeat results from Table II for convenience.

Model - Evaluating π_{who}	Validation	Test - dyads and triads	Test - group of four
Linear - Random Forest - estimators: 100, max depth: 10, minimum samples split: 2, min samples leaf: 2, max features: sqrt, bootstrap: False, splitting criterion: gini impurity	$0.904 \pm .025$	$0.780 \pm .128$	$0.701 \pm .040$
Linear - MLP - Two hidden layer (8-16), batch size: 64, dropout: 0.2	$0.887 \pm .044$	$0.582 \pm .135$	$0.661 \pm .099$
TGM - One layer GNN; ϕ_1^e: MLP, 3 hidden layers (8-16-4); ϕ_1^v: MLP, 2 hidden layers (4-8); dropout: 0.2; batch size: 128	0.904 ± 0.034	0.900 ± 0.046	0.764 ± 0.049

Model - Evaluating π_{what}	Validation	Test - dyads and triads	Test - group of four
Linear - Random Forest - estimators: 100, max depth: 30, minimum samples split: 2, min samples leaf: 2, max features: sqrt/auto, no bootstrapping, splitting criterion: entropy	0.946 ± 0.037	0.185 ± 0.143	0.311 ± 0.04
Linear - MLP - Two hidden layer (4-16), batch size: 64, dropout: 0.2	0.883 ± 0.072	$0.181 \pm .083$	0.526±.066
TGM - Two layer GNN; ϕ_1^e: MLP, 2 hidden layers (8-16); ϕ_1^v: MLP, 2 hidden layers (8-16); ϕ_2^e: MLP, 2 hidden layers (8-16); ϕ_2^v: MLP, 2 hidden layers (4-8); dropout: 0.2; Size global: 6	0.909 ± 0.038	0.890 ± 0.057	0.726 ± 0.066

B. Detailed results

This section presents the hyperparameters of the best models with the F1 scores for TGM compared to the baselines for the respective dataset.

1) Brainstorming: Table V presents the results of training TGM and the baselines on the full dataset. Table VI and VII present the results for training the policies π_{who} and π_{what} on subsets of the dataset.

2) Teenager: The results for training on the full teenager dataset including the selected hyperparameters are given in Table VIII and IX.

TABLE VI
RESULTS FROM TRAINING π_{WHO} – AVERAGED MEAN AND STD OF THE MACRO AVERAGE F1 SCORE OVER 10 SEEDS AND 6 FOLDS AND THE RESPECTIVE HYPERPARAMETERS. WE REPEAT RESULTS FROM TABLE III FOR CONVENIENCE.

Train set	Model	Hyperparameters	Validation	Test ’Dyads and/or Triads’	Test - Groups of 4
				Test ’Triad’	
Set ’Dyad’	Linear - RF	estimators: 100, max depth: 30, minimum samples split: 2, min samples leaf: 1, max features: sqrt/auto, bootstrap: False, splitting criterion: entropy	0.95 ± 0.02	0.50 ± 0.06	0.33 ± 0.05
	Linear - MLP	2 hidden layers (16-8); batch size: 64; dropout: 0.2	0.96 ± 0.02	0.35 ± 0.03	0.27 ± 0.04
	TGM	Two layer GNN; ϕ_1^e: MLP, 2 hidden layers (8-16); ϕ_1^v: MLP, 2 hidden layers (4-8); ϕ_2^e: MLP, 2 hidden layers (4-8); ϕ_2^v: MLP, 2 hidden layers (4-8); batch size: 64; dropout 0.2	0.95 ± 0.02	0.44 ± 0.04	0.23 ± 0.05
				Test ’Dyad’	
Set ’Triad’	Linear - RF	estimators: 500, max depth: 30, minimum samples split: 2, min samples leaf: 1, max features: auto, bootstrap: False, splitting criterion: entropy	0.88 ± 0.03	0.84 ± 0.02	0.72 ± 0.05
	Linear - MLP	2 hidden layers (4-16); batch size: 64; dropout: 0.2	0.83 ± 0.05	0.65 ± 0.04	0.69 ± 0.05
	TGM	One layer GNN; ϕ_1^e: MLP, 2 hidden layers (8-16); ϕ_1^v: MLP, 2 hidden layers (16-4); batch size: 64, dropout 0.2	0.88 ± 0.04	0.84 ± 0.03	0.72 ± 0.05
				Test ’Dyad and Triad’	
Set ’Mixed’	Linear - RF	estimators: 400, max depth: 40, minimum samples split: 2, min samples leaf: 1, max features: sqrt/auto, bootstrap: False, splitting criterion: entropy	0.9 ± 0.02	0.70 ± 0.04	0.61 ± 0.08
	Linear - MLP	2 hidden layers (16-4); batch size: 64; dropout: 0.2	0.83 ± 0.5	0.46 ± 0.06	0.54 ± 0.07
	TGM	Two layer GNN; ϕ_1^e: MLP, 2 hidden layers (4-8); ϕ_1^v: MLP, 2 hidden layers (4-8); ϕ_2^e: MLP, 2 hidden layers (16-4); ϕ_2^v: MLP, 2 hidden layers (16-4); batch size: 64; dropout 0.2	0.88 ± 0.03	0.83 ± 0.02	0.67 ± 0.09

TABLE VII

Results from training π_{WHAT} – Averaged mean and std of the macro average F1 score over 10 seeds and 6 folds and the respective hyperparameters. We repeat results from Table III for convenience.

Train set	Model	Hyperparameters	Validation	Test Dyads and/or Triads	Test - Groups of 4
Set ’Dyad’	Linear - RF	estimators: 100, max depth: 10, minimum samples split: 2, min samples leaf: 2, max features: auto, bootstrap: False, splitting criterion: gini	0.99 ± 0.02	0.17 ± 0.02	0.30 ± 0.02
	Linear - MLP	2 hidden layers (4-16); batch size: 32; dropout: 0.2	0.85 ± 0.14	0.15 ± 0.07	0.36 ± 0.10
	TGM	Two layer GNN; ϕ_1^e: MLP, 2 hidden layers (4-8); ϕ_1^v: MLP, 2 hidden layers (4-8); ϕ_2^e: MLP, 2 hidden layers (4-8); ϕ_2^v: MLP, 2 hidden layers (8-16); batch size: 32; dropout 0.5	0.81 ± 0.16	0.41 ± 0.05	0.36 ± 0.09
Set ’Triad’	Linear - RF	estimators: 400, max depth: 10, minimum samples split: 2, min samples leaf: 1, max features: auto, bootstrap: True, splitting criterion: gini	0.71 ± 0.20	0.08 ± 0	0.29 ± 0.01
	Linear - MLP	2 hidden layers (8-16); batch size: 32; dropout: 0.2	0.64 ± 0.15	0.14 ± 0.03	0.44 ± 0.06
	TGM	One layer GNN; ϕ_1^e: MLP, 2 hidden layers (4-8); ϕ_1^v: MLP, 2 hidden layers (4-8); batch size: 32, dropout 0.2	0.69 ± 0.14	0.45 ± 0.11	0.67 ± 0.12
Set ’Mixed’	Linear - RF	estimators: 100, max depth: 10, minimum samples split: 5, min samples leaf: 1, max features: auto, bootstrap: False, splitting criterion: entropy	0.86 ± 0.12	0.22 ± 0.03	0.29 ± 0.02
	Linear - MLP	2 hidden layers (4-8); batch size: 32; dropout: 0.2	0.71 ± 0.17	0.19 ± 0.04	0.45 ± 0.06
	TGM	Two layer GNN; ϕ_1^e: MLP, 2 hidden layers (4-8); ϕ_1^v: MLP, 2 hidden layers (4-8); ϕ_2^e: MLP, 2 hidden layers (16-4); ϕ_2^v: MLP, 2 hidden layers (8-16); batch size: 32; dropout 0.5	0.73 ± 0.10	0.60 ± 0.09	0.57 ± 0.12

TABLE VIII

Results from training $\pi_{\mathbf{WHO}}$. We trained the full hyperparameter search for the teenager dataset on seed 0. We report the macro average F1 score on the validation and test set.

Hyperparameters		Validation	Test
Chance		0.25	0.25
Linear - $s_{\text{dependent}}$	Two layer architecture; 1. Layer: LSTM wtith 16 hidden units; 2. Layer MLP with 1 hidden layer (4); Batch size: 128	0.274	0.282
Linear - $s_{\text{independent}}$	Three layer architecture; 1. Layer LSTM with 4 hidden units; 2. Layer LSTM with 4 hidden units; 3. Layer MLP with one hidden layer (16); Batch size: 128	0.283	0.211
TGM	Two layer GNN; ϕ_1^e: LSTM, 4 hidden units; ϕ_1^v: LSTM, 16 hidden units; ϕ_2^e: LSTM, 2 hidden units; ϕ_2^v: LSTM, 2 hidden units; ϕ^u: LSTM, 6 hidden units Batch size: 64	0.351	0.306

TABLE IX

Results from training $\pi_{\mathbf{WHAT}}$. We trained the full hyperparameter search for the teenager dataset on seed 0. We report the macro average F1 score on the validation and test set.

	Hyperparameters	Validation	Test
Chance		0.1	0.1
Linear - $s_{\text{dependent}}$	Two layer architecture; 1. Layer LSTM with 4 hidden units; 2. Layer MLP with one hidden layer (16); Batch size: 128+	0.109	0.084
Linear - $s_{\text{independent}}$	Two layer architecture; 1. Layer LSTM with 8 hidden units; 2. Layer MLP with 2 layers (16-4); Batch size: 128	0.108	0.113
TGM	Two layer GNN; ϕ_1^e: LSTM with 8 hidden units; ϕ_1^v: LSTM with 16 hidden units; ϕ_2^e: LSTM, 2 hidden units; ϕ_2^v: LSTM, 4 hidden units; ϕ^u: LSTM with 6 hidden units; Batch size: 128	0.243	0.175