Social Processes: Self-supervised Meta-learning Over Conversational Groups for Forecasting Nonverbal Social Cues

Chirag Raman[1(✉)], Hayley Hung[1], and Marco Loog[1,2]

[1] Delft University of Technology, Delft, The Netherlands
{c.a.raman,h.hung,m.loog}@tudelft.nl
[2] University of Copenhagen, Copenhagen, Denmark

Abstract. Free-standing social conversations constitute a yet underexplored setting for human behavior forecasting. While the task of predicting pedestrian trajectories has received much recent attention, an intrinsic difference between these settings is how groups form and disband. Evidence from social psychology suggests that group members in a conversation explicitly self-organize to sustain the interaction by adapting to one another's behaviors. Crucially, the same individual is unlikely to adapt similarly across different groups; contextual factors such as perceived relationships, attraction, rapport, etc., influence the entire spectrum of participants' behaviors. A question arises: how can we jointly forecast the mutually dependent futures of conversation partners by modeling the dynamics unique to every group? In this paper, we propose the *Social Process* (SP) models, taking a novel meta-learning and stochastic perspective of group dynamics. Training group-specific forecasting models hinders generalization to unseen groups and is challenging given limited conversation data. In contrast, our SP models treat interaction sequences from a single group as a meta-dataset: we condition forecasts for a sequence from a given group on other observed-future sequence pairs from the same group. In this way, an SP model learns to adapt its forecasts to the unique dynamics of the interacting partners, generalizing to unseen groups in a data-efficient manner. Additionally, we first rethink the task formulation itself, motivating task requirements from social science literature that prior formulations have overlooked. For our formulation of *Social Cue Forecasting*, we evaluate the empirical performance of our SP models against both non-meta-learning and meta-learning approaches with similar assumptions. The SP models yield improved performance on synthetic and real-world behavior datasets.

Keywords: Social interactions · Nonverbal cues · Behavior forecasting

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-25066-8_37.

L. Karlinsky et al. (Eds.): ECCV 2022 Workshops, LNCS 13803, pp. 639–659, 2023.
https://doi.org/10.1007/978-3-031-25066-8_37

1 Introduction

Picture a conversing group of people in a free-standing social setting. To conduct such exchanges, we transfer high-order social signals across space and time through explicit low-level behavior cues—examples include our pose, gestures, gaze, and floor control actions [1–3]. Evidence suggests that we employ anticipation of these and other cues to navigate daily social interactions [1,4–8]. Consequently, for machines to truly develop adaptive social skills, they need to have the ability to forecast the future. For instance, foreseeing the upcoming behaviors of partners in advance can enable interactive agents to choose more fluid interaction policies [9], or contend with uncertainties in imperfect real-time inferences surrounding cues [3].

In literature, behavior forecasting works mainly consider data at two representations with an increasing level of abstraction: low-level cues or features that are extracted manually or automatically from raw audiovisual data, and manually labeled high-order events or actions. The forecasting task has primarily been formulated to predict future event or action labels from observed cues or other high-order event or action labels [5,6,9–13]. Moreover, identifying patterns predictive of certain semantic events has been a long-standing topic of focus in the social sciences, where researchers primarily employ a top-down workflow. First, the events of interest are selected for consideration. Then their relationship to preceding cues or other high-order actions are studied in isolation through

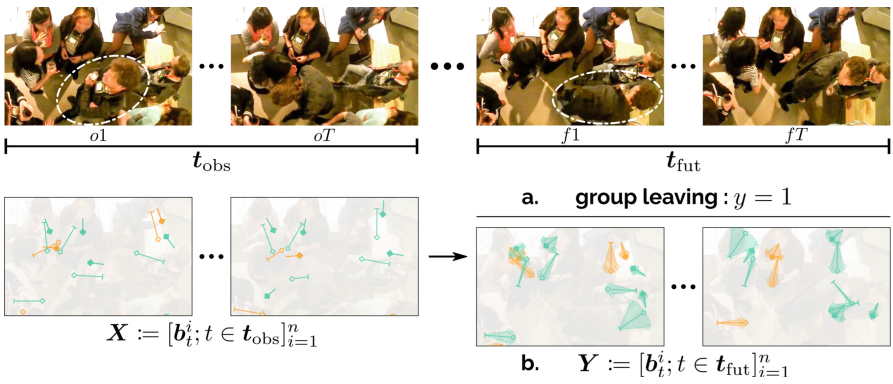

Fig. 1. Conceptual illustration of forecasting approaches on an in-the-wild conversation from the MatchNMingle dataset [16]. **Top.** A *group leaving* event [10]: the circled individual has moved from one group in the observed window $\boldsymbol{t}_{\text{obs}} := [o1 \dots oT]$ to another in a future window $\boldsymbol{t}_{\text{fut}} := [f1 \dots fT]$. **Bottom.** Input behavioral cues $\boldsymbol{b}_t^i$: head pose (solid normal), body pose (hollow normal), and speaking status (speaker in orange). **a.** The top-down approach entails predicting the event label from such cues over $\boldsymbol{t}_{\text{obs}}$, from only 200 instances of group leaving in over 90 min of interaction [10]. **b.** Our proposed bottom-up, self-supervised formulation of *Social Cue Forecasting* involves regressing a future distribution for the same low-level input cues over $\boldsymbol{t}_{\text{fut}}$ (shaded spread). This enables utilizing the full 90 min of event-unlabeled data.

exploratory or confirmatory analysis [14,15]. Examples of such semantic events include speaker turn transitions [5,6], mimicry episodes [13], the termination of an interaction [9,10], or high-order social actions [11,12].

One hurdle in such a top-down paradigm is data efficiency. The labeled events often occur infrequently over the interaction, reducing the effective amount of labeled data. This, combined with the fact that collecting behavior data is cost and labor-intensive, precludes the effective application of neural supervised learning techniques that tend to be data demanding. More recently, some approaches have adopted a more bottom-up formulation for dyadic conversations. The task entails predicting event-independent future cues for a single target participant or virtual avatar from the preceding observed cues of both participants [17,18]. Since training sequences are not limited to windows around semantic events, such a formulation is more data-efficient. Figure 1 illustrates the top-down and bottom-up approaches conceptually.

In practice, however, the concrete formulations within the bottom-up paradigm [17,18] suffer from several conceptual problems: (i) predictions are made for a single individual using cues from both individuals as input; since people behave differently, this entails training one forecasting model per person; (ii) even so, predicting a future for one individual at a time is undesirable as these futures are not independent; and (iii) the prediction is only a single future, despite evidence that the future is not deterministic, and the same observed sequence can result in multiple socially-valid continuations [19–21].

To address all these issues, we introduce a self-supervised forecasting task called Social Cue Forecasting: predicting a *distribution* over future multimodal cues *jointly for all group members* from their same preceding multimodal cues. Note that we use *self-supervised* here to simply distinguish from the formulations where the predicted quantity (e.g. event-labels) is of a different representation than the observed input (e.g. cues). Given the cue data, the inputs and outputs of our formulation are both cues, so we *obtain the supervisory signal from the data itself.*

Furthermore, a crucial characteristic of free-standing conversations is that people sustain the interaction by explicitly adapting to one another's behaviors [1]. Moreover, the way a person adapts to their partners is a function of several complex factors surrounding their interpersonal relationships and the social setting [22, Chap. 1]; [1, p. 237]. The social dynamics guiding such behavior are embedded in the constellation of participant cues and are distinct for every unique grouping of individuals. As such, a model should adapt its forecasts to the group under consideration. (Even in the pedestrian setting where coordination is only implicit, Rudenko et al. [23, Sec. 8.4.1] observe that failing to adapt predictions to different individuals is still a limitation). For our methodological contribution, we propose the probabilistic Social Processes models, viewing each conversation group as a meta-learning *task*. This allows for capturing social dynamics unique to each group without learning group-specific models and generalizing to unseen groups at evaluation in a data-efficient manner. We believe that this framing of SCF as a *few-shot* function estimation problem is especially

suitable for conversation forecasting—a limited data regime where good uncertainty estimates are desirable. Concretely, we make the following contributions:

- We introduce and formalize the novel task of Social Cue Forecasting (SCF), addressing the conceptual drawbacks of past formulations.
- For SCF, we propose and evaluate the family of socially aware probabilistic Seq2Seq models we call Social Processes (SP).

2 Related Work

To aid readers from different disciplines situate our work within the broader research landscape, we categorize behavior-forecasting literature by interaction focus [24]. In a focused interaction, such as conversations, participants explicitly coordinate their behaviors to sustain the interaction. In unfocused interactions, coordination is implicit, such as when pedestrians avoid collisions.

Focused Interactions. The predominant interest in conversation forecasting stems from the social sciences, with a focus on identifying patterns that are predictive of upcoming speaking turns [5–8], disengagement from an interaction [9,10], or the splitting or merging of groups [25]. Other works forecast the time-evolving size of a group [26] or semantic social action labels [11,12]. More recently, there has also been a growing interest in the computer vision community for tasks related to inferring low-level cues of participants either from their partners' cues [27] or raw multimodal sensor data [28]. Here there has also been some interest in forecasting nonverbal behavior, mainly for dyadic interactions [17,18,29]. The task involves forecasting the future cues of a target individual from the preceding cues of both participants.

Unfocused Interactions. Early approaches for forecasting pedestrian or vehicle trajectories were heuristic-based, involving hand-crafted energy potentials to describe the influence pedestrians and vehicles have on each other [30–37]. Recent approaches build upon the idea of encoding relative positional information directly into a neural architecture [38–45]. Some works go beyond locations, predicting keypoints in group activities [46,47]. Rudenko et al. [23] provide a survey of approaches within this space.

Non-interaction Settings. Here, the focus has been on forecasting individual poses from images [48] and video [49,50], or synthesizing poses using high-level control parameters [51,52]. The self-supervised aspects of our task formulation are related to visual forecasting, where the goal has been to predict non-semantic low-level pixel features or intermediate representations [34,50,53–57]. Such learned representations have been utilized for other tasks like semi-supervised classification [58], or training agents in immersive environments [59].

For the interested reader, we further discuss practical considerations distinguishing forecasting in conversation and pedestrian settings in Appendix E.

3 Social Cue Forecasting: Task Formalization

While self-supervision has shown promise for learning representations of language and video data, is this bottom-up approach conceptually reasonable for behavior cues? The crucial observation we make is that the semantic meaning transferred in interactions (the so-called *social signal* [60]) is already embedded in the low-level cues [61]. So representations of this high-level semantic meaning that we associate with actions and events (e.g. *group leaving*) can be learned from the low-level dynamics in the cues.

3.1 Formalization and Distinction from Prior Task Formulations

The objective of SCF is to predict future behavioral cues of *all* people involved in a social encounter given an observed sequence of their behavioral features. Formally, let us denote a window of monotonically increasing observed timesteps as $\boldsymbol{t}_{\text{obs}} := [o1, o2, ..., oT]$, and an unobserved future time window as $\boldsymbol{t}_{\text{fut}} := [f1, f2, ..., fT]$, $f1 > oT$. Note that $\boldsymbol{t}_{\text{fut}}$ and $\boldsymbol{t}_{\text{obs}}$ can be of different lengths, and $\boldsymbol{t}_{\text{fut}}$ need not immediately follow $\boldsymbol{t}_{\text{obs}}$. Given n interacting participants, let us denote their social cues over $\boldsymbol{t}_{\text{obs}}$ and $\boldsymbol{t}_{\text{fut}}$ as

$$\boldsymbol{X} \coloneqq [\boldsymbol{b}_t^i; t \in \boldsymbol{t}_{\text{obs}}]_{i=1}^n, \quad \boldsymbol{Y} \coloneqq [\boldsymbol{b}_t^i; t \in \boldsymbol{t}_{\text{fut}}]_{i=1}^n. \tag{1a, b}$$

The vector $\boldsymbol{b}_t^i$ encapsulates the multimodal cues of interest from participant i at time t. These can include head and body pose, speaking status, facial expressions, gestures, verbal content—any information streams that combine to transfer social meaning.

Distribution Over Futures. In its simplest form, given an $\boldsymbol{X}$, the objective of SCF is to learn a single function f such that $\boldsymbol{Y} = f(\boldsymbol{X})$. However, an inherent challenge in forecasting behavior is that an observed sequence of interaction does not have a deterministic future and can result in multiple socially valid ones—a window of overlapping speech between people may and may not result in a change of speaker [19,20], a change in head orientation may continue into a sweeping glance across the room or a darting glance stopping at a recipient of interest [21]. In some cases, certain observed behaviors—intonation and gaze cues [5,62] or synchronization in speaker-listener speech [63] for turn-taking—may make some outcomes more likely than others. Given that there are both supporting and challenging arguments for how these observations influence subsequent behaviors [63, p. 5]; [62, p. 22], it would be beneficial if a data-driven model expresses a measure of uncertainty in its forecasts. We do this by modeling the distribution over possible futures $p(\boldsymbol{Y}|\boldsymbol{X})$, rather than a single future $\boldsymbol{Y}$ for a given $\boldsymbol{X}$, the latter being the case for previous formulations for cues [18,27,46] and actions [11,12].

Joint Modeling of Future Uncertainty. A defining characteristic of focused interactions is that the participants sustain the shared interaction through explicit,

cooperative coordination of behavior [1, p. 220]—the futures of interacting individuals are not independent given an observed window of group behavior. It is therefore essential to capture uncertainty in forecasts at the *global* level—jointly forecasting one future for all participants at a time, rather than at a *local* output level—one future for each individual independent of the remaining participants' futures. In contrast, applying the prior formulations [17,18,27] requires the training of separate models treating each individual as a target (for the same group input) and then forecasting an independent future one at a time. Meanwhile, other prior pose forecasting works [48–52] have been in non-social settings and do not need to model such behavioral interdependence.

Non-contiguous Observed and Future Windows. Domain experts are often interested in settings where $\boldsymbol{t}_{\text{obs}}$ and $\boldsymbol{t}_{\text{fut}}$ are offset by an arbitrary delay, such as forecasting a time lagged synchrony [64] or mimicry [13] episode, or upcoming disengagement [9,10]. We therefore allow for non-contiguous $\boldsymbol{t}_{\text{obs}}$ and $\boldsymbol{t}_{\text{fut}}$. Operationalizing prior formulations that predict one step into the future [11,12,27,46] would entail a sliding window of autoregressive predictions over the offset between $\boldsymbol{t}_{\text{obs}}$ and $\boldsymbol{t}_{\text{fut}}$ (from oT to $f1$), with errors cascading even before decoding is performed over the window of interest $\boldsymbol{t}_{\text{fut}}$.

Our task formalization of SCF can be viewed as a social science-grounded generalization of prior computational formulations, and therefore suitable for a wider range of cross-disciplinary tasks, both computational and analytical.

4 Method Preliminaries

Meta-Learning. A supervised learning algorithm can be viewed as a function mapping a dataset $C \coloneqq (\boldsymbol{X}_C, \boldsymbol{Y}_C) \coloneqq \{(\boldsymbol{x}^i, \boldsymbol{y}^i)\}_{i \in [N_C]}$ to a predictor $f(\boldsymbol{x})$. Here N_C is the number of datapoints in C, and $[N_C] \coloneqq \{1, \ldots, N_C\}$. The key idea of meta-learning is to learn how to learn from a dataset in order to adapt to unseen supervised tasks; hence the name *meta*-learning. This is done by learning a map $C \mapsto f(\cdot, C)$. In meta-learning literature, a *task* refers to each dataset in a collection $\{\mathcal{T}_m\}_{m=1}^{N_{\text{tasks}}}$ of related datasets [65]. Training is episodic, where each task $\mathcal{T}$ is split into subsets (C, D). A meta-learner then fits the subset of target points D given the subset of context observations C. At meta-test time, the resulting predictor $f(\boldsymbol{x}, C)$ is adapted to make predictions for target points on an unseen task by conditioning on a new context set C unseen during meta-training.

Neural Processes (NPs). Sharing the same core motivations, NPs [66] can be viewed as a family of latent variable models that extend the idea of meta-learning to situations where uncertainty in the predictions $f(\boldsymbol{x}, C)$ are desirable. They do this by meta-learning a map from datasets to stochastic processes, estimating a distribution over the predictions $p(\boldsymbol{Y}|\boldsymbol{X}, C)$. To capture this distribution, NPs model the conditional latent distribution $p(\boldsymbol{z}|C)$ from which a task representation $\boldsymbol{z} \in \mathbb{R}^d$ is sampled. This introduces stochasticity, constituting what

is called the model's *latent path*. The context can also be directly incorporated through a *deterministic path*, via a representation $\boldsymbol{r}_C \in \mathbb{R}^d$ aggregated over C. An observation model $p(\boldsymbol{y}^i|\boldsymbol{x}^i, \boldsymbol{r}_C, \boldsymbol{z})$ then fits the target observations in D. The generative process for the NP is written as

$$p(\boldsymbol{Y}|\boldsymbol{X}, C) := \int p(\boldsymbol{Y}|\boldsymbol{X}, C, \boldsymbol{z}) p(\boldsymbol{z}|C) d\boldsymbol{z} = \int p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{r}_C, \boldsymbol{z}) q(\boldsymbol{z}|\boldsymbol{s}_C) d\boldsymbol{z}, \quad (2)$$

where $p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{r}_C, \boldsymbol{z}) := \prod_{i \in [N_D]} p(\boldsymbol{y}^i|\boldsymbol{x}^i, \boldsymbol{r}_C, \boldsymbol{z})$. The latent $\boldsymbol{z}$ is modeled by a factorized Gaussian parameterized by $\boldsymbol{s}_C := f_s(C)$, with f_s being a deterministic function invariant to order permutation over C. When the conditioning on context is removed ($C = \varnothing$), we have $q(\boldsymbol{z}|\boldsymbol{s}_\varnothing) := p(\boldsymbol{z})$, the zero-information prior on $\boldsymbol{z}$. The deterministic path uses a function f_r similar to f_s, so that $\boldsymbol{r}_C := f_r(C)$. In practice this is implemented as $\boldsymbol{r}_C = \sum_{i \in [N_C]} \text{MLP}(\boldsymbol{x}_i, \boldsymbol{y}_i)/N_C$. The observation model is referred to as the *decoder*, and q, f_r, f_s comprise the *encoders*. The parameters of the NP are learned for random subsets C and D for a task by maximizing the evidence lower bound (ELBO)

$$\log p(\boldsymbol{Y}|\boldsymbol{X}, C) \geq \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{s}_D)}[\log p(\boldsymbol{Y}|\boldsymbol{X}, C, \boldsymbol{z})] - \mathbb{KL}(q(\boldsymbol{z}|\boldsymbol{s}_D)||q(\boldsymbol{z}|\boldsymbol{s}_C)). \quad (3)$$

5 Social Processes: Methodology

Our core idea for adapting predictions to a group's unique behavioral dynamics is to condition forecasts on a context set C of the same group's observed-future sequence pairs. By *learning to learn*, i.e., *meta-learn* from a context set, our model can generalize to unseen groups at evaluation by conditioning on an unseen context set of the test group's behavior sequences. In practice, a social robot might, for instance, observe such an evaluation context set before approaching a new group.

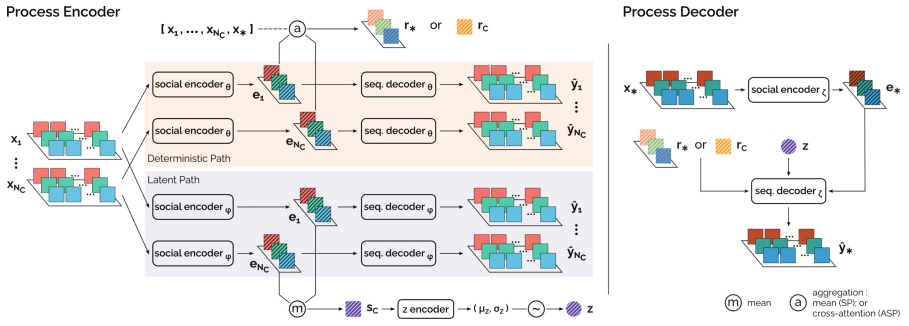

Fig. 2. Architecture of the SP and ASP family.

We set up by splitting the interaction into pairs of observed and future sequences, writing the context as $C := (\boldsymbol{X}_C, \boldsymbol{Y}_C) := (\boldsymbol{X}_j, \boldsymbol{Y}_k)_{(j,k) \in [N_C] \times [N_C]}$,

where every $\boldsymbol{X}_j$ occurs before the corresponding $\boldsymbol{Y}_k$. Since we allow for non-contiguous $\boldsymbol{t}_{\text{obs}}$ and $\boldsymbol{t}_{\text{fut}}$, the jth $\boldsymbol{t}_{\text{obs}}$ can have multiple associated $\boldsymbol{t}_{\text{fut}}$ windows for prediction, up to a maximum offset. Denoting the set of target window pairs as $D \coloneqq (\boldsymbol{X}, \boldsymbol{Y}) \coloneqq (\boldsymbol{X}_j, \boldsymbol{Y}_k)_{(j,k)\in[N_D]\times[N_D]}$, our goal is to model the distribution $p(\boldsymbol{Y}|\boldsymbol{X}, C)$. Note that when conditioning on context is removed ($C = \varnothing$), we simply revert to the non-meta-learning formulation $p(\boldsymbol{Y}|\boldsymbol{X})$.

The generative process for our Social Process (SP) model follows Eq. 2, which we extend to social forecasting in two ways. We embed an observed sequence $\boldsymbol{x}^i$ for participant p_i into a condensed encoding $\boldsymbol{e}^i \in \mathbb{R}^d$ that is then decoded into the future sequence using a Seq2Seq architecture [67,68]. Crucially, the sequence decoder only accesses $\boldsymbol{x}^i$ through $\boldsymbol{e}^i$. So after training, $\boldsymbol{e}^i$ must encode the *temporal* information that $\boldsymbol{x}^i$ contains about the future. Further, social behavior is interdependent. We model $\boldsymbol{e}^i$ as a function of both, p_i's own behavior as well as that of partners $\mathrm{p}_{j,j\neq i}$ from p_i's perspective. This captures the *spatial* influence partners have on the participant over $\boldsymbol{t}_{\text{obs}}$. Using notation we established in Sect. 3, we define the observation model for p_i as

$$p(\boldsymbol{y}^i|\boldsymbol{x}^i, C, \boldsymbol{z}) \coloneqq p(\boldsymbol{b}^i_{f1}, \ldots, \boldsymbol{b}^i_{fT}|\boldsymbol{b}^i_{o1}, \ldots, \boldsymbol{b}^i_{oT}, C, \boldsymbol{z}) = p(\boldsymbol{b}^i_{f1}, \ldots, \boldsymbol{b}^i_{fT}|\boldsymbol{e}^i, \boldsymbol{r}_C, \boldsymbol{z}). \tag{4}$$

If decoding is carried out in an auto-regressive manner, the right hand side of Eq. 4 simplifies to $\prod_{t=f1}^{fT} p(\boldsymbol{b}^i_t|\boldsymbol{b}^i_{t-1}, \ldots, \boldsymbol{b}^i_{f1}, \boldsymbol{e}^i, \boldsymbol{r}_C, \boldsymbol{z})$. Following the standard NP setting, we implement the observation model as a set of Gaussian distributions factorized over time and feature dimensions. We also incorporate the cross-attention mechanism from the Attentive Neural Process (ANP) [69] to define the variant Attentive Social Process (ASP). Following Eq. 4 and the definition of the ANP, the corresponding observation model of the ASP for a single participant is defined as

$$p(\boldsymbol{y}^i|\boldsymbol{x}^i, C, \boldsymbol{z}) = p(\boldsymbol{b}^i_{f1}, \ldots, \boldsymbol{b}^i_{fT}|\boldsymbol{e}^i, r^*(C, \boldsymbol{x}^i), \boldsymbol{z}). \tag{5}$$

Here each target query sequence $\boldsymbol{x}^i_*$ attends to the context sequences $\boldsymbol{X}_C$ to produce a query-specific representation $\boldsymbol{r}_* \coloneqq r^*(C, \boldsymbol{x}^i_*) \in \mathbb{R}^d$.

The model architectures are illustrated in Fig. 2. Note that our modeling assumption is that the underlying stochastic process generating social behaviors

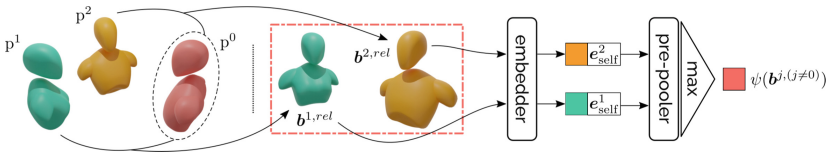

Fig. 3. Encoding partner behavior for participant p^0 for a single timestep. To model the influence partners p^1 and p^2 have on the behavior of p^0, we transform the partner features to capture the interaction from p^0's perspective, and learn a representation of these features invariant to group size and partner-order permutation using the symmetric max function.

does not evolve over time. That is, the individual factors determining how participants coordinate behaviors—age, cultural background, personality variables [22, Chap. 1]; [1, p. 237]—are likely to remain the same over a single interaction. This is in contrast to the line of work that deals with *meta-transfer learning*, where the stochastic process itself changes over time [70–73]; this entails modeling a different $\boldsymbol{z}$ distribution for every timestep.

Encoding Partner Behavior. To encode partners' influence on an individual's future, we use a pair of sequence encoders: one to encode the temporal dynamics of participant p^i's features, $\boldsymbol{e}^i_{\text{self}} = f_{\text{self}}(\boldsymbol{x}^i)$, and another to encode the dynamics of a transformed representation of the features of p^i's partners, $\boldsymbol{e}^i_{\text{partner}} = f_{\text{partner}}(\psi(\boldsymbol{x}^{j,(j\neq i)}))$. Using a separate network to encode partner behavior enables sampling an individual's and partners' features at different sampling rates.

How do we model $\psi(\boldsymbol{x}^{j,(j\neq i)})$? We want the partners' representation to possess two properties: *permutation invariance*—changing the order of the partners should not affect the representation, and *group-size independence*—we want to compactly represent all partners independent of the group size. Intuitively, to model partner influence on p^i, we wish to *capture a view of the partners' behavior as* p^i *perceives it.* Figure 3 illustrates the underlying intuition. We do this by computing pooled embeddings of relative behavioral features, extending Gupta et al. [40]'s approach for pedestrian positions to conversation behavior. Note that our partner-encoding approach is in contrast to that of Tan et al. [28], which is order and group-size dependent, and Yao et al. [46], who do not transform the partner features to an individual's perspective.

Since the most commonly considered cues in literature are pose (orientation and location) and binary speaking status [28, 74, 75], we specify how we transform them. For a single timestep, we denote these cues for p^i as $\boldsymbol{b}^i = [\boldsymbol{q}^i; \boldsymbol{l}^i; s^i]$, and for p^j as $\boldsymbol{b}^j = [\boldsymbol{q}^j; \boldsymbol{l}^j; s^j]$. We compute the relative partner features $\boldsymbol{b}^{j,rel} = [\boldsymbol{q}^{rel}; \boldsymbol{l}^{rel}; s^{rel}]$ by transforming $\boldsymbol{b}^j$ to a frame of reference defined by $\boldsymbol{b}^i$:

$$\boldsymbol{q}^{rel} = \boldsymbol{q}^i * (\boldsymbol{q}^j)^{-1}, \quad \boldsymbol{l}^{rel} = \boldsymbol{l}^j - \boldsymbol{l}^i, \quad s^{rel} = s^j - s^i. \tag{6a-c}$$

Note that we use unit quaternions (denoted $\boldsymbol{q}$) for representing orientation due to their various benefits over other representations of rotation [76, Sec. 3.2]. The operator $*$ denotes the Hamilton product of the quaternions. These transformed features $\boldsymbol{b}^{j,rel}$ for each p^j are then encoded using an *embedder* MLP. The outputs are concatenated with their corresponding $\boldsymbol{e}^j_{\text{self}}$ and processed by a *pre-pooler* MLP. Assuming d_{in} and d_{out} pre-pooler input and output dims and J partners, we stack the J inputs to obtain (J, d_{in}) tensors. The (J, d_{out})-dim output is element-wise max-pooled over the J dim, resulting in the d_{out}-dim vector $\psi(\boldsymbol{b}^{j,(j\neq i)})$ for any value of J, per timestep. We capture the temporal dynamics in this pooled representation over $\boldsymbol{t}_{\text{obs}}$ using f_{partner}. Finally, we combine $\boldsymbol{e}^i_{\text{self}}$ and $\boldsymbol{e}^i_{\text{partner}}$ for p^i through a linear projection (defined by a weight matrix W) to obtain the individual's embedding $\boldsymbol{e}^i_{\text{ind}} = W \cdot [\boldsymbol{e}^i_{\text{self}}; \boldsymbol{e}^i_{\text{partner}}]$. Our intuition is that with information about both p^i themselves, and of p^i's partners from p^i's

point-of-view, $\boldsymbol{e}^i_{\text{ind}}$ now contains the information required to predict p^i's future behavior.

Encoding Future Window Offset. Since we allow for non-contiguous windows, a single $\boldsymbol{t}_{\text{obs}}$ might be associated to multiple $\boldsymbol{t}_{\text{fut}}$ windows at different offsets. Decoding the same $\boldsymbol{e}^i_{\text{ind}}$ into multiple sequences (for different $\boldsymbol{t}_{\text{fut}}$) in the absence of any timing information might cause an averaging effect in either the decoder or the information encoded in $\boldsymbol{e}^i_{\text{ind}}$. One option would be to immediately start decoding after $\boldsymbol{t}_{\text{obs}}$ and discard the predictions in the offset between $\boldsymbol{t}_{\text{obs}}$ and $\boldsymbol{t}_{\text{fut}}$. However, auto-regressive decoding might lead to cascading errors over the offset. Instead, we address this one-to-many issue by injecting the offset information into $\boldsymbol{e}^i_{\text{ind}}$. The decoder then receives a unique encoded representation for every $\boldsymbol{t}_{\text{fut}}$ corresponding to the same $\boldsymbol{t}_{\text{obs}}$. We do this by repurposing the idea of sinusoidal positional encodings [77] to encode window offsets rather than relative token positions in sequences. For a given $\boldsymbol{t}_{\text{obs}}$ and $\boldsymbol{t}_{\text{fut}}$, and d_e-dim $\boldsymbol{e}^i_{\text{ind}}$ we define the offset as $\Delta t = f1 - oT$, and the corresponding offset encoding $OE_{\Delta t}$ as

$$OE_{(\Delta t,2m)} = \sin(\Delta t/10000^{2m/d_e}), OE_{(\Delta t,2m+1)} = \cos(\Delta t/10000^{2m/d_e}). \quad (7\text{a, b})$$

Here m refers to the dimension index in the encoding. We finally compute the representation $\boldsymbol{e}^i$ for Eq. 4 and Eq. 5 as

$$\boldsymbol{e}^i = \boldsymbol{e}^i_{\text{ind}} + OE_{\Delta t}. \quad (8)$$

Auxiliary Loss Functions. We incorporate a geometric loss function for each of our sequence decoders to improve performance in pose regression tasks. For p_i at time t, given the ground truth $\boldsymbol{b}^i_t = [\boldsymbol{q}; \boldsymbol{l}; s]$, and the predicted mean $\hat{\boldsymbol{b}}^i_t = [\hat{\boldsymbol{q}}; \hat{\boldsymbol{l}}; \hat{s}]$, we denote the tuple $(\boldsymbol{b}^i_t, \hat{\boldsymbol{b}}^i_t)$ as B^i_t. We then have the location loss in Euclidean space $\mathcal{L}_{\text{l}}(B^i_t) = ||\boldsymbol{l} - \hat{\boldsymbol{l}}||$, and we can regress the quaternion values using

$$\mathcal{L}_{\text{q}}(B^i_t) = \left\| \boldsymbol{q} - \frac{\hat{\boldsymbol{q}}}{\|\hat{\boldsymbol{q}}\|} \right\|. \quad (9)$$

Kendall and Cipolla [76] show how these losses can be combined using the homoscedastic uncertainties in position and orientation, $\hat{\sigma}^2_{\text{l}}$ and $\hat{\sigma}^2_{\text{q}}$:

$$\mathcal{L}_\sigma(B^i_t) = \mathcal{L}_{\text{l}}(B^i_t) \exp(-\hat{s}_{\text{l}}) + \hat{s}_{\text{l}} + \mathcal{L}_{\text{q}}(B^i_t) \exp(-\hat{s}_{\text{q}}) + \hat{s}_{\text{q}}, \quad (10)$$

where $\hat{s} := \log \hat{\sigma}^2$. Using the binary cross-entropy loss for speaking status $\mathcal{L}_{\text{s}}(B^i_t)$, we have the overall auxiliary loss over $t \in \boldsymbol{t}_{\text{fut}}$:

$$\mathcal{L}_{\text{aux}}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \sum_i \sum_t \mathcal{L}_\sigma(B^i_t) + \mathcal{L}_{\text{s}}(B^i_t). \quad (11)$$

The parameters of the SP and ASP are trained by maximizing the ELBO (Eq. 3) and minimizing this auxiliary loss.

6 Experiments and Results

6.1 Experimental Setup

Evaluation Metrics. Prior forecasting formulations output a single future. However, since the future is not deterministic, we predict a future distribution. Consequently, needing a metric that accounts for probabilistic predictions, we report the log-likelihood (LL) $\log p(\boldsymbol{Y}|\boldsymbol{X}, C)$, commonly used by all variants within the NP family [66,69,70]. The metric is equal to the log of the predicted density evaluated at the ground-truth value. (Note: the fact that the vast majority of forecasting works even in pedestrian settings omit a probabilistic metric, using only geometric metrics, is a limitation also observed by Rudenko et al. [23, Sec. 8.3].) Nevertheless, for additional insight beyond the LL, we also report the errors in the predicted means—geometric errors for pose and accuracy for speaking status—and provide qualitative visualizations of forecasts.

Models and Baselines. In keeping with the task requirements and for fair evaluation, we require that all models we compare against forecast a distribution over future cues.

- To evaluate our core idea of viewing conversing groups as meta-learning tasks, we compare against non-meta-learning methods: we adapt variational encoder-decoder (VED) architectures [78,79] to output a distribution.
- To evaluate our specific modeling choices within the meta-learning family, we compare against the NP and ANP models (see Sect. 5). The original methods were not proposed for sequences, so we adapt them by collapsing the timestep and feature dimensions in the data.

Note that in contrast to the SP models, these baselines have direct access to the future sequences in the context, and therefore constitute a strong baseline. We consider two variants for both NP and SP models: *-latent* denoting only the stochastic path; and *-uniform* containing both the deterministic and stochastic paths with uniform attention over context sequences. We further consider two attention mechanisms for the cross-attention module: *-dot* with dot attention, and *-mh* with wide multi-head attention [69]. Finally, we experiment with two choices of backbone architectures: multi-layer perceptrons (MLP), and Gated Recurrent Units (GRU). Implementation and training details can be found in Appendix D. Code, processed data, trained models, and test batches for reproduction are available at https://github.com/chiragraman/social-processes.

6.2 Evaluation on Synthesized Behavior Data

To first validate our method on a toy task, we synthesize a dataset simulating two glancing behaviors in social settings [21], approximated by horizontal head rotation. The sweeping *Type I* glance is represented by a 1D sinusoid over 20 timesteps. The gaze-fixating *Type III* glance is denoted by clipping the amplitude for the last six timesteps. The task is to forecast the signal over the last 10

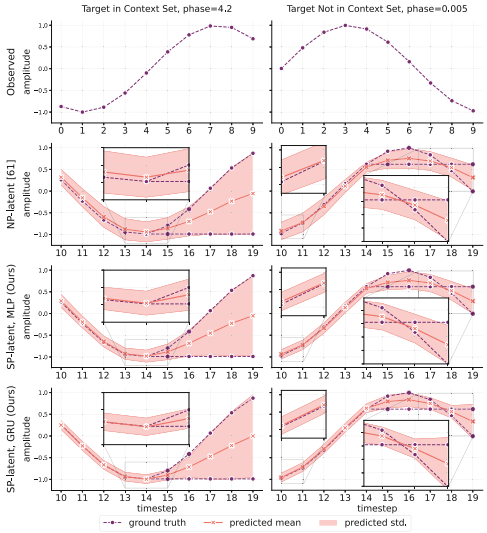

Fig. 4. Ground truths and predictions for the toy task of forecasting simulated glancing behavior. Our SP models learn a better fit than the NP model, SP-GRU being the best (see zoomed insets).

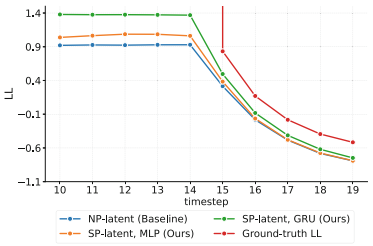

Fig. 5. Mean per timestep LL over the sequences in the synthetic glancing dataset. Higher is better.

Table 1. Mean (Std.) Metrics on the Synthetic Glancing Behavior Dataset. The metrics are averaged over timesteps; mean and std. are then computed over sequences. Higher is better for LL, lower for MAE.

	LL	Head Ori. MAE (°)
NP-latent	0.28 (0.24)	19.63 (7.26)
SP-latent (MLP)	0.36 (0.20)	19.46 (7.05)
SP-latent (GRU)	**0.55 (0.23)**	**18.55 (7.11)**

timesteps ($\boldsymbol{t}_{\text{fut}}$) by observing the first 10 ($\boldsymbol{t}_{\text{obs}}$). Consequently, the first half of $\boldsymbol{t}_{\text{fut}}$ is certain, while the last half is uncertain: every observed sinusoid has two ground truth futures in the data (clipped and unclipped). It is impossible to infer from an observed sequence alone if the head rotation will stop partway through the future. Figure 4 illustrates the predictions for two sample sequences. Table 1 provides quantitative metrics and Fig. 5 plots the LL per timestep. The LL is expected to decrease over timesteps where ground-truth futures diverge, being ∞ when the future is certain. We observe that all models estimate the mean reasonably well, although our proposed SP models perform best. More crucially, the SP models, especially the SP-GRU, learn much better uncertainty estimates compared to the NP baseline (see zoomed regions in Fig. 4). We provide additional analysis, alternative qualitative visualizations, and data synthesis details in Appendices A to C respectively.

6.3 Evaluation on Real-World Behavior Data

Datasets and Preprocessing. With limited behavioral data availability, a common practice in the domain is to solely train and evaluate methods on synthesized behavior dynamics [12,80]. In contrast, we also evaluate on two real-world behavior datasets: the MatchNMingle (MnM) dataset of in-the-wild mingling behavior [16], and the Haggling dataset of a triadic game where two sellers compete to sell a fictional product to a buyer [27]. For MnM, we treat the 42 groups from Day 1

Table 2. Mean (Std.) Log-Likelihood (LL) on the MatchNMingle and Haggling Test Sets. For a single sequence, we sum over the feature and participant dimensions, and average over timesteps. The reported mean and std. are over individual sequences in the test sets. Higher is better. Underline indicates best LL within family.

	MatchNMingle		Haggling	
	Random	Fixed-initial	Random	Fixed-initial
VED family [78,79]				
VED-MLP	8.1 (7.2)	7.9 (7.0)	4.0 (8.3)	4.1 (8.2)
VED-GRU	25.4 (18.0)	25.1 (19.1)	60.3 (2.2)	60.3 (2.1)
NP Family [66,69]				
NP-latent	22.1 (17.8)	21.6 (18.5)	27.2 (17.3)	27.9 (16.3)
NP-uniform	21.4 (18.8)	20.5 (17.8)	24.8 (22.9)	25.0 (22.2)
ANP-dot	22.8 (18.6)	21.0 (18.3)	26.7 (21.4)	24.7 (20.8)
ANP-mh	23.6 (15.6)	20.0 (23.9)	25.1 (23.1)	24.8 (22.4)
Ours (SP-MLP)				
SP-latent	102.1 (29.9)	101.5 (29.2)	136.6 (7.0)	136.7 (7.0)
SP-uniform	112.8 (34.1)	111.4 (33.8)	138.3 (8.0)	137.6 (8.4)
ASP-dot	109.9 (32.9)	107.6 (32.1)	137.8 (7.5)	136.4 (7.6)
ASP-mh	112.9 (34.7)	111.3 (33.6)	146.0 (10.9)	145.7 (10.2)
Ours (SP-GRU)				
SP-latent	86.4 (37.2)	85.4 (37.2)	66.7 (27.4)	66.2 (30.7)
SP-uniform	87.0 (38.4)	85.5 (38.3)	79.9 (50.5)	78.6 (52.2)
ASP-dot	87.6 (39.1)	83.9 (38.1)	38.4 (60.4)	27.2 (93.4)
ASP-mh	85.8 (37.1)	82.3 (36.0)	66.3 (30.3)	59.3 (32.4)

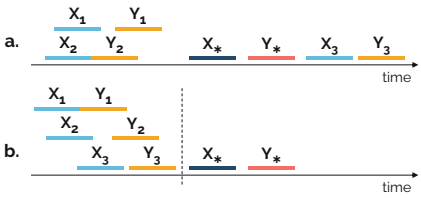

Fig. 6. Context Regimes. For a target sequence pair ($\boldsymbol{X}_*$, $\boldsymbol{Y}_*$), context pairs (here 3) are sampled either **a.** randomly across the lifetime of the group interaction (*random*), or **b.** from a fixed initial duration (*fixed-initial*).

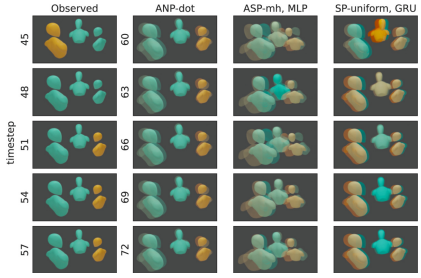

Fig. 7. Forecasts over selected timesteps from the Haggling group *170224-a1-group1*. Speaking status is interpolated between orange (speaking) and blue (listening). Translucent models denote the predicted mean ± std. (Color figure online)

as test sets and a total of 101 groups from the other two days as train sets. For Haggling, we use the same split of 79 training and 28 test groups used by Joo et al. [27]. We consider the following cues: *head pose* and *body pose*, described by the location of a keypoint and an orientation quaternion; and binary *speaking status*. These are the most commonly considered cues in computational analyses of conversations [28,74,75] given how crucial they are in sustaining interactions [1,20,61]. For orientation, we first convert the normal vectors (provided in the horizontal direction in both datasets) into unit quaternions. Since the quaternions $\mathbf{q}$ and $-\mathbf{q}$ denote an identical rotation, we constrain the first quaternion in every sequence to the same hemisphere and interpolate subsequent quaternions to have the shortest distance along the unit hypersphere. We then split the interaction data into pairs of $\boldsymbol{t}_{\text{obs}}$ and $\boldsymbol{t}_{\text{fut}}$ windows to construct the samples for forecasting. We specify dataset-specific preprocessing details in Appendix C.

Context Regimes. We evaluate on two context regimes: *random*, and *fixed-initial* (see Fig. 6). In the *random* regime, context samples (observed-future pairs) are

selected as a random subset of target samples, so the model is exposed to behaviors from any phase of the interaction lifecycle. Here we ensure that batches contain unique $\boldsymbol{t}_{\text{obs}}$ to prevent any single observed sequence from dominating the aggregation of representations over the context split. At evaluation, we take 50% of the batch as context. The *fixed-initial* regime investigates how models can learn from observing the initial dynamics of an interaction where certain gestures and patterns are more distinctive [1, Chap. 6]. Here we treat the first 20% of the entire interaction as context, treating the rest as target.

Conversation Groups as Meta-learning Tasks? While our core idea of viewing groups as meta-learning tasks is grounded in social science literature (see Sect. 5), does it help to improve empirical performance? Comparing the LL of non-meta-learning and meta-learning models in Table 2 by architecture—VED-MLP against NP and SP-MLP, and VED-GRU against SP-GRU—we find that accounting for group-specific dynamics through meta-learning yields improved performance. All best-in-family pairwise model differences are statistically significant (Wilcoxon signed rank test, $p < 10^{-4}$).

Comparing Within Meta-learning Methods. While our SP-MLP models perform the best on LL in Table 2 (pairwise differences are significant), they fare the worst at estimating the mean (Appendix A.2). On the other hand, the SP-GRU models estimate a better LL than the NP models with comparable errors in the mean forecast. The NP models attain the lowest errors in predicted means, but also achieve the worst LL. Why do the models achieving better LL also tend to predict worse means? Upon inspecting the metrics for individual features, we found that the models, especially the MLP variants, tend to improve LL by making the variance over constant features exceedingly small, often at the cost of errors in the means. Note that since the rotation in the data is in the horizontal plane, the qx and qy quaternion dimensions are zero throughout. We do not observe such model behavior in the synthetic data experiments, which do not involve constant features. Figure 7 visualizes forecasts for an example sequence from the Haggling dataset where a turn change has occurred just at the end of the observed window. Here, the SP-GRU model forecasts an interesting continuation to the turn. It anticipates that the buyer (middle) will interrupt the last observed speaker (right seller), before falling silent and looking from one seller to another, both of whom the model expects to then speak simultaneously (see Appendix B for the full sequence). We believe that the forecast indicates that the model is capable of learning believable haggling turn dynamics from different turn continuations in the data. From the visualizations also we observe that the models seem to maximize LL at the cost of orientation errors; in the case of SP-MLP seemingly by predicting the majority orientation in the triadic setting. Also, the NP models forecast largely static futures. In contrast, while being more dynamic, the SP-GRU forecasts contain some smoothing. Overall, the SP-GRU models achieve the best trade-off between maximizing LL and forecasting plausible human behavior.

6.4 Ablations

Encoding Partner Behavior. Modeling the interaction from the perspective of each individual is a central idea in our approach. We investigate the influence of encoding partner behavior into individual representations $\boldsymbol{e}^i_{\text{ind}}$. We train the SP-uniform GRU variant in two configurations: *no-pool*, where we do not encode any partner behavior; and *pool-oT* where we pool over partner representations only at the last timestep (similar to [40]). Both configurations lead to worse LL and location errors (Table 3 and Appendix A).

Table 3. Mean (Std.) LL for the Ablation Experiments with the SP-uniform GRU Model. The reported mean and std. are over individual sequences in the test sets. Higher is better.

		MatchNMingle		Haggling	
		Random	Fixed-initial	Random	Fixed-initial
Full model		87.0 (38.4)	85.5 (38.3)	79.9 (50.5)	78.6 (52.2)
Encoding partner behavior	no-pool	77.8 (31.2)	76.9 (31.0)	54.5 (75.5)	50.1 (97.5)
	pool-oT	82.3 (33.3)	81.0 (33.6)	66.9 (26.0)	66.8 (25.7)
No deterministic decoding	Shared social encoders	88.5 (40.7)	87.6 (39.6)	93.1 (39.3)	91.9 (40.4)
	Unshared social encoders	81.4 (38.1)	80.2 (37.8)	66.6 (24.0)	64.8 (23.4)

Deterministic Decoding and Social Encoder Sharing. We investigate the effect of the deterministic decoders by training the SP-uniform GRU model without them. We also investigate sharing a single social encoder between the Process Encoder and Process Decoder in Fig. 2. Removing the decoders only improves log-likelihood if the encoders are shared, and at the cost of head orientation errors (Table 3 and Appendix A).

7 Discussion

The setting of social conversations remains a uniquely challenging frontier for state-of-the-art low-level behavior forecasting. In the recent forecasting challenge involving dyadic interactions, none of the submitted methods could outperform the naive *zero-velocity* baseline [17, Sec. 5.5]. (The baseline propagates the last observed features into the future as if the person remained static.) Why is this? The predominant focus of researchers working on social human-motion prediction has been pedestrian trajectories [23] or actions such as *punching, kicking, gathering, chasing, etc.* [46,47]. In contrast to such activities which involve pronounced movements, the postural adaptation for regulating conversations is far more subtle (also see the discussion in Appendix E). At the same time, the social intelligence required to understand the underlying dynamics that drive a conversation is comparatively more sophisticated than for an action such as a kick. We hope that the social-science considerations informing the design of SCF (joint

probabilistic forecasting for all members) and the SP models (groups as meta-learning tasks) constitute a meaningful foundation for future research in this space to build upon. Note that for our task formulation, even the performance of our baseline models constitutes new results.

Cross-Discipline Impact and Ethical Considerations. While our work here is an *upstream* methodological contribution, the focus on human behavior entails ethical considerations for downstream applications. One such application involves assisting social scientists in developing predictive hypotheses for specific behaviors by examining model predictions. In these cases, such hypotheses must be verified in subsequent controlled experiments. With the continued targeted development of techniques for recording social behavior in the wild [81], evaluating forecasting models in varied interaction settings would also provide further insight. Another application involves helping conversational agents achieve smoother interactions. Here researchers should be careful that the ability to forecast does not result in nefarious manipulation of user behavior.

Acknowledgements. This research was partially funded by the Netherlands Organization for Scientific Research (NWO) under the MINGLE project number 639.022.606. Chirag would like to thank Amelia Villegas-Morcillo for her input and the innumerable discussions, and Tiffany Matej Hrkalovic for feedback on parts of the manuscript.

References

1. Kendon, A.: Conducting Interaction: Patterns of Behavior in Focused Encounters. Number 7 in Studies in Interactional Sociolinguistics. Cambridge University Press, Cambridge (1990). ISBN 978-0-521-38036-2, 978-0-521-38938-9
2. Vinciarelli, A., Pantic, M., Bourlard, H.: Social signal processing: survey of an emerging domain. Image Vis. Comput. **27**(12), 1743–1759 (2009)
3. Bohus, D., Horvitz, E.: Models for multiparty engagement in open-world dialog. In: Proceedings of the SIGDIAL 2009 Conference on The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue - SIGDIAL 2009, pp. 225–234. Association for Computational Linguistics, London (2009). ISBN 978-1-932432-64-0. https://doi.org/10.3115/1708376.1708409
4. Ishii, R., Kumano, S., Otsuka, K.: Prediction of next-utterance timing using head movement in multi-party meetings. In: Proceedings of the 5th International Conference on Human Agent Interaction, HAI 2017, pp. 181–187. Association for Computing Machinery, New York, October 2017. ISBN 978-1-4503-5113-3, https://doi.org/10.1145/3125739.3125765
5. Keitel, A., Daum, M.M.: The use of intonation for turn anticipation in observed conversations without visual signals as source of information. Front. Psychol. **6**, 108 (2015)
6. Garrod, S., Pickering, M.J.: The use of content and timing to predict turn transitions. Front. Psychol. **6**, 751 (2015)
7. Rochet-Capellan, A., Fuchs, S.: Take a breath and take the turn: how breathing meets turns in spontaneous dialogue. Philos. Trans. Roy. Soc. B Biol. Sci. **369**(1658), 20130399 (2014)

8. Wlodarczak, M., Heldner, M.: Respiratory turn-taking cues. In: INTERSPEECH (2016)
9. Bohus, D., Horvitz, E.: Managing human-robot engagement with forecasts and... um... hesitations. In: Proceedings of the 16th International Conference on Multimodal Interaction, p. 8 (2014)
10. van Doorn, F.: Rituals of leaving: predictive modelling of leaving behaviour in conversation. Master of Science thesis, Delft University of Technology (2018)
11. Airale, L., Vaufreydaz, D., Alameda-Pineda, X.: SocialInteractionGAN: multi-person interaction sequence generation. arXiv:2103.05916 [cs, stat], March 2021
12. Sanghvi, N., Yonetani, R., Kitani, K.: MGPI: a computational model of multiagent group perception and interaction. arXiv preprint arXiv:1903.01537 (2019)
13. Bilakhia, S., Petridis, S., Pantic, M.: Audiovisual detection of behavioural mimicry. In: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, pp. 123–128. IEEE, Geneva, September 2013. ISBN 978-0-7695-5048-0. https://doi.org/10.1109/ACII.2013.27
14. Liem, C.C.S., et al.: Psychology meets machine learning: interdisciplinary perspectives on algorithmic job candidate screening. In: Escalante, H.J., et al. (eds.) Explainable and Interpretable Models in Computer Vision and Machine Learning. TSSCML, pp. 197–253. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98131-4_9
15. Nilsen, E., Bowler, D., Linnell, J.: Exploratory and confirmatory research in the open science era. J. Appl. Ecol. **57** (2020). https://doi.org/10.1111/1365-2664.13571
16. Cabrera-Quiros, L., Demetriou, A., Gedik, E., van der Meij, L., Hung, H.: The matchnmingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates. IEEE Trans. Affect. Comput. (2018)
17. Palmero, C., et al.: Chalearn lap challenges on self-reported personality recognition and non-verbal behavior forecasting during social dyadic interactions: dataset, design, and results. In: Understanding Social Behavior in Dyadic and Small Group Interactions, pp. 4–52. PMLR (2022)
18. Ahuja, C., Ma, S., Morency, L.-P., Sheikh, Y.: To react or not to react: end-to-end visual pose forecasting for personalized avatar during dyadic conversations. arXiv:1910.02181 [cs], October 2019
19. Heldner, M., Edlund, J.: Pauses, gaps and overlaps in conversations. J. Phonet. **38**(4), 555–568 (2010). ISSN 0095-4470. https://doi.org/10.1016/j.wocn.2010.08.002
20. Duncan, S.: Some signals and rules for taking speaking turns in conversations. J. Person. Soc. Psychol. **23**(2), 283–292(1972). ISSN 1939-1315 (Electronic), 0022-3514 (Print). https://doi.org/10.1037/h0033031
21. Moore, M.M.: Nonverbal courtship patterns in women: context and consequences. Ethol. Sociobiol. **6**(4), 237–247 (1985). ISSN 0162-3095. https://doi.org/10.1016/0162-3095(85)90016-0
22. Moore, N.-J., Mark III, H., Don, W.: Stacks. Nonverbal Commun. Stud. Appl. (2013)
23. Rudenko, Palmieri, L., Herman, M., Kitani, K.M., Gavrila, D.M., Arras, K.O.: Human motion trajectory prediction: a survey. Int. J. Robot. Res. **39**(8), 895–935 (2020)
24. Goffman, E.: Behavior in Public Places: Notes on the Social Organization of Gatherings. The Free Press, 1. paperback edn, 24. printing edition, 1966. ISBN 978-0-02-911940-2

25. Wang, A., Steinfeld, A.: Group split and merge prediction with 3D convolutional networks. IEEE Robot. Autom. Lett. **5**(2), 1923–1930, April 2020. ISSN 2377-3766. https://doi.org/10.1109/LRA.2020.2969947
26. Mastrangeli, M., Schmidt, M., Lacasa, L.: The roundtable: an abstract model of conversation dynamics. arXiv:1010.2943 [physics], October 2010
27. Joo, H., Simon, T., Cikara, M., Sheikh, Y.: Towards social artificial intelligence: nonverbal social signal prediction in a triadic interaction. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10865–10875. IEEE, Long Beach, June 2019. ISBN 978-1-72813-293-8. https://doi.org/10.1109/CVPR.2019.01113
28. Tan, S., Tax, D.M.J., Hung, H.: Multimodal joint head orientation estimation in interacting groups via proxemics and interaction dynamics. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 5, no. 1, pp. 1–22, March 2021. ISSN 2474-9567. https://doi.org/10.1145/3448122
29. Tuyen, N.T.V., Celiktutan, O.: Context-aware human behaviour forecasting in dyadic interactions. In: Understanding Social Behavior in Dyadic and Small Group Interactions, pp. 88–106. PMLR (2022)
30. Helbing, D., Molnar, P.: Social force model for pedestrian dynamics. Phys. Rev. E, **51**(5), 4282–4286 (1995). ISSN 1063-651X, 1095-3787. https://doi.org/10.1103/PhysRevE.51.4282
31. Jarosław Wąs, Bartłomiej Gudowski, and Paweł J. Matuszyk. Social Distances Model of Pedestrian Dynamics. In Cellular Automata, volume 4173, pages 492–501. Springer, Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-40929-8 978-3-540-40932-8. https://doi.org/10.1007/11861201_57
32. Antonini, G., Bierlaire, M., Weber, M.: Discrete choice models for pedestrian walking behavior. Transport. Res. Part B Methodol. **40**, 667–687 (2006). https://doi.org/10.1016/j.trb.2005.09.006
33. Treuille, A., Cooper, S., Popović, Z.: Continuum crowds. ACM Trans. Graph./SIGGRAPH 2006 **25**(3), 1160–1168 (2006)
34. Robicquet, A., Sadeghian, A., Alahi, A., Savarese, S.: Learning social etiquette: human trajectory understanding in crowded scenes. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 549–565. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_33
35. Wang, J.M., Fleet, D.J., Hertzmann, A.: Gaussian process dynamical models for human motion. IEEE Trans. Pattern Anal. Mach. Intell. **30**(2), 283–298, February 2008. ISSN 1939-3539. https://doi.org/10.1109/TPAMI.2007.1167
36. Tay, C., Laugier, C.: Modelling smooth paths using gaussian processes. In: Proceedings of the International Conference on Field and Service Robotics (2007)
37. Patterson, A., Lakshmanan, A., Hovakimyan, N.: Intent-aware probabilistic trajectory estimation for collision prediction with uncertainty quantification. arXiv:1904.02765 [cs, math], April 2019
38. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social LSTM: human trajectory prediction in crowded spaces. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 961–971. IEEE, Las Vegas, June 2016. ISBN 978-1-4673-8851-1. https://doi.org/10.1109/CVPR.2016.110
39. Zhang, P., Ouyang, W., Zhang, P., Xue, J., Zheng, N.: SR-LSTM: state refinement for LSTM towards pedestrian trajectory prediction. arXiv:1903.02793 [cs], March 2019

40. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social GAN: socially acceptable trajectories with generative adversarial networks. arXiv:1803.10892 [cs], March 2018
41. Hasan, I., et al.: Forecasting people trajectories and head poses by jointly reasoning on tracklets and vislets. arXiv:1901.02000 [cs], January 2019
42. Huang, Y., Bi, H., Li, Z., Mao, T., Wang, Z.: STGAT: modeling spatial-temporal interactions for human trajectory prediction. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6271–6280. IEEE, Seoul, October 2019. ISBN 978-1-72814-803-8. https://doi.org/10.1109/ICCV.2019.00637
43. Mohamed, A., Qian, K., Elhoseiny, M., Claudel, C.: Social-STGCNN: a social spatio-temporal graph convolutional neural network for human trajectory prediction. arXiv:2002.11927 [cs], February 2020
44. Zhao, H., et al.: TNT: Target-driveN trajectory prediction. arXiv:2008.08294 [cs], August 2020
45. Gilles, T., Sabatini, S., Tsishkou, D., Stanciulescu, B., Moutarde, F.: THOMAS: trajectory heatmap output with learned multi-agent sampling. arXiv:2110.06607 [cs], January 2022
46. Yao, T., Wang, M., Ni, B., Wei, H., Yang, X.: Multiple granularity group interaction prediction. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2246–2254. IEEE, Salt Lake City, June 2018. ISBN 978-1-5386-6420-9. https://doi.org/10.1109/CVPR.2018.00239
47. Vida Adeli, Ehsan Adeli, Ian Reid, Juan Carlos Niebles, and Hamid Rezatofighi. Socially and contextually aware human motion and pose forecasting. IEEE Robotics and Automation Letters, 5 (4): 6033–6040, 2020
48. Chao, Y.-W., Yang, J., Price, B., Cohen, S., Deng, J.: Forecasting human dynamics from static images. arXiv:1704.03432 [cs], April 2017
49. Fragkiadaki, K., Levine, S., Felsen, P., Malik, J.: Recurrent network models for human dynamics. arXiv:1508.00271 [cs], September 2015
50. Walker, J., Marino, K., Gupta, A., Hebert, M.: The pose knows: video forecasting by generating pose futures. arXiv:1705.00053 [cs], April 2017
51. Habibie, I., Holden, D., Schwarz, J., Yearsley, J., Komura, T.: A recurrent variational autoencoder for human motion synthesis. In Procedings of the British Machine Vision Conference 2017, p. 119. British Machine Vision Association, London (2017). ISBN 978-1-901725-60-5. https://doi.org/10.5244/C.31.119
52. Pavllo, D., Grangier, D., Auli, M.: QuaterNet: a quaternion-based recurrent model for human motion. arXiv:1805.06485 [cs], July 2018
53. Ranzato, M.A., Szlam, A., Bruna, J., Mathieu, M., Collobert, R., Chopra, S.: Video (language) modeling: a baseline for generative models of natural videos. arXiv:1412.6604 [cs], December 2014
54. Walker, J., Gupta, A., Hebert, M.: Dense optical flow prediction from a static image. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 2443–2451. IEEE, Santiago, December 2015. ISBN 978-1-4673-8391-2. https://doi.org/10.1109/ICCV.2015.281
55. Dosovitskiy, A., et al.: FlowNet: learning optical flow with convolutional networks. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 2758–2766. IEEE, Santiago, December 2015. ISBN 978-1-4673-8391-2. https://doi.org/10.1109/ICCV.2015.316
56. Walker, J., Gupta, A., Hebert, M.: Patch to the future: unsupervised visual prediction. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3302–3309. IEEE, Columbus, June 2014. ISBN 978-1-4799-5118-5. https://doi.org/10.1109/CVPR.2014.416

57. Vondrick, C., Pirsiavash, H., Torralba, A.: Anticipating visual representations from unlabeled video. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 98–106. IEEE, Las Vegas, June 2016. ISBN 978-1-4673-8851-1. https://doi.org/10.1109/CVPR.2016.18
58. Srivastava, N., Mansimov, E., Salakhutdinov, R.: Unsupervised learning of video representations using LSTMs. arXiv:1502.04681 [cs], February 2015
59. Dosovitskiy, A., Koltun, V.: Learning to act by predicting the future. arXiv:1611.01779 [cs], November 2016
60. Ambady, N., Bernieri, F.J., Richeson, J.A.: Toward a histology of social behavior: judgmental accuracy from thin slices of the behavioral stream. In: Advances in Experimental Social Psychology, vol. 32, pp. 201–271. Elsevier, Amsterdam (2000)
61. Vinciarelli, A., Salamin, H., Pantic, M.: Social signal processing: understanding social interactions through nonverbal behavior analysis (PDF). In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, June 2009. https://doi.org/10.1109/CVPRW.2009.5204290
62. Kalma, A.: Gazing in triads: a powerful signal in floor apportionment. Br. J. Soc. Psychol. **31**(1), 21–39 (1992)
63. Levinson, S.C., Torreira, F.: Timing in turn-taking and its implications for processing models of language. Front. Psychol. **6** (2015). ISSN 1664–1078. https://doi.org/10.3389/fpsyg.2015.00731
64. Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., Cohen, D.: Interpersonal synchrony: a survey of evaluation methods across disciplines. IEEE Trans. Affect. Comput. **3**(3), 349–365 (2012). ISSN 1949–3045. https://doi.org/10.1109/T-AFFC.2012.12
65. Hospedales, T., Antoniou, A., Micaelli, P., Storkey, A.: Meta-learning in neural networks: a survey. arXiv:2004.05439 [cs, stat], November 2020
66. Garnelo, M., et al.: Neural processes. arXiv:1807.01622 [cs, stat] (2018)
67. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, vol. 27, pp. 3104–3112. Curran Associates Inc. (2014)
68. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv:1406.1078 [cs, stat], September 2014
69. Kim, H., et al.: Attentive neural processes. arXiv:1901.05761 [cs, stat], July 2019
70. Singh, G., Yoon, J., Son, Y., Ahn, S.: Sequential neural processes. In: Advances in Neural Information Processing Systems, vol. 32 (2019). https://arxiv.org/abs/1906.10264
71. Yoon, J., Singh, G., Ahn, S.: Robustifying sequential neural processes. In: International Conference on Machine Learning, pp. 10861–10870. PMLR, November 2020
72. Willi, T., Schmidhuber, J.M., Osendorfer, C.: Recurrent neural processes. arXiv:1906.05915 [cs, stat], November 2019
73. Kumar, S.: Spatiotemporal modeling using recurrent neural processes. Master of Science thesis, Carnegie Mellon University, p. 43 (2019)
74. Alameda-Pineda, X., Yan, Y., Ricci, E., Lanz, O., Sebe, N.: Analyzing free-standing conversational groups: a multimodal approach. In: Proceedings of the 23rd ACM International Conference on Multimedia, pp. 5–14. ACM Press (2015). ISBN 978-1-4503-3459-4. https://doi.org/10.1145/2733373.2806238
75. Zhang, L., Hung, H.: On social involvement in mingling scenarios: detecting associates of F-formations in still images. IEEE Trans. Affect. Comput. (2018)

76. Kendall, A., Cipolla, R.: Geometric loss functions for camera pose regression with deep learning. arXiv:1704.00390 [cs], May 2017
77. Vaswani, A., et al.: Attention is all you need. arXiv:1706.03762 [cs], June 2017
78. Ha, D., Eck, D.: A neural representation of sketch drawings. arXiv:1704.03477 [cs, stat], May 2017
79. Bowman, S.R., Vilnis, L., Vinyals, O., Dai, A.M., Jozefowicz, R., Bengio, S.: Generating sentences from a continuous space. arXiv:1511.06349 [cs], May 2016
80. Vazquez, M., Steinfeld, A., Hudson, S.E.: Maintaining awareness of the focus of attention of a conversation: a robot-centric reinforcement learning approach. In: 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 36–43. IEEE, New York, August 2016. ISBN 978-1-5090-3929-6. https://doi.org/10.1109/ROMAN.2016.7745088
81. Raman, C., Tan, S., Hung, H.: A modular approach for synchronized wireless multimodal multisensor data acquisition in highly dynamic social settings. arXiv preprint arXiv:2008.03715 (2020)
82. Raman, C., Hung, H.: Towards automatic estimation of conversation floors within f-formations. In: 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), pp. 175–181. IEEE (2019)
83. Le, T.A., Kim, H., Garnelo, M.: Empirical evaluation of neural process objectives. In: NeurIPS workshop on Bayesian Deep Learning, . 71 (2018)
84. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv:1412.6980 [cs], January 2017
85. Paszke, A., et al.: Pytorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, vol. 32, pp. 8024–8035. Curran Associates Inc. (2019)
86. Falcon, W.A., et al.: Pytorch lightning. GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning, 3, 2019
87. Rienks, R., Poppe, R., Poel, M.: Speaker prediction based on head orientations. In: Proceedings of the Fourteenth Annual Machine Learning Conference of Belgium and the Netherlands (Benelearn 2005), pp. 73–79 (2005)
88. Farenzena, M., et al.: Social interactions by visual focus of attention in a three-dimensional environment. Expert Syst. **30**(2), 115–127 (2013). ISSN 02664720. https://doi.org/10.1111/j.1468-0394.2012.00622.x
89. Ba, S.O., Odobez, J.-M.: Recognizing visual focus of attention from head pose in natural meetings. IEEE Trans. Syst. Man Cybern. Part B (Cybern.) **39**(1), 16–33, February 2009. ISSN 1083–4419. https://doi.org/10.1109/TSMCB.2008.927274