Behavior forecasting for social interactions: a multimodal skeleton-based approach.

German Barquero

Abstract

Many works focus on predicting the motion or trajectory of individuals engaged in a particular action, which intends to reduce the inherent stochasticity of the future. We open a new horizon by aiming at forecasting human behavior in dyadic interactions. In such scenarios, the ability to anticipate human behavior implies an implicit knowledge of the underlying mechanisms of communication involving cognitive, affective, and behavioral perspectives. This knowledge is key for many applications in robotics, medicine and psychology. In this work, we introduce an extended version of the UDIVA dataset which contains automatically extracted face, body and hands landmark annotations for 145 dyadic sessions among 134 participants. We use it to deeply analyze the current limitations of interaction forecasting, most of them derived from the multimodal nature of the future and the huge dimensionality attached to human behavior. In parallel, we propose a multimodal recurrent model based on the popular seq2seq model, which serves as a baseline for future research on this topic. Finally, we present an ablation study to discuss the effects of leveraging multimodal data such as audio and participants metadata.

Index Terms

Human Behavior, Social Interaction, Dyadic Interaction, Behavior Forecasting, Pose Forecasting, Multimodal Network

I. Introduction

TAke a moment to think about a recent conversation you were engaged in. How often did you nod while endorsing the other interlocutor's speech? Or scratch your chin while thinking deeply about an answer to a difficult question? Both your compliance with someone's speech and a demanding question triggered specific behaviors. Now, imagine yourself involved in a conversation where your partner is totally frozen. Weird, right? These conversational behaviors are not only expected but also needed in order to enhance human-to-human communication.

Therefore, understanding the dynamics of individuals' behavior is important for many applications related to human communication. In a unidirectional speech, we all agree that the way someone behaves evolves over time but is exclusively driven by intrapersonal characteristics (e.g., personality, mood, speech) and the scenario context (e.g., thesis defense, informal speech). The body language of an extrovert person talking about themselves will differ from that of a shy person. Similarly, someone's behavior will vary from a job interview to a friendly conversation. In dyadic interactions, mirroring the behavior of a person also needs to contemplate interpersonal cues [1]. These include not only the interlocutor's expression, gestures or body language, but also intrinsic characteristics of the interaction like their social relationship or their role within the conversation. For example, the greeting between two people will depend on their relationship and their meeting context. Similarly, we are more likely to remain static and nod while listening to our partner's arguments. Figure 1 shows some examples of interaction-driven behaviors.

In medicine and psychology, for example, experts analyze the behavior of the patient in social environments in order to detect any communication anomalies. Deficits in eye contact, pointing or expressing emotions represent key markers that might help in the diagnosis of autism spectrum disorders [2], [3]. In other fields like virtual reality or robotics, communication needs to be fully replicated. Therefore, social robots and virtual avatars need to combine all natural modalities involved in a real interaction. In order to induce a human-like social interaction, their facial expressions [4], [5] need to be harmonized with the body pose, the hands gestures and the speech [6]. To do so, we need to fully understand and model the behavior in human-to-human interactions, which represents a challenge involving cognitive, affective, and behavioral perspectives, among others.

In this work, we focus on forecasting human behavior. This problem is especially important in human-robot communication (or personalized assistive agents) where the automated agent may need to predict someone's future behavior in the short-term in order to provide a better and faster response. Similarly, it is of utmost importance to any system to minimise the latencies. We cannot always afford that the person waits until the whole information is processed, so detecting and interpreting the person social signals may help providing with a fast approximation for a starting point, which may be improved as more information is collected. For example, safety driving systems extremely benefit from anticipating pedestrian trajectories, as they may need to make an emergency stop in order to avoid an accident. Also, one may argue that by forecasting human behavior during

Author: German Barquero, barquerogerman@gmail.com
Advisor 1: Dr. Xavier Baró, Department of Computer Science, Multimedia and Telecommunications, Universitat Oberta de Catalunya
Advisor 2: Dr. Sergio Escalera, Department of Mathematics and Computer Science, Universitat de Barcelona
Advisor 3: Cristina Palmero, Department of Mathematics and Computer Science, Universitat de Barcelona
Thesis dissertation submitted: September 2021

Fig. 1. Two examples of conversations where each individual behavior is highly driven by the dyadic interaction. On the top row, the left-most participant nods in answer to the interlocutor question, which results in the latter acquiring a typical posture of someone who is dubious and thinks deeply about something. On the bottom example, both participants engage in a highly interactive conversation which triggers surprise and joy emotions from the right-most participant.

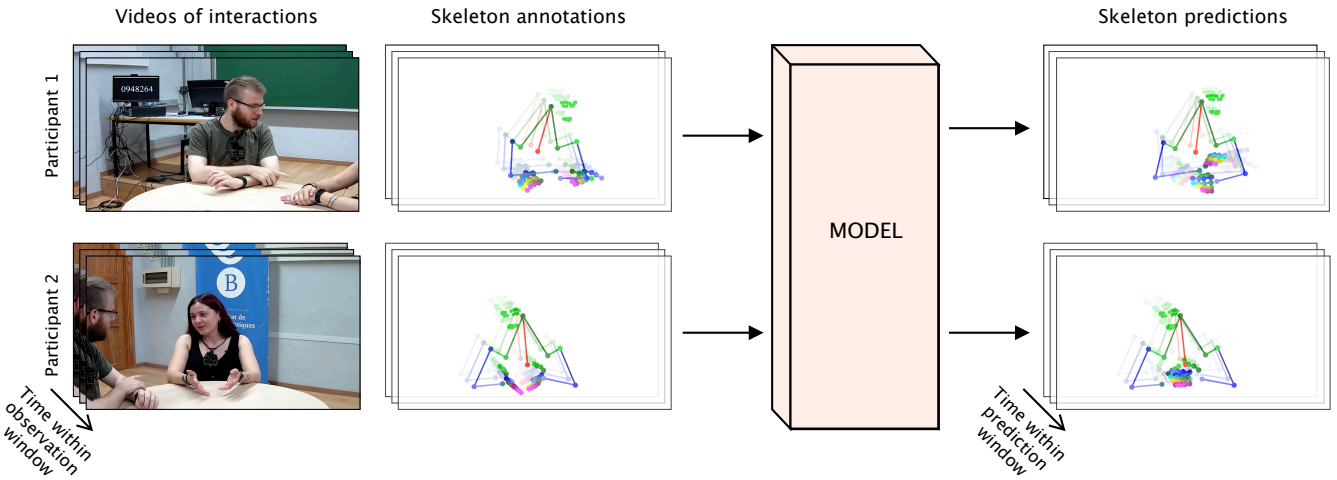

Fig. 2. Representation of the problem tackled in this work. The skeleton annotations including facial landmarks, body joints and hands articulations are available (observation window) and used to predict their movement in the future frames (prediction window).

an interaction, we are also understanding the underlying mechanisms of human interaction and, therefore, uncovering their semantics, context, etc.

To represent the behavior dynamics, we rely on the movement of the skeleton joints of the person, which comprise the main body articulations and facial landmarks, see Figure 2. We have seen that both intra- and interpersonal components affect the human behavior and therefore need to be included in any behavioral model. However, contemplating all possible combinations of components yields a colossal set of scenarios. Additionally, behavior forecasting needs to deal with the stochasticity of the future. In order to alleviate both challenges, many related works focus on predicting the human behavior while performing specific actions [7], [8]. In these scenarios, there are no interpersonal cues and the intrapersonal ones are minimized. Also, the future is narrowed down and becomes more deterministic. Instead, we open a new horizon and define the problem of forecasting the future behavior in dyadic interactions, see Figure 2. The proposed scenarios are unconstrained, action-agnostic, and the participants' behaviors are under the influence of strong interpersonal cues.

The main contributions of this work are summarized as follows.

- We extend the UDIVA dataset [9] with new automatically extracted skeleton annotations for face, body and hands. Annotations underwent a sequence of post-processing steps which reduced errors and improved their quality, especially for hands.
- We propose a recurrent architecture based on the popular Seq2seq architecture [10] to solve the future behavior forecasting problem. We present several experiments where our model outperforms the zero-velocity baseline and also include a first

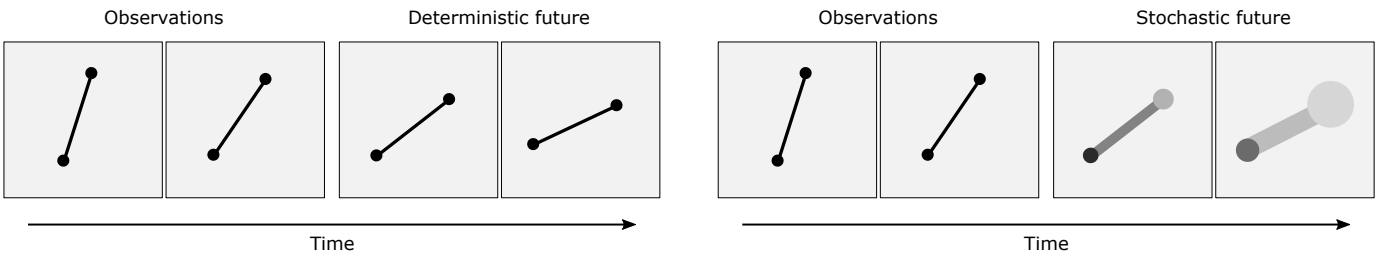

Fig. 3. On the left, a toy example of behavior forecasting in a deterministic environment. On the right, in a stochastic environment. The dots represent two body joints (e.g., the elbow and the wrist), and the line represents the part of the body connecting both joints (e.g., the forearm). Their radius and width correspond to their probability distribution. Darker areas correspond to higher probabilities of the body articulation to be located there at each timestamp. Probabilities get blurrier and averaged as time passes. Figure inspired by [11].

attempt to merge multiple modalities (audio, metadata and dyadic information). We aim at establishing a baseline for future behavior forecasting with UDIVA which we hope will encourage research on this dataset.

- We present a case study where we applied the proposed model to participate in a challenge that was organized in parallel to this work, within the scope of the 2021 International Conference in Computer Vision (ICCV'21). The metrics used in this competition were also explored and defined as part of this work.

II. State of the art

In this section, we review previous research outcomes relevant to human behavior forecasting. We also revise the literature on dyadic human interactions. Finally, we debrief the state-of-the-art on face, body and hand landmarks estimation.

A. Behavior forecasting

The multimodal nature of future prediction makes it an extremely challenging problem. For instance, the much more generic problem of video future prediction often results in blurred predicted sequences of frames [11]. These are representations of the future uncertainty as the average of all the possible futures. A visual representation of this effect can be seen in Figure 3. While being unavoidable, it can be alleviated by reducing the complexity of the prediction and increasing the contextual information provided. In an intent to reduce the future stochasticity, many works focus on leveraging much lower-dimensional data: body poses. In the past years, human body poses and facial expressions have been leveraged to recognize behavior [12], infer intents [13] or detect behavioral and emotional disorders [14]. Similarly, behavior forecasting can be reduced to a future pose estimation problem, which is a relatively new problem compared to predicting image or video pixels.

Stochastic predictions. Some works embrace the future uncertainty and exploit it by predicting multiple futures. In [7], a conditional variational autoencoder (CVAE) takes the trajectories of several joints and treats the task as a pose completion problem. Its main limitation resides in its very nature: future trajectories for at least one joint need to be provided. This is not a limitation for [15], in which the latent vector is brought to the frequency domain to improve the diversity of the predictions while retaining accuracy. The realism of the predicted poses is ensured by recursively fitting a 3D body model to them. The main drawback is that the recursive optimization slows down the inference process and nothing guarantees the motion realism. With a similar idea, a novel sampling strategy, DLow, was proposed to promote the poses variability of future human motions [16]. Conditional Generative Adversarial Networks (GAN) have also been widely used to improve the futures diversity. For example, Long et al. suggest to use a latent code to reflect various behaviors and an attraction point to reflect various trajectories, which need to be provided at inference time [17]. This strategy also alleviates the mode collapse frequently suffered on adversarial training. In [18], a sequence-to-sequence Wasserstein GAN which enforces pose and motion realism by adding a gradient and a bone loss is proposed.

Deterministic predictions. Other works, instead, consider the future to be deterministic. Compared to the stochastic methods, the results of these methods can be easily evaluated by directly comparing them to the ground truth. Most of them add constraints or provide contextual information in order to narrow the future space and therefore reduce its stochasticity. The first methods proposed were encoder-recurrent-decoder architectures which modeled human kinematics [19]. In a similar fashion, a sequence-to-sequence (Seq2seq) recurrent neural network (RNN) with residual connections was proposed as a forceful action-agnostic baseline [20]. Led by their success at natural language processing (NLP), hierarchical multiscale RNNs were favorably translated to human pose forecasting in order to deal with the wide diversity of motion patterns of the body parts and their distinct motion dependencies [21], [22]. However, recurrent approaches tend to suffer from errors accumulation and struggle with long-term predictions. To overcome these challenges, additional strategies have been proposed. In [23], a dropout autoencoder LSTM (Long Short-Term Memory network) and a 3-layer LSTM are combined to implicitly learn and model the structural and the spatio-temporal aspects of the task being performed. Other works apply attention to capture the long-range spatial correlations and temporal dependencies [24], [25]. Very interestingly, Want et al. transformed the human pose prediction into a reinforcement learning problem [26]. They proposed an imitation learning algorithm which learns to make accurate short

and long-term pose predictions with very fast training speed. Instead of predicting the whole sequence of future poses, a recent work has proposed to predict meaningful key moments in a future action [8]. They argue that this formulation is better suited for intent and action forecasting as it disentangles the temporal and intentional aspects of human actions. Unfortunately, this formulation is not applicable to action-agnostic contexts like ours.

Exploiting interactions. The future behavior of individuals is strongly determined by their interaction with other individuals and objects in their surroundings. By definition, graph convolutional networks (GCN) are able to encode such interactions. Some works have successfully used GCNs to generate context-aware encodings which are fed to RNNs decoders [27], [28]. They all come to the same conclusion: the contextual information contains very valuable information regarding the human future behavior. Other solutions proposed went beyond and considered each joint as an individual agent [29], [30]. The rationale behind this idea is that treating the human body as a graph instead of a skeletal kinematic tree helps to capture long range dependencies among joints [29]. Li et al. extended this idea by adding a multiscale graph computational unit which fuses features across different body scales [30]. Very recently, Liang et al. have proposed a novel architecture which merges the history of the agent with the current graph of agent trajectories [31]. Although it was originally designed for autonomous driving, encoding the past history of human motion may become useful to detect and predict person-specific behaviors. Unfortunately, all aforementioned methods skip a very frequent and critical problem observed in many applications: missing or incomplete pose observations. In order to repair such observations, Cui and Sun propose using an additional GCN trained to identify erroneous poses [32].

B. Dyadic interactions

Although the literature review on human behavior forecasting is extensive, few works focus on conversations between two (dyadic) or more interlocutors. In a situation where a group of people is interacting, the positions and orientations of the individuals, their body and hand gestures, gaze, and facial expressions become extremely relevant for behavior forecasting. Also, other semantically important content like speech, voice tone or other information related to the interlocutors may influence their behavior. Naturally, multimodal information needs to be exploited in a specific way in order to fully profit from it.

In order to mix monadic and dyadic information including speech, Ahuja et al. proposed a dyadic residual-attention model [33]. Results show the temporal evolution of the importance of both inter- and intrapersonal dynamics when predicting human behavior. Similar conclusions were extracted from [34], where facial gestures synthesized from dyadic information were preferred over those generated from monadic data. Other studies have recently reported that considering the personality traits of the interlocutors also enhances the prediction of nonverbal behaviors [35], [36]. In order to exploit this, Ahuja et al. present an architecture which encodes behavioral patterns from both individuals and applies style transfer to the raw predicted human behavior [37]. Despite the relevance of the methodology they propose, they aim at synthesizing motion that matches audio for social conversations, which is not our use case. We can assume that leveraging linguistic information helps reducing the stochasticity of the interlocutors future behavior, as it incorporates more semantic information of the interaction taking place. However, the mean-convergence effect observed in the future prediction field is still reported in social behavior prediction [38].

Within the scope of this project, behavior forecasting is limited to the prediction of the image coordinates for facial, body and hand landmarks in the subsequent frames. As reported in most of the literature reviewed, this problem simplification aims at reducing the future uncertainty and narrowing the prediction space to a single solution. In order to support our choice on the methods used for landmarks retrieval, we thoroughly reviewed the state of the art on human landmarks estimation.

C. Landmarks estimation

Human pose estimation has extensively attracted the interest of the computer vision community. Its attractiveness resides in the wide spectrum of fields where it proves useful: robot interaction [39], [40], virtual reality [41], autonomous driving [42] or medicine [43], [44], among others. When it comes to perception and human behavior analysis, the body pose, the facial expression and the hands gestures become of utmost importance.

The estimation of facial landmarks have been intensively studied, yielding very satisfactory results. Recent methods can be split into those regressing 3D morphable model parameters [45], [46], [47] and those directly regressing the 2D/3D coordinates of the points [48], [49]. Feng et al. present a detail-consistency loss which disentangles the expression from the face and helps retrieving more expressive face shapes [50]. At the same time, it allows them to synthesize realistic expressions while keeping person-specific details unchanged. As the popularity of video-based applications increase, new methods incorporate video-based training, which improves the stability of 3D face alignment in videos [51].

The body joints estimation problem has been given roughly the same importance. Methods in the literature have traditionally been enclosed in two categories: bottom-up and top-down approaches. Bottom-up methods leverage heatmaps and anthropometric heuristics to build the human skeleton [52]. The main limitation resides in their inherent bias towards false positives and their low accuracy on truncated (i.e., partially visible) bodies. In [53], the authors work in a metric 3D space instead of the image space (image coordinates in pixels) in order to address the latter limitation (the whole body is always available). Top-down methods, on the other hand, focus on first detecting the location of the person of interest within the image in order

Fig. 4. On the left, setup for the six tripod-mounted cameras of the UDIVA dataset. GF: General frontal camera, GB: General rear camera, HA: individual high angle cameras, FC: individual frontal cameras, E: ego cameras. On the right, synchronized views from the 8 cameras.

to reduce the number of false positives [54]. On the downside, the person detection step may become a bottleneck to handle occlusions and can increase their computational complexity.

However, the translation of these methods to hand pose estimation is not straight-forward due to its unique characteristics. The complex physiology of hands, their rapid movement and their frequent interactions between them or with external objects make it a very challenging problem [55]. As a result, most of them rely on a hand detector so that the hand region can be handled at higher resolutions [56], [57]. Some of these works directly estimate the 3D joints [58], [59], [60] or 3D meshes [57], while many others propose parametric models which estimate configurations of statistical models like MANO [61], [56].

III. The UDIVA dataset

In this section, we briefly describe the UDIVA dataset (Understanding Dyadic Interactions from Video and Audio signals) [9], and present the version v0.5 used in this work.

A. Dataset description

The UDIVA dataset was born as a benchmark to study and understand the mechanisms of influence, perception and adaptation to verbal and nonverbal social signals in dyadic interactions. The dataset was carefully tailored so that the individual and dyad characteristics as well as other contextual factors could be accessed and easily analyzed. The 147 voluntary participants (44.9% female) ranged from 4 to 84 years old (mean=31.29) and came from 22 countries (68% from Spain). They were distributed into 188 dyadic sessions (90.5h of recordings), with an average participation of 2.5 sessions/participant (max. 5 sessions). During the session, they were asked to speak the language both felt most comfortable with. Spanish was the most popular choice (71.8%), followed by Catalan (19.7%) and English (8.5%). Pairs of participants were chosen according to their availability while trying to preserve a close-to-uniform distribution among all possible combinations between variables (gender, age group and relationship between interlocutors).

Participants sat around a table forming a 90º angle with each other to avoid occlusions in the cameras field of view, and close enough to favour the interaction in collaborative activities. The camera setup consisted of six HD tripod-mounted cameras (1280x720px, 25fps): two general frontal and rear cameras, two individual high angle cameras and two frontal cameras, see Figure 4. Additionally, participants wore an egocentric camera (1920x1080px, 30fps) around their necks, and a heart rate monitor on their wrist. The audio was acquired with two individual lapel microphones and an omnidirectional microphone on the table.

Before their first session, participants filled a sociodemographic questionnaire, including: age, gender, ethnicity, occupation, maximum level of education, and country of origin. Personality was also assessed through age-dependent standardized questionnaires. For all sessions, participants completed pre- and post-session mood and fatigue assessments. After each session, participants filled again the personality and mood and fatigue questionnaires about their interlocutor in order to provide their perceived impression.

During the session, participants engaged in five different tasks which were designed by psychologists to elicit distinct behaviors and cognitive workload: *Talk*, *Animals*, *Lego*, *Ghost* and *Gaze*. In *Talk*, participants talked about any subject for 5 minutes. The quality of interaction, empathy and synchrony, among others, can be observed in this task. In *Animals*, participants asked 10 *yes/no* questions each in order to guess the animal from the picture on each other's forehead, which mainly reveals cognitive processes and features gaze events. The cooperative *Lego* building task enforces collaboration, joint attention, and leader-follower behaviors. In the *Ghost* task, participants played the competitive "Ghost blitz" cards game, which fosters cognitive processing speed analysis. Finally, the *Gaze* task was recorded while participants followed directions to look at the interlocutor's face, at static/moving object, or elsewhere. This may represent a useful ground truth for gaze gestures and face modeling with varied head poses.

B. Dataset extension: v0.5

The UDIVA v0.5 dataset is still a preliminary version of the UDIVA dataset [9]. It includes a subset of the participants, sessions, synchronized views, and annotations of the complete UDIVA dataset. This dataset contains 145 dyadic interactive sessions with 134 participants (ranging from 17 to 75 years old, 44.8% female). Each session features two subjects participating together in 4 of the 5 original tasks: *Talk*, *Lego*, *Ghost* and *Animals*. As in the original dataset, the same person was allowed to participate in up to 5 sessions with different interlocutors. The dataset is released divided into training, validation and test sets, which include 116/99, 18/20 and 11/15 sessions/participants, respectively. These splits were selected following a greedy optimization procedure which tried to keep a similar distribution in each split regarding each session and participant characteristics. Participants are not shared among splits.

In addition to the recordings of the 4 tasks, the transcripts of the session conversations and a set of automatically extracted annotations are included. The latter contains the landmarks for face, body and hands and the 3D eye gaze vector. Their extraction, post-processing and cleaning were part of the thesis and they will be thoroughly described next:

1) Landmark extraction and post-processing: Face, body and hands landmarks were extracted from both individual frontal views (*FC1* and *FC2*) of the 4 available tasks: *Talk*, *Lego*, *Ghost* and *Animals*.

- **Face landmarks.** 68 face fiducials were regressed by the 3DDFA_v2 algorithm [51], [62]. This method uses a lightweight backbone together with a landmark-regression regularization to achieve state-of-the-art accuracy at very fast speeds. The method also incorporates a short-video-synthesis training strategy, which helps retrieving stabler landmarks for videos. Additionally to the facial landmarks, the face detection confidence provided by FaceBoxes [63] was stored. The face landmarks retrieval was constrained to the most centered face detected within each frame, discarding false detections product of the occasional profile view of the interlocutor. In order to reduce the jittering, the landmarks coordinates from frame t were average-smoothed with those from frames $t-1$ and $t+1$.
- **Body landmarks.** 24 full-body joints and a detection confidence were retrieved by using the MeTRAbs method [53], [64], which beat all the tested methods thanks to performing particularly well with truncated upper bodies. This top-down algorithm detects the body and leverages volumetric heatmaps to extract 2D landmarks in the image space (image coordinates in pixels) along with their corresponding 3D landmarks in the camera coordinate frame. Additionally, detection mistakes were identified and fixed by leveraging a tracker which enforced a spatio-temporal continuity [65]. An extra post-processing step was applied to translate the 3D coordinates to the image space while preserving the depth value. First, we found the 3D similarity transformation T which minimized the least squares problem $TX_{3D->2D} = Y_{2D}$ where $X_{3D->2D} = (x_{3D}, y_{3D})$ (pose camera coordinates), and $Y_{2D} = (x_{2D}, y_{2D})$ (pose image coordinates). The matrix T was defined to have 8 degrees of freedom: 3 for rotation, 3 for translation and 2 for scaling (this parameter was fixed to 1 for z). Then, T was applied to $X_{3D} = (x_{3D}, y_{3D}, z_{3D})$ vector. The resulting $TX_{3D} = Y_{3D}$ vector was stored as the 3D body joints coordinates. Finally, a one-euro-filter was applied in the temporal axis in order to remove the jitter (cut off to 0.001 and $\beta = 0.005$).
- **Hand landmarks.** 21 hand landmarks were retrieved with the hand estimator module from FrankMocap [56], [66]. This method first detects the hand and then fits a 3D model, which provides 3D hand landmarks. The hand detector leveraged was trained with 100K images featuring a wide range of hand interactions with either objects or themselves [67]. The hands landmarks estimator was trained with the Ho-3D dataset, among others, which contains 60K samples aiming to study the interaction between hands and objects. As a result, the method infers fairly accurate landmarks in the recurrent scenario where hands are interlaced, interacting with objects or mildly occluded, see Figure 5. Similarly to the procedure followed for the body detections, a tracker ensured a spatio-temporal smoothness and filled the gaps of hands missed due to rapid movement or severe occlusions [65]. Unfortunately, this scenario is much more challenging and required of further processing to increase the quality of the annotations, especially to ensure the consistency of the left-right hand associations. This process has been extensively described in Appendix A. The one-euro-filter previously described was likewise applied (0.001 and a $\beta = 0.02$).
- **3D eye gaze vector.** The ETH-XGaze baseline method [68], [69] used the face fiducials previously retrieved to extract the gaze vectors. The method was trained on the largest-scale gaze dataset existing up to date, with huge variability in terms of appearance, head, poses, gaze directions, and accessories like glasses. The reported error of the method is at the level of other state-of-the-art approaches.

Note that the landmarks used for this work were simplified and figures only show 28 landmarks for face (face contour and nose skipped), 10 for body (upper body) and 20 for hands (wrist not considered). More details can be found in Section IV-A.

2) Data cleaning: In order to ensure a fair evaluation, landmarks of the *Talk* task from the validation and test sets underwent a visual inspection process. The selection of this task was driven by its highly interactive nature, which could provide the most valuable insights about the underlying mechanisms of human-human interaction. Also, the landmarks extraction is less challenging due to the lack of occlusions and interactions with objects. The gaze vector was not visually assessed as it will not be evaluated either. All frames from the 18 validation and 11 test sessions were annotated as follows:

Fig. 5. Face, body and hands landmarks, and gaze vector displayed over examples of the *Talk*, *Animals*, *Lego* and *Ghost* tasks from 4 sessions (left to right).

	Validation set						Test set					
	Correct	Mild	Severe	Switched	Fixed Visibility	FBI	Correct	Mild	Severe	Switched	Fixed Visibility	FBI
Face	99.3	0.6	0.1	-	-	-	99.7	0.2	0.1	-	-	-
Body	98.1	1.9	0.0	-	-	-	97.2	2.7	0.1	-	-	-
Left hand	88.9	7.2	3.9	0.4	2.6	6.6	87.0	8.0	5.0	0.2	1.6	7.1
Right hand	90.1	5.4	4.5	0.4	2.0	6.9	82.3	9.7	8.0	0.2	3.4	7.0

Table I. Visual annotation process for validation and test sets: prevalence of each label (% of frames). FBI: Fixed By Interpolation

- **Face annotation.** The quality of the face landmarks was assessed and classified for each frame as: 1) *correct*, if all landmarks faithfully matched their anatomical locations, 2) *mild*, if the retrieved face was slightly translated with respect to its anatomical position but its shape and orientation were correct, and 3) *severe*, if either the shape or the orientation was wrong. Figure 6a shows visual examples for each label.
- **Body annotation.** Similarly, the quality of the body joints was annotated. Comparatively, we were more permissive regarding the quality thresholds as the body joints estimator yielded noisier predictions in our challenging scenario characterized by truncated bodies. Consequently, body landmarks were classified as *correct* provided that their pose was correct and their individual joints locations matched their anatomical position accounting for certain error. The *mild* label was associated to body landmarks with one inaccurate side, and the *severe* label to those either with both sides incorrect or with severe single joints mislocations.
- **Hands annotation.** The quality flag was set to *correct* if their orientation was correct and fingers matched their anatomical position allowing for some error. If the fingers did not match either in shape or position but the overall orientation was still right, they were labeled as *mild*. On the contrary, if neither the hand orientation nor the fingers were correctly inferred, the hand quality was labeled as *severe*. Additionally, the hand visibility (*visible* or *not visible*) and cases where the left hand was detected as the right hand or vice versa (hands *switch*) were annotated. Although the methods presented in Section III-B1 substantially improved the quality of the hand landmarks, wrong or missing hands were still very frequent during hands interactions or occlusions. In order to maximize the number of correct hand annotations, sequences of consecutive frames $(t_0, ..., t_n)$ with $n \geq 2$ with correct hand landmarks at t_0 and t_n but with wrong or missing hands in all frames $\{t_i\}_{1 \leq i \leq n-1}$ were identified. From those, segments for which a linear interpolation between the t_0 and t_n landmarks could generate valid hands for the interval (t_0, t_n) were annotated. Such linear interpolation was applied as an extra post-processing step, see Figure 6b.

Note that, except for the described cases for the hands, the landmarks were not manually fixed. The hands linear interpolation proved extremely useful as it fixed up to 7% of the annotations in both validation and test sets, see Table I. However, the quality of the hands annotations still represents the biggest limitation of the dataset, as the number of *correct* hands annotations varies from 80% to 90% in the validation and test sets.

3) Segments generation: In this work, we consider an observation window of 4 seconds (100 frames), which we consider to provide a reasonable conversational context for behavior forecasting. The length of the window to predict (prediction window) was set to 2 seconds (50 frames), which allows us to evaluate both short- $(0 - 400ms)$ and long-range behaviors $(> 400ms)$ [70], [11], [7]. Each pair of observation window and prediction window will be hereafter called a *segment*. These segments conceptually include information of both participants.

In order to fairly evaluate dyad-driven behaviors, the landmarks in these segments had to be as accurate as possible for both session participants. Ideally, the evaluation could have been constrained to the 150-frames-long sequences with *correct* face, body and hands landmarks for both participants. However, this restrictive scenario implied an upper bound for the amount of

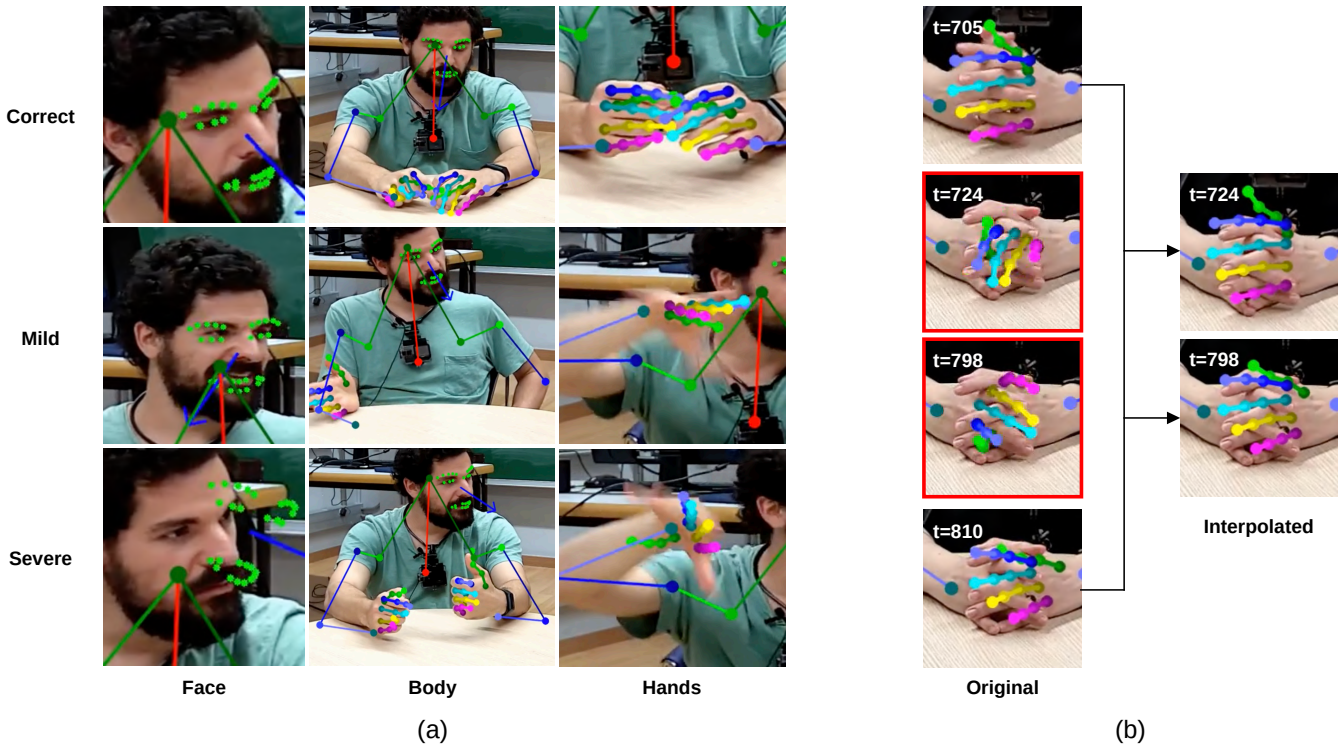

Fig. 6. In (a), examples of *correct*, *mild* and *severe* quality labels for face, body and hands landmarks. In (b), a sequence of 104 frames with wrong left-hand landmarks which was fixed by linearly interpolating the landmarks from the last and the first correct extractions before (frame 705) and after (frame 810) the sequence, respectively.

segments of 288 for the validation set and 97 for the test set (28.2% and 16.2% of total frames used, respectively). In order to keep a good trade-off between landmarks accuracy and number of segments fulfilling the constraints, the *correctness* condition was only required for face, body and at least one hand (if visible) for both participants. This increased the previous upper bound up to 658 and 300 segments for validation and test sets, respectively.

The non-overlapping segments selection can be performed in many different ways. Importantly, the set of selected segments to be predicted needs to be as diverse as possible so that it represents the whole spectrum of human behavior possibilities. Unfortunately, such behavioral diversity can be difficult to model. We simplify it by assuming that such variety can be preserved by enforcing the diversity of movement velocities (magnitude and angle) and hands visibility and correctness. Therefore, for each candidate segment, the angles and speeds of the 28 face landmarks, 6 body arms landmarks and 20x2 hand landmarks were computed. For each participant, two 2D histograms were computed per body part: one for the observation window and one for the prediction window. The histograms were generated with 3 bins for speed ($[0, 3)$, $[3, 6)$ and ≥ 6 px/frame for face and body, and $[0, 4)$, $[4, 8)$ and ≥ 8 px/frame for hands) and 11 bins for angles (equally split from $-\pi$ to π) and normalized. The raveled histograms were concatenated along with the percentage of correct/visible hands (8 values, 2 for each hand of both participants), yielding a feature vector of size 9784 per candidate segment (4892 for each participant).

The feature vectors for all candidate segments from the validation set were grouped into N clusters with $N = 1, ...19$ by the K-means clustering algorithm [71]. The inertia resulting from each clusterization was computed and is shown in Figure 7a. This value reached a plateau for $N \geq 12$, so we chose to classify our segments into 16 clusters. The distribution of candidate segments per cluster was not evenly distributed (Figure 7b, left column), so a greedy sampling strategy was followed: non-overlapping segments were selected starting from the least populated cluster. This strategy yielded 598 validation segments (37.4 ± 10.4 segments per cluster) following a more uniform distribution (Figure 7b, right column). Same clusters were used to classify the test candidate segments, resulting in 278 test segments (18.5 ± 13.6 segments per cluster).

IV. Methodology

In this section, we first describe the pre-processing steps we apply in order to prepare our landmark annotations, metadata and audio feature vectors. Then, we propose a recurrent architecture which attempts to solve the behavior forecasting problem.

A. Pose representation

There are several challenges associated to our automatically extracted skeleton annotations. In the first place, our three skeleton entities (face, body and hands) were independently extracted. As a result, the depth coordinates are not within the

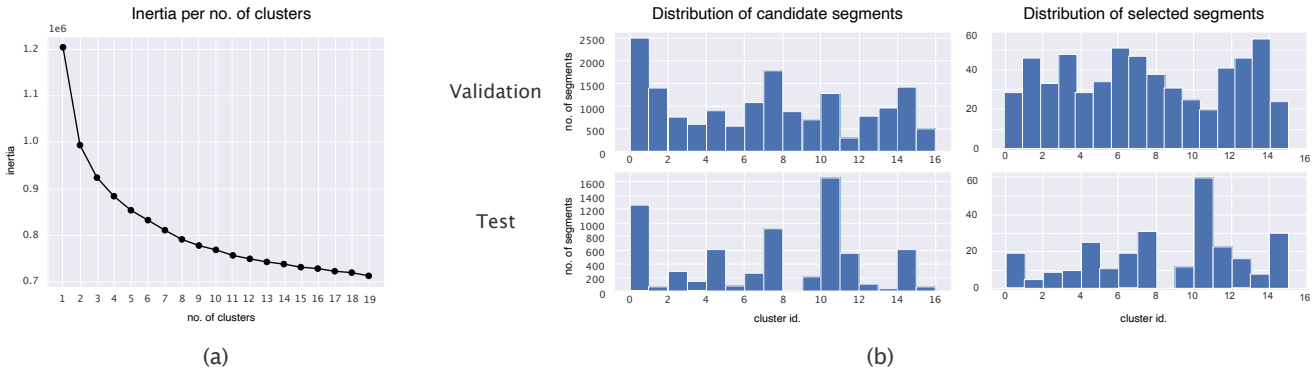

Fig. 7. In (a), the inertia curve when clustering the candidate segments according to their feature vectors into $1 \leq N < 19$ groups. In (b), distribution of the candidate segments (left column) and the sampled final segments (right column) from the validation and test sets (top and bottom rows, respectively).

same plane. This makes it more difficult for the model to extract useful insight from the 3D coordinates. Secondly, we need to detach our predictions from the image space in order to compress the subspace of poses that the network needs to model. At the same time though, the location and trajectory of the person within the video view (e.g., centered, closer to its interlocutor, etc.) is relevant and strongly determines their future behavior. Therefore, our predictions cannot be fully trajectory-agnostic.

The first challenge is not solvable, so we assume that our three spaces of coordinates will be independent and that the depth coordinate may become an important source of noise. In order to alleviate the second problem, some works propose working with offsets instead of poses. This strategy isolates the input and the predictions from the image space, and thus simplifies the subspaces where they belong to [20], [72]. Unfortunately, this solution would aggravate the third problem, as it completely detaches the behavior forecasting from the person position in the scene. In order to find the right trade-off, recent works have shown that mixing root-relative position, velocity and acceleration is beneficial for future poses forecasting [73], [74]. Following this trend, we combine the joints' root-relative positions and their offsets to create our input. To do so, we define three roots: 1) the middle point between both eye pupils, 2) the middle chest joint, and 3) the middle knuckle. They are used to define the face, body and hands root-relative coordinates, respectively, which are computed by subtracting the roots' (x, y, z) coordinates from each joint's (x, y, z) coordinates. The offsets from frame t correspond to the difference between coordinates in frames $t - 1$ and t. We store the values of the roots as the trajectory of each skeleton entity.

In order to reduce the complexity of our problem, we only worked with a subset of landmarks. Landmarks were chosen according to their relevance for human behavior analysis. For face, we only kept 28 landmarks: 5 landmarks per eyebrow, 3 landmarks per eye (4 inner points averaged as the eye center, and both end points), 5 landmarks per lip (computed by averaging the top and bottom edges of each lip), and the 2 mouth extremes. From body, we only kept the 10 upper-body joints (skipping the head joint). We kept all the landmarks of the fingers (4 joints each) and discarded the wrist one, which is especially noisy. Missing face, body or hands were replaced by an array of zeros. Additionally, the trajectory of the roots within the image also needs to be predicted in order to recover the landmarks in the original video sequence. Therefore, the image coordinates of the face root were included as an extra landmark, and the image coordinates of the body and hands roots replace the roots' root-relative coordinates, which are 0. As a result, the full-body skeleton consists of 79 landmarks.

B. Metadata and audio

Our work includes a first attempt to leverage multimodal data to forecast future behavior. To do so, we pre-processed the metadata and the audio so that it could be easily fed into any network.

Metadata. We associate an array of metadata values to each segment which contains two types of information: *participants* and *session* data. For the former, the age, gender, country of origin, education level and self-reported personality (Big Five traits [75]: openmindedness, conscientiousness, extraversion, agreeableness and negative emotionality) are included for each participant. For the latter, the language spoken, the relationship between participants (known or not), and the pre-session mood. The mood was assessed with 8 categorical values: good, bad, happy, sad, friendly, unfriendly, tense, and relaxed. First, categorical variables were transformed to numerical ones by assigning an integer to each possible value. All values were then normalized to the [-1,1] range using their statistics in the training sessions. The metadata arrays contained a total of 39 values (18 for each participant and 3 for the session).

Audio. The audio feature extraction was done with VGGish [76], which was specifically developed for the audio modality. The pre-trained weights were learned on the popular YouTube-8M dataset [77], and the model generates feature vectors of size 128 for audio chunks of 1 second. Therefore, we extracted and concatenated 4 non-overlapping feature vectors from the audio of each segment's observation window of 4 seconds. As a result, the size of the audio feature vector is 512. All audio feature vectors were normalized to the [-1, 1] range using the statistics from the training sessions.

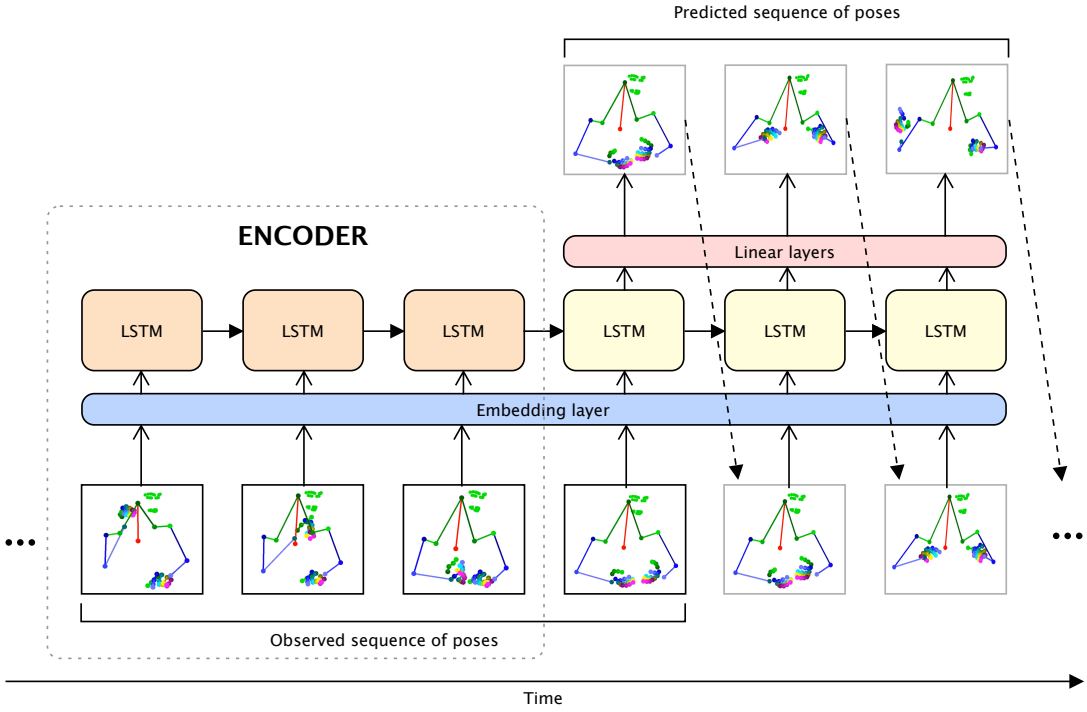

Fig. 8. Proposed Seq2seq architecture for future behavior forecasting based on skeletons. The skeletons are embedded and then recurrently fed into a long short-term memory encoder (LSTM). The beginning of the decoding process is triggered by the last observation. The decoded future poses are recurrently embedded and sent through the decoder until the whole sequence of future poses is predicted.

C. Architecture

The proposed architecture is based on the sequence-to-sequence (Seq2seq) model, originally adopted from the natural language processing field [10], see Figure 8. The choice of this architecture lays on the idea that the skeleton behavior from the observation window can be encoded into a feature vector representation (two if we use a LSTM unit as recurrent cell), which can be merged with the last observed skeleton to effectively predict the next skeleton pose. To do so, the root-relative positions of the participants' skeletons and their offsets are concatenated, flattened and then embedded by going through a dense layer followed by a non-linearity. In our unimodal and main approach, the embedded input from each participant is treated independently and fed to the same encoder. This means that we are not including information from participant 2 to predict the behavior of participant 1 (we include an experiment which does, in subsection V-D). Finally, the last observation (i.e., the landmarks from the last frame of the observable window) is fed to the decoder (also shared between both participants) which also uses the encoder's last hidden and cell states to predict the pose offsets. These offsets are added to the coordinates of the previous skeleton in order to generate the next pose, which is again fed to the decoder. This process is repeated until the whole prediction sequence is generated.

In the final model, both encoder and decoder consist of a one-layered LSTM unit. The choice for the non-linearities is a leaky ReLU with negative slope of 0.01. The layers' configurations have been experimentally set. The output of the embedding layer is 512, and the sizes of the LSTM's hidden and cell states are 1024. The output size is 1024 for the two sequential linear layers. The network has a total of 15.2 millions of parameters.

V. Experiments and Results

In this section, we present the experimental results of this work, which mainly focuses on the recurrent Seq2seq approach. We first show the importance of choosing an appropriate input structure, the effects of dataset scaling and how the recurrent approach outperforms the two considered baselines. Then, we extensively discuss the trade-off between short (< 400ms) and long-term (< 2s) predictions, and the benefits of focusing on one body part at a time instead of predicting all of them at once. Finally, we evaluate the contribution of the different modalities (e.g., dyadic information and audio features) when added to the baseline (i.e., only skeleton landmarks).

Baselines. The simplest baseline for behavior forecasting using skeletons consists in propagating the landmarks from the last observed frame into the future, as if the person froze once the observation window finished (*zero-velocity baseline*), see Figure 9a. While it may seem counterintuitive, the zero-velocity baseline has been proven a very strong and difficult to improve baseline [20]. This is especially true in our use case where most parts of the participants body remain static while listening to their partner's speech. We also include a baseline inspired by the temporal convolutional networks (TCN) [78]. This network

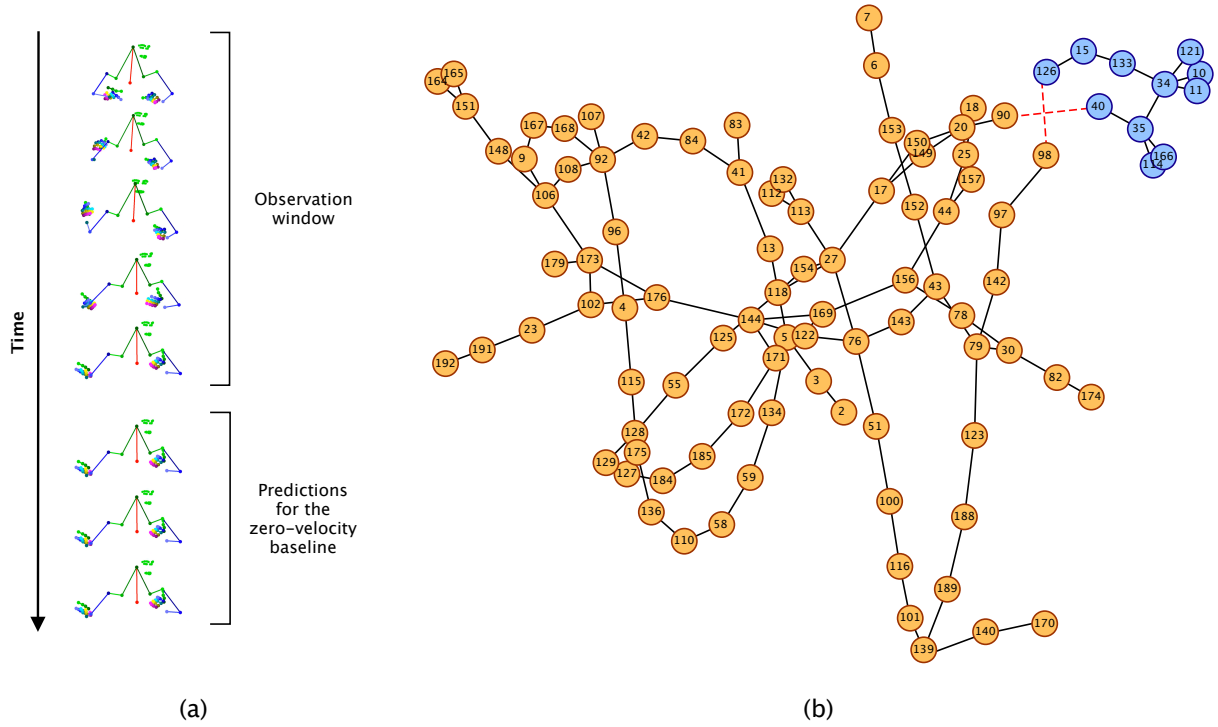

Fig. 9. In (a), a visual example of the zero-velocity baseline. It propagates the last observation into the future predictions as if the participant became frozen immediately after starting to predict. In (b), a graph representing the participants (nodes) and the sessions (edges connecting their two participants) from the training set of the UDIVA dataset v0.5. In red dashed lines, the two sessions removed in order to extract an independent validation set (blue nodes).

is bigger than the Seq2seq (30.9M) as we experimentally found that bigger dense layers at the output yielded better results. TCNs have proven to work as well as recurrent models in many practical applications [79]. Their main advantage is that they can model long-range dynamics with relatively few parameters. However, their main disadvantage is the fixed size of the output sequence. Our TCN architecture is simple and consists of three 1D temporal convolutions (64, 128 and 256 kernels of size 3), followed by a convolution with 1 filter which efficiently fuses the output of the convolutions before feeding it to two sequences of a dense layer followed by a non-linearity (Leaky-ReLU). In this work, most methodological exploration focuses on the recurrent approach, as the length of the sequence to be predicted can be dynamic, which suits very well the needs of this project.

Dataset. We trained our models with the training set from the UDIVA v0.5 dataset. We split its 116 sessions (99 participants) into training (90%) and validation (10%) subsets. In order to avoid any participant appearing in sessions from both sets, two sessions were removed (040090 and 098126), see Figure 9b. As a result, training and validation splits contained 102 and 12 sessions, respectively. From each session of the training set, consecutive segments of 150 frames with a stride of 1 were extracted. For the validation sessions, the stride was set to 100. From both set of segments, segments with a missing hand in the last observed frame but visible in the sequence to predict were filtered out. The rationale behind this decision is purely conceptual. Since the network's output are offsets, we can't reconstruct the full skeleton of the predicted sequence in the image space if the hand is missing in the last frame of the observation window.

Implementation details. The network was trained to predict the displacement offsets for each landmark. It was trained by minimizing the mean squared error (MSE, see Equation 1) between the 2-dimensional joints from the predicted sequence, which is reconstructed from the predicted offsets (Y), and those from the ground truth ($\hat{Y}$). In our scenario, $T = 50$. N varies in function of the experiment, i.e., the parts of the body we predict. For the scenario where all body joints are predicted, $N = N_{face} + N_{body} + 2 \cdot N_{hand} = (28 + 1) + 10 + 2 \cdot 20 = 79$, see Section IV-A for more details.

$$MSE = \frac{1}{N \cdot T} \sum_{t=1}^{T} \sum_{i=1}^{N} (Y_i^t - \hat{Y}_i^t)^2 \tag{1}$$

We train all the instances of our architecture with the AMSGrad variant of the Adam optimizer [80] (weight decay to 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a batch size of 512) with learning rate set to 0.0001, and dropout to 0.5 (0.3 for the TCN). Early stopping was applied to all the models when validation loss stopped decreasing for 10 epochs in a row.

Evaluation metrics. The MSE described above is computed in the root-relative space of coordinates to avoid reconstructing

Architecture	Tasks	Input*	MPJPE	ST	MT	LT	FDE	Δ
Zero-velocity baseline	-	-	24.56	9.61	23.95	32.75	35.33	0.00
TCN	*Talk*	{0,1}	25.22	10.24	24.79	33.23	35.45	70.13
TCN	All	{0,1}	25.45	10.29	25.00	33.57	35.72	66.63
Seq2seq	*Talk*	{0,1}	24.78	9.69	24.19	33.03	35.50	40.92
Seq2seq	All	{0,1}	24.19	9.53	23.62	32.19	34.62	38.62
Seq2seq	All	{0}	24.27	9.71	23.85	32.07	34.37	36.93

Table II. Comparison of the proposed Seq2seq model against the zero-velocity baseline (propagating last observation into the future) and the temporal convolutional baseline (TCN). While TCN does not scale well with the dataset size, the Seq2seq model benefits from training with all the tasks (4 times more data). As a result, our model beats the zero-velocity baseline. Including the first order input helps the network to better model short-term behaviors. * Zero (0, landmark positions) and first (1, velocity offsets) order inputs. MPJPE: Mean Per Joint Position Error, ST: Short-Term (0-400ms), MT: Mid-Term (400ms-1.2s), LT: Long-Term (1.2-2s), FDE: Final Displacement Error, Δ: divergence.

the pose in training time. In exchange for their straight-forward and fast computation, they do not provide us with information on the real distance between the model predictions and the ground truth. Therefore, we evaluate our trained models with the Mean Per Joint Position Error (MPJPE), which has been broadly used in the past for the evaluation of pose estimation methods [81], [82]. It is computed as the mean of the L2 distances between the ground truth and the predicted landmarks in all frames. We also present three MPJPE variants: the short-term (ST), the mid-term (MT), and the long-term (LT) versions by restricting it to the intervals of frames 1-10 (0-400ms), 11-30 (400ms-1.2s) and 31-50 (1.2s-2s), respectively. We also calculate the Final Displacement Error (FDE) [83], which corresponds to the MPJPE of the last predicted frame of the sequence, and helps to depict divergences in the very long-term. Similarly to the loss computation, missing body parts were skipped. Finally, we compute the divergence (Δ) as the norm of the predicted offsets from all frames. This metric quantifies the amount of motion in the predicted sequence of skeletons.

Evaluation segments. The evaluation was performed on the raw annotations of the UDIVA v0.5 validation set (18 sessions), hereafter referred to as *test set*. We grouped all the training segments into 30 clusters with K-means algorithm, using the feature vectors described in Section III-B3 without the *correct* flag. These were used to classify the consecutive overlapping segments extracted from the test set (stride of 1). Then, these were greedily sampled to generate the set of 1616 non-overlapping test segments used to evaluate our models.

A. Input order and dataset scale

As previously discussed, recent works argue on the benefits of including first and second order inputs (landmarks' position and velocity offsets, respectively). This assumption goes along with our experimental results, which show that the proposed Seq2seq approach benefits from the inclusion of first order inputs, see Table II. MPJPE decreases from 24.27 to 24.19 mainly due to the improvement in the short-term period (decreases from 9.71 to 9.53), which may be a sign that the first order information helps the network to model the dynamics of the joints. Although these results may not be meaningful, we include the first order information in our experiments so that the network can choose whether such information is relevant for the task.

Human behavior is very complex and the amount of possible scenarios in a dyadic interaction is huge. As a result, the dimensionality of the observation and prediction spaces that we are trying to map is very high. In order to assess the importance of the training dataset size when modelling such a complex scenario, we compare a model trained with only the *Talk* task against a model trained with all the tasks (roughly four times more segments). The results show how the TCN performance does not scale very well with the dataset size, and even presents a slight decrease of performance (MPJPE from 25.22 to 25.45). This may suggest that the network capacity is limited and that bigger TCNs should be explored for exploiting the extra data, as incorporating new tasks also extends the space of task-specific behaviors. Visually, predictions made by the TCN are very jittery and tend to deform, which is expressed as a high divergence ($\Delta > 65$). This does not match previous works stating the robustness of TCNs against long-term deformations [17]. However, we hypothesize that this effect is caused by our predictions being offsets instead of coordinates, which may eventually force the TCN to act as a recurrent model and therefore complicate its training.

On the contrary, the Seq2seq network, although having roughly the same number of parameters, scales very well with the dataset size and outperforms the TCN in both scenarios (MPJPE of 24.78 against 25.22 with *Talk* only, and 24.19 against 25.45 with all tasks). Interestingly, the main benefit from training the Seq2seq with all tasks resides in the long-term MPJPE (33.03 to 32.19), compared to the small improvement for short-term predictions (9.69 to 9.53). Reasonably, the bigger size of the dataset might help the model to learn long-range behaviors. To state the difficulty of beating the zero-velocity baseline (MPJPE=24.56), note that our model beats it only when trained with all the dataset tasks (MPJPE=24.19).

B. Short/Long-term trade-off.

In this section, we compare our approach to another model (Seq2seq_ST) trained with a less ambitious goal: minimizing the MSE only for the first 10 frames (400ms, compared to the standard MSE considering a future of 2s), therefore ignoring the

	MPJPE	ST	MT	LT	FDE	Δ
Zero-velocity baseline						
All	24.56	9.61	23.95	32.75	35.33	0.00
Face	17.86	6.08	17.19	24.40	26.11	0.00
Body	10.50	3.85	10.12	14.21	15.42	0.00
Hands	37.75	15.88	37.79	49.37	53.01	0.00
Seq2seq_ST - Short-term						
All	25.04	**9.11**	23.74	34.42	38.12	59.08
Face	18.04	5.73	16.98	25.25	27.62	20.19
Body	10.92	3.75	10.17	15.24	17.10	11.06
Hands	38.63	15.05	37.62	52.36	57.99	52.83
Seq2seq_LT - Long-term						
All	**24.19**	9.53	**23.62**	**32.19**	**34.62**	38.62
Face	17.56	6.04	16.98	23.90	25.62	17.62
Body	10.60	3.85	10.20	14.36	15.61	8.00
Hands	37.02	15.72	37.08	48.32	51.56	32.56

Table III. Evaluation results of the proposed Seq2seq model trained by minimizing the short-term mean squared error (MSE for frames within 0-400ms) compared to minimizing the standard MSE (long-term, frames within $0 - 2s$). The results show how our short-term approach (Seq2seq_ST) considerably outperforms the zero-velocity baseline in the short-term MPJPE (ST) in return for a lower performance in the long-term (LT). The opposite effect is observed for the long-term approach, Seq2seq_LT, which also predicts more static behaviors, especially for hands (lower Δ than Seq2seq_ST).

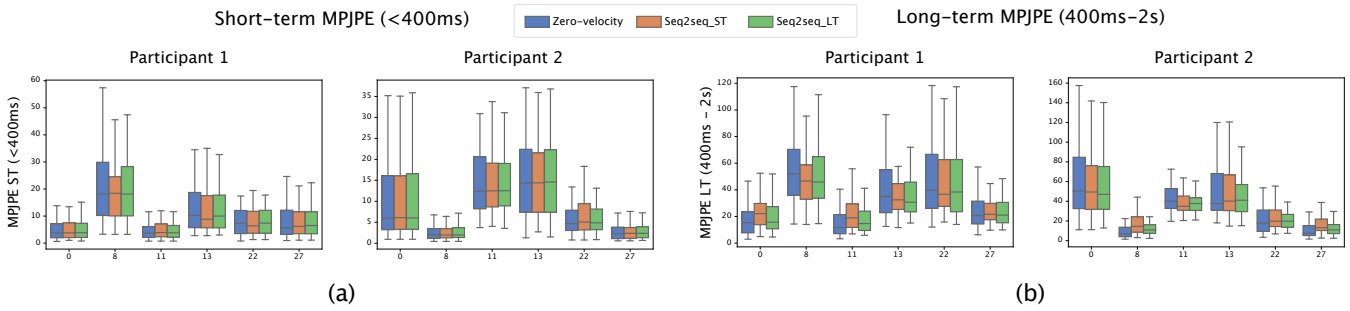

Fig. 10. In (a), boxplot of the short-term MPJPE for the whole skeleton. The zero-velocity baseline is compared to the Seq2seq models trained with short- and long-term losses (Seq2seq_ST and Seq2seq_LT, respectively). Results from the segments of the six most representative test clusters are shown for both session participants. In (b), same results for the long-term MPJPE. Although not being explicitly trained for long-term behavior modeling, the Seq2seq_ST beats the zero-velocity baseline with respect to the LT MPJPE of the full skeleton in segments with more movement (participant 1, clusters 8, 13 and 22). However, the Seq2seq_LT still offers the lowest LT MPJPE in those and outperforms the Seq2seq_ST in more static segments (participant 1, clusters 0, 11 and 27). However, in such static scenarios, the zero-velocity baseline still presents lowest MPJPE.

mid and long-term futures (400ms-2s). Results for the whole skeleton, and split by body part, are shown in Table III. Although the intuition would make us expect a decrease in the short-term MPJPE and a dramatic increse in the mid- and long-term MPJPEs compared to original model (Seq2seq_LT), only the former effect is observed (from 9.53 to 9.11). Interestingly, the MT MPJPE only raises from 23.62 to 23.74, which still outperforms the zero-velocity baseline (MT=23.95), while the LT MPJPE effectively increases by a larger margin (from 32.19 to 34.42). These results are aligned to prior work pointing out that acquiring the ability to predict longer sequences in the future harms the model performance in the short-term, and vice versa. In return though, predictions from the short-term approach tend to be more dynamic and less conservative (Δ is 52.9% higher than Seq2seq_LT). Combined with its considerably higher FDE value (34.62 and 38.12 for long- and short-term goals, respectively), it suggests that predictions from Seq2seq_ST tend to increasingly diverge from the ground truth as we move farther into the future. This effect is validated by visually inspecting the results. On the contrary, when the model aims at longer futures, outputs become more static and less adventurous. This effect is driven by the future stochasticity and resembles the blurring observed in future image prediction [11].

While some of the test segments contain sequences of joints with little to no movement, others present fast motions. It is difficult to assess and compare the performances of our models in such diverse dataset. Clusters used for the retrieval of test segments are used for the sake of a fairer comparison. In Figure 10, we show the distributions of MPJPE per segment and participant (each cluster may be conceptually different in a participant basis) in the six most populated clusters (including at least 80 segments each and a 49% of the whole test set in total). When looking at the ST MPJPE (Figure 10a), its average value for the Seq2seq_LT never beats the baseline, showing its poor short-term prediction capabilities. The Seq2seq_ST, instead, outperforms the zero-velocity baseline in two clusters (participant 1, clusters 13 and 22). On the other hand, the LT MPJPE (Figure 10b) results are clearer: the long-term Seq2seq beats the short-term one in almost all the clusters (9/12). The baseline

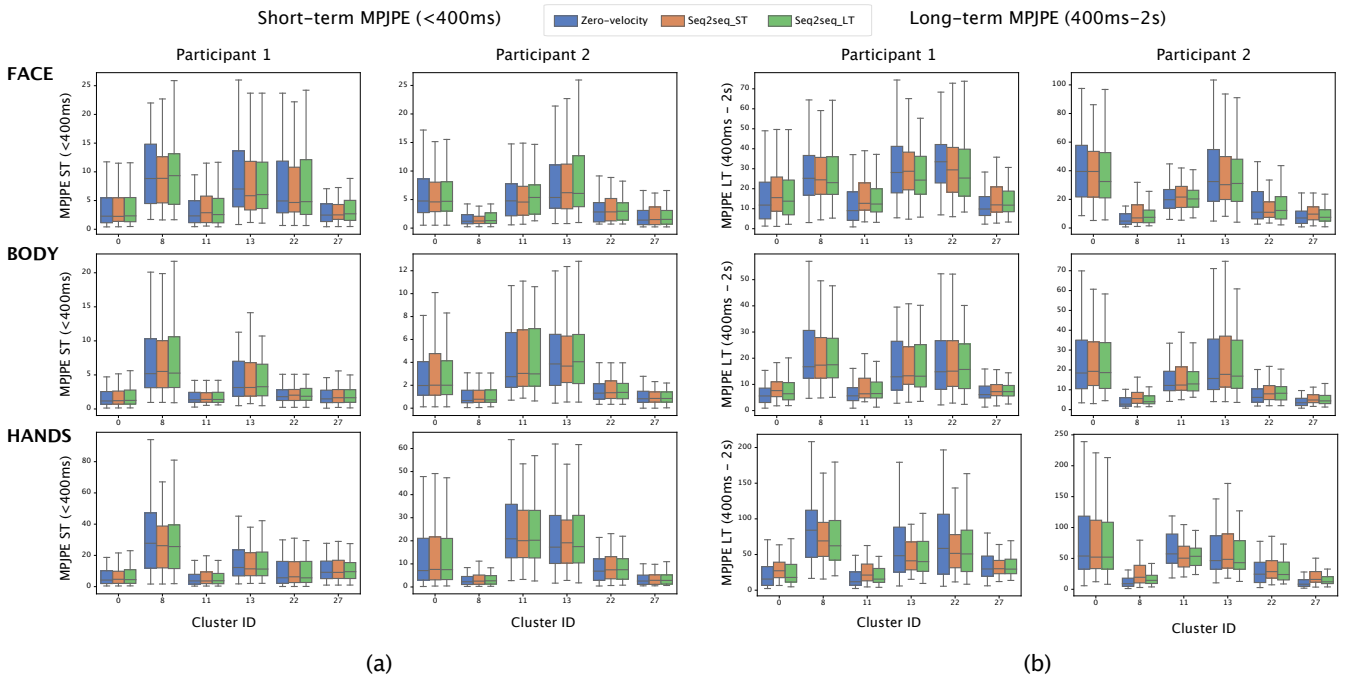

Fig. 11. Similarly to Figure 10, we plot now the MPJPE distributions per part of the body. This fine-grained analysis shows that the differences in MPJPE observed for some clusters in the long-term are originated in the face and hands predictions. For body, the performance from our approach do not differ much from that of the zero-velocity baseline.

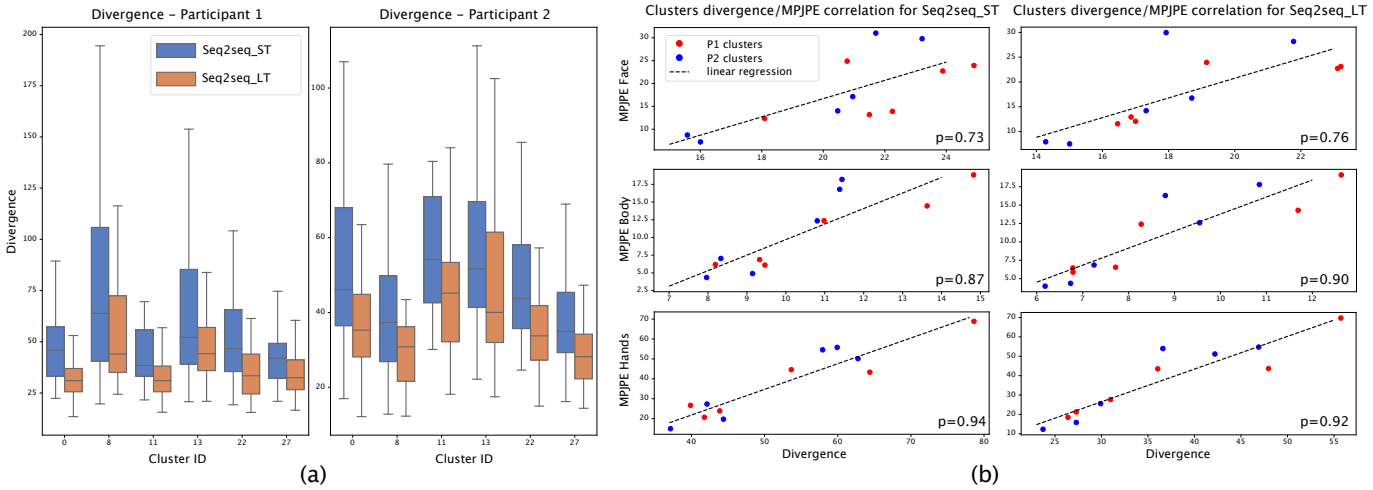

Fig. 12. Analysis of the divergence (quantity of movement) of the predictions from Seq2seq versions trained with short and long-term loss functions (0-400ms and 0-2s, respectively). In (a), divergence boxplots for both models split per participant into the 6 most popular clusters in the test set. The lower divergence of the LT model is consistent in all clusters. In (b), the relationship between divergence and MPJPE per cluster, model (ST on the left and LT on the right) and body part (top, middle and bottom for face, body and hands, respectively). We observe a high linear correlation between divergence and MPJPE, which hints that the more risky predictions are, the more they tend to miss the real future behavior, particularly for hands and body.

keeps its advantage over our models only in clusters with very static segments (low MPJPE for the zero-velocity baseline): clusters 0, 11 for participant 1 and clusters 8, 22 and 22 for participant 2. Figure 11 replicates the analysis for each part of the skeleton. It shows that most of the gains observed in some clusters from the whole picture come from the face and hands. We also discern very limited improvements for the body pose with respect to the zero-velocity baseline, which could be driven by its highly static and fairly noisy nature.

Similarly to our test set containing segments with differences in their motion, our models also predict sequences with diverse motion degrees. We have already seen that training with an eye to the long-term horizon (Seq2seq_LT) produces futures with less motion compared to training aiming at the near future (Seq2seq_ST). In Figure 12a, this effect is exhibited split by participant and cluster, showing its consistency across the six most populated test clusters. In order to determine the accuracy

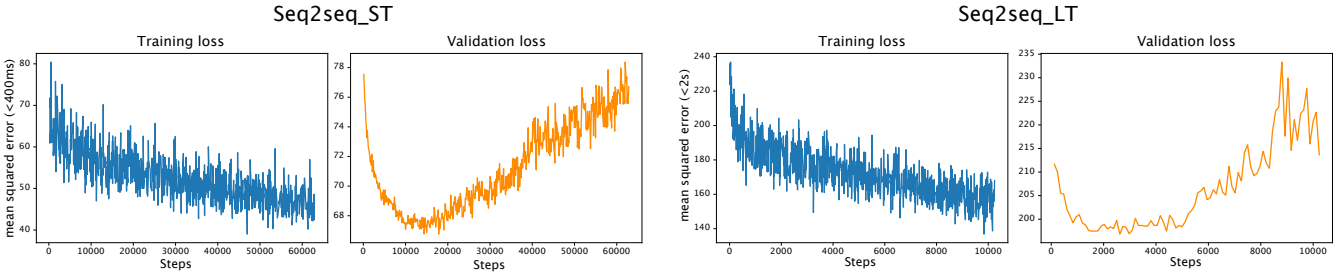

Fig. 13. Training and validation losses for the Seq2seq models trained forecasting a short-term future (Seq2seq_ST) and a long-term future (Seq2seq_LT). For visualization purposes, the early-stopping was deactivated. We observe how training for the long-term future leads to earlier divergence (5000 steps) compared to training for the short-term future, whose validation loss keeps decreasing for a longer time (15000 steps).

Input	Loss	MPJPE$_F$	ST$_F$	MT$_F$	LT$_F$	FDE$_F$
Seq2seq_ST ($<$ 400ms)						
Full	Full	18.04	5.73	16.98	25.25	**27.62**
Full	Face	18.16	5.32	16.40	26.33	30.46
Face	Face	**17.25**	**5.28**	**16.00**	**24.50**	28.03
Seq2seq_LT ($<$ 2s)						
Full	Full	17.56	6.04	16.98	23.90	25.62
Full	Face	16.79	5.71	**16.25**	22.88	24.33
Face	Face	**16.69**	**5.65**	16.28	**22.64**	**24.08**

Input	Goal	MPJPE$_H$	ST$_H$	MT$_H$	LT$_H$	FDE$_H$
Seq2seq_ST ($<$ 400ms)						
Full	Full	38.63	**15.05**	37.62	52.36	57.99
Full	Hand	37.55	15.17	37.18	49.91	54.25
Hand	Hand	**37.46**	15.20	**37.02**	**49.76**	53.59
Seq2seq_LT ($<$ 2s)						
Full	Full	37.02	**15.72**	37.08	48.32	51.56
Full	Hand	**36.91**	**15.72**	**37.06**	48.13	51.50
Hand	Hand	36.94	15.82	37.08	**48.10**	**51.46**

Table IV. On the left, the results of our model trained with the full skeleton compared to the model trained with only the face aiming at predicting the face behavior. On the right, same comparison for hands. For face, training with only face as input yields the best results for both short and long-term goals.

of our models when making risky and dynamic predictions (e.g., clusters 0, 11, 13 for participant 2) compared to static and conservative ones, we plot the pairs of average divergences and MPJPE for each body part (face, body and hands). We find that the MPJPE and the divergence for both ST and LT Seq2seq versions are highly linearly correlated for body ($\rho = 0.87$ and 0.90) and hands ($\rho = 0.94$ and 0.92), which means that risky predictions tend to be wrong. However, this linear correlation is not so strong for face ($\rho = 0.73$ and 0.76, respectively), see Figure 12b. In fact, the plot shows some highly divergent clusters with MPJPE values as low as other less divergent clusters. This represents an indicator that our model behaves remarkably better for face than for body or hands.

Although our best model beats the zero-velocity baseline, we still have a big margin of improvement and are far from predicting sequences which could be mistaken by real behaviors. This is also reinforced by the visualization of the training and validation loss curves, see Figure 13. Although the training loss keeps decreasing, proving the model as capable of modeling the behaviors from the training set, it quickly overfits. Thanks to the early-stopping, the final training and test MPJPE do not differ much for neither the short-term model (21.38 and 25.04) nor the long-term model (20.48 and 24.19).

C. Holistic contributions.

A question that quickly arises when predicting three different body parts is whether training a model to predict the behavior of each part would work better than training a generic model to predict all of them at once. To answer this question, we trained *expert models* for each part of the body. Each of them was trained twice with different inputs: with the past observation of 1) the whole body, and 2) only the part to be predicted. Results are summarized in Table IV. For face forecasting with Seq2seq_LT, the model clearly benefits from restricting the prediction to the face, as the MPJPE decreases from 17.56 (whole skeleton as both input and output) to 16.79 (whole skeleton as input) and 16.69 (only face as input). This improvement is especially boosted for Seq2seq_ST with face as both input and output, improving the MPJPE from the original score of 18.04 to 17.25. Although this phenomenon is also observed for Seq2seq_ST with hands (MPJPE from 38.63 to 37.55 and 37.46 when using the whole skeleton and only hands, respectively), it is not translated to the Seq2seq_LT model (small decrease from 37.02 to 36.91 and 36.94). Again, this could be a sign that the network learns to model the face better than it does for the other parts of the body. The rigidness of the head, the bounded dynamics of the face landmarks, and the limited facial expression captured could explain this superiority.

Interestingly, the higher dimensionality of the full body input made both the short and long-term Seq2seq models overfit faster. Still, they eventually reached a similar minimum in the validation loss curve. This suggests the need of bigger datasets for effectively exploiting the additional knowledge provided by the whole skeleton dynamics.

Architecture	MPJPE	ST	MT	LT	FDE	Δ
Seq2seq_ST: short-term goal ($< 400ms$)						
Unimodal Seq2seq	25.04	9.11	23.74	34.42	38.12	59.08
Seq2seq w/ dyadic	25.02	9.18	23.84	34.22	37.57	58.44
Seq2seq w/ metadata (v1)	24.99	9.14	23.76	34.25	37.60	60.03
Seq2seq w/ metadata (v2)	24.59	9.21	23.62	33.34	36.29	50.80
Seq2seq w/ audio (v1)	24.16	9.13	23.39	32.55	35.29	50.33
Seq2seq w/ audio (v2)	24.70	9.17	23.73	33.55	36.55	56.56
Seq2seq_LT: long-term goal ($< 2s$)						
Unimodal Seq2seq	24.19	9.53	23.62	32.19	34.62	38.62
Seq2seq w/ dyadic	24.13	9.54	23.64	32.01	34.30	37.25
Seq2seq w/ metadata (v1)	24.16	9.56	23.66	32.06	34.50	36.37
Seq2seq w/ metadata (v2)	24.02	9.53	23.54	31.84	34.14	36.45
Seq2seq w/ audio (v1)	24.23	9.54	23.69	32.20	34.62	41.12
Seq2seq w/ audio (v2)	24.15	9.54	23.64	32.05	34.28	38.77

Table V. Comparison of the proposed model with its multimodal versions including dyadic information, participants' metadata and audio. Although small gains are observed when including audio in the Seq2seq_ST model, multimodal approaches do not consistently improve the unimodal model in either the short or long-term goals.

D. Multimodality.

Apart from the landmarks annotations used until this point, the UDIVA dataset includes data from a wide range of sources: speech, transcriptions, metadata, image/video, etc. Merging this data so that the network can leverage it and improve the overall results is challenging. In this section, we make a first attempt to merge some modalities.

Dyadic. In order to include information about the interlocutor's behavior, we assume it to be another modality. To do so, we feed the encoded hidden and cell states from participant 1 into a linear layer (output of size 64) followed by a non-linearity (leaky relu). The resulting hidden and cell embedded vectors of size 64 are concatenated into the encoded hidden and cell states from participant 2 before being fed to the decoder. And vice versa. Therefore, the decoder is aware of the interlocutor's behavior during the observation window.

Metadata. We tested two merging methods. In the first (v1), we simply concatenate the metadata array of participant 1 (2) together with the session metadata array and their embedded pose before being fed to the encoder/decoder of participant 1 (2). The second approach (v2) concatenates the metadata array of participant 1 (2) and the session metadata to the hidden and cell states of the encoder of participant 1 (2) before sending them to the corresponding decoder.

Audio. The audio feature vector of the whole segment (of size 512) is embedded into a feature vector of size 64 by feeding it into a sequence of non-linear (as suggested by the authors in [76]), linear, and non-linear layers. The two versions v1 and v2 merge the audio feature vector in a similar way to the ones described above, replacing the metadata array by the audio feature vector.

Results for all the multimodal models tested are summarized in Table V. The main benefits from including extra information should come in the long-term scenario. However, results are, in overall, inconclusive. It looks like the audio may help a bit for the Seq2seq_ST, but such differences are not translated to the Seq2seq_LT model, where all the tested multimodal models perform equally. We hypothesize that the inclusion of dyadic, metadata and audio information greatly expands the dimensionality of the problem and that more data is needed in order to take advantage of them. Indeed, by plotting the training and validation losses from the Seq2seq_LT, see Figure 14, we observe that the multimodal metadata versions of the model overfit slightly faster than the unimodal model. However, better ways of fusing the multimodal information need to be explored and tested, as how to properly incorporate information from different time scales, semantics, etc, is a huge research topic.

VI. Case study - ICCV'21 DYAD Challenge

In parallel to this work, we co-organized the ICCV'21 DYAD Challenge, which aimed at advancing and motivating the research on visual human behavior analysis in dyadic and small group interactions. The challenge used the here described UDIVA v0.5 dataset. In this section, we present the metrics of the challenge and use them to compare our best models with respect to the baseline and the participants.

A. Evaluation metrics

The objectives of both the DYAD'21 behavior forecasting challenge and this work could be split into 3: accurately predicting face, body and hands trajectory and motion. A suitable metric for such goals should 1) encourage realistic future predictions and 2) avoid penalizing dramatically outlier predictions. The first constraint prevents us from using any distance metric (e.g., MPJPE), as any corrupt prediction would penalize the final score too heavily. Instead, upper bounded metrics could be considered. The second constraint is quite challenging though. While realism can be easily assessed from a qualitative point

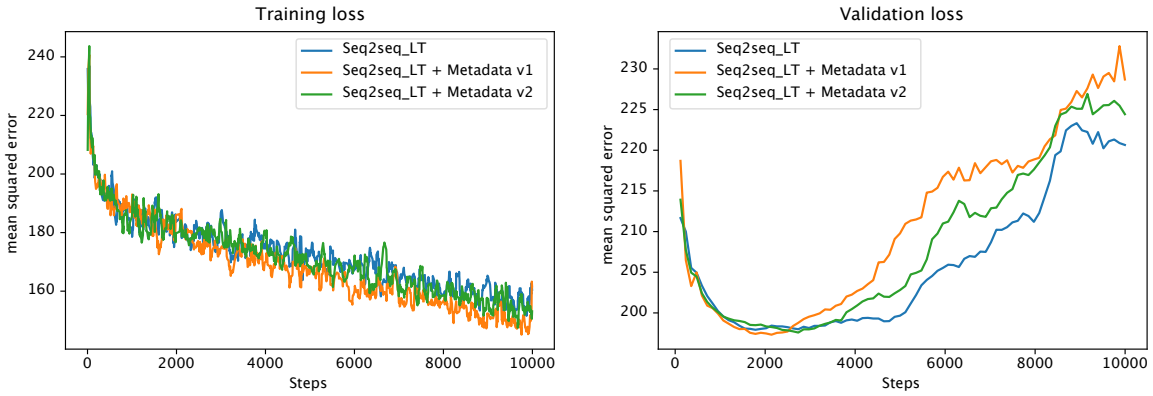

Fig. 14. Training and validation loss while training the Seq2seq_LT and their multimodal versions including metadata. Early stopping was deactivated in order to observe how quickly the validation loss diverges. We observe that by including the metadata makes the model overfits slightly faster..

of view, providing a quantitative score is not trivial. Instead, we propose a set of broadly adopted metrics in pose estimation which already implicitly imply such realism.

1) Face: the area under the curve (AUC) of the cumulative error distribution plot (CED). It is computed up to the 25% of the inter-pupil distance. Such distance is defined by averaging the distance between the landmarks of both pupils in all participant frames. Although the most common upper bound for the CED was 10%, we relax it due to the superior difficulty of our problem [84]. Only frames with *correct* face landmarks are considered (M_F). 100 bins are considered for the AUC calculation.

$$F := AUC^{M_F}_{CED(0:0.25)}$$

2) Body: the percentage of correct keypoints (PCK) up to 50% of the head size [85]. This value is computed for each body joint and averaged to get the body score. While high values for this metric do not implicitly imply body realism, we do not consider this a problem given the static nature of the body in dyadic conversational scenarios. Only frames with *correct* body pose are considered (M_B). In order to contemplate the higher levels of noise present in the body annotations, only 10 bins are considered for the PCK scores computation.

$$B := \frac{1}{P}\sum_{i=0}^{P} AUC^{M_B}_{PCK_i(0:0.5)}, \qquad P := \text{\#body-joints considered}$$

3) Hands: the AUC of the success rate (SR) [86] plot up to the 50% of the palm size. The palm size is defined as the distance between the knuckles of the index and the little fingers. The final hands score is the average of this value for both hands. Only frames with *correct* hand landmarks are considered (M_L and M_R). 100 bins are considered for the AUC calculation.

$$H = \frac{H^L + H^R}{2} = \frac{(AUC^{M_L}_{SR_L(0:0.5)} + AUC^{M_R}_{SR_R(0:0.5)})}{2}$$

In the challenge, participants were ranked according to each of these metrics. The participant with the lowest average rank was declared the winner.

B. Challenge results

In these section, we present the results of our best models with the metrics of the competition. The characteristics of the segments to predict are the same as the ones used in our experiments: an observation window of 100 frames (4 seconds), and a prediction window of 50 frames (2 seconds). The skeleton annotations from the observation window were cleaned as explained in Section III-B2.

Validation set. If we naively used the Seq2seq_LT model to predict the 50 future frames, we would score 0.268, 0.856, 0.340 for face, body and hands, respectively. These are far from the baseline values of 0.292, 0.859, 0.395. Similarly, if we used the Seq2seq_ST model instead, we would get worse results for body and hands (0.272, 0.850, 0.302), which matches the results from our experiments with MPJPE. The reason behind this is that the challenge metrics are very strict and greatly penalize wrong predictions. Unfortunately, given the stochasticity of the future, our model will unavoidably output wrong sequences of future poses from time to time. In order to compensate for those errors, the network would need to perfectly predict many segments. This is not feasible. Instead, we attempted a very conservative strategy. We knew that our models made better predictions for the short-term future. Therefore, we explored whether by using them to predict very few frames,

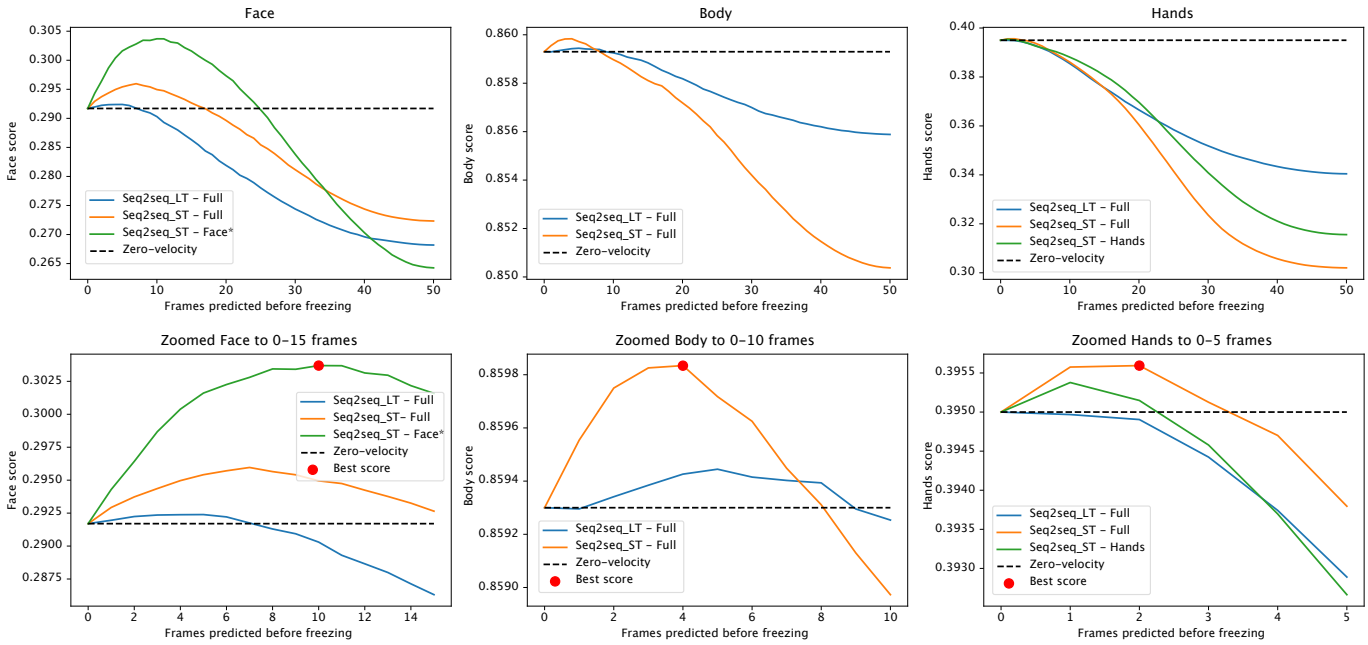

Fig. 15. Performance of the proposed models in the validation stage of the DYAD challenge split by face (left), body (middle) and hands (right). The x axis corresponds to the number of frames predicted before replicating the last predicted pose into the future (in a similar way the zero-velocity baseline is generated). For a better visualization, the bottom row plots zoom in the low values of x from the top row plots. The maximum is reached after predicting few frames, and quickly decreases as we dare to predict further due to the highly penalizing challenge metrics. * Model using only face as input.

	Validation			Test		
Architecture	Face	Body	Hands	Face	Body	Hands
Baseline	0.2917	0.8593	0.3950	0.3458	0.8897	0.5392
Team 1	0.2917	0.8593	0.3955	-	-	-
Team 2	0.1585	0.7991	0.1723	-	-	-
Seq2seq_ST - Face (10-0-0*)	**0.3037**	-	-	**0.3525**	-	-
Seq2seq_ST - Full (7-4-2*)	0.2960	**0.8598**	**0.3956**	0.3465	**0.8898**	**0.5399**
Seq2seq_LT - Full (5-5-0*)	0.2924	0.8594	0.3950	0.3457	0.8896	0.5392
Ensemble of models	0.3037	0.8598	0.3956	0.3525	0.8898	0.5399

Table VI. Final results of our proposed model for the validation and test stages of the DYAD challenge. The best models in the validation stage were used as an ensemble of models to combine the benefits from each of them. Scores for the baseline (propagating last observation to the future frames) and other participating teams are also included. *Number of frames predicted before freezing face, body and hands, respectively.

N, before applying zero-velocity (propagating the last predicted pose into the future), we improved the zero-velocity baseline (propagating the last observed pose into the future). Figure 15 plots the challenge metrics in function of the choice of N for several models. We observe that this strategy indeed improves the baseline for very few frames ($<$ 400ms for face and less than $<$ 200ms for body and hands) before quickly falling as more frames are predicted before the static propagation. By using this strategy along with the Seq2seq_ST, we scored (0.296, 0.858, 0.3956) with $N = 7$, 4 and 2 frames for face, body and hands, respectively, see Table VI. All of them improved the zero-velocity baseline. Additionally, we used the Seq2seq_ST model trained to predict only the face behavior (which improved the face MPJPE, as described in Section V-C) to increase the face score to 0.3037.

Test set. For the test stage, participants only had three attempts to submit their predictions. Therefore, only the best performing models performing at validation stage should be submitted to the test stage. In order to combine the benefits of our best models, we merged the Seq2seq_ST for full body and the Seq2seq_ST for face into an ensemble of models. From the former, we kept its body and hands predictions with $N = 4$ and 2, respectively, which were the best values from the validation stage. Similarly, from the latter, we used its face predictions with $N = 10$. As a result, we scored 0.3525, 0.8898, 0.5399 for face, body and hands, respectively, and outperformed the baseline for all the parts of the body (0.3458, 0.8897, 0.5392), see Table VI.

VII. Conclusions and Future Work

In this work, we introduced an extended version of the UDIVA dataset (v0.5) with automatically extracted face, body and hand annotations which were refined through a series of post-processing steps. These annotations were used to tackle the future

behavior forecasting in dyadic interactions. In an attempt to solve this problem, we proposed a recurrent Seq2seq architecture which learns to predict the future behavior more accurately than the zero-velocity baseline, even for long prediction windows (up to 2 seconds). Although the overall results were not outstanding, this superiority suggests that our model is, in fact, learning behavioral routines up to certain degree. This is especially true for face, whose predictions visually resemble realistic behaviors. We also made a first attempt to use multimodal information to improve our predictions, although results showed no improvements over the monomodal model. Finally, to assess the potential of our model, we tested it in a behavior forecasting competition. Our model beat the zero-velocity baseline and the two teams that were participating until the day of submission of this thesis.

As in many other machine learning applications, the biggest limitation of this work resides in the data. On one side, the complexity associated to modelling human behavior is huge. The number of variables and possible scenarios during a human-to-human interaction is immense. In order to train a model which is able to model all these scenarios, we need to generate a training dataset which illustrates such variability. On the other side, our dataset annotations were automatically extracted and, as a result, are very noisy. This could help to explain the inferior capacity of our model to forecast the hands and body behaviors compared to that of the face, for which annotations were significantly more accurate.

Another limitation lies in the test segments whose last observation does not include the hand landmarks due to either the hand being hidden under the table, or the algorithm not succeeding to retrieve the landmarks. Although much less often, this scenario also happens for the body. As a result, our algorithm predicts meaningless offsets (the future missing body part of the skeleton can not be reconstructed). This is a conceptual problem which is not currently being addressed because those segments are filtered out from the training set. However, it could be mitigated in the future by adding a module which first inferred an approximation of the full skeleton from which the offsets would be subsequently predicted. Similarly to this idea, we could also take advantage of the temporal dimension of our skeleton annotations to train a model which learnt to recover the missing landmarks in a dropout autoencoder fashion. This could be either previously trained and applied as a pre-processing step, or included as an extra module and trained end-to-end.

Future work also includes replicating the automatic skeleton extraction with other similar datasets. Even if such datasets are not filmed during dyadic interactions, the human behavior core is cross-sectional and shared among different contexts. Those generic extra datasets could be used to pre-train our network weights before fine-tuning it with the UDIVA dataset. From the technical point of view, other state-of-the-art architectures like transformers, graph neural networks or generative adversarial networks could also be tested in the context of the UDIVA dataset. In parallel, statistical tests should be incorporated in order to assess the significance of the differences among the performances of the models. Finally, qualitative metrics could be also explored in order to evaluate the realism or the smoothness of the predicted skeleton sequences.

ACKNOWLEDGMENT

First of all, I would like to express the deepest appreciation to Sergio Escalera, who blindly trusted me since the beginning of this thesis and always motivated me along it. I am also deeply grateful to Cristina Palmero, who provided me with priceless knowledge and guided me through this project. Special thanks to Xavier Baró for co-supervising me and their useful insights. I am grateful to the whole HUPBA team for embracing me into their research group.

I would also like to thank the people who helped me along the path which got me here: Petia Radeva, for introducing me to the computer vision, Meritxell Bach, for her unforgettable support and Carles Fernandez and Isabelle Hupont, for their priceless guidance and advice.

Finally, my sincere gratitude to all my family and friends, who always encouraged me to go ahead and helped me whenever I needed it most.

APPENDIX A
HANDS POST-PROCESSING

The appearance of the interlocutor's hands is very frequent in the views used for this work (*FC1* and *FC2*), as can be observed in Figure 5. In such scenarios, the extraction methods may generate landmarks for more than one person. We easily retrieve the face and body of interest by selecting the most centered set of landmarks. Unfortunately, the highly sparse set of locations where the hands of interest appear overlaps with those from the confounding hands (the interlocutor's). This sets off a high number of frames where the hand of interest can be easily mistakenly chosen. The same applies to the right-left association to the hands, which might be wrong if the arms of the person are crossed. By default, the hands estimator module [66] associates the left/right label if the detection is the closest to the left/right body elbow (which is inferred in parallel). This assumption becomes especially erroneous for the aforementioned crossed-arms scenario.

In order to minimize these errors, we sequentially applied several post-processing methods:

1) We extracted hands bounding boxes and their left/right default association.
2) We removed temporally consecutive detections with an intersection-over-union (IoU) value smaller than 0.1 (hand switches/jumps).

3) We looked for left/right hands temporal gaps generated by missed or removed hands. These gaps were tracked forwards and backwards using [65]. If the bidirectional tracker eventually overlapped with the last and first detections before and after the gap (i.e. tracked detection has an IoU with the original detection bigger than 0.25), both sequences of tracked detections from the gap are merged and saved. If only one tracker overlapped with the other side of the gap, the merging was skipped and only the detections from the successful tracker were stored. This step successfully recovered all the correct detections removed in the previous step and fixed sequences of rapidly moving hands.
4) We extracted the hands landmarks inside the bounding boxes from the previous step.
5) We computed the vector from the left wrist to the left palm of the pose from the body of interest, V_{LB}. We also computed the vector from the left/right wrist of the hand to the left/right middle knuckle, V_{LH}. Similarly, we computed V_{RB} and V_{RH} for the right hand. If $V_{LB} \cdot V_{LH} < 0$ and $V_{RB} \cdot V_{RH} < 0$, i.e. angles between the vectors of the body and both hands differed more than 90º, the left/right associations were switched.
6) We removed the hands detections that fulfilled any of the following conditions:
   - Left/right hand wrist was closer to the left/right pose wrist of the confounder body (i.e. interlocutor's body) by a margin of 20 pixels.
   - The left hand wrist was closer to the right pose wrists by a margin of 60 pixels with respect to the left pose wrist. Same for the right hand wrist.
7) We repeated the steps 3 and 4, filling the gaps generated by the previous step, and extracting the new and final hand landmarks.
8) We applied the one-euro-filter to the hand landmarks with a cut-off of 0.001 and a $\beta = 0.02$ [87].

Some of the previous parameters were selected after statistically analyzing the properties of the resulting landmarks (e.g. distances between hands and wrists, angles, etc.) and others by visually inspecting the erroneous results (e.g. intersection-over-union thresholds).

REFERENCES

[1] Judee K Burgoon, Lesa A Stern, and Leesa Dillman. *Interpersonal adaptation: Dyadic interaction patterns*. Cambridge University Press, 2007.
[2] Josephine Barbaro and Cheryl Dissanayake. Early markers of autism spectrum disorders in infants and toddlers prospectively identified in the social attention and communication study. *Autism*, 17(1):64–86, 2013.
[3] E Loth, L Garrido, J Ahmad, E Watson, A Duff, and B Duchaine. Facial expression recognition as a candidate marker for autism spectrum disorder: how frequent and severe are deficits? *Molecular autism*, 9(1):1–11, 2018.
[4] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Transactions on Graphics (TOG)*, 37(4):1–13, 2018.
[5] Maryam Moosaei, Sumit K Das, Dan O Popa, and Laurel D Riek. Using facially expressive robots to calibrate clinical pain perception. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, pages 32–41. IEEE, 2017.
[6] Martin Saerbeck, Tom Schut, Christoph Bartneck, and Maddy D Janse. Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1613–1622, 2010.
[7] Kacper Kania, Marek Kowalski, and Tomasz Trzciński. Trajevae–controllable human motion generation from trajectories. *arXiv preprint arXiv:2104.00351*, 2021.
[8] Christian Diller, Thomas Funkhouser, and Angela Dai. Forecasting characteristic 3d poses of human actions. *arXiv preprint arXiv:2011.15079*, 2020.
[9] Cristina Palmero, Javier Selva, Sorina Smeureanu, Julio Junior, CS Jacques, Albert Clapés, Alexa Moseguí, Zejian Zhang, David Gallardo, Georgina Guilera, et al. Context-aware personality inference in dyadic scenarios: Introducing the udiva dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1–12, 2020.
[10] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.
[11] Sergiu Oprea, Pablo Martinez-Gonzalez, Alberto Garcia-Garcia, John Alejandro Castro-Vargas, Sergio Orts-Escolano, Jose Garcia-Rodriguez, and Antonis Argyros. A review on deep learning techniques for video prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
[12] Amir Nadeem, Ahmad Jalal, and Kibum Kim. Human actions tracking and recognition based on body parts detection via artificial neural network. In *2020 3rd International Conference on Advancements in Computational Sciences (ICACS)*, pages 1–6. IEEE, 2020.
[13] Ue-Hwan Kim, Dongho Ka, Hwasoo Yeo, and Jong-Hwan Kim. A real-time vision framework for pedestrian behavior recognition and intention prediction at intersections using 3d pose estimation. *arXiv preprint arXiv:2009.10868*, 2020.
[14] Kathan Vyas, Rui Ma, Behnaz Rezaei, Shuangjun Liu, Michael Neubauer, Thomas Ploetz, Ronald Oberleitner, and Sarah Ostadabbas. Recognition of atypical behavior in autism diagnosis from video using pose estimation over time. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2019.
[15] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3372–3382, 2021.
[16] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision*, pages 346–364. Springer, 2020.
[17] Naoya Fushishita, Antonio Tejero-de-Pablos, Yusuke Mukuta, and Tatsuya Harada. Long-term video generation of multiple futures using human poses. *CoRR*, abs/1904.07538, 2019.
[18] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1418–1427, 2018.
[19] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4346–4354, 2015.
[20] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2891–2900, 2017.
[21] Junyoung Chung, Sungjin Ahn, and Yoshua Bengio. Hierarchical multiscale recurrent neural networks. In *5th International Conference on Learning Representations, ICLR 2017*, 2019.
[22] Hsu-kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose forecasting. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1423–1432. IEEE, 2019.

[23] Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges. Learning human motion models for long-term predictions. In *2017 International Conference on 3D Vision (3DV)*, pages 458–466. IEEE, 2017.

[24] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *European Conference on Computer Vision*, pages 474–489. Springer, 2020.

[25] Yujun Cai, Lin Huang, Yiwei Wang, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Xu Yang, Yiheng Zhu, Xiaohui Shen, et al. Learning progressive joint propagation for human motion prediction. In *European Conference on Computer Vision*, pages 226–242. Springer, 2020.

[26] Borui Wang, Ehsan Adeli, Hsu-kuang Chiu, De-An Huang, and Juan Carlos Niebles. Imitation learning for human pose prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7124–7133, 2019.

[27] Enric Corona, Albert Pumarola, Guillem Alenya, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6992–7001, 2020.

[28] Vida Adeli, Mahsa Ehsanpour, Ian Reid, Juan Carlos Niebles, Silvio Savarese, Ehsan Adeli, and Hamid Rezatofighi. Tripod: Human trajectory and pose dynamics forecasting in the wild. *arXiv preprint arXiv:2104.04029*, 2021.

[29] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9489–9497, 2019.

[30] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 214–223, 2020.

[31] Ming Liang, Bin Yang, Rui Hu, Yun Chen, Renjie Liao, Song Feng, and Raquel Urtasun. Learning lane graph representations for motion forecasting. In *European Conference on Computer Vision*, pages 541–556. Springer, 2020.

[32] Qiongjie Cui and Huaijiang Sun. Towards accurate 3d human motion prediction from incomplete observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4801–4810, 2021.

[33] Chaitanya Ahuja, Shugao Ma, Louis-Philippe Morency, and Yaser Sheikh. To react or not to react: End-to-end visual pose forecasting for personalized avatar during dyadic conversations. In *2019 International Conference on Multimodal Interaction*, pages 74–84, 2019.

[34] Patrik Jonell, Taras Kucherenko, Gustav Eje Henter, and Jonas Beskow. Let's face it: Probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2020.

[35] Yingying Wang, Jean E Fox Tree, Marilyn Walker, and Michael Neff. Assessing the impact of hand motion on virtual character personality. *ACM Transactions on Applied Perception (TAP)*, 13(2):1–23, 2016.

[36] Ryo Ishii, Chaitanya Ahuja, Yukiko I Nakano, and Louis-Philippe Morency. Impact of personality on nonverbal behavior generation. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2020.

[37] Chaitanya Ahuja, Dong Won Lee, Yukiko I Nakano, and Louis-Philippe Morency. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In *European Conference on Computer Vision*, pages 248–265. Springer, 2020.

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

[39] Christian Zimmermann, Tim Welschehold, Christian Dornhege, Wolfram Burgard, and Thomas Brox. 3d human pose estimation in rgbd images for robotic task learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1986–1992. IEEE, 2018.

[40] Manolis Vasileiadis, Sotiris Malassiotis, Dimitrios Giakoumis, Christos-Savvas Bouganis, and Dimitrios Tzovaras. Robust human pose tracking for realistic service robot applications. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1363–1372, 2017.

[41] Min-Yu Wu, Pai-Wen Ting, Ya-Hui Tang, En-Te Chou, and Li-Chen Fu. Hand pose estimation in object-interaction based on deep learning for virtual reality applications. *Journal of Visual Communication and Image Representation*, 70:102802, 2020.

[42] Renshu Gu, Gaoang Wang, and Jenq-Neng Hwang. Efficient multi-person hierarchical 3d pose estimation for autonomous driving. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 163–168. IEEE, 2019.

[43] Ying Li, Chenxi Wang, Yu Cao, Benyuan Liu, Joanna Tan, and Yan Luo. Human pose estimation based in-home lower body rehabilitation system. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[44] Felix Achilles, Alexandru-Eugen Ichim, Huseyin Coskun, Federico Tombari, Soheyl Noachtar, and Nassir Navab. Patient mocap: Human pose estimation under blanket occlusion for hospital monitoring applications. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 491–499. Springer, 2016.

[45] Feng Liu, Dan Zeng, Qijun Zhao, and Xiaoming Liu. Joint face alignment and 3d face reconstruction. In *European Conference on Computer Vision*, pages 545–560. Springer, 2016.

[46] Hao Liu, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Two-stream transformer networks for video-based face alignment. *IEEE transactions on pattern analysis and machine intelligence*, 40(11):2546–2554, 2017.

[47] Yaojie Liu, Amin Jourabloo, William Ren, and Xiaoming Liu. Dense face alignment. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1619–1628, 2017.

[48] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 534–551, 2018.

[49] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *Proceedings of the IEEE international conference on computer vision*, pages 1031–1039, 2017.

[50] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021.

[51] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pages 152–168. Springer, 2020.

[52] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019.

[53] István Sárándi, Timm Linder, Kai O Arras, and Bastian Leibe. Metrabs: Metric-scale truncation-robust heatmaps for absolute 3d human pose estimation. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2020.

[54] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision*, pages 20–40. Springer, 2020.

[55] Theocharis Chatzis, Andreas Stergioulas, Dimitrios Konstantinidis, Kosmas Dimitropoulos, and Petros Daras. A comprehensive study on deep learning-based 3d hand pose estimation methods. *Applied Sciences*, 10(19):6850, 2020.

[56] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *IEEE International Conference on Computer Vision Workshops*, 2021.

[57] Xingyu Chen, Yufeng Liu, Chongyang Ma, Jianlong Chang, Huayan Wang, Tian Chen, Xiaoyan Guo, Pengfei Wan, and Wen Zheng. Camera-space hand mesh recovery via semantic aggregation and adaptive 2d-1d registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13274–13283, 2021.

[58] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 118–134, 2018.

[59] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4511–4520, 2019.

[60] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*, pages 4903–4911, 2017.

[61] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (ToG)*, 36(6):1–17, 2017.

[62] Jianzhu G. Towards fast, accurate and stable 3d dense face alignment. https://github.com/cleardusk/3DDFA_V2, 2021.

[63] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. Faceboxes: A cpu real-time face detector with high accuracy. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE, 2017.

[64] I Sárándi. Metrabs absolute 3d human pose estimator. https://github.com/isarandi/metrabs, 2021.

[65] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4282–4291, 2019.

[66] Rong Y. Frankmocap: A strong and easy-to-use single view 3d hand+body pose estimator. https://github.com/facebookresearch/frankmocap, 2021.

[67] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9869–9878, 2020.

[68] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *European Conference on Computer Vision*, pages 365–381. Springer, 2020.

[69] Xucong Zhang. Eth-xgaze baseline. https://github.com/xucong-zhang/ETH-XGaze, 2021.

[70] Yongyi Tang, Lin Ma, Wei Liu, and Wei-Shi Zheng. Long-term human motion prediction by modeling motion context and enhancing motion dynamics. In *IJCAI*, pages 935–941, 2018.

[71] John A Hartigan. *Clustering algorithms*. John Wiley & Sons, Inc., 1975.

[72] 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.

[73] Xiaoli Liu, Jianqin Yin, Huaping Liu, and Jun Liu. Deepssm: Deep state-space model for 3d human motion prediction. *arXiv preprint arXiv:2005.12155*, 2020.

[74] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[75] Robert R McCrae and Oliver P John. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215, 1992.

[76] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

[77] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

[78] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017.

[79] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *Universal Language Model Fine-tuning for Text Classification*, 2018.

[80] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

[81] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7035–7043, 2017.

[82] Sijin Li, Weichen Zhang, and Antoni B Chan. Maximum-margin structured learning with deep networks for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2848–2856, 2015.

[83] Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14424–14432, 2020.

[84] Michel Valstar, Brais Martinez, Xavier Binefa, and Maja Pantic. Facial point detection using boosted regression and graph models. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2729–2736. IEEE, 2010.

[85] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[86] Shanxin Yuan, Guillermo Garcia-Hernando, Björn Stenger, Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee, Pavlo Molchanov, Jan Kautz, Sina Honari, Liuhao Ge, et al. Depth-based 3d hand pose estimation: From current achievements to future goals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2636–2645, 2018.

[87] Géry Casiez, Nicolas Roussel, and Daniel Vogel. 1€ filter: a simple speed-based low-pass filter for noisy input in interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2527–2530, 2012.