Context-Aware Human Behaviour Forecasting in Dyadic Interactions

Nguyen Tan Viet Tuyen tan_viet_tuyen.nguyen@kcl.ac.uk
Oya Celiktutan oya.celiktutan@kcl.ac.uk
Centre for Robotics Research
Department of Engineering
King's College London
London WC2R 2LS, United Kingdom

Abstract

Non-verbal behaviours play an indispensable role in social interaction. People tend to use a wide range of non-verbal channels, including eye gaze, body, and facial gestures, to communicate their intentions and emotions to their interacting partners. Such social signals encourage verbal messages of the communicator can be transmitted to other interlocutors in a facile and transparent manner. On the other hand, an essential aspect of communication behaviours is the dynamic exchange of non-verbal signals among interlocutors for adapting current social norms and building a common ground. This factor suggests that data observed from the interacting partners should be considered when modeling the target individual's behaviours. Our paper introduces a generative framework with context awareness that captures the influence of the interacting partner's non-verbal signals on the target individual. The model consists of three components, namely, *Context Encoder*, *Generator*, and *Discriminator*. *Context Encoder* is employed to extract social signals observed from the interacting partner while *Generator* and *Discriminator* are utilized to generate and optimize the target person's gestures. We verify the efficiency of the framework on two different dyadic interaction datasets. The experimental results demonstrate that compared to baselines, our solution can produce human-like gestures better supporting interaction contexts. Undoubtedly, in dyadic interaction, the influence of the interacting partner's social signals on the target individual is observable, and the proposed approach can efficiently capture those effects. The source code of our framework can be found at https://github.com/sairlab/Context-Aware-Human-Behavior-Forecasting.

Keywords: dyadic social interaction, non-verbal behaviour generation, motion forecasting, generative adversarial networks.

1. Introduction

People tend to use a wide range of non-verbal channels, including eye gaze, body, and facial gestures, to communicate their intentions and emotions to their interacting partners. These modalities help to transmit verbal messages to other interlocutors in a facile and transparent manner in social interaction (Knapp et al., 2013). Motivated by the importance of human non-verbal behaviours, considerable attention has been paid to non-verbal behaviour generation tasks for virtual agents (Feng et al., 2017) and social robots (Ahn et al., 2018; Tuyen et al., 2020). Similarly to human interaction, communicative gestures endow virtual agents and robots with abilities to emphasize their speech and express their intentions or emotions. Consequently, non-verbal cues could help to improve the user's perception of robots' behaviours and positively contribute to interaction outcomes (Saunderson and Nejat, 2019).

On the other hand, an essential aspect of communicative behaviours is the dynamic exchange of such non-verbal signals among interlocutors for adapting to interacting social norms (Lakin et al., 2003) and building a common ground (Noy et al., 2011). This factor suggests that data observed from their interacting partners should be considered when modeling the target individual's behaviours. It would ensure output gestures can convey the communicator's intentions, at the same time, conform to the interaction context. The research presented in this paper sheds light on motion forecasting with context awareness. This problem is addressed by a generative framework that captures the influence of the interacting partner's non-verbal signals on the target individual during dyadic interaction.

The main contributions of this work can be summarized as follows:

(i) We suggest a new research question in motion forecasting domain: context-aware human gesture forecasting in dyadic interaction. This topic has not been intensively investigated in previous works in spite of its potential applications in many different areas.

(ii) We introduce a new generative framework that consists of multiple generators to handle the variations of joint distribution across different parts of human body.

2. Related Works

2.1. Non-verbal Behaviour Generation

The problem of non-verbal behaviour generation can be broadly categorized into two groups, namely, motion synthesis and motion forecasting.

Motion Synthesis. In recent years, there has been a growing interest in this research topic, where the connections between non-verbal signals of the communicator (or a robot) and their synthesized data (e.g., speech, emotion, etc.) are determined via a rule-based (Cassell et al., 2004) or data-driven approach (Ahn et al., 2018; Tuyen et al., 2020; Kucherenko et al., 2019; Wu et al., 2021). In particular, co-speech gestures are naturally performed when speaking and they are applied to convey the communicator's emotion, intention, or verbal contents of their speech. The semantic contents of the communicator's speech could be implemented to develop body gestures. This problem is addressed by a Generative Adversarial Network (GAN) based on Sequence to Sequence (Seq2Seq) model (Ahn et al., 2018) or a Conditional Generative Adversarial Network (cGAN) designed with Convolution Neural Network (CNN) operation (Tuyen et al., 2020). In other studies (Kucherenko et al., 2019; Wu et al., 2021), non-verbal gestures are estimated via the communicator's speech features using an auto encoder-decoder (Kucherenko et al., 2019) or a GAN network (Wu et al., 2021).

Motion Forecasting. This task aims to comprehend the non-verbal signals of the individual and generates future motion sequences. Motion forecasting is commonly addressed by a Seq2Seq framework. In Martinez et al. (2017), human motion inputs are encoded into internal representations and forwarded to a decoder to produce a maximum likelihood estimate for prediction. Residual connections are added to the encoder-decoder network for better modeling the velocity of motions. A similar approach can be found in Gui et al. (2018), where discriminators are additionally equipped for the designed framework. It is suggested that the accuracy of predicted motions can be further enhanced by adding the adversarial loss to the training phase.

2.2. Non-verbal Behaviour Generation with Context Awareness

The subsection reviews previous studies in the non-verbal generation task where the data observed from the interacting partners are considered to model gestures of the communicator (called the target person).

Motion Synthesis. Huang and Khan (2017) focused on the problem of facial expressions produced during interactions between an interviewee and an interviewer, and they introduced a framework based on cGAN. The method generated the interviewer's facial gestures that were appropriately contextualized and responsive to the interviewee's facial expressions. Similarly, the authors in Feng et al. (2017) suggested a Variational Auto Encoder-Decoder network to handle the generation of facial cues between a user and an embodied agent. In terms of triadic human communication, the authors (Joo et al., 2019) presented a generative approach that acquires non-verbal signals from interacting partners and encodes them into latent vectors. The encoded features were utilized to estimate the target person's body gestures.

Motion Forecasting. This line of research deals with forecasting short-term future motions of the target person by gathering non-verbal signals exchanged among interlocutors. This problem has been illustrated in Gupta et al. (2018) through a social navigation scenario where two pedestrians need to avoid each other when deciding their future motion paths. The problem is tackled by a GAN framework that observes the motion history of all pedestrians. It should be emphasized that not only in scenarios of social navigation, the dynamic exchange of non-verbal behaviours is unavoidable in scenarios of human social communication (Lakin et al., 2003; Noy et al., 2011). Recently, the authors (Raman et al., 2021) introduced a socially aware sequence-to-sequence model to forecast nonverbal cues (e.g., speaking status, and head pose) of two or more than two people involved in a group conversation. However, to the best of our knowledge, this research topic has not been intensively investigated, especially for producing high-dimensional human-like gestures in dyadic interactions that could be particularly useful for virtual agents and social robots.

3. Methodology

In this section, we first formulate the problem of non-verbal behaviour forecasting in dyadic interaction. Then, we present an overview of the approach, which is followed by details of the individual model components, including *Context Encoder* E, *Generator* G, and *Discriminator* D. We finalise with describing the designed loss functions and training process.

3.1. Problem Formulation

Considering the interaction between a target person S_{fo} and their interacting partner S_{ob}, let $P^{0:k}_{fo}$ denotes the motion data of the target person S_{fo}, within a temporal window,

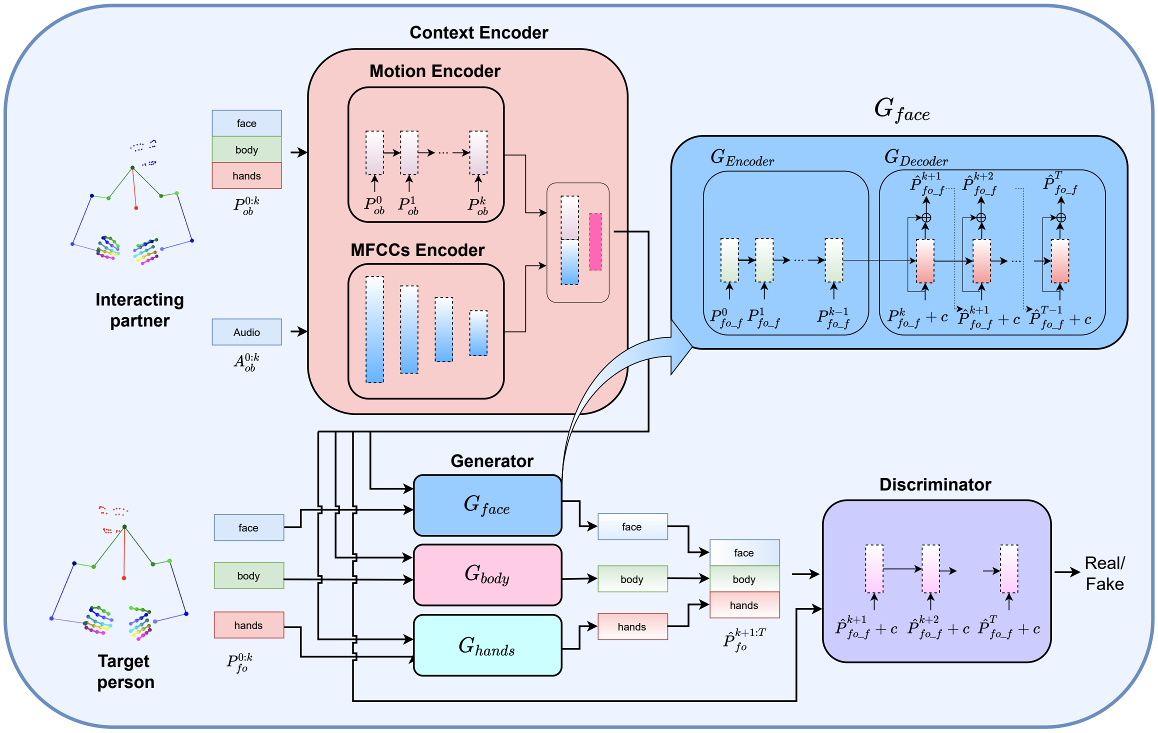

Context Encoder
Motion Encoder
MFCCs Encoder
face
body
hands
$P^{0:k}_{ob}$
Audio
$A^{0:k}_{ob}$
Interacting partner
P^0_{ob} P^1_{ob} P^k_{ob}
G_{face}
$G_{Encoder}$
$G_{Decoder}$
$P^0_{fo\_f}$ $P^1_{fo\_f}$ $P^{k-1}_{fo\_f}$
$\hat{P}^{k+1}_{fo\_f}$ $\hat{P}^{k+2}_{fo\_f}$ $\hat{P}^{T}_{fo\_f}$
$P^k_{fo\_f}+c$ $\hat{P}^{k+1}_{fo\_f}+c$ $\hat{P}^{T-1}_{fo\_f}+c$
Generator
G_{face}
G_{body}
G_{hands}
face
body
hands
$\hat{P}^{k+1:T}_{fo}$
Discriminator
$\hat{P}^{k+1}_{fo\_f}+c$ $\hat{P}^{k+2}_{fo\_f}+c$ $\hat{P}^{T}_{fo\_f}+c$
Real/ Fake
Target person
$P^{0:k}_{fo}$

Figure 1: The proposed framework to forecast the upper body motion $\hat{P}^{k+1:T}_{fo}$ of the target person S_{fo} in dyadic interaction. The model takes into account their current data $P^{0:k}_{fo}$, and the signal $P^{0:k}_{ob}$, $A^{0:k}_{ob}$ collected from the interacting partner S_{ob}.

namely, $(t \in [0,k])$. We can also observe audio $A^{0:k}_{ob}$ and motion features $P^{0:k}_{ob}$ of the interacting partner S_{ob} simultaneously. This paper aims to forecast a possible short-term future motion (a response non-verbal behaviour) $\hat{P}^{k+1:T}_{fo}$ $(t \in [k+1,T])$ of the target person S_{fo}. Hence, the goal of non-verbal behaviour forecasting is to find a mapping function F that receives $P^{0:k}_{fo}$, $P^{0:k}_{ob}$, $A^{0:k}_{ob}$ as inputs, and predicts the output $\hat{P}^{k+1:T}_{fo}$.

3.2. Background: Generative Adversarial Networks

GAN was originally introduced by Goodfellow et al. (2014). The model is designed with a *Generator* G and a *Discriminator* D. GAN operates based on a min-max game of two players. G attempts to capture the real data distribution and create fake data that can fool D. In contrast, D tries to differentiate between real and fake data produced by G. cGAN (Mirza and Osindero, 2014) is an extension of GAN that receives c as a conditional input to regulate generated data. The min-max game between G and D in cGAN can be formulated as follows:

$$\min_{G}\max_{D} V(D,G) = \mathbb{E}_{x,c\sim p_{data}}\left[logD(c,x)\right] + \mathbb{E}_{c\sim p_{data},z\sim p_z}\left[log(1-D(G(c,z),c))\right], \qquad (1)$$

where z is a random noise vector sampled from a prior distribution function.

This adversarial training provides GAN and its extensions capabilities of producing realistic high dimensional outputs, which are difficult to determine via manually designed features (Reed et al., 2016). As a result, GAN has been widely used in many application domains, in particular, non-verbal behaviour generation task (Gui et al., 2018; Ahn et al., 2018; Tuyen et al., 2020; Wu et al., 2021).

3.3. Overview of the Proposed Approach

Fig. 1 illustrates the proposed training framework to forecast $\hat{P}_{fo}^{k+1:T}$ $(t \in [k+1, T])$. The process starts with encoding face, body, and hand landmarks $P_{ob}^{0:k}$ $(t \in [0, k])$ of the interacting partner into c_P, and their corresponding audio features $A_{ob}^{0:k}$ into c_A. c_P is then combined with c_A, and injected into both *Generator* G and *Discriminator* D as the contextual input c. In terms of $P_{fo}^{0:k}$, it is divided into three motion parts, namely face $P_{fo\_f}^{0:k}$, body $P_{fo\_b}^{0:k}$, and hands $P_{fo\_h}^{0:k}$, and fed to G_{face}, G_{body}, and G_{hands}, respectively. Here, *Generator* receives the contextual input c and a corresponding data $P_{fo}^{0:k}$ to forecast a possible motion $\hat{P}_{fo}^{k+1:T}$. Finally, generated data including face $\hat{P}_{fo\_f}^{k+1:T}$, body $\hat{P}_{fo\_b}^{k+1:T}$, and hands $\hat{P}_{fo\_h}^{k+1:T}$ are combined again into a single form $\hat{P}_{fo}^{k+1:T}$. Both $P_{fo}^{k+1:T}$ and $\hat{P}_{fo}^{k+1:T}$ are injected into the *Discriminator* network. In the sequel, we present the details of the proposed approach.

3.4. Context Encoder: Motion Encoder and MFCCs Encoder

Context Encoder E consists of *Motion Encoder* and *MFCCs Encoder*. The network receives the motion $P_{ob}^{0:k}$ and the audio data $A_{ob}^{0:k}$ acquired from S_{ob} as the inputs . More explicitly, the input $P_{ob}^{0:k}$ is fed to *Motion Encoder* while $A_{ob}^{0:k}$ is handled by *MFCCs Encoder*. *Motion Encoder* is constructed with a Long-Short Term Memory (LSTM) layer and a fully connected layer. This network encodes the interacting partner's motion $P_{ob}^{0:k}$ consisting of face, body, and hands into a latent vector c_P. On the other hand, the audio data $A_{ob}^{0:k}$ is fed to *MFCCs Encoder* to create a representation output c_A. Here, we extract MFCC features from $A_{ob}^{0:k}$ (see Section 5.1 for further details). MFCCs are well known to encode signal frequencies according to how humans perceive sounds, these low-level features are widely utilized in speech recognition or identification tasks (Vergin et al., 1999; Murty and Yegnanarayana, 2005). The extracted MFCC features are passed through a series of three $1D$ convolutions and a fully connected layer. On each layer, batch normalization is applied and followed by Rectified Linear Unit (ReLU) activation. Finally, c denotes the contextual vector formed by concatenating c_A encoded by *MFCCs Encoder* with c_P encoded by *Motion Encoder*.

3.5. Generator

Generator G consists of three networks, namely G_{face}, G_{body}, and G_{hand}, to handle the generation of face, body, and hand motions, respectively. We address the problem of variations in joint distribution among different body parts by separating a *Generator* into three smaller networks. Each network concentrates on a specific body area to better imitate a particular joint distribution. Each *Generator* receives the contextual information c provided by *Context Encoder* and the target person's motion as the inputs to predict the short-term future motion, $P_{fo}^{k+1:T}$.

The three *Generator* subnetworks, G_{face}, G_{body} and G_{hands}, are shown in Fig. 1. Each subnetwork consists of $G_{Encoder}$ and $G_{Decoder}$. For instance, in the case of G_{face}, $G_{Encoder}$ is constructed with two LSTM layers, it receives $P_{fo\_f}^{0:k-1}$ as the input and generate the internal representation h_e : $h_e \leftarrow G_{Encoder}(P_{fo\_f}^{0:k-1})$. On the other hand, $G_{Decoder}$ plays the role as a conditional generative network which is built upon two LSTM layers and a fully connected layer. $G_{Decoder}$ receives the resulting vector h_e encoded by $G_{Encoder}$ as an initial hidden state, and $P_{fo\_f}^{k}$ as an initial pose input. At the time stamp t, $G_{Decoder}$ takes the combined vector between its own prediction $\hat{P}_{fo\_f}^{t-1}$ and the contextual information c to forecast the next motion frame $\hat{P}_{fo\_f}^{t}$. Additionally, a residual connection (Martinez et al., 2017) is added between the input and the output of each LSTM cell of $G_{Decoder}$ to foster the continuity of generated motions. The same generation pipeline is implemented in G_{body} and G_{hands}. Finally, face $\hat{P}_{fo\_f}^{k+1:T}$, body $\hat{P}_{fo\_b}^{k+1:T}$, and hands $\hat{P}_{fo\_h}^{k+1:T}$ are concatenated into $\hat{P}_{fo}^{k+1:T}$ representing an action of the target person.

3.6. Discriminator

Both real $P_{fo}^{0:k}$ and fake action $\hat{P}_{fo}^{0:k}$ are injected into the *Discriminator* network, D. The role of D is to tell whether the inputs are sampled from real or generated distribution. Additionally, *Discriminator* also takes the contextual vector c as a conditional input; c delivers information allowing *Discriminator* to validate the synthesis of input motion and the interaction context. This idea has been shared across non-verbal behaviours generation studies (Ahn et al., 2018; Sun et al., 2020; Wu et al., 2021). The motion sequence input and the contextual vector c are concatenated and fed to D, which is designed with two LSTM layers and a fully connected layer. The output value is passed through a sigmoid function to produce a probability indicating whether the input motion is real or fake. Here, D works as a smart adaptive loss function, this adversarial loss encourages *Generator* to produce more realistic motions in alignment with the interaction context.

Overall, the framework demonstrated in Fig. 1 is trained with the loss functions $\mathcal{L}_G$ and $\mathcal{L}_D$, where $\mathcal{L}^{MSE} = \frac{1}{T-k-1}\sum_{t=k+1}^{T} ||P_{fo}^{t} - \hat{P}_{fo}^{t}||_2^2$. We used $\mathcal{L}_G$ to train the *Context Encoder* and *Generator*, while $\mathcal{L}_D$ is utilized for optimizing *Discriminator*. Here, $\mathcal{L}_{face}^{MSE}$, $\mathcal{L}_{body}^{MSE}$, and $\mathcal{L}_{hands}^{MSE}$ are the reconstruction losses indicating mean square errors between the ground truth and the generated data of face, body, and hands, respectively. α_1, α_2, α_3, and β are parameters to control the weights of the loss terms. The training pipeline is summarized in Algorithm 1.

$$\mathcal{L}_G = \alpha_1 * \mathcal{L}_{face}^{MSE} + \alpha_2 * \mathcal{L}_{body}^{MSE} + \alpha_3 * \mathcal{L}_{hand}^{MSE} + \beta * log(1 - D(c, \hat{P}_{fo}^{k+1:T})) \tag{2}$$

$$\mathcal{L}_D = -log(D(c, P_{fo}^{k+1:T})) - log(1 - D(c, \hat{P}_{fo}^{k+1:T})) \tag{3}$$

Algorithm 1: Generating Nonverbal Social Signals with Context-Aware GAN

```
Input: Interacting partner gesture $P_{ob}^{0:k}$, interacting partner audio features $A_{ob}^{0:k}$,
       target person gesture $P_{fo}^{0:k}$, training steps $S$
for $s \leftarrow 1$ to $S$ do
    $c_P \leftarrow MotionEncoder(P_{ob}^{0:k})$;
    $c_A \leftarrow MFCCsEncoder(A_{ob}^{0:k})$;
    $c \leftarrow concat(c_P, c_A)$;
    $\hat{P}_{fo\_f}^{k+1:T} \leftarrow G_{face}(c, P_{fo\_f}^{0:k})$;
    $\hat{P}_{fo\_b}^{k+1:T} \leftarrow G_{body}(c, P_{fo\_b}^{0:k})$;
    $\hat{P}_{fo\_h}^{k+1:T} \leftarrow G_{hand}(c, P_{fo\_h}^{0:k})$;
    $\hat{P}_{fo}^{k+1:T} \leftarrow concat(\hat{P}_{fo\_f}^{k+1:T}, \hat{P}_{fo\_b}^{k+1:T}, \hat{P}_{fo\_h}^{k+1:T})$;
    $y_r \leftarrow D(c, P_{fo}^{k+1:T})$;
    $y_f \leftarrow D(c, \hat{P}_{fo}^{k+1:T})$;
    Update $D$ with $\mathcal{L}_D$;
    Update $G$, $E$ with $\mathcal{L}_G$;
end
```

4. Evaluation Metrics

Generated motions are quantitatively evaluated by analyzing the differences between generated motions $\hat{P}_{fo}^{k+1:T}$ and the ground truth ones $P_{fo}^{k+1:T}$. The closer values to 1, the more similar to the ground truth motions. As given in Eq. 4, the score of *Face* is computed based on the Area Under the Curve (AUC) of the Cumulative Error Distribution (CED) as implemented by Huang et al. (2021). *Body* score is determined as AUC of the Percentage of Correct Keypoints (PCK) (Andriluka et al., 2014). Lastly, the score of *Hands* is calculated based on the AUC of the Success Rate (SR) (Yuan et al., 2018) on both left and right hand.

$$Face = AUC_{CED(0:0.25)}^{P_{fo\_f}^{k+1:T}} \qquad Body = \frac{1}{N}\sum_{i=0}^{N_b} AUC_{PCK_i(0:0.5)}^{P_{fo\_b}^{k+1:T}} \qquad Hands = AUC_{SR(0:0.5)}^{P_{fo\_h}^{k+1:T}} \tag{4}$$

5. Experimental Results on the UDIVA Dataset

5.1. Data Pre-processing

The designed framework is first validated on the UDIVA v0.5 dataset (Palmero et al., 2021, 2022). UDIVA is the time-synchronized multimodal, multi-view dataset of dyadic human interactions recorded in different communication scenarios. In this experiment, we used the data collected from participants involved in talk scenarios. The participants are instructed to talk about any topics while upper body gestures are naturally performed to support their communication. UDIVA can be considered as a communication-oriented dyadic interaction dataset, allowing us to forecast non-verbal gestures performed in social communication

Table 1: Low-level features extracted from the audio input.

Feature	Dimension
Mel Frequency Cepstral Coefficients (MFCC)	13
delta-MFCC (1st)	13
delta-MFCC (2nd)	13
Total	39

contexts. At the same time, we can verify the contribution of the interacting partner's non-verbal signals to the forecasting of the target individual's signals.

The dataset consists of 116 sessions for training, 18 sessions for validating, and 11 sessions for testing. Each interaction session was recorded in 5 minutes with a frame rate of 25 fps. For pre-processing, we segmented interaction sessions into equal size instances of 150 frames (6 seconds) with a sliding window of 20 frames. The forecasting problem was formulated as the prediction of the last 50 frames (2 seconds), given the first 100 frames (4 seconds) in an interaction segment. In terms of the motion data $P^{0:T} \in \mathbb{R}^{150\times2\times78}$, it comprised 150 motion frames and each skeleton frame was constructed by 78 joint coordinates (face = 28, body = 10, and hands = 40) defined in 2D space. Motion data was normalized by taking into account the mean and standard deviation values over the whole time sequence. From the audio $A_{ob}^{0:T}$, we extracted the low-level features with a total dimension of 39 as depicted in Table 1. The audio was processed at a frame rate of 100 fps. The extracted MFCC features were normalized and down-sampled to match the motion frequency of 25 fps. We finally obtained 30964 samples for training, 1196 samples for validating, and 556 samples for testing.

5.2. Implementation Details

The training data was fed to the framework with a batch size of 256. We used the Adam optimizer with a learning rate of 0.0005 and parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ for training. Weights for the loss functions L_G and L_D were chosen empirically ($\alpha_1 = 10$, $\alpha_2 = 5$, $\alpha_2 = 10$, $\gamma = 1$). During the first 50 warm-up epochs, the adversarial loss was not applied in the loss function L_G.

5.3. Ablation Study

In this section, we perform detailed ablation experiments to evaluate the impact of individual model components on the generated motions $P_{fo}^{k+1:T}$ of the target person. Table 2 summarises the key components of seven ablation models implemented in this experiment. The description of individual framework can be detailed as follows:

1. *full model*: This is our proposed solution as illustrated in Section 3. The framework consists of E, G, and D. Here, G is designed with three networks: G_{face}, G_{body}, and G_{hands} to handle the generation of face, body, and hand motions, respectively .

Table 2: The comparison of seven models evaluated in the ablation study.

No	Model	Components			
		Generator	*Context Encoder*		*Discriminator*
			Motion	MFCCs	
1	full model	✓	✓	$A_{ob}^{0:k}$	✓
2	single*Generator*	✓	✓	$A_{ob}^{0:k}$	✓
3	w/o *Discriminator* v1	✓	✓	$A_{ob}^{0:k}$	none
4	full model v2	✓	✓	$A_{fo}^{0:k}$	✓
5	w/o *Discriminator* v2	✓	✓	$A_{fo}^{0:k}$	none
6	w/o Audio *Encoder* Tan Viet Tuyen and Celiktutan (2021)	✓	✓	none	✓ (no c input)
7	Seq2Seq Martinez et al. (2017)	✓	none	none	none

Table 3: Prediction scores of seven models reported on the UDIVA dataset in terms of $Face$, $Body$, and $Hands$ metrics introduced in Eq. 4. The closer values to 1, the more similar to the ground truth values.

No	Model	Face	Body	Hands
1	full model	**0.263**	**0.867**	**0.392**
2	single*Generator*	0.194	0.859	0.319
3	w/o *Discriminator* v1	0.179	0.825	0.209
4	full model v2	0.238	0.863	0.327
5	w/o *Discriminator* v2	0.174	0.814	0.170
6	w/o Audio *Encoder*	0.204	0.850	0.316
7	Seq2Seq	0.050	0.801	0.186
	Ground Truth	1	1	1

2. *singleGenerator*: Rather than dividing the motion data into three different areas (body, face, and hands) in which each of them is handled by a specific *Generator* as employed in *full model*, in *singleGenerator*, the entire motion data of $S_{fo}^{0:k}$ is managed by a single *Generator*.

3. *w/o Discriminator v1*: This network is designed similar to *full model* except that D is withdrawn. In other words, the adversarial loss is not contributed to the training process.

4. *full model v2*: This model investigates a possibility of combining audio features $A_{fo}^{0:k}$ of the target person with the motion data $P_{ob}^{0:k}$ of the interacting partner to create the contextual information c.

5. *w/o Discriminator v2*: It is implemented similar to *w/o Discriminator v1*. However, MFCCs *Encoder* receives $A_{fo}^{0:k}$ collected from S_{fo} as the input, rather than $A_{ob}^{0:k}$ as applied in the *w/o Discriminator v1*.

6. *w/o Audio Encoder* (Tan Viet Tuyen and Celiktutan, 2021): Similar to *full model*, this framework is designed with *Context Encoder* E, *Generator* G, and *Discriminator* D. G also consists of G_{face}, G_{body}, and G_{hand}. However, E is not equipped with MFCCs *Encoder*, and D does not receive the conditional input c provided by E.

7. *Seq2Seq*: Using the motion data $P_{fo}^{0:k}$ only, $P_{fo}^{k+1:T}$ can also be estimated by the well-known *Seq2Seq* network introduced by Martinez et al. (2017). This approach is affiliated with the conventional motion forecasting task mentioned in Section 2.

Table 3 presents the prediction scores of face, body, and hands using the evaluation metrics defined in Section 4, where values closer to 1 (i.e., similar to the ground truth) are desirable. Overall, it can be seen that *full model* yields the best performance with respect to face, body, and hands. Further details about differences between generated motions of *full model* and ground truth are visualised in Fig. 2.

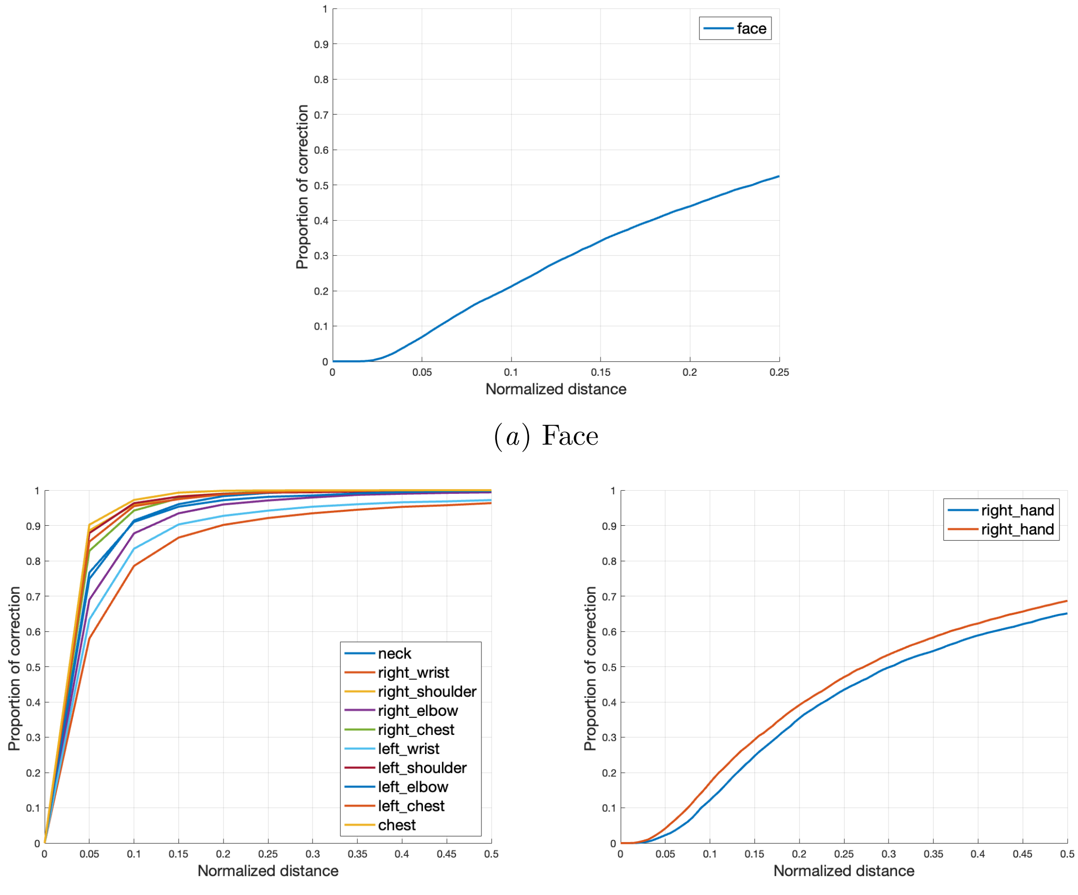
Looking at Table 3, body scores are not much different among ablation models. This result could be interpreted by the nature of the UDIVA dataset in which all interactions were recorded when participants were sitting down around a table. This interaction setup may have constrained participants to perform extensive gestures except for hand and facial movements. As a result, body joints could be easily predicted. The prediction scores of face and hands are highly different among implemented models. In the following, we take into consideration of $Face$ and $Hands$ metrics illustrated in Eq. 4 to discuss in more detail about the model performances.

5.3.1. Generators vs a Single Generator

As compared to *singleGenerator*, *full model* can further improve the accuracy of generated motions. On the UDIVA dataset, joint distributions are significantly different among face, body, and hand areas. For instance, the upper body is sparsely exhibited by 10 markers, while 28 landmarks are densely located in a small area to portray facial movements. This factor implies that different *Generator*s are needed to treat different body parts appropriately. With the *full model* approach, the problem is addressed by creating three *Generator*s, namely, G_{face}, G_{body}, and G_{hands}. Then, the features generated by the three networks are combined again into a completed form $P_{fo}^{k+1:T}$ before feeding to D. In the fully implemented model, *Discriminator* D works as a smart adaptive loss function for the whole training framework. The adversarial loss allows G to optimize realistic features of generated motions. The results in Table 3 suggest that the accuracy of generated gestures is reduced when removing D out the training framework.

5.3.2. The Contribution of Context Encoder to Generated Motions

In this experiment, *full model* and *full model v2* share the same network architecture. However, they receive audio input from different sources. The results demonstrate that using audio features extracted from $A_{ob}^{0:k}$ further improve the model performance as compared to



(a) Face

(b) Body

(c) Hands

Figure 2: Normalized distances between ground truth and the motion features generated by *full model*. The experiment was conducted on the UDIVA dataset.

the audio features provided by S_{fo}. This result implies that the concatenated information encoded from $A_{ob}^{0:k}$ and $P_{ob}^{0:k}$ is more informative for the generative framework than the combination between $A_{fo}^{0:k}$ and $P_{ob}^{0:k}$. This is probably because $A_{ob}^{0:k}$ and $P_{ob}^{0:k}$ are better correlated to each other in both temporal and spatial dimension as those are provided by the same source (the person S_{ob}), and the same time sequence ($t \in [0, k]$).

In the scenario where *Context Encoder* E and *Discriminator* D are removed from the *full model*, the problem of non-verbal behaviours forecasting can be addressed by a well-known *Seq2Seq* approach introduced by Martinez et al. (2017). With *Seq2Seq*, information of the interacting partner is not used to forecast the target user motions, and it is interesting to notice that face and body scores are significantly reduced compared to the *full model*. Fig. 3 and Fig. 4 quantitatively demonstrate differences among predicted actions produced by the two models in two different interaction sessions. A closer look at Fig. 3(c), it is observed that facial landmarks representing eyebrows and mouth are not neutrally displayed as can be seen in the ground truth. Concerning hand markers, both *full model* and *Seq2Seq* fail to predict the future trajectory compared to the true motion presented in Fig. 3(a). However, generated hand gestures created by *full model* appear to be more natural than the ones created by *Seq2Seq*. This problem could be explained by the lack of adversarial loss provided by *Discriminator*, and the drawback of using a single *Generator* network for modeling different body areas.

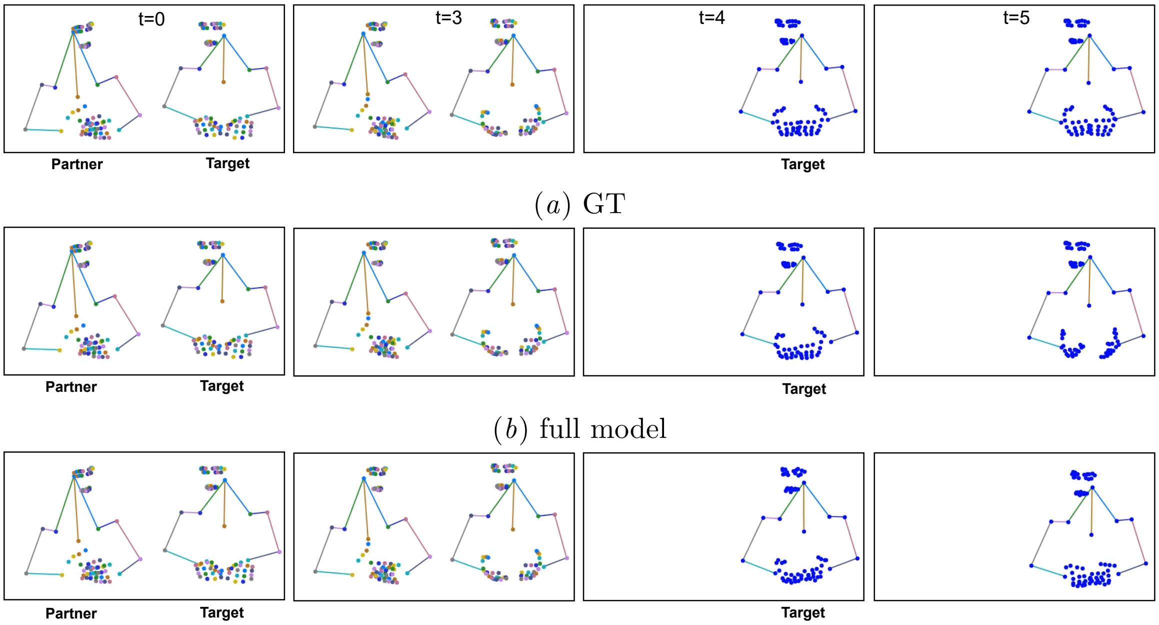
The low accuracy in the facial movements generated by *Seq2Seq* can be further explained by the lack of contextual information provided by *Context Encoder* E. It should be revisited that with the *Seq2Seq* solution, the network only receives the current non-verbal data of the target person S_{fo} to predict a possible short-term future motion. This approach might be inappropriate for scenarios of dyadic social interaction where the dynamic exchange of non-verbal signals among interlocutors is unavoidable. Fig. 4 depicts a simple case that the interacting partner's gesture may contribute to the generation of the target user's motion. In Fig. 4(a), at the time stamp $t = 3$, S_{fo} looks at elsewhere. In the next two seconds ($t \in [4, 5]$), S_{fo} turns their head back to look at the interacting partner, possibly, for maintaining an eye-contact during interactions. At the time stamp $t = 4$, generated facial movements of *full model* and the *Seq2Seq* are almost similar to each other since both receive the same initial pose input. However, differences can be clearly observed at $t = 5$. With the *full model*, S_{fo} also turns their head to look at S_{ob} as observed in the ground truth. Vice versa, the head orientation generated by *Seq2Seq* remains unchanged as it solely relies on the history of motion of S_{fo} to predict the next motion sequences.

6. Experimental Results on the AIR-Act2Act Dataset

6.1. Data Pre-processing

In contrast to the UDIVA communication-oriented dataset discussed in Section 5, in this experiment, we validate the designed framework on the Act2Act (Ko et al., 2021) dataset. Broadly speaking, this dataset includes a series of task-oriented dyadic interactions. Similar to the UDIVA dataset, AIR-Act2Act is the time-synchronized multi-view dataset of human social interaction. The dataset consists of 10 interaction scenarios with 100 participants involved in the data collection phase to create a total of 5000 interaction sessions. Since this dataset features a large number of subjects and interaction sessions, a greater variation in the motion data within interaction scenarios can be observed. As a task-oriented dataset of dyadic interaction, this dataset allows us to examine better the influence of the interacting partner behaviours on the target person.

On each interaction section, we collected the motion data of two interlocutors at a frame rate of 15 fps. By considering the common length of dyadic interactions conducted in AIR-Act2Act, we segmented interaction sessions into equal sizes of 60 (4 secs) frames with a sliding window of 10 frames. The first 30 frames were used as an observed window, and the forecasting task was to predict the remaining, last 30 frames. In AIR-Act2Act, raw motion data included 25 joint coordinates representing the entire body motion defined in 3D. We used 14 joints that correspond to the upper body motion ($P^{0:T} \in \mathbb{R}^{60\times3\times14}$). This dataset contains several locomotion actions. To eliminate the effects of camera distance and body size, we reconstructed and normalized all joint coordinates with respect to the central



(*a*) GT

(*b*) full model

(*c*) Seq2Seq

Figure 3: Visualisation of the ground truth motions (GT) and predicted motions using the fully implemented model (*full model*), and the framework introduced in Martinez et al. (2017) (*Seq2Seq*). The last two windows ($t \in [4, 5]$) depicts generated motions.

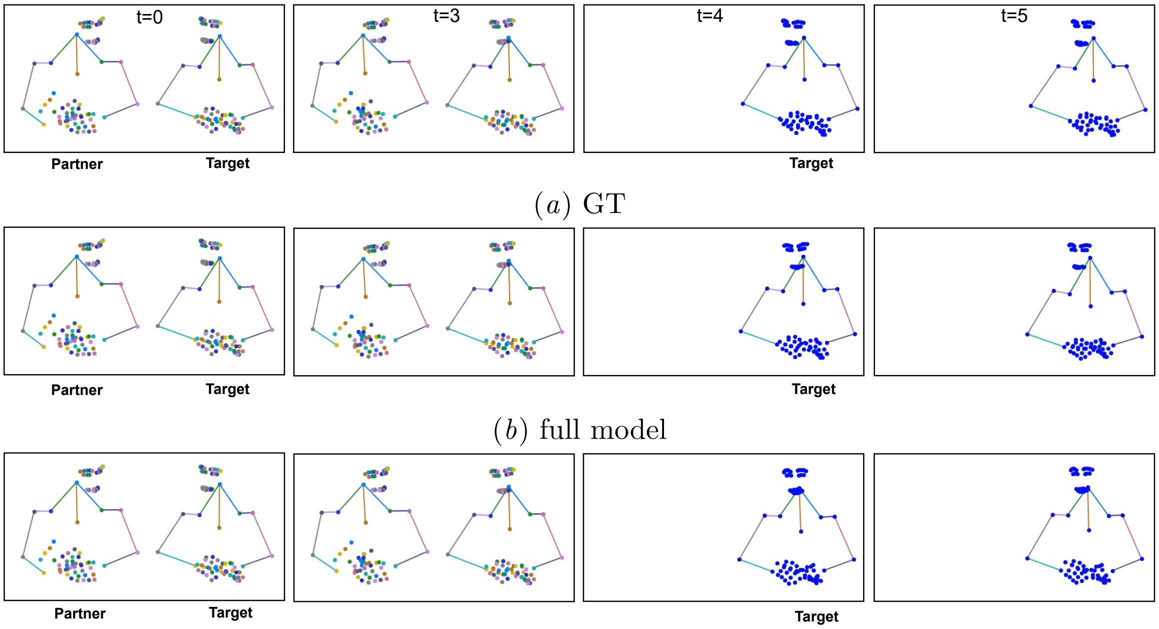
hip. Additionally, motion frames were transformed to the frontal view to disregard camera orientations. The audio source was not provided in this dataset, so MFCCs *Encoder* was not applied in the fully implemented model.

6.2. Implementation Details

The network architecture was designed similar to the fully implemented model illustrated in Section 5. The model consists of *Context Encoder*, *Generator*, and *Discriminator*. In AIR-Act2Act, *Generator* only handles body motions, so it was constructed with a single *Generator* G_{body}. The model was optimized by minimizing the losses L_G and L_D. The network was trained with a batch size of 512. The Adam optimizer was used with a learning rate of 0.001 and parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$. In the loss function L_G, α_2 and β were set to 5 and 1, respectively. In the first 50 training epochs, the adversarial loss was not applied in the loss function L_G.

6.3. Ablation Study

To evaluate the contribution of context awareness, *w/o Context Encoder* model was constructed similarly to *full model* except that Motion *Encoder* was removed. Table 4 summarizes model components of the two designed framework, and their performance on the



(*a*) GT

(*b*) full model

(*c*) Seq2Seq

Figure 4: Visualisation of GT and predictions from *full model* and *Seq2Seq*, demonstrating that *full model* performs better than *Seq2Seq* in predicting facial cues.

Table 4: Ablation study conducted on the AIR-Act2Act dataset.

No	Model	Components				Body
		Generator	*Context Encoder*		*Discriminator*	
			Motion	MFCCs *Encoder*		
1	*full model*	✓	✓	none	✓	**0.772**
2	w/o *Context Encoder*	✓	none	none	✓	0.636
Ground Truth						1

testing data. Further details about accuracy individual joints generated by *full model* are illustrated in Fig 5.

Overall, the body scores produced by the AIR-Act2Act dataset seem to be lower than the ones validated on the UDIVA data. That could be explained by the differences in interaction scenarios between the two datasets, where AIR-Act2Act concentrates on body modality to communicate interlocutors' intentions. Consequently, variations on body channels are more significant than the UDIVA dataset. There are also several differences in the pre-processing steps between the two datasets. Those factors imply that the evaluation scores can only be used to analyze the network performances within the same dataset.

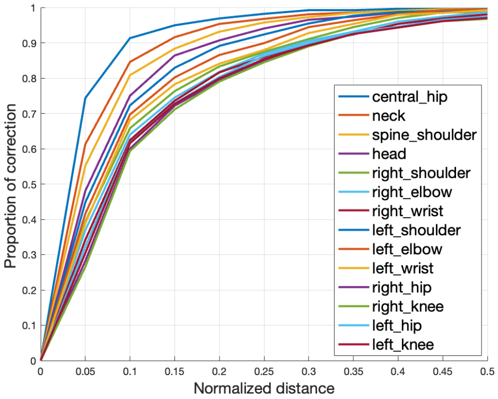
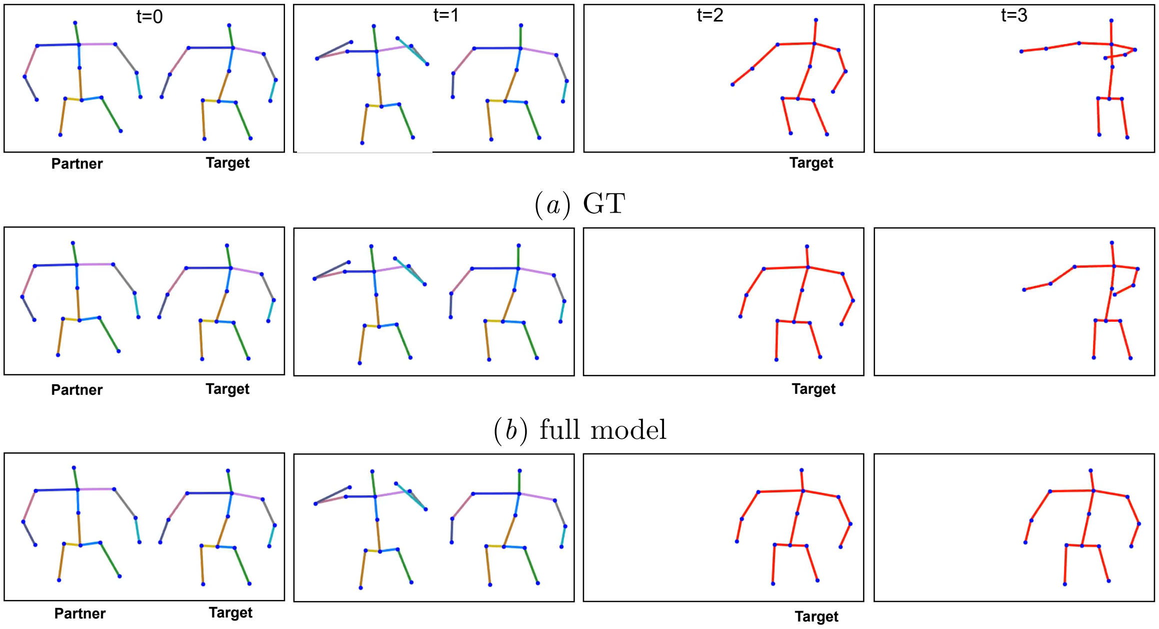


Figure 5: Normalized distances between ground truth and the body joints generated by *full model*. The experiment was carried out on the AIR-Act2Act dataset.



(a) GT

(b) full model

(c) w/o *Context Encoder*

Figure 6: Differences among the ground truth motions (GT), the fully implemented model (*full model*), and the framework without *Context Encoder* (w/o *Context Encoder*). The last two windows ($t \in [2, 3]$) displays generated motions.

6.3.1. The effect of interacting partner gestures on the target person

The results presented in Table 4 demonstrate that the fully implemented model yields better performance than the one without *Context Encoder*. It should be remarked that AIR-Act2Act covers a series of predefined interaction scenarios in which the effects of initiating behaviours performed by S_{ob} on the responsive behaviours of the person S_{fo} are clearly observable. Consequently, the nature of the AIR-Act2Act dataset allows us to better verify the contribution of the context c to the generation of $\hat{P}_{fo}^{k+1:T}$. Fig. 6 presents an interaction snippet from the test data, illustrating this phenomenon. In this example, S_{ob} initializes an interaction by raising their two hands above their shoulder for a "high five". In the next two seconds ($t \in [2, 3]$), S_{fo} responds to the "high five" of S_{ob} by raising their two hands into the air. A closer look at the action $\hat{P}_{fo}^{k+1:T}$ in Fig. $6(b)$, the result indicates that by observing the S_{ob} motion as the contextual input c, $fullmodel$ is able to reason about a generated gesture in the next 2 seconds. In contrast, without obtaining c as a conditional input for the generative framework, it is challenging for *w/o Context Encoder* to accomplish the forecasting task as the temporal information presented in $P_{ob}^{k:0}$ is not sufficient to reason an appropriate motion. As shown in Fig.$6(c)$, *w/o Context Encoder* ends up in a such way that the body movement $\hat{P}_{fo}^{k+1:T}$ remains static over the time sequence.

7. Conclusion

This paper investigated a generative framework with context awareness to forecast human non-verbal gestures in dyadic interaction. The model was constructed with *Context Encoder*, *Generator*, and *Discriminator* networks. We conducted an ablation study on the UDIVA dataset to verify the impact of each model component. The experimental results indicated that by creating a network consisting of three different *Generator*s to handle three different body parts, the accuracy of generated motion can be enhanced. Indeed, the model equipped with *Context Encoder* yielded better performance than the one without observing information encoded from the interacting partner. The proposed model was further evaluated on the AIR-Act2Act dataset, where the impact of the non-verbal signals of the interacting partner on those of the target person were clearly visible. Again, the results confirmed the efficiency of the fully implemented model to capture the interacting contexts.

In social interaction, non-verbal behaviours are essential channels to convey the interlocutor's intentions or emotions. Undoubtedly, the dynamic exchange of social signals among interlocutors should be investigated when modeling their non-verbal gestures. This paper contributes a generative framework that can effectively forecast the human upper body, including face, body, and hand gestures in dyadic interaction. Importantly, our approach treats social signals observed from the interacting partners as essential information to forecast the future motions of the target individual.

There are several unexplored points that could be investigated as future works, including a subjective evaluation of generated motions and modeling a wider range of social signals acquired from two interlocutors (e.g., semantic contents of speech, eye gaze, etc.). Finally, the proposed approach can be implemented for social robots to better support scenarios of human-robot interaction.

Acknowledgments

This work has been supported by the “LISI - Learning to Imitate Nonverbal Communication Dynamics for Human-Robot Social Interaction” project, funded by Engineering and Physical Sciences Research Council (Grant Ref.: EP/V010875/1).

References

Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. Text2action: Generative adversarial synthesis from language to action. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5915–5920. IEEE, 2018.

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.

Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. Beat: the behavior expression animation toolkit. In *Life-Like Characters*, pages 163–185. Springer, 2004.

Will Feng, Anitha Kannan, Georgia Gkioxari, and C Lawrence Zitnick. Learn2smile: Learning non-verbal interaction through observation. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4131–4138. IEEE, 2017.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José MF Moura. Adversarial geometry-aware human motion prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 786–803, 2018.

Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018.

Yangyu Huang, Hao Yang, Chong Li, Jongyoo Kim, and Fangyun Wei. Adnet: Leveraging error-bias towards normal direction in face alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3080–3090, 2021.

Yuchi Huang and Saad M Khan. Dyadgan: Generating facial expressions in dyadic interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 11–18, 2017.

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.

Mark L Knapp, Judith A Hall, and Terrence G Horgan. *Nonverbal communication in human interaction*. Cengage Learning, 2013.

Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, and Jaehong Kim. Air-act2act: Human–human interaction dataset for teaching non-verbal social behaviors to robots. *The International Journal of Robotics Research*, 40(4-5):691–697, 2021.

Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. Analyzing input and output representations for speech-driven gesture generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 97–104, 2019.

Jessica L Lakin, Valerie E Jefferis, Clara Michelle Cheng, and Tanya L Chartrand. The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of nonverbal behavior*, 27(3):145–162, 2003.

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.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

K Sri Rama Murty and Bayya Yegnanarayana. Combining evidence from residual phase and mfcc features for speaker recognition. *IEEE signal processing letters*, 13(1):52–55, 2005.

Lior Noy, Erez Dekel, and Uri Alon. The mirror game as a paradigm for studying the dynamics of two people improvising motion together. *Proceedings of the National Academy of Sciences*, 108(52):20947–20952, 2011.

Cristina Palmero, Javier Selva, Sorina Smeureanu, Julio CS Jacques Junior, Albert Clapés, Alexa Moseguí, Zejian Zhang, David Gallardo, Georgina Guilera, David Leiva, et al. Context-aware personality inference in dyadic scenarios: Introducing the udiva dataset. In *WACV (Workshops)*, pages 1–12, 2021.

Cristina Palmero, German Barquero, Julio C. S. Jacques Junior, Albert Clapés, Johnny Núñez, David Curto, Sorina Smeureanu, Javier Selva, Zejian Zhang, David Saeteros, David Gallardo-Pujol, Georgina Guilera, David Leiva, Feng Han, Xiaoxue Feng, Jennifer He, Wei-Wei Tu, Thomas B. Moeslund, Isabelle Guyon, and Sergio Escalera. Chalearn LAP challenges on self-reported personality recognition and non-verbal behavior forecasting during social dyadic interactions: Dataset, design, and results. In *Understanding Social Behavior in Dyadic and Small Group Interactions*, volume 173 of *Proceedings of Machine Learning Research (PMLR)*, pages 4–52, 2022.

Chirag Raman, Hayley Hung, and Marco Loog. Social processes: Self-supervised forecasting of nonverbal cues in social conversations. *arXiv preprint arXiv:2107.13576*, 2021.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069. PMLR, 2016.

Shane Saunderson and Goldie Nejat. How robots influence humans: A survey of nonverbal communication in social human–robot interaction. *International Journal of Social Robotics*, 11(4):575–608, 2019.

Guofei Sun, Yongkang Wong, Zhiyong Cheng, Mohan S Kankanhalli, Weidong Geng, and Xiangdong Li. Deepdance: music-to-dance motion choreography with adversarial learning. *IEEE Transactions on Multimedia*, 23:497–509, 2020.

Nguyen Tan Viet Tuyen and Oya Celiktutan. Forecasting nonverbal social signals during dyadic interactions with generative adversarial neural networks. *arXiv e-prints*, pages arXiv–2110, 2021.

Nguyen Tan Viet Tuyen, Armagan Elibol, and Nak Young Chong. Conditional generative adversarial network for generating communicative robot gestures. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 201–207. IEEE, 2020.

Rivarol Vergin, Douglas O'Shaughnessy, and Azarshid Farhat. Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition. *IEEE Transactions on speech and audio processing*, 7(5):525–532, 1999.

Bowen Wu, Chaoran Liu, Carlos Toshinori Ishi, and Hiroshi Ishiguro. Modeling the conditional distribution of co-speech upper body gesture jointly using conditional-gan and unrolled-gan. *Electronics*, 10(3):228, 2021.

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.