ISSN NO : 2230-7850 IMPACT FACTOR : 5.1651 (UIF) VOLUME :8, ISSUE : 1, FEBRUARY- 2018

INDIAN STREMS RESEARCH JOURNAL

TAG2G: A Diffusion-Based Approach to Interlocutor-AwareCo-Speech Gesture Generation

Dr. Rekha

Assistant professor in Electronics Government College (Autonomous)
Kalaburagi.

Abstract:

Systems for extended reality (XR) are going to become a part of our everyday life and help a number of industries, including coaching and education. Improving the user experience requires agents to be able to exhibit genuine affective and social behaviours in these systems. Understanding their interaction partner and responding accordingly is a prerequisite for this ability. Our study of recent literature on co-speech gesture production reveals that scientists have created sophisticated models that can produce gestures with a high degree of speaker appropriateness and human-likeness. However, this is only applicable in situations when the agent is speaking actively (that is, when it takes on the role of the speaker) or when it is presenting a monologue in a non-interactive

Keywords: nonverbal behavior generation; co-speech gesture generation; dyadic conversations; diffusion models; human–human interaction modeling

1. Introduction

Large language models (LLMs), one of the most recent developments in machine learning, have greatly improved intelligent virtual assistant technologies and allowed them to successfully engage consumers in unmatched open-ended conversations. The majority of these technologies are embodied and represented in virtual and mixed reality environments, as well as on a variety of XR platforms such computers, cellphones, and freestanding screens. These embodied agents will be more and more integrated into our daily lives to support us in our roles as coaches in physical and mental health activities, personal trainers in virtual and mixed reality systems, and receptionists in public services.

These virtual agents' interaction capabilities need to be more lifelike, mimicking how people engage and communicate with one another, in order to improve human-robot interaction. When simulating human-to-human interaction, vocal involvement is important, but

Nonverbal behavior allows us to support the displayed understanding and verbal con-tent through descriptive and iconic gestures. It also allows to create a connection through empathy and trust between the conversation partners, which further increases the effective- ness of communication. For example, in the domain of education and coaching, several studies show that maintaining an appropriate nonverbal behavior results in an increased effectiveness of the learning process [2–4]. In education, as well as in other contexts such as tailored care-giving, public reception, and information dissemination, progressively widespread and integrated extended reality (XR) and immersive applications [4] allow us to exploit the vast knowledge accessible through IT systems in an increasingly engaging and dedicated way for individual users via specific topics, idioms, and nonverbal behavior. Building upon the importance of generating appropriate nonverbal gestures, we present our work that delves into interactions between humans and virtual agents who take part in a dyadic conversation. Dyadic conversation consists of a two-person interactionvia both verbal and nonverbal communication. With special focus placed on nonverbal communication, existing works tackling this problem are effective at generating gestureswhile the agent is speaking; conversely, the generated motion for a listening agent is often
perceived as inappropriate by human observers [5].

This is mainly due to current methods, which focus on the conversational input fromthe main agent only while partially or totally discarding the interlocutor's information. As a result, an above-chance level of appropriateness with respect to the conversational counterpart is achieved, eventually leading to a low degree of partner consciousness [1]. In these terms, interlocutor awareness is still an open topic deserving of research effort.

Motivated by this, the main purpose of this paper is to increase the capabilities of the virtual agent while listening. To achieve this goal, we aimed to mimic human natural behavior such as nodding while the interlocutor speaks, mirroring their gestures, and other backchanneling behaviors. We achieved this by exploiting the contextual information in the dyadic conversation to better describe the lexical meaning of the dialogue and extractimportant insights from the conversation as a whole. Differently from the existing literature, we incorporated multimodal inputs taken from both participants of the conversation and we introduced a new architecture, TAG2G, combining **t**ext, **a**udio, and **g**estures **to g**enerate new gestures.

1.1. *Related Work*

To be perceived as appropriate in dyadic conversations, virtual agents need to master both verbal and nonverbal communication. When humans interact in pairs, they natu- rally take advantage of both aforementioned communication channels to properly deliver information, making assumptions and sharing ideas while building a strong link with interlocutors based on trust and shared beliefs [3]. Recent works [6,7] reached excellentresults when applying data-driven models to verbal communication, leveraging generative pre-trained transformer (i.e., GPT) architectures to handle natural language processing (NLP) tasks and, thus, accurately generating responses and taking an active part in conver- sations. Conversely, despite its importance, applying data-driven approaches to nonverbal communication is still in its infancy.

Nonverbal behavior is often divided into two different research areas, namely facial expression and body gesture generation. Facial expressions are commonly associated with necessary mouth and facial muscle movements to properly pronounce words and spell letters [1]. Therefore, a strong link exists between speech and facial expressions. Compared to facial expressions, body motion is a multifaceted and complex problem that concerns a large number of gesture movements that do not necessarily exhibit a strong correlationwith the ongoing dialogue. Moreover, there is no mechanical connection between the verbal channel and the body gesture that is employed by a human when interacting with their counterpart. Following this classification, in the following sections, we will further explore and limit our efforts to body gesture generation only.

In the last decade, many works [1] explored the significance of gestures when exploit- ing nonverbal channels during a conversation. At first, rule-based approaches [8–10] were introduced to deploy human-like motions with virtual agents. Such methods rely on a batch of predetermined, hard-coded motion features that can be adopted to link conversation with body motion given a specified set of rules. However, since these control patterns need to be hard-coded, they are limited in terms of the diversity of implemented actions that are perceived, and, in the long run, they become repetitive and usually lack contextual significance. For this reason, data-driven methods were introduced in order to expand variety and quality of motions while, at the same time, gaining the ability to generalize during the learning procedure. This highlighted the demand for more qualitative and sophisticated approaches to tackle the gesture generation problem via data-driven methods to cope with the expected quality of motion when interacting with a human interlocutor.

Data-driven methods for speech-driven gesture generation are enabled by the presence of multiple publicly available datasets, such as BEAT [11] and TalkingWithHands 16.2 M (TWH) [12]. These datasets are commonly used as benchmarks in competitions such as the "Generation and Evaluation of Nonverbal behavior for Embodied Agents" (GENEA) challenge [5,13,14]. These datasets are usually composed of heterogeneous streams of data accurately collected to represent the recorded conversations from a multimodal perspective.Gesture data, especially, are usually provided along with the audio of the speech. Text is often provided, otherwise it can be automatically obtained from audio itself. Moreover, datasets can contain additional information such as unique participant ID, emotional labels of the conversation or ethnicity and other personal details from each single participant. Such supplementary information is highly demanded by researchers to better describe the context of the conversation.

As highlighted by multiple works from the literature [1,15,16], heterogeneous con- versational information in the form of audio and text transcription and additional details acts as complementary blocks in order to successfully root nonverbal behavior into ver- bal communication. Indeed, text has been referred to as the quantitative information needed in order to root a gesture into the contextual meaning of the conversation, thus driving the generation on a long-term span of time. Conversely, audio-controlled gesturegeneration exhibits very good properties in terms of synchronization with the rhythm ofthe speech that we, as humans, are commonly used to deploy via tempo-related move- ments such as hand or head shaking. Consequently, speech-driven gesture generation is demonstrated to rhythmically rely on audio and semantically depend on text. As a result, multimodal input is often considered as the preferred representation of the conversation.In this context, specific procedures are typically used to preprocess raw signals, eventu- ally extracting relevant features [15,16]. Multiple baseline pipelines use properties such as prosody, mel-frequencies, and spectral analysis to deploy audio features into a more complex pipeline. In addition, pre-trained models are used to increase the dimensionality of the audio representation into latent space. WavLM [17] and Hubert [18] are examples of publicly available models suitable for this specific task. Text is also preprocessed using word-to-vector (Word2Vec) pre-trained models such as Crawl-300D_2M [19] in order to achieve a semantically meaningful representation of the dialogue. These models are trained to obtain embedding spaces where semantically coherent words are mapped into close points, therefore rooting the comprehension of the model into contextual information ofthe conversation.

Regarding the neural network architectures, at the very beginning, many researchers [13,15] employed deep learning models leveraging recurrent neural network (RNN) layers such as long-short term memory (LSTM) and gated recurrent units (GRUs) to model complex rela- tionships that link gestures to conversation, looking to predict the most appropriate motion. It is worth noting that multiple works [20,21] tried to represent gestures as encoded mo- tion features leveraging encoder–decoder architectures such as variational auto-encoders (VAEs) and vector-quantized variational auto-encoders (VQVAEs). This approach is par- ticularly interesting because of the capabilities of aforementioned models to learn latentrepresentations that can be further integrated into more complex pipelines [22]. More recently, state-of-the-art generative architectures such as generative adversarial networks (GANs) [23,24] and probabilistic denoising diffusion models (also known as diffusion models) [25–27] have been introduced in the domain of gesture generation, following theexcellent results obtained in other realms of generation such as text-to-image tasks. GAN architectures act as useful tools in online gesture generation tasks; however, they are knownto be hard to train due to the balancing of generator/discriminator loss. Moreover they highly rely on observed data while falling short in the generation of previously unseen samples, thus producing repetitive motions.

As highlighted in the literature, current limitations in the field of nonverbal behavior generation vary across multiple directions; thus, a plenitude of exploration directions are available. State-of-the-art approaches [27] are yielding good results in terms of human- likeness and speaker appropriateness of generated motions. On the other hand, above- chance results have been collected when assessing the appropriateness of a generated gesture when the agent is listening to their interlocutor [5]. Moreover, synthesized motions are scarcely related to an interlocutor's body gestures. Once the target agent is listening, iconic motions such as nodding, backchanneling and mirroring of the interlocutor's gestures are rarely visible, therefore highlighting the poor naturalness of generated motions.

In [28], the authors showed that the nonverbal channel can serve as a research tool to extract interlocutor's unspoken thoughts. Thus, this "hidden information" should be deciphered and better exploited when trying to predict an appropriate gesture for the agent. Nevertheless, only a handful of works actually explored the chance to extend multimodal input from a single agent to all the people taking part into the interaction [5,23,29,30]. Among all of the literature, only in [30] has it been proposed to train two specific models to dynamically interchange depending on whether the main agent is speaking or listening, while the others proposed different approaches to embed the interlocutor's information into gesture generation. Nevertheless, all proposed methods suffered from a lack of appropriateness while the agent is listening.

On the other hand, these methods suffer from below-average results when addressing prolonged speech delivered by the agent. Contrarily, speaker-only based approaches show enhanced capabilities of human-likeness and gesture appropriateness to the speech while the agent is engaged in a protracted monologue. Nevertheless, it is non-trivial for such approaches to generate accurate and appropriate gestures when listening to the interlocutordue to the lack of information from the counterpart [5]. As a result, current methods lack consistency with respect to gesture appropriateness between speaking and listening scenarios. In these terms, interlocutor awareness has the potential for acting as a trade- off between an agent's accurate speech-related gesture and a qualitative interlocutor's appropriate nonverbal behavior [30].

1.2. *Proposed Contribution*

To overcome the limitations of existing works and improve the capabilities of virtual agents towards dyadic conversations, we propose an architecture that combines a VQVAE and a diffusion model. The architecture is called TAG2G, since it leverages conversational information such as text (T), audio (A), and gesture (G), applied to a dyadic setup (2) to address speech-driven gesture (G) generation.

The advantages brought by the proposed approach are twofold. To the best of our knowledge, we are the first to use a dyadic multimodal input to tackle the co-speech gesture generation task, including text, audio, ID, and past observed gestures to predict the next movement. This allows an improvement of the appropriateness of generated gestures. Inaddition, we introduce a VQVAE to learn a latent representation of gestures in the form of a codebook of atomic actions. This allows to speed up training and inference time as compared with other state-of-the-art algorithms.

The rest of the paper is organized as follows: In Section 2.1, we formally describe and introduce the task of gesture generation extended to a dyadic multimodal input setup. Then, in Section 2.2 we delve into the formalization of the proposed architecture, while employed methodologies are treated in Sections 2.4 and 2.5. Materials and software and hardware platforms used during training procedures are highlighted in Sections 2.6 and 2.7; finally, readers can find experimental validation in Section 3. A discussion on experimental validation, results, and future works is presented in Sections 4 and 5.

2. Materials and Methods

2.1. *Problem Definition*

In this work, our main purpose was to extend current state-of-the-art methods in co- speech gesture generation to a dyadic setup, where both agents' information is leveraged to generate the nonverbal behavior of the target agent. Based on the work of [31,32], when taking part in a social interaction, people tend to react according to both their inner state, referred to as stimulus (i.e., the goal of the conversation), and to external information received from the context. The latter is a high-level representation of the verbal, nonverbal, and emotional states of each participant involved in the conversation, which acts as an external force driving the behavior of every single person. This validates the hypothesis that, given a single conversation, a large set of nonverbal behaviors are compatible with the conversation itself. Still, only a sub-set actually matches the context and, thus, will be perceived as real. Therefore, agents will appropriately interact with other people, allowing to build trust and empathizing with humans. In these terms, the gesture generation task can be described as an N-to-N problem, where, given an input stimulus (i.e., conversation to be deployed by the main agent), multiple gestures might be equally correct but none of them is the preferable one. According to the above-described scenario, context needs to be inferred from conversational and emotional state of each of the participants to synthesize and select the most correct subset of gestures that fit a given pair of context–conversation inputs. Consequently, both participants must be taken into account to successfully achieve a broader comprehension of stimuli and contextual information from the environment.

In the scope of this paper, we will refer to the target agent as the main agent while the conversational counterpart will be referred to as the interlocutor.

Our purpose is to understand the complex relationships that link gesture to conversa- tion using a multimodal set of inputs composed of text (T), audio (A), and gestures (Gs).The conversation is represented by audio and its text transcription per agent. Thus, we have A_{ma} and T_{ma} (audio and text, respectively) inputs from the main agent and A_i and T_i from the interlocutor. In addition, we can introduce G_i as the gesture observed from the interlocutor while generating the main agent's gesture G_{ma} during self-supervised training procedures.

Multiple works [25,26] highlighted the strength of diffusion and denoising procedure based architectures applied to conditioned generation tasks, such as co-speech gesture generation. Building upon such works, we employ a denoising diffusion probabilistic model over the above-mentioned latent space representation of motion to generate gesturesmatching the corresponding conversation and the interlocutor's behaviors.

We aim to synthesize gestures using information from both agents by following a sequential scheduling where a sample to be generated is subdivided into shorter sub- samples identified as sequences. In these terms, inspired by [22], our purpose is to learn atomic gestures applying a VQVAE architecture, where the VQ layer will act as a learned codebook of motions expressed through a latent representation.

Given this setup, from the observation of past gestures and audio from both agents, the past text of the main agent, and the next sequence of audio and text of the main agent,we aimed to predict the next sequence of gestures of the main agent. In doing so, we considered two different temporal horizons: while text and audio were observed over

a long-term horizon with sub-sequences of duration h, gestures were observed over a short-term horizon with sub-sequences of duration w, with $w < h$.

Thus, at a given time t, we predict the next main agent's atomic gesture as per Equation (1) and Figure 1:

$$\hat{G}^{t:t+h}_{ma} = P(A^{t:t+h}_{ma}, T^{t:t+h}_{ma}, G^{t-w:t}_{ma}, G^{t-w:t}_{i}, A^{t-h:t}_{i}) \tag{1}$$

where P represents our model, previously trained, at the inference time. The model P includes a short-term learning component to process gestures and a long-term component (diffusion model) to process the conversational inputs (text and audio).

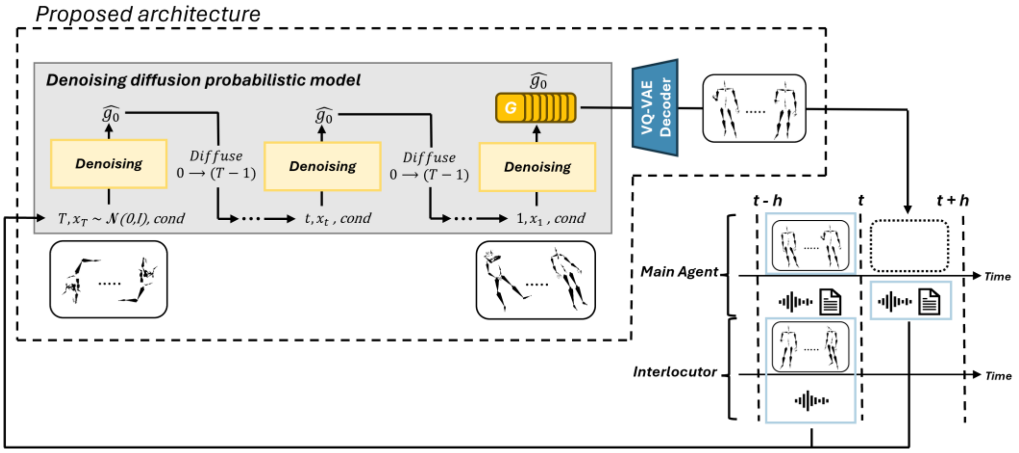

Figure 1. This figure represents the sequential generation obtained combining a VQVAE with a diffusion model from a temporal point of view. The conversational inputs from each agent (light-blue boxes) are fed to the proposed architecture. The gesture elements generated by the diffusion model are mapped back to the agent's joint space via the previously learned VQVAE's decoder.

2.2. Proposed Pipeline

To implement the model P in Equation (1), we propose a pipeline featuring a denoising diffusion probabilistic model, following the development acquired from [25,26] in the usage of such models in the realm of conditioned gesture generation. Figure 2 reports an overview of the proposed pipeline.

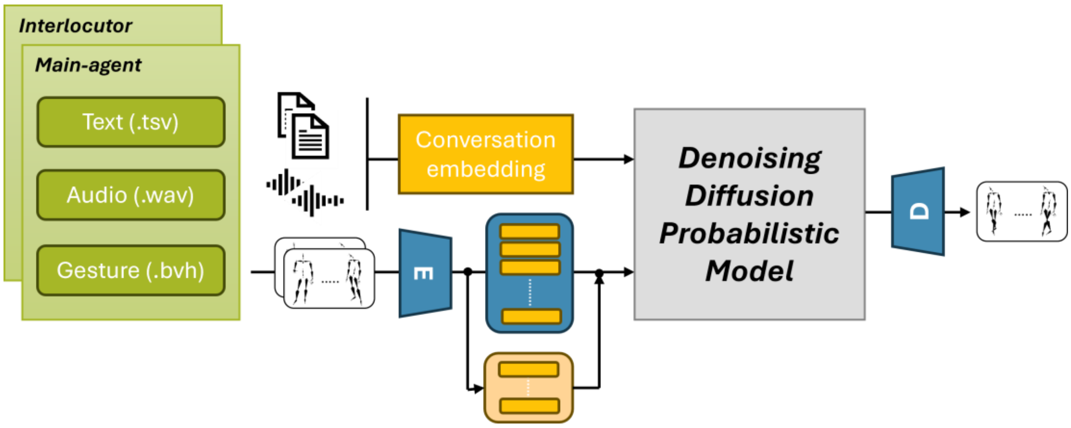

Figure 2. Proposed overall pipeline for dyadic gesture generation. A diffusion model is coupled with a VQVAE architecture. Both agents are included as inputs to predict the next sequence of motion. Text and audio raw data are embedded into latent representations leveraging pre-trained models. Gestures are mapped into a latent space by the encoder E. Encoder is also used to encode interlocutor motion, which is then mapped on a smaller latent space obtained via clustering over the original gesture codebook. This is done to reduce the exploration space of the diffusion model. At the end of denosing–diffusion procedures, the previously learned decoder D maps the generated gestures from the latent space back to the agent's joint space.

As previously stated, the biggest drawback of diffusion models is the large amount of time required during training procedures due to the iterative nature of diffusing and denoising steps. In addition, extending such architecture to multi-agent conversations, thus expanding the amount of conditioning information, would make the method even slower. To overcome these issues, we coupled the generative module with a previously learned embedded space that provides a compressed gesture representation both in joint and time dimensions. This allowed to speed up training and inference time. To map gestures onto the latent space, we employed a VQVAE architecture to learn a latent feature representation starting from short-term sequences of motion of length w.

The main purpose was to learn an agnostic representation of iconic gestures independent from the conversation to be further leveraged when learning longer sequences of dependencies from it. To this end, we used a VQ layer that represents a codebook C of gestures of fixed size $n \times d$. Thus, $C \in \mathrm{R}^{n \times d}$, and each code $c \in \mathrm{R}^{d}$. All the pipeline parameters are summarized in Table 1.

Table 1. Pipeline parameters.

Variable Name	Explanation	Value
w	Short-term window length	18 frames [1]
h	Long-term window length	144 frames [2]
n	Dimension of the codebook	2048
d	Dimension of each code in the codebook	256
J	Dimension of agent's space joint	74
K	Dimension of the clustering from codebook	16

[1] Given 30 fps from the dataset [5], the length is equal to 0.6 s in the time domain. [2] Given 30 fps from the dataset [5], the length is equal to 4.8 s in the time domain.

To deal with audio and text, we introduced a denoising diffusion probabilistic model to learn long-term gesture dependencies from conversation as well as contextual information. To this end, we employed the diffusion model over a longer horizon time equal to h. We also introduced any available interlocutor information to assess some kind of contextual information and let the model synchronize predicted motion with both the agents' speech. In addition, the interlocutor's past gestures and conversation information can be leveraged to predict the context in which the main agent has to deploy a nonverbal behavior to match the expectations and emotional state of their counterpart, eventually properly empathizing with it.

Thus, the main steps to train the above-mentioned pipeline for generating gestures are as follows:

- Encode conversational raw data into text and audio latent representations (see Section 2.3);
- Preprocess gesture data to extract beat gestures for VQVAE training (see Section 2.4);
- Train VQVAE over selected gestures $G_{VQVAE} = \{G_1, G_2, \ldots, G_n\}$ (see Section 2.4);
- Train the diffusion model over VQ latent space using conversational features as conditioning information (see Section 2.5);
- Use the trained diffusion model at inference time, as per Figure 1.

2.3. Multimodal Input Representation

We introduced audio and text to successfully ground nonverbal behavior to the main agent's speech. Therefore, we used pre-trained word-to-vector Crawl300D_2M [2] to obtain an embedded representation from text, thus providing a 300-dimensional feature per each word. The model was trained to project words into a semantically normalized latent space where words with similar meaning are mapped onto two close points. At the same time, we extracted MFCC, mel-spectograms and prosody from audio raw signals. As per multiple works [5], we decided to introduce WavLM_Large [17] in addition to the above-mentioned features with the purpose of obtaining a complete understanding of the conversation. We chose to discard text from the interlocutor while leveraging their audio to depict their emotional state via the extraction of MFCC, prosody, and mel-spectogram features.

In [33], the authors introduced a tri-modal input composed of text, audio, and the interlocutor's ID. Similarly, we decided to introduce the IDs of both participants as one-hot encoding. This information was used to control the style of the conversation, since each participant may interact with a different attitude and behavior.

Gestures were expressed leveraging exponential map rotation representation, discarding the face and fingers joints. Thus, the remaining full-body skeleton was represented by including both upper-body and lower-body joints. Differently from the existing methods in the literature, we aimed to generate full-body motion as a whole, not differentiating between upper- and lower-body gestures. Eventually, each gesture was represented as $G \in R^{J \times T}$, where J is the dimension agent's space joint and was equal to 74 for the considered dataset, and T is the total length of time for each given sample.

To summarize, the proposed pipeline relies upon the following inputs:

- Text is embedded into vectors via Word2Vec [2];
- Audio is embedded using MFCC, mel-spectograms, prosody, and WavLM's features [17];
- Interlocutor's ID, following [33];
- Gestures are represented via exponential map rotations.

2.4. *VQVAE for Representation of Short-Term Gestures*

Following other approaches in the literature [20–22], we employed a VQVAE architecture to learn a compact representation of gestures for both the main agent and interlocutor, as shown in Figure 3. To this end, the codebook C, representing an unbounded discrete latent space, was populated with n atomic actions that represent short-term iconic motions that humans are used to deploy in longer sequences based on verbal communication, namely $C = \{c_1, c_2, \ldots, c_n\}$ where the element c_j denotes the j-th atomic action in the latent space in the codebook.

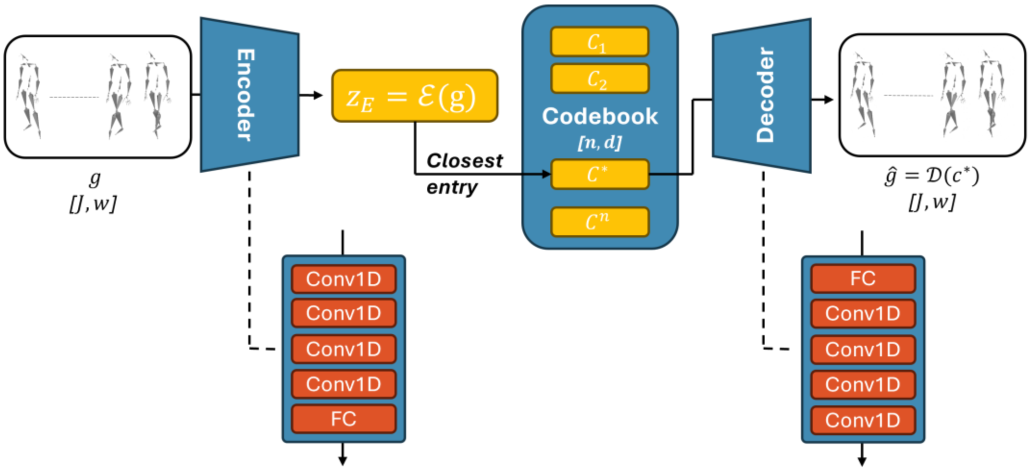

Figure 3. VQVAE architecture used to learn meaningful atomic gestures from raw motion signals. Each portion of a gesture $g \in R^{J \times w}$ is mapped onto its latent representation $z_E(g) = E(g) \in R^d$ via learned encoder. A decoder is then used to project the same gesture embedding back to the agent's joint space as $\hat{g} = D(z_E(g)) \in R^{J \times w}$. Parameters are defined in Table 1. The structure of the encoder and the decoder is defined in Table 2.

The architecture includes an encoder E and a decoder D, both composed of 1-D convolutional layers (Table 2) to successfully achieve time compression; therefore, each single element represents a given time window in the former space. With reference to the parameters in Table 1, the encoder E is trained to learn a function $E : R^{J \times w} \rightarrow R^d$ mapping from the agent's joint space to the latent embedding space of gesture. The decoder D is trained to learn a function $D : R^d \rightarrow R^{J \times w}$ mapping from the latent gesture space back to the agent's joint space. Specifically, as per [34], the input to the encoder is a portion of a gesture $g \in \mathbb{R}^{J\times w}$, which is mapped onto its latent vector $z_E(g) = E(g)$. The latter is then associated with the closest codebook's entry c^* such that

$$z_E = c^*, \quad \text{where} \quad c^* = c_k, \quad k = \underset{j}{\operatorname{argmin}} \|z_E(g) - c_j\|^2, \quad j = 1, \ldots, n. \tag{2}$$

Afterwards, the prediction $\hat{g} = D(c^*)$ is obtained by applying decoder D over the chosen codebook closest entry. The functions $E(\cdot)$ and $D(\cdot)$ and the entries of the codebook have been trained to minimize the mean squared error (MSE) between the reconstructed gesture $\hat{g}$ and the input g over the training set (presented in Section 2.6). Moreover, codebook entries are pushed towards the closest entries produced by the encoder E, while we adopted an update rule based on exponential moving average over not previously used codes to tackle the bottleneck problem, which commonly affects training procedures in vector-quantized architectures [35].

Table 2. VQVAE architecture.

Component	Layer	Type	Hyperparameters [1]
Encoder E	Conv1	Conv1d	input_dim, hidden_dim, kernel_size=3, padding=1
	Conv2	Conv1d	hidden_dim, hidden_dim, kernel_size=3, padding=1
	Conv3	Conv1d	hidden_dim, hidden_dim, kernel_size=3, padding=1
	Conv4	Conv1d	hidden_dim, hidden_dim, kernel_size=3, padding=1
	FC	Linear	hidden_dim × max_frames, embedding_dim
Vector quantizer VQ	Codebook	Embedding	embedding_num, embedding_dim
Decoder D	FC	Linear	embedding_dim, hidden_dim × max_frames
	Conv1	Conv1d	hidden_dim, hidden_dim, kernel_size=3, padding=1
	Conv2	Conv1d	hidden_dim, hidden_dim, kernel_size=3, padding=1
	Conv3	Conv1d	hidden_dim, hidden_dim, kernel_size=3, padding=1
	Conv4	Conv1d	hidden_dim, output_dim, kernel_size=3, padding=1

[1] **Hyperparameters used:** embedding_dim = 256, embedding_num = 2048, hidden_dim = 512, input_dim = output_dim = 74, and max_frames = 18

2.5. *Diffusion Model for Representation of Long-Term Conversational Dynamics*

Inspired by the excellent results proposed by [25,26], our aim was to leverage the latter works extending conditional inputs to a dyadic setup. At the same time, we integrated the diffusion model latent space with the one obtained by VQVAE architecture in Figure 3 to enable faster training and inference procedures.

Probabilistic denoising diffusion models represent Markov chains, where two different algorithms, for the diffusion and denoising steps, alternate. Each step is described as a probability distribution over the previous step. Specifically, the diffusion procedure predicts, from a Gaussian distribution, the amount of noise to be added to a specific set of input features, thus moving from fine-grained ground-truth information to gross-grained, corrupted, and final information. The procedure can be described in Equation (3):

$$\underset{t:1\rightarrow T}{q(x_t|x_{t-1})} = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{3}$$

where β_t is a fixed schedule following a cosine progression used to modulate the amount of noise to be added at a given step, normally yielding better results than linear progression. Conversely, the denoising procedure uses a parametrized probability distribution to predict the amount of noise to be removed from a noisy input in order to move towards a fine-grained output. In formal terms, the procedure can be expressed as

$$\underset{t:T\rightarrow 1}{p_\theta(x_{t-1}|x_t)} = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{4}$$

where μ_θ and Σ_θ, respectively, represent the parametrized mean and variance that need to be optimized in order to predict the amount of noise to be subtracted from the previous step x_t [36].

In this specific context, we aim to learn the denoising procedure p_θ for gesture g represented in the latent space from external conditions, such as conversation and contextual information. Following [25,26], the denoising procedure p_θ is learned via conditioning on multimodal conversational inputs.

The entire procedure is shown in Figure 4. Considering the parameters introduced in Table 1, we employ the diffusion model with verbal and nonverbal conditioning inputs, considered over different time windows, as discussed in Section 2.1. The denoising procedure p_θ is approximated via a deep neural network architecture. Indeed, a cross-local attention layer is used to compute attention over the heterogeneous input signal, concatenating gesture and conditioning information, as per Equation (5), according to [37]:

$$Attention(Q, K, V) = so f tmax \frac{QK^T}{\sqrt{d_k}} V \tag{5}$$

where Q, K, and V, respectively, represent queries, keys, and values computed via an attention mechanism. The output of the latter layer is fed to a transformer encoder [37] used to predict the next sequence of gestures. Rotary positional encoding is used to enhance the transformer architecture's performance [38]. The overall time horizon observed by the diffusion model is $h = l \times w$, considering the quantities reported in Table 1 for the dataset used in this work. This amounts to 144 frames, therefore representing a sequence of 4.8 s. Eventually, a series of gesture codes $\hat{g}_0^{t:t+h} = (g_1, g_2, \ldots, g_l)$ is predicted. The parameters are updated to minimize *Huber* loss over generated gestures with respect to ground-truth ones.

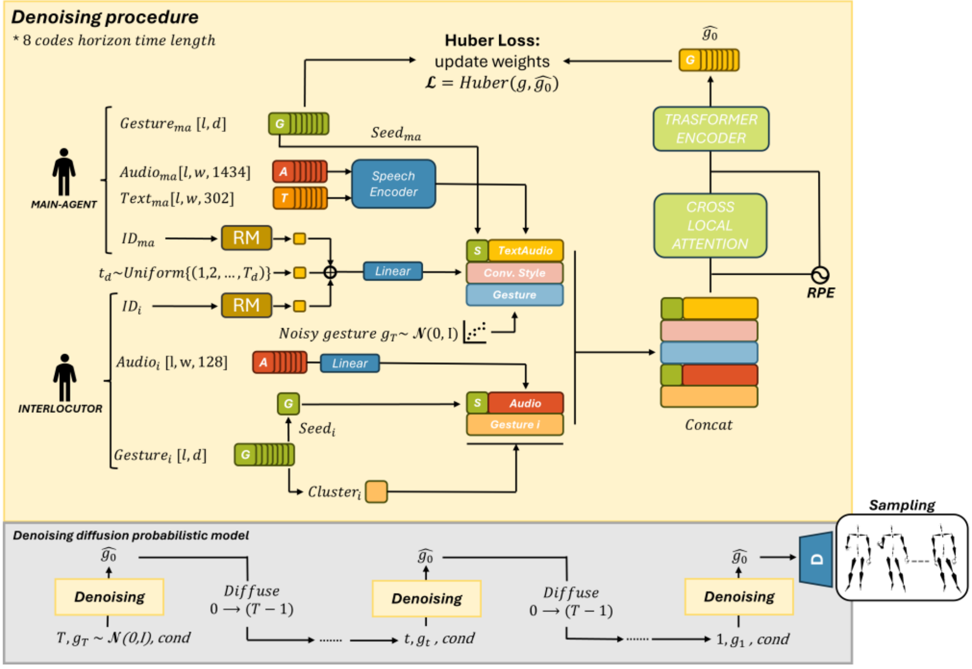

Figure 4. Schematic of the diffusion model, with a focus on the denoising procedure. At the bottom of the figure, the diffuse–denoising procedure is visualized, starting from a noisy gesture and a conditioning input such as "cond". At the end, the pipeline produces a batch of codes that are mapped onto the agent's joint space via the VQVAE decoder. On top of that, the denoising procedure is detailed with respect to every member inside the conditioning input, expressed in Section 2.5.

In particular, we use the main agent's text, audio, and ID features as conditions, adding random masking (RM) at inference time. A speech encoder is used to map the main agent's conversation, composed of text and audio, into the diffusion model's latent space. The same procedure is applied to the interlocutor's audio features, which are composed of MFCC, mel-spectogram, and prosody features. The last gesture code from the interlocutor is used as a seed. Then, we expand such information by adding the interlocutor's beat motion via K-Means clustering over VQVAE's previously learned codebook of gestures.

Indeed, we extracted K (Table 1) centroids from the gestures' codebook to be representative of the most common iconic motion usually exhibited during a conversation, as per Figure 5. A subset $C_c = \{c^1_o\, c^2_o \ldots, c^K_c\}$ is obtained from C; therefore, each cluster entry c^i_c has been mapped to its closest entry in C. These cluster entries are used in the diffusion model's conditioning information to identify interesting fine-grained patterns from the last interlocutor gesture. As a result, the closest centroids is used to enhance the mirroring and backchanneling of the partner's motion.

The conditioning input, namely $cond_{denoise}$, can be represented as in Equation (6):

$$cond_{denoise} = concat(T^{t:t+h}_{ma}, A^{t:t+h}_{ma}, G^{t-w:t}_{ma}, ID_{ma}, A^{t-h:t}_{i}, G^{t-w:t}_{i}, ID_i) \tag{6}$$

The overall procedure involved in the diffusion model training is summarized as pseudocode in Algorithm 1.

Algorithm 1 Diffusion model training procedure. The diffusion model is trained while leveraging all the conditioning information from a given dyadic conversation. Pre-trained models such as Wav_lm, Crawl_300D and the VQVAE are used to produce meaningful embedded representations of former conversational information.

```
Require: $audio, text, gesture\_interlocutor$
Require: $VQVAE\_model, diffusion\_model, Wav\_lm\_model, Crawl\_model$
$Audio_i \leftarrow FeatureEmbedding(Wav\_lm, audio\_interlocutor)$
$Text_{ma}, Audio_{ma} \leftarrow FeatureEmbedding(Wav\_lm, Crawl, audio\_conv\_agent, text\_conv\_agent)$
$Gesture_i \leftarrow VQVAE.Encoder(Gesture_i)$
$ID_{ma}, ID_i \leftarrow load\_ids()$
$cond \leftarrow concat(Gesture_{ma}, Audio_{ma}, Text_{ma}, ID_{ma})$
for $e = 1$ to $Total\_epochs$ do
    for $g = 1$ to $GestureData\_length$ do
        for $t = 1$ to $T$ do
            $g_t \leftarrow gaussian\_diffusion(t_d, g_{t\ \_1})$
        end for
        for $t = T$ to $1$ do
            $\hat{g}_{t-1} \leftarrow denoise(t_d, cond, \hat{g}_t)$
        end for
        $\hat{g}_0 \leftarrow VQVAE.Decoder(\hat{g}_0)$
        $L = HuberLoss(g, \hat{g}_0)$
        $update\_weights(L)$
    end for
end for
```

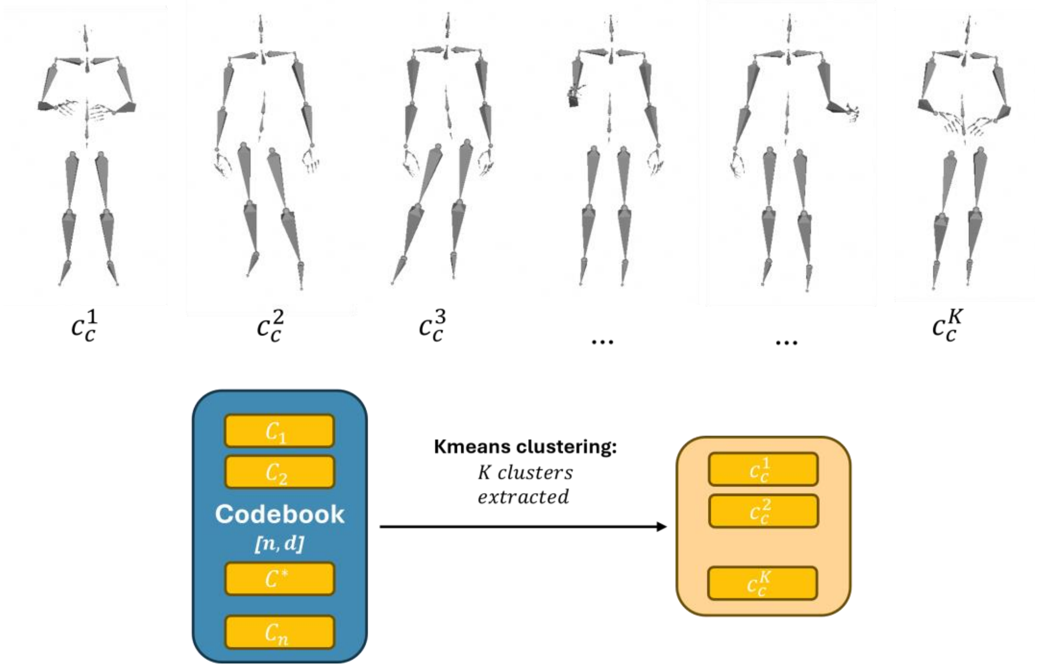

Figure 5. K centroids are extracted, via K-Means clustering, from the VQVAE-learned codebook of gestures. The interlocutor's last code of gesture is mapped to the closest cluster, thus helping the model with backchanneling and mimicking the motion of the counterpart.

2.6. Dataset

To implement the proposed architecture, we considered the GENEA2023 challenge dataset [5], which builds upon the TalkingWithHands 16.2 M (TWH) dataset to train our pipeline [12]. This dataset provides conversational audio paired with time-synchronized text transcription and body motion capture data from both the main agent (i.e., the agent to be generated) and the interlocutor. Two participants are speaking about daily life topics in an informal setup, therefore exchanging comments, past experiences, and personal points of view. The total length of the dataset is equal to 16.2 M frames, recorded at a frame rate of 30 frames per second. The dataset is provided in three splits: training, validation, and testing, composed of 372 samples, 40 samples, and 72 samples, respectively, as per Table 3. Metadata are also provided and include participants' IDs, which are used as a control input for gesture style at inference time.

Table 3. Description of the GENEA Challenge 2023 [5] dataset, the composition of which is derived from [12].

Split	Number of Samples	Main Agent	Interlocutor
Training	372	Text, Audio, Gesture	Text, Audio, Gesture
Validation	40	Text, Audio, Gesture	Text, Audio, Gesture
Testing	72	Text, Audio	Text, Audio, Gesture

We decided to take advantage of the dyadic setup that is present in the dataset to double the data at our disposal when training the model. In other words, the agent and interlocutor were interchanged during the training procedure based on their status, either speaker or listener; thus, we obtained double the observations about the existing correlations between verbal and nonverbal communication channels. This method was employed to learn atomic gestures via VQVAE's training. We derived a smaller dataset of gestures composed of episodes with lengths ranging between 9 and 18 frames representative of beat motions such as raising a hand, which was denoted as G_{VQVAE} and was selected according to audio and text features. Raw audio and text information from the conversation is preprocessed via pre-trained architectures, as explained in Section 2.3.

We used the training split during training procedures, whereas the testing split does not provide the main agent's ground-truth gesture, so we used the validation split to run experimental validation. The dataset was originally provided with agents facing to each other; thus, we introduced a spatial transformation of the interlocutor gesture's root-joint data to obtain a motion representation equivalent to the main agent. In this way, we could effectively learn atomic gesture actions from both agents while interchanging the target agent.

2.7. *Hardware and Software Architectures*

The proposed pipeline, and relative baselines and ablations, have been trained on an NVIDIA RTX A5000 (24 GB) GPU leveraging CUDA version 11.7 paired with an Intel(R) Core(TM) i9-7900X @ 3.30 GHz CPU. It is worth noting that we were also able to run inference procedures on an entry-level laptop GPU such as the NVIDIA GeForce MX330. Results were visually assessed via Blender version 4.0.2.

Using this architecture, we trained the VQVAE with a stop-loss procedure between the training and validation splits. The process lasted approximately 43 h, reaching a total number of 192 epochs with 16,512 iterations using a batch size equal to 1024. During the diffusion model's training procedures, we used a batch size equal to 360.

We compared the proposed dyadic architecture to its derived monadic setup and a dyadic ablation study. Specifically, we excluded the interlocutor's information as blocks, as shown in Table 4:

1. *Monadic* model configuration does not include WavLM features from the main agent; thus, only MFCC, mel-spectrogram, and prosody features are used. The interlocutor's information is discarded.
2. *Convi* model uses the same dyadic approach; nonetheless, the interlocutor's gesture is discarded as this is an ablation study of the latter in which we only explored the usage of the interlocutor's conversation.
3. *Dyadic (TAG2G)* model configuration, which is the proposed approach, uses all conditioning inputs, as per Section 2.5.

Table 4. Training details of pipelines designed during experimental validation.

Model	Main Agent's Data [1]	Interlocutor's Audio	Interlocutor's Gesture	Total Iterations	Training Time
Monadic [2]	Yes	No	No	3 M	43 h
Convi [3]	Yes	Yes	No	500 K	40 h
Dyadic (TAG2G)	Yes	Yes	Yes	500 K	41 h

[1] We refer to complete multimodal data as text, audio, and past gestures as a seed. [2] Monadic configuration does not include WavLM features; thus, only MFCC, mel-spectrogram, and prosody are used. All interlocutors have been discarded. [3] Ablation study from dyadic setup, only conversational inputs are taken from the interlocutor.

2.8. *Evaluation Metrics*

Gesture generation, given its N-to-N formulation, is intrinsically typified by the difficulty of determining whether a given motion is appropriate or not, and which is the most appropriate. Difficulties arise when looking for analytical metrics highly correlated to possible qualitative human evaluation results. In particular, such metrics need to be computed in an agent's joint space, in which the semantic meaning of each configuration is not provided. Thus, computing analytical metrics between ground-truth (considered as the natural motion from the dataset) and generated gestures is trivial. Conversely, human qualitative evaluation via questionnaires is widely employed in challenges and comparisons [5,14] even though it is very expensive to plan and execute, requiring a vast amount of resources and time.

To address this problem, we proposed a quantitative evaluation of generated gestures by following three different dimensions, each corresponding to a specific research question (RQ), similarly to [5]. Our aim was to obtain a multifaceted evaluation similar to the qualitative ones, however applying analytical metrics. The considered evaluation dimensions were then mapped onto a specific research question as follows:

RQ1 Is the quality of generated gestures close to natural motions in terms of human-likeness?
RQ2 Are the generated gestures reflecting the ground truth from the main agent, thus appropriate for the conversation?
RQ3 Are the generated gestures appropriate for the interlocutor's nonverbal behavior, thus to the conversation?

In order to answer the first question (RQ1), we computed acceleration and jerk to estimate the quality of synthesized gestures, as this is the most common approach in these terms. In fact, these physical quantities are often associated with smoothness of motion: the lower these metrics are, the smoother, and, hence, more natural, can the gesture be considered. Results were normalized over the space covered during the motions, to avoid that the optimal motion may be a static one, as per Equations (7) and (8):

$$Acc = \sum_{j=0}^{J} \frac{\ddot{G}^j}{G^j} \tag{7}$$

$$Jerk = \sum_{j=0}^{J} \frac{\dddot{G}^j}{G^j} \tag{8}$$

where J is the total number of joints, while $\ddot{G}^j$ and $\dddot{G}^j$ are, accordingly, acceleration and jerk computed over a given gesture G.

In order to answer the second question (RQ2), we considered the following metrics, where we denoted by G_{ma} the natural gesture, considered as the ground truth, and by $\hat{G}_{\mathrm{ma}}$ the synthetic gesture:

- Average Position Error (*APE*, Equation (9)), to measure the distance between the ground-truth and the generated gesture

$$APE\left(G_{ma}, \hat{G}_{ma}\right) = \sum_{j=0}^{J} \left\| G^j_{ma} - \hat{G}^j_{ma} \right\|_2 \tag{9}$$

where J is the total number of joints;

- Frechét distance for the gesture (*FDG*, Equation (10)), as it has been shown to measure semantic correlation between the two members, as in [39]

$$FDG = FDG\left(G_{ma}, \hat{G}_{ma}\right); \tag{10}$$

- Dynamic Time Warping (*S_DTW*, Equation (11)), to compute similarity between the ground-truth and generated gestures, as suggested in [31]

$$S\_DTW = 1 - \frac{DTW\left(G_{ma}, \hat{G}_{ma}\right)}{max(DTW)}; \tag{11}$$

- Covariance Similarity (*S_COV*, Equation (13)), to compute similarity based on the covariance description in Equation (12), as per [40,41]:

$$C_l = \frac{1}{T-1} \sum_{t=0}^{T} (\mathbf{G}^t_l - \overline{\mathbf{G}}_l)(\mathbf{G}^t_l - \overline{\mathbf{G}}_l)^\top, \quad l \in \{ma, i\} \tag{12}$$

$$S\_COV(C_{ma}, C_i) = \frac{C_{ma} \cdot C_i}{||C_{ma}|| \, ||C_i||} \tag{13}$$

where $\overline{\mathbf{G}_l}$ is the sample mean of G_l computed over the time window length T and $\mathbf{G}_l^t$ refers to the values for each joint at time step t.

The last question (RQ3) was addressed via the computation of *FDG* and previously introduced similarity scores *S_DTW* and *S_COV* of the main agent's generated gesture with respect to the interlocutor's gesture. In this scenario, it was helpful to also compare the natural motions as a baseline.

3. Results

In Table 5, the results of acceleration and jerk are provided in order to address RQ1. Natural motion is used as a baseline, against which the other models are compared. Natural motion significantly outperforms the generative methods (p-value = 0.002837 and p-value = 0.002384 when performing an unpaired t-test). On the other hand, no signifi- cantly different results are reported between different generative methods with respect to acceleration and jerk (respectively, p-value = 0.2838, and p-value = 0.05426 when performing an unpaired t-test at $\rho < 0.05$). This results from the fact that they share the same VQVAE model. Our findings indicate that the learned gesture in the latent space that is exploited by the diffusion model generation yields the same quality of motion at inference time, regardless of which model is used.

Table 5. Gesture's human-likeness evaluation via smoothness assessment in reply to RQ1.

Model	Acc [$\frac{deg}{s^2}$] ↓	Jerk [$\frac{deg}{s^3}$] ↓
Natural [1]	**0.705 ± 0.068**	**1.060 ± 0.140**
Monadic	1.401 ± 0.040 ⋆	2.279 ± 0.147 ⋆
Convi [2]	1.374 ± 0.039	2.438 ± 0.113
Dyadic (TAG2G)	1.373 ± 0.037	2.326 ± 0.197

[1] Natural refers to data from real scenes.[2] Ablation of interlocutor's gesture from dyadic model. ⋆ Statistically significant degradation with respect to previous implementation (shown in previous row).

Furthermore, in Table 6, we present the results obtained while answering RQ2 regarding the ground-truth appropriateness of the generated gesture. In this scenario, the monadic model significantly outperforms the others that take multimodal inputs from both agents (p-value = 0.004798 according to an unpaired t-test). Contrarily, no significant difference is reported (p-value = 0.3589) based on an unpaired t-test between *Convi* and *Dyadic* implementations when evaluating the appropriateness of the gestures generated for the main agent. Nonetheless, since any natural baseline cannot be computed, it is difficult to determine the magnitude of the distance between the proposed models. Consequently, we can only state that the *Monadic* model exhibits a similarity to the ground truth higher than other compared methods.

Table 6. Comparison with main agent's ground truth in reply to RQ2.

Model	APE ↓	FDG ↓	S_DTW ↑	S_COV ↑
Monadic	**3.059 ± 1.406**	**53.099 ± 20.281**	**0.317 ± 0.144**	**0.646 ± 0.158**
Convi [1]	4.285 ± 1.269	60.220 ± 20.304 ⋆	0.190 ± 0.101 ⋆	0.399 ± 0.168 ⋆
Dyadic (TAG2G)	4.759 ± 1.032	60.531 ± 20.421	0.206 ± 0.087	0.327 ± 0.157

[1] Ablation of interlocutor's gesture from dyadic model. ⋆ Statistically significant degradation with respect to previous implementation (shown in previous row)

Finally, in Table 7, both natural and generated gestures from the main agent are compared to the interlocutor's ones, according to RQ3. As highlighted by these results, the proposed model TAG2G significantly outperforms any other proposed method in terms of *FDG* and *DTW*-based similarity (respectively, p-value = 0.002589, and p-value = 0.003578 when performing an unpaired t-test). While natural motion exhibits a slightly better result in terms of covariance-based similarity with respect to the *Monadic* and *Convi* implementations, no significant difference is present with respect to the *Dyadic* method (p-value = 0.49157 when performing an unpaired *t*-test). Incorporating both agents' information leads to a more comprehensive understanding of contextual information, resulting in an increased connection and similarity with interlocutor nonverbal behavior.

Table 7. Comparison to interlocutor's gesture in response to RQ3.

Model	FDG ↓	S_DTW ↑	S_COV ↑
Natural [1]	80.288 ± 29.779	0.200 ± 0.114	**0.418 ± 0.197**
Monadic	65.640 ± 33.038 ◇	0.303 ± 0.111 ◇	0.336 ± 0.171 ⋆
Convi [2]	60.510 ± 37.401	0.407 ± 0.138	0.384 ± 0.178 ◇
Dyadic (TAG2G)	**55.516 ± 34.139** ◇	**0.432 ± 0.112** ◇	0.410 ± 0.169

[1] Natural refers to motion-captured data from real scenes. [2] Ablation of interlocutor's gesture from dyadic model.

⋆, ◇ Statistically significant degradation or improvement with respect to previous implementation (shown at previous row).

4. Discussion

As highlighted by the results concerning RQ1 (Table 5), the same VQVAE model is shared across all the considered methods. Thus, the results yielded by the models are approximately converging. Consequently, these measures seem to be suggesting that the codebook of gestures is actually acting as a bottleneck when looking to generate more natural-like motions. Although atomic gestures learned by the VQVAE result in high quality movements once decoded in agent's joint space, a vast amount of fine-grained motion information is lost between two different short-term windows when the gesture is encoded. Thus, this information is not observed by the diffusion model, eventually resulting in a sequence of fine-grained atomic actions interconnected by gross-grained motion transitions in between two different atomic gestures.

However, as highlighted by training times in Table 4, incorporating VQVAE latent space into the diffusion model leads to a substantial speedup in training and inference times. It is worth noting that the monadic configuration has up to five times shorter training time than current state-of-the-art implementations employing a diffusion model architecture. As a consequence, we can include the interlocutor's information while maintaining the same training efficiency as above-mentioned implementations.

According to the evaluation of the generated results with respect to ground-truth gestures, as per RQ2 in Table 6, the monadic implementation yields higher similarity to natural gestures than dyadic ones. Firstly, *APE* scores highly rely on positional error computation, which is semantically incoherent. As an example, two symmetrical poses, such as either raising the left arm or the right arm, would yield an error, while most of the times they are semantically interchangeable. Nevertheless, monadic implementation also outperforms dyadic approaches in terms of *FDG* and similarity scores. These metrics can be considered more robust than *APE*, thus underlying the quality yielded by focusing only on the main agent information when looking for an increased appropriateness for the main agent. Our findings strongly align with those reported in [5,27], confirming that diffusion model generation conditioned on conversation produces accurate results when compared to the ground truth. However, a qualitative study is required, as future work, to assess how close our results are to natural behavior exhibited by human participants.

From a wider perspective, our findings resemble the ones presented in [5], where most approaches that perform best in terms of main-agent appropriateness and human-likeness yield above-chance results in listener appropriateness evaluation. More specifically, the monadic baseline outperformed the dyadic baseline in terms of appropriateness to the main agent, while the latter outperformed the former regarding appropriateness to the interlocutor. Similar results were obtained in our study regarding RQ3, as summarized in Table 7. The more the interlocutor's information was introduced into conditioning information during the generation procedure, the better the obtained results were with respect to interlocutor appropriateness. Our proposed method, TAG2G, outperformed all other methodologies, natural baseline included. An on-par performance was only observed in terms of covariance-based similarity, while significant differences were observed in terms of *FDG*- and *DTW*-based similarity metrics.

We conjecture that a model trained to receive and process the interlocutor's information at every frame naturally outperforms human capabilities of focusing on their conversational partner during real-time interaction. Indeed, it is non-trivial for a human to deeply focus, as compared to a neural network's capabilities, on the interlocutor's state during a conversation. Moreover, we do not have any information about the level of confidence perceived by participants while taking part in the recordings. Thus, we cannot be certain that they performed as well and as naturally as they would have with previously known people in a more familiar situation. Furthermore, the dataset's authors in [12] verified that high interpersonal covariance scores related to gestures were observed during statistical studies over the dataset, consistent with our findings and confirming the quality of the evaluation procedure employed.

In addition, a subjective evaluation of the proposed methodology, along with its derived approaches used as benchmarks would help to better understand the observed performance. In these regards, the lack of a user study acts as the current limitation of the proposed work. However, the evaluation procedure employed in this study is a collection of established metrics vastly used in the gesture evaluation domain, whose validity is supported by a plenitude of works from the literature [31,39–41].

Nonetheless, our findings are highly in line with previously reported results in [5,12], where top-tier methodologies in terms of human-likeness and appropriateness to the main agent eventually perform below average in terms of appropriateness to the interlocutor. Conversely, an interlocutor-aware approach yields better results in terms of appropriateness to the interlocutor, while below-average performances are found in terms of appropriateness to the main agent. Similar results are observed with the evaluation metrics provided in this work regarding the three different approaches.

In these terms, our results suggest that methodologies such as the one proposed in [30], where two distinct architectures were trained to differentiate between listening and speaking actions, should be investigated. Indeed, the possibility of introducing a mixed approach composed of tailored models for the agent's status, either speaker or listener, remains an open topic that is largely unexplored. Consequently, this direction should be strongly considered.

5. Conclusions and Future Works

In this paper, we proposed a model composed of a diffusion model coupled with a VQVAE to tackle the co-speech gesture generation problem in a dyadic interaction setup. The pipeline takes inputs from both the agents in order to better understand the topic of the verbal communication and the context in which the interaction is established.

We evaluated our method over three significant different dimensions of nonverbal behavior. Firstly, we focused on the quality of movement itself; then, we assessed the appropriateness of the generated gesture when compared to ground-truth motion and interlocutor motion. An ablation study was proposed in order to measure the impact that each block of information has on the results. These studies provide empirical evidence that focusing on both agents and employing a dyadic approach yield the best performance, when evaluating the appropriateness with respect to the interlocutor's gestures. These capabilities should be exploited when the main agent is either listening or involved in a fast-paced conversation where roles (speaker and listener) are repeatedly interchanged. A monadic approach, on the other hand, yields better results in terms of appropriateness when generated gestures are compared to ground-truth motion. These results suggest that focusing on conversational information yields a higher similarity to natural nonverbal behavior, but a total lack of contextual understanding with respect to the interlocutor.

Various potential directions have remained unexplored to date. For instance, a mixed generative approach relying on both the proposed methods should be exploited as a future work. Consequently, we aim to extend this research idea to a mixed approach including both monadic and dyadic methodologies, dynamically switching between the two with respect to the current state of the agent in the conversation, thus either speaking or listening.

References

1. Nyatsanga, S.; Kucherenko, T.; Ahuja, C.; Henter, G.E.; Neff, M. A Comprehensive Review of Data-Driven Co-Speech Gesture Generation. In *Computer Graphics Forum*; Wiley Online Library: Hoboken, NJ, USA, 2023; Volume 42, pp. 569–596.
2. Zeyer, A.; Dillon, J. The role of empathy for learning in complex Science| Environment| Health contexts. *Int. J. Sci. Educ.* **2019**, *41*, 297–315. [CrossRef]
3. Bambaeeroo, F.; Shokrpour, N. The impact of the teachers' non-verbal communication on success in teaching. *J. Adv. Med. Educ. Prof.* **2017**, *5*, 51. [PubMed]
4. Makransky, G.; Petersen, G.B. The cognitive affective model of immersive learning (CAMIL): A theoretical research-based model of learning in immersive virtual reality. *Educ. Psychol. Rev.* **2021**, *33*, 937–958. [CrossRef]
5. Kucherenko, T.; Nagy, R.; Yoon, Y.; Woo, J.; Nikolov, T.; Tsakov, M.; Henter, G.E. The GENEA Challenge 2023: A large-scale evaluation of gesture generation models in monadic and dyadic settings. In Proceedings of the 25th International Conference on Multimodal Interaction, Paris France, 9–13 October 2023; pp. 792–801.
6. Marin Vargas, A.; Cominelli, L.; Dell'Orletta, F.; Scilingo, E.P. Verbal communication in robotics: A study on salient terms, research fields and trends in the last decades based on a computational linguistic analysis. *Front. Comput. Sci.* **2021**, *2*, 591164. [CrossRef]
7. Mahmood, A.; Wang, J.; Yao, B.; Wang, D.; Huang, C.M. LLM-Powered Conversational Voice Assistants: Interaction Patterns, Opportunities, Challenges, and Design Guidelines. *arXiv* **2023**, arXiv:2309.13879.
8. Cassell, J.; Pelachaud, C.; Badler, N.; Steedman, M.; Achorn, B.; Becket, T.; Douville, B.; Prevost, S.; Stone, M. Animated conversation: Rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques, New York, NY, USA, 24 July 1994; pp. 413–420.
9. Cassell, J. A framework for gesture generation and interpretation. In *Computer Vision in Human-Machine Interaction*; Cambridge University Press: Cambridge, MA, USA, 1998; pp. 191–215.
10. Kopp, S.; Krenn, B.; Marsella, S.; Marshall, A.N.; Pelachaud, C.; Pirker, H.; Thórisson, K.R.; Vilhjálmsson, H. Towards a common framework for multimodal generation: The behavior markup language. In Proceedings of the Intelligent Virtual Agents: 6th International Conference, IVA 2006, Marina Del Rey, CA, USA, 21–23 August 2006; Springer: Berlin/Heidelberg, Germany, 2006; pp. 205–217.
11. Liu, H.; Zhu, Z.; Iwamoto, N.; Peng, Y.; Li, Z.; Zhou, Y.; Bozkurt, E.; Zheng, B. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2022; pp. 612–630.
12. Lee, G.; Deng, Z.; Ma, S.; Shiratori, T.; Srinivasa, S.S.; Sheikh, Y. Talking with hands 16.2 m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 763–772.
13. Kucherenko, T.; Jonell, P.; Yoon, Y.; Wolfert, P.; Henter, G.E. A large, crowdsourced evaluation of gesture generation systems on common data: The GENEA Challenge 2020. In Proceedings of the 26th International Conference on Intelligent User Interfaces, College Station, TX, USA, 14–17 April 2021; pp. 11–21.
14. Yoon, Y.; Wolfert, P.; Kucherenko, T.; Viegas, C.; Nikolov, T.; Tsakov, M.; Henter, G.E. The GENEA Challenge 2022: A large evaluation of data-driven co-speech gesture generation. In Proceedings of the 2022 International Conference on Multimodal Interaction, Bengaluru, India, 7–11 November 2022; pp. 736–747.
15. Chiu, C.C.; Morency, L.P.; Marsella, S. Predicting co-verbal gestures: A deep and temporal modeling approach. In Proceedings of the Intelligent Virtual Agents: 15th International Conference, IVA 2015, Delft, The Netherlands, 26–28 August 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 152–166.
16. Liang, Y.; Feng, Q.; Zhu, L.; Hu, L.; Pan, P.; Yang, Y. Seeg: Semantic energized co-speech gesture generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10473–10482.
17. Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 1505–1518. [CrossRef]
18. Hsu, W.N.; Bolte, B.; Tsai, Y.H.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3451–3460. [CrossRef]
19. Mikolov, T.; Grave, E.; Bojanowski, P.; Puhrsch, C.; Joulin, A. Advances in Pre-Training Distributed Word Representations. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018.
20. Ng, E.; Joo, H.; Hu, L.; Li, H.; Darrell, T.; Kanazawa, A.; Ginosar, S. Learning to listen: Modeling non-deterministic dyadic facial motion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 20395–20405.
21. Yazdian, P.J.; Chen, M.; Lim, A. Gesture2Vec: Clustering gestures using representation learning methods for co-speech gesture generation. In Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 23–27 October 2022; IEEE: New York, NY, USA, 2022; pp. 3100–3107.
22. Korzun, V.; Beloborodova, A.; Ilin, A. The FineMotion entry to the GENEA Challenge 2023: DeepPhase for conversational gestures generation. In Proceedings of the 25th International Conference on Multimodal Interaction, Paris, France, 9–13 October 2023; pp. 786–791.
23. Tuyen, N.T.V.; Celiktutan, O. Agree or disagree? Generating body gestures from affective contextual cues during dyadic interactions. In Proceedings of the 2022 31st IEEE International Conference on Robot and Human Interactive Communication

(RO-MAN), Napoli, Italy, 29 August–2 September 2022; IEEE: New York, NY, USA, 2022; pp. 1542–1547.

24. Habibie, I.; Elgharib, M.; Sarkar, K.; Abdullah, A.; Nyatsanga, S.; Neff, M.; Theobalt, C. A motion matching-based framework for controllable gesture synthesis from speech. In Proceedings of the ACM SIGGRAPH 2022 Conference Proceedings, Vancouver, BC, Canada, 7–11 August 2022; pp. 1–9.
25. Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-Or, D.; Bermano, A.H. Human motion diffusion model. *arXiv* **2022**, arXiv:2209.14916.
26. Yang, S.; Wu, Z.; Li, M.; Zhang, Z.; Hao, L.; Bao, W.; Cheng, M.; Xiao, L. Diffusestylegesture: Stylized audio-driven co-speech gesture generation with diffusion models. *arXiv* **2023**, arXiv:2305.04919.
27. Yang, S.; Wang, Z.; Wu, Z.; Li, M.; Zhang, Z.; Huang, Q.; Hao, L.; Xu, S.; Wu, X.; Yang, C.; et al. Unifiedgesture: A unified gesture synthesis model for multiple skeletons. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October 2023; pp. 1033–1044.
28. Goldin-Meadow, S. The role of gesture in communication and thinking. *Trends Cogn. Sci.* **1999**, *3*, 419–429. [CrossRef] [PubMed]
29. Tuyen, N.T.V.; Celiktutan, O. It takes two, not one: Context-aware nonverbal behaviour generation in dyadic interactions. *Adv. Robot.* **2023**, *37*, 1552–1565. [CrossRef]
30. Schmuck, V.; Tuyen, N.T.V.; Celiktutan, O. The KCL-SAIR team's entry to the GENEA Challenge 2023 Exploring Role-based Gesture Generation in Dyadic Interactions: Listener vs. Speaker. In Proceedings of the Companion Publication of the 25th International Conference on Multimodal Interaction, Paris, France, 9–13 October 2023; pp. 214–219.
31. Song, S.; Spitale, M.; Luo, Y.; Bal, B.; Gunes, H. Multiple Appropriate Facial Reaction Generation in Dyadic Interaction Settings: What, Why and How? *arXiv* **2023**, arXiv:2302.06514.
32. Mehrabian, A.; Russell, J.A. *An Approach to Environmental Psychology*; MIT Press: Cambridge, MA, USA, 1974.
33. Yoon, Y.; Cha, B.; Lee, J.H.; Jang, M.; Lee, J.; Kim, J.; Lee, G. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Trans. Graph. (TOG)* **2020**, *39*, 1–16. [CrossRef]
34. Van Den Oord, A.; Vinyals, O. Neural discrete representation learning. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017
35. Łańcucki, A.; Chorowski, J.; Sanchez, G.; Marxer, R.; Chen, N.; Dolfing, H.J.; Khurana, S.; Alumäe, T.; Laurent, A. Robust training of vector quantized bottleneck models. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; IEEE: New York, NY, USA, 2020; pp. 1–7.
36. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
37. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
38. Su, J.; Lu, Y.; Pan, S.; Murtadha, A.; Wen, B.; Roformer, Y.L. Enhanced transformer with rotary position embedding. *arXiv* **2023**, arXiv:2104.09864.
39. Maiorca, A.; Yoon, Y.; Dutoit, T. Evaluating the quality of a synthesized motion with the fréchet motion distance. In Proceedings of the ACM SIGGRAPH 2022 Posters, Vancouver, BC, Canada, 7–11 August 2022; pp. 1–2.
40. Tuyen, N.T.V.; Elibol, A.; Chong, N.Y. A gan-based approach to communicative gesture generation for social robots. In Proceedings of the 2021 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO), Tokoname, Japan, 8–10 July 2021; IEEE: New York, NY, USA, 2021; pp. 58–64.
41. Hussein, M.E.; Torki, M.; Gowayyed, M.A.; El-Saban, M. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, Beijing, China, 3–9 August 2013.