Diffusion models for virtual agent facial expression generation in Motivational interviewing

Nezih Younsi, Catherine Pelachaud, Laurence Chaby

▶ To cite this version:

Nezih Younsi, Catherine Pelachaud, Laurence Chaby. Diffusion models for virtual agent facial expression generation in Motivational interviewing. International Conference on Advanced Visual Interfaces, Jun 2024, Genoa Italy, Italy. hal-04604194

HAL Id: hal-04604194

https://hal.sorbonne-universite.fr/hal-04604194

Submitted on 6 Jun 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Diffusion models for virtual agent facial expression generation in Motivational interviewing

Nezih Younsi
ISIR - Sorbonne university
Paris, France
younsi@isir.upmc.fr

Catherine Pelachaud
CNRS - ISIR - Sorbonne university
Paris, France
Catherine.pelachaud@isir.upmc.fr

Laurence Chaby
ISIR - Paris Cité
Paris, France
Laurence.chaby@isir.upmc.fr

ABSTRACT

Motivational interviewing (MI) is a client-centered counseling style that addresses (the client) user's motivation for behavior change. In this paper, we present a behavior generation model for Socially Interactive Agents (SIA) and apply it to an SIA acting as a virtual therapist in (MI). MI defines different types of dialogue acts for therapist and client. It has been shown that therapist builds rapport with their client by adapting their verbal and nonverbal behaviors. Based on the analysis of a human-human MI dataset (AnnoMI), we found co-occurrences between facial expressions and dialogue acts for both therapist and client. Moreover, the therapist adapts their behavior to their client's behavior to favor rapport. Our behavior generation model embeds these co-occurrences as well as such behavior adaptation. To this aim, we build an observation-to-action framework based on a conditional diffusion approach trained on the AnnoMI corpus. Our model learns to generate the virtual therapist's facial expressions conditioned by MI dialogue acts and the client's nonverbal behaviors. We aim to make SIAs more effective in therapy-like interactions, by using user's behaviors in addition to contextual information (i.e. dialogue acts and nonverbal behaviors of both user and agent) to drive the SIA behavior.

CCS CONCEPTS

• **Human-centered computing**; • **Human computer interaction (HCI)**; • **HCI design and evaluation methods**; • **User models**;

KEYWORDS

Non-verbal behavior adaptation, Conditional Diffusion models, Motivational interviewing

ACM Reference Format:
Nezih Younsi, Catherine Pelachaud, and Laurence Chaby. 2024. Diffusion models for virtual agent facial expression generation in Motivational interviewing. In *Proceedings of (Conference acronym 'AVI)*. ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Adaptation in human-human interactions is a multimodal process that manifests through different levels of communication [10]. Interlocutors adjust their behaviors to each other, creating a smooth exchange that can be observed in terms of verbal behaviors [11], but also non-verbal ones [2, 24, 27]. These behavioral adjustments occur both consciously and unconsciously, serving to enhance the quality of the interaction and achieve common goals [10]. This complex interaction of verbal and non-verbal signals is essential for facilitating effective and high-quality communication. In the field of human-agent interaction, the goal is to emulate these adaptive behaviors in interactions between humans and Socially Interactive Agents (SIAs), whether they are physical (such as social robots) or virtual. Adjusting verbal and non-verbal signals in real time, similar to that observed in humans, is crucial for developing SIAs capable of facilitating natural and high-quality exchanges [1]. This adaptation goes beyond merely replicating human behaviors; it aims to enhance the interaction quality [6]. Our study specifically focuses on the generation of adaptive facial expressions for SIAs, by developing a generative model based on machine learning to produce relevant facial expressions of virtual therapists during motivational interviewing sessions with human clients.

Motivational interviewing (MI) is a client-centered communication approach designed to facilitate their motivation to change behavior [22]. By establishing an interpersonal relationship, therapists seek to optimize the quality of sessions [28][4] [12]. To measure adherence to clients' behavior changes, the Motivational Interviewing Skill Code (MISC) schema [21] is frequently used. It categorizes different aspects of therapist-client interactions, emphasizing strategies that allow therapists to effectively support clients in their change process.

We aim to base our virtual MI therapist model on the behavior of human therapists. To this end, we first analyzed the AnnoMI corpus [31], a collection of therapist-client interview videos with dialogue acts annotated according to the specific MI MISC code [22]. We then analyzed the dynamics of facial expressions between the therapist and the client, according to the behaviors annotated by MI. Following this, we constructed an architecture based on a conditional diffusion model with the goal of learning to reproduce the dynamics of the virtual therapist's facial expressions, taking MI dialogue acts and human client facial expressions as conditions.

2 RELATED WORKS

Numerous approaches and models have been designed to adapt the behaviors of SIAs to those of human interlocutors. Social signals (such as gaze, laughter, or even the impression conveyed by the agent) are considered as rewards for reinforcement learning models

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'AVI, June 03–07, 2024, Genova

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

Figure 1: Interaction setup between a virtual therapist and a human user

to adjust the behaviors of agents [1, 8, 23, 29]. Supervised learning approaches such as neural networks, Transformers, and BI-LSTMs have also been used [13, 17, 26, 30], as well as generative models like Generative Adversarial Networks (GANs) or Variational AutoEncoders (VAEs) to produce behaviors in new situations, often conditioning one behavioral modality on others [7, 15, 16]. Diffusion models, which have recently shown performance surpassing that of their predecessors in many areas (Image Generation [18], Computer Vision [9], Multimodal Modeling [3]) have also been applied to human-machine interaction [20, 25, 33], to clone human behavior or generate communicative gestures conditioned through multimodality. The results obtained by these architectures have motivated the use of a conditional diffusion model to simulate interpersonal adaptation during motivational interviews, using an Observation-Action approach to generate the agent's facial expressions in response to user behaviors.

3 HUMAN-HUMAN CORPUS ANALYSIS

The AnnoMI corpus includes 133 videos of therapist-client interactions [31], annotated according to the Motivational Interviewing Skill Code (MISC) [21] which describes specific dialogue acts for MI, including *Change talk, Sustain talk, Neutral* for the client and *Information, Question, Reflection, Advice* for the therapist. After the extraction and synchronization of Facial Action Units (AUs) [14] using OpenFace [5] with the MI annotations, an initial analysis revealed co-occurrences between certain AUs specific to each interlocutor. This led us to group the AUs into facial expression categories: *Mouth up* (AU12, AU06, and AU25), *Mouth down* (AU14 and AU15), *Nose wrinkle* (AU10 and AU09), and *Neutral*, corresponding to no activation or low intensity [32]. Then, we conducted a sequence extraction analysis that allowed us to identify co-occurrences between these facial expression categories of therapists and clients and the specific dialogue acts of MI. Notably, therapists tend to express 'mouth up' expressions, emphasizing their active role in positive support for clients during the interview. These results motivated the development of a generative model for therapist facial expressions, conditioned on the client's facial expressions and MI dialogue acts. For this purpose, the database had to be restructured into an Observation-Action format representing each speaking turn; the Observation (condition of the generative model) includes the client's facial expressions and dialogue acts, and the Action includes the therapist's facial expression categories.

4 CONDITIONAL DIFFUSION MODEL

The generative model architecture is composed of three components. The DDPM (Denoising Diffusion Probabilistic Model) scheduler [18]., the Noise estimation Model, and the conditional diffusion model. This latter encapsulates both of the two first components, to achieve the diffusion process (Noising phase) and the sampling phase (Denoising phase) that is responsible for the generation of the corresponding action given an observation as a condition. These three components are assembled following a pipeline that goes from transforming data into a latent space and then reverses the process to generate new data samples given unseen conditions.

4.0.1 Noise estimator: The noise estimator is the core module of the reverse Diffusion process responsible for the data generation. It learns to estimate the noise added by the DDPM scheduler to the target. Usually, the noise estimator in the image generation case, is a U-net architecture model, composed of successive Convolutional networks, that takes as input a noisy image, the noising time step, and outputs the noise added to the image at this corresponding noise time step. Since we are working with temporal sequences, using the U-net architecture is not appropriate. Our Noise estimator model employs transformers to process the observation sequences X, effectively capturing the complex temporal relationships and dependencies within the data. This includes multi-head attention mechanisms and positional encoding to maintain the order of events in the sequences. To predict the noise added to the target action Y, the model also takes as input the noising time step n and the observation sequence (the conditional inputs). Before inputting these contextual elements into the noise estimator, we proceed to an embedding step.

Observation sequence embedding: The observation sequence is first split into three elements, the observation sequence, the previous action target, and the turn descriptor. The observation sequence is then processed through the embedding architecture presented in 2. Every tuple of the observation is processed as follows. We have 42 possible behavior types in the dataset. Initially, behavior types undergo one-hot encoding, followed by transformation through a feed-forward neural network into 16-dimensional vectors. The starting time and duration, which are continuous values, are first standardized and normalized relative to the length of the speaking turn; they are then embedded into 16 dimension vectors via another forward network. To maintain temporal consistency among the behavior tuples that represent a speaking turn, these embedded vectors are fed into a Long Short-Term Memory (LSTM) network. This step is crucial for capturing the intra speaking turn temporal relationships between the behaviors within an observation sequence. Finally, the entire observation sequence is represented in a latent space, resulting of vectors of 32 dimensions. This representation serves as a compact and information-rich embedding of the speaking turn, ready to be processed by the next component of the pipeline. The past action is separately processed through the same embedding pipeline, with a skip on the LSTM module. The turn

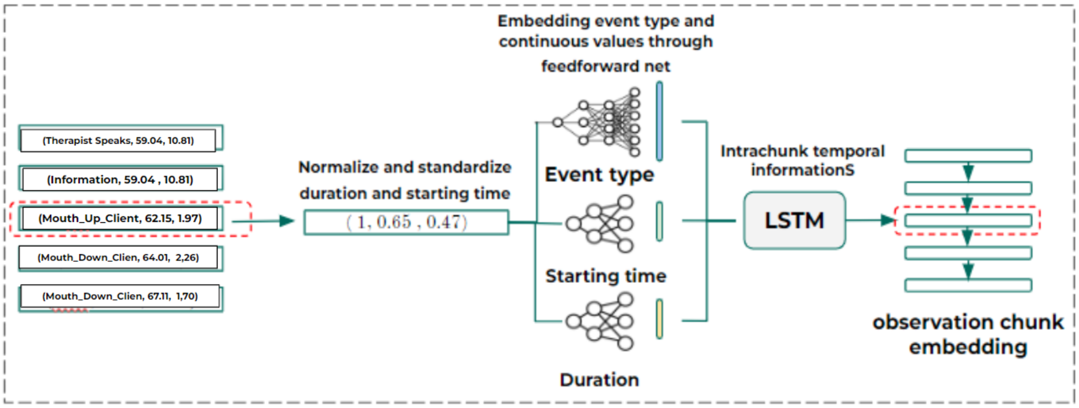
Embedding event type and continuous values through feedforward net
(Therapist Speaks, 59.04, 10.81)
(Information, 59.04 , 10.81)
(Mouth_Up_Client, 62.15, 1.97)
(Mouth_Down_Clien, 64.01, 2,26)
(Mouth_Down_Clien, 67.11, 1,70)
Normalize and standardize duration and starting time
(1, 0.65 , 0.47)
Event type
Starting time
Duration
Intrachunk temporal informationS
LSTM
observation chunk embedding

Figure 2: Observation tuple embedding architecture

descriptor is embedded using a simple feed-forward network into a 16 dimensions vector. The three embedded components are then processed using a Sequence Transformers encoder to extract key features from the observation sequence, crucial for noise estimation in diffusion modeling (see 3) . Temporal context is added through sinusoidal positional encoding. The transformer encoder then refines the sequence through layers of self-attention and feed-forward networks. Finally, the full contextual sequence is condensed into a single vector by averaging then adding a linear projection layer as presented in 3.

Noise Estimation Transformer: This component is designed to integrate and process three vectors: the single vector representation of the observation sequence, the Noisy target, and the Noise time step into a coherent noise estimate (see 3). The architecture of this transformer is composed of specialized 2 successive transformer encoder block modules. Each module is composed of a multihead attention mechanism with 16 heads, enabling the model to focus on different parts of the input sequence simultaneously, to capture a large spectrum of patterns and dependencies. First, the inputs pass through the linear layers that project them into a transformer embedding dimension, set at 64 times the number of heads (16 x 64), to match the model's internal dimensionality. This step ensures that the data is properly formatted for the self-attention operations. The self-attention mechanism within the Transformer Blocks takes the query, the key, and the value inputs that have been transformed through a linear layer. For each input, the self-attention mechanism generates three separate vectors and computes attention scores that indicate the relevance of different parts of the input data. Then, the output data of the self-attention mechanism flow through a fully connected feed-forward network. The final stage of the noise estimation process involves a linear layer that projects the multidimensional output from the transformer blocks into the action target dimension space. This represents the model's final output, providing a precise estimate of the noise distribution required for the reverse diffusion steps.

The training process of the conditional diffusion model is mainly centered on this two transformers pipeline. Once the model learns to properly estimate the added noise on a target depending on the observation sequence and the n noise steps, the reverse diffusion step responsible for the generation process can then be tested.

4.1 Sampling and generation

The sampling phase is the core of the reverse diffusion process responsible for the generation task. It uses the trained noise estimator model to gradually generate a target action starting from Gaussian noise, given an observation in the conditional case. To build our sampling architecture we draw inspiration from the sampling algorithm proposed in [18]. The algorithm starts with pure Gaussian noise, shaped in the same dimension as the target action. Then it iteratively calls the noise estimator model to estimate the noise given the denoising step, and the associated observation.

The sampling phase in diffusion models is indeed crucial, and various techniques have been developed to optimize it. We decided to use Kernel density estimation (KDE). A method where the model generates N samples given one generation condition. After, it fits a kernel-density estimator over all samples, and score the likelihood of each. Then it selects the action with the highest likelihood. This allows avoiding hallucinations and outlier generation by filtering the best likelihood samples from the bad ones.

5 MODEL TRAINING AND EVALUATION

5.1 Noise estimator Training

After restructuring and balancing our dataset, we compiled a database of 20,224 data points, each representing one Observation/Action tuple. Through extensive testing, we organized the dataset into 79 shuffled batches, each containing 256 data points. After adjustments of various hyperparameters (batch size = 128, cyclic learning rate [0.0001, 0.01], number of noise steps = 500), we recorded an MSE loss function of 0.07 on the validation dataset after 1000 training cycles.

The total number of noising steps N, is essential in diffusion models, impacting the transition from original data to Gaussian noise. Initially, with $N = 500$ we saw a loss plateau at 0.20 after 1000 epochs. Increasing this parameter showed diminishing returns in loss reduction beyond 5000 steps, where loss stabilized aroud 0.06. However, more steps mean higher computational costs, crucial for

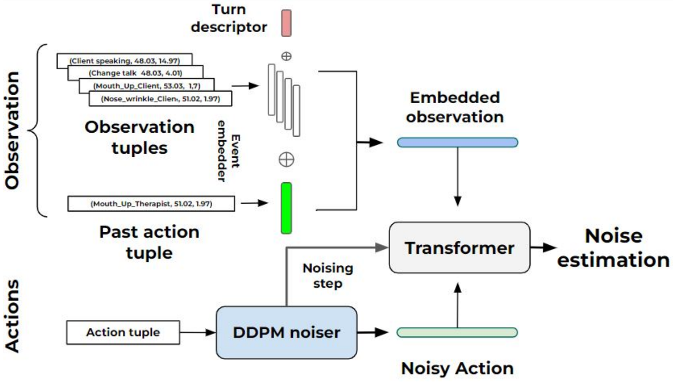

Figure 3: Noise estimation architecture

our goal of real-time deployment for a virtual therapist's behavior. Opting for a balanced approach, we set N= 2500 on trials that showed a promising compromise between low loss (training and validation losses of 0.07 and 0.08, respectively, after 1000 epochs) and computational efficiency, essential for real-time adaptive behavior generation.

5.2 Sampling and generation evaluation

In the sampling phase, we utilized the diffusion model with Kernel Density Estimation (KDE) to improve output accuracy by selecting the best sample for each data point. Testing revealed a trade-off between KDE's precision and computational efficiency. Generating 20 samples per data point balanced KDE's precision benefits with manageable computational resources and time, similar to our approach with noising steps.

Evaluating the sampling phase of a generative model can be challenging. In fields like image generation, subjective measures, including human feedback, are often used to assess the quality of generated outputs. However, in our context, where the generated targets are action vectors comprising facial expression categories, relative starting time, and duration, the evaluation needs a different approach. The first component of these action vectors is categorical, while the other two are continuous. To evaluate the model's performance, we used distinct methods for different components of the action vector. For assessing the accuracy of the facial expression category, we employed categorical model objectives measures (see 1. This allowed us determine how well the model could identify the appropriate facial expression given the observation condition. For the relative starting time and duration, which are continuous variables, we used the Root Mean Square Error (RMSE) to evaluate the model's precision in predicting these aspects.

The results, despite a relatively high validation loss of 0.08, show promising outcomes, particularly in predicting the starting time and duration of facial expression categories. The model achieved a relative average MSE of 0.012 for starting time predictions and 0.005 for duration predictions. In terms of categorizing facial expressions, the model shows a relatively strong predictive power for the *neutral* expression categories, outperforming its predictions for the three other categories.

Table 1: Facial expression categories prediction metrics

Class	Acc (%)	Precision (%)	Recall (%)	F1(%)
Mouth up	60.42	41.27	38.22	39.69
Neutral	85.87	61.44	88.24	72.44
Nose wrinkle	74.21	31.54	19.03	23.74
Mouth down	69.51	36.85	39.50	38.13

Two main factors contribute to these results. First, the data imbalance between the *mouth up* and *neutral* categories and the other categories, a trend already observed in our previous study described in section 2, affects the model's predictive capability. Second, the model's sampling and generation method involves repeatedly calling the noise estimation over N steps (here 2500) to gradually construct the data from pure noise based on the observation condition. This process suggests that prediction errors might accumulate over these steps. Therefore, maintaining a relatively small number of noising steps while striving for a significantly low loss becomes critical for enhancing model performance.

Our ongoing efforts to reduce validation loss and improve model performance include exploring techniques such as Classifier-Free Guidance [19]. In addition, our current evaluation method, which assesses each action vector component separately, provides valuable insights into what the model is learning about each element. However, it does not provide a complete picture of whether the model is effectively capturing the correlations among these components. Addressing these aspects is essential to our future work, as we aim to develop a better understanding of the model's learning process and its ability to understand the correlation between categorical components and their timings, to better replicate the intricate dynamics of facial expression changes during motivational interviewing sessions.

6 CONCLUSION AND FUTURE WORK

In this paper, we presented a conditional diffusion model architecture designed to equip virtual agents with the ability to generate adaptive facial expressions, specifically for their role as virtual therapists in motivational interviewing contexts. This work builds upon upon the analysis of the AnnoMI corpus, utilizing a machine learning approach to create a more dynamic and responsive virtual agent. We have outlined the initial results and preliminary offline architecture of our model, highlighting its real time potential in enhancing virtual counseling sessions. However, several challenges and limitations remain to be addressed. These include achieving a lower loss target, currently aimed at 0.01, resolving the lack of data and imbalance issues, reducing the number of noising time steps to facilitate real-time application, and finding more relevant evaluation techniques regarding the sampling phase. Each of these aspects is crucial for the refinement and practical deployment of our model.

Furthermore, our work represents a step toward integrating adaptive facial expressions into virtual therapists, aligning with motivational interviewing techniques and responding to user behaviors. After addressing the architecture limitations, the next step in this research is to evaluate the model's effectiveness in realistic user interaction scenarios. Testing the model in real-life scenarios will not only help validate its efficacy but also ensure its relevance and applicability in therapeutic settings.

7 ACKNOWLEDGEMENTS

This work was performed as a part of ANR-JST-DFG PANORAMA project.

REFERENCES

[1] Sean Andrist, Bilge Mutlu, and Adriana Tapus. 2015. Look like me: matching robot personality via gaze to increase motivation. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 3603–3612.

[2] Michael Argyle and Mark Cook. 1976. Gaze and mutual gaze. (1976).

[3] Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18208–18218.

[4] Zachary G Baker, Emily M Watlington, and C Raymond Knee. 2020. The role of rapport in satisfying one's basic psychological needs. *Motivation and emotion* 44 (2020), 329–343.

[5] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 59–66.

[6] Atef Ben Youssef, Mathieu Chollet, Hazaël Jones, Nicolas Sabouret, Catherine Pelachaud, and Magalie Ochs. 2015. Towards a socially adaptive virtual agent. In *Intelligent Virtual Agents: 15th International Conference, IVA 2015, Delft, The Netherlands, August 26-28, 2015, Proceedings 15*. Springer, 3–16.

[7] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. 2021. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE virtual reality and 3D user interfaces (VR)*. IEEE, 1–10.

[8] Beatrice Biancardi, Maurizio Mancini, Paul Lerner, and Catherine Pelachaud. 2019. Managing an agent's self-presentational strategies during an interaction. *Frontiers in Robotics and AI* 6 (2019), 93.

[9] Emmanuel Asiedu Brempong, Simon Kornblith, Ting Chen, Niki Parmar, Matthias Minderer, and Mohammad Norouzi. 2022. Denoising pretraining for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4175–4186.

[10] Judee K Burgoon, Lesa A Stern, and Leesa Dillman. 1995. *Interpersonal adaptation: Dyadic interaction patterns*. Cambridge University Press.

[11] Chris Cherpas. 1992. Natural language processing, pragmatics, and verbal behavior. *The Analysis of verbal behavior* 10 (1992), 135–147.

[12] Edward L Deci and Richard M Ryan. 2012. Self-determination theory. *Handbook of theories of social psychology* 1, 20 (2012), 416–436.

[13] Soumia Dermouche and Catherine Pelachaud. 2019. Engagement modeling in dyadic interaction. In *2019 international conference on multimodal interaction*. 440–445.

[14] Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior* (1978).

[15] Mireille Fares. 2020. Towards multimodal human-like characteristics and expressive visual prosody in virtual agents. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 743–747.

[16] David Greenwood, Stephen Laycock, and Iain Matthews. 2017. Predicting head pose from speech with a conditional variational autoencoder. ISCA.

[17] Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. 2018. Evaluation of speech-to-gesture generation using bi-directional LSTM network. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 79–86.

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[19] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

[20] Shivam Mehta, Siyang Wang, Simon Alexanderson, Jonas Beskow, Éva Székely, and Gustav Eje Henter. 2023. Diff-TTSG: Denoising probabilistic integrated speech and gesture synthesis. *arXiv preprint arXiv:2306.09417* (2023).

[21] William R Miller, Theresa B Moyers, Denise Ernst, and Paul Amrhein. 2003. Manual for the motivational interviewing skill code (MISC). *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico* (2003).

[22] William R Miller and Stephen Rollnick. 2012. *Motivational interviewing: Helping people change*. Guilford press.

[23] Noriaki Mitsunaga, Christian Smith, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2008. Adapting robot behavior for human–robot interaction. *IEEE Transactions on Robotics* 24, 4 (2008), 911–916.

[24] Lisette Mol, Emiel Krahmer, Alfons Maes, and Marc Swerts. 2012. Adaptation in gesture: Converging hands or converging minds? *Journal of Memory and Language* 66, 1 (2012), 249–264.

[25] Tim Pearce, Tabish Rashid, Anssi Kanervisto, Dave Bignell, Mingfei Sun, Raluca Georgescu, Sergio Valcarcel Macua, Shan Zheng Tan, Ida Momennejad, Katja Hofmann, et al. 2023. Imitating human behaviour with diffusion models. *arXiv preprint arXiv:2301.10677* (2023).

[26] Najmeh Sadoughi and Carlos Busso. 2017. Joint learning of speech-driven facial motion with bidirectional long-short term memory. In *Intelligent Virtual Agents: 17th International Conference, IVA 2017, Stockholm, Sweden, August 27-30, 2017, Proceedings 17*. Springer, 389–402.

[27] Karen L Schmidt and Jeffrey F Cohn. 2001. Human facial expressions as adaptations: Evolutionary questions in facial expression research. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists* 116, S33 (2001), 3–24.

[28] Peggy Van Minkelen, Carmen Gruson, Pleun Van Hees, Mirle Willems, Jan De Wit, Rian Aarts, Jaap Denissen, and Paul Vogt. 2020. Using self-determination theory in social robots to increase motivation in L2 word learning. In *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*. 369–377.

[29] Klaus Weber, Hannes Ritschel, Ilhan Aslan, Florian Lingenfelser, and Elisabeth André. 2018. How to shape the humor of a robot-social behavior adaptation based on reinforcement learning. In *Proceedings of the 20th ACM international conference on multimodal interaction*. 154–162.

[30] Jieyeon Woo, Catherine Pelachaud, and Catherine Achard. 2023. ASAP: Endowing Adaptation Capability to Agent in Human-Agent Interaction. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 464–475.

[31] Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023. Creation, Analysis and Evaluation of AnnoMI, a Dataset of Expert-Annotated Counselling Dialogues. *Future Internet* 15, 3 (2023), 110.

[32] Nezih Younsi, Catherine Pelachaud, and Laurence Chaby. 2024. Beyond Words: Decoding Facial Expression Dynamics in Motivational Interviewing. In *LREC-COLING*. Turin, Italy.

[33] Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. 2023. Taming Diffusion Models for Audio-Driven Co-Speech Gesture Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10544–10553.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009