2D or not 2D: How Does the Dimensionality of Gesture Representation Affect 3D Co-Speech Gesture Generation?

Téo Guichoux, Laure Soulier, Catherine Pelachaud, Nicolas Obin

2D or not 2D: How Does the Dimensionality of Gesture Representation Affect 3D Co-Speech Gesture Generation?

Téo Guichoux[1,2], Laure Soulier[1], Nicolas Obin[2], Catherine Pelachaud[1,3]
[1] Sorbonne Université, ISIR, F-75005 Paris, France ; [2] Sorbonne Université, IRCAM, Stms Lab, F-75003, Paris France ; [3] CNRS
{teo.guichoux, laure.soulier, catherine.pelachaud}@isir.upmc.fr

ABSTRACT

Co-speech gestures are fundamental for communication. The advent of recent deep learning techniques has facilitated the creation of lifelike, synchronous co-speech gestures for Embodied Conversational Agents. "In-the-wild" datasets, aggregating video content from platforms like YouTube via human pose detection technologies, provide a feasible solution by offering 2D skeletal sequences aligned with speech. Concurrent developments in lifting models enable the conversion of these 2D sequences into 3D gesture databases. However, it is important to note that the 3D poses estimated from the 2D extracted poses are, in essence, approximations of the ground-truth, which remains in the 2D domain. This distinction raises questions about the impact of gesture representation dimensionality on the quality of generated motions — a topic that, to our knowledge, remains largely unexplored. Our study examines the effect of using either 2D or 3D joint coordinates as training data on the performance of speech-to-gesture deep generative models. We employ a lifting model for converting generated 2D pose sequences into 3D and assess how gestures created directly in 3D stack up against those initially generated in 2D and then converted to 3D. We perform an objective evaluation using widely used metrics in the gesture generation field as well as a user study to qualitatively evaluate the different approaches.

CCS CONCEPTS

• **Computing methodologies** → **Shape representations**; **Neural networks**;

KEYWORDS

Co-speech gesture generation, Pose Representation, Diffusion Models, Sequence modeling

1 INTRODUCTION

In human communication, gestures play an integral role by conveying intentions and emphasizing points [40]. Recent studies [4–6, 14, 16, 42, 50, 53–56, 58] aim to create similar gestures for Embodied Conversational Agents (ECA) to make interactions with humans more natural and effective. These new methods use learning algorithms and extensive human motion datasets to generate gestures alongside speech. The representation of co-speech gestures, in 2D or 3D, influences how the agent's non-verbal communication is perceived, especially the speaker's communication style [23].

Multiple learning-based methods for gesture generation focus on generating 2D gestures [2, 16, 21, 45]. Data of 2D motion is easily accessible from "in-the-wild" monocular videos, which are videos freely accessible online, therefore allowing the gathering of large-scale collection of motion data, known as "in-the-wild" datasets. However, most of the recent literature considers 3D motion data [4–6, 14, 42, 50, 53–55, 58], primarily because such data representation contains the depth dimension and is more easily transferable to downstream applications such as 3D virtual agents or social robots [43, 56]. But, it is not easy to collect high-quality 3D motion data, as one needs a motion capture setup in a controlled environment, hence limiting the size and diversity of such datasets. To access 3D motion data and still gather large-scale datasets of diverse and spontaneous gestures, multiple works leverage an estimation of the 3D gestures inferred from 2D poses extracted from "in-the-wild" videos [39, 50, 55]. Nevertheless, to convert extracted 2D keypoints to 3D, one needs a third-party 2D-to-3D lifter, which may be prone to inaccuracies, notably because of the ambiguous nature of 3D pose estimation from 2D keypoints [44].

The co-speech gesture generation field has seen a large shift into 3D gesture modeling. However, methods trained on 3D "in-the-wild" datasets still face the bottleneck of the 2D body pose representation. In "in-the-wild" datasets lifted in 3D, the source is still 2D monocular videos. Thus, the ground-truth gestures are in 2D. Deciding whether to convert 2D poses to 3D beforehand or to train a model to generate 2D gestures and then lift them to 3D in post-processing, needs to be addressed. It remains unclear what the impact of the dimensionality of the skeleton representation is on the training of the generative model and the quality of the generated gestures. To the best of our knowledge, this question has never been extensively studied.

In this work, we compare two training settings to evaluate the influence of data dimensionality on the performance of two speech-to-gesture generative models by either considering 2D or 3D joint coordinates. We use a convolutional neural network [44] to subsequently convert generated 2D gestures to 3D. We chose to evaluate the effect of the pose representation on **1)** a Denoising Diffusion Probabilistic Model (DDPM) [25, 47, 58] and **2)** a recurrent neural generative model [55]. Both approaches have proven their ability to generate natural and diverse gestures aligned with speech.

There is a one-to-many relationship between 2D keypoints and their 3D counterparts. Given the deterministic nature of the 2D-to-3D lifter, it will consistently map any given 2D pose to the same corresponding 3D pose introducing an inductive bias in the process. We aim to measure whether the gesture distribution resulting from lifting will be less human-like, diverse, and in sync with speech than the distribution of 3D gestures generated via a model trained to directly generate 3D gestures. More particularly, our contributions are the following:

- We propose an evaluation pipeline to investigate the impact of the dimensionality of the pose representation on

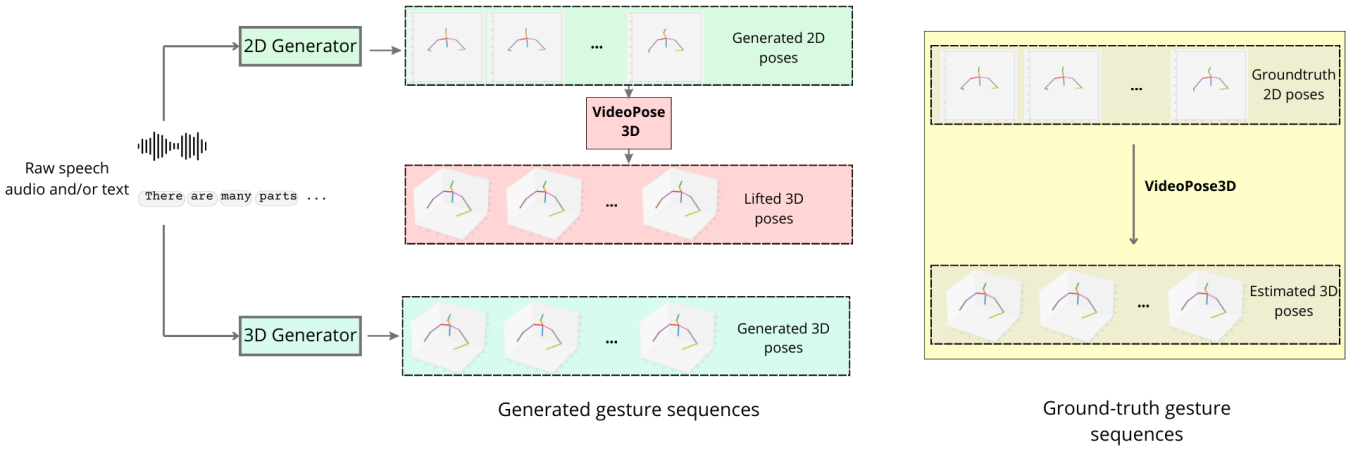

Figure 1: The proposed evaluation pipeline is a combination of a generative model [55, 58] that generates sequences of 2D body poses and VideoPose3D [44] that lifts the generated 2D poses to 3D. The pseudo-ground-truth 3D gesture sequences originate from the TED Gesture-3D dataset [55] and were obtained using VideoPose3D to lift 2D keypoints to 3D. The 2D keypoints were estimated using OpenPose [10] on TED YouTube videos.

the performance of two co-speech gesture generative models [55, 58]. We train both models to generate sequences of body poses represented in 2D coordinates which are then lifted to 3D using VideoPose3D [44]. The pipeline is described in Figure 1.

- We empirically compare the quality of the gestures generated in 2D lifted to 3D to the gestures directly generated in 3D using evaluation metrics commonly used in co-speech gesture generation tasks [39, 55, 58].
- Additionally, we conducted a user study where participants were asked to choose between gestures generated in 2D then lifted to 3D and gestures directly generated in 3D, providing a direct comparison of perceived quality.

The remainder of this paper is organized as follows: first, we present state-of-the-art works focusing on co-speech gesture generation. Secondly, we introduce our methodology and experimental design. Then, we discuss the experimental results of our objective and user studies. We finally end with a conclusion.

2 RELATED WORK

2.1 Learning-based co-speech gesture generation

The co-speech gesture synthesis field has seen an important shift to deep learning approaches for gesture generation due to their effectiveness in creating natural movements that are well-synchronized with speech, with minimal assumptions [43].

Deterministic approaches that directly translate speech to gesture sequences have been proposed. Notable model architectures are multi-layer perceptrons (MLP) [30], convolutional neural networks (CNN) [23], or transformers [9, 51]. Many considered recurrent architectures [9, 39, 55, 56] for their ability to capture long-term temporal dependencies to generate body pose sequences. Yoon et al. [56] proposed a sequence-to-sequence model that was trained to generate 2D gesture sequences from the TED Gesture dataset. The generated pose sequences were then lifted to 3D with a learned MLP to be mapped onto a social robot. Later on, Yoon et al. [55] introduced Trimodal, a multimodal recurrent neural network (RNN) that translates speech audio and text to 3D gestures conditioned on speaker identity. Their model was trained on the TED Gesture-3D dataset [55].

There is a notable interest in non-deterministic generative models such as Variational Autoencoders (VAEs) [36] and diffusion models [5, 6, 11, 14, 25, 48, 49, 57] due to their capacity for producing a wide array of gesture types. Specifically, VAEs are designed to encode gestures into a continuous latent space and subsequently decode these latent representations into speech-conditioned movements [36].

Recently, the gesture generation field has particularly focused on Probabilistic Denoising Diffusion Models [5, 6, 11, 14, 25, 48, 49, 57] due to their capacity to robustly produce diverse and realistic gestures under multiple conditions, including speech, text, speaker identity, and style. In diffusion-based methods audio-driven gesture synthesis is generally executed through classifier-free guidance [5, 6, 26, 58], leveraging both conditional and unconditional generation mechanisms during the sampling process. Alexanderson et al. [5] used Conformers [22] to generate gestures conditioned on behavior style and speech audio. Ao et al. [6] leveraged CLIP [46] to encode speech text and a style prompt. The authors used a combination of AdaIN [27] and classifier-free guidance to generate diverse yet style-conditioned gestures from speech. Zhu et al. [58] proposed DiffGesture, using a Diffusion Audio-Gesture Transformer to guarantee temporally aligned generation. In their work, raw speech audio is concatenated to gesture frames to condition the diffusion process. DiffGesture is trained on the TED Gesture-3D dataset [55] compiling 3D gestures inferred from 2D poses obtained from monocular video.

Table 1: Speech-Gesture datasets since 2018. The collection methods are described in the rightmost column. It can be either Motion Capture (MoCap) or pose estimation. Abbreviations: *rot. rotations coord. coordinates, n.s not specified.* For a more exhaustive list, refer to Nyatsanga et al.[43].

Dataset	Size	# of speakers	Type of motion data	Finger motion	Collection method
TED Gesture, 2018 [56]	52.7 h	1,295	2D joint coord.		OpenPose [10]
Trinity Speech Gesture I, 2018 [17]	6h	1	3D joint rot.		MoCap
SpeechGesture, 2019 [21]	144 h	10	2D joint coord.	Yes	OpenPose [10]
TalkingWithHands, 2019[33]	50h	50	3D joint rot.	Yes	Mocap
TED Gesture 3D, 2020 [55]	97h	n.s.	3D joint coord.		OpenPose [10], VideoPose3D [44]
Trinity Speech Gesture II, 2021 [18]	4h	1	3D joint rot.		Mocap
SpeechGesture 3D , 2021[23]	33h	6	3D joint coord.	Yes	OpenPose [10], XNect [41], [19]
TED Expressive, 2022 [39]	100.8h	n.s.	3D joint coord.	Yes	OpenPose [10], ExPose [12]
PATS, 2020[1]	250h	25	2D joint coord.		OpenPose [10]
BEAT, 2022 [38]	76h	30	3D joint rot.	Yes	MoCap
ZeroEggs, 2022 [20]	2h	1	3D joint rot.	Yes	MoCap
BiGe, 2023 [50]	260h	n.s.	3D joint coord.	Yes	OpenPose [10], VideoPose3D[44]

2.2 Representation and collection of the gesture data

The quality and diversity of the training data are critical for training co-speech gesture generative models. Additionally, to properly ground gestures to speech audio or text, gathering a large quantity of gesture data paired with these modalities is paramount.

Early works mostly considered 2D motion data for training [2, 16, 21, 45, 56]. To collect such datasets, 2D gestures were typically extracted from "in-the-wild" monocular videos using a third-party pose extractor such as OpenPose [1, 10, 21, 56]. We report in Table 1 a list of existing gesture datasets. This collection process makes it possible to gather a large amount of training data with numerous distinct speakers and ensures the diversity and spontaneity of the gestures. However, leveraging such pre-trained pose estimators induces errors resulting in less expressive motion quality, especially for capturing shoulder and finger movements, and limits the pose representation to be two-dimensional.

Most of the recent literature [5, 6, 32, 38, 53, 54, 57] focuses on MoCap datasets [17, 18, 20, 33, 38], in which high-quality motion data is captured in a controlled environment, usually with a limited number of speakers, consequently affecting the variety and spontaneity of the gestures. Multiple works [23, 39, 50, 55, 58] opted for increased diversity and volume of data samples while keeping a 3D representation of gestures, and chose to train their models on datasets of 3D gestures collected from "in-the-wild" videos [23, 39, 50, 55, 58]. To extract 3D body poses from monocular videos, the data collection process typically leverages a pipeline of pose extraction [10] and 2D-to-3D lifting [12, 41, 44]. For instance, the dataset TED Gesture-3D introduced by Yoon et al. [55] leverages VideoPose3D [44] to convert 2D body keypoints extracted by OpenPose [10] to 3D. This dataset is an extension of the previous TED Gesture dataset [56] where the poses are represented in 2D. [1].

There has been an important shift in the field towards 3D gestures generation [5, 6, 23, 38, 39, 50, 53–55, 57, 58]. However, the impact of the pose representation - either 2D or 3D - on the quality of synthesis gestures remains largely unexplored. Kucherenko et al. studied the impact of motion representation on the performances of data-driven co-speech gesture generative models, but their work only focused on the study of a gesture representation learned in a latent space.

In this work, we study how training an audio-driven generative model to synthesize 2D motion data and then post-processing the generated sequences using a 3D lifter impacts the overall quality of the synthesized gestures. We choose to use DiffGesture [58] and Trimodal [55] as baselines for our study as both were initially trained on the TED Gesture-3D dataset and obtained good performances, both qualitatively and objectively. Both models belong to widely used classes of generative models, DDPMs [5, 6, 11, 49, 57] and recurrent encoder-decoders [39, 50, 56]. For the 2D-to-3D lifting model, we employ VideoPose3D [44]. We use the TED Gesture-3D dataset [55] for our evaluation, which is a dataset of 3D gestures extracted from YouTube videos, hence covering a large array of speakers with different gesturing styles.

3 METHODOLOGY

In this section, we present the dataset, our evaluation pipeline, and the gesture generator models DiffGesture [58] and Trimodal [55]. We end this section by presenting our evaluation metrics.

3.1 TED Gesture-3D dataset

TED Gesture-3D [55] is a dataset including pose sequences extracted from in-the-wild videos of TED talkers with the corresponding speaker identity, speech audio, and speech transcription. TED Gesture-3D includes 3D body poses estimated via a combination of a 2D pose extractor from monocular videos [10] and VideoPose3D [44]. The size of the dataset is 97h where the poses are sampled at 15 frames per second with a stride of 10 with a total of 252,109 sequences of 34 frames. TED Gesture-3D is divided into training, validation and test sets which respectively represent 80%, 10% and 10% of the total dataset. Body poses are represented as vectors in $\mathcal{R}^{N\times J\times 3}$ where N is the sequence length and J is the number of body joints. Instead of considering raw joint coordinates for body pose representation, we follow the approach proposed by Yoon et al. [55] where a body pose is represented as nine directional vectors where each direction represents a bone. The vectors are

[1]To avoid confusion we refer to the 3D version of the database used by [55] as *TED Gesture-3D.*

normalized to the unit length and centered on the root joint. This pose representation is invariant to bone length and less affected by root rotations therefore favoring the training. Regarding 2D pose sequences, they are vectors of 3D poses from which the depth axis has been removed.

3.2 Pipeline

To evaluate the inductive bias caused by the dimensionality of the gesture representation (2D or 3D) and the 2D-to-3D conversion, we trained the co-speech gesture generators of [55] and [58] on both 2D and 3D settings. We employed a 3D lifter for post-processing the 2D generated sequences to be able to compare them to the 3D generated sequences. The complete pipeline is described in Figure 1.

3.2.1 Gesture generators. We now present the two gesture generators used as references in this study: DiffGesture [58] and Trimodal [55].

1) DiffGesture is defined as a DDPM that generates sequences of poses out of noise, conditioned on raw speech audio. DDPMs rely on two Markov chains: the forward process that gradually adds noise to the data and the backward process that converts noise to data. In DiffGesture, the backward process is modeled as a deep neural network that synthesizes gestures conditioned on speech. Raw audio is encoded using a convolutional neural network and then concatenated to the noisy pose sequence along the features axis. To synthesize diverse and speech-accurate gestures, DiffGesture uses classifier-free guidance [26]. This approach involves jointly training a conditioned and an unconditioned DDPM, allowing for a trade-off between the quality and diversity of the generated poses at inference time.

2) Trimodal is an encoder-decoder model trained in an adversarial scheme, that translates speech audio and text into 3D gestures, conditioned on speaker identity. It employs three distinct neural network encoders to process the three input modalities: audio, text, and speaker identity. The gesture generation utilizes a bi-directional gated recurrent unit (GRU) [7] to maintain temporal consistency. Audio and text inputs are handled by separate convolutional networks, while speaker identity is encoded into a style vector using a VAE. This VAE constructs a latent 'style' embedding space that captures the unique characteristics of each speaker. The style feature vector derived from this space is consistently applied across all synthesis time steps, ensuring coherent gesture representation throughout the sequence.

DiffGesture and Trimodal were first designed to generate 3D gestures. We adapted these architectures to account for 2D body pose sequences by changing the input and output dimensions of the denoising network and recurrent decoder network respectively. Specifically, we removed the depth axis of the body pose coordinates thus considering poses in $\mathcal{R}^{J\times 2}$. We refer to these versions as *DiffGesture 2D* and *Trimodal 2D* and the original versions are referred to as *DiffGesture 3D* and *Trimodal 3D*. We trained DiffGesture and Trimodal in two different settings: 2D motion generation and 3D motion generation, as described in Figure 1. For all experiments, we follow the original implementation proposed by Zhu et al. [58] and Yoon et al. [55]. In our objective study, we obtained similar results as Zhu et al. and Yoon et al. when retraining *DiffGesture 3D* and *Trimodal 3D* demonstrating the validity of our evaluation protocol (see Table 2).

3.2.2 2D-3D Lifter. We employed a 2D-3D lifter defined by a temporal convolutional network (TCN). Specifically, we used VideoPose3D [44] to lift 2D pose sequences to 3D as illustrated in Figure 1. The lifting process is defined as a mapping problem, in which the TCN employs 1-D convolutions along the temporal axis to transform 2D full body poses into a temporally consistent sequence of 3D body poses. VideoPose3D utilizes dilated temporal convolutions to capture long-term information.

We retrained VideoPose3D [44] on the TED Gesture-3D dataset to be able to input body poses in $\mathcal{R}^{2\times 9}$ i.e when only the upper part of the body is considered. We obtained a slightly better mean per joint positional error (MPJPE) when the sequences were up-scaled to 273 frames to exceed the receptive field of VideoPose3D instead of the original 34 frames. The final MPJPE of VideoPose3D was **13.4** on the test set of TED Gesture-3D. We kept the model architecture and training hyper-parameters consistent with the original implementation.

Supplementary videos used in the subjective evaluation study are provided at the following website:
`https://sites.google.com/view/iva-2d-or-not-2d`.

3.3 Comparative setting

To study the impact of the dimensionality of the motion data on the performance of gesture generative models, we considered two experimental settings.

3.3.1 Evaluation in the 3D gesture space. To evaluate the impact of training on 2D motions on the quality of 3D gesture sequences, we define *DiffGesture 2D + VP3D* and *Trimodal 2D + VP3D* as DiffGesture and Trimodal trained on 2D motion data whose outputs are then lifted to 3D using VideoPose3D and we compare them to the original models, *DiffGesture 3D* and *Trimodal 3D* [55, 58].

3.3.2 Evaluation in the 2D gesture space. To further explore the impact of motion dimensionality on the generated gesture, we also compare *DiffGesture 2D* and *Trimodal 2D* to *DiffGesture 3D* and *Trimodal 3D* but where the 3D generated motion is narrowed to 2D by removing the depth axis, we refer to these models as *DiffGesture 3D→2D* and *Trimodal 3D→2D*.

3.4 Objective Evaluation metrics

We numerically evaluate our models with three commonly used metrics in the co-speech gesture generation field.

3.4.1 The Fréchet Gesture Distance. The Fréchet Gesture Distance (FGD) defined by Yoon et al. [55] is an adaptation of the Fréchet Inception Distance (FID) [24]. The FGD computes the 2-Wasserstein distance between two distributions leveraging latent features extracted with a pose encoder. Similar distributions result in a low FGD value. The FGD is defined as follows:

$$FGD(X, \hat{X}) = \|\mu_r - \mu_g\| + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^2) \qquad (1)$$

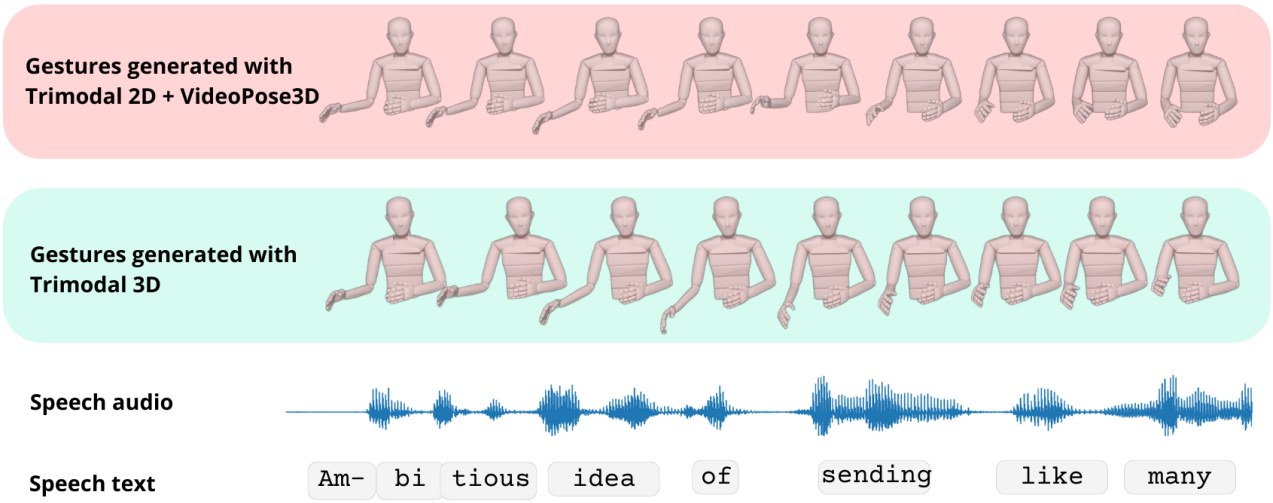

Figure 2: Keyframes of an animation generated with *Trimodal 2D + VP3D* (up) and *Trimodal 3D* (down).

Where: $X, \hat{X}$ are the real and generated distributions respectively; $\mu_r, \mu_g, \Sigma_r, \Sigma_g$ are the mean and covariance of the latent distributions extracted from the real and generated distributions.

3.4.2 The Beat Consistency Score. The Beat Consistency Score (BC) measures the temporal consistency between kinematic and audio beats of a paired audio-motion sequence. This measure, first introduced for evaluating the synchrony of dance with music [35], has been adapted to speech and gestures [37]. First, kinematic beats are extracted from a pose sequence by selecting the time steps of the sequence where the average angle velocity is higher than a certain threshold. Angle velocity is computed using the variation in angle between two successive frames. Intuitively, BC measures the average distance between the time steps corresponding to an audio beat and the closest time steps corresponding to a kinematic beat. The audio beats are extracted using a pre-trained detection model and the BC score is computed as follows:

$$BC = \frac{1}{n}\sum_{i=0}^{n} exp\Big(-\frac{min_{\forall t_j^y \in \mathcal{B}^y}\|t_j^y - t_i^x\|^2}{2\sigma^2}\Big) \quad (2)$$

Where: t_i^x is the $i-th$ audio beats, $B_y = t_i^y$ is the set of the kinematic beats of the $i-th$ sequence, and σ is a parameter to normalize sequences, set to 0.1 empirically as in [58].

3.4.3 The Diversity measure. The Diversity measure also leverages the latent features extracted with a pose encoder [34]. Diversity is computed by randomly selecting two sets of N features from the generated distribution and calculating the distance between the mean of both sets in the feature space. Typically, if a model generates similar gestures all gestures will be close to the average gesture sequence, resulting in a small distance between the two sets, as formalized below:

$$Div(X) = \|\mu_A - \mu_B\|_2 \quad (3)$$

Where: X is a distribution of gestures, A and B are sets of gestures randomly sampled from X, and μ_A and μ_B are the mean of the gesture features in both sets.

The FGD and diversity metrics are calculated using an auto-encoder designed to encode 3D gestures into latent space. Yoon et al. [55] developed this auto-encoder using the human3.6m dataset [28]. Similarly, we adapted this model to encode 2D gestures by eliminating the depth axis and training it on the same dataset.

4 USER STUDY

Properly evaluating generative models is a challenging task, especially in the co-speech gesture generation field partly because of the subjective nature of human communication [5, 43, 55]. In this section, we present our user study protocol to qualitatively compare the gestures directly generated in 3D to those generated in 2D and subsequently lifted to 3D.

4.1 Protocol

We created videos showing the animation of the upper body of an articulated humanoid. Each video features two stimuli for pairwise comparison as it has been shown to reduce the cognitive load of the users [13, 52, 55]. The first animation is displayed on the left side of the screen with the second animation masked. The second animation is shown while the first is masked. In each video, both animations used the same model (either DiffGesture or Trimodal), one with direct 3D gestures and the other with 2D gestures converted to 3D. We qualitatively evaluated the impact of the 2D-to-3D lifting, by including baseline videos (referred to as *Human GT*). These videos paired 3D pseudo-ground-truth gestures to those created by converting the 2D versions of these gestures to 3D using the retrained version of VideoPose3D. The pseudo-ground-truth gestures originate from the test set of the TED Gesture-3D dataset.

After viewing each video, participants were asked to answer three questions:

- "*Select the video in which the articulated figure is more **human-like***."
- "*Select the video in which the articulated figure looks more **alive***."
- "*Select the video in which the articulated figure looks more **in sync with the speech***."

We selected the terms "human-like" and "alive" to evaluate two distinct dimensions: the anthropomorphism and animacy of the agent, following the semantics of the Godspeed Questionnaire [8]. For each question, there were four possible answers: "*Clearly left*", "*Fairly left*", "*Fairly right*", and "*Clearly right*". The options "Clearly" and "Fairly" correspond to the degree of confidence the participants had in their choice. Each response is assigned an integer value: +2 for a clear preference for gestures directly generated in 3D, +1 for a slight preference, -1 for a slight preference for lifted 3D gestures, and -2 for a clear preference for lifted 3D gestures.

In our subjective study, we created 14 stimuli for each condition: *Human GT*, *DiffGesture*, and *Trimodal*, resulting in 42 pairwise comparisons of the 3D model and its 2D counterpart. To reduce the length of the questionnaires and keep the participants focused during the study, we conducted two evaluation sessions featuring 21 stimuli. Hence each participant saw 7 stimuli for each condition. To prevent ordering bias, we created four distinct questionnaires for each session where the order of appearance of each stimuli has been randomized. In addition to the main stimuli, we included a first example video to familiarize the users with the task and check for technical issues such as no audio or video. Our questionnaire also featured two attention checks. In the first attention check, the stimuli were replaced by a black-screen video and users were asked to select a specific option three times instead of the questions about human likeness, aliveness, and speech synchrony. In the second attention check, we replaced the original stimuli with a modified version where the audio indicated the users to choose the rightmost option for each question.

4.2 Gestures rendering

For visualization, we used the 3D character from the user evaluation in the Genea Challenge 2021 [31]. The gestures were rendered with Blender. To create the animations, we selected 14 speech samples from the test set of the TED Gesture-3D dataset. Each sample lasts around 10 seconds, which is twice as long as the samples used for the subjective evaluation in Yoon et al. [55], Zhu et al. [58] did not report the duration of their stimuli. We selected the samples based on the audio quality and for each sample, we generated four 3D co-speech gesture sequences with *DiffGesture 3D*, *DiffGesture 2D + VP3D*, *Trimodal 3D*, and *Trimodal 2D + VP3D*. As a baseline, we also included gesture sequences directly from the TED Gesture-3D dataset and those were reduced to 2D and then converted back to 3D. Hence, for each audio sample, we obtained 6 sequences for the gestures generated in 3D and their 2D-to-3D counterpart. We created videos that pair the gestures generated in 3D and those converted from 2D, resulting in 42 animation pairs. The order of animations (either direct 3D gestures first or lifted gestures first) was randomized to prevent ordering bias.

Table 2: Objective results of the experiments on the TED Gesture-3D dataset [55]. These results correspond to the experiments (1) and (2) (see section 3.3). Up arrows indicate that a higher result is better whereas down arrows indicate that a lower result is better. * means that the results are reported from [58].

	TED Gesture		
Methods	FGD ↓	BC ↑	Diversity ↑
Evaluation on the 3D gesture space			
Ground Truth 3D	0	0.702	102.339
DiffGesture 3D [58]	1.947	0.678	101.436
DiffGesture 2D + VP3D	3.121	0.551	100.822
Trimodal 3D [55]	3.964	0.733	95.253
Trimodal 2D + VP3D	6.374	0.610	93.017
Evaluation on the 2D gesture space			
Ground Truth 2D	0	0.689	112.76
DiffGesture (3D→2D)	2.452	0.661	109.978
DiffGesture 2D	2.971	0.644	111.869
Trimodal (3D→2D)	5.295	0.724	104.072
Trimodal 2D	6.227	0.706	102.100
Reported results from [58]			
Ground Truth 3D	0	0.698	108.525
DiffGesture * [58]	1.506	0.699	106.722
Attention Seq2Seq* [56]	18.154	0.196	82.776
Speech2Gesture* [21]	19.254	0.668	93.802
Joint Embedding* [3]	22.083	0.200	90.138
Trimodal* [55]	3.729	0.667	101.247
HA2G* [39]	3.072	0.672	104.322

4.3 Participants

The participants in the evaluation were recruited online on the Prolific platform [15, 29]. Among the 67 participants who took the test, 7 failed the attention checks. The users were 36.7+/-11.8 years old and there were 37 females and 30 males and the median completion time was 17 minutes. As we performed two sessions with 7 stimuli for each condition, we obtained 30 valid responses for each stimulus. The participants were paid 3£ if they passed the attention checks.

5 EXPERIMENTAL RESULTS

5.1 Objective evaluation

The results of our objective experiments are reported in Table 2. In the table's upper section, we present outcomes from our experiments evaluating gestures in 3D. The middle section details the results from our experiments assessing gestures in 2D. The results from Zhu et al. and Yoon et al. [55, 58] are reported in the table's lower section. It is important to note that we retrained the motion encoder used to compute the FGD and diversity score. The reported results from Zhu et al. and Yoon et al. were obtained using their encoder.

5.1.1 Evaluation of lifted generated gestures. When comparing the results of *DiffGesture 2D + VP3D* and *Trimodal 2D + VP3D* to

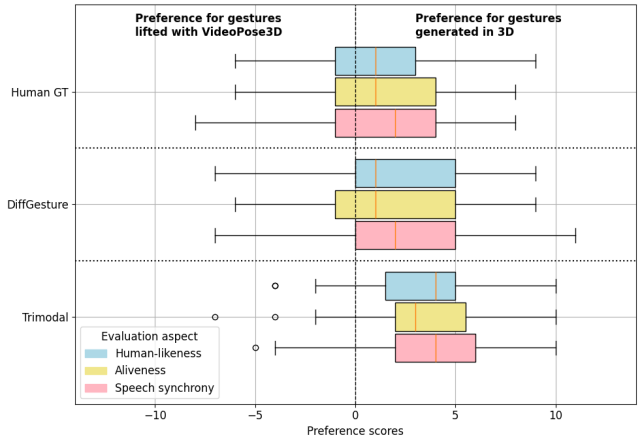

Figure 3: Results of our evaluation study. A positive score means that gestures generated directly in 3D are preferred over 2D gestures lifted to 3D. Reciprocally, a negative score means that 2D gestures lifted to 3D are preferred over direct 3D gestures. A score close to 0 means that the preference is unclear.

those of *DiffGesture 3D* and *Trimodal 3D*, we can notice that the models trained on 2D gestures perform worse than the original 3D models in terms of FGD, and BC. There is also a slight drop in diversity for Trimodal. We assume that the one-to-many relationship between 2D and 3D keypoints is mostly responsible for the performance drop of *DiffGesture 2D + VP3D* and *Trimodal 2D + VP3D* for the FGD. As VideoPose3D is deterministic, to one 2D pose it will systematically predict the same 3D pose although there exists multiple possibilities. Hence, the distribution resulting from lifting 2D sequences is tighter than the distribution directly generated in 3D, explaining the high FGD of the gestures generated in 2D lifted to 3D. There is a drop in BC between the 3D models and their 2D counterparts. It can be that post-processing 2D gestures using VideoPose3D tends to over-smooth the resulting 3D gestures therefore reducing the number of kinematic beats. While the overall quality diminishes when generating gestures in 2D and then lifting them to 3D, *DiffGesture 2D + VP3D* and *Trimodal 2D + VP3D* remain competitive compared to the other baselines reported in the table lower section.

5.1.2 Evaluation of the quality of gestures generated in 2D. When evaluating in the 2D motion space, both *DiffGesture 3D⟶2D* and *Trimodal 3D⟶2D* perform better than *DiffGesture 2D* and *Trimodal 2D* respectively in terms of FGD. Hence, training the generative models to generate 3D motion sequences seems to behave better than training the model on 2D motion data. This outcome was anticipated since the representation of poses in 3D is richer than the 2D version. The BC and diversity scores do not seem to be influenced by the dimensionality of the gestures used to train the generative models.

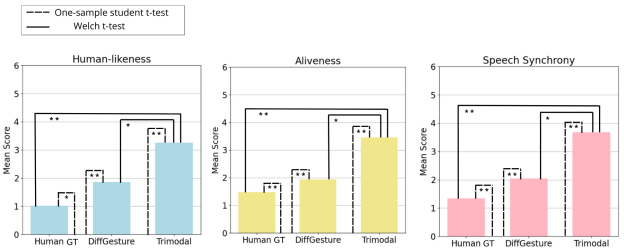

Figure 4: Statistical comparison of mean scores for each model and each aspect (Human-likeness, Aliveness, and Speech synchrony). The lines represent a significant superiority between the two values. Dotted lines correspond to Student t-tests and plain lines to Welch t-tests. * means p-value < 0.05 while ** means p-value < 0.01

Table 3: Confidence Intervals of the Mean Scores

Approach	Human-likeness	Aliveness	Speech Synchrony
Human GT	[-0.22, 2.26]	[0.29, 2.66]	[0.08, 2.56]
DiffGesture	[0.59, 3.14]	[0.68, 3.18]	[0.83, 3.27]
Trimodal	[2.27, 4.23]	[2.46, 4.46]	[2.61, 4.75]

5.2 Subjective evaluation

The results of our user study are presented in Figure 3. The figure highlights pairwise preference between gestures generated directly in 3D or gestures generated in 2D lifted to 3D using VideoPose3D. We conducted a statistical analysis to determine the significance of our user study results. We used Student t-tests to evaluate the three different techniques (Human GT, DiffGesture, Trimodal) for generating gestures in 3D. The Student t-test checked if the average scores for human likeness, aliveness, and speech synchrony were significantly different from zero. A score significantly greater than zero indicates that gestures created directly in 3D are better perceived than those initially generated in 2D and subsequently converted to 3D using VideoPose3D. We also conducted Welch t-tests to determine the significance level of the model-wise mean score comparisons. The results of the Student t-tests and Welch tests are depicted in Figure 4. We calculated the confidence intervals of the mean scores for Human GT, DiffGesture, and Trimodal for the human-likeness, aliveness and speech synchrony. These intervals are depicted in Table 3.

First, for the DiffGesture and Trimodal techniques, all scores are significantly higher than zero (Figure 4), with a p-value less than 0.01. This suggests that gestures created directly in 3D by these methods are more effective compared to those initially created in 2D and then converted to 3D for all three aspects. In Table 3, the confidence intervals of the score of Trimodal show that *Trimodal 3D* is preferred over *Trimodal 2D + VP3D* whereas the preference for *DiffGesture 3D* over *DiffGesture 2D + VP3D* is slighter. It is important to note that the confidence intervals of the scores overlap between DiffGesture and Trimodal, but the human-likeness, aliveness, and speech-synchrony means of Trimodal are significantly greater than

those of DiffGesture with a p-value of 0.028, 0.016 and 0.011 respectively. A comparative example between gestures generated with *Trimodal 3D* and *Trimodal 2D + VP3D* is depicted in Figure 2.

For the Human baseline, converting 2D gestures to 3D demonstrates a minimal yet statistically significant impact on the perception of their human-likeness animation, with a score above zero (p-value of 0.020) and a confidence interval of the mean score close to zero. This suggests that while VideoPose3D influences the perceived human likeness, the effect is subtle. In contrast, the conversion process significantly affects gesture quality in terms of aliveness and speech synchrony, as evidenced by scores significantly higher than zero (p-values of 0.001 and 0.004, respectively). This indicates a notable degradation in these aspects due to the use of VideoPose3D for lifting 2D gestures to 3D.

From these results, we can conclude that training a model to generate 2D gestures and then converting these gestures to 3D deteriorates the overall animation quality in terms of human likeness, aliveness, and speech synchrony. The 2D-to-3D conversion of gestures has a small yet significant impact on the perception of human-likeness. Hence, the drop in human-likeness quality for gestures generated in 2D and then lifted to 3D may come from the training of the generative model itself since the 2D gesture representation may not allow the generation of highly human-like gestures once converted to 3D.

6 CONCLUSION

In this study, we explored how training two co-speech gesture generators with 2D data affects the overall performance of the models and the perceived quality of the synthesized gestures. We introduced a pipeline that pairs a gesture generator — either DiffGesture [58] or Trimodal [55] — with a 2D-to-3D lifting model, specifically VideoPose3D [44]. Our objective results reveal that using this pipeline negatively impacts overall performance. 3D gestures lifted from 2D generated gestures are less similar to the target 3D gesture distribution in comparison to gestures generated directly in 3D and lifting 2D gestures reduces the BC score. To further confirm these results, we performed a large-scale user study involving 60 participants. The goal of this human evaluation was to assess the impact of 2D motion representation on the perceived human likeness, aliveness, and speech synchrony of the generated gestures. Our findings show that direct 3D gestures are preferred over gestures lifted using VideoPose3D for both Trimodal [55] and DiffGesture [58]. Our findings also indicate that converting 2D gestures to 3D slightly reduces their perceived human likeness, aliveness, and speech synchrony. These results suggest that generating 2D body poses and then lifting them in 3D produces gesture animations that are less human-like, alive, and in sync with speech than those created directly in 3D. Further, they show that using 3D representations for co-speech gesture generation enhances their quality and relation to speech.

7 LIMITATIONS AND FUTURE WORK

It is worth noting that our study is not without limitations. In the TED Gesture-3D dataset, the 3-dimensional gestures are lifted from 2D body poses. Our evaluation is therefore biased as we do not have access to the real 3D ground truth data. For future research, we plan to conduct a similar analysis using Mocap data to have access to the real 3D ground-truth gestures.

In the near future, we will test the generalization of our approach on similar "in-the-wild" datasets such as PATS, SpeechGesture and TED Expressive [1, 21, 23, 39]. Additionally, we will consider finger motions which convey a lot of information in human communication [40]. Transforming 2D finger motion into 3D is a complex task, and our focus will be on exploring the best data representation for accurately generating such fine-grained gestures.

REFERENCES

[1] Chaitanya Ahuja, Dong Won Lee, Ryo Ishii, and Louis-Philippe Morency. 2020. No Gestures Left Behind: Learning Relationships between Spoken Language and Freeform Gestures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 1884–1895.

[2] Chaitanya Ahuja, Dong Won Lee, Yukiko I. Nakano, and Louis-Philippe Morency. 2020. Style Transfer for Co-Speech Gesture Animation: A Multi-Speaker Conditional-Mixture Approach. https://arxiv.org/abs/2007.12553

[3] Chaitanya Ahuja and Louis-Philippe Morency. 2019. Language2Pose: Natural Language Grounded Pose Forecasting. *CoRR* abs/1907.01108 (2019). arXiv:1907.01108 http://arxiv.org/abs/1907.01108

[4] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows. *Computer Graphics Forum* 39, 2 (2020), 487–496. https://doi.org/10.1111/cgf.13946 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13946

[5] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. 2023. Listen, Denoise, Action! Audio-Driven Motion Synthesis with Diffusion Models. *ACM Transactions on Graphics* 42, 4 (2023). https://doi.org/10.1145/3592458

[6] Tenglong Ao, Zeyi Zhang, and Libin Liu. [n. d.]. GestureDiffuCLIP: Gesture Diffusion Model with CLIP Latents. *ACM Trans. Graph.* ([n. d.]), 18 pages. https://doi.org/10.1145/3592097

[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR* abs/1409.0473 (2014). https://api.semanticscholar.org/CorpusID:11212020

[8] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics* 1, 1 (01 Jan 2009), 71–81. https://doi.org/10.1007/s12369-008-0001-3

[9] Uttaran Bhattacharya, Elizabeth Childs, Nicholas Rewkowski, and Dinesh Manocha. 2021. Speech2AffectiveGestures: Synthesizing Co-Speech Gestures with Generative Adversarial Affective Expression Learning. *CoRR* abs/2108.00262 (2021). arXiv:2108.00262 https://arxiv.org/abs/2108.00262

[10] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

[11] Ankur Chemburkar, Shuhong Lu, and Andrew Feng. 2023. Discrete Diffusion for Co-Speech Gesture Synthesis. In *Companion Publication of the 25th International Conference on Multimodal Interaction* (, Paris, France,) *(ICMI '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 186–192. https://doi.org/10.1145/3610661.3616556

[12] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. 2020. Monocular Expressive Body Regression through Body-Driven Attention. In *European Conference on Computer Vision (ECCV)*. https://expose.is.tue.mpg.de

[13] Andrew P. Clark, Kate L. Howard, Andy T. Woods, Ian S. Penton-Voak, and Christof Neumann. 2018. Why rate when you could compare? Using the "EloChoice" package to assess pairwise comparisons of perceived physical strength. *PLOS ONE* 13, 1 (01 2018), 1–16. https://doi.org/10.1371/journal.pone.0190393

[14] Anna Deichler, Shivam Mehta, Simon Alexanderson, and Jonas Beskow. 2023. Diffusion-Based Co-Speech Gesture Generation Using Joint Text and Audio Representation. In *Proceedings of the 25th International Conference on Multimodal Interaction (ICMI '23)*. Association for Computing Machinery, New York, NY, USA, 755–762. https://doi.org/10.1145/3577190.3616117

[15] Benjamin D Douglas, Patrick J Ewell, and Markus Brauer. 2023. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLoS One* 18, 3 (March 2023), e0279720.

[16] Mireille Fares, Catherine Pelachaud, and Nicolas Obin. 2023. Zero-Shot Style Transfer for Multimodal Data-Driven Gesture Synthesis. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. 1–4. https://doi.org/10.1109/FG57933.2023.10042658

[17] Ylva Ferstl and Rachel McDonnell. 2018. Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (Sydney, NSW, Australia) *(IVA '18)*. Association for Computing Machinery, New York, NY, USA, 93–98.

https://doi.org/10.1145/3267851.3267898

[18] Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2021. ExpressGesture: Expressive gesture generation from speech through database matching. *Computer Animation and Virtual Worlds* 32, 3-4 (2021), e2016. https://doi.org/10.1002/cav.2016 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/cav.2016

[19] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. 2016. Reconstruction of Personalized 3D Face Rigs from Monocular Video. *ACM Trans. Graph.* 35, 3, Article 28 (may 2016), 15 pages. https://doi.org/10.1145/2890493

[20] Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F. Troje, and Marc-André Carbonneau. 2022. ZeroEGGS: Zero-shot Example-based Gesture Generation from Speech. arXiv:2209.07556 [cs.GR]

[21] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning Individual Styles of Conversational Gesture. *CoRR* abs/1906.04160 (2019). arXiv:1906.04160 http://arxiv.org/abs/1906.04160

[22] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. arXiv:2005.08100 [eess.AS]

[23] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. 2021. Learning Speech-driven 3D Conversational Gestures from Video. In *ACM International Conference on Intelligent Virtual Agents (IVA)*. arXiv:Todo

[24] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Nash Equilibrium. *CoRR* abs/1706.08500 (2017). arXiv:1706.08500 http://arxiv.org/abs/1706.08500

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS'20)*. Curran Associates Inc., Red Hook, NY, USA, Article 574, 12 pages.

[26] Jonathan Ho and Tim Salimans. 2022. Classifier-Free Diffusion Guidance. arXiv:2207.12598 [cs.LG]

[27] Xun Huang and Serge J. Belongie. 2017. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. *CoRR* abs/1703.06868 (2017). arXiv:1703.06868 http://arxiv.org/abs/1703.06868

[28] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 7 (jul 2014), 1325–1339.

[29] Patrik Jonell, Taras Kucherenko, Ilaria Torre, and Jonas Beskow. 2020. Can we trust online crowdworkers?: Comparing online and offline participants in a preference test of virtual agents. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20)*. ACM. https://doi.org/10.1145/3383652.3423860

[30] Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexanderson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. *CoRR* abs/2001.09326 (2020). arXiv:2001.09326 https://arxiv.org/abs/2001.09326

[31] Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, Zerrin Yumak, and Gustav Henter. 2021. GENEA Workshop 2021: The 2nd Workshop on Generation and Evaluation of Non-verbal Behaviour for Embodied Agents. In *Proceedings of the 2021 International Conference on Multimodal Interaction* (Montréal, QC, Canada) *(ICMI '21)*. Association for Computing Machinery, New York, NY, USA, 872–873. https://doi.org/10.1145/3462244.3480983

[32] Taras Kucherenko, Rajmund Nagy, Youngwoo Yoon, Jieyeon Woo, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2023. The GENEA Challenge 2023: A large-scale evaluation of gesture generation models in monadic and dyadic settings. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI '23)*. ACM.

[33] Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha Srinivasa, and Yaser Sheikh. 2019. Talking With Hands 16.2M: A Large-Scale Dataset of Synchronized Body-Finger Motion and Audio for Conversational Motion Analysis and Synthesis. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 763–772. https://doi.org/10.1109/ICCV.2019.00085

[34] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. 2019. Dancing to Music. *CoRR* abs/1911.02001 (2019). arXiv:1911.02001 http://arxiv.org/abs/1911.02001

[35] Buyu Li, Yongchi Zhao, and Lu Sheng. 2021. DanceNet3D: Music Based Dance Generation with Parametric Motion Transformer. *CoRR* abs/2103.10206 (2021). arXiv:2103.10206 https://arxiv.org/abs/2103.10206

[36] Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. 2021. Audio2Gestures: Generating Diverse Gestures from Speech Audio with Conditional Variational Autoencoders. *CoRR* abs/2108.06720 (2021). arXiv:2108.06720 https://arxiv.org/abs/2108.06720

[37] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. 2021. Learn to Dance with AIST++: Music Conditioned 3D Dance Generation. *CoRR* abs/2101.08779 (2021). arXiv:2101.08779 https://arxiv.org/abs/2101.08779

[38] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022. BEAT: A Large-Scale Semantic and Emotional Multi-Modal Dataset for Conversational Gestures Synthesis. arXiv:2203.05297 [cs.CV]

[39] Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. 2022. Learning Hierarchical Cross-Modal Association for Co-Speech Gesture Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10462–10472.

[40] David Mcneill. 1992. Hand and Mind: What Gestures Reveal About Thought. *University of Chicago Press* 27 (1992). https://doi.org/10.2307/1576015

[41] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. 2019. XNect: Real-time Multi-person 3D Human Pose Estimation with a Single RGB Camera. *CoRR* abs/1907.00837 (2019). arXiv:1907.00837 http://arxiv.org/abs/1907.00837

[42] Shivam Mehta, Siyang Wang, Simon Alexanderson, Jonas Beskow, Eva Szekely, and Gustav Eje Henter. 2023. Diff-TTSG: Denoising probabilistic integrated speech and gesture synthesis. In *12th ISCA Speech Synthesis Workshop (SSW2023)*. ISCA. https://doi.org/10.21437/ssw.2023-24

[43] S. Nyatsanga, T. Kucherenko, C. Ahuja, G. E. Henter, and M. Neff. 2023. A Comprehensive Review of Data-Driven Co-Speech Gesture Generation. *Computer Graphics Forum* 42, 2 (May 2023), 569–596. https://doi.org/10.1111/cgf.14776

[44] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[45] Shenhan Qian, Zhi Tu, Yihao Zhi, Wen Liu, and Shenghua Gao. 2021. Speech Drives Templates: Co-Speech Gesture Synthesis with Learned Templates. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 11057–11066.

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. *CoRR* abs/2103.00020 (2021). arXiv:2103.00020 https://arxiv.org/abs/2103.00020

[47] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. *CoRR* abs/1503.03585 (2015). arXiv:1503.03585 http://arxiv.org/abs/1503.03585

[48] Yang Song and Stefano Ermon. 2019. Generative Modeling by Estimating Gradients of the Data Distribution. *CoRR* abs/1907.05600 (2019). arXiv:1907.05600 http://arxiv.org/abs/1907.05600

[49] Rodolfo Luis Tonoli, Leonardo Boulitreau de Menezes Martins Marques, Lucas Hideki Ueda, and Paula Paro Dornhofer Costa. 2023. Gesture Generation with Diffusion Models Aided by Speech Activity Information. In *GENEA: Generation and Evaluation of Non-verbal Behaviour for Embodied Agents Challenge 2023*. https://openreview.net/forum?id=S9Efb3MoiZ

[50] Hendric Voß and Stefan Kopp. 2023. AQ-GT: a Temporally Aligned and Quantized GRU-Transformer for Co-Speech Gesture Synthesis. *arXiv preprint arXiv:2305.01241* (2023).

[51] Jonathan Windle, Iain Matthews, Ben Milner, and Sarah Taylor. 2023. The UEA Digital Humans entry to the GENEA Challenge 2023. In *GENEA: Generation and Evaluation of Non-verbal Behaviour for Embodied Agents Challenge 2023*. https://openreview.net/forum?id=bBrebR1YpXe

[52] Pieter Wolfert, Nicole Robinson, and Tony Belpaeme. 2022. A Review of Evaluation Practices of Gesture Generation in Embodied Conversational Agents. *IEEE Transactions on Human-Machine Systems* 52, 3 (June 2022), 379–389. https://doi.org/10.1109/thms.2022.3149173

[53] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, Ming Cheng, and Long Xiao. 2023. DiffuseStyleGesture: Stylized Audio-Driven Co-Speech Gesture Generation with Diffusion Models. arXiv:2305.04919 [cs.HC]

[54] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, and Haolin Zhuang. 2023. QPGesture: Quantization-Based and Phase-Guided Motion Matching for Natural Speech-Driven Gesture Generation. arXiv:2305.11094 [cs.HC]

[55] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity. *ACM Transactions on Graphics* 39, 6 (2020).

[56] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2018. Robots Learn Social Skills: End-to-End Learning of Co-Speech Gesture Generation for Humanoid Robots. *CoRR* abs/1810.12541 (2018). arXiv:1810.12541 http://arxiv.org/abs/1810.12541

[57] Weiyu Zhao, Liangxiao Hu, and Shengping Zhang. 2023. DiffuGesture: Generating Human Gesture From Two-person Dialogue With Diffusion Models. In *Companion Publication of the 25th International Conference on Multimodal Interaction* (<conf-loc>, <city>Paris</city>, <country>France</country>, </conf-loc>) *(ICMI '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 179–185. https://doi.org/10.1145/3610661.3616552

[58] 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 12th April, 2024