Investigating the impact of 2D gesture representation on co-speech gesture generation

Téo Guichoux, Laure Soulier, Nicolas Obin, Catherine Pelachaud

Investigating the impact of 2D gesture representation on co-speech gesture generation

Téo Guichoux
Sorbonne Université
ISIR
STMS lab IRCAM
CNRS
Paris, France
teo.guichoux@isir.upmc.fr

Laure Soulier
Sorbonne Université
ISIR, CNRS
Paris, France
laure.soulier@isir.upmc.fr

Nicolas Obin
STMS lab
IRCAM, CNRS, Sorbonne Université
Paris, France

Catherine Pelachaud
Sorbonne Université
ISIR, CNRS
Paris, France
catherine.pelachaud@isir.upmc.fr

ABSTRACT

Co-speech gestures play a crucial role in the interactions between humans and embodied conversational agents (ECA). Recent deep learning methods enable the generation of realistic, natural co-speech gestures synchronized with speech, but such approaches require large amounts of training data. "In-the-wild" datasets, which compile videos from sources such as YouTube through human pose detection models, offer a solution by providing 2D skeleton sequences that are paired with speech. Concurrently, innovative lifting models have emerged, capable of transforming these 2D pose sequences into their 3D counterparts, leading to large and diverse datasets of 3D gestures. However, the derived 3D pose estimation is essentially a pseudo-ground truth, with the actual ground truth being the 2D motion data. 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. In this work, we evaluate the impact of the dimensionality of the training data, 2D or 3D joint coordinates, on the performance of a multimodal speech-to-gesture deep generative model. We use a lifting model to convert 2D-generated sequences of body poses to 3D. Then, we compare the sequence of gestures generated directly in 3D to the gestures generated in 2D and lifted to 3D as post-processing.

CCS CONCEPTS

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

KEYWORDS

Co-speech gesture generation, Pose Representation, Diffusion Models

WACAI, 12-14 juin 2024, Bordeaux (France)
2024.

ACM Reference Format:
Téo Guichoux, Laure Soulier, Nicolas Obin, and Catherine Pelachaud. 2024. Investigating the impact of 2D gesture representation on co-speech gesture generation. In *Workshop Affects, Compagnons Artificiels et Interactions (WACAI)*. 8 pages.

1 INTRODUCTION

In human communication, gestures play an integral role by conveying intentions and emphasizing points [33]. Recent studies [3–5, 10, 12, 35, 42, 44–47, 49] 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 [19]. Most recent work on co-speech gesture synthesis considers 3D motion data [3–5, 10, 35, 42, 44–46, 49], primarily because such data representation is more expressive and more easily transferable to downstream applications such as animation, virtual reality or social robots [36, 47]. However, the quality of learning-based co-speech gesture synthesis in terms of naturalness, speech synchrony, and diversity heavily relies on the quantity and quality of the training data which may be costly to collect. In recent works, mostly two kinds of datasets are considered: motion-capture (Mocap) [13, 14, 16, 26, 31] and pose estimation from "in-the-wild" videos [1, 17, 19, 42, 47] which are videos freely accessible online.

Mocap datasets show obvious superiority in terms of fine-grained motion and annotation quality over "in-the-wild" datasets. However, they are very costly to collect and lack **1)** diversity because of the reduced number of considered speakers and **2)** naturalness because gestures and expressed emotions are acted in a controlled studio environment. Datasets based on "in-the-wild" videos, on the contrary, usually gather data from many different speakers with different speech and gesture profiles. They offer the possibility to access large raw datasets of online videos but with heterogeneous and uncontrolled situations and environments. Such datasets require robust and accurate pose estimation algorithms to estimate

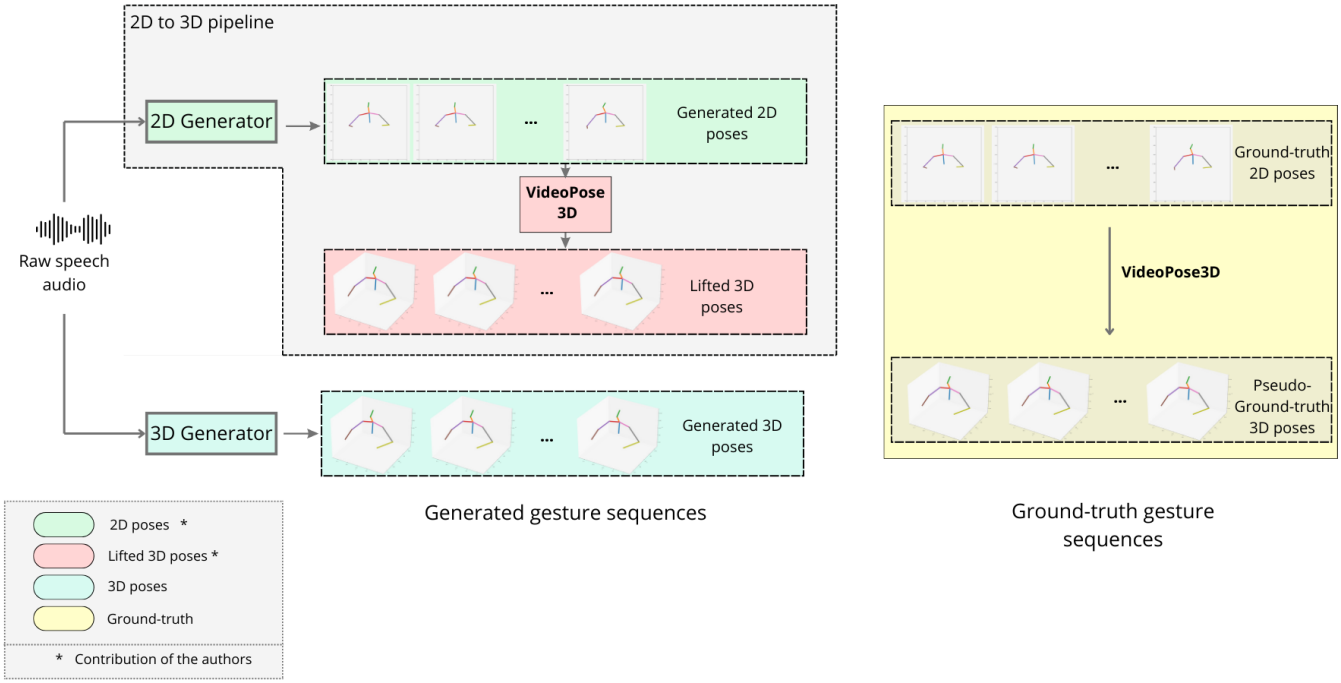

Figure 1: The proposed evaluation pipeline is a combination of DiffGesture [49] that generates sequences of 2D body poses and VideoPose3D [37] that lifts the generated 2D poses to 3D. The pseudo-ground-truth 3D gesture sequences originate from the TED Gesture-3D dataset [46] and were obtained using VideoPose3D to lift 2D keypoints to 3D. The 2D keypoints were estimated using OpenPose [7] on TED YouTube videos.

2D keypoints coordinates from monocular videos [7, 11]. 3D body poses can further be inferred from these estimated 2D key-points using a third-party 2D to 3D lifting model [9, 34, 37]. Nevertheless, relying on such 3D lifters induces error inherently dependent on the ambiguous nature of 2D to 3D lifting problems, as one 2D pose can correspond to multiple 3D poses.

Due to the ambiguous nature of 3D pose estimation from 2D keypoints, datasets leveraging 2D to 3D lifting models are prone to inaccuracies and lack of fine-grained motions such as hand joints and finger positions. Recent works on co-speech gesture synthesis from "in-the-wild" datasets mostly consider 3D data lifted from estimated 2D keypoints [19, 32, 42, 46, 49], but it remains unclear what is the impact of the dimensionality of the skeleton representation on the training of the generative model and the quality of the generated gestures.

The influence of the dimensionality of the gesture representation is clearly under-studied. Since "in-the-wild" datasets all use a 2D representation before estimating the 3D poses, we believe that the question of the data representation is fundamentally important.

In this work, we compare two training settings to evaluate the influence of data dimensionality on the performance of a speech-to-gesture generative model by either considering 2D or 3D joint coordinates. We chose to evaluate the effect of the pose representation on a Denoising Diffusion Probabilistic Model (DDPM) because such models have proven their ability to generate natural and diverse gestures aligned with speech and are widely used in the co-speech gesture synthesis field [4, 5, 10, 44, 48, 49]. 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, therefore, formulate the following hypotheses:

- **H1)** The distribution of lifted gestures cannot perfectly match the original 3D gesture distribution.
- **H2)** There is a drop in the consistency between speech and gestures when generating 2D gestures and lifting them to 3D.
- **H3)** The gestures generated in 2D and subsequently lifted to 3D are less diverse than those directly generated in 3D.

To verify these hypotheses, we propose the following contributions:

- We propose an evaluation pipeline to investigate the impact of the dimensionality of the pose representation on the performance of a DDPM [21, 39, 40] for gesture generation. We train a speech-to-gesture DDPM [49] to generate sequences of body poses represented in 2D coordinates which are then lifted to 3D using VideoPose3D [37]. 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 [32, 46, 49]. Specifically, we conduct a series of experiments using a 2D dataset from "in-the-wild" videos and its 3D counterparts.

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

2 RELATED WORK

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 [36].

Deterministic approaches that directly translate speech to gesture sequences have been proposed. To this end, one can choose different neural network architectures such as multi-layer perceptrons [24], convolutional neural networks [19], recurrent neural network [6, 32, 46, 47] or transformers [6, 43]. Yoon et al. [47] 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 to be mapped onto a social robot.

In recent works, there is a notable interest in non-deterministic generative models such as Variational Autoencoders (VAEs) [29] and diffusion models [4, 5, 8, 10, 21, 40, 41, 48] due to their capacity for producing a wide array of gestures.

Specifically, VAEs are designed to encode gestures into a continuous latent space and subsequently decode these latent representations into speech-conditioned movements [29]. Recently, the gesture generation field has particularly focused on Probabilistic Denoising Diffusion Models [4, 5, 8, 10, 21, 40, 41, 48] 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 [4, 5, 22, 49], leveraging both conditional and unconditional generation mechanisms during the sampling process. Alexanderson et al. [4] used Conformers [18] to generate gestures conditioned on behavior style and speech audio. Ao et al. [5] leverage CLIP [38] to encode speech text and a style prompt and use a combination of AdaIN [23] and classifier-free guidance to generate diverse yet style-conditioned gestures from speech. Zhu et al. [49] 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 was trained on the TED Gesture-3D dataset [46], which compiles 3D gestures inferred from 2D poses obtained from monocular video. We hence used their model as a baseline for our study.

Representation and collection of the gesture data. The quality and diversity of the training data are crucial for training co-speech gesture generative models. Early works mostly considered 2D motion data [12, 17, 47]. 2D gestures were typically extracted from "in-the-wild" monocular videos using a third-party pose extractor such as OpenPose [1, 7, 17, 47]. See Table 1 for a list of existing gesture datasets. This collection process allows the gathering of a large amount of training data with numerous different speakers and ensures the diversity and spontaneity of the gestures. However, leveraging such pre-trained pose estimators induces errors resulting in low motion quality, especially for fine-grained motion such as fingers, and limits the pose representation to be two-dimensional. Most of the recent literature [4, 5, 25, 31, 44, 45, 48] focuses on MoCap datasets [13, 14, 16, 26, 31]. MoCap datasets capture detailed 3D movements in a studio, ensuring high-quality motion capture, including detailed finger movements and precise full-body keypoint positions. However, the limited number of speakers in the dataset and the controlled studio environment for data capture diminish the diversity and spontaneity of the training data, consequently affecting the variety of the gestures generated by models trained on such datasets [4, 5, 25, 31, 36, 44, 45, 48]. Recently, multiple works [19, 32, 42, 46, 49] opt for increased diversity and volume of data samples while keeping a 3D representation of gestures, choosing to train their models on datasets of 3D gestures collected from "in-the-wild" videos [19, 32, 42, 46, 49]. To extract 3D body poses from monocular videos, the data collection process typically leverages a pipeline of pose extraction [7] and 2D-to-3D lifting [9, 34, 37]. For instance, the dataset TED Gesture-3D introduced by Yoon et al. [46] leverages VideoPose3D [37] to convert 2D body keypoints extracted by OpenPose [7] to 3D. This dataset is an extension of the previous TED Gesture dataset [47] where the pose where represented in 2D. [1]

In this work, we study how training an audio-driven diffusion model to generate 2D motion data and then post-processing the generated sequences using a 3D lifter impacts the overall quality of the synthesized gestures. Specifically, we use DiffGesture [49] as a gesture generator which obtained state-of-the-art results on the TED Gesture-3D and TED Expressive datasets [32, 46]. For the 2D to 3D lifting model, we employ VideoPose3D [37]. We use the TED Gesture 3D dataset [46] for our evaluation, which is a dataset of 3D gestures extracted from YouTube videos. We aim to highlight the presence of an inductive bias due to using a deterministic 2D-to-3D lifting model, and we evaluate its impact on the generated gestures' diversity, naturalness, and synchrony.

3 METHODOLOGY

The objective of this paper is to analyze how the dimensionality of body pose data influences the behavior of the generation of co-speech gestures. In this section, we present our evaluation pipeline, the model and the dataset we used as well as the chosen evaluation metrics.

3.1 Data

Our analysis is based on the TED Gesture-3D dataset [46]. TED Gesture-3D is a dataset including pose sequences extracted from in-the-wild videos of TED talkers with the corresponding speaker identity, speech, and speech transcription. TED Gesture-3D includes 3D body poses estimated via a combination of a 2D pose extractor from monocular videos [7] and VideoPose3D [37]. 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. 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

[1] To avoid confusion we refer to the 3D version of [46] as *TED Gesture-3D*

Table 1: Speech-Gesture datasets since 2019. The collection methods are described in the rightmost column. It can be either Motion Capture (MoCap) or pose estimation. Abbreviations: *up. upper, rot. rotations coord. coordinates, n.s not specified.* This list is an update of the one provided by Nyatsanga et al.[36].

Dataset	Size	# of speakers	Type of motion data	Finger motion	Collection method
TED Gesture [47]	52.7 h	1,295	2D joint coord.		OpenPose [7]
TED Gesture 3D [46]	97h	n.s.	3D joint coord.		OpenPose [7], VideoPose3D [37]
BiGe [42]	260h	n.s.	3D joint coord.	Yes	OpenPose [7], VideoPose3D[37]
TED Expressive [32]	100.8h	n.s.	3D joint coord.	Yes	OpenPose [7], ExPose [9]
PATS [1]	250h	25	2D joint coord.		OpenPose [7]
SpeechGesture [17]	144 h	10	2D joint coord.	Yes	OpenPose [7]
SpeechGesture 3D [19]	33h	6	3D joint coord.	Yes	OpenPose [7], XNect [34], [15]
BEAT [31]	76h	30	3D joint rot.	Yes	MoCap
Trinity Speech Gesture I [13]	6h	1	3D joint rot.		MoCap
Trinity Speech Gesture II [14]	4h	1	3D joint rot.		Mocap
ZeroEggs [16]	2h	1	3D joint rot.	Yes	MoCap
TalkingWithHands [26]	50h	50	3D joint rot.	Yes	Mocap

considering raw joint coordinates for body pose representation, we follow the approach proposed by Yoon et al. [46] where a body pose is represented as nine directional vectors where each direction represents a bone. The vectors are 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. In this work, 2D pose sequences 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 a co-speech gesture generator on both 2D and 3D settings and 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. To better understand the influence of the motion dimensionality on the relationship between speech and gesture we also perform an ablation on the gesture generator where the speech condition is removed.

3.2.1 Gesture generator. The co-speech gesture generator used as a reference in this study is defined as a DDPM which 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. 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. We used DiffGesture proposed by Zhu et al [21, 49] trained on the TED Gesture-3D dataset [46]. The body poses are represented in $\mathcal{R}^{J\times 3}$ where J is the number of considered body joints. To synthesize diverse and speech-accurate gestures, DiffGesture uses classifier-free guidance [22]. 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.

DiffGesture was first designed to generate 3D gestures, we straightforwardly adapted the architecture to account for 2D body pose sequences by changing the input and output dimensions of the denoising network. Specifically, we removed the depth axis of the body pose coordinates thus considering poses in $\mathcal{R}^{J\times 2}$, we refer to this version as *DiffGesture 2D* and the original version is referred as *DiffGesture 3D*. We trained DiffGesture in four different settings: 2D motion generation, and 3D motion generation, with and without speech condition. For unconditional generation, we simply masked the speech input, forcing DiffGesture to directly generate gestures out of noise, without further guidance. We obtained similar results as Zhu et al. [49] when retraining *DiffGesture 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 [37] to lift 2D pose sequences to 3D. 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 [37] 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 upscaled to 273 frames per second (fps) to exceed the receptive field of VideoPose3D instead of the original 15 fps. The final MPJPE of VideoPose3D was **11.1** on the test set of TED Gesture-3D. We kept the model architecture and training hyper-parameters consistent with the original implementation, except for the learning rate decay, which we found did not enhance the training process.

3.3 Experimental set-up

3.3.1 Comparative settings. We aim to study the impact of the dimensionality of the motion data on the performance of a diffusion-based generative model. To this end, we considered three experimental settings.

1) We want to evaluate the impact of training on 2D motions on the quality of 3D gesture sequences. For this experiment, we define *DiffGesture 2D + VP3D* as DiffGesture trained on 2D motion data whose outputs are then lifted to 3D using VideoPose3D and we compare it to the original DiffGesture [49].

2) To further explore the impact of motion dimensionality on the generated gesture, we also compare *DiffGesture 2D* to DiffGesture but where the 3D generated motion is narrowed to 2D by removing the depth axis, we refer to this model as *DiffGesture 3D->2D*.

3) To evaluate the effect of motion dimensionality on multimodality we proceeded to an ablation of DiffGesture where the input speech is masked during training and inference. In practice, this is equivalent to the original conditional training of DiffGesture with a masking probability p_{uncond} always set to 1. We performed these experiments in both 2D and 3D settings. We refer to these models as *Uncond. DiffGesture 2D* and *Uncond. DiffGesture 3D*.

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

The **Fréchet Gesture Distance** (FGD) defined by Yoon et al. [46] is an adaptation of the Fréchet Inception Distance (FID) [20]. The FGD computes the 2-Wasserstein distance between two distributions leveraging latent features extracted with a pose encoder. Similar distributions will result in a high FGD value. The FGD is defined as follows:

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

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.

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 [28], has been adapted to speech and gestures [30]. 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 [49].

The **Diversity** measure also leverages the latent features extracted with a pose encoder [27]. 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.

Table 2: Experimental results of experiments on the TED Gesture-3D dataset [46]. These results correspond to the experiments (1) and (2) in section 3.3.1. Up arrows indicate that a higher result is better whereas down arrows indicate that a lower result is better. * means reported results from [49]

	TED Gesture		
Methods	FGD ↓	BC ↑	Diversity ↑
Evaluation on the 3D gesture space			
Ground Truth 3D	0	0.702	102.339
DiffGesture 3D [49]	1.370	0.659	102.586
DiffGesture 2D + VP3D	9.833	0.571	92.136
Evaluation on the 2D gesture space			
Ground Truth 2D	0	0.689	112.76
DiffGesture (3D->2D)	1.722	0.645	110.649
DiffGesture 2D	3.279	0.643	112.165
Reported results from [49]			
Ground Truth 3D	0	0.698	108.525
DiffGesture * [49]	1.506	0.699	106.722
Attention Seq2Seq* [47]	18.154	0.196	82.776
Speech2Gesture* [17]	19.254	0.668	93.802
Joint Embedding* [2]	22.083	0.200	90.138
Trimodal* [46]	3.729	0.667	101.247
HA2G* [32]	3.072	0.672	104.322

Table 3: Ablation study of DiffGesture [49] on the TED Gesture-3D dataset [46] where the speech condition has been removed. Up arrows indicate that a higher result is better whereas down arrows indicate that a lower result is better.

	TED Gesture		
Methods	FGD ↓	BC ↑	Diversity ↑
Evaluation on the 3D gesture space			
Ground Truth 3D	0	0.702	102.339
Uncond. DiffGesture 3D	3.288	0.683	98.905
Uncond. Diff Gesture 2D + VP3D	10.009	0.595	93.945
Evaluation on the 2D gesture space			
Ground Truth 2D	0	0.689	112.76
Uncond. DiffGesture (3D->2D)	5.529	0.667	111.599
Uncond. DiffGesture 2D	1.757	0.653	113.304

4 RESULTS AND DISCUSSION

The results of our 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. [49] 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. were obtained using their own encoder. We reported the results of our ablation study in Table 3.

Evaluation of lifted generated gestures. When comparing the results of *DiffGesture 2D + VP3D* to those of *DiffGesture 3D*, we can notice that *DiffGesture 2D + VP3D* performs worse than the original *DiffGesture 3D* in terms of FGD, BC, and diversity. The drop in FGD validates the **H1** hypothesis as the FGD measures the distance between two distributions of gestures. Similarly, the drop in diversity confirms the hypothesis **H3**. We assume that the one-to-many relationship between 2D and 3D keypoints is mostly responsible for the performance drop of *DiffGesture 2D + VP3D* for the FGD and diversity. 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 and low diversity of the gestures generated in *DiffGesture 2D + VP3D*. There is a drop in BC between *DiffGesture 3D* and *DiffGesture 2D + VP3D* which corroborates the **H2** hypothesis. We think that post-processing 2D gestures using VideoPose3D tends to over-smooth the resulting 3D gestures, reducing the number of kinematic beats.

Evaluation of the quality of gestures generated in 2D. When evaluating in the 2D motion space, *DiffGesture 3D->2D* performs better than *DiffGesture 2D* in terms of FGD. Hence, training DiffGesture 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 more detailed compared to the 2D version. The BC and diversity scores do not seem to be influenced by the dimensionality of the gestures that were used to train the generative model.

Ablation of the speech condition. The results of the ablation study are reported in Table 3. *Uncond. DiffGesture 3D* shows worse FGD and diversity than *DiffGesture 3D*. This outcome demonstrates that the speech condition helps *DiffGesture 3D* to synthesize gesture sequences that are both diverse and similar to the target distribution. As depicted in Table 2, *DiffGesture 2D* shows a higher FGD value than *DiffGesture (3D->2D)* and *Uncond. DiffGesture 2D*, suggesting that the speech condition reduces the quality of the gestures generated in 2D. We are led to think that the integration of a speech condition adds ambiguity to the generation of 2D gestures. A 2D gesture can correspond to multiple 3D gestures, and there is a one-to-many relationship between 3D gestures and speech. Therefore reducing the gesture dimension may lead to a more pronounced one-to-many relationship between gestures and speech. Such findings suggest that incorporating a speech condition enhances the diversity of gestures generated by *DiffGesture 3D*. In contrast, *Uncond. DiffGesture 2D* manages to produce samples that maintain a diversity level comparable to that of the ground truth. This indicates that 2D gestures can be synthesized without the speech condition while still closely aligning with the target distribution in terms of FGD and diversity. This needs to be more studied with other models and other databases.

It is important to note that the BC score is almost unaffected when removing the speech condition as depicted in Table 3. More experiments involving other datasets and other models are needed to verify this behavior and are left for future work.

Our objective evaluation validated our three hypotheses **H1**, **H2** and **H3** (c.f section 1). We can conclude that generating gestures in 2D and subsequently lifting them to 3D significantly impairs the overall quality of the gestures in terms of FGD, BC, and diversity. We believe that the 2D-to-3D conversion is mainly responsible for the performance drop. However, as gestures generated in 2D show worse FGD than gestures generated in 3D and narrowed to 3D, we think that the 3D representation of gestures favors the training of the generative model. This also shows that, even when generating 2D gestures, it is better to train DiffGesture to generate 3D gestures and narrow them to 2D as post-processing.

5 CONCLUSION

In this study, we explored how training a diffusion-based co-speech gesture generator with 2D data affects its performance. We introduced a pipeline that pairs a gesture generator with a 2D-to-3D lifting model, specifically VideoPose3D [37]. Our findings reveal that using this pipeline negatively impacts overall performance. Using a deterministic lifting model, such as VideoPose3D, reduces the diversity of the generated gestures. 3D gestures lifted from 2D generated gestures are also less similar to the target 3D gesture distribution in comparison to gestures generated directly in 3D. However, using such a 2D-to-3D lifter remains a feasible strategy when retraining the model with 3D data (e.g. MoCap data) is not an option. We also found that DiffGesture faces challenges in accurately modeling the relationship between 2D motion data and speech, a problem that does not occur with 3D motion data. We attribute this issue to the inherent ambiguity of 2D coordinates compared to 3D, which hinders the model's ability to synchronize the two modalities effectively. However, further evaluations on other datasets are needed to confirm this tendency. 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. For future research, we plan to conduct a similar analysis using a Mocap dataset. Additionally, we aim to employ an alternative generative model to validate our findings.

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 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

[3] 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

[4] 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

[5] 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

[6] 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

[7] 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).

[8] 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

[9] 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

[10] 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

[11] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2017. RMPE: Regional Multi-person Pose Estimation. In *ICCV*.

[12] 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

[13] 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

[14] 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

[15] 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

[16] 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]

[17] 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

[18] 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]

[19] 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

[20] 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

[21] 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.

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

[23] 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

[24] 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

[25] 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.

[26] 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

[27] 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

[28] 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

[29] 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

[30] 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

[31] 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]

[32] 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.

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

[34] 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

[35] 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

[36] 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

[37] 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)*.

[38] 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

[39] 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

[40] 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

[41] 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

[42] 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).

[43] 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

[44] 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]

[45] 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]

[46] 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).

[47] 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

[48] 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

[49] 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 8th March, 2024