AQ-GT: a Temporally Aligned and Quantized GRU-Transformer for Co-Speech Gesture Synthesis

Hendric Voß
hvoss@techfak.uni-bielefeld.de
Social Cognitive Systems Group
Bielefeld University
Germany

Stefan Kopp
skopp@techfak.uni-bielefeld.de
Social Cognitive Systems Group
Bielefeld University
Germany

ABSTRACT

The generation of realistic and contextually relevant co-speech gestures is a challenging yet increasingly important task in the creation of multimodal artificial agents. Prior methods focused on learning a direct correspondence between co-speech gesture representations and produced motions, which created seemingly natural but often unconvincing gestures during human assessment. We present an approach to pre-train partial gesture sequences using a generative adversarial network with a quantization pipeline. The resulting codebook vectors serve as both input and output in our framework, forming the basis for the generation and reconstruction of gestures. By learning the mapping of a latent space representation as opposed to directly mapping it to a vector representation, this framework facilitates the generation of highly realistic and expressive gestures that closely replicate human movement and behavior, while simultaneously avoiding artifacts in the generation process. We evaluate our approach by comparing it with established methods for generating co-speech gestures as well as with existing datasets of human behavior. We also perform an ablation study to assess our findings. The results show that our approach outperforms the current state of the art by a clear margin and is partially indistinguishable from human gesturing. We make our data pipeline and the generation framework publicly available.

CCS CONCEPTS

• **Human-centered computing → Interactive systems and tools**; *Empirical studies in interaction design*; **HCI theory, concepts and models**; • **Computing methodologies → Neural networks**; **Learning latent representations**; *Unsupervised learning*.

KEYWORDS

machine learning; deep learning; co-speech gesture; gesture synthesis; multimodal data; quantization; transformer; gated recurrent units; temporal alignment

ICMI '23, October 09–13, 2023, Paris, France

ACM ISBN 979-8-4007-0055-2/23/10…$15.00
https://doi.org/10.1145/3577190.3614135

ACM Reference Format:
Hendric Voß and Stefan Kopp. 2023. AQ-GT: a Temporally Aligned and Quantized GRU-Transformer for Co-Speech Gesture Synthesis. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23), October 09–13, 2023, Paris, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3577190.3614135

1 INTRODUCTION

Effective communication between humans encompasses various modalities, such as speech, facial expressions, and bodily gestures. The ability to comprehend and generate these multimodal signals enables us to engage in meaningful and nuanced conversation and is commonly employed in everyday interactions [10, 49]. Consequently, researchers have devoted considerable efforts to accurately process, interpret, and generate these communicative cues to facilitate natural and seamless human-machine interaction [33].

Generating non-verbal cues, such as body language and hand gestures, to accompany spoken language poses a particularly challenging task for embodied agents. In recent years, machine learning approaches have been developed to predict gestures based on a given linguistic input. Such co-speech-driven gesture synthesis frameworks achieve impressive outcomes in generating realistic gestural motion that aligns with a given spoken utterance or text but still face difficulties in generating movements that meet the communicative effectiveness and richness of human gesture [40].

In this paper, we introduce a framework that aims to learn a discrete latent space of gesture sequences. The proposed AQ-GT approach utilizes a combination of Gated Recurrent Units (GRU) and Transformers to learn an intermediate representation of vectors, which, through a temporally aligning network architecture, constructs new novel co-speech gestures. We thereby make several contributions to the field of gesture synthesis. Firstly, we propose a novel data processing pipeline that enables the tracking of full 3D body joints for arbitrary monocular videos. This pipeline enables the automatic acquisition of precise gesture information, facilitating the learning of natural gestures. Secondly, we introduce a novel approach to co-speech gesture synthesis. This approach leverages a discrete latent space of gestures and combines the strengths of gated recurrent units (GRUs) and Transformers to generate naturalistic gestures that are closely aligned with accompanying speech. Thirdly, we present a thorough evaluation of our framework, using both objective and subjective measures, demonstrating that it outperforms current state-of-the-art approaches and generates gestures that are partially indistinguishable from human gesturing. Finally, we present an ablation study that isolates and analyzes the contribution of single components of the proposed model to the overall output quality. A video with examples of generated co-speech gestures is made available online[1].

2 RELATED WORK

The generation of co-speech gestures has been a long-standing goal for socially interactive multimodal agents such as embodied conversational agents, 3D avatars, or socially assistive robots. Early approaches mainly relied on rule-based models that were crafted by hand, based on empirical or theoretical insights. One such example is the Behaviour Markup Language [33] that allowed for the generation of multimodal behavior by specifying function-to-form mappings in an XML format. Similarly, the BEAT toolkit [11] employed expandable rule sets, that were derived from linguistic and contextual analyses of human conversational behavior. Later methods applied dynamic learning approaches that mainly used kernel-based probability distribution models or stochastic Markov models to learn motion patterns from motion capture sequences or high-level structures from temporal segmentation [9, 22]. Due to their high degree of customization and low computational requirements, such techniques have also been used in commercially available products such as the Pepper and Nao robots [41].

In recent years, the development of deep machine learning has significantly impacted the field of multimodal behavior processing and virtual agent behavior generation [12, 13, 20, 31]. As a result, data-driven approaches have become the primary method for generating complex behaviors in various human-agent and human-human interaction tasks [24, 45, 55].

During the early stages of deep learning, the main objective was to train virtual agents to demonstrate natural behaviors and create visually appealing co-speech gestures, without the need for time-consuming hand-crafted techniques. The main focus was on using a single input modality. Initial attempts such as a study conducted by Fan et al. [18] utilized a Bidirectional Long Short-Term Memory (LSTM) model to generate new sequences of co-speech gesture solely from text input and a given initial gesture input. Similarly, Ferstl and McDonnell [21] employed a recurrent neural network with an encoder-decoder architecture, trained on prosodic speech features to generate short sequences of motion, while Ginosar et al. [25] proposed a model to generate gestures from speech audio data along with an initial pose.

More recent work includes multiple input modalities, each exhibiting different dimensionalities and structures, and has prompted the development of more complex data-driven methodologies. In their investigation of co-speech gestures, Holden et al. [29] employed feed-forward networks, training them with a cyclical function to generate gestures. In contrast, Henter et al. [28] applied a distinct deep-learning technique, implementing a glow network with invertible 1x1 convolutions for co-speech gesture generation. Ling et al. [36] adopted a related approach, utilizing a variational autoencoder combined with deep reinforcement learning for goal-oriented control of co-speech gestures.

Current research highlights the potential of general adversarial networks (GANs) to produce highly realistic results, with the capacity to merge multiple modalities cohesively, thus enhancing the representation of human communication. Ahuja et al. [4] devised a method for learning style embeddings, enabling the generation of unique gesture styles for individual speakers. This approach enables a diverse range of gesture variations, which could be dynamically adjusted during the generation process. To integrate audio, text, and speaker identity within a GAN framework, Yoon et al. [52] designed a technique that effectively generates hand gestures closely resembling those observed in natural human communication. This method produces gestures that reflect the semantic meaning and rhythm of spoken words, thus simulating authentic human interaction in virtual environments. Likewise, Ao et al. [6] recently introduced an approach to gesture synthesis that accounts for both rhythmic and semantic aspects. Specifically, they integrated a rhythm-based segmentation pipeline with neural embeddings of speech and motion, drawing on insights from linguistic theory.

3 DATA

Several datasets are currently available for training co-speech gesture generation algorithms. However, they often have certain limitations [40]. For instance, many datasets contain only a small number of gesture sequences, while others provide solely 2D tracking of body joints or do not incorporate tracked hand information such as finger movements. To the best of our knowledge, only two datasets provide 3D body joints and are sufficiently large for our purposes: the BEAT dataset [37] and the Talking with Hands dataset [34, 54]. Unfortunately, both of these datasets cover only a limited number of speakers, which may restrict a model's ability to generalize across speaker styles. Therefore, we decided to collect and build our own dataset, called the BiGe dataset.

3.1 Data Acquisition

In order to successfully train co-speech gesture generation algorithms, the given data must meet certain requirements. In particular, large amounts of noise or significant variability in the data can have an adverse effect on the generated gestures. Since we aimed to capture full-body gestures and especially hand gestures of the presenter, we needed to ensure that the entire person is visible. Additionally, the person had to face the camera to avoid obscuring their gestures. Given the challenge of eliminating camera movements, we also required a video source with stationary cameras. To meet these criteria, we opted to use the same data source as Yoon et al. [53], namely the official TED channel on Youtube [1]. To diversify our dataset with multi-language presentations, we also selected the TEDx Talks Youtube channel which features international presentations in different languages [2]. All videos are licensed under the "CC BY – NC – ND 4.0 International" license. In total, we collected 6945 videos with the highest possible video and audio format, as well as automatically and manually annotated subtitles. Next, we filtered out any videos that did not have available subtitles, featured a resolution lower than 1280x720 pixels, or lacked subtitles with word-for-word timing information. After the filtering process, we were left with 4327 videos for our dataset, with a total length of 1021 hours.

[1] https://vimeo.com/823756031

Table 1: Overview of the BiGe dataset and the Shots of Interest (SoT)

Number of videos	4327
Total length of videos	1021 h
Videos with SoT	2756
Total length of SoT	260.6 h
Average length of SoT	17 s
Number of SoT	54.360 (20 per video on average)

3.2 Data Processing

To extract body joint keypoint information from the collected videos, a modified data processing pipeline was created, based on the work of Yoon et al. [53]. In order to track 59 3D full-body joints for each individual in the videos, we employed a combination of the AlphaPose library [19], the VideoPose3D model [42], and the MediaPipe hands model [56]. The pipeline first utilized a YOLOv3 model [44], through the AlphaPose library, to detect all persons in each frame of the video. Then, the 2D position of each joint was estimated through the implementation of a FastPose model [57]. The tracked joints, excluding the hands, were subsequently transformed into 3D coordinates by employing the VideoPose3D model [42]. Finally, the MediaPipe Hands model was used to determine the 3D positions of the hands by elevating the previously tracked 2D hand joints to 3D coordinates. After tracking, we cut the videos into multiple clips based on detected camera cuts. We then determined Shots of Interest (SoT) for the entire dataset, which are any clips that fulfill the removal conditions established by Yoon et al. [53]. Namely, all clips were removed in which joints were missing, occupied less than thirty percent of the image area, or in which joints were static for the entire duration of the clip. As the tracking pipeline provided us with a confidence score for each joint, we used this score to further refine our selection of SoT by removing any SoT where the median value falls below a threshold τ, which we set to 0.05. Table 1 gives an overview of the characteristics of the resulting BiGe dataset: It comprises 2756 videos, out of which we extracted 54.360 SoTs with an average length of 17 seconds. To create our dataset, we split these data into 2156 training videos, 300 validation, and 300 test videos.

4 AQ-GT ARCHITECTURE

The overall structure of the proposed model architecture is shown in Figure 1. The model consists of two parts: The first one (colored in yellow) is a quantized variational autoencoder that yields a structured embedding space representation of sequential data, both for the raw audio as well as the gestural data. The second one (colored in green) is the model for autoregressively generating gesture frames from audio, text, and speaker ID input. This part comprises a preprocessing and integration of the input data, a GRU-Transformer, a network for speaker prediction, and a temporal aligner. We emphasize the use of the temporal aligner as well as the quantized latent space for the GRU-transformer in the name of the model, "AQ-GT". Each of the components and their training is described in the following.

4.1 Quantization in Discrete Latent Space

Drawing inspiration from the work of Esser et al. [16], we utilize a combination of a Vector Quantized Variational Autoencoder (VQ-VAE) [46] and a Generative Adversarial Network (GAN) [26, 38]. We employ the VQ-VAE-2 model [43] in conjunction with a Wasserstein Generative Adversarial Network with Divergence penalty (WGAN-div) [51] to learn a discrete latent space of concise gesture sequences. The original VQ-VAE model [46] consists of two main components, an encoder and a decoder, both of which employ a shared "codebook". The encoder maps observations onto a sequence of discrete latent variables, while the decoder reconstructs the original observations using these discrete variables. To achieve this, the encoder uses a non-linear mapping from the input space, x, to a vector $E(x)$, which is then quantized based on its distance to the prototype vectors $c_i, i \in 1...C$ in the shared codebook. The VQ-VAE-2 model introduced by Razavi et al. [43] can be seen as an extension of the VQ-VAE, which incorporates a hierarchical structure. In this model, a codebook is learned for each hierarchical level, which we denote as C_{top}, C_{bottom}, and the combination of both codebooks as C. The loss function used for training the VQ-VAE-2 is given by:

$$\mathcal{L}_{vq}(\mathbf{x}, D(\mathbf{z})) = ||\mathbf{x} - D(\mathbf{z})||_2^2 + ||sg[E(\mathbf{x})] - \mathbf{z}||_2^2 + \alpha||sg[\mathbf{z}] - E(\mathbf{x})||_2^2 \tag{1}$$

where E represents the Encoder, D represents the Decoder, sg refers to the stop-gradient, x represents the data input, z represents the model output, and α is a hyperparameter that controls the impact of the regularisation term. The first term in Equation 1 measures the reconstruction error between the input and the output of the decoder, while the second and third terms encourage the model to use the codes from the codebook and to minimize the distance between the encoder output and the selected codes, respectively.

One notable issue encountered in quantized models is the occurrence of codebook collapse, wherein a substantial portion of the codebook is disregarded, and only a limited subset is utilized. Although the VQ-VAE-2 model enhances the stability of the codebook by incorporating hierarchical learning, this problem may still manifest. Recent work by Esser et al. [16] has demonstrated that the combination of a VQ-VAE with a GAN results in a perceptually more diverse codebook and more stable training. In light of these findings, we add a discriminator for the adversarial training of the VQ-VAE-2 model, which processes the decoded output of the VQ-VAE-2 and outputs an unbounded scalar value. Leveraging the insights from the research of Wu et al. [51], we utilize a Wasserstein divergence objective for the training which enhances the fidelity and stability of the output, as compared to Wasserstein loss with gradient penalty [27]. We call this combination of the VQ-VAE-2 and the WGAN-div the VQ-VAE-2_{wdiv} model.

The loss function for our discriminator is given by:

$$\mathcal{L}_{D_{wdiv}}(\mathbf{x}, D(\mathbf{z})) = Dis(\mathbf{x}) - Dis(D(\mathbf{z})) + \delta\|\nabla_{\hat{\mathbf{x}}} Dis(\hat{\mathbf{x}})\|^p \tag{2}$$

where Dis is the discriminator and δ is a hyperparameter that determines the strength of the divergence penalty. The first part of the loss $Dis(\mathbf{x}) - Dis(D(\mathbf{z}))$ evaluates the difference between the real sample $\mathbf{x}$ and the output of our VQ-VAE-2 $D(\mathbf{z})$. The second term $\delta\|\nabla_{\hat{\mathbf{x}}} Dis(\hat{\mathbf{x}})\|^p$ is the divergence penalty, which encourages the generated sample $D(\mathbf{z})$ to be close to the real data distribution.

The generator loss function is simply given by

$$\mathcal{L}_{G_{wdiv}} = Dis(D(\mathbf{z})) \tag{3}$$

which minimizes the result of the discriminator output. Therefore, our complete loss function for the generator of the VQ-VAE-2 model is the combination of the $\mathcal{L}_{vq}$ and the $\mathcal{L}_{G_{wdiv}}$ loss functions:

$$\mathcal{L} = \beta\mathcal{L}_{vq} + \gamma\mathcal{L}_{G_{wdiv}} \tag{4}$$

We additionally introduce the hyperparameter β and γ to adjust the contribution of both loss functions.

4.2 Input Modalities

In our proposed framework, we leverage a combination of three input modalities (initial or previous gesture position, text, and audio), along with a speaker ID. Each of these modalities is pre-processed in a specific way. To facilitate the flexibility of our framework, we allow for arbitrary lengths of initial information and generated gesture sequences. To this end, we introduce two parameters: N represents the number of frames in the initial information, which we set to four; M denotes the number of frames generated by the gesture sequence, which we set to 30. To ensure consistency in our pre-processing pipeline, we sample gestures at a rate of 15 frames per second, resulting in a 2-second sequence of newly generated gestures.

For the pre-processing of previous gesture frames g, we train a VQ-VAE-2_{wdiv} model on the entire gesture training set, with a learning rate of $2e-5$ and a batch size of 256. Henceforth, we will call this model VQ_G. For the introduced hyperparameters we perform a hyperparameter Bayesian search and, based on the results, set α, β, and γ to 0.25, 1, and 1, respectively. After training for 200 epochs we restore the weights with the lowest validation loss and freeze the model weights. Instead of using the absolute position in 3D space, we use the relative position of the parent bone as our input and output vectors. Finally, we construct an embedding space for each of the 512 possible codebook vectors and, based on the indices of the codebooks, return the selected embeddings.

The text pre-processing uses the pre-trained BERT model [15], which converts the given text tex to tokens and extracts a high-level representation of it. For our framework, we return the representation vector with a fixed length of up to 300 tokens.

The audio processing pipeline employs three distinct models to effectively process and extract meaningful features from the raw audio data a. These models include the Wav2Vec2 framework [7], our custom pre-trained VQ-VAE-2_{wdiv} model, and an Onset Encoder model [35]. The Wav2Vec2 framework is a semi-supervised model trained on speech input, which we use to process the raw audio data and obtain the last hidden state of the model. However, as the focus of the Wav2Vec2 framework is primarily on reconstructing speech from given audio data, it may result in the loss of crucial prosodic features during training. To address this, we additionally train our own VQ-VAE-2_{wdiv} model on 0.25-second segments of downsampled 16kHz raw audio data. We collect 1500 hours of audio data from randomly selected YouTube videos, which we split into 1200 hours of training data and 150 hours of validation and test data. We use the same learning rate, batch size, and hyperparameter configuration as for the VQ_G. As the dataset is far larger than our gesture dataset, we train the model for 100 epochs and afterward freeze the weights with the best validation loss. We refer to this model as the $\mathrm{VQ_A}$ model. During the training of the entire framework, we process the audio with this model and return the vector E_x.

Recent studies have highlighted the significance of sound onset in co-speech gesture generation [6, 35]. By leveraging this information, the model can synchronize the generated gestures with the audio input more effectively, resulting in a closer adherence to the rhythm of the speech. To this end, we adopted the Onset Encoder model proposed by Liang et al. [35]. This model first uses a filter to detect the onset of the audio and then decouples the speech input into semantically relevant and semantically irrelevant cues. Unlike the other two models, the weights of the Onset Encoder model are not frozen and updated during the framework training.

To account for different speaker styles, we use the speaker identity sp and create an embedding space for each possible speaker in the dataset. After passing the embeddings through a multilayer perceptron (MLP), we employ the "Reparameterization Trick" as proposed by Kingma and Welling [32] to establish a multivariate normal distribution within the latent space. This technique can be leveraged to generate new, unseen gesture styles, as demonstrated in prior works [4, 23]. In addition to the embedding space, we define a "Speaker Decision Network" consisting of an MLP network, which predicts the current speaker identity sp^* given the input vector W and returns a corresponding embedding layer based on the result. We train this network using a Softmax contrastive loss as described in Section 5.

4.3 GRU-Transformer

The data-based generation of co-speech gestures requires the modeling of temporal dependencies in sequential gesture data. A popular choice for this are Gated Recurrent Units (GRUs) [14] because of their high versatility and ease of use [6, 35, 39, 52]. The particular effectiveness of GRUs lies in their ability to capture non-linear dependencies, which is particularly important in modeling sequential gesture data that often exhibit complex temporal dynamics. However, the performance of GRUs is often hindered by challenges such as slow convergence rates, limited learning efficiency, and vanishing gradient problems [50]. At the same time, Transformer architectures [47] have been introduced to learn long-term dependencies in structured input data. Despite their potential, transformers are not commonly utilized in co-speech gesture generation for their high computational demands and limited capabilities of processing complex, non-linear gesture data [3, 8, 58].

We propose to fuse these two approaches to comprehensively and robustly model the complex temporal (sequential) structure of co-speech gestures. In particular, we use the Transformer to capture the broader global dependencies within the input sequence, while the GRU enables the learning of non-linear, localized dependencies. As shown in Figure 2, we propose a GRU-Transformer model that processes the data input both with a GRU and a transformer pipeline. The transformer blocks are comprised of a multi-headed self-attention block with temporal information pre-processing, followed by a multilayer perceptron (MLP) to reduce the dimension of the output. As the maximum number of heads has to be divisible by the input, we set the number of heads to $n_{heads} \leftarrow max(\{n \in$

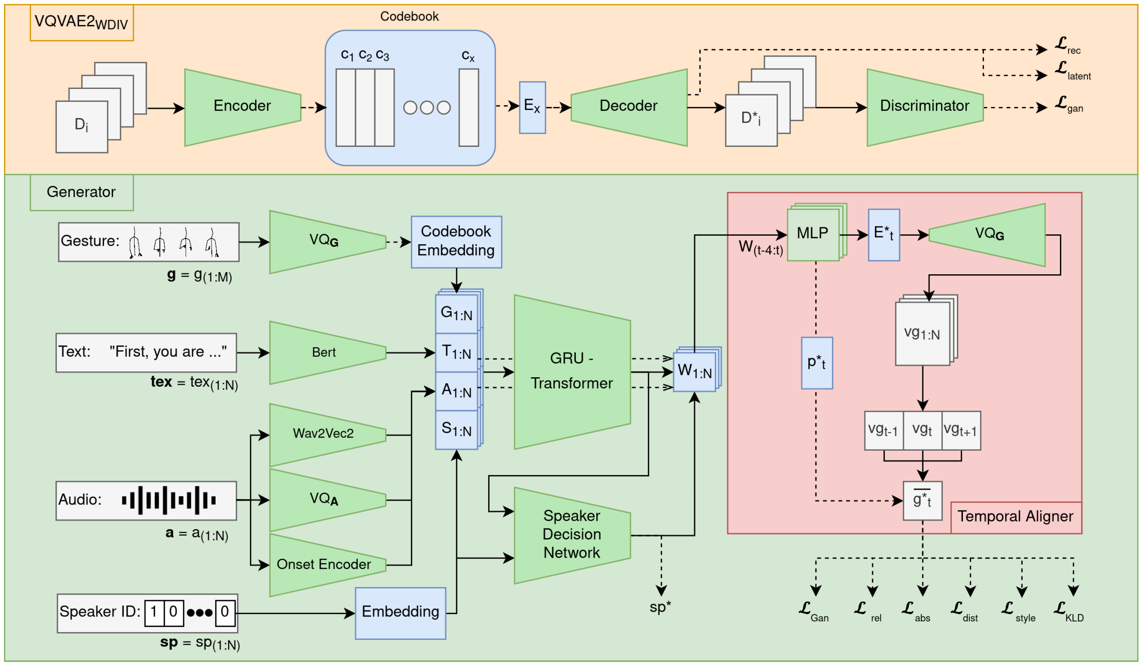

Figure 1: Overview of the AQ-GT model. Top: the VQVAE2 network with the added GAN discriminator. Bottom: The generator network with input modalities, pre-processing, GRU-Transformer and Temporal Aligner network (in red).

$1, 2, 3, ..., 12 : n \mod n_i = 0\})$, where n_i is equal to the input dimension. The GRU_1 block is comprised of a single GRU layer followed by a layer normalization. For the GRU_2 block we use four stacked GRU layers. Using skip connections, the intermediate output of all blocks is added to the output vector.

4.4 Temporal Aligner

One of the challenging tasks in generating co-speech gestures is creating temporal synchronicity between the relevant verbal and gestural events. Although both transformers and GRUs are able to learn temporal information from the input, in practice these layers still struggle to create convincing gestures that are free of sudden

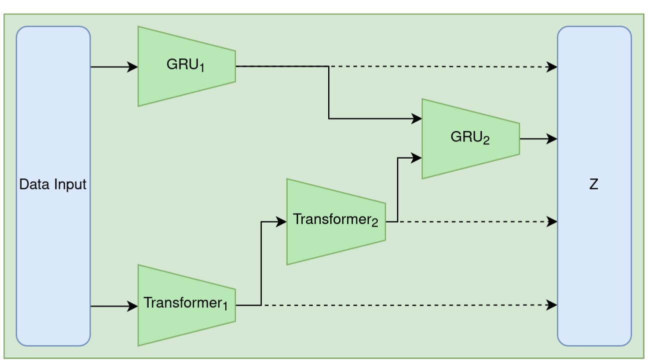

Figure 2: Architecture of GRU-Transformer component. The input and output data are colored in blue; model layers are colored green; dotted lines represent skip connections.

unwanted gesture artifacts [17, 39]. Therefore, we add a temporal aligner network, which learns the temporal dependencies of the previously generated frames and aligns the reconstructed sequences to remove sudden noise in the generated gestures. The structure of our Temporal Aligner is visualized in Figure 1 (red square). To construct the aligner, we reuse the VQ_g, described in Section 4.2. After receiving the combined vector W, we apply a sliding window to iterate over the preceding three frames and concatenate them with the current frame and the result of the previous iteration. We then pass the combined vector into an MLP and save the result for the next iteration. Formally, we compute the temporal vector

$$f_c(\mathbf{W}_t) = MLP(f_c(\mathbf{W}_{t-1}) + \bigg\Vert_{i=0}^{-3} \mathbf{W}_{t-i}) \tag{5}$$

From this, we reconstruct the vector E_x of the network VQ_G, denoted as E_t^*, for each time step t. Using this vector E_t^* we reconstruct the next gesture frame vg_t with the VQ_G network. As the VQ_G is trained on a sequence of four frames, we compose the gesture vector g_t^* out of the third element of vg_{t-1}, the second element of vg_t, and the first element of vg_{t+1}. Formally, we perform:

$$g_t^* = \frac{1}{3}\sum_{i=0}^{2} vg_{[t-1+i,2-i]} \tag{6}$$

Once the vector g_t^* has been constructed, we average it with the vector p_t^*, which is computed using an MLP given the output of the temporal vector $f_c(\mathbf{W}_t)$. This ensures that the network can correct noise remaining in the reconstruction of the vector g_t^*.

5 TRAINING THE GENERATOR MODEL

The generator model is trained as illustrated in Figure 1. In the model, the separately pre-processed input data is combined into an input vector for the GRU-Transformer. Using its output we predict the current speaker identity using a Speaker Decision Network. We combine the output of the GRU-Transformer, the Speaker Decision Network, and the input text and audio into a vector W. Using this vector, we synthesize the gesture G^* through a Temporal Aligner.

We train this model with the same WGAN-div approach as described in Section 4.1. For the training of the framework, we define multiple loss functions to guide the different parts of the training. When calculating a distance between two sets of vectors, we use the Huber Loss [30], which we denote as HL.

First, we define the absolute and relative reconstructive loss. For this, we compare the Huber distance given by the output of our generator model compared to the ground truth, as well as the absolute Huber distance given by reconstructing the skeleton structure of the pose data, both for the generated gestures and the ground truth. Formally, we define the reconstructive loss as the average Huber distance between the skeleton of the generated poses and the ground truth skeleton:

$$\mathcal{L}_{abs} = \frac{1}{N}\sum_{i=1}^{N} HL(Rc(g_i), Rc(g_i^*)) \tag{7}$$

$$\mathcal{L}_{rel} = \frac{1}{N}\sum_{i=1}^{N} HL(g_i, g_i^*) \tag{8}$$

where Rc is the reconstruction function for the absolute joint position, N is the sample size and g_i^* is the gesture vector from the generator model.

To ensure correct gestures involving both hands, we define a loss that measures the pairwise distance between all bones in the left and right arms and compares it to the ground truth. Formally, we define the distance loss as:

$$\mathcal{L}dist = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} HL(\mathrm{g_i}_{l,j} - \mathrm{g_i}_{r,j}, \mathrm{g_i^*}_{l,j} + \mathrm{g_i^*}_{r,j}) \tag{9}$$

where K is the number of arm joints.

We also add a speaker contrastive loss, employing a combination of the speaker style loss proposed by Yoon et al. [52] and our speaker decision softmax loss. Our training approach involves generating gestures based on input T and the original speaker identity $\boldsymbol{sp}1$, while maximizing the distance for a random speaker identity $\boldsymbol{sp}_2$ with the same input T. Formally, we define:

$$f_{st} = \mathrm{CE}(\mathbf{G}(\boldsymbol{T}, \boldsymbol{sp}_1), \mathrm{G}(\boldsymbol{T}, \boldsymbol{sp}_2)) \tag{10}$$

$$f_{sp} = \mathrm{HL}(\mathbf{SPD}(\boldsymbol{T}, \boldsymbol{sp}_1), \mathbf{SPD}(\boldsymbol{T}, \boldsymbol{sp}_2)) \tag{11}$$

$$\mathcal{L}_{\text{style}} = -\mathbb{E}\left[\min\left(\frac{f_{st} + f_{sp}}{\|\boldsymbol{sp}_1 - \boldsymbol{sp}_2\|_1}, \epsilon\right)\right] \tag{12}$$

where $SPD(T, sp)$ is the softmax output of the Speaker Decision Network and CE is the Cross-Entropy Loss. To make it easier to sample different styles, we use the Kullback-Leibler (KL) divergence loss ($\mathcal{L}_{KLD}$) between the feature space of the speaker identity and $N(0, I)$ to assume a Gaussian distribution for the style embedding.

Finally, we need to account for the position and temporal dependencies of joint vectors in natural gesturing. To that end, we include the first to sixth derivatives of the joint position vector with respect to time in a loss function. Formally, we define:

$$f_{dop}(g) = \frac{1}{6}\sum_{i=1}^{6}\frac{d^i g}{dt^i} \tag{13}$$

$$\mathcal{L}_{DoP} = \frac{1}{N}\sum_{i=1}^{N} LH(f_{dop}(g_i), f_{dop}(g_i^*)) \tag{14}$$

The overall loss function combines the components as follows:

$$\begin{aligned}\mathcal{L}total = \pi_1\mathcal{L}_{gan} + \pi_2(\mathcal{L}_{rel} + \mathcal{L}_{abs})\\ + \pi_3\mathcal{L}dist + \pi_4\mathcal{L}_{\text{style}} + \pi_5\mathcal{L}_{KLD} + \pi_6\mathcal{L}_{DoP}\end{aligned} \tag{15}$$

with $\pi_1, \pi_2, \pi_3, \pi_4, \pi_5, \pi_6$ being unbounded hyperparameters to adjust the loss strengths. We performed a hyperparameter search for all parameters of our framework. Based on the results, we use a batch size of 190, π_1 of 2, π_2, π_3, π_4, and π_6 of 20 and π_5 of 0.004. We use an unlimited number of epochs combined with a stop criteria to finish the training and restore the best-performing weights if there is no decrease in the validation loss after 50 epochs. Subsequently, we train our model for 368 epochs and restore the weights from epoch 318. We adopt a step-wise decreasing learning rate schedule to optimize the performance of our model. Specifically, we initialize the learning rate to $1e-4$. Every 20 epochs we multiply the learning rate by a factor of 0.75, to a minimum of $1e-5$. To reduce the possibility of overfitting, we use dropout with a step-wise increment of 0.05 every 25 epochs, up to a maximum of 0.3, between every layer and for the input of the network.

Table 2: Comparison of the proposed model with three state-of-the art models for the BiGe dataset. For FGD and MAJE scores lower is better; for Diversity distance higher is better.

Methods	FGD ↓	Diversity ↑	MAJE ↓
Ground Truth	0.00	42.128	0.00
Trimodal [52]	2.023	38.980	0.0119
SEEG [35]	2.018	38.660	0.0121
HA2G [39]	1.346	41.332	0.0096
Ours (AQ-GT)	**0.3977**	**42.885**	**0.0078**

6 EVALUATION

To evaluate the proposed framework, we compare the generated gestures with what other state-of-the-art frameworks produce, both using available objective measures and with a user study to investigate subjectively perceived differences between the generated output.

6.1 Objective Evaluation

To ensure a fair comparison, we trained three available frameworks on our large BiGe dataset and compare the results. Specifically, we evaluate the Trimodal framework [52], the SEEG framework [35], and the HA2G framework [39] against our proposed model. Regrettably, we were unable to compare our model to the model proposed by Ao et al. [6] as we did not get access to their implementation.

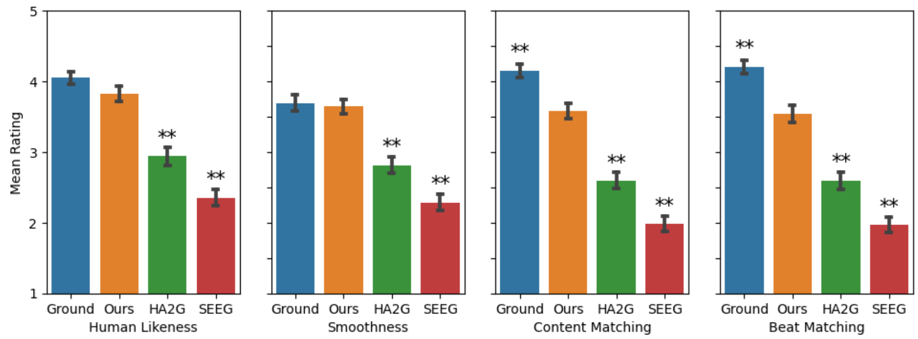

Figure 3: The results of the subjective evaluation study. The ground truth is marked in blue. Our framework is marked in yellow. The HA2G framework is marked in green. The SEEG framework is marked in red. Asterisks indicate significant effects in comparison to our proposed framework (* : p < 0.05, ** : p < 0.005).

We adhered to the original configurations for each framework and only altered the output dimension to accommodate the higher joint dimensionality in our dataset. We measured the performance of each model using the Fréchet Gesture distance (FGD) [52], the Diversity distance [39], and the mean of the absolute errors between the generated joint positions and ground truth (MAJE). Please note, that we only report the MAJE as a legacy value, as the MAJE has shown little to no correlation to gesture quality [52]. As the Diversity distance exhibited high variance, we report the average of one thousand calculations. The results are reported in Table 2.

The experimental results demonstrate that our model surpasses the existing frameworks in all evaluated metrics. Notably, the AQ-GT framework achieved a three times lower FGD than the HA2G. For the Diversity distance, our framework surpasses the distance of all other frameworks and even exceeds the human ground truth. It is important to note, that the diversity measures the absolute distance between random pairs of gesture samples based on their latent representations. Our evaluation result, while encouraging, may thus imply the possibility that the discrete representation of gestures could introduce outliers that are mapped in different regions of the latent space, therefore increasing the Diversity distance.

To compare our model to a wider range of published frameworks, we additionally trained the model on the TED gesture dataset [53]. Our model architecture and configuration were kept unchanged, only the input and output dimensions were adjusted to accommodate the differing body joint dimensions. To prevent an unfair advantage, we also pre-trained a new VQ_G model on the TED dataset and incorporated it during the training process of our framework. The results are given in Table 3, again illustrating a significant improvement in FGD compared to all other approaches. Additionally, our framework achieves the second-highest Diversity distance score, closely trailing the HA2G framework.

Table 3: Comparison of the proposed model with state-of-the art systems on the TED dataset. The MAJE was omitted, to ensure a fair comparison. SEEG and Rythmic Gesticulator did not disclose their diversity score. For FGD lower is better; for the Diversity distance higher is better (SEEG and Rythmic Gesticulator do not report a Diversity distance).

Methods	FGD ↓	Diversity ↑
Ground Truth	0.00	110.821
Joint Embedding [5]	22.083	91.223
Speech2Gesture [25]	19.254	98.095
Attention Seq2Seq [53]	18.154	92.176
SEEG [35]	3.751	-
Trimodal [52]	3.729	102.539
HA2G [39]	3.072	**108.086**
Rythmic Gesticulator [6]	2.04	-
Ours (AQ-GT)	**1.612**	104.76

6.2 Human Rater Study

As assessing the quality of co-speech gestures is highly subjective, gesture generation frameworks are commonly evaluated, by employing human raters [40]. We therefore conducted a human rater study to evaluate our proposed framework in comparison to the frameworks by Liu et al. [39] and Liang et al. [35]. To this end, we randomly selected ten 20-second-long audio/text sequences from our test set and generated co-speech gestures using all three models. The study was conducted as an online evaluation with 70 English-speaking participants (50% female and 50% male). Participants had to watch the videos in random order and, using a 5-point Likert scale, rate the co-speech gestures concerning human likeness and smoothness of motion as well as how much it matched the accompanied speech in content and timing (beat). Prior to the main study, we conducted a pre-survey with 10 participants to estimate the time required for each sequence. Based on the results of this pre-study, we excluded in the analysis any participant who remained less than 30 seconds on the survey pages, as well as any participant that deviated more than two times the standard deviation from the mean time. After applying these conditions, 39 participants remained. As our results did not fulfill the normality and homogeneity of variances tests, we used a generalized linear mixed-effects model (GLMM) for a repeated measure analysis. To

Table 4: Comparison of the proposed original model with four different versions where single components were left out (for FGD and MAJE lower is better; for Diversity higher is better).

Methods	FGD ↓	Diversity ↑	MAJE ↓
Ground Truth	0.00	42.128	0.00
w/o GRU	57.75	16.235	0.0330
w/o Temporal Aligner	1.653	39.054	0.0157
w/o VQ_G	1.342	41.698	0.0102
w/o Transformer	0.777	43.592	0.0086
Original	**0.3977**	**42.885**	**0.0078**

compare the pairwise significance effect of the different models, we then used the Dunn posthoc test. The results are presented in Figure 3, demonstrating that the AQ-GT approach proposed here significantly outperforms both the HA2G and SEEG frameworks in all rated categories. Notably, our framework achieves ratings of human likeness and smoothness comparable to human ground truth (no significant difference), while the ratings of content and beat matching are still significantly lower than ground truth. These findings suggest that while our approach is capable of generating appropriate, smooth, and natural gestures, it still falls short in terms of generating convincing gestures that accurately reflect the content of speech or convey substantial additional meaning.

7 ABLATION STUDY

In order to assess the contribution of individual components of our proposed framework, we conducted an ablation study employing four different architectural configurations. The impact of the GRU-Transformer was evaluated by selectively removing all GRU layers (w/o GRU) or all Transformer layers (w/o Transformer). The Temporal Aligner was evaluated by removing either the final reconstruction of gestures with the VQ_G (w/o VQ_G) or the entire Temporal Aligner (w/o Temporal Aligner). Each modified architecture was re-trained for 50 epochs, and the best-performing validation loss weights were restored. The performance on the test set is reported in Table 4. As can be seen, the removal of the GRU layers had the largest impact on all reported measures. Visual inspection of the generated output did also show strongly deformed joints without any visible gestures. On the other hand, the removal of the Transformer layer did lead to a small degradation in the FGD and MAJE but led to a higher Diversity distance value. As argued above, this could indicate problems with the Diversity distance, as it is highly susceptible to noise and outliers in the gestures. Nevertheless, this shows that the Transformer layer did have the intended effect of learning long-term dependencies between frames and reducing noise in the generated gestures. The removal of the VQ_G reconstruction led to a strong degradation in all measures but still achieved results similar to the HA2G framework (see Table 2). Removing the entire Temporal Aligner leads to an even more substantial decrease in performance, achieving results more similar to the Trimodal and SEEG frameworks. This clearly indicates that learning the temporal dependencies of the resulting intermediate sequence helps to create more coherent and natural gestures and generally should be incorporated in future co-speech gesture synthesis frameworks.

8 CONCLUSION

In this paper, we proposed a novel co-speech gesture synthesis framework, called AQ-GT, which combines and utilizes several approaches and techniques for the generation of co-speech gestural motion. Most notably, we combine GRU and transformer layers to model the non-linear temporal dependencies in multimodal behavior. In addition, we use a vector quantized autoencoder to allow for a discrete latent space representation for sequential data in both modalities (speech and gesture), and employ a dedicated Temporal Aligner network to reconstruct natural gestures. For training this complex model, we collected a large set of highly accurate multimodal data, using an extended data processing pipeline to track full 3D body joints from monocular videos. Our evaluations study results show that this approach is able to synthesize co-speech gestures that surpass the current state-of-the-art both in objective and subjective measures, and are in parts indistinguishable from human gesturing in terms of human-likeness and smoothness. By performing the ablation study, we were able to show that every component has a beneficial influence on the objective measures.

Overall, the proposed AQ-GT framework contributes to the field and advances the current state-of-the-art in co-speech gesture synthesis. Although our framework is able to create convincing natural gestures, there are still challenges. For one, the Diversity distance showed that there are possibly artifacts in the gestures generated from the discrete latent representation. In addition, the subjective study revealed that the present framework (just as all other learning-based approaches) is still unable to cover the meaning dimension of human communication and therefore generates primarily conversational "motor gestures" that support the act of speaking but rarely provide additional information. Another drawback that is inherent in learning-based co-speech gesture synthesis, is the inability to guide or control the generation of gestures. In future work we will both look at the drawbacks of discrete representations and how these can be improved, as well as if it is possible to guide the generation of gestures to more closely match the content of the conversation by conveying additional meaning. In current work, published elsewhere [48], we have started to extend the AQ-GT model to support form and meaning target features in gesture synthesis, learned from a richly annotated but smaller dataset. This shows that the proposed AQ-GT model can provide a suitable basis for further explorations into the generation of not just co-speech motion, but meaningful gestures as an integral part of multimodal communicative behavior.

REFERENCES

[1] [n. d.]. TED — youtube.com. https://www.youtube.com/c/TED/videos. [Accessed 16-Feb-2023].

[2] [n. d.]. TEDx Talks — youtube.com. https://www.youtube.com/channel/UCsT0YIqwnpJCM-mx7-gSA4Q. [Accessed 16-Feb-2023].

[3] 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 *Findings of the Association for Computational Linguistics: EMNLP 2020*. 1884–1895.

[4] 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. http://arxiv.org/abs/2007.12553

arXiv:2007.12553 [cs].

[5] Chaitanya Ahuja and Louis-Philippe Morency. 2019. Language2Pose: Natural Language Grounded Pose Forecasting. http://arxiv.org/abs/1907.01108 arXiv:1907.01108 [cs].

[6] Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. 2022. Rhythmic Gesticulator: Rhythm-Aware Co-Speech Gesture Synthesis with Hierarchical Neural Embeddings. *ACM Transactions on Graphics* 41, 6 (Dec. 2022), 1–19. https://doi.org/10.1145/3550454.3555435 arXiv:2210.01448 [cs, eess].

[7] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. https://doi.org/10.48550/arXiv.2006.11477 arXiv:2006.11477 [cs, eess].

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

[9] Matthew Brand and Aaron Hertzmann. 2000. Style machines. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques (SIGGRAPH '00)*. ACM Press/Addison-Wesley Publishing Co., USA, 183–192. https://doi.org/10.1145/344779.344865

[10] Justine Cassell, David McNeill, and Karl-Erik McCullough. 1999. Speech-gesture mismatches: Evidence for one underlying representation of linguistic and non-linguistic information. *Pragmatics & cognition* 7, 1 (1999), 1–34.

[11] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2004. BEAT: the Behavior Expression Animation Toolkit. In *Life-Like Characters: Tools, Affective Functions, and Applications*, Helmut Prendinger and Mitsuru Ishizuka (Eds.). Springer, Berlin, Heidelberg, 163–185. https://doi.org/10.1007/978-3-662-08373-4_8

[12] Changchun Liu, P. Rani, and N. Sarkar. 2005. An empirical study of machine learning techniques for affect recognition in human-robot interaction. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, Edmonton, Alta., Canada, 2662–2667. https://doi.org/10.1109/IROS.2005.1545344

[13] Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella. 2015. Predicting Co-verbal Gestures: A Deep and Temporal Modeling Approach. In *Intelligent Virtual Agents*, Willem-Paul Brinkman, Joost Broekens, and Dirk Heylen (Eds.). Vol. 9238. Springer International Publishing, Cham, 152–166. https://doi.org/10.1007/978-3-319-21996-7_17 Series Title: Lecture Notes in Computer Science.

[14] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/arXiv.1810.04805 arXiv:1810.04805 [cs].

[16] Patrick Esser, Robin Rombach, and Björn Ommer. 2021. Taming Transformers for High-Resolution Image Synthesis. http://arxiv.org/abs/2012.09841 arXiv:2012.09841 [cs].

[17] Angela Fan, Thibaut Lavril, Edouard Grave, Armand Joulin, and Sainbayar Sukhbaatar. 2020. Addressing some limitations of transformers with feedback memory. *arXiv preprint arXiv:2002.09402* (2020).

[18] Yuchen Fan, Yao Qian, Feng-Long Xie, and Frank K. Soong. 2014. TTS synthesis with bidirectional LSTM based recurrent neural networks. In *Interspeech 2014*. ISCA, 1964–1968. https://doi.org/10.21437/Interspeech.2014-443

[19] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. 2022. AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[20] Mireille Fares, Catherine Pelachaud, and Nicolas Obin. 2022. Transformer Network for Semantically-Aware and Speech-Driven Upper-Face Generation. https://doi.org/10.48550/arXiv.2110.04527 arXiv:2110.04527 [eess].

[21] 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*. ACM, Sydney NSW Australia, 93–98. https://doi.org/10.1145/3267851.3267898

[22] Aphrodite Galata, Neil Johnson, and David Hogg. 2001. Learning Variable-Length Markov Models of Behavior. *Computer Vision and Image Understanding* 81, 3 (March 2001), 398–413. https://doi.org/10.1006/cviu.2000.0894

[23] Saeed Ghorbani, Ylva Ferstl, and Marc-André Carbonneau. 2022. Exemplar-based stylized gesture generation from speech: An entry to the GENEA Challenge 2022. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 778–783.

[24] Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F. Troje, and Marc-André Carbonneau. 2022. ZeroEGGS: Zero-shot Example-based Gesture Generation from Speech. https://doi.org/10.48550/arXiv.2209.07556 arXiv:2209.07556 [cs].

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

[26] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

[27] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. *Advances in neural information processing systems* 30 (2017).

[28] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. 2020. MoGlow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics* 39, 6 (Dec. 2020), 1–14. https://doi.org/10.1145/3414685.3417836 arXiv:1905.06598 [cs, eess, stat].

[29] Daniel Holden, Taku Komura, and Jun Saito. 2017. Phase-functioned neural networks for character control. *ACM Transactions on Graphics* 36, 4 (Aug. 2017), 1–13. https://doi.org/10.1145/3072959.3073663

[30] Peter J. Huber. 1964. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics* 35, 1 (March 1964), 73–101. https://doi.org/10.1214/aoms/1177703732 Publisher: Institute of Mathematical Statistics.

[31] Kyung-Min Kim, Chang-Jun Nan, Jung-Woo Ha, Yu-Jung Heo, and Byoung-Tak Zhang. 2015. Pororobot: A Deep Learning Robot that Plays Video Q&A Games. (2015).

[32] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[33] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N. Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R. Thórisson, and Hannes Vilhjálmsson. 2006. Towards a Common Framework for Multimodal Generation: The Behavior Markup Language. In *Intelligent Virtual Agents (Lecture Notes in Computer Science)*, Jonathan Gratch, Michael Young, Ruth Aylett, Daniel Ballin, and Patrick Olivier (Eds.). Springer, Berlin, Heidelberg, 205–217. https://doi.org/10.1007/11821830_17

[34] Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha S. Srinivasa, and Yaser Sheikh. 2019. Talking with hands 16.2 m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 763–772.

[35] Yuanzhi Liang, Qianyu Feng, Linchao Zhu, Li Hu, Pan Pan, and Yi Yang. 2022. SEEG: Semantic energized Co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10473–10482.

[36] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel van de Panne. 2020. Character Controllers Using Motion VAEs. *ACM Transactions on Graphics* 39, 4 (Aug. 2020). https://doi.org/10.1145/3386569.3392422 arXiv:2103.14274 [cs].

[37] 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. In *Computer Vision – ECCV 2022 (Lecture Notes in Computer Science)*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer Nature Switzerland, Cham, 612–630. https://doi.org/10.1007/978-3-031-20071-7_36

[38] Jinlin Liu, Yuan Yao, and Jianqiang Ren. 2019. An acceleration framework for high resolution image synthesis. *arXiv preprint arXiv:1909.03611* (2019).

[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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, New Orleans, LA, USA, 10452–10462. https://doi.org/10.1109/CVPR52688.2022.01021

[40] Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. 2023. A Comprehensive Review of Data-Driven Co-Speech Gesture Generation. https://doi.org/10.1111/cgf.14776 arXiv:2301.05339 [cs].

[41] Amit Kumar Pandey and Rodolphe Gelin. 2018. A Mass-Produced Sociable Humanoid Robot: Pepper: The First Machine of Its Kind. *IEEE Robotics & Automation Magazine* 25, 3 (Sept. 2018), 40–48. https://doi.org/10.1109/MRA.2018.2833157 Conference Name: IEEE Robotics & Automation Magazine.

[42] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7753–7762.

[43] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. 2019. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems* 32 (2019).

[44] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[45] Adriana Tapus, Antonio Bandera, Ricardo Vazquez-Martin, and Luis V. Calderita. 2019. Perceiving the person and their interactions with the others for social robotics – A review. *Pattern Recognition Letters* 118 (Feb. 2019), 3–13. https://doi.org/10.1016/j.patrec.2018.03.006

[46] Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems* 30 (2017).

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[48] Hendric Voß and Stefan Kopp. 2023. Augmented Co-Speech Gesture Generation: Including Form and Meaning Features to Guide Learning-Based Gesture Synthesis.

arXiv preprint arXiv:2307.09597 (2023).
[49] Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview. , 209–232 pages.
[50] Xin Wang, Jiabing Xu, Wei Shi, and Jiarui Liu. 2019. OGRU: An optimized gated recurrent unit neural network. In *Journal of Physics: Conference Series*, Vol. 1325. IOP Publishing, 012089.
[51] Jiqing Wu, Zhiwu Huang, Janine Thoma, Dinesh Acharya, and Luc Van Gool. 2018. Wasserstein divergence for gans. In *Proceedings of the European conference on computer vision (ECCV)*. 653–668.
[52] 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 (Dec. 2020), 1–16. https://doi.org/10.1145/3414685.3417838 arXiv:2009.02119 [cs] archive:https://web.archive.org/web/20230405100117/https://arxiv.org/abs/2009.02119.
[53] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. Robots Learn Social Skills: End-to-End Learning of Co-Speech Gesture Generation for Humanoid Robots. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, Montreal, QC, Canada, 4303–4309. https://doi.org/10.1109/ICRA.2019.8793720
[54] Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2022. The GENEA Challenge 2022: A large evaluation of data-driven co-speech gesture generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 736–747.
[55] Chuang Yu and Adriana Tapus. 2019. Interactive Robot Learning for Multimodal Emotion Recognition. In *Social Robotics (Lecture Notes in Computer Science)*, Miguel A. Salichs, Shuzhi Sam Ge, Emilia Ivanova Barakova, John-John Cabibihan, Alan R. Wagner, Álvaro Castro-González, and Hongsheng He (Eds.). Springer International Publishing, Cham, 633–642. https://doi.org/10.1007/978-3-030-35888-4_59
[56] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. 2020. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214* (2020).
[57] Feng Zhang, Xiatian Zhu, and Mao Ye. 2019. Fast human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3517–3526.
[58] Wenlin Zhuang, Jinwei Qi, Peng Zhang, Bang Zhang, and Ping Tan. 2022. Text/Speech-Driven Full-Body Animation. *arXiv preprint arXiv:2205.15573* (2022).