Emotional Speech-driven 3D Body Animation via Disentangled Latent Diffusion

Kiran Chhatre[1] Radek Daněček[2] Nikos Athanasiou[2]
Giorgio Becherini[2] Christopher Peters[1] Michael J. Black[2] Timo Bolkart[2*]
[1]KTH Royal Institute of Technology, Sweden [2]Max Planck Institute for Intelligent Systems, Germany

Abstract

Existing methods for synthesizing 3D human gestures from speech have shown promising results, but they do not explicitly model the impact of emotions on the generated gestures. Instead, these methods directly output animations from speech without control over the expressed emotion. To address this limitation, we present AMUSE, an emotional speech-driven body animation model based on latent diffusion. Our observation is that content (i.e., gestures related to speech rhythm and word utterances), emotion, and personal style are separable. To account for this, AMUSE maps the driving audio to three disentangled latent vectors: one for content, one for emotion, and one for personal style. A latent diffusion model, trained to generate gesture motion sequences, is then conditioned on these latent vectors. Once trained, AMUSE synthesizes 3D human gestures directly from speech with control over the expressed emotions and style by combining the content from the driving speech with the emotion and style of another speech sequence. Randomly sampling the noise of the diffusion model further generates variations of the gesture with the same emotional expressivity. Qualitative, quantitative, and perceptual evaluations demonstrate that AMUSE outputs realistic gesture sequences. Compared to the state of the art, the generated gestures are better synchronized with the speech content, and better represent the emotion expressed by the input speech. Our code is available at amuse.is.tue.mpg.de.

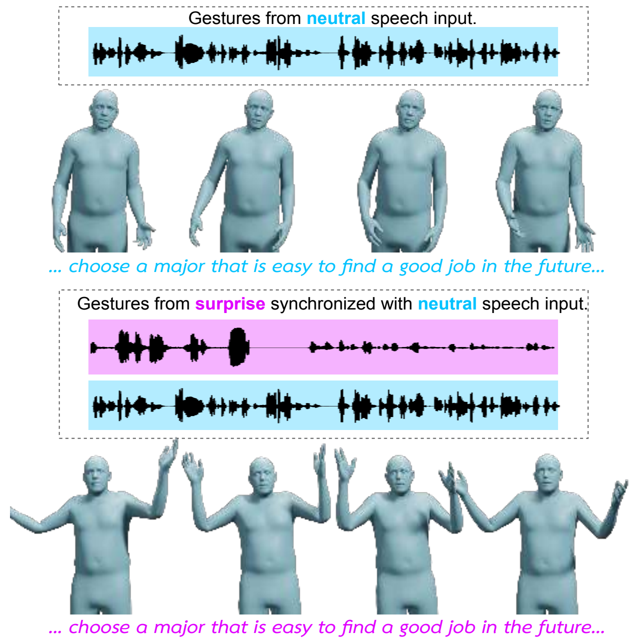

Figure 1. **Goal.** AMUSE generates realistic emotional 3D body gestures directly from a speech sequence (top). It provides user control over the generated emotion by combining the driving speech sequence with a different emotional audio (bottom).

1. Introduction

Animating 3D bodies from speech has a wide range of applications, such as telepresence in AR/VR, avatar animation in games and movies, and to embody interactive digital assistants. While methods for speech-driven 3D body animation have recently shown great progress [5, 7, 31, 56, 101], existing methods do not adequately address one crucial factor: the impact of emotion from the driving speech signal on the generated gestures. Emotions and their expressions play a fundamental role in human communication [29, 35, 65] and have become an important consideration when designing computer systems that interact with humans in a natural manner [78, 79]. They are of central concern when synthesizing human animations for a wide variety of application contexts, such as Socially Interactive Agents [61]. Because of this, speech-driven animation systems must not only align movement with the rhythm of the speech, but should also be capable of generating gestures that are perceived as expressing the suitable emotion.

Many factors contribute to the perception of emotion and personal idiosyncrasies, such as facial expressions [19], gaze and eye contact [42], physiological responses [47], tone of voice [87], body language [66], and gestures [39]. When it comes to 3D animation, the most relevant factors

*Now at Google.

are facial expressions, gestures, and body language [95]. While emotional speech-driven animation methods have recently been proposed for 3D faces [18, 74, 90, 107], animating emotional bodies from speech remains under-explored.

Generating gestures solely from speech with emotional control is a difficult task. First, the mapping from audio to body motion is a non-deterministic many-to-many mapping, which is difficult to model. Gestures across subjects can vary when uttering the same sentence, and a single individual's motions can change significantly across repetitions. Second, factoring out the impact of emotional state on the body motion from other, unknown factors, is difficult. This requires disentangling the effects of three different factors on the generated motion, namely content-based (i.e., gestures related to speech rhythm and word utterances), emotion-based, and those based on personal style. AMUSE addresses this by separating a speech sequence into content, emotion, and style latent vectors, which are then used to condition a latent diffusion model. Specifically, AMUSE consists of three main components: (1) an audio autoencoder trained to produce disentangled vectors of content, emotion, and style, (2) a 3D body motion prior in the form of a temporal variational autoencoder (VAE) to generate smooth and realistic gestures, and (3) a latent diffusion model, which generates 3D body motion given the input content, emotion, and style latent vectors.

Training such a model requires a speech-to-3D body dataset of sufficient scale, which is rich and diverse in speakers and emotions. BEAT [55] is a good candidate because it provides a large set of 3D gestures associated with single-person monologues. Unfortunately, the bodies are represented as skeletons, and it lacks face mocap markers and FLAME expressions. Instead, to produce realistic body animations, we require articulated 3D body surfaces. To overcome this, we convert BEAT sequences to SMPL-X [73] format using MoSh++ [62] and use the SMPL-X parameters for training. See [56] for comparison.

Our contributions are: (1) We present a framework to synthesize emotional 3D body gestures directly from speech. (2) We factor an input audio into disentangled content, emotion and style vectors, which enables us to separately control emotion in generated gestures. (3) We adapt temporal latent diffusion for multiple target conditions.

2. Related Work

2.1. 3D Conditional Human Motion Generation

Early works focus mostly on predicting [10, 16, 33, 41, 57, 64, 70, 86, 106, 109] or generating human motion [30, 49], but do not consider multi-modal control. Recently, conditional motion generation through other modalities, such as text [2, 8, 9, 17, 22, 28, 77], music [50, 68, 94], speech [32], or action labels [27, 75], has gained more attention. Below, we focus on speech-driven motion generation methods, since they are the most relevant to our work.

2.2. Gesture Generation from Speech

Rule-based gesture synthesis. Embodied conversational agents (ECA) are designed to interact and communicate with humans. Using the Behavior Markup Language (BML) [44] one can build rule-based systems for humanoids based on predefined behaviors [80]. This is used for completion of a storytelling task in an expressive manner [45]. The BEAT rule-based toolkit [14] enables adding non-verbal behavior on top of a pre-animated figure. Thiebaux et al. [92] develop an ECA by using procedural animation techniques and keyframe interpolation. Marsella et al. [63] design a generalized rule-based agent to generate expressions, eye gaze, and gestures from speech. Each of these approaches are based on non-trainable, rule-based techniques that may require substantial manual modelling effort to adapt to new tasks.

Data-driven gesture synthesis. More recently, data-driven methods have superseded rule-based systems. Yoon et al. [104] use a fusion of text, audio and upper body gestures to learn an upper body gesture avatar, but can only control the style of individual speakers by sampling from their latent space. SpeechGestureMatching [32] generates 3D facial meshes and 3D keypoints of the body and hands from speech, but the outputs are separated and the method does not provide control over the generations. QPGesture [98] uses phase to better align the generated 3D skeleton-based gesturing avatars with the audio input. Ginosar et al. [23] and Diverse-3D-Hand-Gesture-Prediction [82] generate hand and arm motions only. Audio2Gestures [48] encode motion and audio to a low-dimensional latent space and generate gestures. SEEG [53] aims to generate gestures that align well with the semantics of the speech. DiffTTSG [67] regresses speech and gestures at the same time, joining the two modalities in a single system. DiffGAN [3] retargets gestures across speakers in a low-resource setting. The GENEA challenge [105] tackles gesticulation from speech alone using the Talking-with-Hands dataset [46]. Gesture2Vec [99] uses a machine translation model to translate text into gesture chunks and output full sequences using such quantized representations. TalkSHOW [101] uses a VQ-VAE to generate 3D human bodies gesturing with facial expressions from speech segments, but in an uncontrolled manner. Similarly, Co-speech gesture [60] uses an RQ-VAE to generate different gestures from speech. Alternative gesture generation from speech methods have been proposed such as reinforcement learning [91], self-supervised pretraining [40], and diffusion [67, 110]. BodyFormer [71] introduces a dataset of pseudo-groundtruth and a transformer-based method for generating gestures from speech. However, none of these methods provide explicit emotional control over the generated motion.

For controllable generation, GestureDiffuCLIP [7] incorporates multiple conditions including CLIP [83] text features, video, or motion prompts via AdaIn [37] layers to generate gestures from speech, however, it does not allow explicit control over the emotion conveyed by the driving audio. ListenDenoiseAction [5] combines conformers and the DiffWave [43] architecture to generate gestures that can be controlled by a style vector, RhythmicGesticulator [6] disentangles the latent space into a vector related to the semantics of the gesture and one related to the subtle variations, while DisCo [54] models content and rhythm. StyleGestures [4] adapts MoGlow [34], demonstrating limited control over some motion attributes like the speed and expressiveness of gestures. DiffuseStyleGesture [97] uses diffusion to generate diverse gestures from speech.

2.3. Emotion Control

Emotion classification and control has been little studied in 3D human motion generation with only a a few methods using skeletal motion in multi-class classification. Ghaleb et al. [20] employ a spatio-temporal graph convolution network to classify gestures into four classes: preparation, stroke, retraction, and neutral. Li et al. [52], on the other hand, use hidden Markov models for emotion classification of human movement mocap data. Karras et al. [38] learn face animations of a single actor, and test their method on different tasks by modifying the latent vectors. However, there is no disentanglement mechanism, and they do not model the synchronization of the emotion with the with the facial motions. Recently, EmoTalk [74], animates emotional 3D faces from speech input with control over the emotion intensity and EMOTE [18] disentangles emotion and speech to allow emotion editing at test time. However, models solely intended for facial tasks like lip syncing and capturing expressions might not smoothly adapt to the complexity of whole-body movements and distinct articulation. Regarding emotion-conditioned motion generation, Aberman et al. [1] show style-transfer from video data to motion and provide some style-based control, but do not address speech-driven emotional gestures. Similarly, the ZeroEGGs [21] dataset contains some emotional gesture controls but also includes more generic styles of motion. The method requires the input of arbitrary frames of desired motion to encode a style, thereby relying on motions and speech as conditions during inference. Text-driven emotional gesticulation, as explored by Bhattacharya et al. [11, 12], emphasizes the generation of gestures based on textual cues, incorporating additional conditions such as speech, speaker ID, seed poses, as well as valence, arousal, and dominance triplets. However, these approaches do not provide the means to distill explicit emotion features, limiting free control over the generated gestures. Closer to our work, EMoG [102] incorporates emotion cues from the BEAT dataset [55] to generate improved gesture quality without explicit emotion control. EmotionGesture [81] uses a TED Emotion Dataset and BEAT to incorporate emotion features in gesture generation and generate emotional gestures. Although they can generate emotional gestures, their method is not end-to-end and has no explicit motion control. Specifically, it uses an emotion-conditioned VAE after training to acquire diverse emotion features that are used to generate gestures without guarantees and control over emotion types. Wu et al. [96] introduce the first multi-cultural gesture dataset containing 200 individuals of 10 different cultures. In contrast to prior work, we explicitly control the emotions conveyed by the generated gestures solely through emotional speech without relying on additional conditions.

3. Method

The AMUSE pipeline consists of two separately trained networks. The audio disentanglement module, which encodes input speech into latent vectors for content, emotion, and style is described in Sec. 3.2. The main architecture is described in Sec. 3.3. It consists of a 3D human motion prior coupled with a latent diffusion model. It takes random noise (or partially denoised latent vectors) on the input and outputs a human motion sequence. We introduce broader applications in gesture editing in Sec. 3.4.

3.1. Preliminary: Expressive 3D Body Model

SMPL-X [73] is a 3D model of the body surface. SMPL-X is defined as function $M(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi})$ that produces a 3D body mesh. It is parameterized by identity shape $\boldsymbol{\beta} \in \mathbb{R}^{300}$, pose $\boldsymbol{\theta} \in \mathbb{R}^{J \times 3}$ including finger articulation for rotations around J joints, and facial expression $\boldsymbol{\psi} \in \mathbb{R}^{100}$. We adopt the continous 6D rotation representation for training following Zhou et al. [108], making $\boldsymbol{\theta} \in \mathbb{R}^{J \times 6}$. Given pose parameters and any shape parameter, we can obtain body mesh vertices V using the differentiable SMPL-X layer [73]. As the focus of our paper is on synthesizing body gestures and not locomotion, we disregard 8 joints that correspond those of the lower body joint poses, leaving $J = 47$. Further, we omit the facial expression parameters, i.e., set $\boldsymbol{\psi} = \mathbf{0}$.

3.2. Speech Disentanglement Model

Architecture. The goal of the this model is to factor an input speech into three disentangled latent representations, one for content (i.e., the words spoken), one for emotion, and one for personal style. To do so, we devise a specialized encoder–decoder architecture with three separate encoders, one for each latent space. We denote the encoders as: $E_c(a) = c, \quad E_e(a) = e, \quad E_s(a) = s$, where a is the input filterbank, c, e and s denote the latent vectors for content, emotion and style and E_c, E_e and E_s are their encoders. The architecture of the three encoders follows the

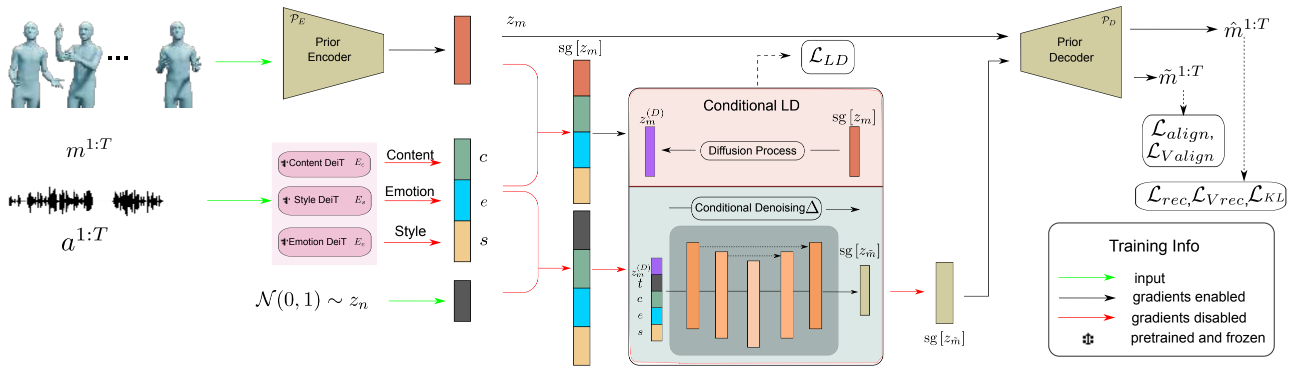

Figure 2. **Training.** We train the motion prior ($\mathcal{P}_E, \mathcal{P}_D$) and the latent denoiser Δ jointly, while keeping the audio encoding networks frozen. In the forward pass, we take an input audio $a^{1:T}$ and pose sequence $m^{1:T}$. Firstly, we do a forward pass of $m^{1:T}$ through $\mathcal{P}_E$ and $\mathcal{P}_D$ and compute $\mathcal{L}_{rec}$, $\mathcal{L}_{Vrec}$, and $\mathcal{L}_{KL}$. Then, we apply the diffusion process to a gradient-detached sg $[z_m]$ obtaining the noisy $z_m^{(D)}$, which is then denoised with Δ and $\mathcal{L}_{LD}$ is computed. Finally, we use Δ to fully denoise z_n into gradient-detached sg $[z_{\tilde{m}}]$, further decode $\tilde{m}^{1:T}$ using $\mathcal{P}_D$, and compute $\mathcal{L}_{align}$ and $\mathcal{L}_{Valign}$.

design by Gong et al. [25, 26] (i.e., leveraging the DeiT visual transformer [93] adapted for processing filterbank images extracted from the input audio). The decoder takes the three latent vectors and produces a reconstructed filterbank. Formally $D(c, e, s) = \widehat{a}$, where $\widehat{a}$ denotes the reconstructed filterbank. The decoder architecture consists of a fusion module and transformer-encoder layers.

Training. The audio module is trained with a multiple loss terms that ensure that the three latent spaces are properly disentangled. In addition to the standard autoencoder reconstruction loss, we also employ three cross-reconstruction losses, in which we enforce the correct reconstruction of the audio signal where we modify one of the content, style or emotion latents. Additionally, we employ three loss terms on the latent vector predictions – namely emotion and style classification losses over e and s, and a content similarity loss between pairs of two content latent vectors extracted from audios that have the same spoken content. For a detailed description of the encoder–decoder architecture, a formal definition of the loss functions and a detailed description of the training process please refer to the Sup. Mat.

3.3. Gesture Generation Model

Motion prior. Similar to [15, 76], our motion prior network is a VAE transformer architecture with encoder $\mathcal{P}_E$ and decoder $\mathcal{P}_D$. Specifically, both $\mathcal{P}_E$ and $\mathcal{P}_D$ follow a U-Net-like [85] structure with skip connections between transformer blocks (see Sup. Mat. for details). The positional embeddings are learnable and injected into each multi-head attention layer, following the design of Carion et al. [13]. Formally, the encoder takes a sequence of T frames of the SMPL-X pose vectors $m^{1:T} \in \mathbb{R}^{6J \times T}$ and the first two tokens of its output, $\mu \in \mathbb{R}^{d_m}$ and $\Sigma \in \mathbb{R}^{d_m \times d_m}$ are used to extract the motion latent $z_m \in \mathbb{R}^{d_m}$ via the reparametrization trick. The decoder takes zero positional encodings as query input and the motion latent is fed as memory to every cross-attention transformer layer, producing the reconstructed motion $\hat{m}^{1:T}$.

Diffusion process. The forward diffusion process is similar to [36, 69]. We employ fixed variance and linearly scaled noise scheduler. We add noise to the motion latent z_m for D diffusion timesteps to obtain $z^{(D)}$ following:

$$q(z_m^{(t_d)} \mid z_m^{(0)}) = \mathcal{N}(z_m^{(t_d)}; \sqrt{\bar{\alpha}_{t_d}} z_m^{(0)}, (1 - \bar{\alpha}_{t_d})\mathbf{I}),$$

with $\alpha_{t_d} = 1 - \beta_{t_d}$, $\bar{\alpha}_{t_d} = \prod_{s=1}^{t_d} \alpha_s$, and β_{t_d} denotes diffusion process variance.

Conditional denoising process. The denoising process consists of iteratively denoising a conditioned noisy motion latent vector to obtain the denoised motion latent $z_{\tilde{m}^{1:T}}$. Our denoiser Δ is a latent variable model [84] and its architecture is similar to the U-Net-like structure of the motion prior encoder $\mathcal{P}_E$. The input of the model is a concatenation of: $z_m^{(t_d)}, \mathrm{SE}(t_d), c, e, s \in \mathbb{R}^{256}$, where $\mathrm{SE}(t_d)$ is a sinusoidal positional encoding of diffusion timestep t_d as defined in [36]. Δ iteratively denoises through each reversed diffusion step:

$$z_m^{(t_d - 1)} = \Delta([z_m^{(t_d)}, \mathrm{SE}(t_d), c, e, s]).$$

Training. We optimize the motion prior and the latent denoiser jointly to ensure audio–motion latent code alignment during conditional fusion in the denoising process using a 3-step forward pass through the gesture generation model. First, following standard VAE practice, we reconstruct $\hat{m}^{1:T}$ by the motion prior forward pass. As shown in Fig. 2, we then disable gradient calculation in $\mathcal{P}_E$ to infer the intermediate motion latent sg $[z_m]$, which serves as input to the denoiser. At this stage, we obtain the denoiser noise prediction, δ and use to compute the diffusion model gradients. Finally, in the third step we compute $\tilde{m}^{1:T} = \mathcal{P}_D(\text{sg}[z_{\tilde{m}}])$, where $z_{\tilde{m}}$ is obtained by iteratively using the Δ to obtain a fully denoised latent from $z_n^{(t_D)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We indicate computations done without gradients with a stop-gradient operation sg [.].

Losses. To train the motion prior, we include the standard VAE losses, namely the reconstruction loss on pose parameters $\mathcal{L}_{rec}$ and on vertex coordinates $\mathcal{L}_{Vrec}$ using the smooth L1 metric introduced in [24], which we denote as L_1^s:

$$\mathcal{L}_{rec} = L_1^s(m^{1:T}, \hat{m}^{1:T}), \quad \mathcal{L}_{Vrec} = L_1^s(V^{1:T}, \hat{V}^{1:T}),$$

where the root-centered vertices V are obtained by feeding in pose parameters m to a differentiable SMPL-X layer (without learnable parameters) and a mean shape $\boldsymbol{\beta} = \vec{0}$. The KL divergence loss of the motion prior is:

$$\mathcal{L}_{KL} = \frac{1}{2}\left[\sum_{i=1}^{z}(\mu_i^2 + \sigma_i^2) - \sum_{i=1}^{z}\left(log(\sigma_i^2) + 1\right)\right].$$

To ensure the alignment of the diffusion-generated motions and the input audio, we apply the alignment reconstruction loss on the inferred motion pose parameters and the vertex coordinates:

$$\mathcal{L}_{align} = L_1^s(m^{1:T}, \tilde{m}^{1:T}), \quad \mathcal{L}_{Valign} = L_1^s(V^{1:T}, \tilde{V}^{1:T}).$$

Finally, we utilize the objective similar to [15, 36, 84] to supervise the denoiser:

$$\mathcal{L}_{LD} = \left\|\delta^{(t_d)} - \Delta(z_m^{(t_d)}, \text{SE}(t_d), c, e, s)\right\|_2^2,$$

where $\delta^{(t_d)}$ is the noise vector sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ in the corresponding diffusion step t_d. The combined gesture model loss is:

$$\mathcal{L}_{ges} = \mathcal{L}_{rec} + \mathcal{L}_{Vrec} + \mathcal{L}_{KL} + \mathcal{L}_{align} + \mathcal{L}_{Valign} + \mathcal{L}_{LD}$$

Inference. We employ DDIM [89] to infer high quality conditional motion samples with a small number of denoising timesteps. During inference we draw a sample vector from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ to iteratively denoise in reversed timesteps. The denoised sample is then passed through the decoder $\mathcal{P}_D(z_{\tilde{m}^{1:T}})$ to obtain motion $\tilde{m}^{1:T}$.

3.4. Gesture Editing

Due to the disentangling of the inputs, AMUSE achieves semantic gesticulation control using two driving input audios. Specifically, given two input audio signals a_1 and a_2, we extract their latent representations of content c_1, c_2, emotion e_1, e_2, and style s_1, s_2. Then, we simply initialize the denoising procedure of Δ with the triplet (c_1, e_2, s_1), generating the gesture with the content and style of a_1 but the emotion of input audio a_2. Similarly, instead of emotion we can also change the gesticulation style to that of the speaker of a_2 by initializing with (c_1, e_1, s_2).

4. Implementation Details

MoCap data preparation. The BEAT [55] mocap sequences, captured in a Vicon system at 120 Hz, are downsampled to 30 Hz and processed using MoSh++ [58, 62] to obtain SMPL-X parameters. Given a sequence of 3D mocap marker positions, we jointly optimize SMPL-X shape and pose parameters, 3D body translation, and embedding of the mocap markers in the SMPL-X surface. Once processed, the sequences are then divided according to the emotion annotations in the BEAT dataset. We use sequences of English speaking subjects in monologue speaking style for training and evaluating AMUSE. For each sequence we draw $m^{1:L}$ at 30 FPS and concatenate with audio content c, emotion e, and style s latent vectors. Then, we segment it to 10-sec windows T, beginning from the timestamp 0 and discarding additional unaligned information at the end. This preprocessing choice allows us to train transformer networks without masking. We provide additional data processing information in the Sup. Mat.

Audio preprocessing. We use audio sequences belonging to eight categorical emotion labels (neutral, happy, angry, sad, contempt, surprise, fear, and disgust). Each audio chunk of 10s is converted into a filter bank with 128 mel-frequency bins with a 25ms Hamming frame window and 10ms frame shift. We mask each sample with a maximum length of 24 in the frequency domain and a maximum length of 96 in the time domain, employing Park et al. [72]. Following [25, 26], we standardize the filter bank and augment it via noise injection and circular shifting. Before feeding in our speech disentanglement model, each filter bank is split into a sequence of fixed 1209 patches of 16 x 16 each having 6 units overlap in frequency and time domain.

Motion prior. The motion prior is a VAE encoder–decoder with 9 layers and 4 heads, following Chen et al. [15]. The encoder–decoder is a U-Net-like transformer with residual connections. Learnable positional embeddings are injected in each multi-head attention layer. We have a linear projection at the start and the end of our motion prior network. The KL divergence term is weighted with a factor of $1e-4$.

Denoiser. The denoiser follows the same network architecture as our prior encoder. The hidden dimension of all transformer layers is 1024. We use 1000 diffusion steps D during training and 50 during inference. Noise betas are in range $[0.00085, 0.012]$. We jointly optimize the prior and denoiser networks for 5000 epochs with batch size of 64, learning rate 0.0001, and the AdamW optimizer [59].

5. Experiments

Speech disentanglement model. We evaluate the performance of the speech disentanglement model quantitatively using classification accuracy and F1 scores on emotion and style. The accuracy is computed as average scores for all 8 emotion as well style categories that are part of the test dataset. The emotion and style accuracy is 91.53% and 96.06%, respectively. The emotion F1 score and style F1 scores are 0.914 and 0.960, respectively. See the Sup. Mat. for ablations and a detailed metric analysis.

Gesture generation model. We evaluate the performance of our gesture generation model quantitatively, qualitatively, and perceptually against following methods: TalkSHOW [101] and the re-implementation of Habibie et al. [31] provided by the TalkSHOW authors in the official TalkSHOW release [100], DiffuseStyleGesture (DSG) [97], MoGlow [34], and CaMN [55]. Additionally, we adapt TalkSHOW to include categorical emotion labels as input along with the existing architecture that only allows one-hot encodings of personal style. We then retrain it on our training data. We refer to it as TalkSHOW-BEAT. There are some concurrent works [5, 7, 56], which introduce methods for gesture generation from speech, however, direct comparison is hindered by the unavailability of released code our task. Refer to the Sup. Mat. for the ablation experiments, the emotion and style editing experiments, and their quantitative evaluation.

5.1. Quantitative Evaluation

To quantitatively evaluate our method's gesture generations and edited gesture generations, we train a transformer-based encoder architecture (denoted as M) similar to Petrovich et al. [76] in an autoencoder setting, where we append a CLS token at the beginning of the motion sequence. M is trained with a cross-entropy emotion classification objective applied to the output CLS token. We train M on the BEAT training dataset and use its features to compute the following metrics: (1) Fréchet gesture distance (FGD): We follow [88, 103, 104] to compute the feature distance between generated and ground truth motion features. (2) Gesture diversity (Div): Similarly to Chen et al. [15], we compute variance across generated features. (3) Gesture emotion accuracy (GA): We report top-1 emotion classification accuracy predicted by a classifier trained on the motion M-predicted latents. (4) Beat align (BA): We follow [51, 55], to evaluate the motion-speech correlation in terms of the similarity between the kinematic motion beats and speech audio beats. The kinematic motion beats are directly computed from the generated motion sequences. (5) Semantic-Relevant Gesture Recall (SRGR): We follow Liu et al. [55], to evaluate the semantic relevancy of gestures with GT motion. We use the ground truth semantic scores to compute this metric. The scores are obtained from the BEAT authors, representing a continuous score on a scale 0-1 per gesture style for 4 gesture semantic categories: beat, deictic, iconic, and metaphoric. While comparing with methods that output coarse skeletal data (DSG [97], MoGlow [34], and CaMN [55]), we convert the skeleton motion data into the SMPL-X axis angle representation. For details on the architectures and training of M, and the losses, please refer to the Sup. Mat.

Method	SRGR↑	BA↑	FGD↓	Div→	GA[a]↑
GT	—	0.83	—	27.83	64.04
Ours	0.36	0.81	388.63	25.06	46.76
Ours-EmoEdit[b]	—	0.79	792.58	24.68	34.18
TalkSHOW-BEAT	0.31	0.64	808.99	24.16	22.71
TalkSHOW [101]	0.30	0.60	762.15	23.19	29.41
DSG [97]	0.23	0.40	763.10	19.77	22.70
Habibie et al. [31]	0.23	0.39	809.17	21.34	16.67
MoGlow [34]	0.21	0.35	1097.03	19.50	16.62
CaMN [55]	0.21	0.39	1063.87	18.90	14.17

[a] GA is average of all 8 emotions.
[b] GA for these are average accuracy for all generations with 7 edited audio sequences.

Table 1. **Gesture quantitative results.** We compare our methods against several SOTA methods using metrics explained in Sec. 5.1. We observe that AMUSE outperforms in all scores compared to baseline methods. Additionally, AMUSE-EmoEdit outperforms in Beat Align, Diversity, and Gesture Emotion Accuracy scores compared to the baseline methods.

We prepare the evaluation data by randomly selecting 72 unique motion sequences each of length 10s and comprising 8 emotions across test subjects and compute the aforementioned metrics. We use 9 sequences for each emotion per subject. The results are reported in Tab. 1. All best scores are highlighted in green and second best in blue. AMUSE outperforms the baseline methods in all given metrics. To validate the performance of gesture emotion editing, we also report the same metrics for the emotion editing task (Ours-EmoEdit). During inference, the input style and content latents are extracted from neutral-emotion audio, while the emotion latent comes from a different audio of different emotion. These emotional edits offer numerous possibilities, allowing for transitions from any to any emotion. Tab. 1 shows the average for editing from neutral to other emotions. Since we require the GT gesture semantics score to compute SRGR metric, it is not possible to compute the SRGR for the synthetic edited-emotion gestures as they are not part of the original BEAT dataset. Ours-EmoEdit outperforms the baseline methods in BA, Div, and GA metrics. This demonstrates the capability of our model to maintain highly discriminative cues when switching between different emotions. TalkSHOW-BEAT has the second best score for SRGR whereas TalkSHOW demonstrates second best FGD score. Although, our model and ours-EmoEdit show improvements over the baseline methods, GT motions have higher diversity, Beat alignment score, and are easier to classify than generations of AMUSE, highlighting the challenging nature of the problem.

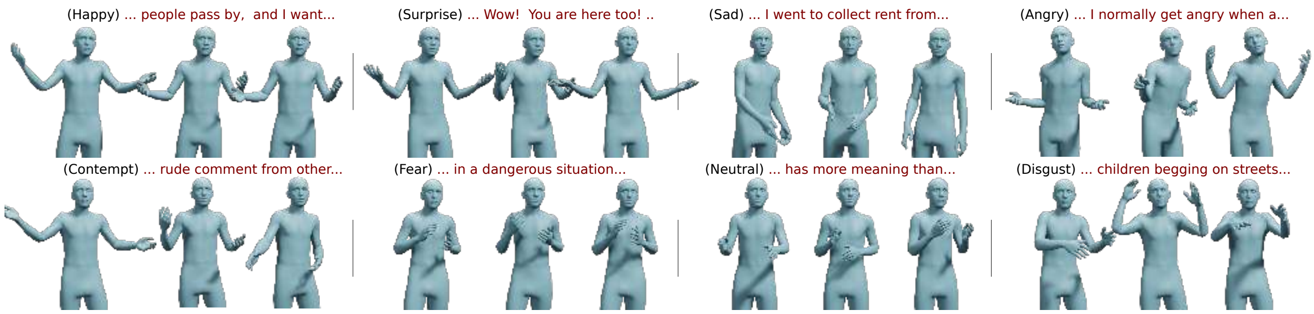

Figure 3. **Qualitative comparison across all emotions.** We evaluate generation on different test audios. AMUSE exhibits well-synchronized beat gestures and consistently produces gestures that accurately convey the emotional content expressed in the input speech.

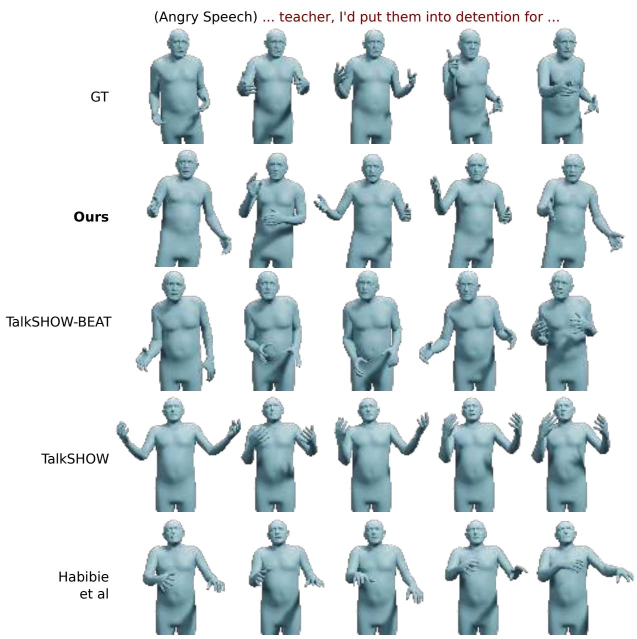

Figure 4. **Qualitative comparison with baseline methods**. The speech segment describes intense angry speech.

5.2. Qualitative Evaluation

Comparison with baseline methods. In Fig. 4, we demonstrate comparison with baseline methods that output a 3D body mesh: Habibie et al. [31], TalkSHOW [101], TalkSHOW-BEAT, and the BEAT ground truth (GT) [55]. We observe that AMUSE generates gestures that are semantically closer to the speech content and produces expressive emotional gestures closer to the perceived emotion. For example, the GT motion exhibits anger when saying "*put them into detention*". AMUSE demonstrates tense posture and aggressive movements comparable with the ground truth data and accurate synchronization with the spoken words. TalkSHOW [101] and Habibie et al. [31] exhibit limited movement and display inferior and static gestures on test audios as seen in the last two rows

Figure 5. **Qualitative evaluation of diverse generations**. Multiple generations overlayed.

of Fig. 4. TalkSHOW-BEAT slightly outperforms other baseline methods by demonstrating enhanced synchronized gestures, but it still does not perform as well as AMUSE.

Diverse emotional gestures. In Fig. 5, our probabilistic model can generate diverse gestures for same input audio.

Emotional gesture generation. In Fig. 3 AMUSE demonstrates strong correlation with the spoken utterances as well as different emotions. We observe that our model is able to correlate semantic words to associated gestures. For example, gestures demonstrate forceful actions and tense stance with angry audio "*normally get angry*" whereas it generates lowered and calm hand positions for sad audio "*I went to collect*". Similarly, our generations show hands that are closer to body for fearful audio "*in a dangerous situation*" while widely open expressing astonishment for happy and surprised audio "*people pass by*" and "*Wow! You are here*".

Emotion editing. We use two audio streams of a female subject for neutral and sad emotion. This experiment edits the subject's gesture style from moderately controlled hand movements to a sad style with lethargic posture conveying a sense of heaviness, as seen in Fig. 7 (top).

Gesture style editing. We use audio streams of two male subjects for the happy (ID - 13) and angry (ID - 2) emotion. With the emotion, style and content latent fusion mechanism from two driving audio streams, AMUSE is able to adapt the male (ID - 13) subject's body gestures from being close to their body to more open with squared tightened shoulders, expressing a shift from happy to angry emotions of a different subject (ID - 2), as shown in Fig. 7 (bottom). Please refer to the supplemental video for qualitative results and comparisons to additional gesture genera-

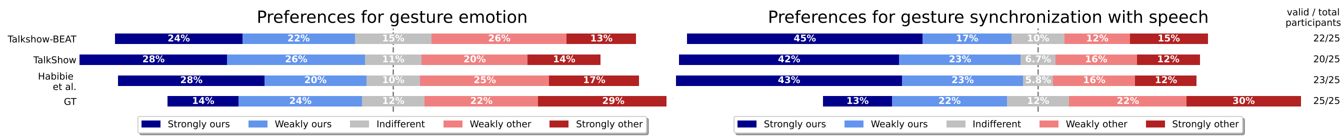

Figure 6. **The perceptual study results for gesture emotion preference (left) and synchronization with speech (right)**. The number of attentive participants that passed the catch trials is indicated on the right and the reported results only consider these participants.

Figure 7. **Gesture editing.** Top: We modify style from being neutral (left) to being sad (right) by combining the emotion latent from sad audio with the content latent from neutral audio. Bottom: We transform the style from Subject 13 being happy (left) to being angry (right) by merging the content latent from happy audio with the style and emotion latents from an angry audio of Subject 2.

tion methods [4, 55, 110] trained on coarse skeletal data.

5.3. Perceptual Study

Design. Our perceptual study is designed as a side-by-side comparison of two gesture videos generated with the same audio as input but by two different methods (AMUSE and another model or GT). The participants are asked to rate their preference of the methods on a five-point Likert scale for "synchronization with speech" and "gesture emotion appropriateness" given the GT emotion label of the input audio. We recruit 25 participants per method-to-method comparison on Amazon Mechnical Turk. Each participant is shown 24 pairs of randomly selected test set animations, 3 per emotion (neutral, happy, angry, sad, disgust, fear, surprise, and contempt). To allow the participant to get used to the task, we discard the answers of the first three comparisons and repeat these at the end. We incorporate three catch trials and responses from participants that fail on more than one are filtered out, as shown in Fig. 6 (right).

Results. The results of the study are shown in Fig. 6. AMUSE outperforms all competing methods by a considerable margin on both tasks, suggesting that AMUSE's generations are more appropriate for both the content of the input speech and its emotion compared to the baselines. However, it must be noted that there is still a significant gap between AMUSE and the GT. Please refer to the Sup. Mat. for details about the perceptual study.

5.4. Discussion and Future Work

Upper-body motion. We focus on the smooth coordination between the pelvis and upper body animation for side-by-side comparisons with other methods, as all other methods primarily focus on upper body movements. Future work should include lower-body motion and locomotion as these impact the perceived emotional state of a sequence.

Semantics. While the generated gestures, synchronized with the driving speech sequence, do not account for semantics such as deictic and metaphoric gestures, incorporating the text/language modality could help further improve in this direction.

Facial expressions. While emotional speech-driven face animation methods [18, 74] can be combined with bodies generated from AMUSE, jointly learning to generate emotional 3D bodies from speech is a topic that needs attention.

End-to-end training. Joint audio-gesture training may enhance results but requires careful loss term balancing and increased GPU memory. Therefore, we opted for separate training.

6. Conclusion

We present AMUSE, a framework to generate emotional body gestures from speech. The emotions and personal styles of the synthesized gestures can be controlled, thanks to the disentanglement of content, emotion, and style directly from the speech. The latent diffusion-based framework can further generate variations of the same gesture with the same emotion. Our quantitative evaluations show that AMUSE achieves state of the art performance on a variety of metrics: diversity, gesture emotion classification accuracy, Frechét gesture distance, beat alignment score, and semantic relevant gesture recall. Finally, our perceptual study demonstrates that AMUSE generates motions that are better synchronized and better match the emotion expressed of the input speech than previous state of the art.

Acknowledgments. We thank A Cseke, T McConnell and T Alexiadis for help with the perceptual study, B Pellkofer, J Piles-Contreras, and E Fritzler for IT support, and P Kulits, M Petrovich, S Kokane, and S Zojaji for proofreading and insightful discussions. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 860768.

Disclosure. https://download.is.tue.mpg.de/amuse/Disclosure.txt

References

[1] Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Unpaired motion style transfer from video to animation. *Transactions on Graphics (TOG)*, 2020. 3

[2] Chaitanya Ahuja and Louis-Philippe Morency. Language2Pose: Natural language grounded pose forecasting. *International Conference on 3D Vision (3DV)*, 2019. 2

[3] Chaitanya Ahuja, Dong Won Lee, and Louis-Philippe Morency. Low-resource adaptation for personalized co-speech gesture generation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 20566–20576, 2022. 2

[4] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. Style-controllable speech-driven gesture synthesis using normalising flows. *Computer Graphics Forum (CGF)*, 39(2):487–496, 2020. 3, 8

[5] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, Denoise, Action! Audio-driven motion synthesis with diffusion models. *Transactions on Graphics (TOG)*, 42(4):1–20, 2023. 1, 3, 6

[6] Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. Rhythmic gesticulator. *Transactions on Graphics (TOG)*, 41(6):1–19, 2022. 3

[7] Tenglong Ao, Zeyi Zhang, and Libin Liu. GestureDiffuCLIP: Gesture diffusion model with CLIP latents. *Transactions on Graphics (TOG)*, 2023. 1, 3, 6

[8] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. TEACH: Temporal Action Compositions for 3D Humans. In *International Conference on 3D Vision (3DV)*, 2022. 2

[9] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. SINC: Spatial composition of 3D human motions for simultaneous action generation. *International Conference on Computer Vision (ICCV)*, 2023. 2

[10] Emad Barsoum, John Kender, and Zicheng Liu. HP-GAN: probabilistic 3d human motion prediction via GAN. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017. 2

[11] Uttaran Bhattacharya, Elizabeth Childs, Nicholas Rewkowski, and Dinesh Manocha. Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning. In *ACM International Conference on Multimedia (MM)*, 2021. 3

[12] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *IEEE Virtual Reality and 3D User Interfaces, VR*, 2021. 3

[13] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, pages 213–229. Springer, 2020. 4

[14] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. *BEAT: The Behavior Expression Animation Toolkit*, pages 163–185. Springer, 2004. 2

[15] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Computer Vision and Pattern Recognition (CVPR)*, pages 18000–18010, 2023. 4, 5, 6

[16] Enric Corona, Albert Pumarola, G. Alenyà, and F. Moreno-Noguer. Context-aware human motion prediction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6990–6999, 2020. 2

[17] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[18] Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael Black, and Timo Bolkart. Emotional speech-driven animation with content-emotion disentanglement. In *International Conference on Computer Graphics and Interactive Techniques in Asia (SIGGRAPH ASIA)*, 2023. 2, 3, 8

[19] Paul Ekman. Facial expression and emotion. *American Psychologist*, 48(4):384–392, 1993. 1

[20] Esam Ghaleb, Ilya Burenko, Marlou Rasenberg, Wim Pouw, Peter Uhrig, Judith Holler, Ivan Toni, Aslı Özyürek, and Raquel Fernández. Co-speech gesture detection through multi-phase sequence labeling. In *Winter Conference on Applications of Computer Vision (WACV)*, 2024. 3

[21] Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F. Troje, and Marc-André Carbonneau. Zeroeggs: Zero-shot example-based gesture generation from speech. *Computer Graphics Forum (CGF)*, 42(1):206–216, 2023. 3

[22] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[23] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3497–3506, 2019. 2

[24] Ross B. Girshick. Fast R-CNN. In *International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015. 5

[25] Yuan Gong, Yu-An Chung, and James R. Glass. AST: audio spectrogram transformer. In *Interspeech 2021*, pages 571–575, 2021. 4, 5

[26] Yuan Gong, Cheng-I Lai, Yu-An Chung, and James R. Glass. SSAST: self-supervised audio spectrogram transformer. In *AAAI Conference on Artificial Intelligence*, pages 10699–10709, 2022. 4, 5

[27] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2Motion: Conditioned generation of 3d human motions. In *ACM International Conference on Multimedia (MM)*, 2020. 2

[28] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3D human motions from text. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[29] Nejla Gürefe. The role of gestures in mathematical discourse of hard-hearing students: Prism example. *Acta Didactica Napocensia*, 11:125–140, 2018. 1

[30] I. Habibie, Daniel Holden, Jonathan Schwarz, J. Yearsley, and T. Komura. A recurrent variational autoencoder for human motion synthesis. In *British Machine Vision Conference (BMVC)*, 2017. 2

[31] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. Learning speech-driven 3d conversational gestures from video. In *ACM International Conference on Intelligent Virtual Agents (IVA)*, 2021. 1, 6, 7

[32] Ikhsanul Habibie, Mohamed A. Elgharib, Kripasindhu Sarkar, Ahsan Abdullah, Simbarashe Linval Nyatsanga, Michael Neff, and Christian Theobalt. A motion matching-based framework for controllable gesture synthesis from speech. *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 2022. 2

[33] Félix G. Harvey, Mike Yurick, Derek Nowrouzezahrai, and Cristopher Joseph Pal. Robust motion in-betweening. *Transactions on Graphics (TOG)*, 2020. 2

[34] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. MoGlow: Probabilistic and controllable motion synthesis using normalising flows. *Transactions on Graphics (TOG)*, 39(4):236:1–236:14, 2020. 3, 6

[35] Sandra Herbert. Gesture types for functions. *Mathematics Education Research Group of Australasia*, pages 322–329, 2012. 1

[36] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 4, 5

[37] Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *International Conference on Computer Vision (ICCV)*, pages 1510–1519, 2017. 3

[38] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *Transactions on Graphics (TOG)*, 2017. 3

[39] Adam Kendon. *Gesture: Visible action as utterance*. Cambridge University Press, 2004. 1

[40] Gwantae Kim, Seonghyeok Noh, Insung Ham, and Hanseok Ko. MPE4G : Multimodal pretrained encoder for co-speech gesture generation. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 2

[41] Jihoon Kim, Taehyun Byun, Seungyoung Shin, Jungdam Won, and Sungjoon Choi. Conditional motion in-betweening. *Pattern Recognition*, 2022. 2

[42] Chris L Kleinke. Gaze and eye contact: A research review. *Psychological Bulletin*, 100(1):78–100, 1986. 1

[43] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. DiffWave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations (ICLR)*, 2021. 3

[44] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N. Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R. Thórisson, and Hannes Vilhjálmsson. Towards a common framework for multimodal generation: The behavior markup language. In *Intelligent Virtual Agents (IVA)*, pages 205–217. Springer, 2006. 2

[45] Quoc Anh Le, Souheil Hanoune, and Catherine Pelachaud. Design and implementation of an expressive gesture model for a humanoid robot. In *IEEE-RAS International Conference on Humanoid Robots*, pages 134–140, 2011. 2

[46] Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha S. Srinivasa, and Yaser Sheikh. Talking With Hands 16.2M: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In *International Conference on Computer Vision (ICCV)*, pages 763–772, 2019. 2

[47] Robert W Levenson. Blood, sweat, and fears: The autonomic architecture of emotion. *Annals of the New York Academy of Sciences*, 1000:348–366, 2003. 1

[48] Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11293–11302, 2021. 2

[49] Peizhuo Li, Kfir Aberman, Zihan Zhang, Rana Hanocka, and Olga Sorkine-Hornung. GANimator: Neural motion synthesis from a single sequence. *Transactions on Graphics (TOG)*, 2022. 2

[50] Ruilong Li, Shan Yang, D. A. Ross, and Angjoo Kanazawa. AI choreographer: Music conditioned 3D dance generation with AIST++. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[51] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. AI choreographer: Music conditioned 3d dance generation with AIST++. In *International Conference on Computer Vision (ICCV)*, pages 13381–13392, 2021. 6

[52] William Li and Philippe Pasquier. Automatic affect classification of human motion capture sequences in the valence-arousal model. In *International Symposium on Movement and Computing*, 2016. 3

[53] Yuanzhi Liang, Qianyu Feng, Linchao Zhu, Li Hu, Pan Pan, and Yi Yang. Seeg: Semantic energized co-speech gesture generation. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[54] Haiyang Liu, Naoya Iwamoto, Zihao Zhu, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. DisCo: Disentangled implicit content and rhythm learning for diverse co-speech gestures synthesis. In *ACM International Conference on Multimedia (MM)*, pages 3764–3773, 2022. 3

[55] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. BEAT: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. *European Conference on Computer Vision (ECCV)*, 2022. 2, 3, 5, 6, 7, 8

[56] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo

Zheng, and Michael J. Black. EMAGE: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 6

[57] Yuejiang Liu, Riccardo Cadei, Jonas Schweizer, Sherwin Bahmani, and Alexandre Alahi. Towards robust and adaptive motion forecasting: A causal representation perspective. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[58] Matthew M. Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion and shape capture from sparse markers. *Transactions on Graphics (TOG)*, 33(6):220:1–220:13, 2014. 5

[59] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. 5

[60] Shuhong Lu, Youngwoo Yoon, and Andrew W. Feng. Co-speech gesture synthesis using discrete gesture token learning. *International Conference on Intelligent Robots and Systems (IROS)*, pages 9808–9815, 2023. 2

[61] Birgit Lugrin, Catherine Pelachaud, and David Traum, editors. *The Handbook on Socially Interactive Agents: 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition*. 2021. 1

[62] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*, 2019. 2, 5

[63] Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. Virtual character performance from speech. In *SIGGRAPH/Eurographics Symposium on Computer Animation (SCA)*, page 25–35, 2013. 2

[64] Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[65] David Mcneill. Hand and mind: What gestures reveal about thought. *Bibliovault OAI Repository, the University of Chicago Press*, 27, 1994. 1

[66] Albert Mehrabian. *Nonverbal communication*. Aldine-Atherton Chicago, 1972. 1

[67] Shivam Mehta, Siyang Wang, Simon Alexanderson, Jonas Beskow, Éva Székely, and Gustav Eje Henter. Diff-TTSG: Denoising probabilistic integrated speech and gesture synthesis. In *Proc. ISCA Speech Synthesis Workshop (SSW)*, pages 150–156, 2023. 2

[68] Davide Moltisanti, Jinyi Wu, Bo Dai, and Chen Change Loy. BRACE: the breakdancing competition dataset for dance motion synthesis. In *European Conference on Computer Vision (ECCV)*, pages 329–344, 2022. 2

[69] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*, pages 8162–8171, 2021. 4

[70] Dirk Ormoneit, Michael J. Black, T. Hastie, and H. Kjellström. Representing cyclic human motion using functional analysis. *Image and Vision Computing (IVC)*, 2005. 2

[71] Kunkun Pang, Dafei Qin, Yingruo Fan, Julian Habekost, Takaaki Shiratori, Junichi Yamagishi, and Taku Komura. BodyFormer: Semantics-guided 3d body gesture synthesis with transformer. *Transactions on Graphics (TOG)*, 42(4), 2023. 2

[72] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*. ISCA, 2019. 5

[73] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2, 3

[74] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. EmoTalk: Speech-driven emotional disentanglement for 3d face animation. In *International Conference on Computer Vision (ICCV)*, pages 20687–20697, 2023. 2, 3, 8

[75] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[76] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *International Conference on Computer Vision (ICCV)*, 2021. 4, 6

[77] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[78] Paolo Petta, Catherine Pelachaud, and Roddy Cowie. *Emotion-Oriented Systems: The Humaine Handbook*. Springer Publishing Company, Incorporated, 2011. 1

[79] Rosalind W. Picard. *Affective Computing*. MIT Press, 1997. 1

[80] I. Poggi, C. Pelachaud, F. de Rosis, V. Carofiglio, and B. De Carolis. *Greta. A Believable Embodied Conversational Agent*, pages 3–25. Springer, 2005. 2

[81] Xingqun Qi, Chen Liu, Lincheng Li, Jie Hou, Haoran Xin, and Xin Yu. EmotionGesture: Audio-driven diverse emotional co-speech 3d gesture generation, 2023. 3

[82] Xingqun Qi, Chen Liu, Muyi Sun, Lincheng Li, Changjie Fan, and Xin Yu. Diverse 3D hand gesture prediction from body dynamics by bilateral hand disentanglement. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4616–4626, 2023. 2

[83] 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. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. 3

[84] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2022. 4, 5

[85] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI*, pages 234–241, 2015. 4

[86] Tim Salzmann, Marco Pavone, and Markus Ryll. Motron: Multimodal Probabilistic Human Motion Forecasting. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[87] Klaus R Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40 (1-2):227–256, 2003. 1

[88] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. `https://github.com/mseitzer/pytorch-fid`, 2020. Version 0.3.0. 6

[89] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021. 5

[90] Kim Sung-Bin, Lee Hyun, Da Hye Hong, Suekyeong Nam, Janghoon Ju, and Tae-Hyun Oh. Laughtalk: Expressive 3d talking head generation with laughter. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 6404–6413, 2024. 2

[91] Mingyang Sun†, Mengchen Zhao†, Yaqing Hou*, Minglei Li, Huang Xu, Songcen Xu, and Jianye Hao. Co-speech gesture synthesis by reinforcement learning with contrastive pre-trained rewards. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[92] Marcus Thiébaux, Stacy Marsella, Andrew N. Marshall, and Marcelo Kallmann. SmartBody: behavior realization for embodied conversational agents. In *Adaptive Agents and Multi-Agent Systems (AAMAS)*, 2008. 2

[93] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, pages 10347–10357, 2021. 4

[94] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[95] R. Williams. *The Animator's Survival Kit: A Manual of Methods, Principles and Formulas for Classical, Computer, Games, Stop Motion and Internet Animators*. Farrar, Straus and Giroux, 2012. 2

[96] Jingyu Wu, Shi Chen, Shuyu Gan, Weijun Li, Changyuan Yang, and Lingyun Sun. Cultural self-adaptive multimodal gesture generation based on multiple culture gesture dataset. In *ACM International Conference on Multimedia (MM)*, page 3538–3549, 2023. 3

[97] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, Ming Cheng, and Long Xiao. Diffusestylegesture: Stylized audio-driven co-speech gesture generation with diffusion models. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2023. 3, 6

[98] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, and Haolin Zhuang. Qpgesture: Quantization-based and phase-guided motion matching for natural speech-driven gesture generation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2321–2330, 2023. 2

[99] Payam Jome Yazdian, Mo Chen, and Angelica Lim. Gesture2vec: Clustering gestures using representation learning methods for co-speech gesture generation. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 3100–3107, 2022. 2

[100] Hongwei Yi. Show github. `https://github.com/yhw-yhw/SHOW`, 2023. 6

[101] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *Computer Vision and Pattern Recognition (CVPR)*, pages 469–480, 2023. 1, 2, 6, 7

[102] Li-Ping Yin, Yijun Wang, Tianyu He, Jinming Liu, Wei Zhao, Bohan Li, Xin Jin, and Jianxin Lin. Emog: Synthesizing emotive co-speech 3d gesture with diffusion model. *ArXiv*, abs/2306.11496, 2023. 3

[103] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *International Conference on Robotics and Automation (ICRA)*, 2019. 6

[104] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *Transactions on Graphics (TOG)*, 39(6), 2020. 2, 6

[105] Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. The genea challenge 2022: A large evaluation of data-driven co-speech gesture generation. In *International Conference on Multimodal Interaction*, 2022. 2

[106] Ye Yuan and Kris M. Kitani. DLow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[107] Yicheng Zhong, Huawei Wei, Peiji Yang, and Zhisheng Wang. Expclip: Bridging text and facial expressions via semantic alignment. In *AAAI Conference on Artificial Intelligence*, pages 7614–7622, 2024. 2

[108] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5745–5753, 2019. 3

[109] Yi Zhou, Jingwan Lu, Connelly Barnes, Jimei Yang, Sitao Xiang, and Hao Li. Generative tweening: Long-term inbetweening of 3d human motions. *arXiv:2005.08891*, 2020. 2

[110] Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. Taming diffusion models for audio-driven co-speech gesture generation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10544–10553, 2023. 2, 8