No Gestures Left Behind: Learning Relationships between Spoken Language and Freeform Gestures

Chaitanya Ahuja, Dong Won Lee, Ryo Ishii, Louis-Philippe Morency
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA
{cahuja,dongwonl,rishii,lmorency}@cs.cmu.edu

Abstract

We study relationships between spoken language and co-speech gestures in context of two key challenges. First, distributions of text and gestures are inherently skewed making it important to model the long tail. Second, gesture predictions are made at a subword level, making it important to learn relationships between language and acoustic cues. We introduce Adversarial Importance Sampled Learning (or AISLe), which combines adversarial learning with importance sampling to strike a balance between precision and coverage. We propose the use of a multimodal multiscale attention block to perform subword alignment without the need of explicit alignment between language and acoustic cues. Finally, to empirically study the importance of language in this task, we extend the dataset proposed in Ahuja et al. (2020) with automatically extracted transcripts for audio signals. We substantiate the effectiveness of our approach through large-scale quantitative and user studies, which show that our proposed methodology significantly outperforms previous state-of-the-art approaches for gesture generation. Link to code, data and videos: https://github.com/chahuja/aisle

1 Introduction

Spoken language has gained more traction in the past decade due to improvements in natural language understanding and speech recognition. With an eye on the future, technologies such as intelligent personal assistants (e.g. Alexa, Siri, Cortana) are likely to also include embodiment to take advantage of the non-verbal communication that people naturally use in face-to-face interactions. As a stepping stone in this direction, it is important to study the relationship between spoken language (which also includes acoustic information) and free form gestures (which go beyond just a pre-defined dictionary of gesture animations). In other words, how can we automatically generate human body pose (gestures) from language and acoustic inputs?

An important technical challenge in such a natural language processing task, is modeling the long tail of the language-gesture distribution (see Figure 1). If not addressed directly, computational models will likely focus on the common gestures (e.g beat gestures) as a way to improve precision at the cost of reduced coverage for less frequent words and gestures (Ginosar et al., 2019). Hence, when learning these models, we need to not only be accurate for gesture generation, but also handle coverage of both linguistic and visual distributions (Pelachaud, 2009; Kucherenko et al., 2019). In other words, we need models that can balance precision and coverage. Another technical challenge comes from the differences in granularity between language and gestures. Gestures can be triggered at the sub-word level; for example, by a change of intonation in acoustics. Thus, it is important to have sub-word level alignment between language and acoustics to generate the freeform gestures.

In this paper, we study the link between spoken language and free form gestures. As a first contribution, we propose Adversarial Importance Sampled Learning(or AISLe), an approach whose main novelty is to bring adversarial learning and importance sampling together to improve coverage of the generated distribution without compromising on the precision at no extra computational cost. As a second contribution, we introduce the use of neural cross-attention architecture (Vaswani et al., 2017; Tsai et al., 2019) for gesture generation conditioned on spoken language. This idea allows transformer blocks to help with subword alignment between language and acoustic signals. A third contribution is the extension of dataset proposed in Ahuja et al. (2020) with automatically

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1884–1895
November 16 - 20, 2020.

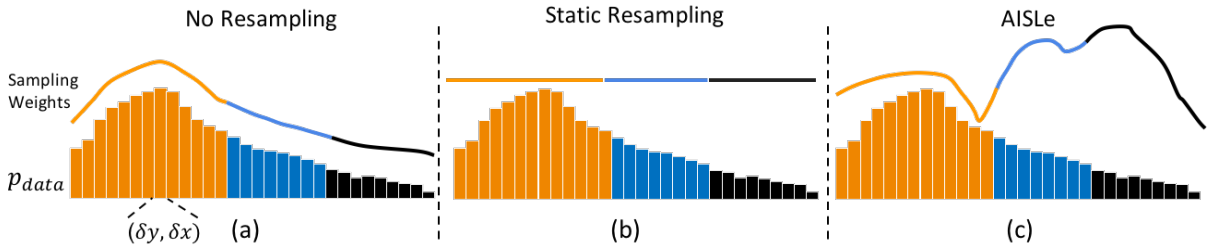

Figure 1: A toy representation of data distribution p_{data} as a histogram. Colours ■, ■, ■ represent bins from the mode, heavy tail and long tail of p_{data} respectively. The color coded envelope covering p_{data} is the distribution of weights across bins $(\delta y, \delta x)$ for the following resampling techniques: (a) No Resampling, (b) Static Resampling, and (c) AISLe. While p_{data} is a multivariate distribution, we use a 1-dimensional histogram for the sake of demonstration.

extracted transcripts for audio signals corresponding to 250+ hours of freeform gesture information and 25 speakers. Our experiments study the effectiveness of our proposed method with a focus on precision-coverage trade-off. These quantitative experiments are complimented with important subjective human studies as the englobing judges of the generation quality.

2 Related Work

Language and Speech for Gesture Generation An early study by Cassell et al. (2001) proposed the behavior expression animation toolkit (BEAT) that can select and schedule behaviors, such as hand gestures, head nods and gaze, which was extended by applying behavior decision rules to the linguistic information obtained from input text (Lee and Marsella, 2006; Marsella et al., 2013; Lhommet et al., 2015; Lhommet and Marsella, 2016; Xu et al., 2014). Rule based approaches were replaced by deep conditional neural fields (Chiu et al., 2015; Chiu and Marsella, 2014) and Hidden Markov Models for prosody-driven head motion generation (Sargin et al., 2008) and body motion generation (Levine et al., 2009, 2010). These use a dictionary of predefined animations, limiting the diversity of generated gestures.

Moving forward, neural networks were employed to predict a sequence of frames for gestures (Hasegawa et al., 2018), head motions (Sadoughi and Busso, 2018) and body motions (Shlizerman et al., 2018; Ahuja et al., 2019; Ginosar et al., 2019; Ferstl et al., 2019) conditioned on a speech input while Yoon et al. (2019) uses only a text input. Unlike these approaches, Kucherenko et al. (2020) rely on both speech and language for gesture generation. But their choice of early fusion to combine the modalities ignores multi-scale correlations (Tsai et al., 2019) between speech and language.

While publicly datasets of co-speech gestures are available, they are either small (Sadoughi et al., 2015; Tolins et al., 2016; Yoon et al., 2019) or do not contain language information (Ginosar et al., 2019; Joo et al., 2015; Lee et al., 2019), which motivates for a dataset that resolves these shortcomings.

Distribution Coverage in Generative Modeling Implicit generative models have seen a lot of progress in the past decade with the introduction of GANs (Goodfellow et al., 2014; Yan and Wang, 2017). Especially two aspects of distribution estimation, (1) conditional generation *precision* (Zhang et al., 2017; Ginosar et al., 2019; Isola et al., 2017; Mirza and Osindero, 2014) and (2) *coverage* of the entire underlying distribution (Sharma and Namboodiri, 2018; Zhong et al., 2019; Tolstikhin et al., 2017; Arjovsky et al., 2017) have gained traction.

To tackle the precision-coverage trade-off, methods have been introduced for out-of-distribution detection but they do not work for implicit models like GANs (Nalisnick et al., 2019). These approaches have similarities to importance weighting (Byrd and Lipton, 2018; Katharopoulos and Fleuret, 2018), which are often used for post-hoc debiasing of the learnt model (Domke and Sheldon, 2018; Grover et al., 2019; Turner et al., 2018), correcting covariate shift (Shimodaira, 2000), label shift (Lipton et al., 2018; Garg et al., 2020), imitation learning (Murali et al., 2016; Kostrikov et al., 2018) and curriculum learning (Jiang et al., 2015; Bengio et al., 2009; Matiisen et al., 2019). Byrd and Lipton (2018) observe that sub-sampling from unbalanced categorical classes demonstrates a significant effect on the network's predictions. Importance sampling

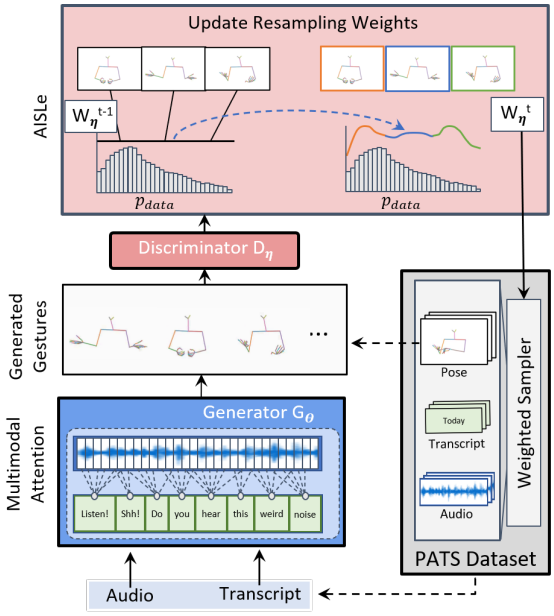

Figure 2: Overview of the key components of our model. Starting at the dataset and going clockwise, audio and transcripts go through sub-word alignment in the generator G_θ and are decoded to generate a freeform gesture animation. Next, the AISLe updates the weighted sampler of the dataset based on the output of the discriminator D_η to complete the loop.

in GANs (Diesendruck et al., 2019; Li et al., 2017; Yi et al., 2019), which uses re-weighting of maximum mean discrepancy between source and target distributions, has shown to improve the coverage in cases of unbalanced datasets, but do not provide insights on precision and coverage in the presence of conditional inputs.

3 Problem Statement

The goal of this cross-modal translation task is to generate a series of freeform gestures that are aligned with the spoken sentence (see Figure 2). By free form gestures, we refer to a sequence of joint positions (a.k.a. poses) of the upper human body including neck, torso, arms, hands and fingers. On our way to achieving this goal we work towards solving two challenges: (1) generating gestures from the long-tail of the language-gesture distribution while maintaining high precision of these generated gestures and, (2) sub-word level alignment of language, acoustic cues and gestures to account for the differences in frame rates between among these modalities.

Formally, we are given a sentence of K language tokens $\mathbf{X}^w = \left[x_0^w, x_1^w, \ldots x_{K-1}^w\right]$ which has a dynamic frame rate -i.e. each token has a variable time duration dependent on its context- as compared to the fixed frame rate of a sequence of speech features, $\mathbf{X}^a = \left[x_0^a, x_1^a, \ldots x_{T-1}^a\right]$. We want to predict a sequence of T gesture poses $\mathbf{Y}^p = \left[y_0^p, y_1^p, \ldots y_{T-1}^p\right]$ that co-occur with $\mathbf{X}^a$ and $\mathbf{X}^w$. Here $y_t^p \in \mathcal{R}^{J\times 2}$ are the xy-coordinates for t^{th} frame for J joints of the body skeleton.

This problem can be formalized as learning a true conditional probability distribution $p_{data}(y|x)$ of output $y = \mathbf{Y}^p$, given input $x = \{\mathbf{X}^a, \mathbf{X}^w\}$ consisting of text and speech. We write this in form of a generator function G_θ with trainable parameters θ as:

$$\hat{\mathbf{Y}}^p = G_\theta(\mathbf{X}^a, \mathbf{X}^w) \quad (1)$$

$$= G_{dec}\left(G_{attn}\left(G_{enc}^a(\mathbf{X}^a), G_{enc}^w(\mathbf{X}^w)\right)\right) \quad (2)$$

where $\hat{\mathbf{Y}}^p$ are generated poses from the learnt conditional distribution $p_\theta(y|x)$, which is an approximation of p_{data}. G_{enc}^a and G_{enc}^w are the acoustic and language encoders, G_{attn} is the multimodal attention block and G_{dec} is the pose decoder.

All our experiments are in an adversarial set-up to alleviate the challenge of overly smooth generation (Ginosar et al., 2019) caused by the reconstruction loss $\mathcal{L}_{rec} = \mathbb{E}_{\mathbf{Y}^p,\mathbf{X}^a,\mathbf{X}^w}\|\mathbf{Y}^p - G_\theta(\mathbf{X}^a, \mathbf{X}^w)\|_1$. The generated pose sequence $\hat{\mathbf{Y}}^p$ is fed as a signal for the adversarial discriminator D_η, which tries to classify the true pose $\mathbf{Y}^p$ from the generated pose $\hat{\mathbf{Y}}^p$. This is jointly trained with the generator, which learns to fool the discriminator by generating realistic poses. This adversarial loss (Goodfellow et al., 2014) is written as:

$$\begin{aligned}\mathcal{L}_{adv} \quad = \quad & \mathbb{E}_{\mathbf{Y}^p} \log D_\eta\left(\mathbf{Y}^p\right) \\ & + \mathbb{E}_{\mathbf{X}^a,\mathbf{X}^w} \log\left(1 - D_\eta(G_\theta\left(\mathbf{X}^a, \mathbf{X}^w\right)\right)\end{aligned} \quad (3)$$

The model is jointly trained to optimize the overall loss function $L(y, x)$,

$$\max_\eta \min_\theta \mathcal{L}_{rec} + \mathcal{L}_{mix} + \mathcal{L}_{adv} \quad (4)$$

where $\mathcal{L}_{mix}$ is a loss for training mixture of generators and defined in Section 4.3.

4 Model

In this section, we present our Adversarial Importance Sampled Learning (or AISLe) paradigm which is designed to improve coverage while learning accurate relationships between spoken language and gestures. This contribution is described

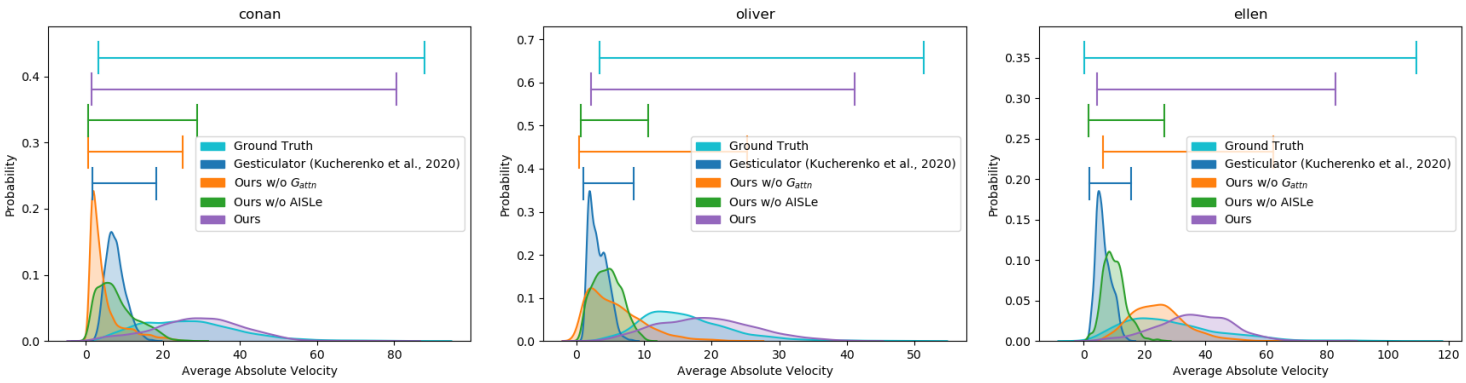

Figure 3: Distribution of the generated gestures with average absolute velocity as the statistic for three different speakers. The support (or coverage) of the distribution is denoted with the colour coded lines at the top of each plot. Larger overlap of a model's distribution with the ground truth distribution is desirable.

in Section 4.1. Our second contribution is the application of a transformer architecture to the problem of sub-word alignment between language and acoustic features. This model Multimodal Multi-Scale Transformer (MMS-Transformer) is presented in Section 4.2. The remaining components of our full model; pose decoder G_{dec}, language encoder G^{w}_{enc} and acoustic encoder G^{a}_{enc} are presented in Section 4.3. The key contributions are illustrated in Figure 2 and can be summarized by optimizing the overall loss function $\mathcal{L}(y, x)$ with AISLe in Algorithm 1.

4.1 Adversarial Importance Sampled Learning (or AISLe)

To improve coverage, we want to be sure that the learnt distribution $p_\theta(y|x)$ is a good approximation of the underlying distribution $p_{data}(y|x)$, including the long tail. Our intuition to solve this problem is to have our model give adaptive importance to the long tail of the gesture distribution while still allowing access to the more likely regions (i.e. modes) of the distribution (see Figure 1). This can be achieved by introducing a multiplicative weight factor $w_\eta(x) = \frac{p_\theta(\tilde{y}|x)}{p_{\text{data}}(\tilde{y}|x)}$ to the expected loss function,

$$\mathop{\mathbb{E}}_{x\sim p(.)} \mathop{\mathbb{E}}_{\substack{y\sim p_{\text{data}}(.|x)\\ \tilde{y}\sim p_\theta(.|x)}} \frac{p_\theta(\tilde{y}|x)}{p_{\text{data}}(\tilde{y}|x)} \mathcal{L}(y, x) \qquad (5)$$

where $\mathcal{L}(y, x)$ is the overall loss function and $p(x)$ is the marginal distribution of the input (i.e. language and acoustics). At a high level, as training progresses, if the generated sample has more likelihood of being generated by the learnt distribution than the true data distribution, it is given more importance. As this process reaches a desired equilibrium, where $p_\theta \xrightarrow{p} p_{data}$, $w_\eta(x)$ will approach 1 and revert back to the unweighted loss function.

We first derive this weighted function, then show how w_η can be estimated practically in tandem with the adversarial setup of our problem without any additional computational cost, Finally, we tie it all up with an algorithm for AISLe.

Deriving the Weighted Loss Function: Unlike prior work (Katharopoulos and Fleuret, 2018; Diesendruck et al., 2019), we derive the weighted cost function in Equation 5 using first principles. As illustrated in Figure 1, we divide the support of p_{data} into a grid of multi-dimensional bins of size $(\delta y, \delta x) \in \mathbb{R}^{\dim(y)+\dim(x)}$ where $\dim(.)$ gives dimensions of a variable. If $(\delta y, \delta x)$ is sufficiently small, it is a reasonable assumption that all samples (i.e. pair of poses and spoken words) in this bin will be close to each other. Hence, if the model was to see some, and not all of the samples in this bin, it would still be able to learn the dynamics between poses and spoken words. As bins in the mode of the distribution have more samples than bins in the tail, the model would learn from samples in the tail less often if we optimize over an unweighted loss function given by $\mathbb{E}_{x\sim p(.)}\, \mathbb{E}_{y\sim p_{\text{data}}(.|x)}\, \mathcal{L}(y, x)$. This is visually illustrated by the weights proportional to bin frequency in Figure 1(a).

To counteract this imbalance, we first perform a *static rebalance* of the expected cost by assigning the same weight to each bin as shown in Figure 1(b). This encourages that equal number of samples are

drawn from each bin while training,

$$\mathop{\mathbb{E}}_{x\sim p(.)} \mathop{\mathbb{E}}_{\substack{y\sim p_{\text{data}}(.|x)\\ \tilde{y}\sim p_\theta(.|x)}} \frac{1}{p_{\text{data}}(\tilde{y}|x)}\mathcal{L}(y,x) \tag{6}$$

Second, the importance of each bin is proportional to the likelihood of generated sample belonging to the proposal distribution p_θ, i.e. if a sample is more likely to have been generated by p_θ than p_{data}, then the model has yet to learn the corresponding bin. Multiplying p_θ to the numerator in Equation 6 gives us Equation 5. This appears as adaptive weighting across the support of the data distribution as shown in Figure 1(c).

Estimation of Importance Weights: We follow a likelihood-free approach (Grover et al., 2019; Turner et al., 2018) to estimate w_η by computing the outputs of the discriminator D_η. Rewriting w_η in Equation 5 as,

$$w_\eta(x) = \frac{1 - D_\eta(G_\theta(x))}{D_\eta(G_\theta(x))} \tag{7}$$

As D_η is learnt while optimizing $\mathcal{L}(y,x)$ and is computed for every training iteration, there is no additional computational cost in estimating weights while training. The estimated importance weights are used for data duplication while training (Diesendruck et al., 2019), which is an equivalent alternative to optimize weighted loss functions. We illustrate the weight update cycle in Algorithm 1.

Algorithm 1: Adversarial Importance Sampled Learning

initialization;
$w_\eta(\tilde{y}) \leftarrow 1, \forall \tilde{y}$;
datasetSampler.updateWeights(w_η);
for *count in numEpochs* **do**
  for x_{batch} *in datasetSampler* **do**
    $w_\eta(batch) \leftarrow \frac{1-D_\eta(G_\theta(x_{batch}))}{D_\eta(G_\theta(x_{batch}))}$;
    ⋮
    Model Training;
  end
  # keep weights around 1;
  $w_\eta \leftarrow \frac{w_\eta - \text{mean}(w_\eta)}{\text{std}(w_\eta)} + 1$;
  # clip weights to lie in (0.1, 10) ;
  $w_\eta \leftarrow \text{clip}(w_\eta, 0.1, 10)$;
  datasetSampler.updateWeights(w_η);
end

4.2 Multimodal Multiscale Attention Block

To address the challenge of sub-word alignment, we take inspiration from recent work self-attention (Vaswani et al., 2017) and cross-attention models (Tsai et al., 2019) to alleviate the need of explicit alignment between audio and language embeddings. Note that these modalities provide complimentary information for gesture prediction: audio estimates rhythm, pauses and speed of the gestures (i.e. beat gestures) while language can be helpful for iconic or metaphoric gestures (Cassell, 2001). A multimodal attention mechanism can make use of sub-word information from the audio to drive well-timed and meaningful gesture animation.

Consider a temporal sequence of audio embeddings $G^a_{enc}(\mathbf{X}^a) = \mathbf{Z}^a \in \mathcal{R}^{T\times h^a}$ and language embeddings $G^w_{enc}(\mathbf{X}^w) = \mathbf{Z}^w \in \mathcal{R}^{N\times h^w}$. We define audio query as $\mathbf{Q}^a = \mathbf{Z}^a\mathbf{W}_{\mathbf{Q^a}}$, language key as $\mathbf{K}^w = \mathbf{Z}^w\mathbf{W}_{\mathbf{K^w}}$ and language values as $\mathbf{V}^w = \mathbf{Z}^w\mathbf{W}_{\mathbf{V^w}}$. Here $\mathbf{W}_{\mathbf{Q^a}} \in \mathcal{R}^{h^a\times h}$, $\mathbf{W}_{\mathbf{K^w}} \in \mathcal{R}^{h^w\times h}$ and $\mathbf{W}_{\mathbf{V^w}} \in \mathcal{R}^{h^w\times h}$ are trainable weights. Sub-word information from audio is learnt via a cross modal attention CM.

$$\mathbf{Z}^{aw} = \text{CM}(\mathbf{Z}^a, \mathbf{Z}^w) = \text{softmax}\left(\frac{\mathbf{Q}^a\mathbf{K}^{wT}}{\sqrt{h^a}}\right)\mathbf{V}^w \tag{8}$$

Unlike (Tsai et al., 2019), we precede cross-modal attention with a layer of self attention (Vaswani et al., 2017) which learns correlations between the low-level language features before assessing sub-word information from the audio modality. After cross-modal attention, we add layer normalization (Ba et al., 2016) followed by a point-wise feedforward layer along with residual connections as described in (Vaswani et al., 2017; Tsai et al., 2019; Devlin et al., 2018). Z^{aw} is now the same scale as the audio input and hence is concatenated with Z^a. This completes the multimodal multiscale attention block G_{attn}.

4.3 Other Network Components

Decoder G_{dec}: The decoder G_{dec} takes aligned multimodal representations from G_{attn} to generate output pose sequences. We start with a 1D U-Net (Ronneberger et al., 2015) following suit in (Ginosar et al., 2019) to get $\mathbf{Z} = \text{U-Net}([\mathbf{Z}^{aw}\mathbf{Z}^a])$. In addition, the distribution of gestures contains multiple modes. Hence, to prevent mode collapse we use mixture-model guided sub-generators (Ahuja et al., 2020; Hao et al., 2018; Arora et al., 2017;

Models	Expressivity	Naturalness	Relevance	Timing
S2G (Ginosar et al., 2019)	24.6 ± 3.1	22.1 ± 1.8	22.4 ± 1.7	27.6 ± 1.7
Gesticulator (Kucherenko et al., 2020)	31.9 ± 2.0	32.1 ± 1.7	31.4 ± 1.8	31.1 ± 1.7
Ours w/o G_{attn}	35.0 ± 2.3	29.2 ± 1.7	30.9 ± 1.8	30.8 ± 1.7
Ours w/o AISLe	35.8 ± 2.9	$\mathbf{35.7 \pm 1.7}$	33.7 ± 1.7	32.1 ± 1.7
Ours	$\mathbf{38.9 \pm 1.7}$	$\mathbf{36.7 \pm 1.6}$	$\mathbf{37.1 \pm 1.7}$	$\mathbf{35.3 \pm 1.7}$

Table 1: Human perceptual study comparing our model with prior work and strong baselines over four criteria measuring quality of co-speech gestures. we report the preference scores (higher is better) of a model as compared to the ground truth gestures. 90% confidence intervals around the mean performance and calculated by a bootstrapped t-test are also reported.

Hoang et al., 2018),

$$\hat{\mathbf{Y}}_p = \sum_{m=1}^{M} \phi_m G_m(\mathbf{Z}) \tag{9}$$

where $\forall m, G_m$ is the sub-generator function and ϕ_m is the corresponding mixture model prior. While training, the true value of ϕ_m can be estimated based on which sub-distribution the pose belongs to. At inference time, we do not have the ground truth pose to make such estimation. Instead, we train a classification network H to estimate ϕ_m at inference time based on the input embedding $\mathbf{Z}$. H is optimized via a mode regularization loss $\mathcal{L}_{mix} = \mathbb{E}_{\Phi,\mathbf{Z}}\text{CCE}(\Phi, H(\mathbf{Z}))$, where CCE is categorical cross-entropy and $\Phi = [\phi_1, .., \phi_M]$.

Language Encoder G^w_{enc}: In order to utilize the semantic and contextual information of language, we fine-tune BERT for the task of gesture generation (Devlin et al., 2018) using an existing implementation with pre-trained weights (Wolf et al., 2019). The contextual dependence allows the model to be exposed to semantic differences in the meaning of the same word. These embeddings at model contextual dependence only at the word level leaving sub-word level dynamics to the multimodal attention block G_{attn}.

Audio Encoder G^a_{enc}: For audio embeddings, we use a Temporal Convolutional Network (or TCNs), which has shown to perform well in speech-conditioned pose generation task (Ginosar et al., 2019; Ahuja et al., 2019). In our experiments, we use an audio encoder based on Temporal Convolution Networks consisting of a convolution layer, followed by batch normalization (Ioffe and Szegedy, 2015), and ReLU (Nair and Hinton, 2010). We use a similar TCN network for the discriminator $\mathbf{D}_\eta$[1].

[1]We refer the readers to the appendix for exact implementation and hyperparameters.

5 Experiments

5.1 Baseline Models

Speech2Gesture (Ginosar et al., 2019): Speech2Gesture does not use the text modality (i.e. no multimodal attention block) and any form of re-sampling while training.

Gesticulator (Kucherenko et al., 2020): Unlike MMS-Transformer , Gesticulator has a set of fully connected layer followed by autoregressive fully connected layers which are FiLM conditioned (Perez et al., 2018). In addition to audio and text, features of duration of each word (i.e. start, end, percentage completed and so on) are used as inputs. To align audio and text, each token (i.e. text) is replicated to match its duration, hence performing an explicit alignment between text and audio.

Ablation Models: Components **AISLe** and $\mathbf{G_{attn}}$ are removed from the model one at a time to measure its contribution in gesture generation for the first set of ablation models. **Static Rebalancing** (Equation 6), which is one step before AISLe, is also used as an ablation model. Finally, **top k%** highest velocity regions (or tails) are used as a sub-sampled dataset. This is a manual method of importance sampling high velocity gestures.

5.2 Evaluation Metrics

Human Perceptual Study: We conduct a human perceptual study on Amazon Mechanical Turk (AMT) to measure human preference towards generated animations on four criteria, (1) **naturalness**, (2) **expressivity**, (3) **timing** and (4) **relevance**. We show a pair of videos with skeletal animations to the annotators. One of the animations is from the ground-truth set, while the other is a generation from our proposed model or a baseline. With unlimited time and for each criterion, users have to choose one video which they felt was better. We

Model	Modality	Coverage ↓			Precision ↑	
		FID	W1 (vel.)	W1 (acc.)	PCK	F1
S2G (Ginosar et al., 2019)	**A**	68.1	12.5	15.8	**0.374**	0.189
Gesticulator (Kucherenko et al., 2020)	**A + T**	49.5	20.6	27.2	0.350	0.268
Ours	**A + T**	**27.8**	**7.3**	**10.6**	**0.376**	**0.317**
Ours w/o AISLe	**A + T**	55.3	12.4	22.2	**0.375**	**0.312**
Ours w/o G_{attn}	**A + T**	34.8	8.1	11.3	0.363	0.298

Table 2: Quantitative comparison of our model as compared to existing work, and ablations with one component missing at a time. Comparisons in ■ shows the impact of AISLe on coverage, while ■ shows the impact of G_{attn} in our model on precision

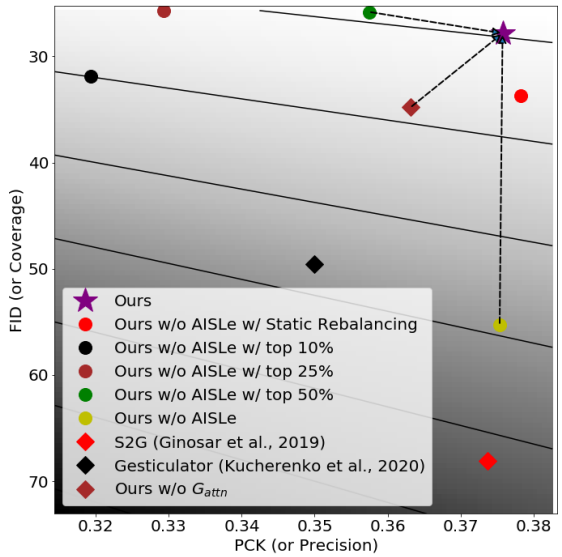

Figure 4: Precision Coverage Tradeoff for all models. Lighter areas represent high PCK and low FID which is favourable for the model. Contour lines corresponds constant values of $\frac{PCK}{FID}$. We show impacts of AISLe, G_{attn} and dataset subsampling with dotted lines traversing the PCK-FID plane, with our model enjoying the best of both worlds.

run this study for randomly selected with 20 pairs of videos per model per speaker from the held-out set, giving a total of 1500 sample points for each model. We refer the readers to the appendix for more details of the setup.

Precision: To measure the accuracy of the generated gesture we use two metrics, (1) **Probability of Correct Keypoints (PCK)** (Andriluka et al., 2014; Simon et al., 2017): the values are averaged over $\alpha = 0.1, 0.2$ as suggested in (Ginosar et al., 2019) and (2) **Mode Classification F1**: if the generated pose ($\hat{Y}^p$) lies in the same cluster as the ground truth, it was sampled from the correct mode. F1 measure, for this classification task, is used to measure correctness of gesture generation.

Coverage: to measure the coverage of the generated distribution we use two metrics, (1) **Fréchet Inception Distance (FID)**: distance between distributions of generated and ground truth poses(Heusel et al., 2017). (2) **Wasserstein-1 distances (or W1)**: distance between distribution of generated and ground truth average velocity. The same distance is calculated for average acceleration.

5.3 Pose, Audio, Transcripts and Style (PATS) dataset

We extend the Pose, Audio, Transcripts and Style (PATS) dataset (Ahuja et al., 2020) with automatically extracted transcripts for audio signals to study the effect of language and speech on co-speech gesture generation. It offers data for 25 speakers with diverse gestures and linguistic content (Ahuja et al., 2020; Ginosar et al., 2019). Specifically, it contains 15 talk show hosts, 5 lecturers, 3 YouTubers, and 2 televangelists, providing a total of 251 hours of video clips, with a mean of 10.7 seconds and a standard deviation of 13.5 seconds per clip.

5.3.1 Dataset Features

Aligned Transcriptions: As manual transcriptions are often not aligned and not readily available, we use Google Automatic Speech Recognition (Chiu et al., 2018) to collect subtitles and aligned timings of each spoken word. The average Word Error Rate of the transcriptions, calculated on the set of available transcriptions (i.e. subtitles), using the Fisher-Wagner algorithm is 0.29 (Navarro, 2001).

Pose: Each speaker's pose is represented via skeletal keypoints collected via OpenPose (Cao et al., 2018) following the approach in Ginosar et al. (2019). It consists of of 52 coordinates of an indi-

Model	Coverage ↓			Precision ↑	
	FID	W1 (vel.)	W1 (acc.)	PCK	F1
Ours	27.8	7.3	10.6	**0.376**	**0.317**
Ours w/o AISLe w/ Static Rebalancing	33.7	12.2	21.6	**0.378**	**0.314**
Ours w/o AISLe w/ top 100%	55.3	12.4	22.2	**0.375**	**0.312**
Ours w/o AISLe w/ top 50%	**25.8**	**5.2**	**7.6**	0.357	0.303
Ours w/o AISLe w/ top 25%	**25.7**	6.8	9.2	0.329	0.285
Ours w/o AISLe w/ top 10%	31.9	6.9	8.6	0.319	0.269

Table 3: Quantitative comparison of AISLe in our model with strong rebalancing baselines. Comparisons in ■ demonstrate the impact of adaptive sampling in AISLe on coverage, while ■ demonstrates robustness of AISLe in precision

vidual's major joints for each frame at 15 frames per second, which we rescale by holding the length of each individual's shoulder constant.

Audio: Following prior work (Kucherenko et al., 2019; Ginosar et al., 2019), we represent audio features as spectrograms, which is a rich input representation shown to be useful for gesture generation.

6 Results and Discussions

First, we study the effect of different components of our model on **coverage** and **precision**. We follow this up with the quantitative effects of dataset subsampling. Finally, we conclude with a discussion on the need of a precision-coverage trade-off for co-speech gesture generation. All models are trained separately for each of 25 speakers in PATS dataset and we report scores averaged over all speakers for comparison.

Comparison with previous baselines: We focus first on the human perceptual study in Table 1, since it is arguably the most important metric. We see a significantly[2] larger preference for our model as compared to S2G and Gesticulator for all four criteria. Specifically, *expressivity* sees the largest jump, indicating improved coverage in the generated gestures. A similar trend is seen on the objective scores for coverage in Table 2 which indicates a possible correlation between high coverage and human-judged expressivity of gestures. Interestingly, PCK score for S2G is not significantly different from ours, indicating that a simple accuracy metric may not be sufficient to judge performance in a co-speech gesture generation task.

Impact of AISLe on Coverage: Incorporating AISLe while training a generative model shows significant gains for coverage metrics in Table 3 ■. We observe that the use of Static Rebalancing (Equation 6) instead, which is an extreme version of AISLe, is better than not resampling at all. However, it is unable to reach the performance of AISLe on coverage metrics. A similar trend can be seen in the perceptual study scores in Table 1, where the addition of AISLe makes the generations preferable for most criteria. We also note that, while AISLe generates significant gains for coverage metrics, it still maintains the same level of *precision* as compared to Static Rebalancing.

Next, we visually compare the distribution of the generated gestures. We use average velocity of the body as a statistic as motion (or energy (Pelachaud, 2009)), which is one of the key indicators of naturalistic gestures. In Figure 3, we observe that our model(⊢) is able to (nearly) generate the velocity distribution of the ground truth. Models without AISLe shift the velocity of the generated distribution closer to zero indicating more gestures were generated with no or little motion, unlike the true data distribution (compare ⊢ and ⊢).

Impact of G_{attn} on precision: Removal of Multimodal Multiscale Attention Block (G_{attn}) from our model results in significant performance dip of precision metrics in Table 2 ■. Relevance of generated gestures to the corresponding spoken language also suffers a significant decrease without G_{attn} in Table 1. These support our hypothesis that a representation which explicitly learns subword attentions between text and audio is a better predictor of the corresponding gestures.

Impact of a Sub-sampled dataset on Precision and Coverage: We find, in Table 3 ■, that pruning the dataset to select samples which have a high

[2] significance refers to statistical significance inferred using a 90% confidence interval estimated by a 2-sided t-test

average velocity (or Ours w/o AISLe w/ top $x\%$), is a simple way of improving the support of the generated distribution. While this approach of re-sampling is a strong baseline for distribution *coverage*, it reduces the generalizability of the model -i.e. sharp decrease in PCK and F1 scores- probably due to the missing low velocity examples during training which is undesirable.

Precision Coverage Trade-off: We observe that models without AISLe may have comparable PCK scores to our model but have significantly worse coverage and hence are not close to the true gesture distribution. Furthermore, models with static rebalancing have improved FID scores, but fail to generalize over precision. In Figure 4, the lighter regions have better PCK and FID scores indicating both high precision and high coverage of a given model. It would make the evaluation more robust, if we consider precision and coverage as a trade-off instead of two independent criteria. We observe that employing AISLe and G_{attn} helps our model (⬟) to enjoy the best of both worlds by striking a balance between precision and coverage.

7 Conclusions

In this paper, we studied the relationship between spoken language and free-form gestures. First, we introduced Adversarial Importance Sampled Learning, which combines adversarial learning with importance sampling to strike a balance between precision and coverage at no extra computational cost. Second, this work also introduced the use of transformers for gesture generation conditioned on spoken language. Third, we extended the PATS dataset in (Ahuja et al., 2020) by extracting transcripts for audio signals to study the effect of language in co-speech gesture generation. We substantiated the effectiveness of our approach through large-scale quantitative and user studies and show significant improvements over previous state-of-the-art approaches on both precision and coverage.

Acknowledgements

This material is based upon work partially supported by the National Science Foundation (Awards #1750439 #1722822), National Institutes of Health and the InMind project. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of National Science Foundation or National Institutes of Health, and no official endorsement should be inferred.

References

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. *Proceedings of the European Conference on Computer Vision*.

Chaitanya Ahuja, Shugao Ma, Louis-Philippe Morency, and Yaser Sheikh. 2019. To react or not to react: End-to-end visual pose forecasting for personalized avatar during dyadic conversations. In *2019 International Conference on Multimodal Interaction*, pages 74–84. ACM.

Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2014. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875*.

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. 2017. Generalization and equilibrium in generative adversarial nets (gans). In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 224–232. JMLR. org.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Jonathon Byrd and Zachary C Lipton. 2018. What is the effect of importance weighting in deep learning? *arXiv preprint arXiv:1812.03372*.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*.

Justine Cassell. 2001. Embodied conversational agents: representation and intelligence in user interfaces. *AI magazine*, 22(4):67.

Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2001. BEAT: the Behavior Expression Animation Toolkit. In *the 28th annual conference on Computer graphics and interactive techniques (SIGGRAPH '01)*, pages 477–486.

Chung-Cheng Chiu and Stacy Marsella. 2014. Gesture generation with low-dimensional embeddings. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 781–788.

Chung Cheng Chiu, Louis Philippe Morency, and Stacy Marsella. 2015. Predicting co-verbal gestures: A deep and temporal modeling approach. In *Proceedings of the 15th international conference on Intelligent virtual agents (IVA2015)*, volume 9238, pages 152–166.

Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al. 2018. State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4774–4778. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Maurice Diesendruck, Ethan R Elenberg, Rajat Sen, Guy W Cole, Sanjay Shakkottai, and Sinead A Williamson. 2019. Importance weighted generative networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 249–265. Springer.

Justin Domke and Daniel R Sheldon. 2018. Importance weighting and variational inference. In *Advances in neural information processing systems*, pages 4470–4479.

Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2019. Multi-objective adversarial gesture generation.

Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary C Lipton. 2020. A unified view of label shift estimation. *arXiv preprint arXiv:2003.07554*.

Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning individual styles of conversational gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3497–3506.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Aditya Grover, Jiaming Song, Ashish Kapoor, Kenneth Tran, Alekh Agarwal, Eric J Horvitz, and Stefano Ermon. 2019. Bias correction of learned generative models using likelihood-free importance weighting. In *Advances in Neural Information Processing Systems*, pages 11056–11068.

Guang-Yuan Hao, Hong-Xing Yu, and Wei-Shi Zheng. 2018. Mixgan: learning concepts from different domains for mixture generation. *arXiv preprint arXiv:1807.01659*.

Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. 2018. Evaluation of Speech-to-Gesture Generation Using Bi-Directional LSTM Network. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents (IVA18)*, pages 79–86.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637.

Quan Hoang, Tu Dinh Nguyen, Trung Le, and Dinh Phung. 2018. Mgan: Training generative adversarial nets with multiple generators.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G Hauptmann. 2015. Self-paced curriculum learning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. 2015. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342.

Angelos Katharopoulos and François Fleuret. 2018. Not all samples are created equal: Deep learning with importance sampling. *arXiv preprint arXiv:1803.00942*.

Ilya Kostrikov, Kumar Krishna Agrawal, Sergey Levine, and Jonathan Tompson. 2018. Addressing sample inefficiency and reward bias in inverse reinforcement learning. *arXiv preprint arXiv:1809.02925*.

Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. 2019. Analyzing input and output representations for speech-driven gesture generation. *arXiv preprint arXiv:1903.03369*.

Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexanderson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A framework for semantically-aware

speech-driven gesture generation. *arXiv preprint arXiv:2001.09326*.

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 *ICCV*, pages 763–772.

Jina Lee and Stacy Marsella. 2006. Nonverbal behavior generator for embodied conversational agents. In *Proceedings of the 6th international conference on Intelligent virtual agents (IVA2006)*, pages 243–255.

Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun. 2010. Gesture controllers. *ACM Trans. Graph.*, 29(4):124:1–124:11.

Sergey Levine, Christian Theobalt, and Vladlen Koltun. 2009. Real-time prosody-driven synthesis of body language. *ACM Trans. Graph.*, 28(5):172:1–172:10.

Margot Lhommet and Stacy Marsella. 2016. From embodied metaphors to metaphoric gestures. *CogSci*, pages 788–793.

Margot Lhommet, Yuyu Xu, and Stacy Marsella. 2015. Cerebella: Automatic Generation of Nonverbal Behavior for Virtual Humans. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 4303–4304.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. 2017. Mmd gan: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, pages 2203–2213.

Zachary C Lipton, Yu-Xiang Wang, and Alex Smola. 2018. Detecting and correcting for label shift with black box predictors. *arXiv preprint arXiv:1802.03916*.

Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. 2013. Virtual character performance from speech. In *Symposium on Computer Animation*, pages 25–35.

Tambet Matiisen, Avital Oliver, Taco Cohen, and John Schulman. 2019. Teacher-student curriculum learning. *IEEE transactions on neural networks and learning systems*.

Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Adithyavairavan Murali, Animesh Garg, Sanjay Krishnan, Florian T Pokorny, Pieter Abbeel, Trevor Darrell, and Ken Goldberg. 2016. Tsc-dl: Unsupervised trajectory segmentation of multi-modal surgical demonstrations with deep learning. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4150–4157. IEEE.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. 2019. Detecting out-of-distribution inputs to deep generative models using a test for typicality. *arXiv preprint arXiv:1906.02994*.

Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1):31–88.

Catherine Pelachaud. 2009. Studies on gesture expressivity for a virtual agent. *Speech Communication*, 51(7):630–639.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Najmeh Sadoughi and Carlos Busso. 2018. Novel realizations of speech-driven head movements with generative adversarial networks. pages 6169–6173.

Najmeh Sadoughi, Yang Liu, and Carlos Busso. 2015. Msp-avatar corpus: Motion capture recordings to study the role of discourse functions in the design of intelligent virtual agents. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 7, pages 1–6. IEEE.

Mehmet E. Sargin, Yucel Yemez, Engin Erzin, and Ahmet M. Tekalp. 2008. Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30:1330–1345.

Shashank Sharma and Vinay P Namboodiri. 2018. No modes left behind: Capturing the data distribution effectively using gans. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Hidetoshi Shimodaira. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244.

Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher. 2018. Audio to body dynamics. *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1145–1153.

Jackson Tolins, Kris Liu, Yingying Wang, Jean E Fox Tree, Marilyn Walker, and Michael Neff. 2016. A multimodal motion-captured corpus of matched and mismatched extravert-introvert conversational pairs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3469–3476.

Ilya O Tolstikhin, Sylvain Gelly, Olivier Bousquet, Carl-Johann Simon-Gabriel, and Bernhard Schölkopf. 2017. Adagan: Boosting generative models. In *Advances in Neural Information Processing Systems*, pages 5424–5433.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. *arXiv preprint arXiv:1906.00295*.

Ryan Turner, Jane Hung, Eric Frank, Yunus Saatci, and Jason Yosinski. 2018. Metropolis-hastings generative adversarial networks. *arXiv preprint arXiv:1811.11357*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Thomas Wolf, L Debut, V Sanh, J Chaumond, C Delangue, A Moi, P Cistac, T Rault, R Louf, M Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv, abs/1910.03771*.

Yuyu Xu, Catherine Pelachaud, and Stacy Marsella. 2014. Compound gesture generation: A model based on ideational units. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8637 LNAI:477–491.

Qiaojing Yan and Wei Wang. 2017. Dcgans for image super-resolution, denoising and debluring. *Advances in Neural Information Processing Systems*, pages 487–495.

Shiyu Yi, Donglin Zhan, Zhengyang Geng, Wenqing Zhang, and Chang Xu. 2019. Fis-gan: Gan with flow-based importance sampling. *arXiv preprint arXiv:1910.02519*.

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)*, pages 4303–4309. IEEE.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915.

Peilin Zhong, Yuchen Mo, Chang Xiao, Pengyu Chen, and Changxi Zheng. 2019. Rethinking generative mode coverage: A pointwise guaranteed approach. In *Advances in Neural Information Processing Systems*, pages 2086–2097.