Communication Beyond Words:
Grounding Visual Body Motion with Language

Chaitanya Ahuja

CMU-LTI-22-004

April 2022

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
www.lti.cs.cmu.edu

Thesis Committee:

Louis-Philippe Morency (Chair)	Carnegie Mellon University
Alan W. Black	Carnegie Mellon University
Yaser Sheikh	Carnegie Mellon University Facebook Reality Labs
Michael Black	MPII Tübingen

*Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in Languages and Information Technologies.*

Keywords: grounding, co-speech gestures, style, style transfer, low resource learning, generative modeling

Dedicated to my Nana,
who always motivated me to reach for the stars.

Abstract

Communication is essential in sharing knowledge and ideas. It encourages collaboration and teamwork, an important step towards inducing positive change in human societies. It is also a key building block for forging new relationships through self-expression as well as understanding others' emotions and thoughts. Communication is often categorized with verbal and nonverbal messages where nonverbal includes both vocal (e.g. prosody) and visual modalities (e.g. hand gestures, facial expressions). These three modalities have a fruitful and complex relationship with each other when communicating. Evolving technologies for online communication such as virtual reality have created a need for generating high-fidelity nonverbal communication along with verbal and vocal cues (e.g. communication in a virtual space). One key communicative cue is visual body motions which can express a wide range of messages across arms, hands, gait, physical skills (such as jumping, running, and so on) and interaction with the environment. Body motions also include gestures that accompany spoken language. These co-speech gestures allow speakers to articulate the intent and express emphasis.

The central theme of this thesis is to understand the two-way relationship (a.k.a. grounding) between human body motions and its associated *spoken language*, which includes both verbal and vocal cues. Understanding this complex relationship will help us to both better understand the meaning intended by body gestures and provide us with the knowledge necessary to generate more realistic nonverbal body animations with interactive technologies. With these motivations in mind, we propose three key challenges: (1) *Nonverbal Grounding* as the core component of this thesis to study the close relationship between spoken language and motion, (2) *Personalization* to better understand idiosyncrasies and commonalities on how people gesture, and (3) *Low Resource Learning* when gestures occur infrequently or the amount of labeled data is limited and often unbalanced. These challenges investigate the commonalities, uniqueness and generalizability of visual body communication respectively in the presence of verbal and vocal information.

This thesis makes some significant contributions for all three technical challenges starting with a contribution studying nonverbal grounding for a slightly easier problem: *descriptive language*, where the language is a direct description of the body motion. For the next contribution, we transition to the more challenging problem which is central to this thesis: *co-speech gestures*, where body motion naturally happens during spoken language (which is a combination of both vocal and verbal cues). Spoken language often has a long-tailed distribution which can prevent modeling for uncommon words and gestures. Hence, we study grounding of co-speech gestures in spoken language in presence of a long-tail distribution in the data. In the next contribution, we study the different styles of gesturing across many individuals in tandem with an adaptation of our ideas of grounding co-speech gestures for this context. Our next contribution revolves around the idea that humans can quickly understand gestures and body motion of a new person with only a few minutes of interaction with this person motivating the study of personalization of grounding and gesture style in a low resource setting. For our final contribution, we extend the ideas of low resource personalization to a more practical setting of continual learning. Here, the personalization still needs only a few minutes of training data, but in the process of learning new speakers, the model does not forget the old ones.

Acknowledgments

I have been really fortunate to have had LP Morency as my advisor without whom the thesis would not have been possible. The first time that I met LP, I was in awe of his ability to communicate research as a beautiful story. This coupled with the excitement of brainstorming new ideas and working toward impactful research has been a constant source of inspiration for me. LP, you have been an amazing mentor through all the highs and lows. When plans go south (and they almost always do), your patience and encouragement is unparalleled. And sometimes when things do go according to plan, your effort to celebrate these successes are also unparalleled. I hope to emulate some of these wonderful qualities in my research and mentorship which will continue to guide me well beyond my PhD. I started out as your student, but have been lucky to find an amazing mentor, brilliant colleague and lifelong friend.

I would like to thank my committee members - Alan W. Black, Yaser Sheikh, and Michael J. Black. I really appreciate their scientific curiosity and insightful comments and questions which gave me additional perspectives of my work and positively impacted the final thesis.

I have been really lucky to have crossed paths with many talented collaborators who have taught me a lot and made the projects a lot of fun - Dong Won Lee, Yukiko Nakano, Ryo Ishii, Paul Liang, Sanika Natu, Pratik Joshi, Shradha Sehgal, Sharath Rao and Michal Muszynski. I would also like to thank Stacey Young, John Friday and Nicki Silverling who have tirelessly worked behind the scenes for seamless processes at LTI. I have been fortunate to work with Shugao Ma, Jason Saragih and Yaser Sheikh during my internship which has added a lot of value to my thesis.

My time in Pittsburgh was made even more enjoyable by the group of friends I made: Ankush Das, Shrimai Prabhumoye, Bhuwan Dhingra, Rolly Mantri, Dheeraj Rajagopal, Vidhisha Balachandran, Shruti Palaskar, Venkat Perumal, Devendra Chaplot, Kanthashree Mysore, Vikesh Siddhu, Rama Kumar Pasumarthi, Sreecharan Sankaranarayanan, Kundan Krishna, Vasu Sharma, Sushil Lathwal, Priyank Lathwal, Dipan Pal, Alexandria K Vail, Alex Wilf, Dong Wong Lee, Amir Zadeh, Paul Liang, Hiroaki Hayashi, Salvador Medina, Torsten Wörtwein, Volkan Cirik, and Michael Miller Yoder; and the long distance friendships that kept me going: Jayesh K Gupta, Varun Harbola, Manish Mehta, Amit Munje, Khagesh Patel, Atri Bhattacharyya, and Ashudeep Singh.

I would also like to thank my family - my mother Alpana, my father Bharat, my brother Yajur, my cousins Aastha, Atreya, Kaya, Ayan, my aunts Anupama and Rachana, my uncles Guru and Ashish, and my grandparents Jagadish, Kamal and Nand Kishore for supporting me throughout this endeaver. And finally, I was incredibly lucky to have met Bhavya Balu, *mi preciosa*, during my time in Pittsburgh. Bhavya, thank you for your unconditional love and support that has filled my life with laughter, optimism and joy.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> Language is not merely a set of unrelated sounds, clauses, rules, and meanings; it is a total coherent system of these integrating with each other, and with behavior, context, universe of discourse, and observer perspective.
>
> *Kenneth Pike* [142]

Communication, which is essential to sociocultural evolution [98], frequents our day to day lives. It facilitates a conscious or unconscious decision to express emotions, discuss ideas, convey messages or share knowledge with others. Hence it encourages collaboration and forging new relationships through self-expression. Communication is often expressed in two different forms: (1) *Verbal* cues are expressed in form of spoken language that demonstrate meaning and intent. (2) *Nonverbal* cues appear as a variety of visual body motions such as hand gestures, locomotion and other actions like dancing, jumping, typing and so on. Nonverbal also includes vocal cues that express timing, tone, stress, intonation and rhythm in the form of acoustic signals. The system of coherent interplay between natural language (i.e. verbal), acoustics and visual body gestures (i.e nonverbal) can be quite complex. Technologies like virtual reality [37, 68, 72, 110] are constantly evolving for a more realistic online communication platform. This research has the potential to enable *communication beyond words* by being inclusive of high fidelity nonverbal communication along with vocal and verbal information in such virtual spaces. Furthermore, it gives us the ability to design AI systems to encourage effective communication by intelligent tutoring systems in classroom settings [176], empathy in clinical psychology [40, 124], tools

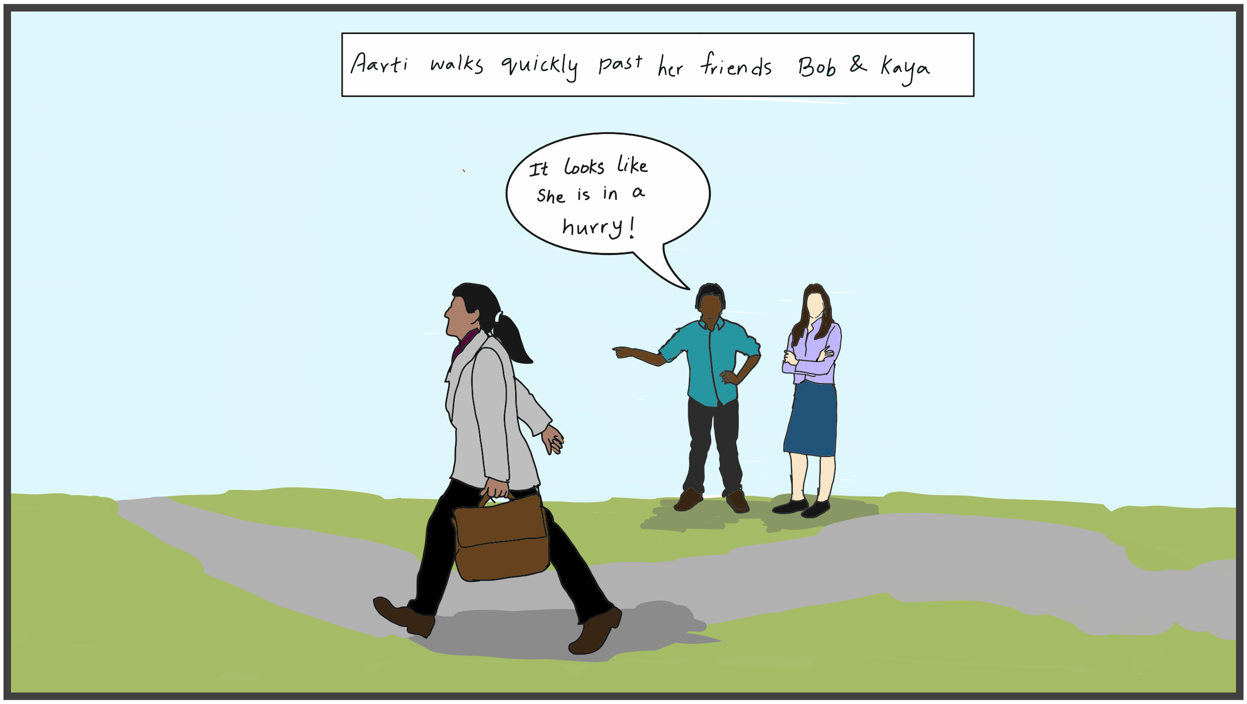

Aarti walks quickly past her friends Bob & Kaya
It looks like She is in a hurry!

Figure 1.1: A visual depiction of the links between visual body motion and language. "Aarti walks quickly past her friends Bob and Kaya" is a descriptive sentence illustrating Aarti's action of quickly walking. "It looks like she is in a hurry" is spoken by Bob and is accompanied by a pointing action i.e. co-speech gesture.

to aid animation generation [28], and enables realism in remote work collaborations via virtual communication with relevant and diverse gestures.

To build systems that can naturally interact with humans, we must understand the link between language and visual body communication. As an illustrative example, consider a scenario depicted in in Figure 1.1. A direct description of Aarti's motion (a.k.a. descriptive language) is the first sentence: *Aarti* walks quickly past her friends *Bob* and *Kaya*. This sentence directly translates to Aarti's swift locomotion. While *Aarti* did not say anything, her speed and the fact that she did not greet her friends indicates that she was in a hurry. This leads to a spoken utterance by *Bob* accompanied by a co-speech gesture: *Bob* points at *Aarti* while exclaiming to *Kaya*, "It looks like she is in a hurry!". This body motion in tandem with spoken language relays a message to *Kaya* that 'she' refers specifically to *Aarti*. This simple example illustrates interesting interactions between visual body motions and language. These gestures are often performed unconsciously and their links with spoken language are naturally understood by people. Although intuitive for humans, computers are still not able to properly understand and generate relevant gestures and body motions.

The central theme of this thesis is to understand the two-way relationship between visual body motions and spoken language, where spoken language includes both verbal and vocal cues. This relationship is often referred to as *grounding* between language and motion. To approach this complex problem, we propose to study the body motion and language relationships under two scenarios, both of which were illustrated in Figure 1.1. First, we study descriptive language as it offers a more direct link between language and motion. This is a stepping stone towards the study of the second scenario, *co-speech gestures*. These gestures bring a more complex relationship (or grounding) with language and, for this reason, a larger portions of this thesis will study co-speech gesture modeling.

Descriptive Language and Motion. Visual body motion (such as locomotion and physical skills) can be studied more directly with descriptive language. Consider a descriptive sentence such as, ”After receiving a phone call, she jumped with joy”. This act of jumping is physical motion which is directly referred in the language of this sentence. While some actions are directly stated, other aspects of the motions may be more indirect. In this example, the type of jumping should be an expression of joy (most likely from getting some good news over the phone). The intent and reason for the action may have to be inferred from context. So although, the relationship of body motion is more direct with descriptive language, it still brings some interesting challenges. Studying descriptive language has great applications in the animation and gaming industries. For example, descriptive sentences in a movie script can be quickly translated to canonical animations, which could be a valuable addition to an animator’s toolkit. In this thesis, we study descriptive language as a stepping stone for more complex relationships between co-speech gestures and spoken language.

Spoken Language and Co-speech Gestures. During spoken language, people naturally gesture with their body to add information about their intent and meaning to spoken utterances [1, 84, 119]. These predominantly upper body motions bring interesting challenges of studying the relationship between body motion and spoken language. One other interesting aspect of spoken language is that it is also accompanied with vocal cues which play a key role in modeling prosody, synchrony and emphasis. These vocal cues are often reflected in the co-speech gestures of the speaker [119]. We are interested in building intelligent systems which can understand the interactions between co-speech gestures, language and acoustics as a combined communicative process. We are particularly interested in better understanding the connections between gestures and language, often referred to as grounding. One approach when studying this problem is to formalize it as a cross-modal translation task of generating a sequence of co-speech gestures using as input the spoken language and vocal cues. From the perspective of animation generation, co-speech gestures have a complex relationship with language and acoustics. For example, there can be many correct gestures corresponding to the same language and acoustics. Additionally, a speaker can always decide not to gesture while speaking and in some cases speakers perform body motion without synchronous spoken language [1, 119].

One very interesting challenge with co-speech gestures is that they are often person specific. In other words, gestures are often idiosyncratic in nature. These idiosyncrasies, or more broadly *gesture style*, result in the production of different gestures when delivered by a different speaker, even if the expressed message was the same. Furthermore, these idiosyncrasies can act as identity markers of an individual [67]. For example, an extroverted person can be more grandiose with their gestures and may produce gestures more frequently as compared to an introverted person. This is an example where style could generalize over a group of individuals [86, 125]. Our goal is to integrate the notion of speaker style when modeling the relationships between spoken language and gestures. Hence, to drive realistic personalized avatars on virtual reality platform, a long term goal is garner the ability to learn a person's style promptly (and hopefully be able to do so with only a few examples). This is similar to a scenario when people quickly learn and understand gesture styles of a person they meet for the first time.

In the next section we describe the main technical challenges and core research questions to move forward our research. Section 1.2 will then describe the key contributions of this thesis followed by the overview of the structure of this thesis in Section 1.3.

1.1 Technical Challenges

In this section, we define and elucidate three key technical challenges which are the driving forces of this thesis. First, we study the central challenge of **nonverbal grounding** which allows us to better understand and model the relationships between visual body motion and language. Since every individual has a different way of gesturing, we need to take into consideration these idiosyncrasies and uniquenesses in gesturing. Hence, in our second challenge, we study **personalization** of gestures at the individual level. The complexity of learning these relationships and personalization can be magnified if the data distribution has long tails or there isn't much data to begin with. Hence, we explore the challenge of **low resource learning** in context of both nonverbal grounding and personalization as the final challenge.

Nonverbal Grounding

How can we better understand and model relationships between visual body motion and language? This challenge is especially stimulating because human body motions have a complex relationship with language and acoustic cues. For example, a hand ”wave” is directly grounded to a linguistic concept of greeting, but the adverb ”very” can be gestured in different ways to express emphasis on the adjective or verb that follows. Furthermore the same gesture might be linked to different linguistic concepts like in the case of pointing back which could be referring to an object in that direction or to an event that occurred in the past. In some cases the speaker might not gesture at all. These examples indicate the importance of the expressivity distribution of the speaker (i.e. how often and what kind of gestures does the speaker generate?) along with the relevance of gestures accompanying spoken utterances for nonverbal grounding. One of the goals of nonverbal grounding is generating plausible visual body motion that either represents the mental state of the speaker's spoken utterances or directly corresponds to the descriptive language. We first study generation of body motion from descriptive language as it represents a more direct mapping. We follow this up with a more complex task of co-speech gesture generation from spoken language. The complexity of both these cross-modal translation tasks increase with vocabulary of either modalities. Coarse-grained body motions (i.e. small vocabulary for visual body motions) can be modeled as structured prediction tasks [33, 74, 197] and coarse-grained language (i.e. small vocabulary for language) can be modeled to generate gestures via rule based approaches [26, 54, 95], but the rules may fail to generalize in new situations and a fixed vocabulary of body motions may not be representative of all speakers. To overcome these limitations, we relax the assumption of a fixed vocabulary for visual body motions by working in a continuous space of joint positions with an orders of magnitudes larger vocabulary for language information. A challenge of working with a continuous space is evaluating the generated body motions with objective metrics. The generated animations may be correct for the given input language, but could be different from ground truth animations. Hence, we need to go beyond accuracy metrics and develop both objective and human study based indicators for progress.

Concretely, we study the following research questions,

Q1.1 How are human body motion grounded on language and acoustic cues?

Q1.2 How can we generate animations of human body motions which are grounded in language and acoustics?

Personalization

Co-speech gestures are idiosyncratic. Each person has a different style of delivering a message with gestures. While there are idiosyncrasies in gestures, many commonalities also exist when studying gestures across multiple individuals [8, 196]. A key challenge is to be able to personalize co-speech gesture generation by leveraging these commonalities, but without losing out on the uniqueness of each individual's style. A parallel goal is to be able to control the personalization of the generated co-speech gestures. For example, in Figure 1.1, Bob has his own style of gesturing. If Kaya was speaking instead, the accompanying gestures would probably have been different. The goal of style transfer would be to personalize gestures that follow Kaya's style while using as input Bob's spoken utterances. To study these challenges it may be necessary to learn a common style space which decouples gesticulation style from the gestures accompanying spoken language. Learning idiosyncrasies and gesture style will rely on examples of body motion examples from multiple individuals. A benefit of learning jointly is that the model will be able to generate gestures from multiple different styles. Moreover, leveraging the commonalities amongst different styles can lead to computationally efficient models that are able to generate relevant co-speech gestures while personalizing to many different styles.

Concretely, we study the following research questions,

Q2.1 How can we generate gestures relevant to a speaker's utterances and in an another speaker's style?

Q2.2 How can we personalize gesture style for many speakers together?

Q2.3 How can we leverage the prior knowledge of old speakers to personalize the style of new speakers?

Low Resource Learning

Modeling the relationships between co-speech gestures and body motion requires significant amount of paired data with current methods. But datasets for co-speech gestures are scarce, much like real-world scenarios where an individual may only have a few minutes of interaction with a new person. These interactions are generally enough for the individual to understand the gestures of the new person. Motivated with this we would like to study modeling of visual body motion in presence of only a few examples. It has been observed that visual body motions lie on a low dimensional space despite the high dimensionality of the body joints' configuration [44, 64]. As the dimensionality of the visual body motion space decreases, it reduces the number of examples needed to reliably construct this latent space -i.e. low resource learning. Furthermore, we would also like to ground this denser latent space in spoken language, achieving these goals in parallel with the first challenge. To formalize this problem, we want to personalize both nonverbal grounding and gesture style of a co-speech gesture generation source model with only a few minutes of gesture and spoken language of multiple new speakers. The source model here is trained with a relatively larger set of gestures than the new speakers.

Another challenge may arise in this process of personalizing grounding and gesture style to new speakers. The gesture generation model for a new speaker may lose the ability to generate gestures in the style of the source speaker. This phenomenon is commonly known as catastrophic forgetting [46] and can be quite limiting for an intelligent agent [24, 103, 136, 137] that intends to continually adapt as it interacts with different speakers at different points in time. This continuous adaptation of an agent (or model) is often studied in a continual learning paradigm. Additionally, these continuously changing agents have only a few examples to learn the new speakers' behaviours from, hence increasing the complexity of this already challenging task. To replicate a practical scenario of an continuously adapting agent in the wild, we study the models of gesture generation in low resource and continual learning setting. Formally stated, we want to continuously personalize our gesture generation model to many different new speaker styles without forgetting the older speakers and constrained with only a few minutes of training data for each speaker.

Another related challenge is that the space of grounded gestures is large and can be underrepresented both on the gesture and language side. For example, a waving gesture might occur infrequently because it is generally performed at the start or end of a conversation. Additionally, language representation depends on the topic of the conversation. For example, a conversation on chemistry would often make references to elements like hydrogen and not about topics like movies. Furthermore, there are many words that appear only a few times which is indicative of long-tailed language distributions. Long-tailed distributions implicitly create a low-resource setting for the words and gestures that occur only a few times. For the task of cross-modal translation, uncommon words in the long-tail may produce generic gestures or no gestures at all. Also, uncommon gestures can be forgotten by such models resulting in reduced diversity of the generated gestures.

We are interested in designing low resource training paradigms for personalization and nonverbal grounding which can potentially leverage prior knowledge of other speakers' behaviours to generate diverse gestures which are relevant to the spoken language.

Concretely, we study the following research questions,

Q3.1 How can we learn a generative model from unbalanced data which produces both precise and diverse gestures?

Q3.2 How can we personalize the nonverbal grounding and gesture style of new speakers with few minutes of data?

Q3.3 How can we personalize the nonverbal grounding and gesture style of new speakers without forgetting the old speakers with few minutes of data?

1.2 Thesis Contributions

The central theme of this thesis is **grounding** visual body motions in language, also referred to as Nonverbal Grounding. As a first contribution, we ground visual body motions in descriptive language. This cross modal task of generating body motions given a descriptive sentence requires a direct translation of low-level concepts like action, speed and direction. This work is a stepping stone for our next contribution of grounding co-speech gestures in spoken language. The link between co-speech gestures and spoken language brings an extra set of challenges of grounding high-level concepts such meaning, intent, timing, intonation and rhythm. We study these in context of long-tail distributions in gesture and language space. For our next contribution, we study co-speech gesture grounding in context of **personalizing** to different individual styles of gesturing. Nonverbal grounding and personalization require large datasets of gestures along with the corresponding spoken language for which we propose a large scale gesture-language dataset to study the technical challenges as our fourth contribution. While a large scale dataset is crucial to study the two technical challenges, practical scenarios may demand learning gesture generation models from a few minutes of data. We study nonverbal grounding and personalization in a **low resource** setting in our next contribution. As our final contribution, we extend the ideas for low resource learning to nonverbal grounding and personalization in a continual learning setting which mimics an intelligent agent continuously adapting to multiple gesture styles by sequentially exposing it to new speakers. We also link each of our main contributions to the listed research questions in the previous section.

Descriptive Language: Grounding (Chapter 2)

Visual body motions **grounded** in descriptive language are a direct translation of low-level concepts like action, speed and direction. We present an approach, called *Joint Language-to-Pose (JL2P)*, which grounds visual body motions in descriptive and complete sentences. We learn a multimodal joint embedding space to learn the connections between body motions and language in line with the research question **Q1.1**. With these learnt representations, a goal is to generate animations of third-person human body motion conditioned on natural language sentences. For cross-modal translation tasks, such as this contribution, we show that projecting the linguistic cues and corresponding movement vectors to a joint embedding space is key in learning low-level concepts like action, speed, and direction from descriptive sentences as an answer to the question **Q1.2**. To learn the joint embedding space, we propose the use of curriculum learning which guides the model to learn grounding shorter body motion sequences first before moving on to the longer and harder ones. We show that this approach's strong empirical performance on a standard benchmark for both objective scores and human judgements. Part of this chapter appeared in the proceedings of International Conference on 3D Vision (3DV) 2019 [5].

Co-Speech Gestures: Grounding and Long Tail Distributions (Chapter 3)

For this contribution, we further study the research questions **Q1.1** & **Q1.2** in context of a more interesting task of **grounding** high-level concepts such meaning, intent, timing, intonation and rhythm from spoken language in co-speech gestures. As a part of this study, we observe that distributions of language and gestures are inherently skewed making it important to model **long tail** distributions along with **grounding**. Additionally, 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) as a solution to the research question **Q3.1**, 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. 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. Part of this chapter appeared in the proceedings of Findings at Empirical Methods on Natural Language Processing (EMNLP) 2020 [7].

Co-Speech Gestures: Grounding and Gesture Style (Chapter 4)

Each person has a different **style** of conveying a message with their co-speech gestures. A key challenge described in the research question **Q2.1**, called gesture style transfer, is to learn a model that generates gestures for a speaking agent 'A' in the gesturing style of a target speaker 'B'. A secondary goal is to simultaneously learn to generate co-speech gestures for multiple speakers while remembering what is unique about each speaker also known as style preservation or **grounding**. In this chapter, we propose a new model as a solution to the research question **Q2.2**, named Mix-StAGE, which trains a single model for multiple speakers while learning unique style embeddings for each speaker's gestures in an end-to-end manner. A novelty of Mix-StAGE is to learn a mixture of generative models which allows for conditioning on the unique gesture style of each speaker. As Mix-StAGE disentangles style and content of gestures, gesturing styles for the same input speech can be altered by simply switching the style embeddings. Mix-StAGE also allows for style preservation when learning simultaneously from multiple speakers. Our proposed Mix-StAGE model significantly outperforms the previous state-of-the-art approach for gesture generation and provides a path towards performing gesture style transfer across multiple speakers. Part of this chapter appeared in the proceedings of European Conference on Computer Vision (ECCV) 2020 [8].

Co-speech Gestures: Low Resources for Grounding and Gesture Style (Chapter 5)

Personalizing an avatar for co-speech gesture generation from spoken language requires learning the idiosyncrasies of a person's gesture style from a small amount of data. Previous methods in gesture generation require large amounts of data for each speaker, which is often infeasible motivating the research questions **Q2.3**& **Q3.2**. We propose an approach, named DiffGAN, that efficiently personalizes co-speech gesture generation models of a high-resource source speaker to target speaker with just 2 minutes of target training data. A unique characteristic of DiffGAN is its ability to account for the crossmodal grounding shift, while also addressing the distribution shift in the output domain. We substantiate the effectiveness of our approach a large scale publicly available dataset through quantitative, qualitative and user studies, which show that our proposed methodology significantly outperforms prior approaches for low-resource adaptation of gesture generation. Part of this chapter appeared in the proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) 2022 [9].

Co-Speech Gestures: Low Resource Continual Learning for Grounding and Gesture Style (Chapter 6)

Personalization of co-speech gestures in realistic scenarios can often only be afforded a small amount of data for the new speaker. But in the process of learning a new speaker’s gestures, the model catastrophically forgets the grounding and personalization of the original speakers’. This can be quite limiting for an intelligent agents that intend to continuously learn and adapt new nonverbal behaviours of new speakers at different points in time. The continual learning setting for a crossmodal translation task brings a unique technical challenge of *crossmodal catastrophic forgetting* in line with the research question **Q3.3**. To tackle this challenge, we propose an approach, named C-DiffGAN, that can efficiently personalize co-speech gesture generation models from a high-resource source speaker to multiple low-resource target speakers. To the best of our knowledge, this is the first approach that is able to learn a personalized model for multiple speakers with only 2 minutes each of speaker data (i.e. as opposed to 10 hours [7, 50, 60, 89]) in a continual learning setting. We substantiate the effectiveness of our approach a large scale publicly available dataset through quantitative and qualitative studies, which show that our proposed methodology significantly outperforms prior approaches for low resource continual learning of nonverbal grounding and personalization of gesture generation models.

Co-speech Gestures: PATS Dataset (Chapter 4)

We introduce Pose-Audio-Transcript-Style (PATS)dataset, to study the three technical challenges of this thesis. It contains a total of 250+ hours of pose, audio and text for 25 different speakers delivering monologues. These spoken language and co-speech gestures provide a unique and versatile test bench to study nonverbal grounding. This dataset can especially be used in context of a cross-modal task of co-speech gesture generation. A unique quality of PATS is the diversity among the 25 speakers (15 talk show hosts, 5 lecturers, 3 YouTubers, and 2 televangelists), which is necessary to study many different individual styles of gesture. Furthermore, we can augment the data by removing a large chunk of examples from the training set to emulate a few-shot learning paradigm. Additionally, the size and diversity of PATS gives us the ability to study the amount and quality of gestures needed to efficiently learn grounding and style.

1.2.1 Other Contributions

In this section we outline the contributions which are not directly associated with this thesis but were insightful for construction of the previously stated technical challenges.

Grounding Co-speech Gestures in a Dyadic Conversation

Creating personalized avatars for dyadic interactions not only requires to model intrapersonal dynamics between a avatar's speech and their body pose, but it also needs to model interpersonal dynamics with the interlocutor present in the conversation. In this work, we introduce a neural architecture named Dyadic Residual-Attention Model (DRAM), which integrates intrapersonal (monadic) and interpersonal (dyadic) dynamics using selective attention to generate sequences of body pose conditioned on audio and body pose of the interlocutor and audio of the human operating the avatar. We evaluate our proposed model on dyadic conversational data consisting of pose and audio of both participants, confirming the importance of adaptive attention between monadic and dyadic dynamics when predicting avatar pose. We also conduct a user study to analyze judgments of human observers. Our results confirm that the generated body pose is more natural, models intrapersonal dynamics and interpersonal dynamics better than non-adaptive monadic/dyadic models. This work appeared in the proceedings of International Conference on Multimodal Interaction (ICMI) 2019 [6].

Impact of Personality on Nonverbal Behavior Generation

Personality of an individual can modify the production of nonverbal behaviors, hence it is important to generate such behaviors that match the expected personality of a virtual agent. In this work, we study how personality traits relate to the process of generating individual nonverbal behaviors of the whole body, including the head, eye gaze, arms, and posture. With collaborators, we first created a dialogue corpus including transcripts, a broad range of labeled nonverbal behaviors, and Big Five scores which are self-reported personality scores of participants in dyadic interactions. My contribution in this work was to design models that can predict each nonverbal behavior label grounded in spoken language and while matching the personality of the individual. Our experimental results show that personality can help improve the prediction of nonverbal behaviors. This work appeared in the proceedings of Intelligent Virtual Agents (IVA) 2020 [74].

Lattice Recurrent Units for Language Modeling

Recurrent neural networks have shown remarkable success in modeling sequences. However low resource situations still adversely affect the generalizability of these models. We introduce a new family of models, called Lattice Recurrent Units (LRU), to address the challenge of learning deep multi-layer recurrent models with limited resources. LRU models achieve this goal by creating distinct (but coupled) flow of information inside the units: a first flow along time dimension and a second flow along depth dimension. It also offers a symmetry in how information can flow horizontally and vertically. We analyze the effects of decoupling three different components of our LRU model: Reset Gate, Update Gate and Projected State. We evaluate this family of new LRU models on computational convergence rates and statistical efficiency. Our experiments are performed on four publicly-available datasets, comparing with Grid-LSTM and Recurrent Highway networks. Our results show that LRU has better empirical computational convergence rates and statistical efficiency values, along with learning more accurate language models. This work appeared in the proceedings of Association for the Advancement of Artificial Intelligence (AAAI) 2018 [4].

Multimodal Machine Learning: A taxonomy and survey

Our experience of the world is multimodal - we see objects, hear sounds, feel texture, smell odors, and taste flavors. Modality refers to the way in which something happens or is experienced and a research problem is characterized as multimodal when it includes multiple such modalities. In order for Artificial Intelligence to make progress in understanding the world around us, it needs to be able to interpret such multimodal signals together. Multimodal machine learning aims to build models that can process and relate information from multiple modalities. It is a vibrant multi-disciplinary field of increasing importance and with extraordinary potential. Instead of focusing on specific multimodal applications, this work surveys the recent advances in multimodal machine learning itself and presents them in a common taxonomy. We go beyond the typical early and late fusion categorization and identify broader challenges that are faced by multimodal machine learning, namely: representation, translation, alignment, fusion, and co-learning. This new taxonomy will enable researchers to better understand the state of the field and identify directions for future research. This work appeared in the proceedings of Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 2017 [15] and The Handbook of Multimodal-Multisensor Interfaces 2018 [16].

1.3 Thesis Outline

We begin with the study of nonverbal **grounding** with a focus on descriptive language in Chapter 2. We follow this up with the study of a more complex **grounding** of co-speech gestures in spoken language in Chapter 3. This work also takes into account a likely scenario of long-tail distribution in the language and gesture information. Next, we focus our attention on the **personalization** of the delivery the same message via co-speech gestures in Chapter 4. We study this in tandem with the challenge of **grounding** co-speech gestures of many different styles along with the ability to perform gesture style transfer. We follow this up by a study of nonverbal grounding and personalization in a **low resource** setting in Chapter 5. We extend these ideas in Chapter 6 to a more practical setting of continual learning where we personalize nonverbal grounding and gesture style with low resources of multiple new speakers without forgetting the old speakers. To evaluate our contributions, we also propose a dataset PATS in Chapter 3 which consists of 250+ hours of co-speech gestures, and aligned spoken language in 25 different styles. Finally, in Chapter 7 we conclude with the contributions of this thesis, its limitations, broader impact and where we see this field progressing in the future.

Chapter 2

Descriptive Language: Grounding

A key challenge in human communication lies in nonverbal grounding. How are visual body motions related to verbal cues? In this chapter, we first study the relationships between descriptive language, which is a subset of verbal cues, and corresponding visual body motions which are a combination of concepts like action (i.e. walking, running and so on), speed and direction. Next, we explore these relationships in form of a cross-model translation task where descriptive sentences as in input generate a relevant animation.

Code for reproducing the experiments in this chapter is available on GitHub[1] and some generated videos are available here[2].

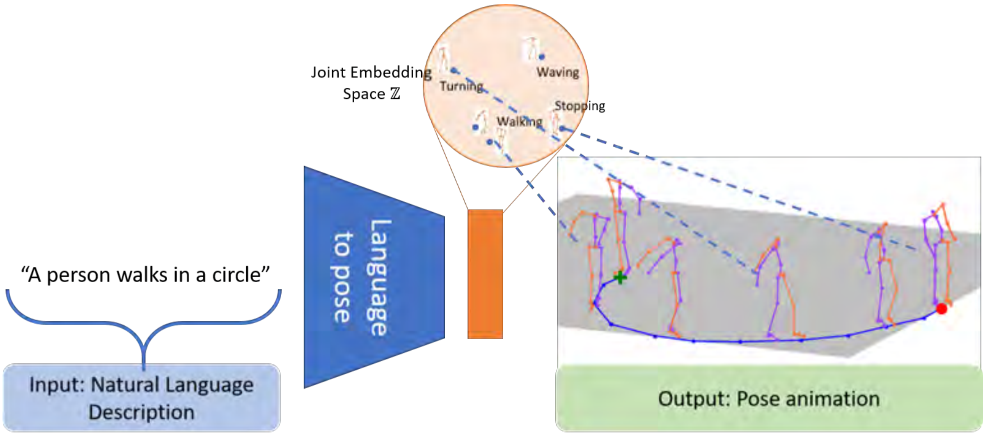
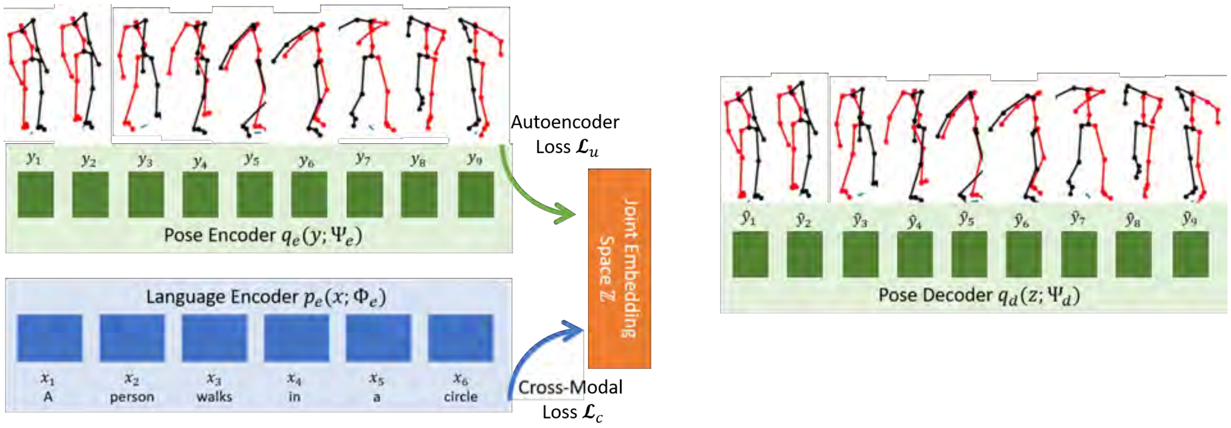


Figure 2.1: Overview of our model which uses joint multimodal space of language and pose to generate an animation conditioned on the input sentence.

[1] `https://github.com/chahuja/language2pose`
[2] `http://chahuja.com/language2pose`



y_1 y_2 y_3 y_4 y_5 y_6 y_7 y_8 y_9
Pose Encoder $q_e(y;\Psi_e)$
Autoencoder Loss $\mathcal{L}_u$
Joint Embedding Space $\mathbb{Z}$
Language Encoder $p_e(x;\Phi_e)$
x_1 A
x_2 person
x_3 walks
x_4 in
x_5 a
x_6 circle
Cross-Modal Loss $\mathcal{L}_c$
$\hat{y}_1$ $\hat{y}_2$ $\hat{y}_3$ $\hat{y}_4$ $\hat{y}_5$ $\hat{y}_6$ $\hat{y}_7$ $\hat{y}_8$ $\hat{y}_9$
Pose Decoder $q_d(z;\Psi_d)$

Figure 2.2: Overview of our proposed model Joint Language-to-Pose (or JL2P). Language and pose are mapped to a joint embedding space $\mathcal{Z}$, which can now be used by a trained pose decoder q_d to generate a pose sequence. At train time both p_e and q_e are used to create the joint embedding using a training curriculum. But at inference time $z \in \mathcal{Z}$ is encoded by p_e and decoded by q_d, giving us a model which can generate a animation (or sequence of poses) from a free form description (or language).

2.1 Overview

Generating animations from natural language descriptions is a first step for movie script visualization [57, 113] which can later be stitched together while maintaining co-references in the story-line [200]. These language grounded animations can also be useful in cases like virtual human animation [30, 33, 167], robot motion and task planning [3, 71].

An animation consists of a sequence of poses, which can be represented by positions of different joints in the body such as *Root* (base of spine), *head*, *shoulder*, *wrist*, *knee* and so on. Pose forecasting conditioned on natural language has 3 major challenges. First, pose and natural language are very different modalities. The model needs a joint space where both natural language sentences and poses can be mapped. The model should also be able to decode animations from this embedding space. Second, different words of a sentence represent different qualities about the animation. Verbs and adverbs describe the action and speed/acceleration of the action; nouns and adjectives describe locations and directions respectively. The model has to map these concepts to small pose sequences and then stitch them to render convincing animations. Third, we want to see if objective metrics correlate with subjective metrics for this task as our models are trained using objective distance metrics, but the quality of generated animations can only be judged by humans.

In this chapter, our two main contributions tackle the modeling challenges of pose and natural language. First, we propose a model Joint Language-to-Pose (or JL2P) that learns a joint embedding space of these two modalities. Second, we use a training curriculum to help the model emphasize more on shorter and and easier sequences first and longer and harder sequences later. Additionally, to make the training regimen robust to the outliers in the dataset, we use Smoooth L1 as the distant metric in our loss function. Through multiple objective and subjective experiments, we show that our model can generate more accurate and natural animations from natural language sentences than other data driven models.

2.2 Problem Statement

As an example, consider a natural language sentence which describes a human's motion: *"A person walks in a circle"*. The goal of this cross-modal language-to-pose translation task is to generate an animation representing the sentence; i.e. an animation that shows a person following a trajectory of a circle with a walking motion (see figure 2.1).

Formally, given a sentence, represented by an N-sized sequence of words $X_{1:N} = [x_1, x_2, \dots x_N]$, we want to predict a T-sized sequence of 3D poses $Y_{1:T} = [y_1, y_2, \dots y_T]$ that are coherent with the semantics in the sentence. $x_i \in \mathcal{R}^K$ is the i^{th} word vector with dimension K. $y_t \in \mathcal{R}^{J \times 3}$ is the pose matrix at time t. Rows of y_t represent joints of the skeleton and columns are the xyz-coordinates of each joint. Tensors X and Y are elements of sets $\mathcal{X}$ and $\mathcal{Y}$ respectively. Modeling language-to-pose is done by training a model $f : \mathcal{R}^{K \times N} \longrightarrow \mathcal{R}^{J \times 3 \times T}$ to predict a pose sequence $\hat{Y}_{1:T}$

$$\hat{Y}_{1:T} = f\left(X_{1:N}; \Theta\right) \tag{2.1}$$

where Θ are trainable parameters of the model f.

2.3 Joint Language-to-Pose

Language-to-pose models should be able to grasp nuanced concepts like speed, direction of motion and the kind of actions from the language and translate them to pose sequences (or animations). This requires the model to learn a multimodal joint space of language and pose. In doing so, it should also be able to generate sequences that are deemed correlated to the sentence by humans. To achieve that objective, we propose Joint Language-to-Pose (or JL2P) model to learn the joint embedding space. Given an input sentence, an animation can be sampled from this model at inference stage.

In this section, a joint embedding space of language and pose is formalized. This is followed by an algorithm to train for the joint embedding space and a discussion on the practical edge cases at inference time for our Joint Language-to-Pose model.

2.3.1 Joint Embedding Space for Language and Pose

To learn a joint embedding space of language and pose, the sentence $X_{1:N}$ and pose $Y_{1:T}$ are first mapped to a latent representation using a sentence encoder $p_e(X_{1:N}; \Phi_e)$ and a pose encoder $q_e(Y_{1:T}; \Psi_e)$ respectively. These estimate the latent representation or embeddings z_x and z_y respectively in the embedding space $\mathcal{Z} \subset \mathcal{R}^h$,

$$z_x = p_e\left(X_{1:N}; \Phi_e\right) \tag{2.2}$$

$$z_y = q_e\left(Y_{1:T}; \Psi_e\right) \tag{2.3}$$

z_x, z_y should lie close to each other in $\mathcal{Z}$ as they represent the same concept. To ensure that they do lie close together, a joint translation loss is constructed (refer to Figure 2.2) and trained end to end with a training curriculum.

2.3.2 Joint Loss Function

Once we have the embedding z_x or z_y, a pose decoder $q_d(; \Psi_d)$ is used to generate an animation from the joint embedding space $\mathcal{Z}$. The output of the pose decoder must now lie close to the pose sequence $Y_{1:T}$. Hence, using $X_{1:N}$ as inputs and $Y_{1:T}$ as outputs, the cross-modal translation loss is defined as,

$$\mathcal{L}_c = d\left(q_d\left(z_x; \Psi_d\right), Y_{1:T}\right) \tag{2.4}$$

and using $Y_{1:T}$ as inputs and $Y_{1:T}$ as outputs, the uni-modal translation (or autoencoder) loss is defined as,

$$\mathcal{L}_u = d\left(q_d\left(z_y; \Psi_d\right), Y_{1:T}\right) \tag{2.5}$$

where $d(x, y)$ is a function to calculate the distance between the predicted values and ground truth of pose. Φ_e, Ψ_e and Ψ_d are trainable parameters of the sentence encoder, pose encoder and pose decoder respectively.

Combining equations 2.4 and 2.5 we get a joint translation loss,

$$\mathcal{L}_j = \mathcal{L}_c + \mathcal{L}_u \tag{2.6}$$

Jointly optimizing the loss $\mathcal{L}_j$ pushes z_x and z_y closer together improving generalizability and additionally trains the pose decoder which is useful for inference from the joint embedding space.

As $\mathcal{L}_j$ is a mutivariate function in $X_{1:N}$ and $Y_{1:T}$, coordinate descent [186] for optimizing the loss function is a natural choice and is described in Algorithm 1.

2.3.3 Training Curriculum

Cross modal pose forecasting can be a challenging task to train [29]. Starting with simpler examples before moving on to tougher ones can be beneficial to the training process [19, 193, 198]. The curriculum design commonly used for pose forecasting [29] is adapted for our joint model. We first optimize the model to predict 2 time steps conditioned on the complete sentence. This easy task helps the model learn very short pose sequences like leg motions for walking, hand motions for waving and torso motions for bending. Once the loss on the validation set starts increasing, we move on to the next stage in the curriculum. The model is now given twice the amount of poses for prediction. The complexity of the task is increased in every stage till the maximum time-steps (T) of prediction is reached. We describe the complete training process in Algorithm 1.

2.3.4 Optimization

For the distance metric $d(x, y)$ in Equation 2.4, 2.5 and 2.6, Smooth L1 loss (similar to Huber Loss [70]) is used which is defined as,

$$SmoothL1(x, y) = \begin{cases} 0.5(x-y)^2 & \text{for } |x-y| < 1 \\ |x-y| - 0.5 & \text{otherwise} \end{cases} \tag{2.7}$$

In contrast, Lin et. al.[104] uses L2 loss for $d(x, y)$. L2 loss is more sensitive to outliers than L1 loss due to its linearly proportional gradient with respect to the error, while L1 loss has a constant gradient of 1 or -1. But L1 Loss can become unstable when $|x - y| \approx 0$, due to oscillating gradients between 1 and -1. On the other hand, Smooth L1 is continuous and smooth near 0 and more generally for all $x, y \in \mathcal{R}$, hence it is more stable than L1 as a loss function.

Algorithm 1: Learning language-pose joint embedding

```
initialization;
$\mathcal{X}_{train}, \mathcal{X}_{val}, \mathcal{Y}_{train}, \mathcal{Y}_{val} \leftarrow$ SplitData($\mathcal{X}, \mathcal{Y}$);
MaxValLoss $\leftarrow$ inf;
$t \leftarrow 2$;
while $t \leq T$ do
    for $X_{1:N}, Y_{1:t} \in \mathcal{X}_{train}, \mathcal{Y}_{train}$ do
        $r \leftarrow$ CoinFlip() ;                                  // For Coordinate Descent
        if $r == 0$ then
            $z \leftarrow p_e(X_{1:N}; \Phi_e)$ ;                    // Encoder
        else
            $z \leftarrow q_e(Y_{1:t}; \Psi_e)$ ;                    // Encoder
        end
        $\hat{Y_{1:t}} \leftarrow q_d(z; \Psi_d)$ ;                  // Decoder
        loss $\leftarrow d(Y_{1:t}, \hat{Y}_{1:t})$ ;
        $\Phi_e, \Psi_e, \Psi_d \leftarrow$ UpdateModelParams(loss) ;
    end
    ValLoss $\leftarrow CalcValLoss(\mathcal{X}_{val}, \mathcal{Y}_{val})$;
    if *ValLoss < MaxValLoss ;*                                      // Checkpoint best models
     then
        ($p_e$,$q_e$,$q_d$).checkpoint();
        MaxValLoss $\leftarrow$ ValLoss;
    else if *not IsModelImproving(ValLoss, MaxValLoss) ;*            // Move to a harder task
     then
        $t \leftarrow 2t$;
        MaxValLoss $\leftarrow$ inf;
    end
end
```

2.4 Experiments

Joint language to pose modeling can be broken down into three core challenges,

1. **Prediction Accuracy by Joint Space**: How accurate is pose prediction from the joint embedding ?
2. **Human Judgment**: Which of the generated animation is more representative of the input sentence? Does the subjective evaluation correlate with the results from the objective evaluations?
3. **Modeling nuanced language concepts**: Is the model able to map nuanced concepts such as speed, direction and action in the generated animations?

Experiments are designed to evaluate these challenges of language grounded pose forecasting.

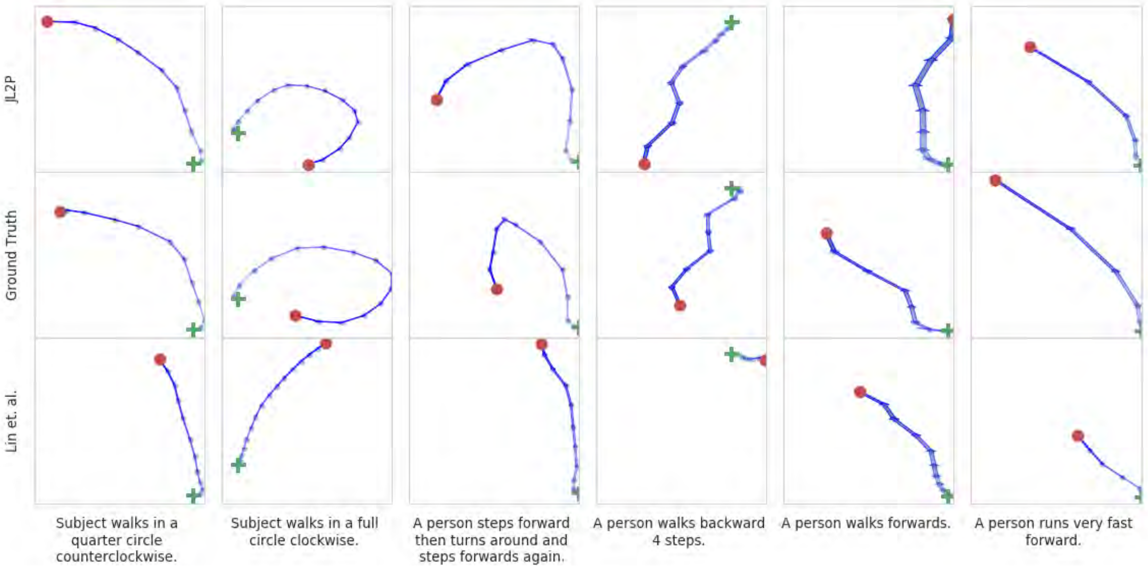


Figure 2.3: Trajectory plots of the generated pose (i.e. *Root*'s position) viewed from the top. Each box represents a generated trajectory of the model on the vertical axis and sentence on the horizontal axis. The person starts at the green cross (x) and ends at the red circle (•) with blue dots (•) denoting equally placed time-steps. All trajectories in each column have the same scale for fair comparison across models.

In the following subsections, the dataset and its pre-processing is briefly discussed which is followed by the evaluation metrics for both objective and subjective evaluations. Finally, design choices of the encoder and decoder models are described which are used to construct the baselines in the final subsection.

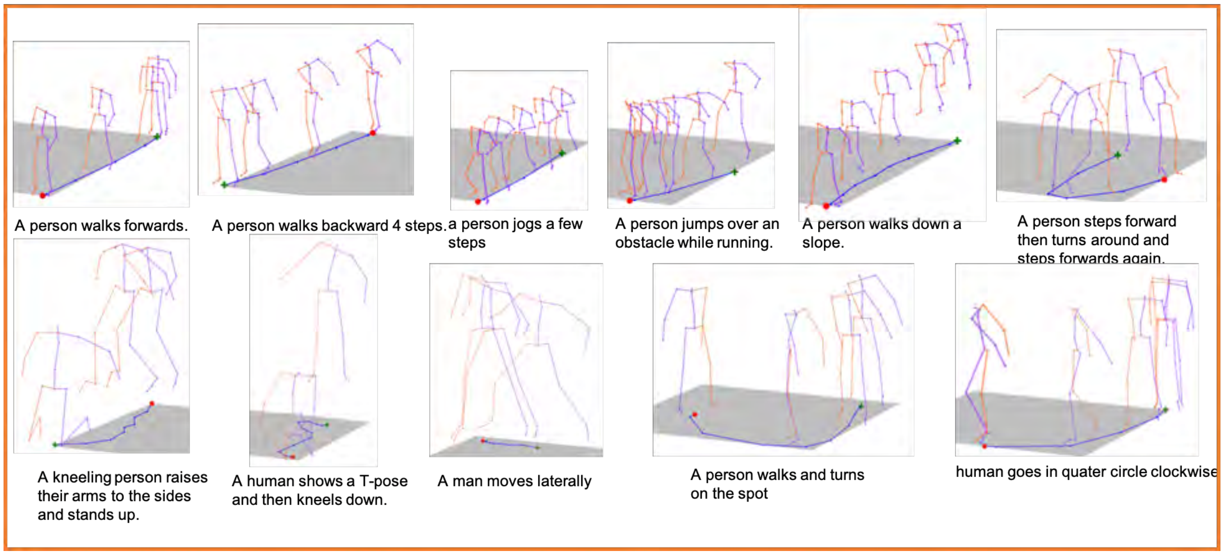
2.4.1 Dataset

Our models are trained an evaluated on KIT Motion-Language Dataset [143] which combines human motion with natural language descriptions. It consists of 3911 recordings (approximately 11.23 hours) which are re-targeted to a kinematic model of a human skeleton with 50 DoFs (6 DoF for the *Root* joint's orientation and position, while remaining 44 DoFs for arms legs, head and torso). The dataset also consists of 6278 English sentences (approximately 8 words per sentence) describing the recordings. This is more than the number of recordings as each recording has one or more descriptions which are annotated by human volunteers. We use 20% of the data as a randomly sampled held-out set for evaluating all models. There is wide variety of motions in this dataset ranging from locomotion (e.g. walking, running, jogging), performing (e.g. playing violin/guitar), and gesticulation (e.g. waving). Many recording have adjectives to further describe the motion like speed (e.g. fast and slow), direction (left, right and forward), and number for periodic motions (e.g. walk 4 steps).

We use the pre-processing steps used in Holden et. al. [65]. All the frames of the motion are transformed such that body always faces the Z-axis. Joint rotation angles are transformed to 3D positions is the skeleton's local frame of reference with *Root* as the origin. *Root*'s position on XZ-plane and orientation along Y-axis is represented as velocity instead of absolute values. Motion sequences are then sub-sampled to a frequency of 12.5 Hz down from 100Hz. This is low enough to bring enough variance between 2 time-steps for the decoder to train for a regression task, while not compromising on the human's perception of the animation [29].

2.4.2 Implementation Details

For pose encoder (q_e) a network of Gated Recurrent Units (GRUs) [36] is used in our model JL2P. The pose decoder (q_d) is the same except it has residual connection from the input to the output layer. This is similar to the pose decoder in Lin et. al. [104], except an extra layer to predict the trajectory (or Trajectory Predictor) is discarded in our model. For langauge encoder (p_e), a network of Long-Short Term Memory Units (LSTMs) [63] is used. Each token of the sentence is converted into a distributional embedding using a pre-trained Word2Vec model [121].



A person walks forwards.

A person walks backward 4 steps.

a person jogs a few steps

A person jumps over an obstacle while running.

A person walks down a slope.

A person steps forward then turns around and steps forwards again.

A kneeling person raises their arms to the sides and stands up.

A human shows a T-pose and then kneels down.

A man moves laterally

A person walks and turns on the spot

human goes in quater circle clockwise

Figure 2.4: Renders of generated animations with a diverse set of sentences as input by our proposed model. Our model is able to change speed, direction and actions based on changes in the input sentence. Trajectory of the character is drawn with a blue line which starts at the green cross (x) and ends at the red circle (●).

2.4.3 Baselines

There has been limited work done in the domain of data-driven cross-modal translation from natural language descriptions to pose sequence generation. The closest work to our proposed approach is by **Lin et. al.** [104][3]. As mentioned in Section 2.6, their model does not follow a training curriculum and uses L2 loss as the loss function. Their model also maps the language embeddings to an existing embedding space of poses instead of jointly learning it.

We also compare our model **JL2P** (see Section 2.3) with three ablations derived from itself. These ablations study the 3 main components of the model, joint embedding space, curriculum learning and Smooth L1 loss:

- **JL2P w/o Curriculum** - Training curriculum in Section 2.3.3 is dropped.
- **JL2P w/o L1** - L2 loss is used instead of Smooth L1 loss as the distance metric $d(x, y)$.
- **JL2P w/o Joint Emb.** - Instead of joint training as described in Section 2.3.2, autoencoder loss $\mathcal{L}_u$ minimized first followed by optimization of the cross-translation loss $\mathcal{L}_c$.

[3] At the time of publication we could not find publicly available code or pre-trained models for this work, hence we use our own implementation and training on the same data as all other baselines

2.4.4 Objective Evaluation Metrics

All models are evaluated on the held-out set with a metric Average Position Error (APE). Given a particular joint j, it can be denoted as APE(j),

$$\text{APE}(p) = \frac{1}{\mathcal{Y}} \sum_{\mathcal{Y}} \|\hat{y}_t[j] - y_t[j]\|_2 \tag{2.8}$$

where $y_t[j]$ is the true location and $\hat{y}_t[j] \in \mathcal{Y}$ is the predicted location of joint j at time t

Another metric, Probability of Correct Keypoints (PCK) [10, 160], is also used as an evaluation metric. If a predicted keypoint (or joint) lies inside a sphere (of radius σ) around the ground truth, the prediction is deemed correct. Given a particular joint j, PCK$_\sigma(j)$ is defined as follows,

$$\text{PCK}_\sigma(j) = \frac{1}{\mathcal{Y}} \sum_{\mathcal{Y}} \delta\left(\|\hat{y}_t[j] - y_t[j]\|_2 \leq \sigma\right) \tag{2.9}$$

where δ is the indicator function.

2.4.5 User Study: Subjective Evaluation Metric

Joint language to pose generation is a subjective task, hence a human's subjective judgment on the quality of prediction is an important metric for this task.

To achieve this, we design a user study which asked human annotators to rank two videos generated by 2 different models but with same sentence as the input. One of the videos is generated by Lin et. al. and the other is either ground truth or generated by JL2P , JL2P w/o Curriculum, JL2P w/o Joint Emb., or JL2P L1. The annotators answer the following question for each pair of videos, *Which of the 2 generated animations is better described by "<sentence>"?*. To ensure that annotators spend enough time to decide, any annotations which took less than 20 seconds were rejected. This study subjectively evaluates the preference of humans for generated animations by different models.

2.5 Results and Discussion

In this section we first use objective measures and then conduct a user study to get a subjective evaluation. Finally, we probe some qualitative examples to understand the effectiveness of the model in tackling the core challenges described in Section 2.4.

Models	**Average Positional Error (APE) in mm**										
	Mean	**Mean w/o Root**	**Root**	**Torso**	**Head**	**LArm**	**RArm**	**LHip**	**RHip**	**LFoot**	**RFoot**
Lin et. al.[104]	54.9^{***}	50.0	151.6	26.6	35.4	61.3	61.6	32.2	32.1	63.3	63.2
JL2P w/o Curriculum	52.2^{***}	47.9	139.2	24.2	32.5	57.3	57.2	30.6	30.7	62.9	63.2
JL2P w/o L1	51.7^{**}	47.0	145.0	24.4	32.8	58.0	57.6	29.9	30.7	59.3	59.8
JL2P w/o Joint Emb.	50.4	45.7	143.3	24.0	**31.0**	55.6	**54.5**	29.7	29.5	**59.0**	59.5
JL2P	**49.5**	**45.4**	**131.1**	**23.0**	31.4	**55.3**	55.0	**28.6**	**29.0**	59.2	**58.8**

Table 2.1: Average positional error (APE) for JL2P , JL2P w/o Joint Emb., JL2P w/o L1, JL2P w/o Curriculum and Lin et. al.. Lower is better. Our models (JL2P and variants) show consistent increase in accuracy over Lin et. al. across all joints with the addition of components joint embedding, smooth L1 loss and curriculum learning. Two-tailed pairwise t-test between all models and JL2P where $* * *$- $p<0.001$, and $**$- $p<0.01$.

2.5.1 Prediction Accuracy by Joint Space

JL2P demonstrates at least a 9% improvement over Lin et. al. (see Table 2.1) for all joints. The maximum improvement around 15% is seen in the *Root* joint. Errors in *Root* prediction can lead to a ”sliding-effect” of the feet; when the generation is translating faster than the frequency of the feet. Improvements in APE scores for long-term prediction, especially for *Root*, can help get rid of these artifacts in the generated animation.

When compared to its variants, JL2P loses maximum APE value when it is trained without curriculum (or JL2P w/o Curriculum). As discussed in Section 2.3.3, learning to predict shorter sequences before moving on to longer ones proves beneficial for pose generation. APE scores go down by 4%, if L2 loss is used instead of Smooth L1. In an output space as diverse as pose sequences, it becomes important to ignore outliers which may drive model to overfit. APE scores go down only by 1%, if the embedding space is not trained with the joint loss $\mathcal{L}_j$

APE values across time for JL2P of different parts of the body (*Root, Legs, Arms, Torso* and *Head*) are plotted in Figure 2.6. *Root*'s APE scores have the fastest rate of increase, followed by *Arms, Legs* and then *Head, Torso*. Two out of three coordinates of *Root* are represented as velocity which accumulates errors when integrated back to absolute positions; this is probably a contributing factor to the rapid increase of prediction error over time.

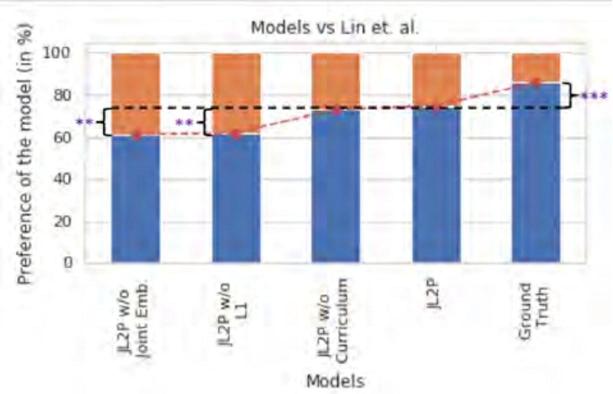


Figure 2.5: Preference scores of baseline models vs Lin et. al.. Blue bars denote the preference percentage of models marked on the horizontal axis. Our models (JL2P and variants) show consistent rise in preference over Lin et. al. with the addition of components joint embedding, smooth L1 loss and curriculum learning. **- $p < 0.01$ and ***- $p < 0.001$ for McNemar's test [120] on paired nominal data.

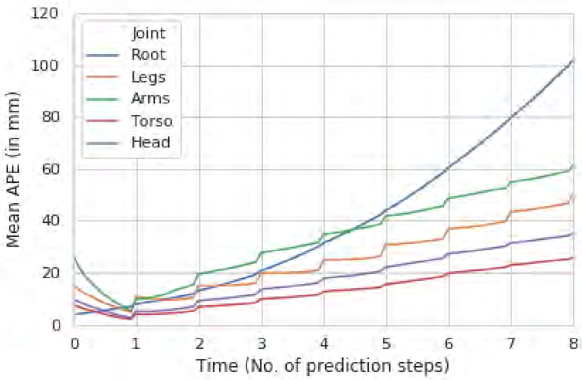


Figure 2.6: Plot of mean APE values across time for different parts of the body (*Root, Legs, Arms, Torso* and *head*) for JL2P. Lower is better. Generating trajectory of the animating character is harder than the other joints as *Root*'s APE blows up after around 500ms

Our final objective metric is PCK. PCK values (for 35$\leq \sigma \leq$55) on generated animations are compared among JL2P , its variants and Lin et. al. in Figure 2.7. JL2P and its ablations show a consistent improvement over Lin et. al. which further strengthen the claim about the prediction accuracy by our model's joint space.

2.5.2 Human Judgment

Human judgment is quantified by preference scores in Figure 2.5. Human preference of all our baseline models and ground truth are compared against Lin et. al. animations. JL2P has a

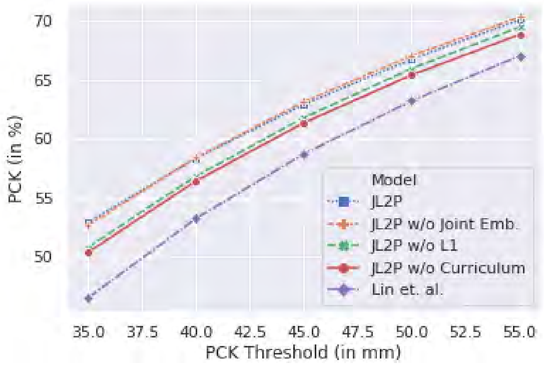


Figure 2.7: Plots of average Probability of Correct Keypoint (PCK) values over multiple values of thresholds (σ) for JL2P , JL2P w/o Joint Emb., JL2P w/o L1, JL2P w/o Curriculum and Lin et. al.. Our model JL2P shows consistent improvements over other baselines across a large range of thresholds. Higher values are better.

preference of 75% which is shy of ground truth by 10%. Preference scores consistently drop with all the other variants of JL2P .

JL2P w/o Joint Emb. has the lowest preference score of 60% when ranked against Lin et. al. It is still more preferred than Lin et. al. but far more unlikely to be picked when pitted against JL2P . This is an interesting change in trend, as removing joint loss from JL2P did not affect the objective scores significantly, but have lowered its human preference by a significant fraction. This leads us to conclude that objective metrics are not enough to judge the performance of a model. Instead a combination of human judgment and objective metrics is necessary for evaluating pose generation models.

2.5.3 Modeling nuanced language concepts

Root joint decides the trajectory of the animation which is crucial for translating concepts like speed (e.g. fast, and slow), direction (e.g. left, right, forward and backward) from natural language to animation. We plot the trajectories generated by JL2P, ground truth and Lin et. al. for different sentences in Figure 2.3.

Modeling direction: Animations' trajectory for these sentences for JL2P is similar to that of the ground truth trajectories. In contrast, Lin et. al.'s trajectories tend to be semantically incorrect and have a slightly curved forward motion for these sentences.

Modeling speed: In the sentence, ”A person runs very fast forward”, JL2P is able to understand that the animation has to move faster. It is able to walk approximately the same distance as the ground truth in the same amount of time, hence has the same speed. In contrast, even though Lin et. al.'s motion is in the forward direction, it is not able to maintain the same speed as required by the sentence.

Modeling actions: In figure 2.4, we plot animations generated by a diverse set of sentences. JL2P is able to understand the action from the sentences, and is able to generate an animation corresponding to the action. JL2P is able to handle many actions ranging from *kneeling* (with complex leg motions) to *jogging* (with periodic hand and leg motion).

We show, via qualitative examples, that our model JL2P is able to model nuanced language concepts which are then reproduced in the animations generated at inference time.

2.6 Related Work

Pose Forecasting: Data driven human pose forecasting attempts to understand the behaviours of the subject from its history of poses and generates the next sequence of poses. Short-term predictions [139] focus on modeling joint angles corresponding to hands, legs, head and torso. Long-term predictions [49, 139, 168] additionally model the positions of the human character to generate animations like walking, running, jumping and crawling. While some works use different actions (such as running, kicking, and more) as conditioning variables to generate the future pose [105, 168], others rely solely on the history of poses to predict what kind of motion will follow [35]. Pose forecasting for locomotion is a more commonly researched topic, where models decide where and when to run/walk based on low-level control parameters such as trajectory and terrain [66]. Task based locomotion (such as writing on a whiteboard, moving a box, and sitting on a box) add the nuances of transitioning from one task to another, but pose generation is based on task-specific footstep plans that act as motion templates [2]. All these approaches are either action specific, or require a set of low-level control parameters to forecast future pose. In this work, we aim replace low-level control parameters with high-level control parameters (e.g. natural language) to control actions and their speed and direction for the generated pose.

Image or Speech conditioned pose forecasting: Images with a human can act as a context to forecast what comes next. Chao et. al. [29] use one image frame to predict the next few poses. These generated poses can now be used to aid the generation of a video [192] or a sequence of images [112]. An image, a high-level control parameter, has action information for pose generation, but it does not provide a fine-grained control on the speed and acceleration of the motion trajectory. Speech can also be used to control animations of virtual characters. Taylor et. al.[169]

use a data driven approach to model facial animation, while upper body pose forecasting conditioned on speech inputs has been tackled by Takeuchi et. al.[167]. But, these pose sequences model the non-verbal behaviours (such as head nods, pose switches, hand waving and so on) of the character and do not offer fine-grained control over the characters next movements.

Language conditioned pose forecasting: Natural language sentences consists of verbs describing the actions, adverbs describing the speed/acceleration of the action, and nouns with adjectives to describe the direction or target. This information can help provide a more fine-grained control over pose generations compared to image or speech. Statistical models [165, 166] which use bigram models for natural language have been trained to encode motion sequences from sentences. Ahn et. al. [3] use around 2100 hours of youtube videos with annotated text descriptions to train a pose generation model. Pose sequences extracted from videos have limited translation and occluded lower bodies, hence their model only predicts the upper body with a static *Root* joint. Some works use 3D motion capture data instead [144, 190]. Human motions generally have translation of the *Root* joint, hence forecasting trajectory is important to get natural looking animations. Lin et. al [104] generates pose of all the joints of the body by pretraining a pose2pose autoencoder model before mapping language embeddings on the learned pose space. But the embedding space is not learned jointly [138] which may limit the generative powers of the pose decoder. In contrast, our proposed approach learns a joint embedding space of language and pose using a curriculum learning training regime.

Chapter 3

Co-Speech Gestures: Grounding and Long Tail Distributions

In this chapter, we further study grounding in context of more high-level concepts such as meaning, intent, timing, intonation and rhythm in spoken language and their effects on the accompanying gestures, or co-speech gestures. Furthermore, we also tackle the likely scenario of long-tail distributions for both spoken language and co-speech gestures.

Code for reproducing the experiments in this chapter is available on GitHub[1] and the PATS dataset can be downloaded here[2]

3.1 Overview

Spoken language has gained more traction in the past decade due to improvements in natural language understanding and speech recognition. 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

[1] https://github.com/chahuja/aisle
[2] http://chahuja.com/pats

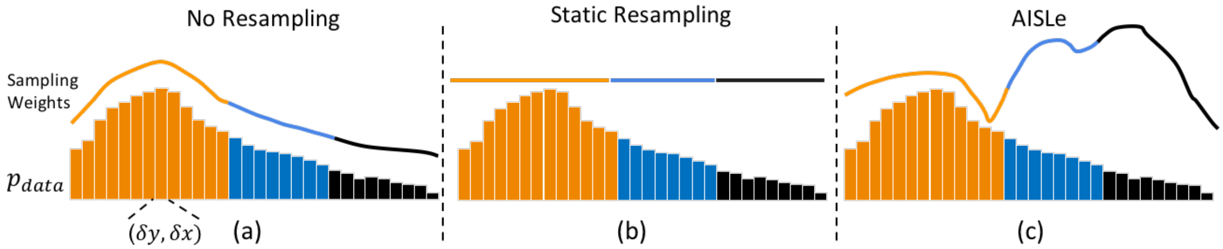


Figure 3.1: A toy representation of data distribution p_{data} as a histogram. Colours 🟧, 🟦, ⬛ represent bins from the mode, heavy tail and long tail of p_{data} respectively. The color coded envelope covering p_{data} is the distribution of weights across bins $(\delta y, \delta x)$ for the following 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.

the long tail of the language-gesture distribution (see Figure 3.1). 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 [50]. 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 [88, 140]. 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 chapter, 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 [172, 175] for gesture generation conditioned on spoken language. This idea allows transformer blocks to help with subword alignment between language and acoustic signals. As a third contribution, we curated Pose-Audio-Transcript-Style (PATS) dataset, designed to study gesture generation and style transfer. It consists of 25 speakers (15 new speakers and 10 speakers from Ginosar [50]) for a total of 250+ hours of gestures and aligned audio signals. 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.

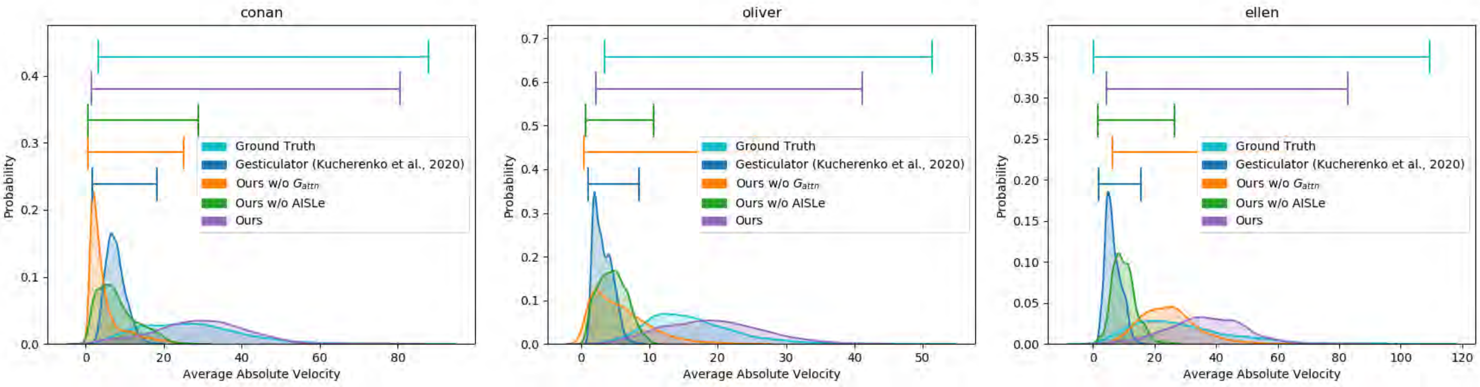


Figure 3.2: Distribution of the generated gestures with average absolute velocity as the statistic for three different speakers. The support (or coverage) of the distribution is denoted with the colour coded lines at the top of each plot. Larger overlap of a model's distribution with the ground truth distribution is desirable.

3.2 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 3.3). 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 $\mathbf{X}^w = \left[x^w_0, x^w_1, \dots x^w_{K-1}\right]$ 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, $\mathbf{X}^a = \left[x^a_0, x^a_1, \dots x^a_{T-1}\right]$. We want to predict a sequence of T gesture poses $\mathbf{Y}^p = \left[y^p_0, y^p_1, \dots y^p_{T-1}\right]$ that co-occur with $\mathbf{X}^a$ and $\mathbf{X}^w$. Here $y^p_t \in \mathcal{R}^{J\times 2}$ 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:

$$\hat{\mathbf{Y}}^p = G_\theta(\mathbf{X}^a, \mathbf{X}^w) \tag{3.1}$$

$$= G_{dec}\left(G_{attn}\left(G^a_{enc}(\mathbf{X}^a), G^w_{enc}(\mathbf{X}^w)\right)\right)) \tag{3.2}$$

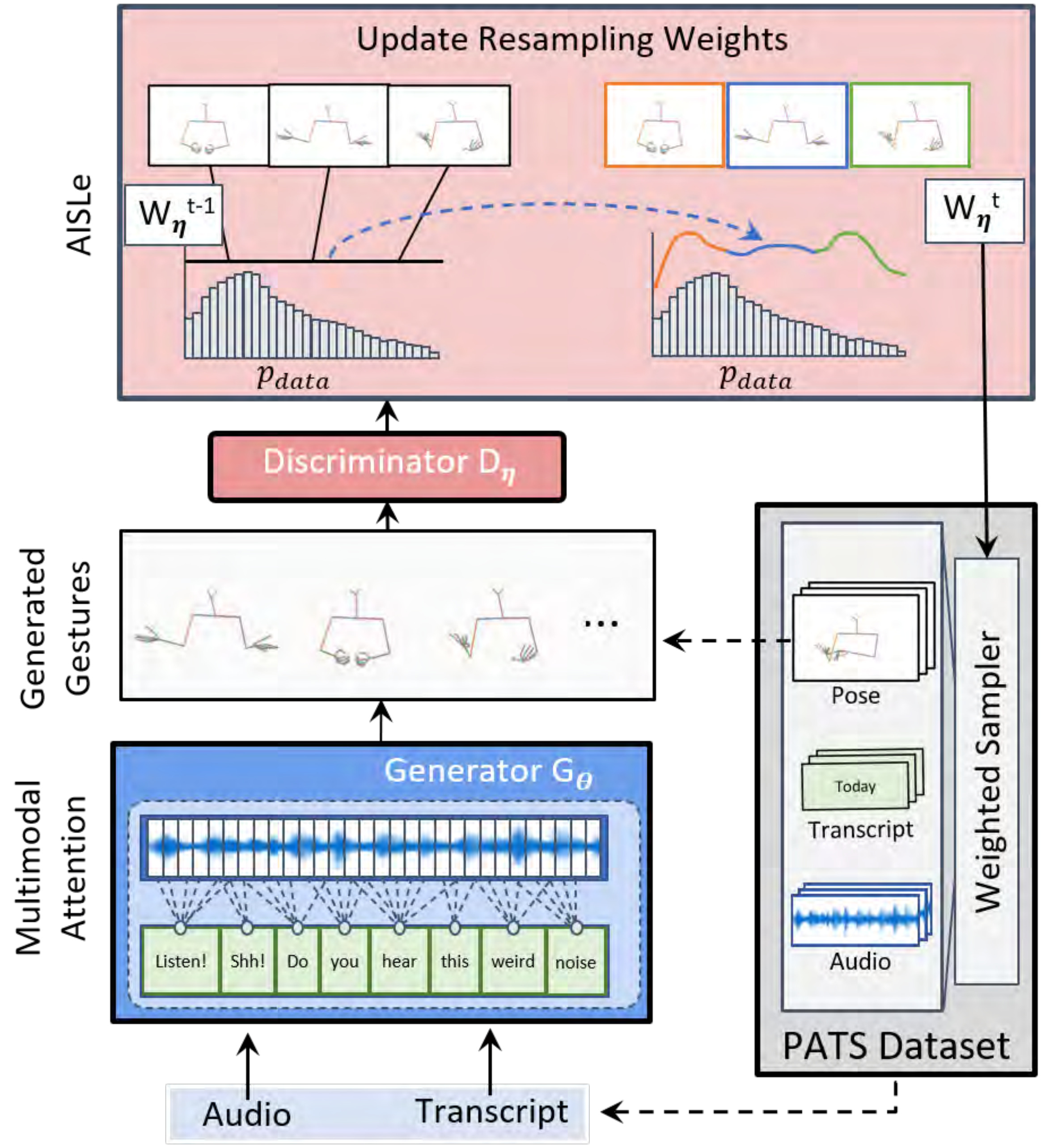


Figure 3.3: Overview of the key components of our model. Starting at the dataset and going clockwise, audio and transcripts go through sub-word alignment in the generator G_θ and are decoded to generate a freeform gesture animation. Next, the AISLe updates the weighted sampler of the dataset based on the output of the discriminator D_η to complete the loop.

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^a_{enc} and G^w_{enc} 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 [50] 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 [52] is written as:

$$\begin{aligned}\mathcal{L}_{adv} \;&=\; \mathbb{E}_{\mathbf{Y}^p} \log D_\eta\left(\mathbf{Y}^p\right) \\ &\quad+\mathbb{E}_{\mathbf{X}^a,\mathbf{X}^w} \log\left(1 - D_\eta(G_\theta\left(\mathbf{X}^a, \mathbf{X}^w\right)\right)\end{aligned} \tag{3.3}$$

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

$$\max_{\eta} \min_{\theta} \mathcal{L}_{rec} + \mathcal{L}_{mix} + \mathcal{L}_{adv} \tag{3.4}$$

where $\mathcal{L}_{mix}$ is a loss for training mixture of generators and defined in Section 3.3.3.

3.3 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 in Section 3.3.1. 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 3.3.2. 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 3.3.3. The key contributions are illustrated in Figure 3.3 and can be summarized by optimizing the overall loss function $\mathcal{L}(y, x)$ with AISLe in Algorithm 2.

3.3.1 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 3.1). 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,

$$\mathop{\mathbb{E}}_{x \sim p(.)} \mathop{\mathbb{E}}_{\substack{y \sim p_{\text{data}}(.|x) \\ \tilde{y} \sim p_\theta(.|x)}} \frac{p_\theta(\tilde{y}|x)}{p_{\text{data}}(\tilde{y}|x)} \mathcal{L}(y, x) \tag{3.5}$$

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 [42, 82], we derive the weighted cost function in Equation 3.5 using first principles. As illustrated in Figure 3.1, we divide the support of p_{data} into a grid of multi-dimensional bins of size $(\delta y, \delta x) \in \mathbb{R}^{\text{dim}(y)+\text{dim}(x)}$ 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 3.1(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 3.1(b). This encourages that equal number of samples are drawn from each bin while training,

$$\mathop{\mathbb{E}}_{x \sim p(.)} \mathop{\mathbb{E}}_{\substack{y \sim p_{\text{data}}(.|x) \\ \tilde{y} \sim p_\theta(.|x)}} \frac{1}{p_{\text{data}}(\tilde{y}|x)} \mathcal{L}(y, x) \tag{3.6}$$

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 3.6 gives us Equation 3.5. This appears as adaptive weighting across the support of the data distribution as shown in Figure 3.1(c).

Estimation of Importance Weights: We follow a likelihood-free approach [53, 173] to estimate w_η by computing the outputs of the discriminator D_η. Rewriting w_η in Equation 3.5 as,

$$w_\eta(x) = \frac{1 - D_\eta(G_\theta(x))}{D_\eta(G_\theta(x))} \tag{3.7}$$

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 [42], which is an equivalent alternative to optimize weighted loss functions. We illustrate the weight update cycle in Algorithm 2.

Algorithm 2: Adversarial Importance Sampled Learning

```
initialization;
$w_\eta(\tilde{y}) \leftarrow 1, \forall \tilde{y}$;
datasetSampler.updateWeights($w_\eta$);
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
    $w_\eta \leftarrow \frac{w_\eta - \text{mean}(w_\eta)}{\text{std}(w_\eta)} + 1$ ;          // keep weights around 1
    $w_\eta \leftarrow \text{clip}(w_\eta, 0.1, 10)$ ;          // clip weights to lie in (0.1, 10)
    datasetSampler.updateWeights($w_\eta$);
end
```

3.3.2 Multimodal Multiscale Attention Block

To address the challenge of sub-word alignment, we take inspiration from recent work self-attention [175] and cross-attention models [172] 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 [25]. A multimodal attention mechanism can make use of sub-word information from the audio to drive well-timed

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	**35.7 ± 1.7**	33.7 ± 1.7	32.1 ± 1.7
Ours	**38.9 ± 1.7**	**36.7 ± 1.6**	**37.1 ± 1.7**	**35.3 ± 1.7**

Table 3.1: 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.

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.

$$\mathbf{Z}^{aw} = \text{CM}(\mathbf{Z}^a, \mathbf{Z}^w) = \text{softmax}\left(\frac{\mathbf{Q}^a\mathbf{K}^{wT}}{\sqrt{h^a}}\right)\mathbf{V}^w \tag{3.8}$$

Unlike [172], we precede cross-modal attention with a layer of self attention [175] 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 [13] followed by a pointwise feedforward layer along with residual connections as described in [41, 172, 175]. 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}.

3.3.3 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 [150] following suit in [50] 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 [8, 12, 58, 62],

$$\hat{\mathbf{Y}}_p = \sum_{m=1}^{M} \phi_m G_m(\mathbf{Z}) \tag{3.9}$$

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 [41] using an existing implementation with pre-trained weights [184]. 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 [6, 50]. In our experiments, we use an audio encoder based on Temporal Convolution Networks consisting of a convolution layer, followed by batch normalization [73], and ReLU [128]. We use a similar TCN network for the discriminator $\mathbf{D}_\eta$.

3.4 Experiments

3.4.1 Baseline Models

Speech2Gesture [50]: Speech2Gesture does not use the text modality (i.e. no multimodal attention block) and any form of re-sampling while training.

Gesticulator [89]: Unlike MMS-Transformer , Gesticulator has a set of fully connected layer followed by autoregressive fully connected layers which are FiLM conditioned [141]. In addition to audio and text, features of duration of each word (i.e. start, end, percentage completed and so

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 3.2: 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

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 3.6), 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.

3.4.2 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 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)** [10, 160]: the values are averaged over $\alpha = 0.1, 0.2$ as suggested in [50] 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

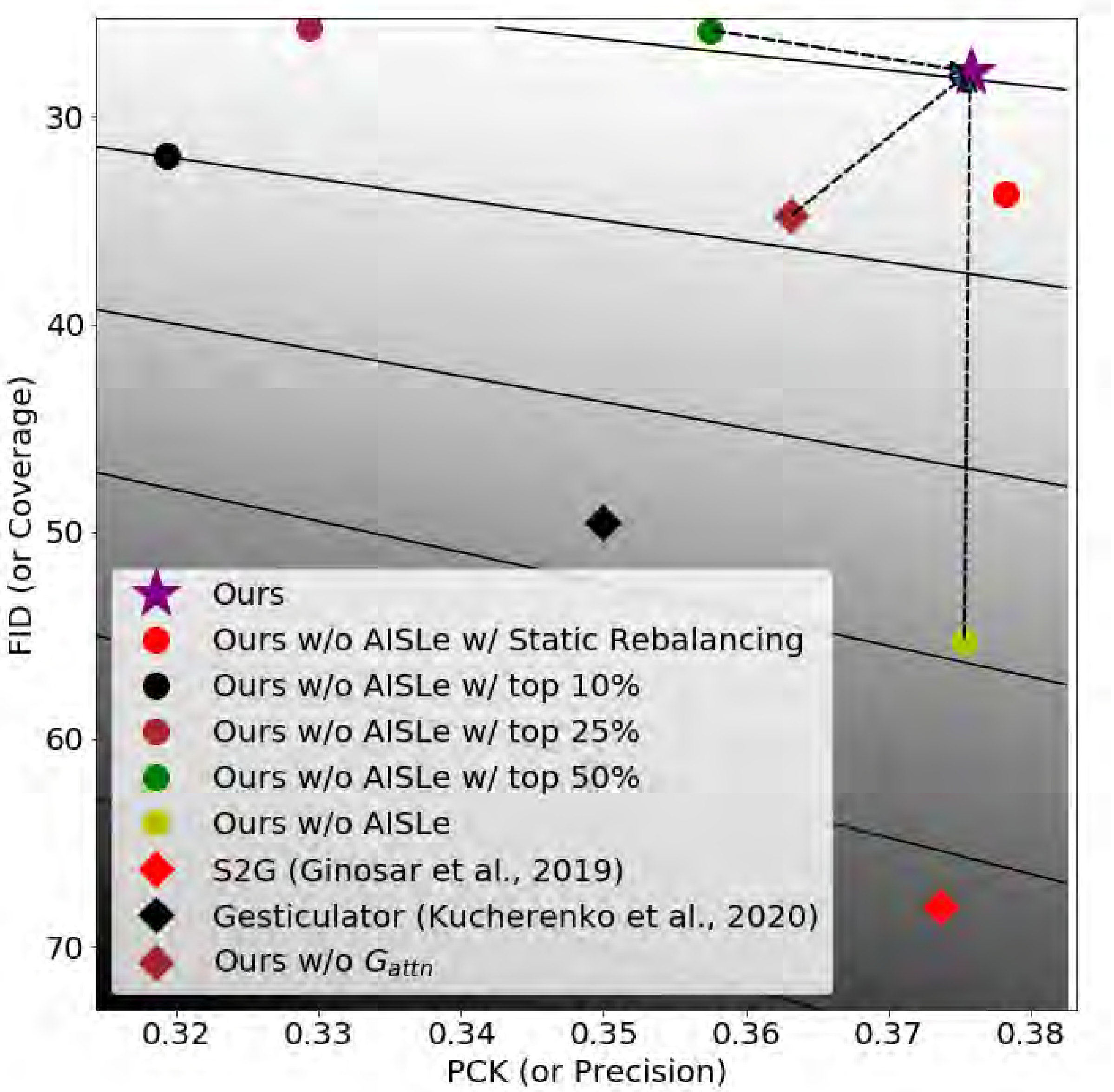


Figure 3.4: Precision Coverage Tradeoff for all models. Lighter areas represent high PCK and low FID which is favourable for the model. Contour lines corresponds constant values of $\frac{PCK}{FID}$. We show impacts of AISLe, G_{attn} and dataset subsampling with dotted lines traversing the PCK-FID plane, with our model enjoying the best of both worlds.

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[61]. (2) **Wasserstein-1 distances (or W1)**: distance between distribution of generated and ground truth average velocity. The same distance is calculated for average acceleration.

3.4.3 Pose-Audio-Transcripts-Style (PATS) dataset

We create a new dataset, Pose-Audio-Transcripts-Style (PATS) dataset with 250+ hours of co-speech gestures, audio and automatically extracted transcripts 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 [8, 50]. Specifically, it contains 15 talk show hosts, 5 lecturers, 3 YouTubers, and 2 televangelists with a mean of 10.7 seconds and a standard deviation of 13.5 seconds per video clip.

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 3.3: 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

Dataset Features

Aligned Transcriptions: As manual transcriptions are often not aligned and not readily available, we use Google Automatic Speech Recognition [34] 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 [130].

Pose: Each speaker's pose is represented via skeletal keypoints collected via OpenPose [23] following the approach in Ginosar et al. [50]. It consists of of 52 coordinates of an individual'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 [50, 88], we represent audio features as mel-spectrograms, which is a rich input representation shown to be useful for gesture generation.

3.5 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 sub-sampling. 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 3.1, since it is arguably the most important metric. We see a significantly[3] 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 3.2 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 3.3 🟩. We observe that the use of Static Rebalancing (Equation 3.6) 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 3.1, 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 [140]), which is one of the key indicators of

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

naturalistic gestures. In Figure B.9, 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 3.2 🟧. Relevance of generated gestures to the corresponding spoken language also suffers a significant decrease without G_{attn} in Table 3.1. These support our hypothesis that a representation which explicitly learns sub-word 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 3.3 🟧, that pruning the dataset to select samples which have a high 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 resampling 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 3.4, 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.

While effective, a limitation of this work is that the learnt grounded gesture space is speaker specific. Hence, we would need to train a new model for every new speaker. In the next chapter, we overcome this limitation by explicitly modeling style of every speaker.

3.6 Related Work

Language and Speech for Gesture Generation An early study by Cassell et al. [27] 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 [94, 99, 100, 116, 189]. Rule based approaches were replaced by deep conditional neural fields [31, 32] and Hidden Markov Models for prosody-driven head motion generation [156] and body motion generation [96, 97]. 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 [59], head motions [153] and body motions [6, 45, 50, 159] conditioned on a speech input while Yoon et al. [195] uses only a text input. Unlike these approaches, Kucherenko et al. [89] rely on both speech and language for gesture generation. But their choice of early fusion to combine the modalities ignores multi-scale correlations [172] between speech and language.

While publicly datasets of co-speech gestures are available, they are either small [154, 170, 195] or do not contain language information [50, 78, 91], 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 [52, 191]. Especially two aspects of distribution estimation, (1) conditional generation *precision* [50, 75, 122, 199] and (2) *coverage* of the entire underlying distribution [11, 157, 171, 201] 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 [129]. These approaches have similarities to importance weighting [22, 82], which are often used for post-hoc debiasing of the learnt model [43, 53, 173], correcting covariate shift [158], label shift [47, 106], imitation learning [87, 126] and curriculum learning [19, 76, 117]. Byrd and Lipton [22] observe that sub-sampling from unbalanced categorical classes demonstrates a significant effect on the network's predictions. Importance sampling in GANs [42, 101, 194], 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.

Chapter 4

Co-Speech Gestures: Grounding and Gesture Style

Every person has a different way of delivering the same message via co-speech gestures. These idiosyncrasies, also known as gesture style, are markers of individual identity and a key component of co-speech gestures. In this chapter, we formalize an approach to model gesture style in parallel with grounding with spoken language. We also develop an approach that generates gestures for a speaking agent 'A' in the gesturing style of a target speaker 'B', -i.e. style transfer.

Code for reproducing the experiments in this chapter is available on GitHub[1] and some generated videos are available here[2].

4.1 Overview

Nonverbal behaviours such as body posture, hand gestures and head nods play a crucial role in human communication [134, 178]. Pointing at different objects, moving hands up-down in emphasis, and describing the outline of a shape are some of the many gestures that co-occur with the verbal and vocal content of communication. These are known as co-speech gestures [83, 118]. When creating new robots or embodied virtual assistants designed to communicate with humans, it is important to generate *naturalistic* looking gestures that are meaningful with the speech [14]. Some recent works have proposed speaker-specific gesture generation models [28, 32, 45, 50] that are both trained and tested on the same speaker. The intuition behind this prior work is

[1] https://github.com/chahuja/mix-stage

[2] http://chahuja.com/mix-stage

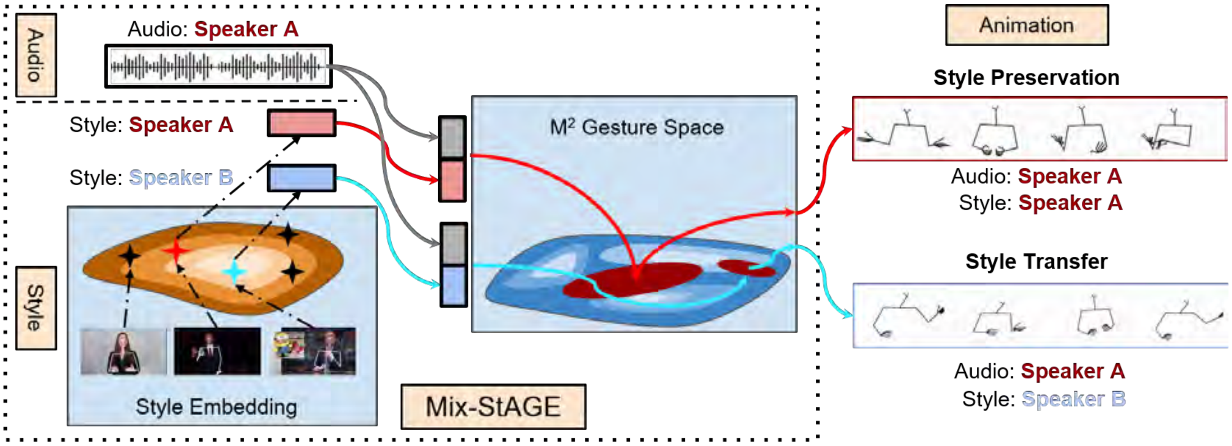


Figure 4.1: Overview of co-speech gesture generation and gesture style transfer/preservation task. The models learns a style embedding for each speaker, which can be be mapped to a gesture space with either the same speaker's audio to generate style preserved gestures or a different speaker's audio to generate style transferred gestures.

that co-speech gestures are idiosyncratic [118, 188]. There is an unmet need to learn generative models that are able to learn to generate gestures simultaneously from multiple speakers (→ in Figure 4.1) while at the same time remembering what is unique for each speaker's gesture style. These models should not simply remember the "average" speaker. A bigger technical challenge is to be able to transfer gesturing style of speaker 'B' to speaker 'A' (→ in Figure 4.1).

The gesturing style can defined along two dimensions which is a result of (a) the speaker's idiosyncrasy (or speaker-level style), and (b) due to some more general attributes such as standing versus sitting, or the body orientation such as left versus right (or attribute-level style). For both gesture style types, the generation model needs to be able to learn the diversity and expressivity [20, 140] present in the gesture space, within and amongst speakers. The gesture distribution is likely to have multiple modes, some of them shared among speakers and some distinct to a speaker's prototypical gestures.

In this chapter, we introduce the Mixture-Model guided Style and Audio for Gesture Generation (or Mix-StAGE) approach which trains a single model for multiple speakers while learning unique style embeddings for each speaker's gestures in an end-to-end manner (see Figure 4.1). We use this model to perform two tasks for gesture generation conditioned on the input audio signal, (1) **style preservation** which ensures that while learning from multiple speakers we are still able to preserve unique gesturing styles of each speaker, and (2) **style transfer** where generated gestures are from a new style that was not the same as the source of the speech. A novelty of Mix-StAGE is to learn a mixture of generative models which allows for conditioning on the unique gesture style of each speaker. Our experiments study the impact of multiple speakers on both style transfer and preservation. Our study focuses on the non-verbal components of speech asking the research question if we can predict gestures without explicitly modeling verbal signals.

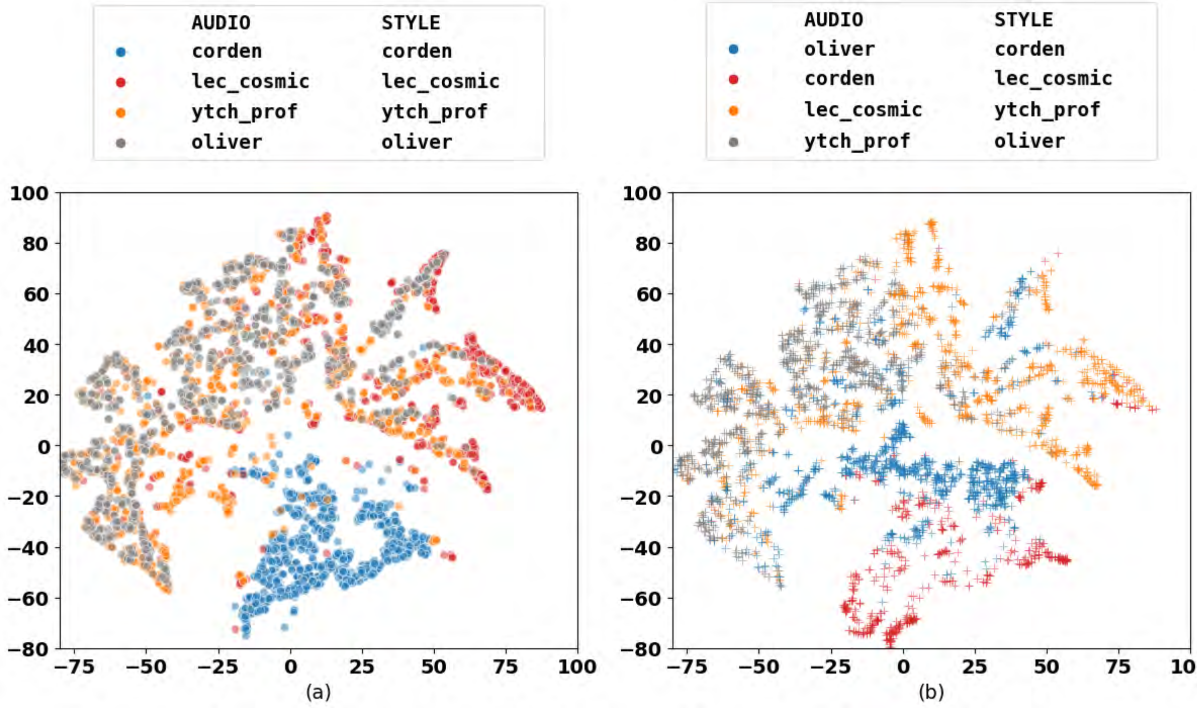


Figure 4.2: t-SNE[115] representation of the Multi-mode Multimodal Gesture Space (Section 4.3.1). Each color represents a style, which is fixed for both plots. The plot on the left visualizes the gesture space generated from the audio content and style of the same speaker. The plot on the right shows the generated gesture space where the audio content and style are not from the same speaker. It can be observed that a similar gesture space is occupied by each speaker's style even when the audio content is not of their own.

4.2 Stylized Co-Speech Gesture Animation

We define the problem of stylized co-speech gesture animation with two main goals, (1) generation of an animation which represents the gestures that would co-occur with the the spoken segment and (2) modification of the style of these gestures. Figure 4.1 shows the first goal (→) exemplified with the style preservation scenario, while the second goal (→) exemplifies with the style transfer scenario.

Formally, given a sequence of T audio frames $\mathbf{X}_a \sim F_a$ and i^{th} speaker's style $S(i)$, the goal is to predict a sequence of T frames of 2-D poses $\mathbf{Y}_p \sim F_p$. Here F_a and F_p are the marginal distributions of the content of input audio and style of output pose sequences. To control pose generation by both style and audio, we learn a joint distribution over pose, audio and style $F_{p,a,s}$ which can be broken down into 3 parts

$$F_{p,a,s} = F_{p|\Phi} F_{\Phi|a,s} \cdot F_s \cdot F_a \tag{4.1}$$

where $F_{\Phi|a,s}$ is the distribution of the gesture space Φ conditioned on the audio and style of pose (Figure 4.1). We discuss the modelling of $F_{p|\Phi} F_{\Phi|a,s}$, F_a, and F_s in Section 4.3.1, 4.3.2 and 4.3.3 respectively.

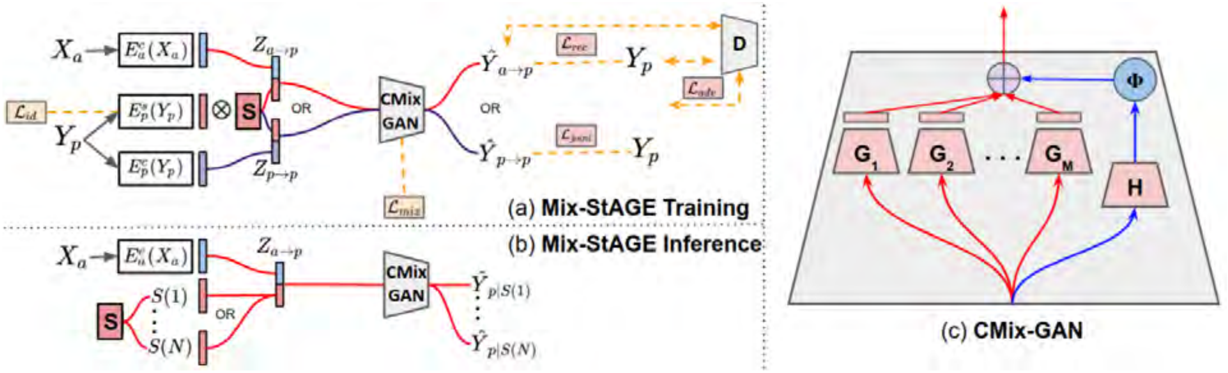


Figure 4.3: (a) Overview of the proposed model Mix-StAGE in training mode, where audio X_a and pose Y_p are fed as inputs to learn a style embedding and concurrently generate a gesture animation. **S** represents the style matrix, which is multiplied with a separately encoded pose. $\bigotimes$ represents argmax for style ID followed by matrix multiplication. Discriminator D is used for adversarial training. All the loss functions are represented with dashed lines. (b) Mix-StAGE in inference mode, where any speaker's style embedding can be used on an input audio X_a to generate gesture style-transferred or style-preserved animations (c) CMix-GAN generator: a visual representation of the conditional Mix-GAN model, where the $\bigoplus$ represents a weighted sum of the model priors Φ with the generated outputs by the sub-generators.

4.3 Mix-StAGE: Mixture-Model guided Style and Audio for Gesture Generation

Figure 4.3 shows an overview of our Mix-StAGE model, including the training inference pathways. A first component of our Mix-StAGE model is the audio encoder E_a^c, which takes as input the spoken audio X_a. During training, we also have the pose sequence of the speaker Y_p. This pose sequence is decomposed into content and style, with two specialized encoders E_p^c and E_p^s. During training, the style for the pose sequence can either be concatenated with the audio or the pose content.

The pose sequences for multiple speakers are represented as a distribution with multiple modes [58]. To decode from this multi-mode multimodal gesture space, we use a common generator G with multiple sub-generators (or CMix-GAN) conditioned on input audio and style to decode both these embeddings to output pose $\mathbf{Y}_p$.

Our loss function comprises of a mode-regularization loss (Section 4.3.1) to ensure that audio and style embedding can sample from the appropriate marginal distribution of poses, a joint loss (Section 4.3.2) to ensure latent distribution matching for content in a cross-modal translation task, a style consistency loss (Section 4.3.3) to ensure that the correct style is being generated and an adversarial loss (Section 4.3.4) that matches the generated pose distribution to the target pose distribution.

4.3.1 M^{2}GS: Multi-mode Multimodal Gesture Space

Humans perform different styles of gestures, where each style consists of different kinds of gestures (i.e beat, metaphorical, emblematic, iconic and so on)[118]. Learning pose generators for multiple speakers, each with their own style of gestures, presents a distribution with multiple modes. These gestures have a tendency of switching from one mode to the other over time, which depends on style embeddings and content of the audio.

To prevent mode collapse [11] we propose the use of mixture-model guided sub-generators [12, 58, 62], each learning a different mode of M^2 gesture space $F_{\Phi|a,s}$.

$$\hat{\mathbf{Y}}_p = \sum_{m=1}^{M} \phi_m G_m(\mathbf{Z}) = G(\mathbf{Z}) \tag{4.2}$$

where $\mathbf{Z} \in \{\mathbf{Z}_{a\to p}, \mathbf{Z}_{p\to p}\}$ are cross-modal and self-modal latent spaces respectively. They are defined as $\mathbf{Z}_{a\to p} = \left[E_a^c(\mathbf{X}_a), E_p^s(\mathbf{Y}_p) \bigotimes \mathbf{S}\right]$ and $\mathbf{Z}_{p\to p} = \left[E_p^c(\mathbf{Y}_p), E_p^s(\mathbf{Y}_p) \bigotimes \mathbf{S}\right]$ where $\mathbf{S}$ is the style embedding matrix (See Section 4.3.3) and $\bigotimes$ is argmax for style ID followed by matrix multiplication. Pose sequence $\hat{Y}_p \sim F_{p|\Phi} F_{\Phi|a,s}$ represents the pose probability distribution conditioned on audio and style. $G_m \sim F^m_{p|a,s}\ \forall m \in [1, 2, \ldots M]$ are sub-generator functions with corresponding mixture-model priors $\Phi = \{\phi_1, \phi_2, \ldots \phi_M\}$. These mixture model priors represent the M^2 gesture space and are estimated at inference time conditioned on the input audio and style.

Estimating Mixture Model Priors: During training, we partition poses Y_p into M clusters using an unsupervised approach, Lloyd's algorithm [109]. While other unsupervised clustering methods [148] can also be used at this stage, we choose Lloyd's algorithm for its simplicity and speed. Each of these clusters represent samples from probability distributions $\{F^1_{p|a,s}, F^2_{p|a,s}, \dots F^M_{p|a,s}\}$. If a sample belongs to the m^{th} cluster, $\phi_m = 1$, otherwise $\phi_m = 0$, making Φ a sequence of one-hot vectors. While training the generator G with loss function $\mathcal{L}_{\text{rec}}$, if a sample belongs to the distribution $F^m_{p|a,s}$, only parameters of sub-generator G_m are updated. Hence, each sub-generator learns different components of the true distribution, which are combined using Equation 4.2 to give the generated pose.

At inference time, we do not have the true values of mixture-model priors Φ. As mixture model priors modulate based on the style of the speaker and audio content at any given moment, we jointly learn a classification network $H \sim F_{\Phi|a,s}$ to estimate values of Φ in form of a mode regularization loss function

$$\mathcal{L}_{mix} = \mathbb{E}_{\Phi,\mathbf{Z}} \text{CCE}(\Phi, H(\mathbf{Z})) \tag{4.3}$$

where CCE is categorical cross-entropy.

4.3.2 Joint Space of Style, Audio and Pose

A set of marginal distributions F_a and F_s are learnt by our content encoders E^c_a and E^c_p, which together define the joint distribution of the generated poses: $F_{p,a,s}$. Since both cross-modal $\mathbf{Z}_{a\rightarrow p}$ and self-modal $\mathbf{Z}_{p\rightarrow p}$ latent spaces are designed to represent the same underlying content distribution, they should be consistent with each other. Using the same generator G for decoding both of these embeddings[107] yields content invariant generator. We enforce a reconstruction and joint loss [5] which encourages a reduction in distance between $\mathbf{Z}_{a\rightarrow p}$ and $\mathbf{Z}_{p\rightarrow p}$. As cross-modal translation is not reversible for this task (i.e. audio signal cannot be generated with pose input), a bi-directional reconstruction loss [92] for latent distribution matching cannot be directly used. This joint loss achieves the same goal of latent distribution matching in a uni-modal translation task [69, 151, 152] but for a cross-modal translation task.

$$\mathcal{L}_{joint} = \mathbb{E}_{\mathbf{Y}_p} \|\mathbf{Y}_p - G(\mathbf{Z}_{p \to p})\|_1 \tag{4.4}$$

$$\mathcal{L}_{rec} = \mathbb{E}_{\mathbf{Y}_p, \mathbf{X}_a} \|\mathbf{Y}_p - G(\mathbf{Z}_{a \to p})\|_1 \tag{4.5}$$

4.3.3 Style Embedding

We represent style as a collection of embeddings $S(i) \in \mathbf{S} \sim F_s$, where $S(i)$ is the style of the i^{th} speaker in the style matrix **S**. Style space and embeddings are conceptually similar to the GST (Global Style Token) layer [182] which decomposes the audio embedding space into a set of basis vectors or style tokens, but only one out of the two modalities in the stylized audio generation task [114, 182] have both style and content. In our case, both audio and pose have style and content. To ensure that generator G is attending only to style of pose while ignoring style of the audio, a style consistency loss is enforced on input $\mathbf{Y}_p$ and generated $\hat{\mathbf{Y}}_p$.

$$\mathcal{L}_{id} = \mathbb{E}_{Y \in \{\mathbf{Y}_p, \hat{\mathbf{Y}}_p\}} \text{CCE}\left(\text{Softmax}\left(E_p^s(Y)\right), \mathbf{ID}\right) \tag{4.6}$$

where **ID** is a one-hot vector denoting the speaker level style.

4.3.4 Total Loss with Adversarial Training

To alleviate the challenge of overly smooth generation caused by L1 reconstruction and joint losses in Equation 4.4,4.5, we use the generated pose sequence $\hat{\mathbf{Y}}^p$ as a signal for the adversarial discriminator D [50]. The discriminator tries to classify the true pose $\mathbf{Y}^p$ from the generated pose $\hat{\mathbf{Y}}^p$, while the generator learns to fool the discriminator by generating realistic poses. This adversarial loss[52] is written as:

$$\mathcal{L}_{adv} = \mathbb{E}_{\mathbf{Y}_p} \log D\left(\mathbf{Y}_p\right) + \mathbb{E}_{\mathbf{X}_a, \mathbf{Y}_p} \log\left(1 - D(G\left(\left[E_a^c(\mathbf{X}_a), E_p^s(\mathbf{X}_p)\right]\right))\right) \tag{4.7}$$

The model is jointly trained to optimize the overall loss function:

$$\max_{D} \min_{E_a^c, E_p^c, E_p^s, G} \mathcal{L}_{mix} + \mathcal{L}_{joint} + \mathcal{L}_{rec} + \lambda_{id}\mathcal{L}_{id} + \mathcal{L}_{adv} \tag{4.8}$$

where λ_{id} controls the weight of the style consistency loss term.

4.3.5 Network Architectures

Our proposed approach can work with any temporal network, giving it the flexibility of incorporating domain dependent or pre-trained temporal models.

In our experiments we use a Temporal Convolution Network (TCN) module for both content and style encoders. The style space is a matrix $\mathbf{S} \in \mathbb{R}^{N \times D}$ where N is the number of speakers and D is the length of the style embeddings. The generator $G(.)$ consists of a 1D version of U-Net [50, 150] followed by M TCNs as sub-generator functions. The discriminator is also a TCN module with lower capacity than the generators. A more detailed architecture can be found in the supplementary.

		Single-Speaker Models				**Multi-Speaker Models**					
No. of Speakers	**Speaker**	**S2G**[50]		**CMix-GAN**		**MUNIT**[69]		**StAGE**		**Mix-StAGE**	
		PCK	**F1**	**PCK**	**F1**	**PCK**	**F1**	**PCK**	**F1**	**PCK**	**F1**
2	**Mean**	0.25	0.08	0.26	**0.27**	0.24	0.06	**0.36**	0.21	0.34	0.22
	Corden	0.30	0.05	0.32	0.21	0.25	0.06	0.36	0.21	0.34	0.24
	lec_cosmic	0.19	0.12	0.19	0.33	0.15	0.19	0.20	0.48	0.24	0.49
4	**Mean**	0.37	0.18	0.37	0.27	0.22	0.05	0.38	**0.34**	**0.39**	**0.35**
	Corden	0.30	0.05	0.32	0.21	0.24	0.07	0.35	0.27	0.35	0.30
	lec_cosmic	0.19	0.12	0.19	0.33	0.19	0.16	0.18	0.23	0.20	0.19
8	**Mean**	0.36	0.14	0.37	0.26	0.31	0.21	0.38	**0.32**	**0.40**	**0.33**
	Corden	0.30	0.05	0.32	0.21	0.23	0.03	0.32	0.28	0.36	0.27
	lec_cosmic	0.19	0.12	0.19	0.33	0.13	0.09	0.23	0.34	0.24	0.32

Table 4.1: **Style Preservation**: Objective metrics for pose generation of single-speaker and multi-speaker models as indicated in the columns. Each row refers to the number of speakers the model was trained, with the average performance indicated at the top. The scores for common individual speakers are also indicated below alongside. For detailed results on other speakers please refer to the supplementary. Bold numbers indicate $p < 0.1$ in a bootstrapped two sided t-test.

4.4 Experiments

Our experiments are divided into 2 sections, (1) **Style Preservation:** Generating co-speech gestures for multiple speakers with their own individualistic style, (2) **Style Transfer:** Generating co-speech gestures with content (or audio) of a speaker and gesture style of another speaker. Additionally, style transfer can be speaker-level as well as attribute-level. We choose visually distinguishable attribute-level styles: (1) body orientation, (2) gesture frequency, (3) primary arm function and (4) sitting/standing posture. We start by describing the baseline models followed by the evaluation metrics, which we will use to compare our model. We end this section with the description of our proposed dataset.

4.4.1 Baseline Models

Single-Speaker Models: These models are not designed to perform style transfer and hence are not included for those experiments.

- **Speech2Gesture [50]**: The closest work to co-speech gesture generation is one that only generates individualistic styles. We use the pre-trained models available from their codebase to render the videos. For rest of the speakers in PATS, we replicate their model, hyper-parameters and train speaker specific models.
- **CMix-GAN** (variant of our model): As an ablation, we remove the style embedding module and style consistency losses from our model Mix-StAGE . Hence, a separate model is required to be trained for each speaker for style preservation experiments.

Multi-speaker Models

- **MUNIT [69]**: The closest work to our style-transfer task is MUNIT which takes multiple domains of images (i.e. uni-modal). We modify the encoders and decoders to domain specific architectures (i.e. 1D convolutions for audio instead of 2D convolutions for images) while retaining the loss functions.
- **StAGE** (variant of our model): As an ablation, we fix the number of sub-generators in our model Mix-StAGE to one. This is equivalent to setting $M = 1$ in equation 4.2.

4.4.2 Evaluation Metrics

Human Perceptual Study:

We conduct a human perceptual study on Amazon Mechanical Turk (AMT) for co-speech gesture generation (or style preservation) and style transfer (speaker-level and attribute-level) and measure preferences in two aspects of the generated animations, (1) **naturalness**, and (2) **style transfer correctness** for animation generation with content (i.e. audio) of speaker A and style of speaker B. 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 generated using our proposed model. The generated animation could either have the same style or a different style as the original speaker. With unlimited time, the annotator has to answer two questions, (1) Which of the videos has more natural gestures? and (2) Do these videos have the same attribute-level style (or speaker-level style)? The first question is a real vs. fake perceptual study against the ground truth, while the second question measures how often the algorithm is able to visually preserve or transfer style (attribute or individual level). We run this study for randomly selected 100 pairs of videos from the held-out set. .

Probability of Correct Keypoints (PCK):

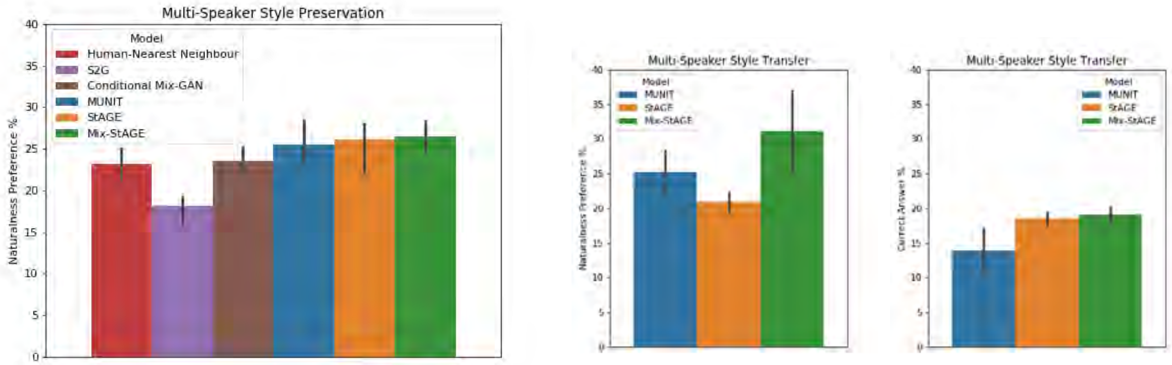
To measure the accuracy of the gesture generation, PCK [10, 160] is used to evaluate all models. PCK values are averaged over $\alpha = 0.1, 0.2$ as suggested in [50].

Mode Classification F1:

Correctness of shape of a gesture can be quantified by measuring the number of times the model has sampled from the correct mode of the pose distribution. Formally, we use the true ($\mathbf{Y}_p$) and generated ($\hat{\mathbf{Y}}^p$) pose to find the closest cluster $\hat{m}$ and m respectively. If $m = \hat{m}$, the generated pose was sampled from the correct mode. F1 score of this M-class classification problem is defined as Mode Classification F1, or simply F1.

Inception Score (IS):

Generated pose sequences with the audio of speaker A and style of speaker B does not have a ground truth reference. To quantitatively measure the correctness and diversity of generated pose sequence we use the inception score [155]. For generative tasks such as image generation, this metric has been used with a pre-trained classification network such as Inception Model [164]. In our case, the generated samples are not images, but a set of 2D keypoints. Hence, we train a network which classifies a sequence of poses to its corresponding speaker which estimates the conditional likelihood to calculate IS scores.



(a) Style Preservation Naturalness

(b) Style Transfer Naturalness

(c) Style Transfer Correctness

Figure 4.4: Perceptual Study for speaker-level style preservation in (a) and speaker level style transfer in (b), (c). We have naturalness preference for both style transfer and preservation, and style transfer correctness scores for style transfer. Higher is better. Error bars calculated for $p < 0.1$ using a bootstrapped two sided t-test.

4.4.3 Pose-Audio-Transcript-Style (PATS) dataset

Gesture styles, which may be defined by attributes such as type, frequency, orientation of the body, is representative of the idiosyncrasies of the speaker [38]. For our experiments we use Pose-Audio-Transcript-Style (PATS)which was introduced in Chapter 3, to study various styles of gestures for a large number of speakers in diverse settings. PATS contains pose sequences aligned with corresponding audio signals and transcripts[3] for 25 speakers (including 10 speakers from [50]) to offer a total of 251 hours of data, with a mean of 10.7 seconds and a standard deviation of 13.5 seconds per interval. The demographics of the speakers include 15 talk show hosts, 5 lecturers, 3 YouTubers, and 2 televangelists.

Each speaker's pose is represented via skeletal keypoints collected via OpenPose [23] similar to [50]. It consists of 52 coordinates of an individual's major joints for each frame at 15 frames per second, which we rescale by holding the length of each individual's shoulder constant. This prevents the model from encoding limb length in the style embeddings. Following prior work [50, 88], we represent audio features as mel-spetrograms, which is a rich input representation shown to be useful for gesture generation.

[3] While transcripts are a part of this dataset, they are ignored for the purposed of this chapter.

Model	Number of Speakers			Attributes			
	2 Speakers	4 Speakers	8 Speakers	Sitting vs Standing	Gesture Frequency	Body Orientation	Primary Arm Func.
MUNIT [69]	1.11	1.90	2.06	1.10	2.49	1.05	3.32
StAGE	2.17	**2.85**	3.89	1.68	4.38	**6.81**	3.14
Mix-StAGE	**2.61**	**2.85**	**4.48**	**3.08**	**4.50**	6.69	**3.32**

Table 4.2: **Style Transfer**: Inception scores for style transfer on multi-speaker models (indicated in each row). Columns on the left refer to the speaker-level style transfer task while those on the right refer to the specific attribute-level style task. Bold numbers indicate $p < 0.1$ in a bootstrapped two sided t-test.

4.5 Results and Discussion

We group our results and discussions in (1) a first set of experiments studying style preservation (when output gesture styles are the same the original speaker) and (2) a second set of experiments studying transfer of gesture styles.

4.5.1 Gesture Animation and Style Preservation

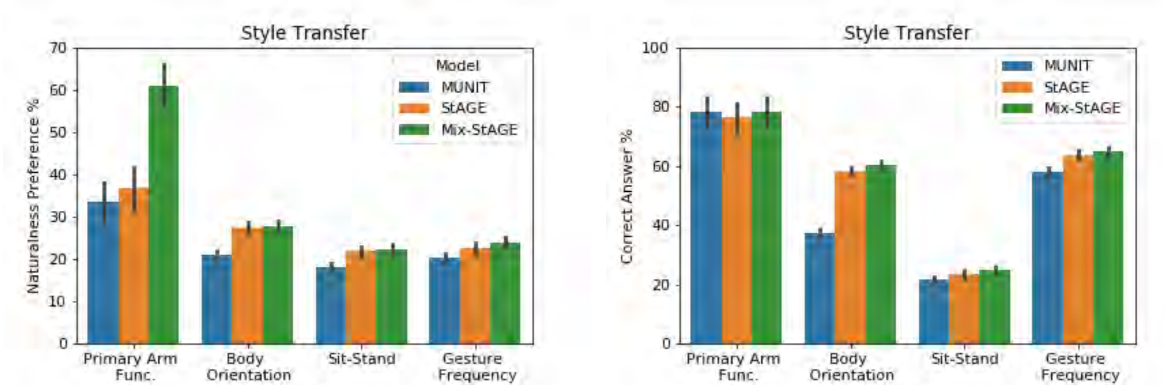
To understand the impact of adding more speakers, we select a random sample of 8 speakers for the largest 8-speaker multi-speaker model, and train smaller 4-speaker and 2-speaker models where the speakers trained are always a subset of the speakers that were trained in a larger model. This allows to compare the performance on the same two initial speakers which are '*Corden*' and '*lec_cosmic*' in our case. We also compare with single-speaker models trained and tested on one speaker at a time.

Impact of training with Multiple Speakers

Results from Table 4.1 show that multi-speaker models outperform single-speaker models especially for pose accuracy (i.e. PCK), shape and timing (i.e. F1). We find that increasing the number of speakers could sometimes reduce the performance of individual speakers but the overall performance generally shows improvement.

Comparison with previous baselines

To compare with prior baselines, we focus first on the subjective evaluation shown in Figure 4.4a, since it is arguably the most important metric. The results show consistent improvements on the naturalness rating for our proposed model Mix-StAGE and also our single-speaker variant CMix-GAN over the previous state of the art approach S2G [50]. We also observe that multi-speaker models perform better than single speaker-models. In Table 4.1, we show similar quantitative improvements of Mix-StAGE and CMix-GAN over S2G for both PCK and F1 scores.



(a) Naturalness Preference

(b) Style Transfer Correctness

Figure 4.5: A visualization of the perceptual human study for attribute-level style transfer with (a) naturalness preference, and (b) style transfer correctness scores for the generated animations for a different style than the speaker. Higher is better. Error bars calculated for $p < 0.1$ using a bootstrapped two sided t-test.

Impact of Multiple Generators for Decoding

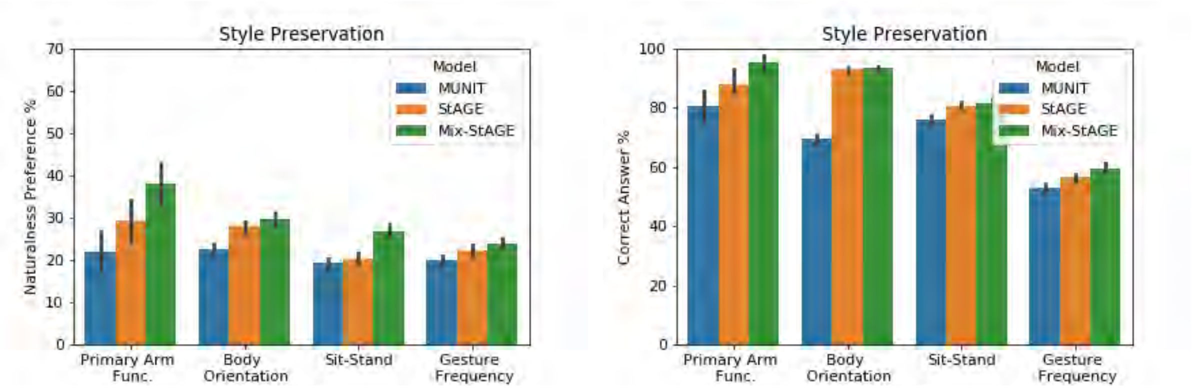
Mix-StAGE's gesture space models multiple modes, as seen in Figure 4.2. Its importance is shown in Table 4.1 where models with single generators as the decoder (i.e. S2G, MUNIT and StAGE) showed lower F1 scores, most likely due to mode collapse while training. Multiple generators in CMix-GAN and Mix-StAGE boost F1 scores as compared to other models in the single-speaker and multi-speaker regimes respectively. A similar trend was observed in the perceptual study in Figure 4.4.

We also study the impact of the number of generators (hyperparameter M) in our Mix-StAGE model. While for small number of speakers (i.e. 2 speakers) a single generator is good enough, the positive effect of multiple generators can be observed as the number of speakers increase (see Table 4.1). We also vary $M \in \{1, 2, 4, 8, 12\}$ and observe that improvements seem to plateau at

$M = 8$ with only marginal improvements for larger number of sub-generators. For the ablation study we refer the readers to the supplementary.

Attribute-level Style Preservation in Multi-Speaker Models

We also study style preservation for attributes in Section 4.4 as a perceptual study in Figure 4.6. We observe that humans deem animations generated by Mix-StAGE significantly more natural in most cases. High scores ranging 60-90% for style preservation correctness, with Mix-StAGE outperforming others, are observed for pairs of speakers in Figure 4.6b. This indicates that style preservation may be a relatively easy task as compared to style transfer for multi-speaker models. With this, we now shift our focus to style transfer.



(a) Naturalness Preference

(b) Style Preservation Correctness

Figure 4.6: A visualization of the perceptual human study for attribute-level style preservation with (a) naturalness preference, and (b) style preservation correctness scores for the generated animations for the same style as the speaker. Higher is better. Error bars calculated for $p < 0.1$ using a bootstrapped two sided t-test.

4.5.2 Style Transfer

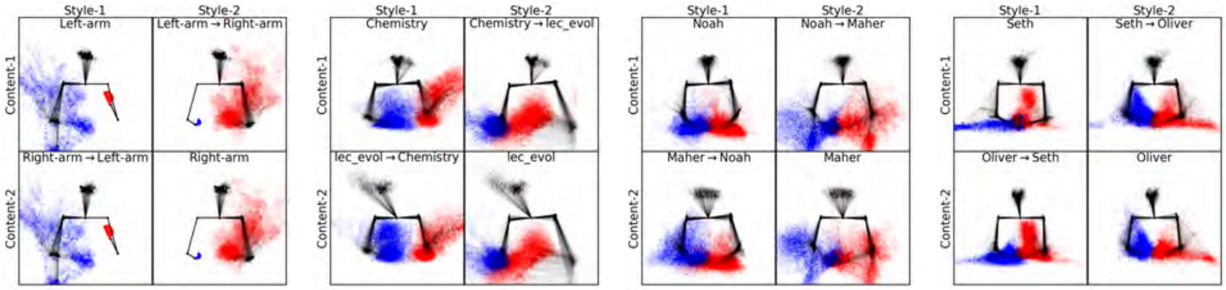
Speaker-level Style Transfer

To study our capability to transfer style of a specific speaker to a new speaker, we will compare the gesture spaces between the original speakers and the transferred speakers. Figure 4.2a shows that each original speaker occupies different regions in the M^2 gesture space. Using our Mix-StAGE model to transfer style, we can see the new gesture space in Figure 4.2b. For the transferred speakers the 2 spaces look quite similar. For instance, '*Corden*' style (a speaker in our dataset) is represented by the color blue in Figure 4.2a and occupies the lower region of the gesture space. When Mix-StAGE generates co-speech gestures using audio of '*Oliver*' and the style of '*Corden*', it occupies a subset of '*Corden's*' region in the gesture space, also represented by blue in Figure 4.2b. We see a similar trend for styles of '*Oliver*' and '*ytch_prof*'. This is an indication of a successful style transfer across different speakers. We note the lack of clean separation in the gesture space among different styles as there could common gestures across multiple speakers.

For the perceptual study, we want to know if humans can distinguish the generated speaker styles. For this, we show human annotators two videos: a ground truth video in a specific style, and a generated video which is either from the style of the same speaker or a different speaker. Annotators have to decide if this is the same style or not. We use the 4-speaker model for this experiment. Figure 4.4b shows naturalness preference and 4.4c shows percentage of the time style was transferred correctly. Our model Mix-StAGE performs best in both cases. This trend is corroborated with higher inception scores in Table 4.2.

Impact of Number of Speakers for Style Transfer

In Table 4.2, we observe that increasing the number of speakers used for training also increases the average inception score for the stylized gesture generations. This is a welcome effect as it indicates increases in the diversity and the accuracy of the generations.



(a) Primary Arm Func. (b) Body Orientation (c) Sitting vs Standing (d) Gesture Frequency

Figure 4.7: Style-Content Heatmaps for attribute-level style transfer. Each column represents the same style, while rows have input audio from different speakers. These heatmaps show that gestures are consistent across audio inputs but different between styles. Red regions correspond to the motion of the right arm, while blue corresponds to the left.

Attribute-level Style Transfer in Multi-Speaker Models

We study four common attributes of gesture style which are also visually distinguishable by humans: (1) sitting vs. standing, (2) high vs low gesture frequency, (3) left vs right body orientation and (4) left vs right primary arm. Speakers were selected carefully to represent each extremes of these four attributes. We run a perceptual study similar to the one for speaker-level styles. However, we ask the annotators to judge if the attribute is the same in both of the videos (e.g. are both the people gesturing with the same arm?). Results from Figure 4.5 show that Mix-StAGE generates more (or similar) number of natural gestures with the correct attribute-level style compared to the other baselines. We also observe that it is harder for humans to determine if a person is standing or sitting, which we suspect is due to the missing waistline in the animation.

For a visual understanding of the generated gestures and stylized gestures, we plot a style-content heatmap in Figure 4.7, where columns represent generations for a specific style, while rows represent different speaker's audio as input. These heatmaps show that gestures are consistent across audio inputs but different between styles. Accuracy and diversity of style transfer is corroborated by inception scores in Table 4.2.

The Mix-StAGe approach is effective for many gesture styles, a limitation of this work is that it requires a significant amount of data for each speaker. Sometimes, we may only be able to curate a few minutes of data for a new speaker. In the next chapter, we overcome this limitation by learning a low resource generative modeling approach that can personalize both grounding and gesture style with only a few minutes of data of the new speaker.

4.6 Related Work

Speech driven Gesture Generation: For prosody-driven head motion generation [156] and body motion generation [96, 97], Hidden Markov Models were used to predict a sequence of frames. Chiu & Marsella [31] proposed a two-step process: predicting gesture labels from speech signal using conditional random fields (CRFs) and converting the label to gesture motion using Gaussian process latent variable models (GPLVMs). More recently, an LSTM network was applied to MFCC features extracted from speech to predict a sequence of frames for gestures [59] and body motions [6, 159]. Generative adversarial networks (GAN) were used to generate head motions [153] and body motions[45]. Gestures driven by an audio signal[50] is the closest approach to our task of style preservation but it uses models trained on single speakers unlike our multi-speaker models.

Disentanglement and Transfer of Style: Style extraction and transfer have been studied in context of image artistic style [48, 77], factorizing foreground and background in videos[39, 177], disentanglement in speech [21, 55, 182]. These approaches were extended to translation between properties of style such as map edges and real photos using paired samples [75]. Paired data limits the variety of attributes of source and target, which encouraged unsupervised domain translation for images [202, 203] and videos[17]. Style was disentangled from content using a shared latent space[108], a cycle consistency loss [202] and contrastive learning [127]. Cycle consistency losses were shown to limit diversity in the generated outputs as opposed to a weak consistency loss [69] and shared content space [92]. Cycle consistency in cross-domain translation assumes reversibility (i.e. domain A can be translated to domain B and vice-versa). These assumptions are violated in cross-modal translation [114] and style control [182] tasks where information in modality B (e.g. pose) is a subset of that in modality B (e.g. audio). Style transfer for pose has been studied in context of generating dance moves based on the content of the audio [93] or walking styles [161]. Generated dance moves are conditioned on both the style and content of the audio (i.e. kind of music like ballet or hip-hop), unlike co-speech gesture generation which requires only the content and not the style of the audio (i.e. speaker specific style like identity or fundamental frequency). Co-speech gesture styles have been studied in context of speaker personalities [131], but requires a long annotation process to create a profile for each speaker. To our knowledge, this is the first fully data-driven approach that learns gesture style transfer for multiple speakers in a co-speech gesture generation setting.

Chapter 5

Co-Speech Gestures: Low Resources for Grounding and Gesture Style

5.1 Overview

Technologies to assist human communication, both verbal (e.g. spoken language) and nonverbal (e.g. co-speech gestures), have gained more traction in the past decade. One promising direction is virtual reality [37, 68, 72, 110] which aims at creating a more realistic online communication platform for embodied virtual agents [24, 103] and remote avatars [136, 137]. These advancements could be seen as a normal progression to speech-based technologies such as intelligent personal assistants (e.g. Alexa, Siri, Cortana). These agents, in the future, could also communicate more naturally with a nonverbal embodiment that complements the verbal communication [119]. To enable this vision of immersive verbal and nonverbal communication through an avatar, one technical challenge is generating visual gestures based on input speech and language [7, 50, 60, 89]. An even more challenging task is the generation of personalized visual gestures, which reflects the idiosyncratic behaviours of a specific person [119]. The main goal of our chapter is to create a personalized gesture generation model (e.g. as part of a personalized avatar) with limited data from a new speaker. In technical terms, this problem requires an adaptation of crossmodal generative models in a low-resource setting as illustrated in Figure 5.1. Leveraging an existing source model, pretrained on a large dataset of one speaker (i.e. source domain), our goal is to personalize to a new speaker (i.e. target domain) with only 2 minutes of the target data.

The problem setting brings a unique challenge, typically not studied in typical domain adapta-

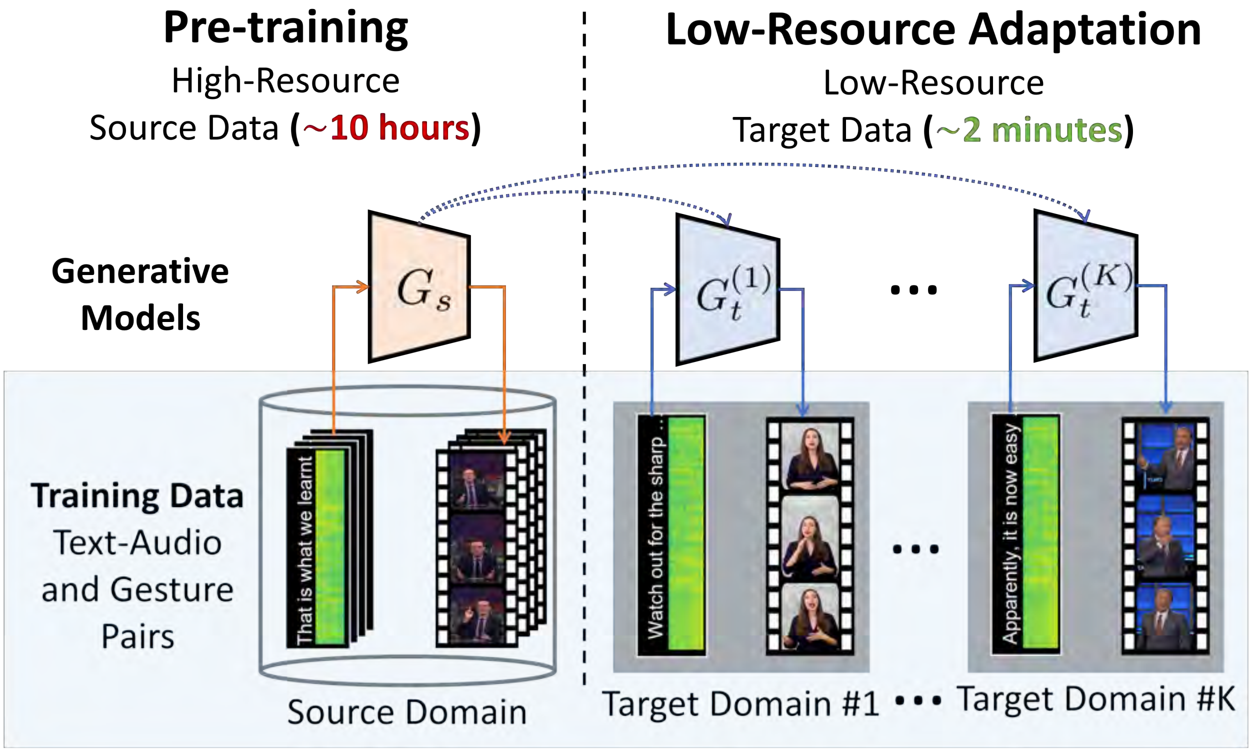


Figure 5.1: Overview of the co-speech gesture personalization task. On the left is a generative source model G_s pre-trained on a source speaker. We adapt G_s to multiple target models G_t using low-resource data for each of the target speaker.

tion settings: *crossmodal grounding shift*. Due to the crossmodal nature of our task, crossmodal grounding shift refers to the distributional shift of the relationships between the input spoken language modalities and output gesture modality. For example, consider a speaker, *Aarti*, who waves their right hand while greeting a friend. These gestures are conditionally dependent on what the speaker says. We define such relationships between the gestures and spoken language as crossmodal grounding. While these relationships are conditioned on spoken language, they are also heavily influenced by the speaker's idiosyncrasies. Now, consider a new speaker, *Bob*, who chooses from a set of two gestures while greeting: a left handed wave or waving both their hands vigorously. As is the case for this speaker, typically the conditional gesture spaces have a larger support (i.e. different kinds of gestures) for the same language context. Such differences between conditional gestures of source and target speakers are very common, especially because the conditional variable is language which is a very large space. These common, yet complex differences represent crossmodal grounding shift.

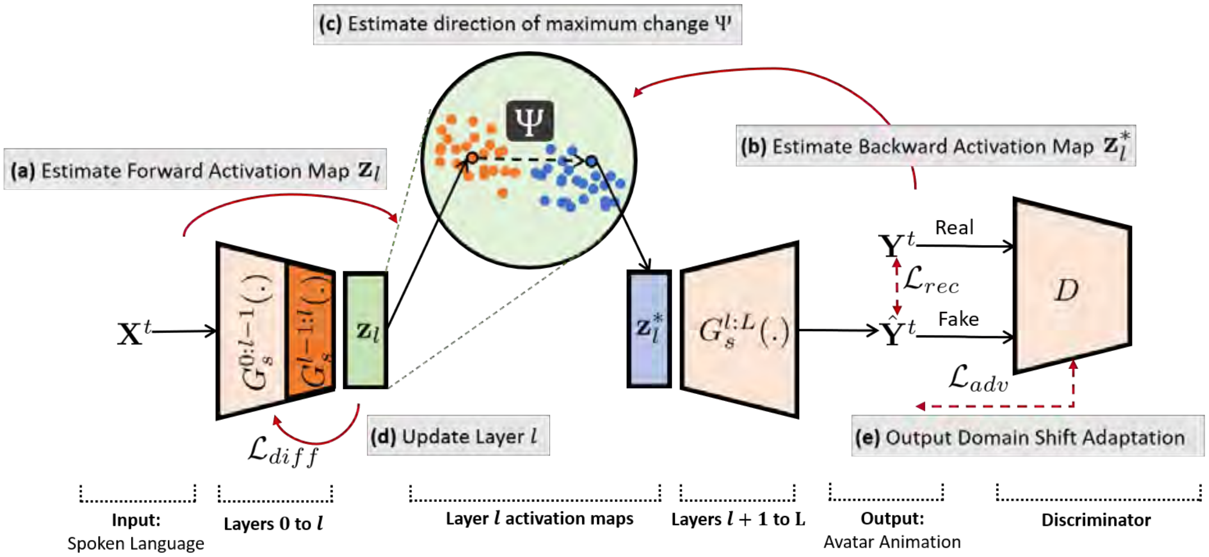


Figure 5.2: Overview of the key components of our proposed model DiffGAN. (a)-(d): Low-resource adaptation of the crossmodal grounding relationships from source to target domain. (e): Modeling output domain shift from source to target domain

In this chapter, we propose an approach, named DiffGAN, that can efficiently personalize co-speech gesture generation models from a high-resource source speaker to a low-resource target speaker. To the best of our knowledge, this is the first approach that is able to learn a personalized model with only 2 minutes of speaker data (i.e. as opposed to 10 hours [7, 50, 60, 89]). Our DiffGAN approach does not require access to source training data. Instead, DiffGAN directly identifies shifts in crossmodal grounding relationships along with the shifts in the output domain from the pretrained source model. Based on these identified distribution shifts, DiffGAN updates a few necessary parameters in a single layer of the source model, allowing efficient adaptation with low resources. Our experiments study the effectiveness of our DiffGAN approach on a diverse publicly available dataset. As part of our evaluation methodology, we report that DiffGAN produces a consistent improvement of around 10% preference scores of human judgments over strong baselines among other quantitative improvements. Furthermore, DiffGAN extrapolates to gestures in the target distribution without ever having seen them in the source distribution.

5.2 Problem Statement

We are given a pretrained gesture generation model of a source speaker as a generator-discriminator pair (G_s, D_s) [52] that is trained on a large gesture source dataset $\mathcal{D}_s = \{\mathbf{X}_i^s, \mathbf{Y}_i^s\}_{i=1}^N$. It gen-

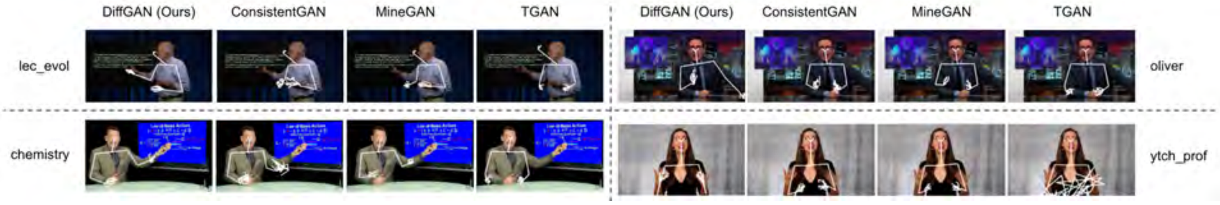


Figure 5.3: Qualitative comparison of our DiffGAN with prior work over shape of generated gestures. With **maher** as the source domain, target model outputs over the target domain are superimposed over ground truth video frames for easy comparison.

		Gesture Quality		Crossmodal Grounding		Output Domain
Pre-trained G_s	Models	Naturalness	Expressivity	Timing	Relevance	Style
✗	**AISLe [7]**	7.3 ± 2.9	15.3 ± 7.6	9.2 ± 2.4	8.3 ± 4.1	16.3 ± 5.3
✓	**TGAN [179]**	9.4 ± 3.8	12.7 ± 4.6	12.1 ± 2.3	10.8 ± 4.0	13.7 ± 2.9
✓	**MineGAN [180]**	13.0 ± 2.9	16.6 ± 4.8	16.0 ± 2.3	14.1 ± 4.2	29.6 ± 7.3
✓	**ConsistentGAN [135]**	9.0 ± 1.9	17.7 ± 3.1	10.9 ± 1.9	9.3 ± 1.6	17.6 ± 6.3
✓	**DiffGAN (Ours)**	**21.9 ± 2.5**	**27.6 ± 6.5**	**26.2 ± 2.1**	**23.9 ± 4.8**	**46.3 ± 9.2**
✓	**DiffGAN w/o $\mathcal{L}_{diff}$**	**19.8 ± 1.3**	**24.4 ± 5.0**	**22.1 ± 3.3**	21.0 ± 3.0	**47.8 ± 4.8**
✓	**DiffGAN w/o $\mathcal{L}_{adv}$**	12.3 ± 4.1	**20.0 ± 5.9**	15.8 ± 3.9	13.6 ± 3.8	26.5 ± 6.4

Table 5.1: Human perceptual study comparing our model with prior work and strong baselines over five criteria measuring **quality**, **crossmodal grounding** and **output domain shift** of generated gestures. We report the preference scores of a model as compared to the ground truth gestures. Confidence intervals reported as standard deviation across experiments on all source-target pairs. All models are adapted/trained with 2 minutes of target data. Higher is better and for scores beyond 50 % the model passes the Turing test as it fools humans on average.

erates gestures as a sequence of body poses $\mathbf{Y}_i^s$ that is driven by both language and speech as the input modalities $\mathbf{X}_i^s$. A goal is to adapt parameters of the pretrained generator G_s to a target model G_t by using a much smaller target dataset of the target speaker $\mathcal{D}_t = \{\mathbf{X}_i^t, \mathbf{Y}_i^t\}_{i=1}^P$ where $P << N$.

5.3 Method

We propose a new approach, DiffGAN, that learns a target model G_t by adapting a pre-trained source model G_s in a low-resource setting. This approach is a two-step process illustrated in Figure 5.2. First, in section 5.3.1, the model learns to identify the crossmodal grounding shifts through a novel loss function $\mathcal{L}_{diff}$ and low-resource target data. Second, in section 5.3.2, we discuss use of a loss function $\mathcal{L}_{shift}$, which encourages the target model to shift the output domain distribution to be closer to that of the target's. Optimization of the combined loss function describes the complete model,

$$G_t^* = \mathbb{E}_{\mathbf{X},\mathbf{Y} \in \mathcal{D}_t} \underset{G_s, \theta_{l-1:l}}{\text{argmin}} \max_{D_s} \mathcal{L}_{diff}(\theta_{l-1:l}) + \mathcal{L}_{shift}(G_s, D_s) \tag{5.1}$$

where $\theta_{l-1:l}$ are parameters of a layer l in G_s (discussed in Section 5.3.1).

5.3.1 Crossmodal Grounding Shift

The crossmodal grounding relationships between input and outputs spaces of source data are already encoded in the source model, G_s. Instead of modifying all of these relationships in the source model, we first discover the shift in grounding from the source to the the target dataset. This is followed by the adaptation of the model parameters which account for only the shifted relationships in the target model. This approach has a two key advantages. First, adapting to only the differences suggested by the discrepancies between target and source data, allows the model to retain the previously learnt essential grounding relationships intact. Second, in a low-resource setting, updating only a few layers instead of the complete model is sufficient [18, 123]. Doing so allows the target model to learn new grounding relationships while preventing overfitting.

Notations: Source G_s and target G_t models both have a total of L layers. Function $G_s^{l:m}(.)$ represents layers l through m in G_s. For example, $G_s^{l:m}(.)$ takes activation maps of layer l (or $\mathbf{z}_l$) in G_s as the input and returns activation maps of layer m (or $\mathbf{z}_m$) in G_s as the output. Parameters of layers l through m can be explicitly specified in function $G_s^{l:m}(.; \theta_{l:m})$, but may be skipped for brevity.

Discovering shift in crossmodal grounding relationships: Crossmodal grounding represents relationships between the input and output modalities. For the source data. these relationships are already encoded in the latent spaces [8, 77, 196] of the source model, G_s. In the adapted target model G_t, this latent space will shift creating new grounding relationships. More concretely, this latent space represents the activation maps at layer l for source and target models, or z_l and z_l^* respectively. To estimate the direction along which the grounding relationships have shifted, we compute the vector difference $\Psi = |z_l - z_l^*|$ between the activation maps at layer l. We can now update the parameters of layer l in the direction Ψ to produce the required grounding shift.

Computing direction of grounding shift Ψ: To compute $\Psi = |z_l - z_l^*|$, we need both z_l and z_l^*. As z_l is an activation map of the source model with a target sample $\mathbf{X}^t$ as input, we can compute it as $z_l = G_s^{0:l}(\mathbf{X}^t)$ as shown in Figure 5.2a. Estimating z_l^* is tricky as the target model is not available yet. As we are only updating the parameters of layer l in this step, the parameters of $G_s^{l:L}$ do not change. As a result, we can use values of the target output modality $\mathbf{Y}^t$ to optimize $\text{argmin}_{\mathbf{z}} \|G_s^{l:L}(\mathbf{z}) - \mathbf{Y}^t\|_2$ as shown in Figure 5.2b. This minimization objective serves as an accurate estimate of z_l^*, and consequently an accurate estimate of the direction of grounding shift Ψ. To prevent overfitting due to limited amount of data, we concentrate the gradient update to the directions (i.e. channel dimensions) with top-k grounding shifts represented as Ψ_k. The criteria for choosing l and k is discussed in Section 5.5.

Updating Crossmodal Grounding in layer l: To encourage generation of the shifted latent space z_l^* for target domain inputs $\mathbf{X}_t$, we update the weights of only layer l (or $\theta_{l-1:l}$) through an L2 loss. Furthermore, as Ψ_k is the measure of grounding shift for each parameter, we use it as a weighting function to guide parameter updates of layer l,

$$\mathcal{L}_{diff} = \|\Psi_k \odot \mathbf{z}_l^* - \Psi_k \odot G_s^{l-1:l}\left(G_s^{0:l-1}(\mathbf{X}^t); \theta_{l-1:l}\right)\|_2, \tag{5.2}$$

where $\odot$ is element-wise product. As the training progresses, both Ψ_k and z_l^* are re-estimated based on the updated parameters of the source model. Hence, as the latent space of the adapted source model shifts closer to that of the target domain, Ψ_k will re-adjust until convergence, arriving at the target model. Please note that while this approach is described for a single layer l, it can easily be adjusted to update any sequence of layers l through m without loss of generality.

5.3.2 Output Domain Shift

The second step is to shift the output domain of the source model G_s toward that of the target gesture distribution. We follow the fine-tuning approach suggested in [135, 179] of optimizing for the adversarial loss function L_{adv},

$$\mathcal{L}_{adv} = \mathbb{E}_{\mathbf{X}^t, \mathbf{Y}^t \in \mathcal{D}_t} \log D_s\left(\mathbf{Y}^t\right) + \log\left(1 - D_s\left(G_s\left(\mathbf{X}^t\right)\right)\right), \tag{5.3}$$

where the discriminator $D_s(.)$ measures domain correctness of the output modality. This adversarial loss encourages the model to generate gesture sequences whose structure represents the target distribution. We would also like to encourage generation of output sequences that

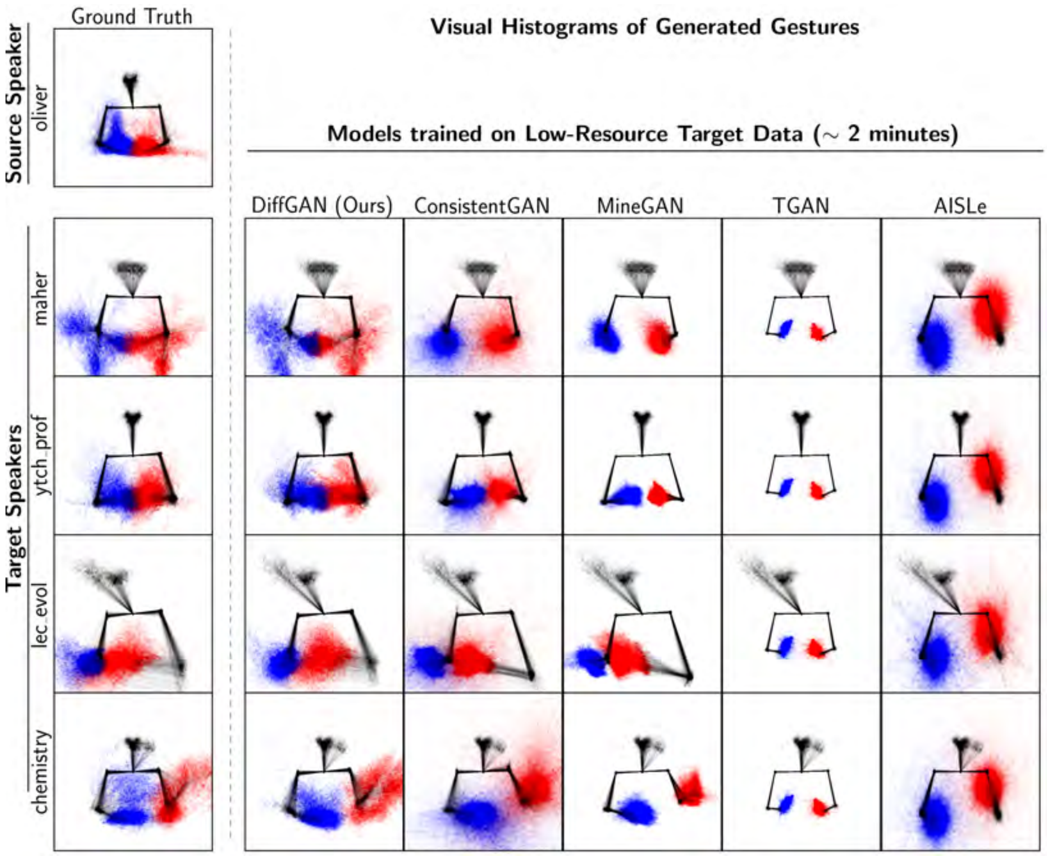


Figure 5.4: Visual Histograms of generated gestures visually describe the distribution of hand gestures in space. Red and blue colors denote the left and right arms respectively. First row is the source speaker, below which we have all the target speakers. Each column denotes a model which adapts output distribution of the source domain to the target domain. Qualitatively, DiffGAN is successful in modeling the distribution of the source speaker with just 2 minutes of data.

temporally match the target ground truth sequences $\mathbf{Y}^t \in \mathcal{D}_t$ for which we use a reconstruction loss [7, 8, 50],

$$\mathcal{L}_{rec} = \mathbb{E}_{\mathbf{X}^t, \mathbf{Y}^t \in \mathcal{D}_t} ||\mathbf{Y}^t - \hat{\mathbf{Y}}^t||_1, \tag{5.4}$$

where $\hat{\mathbf{Y}^t} = G_t(\mathbf{X}^t)$. The combination of the adversarial and reconstruction loss, $\mathcal{L}_{adv} + \mathcal{L}_{rec}$, is defined as $\mathcal{L}_{shift}$ and encourages the output domain to shift toward the target distribution (see Figure 5.2e).

				FID ↓				**PCK ↑**			
Amount of data (minutes)	**Pre- trained** G_s	**Models**	**source ↓ target**	**maher ↓ oliver**	**maher ↓ chemistry**	**oliver ↓ maher**	**oliver ↓ chemistry**	**maher ↓ oliver**	**maher ↓ chemistry**	**oliver ↓ maher**	**oliver ↓ chemistry**
2	✗	**AISLe [7]**		49.2 ± 0.8	84.5 ± 3.1	83.7 ± 2.6	84.5 ± 3.1	0.18 ± 0.01	0.2 ± 0.0	0.18 ± 0.0	0.2 ± 0.0
	✓	**TGAN [179]**		57.5 ± 2.4	184.3 ± 5.4	339.1 ± 1.2	323.5 ± 1.9	0.31 ± 0.01	0.23 ± 0.0	0.2 ± 0.0	0.25 ± 0.0
	✓	**MineGAN [180]**		42.5 ± 2.5	157.5 ± 10.6	290.3 ± 7.2	302.5 ± 6.8	0.38 ± 0.03	0.26 ± 0.02	0.21 ± 0.01	**0.31 ± 0.01**
	✓	**ConsistentGAN [135]**		61.0 ± 3.2	194.2 ± 15.6	320.1 ± 11.5	325.6 ± 44.3	0.39 ± 0.01	0.27 ± 0.01	0.21 ± 0.01	0.25 ± 0.01
	✓	**DiffGAN (Ours)**		**25.0 ± 3.7**	**42.8 ± 5.1**	**47.9 ± 25.5**	**48.2 ± 15.9**	**0.45 ± 0.02**	**0.31 ± 0.01**	**0.26 ± 0.01**	**0.29 ± 0.01**
10	✗	**AISLe [7]**		47.1 ± 0.2	85.0 ± 1.9	80.3 ± 0.8	85.0 ± 1.9	0.18 ± 0.0	0.21 ± 0.0	0.19 ± 0.0	0.21 ± 0.0
	✓	**TGAN [179]**		62.9 ± 1.8	191.7 ± 1.2	341.5 ± 0.4	326.8 ± 1.6	0.3 ± 0.01	0.22 ± 0.01	0.2 ± 0.0	0.24 ± 0.0
	✓	**MineGAN [180]**		41.4 ± 3.1	145.0 ± 14.1	293.5 ± 12.7	318.2 ± 5.4	0.4 ± 0.01	0.24 ± 0.03	0.22 ± 0.02	**0.31 ± 0.01**
	✓	**ConsistentGAN [135]**		63.1 ± 1.9	188.2 ± 17.0	322.2 ± 2.4	327.0 ± 18.7	0.41 ± 0.03	**0.28 ± 0.04**	0.22 ± 0.01	0.27 ± 0.01
	✓	**DiffGAN (Ours)**		**15.0 ± 5.2**	**31.7 ± 3.8**	**30.3 ± 6.9**	**24.3 ± 4.2**	**0.46 ± 0.01**	**0.32 ± 0.02**	**0.26 ± 0.01**	**0.3 ± 0.01**
Full	✗	**AISLe [7]**		16.1	8.7	10.2	8.7	0.49	0.39	0.27	0.39

Table 5.2: Comparison of our DiffGAN with prior work for low-resource crossmodal generative modeling from source to target speakers to evaluate output domain shift (i.e. FID) and crossmodal grounding (i.e. PCK)

5.4 Experiments

Dataset: We use the PATS dataset [7, 8, 50] as the benchmark to measure performance. It consists of around 10 hours of aligned body pose, audio and transcripts for each of the 25 speakers. We choose five speakers (`oliver`, `maher`, `chemistry`, `ytch_prof` and `lec_evol`) with visually different gesture styles and diverse linguistic content for our experiments, in which `source` → `target` denotes the source and target domain. For speakers in the target domain, we simulate a low-resource setting by randomly sampling 2 or 10 minutes of data for all experiments.

Baseline Models: We compare our proposed model with a family of baselines that adapts the same source model to a target domain in a low-resource setting. (a) **TGAN** [179] finetunes all layers of source model with the low-resource target data, (b) **MineGAN** [180] projects the source latent space of the noise input onto a latent space representative of the target data, (c) **ConsistentGAN** [135] uses a cross-domain consistency loss which regularizes the tuning process and a patch discriminator which encourages different levels of realism over different image patches, and (d) **AISLe** [7] learns the target model without a pretrained source model. We also run ablation studies on two versions of our model (e) **DiffGAN w/o $\mathcal{L}_{\text{shift}}$** and (f) **DiffGAN w/o $\mathcal{L}_{\text{diff}}$**.

Human Perceptual Study: We conduct a human perceptual study on Amazon Mechanical Turk (AMT) to measure human preference towards generated animations. Given a pair of videos, one of which is from the ground truth and the other is generated by a model, the annotators have to choose either one of the videos based on the five criterion: gesture quality (**naturalness** and **expressivity**), crossmodal grounding (**timing** and **relevance**) and output domain shift (**style**) of generated gestures. The correctness of output domain shift is measured by the reflection of the true gesture style of a target speaker in the gestures generated by a target model [8]. We report the average preference % of human annotators as a score for comparison. We refer the readers to the appendix for more detailed definitions and setup.

Quantitative Metrics: (a) To measure relevance and timing of gestures with respect to spoken language we use two metrics, **Probability of Correct Keypoints (PCK)** [10, 160] where values are averaged over $\alpha = 0.1, 0.2$ as suggested in [50] and **L1 distance** between generated and ground truth gestures. To measure the distribution of the output domain we use **Fréchet Inception Distance (FID)** which is the distance between distributions of generated and ground truth poses [61]

Qualitative Visualization: To judge the quality of the generated spatio-temporal outputs, we would encourage the readers to see the supplementary video. Other than that, we qualitatively visualize three key properties of gestures [118, 140] (1) distribution, (2) velocities and (3) shapes of gestures

Implementation Details: For our pretrained source models, we use publicly available models by Ahuja et al. [7] for all experiments. We trained all the baselines with the reported hyperparameters. All our models were trained for 4000 iterations with a batch size of 32. Either 2 minutes or 10 minutes of video recordings were used as the target data. Each model was trained over three such randomly chosen target sets and quantitative metrics were averaged across these runs. We refer the readers to the supplementary materials for more implementation details.

5.5 Results and Discussion

In this section, we discuss the qualitative, quantitative and user study results from our experiments. As the output modality is spatio-temporal, we highly encourage the readers to **check out the supplementary video** to get a better idea of the quality of the generated gestures. We also report additional experimental results on source-target combinations, that could not make in the main chapter in the supplementary.

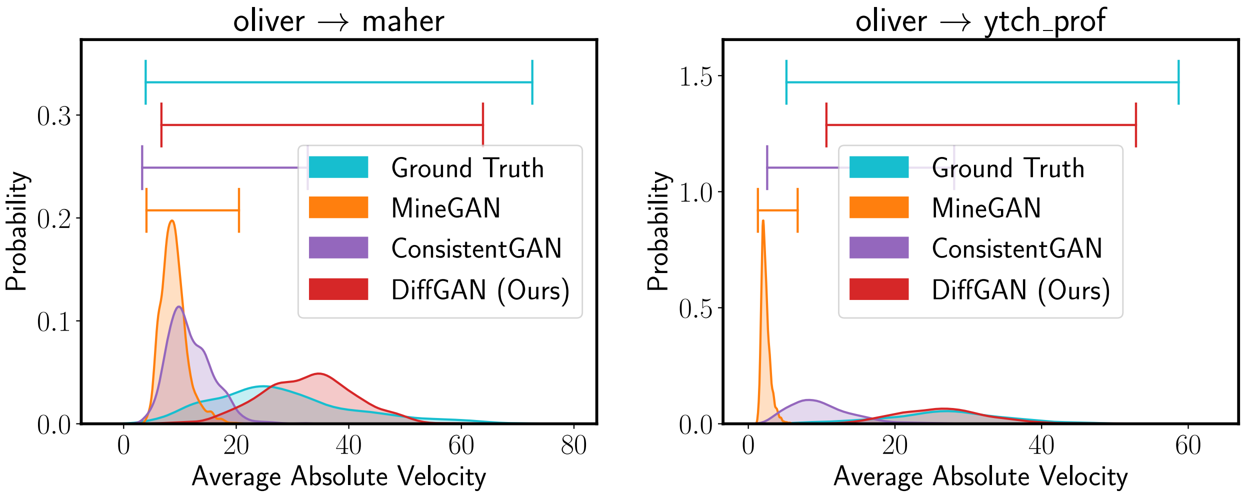


Figure 5.5: Distribution of the generated gestures with average absolute velocity as the statistic for source to target domain adaptation. The support (or coverage) of the distribution is denoted with the colour coded lines at the top of each plot. Larger overlap of a model's distribution with the ground truth distribution is desirable.

				FID ↓				L1 ↓			
Amount of data (minutes)	Pre-trained G_s	Models	source ↓ target	oliver ↓ maher	oliver ↓ chemistry	maher ↓ oliver	maher ↓ chemistry	oliver ↓ maher	oliver ↓ chemistry	maher ↓ oliver	maher ↓ chemistry
10	✓	**DiffGAN (Ours)**		**30.3 ± 6.9**	**24.3 ± 4.2**	**15.0 ± 5.2**	**31.7 ± 3.8**	1.48 ± 0.01	1.36 ± 0.03	**0.53 ± 0.02**	0.88 ± 0.03
	✓	**DiffGAN w/o $\mathcal{L}_{diff}$**		**25.9 ± 4.3**	**27.5 ± 6.7**	**19.0 ± 7.7**	**29.0 ± 1.6**	1.57 ± 0.03	1.43 ± 0.01	0.61 ± 0.02	0.96 ± 0.02
	✓	**DiffGAN w/o $\mathcal{L}_{shift}$**		119.1 ± 7.0	131.0 ± 5.1	22.4 ± 2.1	54.1 ± 1.9	**1.33 ± 0.01**	**1.26 ± 0.01**	**0.53 ± 0.01**	**0.84 ± 0.0**
Full	✗	**AISLe [7]**		10.2	8.7	16.1	8.7	0.91	0.73	0.71	0.73

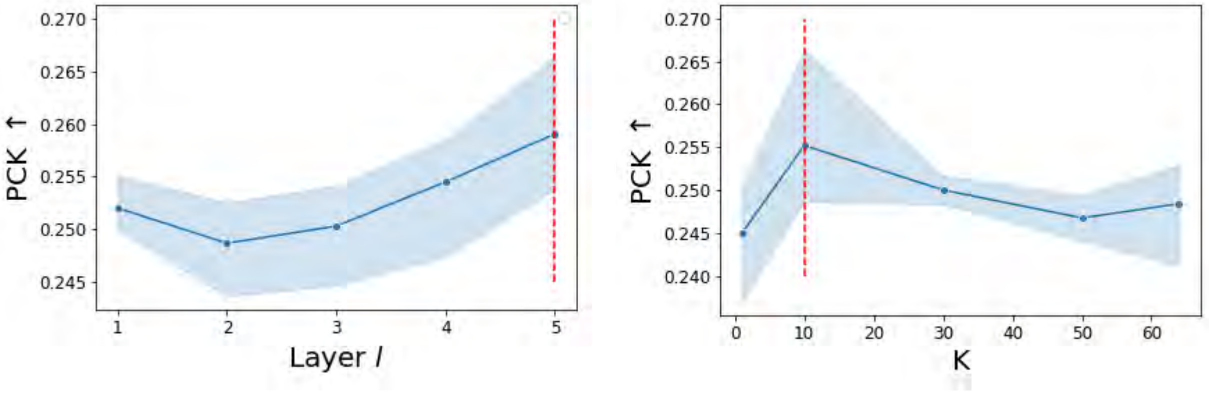
Table 5.3: **Evaluating impact of loss functions**: DiffGAN and its ablations are trained on 10 minuites of low-resource data. The metrics measure the impact of $\mathcal{L}_{shift}$ and $\mathcal{L}_{diff}$ on both output domain shift (i.e. FID) and crossmodal grounding (i.e. L1).

Comparison with prior work When evaluating generative models human judgements are often seen as a de facto evaluation [7, 185]. The results of out human perceptual study are summarized in in Table 5.1. We see a 10% larger preference, if not more, for our model DiffGAN as compared to the baseline models across all five criteria.

Crossmodal grounding shift is evaluated by measuring the relevance, timing and correctness of the generated gesture in context of spoken language. In the human perceptual study in Table 5.1, higher preference scores for DiffGAN over *relevance* and *timing* criteria are indicative of the positive impact of modeling the grounding shift explicitly which is not the case for the other baselines. Qualitatively, in Figure 5.3, we observe gesture shapes that are closer to the ground truth for DiffGAN than the other baselines, further indicating that the generated gesture shapes are more relevant to the input modality for DiffGAN. This is further corroborated by significantly higher PCK values for DiffGAN when compared with TGAN [179], MineGAN [180] and, ConsistentGAN [135] in Table 5.2.

Output domain shift (i.e. gesture style) of the generated gestures was especially convincing, as DiffGAN was preferred by human annotators (in Table 5.1) over the ground truth motion 46% of the time. A similar trend is also seen qualitatively in Figure 5.4 where we are comparing the pose histograms for both ground truth of the target speaker and the generated animations. DiffGAN is able to adapt the source model such that it generates a distribution of gestures similar to the target domain. Even though source `oliver` typically has gestures close to his body with hands moving up and down, DiffGAN is able to extrapolate to a target model for `maher` with his prototypical side to side arm movements. For TGAN [179] and MineGAN [180], the hand positions are concentrated in only one region indicating reduced diversity in the generated gestures and therefore a mode collapse. In column three of Figure 5.4, ConsistentGAN [135] is able to learn the correct rest pose for the target speakers, potentially due to its approach of encouraging distance consistency in the output domain. But the distribution of output gestures does not correctly match the distribution of true distribution of target speakers illustrated in column one of Figure 5.4. These trends are further corroborated by significantly better values of FID for DiffGAN when compared with TGAN [179], MineGAN [180] and ConsistentGAN [135] in Table 5.2.

The distribution of gesture velocities is another statistic with which we can examine the correctness of output domain shift [140]. In Figure 5.5, we find that our model DiffGAN (⊢⊣) is closely able to generate a velocity distribution that is similar to the true distribution of the target domain. However, distribution modes of MineGAN [180] and, ConsistentGAN [135] are close



(a) Choice of Layer l

(b) Choice of k in Grounding Shift

Figure 5.6: (a) Impact of choice of Layer l on crossmodal grounding shift (i.e. PCK ↑). Layer numbers are in increasing order from input to output. Layers corresponding to indices on the X-axis are defined in supplementary. (b) Impact of choice of k in grounding shift direction Ψ_k on crossmodal grounding shift (i.e. PCK ↑).

to zero indicating that the generated gestures are have very little or no movements.

Impact of $\mathcal{L}_{diff}$ on crossmodal grounding shift In Table 5.1, we observe that the removal of $\mathcal{L}_{diff}$ from DiffGAN reduces human preference for the gestures generated by the model with respect to *timing* and *relevance* criteria. We observe a similar trend in the accuracy metric L1, which significantly worsens in Table 5.3. This supports our hypothesis that optimizing $\mathcal{L}_{diff}$ can improve the discovery of new grounding relationships in a low-resource setting.

Impact of $\mathcal{L}_{shift}$ on output domain shift On the other hand, removal of $\mathcal{L}_{shift}$ reduces correctness of the style of generated gestures (i.e. output domain shift) as indicated by human preference in Table 5.1. This is most likely due to the adversarial component of $\mathcal{L}_{shift}$ which adapts the output domain with the help of gesture sequences from the target data. In parallel, FID values in Table 5.3 undergo the same effect indicating that the target model does not generate a large variety of gestures without $\mathcal{L}_{shift}$, even though the gestures are well-grounded in the input. We note here that human annotators' judgements can be strongly influenced by the naturalness of generated gestures [185]. This is a likely reason for decrease in preference of crossmodal grounding metrics for DiffGAN w/o $\mathcal{L}_{shift}$, however it is still preferred more often than other baseline models in Table 5.1.

Impact of number of training examples on model gesture style and grounding The amount of data has a large part to play in generative modeling and adaptation [163, 204]. We vary the

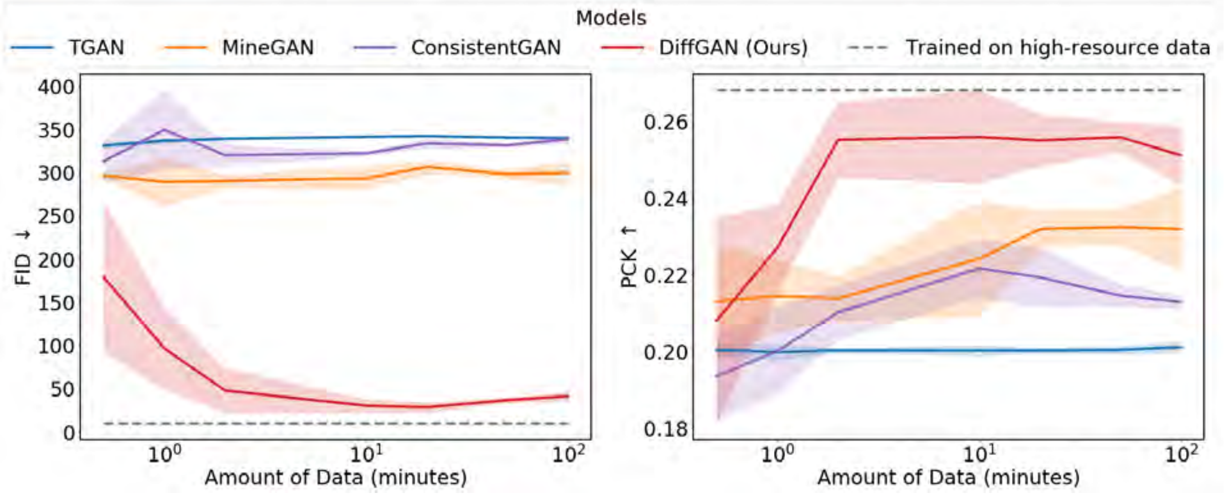


Figure 5.7: Impact of amount of data on crossmodal grounding shift (i.e. PCK ↑) and output domain shift (i.e. FID ↓) in crossmodal generative models. Note that X-Axis is logarithmic and error bars are standard deviation over three randomly sampled training sets.

amount of training data from 30 seconds to 100 minutes in Figure 5.7. We observe for our model DiffGAN, the output domain shift and crossmodal grounding adapts faster than all the baselines. The variance of these metrics decreases with increasing amount of data indicates a more stable training. Our choice of low-resource datasets is completely random.

Which weights should be updated? For a successful low-resource adaptation, a challenge is to select the best weights to update. Our DiffGAN approach requires a choice of trainable layer l and a choice of k number of channels that get updated in each iteration. Through hyperparameter tuning, we observe that layers closer to the output are typically able to model a better crossmodal grounding shift as seen in Figure 5.6a. The ideal choice of k is trickier. We want to update enough parameters to model the grounding shifts, but not so many that the model would overfit on the target data. For our choice of pretrained source models, we conduct an experiment with varying number of ks in Figure 5.6b. At $k = 0$ we see a drop in performance, probably due to the inactivity of $\mathcal{L}_{diff}$. At $k = 64$, we find the performance saturates indicating an overfitted model. We find a balance somewhere in the middle at $k = 10$.

Visualizing crossmodal grounding shift For the same input, we probe the output spaces of the source and target model in Figure 5.8. We find that distribution gestures corresponding to the input can be sparse (i.e. visually different gestures) or dense (i.e. visually similar gestures). In other words, a verbal concept such as a greeting has a single way of gesturing when the conditional output distribution is dense, but has multiple possible gestures if the conditional output distribution is sparse. As the output distribution is conditioned not only on the input but also on the speaker, we can visually observe crossmodal grounding shift in form of expansion or contraction as we traverse from the source to target output space.

Limitations and future work: While our method generates compelling results, it is not without limitations. Our choice of source models were trained on a single source domain (i.e. speaker), which may can sometimes have a smaller overlap with the target domain. This poses a trade-off between the complexity of the source model and the amount of target data. Another challenge with our approach, is that the choice of layer(s) l and k grounding shift channels can potentially change depending on the choice of the pre-trained model architecture. Hence, these hyperparameters may need some tuning.

5.6 Related Work

Language and speech for gesture generation A rule-based approach was proposed in an earlier study by Cassell et al. [27], where the behavior expression animation toolkit (BEAT) was developed to schedule behaviors, such as hand gestures, head nods and gaze. This approach was extended to utilize linguistic information from input text for decision making [94, 99, 100, 116, 189].

Rule based approaches were replaced by deep conditional neural fields [31, 32] and Hidden Markov Models for prosody-driven head motion generation [156] and body motion generation [96, 97]. These use a dictionary of predefined animations, limiting the diversity of generated gestures. Soon, neural network based models were introduced, using unimodal inputs, specifically speech, to generate a sequence of gestures [59], head motions [153] and body motions [6, 8, 45, 50, 159]. On the other hand, Yoon et al. [195] uses only a text input for gesture generation. More recently, multimodal models relying both speech and language were developed. Kucherenko et al. [89] uses early fusion to combine the two representations, Ahuja et al. [7] utilizes a cross-modal attention mechanism to account for correlations between speech and language. These approaches typically require many hours of multimodal data to train a single speaker-specific model. We propose an approach that can adapt a single-speaker generative model to a new speaker domain after being exposed to only a few minutes of data.

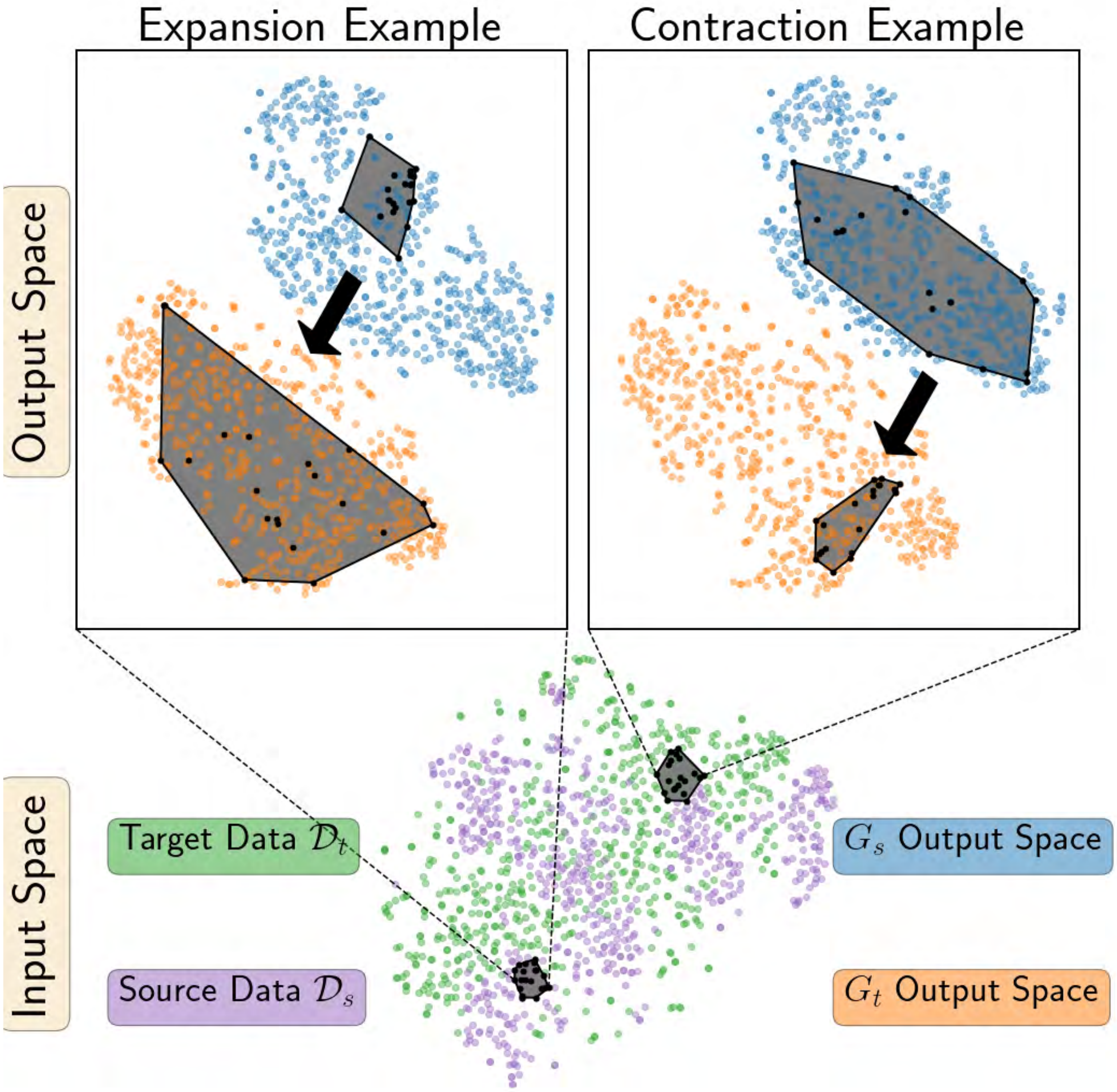


Figure 5.8: At the bottom, we display a t-SNE [115] plot of the input space for both the target and source data. We choose a region which contains both source and target input samples. At the top, we display the t-SNE plots corresponding to where these samples map to in the output space (indicated in black). On top left, similar inputs produces a compact source output space (i.e. visually similar gestures) but a sparse target output space (i.e. visually different gestures). On top right, the opposite effect is observed. These contractions and expansions of the output space conditioned on the input space represent crossmodal grounding shift.

Low-resource adaptation of generative models Prior works similary focus on pre-training a source model on large source data, then adapting it to a low-resource setting. Nichol et al. [132], Wang et al. [180] introduce new parameters in the model, whereas Karras et al. [81], Wang et al. [179] fine-tunes the complete model on the target data or to specific layers or modules are applied [123, 133]. Li et al. [102], Wang et al. [180] utilize importance sampling to transform the original latent space of the source to a space which is more relevant to the target. While this approach can be effective when the source distribution and the target distributions share support, it may not be well-generalizable when their supports are disjoint. To address this concern, Ojha et al. [135] introduces a contrastive learning approach to preserve the similarities and differences in the source, and then adapting to the target domain. These methods focus on adapting only the output domain of unimodal generative models (i.e. generate one modality with noise or a small set of discrete classes as the input). However, we believe that for crossmodal generative modeling tasks, we need to explicitly model complex relationships between the input modalities and the generated output modality, both of which have a spatial and/or temporal structure.

Chapter 6

Co-Speech Gestures: Low Resource Continual Learning for Grounding and Gesture Style

Advancements in verbal communicative technologies such as speech-based intelligent personal assistants (e.g. Alexa, Siri, Cortana) have naturally segued into nonverbal embodiment [119] that complements verbal communication. To facilitate this vision of a combined communicative entity of verbal and nonverbal communication, a technical challenge is to generate plausible co-speech gestures that are grounded in the verbal signals (i.e. language and speech) which we introduce in Chapter 3. This challenge was pushed further in Chapter 4 with an even more challenging task of generation of personalized visual gestures, which reflects the idiosyncratic behaviors of a specific person. Typically, a significant amount of parallel data for each speaker (i.e. 5-10 hours of speech, language, and gesture data) is required to learn the grounding relationships between verbal and nonverbal signals and learn the personalization of visual gestures. While a supervised approach is effective in tackling the grounding and personalization challenges, it is not trivial to gather 10 hours of data for each speaker in real-world situations. Furthermore, even if 10 hours of data is available, training a new model from scratch for each speaker necessitates long compute times ($\sim$5 hours). In Chapter 5, these challenges are tackled by creating a personalized gesture generation model with 2 minutes of parallel data of a new speaker which additionally absorbs a significant chunk of computational resource burdens by cutting the training time by five times. But in the process of learning a new speaker's gestures, the model completely forgets the grounding and personalization of the original speakers'. This can be quite limiting for an

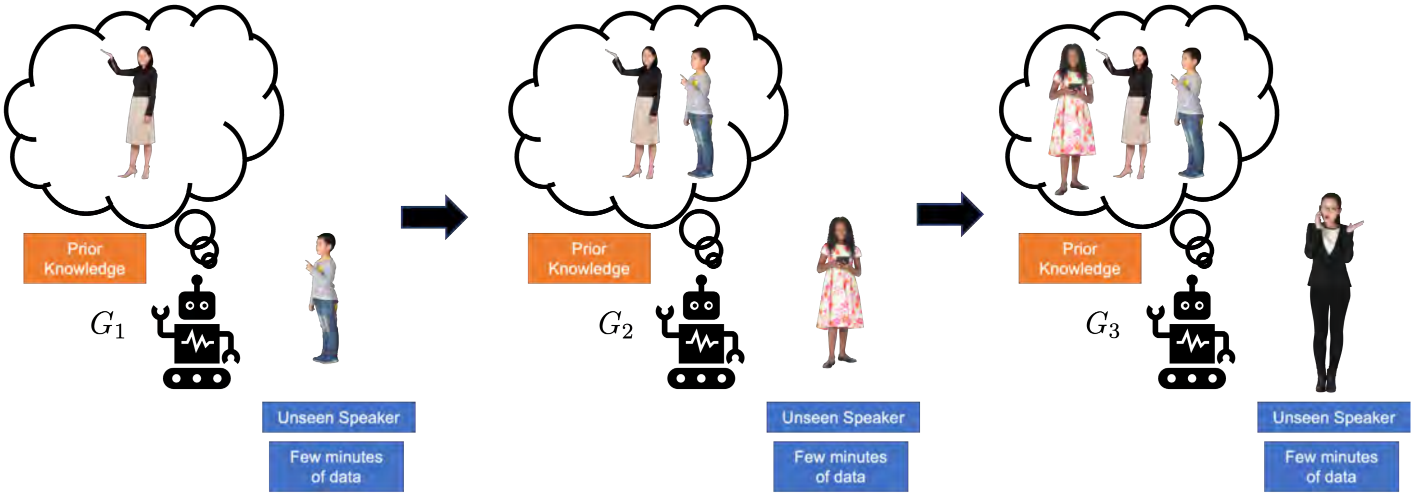


Figure 6.1: Overview of the continual learning paradigm of co-speech gesture personalization task. We start with a source model G_1 pre-trained on a source speaker. We personalize G_1 to target models G_2, G_3 and so on in a sequential manner using low-resource data ($\sim$2 minutes) for each of the target speaker.

intelligent virtual agent [24, 103, 136, 137] that intends to continually learn nonverbal behavior generation as it interacts with many different speakers at different points in time. A useful property of this virtual agent would be to learn a new speaker's behavior while not forgetting the behaviors of the older speakers. As illustrated in Figure 6.1, the main goal of this chapter is to create a gesture generation model which has the ability to learn to generate co-speech gestures of many different speakers as it sequentially experiences a limited amount (i.e. 2 minutes) of these nonverbal behaviors over time.

6.1 Overview

This problem setting brings a unique technical challenge, typically not studied in continual learning for generative modeling settings: *crossmodal catastrophic forgetting*. Due to the crossmodal nature of our task, crossmodal catastrophic forgetting refers to the forgetting of the crossmodal grounding relationships between input spoken language modalities and output gesture modality of speakers that the model interacted with earlier in time. For example, consider a virtual agent that has the knowledge of generating gestures of one speaker. As it starts to interact with the new speakers in the world, it experiences new crossmodal grounding relationships between gestures and spoken language. While the gestures are heavily dependent on the spoken language, they are also heavily influenced by the new speakers' idiosyncrasies. A practical challenge here is that these interactions are often short, hence creating a low-resource setting for this agent. Another challenge is that all the data is not available at once. Instead, the agent sequentially receives new data as it interacts with multiple new speakers over time. The goal of this virtual agent is to learn to generate personalized gestures for many different speakers without forgetting the crossmodal grounding of the speakers that it interacted with earlier in its life. The agent should be able to achieve these goals with the practical constraints of low-resource data, limited storage space, and faster training.

In this chapter, we propose an approach, named C-DiffGAN, that can efficiently personalize co-speech gesture generation models from a high-resource source speaker to multiple low-resource target speakers. To the best of our knowledge, this is the first approach that is able to learn a personalized model for multiple speakers with only 2 minutes each of speaker data (i.e. as opposed to 10 hours [7, 50, 60, 89]) in a continual learning setting. Our C-DiffGAN approach requires access to only 2 minutes of the input data (i.e. language and speech) for the prior speakers and 2 minutes of paired data (i.e. language, speech, and gestures) for the new speaker. For continually learning new speakers' behaviors while not forgetting the prior speakers', C-DiffGAN follows two steps: First, it directly identifies shifts in crossmodal grounding relationships along with the shifts in the output domain from the pretrained source model. Based on these identified distribution shifts, C-DiffGAN updates a few necessary parameters in a single layer of the source model, allowing efficient adaptation with low resources. Second, it utilizes the low-resource input data of prior speakers in tandem to prevent the model from drifting from the prior speakers' crossmodal grounding, hence preventing crossmodal catastrophic forgetting. Our experiments study the effectiveness of our C-DiffGAN approach on a diverse publicly available dataset and is substantiated through a myriad of quantitative and qualitative studies, which show that our proposed methodology significantly outperforms prior approaches for low resource continual learning of nonverbal grounding and personalization of gesture generation models.

6.2 Problem Statement

We are given a set of training datasets for each speaker (or experiences) $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_M\}$. Here $\mathcal{S}_j = \{j, \mathbf{X}_i^j, \mathbf{Y}_i^j\}_{i=1}^{Nj}$ is paired data of input language and speech $\mathbf{X}_i^j$ and output sequence of body poses $\mathbf{Y}_i^j$ for speaker j. A goal here is to learn a sequence of gesture generation models, represented as Generator-Discriminator pairs [52], that are adapted through

$(G_1, D_1) \rightarrow (G_2, D_2) \rightarrow \ldots \rightarrow (G_M, D_M)$ by sequentially training on speakers 1 through M in a continual learning setting. Here G_j is a model that is able to generate personalized gestures of speakers 1 through j that are driven by both language and speech. D_j is a model that can distinguish between real and fake sequence of gestures for the jth experience. In addition to sequential training, the number of training samples N^j is often small in practical scenarios which emulates a low-resource continual learning setting. In this setup, only the final version of the model G_M is deployed.

6.3 Method

We propose a new approach, Continual-DiffGAN (or C-DiffGAN), that learns a target model G_j for speaker (or experience) j while not forgetting the crossmodal grounding knowledge of speakers 1 through j in the low-resource adaptation from the source model G_{j-1}. This approach is sequential and applies for speakers 1 through M. C-DiffGAN has three components: (1) *Generative Modeling* (Section 6.3.3) which learns to generate gestures that are driven by input spoken language along with personalizing to multiple speakers through a loss function $\mathcal{L}_{gen}$, (2) *Crossmodal Adaptation* (Section 6.3.2) which learns to adapt from a source model to a target model through a loss function $\mathcal{L}_{diffgan}$, and (3) *Crosmodal Catastrophic Forgetting* (Section 6.3.1) that prevents the new target model G_j from catastrophically forgetting the crossmodal grounding knowledge earlier speakers (i.e. 1 through $j-1$) through a loss function $\mathcal{L}_{ccf}$. Optimization of the combined loss function describes the complete model,

$$G_j^* = \mathbb{E}_{\mathcal{S}_j^{exp}} \underset{G_{j-1}}{\text{argmin}} \max_{D_{j-1}} \mathcal{L}_{gen} + \mathcal{L}_{diffgan} + \mathcal{L}_{ccf} \tag{6.1}$$

where $\mathcal{S}_j^{exp}$ is a low-resource training dataset for the jth experience.

6.3.1 Crossmodal Catastrophic Forgetting

A key technical challenge in continual learning paradigms is the phenomenon of catastrophic forgetting. Neural networks typically forget previously learnt domain knowledge when they are fine-tuned on new experiences [149]. This challenge becomes even more complex as our task is generative crossmodal, which means that both input and output modalities are part of a large and continuous representation space. To tackle this *crossmodal catastrophic forgetting* challenge, an approach is to fine-tune the source model with previous experiences' training data [147] along with the new experience's training data. But this approach is not scalable, as the memory and computational footprint of the continual learning models will linearly increase with the number of experiences. To mitigate the scalability challenge along with *crossmodal catastrophic forgetting*, we propose a two step approach: (1) Reminiscence and (2) Crossmodal Alignment.

Low-resource Reminiscences We leverage the source model G_{j-1} to create an extended dataset $\mathcal{S}_j^{exp} = \mathcal{S}_j \bigcup_{e=1}^{j-1} \tilde{\mathcal{S}}_e$ that consists both real training data of the current experience as well as memory reminiscences from the previous experiences. The set $\tilde{\mathcal{S}}_e$ for a given experience e is constructed using just the input examples of all previous experiences and the generated gestures by source model G_{j-1} as $\bigcup_{e=1}^{j-1} \bigcup_{i=1}^{N_e} \{e, \mathbf{X}_i^e, G_{j-1}(\mathbf{X}_i^e, j)\}$. Due to the low-resource and crossmodal nature of this task N_j is typically quite small ranging from 2-10 minutes of language and speech inputs. This makes the utilization of memory reminiscences especially challenging when compared to memory replays [187] where unlimited amount of replay data can be generated due to the unimodal nature of the tasks. With the constructed memory reminiscence, the target model G_j can be trained by optimizing the overall loss function, defined in Equation 6.1, on the source model G_{j-1}.

Crossmodal Alignment While learning the target model G_j, the crossmodal relationships between spoken language gestures for speakers 1 through $j-1$ are catastrophically forgotten. We propose an approach where the target model is encouraged to remember this knowledge by explicitly through the loss function $\mathcal{L}_{ccf}$ defined as follows:

$$\mathcal{L}_{ccf} = \mathbb{E}_{k, \tilde{\mathbf{X}}, \tilde{\mathbf{Y}} \in \bigcup_{e=1}^{j-1} \tilde{\mathcal{S}}_e} \|\tilde{\mathbf{Y}} - G_j(\tilde{\mathbf{X}}, k)\|_2 \tag{6.2}$$

6.3.2 Low-resource Crossmodal Adaptation

A practical challenge in the continual learning paradigm is availability of training data for new experiences, often making it a low-resource paradigm. Learning a crossmodal generative model is quite challenging in a low-resource setting, but it is achievable if we learn a target model G_2 for the new experience by drawing on from the knowledge of a source model G_1 that is well trained on high-resource data. We use a two step approach for low-resource crossmodal adaptation as discussed in Chapter 5.3 [9]. First, the model learns to identify the crossmodal grounding shifts through a loss function $\mathcal{L}_{diff}$ and low-resource target data. Second, the target model is encouraged to shift the output domain distribution to be closer to that of the targets' through the use of a loss function $\mathcal{L}_{shift}$. The combined loss function $\mathcal{L}_{diffgan} = \mathcal{L}_{diff} + \mathcal{L}_{shift}$ encourages low-resource crossmodal adaptation of our C-DiffGAN model.

This low-resource crossmodal adaptation in tandem with the mitigation of crossmodal catastrophic forgetting allows us to learn a well-trained target model G_2 that is able to generate gestures of speakers from the first two experiences. Now, the role of G_2 switches to a well-trained source model which can now be used to learn the new target model G_3 with a new experience, continuing the training cycle.

6.3.3 Generative Modeling

The final challenge is to generate plausible gestures that correspond to the input spoken language for multiple speakers. As a first step we encourage the model to generate correct gestures via a reconstruction loss for every experience j,

$$\mathcal{L}_{rec} = \mathbb{E}_{e,\mathbf{X},\mathbf{Y} \in \mathcal{S}_j^{exp}} \|\mathbf{Y} - G_{j-1}(\mathbf{X}, e)\|_1 \tag{6.3}$$

To alleviate the challenge of overly smooth generation caused by L1 reconstruction in Equation 6.3, we use the generated pose sequence $\hat{\mathbf{Y}} = G_{j-1}(\mathbf{X}, j)$ as a signal for the adversarial discriminator D_{j-1}. The discriminator tries to classify the true pose $\mathbf{Y}$ from the generated pose $\hat{\mathbf{Y}}$, while the generator G_{j-1} is encouraged to fool the discriminator by generating realistic poses. The adversarial loss is written as [52],

$$\mathcal{L}_{adv} = \mathbb{E}_{e,\mathbf{X},\mathbf{Y} \in \mathcal{S}_j^{exp}} \log D_{j-1}\left(\mathbf{Y}\right) + \log\left(1 - D_{j-1}\left(G_{j-1}\left(\mathbf{X}, e\right)\right)\right). \tag{6.4}$$

Our task requires the model to generate gestures in multiple speakers styles, but often the gesture styles of the gesture distribution is different across different speakers. As networks learnt with adversarial losses are prone to mode collapse [11], we explicitly learn multiple subgenerators as part of the main generator G_j also called Mix-GAN in Chapter 4 [8]. Each of these generators learn different modes of the distribution and together are able to represent multiple speaker gesture styles. To train the multiple generators of the Mix-GAN we also optimize for a loss $\mathcal{L}_{mix}$, defined in Equation 4.3, which encourages the spoken language encoder to choose the correct set of sub-generators at inference time. Another challenge with multiple speakers is that the spoken language data for each speaker is likely to be different and in some cases completely non-overlapping. To prevent the gestures dependence on the content of the spoken language, a goal here is to disentangle the gesture style from the content of the gesture. To achieve style disentanglement we use the style consistency loss function discussed in Chapter 4.3.3,

$$\mathcal{L}_{id} = \mathbb{E}_{Y \in \{\mathbf{Y}, \hat{\mathbf{Y}}\}} \text{CCE} \left(\text{Softmax} \left(P_{enc}(Y) \right), e \right), \tag{6.5}$$

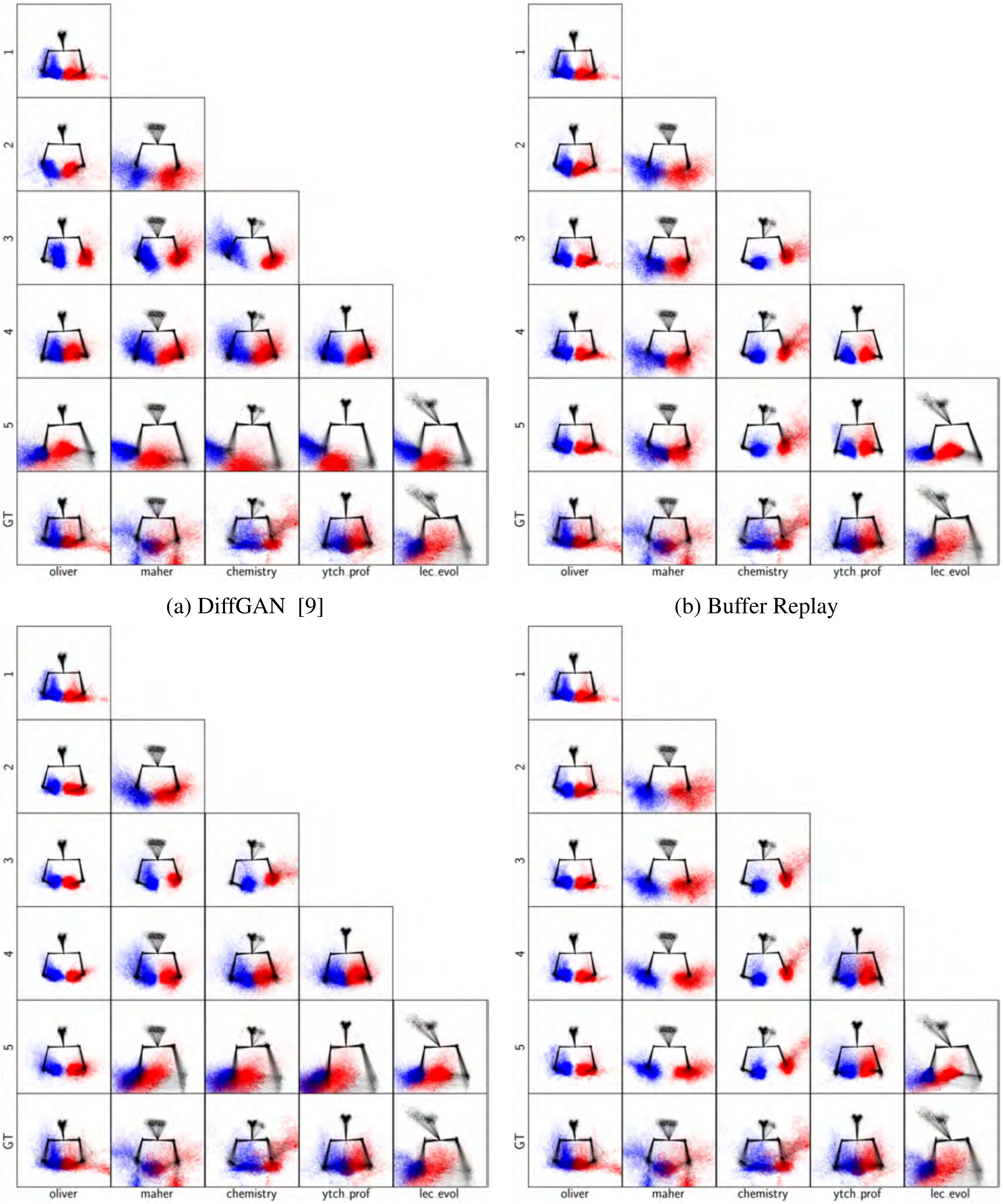
where P_{enc} is a gesture style encoder which is encouraged to generate features representing gesture style. CCE is a categorical cross-entropy loss optimizing which maintains consistency between in the true and the generated gestures.

The combination of the loss functions in this section is defined as the generative modeling loss $\mathcal{L}_{gen} = \mathcal{L}_{rec} + \mathcal{L}_{adv} + \mathcal{L}_{mix} + \mathcal{L}_{id}$, which is trained together with the crossmodal adaptation and crossmodal catastrophic forgetting losses in Equation 6.1.

6.4 Experiments

Dataset: We use the PATS dataset introduced in Chapter 3 [7, 8, 50] as the benchmark to measure performance. It consists of around 10 hours of aligned body pose, audio and transcripts for each of the 25 speakers. For all experiments we use a sequence of five randomly chosen speakers (`oliver`, `maher`, `chemistry`, `ytch_prof` and `lec_evol`) that have visually different gesture styles and diverse linguistic content for our experiments. Unless specified otherwise, we start with a model described in Chapter 4 trained on a high-resource dataset of the speaker `oliver`. For speakers in the subsequent experiences, we limit the training data to 2 or 10 minutes of to simulate a low-resource continual learning setting.

Baseline Models: To the best of our knowledge this crossmodal continual low-resource generative modeling task has not been explored before, hence there are no baselines that directly associated with this task. We use a family of strong baselines most relevant to the challenges posed by this task: (a) **DiffGAN** [9] performs adaptation from a high-resource trained source model to a target model in a low-resource co-speech gesture generation setting. (b) **MeRGAN-JTR** and **MeRGAN-RA** [187] performs continual learning (or CL) for unimodal generative modeling in a high-resource setting without the need to explicitly store training examples of the previous experiences. We modify it to work in our crossmodal task in a low-resource setting. (c) **Buffer Replay** explicitly saves training examples from the previous experiences in a buffer, which becomes a part of the subsequent training cycles. This strong baseline requires extra stor-



(a) DiffGAN [9]

(b) Buffer Replay

(c) MeRGAN-RA [187]

(d) C-DiffGAN (Ours)

Figure 6.2: Visual Histograms of generated gestures visually describe the distribution of hand gestures in space. Red and blue colors denote the left and right arms respectively. Each row represents the gesture distribution at the end of each continual learning experience. The final row denotes the true distribution of gestures for each speaker. The columns represent the speakers in the order they were exposed to the model in the continual learning paradigm. As we go from top to bottom in each column, we can see how the distribution of each speaker changes over the number of experiences. In Fig. (a) the baseline model DiffGAN [9] forgets the distribution of gestures for all the older speakers, while just retaining the information of the newest speaker. On the other hand, for our C-DiffGAN model in Fig. (d), we see that the model remembers the distribution of gestures for all the speakers through all the training experiences

age memory and training time making it less scalable. (d) **MixStAGe** [8] learns a common model for multiple speaker styles by jointly training (or JT) for multiple speakers in a high-resource setting. We also run ablation studies on two versions of our model: (e) **C-DiffGAN w/o** $\mathcal{L}_{\textbf{diffgan}}$ and (f) **C-DiffGAN w/ K starting speakers** where the first source model G_1 is jointly trained on K speakers as a first experience followed by continual learning over the rest of the speakers.

Quantitative Metrics: In a continual learning setting we typically measure two performance criteria [187]. (1) **Average Final Accuracy** measures the average metrics of the final model over the examples of all the experiences and is defined as

$$\text{Average-Final-Accuracy}(R, M) = \frac{1}{M}\sum_{j=1}^{M} R_{M,j}, \tag{6.6}$$

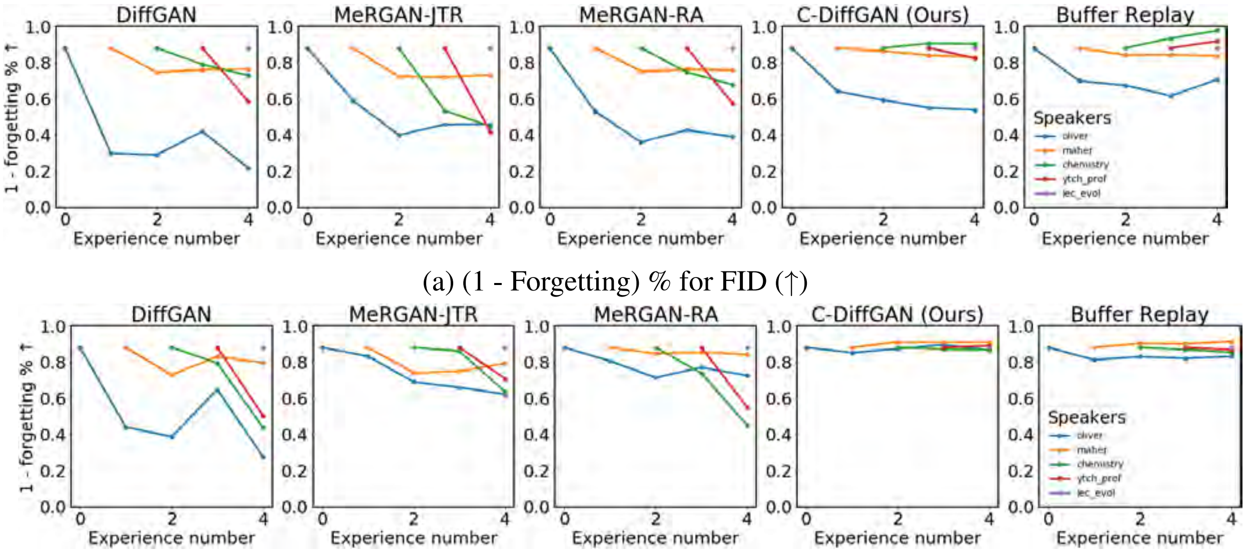
where $R_{M,j}$ is a metric R measured on a model at experience M on data from experience j, and M is the total number of experiences. (2) **Average Forgetting** measures the extent to which the final model has forgotten about the prior experiences and is defined as

$$\text{Average-Forgetting}(R, M) = \frac{\delta}{M-1}\sum_{j=1}^{M-1} \max_{j \le e \le M-1} R_{e,j} - R_{M,j}, \tag{6.7}$$

where $\delta = +1$ if a higher value of metric R denotes better performance and $\delta = -1$ if a lower value of metric R denotes better performance. A model is said to perform better in a continual learning setting when the Average Forgetting is lower.

We use two metrics that are useful to measure performance for co-speech gesture generation tasks: (a) **Probability of Correct Keypoints (PCK)** [10, 160] measures relevance and timing of gestures with respect to spoken language. Here the PCK values are averaged over $\alpha = 0.1, 0.2$ as suggested in [50]. (2) **Fréchet Inception Distance (FID)** is the distance between distributions of generated and ground truth poses [7, 61] which is used to measure the diversity in the generated gestures.

Implementation Details: For our pretrained source models, we use publicly available models by Ahuja et al. [7] for all experiments. We trained all the baselines with the reported hyper-parameters. All our models were trained for 4000 iterations with a batch size of 32. Either 2 minutes or 10 minutes of video recordings were used as the target data. To alleviate sample bias,



(a) (1 - Forgetting) % for FID (↑)

(b) (1 - Forgetting) % for PCK (↑)

Figure 6.3: Comparing our C-DiffGAN with baselines on the measure of forgetting across number of experiences with 2 minutes of training data for each speaker. We plot (1-Forgetting)% for both FID and PCK for all speakers. Hence higher is better. The sudden dips of the measures for the older speakers indicate catastrophic forgetting and can be observed cleary in DiffGAN [7], MeRGAN-JTR [187] and MeRGAN-RA [187]. C-DiffGAN, on the other hand, is able to retain the performance over all the 5 experiences reasonably well.

each model was trained over three such randomly chosen target sets and quantitative metrics were averaged across these runs. We refer the readers to the supplementary materials for more implementation details.

6.5 Results and Discussion

In this section, we discuss the qualitative and quantitative results from our experiments.

Comparison with prior work When evaluating models trained in a continual learning paradigm, we compare the performance of our proposed C-DiffGAN with the other baseline models using two criteria, *Average Final Accuracy* and *Average Forgetting*.

Average Final Accuracy is a measure of the average performance of the final model G_M over all 1 through M experiences. A goal of this task is to be able to consistently generate relevant and diverse gestures corresponding to the input spoken language for all speakers the model was exposed to. The significantly better PCK and FID *Average Final Accuracy* in Table 6.1 for our C-

DiffGAN is indicative of the positive impact of crossmodal adaptation in learning new speaker personalizations from new experiences in a continual learning setting. Our C-DiffGAN only requires 2-10 minutes of low-resource data with no buffer memory hence making this approach scalable to a larger number of experiences. This trend is qualitatively corroborated in the visual pose histograms of the last two rows of Figure 6.2d. Here, C-DiffGAN after 5 experiences is able to recreate the true distribution of the body poses for all 5 speakers.

Average Forgetting is a measure of the knowledge that the model forgets after experiencing new speakers' data. A goal of this task is to be able to retain the knowledge of generating both relevant and diverse gestures corresponding to the input spoken language, especially for speakers that were seen earlier in the training cycle. The significantly better *Average Forgetting* for both PCK and FID in Table 6.1 for our C-DiffGAN is indicative of the positive impact of low-resource reminiscences in retaining old speakers personlizations even though the model is exposed to new experiences as the training progresses. Our C-DiffGAN does not require any buffer memory, instead it utilizes the previous experience's model G_{j-1} to generate reminiscences which is useful for retaining older knowledge, while also learning a new speaker's personalized gesture model. In Figure 6.4 and 6.3, we observe that DiffGAN [9], MerGAN-JTR and MerGAN-RA [187] forgets the crossmodal grounding relationships (i.e. PCK) and output distribution (i.e. FID) by significant amounts over 1-2 new experiences. This is unlike our C-DiffGAN model which is completely able to retain the output distribution information in Figure 6.4a and 6.3a. Crossmodal grounding is also retained to a high degree in Figure 6.4b and 6.3b. The amount of knowledge lost could potentially be attributed to crossmodal grounding being a harder challenge and could benefit from more examples. This trend is qualitatively corroborated in the visual pose histograms in each column of Figure 6.2d. Here, C-DiffGAN is able to recreate the true distribution of the body poses for all 5 speakers through all the training experiences. In contrast, DiffGAN [9] and MeRGAN-RA [187] are not able to recreate the distribution of the body poses of the older speakers as the training experiences progresses. Instead the distribution of the body poses of the older speakers resembles the newest speaker indicating that the knowledge about the previous experiences gets overwritten by the new speaker.

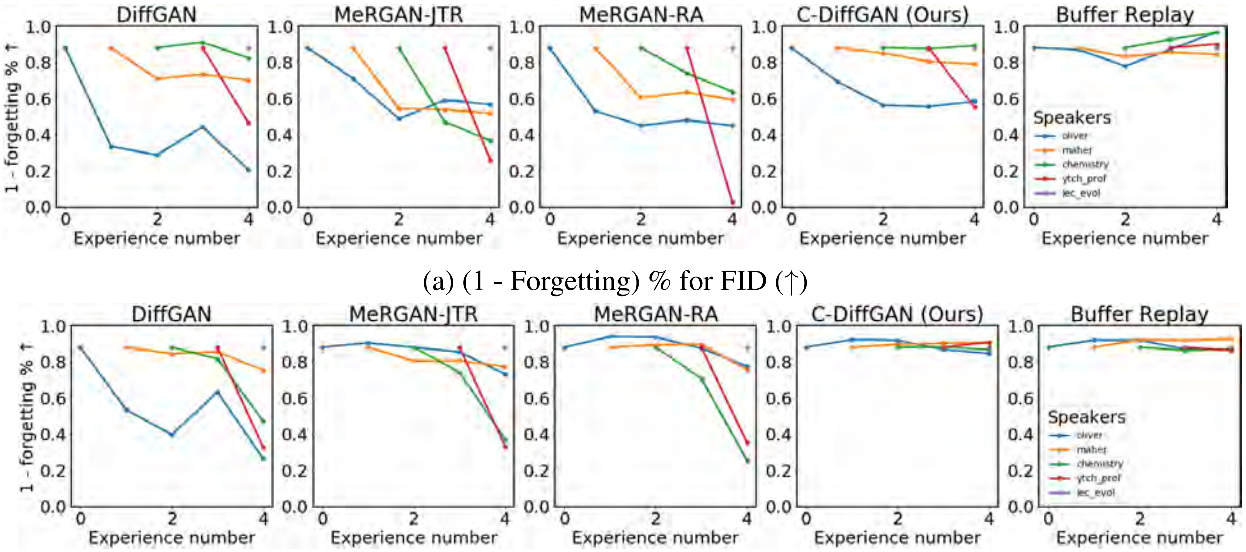
Impact of Crossmodal Adaptation $\mathcal{L}_{diffgan}$ on Average Accuracy and Forgetting The crossmodal adaptation loss $\mathcal{L}_{diffgan}$ positively impacts both crossmodal grounding shift and output domain shift for low-resource adaptation as discussed in Chapter 5 [9]. In table 6.2, we observe that learning without $\mathcal{L}_{diffgan}$ significantly degrades both crossmodal grounding (PCK) and out-

Amount of Data (minutes)	Training	Buffer Memory	Models	Average Final Accuracy		Average Forgetting	
				FID↓	PCK↑	FID↓	PCK↓
2	CL	✗	DiffGAN [9]	350.3	0.20	343.3	0.16
	CL	✗	MeRGAN-JTR [187]	171.9	0.27	158.6	0.08
	CL	✗	MeRGAN-RA [187]	309.1	0.25	271.8	0.10
	CL	✗	C-DiffGAN (Ours)	**114.9**	**0.35**	**16.2**	**0.00**
	CL	✓	Buffer Replay	90.5	0.35	0.6	0.01
10	CL	✗	DiffGAN [9]	613.6	0.16	674.5	0.18
	CL	✗	MeRGAN-JTR [187]	316.9	0.24	355.0	0.13
	CL	✗	MeRGAN-RA [187]	494.1	0.23	561.1	0.15
	CL	✗	C-DiffGAN (Ours)	**55.6**	**0.35**	**12.7**	**0.01**
	CL	✓	Buffer Replay	61.6	0.37	-2.2	0.01
Full	JT	✓	MixStAGe [8]	22.0	0.40	5.6	0.01

Table 6.1: Comparison of our C-DiffGAN with prior work for low-resource continual learning (CL) and joint training (JT) for crossmodal generative modeling. We use the Average Final Accuracy and Average Forgetting as the continual learning metrics for FID and PCK. Buffer Memory indicates if the method requires additional storage memory

Amount of Data (minutes)	Models	Average Final Accuracy		Average Forgetting	
		FID↓	PCK↑	FID↓	PCK↓
10	C-DiffGAN w/o $\mathcal{L}_{diffgan}$	70.8	0.34	14.0	**0.01**
	C-DiffGAN (Ours)	**55.6**	**0.35**	**12.7**	**0.01**
Full	MixStAGe [8]	22.0	0.40	5.6	**0.01**

Table 6.2: Comparison of our C-DiffGAN with an ablation C-DiffGAN w/o $\mathcal{L}_{diffgan}$. We use the Average Final Accuracy and Average Forgetting as the continual learning metrics for FID and PCK.



(a) (1 - Forgetting) % for FID (↑)

(b) (1 - Forgetting) % for PCK (↑)

Figure 6.4: Comparing our C-DiffGAN with baselines on the measure of forgetting across number of experiences with 10 minutes of training data for each speaker. We plot (1-Forgetting)% for both FID and PCK for all speakers. Hence higher is better. The sudden dips of the measures for the older speakers indicate catastrophic forgetting and can be observed cleary in DiffGAN [7], MeRGAN-JTR [187] and MeRGAN-RA [187]. C-DiffGAN, on the other hand, is able to retain the performance over all the 5 experiences reasonably well.

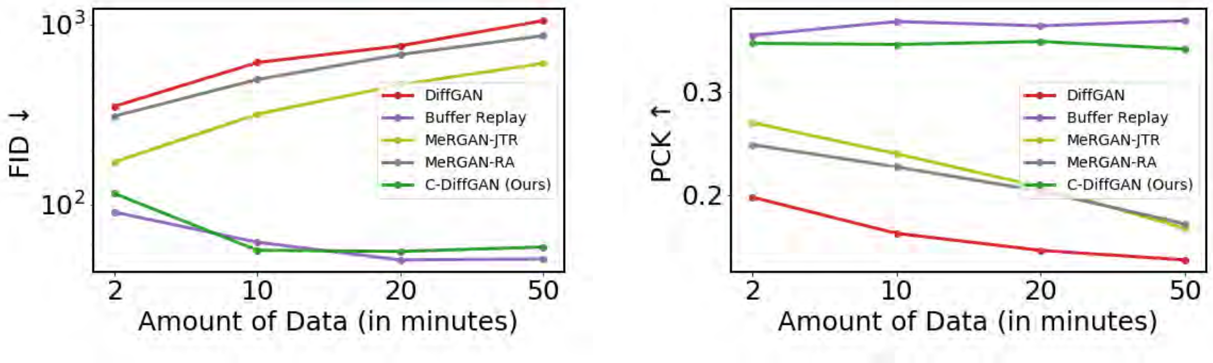


Figure 6.5: Trends of Average Final Accuracy (FID and PCK) for our C-DiffGAN when compared to other baselines. Lower is better for FID and higher is better for PCK.

Amount of Data (minutes)	**Models**	**Starting Speakers**	**Average Final Accuracy**		**Average Forgetting**	
			FID↓	PCK↑	FID↓	PCK↓
2	**C-DiffGAN (Ours)**	**1**	114.9	0.35	16.2	**0.00**
		2	98.6	0.32	28.1	**0.01**
		3	**36.5**	**0.35**	**15.2**	**0.01**

Table 6.3: Comparison of our C-DiffGAN with its ablations on number of starting speakers. We train 3 initial models with 1, 2, and 3 speakers respectively. With these as the source models, we train them on new experiences in a continual learning manner as usual. We observe that an initial model with a more diverse knowledge is better suited to learn better models through the help of future experiences.

put domain (FID) indicating that its impact is complimentary to the challenge of Crossmodal Catastrophic forgetting.

Impact of number of training examples on model gesture style and grounding As the amount of data increases, we observe in Figure 6.5 that our C-DiffGAN is able to model the output domain (FID) significantly better before plateauing. The crossmodal grounding (PCK) is remains fairly stable even with an increase in training data. In contrast, we observe an opposite effect for DiffGAN [9], MeRGAN-JTR and MeRGAN-RA [187] where the modeling ability of output domain (FID) and crossmodal grounding (PCK) consistently gets worse. This is indicative of these baselines easily learning to personalize to new speakers and just as easily forgetting old speakers. Adding more examples to the training experiences only speeds up this process.

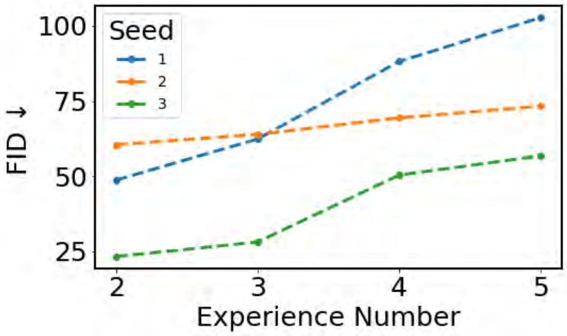


Figure 6.6: Measuring variability of FID scores across experiences and for three different seed values for the low-resource training data of `oliver` in our C-DiffGAN. The choice of the low resource training dataset can potentially have an impact on the distribution of the generated gestures.

Impact of number of speakers in the first Source Model G_1 We experimented with different number of speakers in the first source model G_1. As observed in Table 6.3, if our source model is trained with a larger number of speakers it has a positive impact on both Average Final Accuracy and Average Forgetting. This is in accordance with the findings in Chapter 4 [8] where increasing the number of speakers consistently improves model performance over all speakers. Here, a larger diverisity of knowledge from different speakers starts the continual learning process off with a well trained generator-discriminator pair which is likely a reason for the performance boost.

Reminiscence vs Buffer Memory vs Joint Training trade-off For most of our experiments we avoid using a *Buffer Memory*, which is not scalable in long-term continual learning settings. Instead we use *Reminiscence* to reconstruct some of the data for older experiences, bypassing the need for extra storage and compute. While reminiscence along with $\mathcal{L}_{ccf}$ is quite successful in preventing catastrophic forgetting as observed in our discussions, explicitly storing examples in the buffer can potentially boost performance as seen in Table 6.1. In the extreme case of storing all available examples of all the speakers in the buffer (i.e. Joint Training for MixStAGe [8] in Table 6.1) can further boost the performance. The trade-off here is the need for extra storage, computational resources and training time for better performance. As there is no one correct solution, we advise the readers to be aware of the trade-off and choose the methodology that fits better in their scenario.

Impact of choice of low-resource training data The choice of training samples in the low-resource data can potentially impact performance as seen in Figure 6.6. It is interesting to note that the model FID scores have a significant variance from 25 to 100 through the continual learning cycle. As we are working in a low-resource setting, the choice of sample points becomes even more crucial. If we are not careful, we might end up ignoring samples from one or more critical regions of the gesture distribution. While the choice of training samples is not in the scope of this chapter, but we would like our readers to be aware of this factor when dealing with crossmodal generative modeling in low-resource continual learning settings.

Chapter 7

Conclusions

In this thesis we studied the core technical challenge of learning a two way relationship between visual body motion and language, aka nonverbal grounding. We developed methods for the cross-modal translation tasks of generating body motion or co-speech gestures that plausibly co-occur with a given language and/or speech signal. We further study the challenge of grounded gesture generation in context of personalized co-speech gestures by proposing an individual gesture style transfer task. By leveraging the commonalities amongst multiple individuals we were also able to improve the grounding abilities of our models. Next, we also proposed a low resource adaptation task for gesture generation to mimic a practical scenario when only few minutes of training data is available for a new speaker. We proposed a novel approach that can leverage an existing gesture generation model to learn a personalized gesture generation model in this low-resource setting. We also extended this low resource adaptation task to a new paradigm of crossmodal continual learning which mimics an intelligent agent continuously learning and adapting to styles of gesture generation. The tasks and models proposed in this these have advanced our understanding of the relationships between human nonvberbal behaviour and spoken language.

In this chapter, we start by summarizing the key contributions of this thesis in Chapter 7.1, followed by a discussion of some key insights from the thesis in Chapter 7.2. We conclude by outlining some of limitations that open up new directions for future work in 7.3.

7.1 Summary of Contributions

Descriptive Language: Grounding In Chapter 2, we proposed a neural architecture called Joint Language-to-Pose (or JL2P), which integrates language and pose to learn a joint embedding space in an end-to-end training paradigm. This embedding space can now be used to generate animations conditioned on an input description. We also proposed the use of curriculum learning approach which forces the model to generate shorter sequences before moving on to longer ones. We evaluated our proposed model on a parallel corpus of 3D pose data and human-annotated sentences with objective metrics to measure prediction accuracy, as well as with a user study to measure human judgment. Our results confirm that our approach, to learn a joint embedding in a curriculum learning paradigm by JL2P, was able to generate more accurate animations and are deemed more visually represented by humans than the state-of-the-art model. This chapter is a stepping stone toward understanding more complex relationships between spoken language, which is another form of verbal communication, and gestures that co-occur with them.

Co-Speech Gestures: Grounding and Long Tail Distributions In Chapter 3, 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 curated Pose-Audio-Transcript-Style (PATS), designed to study gesture generation and style transfer. It consists of 25 speakers (15 new speakers and 10 speakers from Ginosar [50]) for a total of 250+ hours of gestures and aligned audio signals. 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.

Co-Speech Gestures: Grounding and Gesture Style In Chapter 4, we proposed a new model, named Mix-StAGE, which learns a single model for multiple speakers while learning unique style embeddings for each speaker's gestures in an end-to-end manner. A novelty of Mix-StAGE was to learn a mixture of generative models conditioned on gesture style while the audio drives the co-speech gesture generation. Our proposed Mix-StAGE model significantly outperformed previous state-of-the-art approach for gesture generation and provided a path towards performing gesture style transfer across multiple speakers. We also demonstrated, through human perceptual studies, that the generated animations by our model are more natural whilst being able to retain or transfer style.

Co-Speech Gestures: Low Resources for Grounding and Gesture Style In Chapter 5, we studied low-resource adaptation of crossmodal generative models for gesture generation. We introduced a new generative model DiffGAN, that can efficiently address the shift in crossmodal grounding and the output distribution from the source to target speaker with only a few minutes of data. DiffGAN explicitly identifies shifts in crossmodal grounding relationships along with the shifts in the output domain from the pretrained source model. Based on these identified distribution shifts, DiffGAN updates a few necessary parameters in a single layer of the source model, allowing efficient adaptation with low resources. We benchmarked the effectiveness of our approach on a publicly available dataset through quantitative, qualitative and human studies. To our knowledge, this is the first approach that is able to learn a personalized gesture generation model with only 2 minutes of speaker data.

Co-Speech Gestures: Low Resource Continual Learning for Grounding and Gesture Style In Chapter 6, we studied low-resource adaptation of crossmodal generative models for gesture generation in a continual learning setting. We started with a gesture generation model trained on one speaker with high-resource data (i.e. 5-10 hours). We then sequentially exposed this model to multiple new speakers to simulate a continual learning scenario. We proposed an approach, named C-DiffGAN, that can efficiently leverage the continually arriving data to personalize the grounding and gesture style of the model to that of the new speakers. By utilizing a crossmodal catastrophic forgetting loss, the model is able to retain the grounding and style knowledge of the older speakers. To the best of our knowledge, this is the first approach that is able to learn a personalized model for multiple speakers with only 2 minutes each of speaker data (i.e. as opposed to 10 hours [7, 50, 60, 89]) in a continual learning setting. We substantiated the effectiveness of our approach a large scale publicly available dataset through quantitative and qualitative studies, which show that our proposed methodology significantly outperforms prior approaches for low resource continual learning of nonverbal grounding and personalization of gesture generation models.

7.2 Key Insights

Inductive Biases A recurring theme in the methods proposed in the thesis is to encode inductive biases that better fit with the corresponding task. These inductive biases stem from the patterns observed in the task, data, and/or relationships between the inputs and outputs of a model. By inducing relevant biases can help reduce the solution space, which is especially useful if the amount of training data available is limited. A common example is the use of convolutions for images and audio that encodes both translation invariance [90] and fixed contextual information as opposed to fully connected layers that make no such assumption or transformers [175] that have no constraints on the context length. The latter models are more general as they make lesser assumptions about the task, and hence heavily rely on massive amounts of data to find its way through a much larger solution space.

We have introduced many inductive biases for representing audio, language, language-audio fusion, multiple modes in the body pose generation. The AISLe approach in Chapter 3 is based of the assumption that the gesture and text have a long-tailed distribution. In Chapter 4, we proposed an approach that uses multiple generators to model each mode of the distribution separately. This inductive bias is especially helpful to personalize a single model with multiple speakers. In Chapter 5&6, the formulation of DiffGAN relies on the assumption that personalization of grounding and gesture styles of different speakers can be represented as a low-dimensional transformation which would require fewer examples.

Crossmodal Generative Modeling The core task in this thesis is to learn the relationships between body motion and language. Both of these modalities are temporal and lie in a high dimensional space, but the information that they contain is broadly different. Body motion is a visual space which can be heavily dependent on the surroundings and fundamental laws of physics, while language is a modality constructed by humans and does not have any connection to physics (except when it is describing the laws of physics). And, while it is possible to learn a common space between these two seemingly different modalities, this common latent space might not be enough to reconstruct the either modalities making the task of translation tricky. At the most, the relationships between body motion and language are weakly correlated. Hence to generate body motion that is translated from language, we not only rely on the weakly correlated space but also learn how the output distribution looks like. This knowledge of the output distribution helps to fill in the gaps that the language cannot provide. This phenomenon leads to an aleatoric uncertainty in the generative process. Aleatoric uncertainty is the intrinsic randomness in the data/model that cannot be resolved no matter how much information is added, unlike epistemic uncertainty which is caused due to missing information [56]. This results in the generation of more than one plausible body motion sequence, indicating that there is no one correct solution but a distribution of solutions.

Evaluation of Generative Models Due to the uncertainty of generation in the setting of weak correlation between language and body motion, it is a challenge to evaluate and compare models. Much like other generative models in literature for image [79, 80], audio [174, 181], and text [145, 146], qualitative visualization is one approach to verify the quality of generation. But this approach could involve a lot of cherry picking and making the evaluation quite random and hard to reproduce. Quantitative evaluation measuring correctness can often go wrong as the model task itself expects multiple plausible generations. Hence, diversity of generation becomes a more important metric. One way to measure that would be to distance between the generated body motion distribution and true distribution, such as a Fréchet Distance as proposed in Chapter 3. Distribution metrics give a clearer idea about the quality of the model's generations, especially in cases where the body motion generations are static (i.e. no diversity) or off the manifold of possible body motions. Even then, the quantitative metrics fall short of measuring the naturalness of the generations. Hence making human studies an essential part of the evaluation in this thesis.

7.3 Limitations $\leftrightarrow$ Future Work

The thesis of this research revolves around the discovery of the grounding relationships between two seemingly different modalities of body motion and language. These relationships can help guide generation of personalized gestures the are relevant to the input language and can do so with only few minutes of training data. While we have made some strides in these core technical challenges, there is a lot more to uncover, some of which we highlight in this section. We hope that these ideas could serve as a spark for some cool new research direction.

Weak correlation between verbal and nonverbal channels of communication The underlying differences between the verbal an nonverbal channels of communications result in a weakly correlated latent space of these two channels. For example, beat gestures are highly correlated with the prosody, but if there are gestures without any spoken language (i.e. miming) it is not easy to predict or generate such gestures. In this thesis, the key idea was to find and learn the stronger relationships between the verbal and nonverbal channels and fill in the gaps with probable hallucinations for relatively weaker relationships. While this approach helps generate plausible gestures, it is possible that some of these hallucinated behaviours may not be expected by an end-user. This could potentially reduce the naturalness and correctness of the interactions. This is an interesting inflexion point for a new direction of research which could try to answer questions on modeling such weak correlations in a multimodal generative setting.

Scaling challenges Human behaviour is idiosyncratic. Correctly modeling the behaviour of every person is a key goal of the *personalization* challenge, but it is certainly not an easy one. At the time of writing this thesis, the human population of the world is upwards of 7 billion people. At this level, collecting and learning the behaviour of every person becomes a scaling challenge. This is especially hard, because behaviour of humans can change over their lifetime and needs to be updated. But a ray of hope in this mammoth of a challenge is that while every person may have different behaviour patterns, these patterns a influenced by a combination of multiple people. Let's assume, at an average, every person is friends with 10 other people. In this graph of social relationships, if we conservatively restrict the influence of only second degree connections on a person's behaviour, each human could largely be represented by 100 other humans' behaviours. Hence, we would potentially require modelling the behaviour of tens of millions of people. This is a still a large number and a conservative upper bound based on the assumption that our social networks are uniform. In reality, our social networks are heavily centered toward a few people who have significantly more influence than others. Modeling such clusters of similar behaviours can further reduces the complexity of scaling up. A trade-off here is that these clusters are dynamic and can be formed based on many criteria such as individuals, friends, family, school, university, cultural leanings and many more. A solution to this challenge of scaling lies somewhere at the intersection of modeling the social network and connecting those to individual nonverbal behaviour.

Beyond Monologues In this thesis, we have mostly focused on modeling body motion for only one individual in the scene. However, human behaviours, are more often than not, guided by many other factors such as the situation (formal or informal), location (home, vacation, or work) and other living beings (friends, family or pets). These factors are not static and can quickly change with an expectation of an appropriate reaction. For example, a person's body motion dynamics are significantly different if they are giving a monologue or are having a dialogue or communicating in a group setting. Modeling these group dynamics can be especially challenging form the viewpoint of the "curse of dimensionality" because behaviours of the interlocutors constitute another complex dimension. Adding multiple people to the conversation could also create the need to study the reasoning behind the conducted behaviour.

Beyond Individual Gesture Style The discussion on gesture style and personalization has been limited to an individual level in this thesis. However the definition of style can be extended to four broad categories based on granularity, (1) *Culture*, where a group of people share a set of common gestures, (2) *Individual*, where a single person has their own idiosyncrasies which constitutes to their gesture style, (3) *Situation*, where the current environment dictates the kinds of gestures a person might make. For example, the gestures in an office setting can be very different from the gestures made in an informal setting with friends and (4) *Emotion*, where the current thought process could dictate the distribution of gestures. An important direction of research would be to personalize human behaviour according all dimensions of style which can be accelerated by the curation of relevant annotations of these style dimensions.

Beyond Supervised Crossmodal Translation Throughout, we have worked with supervised datasets for the crossmodal translation task of gesture generation. However, supervised datasets are hard to come by in real world situation, especially for new speakers. We have explored low resource and continual learning settings that rely on a well trained source model as a starting knowledge base and a few supervised examples for a new speaker. An interesting research direction would be to explore a semi-supervised low resource setting where we have access to a few supervised examples for multiple speakers, but a relatively larger set of unsupervised examples (i.e. unimodal gesture and language data that are not related to each other). Using the few supervised examples, it may be possible to learn a low-dimensional transformation between the latent spaces of two different speakers. Furthermore, this direction is a potential stepping stone towards a fully unsupervised methodology for gesture generation.

Beyond Rehearsed Datasets Rehearsed datasets consists of speakers who are professionals and are able to generate a similar distribution of gestures every time. The data in such datasets have a pattern and hence are good test bed to study the grounding and personalization relationships. But, in the wild datasets have no such constraints making it non-trivial to translate the findings on the rehearsed datasets. An interesting direction of research lies in the curation and study of such in-the-wild datasets. The hodgepodge of different styles, situations and types of gestures adds an additional layer of complexity in lieu of a more practically usable outcome.

Additional Low Resource Challenges Low resource settings in generative modeling has different challenges when compared to a classification paradigm. One approach in low-resources for classification is to find the decisive examples (i.e. examples close to the decision boundary) which should, in theory, be able to reconstruct the decision boundary. Generative modeling, on the other hand, should be able to reconstruct the complete modality space. Hence, in a low resource settings, the important examples lie all over the distribution space. In this thesis, we focus on how to utilize a few examples to transform an existing latent space to a new domain. While this approach is effective, selecting the best examples to make this transformation can optimize the effectiveness of the adaptation approaches. This direction of research has a longer term goal of being able to specify a small set of sentences that a speaker needs to record with which a well-trained gesture generation model can be constructed. This is akin to a personalized text-to-speech system which can be constructed using a small set of sentences that cover all the phoneme and phoneme transition pronunciations.

Beyond Co-speech Gestures Currently the grounding and personalization research is limited to co-speech gestures and location based body motion. An interesting future direction could be the exploration of domains such as dancing, sign language and so on and develop approaches to unify them at the level of grounding and generative modeling. Specifically, a goal could be to learn relationships between co-speech gestures and spoken language which could be useful in learning relationships between another domain of visual body motion and verbal/vocal cues with lesser supervised data. A long term goal is to learn universal decoder which can learn to generate visual body for many domains with minimal supervision.

Acknowledgements

This material is based upon work partially supported by the National Science Foundation (Awards #1750439 #1722822), National Institutes of Health, the InMind project, Oculus VR and NTT Japan. 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 the aforementioned organizations, and no official endorsement should be inferred.

Appendices

Appendix A

Descriptive Language: Grounding

A.1 How does the choice of encoder/decoder models affect performance on the task?

Choice of Language Encoders: As an ablation to the models in Table A.2, a pre-trained model of BERT [41] is used as the language encoder. A smaller network, with 2 fully-connected layers, ReLU [51] and Dropout [162] on top of BERT is fine-tuned jointly with the pose encoder and decoder.

Switching from LSTM to BERT in the encoder gives a small improvement to APE values but they are not significant. Although, in Figure A.3, it is surprising to see that humans prefer models with language encoders as LSTMs (30% vs ground truth) rather than BERT(20% vs ground truth), which strengthens the need of user studies for evaluating pose generation models.

Choice of Decoders: The choice of pose decoder can influence the generation of animations. We run an ablation study comparing different decoder models,

- **Emphasizing Language Context**: Concatenate z_x to every input while decoding
- **Emphasizing Pose Context**: concatenating hidden states $c \in \mathcal{C}$ of the decoder from previous time step to the current input.

Giving the decoder inputs an addition feedback of joint embedding z or pose context c gives a small improvement to the objective metrics, but it is not significant (see Table A.1).

A.2 Additional Objective and Subjective Results

For the sake of brevity, we only show APE scores for different parts of the body (like LHand, RHip and so on) in the main paper. We complement that with a detailed table of APE scores for each separate joint in Table A.2 and APE scores changing across time for the same models in Figure A.1. We also conduct a user study which compares preference of all baseline models and Lin et. al. [104] against ground truth videos. All of them show the same trends of improving performance due to addition of training curriculum, joint embedding and Smooth L1 loss.

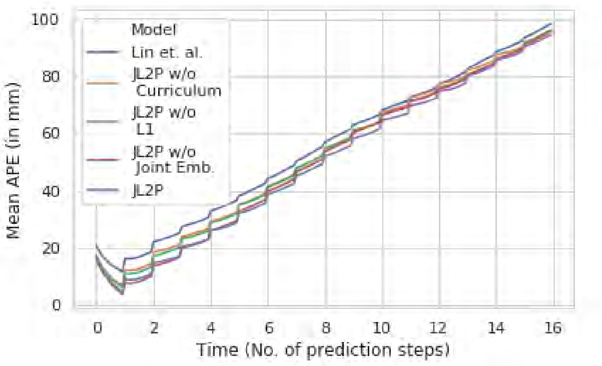


Figure A.1: Plot of mean APE values across time steps for all baseline models. Each time-step is approximately 80ms long. Lower is better.

A.3 Some failure cases

While, the model is able to generalize over a variety of different actions, currently there are some kinds of actions which have not been convincingly modeled by this approach. Some examples include, *chicken wings flapping*, *moving forward then backward* (backward gets forgotten), *waltz/dance*, *dancing a cool move* (too ambiguous to learn), *kicking* (the kick is visible but not very well formed).

A.4 Training Hyperparameters

We use a size of 1024 for the joint embedding, pose encoder's GRU embedding size, decoder's GRU embedding size and language encoder's LSTM embedding size. This model is trained with Adam [85] Optimizer with a starting learning rate of 0.001 which decays exponentially over time

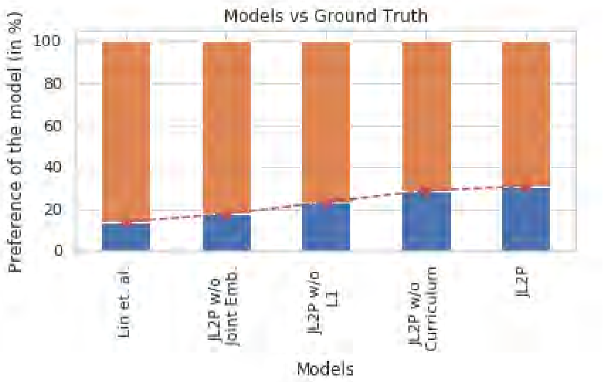


Figure A.2: Preference scores of baseline models vs Ground Truth. Blue bars denote the preference percentage of models marked on the horizontal axis. Our models (JL2P and variants) show consistent rise in preference over Lin et. al. with the addition of components joint embedding, smooth L1 loss and curriculum learning.

by a factor of 0.99. Training is stopped once the loss over the validation set increases for 3 epochs at a stretch.

Language Encoder (p_e)	Feedback		Average Positional Error (APE) in mm		
	$\mathcal{Z}$	$\mathcal{C}$	Mean w/o Root	Root	Mean
LSTM	✗	✗	41.6	131.1	49.5
BERT	✗	✗	40.4	129.0	48.2
LSTM	✓	✗	**39.9**	**126.2**	**47.5**
BERT	✓	✗	40.9	137.7	49.1
LSTM	✗	✓	41.0	134.3	49.1
BERT	✗	✓	**39.9**	**126.7**	**47.5**

Table A.1: Average Positional Error (APE) for our model JL2P and its variants by tweaking language encoder (LSTM vs BERT) and pose decoder (**emphasizing language context** ($z \in \mathcal{Z}$) or **emphazising pose context** ($c \in \mathcal{C}$).)

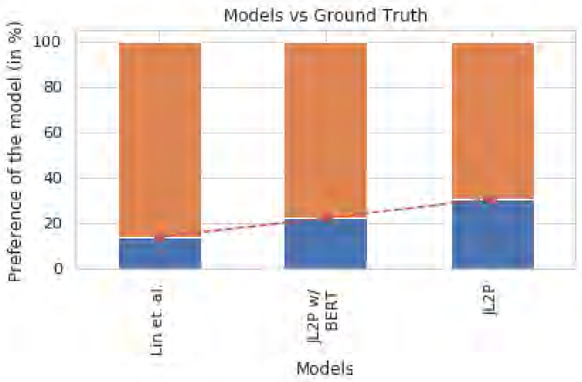


Figure A.3: Preference scores of ablation models vs ground truth. Blue bars denote the preference percentage of models marked on the horizontal axis. JL2P loses human preference if the language encoder p_e is switched to BERT, from LSTMs but it is still better than Lin et. al..

Models	Average Positional Error (APE) in mm																						
	Mean	Mean w/o Root		Torso		Head		LHand			RHand			LHip		LFoot			RHip		RFoot		
			Root	BP	BT	BLN	BUN	LS	LE	LW	RS	RE	RW	LH	LK	LA	LM	LF	RH	RK	RA	RM	RF
Lin et. al. [104]	54.9	50.0	151.6	21.5	31.7	34.7	36.1	36.8	59.8	87.3	38.8	59.6	86.4	20.8	43.7	60.9	63.0	66.0	20.6	43.6	61.3	62.8	65.4
JL2P w/o Curriculum	52.2	47.9	139.2	19.1	29.2	31.6	33.5	35.1	55.2	81.7	35.7	55.8	80.1	19.2	42.0	60.4	62.6	65.8	18.8	42.7	60.9	62.7	66.0
JL2P w/o L1	51.7	47.0	145.0	19.1	29.8	32.2	33.4	35.4	55.9	82.7	35.6	56.2	81.1	18.7	41.2	57.5	59.0	61.5	19.3	42.1	58.1	59.3	62.2
JL2P w/o Joint Emb.	50.4	45.7	143.3	19.8	**28.2**	**30.2**	**31.7**	**33.8**	54.2	78.9	**33.4**	**53.8**	**76.4**	18.8	40.6	**57.1**	58.7	**61.1**	17.9	41.2	57.4	59.1	62.1
JL2P	**49.5**	**45.4**	**131.1**	**17.8**	28.3	30.4	32.4	**33.8**	**53.6**	**78.4**	34.2	**53.8**	77.0	**17.4**	**39.8**	**57.1**	**58.6**	61.8	**17.5**	**40.4**	**56.8**	**58.4**	**61.3**

Table A.2: A more detailed version, with all joints separate, of Average positional error (APE) for JL2P , JL2P w/o Joint Emb., JL2P w/o L1, JL2P w/o Curriculum and Lin et. al.. Lower is better. Our models (JL2P and variants) show consistent increase in accuracy over Lin et. al. across all joints with the addition of components joint embedding, smooth L1 loss and curriculum learning.

Appendix B

Co-Speech Gestures: Grounding and Long Tail Distributions

B.1 More analyses

Weight updates on a toy example For a toy example that learns to generate samples from a guassian distribution, we show how the resampling weights look like for each bin over the course of optimization in Figure B.1. We also note that AISLe recognizes discrepancies throughout the proposal distribution and assigns larger weights (local maximas denoted by —) to the samples in those bins. Initially, the weights are almost equal for all bins, because the discriminator is still learning to predict the correct likelihood. As the training progresses, the network learns to first focus on the center of the distribution, before shifting gears towards the tail close to the end of training.

Coverage vs frequency of occurring words While it is expected that rarely occurring words will be less likely to generate gestures that represent the true distribution (right half of Figure B.2), we see that AISLe is able to push boundaries of MMS-Transformer to generate more correct distributions for words in the long tail.

More qualitative analysis We compare the coverage of the generated gestures for our model and baselines in Figure B.9. We also show generated gesture as a skeleton plot over the ground truth and compared with previous work in B.8. While, these images give some idea about the qualitative performance, we would recommend looking at the attached video for a better under-

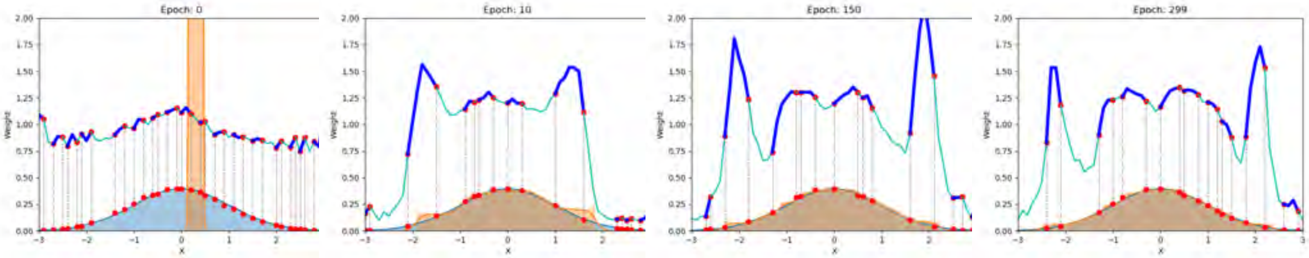


Figure B.1: Progress of learning a Gaussian distribution using AISLe on top of a vanilla GAN. The top line plot refers to the weights assigned adaptively assigned to the samples corresponding to the bins on the X axis. Initially, the model focuses on the samples close around zero and gradually moves on to focusing on the heavy tail of the distribution. The dark-blue segments of the line plot refer to local maximas and segments that require more attention. The dotted vertical lines correspond to inflection points of the weight line plot.

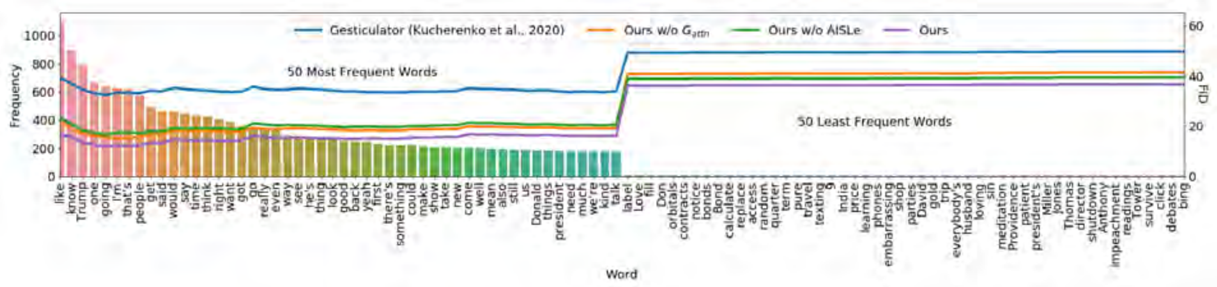


Figure B.2: Words vs FID. The top 100 occurring words are shown

standing.

Speaker-wise objective results We also have the speaker-wise objective results in Figures B.10-B.14.

B.2 Model

B.2.1 Estimating Mixture Model Priors during Training

During training, we partition poses Y_p into M clusters using an unsupervised approach, Lloyd's algorithm [109]. While other unsupervised clustering methods [148] can also be used at this stage, we choose Lloyd's algorithm for its simplicity and speed. Each of these clusters represent samples from probability distributions $\{p^1(y|x), p^2(y|x), \dots p^M(y|x)\}$. If a sample belongs to the m^{th} cluster, $\phi_m = 1$, otherwise $\phi_m = 0$, making Φ a sequence of one-hot vectors. While training the generator G_θ, if a sample belongs to the distribution $p^m(y|x)$, only parameters of sub-generator G_m are updated. Hence, each sub-generator learns different components of the true distribution, which are combined using Equation 9 (main paper) to give the generated pose.

B.2.2 Aligning Multi-Scale Embeddings

As language and audio have different scales, we augment the idea of positional embeddings proposed in [175] to provide the information of word-level ordering as well as sub-word frame-level ordering.

Word-level Ordering: Given language embeddings $\mathbf{Z}^w \in \mathcal{R}^{N \times h^w}$, where N represents the number of words in a sampled sequence, and $pos \in N$ is the dimensional position of the word in the sequence. The term $i \in h^w$ represents the i-th position word embedding and is used to ensure that each positional encoding corresponds to a sinusoid. We add the corresponding word-level positional embedding for each word embedding.

The positional embedding is derived as the following:

$$\mathbf{PE}_{pos,i} = \sin(pos/10000^{\frac{2i}{h^w}}), \qquad \text{if } i \text{ even} \qquad \text{(B.1)}$$

$$\mathbf{PE}_{pos,i} = \cos(pos/10000^{\frac{2i-1}{h^w}}), \qquad \text{if } i \text{ odd} \qquad \text{(B.2)}$$

Frame-level Ordering: Given a single language embeddings $\mathbf{Z}^w \in \mathcal{R}^{N \times h^w}$, $\mathbf{Z}^w$ occupies multiple time-frames. In order to account for the frame-wise progression of the word, we use the same positional embedding as shown in the above equation. We additionally process the word duration of each word, which represents the number of frames each word occupies. Then, we replace $pos \in$ Word Duration with the position of the frame for the word. We add the corresponding frame-wise positional embedding for each frame-level word embedding.

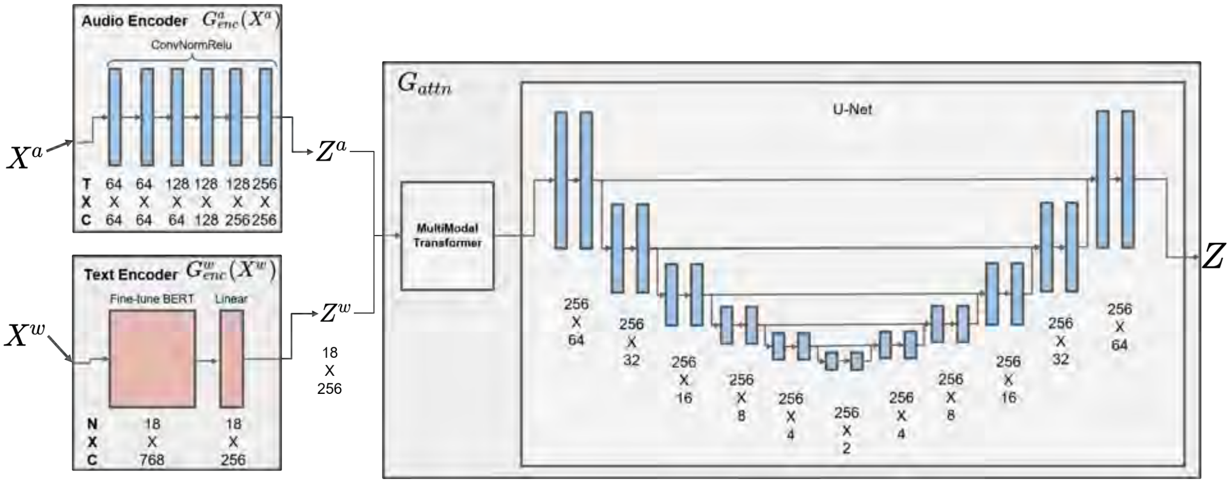


Figure B.3: Encoder Architecture

B.3 Experiments

B.3.1 Baselines

Gesticulator[89]: Unlike MMS-Transformer , Gesticulator is an autoregressive model for generating gestures using text and speech. The audio inputs are represented via log-mel-spectrograms. For text features, in comparison to multi-scale BERT embeddings used for MMS-Transformer, single scale BERT embeddings are used for the Gesticulator. The text features are repeated to align with audio frames. Furthermore, text features corresponding to filler words and features for silence ,which do not contain semantic information, is additionally processed. Using WebRTC Voice Activity Detection [183] to find timesteps with silence, all elements of the in the embeddings corresponding to silence is set to -15 and made distinct from all other audio encodings. Additionally, filler words are found for each speaker's transcripts using the NLTK package. Then, the weighted average of the BERT embeddings of all filler words spoken per speaker are calculated. The averaged filler BERT embedding replaces each filler word spoken by the speaker. After processing the data to account for silence and filler words, in order to provide more contextual information, a sliding window of audio features, including 7 and 15 future time steps are concatenated with the current time step features (audio and text) to produce a long vector.

In comparison to MMS-Transformer, implemented with cross-modal multihead attention, CNNs and adversarial training, the Gesticulator's model architecture relies solely on fully connected layers. Given the processed input as described above, 3 fully connected layers are applied to reduce dimensionality and producing an output x. Furthermore, unlike our model, autoregression is applied via FiLM conditioning, where previous 3 poses are taken as input and fed into fully connected layers to produce scaling α and offset vectors β. Then the out put is applied to element-wise affine transformations: $x * \alpha + \beta$. For the first 7 epochs, no FilM conditioning is applied. Then for the proceeding 5 epochs, varying teacher forcing is applied where the number of times the model receives ground-truth poses is annealed over time. By the 12th epoch, the model uses its own generated poses in FiLM conditioning [141]. Finally, the loss function is a sum of MSE between the poses and the velocities of the gestures.

B.3.2 Implementation details

We use PyTorch as the auto-differentiation library to train all our models. The detailed description of our model, with layer sizes, is described in Figures B.3, B.4 and B.5. In our experiments, we use use the following hyperparameter settings: Our batch size is 32, sampling intervals of approximately 4.0 seconds for each batch. We use a overlapping windows during sampling with step-size of 5. In order to find the optimal learning rate within the range of 0.00001 to 0.00005, we uniformly sampled with an increment of 0.00001 and ran an hyperparameter search on one model for one speaker. We found that the learning rate 0.00003 was marginally better than the others, making it our choice for all models. Furthermore, in training, we use Adam with rectified weight decay [111] with a linearly decaying learning rate schedule. The number of training iterations are 40000 and we check the validation score at every 400 iterations, making sure that the model runs for a minimum of 20000 iterations before it considers early stopping. We use M=8 for the mixture of GANs. We choose 8 by running an ablation which shows that the performance plateaus after 8. The average model train runtime was around 24 hours (+ 6 hours if it decided to run the complete 40000 iterations) on Titan X 1080 GPUs.

The following evaluation metrics were used, with links provided:

- FID: https://github.com/mseitzer/pytorch-fid/blob/master/fid_score.py
- WD1: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wasserstein_distance.html
- PCK: https://github.com/amirbar/speech2gesture/blob/master/common/evaluation.py

B.4 PATS Dataset

We used a train/validation/test split of 80/10/10 which is fixed to ensure consistency across experiments. A visual description of the dataset, which compares the lexical and gesture diversity of each individual speaker, can be found in Figure B.6. Link to dataset: http://chahuja.com/pats

B.4.1 Human Perceptual Study

We attach a screenshot of a sample study and the questions asked to the users in Figure B.7

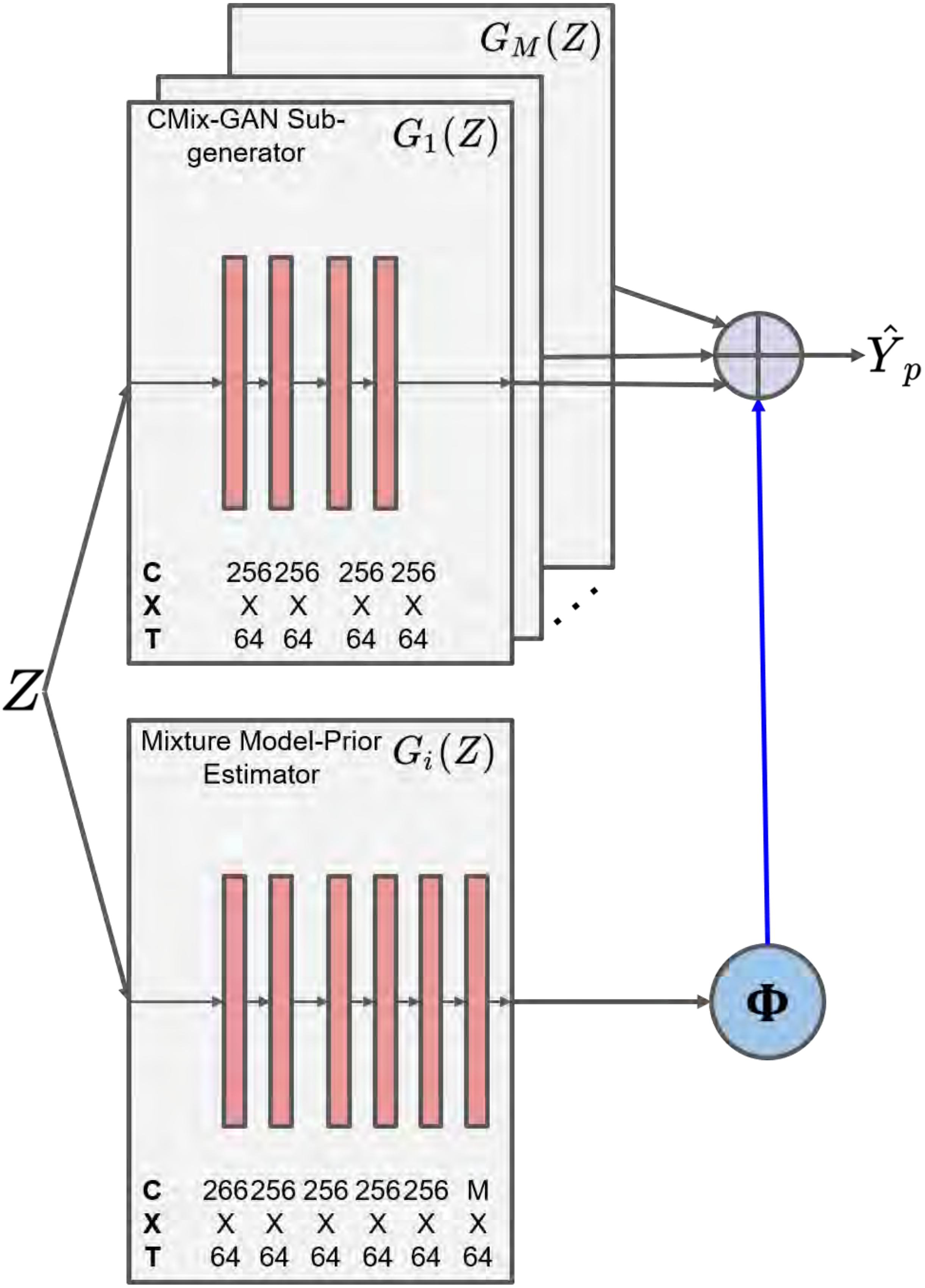


Figure B.4: Decoder Architecture

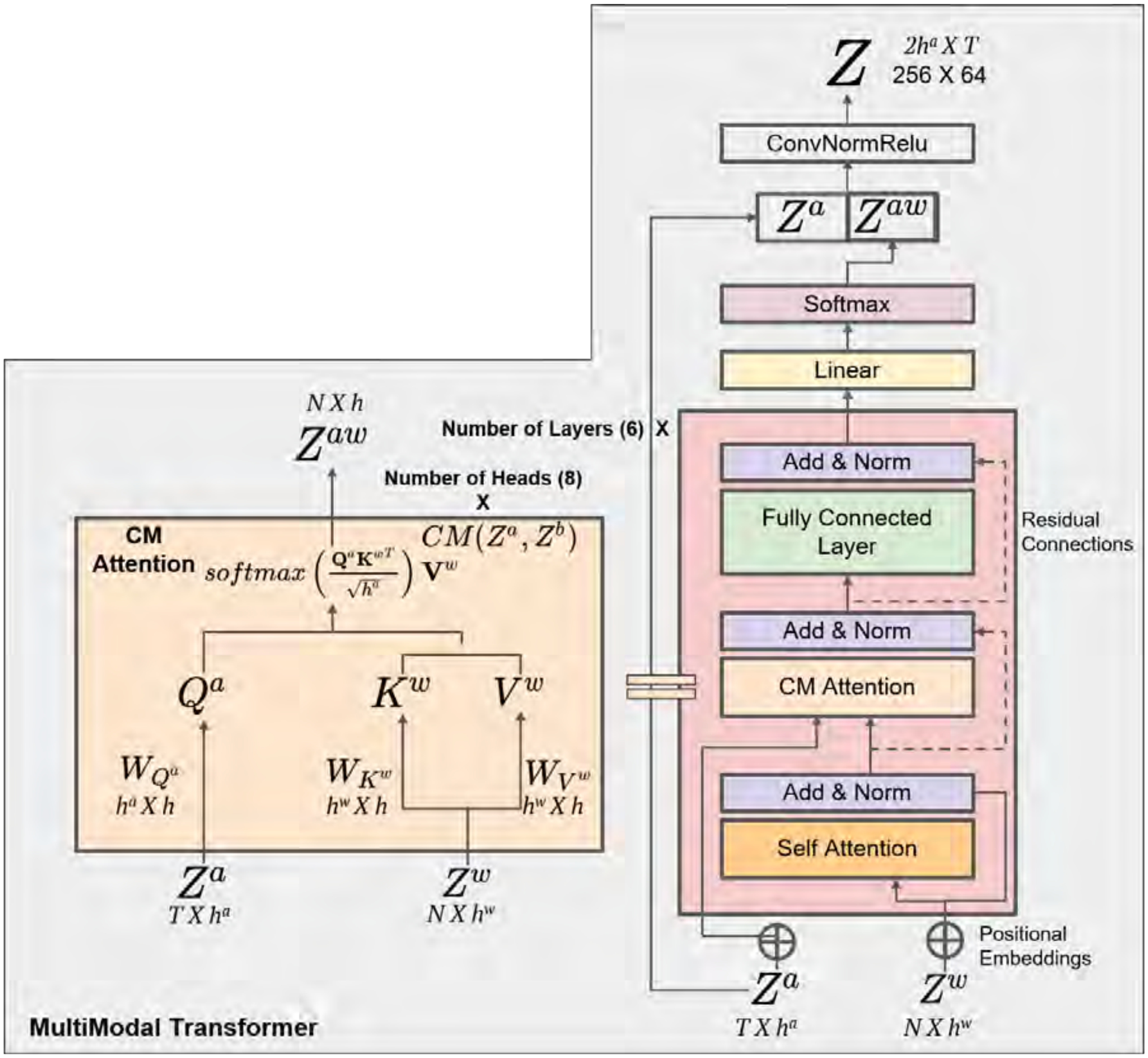


Figure B.5: MultiModal Transformer Architecture

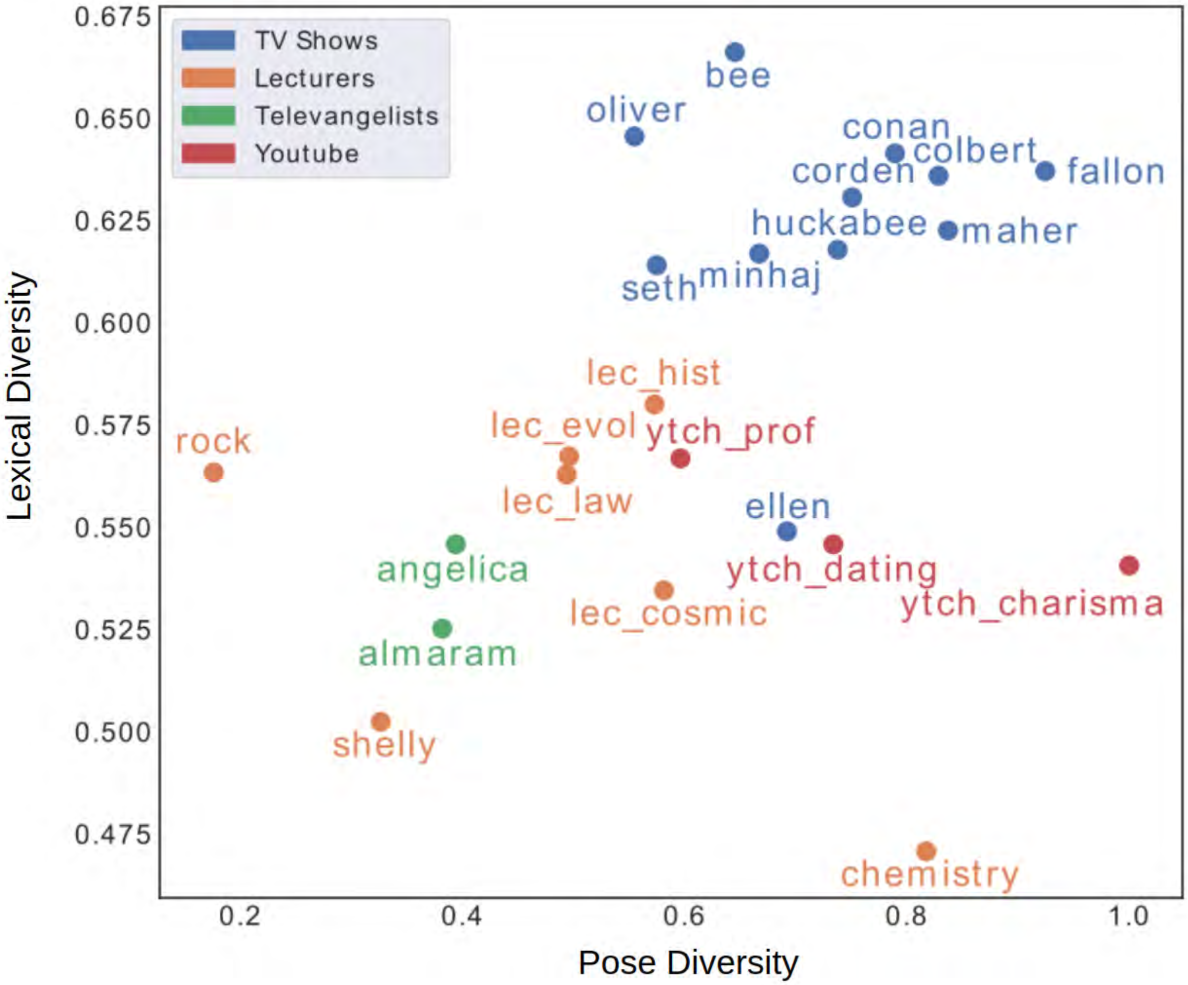
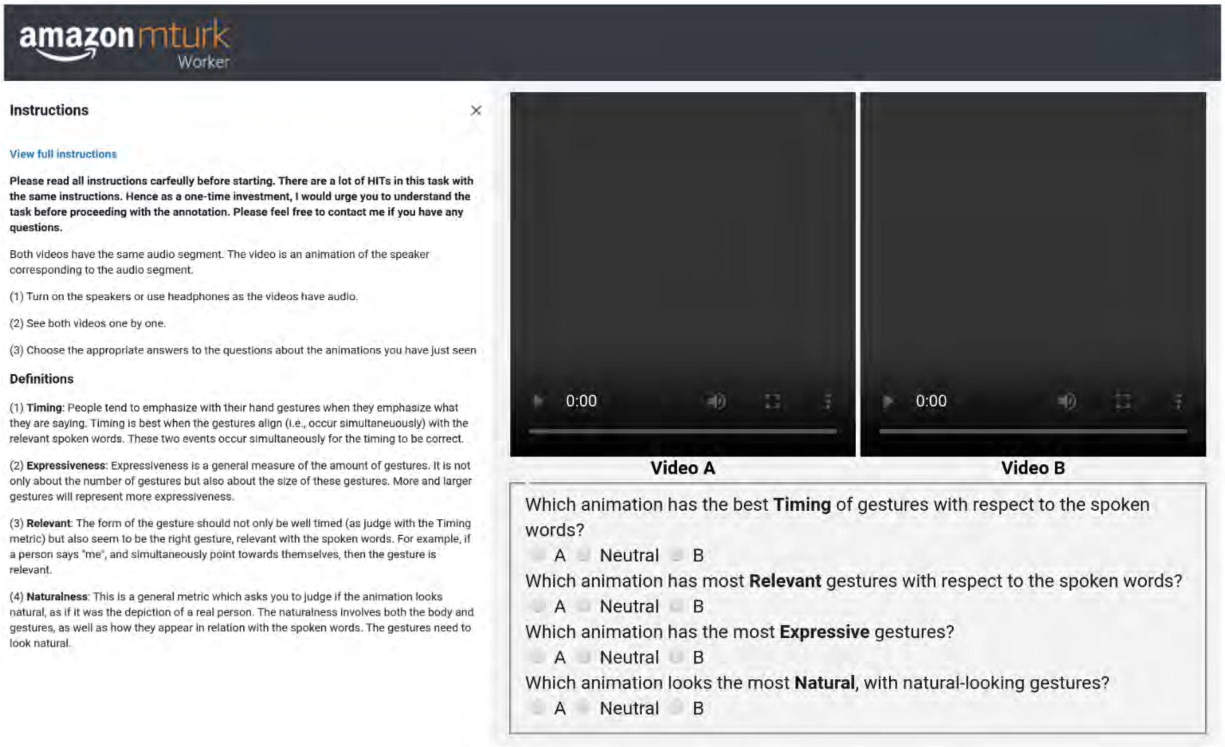


Figure B.6: A visual representation of all speakers in PAT+ dataset. X-axis represents the average diversity of gestures while Y-axis denotes the lexical diversity in the speakers transcripts.



Instructions

View full instructions

Please read all instructions carfeully before starting. There are a lot of HITs in this task with the same instructions. Hence as a one-time investment, I would urge you to understand the task before proceeding with the annotation. Please feel free to contact me if you have any questions.

Both videos have the same audio segment. The video is an animation of the speaker corresponding to the audio segment.

(1) Turn on the speakers or use headphones as the videos have audio.

(2) See both videos one by one.

(3) Choose the appropriate answers to the questions about the animations you have just seen

Definitions

(1) **Timing**: People tend to emphasize with their hand gestures when they emphasize what they are saying. Timing is best when the gestures align (i.e., occur simultaneuously) with the relevant spoken words. These two events occur simultaneously for the timing to be correct.

(2) **Expressiveness**: Expressiveness is a general measure of the amount of gestures. It is not only about the number of gestures but also about the size of these gestures. More and larger gestures will represent more expressiveness.

(3) **Relevant**: The form of the gesture should not only be well timed (as judge with the Timing metric) but also seem to be the right gesture, relevant with the spoken words. For example, if a person says "me", and simultaneously point towards themselves, then the gesture is relevant.

(4) **Naturalness**: This is a general metric which asks you to judge if the animation looks natural, as if it was the depiction of a real person. The naturalness involves both the body and gestures, as well as how they appear in relation with the spoken words. The gestures need to look natural.

Video A **Video B**

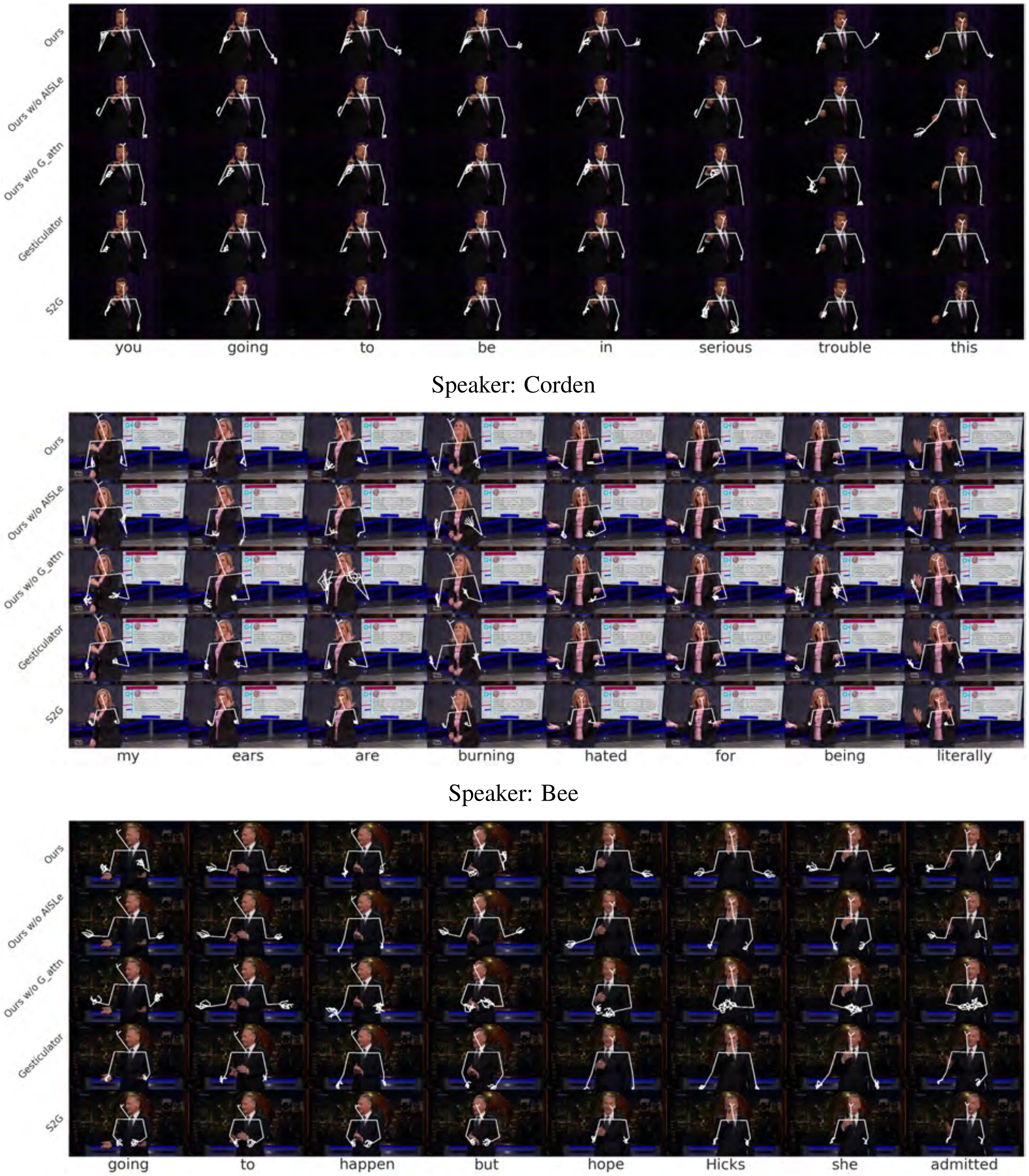
Which animation has the best **Timing** of gestures with respect to the spoken words?
○ A ○ Neutral ○ B

Which animation has most **Relevant** gestures with respect to the spoken words?
○ A ○ Neutral ○ B

Which animation has the most **Expressive** gestures?
○ A ○ Neutral ○ B

Which animation looks the most **Natural**, with natural-looking gestures?
○ A ○ Neutral ○ B

Figure B.7: Screenshot of MTurk Experiment used to measure subjective metrics

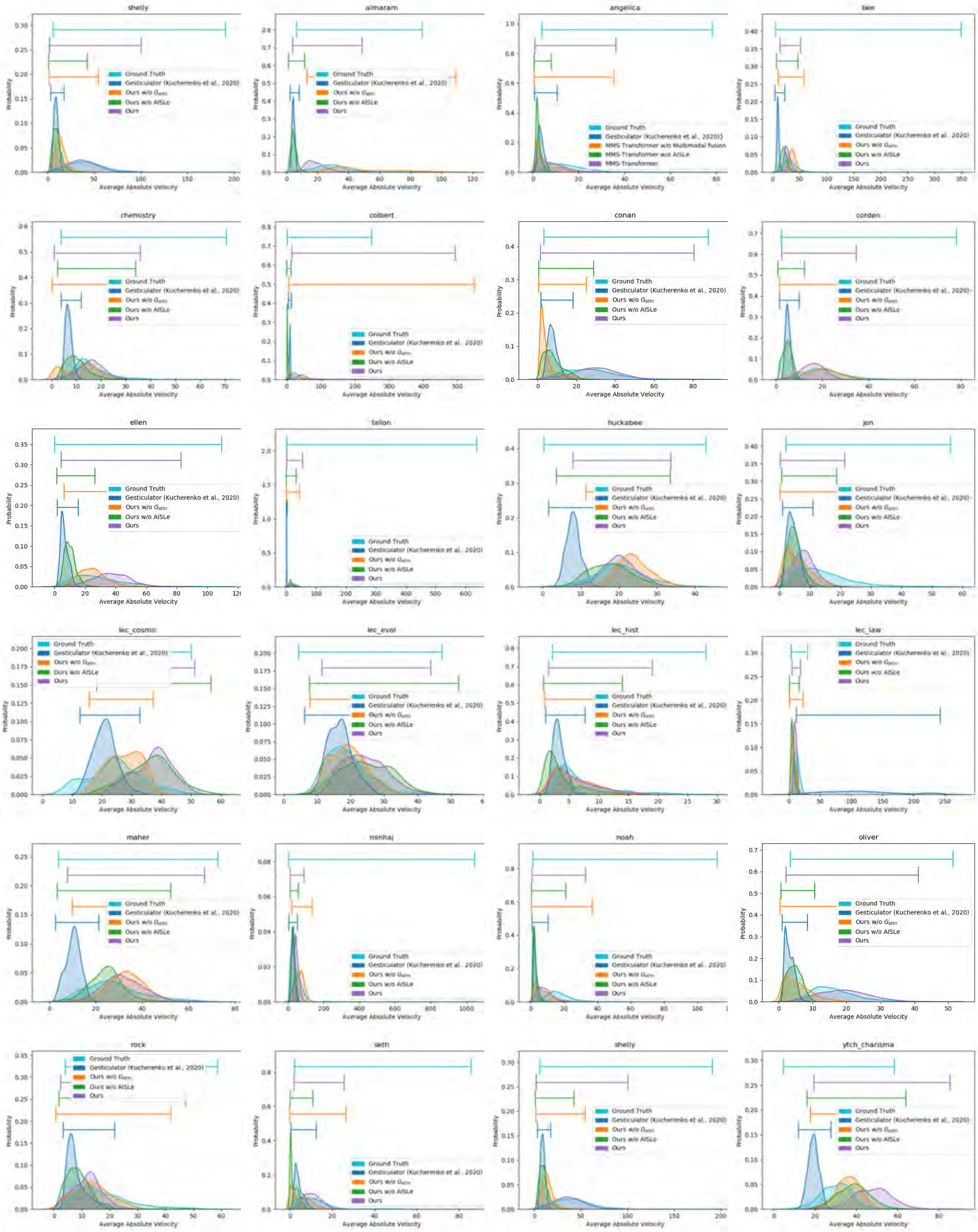


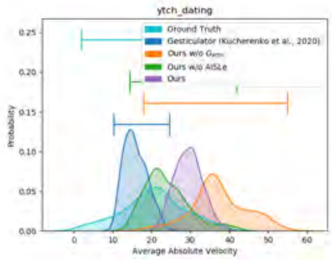
Speaker: Corden

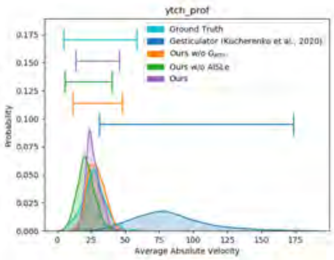
Speaker: Bee

Speaker: Maher

Figure B.8: Generated animations are plotted as frames over the ground truth video frames. The text at the bottom refers to the context of the generation. While, these images give some idea about the qualitative performance, we would recommend looking at the attached video for a better understanding.


shelly
almaram
angelica
bee
chemistry
colbert
conan
corden
ellen
fallon
huckabee
jon
lec_cosmic
lec_evol
lec_hist
lec_law
maher
minhaj
noah
oliver
rock
seth
shelly
ytch_charisma
Ground Truth
Gesticulator (Kucherenko et al., 2020)
Ours w/o G_{attn}
Ours w/o AISLe
Ours
Gesticulator (Kucherenko et al., 2020))
MMS-Transformer w/o Multimodal fusion
MMS-Transformer w/o AISLe
MMS-Transformer
Probability
Average Absolute Velocity


ytch_dating
Ground Truth
Gesticulator (Kucherenko et al., 2020)
Ours w/o G_{attn}
Ours w/o AISLe
Ours
Probability
Average Absolute Velocity


ytch_prof
Ground Truth
Gesticulator (Kucherenko et al., 2020)
Ours w/o G_{attn}
Ours w/o AISLe
Ours
Probability
Average Absolute Velocity

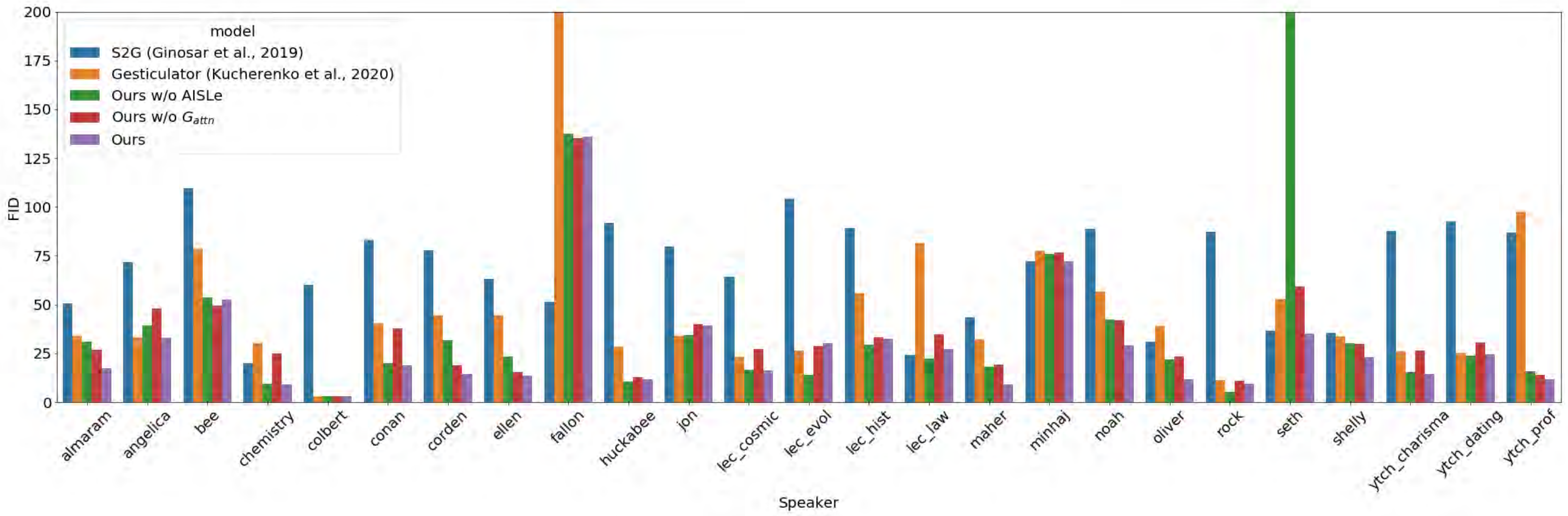


Figure B.10: FID

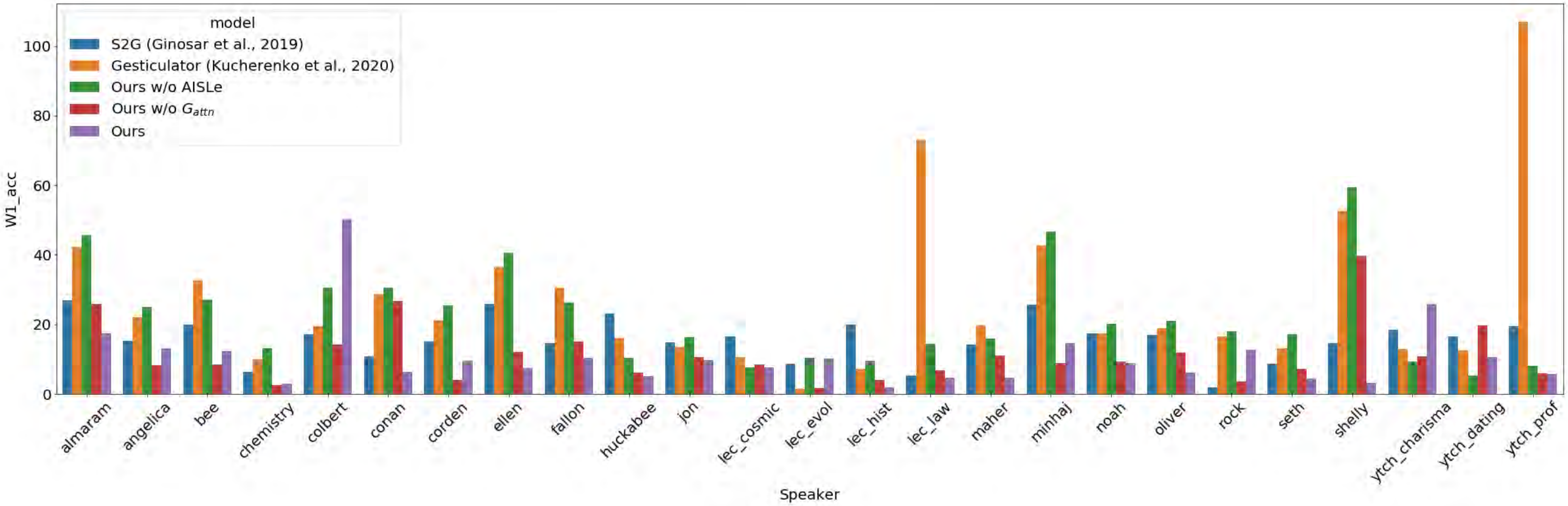


Figure B.11: W1 (acc.)

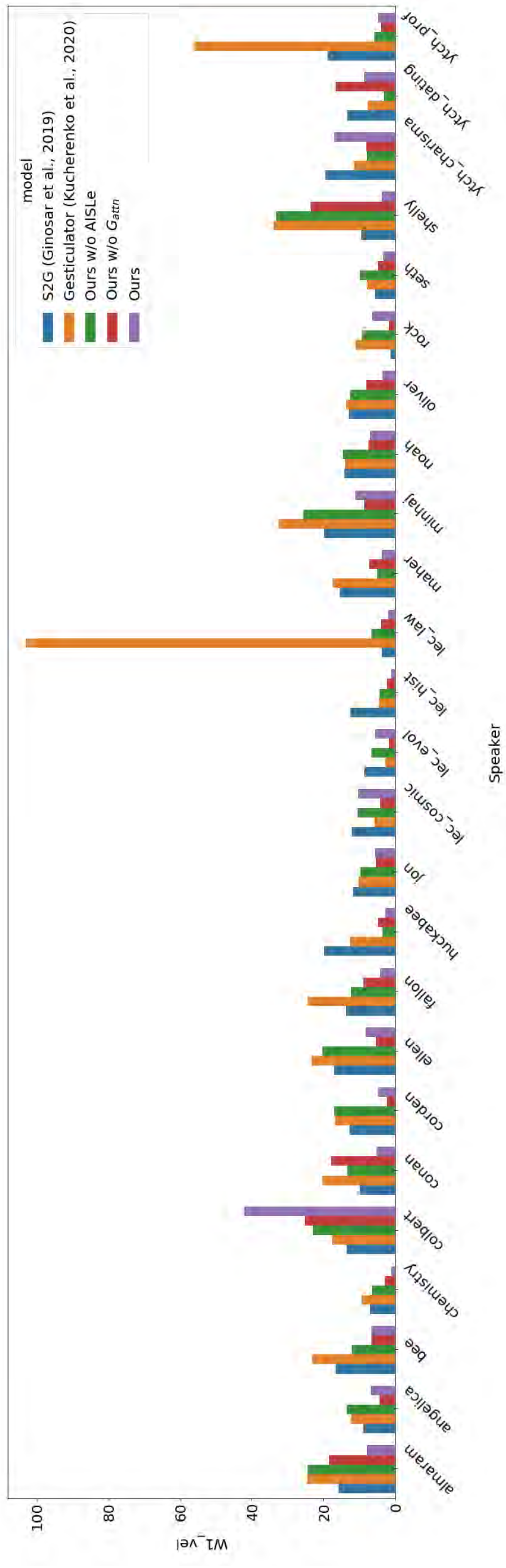


Figure B.12: W1 (vel.)

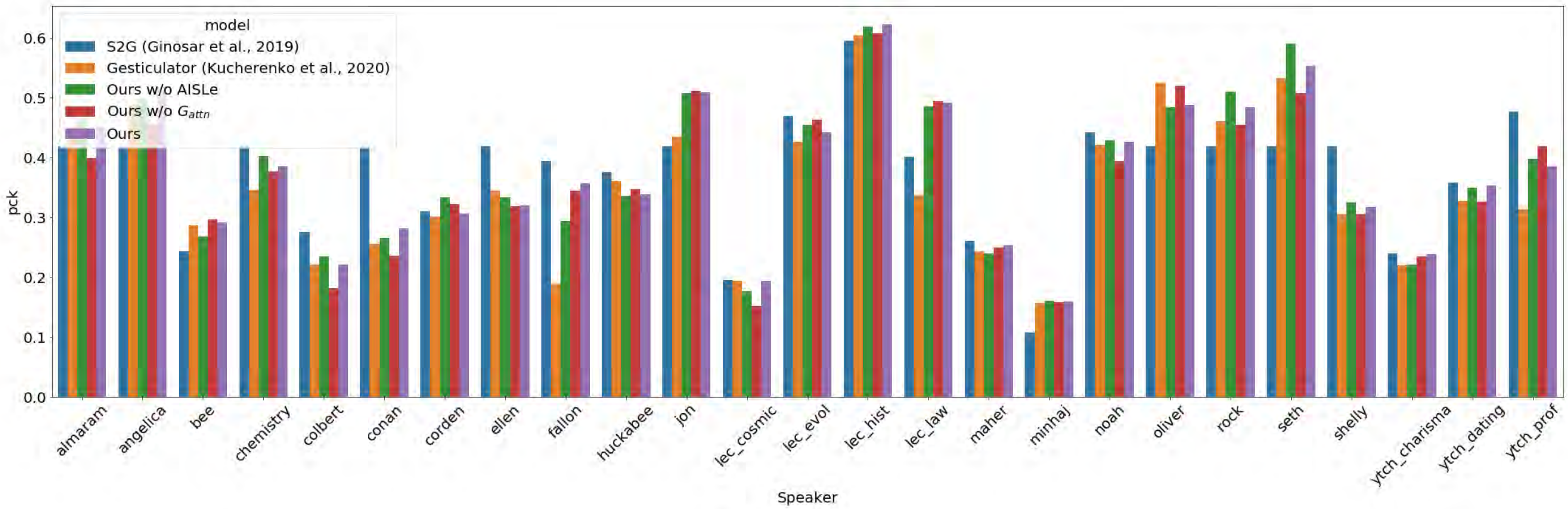


Figure B.13: PCK

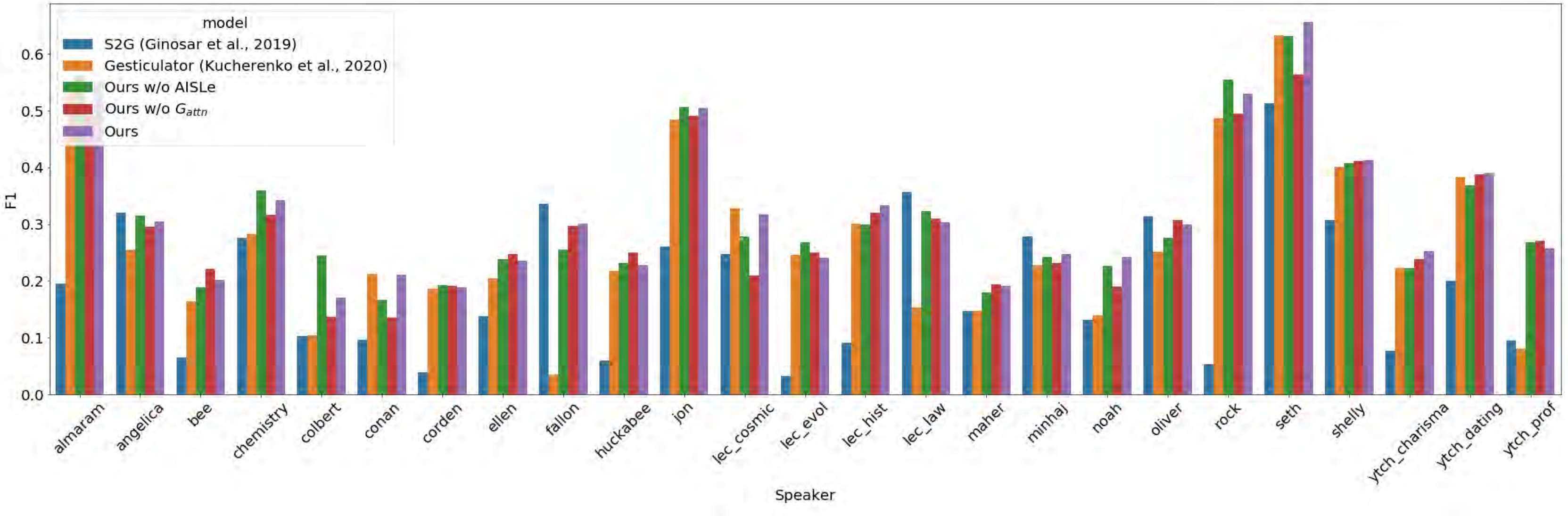


Figure B.14: F1

Appendix C

Co-Speech Gestures: Grounding and Gesture Style

C.1 PATS dataset

C.1.1 Speaker List

The list of speakers in the dataset are in Figure C.1 as a dendrogram. This dendogram was created using text as the discriminating features. Speakers within the same cluster have a similar vocabulary. For the purposes of our experiments we use the speakers listed in Table C.2 and C.3.

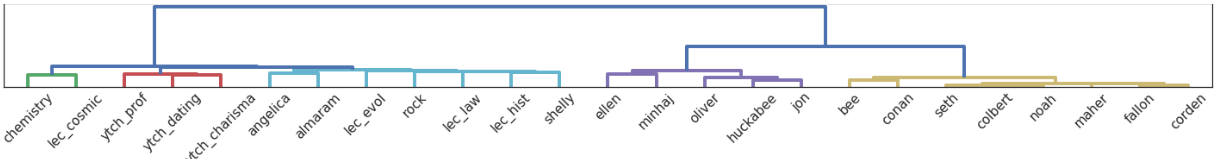


Figure C.1: List of speakers in the dataset as a dendrogram based on the content of the speech.

C.1.2 Attributes

We define 4 different attributes in Table C.1 and select pairs of speakers that demonstrate visually striking differences with respect to those attributes.

Sitting/Standing	**Gesture Frequency**	**Body Orientation**	**Primary Arm Func.**
Sitting: Noah	**Low:** Seth	**Right:** Chemistry	**Right Arm:** lec_cosmic
Standing: Maher	**High:** Oliver	**Left:** Oliver	**Left Arm:** lec_cosmic

Table C.1: Selection of speakers for attribute-level style modeling

Attributes	**Speakers**	**Single-Speaker Models**				**Multi-Speaker Models**					
		S2G		**CMix-GAN**		**MUNIT**		**StAGE**		**Mix-StAGE**	
		PCK	**F1**	**PCK**	**F1**	**PCK**	**F1**	**PCK**	**F1**	**PCK**	**F1**
Sitting/Standing	**Mean**	0.35	0.12	0.35	0.25	0.36	0.07	**0.42**	0.18	**0.42**	**0.25**
Sitting	Noah	0.45	0.11	0.45	0.28	0.34	0.09	0.44	0.14	0.44	0.26
Standing	Maher	0.25	0.13	0.24	0.22	0.22	0.07	0.28	0.26	0.26	0.25
Gesture Frequency	**Mean**	0.55	0.41	0.56	0.44	0.34	0.14	**0.58**	0.51	**0.58**	**0.53**
Low	Seth	0.56	0.50	0.58	0.54	0.22	0.02	0.58	0.54	0.59	0.57
High	Oliver	0.54	0.32	0.54	0.34	0.35	0.22	0.54	0.38	0.56	0.42
Body Orientation	**Mean**	0.39	0.14	**0.43**	0.25	0.14	0.05	0.40	**0.42**	0.40	**0.40**
Right	Chemistry	0.35	0.23	0.36	0.27	0.15	0.05	0.37	0.39	0.40	0.39
Left	lec_evol	0.44	0.05	0.50	0.23	0.28	0.34	0.50	0.44	0.49	0.46
Primary Arm Func.	**Mean**	0.43	0.12	0.43	0.33	0.35	0.02	0.59	0.30	**0.61**	**0.37**
Left Arm	lec_cosmic	0.41	0.08	0.41	0.28	0.32	0.06	0.60	0.27	0.62	0.36
Right Arm	lec_cosmic	0.45	0.18	0.45	0.38	0.44	0.12	0.58	0.31	0.60	0.35

Table C.2: Objective metrics for attribute-level style preservation of single-speaker and multi-speaker models as indicated in the columns. Each row refers to the number of speakers the model was trained, with the average performance indicated at the top. The scores for common individual speakers are also indicated below alongside. For detailed results on other speakers please refer to the supplementary

C.2 Other Results, Discussions and Future Directions

These results complement Section 6 of the main paper with a few more observations, explorations and ablation studies. This is followed by potential future directions.

Attribute-Level Style Preservation

Gesture generation for pairs of speakers shows improvements in PCK and F1 scores (see Table C.2) following the trend of perceptual study in Figure 6 of the main paper.

Speaker-Level Style Preservation

Complete numerical results for speaker-level style preservation (for Table 1 in the main paper) are listed in Table C.3. The PCK and F1 scores of the individual speakers show the same trend

No. of Speakers	**Speaker**	**Single-Speaker Models**				**Multi-Speaker Models**					
		S2G		**CMix-GAN**		**MUNIT**		**StAGE**		**Mix-StAGE**	
		PCK	**F1**	**PCK**	**F1**	**PCK**	**F1**	**PCK**	**F1**	**PCK**	**F1**
2	**Mean**	0.25	0.08	0.26	**0.27**	0.24	0.06	0.36	0.21	**0.34**	0.22
	Corden	0.30	0.05	0.32	0.21	0.25	0.06	0.36	0.21	0.34	0.22
	lec_cosmic	0.19	0.12	0.19	0.33	0.15	0.19	0.20	0.48	0.24	0.49
4	**Mean**	0.37	0.18	0.37	0.27	0.22	0.03	0.38	**0.34**	**0.39**	**0.35**
	Corden	0.30	0.05	0.32	0.21	0.24	0.07	0.35	0.27	0.35	0.30
	lec_cosmic	0.19	0.12	0.19	0.33	0.19	0.16	0.18	0.23	0.20	0.19
	ytch_prof	0.43	0.22	0.43	0.22	0.15	0.02	0.42	0.34	0.40	0.32
	Oliver	0.54	0.32	0.54	0.34	0.20	0.09	0.54	0.47	0.55	0.52
8	**Mean**	0.36	0.14	0.37	0.26	0.31	0.15	0.38	**0.32**	**0.40**	**0.33**
	Corden	0.30	0.05	0.32	0.21	0.23	0.03	0.32	0.28	0.36	0.27
	lec_cosmic	0.19	0.12	0.19	0.33	0.13	0.09	0.23	0.34	0.24	0.32
	ytch_prof	0.43	0.22	0.43	0.22	0.39	0.37	0.44	0.39	0.45	0.39
	Oliver	0.54	0.32	0.54	0.34	0.35	0.30	0.54	0.39	0.54	0.46
	Ellen	0.29	0.13	0.30	0.23	0.33	0.17	0.34	0.21	0.33	0.25
	Noah	0.45	0.11	0.45	0.28	0.40	0.23	0.44	0.24	0.44	0.27
	lec_evol	0.44	0.05	0.50	0.23	0.33	0.42	0.45	0.66	0.48	0.66
	Maher	0.25	0.13	0.24	0.22	0.23	0.17	0.25	0.25	0.25	0.25

Table C.3: Objective metrics for speaker-level style preservation of single-speaker and multi-speaker models as indicated in the columns. Each row refers to the number of speakers the model was trained, with the average performance indicated at the top. The scores for individual speakers are also indicated below alongside. * refers to a Single-speaker Model

as the average score for each model.

Impact of value of M on gesture generation

We run an ablation study on the choice of M for the pose decoder. We report the average of PCK and F1 scores in Table C.4 which were calculated for each speaker in single-speaker models. We find the the scores plateau with increasing values of M for single speaker models unlike multi-speaker models like Mix-StAGE .

Exploring Style Control

As a preliminary experiment, we modified the style vector to $[0.5, 0.5]$ in order to mix the styles of two speakers with different Primary Arm Functions. The generated gesture space in Figure C.2 indicates that different speaker styles could be interpolated into a completely new style.

Single Speaker Models	**Metrics**	
	F1 ↑	**PCK** ↑
S2G	18.9	36.6
CMix-GAN ($M = 1$)	26.6	**37.9**
CMix-GAN ($M = 4$)	27.7	36.6
CMix-GAN ($M = 8$)	**28.0**	36.7
CMix-GAN ($M = 12$)	**27.8**	**37.0**

Table C.4: Comparision of Mix-StAGE with different values of M over F1 and PCK. The results are reported as a mean over all speakers in PATS. We can see that the performance for single speaker models does not improve by increasing the number of modes M. This is unlike multi-speaker models, where the addition of sub-generators gives the model an edge over single-speaker models.

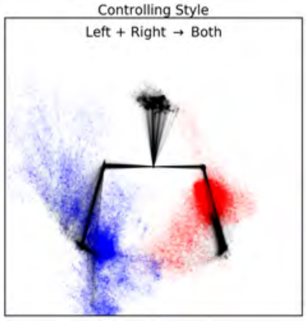


Figure C.2: Heat map of mixture of two styles: primary arm function of left and right mixed to give motion for both hands. Red represents the left hand and blue represents the right hand.

Future Directions

Our efforts were aimed at modeling, disentangling and transferring gesture style under the assumption that the emotional state of the speaker does not affect the gestures. While this is reasonable for speakers in PATS, which are mostly scripted monologues, it may not be true in general hence motivating an interesting future direction. Another direction, that might induce diversity in the generated gestures, is the inclusion of verbal information (i.e. natural language). This may not be trivial in context of style transfer as the difference in vocabulary of different speakers could create an unwanted bias - some words might get associated with certain styles of gesturing.

C.3 Implementation Details

This section gives more detail about the exact architectures used for our model also described in Section 4 of the main paper.

C.3.1 Network Architectures

Figure C.3 and C.4 consists of the visual representation of the architectures used for our model Mix-StAGE . For the decoder $\bigoplus$ is a weighted sum as described in Equation (2) of the main paper. Every operation is a **1D-convolution** followed by a **Batch-Norm** and finally **ReLU**. Each convolution uses a kernal size of 3 and hop length of 1, except for cases where temporal dimension is downsampled where the kernal size is 4 and hop length is 2.

C.3.2 Training Details

We use Adam [85] to optimize the model with a exponentially decaying learning rate of 0.001. We train each model for 60000 iterations while check-pointing every 3000 iterations. Finally, we choose the best model based on loss on the development set. We use $\lambda_{id} = 0.1$ to prevent the style consistency loss from stealing focus while training the pose gesture generator.

C.3.3 Human study experiments

We conducted our human studies on Amazon Mechanical Turk (or AMT), for which we used 100 random videos for each speaker which gave us 2400 pairs of comparisons per model for

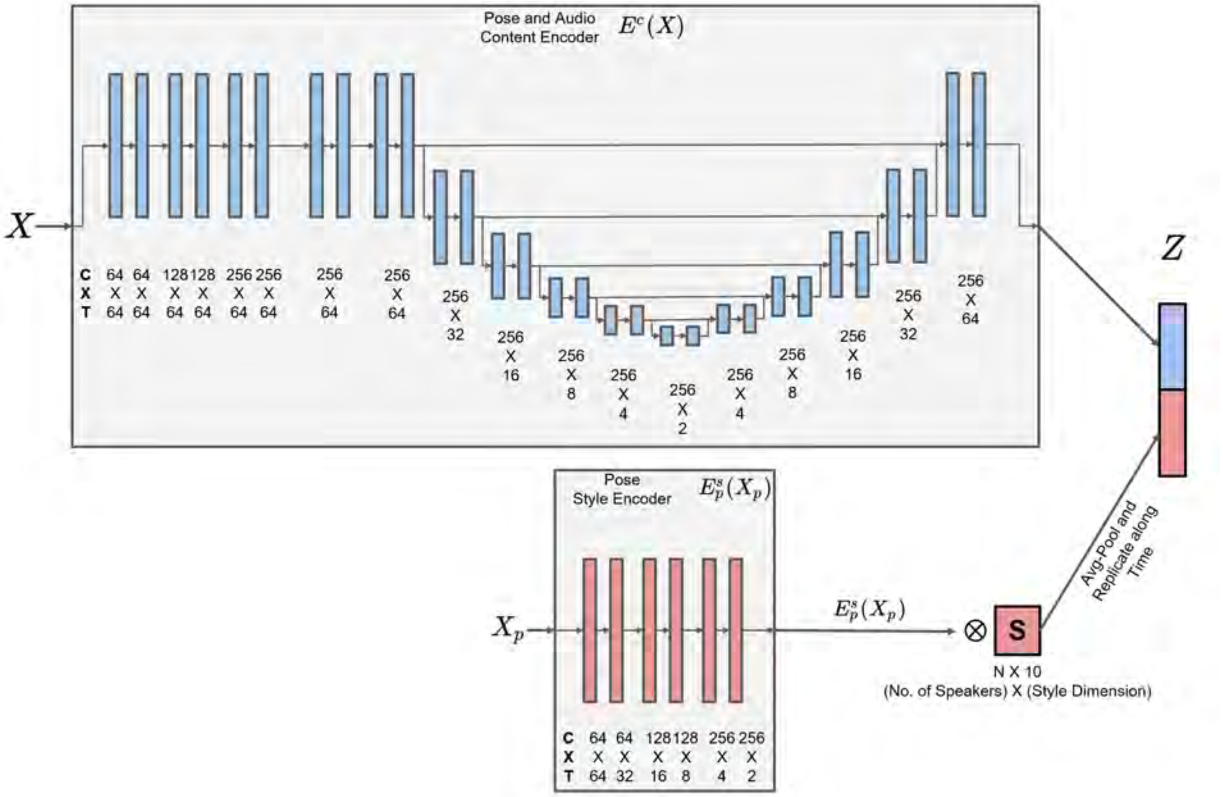


Figure C.3: Encoder Architecture

each study (out of 5). This is a significant number of comparisons and helps with reliability of the results. Each annotation task contained 20 videos and is performed by 3 different users; hence we had approximately (2400*3/20) = 360 participants.

To help filter unreliable annotators, we use two ground truth videos from the same speaker with the same style as control samples. If annotators tag these two videos as different styles, then we disregard this annotation set as unreliable.

Sample study for style transfer (other studies also follow similar method)
Two videos are shown to the user. One video is a ground truth(Speaker A Style A) and the other is generated by a model. The generated video could be either of (a) Speaker A Style A or (b) Speaker B Style A. We ask two questions to measure correctness of style transfer and naturalness:

1. Do the animations have different styles of gestures?
2. Which of the videos 1 or 2 has more Natural gestures with respect to the audio?

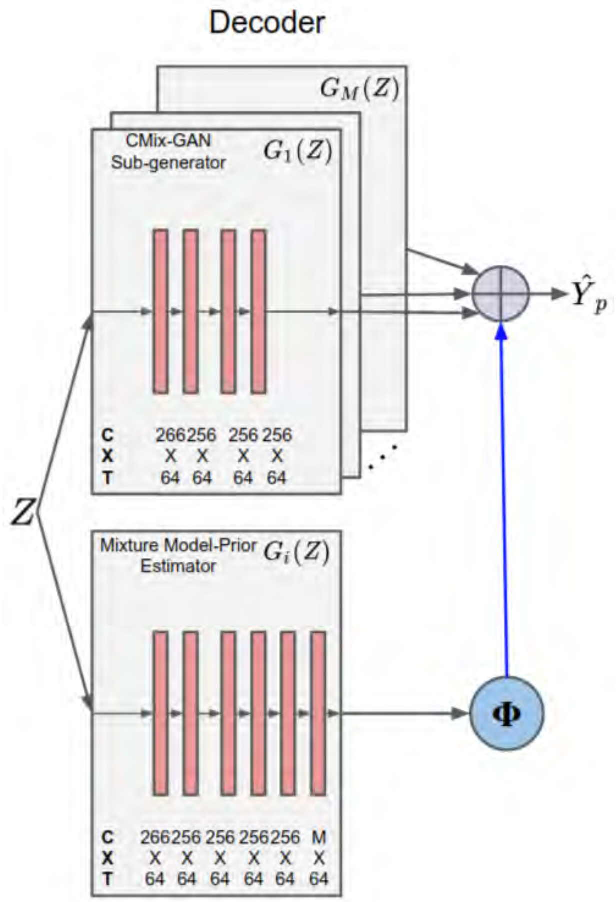


Figure C.4: Decoder Architecture

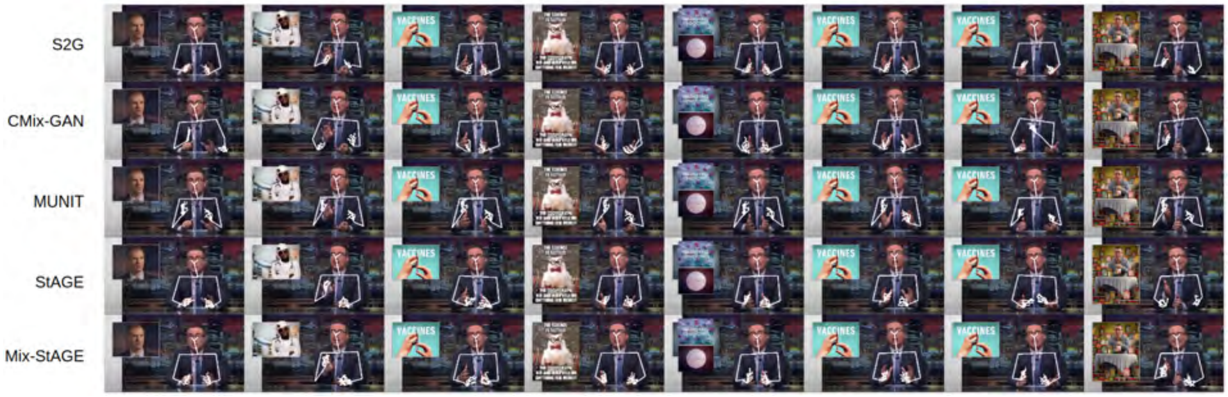
C.4 Videos: Style Transfer and Preservation

We refer the readers to `http://chahuja.com/mix-stage` for demo videos.

C.5 Video Frames: Style Preservation Qualitative Results

These results complement Figure 8 in the main paper. We plot some more animation figures generated by random audio samples in the test-set to provide some more samples for qualitative judgment in Figures C.5 and C.6.

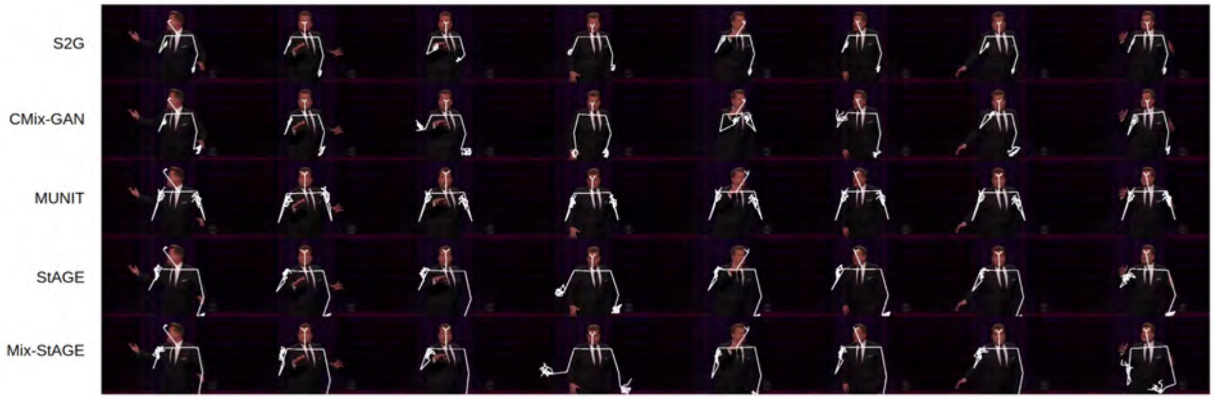
Figure C.5: Animation depicted as a series of frames for different speakers. The vertical axis is labeled as models and horizontal axis is time. The generated animation is superimposed over the ground truth video.



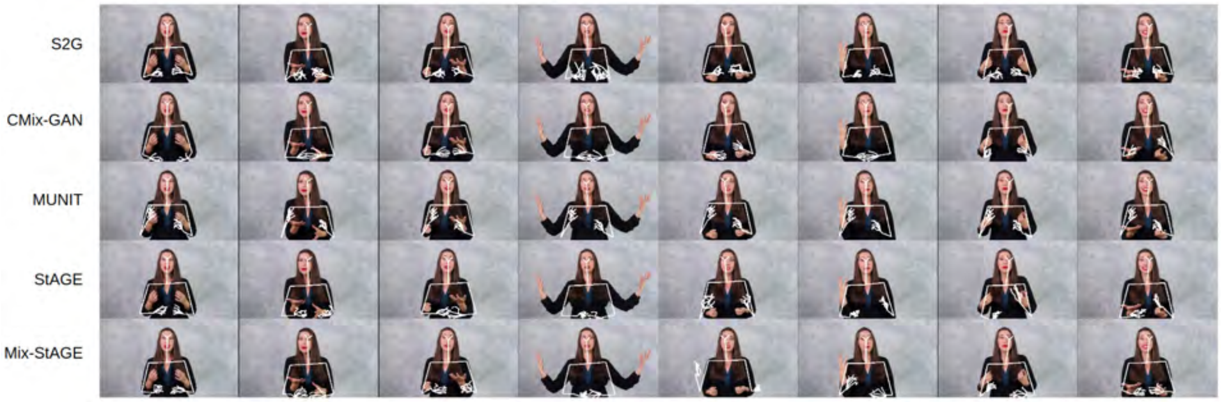
oliver

lec_cosmic

Figure C.6: Animation depicted as a series of frames for different speakers. The vertical axis is labeled as models and horizontal axis is time. The generated animation is superimposed over the ground truth video.



corden



ytch_prof

Appendix D

Co-Speech Gestures: Low Resources for Grounding and Gesture Style

D.1 Additional results

D.1.1 Impact of additional unsupervised input data on the quality of outputs

We also run some experiments with additional unsupervised data in form of three semi-supervised training paradigms: (1) *Low-resource + Unsupervised Source Data*, (2) *Low-resource + Unsupervised Target Data* and (3) *Low-resource + Unsupervised Source and Target Data*.

Additional unsupervised target or source data exposes the model to a larger variety of data in the input space. This makes the model more robust to changes in the input space and consequently induces a positive effect on quality of the gestures. Table D.1 has experiments with additional unsupervised data where we observe relatively better values of FID when additional unsupervised source and target data is also used. Furthermore, the gestures are deemed more natural by human annotators in Table D.1 when additional unsupervised target data is used. Unsurprisingly, we do not observe any practical change in PCK values as no additional information about grounding is being provided by unsupervised input data.

	FID ↓		PCK ↑		Human Perceptual Study ↑	
Training Data	**maher** **↓** **oliver**	**maher** **↓** **chemistry**	**maher** **↓** **oliver**	**maher** **↓** **chemistry**	**Naturalness**	**Style**
Low-resource	15.0 ± 5.2	31.7 ± 3.8	0.46 ± 0.01	0.32 ± 0.02	**21.9 ± 2.5**	46.3 ± 9.2
Unsp. Source + Low-resource	18.8 ± 7.4	**27.7 ± 1.5**	0.44 ± 0.02	0.33 ± 0.01	18.5 ± 9.9	44.4 ± 16.2
Unsp. Target + Low-resource	14.1 ± 4.7	**27.2 ± 3.1**	0.46 ± 0.01	0.34 ± 0.01	**26.7 ± 9.5**	**60.0 ± 17.7**
Unsp. Source + Unsp. Target + Low-resource	**12.6 ± 4.1**	28.2 ± 1.8	0.46 ± 0.0	0.33 ± 0.01	20.8 ± 10.5	48.3 ± 21.8
High-resource	16.1	8.7	0.49	0.39	-	-

Table D.1: **Evaluating impact of additional unsupervised data**: DiffGAN is trained on 10 minuites of low-resource data, followed by additional unsupervised data of the source and/or target domain. The metrics measure its impact on output domain shift (i.e. FID and Style), crossmodal grounding (i.e. PCK) and quality of gestures (i.e. naturalness).

D.1.2 Qualitative visualizations

A more exhaustive set of qualitative results is shown in form of visual histograms in Figures D.2, D.3, D.4. Furthermore, we also plot of the velocity distributions across our model and different baselines in Figure D.5. Additionally, we overlay the generated gestures by different models on the original video to qualitatively verfiy the correctness of the generated gestures in Figures D.7, D.6, D.8 and D.9.

D.1.3 Quantitative results

A more exhaustive set of quantitative results on FID scores [7, 61, 196] (i.e., measure of distribution of generated gestures) and PCK scores [10, 160] (i.e., measure of relevance of generated gestures to input language) is shown in Table D.2.

D.2 Experimental setup details

D.2.1 Model hyperparameters

We use, as our source model, an architecture described in Ahuja et al. [7] and illustrated in Figure D.1. We optimize our model using Adam [85] with a learning rate of 0.0001, decay rates $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We update the source model for 4000 iterations with a batch size of 32, which is more that enough to reach convergence. Our experiments were all run on either NVIDIA Titan 1080 or NVIDIA GeForce RTX 2080 and took around 1-2 hours. Unless explicitly specified, the adaptation experiment was performed on 10 minutes of low-resource target supervised data with

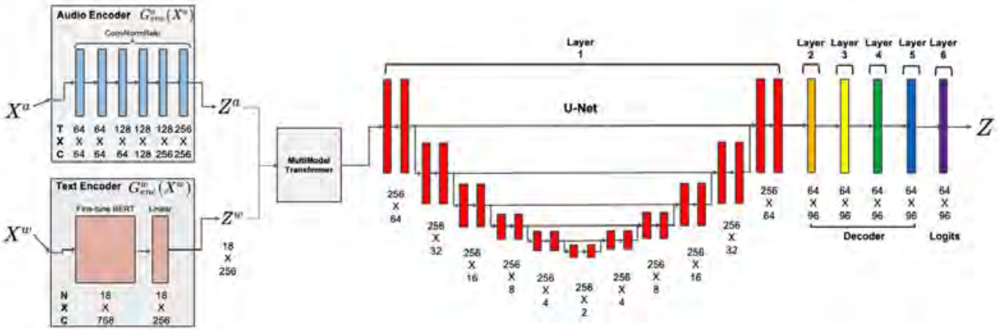


Figure D.1: Illustration of the source model [7] for our experiments in the main paper. The layer IDs from 1 through 6 are the possible choices for layer l for the low resource adaptation

our full DiffGAN model. We provide a copy of the code that we use to run all experiments in `supplementary_code.zip`

Choice of Layer l The choice of layer l as the ideal set of parameters for optimal low-resource adaptation is tricky. We show through an ablation study (in the main paper) that the second-to-last layer (i.e. layer 5 in Figure D.1) of the source model generator [7] is the best choice.

D.2.2 Dataset

Pose Audio Transcripts (PATS) dataset [7, 8, 50] contains aligned transcribed language, audio, gesture data for 25 speakers from different domains in academia, social media, television. In consists of 251 hours of data, with a mean of 10.7 seconds and a standard deviation of 13.5 seconds per interval. This dataset provides a train, test and validation split which we use for our experiments. PATS dataset is licensed under a Creative Commons Attribution-NonCommercial 2.0 Generic License.

We choose five speakers `oliver`, `maher`, `chemistry`, `ytch_prof` and `lec_evol`, that have different domains of output gestures as well as a diverse linguistic input domain. For the `source` domain, we use the speakers `oliver` and `maher` individually. The full list of combinations is the following (`source` → `target` denotes the source to target domain):

- `oliver` → `maher`
- `oliver` → `chemistry`
- `oliver` → `ytch_prof`

- `oliver` → `lec_evol`
- `maher` → `oliver`
- `maher` → `chemistry`
- `maher` → `ytch_prof`
- `maher` → `lec_evol`

D.2.3 Human perceptual study setup

Sample study for Style (aka Domain Shift) We show 3 reference ground videos of a target speaker. Then, two videos are shown to the user, one video is a ground truth and the other is generated by a model (in a low-resource setting). The generated video does not have audio. We ask a single question to measure the correctness of the domain shift by looking at the style: Gesture style is defined by the gesture's extent, frequency, timing, and position of the body in relation to speech. Which video has the same style of gestures as the style shown in the Reference Videos?

Sample study for Subjective Metrics 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 in terms of subjective metrics (Timing, Relevance, Expressiveness and Naturalness) [7].

We attach a screenshot of a sample study and the questions asked to the users. The estimated hourly wage for the annotators is around **9 USD an hour**. The definitions of the subjective metrics are listed below, and a screenshot of this experiment is shown in Figure D.10 and D.11.

Definitions:

- **Style:** Gesture style is defined by the gesture's extent, frequency, timing, and position of the body in relation to speech.
- **Extent:** Gesture extent is the space around the speaker that the speakers' gestures (hand/arms) cover.
- **Frequency:** Gesture frequency is the rate at which the speakers use gestures.
- **Timing:** People tend to emphasize on their hand gestures when they emphasize what they are saying. Timing is best when the gestures align (i.e., occur simultaneously) with the relevant spoken words. These two events occur simultaneously for the timing to be correct.

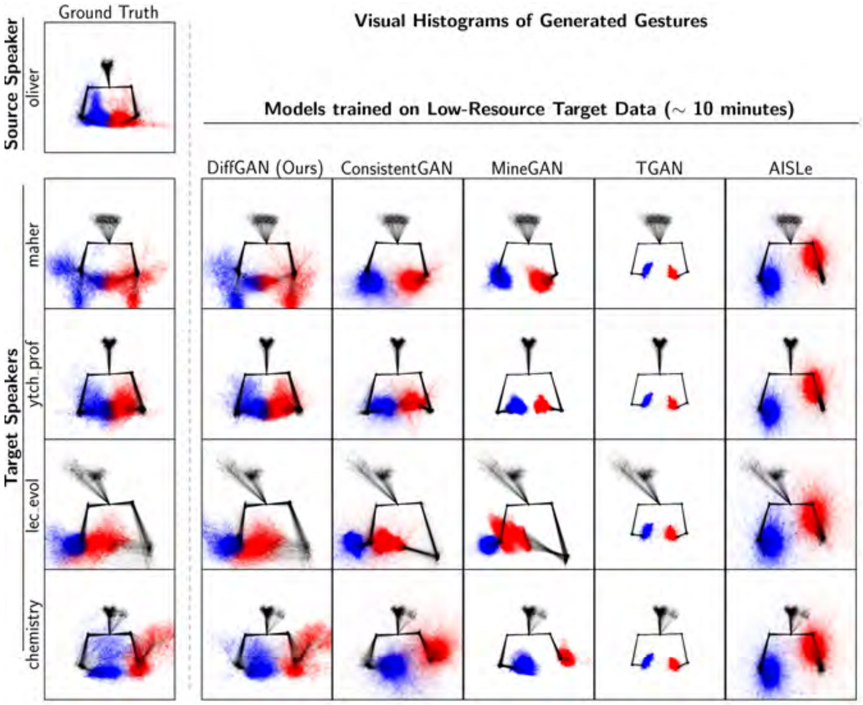


Figure D.2: Visual Histograms of generated gestures visually describe the distribution of hand gestures in space. Red and blue colors denote the left and right arms respectively. First row is the source speaker, below which we have all the target speakers. Each column denotes a model which adapts output distribution of the source domain to the target domain.

- **Relevance:** The form of the gesture should not only be well timed (as judge with the Timing metric) but also seem to be the right gesture, relevant to the spoken words. For example, if a person says ”me”, and simultaneously points towards themselves, then the gesture is relevant.
- **Expressiveness:** Expressiveness is a general measure of the amount of gestures. It is not only about the number of gestures but also about the size of these gestures. More and larger gestures will represent more expressiveness.
- **Naturalness:** This is a general metric which asks you to judge if the animation looks natural, as if it was the depiction of a real person. Naturalness involves both the body and gestures, as well as how they appear in relation with the spoken words. The gestures need to look natural.

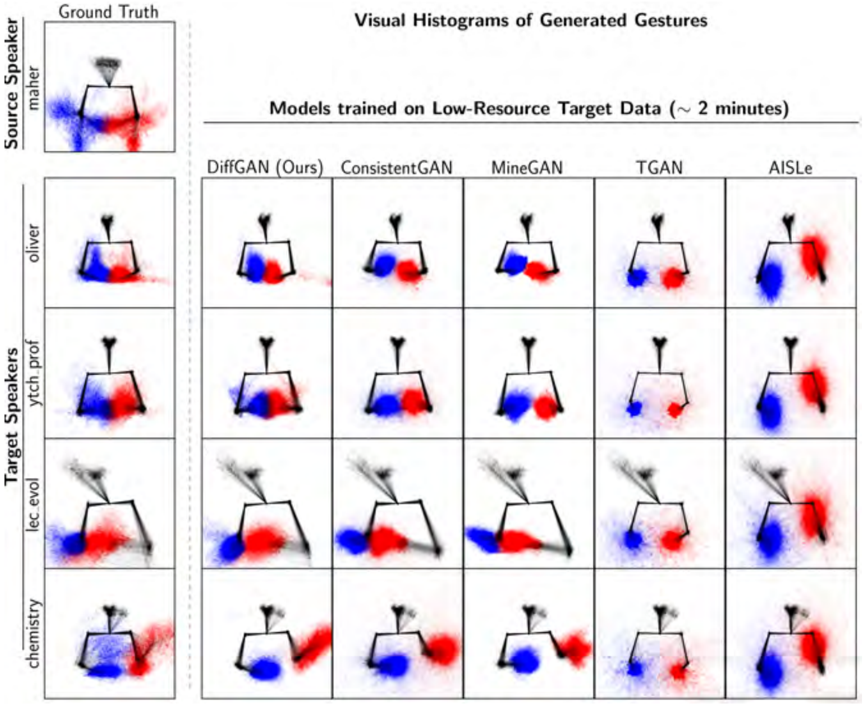


Figure D.3: Visual Histograms of generated gestures visually describe the distribution of hand gestures in space. Red and blue colors denote the left and right arms respectively. First row is the source speaker, below which we have all the target speakers. Each column denotes a model which adapts output distribution of the source domain to the target domain.

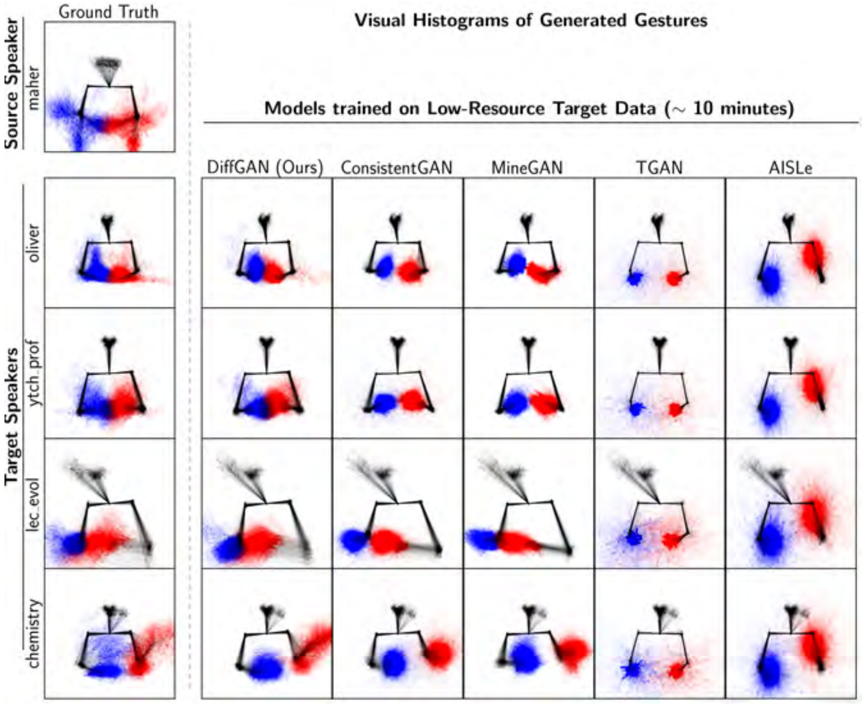


Figure D.4: Visual Histograms of generated gestures visually describe the distribution of hand gestures in space. Red and blue colors denote the left and right arms respectively. First row is the source speaker, below which we have all the target speakers. Each column denotes a model which adapts output distribution of the source domain to the target domain.

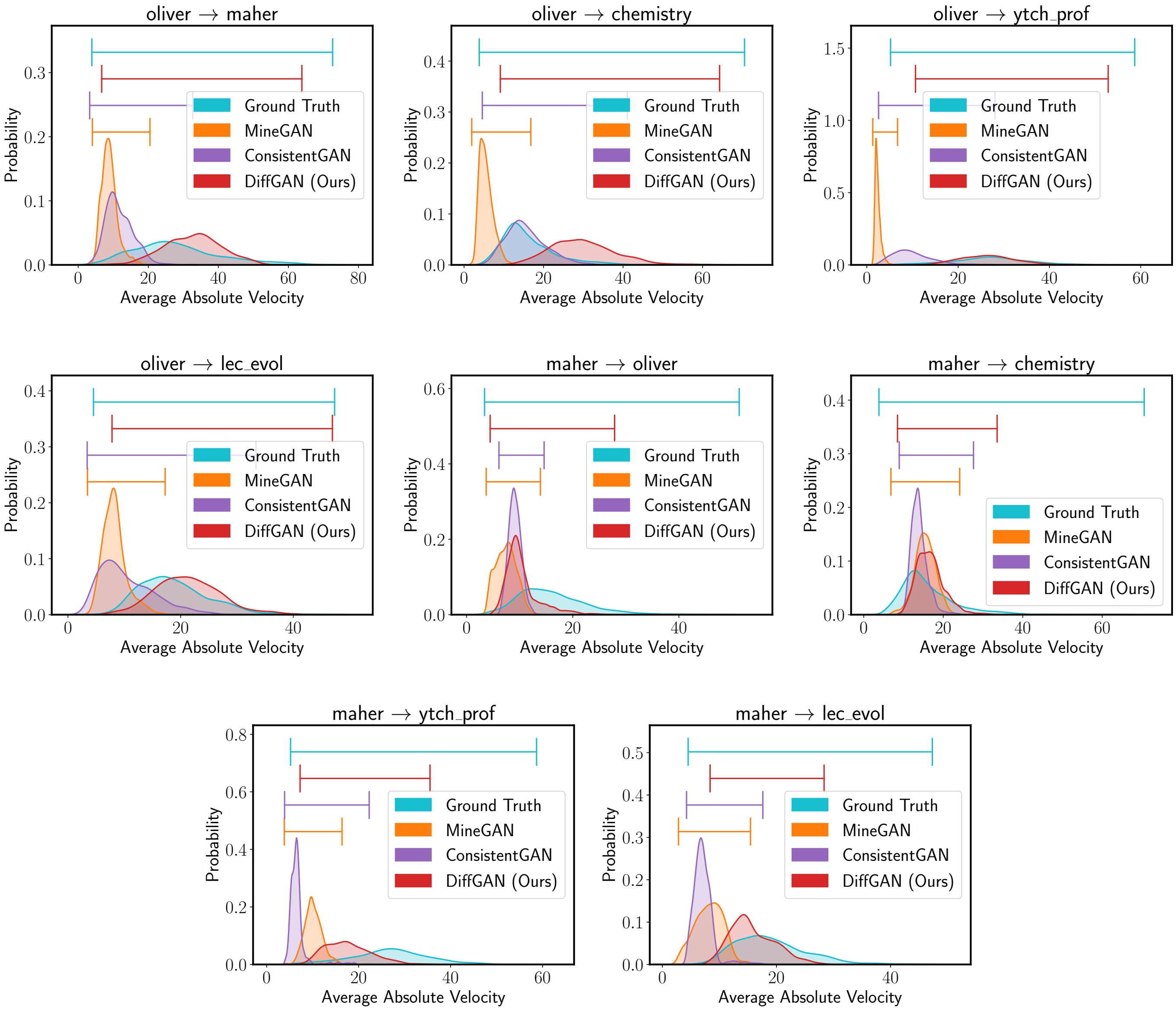


Figure D.5: Distribution of the generated gestures with average absolute velocity as the statistic for source to target domain adaptation. The support (or coverage) of the distribution is denoted with the colour coded lines at the top of each plot. Larger overlap of a model's distribution with the ground truth distribution is desirable. All the experiments for these plots were conducted with 10 minutes of target domain data.

		FID ↓								PCK ↑							
Amt. of Data (minutes)	**Models**	**maher ↓ oliver**	**maher ↓ chemistry**	**maher ↓ ytch_prof**	**maher ↓ lec_evol**	**oliver ↓ maher**	**oliver ↓ chemistry**	**oliver ↓ ytch_prof**	**oliver ↓ lec_evol**	**maher ↓ oliver**	**maher ↓ chemistry**	**maher ↓ ytch_prof**	**maher ↓ lec_evol**	**oliver ↓ maher**	**oliver ↓ chemistry**	**oliver ↓ ytch_prof**	**oliver ↓ lec_evol**
2	**Overfit**	49.2 ± 0.8	84.5 ± 3.1	45.5 ± 1.6	205.8 ± 4.0	83.7 ± 2.6	84.5 ± 3.1	45.5 ± 1.6	205.8 ± 4.0	0.18 ± 0.01	0.2 ± 0.0	0.24 ± 0.0	0.17 ± 0.01	0.18 ± 0.0	0.2 ± 0.0	0.24 ± 0.0	0.17 ± 0.01
	TGAN [179]	57.5 ± 2.4	184.3 ± 5.4	79.1 ± 0.3	247.0 ± 11.3	339.1 ± 1.2	323.5 ± 1.9	138.0 ± 0.3	737.5 ± 2.8	0.31 ± 0.01	0.23 ± 0.0	0.29 ± 0.0	0.27 ± 0.01	0.2 ± 0.0	0.25 ± 0.0	0.38 ± 0.0	0.14 ± 0.0
	MineGAN [180]	42.5 ± 2.5	157.5 ± 10.6	75.8 ± 4.4	237.7 ± 15.4	290.3 ± 7.2	302.5 ± 6.8	119.4 ± 5.8	725.1 ± 1.5	0.38 ± 0.03	0.26 ± 0.02	0.38 ± 0.02	0.41 ± 0.04	0.21 ± 0.01	0.31 ± 0.01	0.45 ± 0.01	0.41 ± 0.03
	ConsistentGAN [135]	61.0 ± 3.2	194.2 ± 15.6	83.1 ± 2.3	247.1 ± 10.0	320.1 ± 11.5	325.6 ± 44.3	132.0 ± 6.2	735.0 ± 9.2	0.39 ± 0.01	0.27 ± 0.01	0.4 ± 0.01	0.39 ± 0.03	0.21 ± 0.01	0.25 ± 0.01	0.4 ± 0.01	0.38 ± 0.03
	DiffGAN (Ours)	25.0 ± 3.7	42.8 ± 5.1	18.2 ± 5.3	79.0 ± 14.0	47.9 ± 25.5	48.2 ± 15.9	24.7 ± 5.0	181.3 ± 42.0	0.45 ± 0.02	0.31 ± 0.01	0.4 ± 0.03	0.42 ± 0.01	0.26 ± 0.01	0.29 ± 0.01	0.35 ± 0.02	0.39 ± 0.01
	DiffGAN w/o $\mathcal{L}_{diff}$	30.6 ± 4.8	38.7 ± 4.3	14.9 ± 1.9	81.7 ± 4.9	42.6 ± 24.2	45.1 ± 12.8	24.5 ± 4.2	184.3 ± 51.2	0.46 ± 0.01	0.3 ± 0.02	0.38 ± 0.02	0.42 ± 0.02	0.25 ± 0.01	0.3 ± 0.01	0.34 ± 0.02	0.38 ± 0.02
	DiffGAN w/o $\mathcal{L}_{shift}$	16.0 ± 1.2	41.2 ± 5.2	13.4 ± 0.2	103.4 ± 9.1	93.8 ± 14.3	111.2 ± 6.6	34.5 ± 4.3	267.3 ± 40.9	0.38 ± 0.02	0.27 ± 0.01	0.35 ± 0.01	0.35 ± 0.0	0.22 ± 0.01	0.28 ± 0.0	0.35 ± 0.02	0.31 ± 0.0
10	**Overfit**	47.1 ± 0.2	85.0 ± 1.9	44.5 ± 0.5	204.1 ± 1.6	80.3 ± 0.8	85.0 ± 1.9	44.5 ± 0.5	204.1 ± 1.6	0.18 ± 0.0	0.21 ± 0.0	0.24 ± 0.0	0.17 ± 0.0	0.19 ± 0.0	0.21 ± 0.0	0.24 ± 0.0	0.17 ± 0.0
	TGAN [179]	62.9 ± 1.8	191.7 ± 1.2	79.0 ± 2.0	248.4 ± 8.5	341.5 ± 0.4	326.8 ± 1.6	139.1 ± 0.2	739.4 ± 2.5	0.3 ± 0.01	0.22 ± 0.01	0.29 ± 0.01	0.27 ± 0.02	0.2 ± 0.0	0.24 ± 0.0	0.38 ± 0.0	0.14 ± 0.0
	MineGAN [180]	41.4 ± 3.1	145.0 ± 14.1	67.2 ± 2.4	233.9 ± 13.3	293.5 ± 12.7	318.2 ± 5.4	127.0 ± 2.0	736.7 ± 12.2	0.4 ± 0.01	0.24 ± 0.03	0.4 ± 0.01	0.42 ± 0.01	0.22 ± 0.02	0.31 ± 0.01	0.45 ± 0.02	0.4 ± 0.01
	ConsistentGAN [135]	63.1 ± 1.9	188.2 ± 17.0	76.6 ± 2.9	259.5 ± 10.5	322.2 ± 2.4	327.0 ± 18.7	135.8 ± 2.5	740.5 ± 5.0	0.41 ± 0.03	0.28 ± 0.04	0.39 ± 0.01	0.4 ± 0.02	0.22 ± 0.01	0.27 ± 0.01	0.4 ± 0.01	0.4 ± 0.01
	DiffGAN (Ours)	15.0 ± 5.2	31.7 ± 3.8	15.9 ± 4.2	40.4 ± 13.0	30.3 ± 6.9	24.3 ± 4.2	19.1 ± 3.7	54.2 ± 17.5	0.46 ± 0.01	0.32 ± 0.02	0.42 ± 0.01	0.44 ± 0.01	0.26 ± 0.01	0.3 ± 0.01	0.38 ± 0.01	0.43 ± 0.01
	DiffGAN w/o $\mathcal{L}_{diff}$	19.0 ± 7.7	29.0 ± 1.6	18.5 ± 6.5	26.7 ± 13.7	25.9 ± 4.3	27.5 ± 6.7	19.5 ± 4.0	48.6 ± 12.3	0.46 ± 0.01	0.32 ± 0.01	0.43 ± 0.02	0.43 ± 0.01	0.26 ± 0.01	0.3 ± 0.01	0.38 ± 0.01	0.42 ± 0.0
	DiffGAN w/o $\mathcal{L}_{shift}$	22.4 ± 2.1	54.1 ± 1.9	19.6 ± 0.1	100.5 ± 2.2	119.1 ± 7.0	131.0 ± 5.1	56.8 ± 2.8	250.5 ± 24.7	0.42 ± 0.02	0.31 ± 0.01	0.37 ± 0.0	0.37 ± 0.0	0.23 ± 0.0	0.31 ± 0.0	0.39 ± 0.0	0.33 ± 0.01
	Train on high-resource data	16.1	8.7	12.5	25.6	10.2	8.7	12.5	25.6	0.49	0.39	0.39	0.45	0.27	0.39	0.39	0.45

Table D.2: Comparison of our DiffGAN with prior work for low-resource crossmodal generative modeling from source to target speakers to evaluate output domain shift (i.e. FID) and crossmodal grounding (i.e. PCK)

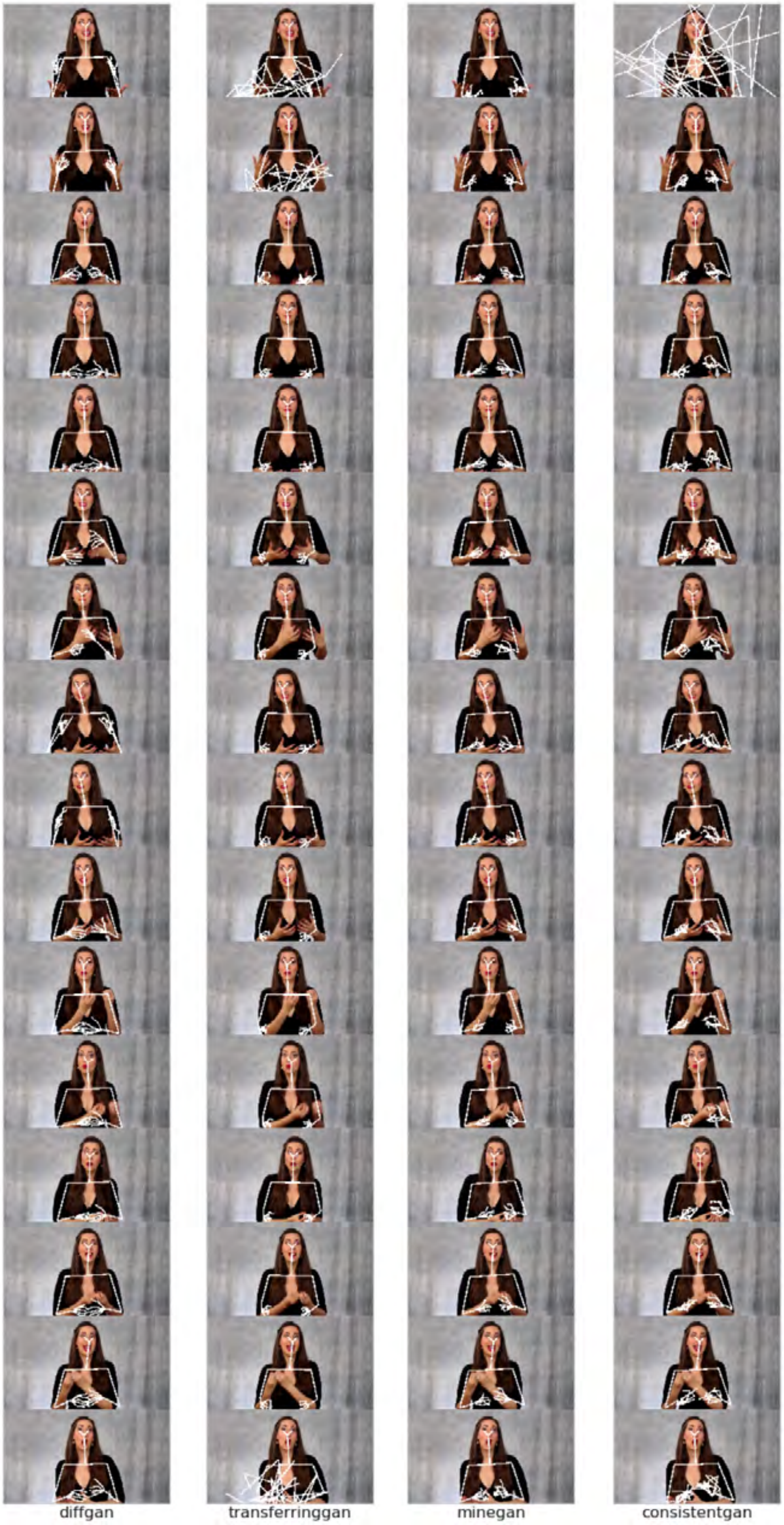


Figure D.6: Qualitative comparison of our DiffGAN with prior work over shape of generated gestures. With maher as the source domain, target model outputs over the target domain ytch_prof are superimposed over ground truth video frames for easy comparison.

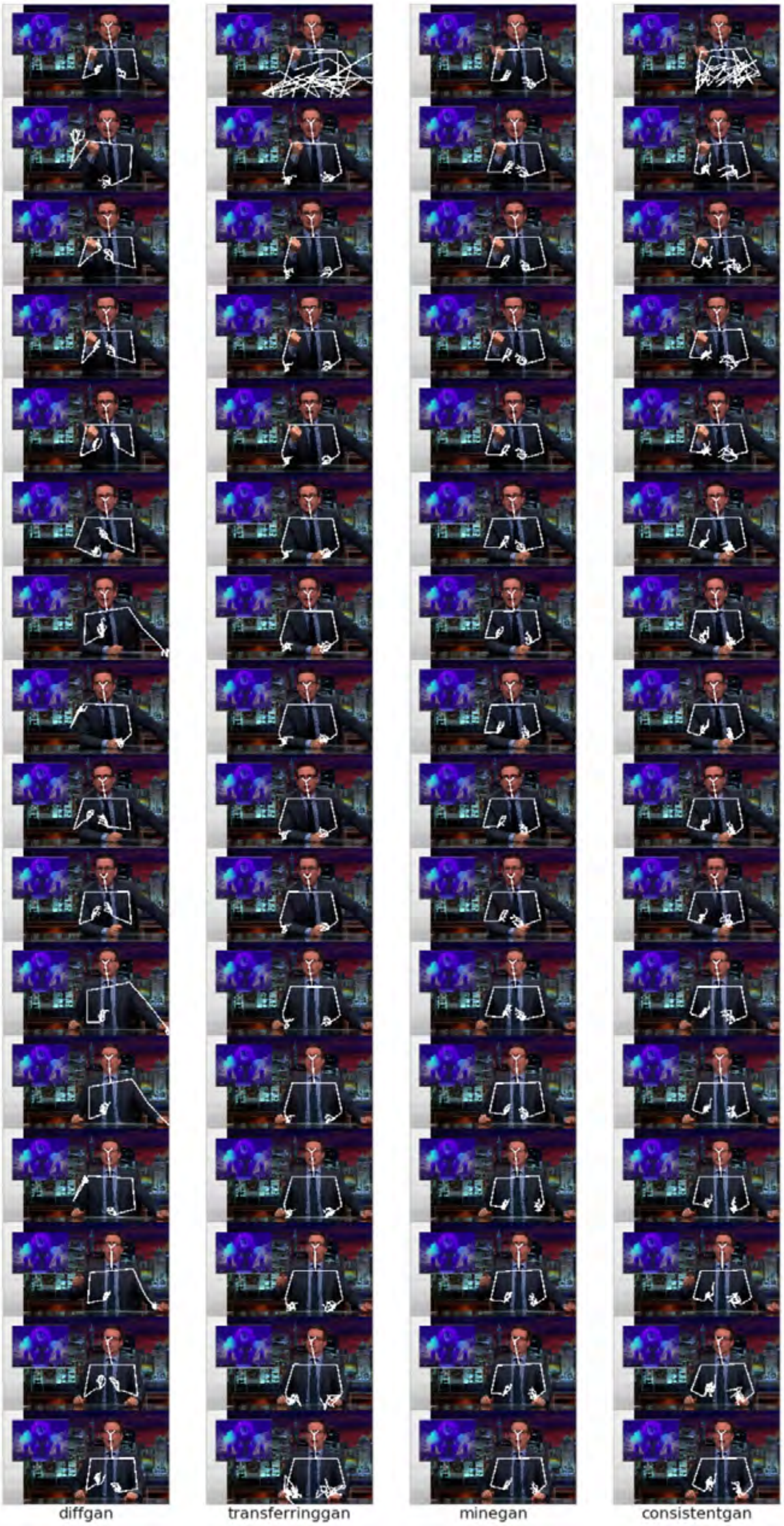


Figure D.7: Qualitative comparison of our DiffGAN with prior work over shape of generated gestures. With `maher` as the source domain, target model outputs over the target domain `oliver` are superimposed over ground truth video frames for easy comparison.

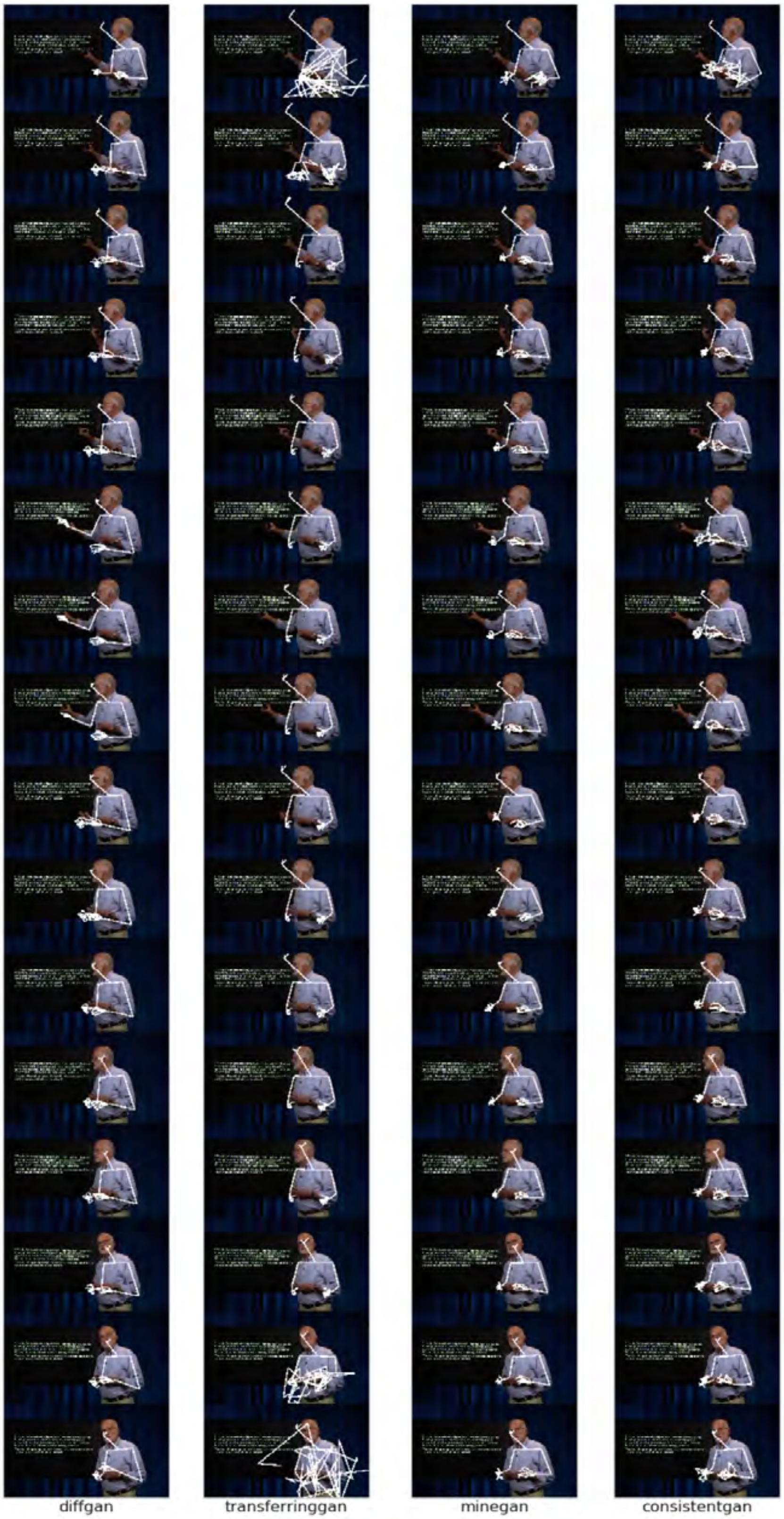


Figure D.8: Qualitative comparison of our DiffGAN with prior work over shape of generated gestures. With `maher` as the source domain, target model outputs over the target domain `lec_evol` are superimposed over ground truth video frames for easy comparison.

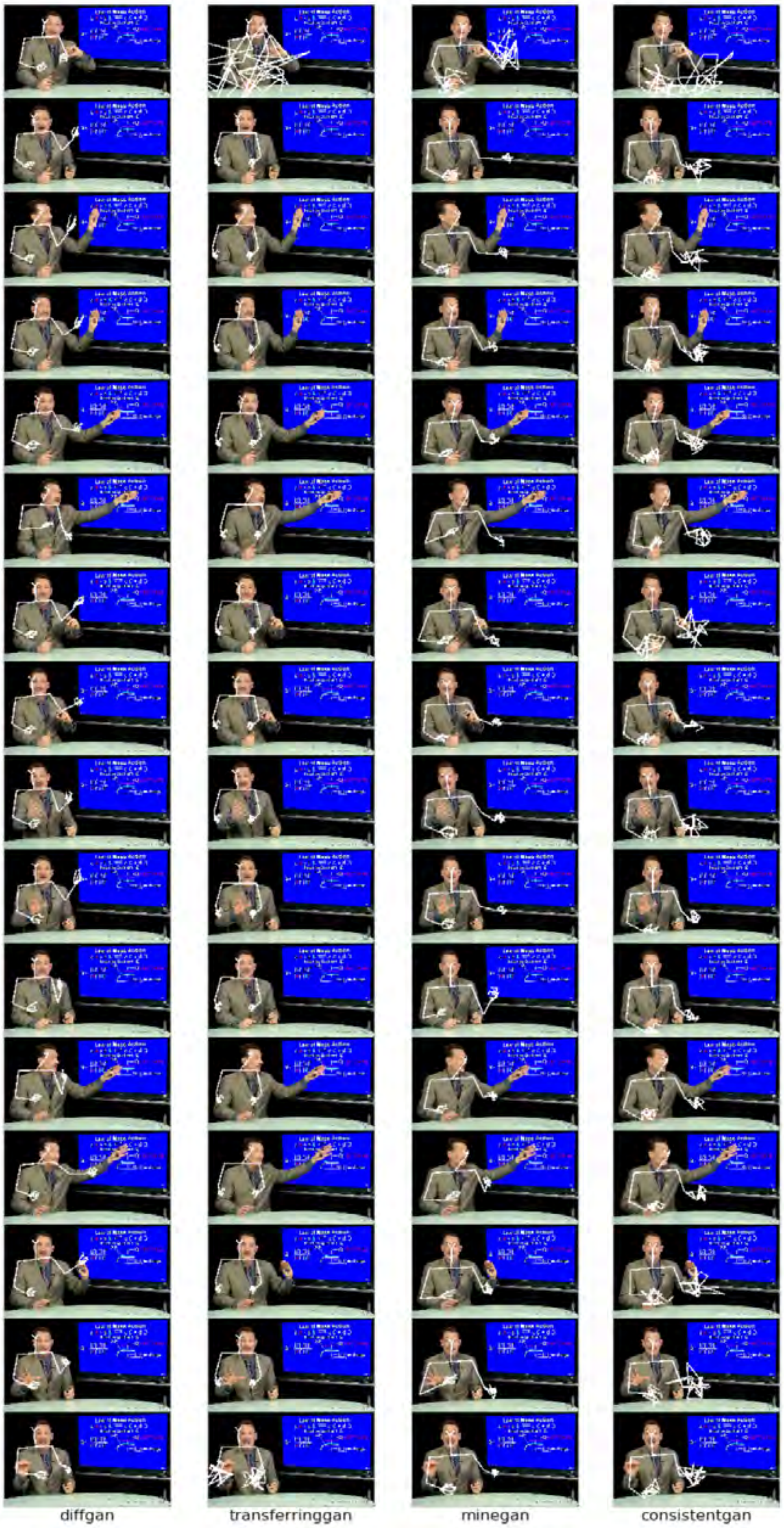
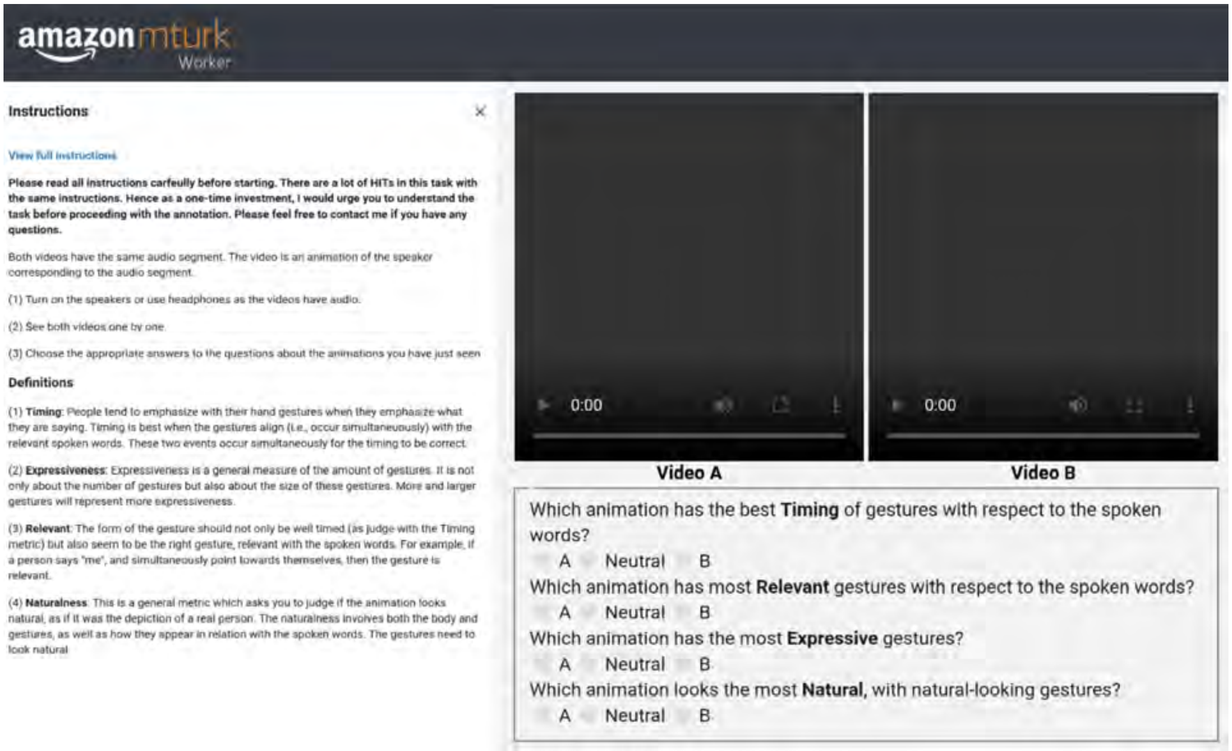


Figure D.9: Qualitative comparison of our DiffGAN with prior work over shape of generated gestures. With `maher` as the source domain, target model outputs over the target domain `chemistry` are superimposed over ground truth video frames for easy comparison.



Instructions

View full instructions

Please read all instructions carfeully before starting. There are a lot of HITs in this task with the same instructions. Hence as a one-time investment, I would urge you to understand the task before proceeding with the annotation. Please feel free to contact me if you have any questions.

Both videos have the same audio segment. The video is an animation of the speaker corresponding to the audio segment.

(1) Turn on the speakers or use headphones as the videos have audio.

(2) See both videos one by one.

(3) Choose the appropriate answers to the questions about the animations you have just seen

Definitions

(1) **Timing**: People tend to emphasize with their hand gestures when they emphasize what they are saying. Timing is best when the gestures align (i.e., occur simultaneuously) with the relevant spoken words. These two events occur simultaneously for the timing to be correct.

(2) **Expressiveness**: Expressiveness is a general measure of the amount of gestures. It is not only about the number of gestures but also about the size of these gestures. More and larger gestures will represent more expressiveness.

(3) **Relevant**: The form of the gesture should not only be well timed (as judge with the Timing metric) but also seem to be the right gesture, relevant with the spoken words. For example, if a person says "me", and simultaneously point towards themselves, then the gesture is relevant.

(4) **Naturalness**: This is a general metric which asks you to judge if the animation looks natural, as if it was the depiction of a real person. The naturalness involves both the body and gestures, as well as how they appear in relation with the spoken words. The gestures need to look natural

Video A **Video B**

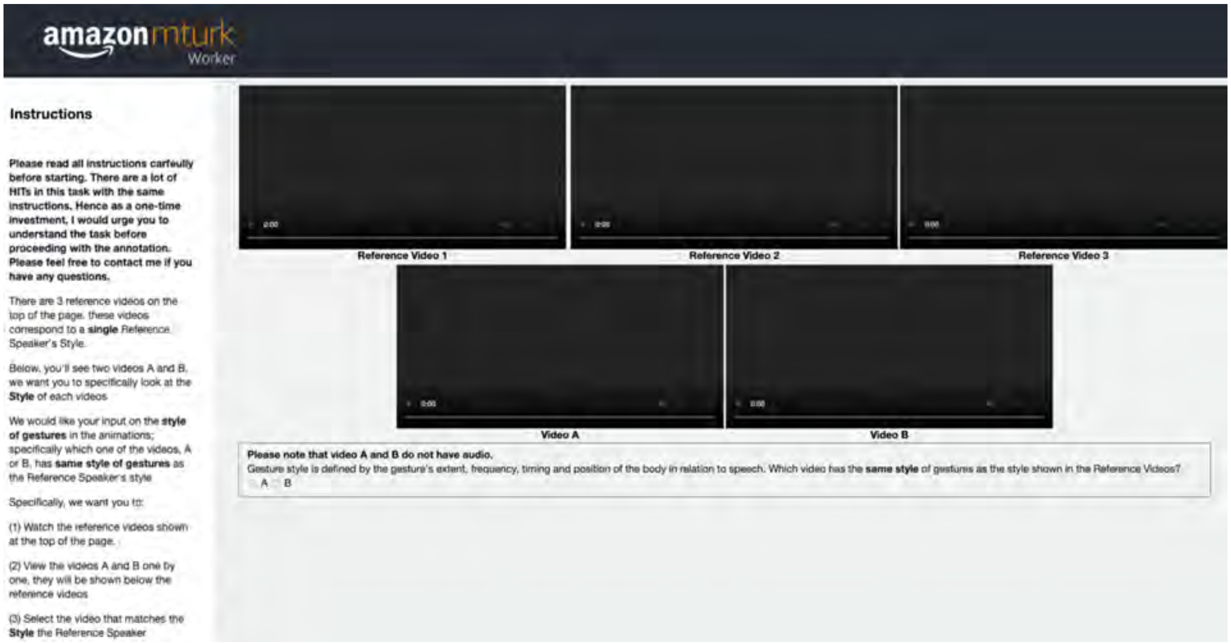
Which animation has the best **Timing** of gestures with respect to the spoken words?
○ A ○ Neutral ○ B

Which animation has most **Relevant** gestures with respect to the spoken words?
○ A ○ Neutral ○ B

Which animation has the most **Expressive** gestures?
○ A ○ Neutral ○ B

Which animation looks the most **Natural,** with natural-looking gestures?
○ A ○ Neutral ○ B

Figure D.10: Screenshot of MTurk Experiment used to measure subjective metrics



Instructions

Please read all instructions carfeully before starting. There are a lot of HITs in this task with the same instructions. Hence as a one-time investment, I would urge you to understand the task before proceeding with the annotation. Please feel free to contact me if you have any questions.

There are 3 reference videos on the top of the page. these videos correspond to a **single** Reference Speaker's Style.

Below, you'll see two videos A and B. we want you to specifically look at the **Style** of each videos.

We would like your input on the **style of gestures** in the animations; specifically which one of the videos, A or B, has **same style of gestures** as the Reference Speaker's style

Specifically, we want you to:

(1) Watch the reference videos shown at the top of the page.

(2) View the videos A and B one by one, they will be shown below the reference videos

(3) Select the video that matches the **Style** the Reference Speaker

Reference Video 1 **Reference Video 2** **Reference Video 3**

Video A **Video B**

Please note that video A and B do not have audio.
Gesture style is defined by the gesture's extent, frequency, timing and position of the body in relation to speech. Which video has the **same style** of gestures as the style shown in the Reference Videos?
○ A ○ B

Figure D.11: Screenshot of MTurk Experiment used to measure style metrics

Bibliography

[1] Natasha Abner, Kensy Cooperrider, and Susan Goldin-Meadow. Gesture for linguists: A handy primer. *Language and linguistics compass*, 9(11):437–451, 2015. 1

[2] Shailen Agrawal and Michiel van de Panne. Task-based locomotion. *ACM Transactions on Graphics (TOG)*, 35(4):82, 2016. 2.6

[3] Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. Text2action: Generative adversarial synthesis from language to action. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–5. IEEE, 2018. 2.1, 2.6

[4] Chaitanya Ahuja and Louis-Philippe Morency. Lattice recurrent unit: Improving convergence and statistical efficiency for sequence modeling. In *AAAI-18*, pages 4996–5003, 2018. URL `https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17394`. 1.2.1

[5] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. *2019 3DV,*, 2019. `http://chahuja.com/language2pose`. 1.2, 4.3.2

[6] Chaitanya Ahuja, Shugao Ma, Louis-Philippe Morency, and Yaser Sheikh. 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, 2019. 1.2.1, 3.3.3, 3.6, 4.6, 5.6

[7] Chaitanya Ahuja, Dong Won Lee, Ryo Ishii, and Louis-Philippe Morency. No gestures left behind: Learning relationships between spoken language and freeform gestures. *Proceedings of Findings at the Conference on Empirical Methods in Natural Language Processing*, 2020. (document), 1.2, 1.2, 5.1, 5.1, **??**, 5.3.2, **??**, **??**, **??**, 5.4, 5.4, 5.4, **??**, 5.5, 5.6, 6.1, 6.4, 6.4, 6.4, 6.3, 6.4, 7.1, D.1.3, D.2.1, D.1, D.2.1, D.2.2, D.2.3

[8] Chaitanya Ahuja, Dong Won Lee, Yukiko I Nakano, and Louis-Philippe Morency. Style

transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. *Proceedings of the European Conference on Computer Vision*, 2020. 1.1, 1.2, 3.3.3, 3.4.3, 5.3.1, 5.3.2, 5.4, 5.4, 5.6, 6.3.3, 6.4, 6.4, **??**, **??**, 6.5, 6.5, D.2.2

[9] Chaitanya Ahuja, Dong Won Lee, and Louis-Philippe Morency. Low-resource adaptation for personalized co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. (document), 1.2, 6.3.2, 6.4, 6.2a, 6.2, 6.5, 6.5, **??**, **??**, 6.5

[10] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 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, 2014. 2.4.4, 3.4.2, 4.4.2, 5.4, 6.4, D.1.3

[11] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. 3.6, 4.3.1, 6.3.3

[12] Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. 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, 2017. 3.3.3, 4.3.1

[13] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3.3.2

[14] Jeremy N Bailenson, Nick Yee, Dan Merget, and Ralph Schroeder. The effect of behavioral realism and form realism of real-time avatar faces on verbal disclosure, nonverbal disclosure, emotion recognition, and copresence in dyadic interaction. *Presence: Teleoperators and Virtual Environments*, 15(4):359–372, 2006. 4.1

[15] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018. 1.2.1

[16] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. *Challenges and Applications in Multimodal Machine Learning*, page 17–48. Association for Computing Machinery and Morgan and Claypool, 2018. ISBN 9781970001716. URL `https://doi.org/10.1145/3107990.3107993`. 1.2.1

[17] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. Recycle-gan: Unsuper-

vised video retargeting. In *Proceedings of the European conference on computer vision (ECCV)*, pages 119–135, 2018. 4.6

[18] David Bau, Steven Liu, Tongzhou Wang, Jun-Yan Zhu, and Antonio Torralba. Rewriting a deep generative model. In *European Conference on Computer Vision*, pages 351–369. Springer, 2020. 5.3.1

[19] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 2.3.3, 3.6

[20] Kirsten Bergmann and Stefan Kopp. Increasing the expressiveness of virtual agents: autonomous generation of speech and gesture for spatial description tasks. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 361–368, 2009. ISBN 978-0-9817381-6-1. 4.1

[21] Yanyao Bian, Changbin Chen, Yongguo Kang, and Zhenglin Pan. Multi-reference tacotron by intercross training for style disentangling, transfer and control in speech synthesis. *arXiv preprint arXiv:1904.02373*, 2019. 4.6

[22] Jonathon Byrd and Zachary C Lipton. What is the effect of importance weighting in deep learning? *arXiv preprint arXiv:1812.03372*, 2018. 3.6

[23] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018. 3.4.3, 4.4.3

[24] Justin Cassell. More than just another pretty face: Embodied conversational agents. *Communications of the ACM*, 43(4), 2001. 1.1, 5.1, 6

[25] Justine Cassell. Embodied conversational agents: representation and intelligence in user interfaces. *AI magazine*, 22(4):67, 2001. 3.3.2

[26] Justine Cassell, Matthew Stone, Brett Douville, Scott Prevost, Brett Achorn, Mark Steedman, Norman I Badler, and Catherine Pelachaud. Modeling the interaction between speech and gesture. *Technical Reports (CIS)*, page 341, 1994. 1.1

[27] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. BEAT: the Behavior Expression Animation Toolkit. In *the 28th annual conference on Computer graphics and interactive techniques (SIGGRAPH '01)*, pages 477–486, 2001. doi: https://doi.org/10.1145/383259.383315. 3.6, 5.6

[28] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. Beat: the behavior expression animation toolkit. In *Life-Like Characters*, pages 163–185. Springer, 2004. 1, 4.1

[29] Yu-Wei Chao, Jimei Yang, Brian L Price, Scott Cohen, and Jia Deng. Forecasting human dynamics from static images. 2.3.3, 2.4.1, 2.6

[30] Chung-Cheng Chiu and Stacy Marsella. How to train your avatar: A data driven approach to gesture generation. In *International Workshop on Intelligent Virtual Agents*, pages 127–140. Springer, 2011. 2.1

[31] Chung-Cheng Chiu and Stacy Marsella. Gesture generation with low-dimensional embeddings. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 781–788, 2014. 3.6, 4.6, 5.6

[32] Chung Cheng Chiu, Louis Philippe Morency, and Stacy Marsella. 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, 2015. ISBN 9783319219950. doi: 10.1007/978-3-319-21996-7_17. 3.6, 4.1, 5.6

[33] Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella. Predicting co-verbal gestures: a deep and temporal modeling approach. In *International Conference on Intelligent Virtual Agents*, pages 152–166. Springer, 2015. 1.1, 2.1

[34] 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. 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, 2018. 3.4.3

[35] Hsu-kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose forecasting. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1423–1432. IEEE, 2019. 2.6

[36] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014. 2.4.2

[37] Pietro Cipresso, Irene Alice Chicchi Giglioli, Mariano Alcañiz Raya, and Giuseppe Riva. The past, present, and future of virtual and augmented reality research: a network and

cluster analysis of the literature. *Frontiers in psychology*, 9:2086, 2018. 1, 5.1

[38] Robert O Davis and Joseph Vincent. Sometimes more is better: Agent gestures, procedural knowledge and the foreign language learner. *British Journal of Educational Technology*, 50(6):3252–3263, 2019. 4.4.3

[39] Emily L Denton et al. Unsupervised learning of disentangled representations from video. In *Advances in neural information processing systems*, pages 4414–4423, 2017. 4.6

[40] David DeVault, Kallirroi Georgila, Ron Artstein, Fabrizio Morbini, David Traum, Stefan Scherer, Albert A Rizzo, and Louis-Philippe Morency. Verbal indicators of psychological distress in interactive dialogue with a virtual human. In *Proceedings of the SIGDIAL 2013 Conference*, pages 193–202, 2013. 1

[41] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3.3.2, 3.3.3, A.1

[42] Maurice Diesendruck, Ethan R Elenberg, Rajat Sen, Guy W Cole, Sanjay Shakkottai, and Sinead A Williamson. Importance weighted generative networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 249–265. Springer, 2019. 3.3.1, 3.3.1, 3.6

[43] Justin Domke and Daniel R Sheldon. Importance weighting and variational inference. In *Advances in neural information processing systems*, pages 4470–4479, 2018. 3.6

[44] Ahmed Elgammal and Chan-Su Lee. The role of manifold learning in human motion analysis. In *Human Motion*, pages 25–56. Springer, 2008. 1.1

[45] Ylva Ferstl, Michael Neff, and Rachel McDonnell. Multi-objective adversarial gesture generation. 2019. 3.6, 4.1, 4.6, 5.6

[46] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999. 1.1

[47] Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary C Lipton. A unified view of label shift estimation. *arXiv preprint arXiv:2003.07554*, 2020. 3.6

[48] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015. 4.6

[49] Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges. Learning human motion models

for long-term predictions. In *2017 International Conference on 3D Vision (3DV)*, pages 458–466. IEEE, 2017. 2.6

[50] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3497–3506, 2019. 1.2, 3.1, 3.2, 3.3.3, 3.3.3, 3.4.1, 3.4.2, 3.4.3, 3.4.3, 3.6, 3.6, 4.1, 4.3.4, 4.3.5, **??**, 4.4.1, 4.4.2, 4.4.3, 4.5.1, 4.6, 5.1, 5.1, 5.3.2, 5.4, 5.4, 5.6, 6.1, 6.4, 6.4, 7.1, 7.1, D.2.2

[51] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323, 2011. A.1

[52] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 3.2, 3.6, 4.3.4, 5.2, 6.2, 6.3.3

[53] Aditya Grover, Jiaming Song, Ashish Kapoor, Kenneth Tran, Alekh Agarwal, Eric J Horvitz, and Stefano Ermon. Bias correction of learned generative models using likelihood-free importance weighting. In *Advances in Neural Information Processing Systems*, pages 11056–11068, 2019. 3.3.1, 3.6

[54] 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 *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*, 2020. 1.1

[55] Nishant Gurunath, Sai Krishna Rallabandi, and Alan Black. Disentangling speech and non-speech components for building robust acoustic models from found data. *arXiv preprint arXiv:1909.11727*, 2019. 4.6

[56] Ian Hacking et al. *The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference*. Cambridge University Press, 2006. 7.2

[57] Eva Hanser, Paul Mc Kevitt, Tom Lunney, and Joan Condell. Scenemaker: Intelligent multimodal visualisation of natural language scripts. In *Irish Conference on Artificial Intelligence and Cognitive Science*, pages 144–153. Springer, 2009. 2.1

[58] Guang-Yuan Hao, Hong-Xing Yu, and Wei-Shi Zheng. Mixgan: learning concepts from different domains for mixture generation. *arXiv preprint arXiv:1807.01659*, 2018. 3.3.3, 4.3, 4.3.1

[59] Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. 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, 2018. 3.6, 4.6, 5.6

[60] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020. 1.2, 5.1, 5.1, 6.1, 7.1

[61] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 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, 2017. 3.4.2, 5.4, 6.4, D.1.3

[62] Quan Hoang, Tu Dinh Nguyen, Trung Le, and Dinh Phung. Mgan: Training generative adversarial nets with multiple generators. 2018. 3.3.3, 4.3.1

[63] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2.4.2

[64] Daniel Holden, Jun Saito, Taku Komura, and Thomas Joyce. Learning motion manifolds with convolutional autoencoders. In *SIGGRAPH Asia 2015 Technical Briefs*, pages 1–4. 2015. 1.1

[65] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 35(4):138, 2016. 2.4.1

[66] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)*, 36(4):42, 2017. 2.6

[67] Autumn B Hostetter and Andrea L Potthoff. Effects of personality and social situation on representational gesture production. *Gesture*, 12(1):62–83, 2012. 1

[68] Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. Avatar digitization from a single image for real-time rendering. *ACM Transactions on Graphics (ToG)*, 36(6):1–14, 2017. 1, 5.1

[69] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018. 4.3.2, **??**, 4.4.1, **??**, 4.6

[70] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992. 2.3.4

[71] Yong K Hwang, Pang C Chen, and Peter A Watterberg. Interactive task planning through natural language. In *Proceedings of IEEE International Conference on Robotics and Automation*, volume 1, pages 24–29. IEEE, 1996. 2.1

[72] Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. Dynamic 3d avatar creation from hand-held video input. *ACM Transactions on Graphics (ToG)*, 34(4):1–14, 2015. 1, 5.1

[73] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 3.3.3

[74] Ryo Ishii, Chaitanya Ahuja, Yukiko I. Nakano, and Louis-Philippe Morency. Impact of personality on nonverbal behavior generation. *20th ACM International Conference on Intelligent Virtual Agents*, 2020. 1.1, 1.2.1

[75] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 3.6, 4.6

[76] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G Hauptmann. Self-paced curriculum learning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. 3.6

[77] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 4.6, 5.3.1

[78] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015. 3.6

[79] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 7.2

[80] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 7.2

[81] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *arXiv preprint arXiv:2006.06676*, 2020. 5.6

[82] Angelos Katharopoulos and François Fleuret. Not all samples are created equal: Deep learning with importance sampling. *arXiv preprint arXiv:1803.00942*, 2018. 3.3.1, 3.6

[83] Adam Kendon. Gesture and speech: two aspects of the process of utterance. In M. R. Key, editor, *Nonverbal Communication and Language*, pages 207–227. 1980. 4.1

[84] Adam Kendon. *Sign languages of Aboriginal Australia: Cultural, semiotic and communicative perspectives*. Cambridge University Press, 1988. 1

[85] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. A.4, C.3.2, D.2.1

[86] Sotaro Kita. Cross-cultural variation of speech-accompanying gesture: A review. *Language and cognitive processes*, 24(2):145–167, 2009. 1

[87] Ilya Kostrikov, Kumar Krishna Agrawal, Sergey Levine, and Jonathan Tompson. Addressing sample inefficiency and reward bias in inverse reinforcement learning. *arXiv preprint arXiv:1809.02925*, 2018. 3.6

[88] Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. Analyzing input and output representations for speech-driven gesture generation. *arXiv preprint arXiv:1903.03369*, 2019. 3.1, 3.4.3, 4.4.3

[89] Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexanderson, Iolanda Leite, and Hedvig Kjellström. Gesticulator: A framework for semantically-aware speech-driven gesture generation. *arXiv preprint arXiv:2001.09326*, 2020. 1.2, 3.4.1, 3.6, 5.1, 5.1, 5.6, 6.1, 7.1, B.3.1

[90] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995. 7.2

[91] Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha S Srinivasa, and Yaser Sheikh. 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, 2019. 3.6

[92] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation via disentangled representations. *arXiv preprint arXiv:1905.01270*, 2019. 4.3.2, 4.6

[93] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. In *Advances in Neural Information Processing Systems*, pages 3581–3591, 2019. 4.6

[94] Jina Lee and Stacy Marsella. Nonverbal behavior generator for embodied conversational agents. In *Proceedings of the 6th international conference on Intelligent virtual agents (IVA2006)*, pages 243–255, 2006. 3.6, 5.6

[95] Jina Lee and Stacy Marsella. Nonverbal behavior generator for embodied conversational agents. In *International Workshop on Intelligent Virtual Agents*, pages 243–255. Springer, 2006. 1.1

[96] Sergey Levine, Christian Theobalt, and Vladlen Koltun. Real-time prosody-driven synthesis of body language. *ACM Trans. Graph.*, 28(5):172:1–172:10, December 2009. ISSN 0730-0301. 3.6, 4.6, 5.6

[97] Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun. Gesture controllers. *ACM Trans. Graph.*, 29(4):124:1–124:11, July 2010. ISSN 0730-0301. 3.6, 4.6, 5.6

[98] Loet Leydesdorff. *A sociological theory of communication: The self-organization of the knowledge-based society*. Universal-Publishers, 2001. 1

[99] Margot Lhommet and Stacy Marsella. From embodied metaphors to metaphoric gestures. *CogSci*, pages 788–793, 2016. 3.6, 5.6

[100] Margot Lhommet, Yuyu Xu, and Stacy Marsella. Cerebella: Automatic Generation of Nonverbal Behavior for Virtual Humans. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 4303–4304, 2015. 3.6, 5.6

[101] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, pages 2203–2213, 2017. 3.6

[102] Y. Li, Richard Zhang, Jingwan Lu, and E. Shechtman. Few-shot image generation with elastic weight consolidation. *ArXiv*, abs/2012.02780, 2020. 5.6

[103] Tze Wei Liew and Su-Mae Tan. Exploring the effects of specialist versus generalist embodied virtual agents in a multi-product category online store. *Telematics and Informatics*, 35(1):122–135, 2018. 1.1, 5.1, 6

[104] Angela S Lin, Lemeng Wu, Rodolfo Corona, Kevin Tai, Qixing Huang, and Raymond J Mooney. 1. generating animated videos of human activities from natural language descriptions. *Learning*, 2018, 2018. 2.3.4, 2.4.2, 2.4.3, **??**, 2.6, A.2, **??**

[105] Xiao Lin and Mohamed R Amer. Human motion modeling using dvgans. *arXiv preprint arXiv:1804.10652*, 2018. 2.6

[106] Zachary C Lipton, Yu-Xiang Wang, and Alex Smola. Detecting and correcting for label shift with black box predictors. *arXiv preprint arXiv:1802.03916*, 2018. 3.6

[107] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Advances in neural information processing systems*, pages 469–477, 2016. 4.3.2

[108] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017. 4.6

[109] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982. 4.3.1, B.2.1

[110] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Transactions on Graphics (TOG)*, 37(4):68, 2018. 1, 5.1

[111] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. B.3.2

[112] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems*, pages 406–416, 2017. 2.6

[113] Minhua Ma and Paul Mc Kevitt. Virtual human animation in natural language visualisation. *Artificial Intelligence Review*, 25(1-2):37–53, 2006. 2.1

[114] Shuang Ma, Daniel Mcduff, and Yale Song. Neural tts stylization with adversarial and collaborative games. 2018. 4.3.3, 4.6

[115] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. (document), 4.2, 5.8

[116] Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. Virtual character performance from speech. In *Symposium on Computer Animation*, pages 25–35, 2013. ISBN 9781450321327. doi: 10.1145/2485895.2485900. 3.6, 5.6

[117] Tambet Matiisen, Avital Oliver, Taco Cohen, and John Schulman. Teacher-student curriculum learning. *IEEE transactions on neural networks and learning systems*, 2019. 3.6

[118] David McNeill. *Hand and mind: What gestures reveal about thought.* University of Chicago Press, 1992. 4.1, 4.3.1, 5.4

[119] David McNeill. *Hand and mind: What gestures reveal about thought.* University of Chicago press, 1992. 1, 5.1, 6

[120] Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. (document), 2.5

[121] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013. 2.4.2

[122] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 3.6

[123] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze discriminator: A simple baseline for fine-tuning gans. *ArXiv*, abs/2002.10964, 2020. 5.3.1, 5.6

[124] Louis-Philippe Morency, Giota Stratou, David DeVault, Arno Hartholt, Margot Lhommet, Gale M Lucas, Fabrizio Morbini, Kallirroi Georgila, Stefan Scherer, Jonathan Gratch, et al. Simsensei demonstration: A perceptive virtual human interviewer for healthcare applications. In *AAAI*, pages 4307–4308, 2015. 1

[125] Desmond Morris. *Gestures, their origins and distribution.* Stein &amp; Day Pub, 1979. 1

[126] Adithyavairavan Murali, Animesh Garg, Sanjay Krishnan, Florian T Pokorny, Pieter Abbeel, Trevor Darrell, and Ken Goldberg. 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, 2016. 3.6

[127] Arsha Nagrani, Joon Son Chung, Samuel Albanie, and Andrew Zisserman. Disentangled speech embeddings using cross-modal self-supervision. In *ICASSP 2020-2020 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6829–6833. IEEE, 2020. 4.6

[128] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010. 3.3.3

[129] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using a test for typicality. *arXiv preprint arXiv:1906.02994*, 2019. 3.6

[130] Gonzalo Navarro. A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1):31–88, 2001. 3.4.3

[131] Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics (TOG)*, 27(1):1–24, 2008. 4.6

[132] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 5.6

[133] Atsuhiro Noguchi and Tatsuya Harada. Image generation from small datasets via batch statistics adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2750–2758, 2019. 5.6

[134] Christian Obermeier, Spencer D Kelly, and Thomas C Gunter. A speaker's gesture style can affect language comprehension: Erp evidence from gesture-speech integration. *Social cognitive and affective neuroscience*, 10(9):1236–1243, 2015. 4.1

[135] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. *arXiv preprint arXiv:2104.06820*, 2021. **??**, 5.3.2, **??**, **??**, 5.4, 5.5, 5.6, **??**, **??**

[136] Minna Pakanen, Paula Alavesa, Niels van Berkel, Timo Koskela, and Timo Ojala. "nice to see you virtually": Thoughtful design and evaluation of virtual avatar of the other user in ar and vr based telexistence systems. *Entertainment Computing*, 40:100457, 2022. 1.1, 5.1, 6

[137] Ye Pan and Anthony Steed. The impact of self-avatars on trust and collaboration in shared virtual environments. *PloS one*, 12(12):e0189078, 2017. 1.1, 5.1, 6

[138] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embed-

ding and translation to bridge video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4594–4602, 2016. 2.6

[139] Dario Pavllo, David Grangier, and Michael Auli. Quaternet: A quaternion-based recurrent model for human motion. *arXiv preprint arXiv:1805.06485*, 2018. 2.6

[140] Catherine Pelachaud. Studies on gesture expressivity for a virtual agent. *Speech Communication*, 51(7):630–639, 2009. 3.1, 3.5, 4.1, 5.4, 5.5

[141] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 3.4.1, B.3.1

[142] Kenneth Lee Pike. *Linguistic concepts: An introduction to tagmemics*, volume 790. University of Nebraska Press Lincoln, 1982. 1

[143] Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT motion-language dataset. *Big Data*, 4(4):236–252, dec 2016. doi: 10.1089/big.2016.0028. URL `http://dx.doi.org/10.1089/big.2016.0028`. 2.4.1

[144] Matthias Plappert, Christian Mandery, and Tamim Asfour. Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. *Robotics and Autonomous Systems*, 109:13–26, 2018. 2.6

[145] Shrimai Prabhumoye, Alan W Black, and Ruslan Salakhutdinov. Exploring controllable text generation techniques. *arXiv preprint arXiv:2005.01822*, 2020. 7.2

[146] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 7.2

[147] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017. 6.3.1

[148] Douglas A Reynolds. Gaussian mixture models. *Encyclopedia of biometrics*, 741, 2009. 4.3.1, B.2.1

[149] Anthony Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146, 1995. 6.3.1

[150] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image comput-*

ing and computer-assisted intervention, pages 234–241. Springer, 2015. 3.3.3, 4.3.5

[151] Mihaela Rosca, Balaji Lakshminarayanan, David Warde-Farley, and Shakir Mohamed. Variational approaches for auto-encoding generative adversarial networks. *arXiv preprint arXiv:1706.04987*, 2017. 4.3.2

[152] Amélie Royer, Konstantinos Bousmalis, Stephan Gouws, Fred Bertsch, Inbar Mosseri, Forrester Cole, and Kevin Murphy. Xgan: Unsupervised image-to-image translation for many-to-many mappings. In *Domain Adaptation for Visual Understanding*, pages 33–49. Springer, 2020. 4.3.2

[153] Najmeh Sadoughi and Carlos Busso. Novel realizations of speech-driven head movements with generative adversarial networks. pages 6169–6173, 04 2018. doi: 10.1109/ICASSP.2018.8461967. 3.6, 4.6, 5.6

[154] Najmeh Sadoughi, Yang Liu, and Carlos Busso. 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, 2015. 3.6

[155] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016. 4.4.2

[156] Mehmet E. Sargin, Yucel Yemez, Engin Erzin, and Ahmet M. Tekalp. Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30:1330–1345, 2008. doi: 10.1109/TPAMI.2007.70797. 3.6, 4.6, 5.6

[157] Shashank Sharma and Vinay P Namboodiri. No modes left behind: Capturing the data distribution effectively using gans. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 3.6

[158] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000. 3.6

[159] Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher. 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, 06 2018. 3.6, 4.6, 5.6

[160] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1145–1153, 2017. 2.4.4, 3.4.2, 4.4.2, 5.4, 6.4, D.1.3

[161] Harrison Jesse Smith, Chen Cao, Michael Neff, and Yingying Wang. Efficient neural networks for real-time motion style transfer. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 2(2):1–17, 2019. 4.6

[162] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. A.1

[163] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 5.5

[164] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 4.4.2

[165] Wataru Takano and Yoshihiko Nakamura. Bigram-based natural language model and statistical motion symbol model for scalable language of humanoid robots. In *2012 IEEE International Conference on Robotics and Automation*, pages 1232–1237. IEEE, 2012. 2.6

[166] Wataru Takano and Yoshihiko Nakamura. Statistical mutual conversion between whole body motion primitives and linguistic sentences for human motions. *The International Journal of Robotics Research*, 34(10):1314–1328, 2015. 2.6

[167] Kenta Takeuchi, Dai Hasegawa, Shinichi Shirakawa, Naoshi Kaneko, Hiroshi Sakuta, and Kazuhiko Sumi. Speech-to-gesture generation: A challenge in deep learning approach with bi-directional lstm. In *Proceedings of the 5th International Conference on Human Agent Interaction*, pages 365–369. ACM, 2017. 2.1, 2.6

[168] Yongyi Tang, Lin Ma, Wei Liu, and Weishi Zheng. Long-term human motion prediction by modeling motion context and enhancing motion dynamic. *arXiv preprint arXiv:1805.02513*, 2018. 2.6

[169] Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)*, 36(4):93, 2017. 2.6

[170] Jackson Tolins, Kris Liu, Yingying Wang, Jean E Fox Tree, Marilyn Walker, and Michael Neff. 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, 2016. 3.6

[171] Ilya O Tolstikhin, Sylvain Gelly, Olivier Bousquet, Carl-Johann Simon-Gabriel, and Bernhard Schölkopf. Adagan: Boosting generative models. In *Advances in Neural Information Processing Systems*, pages 5424–5433, 2017. 3.6

[172] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. *arXiv preprint arXiv:1906.00295*, 2019. 3.1, 3.3.2, 3.3.2, 3.6

[173] Ryan Turner, Jane Hung, Eric Frank, Yunus Saatci, and Jason Yosinski. Metropolis-hastings generative adversarial networks. *arXiv preprint arXiv:1811.11357*, 2018. 3.3.1, 3.6

[174] Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. 7.2

[175] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3.1, 3.3.2, 3.3.2, 7.2, B.2.2

[176] Ingrid Vilà-Giménez and Pilar Prieto. Encouraging kids to beat: Children's beat gesture production boosts their narrative performance. *Developmental Science*, page e12967, 2020. 1

[177] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. *arXiv preprint arXiv:1706.08033*, 2017. 4.6

[178] Petra Wagner, Zofia Malisz, and Stefan Kopp. Gesture and speech in interaction: An overview, 2014. 4.1

[179] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring gans: generating images from limited data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 218–234, 2018. **??**, 5.3.2, **??**, **??**, 5.4, 5.5, 5.6, **??**, **??**

[180] Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: effective knowledge transfer from gans to target domains with few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9332–9341, 2020. **??**, **??**, **??**, 5.4, 5.5, 5.6, **??**, **??**

[181] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017. 7.2

[182] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. *arXiv preprint arXiv:1803.09017*, 2018. 4.3.3, 4.6

[183] WebRTC. "webrtc," 2017. [online]. available: https://webrtc.org/. B.3.1

[184] Thomas Wolf, L Debut, V Sanh, J Chaumond, C Delangue, A Moi, P Cistac, T Rault, R Louf, M Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv, abs/1910.03771*, 2019. 3.3.3

[185] Pieter Wolfert, Nicole Robinson, and Tony Belpaeme. A review of evaluation practices of gesture generation in embodied conversational agents. *arXiv preprint arXiv:2101.03769*, 2021. 5.5, 5.5

[186] Stephen J Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1): 3–34, 2015. 2.3.2

[187] Chenshen Wu, Luis Herranz, Xialei Liu, Joost van de Weijer, Bogdan Raducanu, et al. Memory replay gans: Learning to generate new categories without forgetting. *Advances in Neural Information Processing Systems*, 31, 2018. (document), 6.3.1, 6.4, 6.2c, 6.4, 6.3, 6.5, **??**, **??**, **??**, **??**, 6.4, 6.5

[188] Jiang Xu, Patrick J Gannon, Karen Emmorey, Jason F Smith, and Allen R Braun. Symbolic gestures and spoken language are processed by a common neural system. *Proceedings of the National Academy of Sciences*, 106(49):20664–20669, 2009. 4.1

[189] Yuyu Xu, Catherine Pelachaud, and Stacy Marsella. 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, 2014. ISSN 16113349. doi: 10.1007/978-3-319-09767-1_58. 3.6, 5.6

[190] Tatsuro Yamada, Hiroyuki Matsunaga, and Tetsuya Ogata. Paired recurrent autoencoders for bidirectional translation between robot actions and linguistic descriptions. *IEEE Robotics and Automation Letters*, 3(4):3441–3448, 2018. 2.6

[191] Qiaojing Yan and Wei Wang. Dcgans for image super-resolution, denoising and debluring. *Advances in Neural Information Processing Systems*, pages 487–495, 2017. 3.6

[192] Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. Pose guided human video generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018. 2.6

[193] Jimei Yang, Scott E Reed, Ming-Hsuan Yang, and Honglak Lee. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In *Advances in Neural Information Processing Systems*, pages 1099–1107, 2015. 2.3.3

[194] Shiyu Yi, Donglin Zhan, Zhengyang Geng, Wenqing Zhang, and Chang Xu. Fis-gan: Gan with flow-based importance sampling. *arXiv preprint arXiv:1910.02519*, 2019. 3.6

[195] 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 *2019 International Conference on Robotics and Automation (ICRA)*, pages 4303–4309. IEEE, 2019. 3.6, 5.6

[196] 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. *ACM Transactions on Graphics (TOG)*, 39(6):1–16, 2020. 1.1, 5.3.1, D.1.3

[197] Fajrian Yunus, Chloé Clavel, and Catherine Pelachaud. Sequence-to-sequence predictive model: From prosody to communicative gestures. *arXiv preprint arXiv:2008.07643*, 2020. 1.1

[198] Wojciech Zaremba and Ilya Sutskever. Learning to execute. *arXiv preprint arXiv:1410.4615*, 2014. 2.3.3

[199] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang,

and Dimitris N Metaxas. 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, 2017. 3.6

[200] Yeyao Zhang, Eleftheria Tsipidi, Sasha Schriber, Mubbasir Kapadia, Markus H. Gross, and Ashutosh Modi. Generating animations from screenplays. *CoRR*, abs/1904.05440, 2019. 2.1

[201] Peilin Zhong, Yuchen Mo, Chang Xiao, Pengyu Chen, and Changxi Zheng. Rethinking generative mode coverage: A pointwise guaranteed approach. In *Advances in Neural Information Processing Systems*, pages 2086–2097, 2019. 3.6

[202] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 4.6

[203] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in neural information processing systems*, pages 465–476, 2017. 4.6

[204] Xiangxin Zhu, Carl Vondrick, Charless C Fowlkes, and Deva Ramanan. Do we need more training data? *International Journal of Computer Vision*, 119(1):76–92, 2016. 5.5