Supplementary Materials:
Low-Resource Adaptation for Personalized Co-speech Gesture Generation

Chaitanya Ahuja[1], Dong Won Lee[2] & Louis-Philippe Morency[1]
[1]Language Technologies Institute, CMU & [2]Machine Learning Department, CMU
{cahuja, dongwonl}@andrew.cmu.edu, morency@cs.cmu.edu

A. Additional results

A.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 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 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.

A.2. Qualitative visualizations

A more exhaustive set of qualitative results is shown in form of visual histograms in Figures 2, 3, 4. Furthermore, we also plot of the velocity distributions across our model and different baselines in Figure 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 7, 6, 8 and 9.

A.3. Quantitative results

A more exhaustive set of quantitative results on FID scores [1, 5, 11] (i.e., measure of distribution of generated gestures) and PCK scores [3, 8] (i.e., measure of relevance of generated gestures to input language) is shown in Table 2.

Training Data	FID ↓		PCK ↑		Human Perceptual Study ↑	
	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 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).

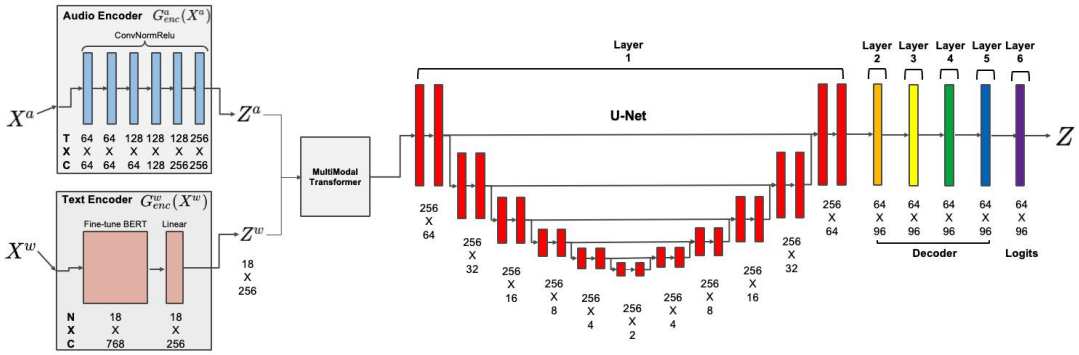

Figure 1. Illustration of the source model [1] for our experiments in the main paper. The layer IDs from 1 through 6 are the possible choices for layer l for the low resource adaptation

B. Experimental setup details

B.1. Model hyperparameters

We use, as our source model, an architecture described in Ahuja et al. [1] and illustrated in Figure 1. We optimize our model using Adam [6] 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 our full DiffGAN model.

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 1) of the source model generator [1] is the best choice.

Construction of ϕ_k The intuition here is that directions large magnitudes are considered more important. Hence, for **each sample in the batch** and **each time step in** Y^t, the k highest (i.e. top-k) values in the channel dimension are selected. Other values are zeroed out, giving ϕ_k which is used as a mask in Equation 2 of the main paper.

Wall times of Low-Resource Adaptation The time for low-resource adaptation is typically much lesser than training a model from scratch. In this instance, our DiffGAN takes around 0.5 hours (equivalent of 4ooo iterations), which is significantly lesser than 5 hours taken by AISLe [1].

B.2. Dataset

Pose Audio Transcripts (PATS) dataset [1, 2, 4] 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` $\rightarrow$ `target` denotes the source to target domain):

- `oliver` $\rightarrow$ `maher`
- `oliver` $\rightarrow$ `chemistry`
- `oliver` $\rightarrow$ `ytch_prof`
- `oliver` $\rightarrow$ `lec_evol`

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

B.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) [1].

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 10 and 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.
- **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.

References

[1] Chaitanya Ahuja, Dong Won Lee, Ryo Ishii, and Louis-Philippe Morency. No gestures left behind: Learning relationships between spoken language and freeform gestures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1884–1895, 2020. 1, 2, 3

[2] 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. 2

[3] 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. 1

Amt. of Data (minutes)	Models	FID ↓								PCK ↑							
		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 [9]	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 [10]	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 [7]	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 [9]	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 [10]	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 [7]	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 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)

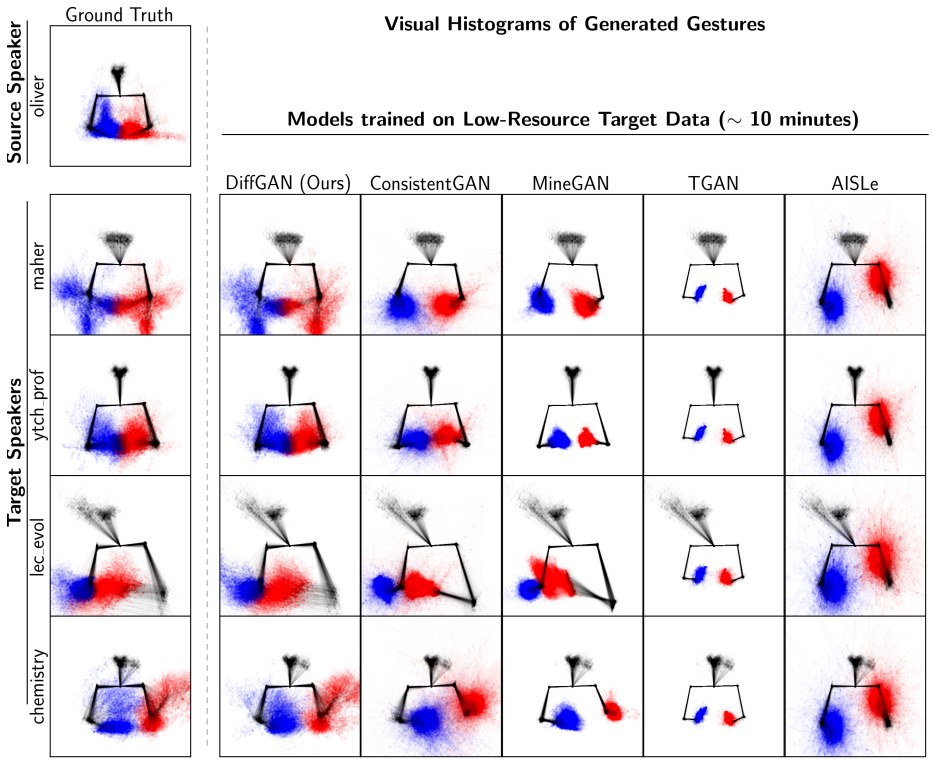

Figure 2. Visual Histograms of generated gestures visually describe the distribution of hand gestures in space. Red and blue colors denote the left and right arms respectively. First row is the source speaker, below which we have all the target speakers. Each column denotes a model which adapts output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

[4] 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. 2

[5] 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. 1

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[7] 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. 4

[8] 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. 1

[9] 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. 4

[10] 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. 4

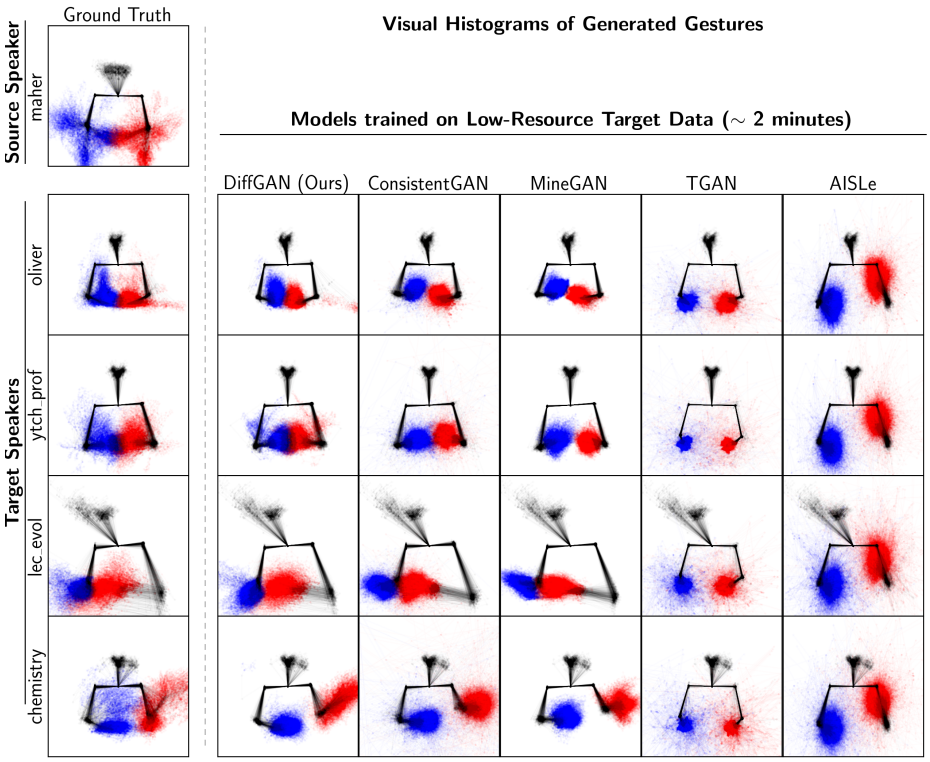

Figure 3. Visual Histograms of generated gestures visually describe the distribution of hand gestures in space. Red and blue colors denote the left and right arms respectively. First row is the source speaker, below which we have all the target speakers. Each column denotes a model which adapts output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

[11] 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

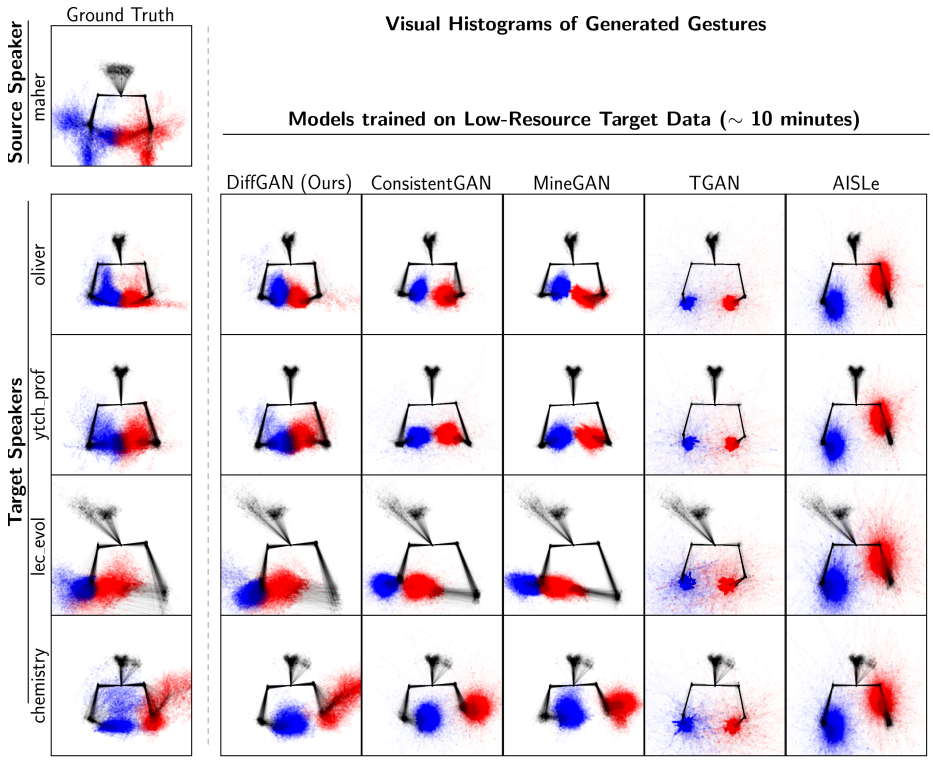

Figure 4. Visual Histograms of generated gestures visually describe the distribution of hand gestures in space. Red and blue colors denote the left and right arms respectively. First row is the source speaker, below which we have all the target speakers. Each column denotes a model which adapts output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

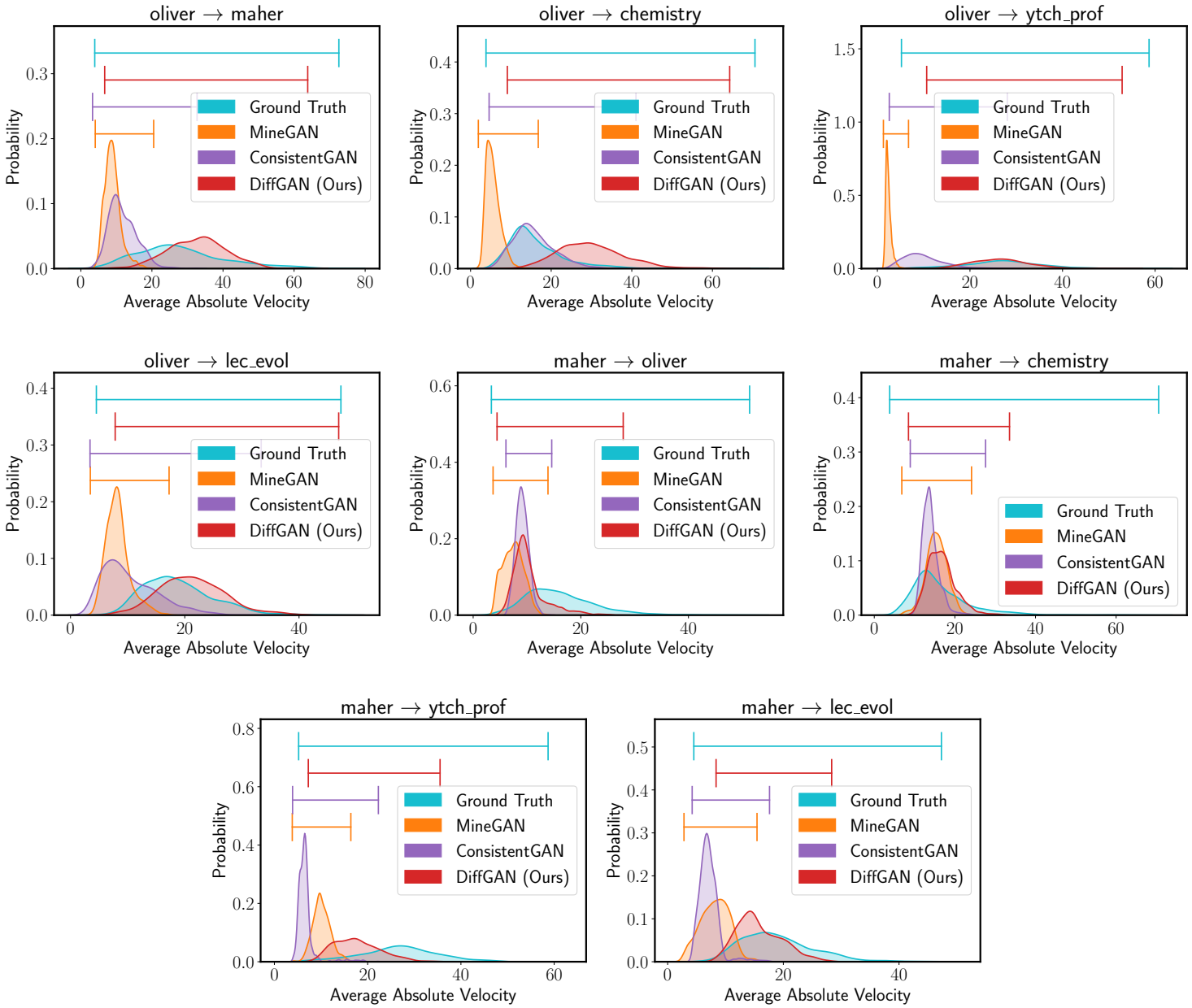

Figure 5. Distribution of the generated gestures with average absolute velocity as the statistic for source to target domain adaptation. The support (or coverage) of the distribution is denoted with the colour coded lines at the top of each plot. Larger overlap of a model's distribution with the ground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

Figure 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.

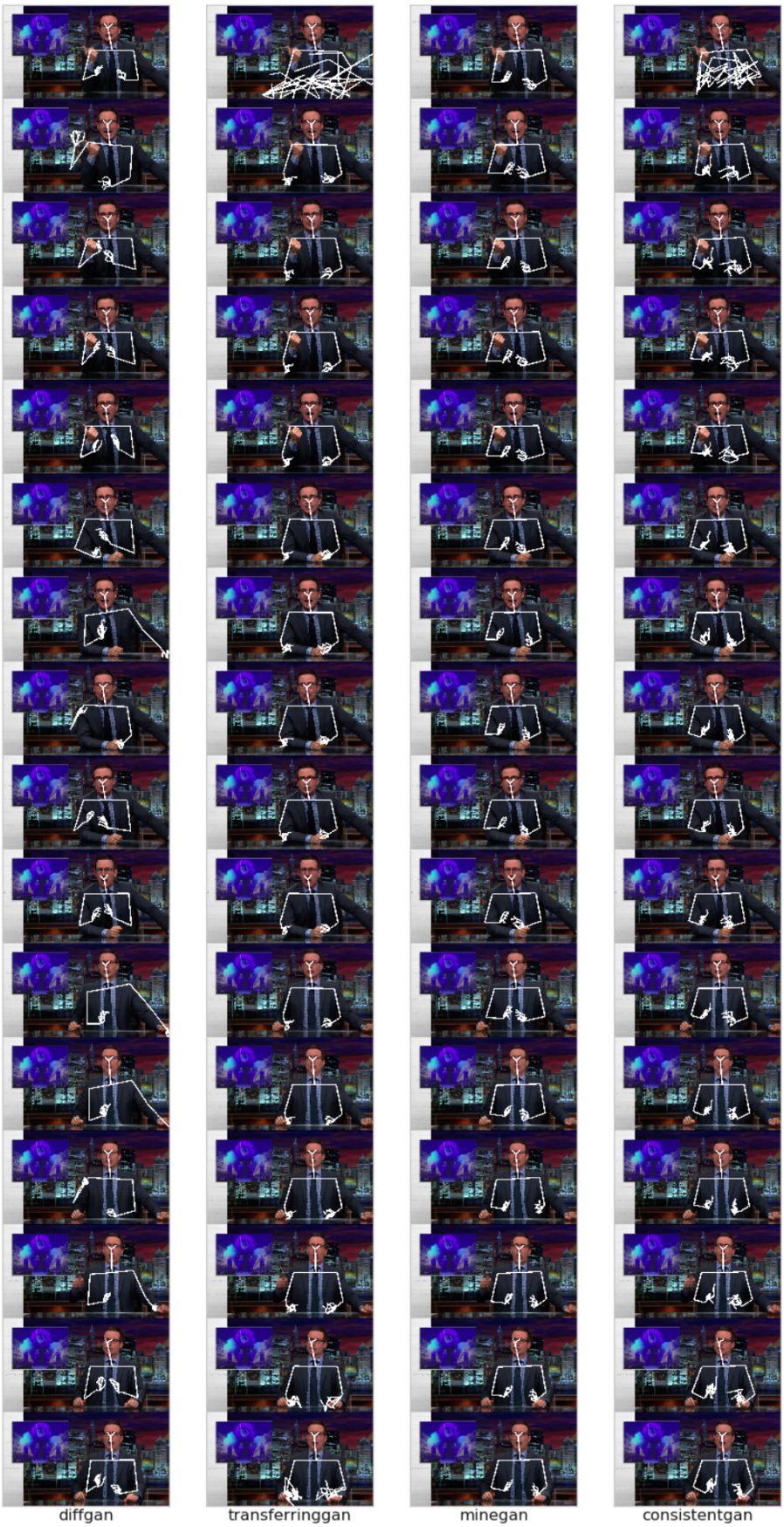

Figure 7. Qualitative comparison of our DiffGAN with prior work over shape of generated gestures. With `maher` as the source domain, target model outputs over the target domain `oliver` are superimposed over ground truth video frames for easy comparison.

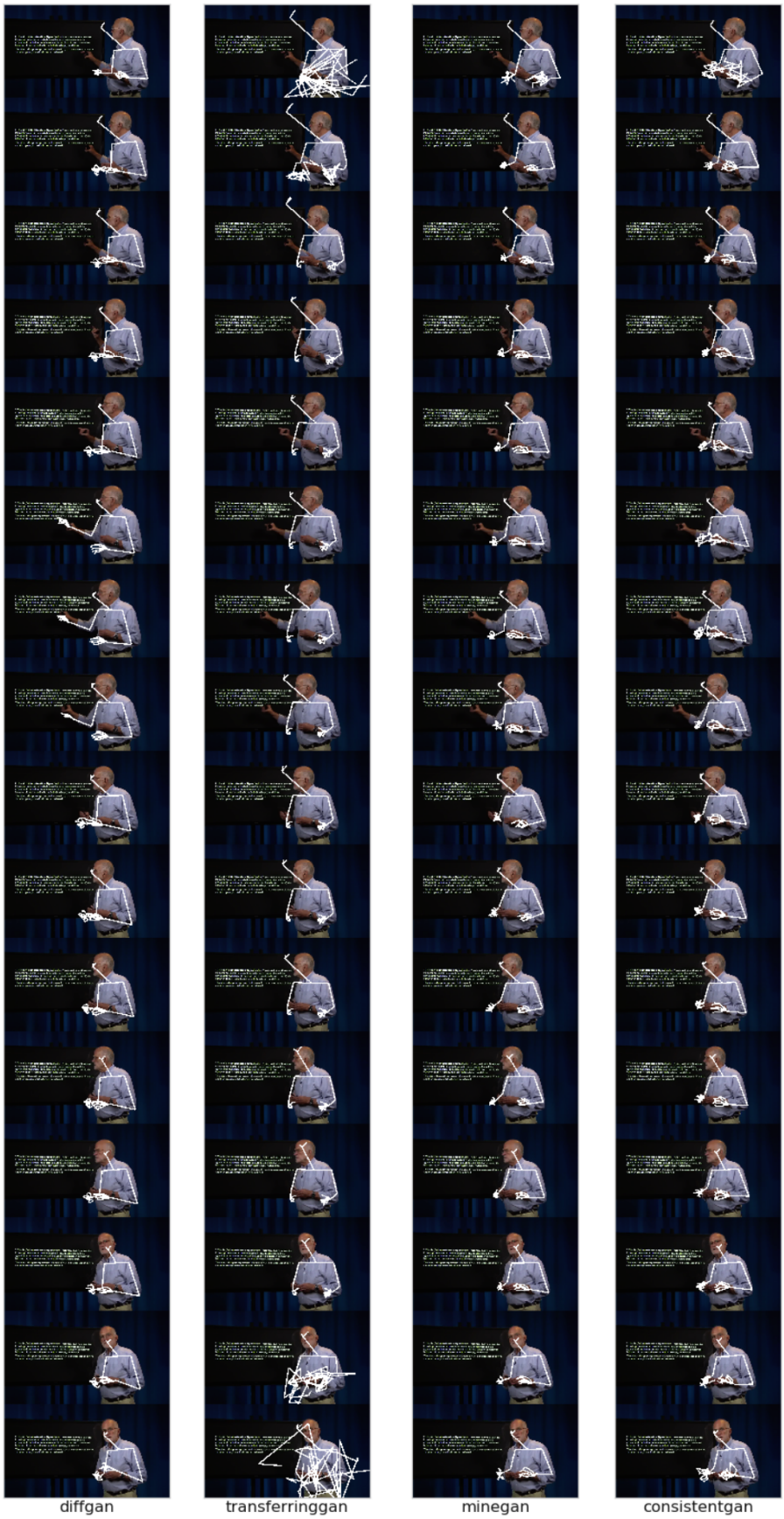

Figure 8. Qualitative comparison of our DiffGAN with prior work over shape of generated gestures. With `maher` as the source domain, target model outputs over the target domain `lec_evol` are superimposed over ground truth video frames for easy comparison.

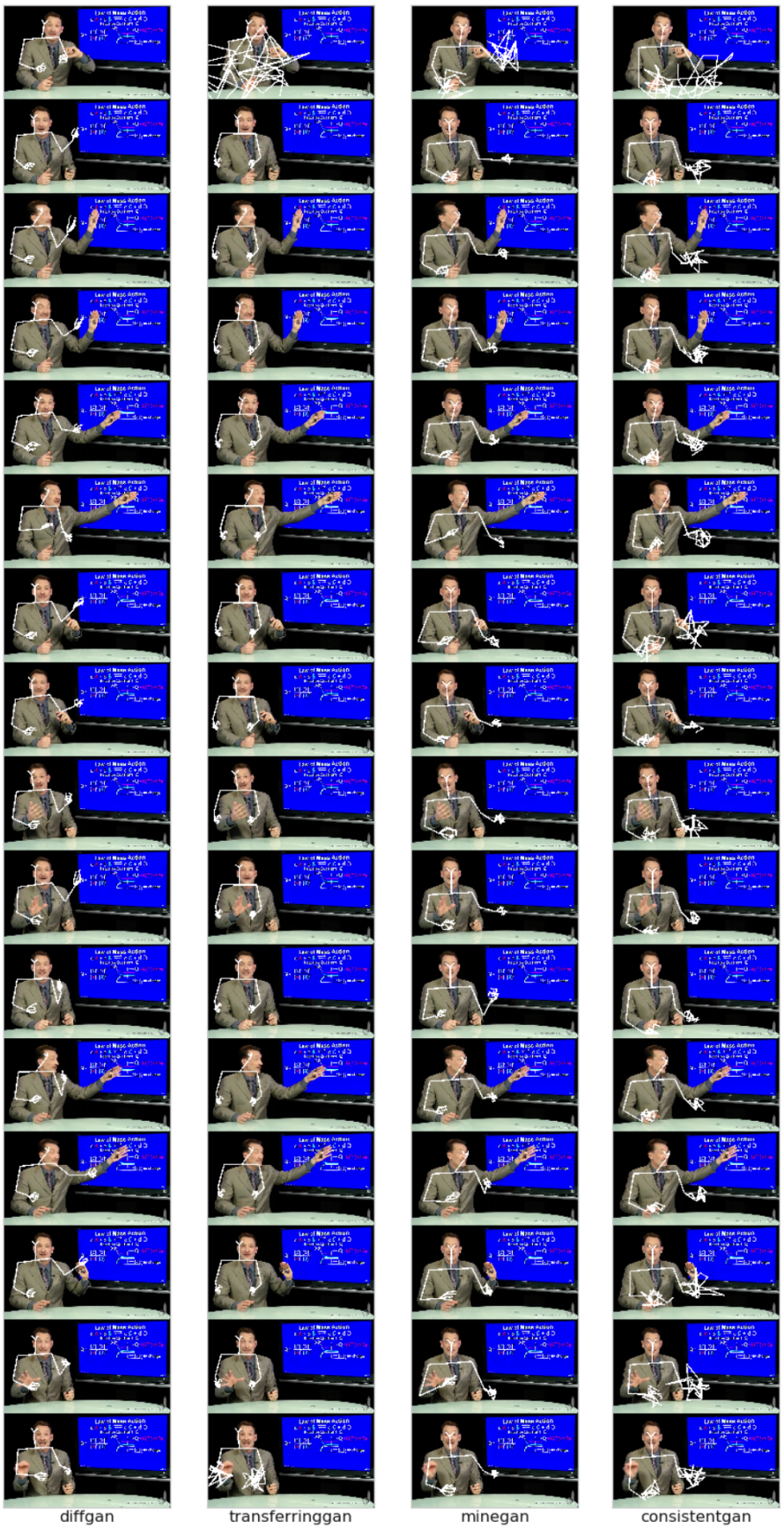

Figure 9. Qualitative comparison of our DiffGAN with prior work over shape of generated gestures. With `maher` as the source domain, target model outputs over the target domain `chemistry` are superimposed over ground truth video frames for easy comparison.

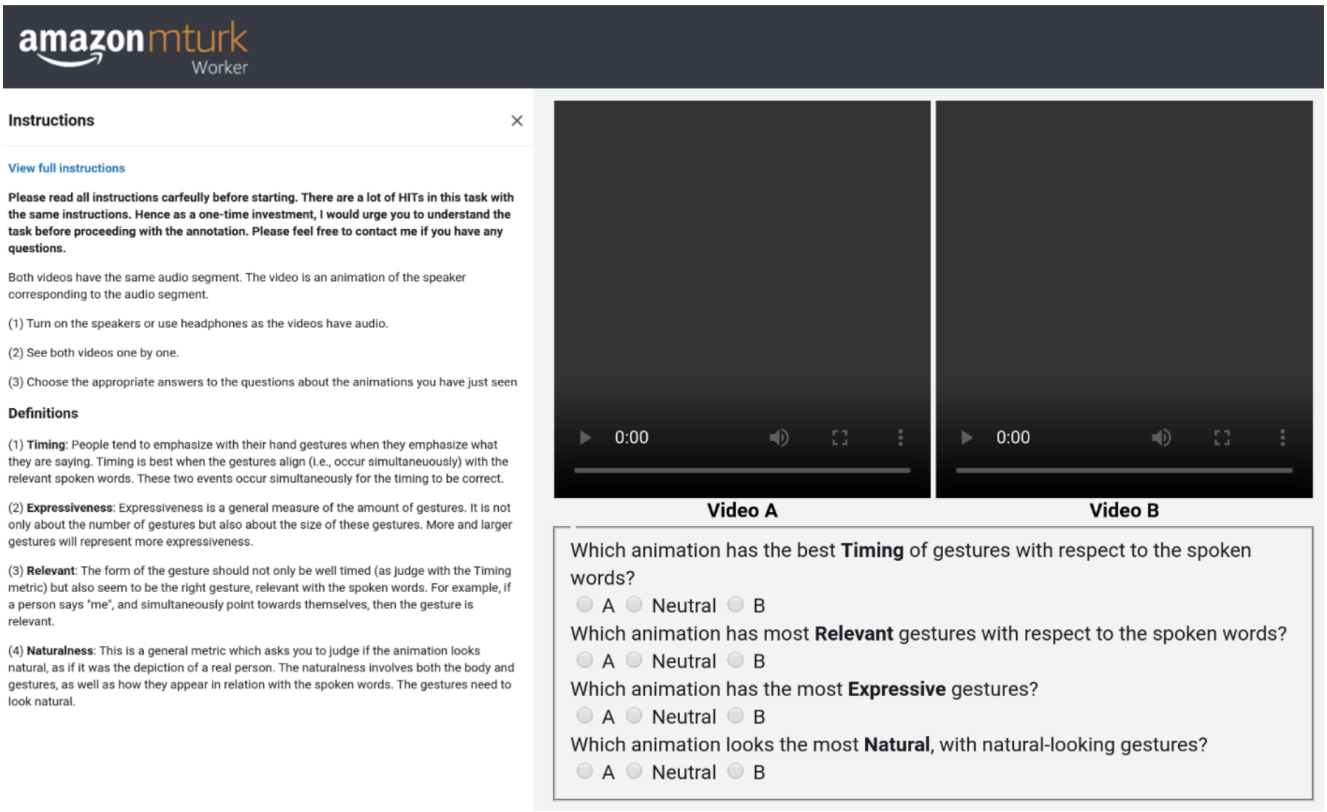

Figure 10. Screenshot of MTurk Experiment used to measure subjective metrics

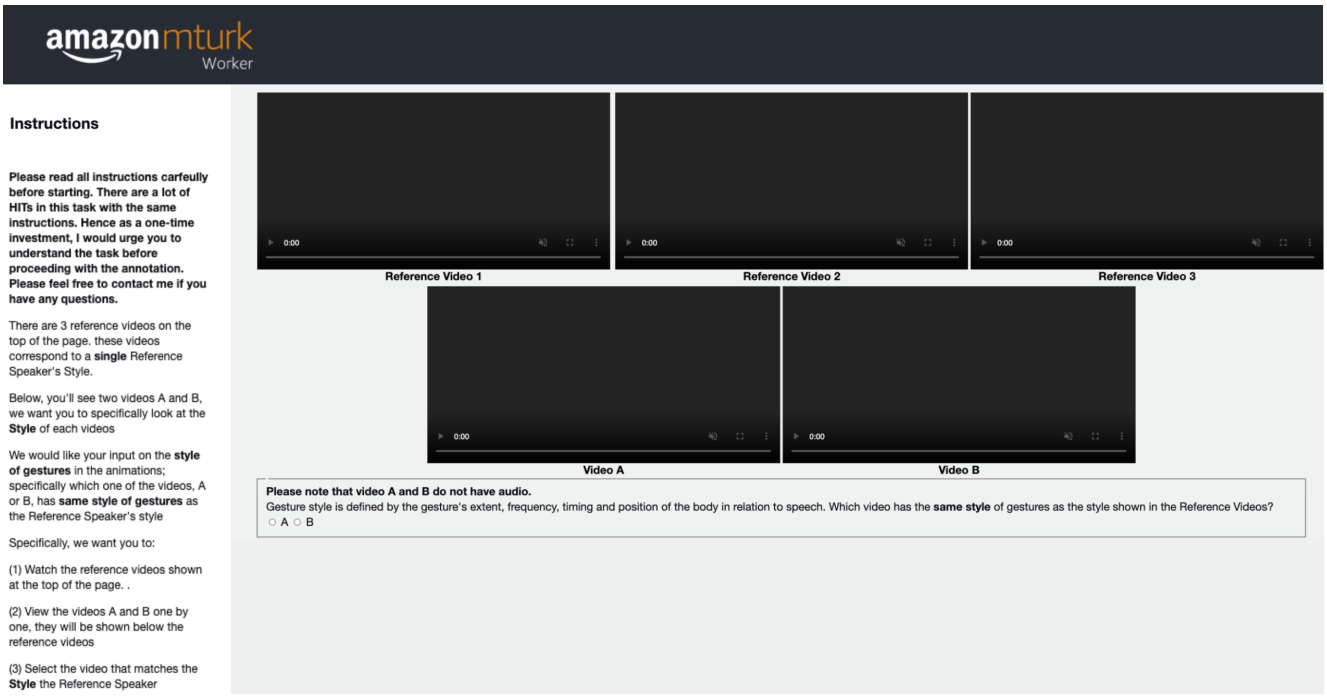

Figure 11. Screenshot of MTurk Experiment used to measure style metrics