Supplementary Material:
Continual Learning for Personalized Co-Speech Gesture Generation

Chaitanya Ahuja[1], Pratik Joshi[1], Ryo Ishii[2] & Louis-Philippe Morency[1]
[1]Language Technologies Institute, CMU & [2]NTT Human Informatics Laboratories
ahujachaitanya@gmail.com, pratikmjoshi123@gmail.com, ryoct.ishii@ntt.com, morency@cs.cmu.edu

A. Results and Discussion (contd.)

Qualitative evaluation Another aspect of gesture generation lies in the quality of the generation. We recommend the readers to view the supplementary video attached to get a better idea of quality of the generated gestures for our C-DiffGAN model as compared to other baselines.

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 1, 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 similar to the findings in [5, 3], where increasing the number of speakers consistently improves model performance over all speakers. Here, a larger diversity 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.

Impact of speaker order on performance Shuffling the order of speakers in the sequential learning process doesn't result in a significant performance difference for all the models, making our gains robust to any peturbation or differences in speaker order. Shuffling the order 4 times and retraining our model resulted in similar FID (i.e. 55.6, 44.0, 58.6, 46.8) and PCK (i.e. 0.35, 0.29, 0.32, 0.27). All of them strongly outperform all baselines.

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 3. 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.

Consistent baseline results with different training data sizes In Table 2, we show that even with just 2 minutes of data, our model shows significant improvement over the baselines. This is also corroborated by the forgetting curves in Figures 1 and 2

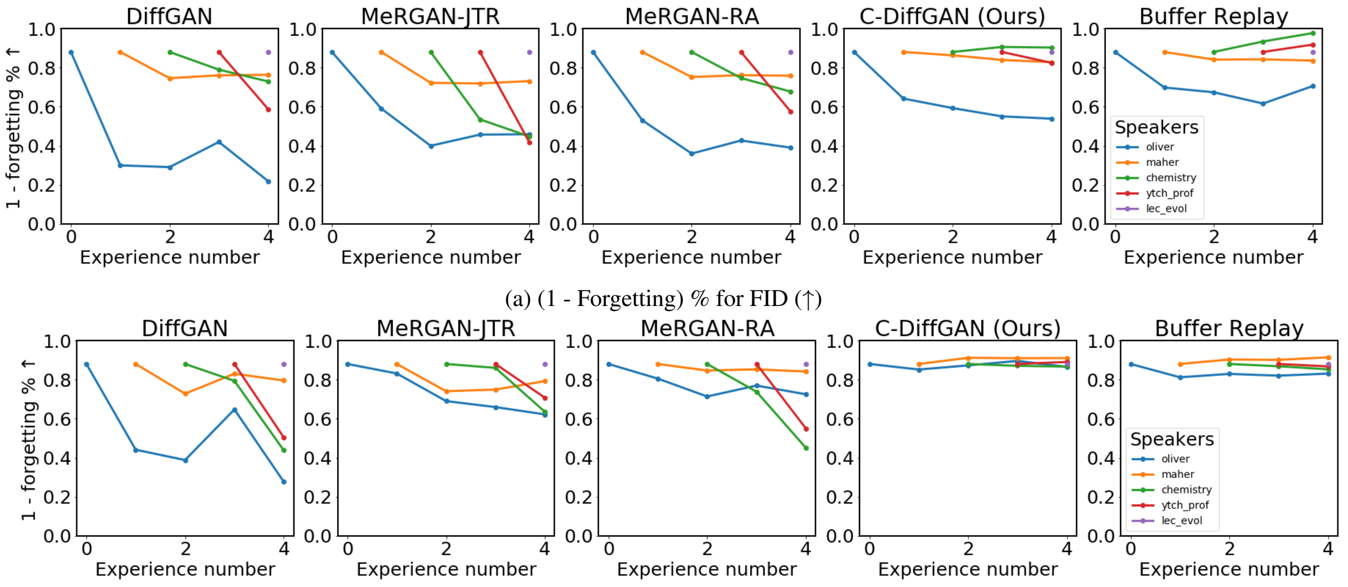
Human Perceptual Study We used Amazon MTurk to get a quantitative understanding of the quality of generated gestures for our C-DiffGAN and other models with respect to the ground truth.

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]. For measuring the Style, we asked the users if both videos had the same gesture style and whether they appeared to be performed by the same speaker. This metric measures was the model able to remember the speakers it had seen in the past.

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 4.

Definitions:

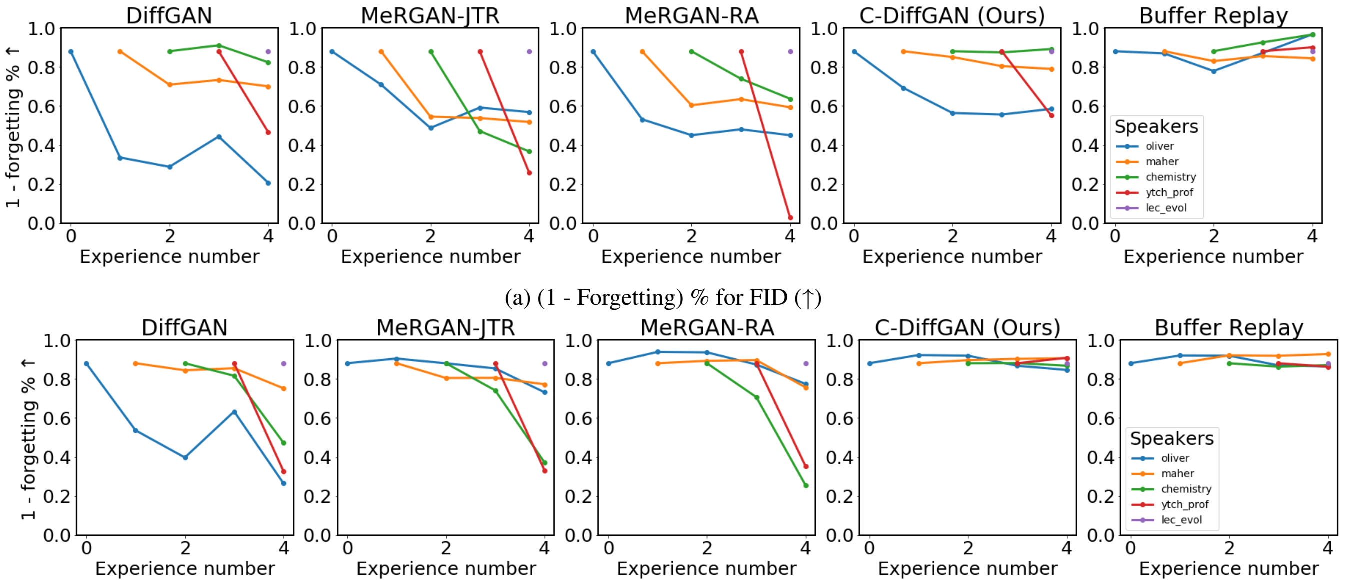
- **Style:** Gesture style is defined by the gesture's extent, frequency, timing, and position of the body in relation to speech.
- **Extent:** Gesture extent is the space around the speaker that the speakers' gestures (hand/arms) cover.



(a) (1 - Forgetting) % for FID (↑)

(b) (1 - Forgetting) % for PCK (↑)

Figure 1: 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 [1], MeRGAN-JTR [4] and MeRGAN-RA [4]. C-DiffGAN, on the other hand, is able to retain the performance over all the 5 experiences reasonably well.



(a) (1 - Forgetting) % for FID (↑)

(b) (1 - Forgetting) % for PCK (↑)

Figure 2: 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 clearly in DiffGAN [1], MeRGAN-JTR [4] and MeRGAN-RA [4]. C-DiffGAN, on the other hand, retains the performance over all the 5 experiences reasonably well.

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 1: 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.

Amount of Data (minutes)	**Training**	**Buffer Memory**	**Models**	**Average Final Accuracy**		**Average Forgetting**	
				FID↓	**PCK↑**	**FID↓**	**PCK↓**
2	**CL**	✗	**DiffGAN [2]**	350.3	0.20	343.3	0.16
	CL	✗	**MeRGAN-JTR [4]**	171.9	0.27	158.6	0.08
	CL	✗	**MeRGAN-RA [4]**	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 [2]**	613.6	0.16	674.5	0.18
	CL	✗	**MeRGAN-JTR [4]**	316.9	0.24	355.0	0.13
	CL	✗	**MeRGAN-RA [4]**	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 [3]**	22.0	0.40	-	-

Table 2: 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.

- **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.

Broader Impact The overarching aim of our work is to improve human-agent communication by improving the non-verbal gestures that agents can make when conversing with humans. The goal is to make agents more personable and natural, without requiring large amounts of data and re-training. While generated gestures are just joint movements and skeletal keypoints and can't be used for impersonation, they could potentially be used to enhance impersonations, such as Deep-

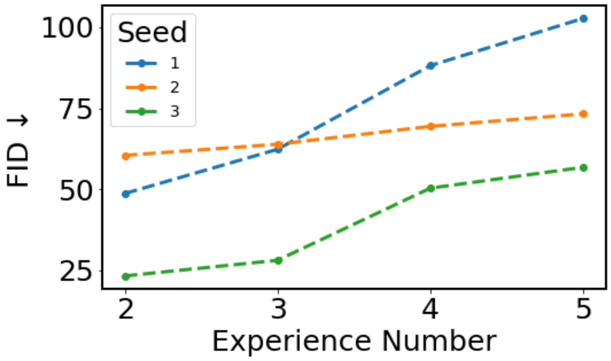


Figure 3: Measuring variability of FID scores across experiences and for three different seed values for the low-resource training data of `oliver` in our C-DiffGAN. The choice of the low resource training dataset can potentially have an impact on the distribution of the generated gestures.

fakes. Deepfakes can be used to spread misinformation and perpetrate scams and identity theft. We release our work under an ethical license as a starting point to discourage and prevent anyone from using our work to contribute to misinformation or hate speech (Do No Harm, Nonviolent Public or Hippocratic License).

References

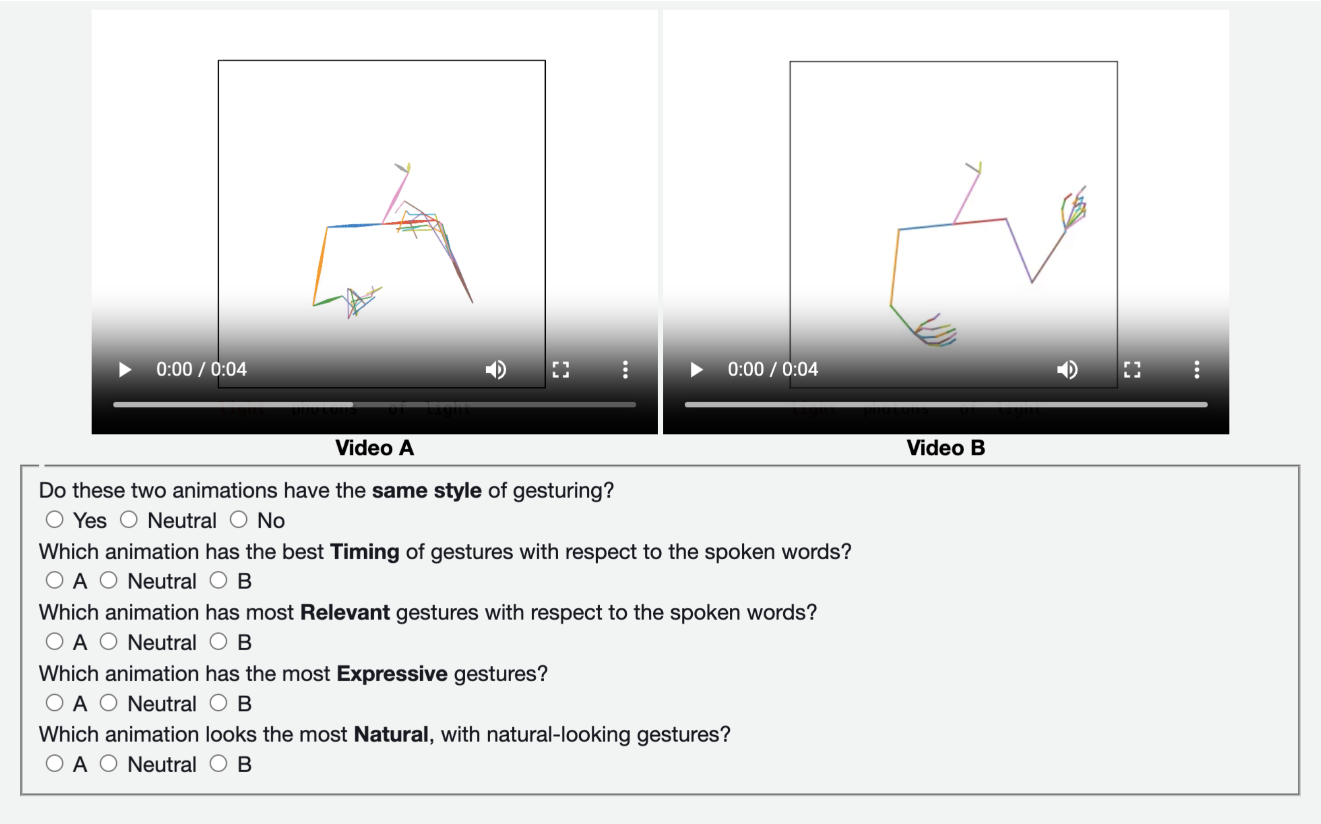
[1] Chaitanya Ahuja, Dong Won Lee, Ryo Ishii, and Louis-Philippe Morency. No gestures left behind: Learning relationships between spoken language and freeform gestures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1884–1895, 2020. 1, 2

[2] 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. 3

[3] 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, 3

[4] 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. 2, 3

[5] 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



Do these two animations have the **same style** of gesturing?

○ Yes ○ Neutral ○ No

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 4: Snapshot of the Mturk study with two videos, one is the ground truth while the other is a model generation. The order of the videos is randomized.