Style Conditioned Diffusion for Speech-To-Gesture Generation With Long-Term Context

Jonathan Windle[1] Sarah Taylor[2] David Greenwood[1] Iain Matthews[1]

INTRODUCTION

We approach the task of speech-driven gesture generation by introducing a new model that integrates a diffusion model with the Transformer-XL architecture which allows the long-term context of motion to influence predictions. We demonstrate effective style control using our approach and show how styles can be interpolated and varied over a single animated sequence.

METHOD

We introduce a ***Gesture Generation Network (GGN)*** (Figure 1) that combines a diffusion model with a Transformer-XL for predicting a sequence of poses, x, from a stream of audio, a, and a given style s.
We split the sequence into non-overlapping segments of length w frames, and compute Frame Feature Vectors (FFVs) to encode the speech using PASE+ and a learned style embedding.
The ***GGN*** gradually *denoises* a noisy sequence and generates a sequence of poses conditioned on speech and style.
The Transformer-XL uses an attention mechanism not only on the current sequence, but also the past context.
This knowledge is given in the form of reusable states from earlier segments, defined as $\boldsymbol{W_{t-m}}$ where m is the memory length.
We can use a ***GGN*** to remember long-term context of variable length that can extend beyond a single segment.
Through recurrent state reuse (Figure 2), our model overcomes the common sequence length limitation with Transformer based models while retaining knowledge of a long-term context.

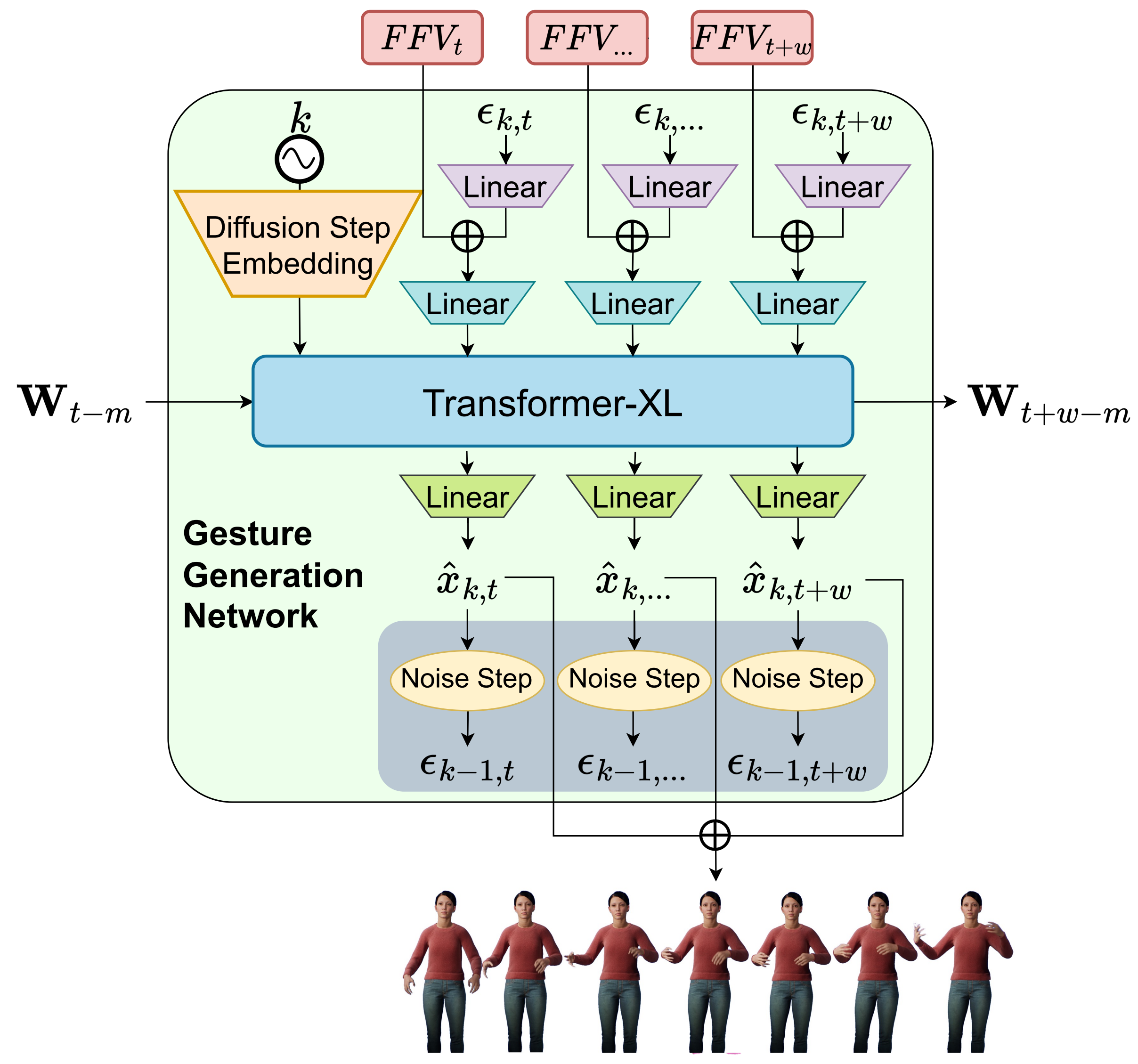

Figure 1: Gesture Generation Network. The diffusion process runs for $k = 1000:1$ steps. Given a sequence of feature vectors, FFV, and noisy pose vectors, ε, of length w, each step predicts the corresponding denoised pose sequence $\hat{x}$. For all steps $k > 1$, the prediction $\hat{x}$ is subsequently noised and fed to the next denoising step, concatenated with the same FFV. Colours indicate the same layer being used for each input when applicable.

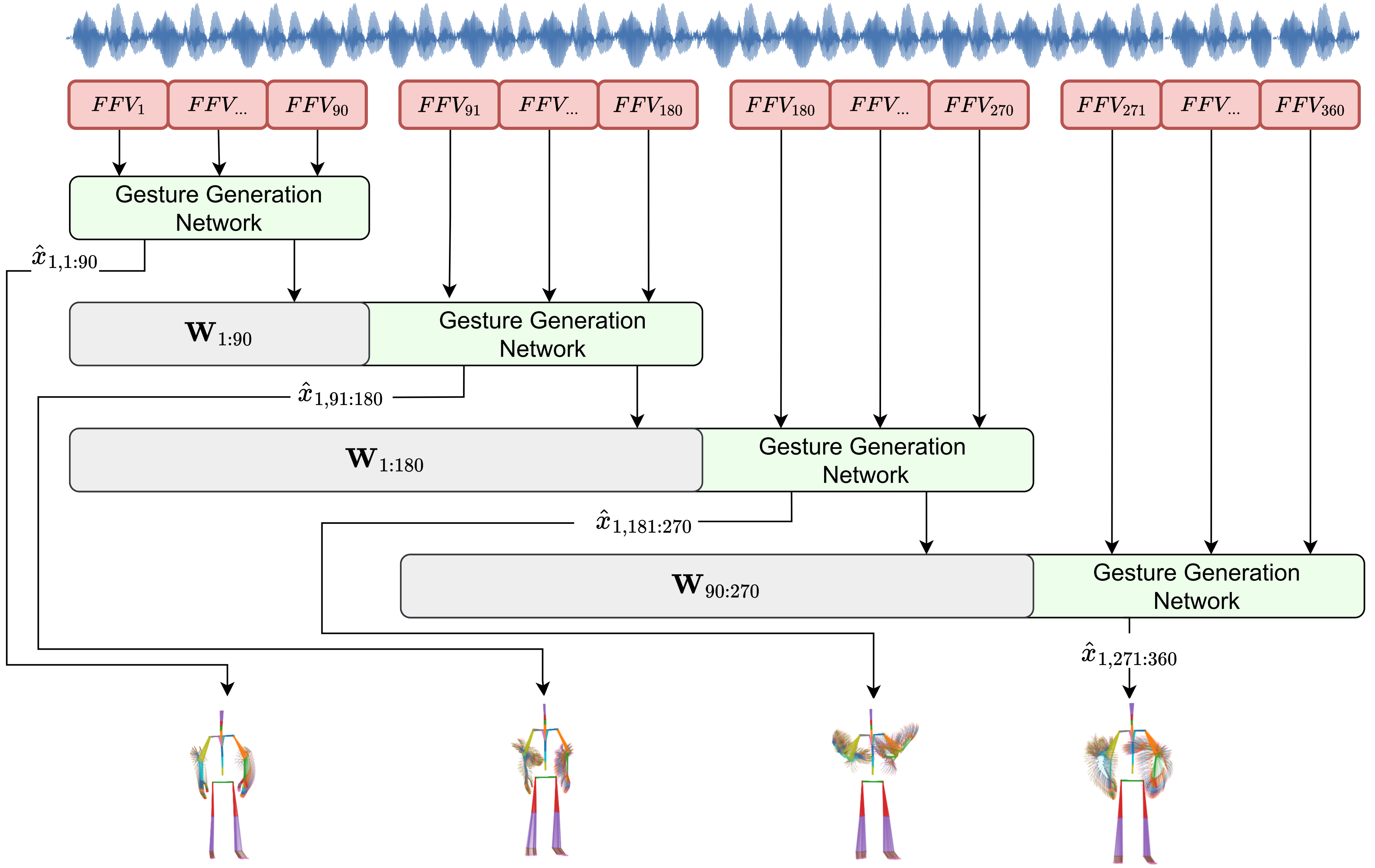

Figure 2: Overview of our long-term context model at inference time. Audio is split into segments of length 90 where each frame corresponds to a motion frame window (sampled at 30fps). Frame Feature Vectors (FFV) are derived from the audio segment. Each segment of FFV values are passed to a Gesture Generation Network as described in Figure 1. Each Generation Network will output 90 frames of motion and the previous states from the Transformer-XL model as $\boldsymbol{W}$ which stores the long-term context of up to 180 previous frames.

RESULTS

Scan For Video Results!

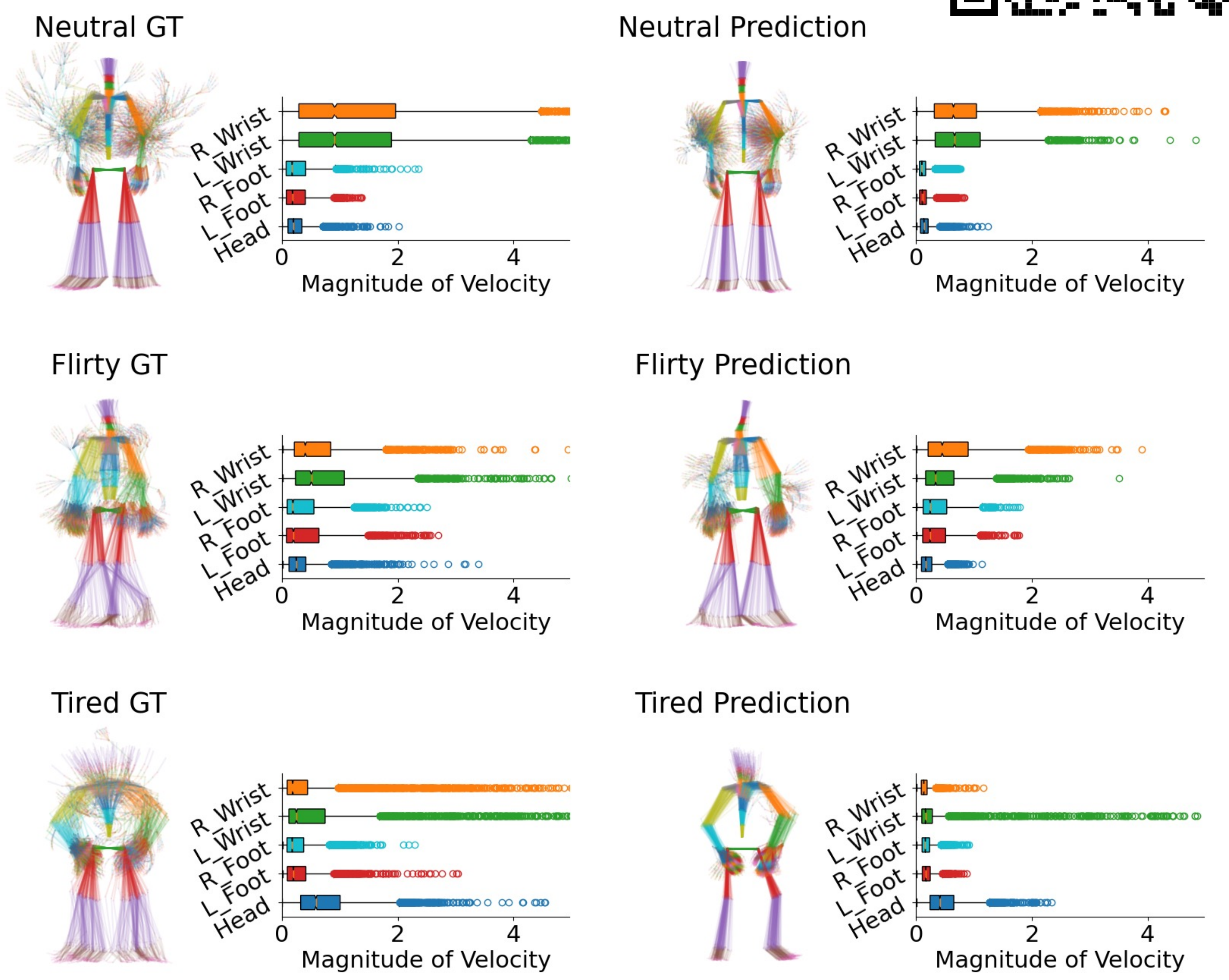
Neutral GT
Neutral Prediction
Flirty GT
Flirty Prediction
Tired GT
Tired Prediction

Figure 3: A test sequence from 3 different style categories shown for ground truth (left) and predicted sequences from the same audio (right) sampled every 1 second. Each example is shown with the distribution of the velocity magnitude for the wrists, feet and head to indicate the amount and speed of motion across the sequence.

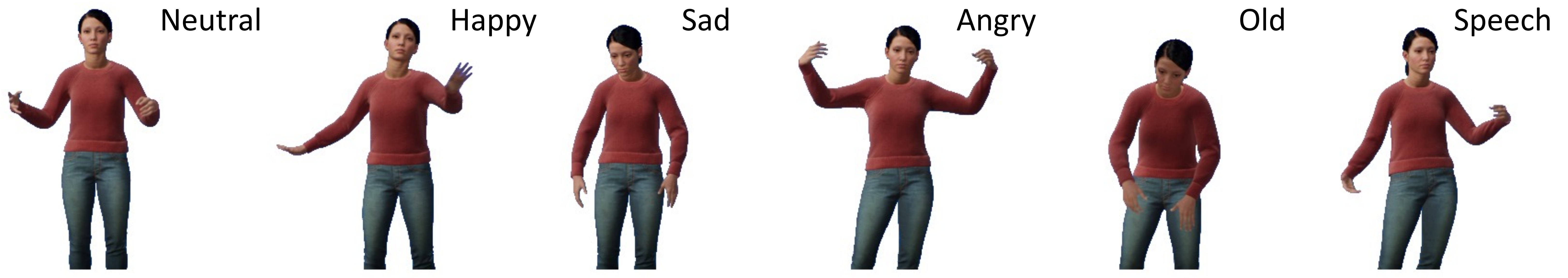

Figure 4: Rendered frames from animation generated from the same audio and noise sample but conditioned on different styles.

University of East Anglia

[1] University of East Anglia
[2] Independent Researcher