Supplemental Material for "From Audio to Photoreal Embodiment: Synthesizing Humans in Conversations"

Project Page Results Video Dataset, Code, Demo

1. Results Video

Our results video shows sequences of various individuals in different conversational settings from our dataset. Below, we denote the time stamp range associated with the discussion - (@mm:ss-mm:ss).

The results show that our model successfully models plausible, motion that is synchronous with the ongoing conversational dynamics. For instance, it correctly generates facial expressions and body language of someone feeling disgruntled *e.g*. dismissive hand wave and turning away (@02:58-03:11). The generated gestures are well-timed with the conversation *e.g*. raised finger with "I think" (@02:30-@02:50). Additionally, our approach can produce multiple plausible motion trajectories based on a single conversational audio input, each with distinct variations (@03:15-@03:55).

Compared against baselines and prior works, our method generates more "peaky" motion such as wrist flicks while listing (@04:16), and finger pointing (@04:50), which are both missed by a diffusion-based method LDA [Alexanderson *et al*. 2023]. In comparison to a VQ-based method SHOW [Yi *et al*. 2023], ours produces more dynamic motion with increased arm movement (@04:52), and seamless transitions between poses when switching from asking a question, to listening, to responding (@05:12-05:30). In contrast, SHOW moves to the audio but hovers around the same pose throughout. In comparison to both Random and KNN, gestures by our approach match the audio far better.

Notably, without any retraining, our method generalizes to conversational audio not seen in the dataset, such as a random movie clip audio (@05:44-@06:03). This is possibly due to the identity-agnostic training of Wav2Vec. We can also extend our method to the application of video editing, where we can reanimate a target person with a different motion trajectory by swapping guide poses (@06:10-06:27).

2. Contributions

We would like to emphasize the scientific value our work provides to the community: currently, motion generation researchers are stuck with mesh-only approaches, for which it is challenging to discern whether the issue lies in the quality of the avatar model or in the generated motion. For the first time, we (1) provide evidence that evaluators are more perceptive of subtle gestures in the photoreal space than in mesh space, which hides deficiencies (Fig. 8), (2) publish a photoreal gesture dataset that has not existed or was not widely available, allowing the community to make the necessary pivotal jump from mesh-based to photoreal representations. Public availability of this high quality data has the potential to cause a major leap forward in motion synthesis. (3) Our method provides a strong baseline that serves as a starting point for those looking to use our data. We provide a gradio app, code, and dataset on our project github.

3. Method

3.1. Pose representation

While we use a standard SO(3) representation for the joint angles, we note that not all joints are parameterized with 3 degrees of freedom (*e.g*. arm twist is only represented with roll, head bend with yaw, etc.). In total, we have 104 rotation angles across all of the joints.

3.2. Residual VQ-VAE

The residual VQ-VAE allows us to capture finer-grain details by employing a cascade of codebooks to capture progressively finer approximations. We use residual length of 4. In practice, this means we need a sequence of 4 VQ tokens to represent a single pose. To generate poses during test time for the diffusion model, we autoregressively output $4 \times K$ tokens one at a time, where K is the length of the downsampled sequence. For the both the encoder and decoder, we use a series of 1D convolutions of kernel size 2. The total receptive field for both the encoder and decoder is 8. We use a codebook size of 1024, and embedding size of 64. We train for 300k steps.

3.3. Guide pose Transformer

We adapt the diffusion model's architecture for the guide pose network. The transformer architecture is composed of masked self-attention layers that focuses only on previous timesteps to enable autoregressive prediction. The audio

is then incorporated using non-causal cross attention layers. This means the network doesn't see past motion, but sees the full context of audio. We then remove the diffusion timestep τ conditioning, and instead feed in an audio embedding (averaged over the whole time series) to the FiLM layers. While not necessary, this slightly helps the transformer to generate more plausible poses on the very first time-step. We use 2 masked self-attention layers and 6 cross-attention layers, all with 8 heads. We train for $\approx 100k$ iterations depending on the individual.

3.4. Comparisons on other data

We emphasize that our primary scientific contribution lies in highlighting the need for photorealistic avatars when evaluating fine-grain motion in gesture generation. Since prior works are constrained to datasets with untextured meshes, we publicly release a photoreal dataset to facilitate this transition. We propose a baseline method to set a benchmark for using this novel dataset. In Tables 1 and 2, we show that our ideas (e.g. guide poses) work well vs. prior motion generation works.

Training on other data is nontrivial as they rarely include full face, body, and hands. Even when they do, transforming the joints from one skeleton to another is nontrivial. Furthermore, since prior datasets are all monologues (one speaker only), we retrain prior SOTA (LDA [2] and SHOW [40]) on our dyadic data, where we observe back-and-forth, listening/speaking dynamics missing from monologues.

Still, to provide evidence that our method is well-justified, we evaluate the full motion (Tab. 1) and isolated face (Tab. 2) against prior SOTA. We also ran a small experiment with CodeTalker [Xing *et al*. CVPR 2023], which is a purely lip-sync method. A new row in Table 2 for *CodeTalker* would be 2.59, 2.01, 1.90 (vs. Ours 2.29, 1.89, 1.76), which has similar failure cases as SHOW (L805-08).

3.5. Dataset info

For the data collection process, we had participants in a multi-view capture dome. Our processing pipeline is a standard tracking pipeline which closely follows the procedure described in Multiface [Wuu et. al. 2023]. For full body reconstruction, we extended the pipeline from face to body accordingly. The individuals are all captured at a resolution of 2048×1334 pixels. From the multi-view cameras, we then turn them into tracked meshes including headposes and unwrapped textures of 1024×1024 pixels. The capture system also tracks metadata including camera intrinsics and extrinsics, along with the audio of each person. The code and dataset of the original capture system is released in this repository.

For our conversational dataset, we use the above approach to capture 4 participants. We take pairs of participants and ask them to discuss a vague topic. The topics we assign them range from selling an object to interviewing a candidate. However, in order to make sure the interactions do not feel too canned, we keep the prompt vague.

For each conversation, one person is standing in a multiview dome while the other is sitting in a separate dome. They view each other from a screen and have microphones to speak to one another in real time. However, we do note that this setup is a bit limiting in that it allows us to reconstruct the full body of one person, and only the face of the other. Furthermore, having them sit in separate domes may affect the interactions. However, we leave these improvements for future work.

3.6. Implementation details

We use a max sequence length of 600 frames at 30 fps (20 second videos). During training, we randomly sample a sequence between 240 frames and 600 frames. We then train on padded sequences of random lengths for all of our networks. This allows us to generate sequences of arbitrary length during test time. We train each network for each subject in the data separately. All networks are trained on a single A100. Approximate train times: face diffusion model (8 hr), VQ + coarse pose predictor (5 hr), pose diffusion model (8 hr).

4. Results

4.1. Quantitative metrics

Proposing an effective quantitative metric for evaluating dyadic motion is still an open problem [28]. For instance, **NN** and **Random** have better FD_k and Div_g because they are simply samples from the train set (not generated). Thus, it is expected they would have the diversity and velocity of motion that match the test set. However FD_g, measuring the "realism" of the generated static poses/expressions, is bad for both because they do not appropriately match the conversational dynamics, and thus fail to match the true expressions/motion of ground truth (L732-36). It is therefore difficult to reduce the quantitative metric down to a single number. Like prior work [2], we break down the metrics into different axes- realism and diversity- and find the best balance across all metrics (Table 1). Similar to [2, 3, 28], a well-designed perceptual evaluation is more reliable. Here, we significantly outperform prior work. We have released our evaluation setup in our codebase.

4.2. Perceptual evaluation

For each Ours vs. GT (mesh), vs. GT (photoreal), vs. LDA (mesh), vs. LDA (photoreal), we generate 50 A-B tests. For each test, we ask 3 different evaluators, totalling to 600 evaluators. Each A-B test contained 14 questions. Prior to the actual test, we provide a headphone check to

make sure the evaluators are listening to audio. However, we do not ask additional questions that check to see if they are actually listening to the speech. The landing page describes the task and walks evaluators through 2 examples. To ensure the evaluators are not randomly clicking, we include 3 questions with an obvious mismatch (one speaker laughing while the listener is neutral) twice. If the evaluator selects a different response for these duplicated questions, we do not allow them to submit.

4.3. Ablation with VQ-only method

In the main paper, the VQ-only baseline is represented with prior work SHOW [Alexanderson *et al*. 2023], which is very similar to our guide pose network. For completeness, we also train a VQ-only baseline using our network architecture. We see very similar results to SHOW and similar limitations. Quantitatively, $\text{FD}_g = 5.00$, $\text{FD}_k = 2.80$, $\text{Div}_g = 2.20$, $\text{Div}_k = 1.89$. Note the higher FD and lower diversity compared to our complete method. We notice that after many timesteps, drift often happens which causes the method to either get stuck in a local minima (no motion).

4.4. Unified model

We conjecture that pretraining on an outside larger dataset or on the full dataset would improve results. While we briefly attempted training on the full 8 hours of data across all IDs, this did not significantly improve results. We conjecture that 6 hours of additional data is not too significant. As such, we opted to keep things simple and trained ID specific models.