AV-Flow: Transforming Text to Audio-Visual Human-like Interactions

Supplementary Material

Contents

The supplementary document is organized as follows:

A. Additional Results
B. Implementation Details
C. Ethical Considerations

We strongly encourage the readers to watch our supplementary video.

A. Additional Results

Audio-Visual Guidance. As mentioned in Sec. 3.5, we propose to provide audio-visual guidance from a participant (user) in dyadic conversations. We condition the model to visual information, by extracting features $\boldsymbol{s}_i$ using SMIRK [10] from the monocular video of the user. For the audio, we further extract ASR tokens $\boldsymbol{a}_i^p$ from the audio channel of the user. We notice that both modalities are important, in order to produce realistic and meaningful interactions. However, overall for our conversational data, the results feel realistic even if we condition on only one modality. In Fig. 1, we demonstrate some cases where we notice some difference when only one modality is available. In the first row, the avatar better mirrors the expression, with a wider smile, when visual information is present. In the second row, it produces a realistic but unnecessary smile with only visual guidance, whereas it loses eye contact with only audio guidance. In the third row, the audio seems to play an important role that makes the user to smile. Overall, since we have available both audio and video, we propose to condition AV-Flow to both modalities to produce our photo-realistic *always-on* avatar.

Tab. 1 shows the quantitative results for the basic AV-Flow without guidance and with audio-visual guidance. We compute the F1-score for the lip closures, as well as the F1-score for the smiles. The lip closures are detected by measuring the distance of the vertices of the inner upper and lower lips for the 3D mesh per frame. Similarly, smiling is detected when the distance of the left and right corners of the mouth is larger than a threshold. We also compute the Fréchet distance between ground truth and generated face expressions (FD_e) to estimate the distribution distance. With our proposed guidance, we notice an increase in F1-score for the smiles and a decrease in FD_e for the dyadic setting, as the avatar produces more realistic reactions and facial expressions while listening to the user. Our basic AV-Flow achieves a slightly more accurate lip synchronization.

Fig. 3 shows the L2 norm of the SMIRK [10] expression codes over time for the ground truth actor, participant, and generated actor with and without guidance. The graph shows how the actor reacts at the same time or before/after the participant, while they interact. Notice how the L2 norm of expression codes at around 400 frames (peak) is significantly lower without guidance (*i.e.*, more neutral expression), and correlation differences appear at 550-650 and 150-250 frames. See also suppl. video.

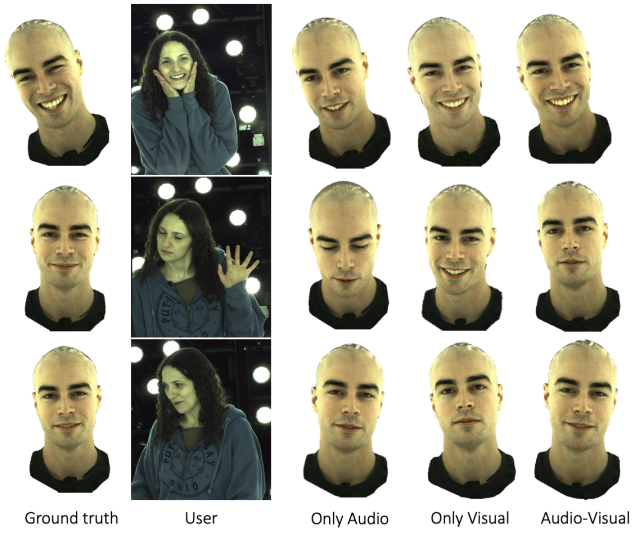

Figure 1. **Audio-Visual Guidance.** AV-Flow with audio-visual guidance (user input) produces more realistic expressions while listening. In audio-only guidance, the avatar might lose eye contact or not mirror a smile. In visual-only guidance, it might produce more smiles than needed.

Method	$\text{F1}_{lips}\uparrow$	$\text{F1}_{smiles}\uparrow$	$\text{FD}_e\downarrow$
AV-Flow w/o guidance	**0.964**	0.611	0.861
AV-Flow w/ guidance	0.933	**0.685**	**0.845**

Table 1. **AV-Flow with or without Guidance.** In dyadic conversations, the proposed audio-visual guidance leads to more realistic reactions and facial expressions, while the avatar is listening to the user. Without guidance, our basic AV-Flow achieves slightly more accurate lip synchronization.

Audio-Visual Alignment. We design intermediate highway connections that enable communication between the audio and visual diffusion transformers. In our ablation study in Sec. 4.1, we notice that our proposed audio-visual fusion leads to the best correlation between audio and motion, as measured by the BC_h and BC_e metrics. This audio-visual correlation is also shown in Fig. 4. We observe patterns where the energy of the facial motion matches the energy of the corresponding synthesized audio (plotted as

Figure 2. **Qualitative Results of AV-Flow.** We show frontal and side views for corresponding phonemes. We use pre-trained personalized renderers [1] that synthesize photo-realistic 3D avatars.

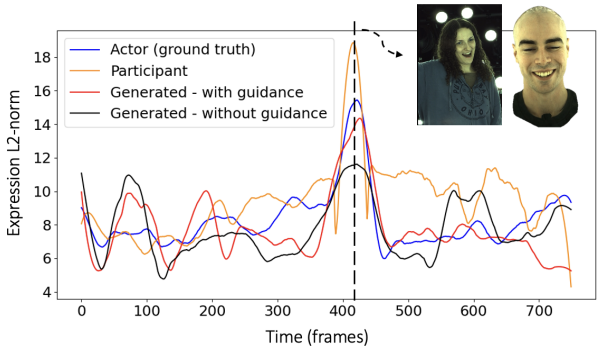

Figure 3. **Guidance over time**. L2 norm of expression codes for the ground truth actor, the participant, and the generated actor with or without guidance. With guidance, AV-Flow produces the appropriate reaction in dyadic interactions.

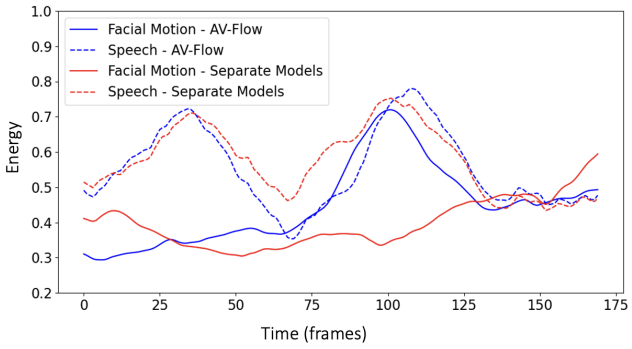

Figure 4. **Audio-Visual Alignment.** Correlation between synthesized speech and facial motion by AV-Flow, compared to the variant of separate models without any connections. Energy is estimated as the normalized squared L2-norm of the generated facial dynamics and mel-spectrogram over time.

the normalized squared L2-norm of the generated facial dynamics and mel-spectrogram over time). We compare with the variant of the separate models, without any cross-modal connections, where the correlation is lower.

Additional Qualitative Results. Fig. 2 shows additional qualitative results of our method, rendered in frontal and side views. We use pre-trained personalized renderers [1, 7] that produce photo-realistic 3D avatars. Therefore, we can render the generated avatars from any viewpoint. In this work, we are mostly interested in the facial expression, lip synchronization and realistic head motion over time. We only provide the side views for completeness. We refer the interested reader to [1, 4] for more details in Codec Avatars.

Video Results. We encourage the readers to watch our supplementary video.

B. Implementation Details

Text-to-Tokens. Since our training dataset does not include any text annotations, we extract tokens (logits) from the raw audio using an ASR model (see Sec. 3.1). At inference time, in order to be able to synthesize audio-visual content directly from text characters, we learn a text-to-tokens model. Inspired by the architecture proposed by Matcha-TTS [6], we first map the input text to phonemes. We learn phoneme embeddings (192-dimensional) that are passed through a text encoder of 3 1D convolutional layers. A duration predictor gives their duration. A diffusion transformer of 3 layers maps the features to logits, which can be used as input tokens to our model. We follow the rest hyperparameters, architecture and training with flow matching of Matcha-TTS [6]. Our main difference is that we predict character-level logits, not mel-spectrograms. We use LJSpeech [2] to train our text-to-tokens model.

Architecture. We use $N = 8$ blocks for our audio and visual DiTs. We first project the inputs to 512-dimensional through a linear layer. Each transformer block has input

and output dimensions of 512, hidden size of 1024, and 4 heads for the multi-head self-attention. We use windows of 10 frames, looking only 2 frames in the future. With this windowing, we achieve only 120ms latency, as mentioned in Sec. 4. We have also tried windows of 20 frames, getting similar results, but a bit higher latency. As described in Sec. 3.2, we upsample the data at 86fps to achieve exact correspondence between audio and video. We extract mel-spectrograms following the same extraction as BigVGAN [3]. In this way, we directly use the pre-trained BigVGAN as our vocoder to get the output speech signal. Our input tokens are extracted from a pre-trained Wav2Vec2 model with the base architecture, that is trained for ASR using 960 hours of unlabeled audio from the LibriSpeech dataset [8, 9, 12].

Training. During training, we use a batch size of 16 segments. Each segment corresponds to a duration of 20 seconds. We set $\sigma_{\text{min}} = 10^{-6}$. Our implementation is based on PyTorch [9]. We use AdamW optimizer [5] with a learning rate of 10^{-4}, and hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-9}$. We train AV-Flow for about 36 hours (1 million iterations) on a single A100 GPU. For Audio2Photoreal, we use the provided train-test split. For our 50h dataset, we keep $\sim 20\%$ for test set, explicitly separating 16 users for training and 4 for test.

Pre-trained Models. In the main paper, we provided the corresponding references of pre-trained models used in our method. For completeness, we include here specific URLs:

- Audio2Photoreal data, decoders, renderers: https://github.com/facebookresearch/audio2photoreal.
- Wav2Vec2 model: https://docs.pytorch.org/audio/stable/generated/torchaudio.pipelines.WAV2VEC2_ASR_BASE_960H.html#torchaudio.pipelines.WAV2VEC2_ASR_BASE_960H.
- BigVGAN vocoder: https://huggingface.co/nvidia/bigvgan_base_22khz_80band.
- SMIRK model: https://github.com/georgeretsi/smirk.

C. Ethical Considerations

We use the publicly available dataset proposed by Audio2Photoreal [7]. We also collected an additional dataset of 50 hours in a similar setting. Both datasets include dyadic conversations between individuals. The data include raw audio and video, as well as face expression codes and head poses for the actors, paired with pre-trained personalized renderers [1, 4]. During collection of the data, we have followed appropriate procedures and all individuals have provided their full consent for our research work. Our model is identity-specific and thus, only those individuals can be rendered and no one else. This addresses ethical concerns of generating subjects without their consent, or generating misleading content. We have also used audio from the EARS dataset [11] to test lip synchronization to custom input speech and the widely used LJSpeech [2] to train our text-to-tokens module.

Although we have strictly followed all these procedures in collecting and using our data, we would like to note the potential misuse of similar technologies in generating photo-realistic human avatars. Apart from the benefits for education, virtual communication, healthcare, etc, there is still the possibility of generating misleading content. Research on fake content detection and forensics is crucial. We intend to release our source code to help improving such research.

References

[1] Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabián Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. Driving-signal aware full-body avatars. *ACM Transactions on Graphics (TOG)*, 40(4):1–17, 2021. 2, 3

[2] Keith Ito and Linda Johnson. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017. 2, 3

[3] Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. *arXiv preprint arXiv:2206.04658*, 2022. 3

[4] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Trans. Graph.*, 37(4):68:1–68:13, 2018. 2, 3

[5] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[6] Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. Matcha-TTS: A fast TTS architecture with conditional flow matching. In *Proc. ICASSP*, 2024. 2

[7] Evonne Ng, Javier Romero, Timur Bagautdinov, Shaojie Bai, Trevor Darrell, Angjoo Kanazawa, and Alexander Richard. From audio to photoreal embodiment: Synthesizing humans in conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1001–1010, 2024. 2, 3

[8] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015. 3

[9] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 3

[10] George Retsinas, Panagiotis P. Filntisis, Radek Danecek, Victoria F. Abrevaya, Anastasios Roussos, Timo Bolkart,

and Petros Maragos. 3d facial expressions through analysis-by-neural-synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[11] Julius Richter, Yi-Chiao Wu, Steven Krenn, Simon Welker, Bunlong Lay, Shinjii Watanabe, Alexander Richard, and Timo Gerkmann. EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation. In *Interspeech*, 2024. 3

[12] Yao-Yuan Yang, Moto Hira, Zhaoheng Ni, Anjali Chourdia, Artyom Astafurov, Caroline Chen, Ching-Feng Yeh, Christian Puhrsch, David Pollack, Dmitriy Genzel, Donny Greenberg, Edward Z. Yang, Jason Lian, Jay Mahadeokar, Jeff Hwang, Ji Chen, Peter Goldsborough, Prabhat Roy, Sean Narenthiran, Shinji Watanabe, Soumith Chintala, Vincent Quenneville-Bélair, and Yangyang Shi. Torchaudio: Building blocks for audio and speech processing. *arXiv preprint arXiv:2110.15018*, 2021. 3