Supplementary Material
Responsive Listening Head Generation: A Benchmark Dataset and Baseline

Mohan Zhou[1*†], Yalong Bai[2†], Wei Zhang[2], Ting Yao[2], Tiejun Zhao[1§], and Tao Mei[2]

[1]Harbin Institute of Technology [2]JD Explore Academy, Beijing, China
{mhzhou99, ylbai}@outlook.com, {wzhang.cu, tingyao.ustc}@gmail.com,
tjzhao@hit.edu.cn, tmei@jd.com

This supplementary material provides additional information about our ViCo dataset and our responsive listening head generation baseline. In Appendix A, we provide the more detailed statistical information about the ViCo dataset. In Appendix B, we further include the YouTube copyright and discuss the IRB approval. In Appendix C, we show more details of our responsive listening head generation pipeline. In Appendix D, we further include more additional experimental results. Applications and limitations are discussed in Appendix E.

A Dataset Details

A.1 Clip Length / Duration Distribution

We counted the distribution of clip lengths and the corresponding duration percentage in ViCo, and the results are shown in Fig. A1.

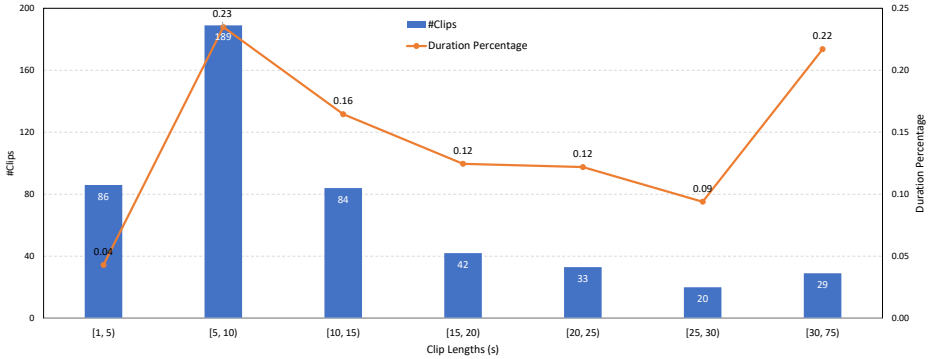

Fig. A1: The clip length distribution and the corresponding duration percentage in ViCo dataset.

∗ This work was done at JD Explore Academy.
† Equal contribution. § Corresponding author.

A.2 Identities

The ViCo dataset contains 92 identities. In the Fig. A2, we picked a random image for each identity. It can be found that there are different genders and races; thus, the potential ethics problems can be reduced, *i.e.*, the dataset represent the diversity of the community.

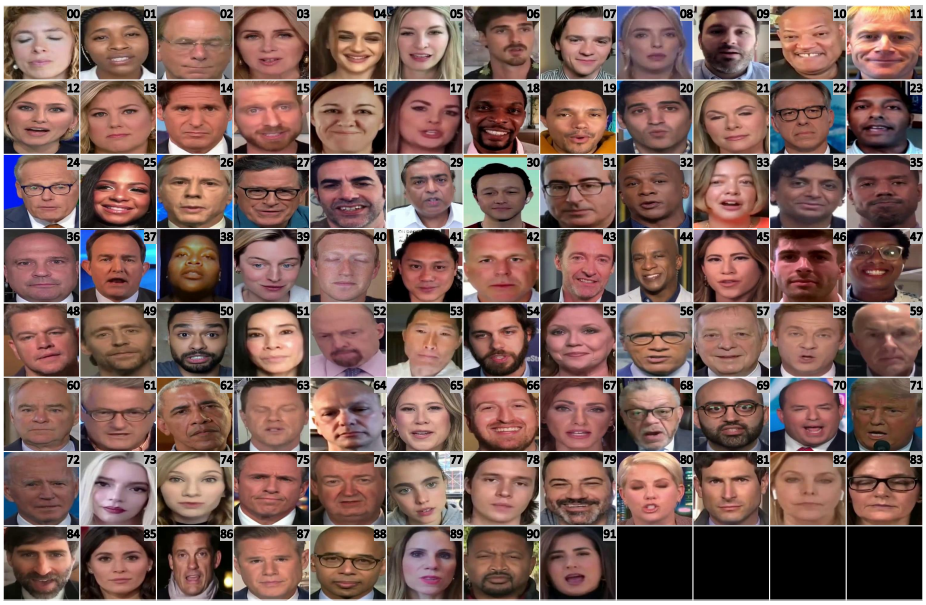

Fig. A2: Ninty-two identities with different genders and races in our dataset.

Some identities in our dataset may act as different roles (speaker/listener) and present different listener patterns under different attitudes. So, we statistic the identity distribution in Fig. A3. This demonstrates the diversity of our dataset.

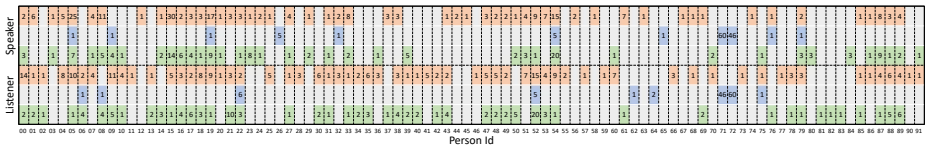

Fig. A3: The identity distribution in our dataset. If an identity in a certain role (speaker/listener) shows a certain attitude (positive, negative and neutral), then the corresponding cell is colored, and the number of clips is also attached.

B IRB Approval

The YouTube community has a strict censorship mechanism to avoid violent or dangerous content [2]. And as shown in the copyright[1], research purpose is considered fair use, which is allowed the reuse of copyright-protected material without getting permission from the copyright owner. This guarantees our dataset is congruence with the ethics guidelines.

C Pipeline Details

C.1 3DMM Coefficients

We extract the 3DMM coefficients following [1,5], with the commonly used toolkit[2] and the guides of PIRender[3], we can obtain a parametric representation of the face: $\{\alpha, \beta, \delta, p, \gamma\}$ which denote the identity, expression, texture, pose and lighting, respectively. Here, $\alpha \in \mathbb{R}^{80}, \beta \in \mathbb{R}^{64}, \delta \in \mathbb{R}^{80}$, and $\gamma \in \mathbb{R}^{27}$ for RGB channels in three-bands Spherical Harmonics [3,4] representations, $p \in \mathbb{R}^6$ to represent rotations with SO(3)$\in \mathbb{R}^3$ and translations in $\mathbb{R}^3$.

Therefore, the relative fixed, identity-dependent features $\mathcal{I} = (\alpha, \delta, \gamma)$ is in $\mathbb{R}^{187}$, and the relative dynamic, identity-independent features $m = (\beta, p)$ is in $\mathbb{R}^{70}$. Additionally, to better model the head movements and make it compatible with PIRender [5], we use a new "crop" parameter of $\mathbb{R}^3$ in practice. This guides where we will place and size the parametric 3D face in the original image.

C.2 Model Complexity

Our proposed model is lightweight (248K params) and efficient (941K flops for 30 frames input) with the 3DMM coefficients.

D Additional Experimental Results

Qualitative Results with Different Modality Inputs We conduct an ablation study on the impact of different speaker signals for listening head modeling. The qualitative results are shown in Fig. A4. We sample three frames from $\mathcal{D}_{test}$ (Fig. A4a) and $\mathcal{D}_{ood}$ (Fig. A4b) to make a simplified but clear visualization. From Fig. A4a, show that models with only a single input have less head motion and expression variations, and furthermore, the audio only model is able to express some expressions (columns 1, 2) as full input model, while the video only model performs even worse in expression modeling, which may indicate that the speaker's audio signals contributes more on listener's expression. From Fig. A4b, we can find that the listener is more likely to lose focus in the absence of video input (columns 2, 3), which may indicate that the speaker's visual signals can guide the listener where to focus.

[1] https://www.youtube.com/howyoutubeworks/policies/copyright/#fair-use
[2] https://github.com/microsoft/Deep3DFaceReconstruction
[3] https://github.com/RenYurui/PIRender

(a) (b)

Fig. A4: Qualitative comparisons of listening head generation with different speaker signals (shown in rows). The attitude is positive for all listening heads

Table A1: The Feature Distance (×100) of generations with different modality inputs and architectures **across all attitudes** (Averaged) on $\mathcal{D}_{test}$.

Method	angle	exp	trans
Audio only	8.69	17.19	8.49
Visual only	8.06	18.85	7.54
Non-sequential Model	8.40	18.50	7.06
Ours	**7.79**	**15.04**	**6.52**

Comparisons with Non-sequential Model Furthermore, we also experimented generate the listener frames in a "purely parallel (non-sequential) manner" rather than our "auto-regressive decoder manner", by removing the temporal connections and converting the LSTM cells to fully-connected layers with similar #params. The results are shown in Tab. A1 which worse than our model, since the "non-sequential model" results in noise and unnatural head motion caused by the lack of temporal constraint, as shown in Fig. A5.

E Applications and Limitations

Applications Active listening faces ineffective communication and plays an important role in human-to-human interaction. People are encouraged to learn to actively listen to others in many scenarios, such as doctor - patient, teacher

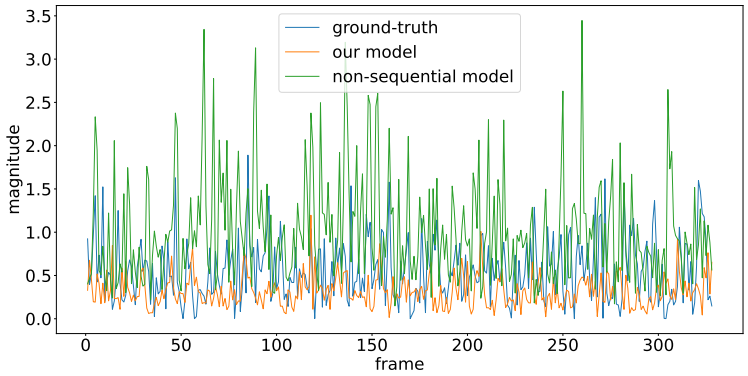

Fig. A5: The magnitude of variation for feature `angle` and `trans` between frame t and frame $t+1$ with L_1 distance

- student, salesperson - customer, *etc*. Our proposed responsive listening head generation task fills a gap in modeling face-to-face communication in the computer vision area. It can be applied to many scenarios, such as human-computer interaction, intelligence assistance, virtual human, *etc*. It would contribute to the mutual interaction between the virtual human and the real human. It can also be adopted to virtual audience modeling or providing guidance about how to act as an active listener.

Limitations One potential limitation of our constructed dataset is that the attitude is assumed to be consistent across the clips, since we cut and annotate short clips from the candidate video. However, as shown in the qualitative results, the generated head sequences can vary from one common head respecting different attitudes. We may deduce that our model is feasible to the attitude-transferable conversations.

References

1. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: IEEE Computer Vision and Pattern Recognition Workshops (2019)
2. Kington, R.S., Arnesen, S., Chou, W.Y.S., Curry, S.J., Lazer, D., Villarruel, A.M.: Identifying credible sources of health information in social media: Principles and attributes. NAM perspectives **2021** (2021)
3. Ramamoorthi, R., Hanrahan, P.: An efficient representation for irradiance environment maps. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques. pp. 497–500 (2001)
4. Ramamoorthi, R., Hanrahan, P.: A signal-processing framework for inverse rendering. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques. pp. 117–128 (2001)
5. Ren, Y., Li, G., Chen, Y., Li, T.H., Liu, S.: Pirenderer: Controllable portrait image generation via semantic neural rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13759–13768 (2021)