# CustomListener: Text-guided Responsive Interaction for User-friendly Listening Head Generation

Supplementary Material

## A. Implementation Details

### A.1. Network Details

We introduced Audio Encoder, Mapping Net, Motion Encoder and Transformer Decoder-based Diffusion Model in Section 3 of our main paper. Due to the page limitation, we show implementation details of these modules in this Appendix.

**Audio Encoder** For audio preprocessing, we first extract 45-dim acoustic features from speaker audio, and then encode these acoustic features into audio features. The structure of Audio Encoder is shown in Figure 8, which consists of two convolution layers, a GELU activation function, six 768-dim transformer layers, and a linear layer.

**Mapping Net** As discussed in Section 3.2 of the main paper, in order to obtain portrait-related tokens, we further employ a Mapping Net to convert text embeddings from RoBERTa [22] into static portrait tokens, which describes the listener's static portrait (e.g. expressions, head movements) in the entire video segment. The structure of Mapping Net is shown in Figure 8.

**Motion Encoder** We only use a single linear layer to encode the 70-dim noised listener motions to 768-dim motion tokens (denoted with small blue squares in Figure 2 of our main paper).

**Transformer Decoder-based Diffusion Model** We employ transformer decoder as our basic denoising module for two reasons: First, compared to U-Net [33] used in traditional diffusion model, transformer [38] can better capture temporal information in listener motion sequences. Furthermore, compared to transformer encoder which only contains self-attention, transformer decoder is capable of facilitating information interaction better between noised listener motion features as well as concatenated dynamic portrait token and motion prior. Specifically, we use an 8-layer transformer decoder, with 8 heads and 2048-dim fully forward layers. Additionally, we concatenate the input motion tokens with diffusion time-step token (denoted with small pink squares in Figure 2 of our main paper) to inject time-step information.

### A.2. Inference Details

**Long-video Inference** During inference, we generate long-term video in a segment-by-segment way. Specifically, we first sequentially clip the original long video to several 60-frame video segments, containing video frames, audios and speech content. We then input user-customized listener attributes. Subsequently, by incorporating speech content of each segment as well as pre-customized attribute description, text priors of each segment will be produced with the help of GPT. Finally, our proposed CustomListener will generate long-term listener motions clip-by-clip.

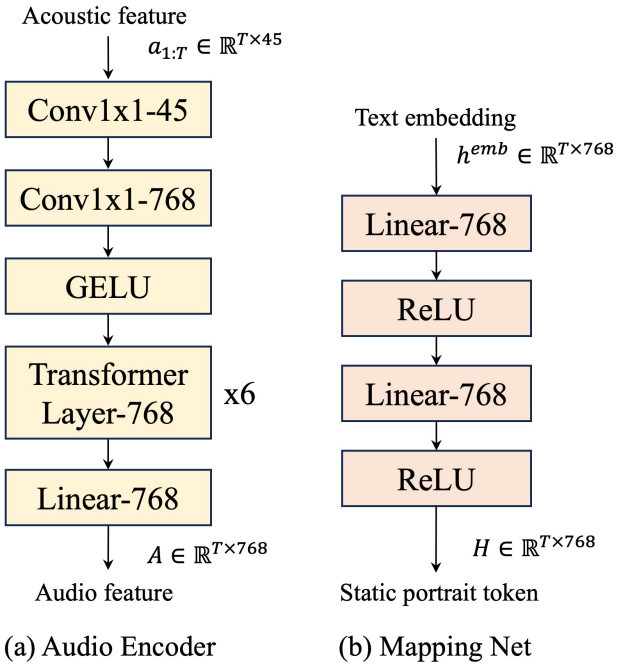


Figure 8. Stucture of Audio Encoder and Mapping Net.

**Text Prior Generation** To generate text prior in specific format like [A person <EMOTION> and listens with <AU> (and <HEAD MOTION>)]), inspired by [12], we use chain-of-thought prompting [40] when querying GPT. Specifically, before asking question, we first provide GPT with a few examples showing the desired answer format as well as reasoning process. This strategy can also enable GPT to provide more accurate text prior about facial expressions and head motions while reducing the misunderstanding probability towards the speaker's content. We present some examples of queries in Figure 9 and Figure 10.

After getting the text prior, we randomly replace the adjectives and adverbs to enhance the diversity of description. For example, for <EMOTION>, if it is 'happy', we replace it by randomly selecting one from a group <looks happy, seems joyful, ...>. For <AU>, following Facial Action

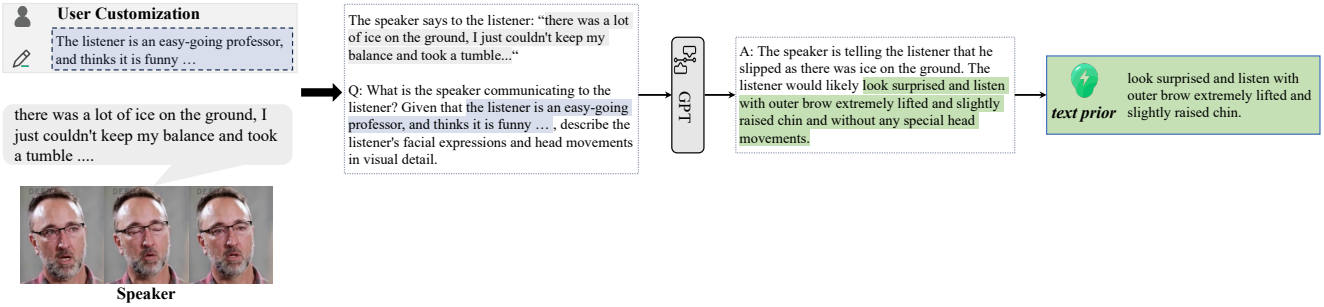


Figure 9. Details of Text Prior Generation.

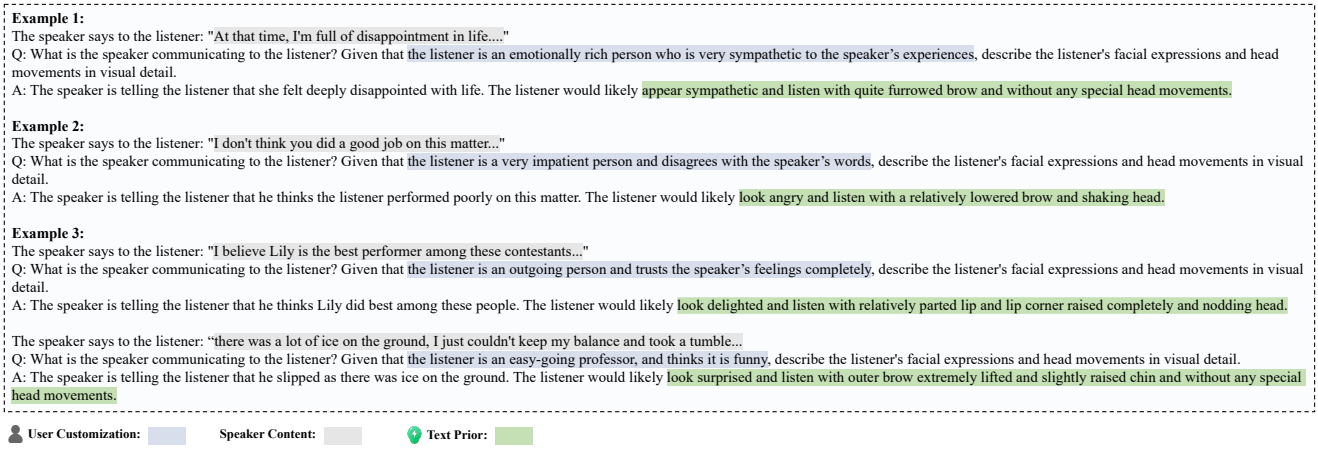


Figure 10. Illustration of Chain-of-thought Prompting.

Coding System, if AU12 (lip corner puller) is activated and level=3, we randomly choose an adj. from <raised, pulled, ...>, an adv. from <fully, extremely, ...>, then combine the above and get [ A person seems joyful and listens with fully raised lip corners.].

## B. Quantitative Results on RealTalk

We did not present quantitative results on RealTalk in main paper due to page limitation, thus we give these results in the Appendix B. For a fair comparison, we retrain PCH [15], RLHG [46] and L2L [25] on RealTalk dataset. Due to the unavailability of source codes, we cannot compare with MFR-Net [21] and ELP [35] on RealTalk. Notably, since PCH [15] and RLHG [46] need listener's attitude label as input, we use [6] to obtain emotion labels of each video clip. Then, we group these emotion labels into three basic attitudes: positive, negative and neutral. As shown in Table 5, we present the FD, RTLCC, RWTLCC, FID$_{\triangle fm}$, SND and V-D of our model as well as other methods. Our proposed CustomListener achieves best performance on RealTalk across all metrics, which demonstrates the superiority of CustomListener in generating realistic and coherent listener motions that highly synchronous with the speaker's motions. In Table 6, we also present evaluation results related to image-level quality on RealTalk. Our model achieves optimal results, which indicates CustomListener can generate relatively high-quality listener images without special design in the render module, thus justifies our generated motions are highly precise.

## C. Supplementary Visual Results

### C.1. Visual Comparisons with Other Methods

In Figure 11, we compared our results with PCH [15], RLHG [46] and ELP [35]. As there is no source code provided by ELP [35], for fairly comparison, we first find the speaker video showed in ELP [35], and then generate listener head based on the speaker video and corresponding description texts instead of an attitude label. Then, we displayed four generated listener frames, which are aligned with the four speaker frames in time. As shown in the visual comparisons, our method has two advantages. Firstly, without any special design in render model, our generated listener images showed less facial artifacts and more closed to the ground truth, which indicates the listener motions produced by CustomListener are highly aligned with the ground-truth motions. Secondly, without a blink modeling like ELP [35], our generated listener can perform blinking action, which

| Methods | FD ↓ | | | RTLCC ↓ | | RWTLCC↓ | | FID$_{\triangle fm}$↓ | | SND ↓ | | V-D ↑ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | exp | angle | trans | exp | pose | exp | pose | exp | pose | exp | pose | exp | pose |
| RLHG* [46] | 20.11 | 10.32 | 7.16 | 0.162 | 0.365 | 0.160 | 0.363 | 12.42 | 0.76 | 4.75 | 0.09 | 0.54 | 0.14 |
| PCH* [15] | 23.07 | 13.46 | 8.04 | 0.170 | 0.368 | 0.163 | 0.391 | 12.38 | 0.81 | 5.02 | 0.11 | 0.29 | 0.10 |
| L2L* [25] | 19.01 | 10.16 | 7.03 | 0.158 | 0.319 | 0.156 | 0.315 | 10.80 | 0.71 | 4.67 | 0.10 | 0.56 | 0.40 |
| Ours | **17.63** | **9.30** | **6.49** | **0.138** | **0.218** | **0.134** | **0.207** | **7.70** | **0.14** | **4.37** | **0.07** | **2.90** | **1.01** |

Table 5. Comparisons of our model with other methods on $\mathcal{D}_{test}$ of RealTalk. **Bold** represents the best. The * denotes we retrain the model. The ↓ indicates lower is better and the ↑ indicates higher is better. The values of FD and FID$_{\triangle fm}$ are multiplied by 100.

| Method | SSIM ↑ | CPBD ↑ | PSNR ↑ | FID ↓ |
|---|---|---|---|---|
| RLHG* [46] | 0.51 | 0.29 | 15.34 | 30.22 |
| PCH* [15] | 0.56 | 0.32 | **16.13** | 24.68 |
| L2L* [25] | 0.56 | 0.31 | 15.98 | 25.68 |
| Ours | **0.58** | **0.34** | 16.01 | **23.86** |

Table 6. Quantitative results with other methods on image quality. The * denotes we retrain the model. The best results are highlighted in bold.

demonstrates that the listener motions are natural-looking and photorealistic. More visual comparisons with PCH [15], RLHG [46] and L2L [25] on ViCo and RealTalk are presented in Figure 12 and Figure 13, respectively.

## C.2. Visual Results of Ablation

In Section 4.5 of our main paper, we presented visual ablation of PGM module. Due to the page limitation, we show the visual ablation of SDP module in this subsection. As shown in Figure 14, with the help of SDP module, our generated listener motions can not only realize progressive motion changes, but also fluctuate with the speaker's semantics, intonation and movement amplitude, and thus achieve the speaker-listener coordination.

## C.3. Supplementary Video

We provide a supplementary video to present more visual results, which include visual videos on ViCo and RealTalk, more visual comparisons with other methods as well as ablation study of each component. The video can be found in the "Supplementary_Videos.mp4" in the Supplementary Materials. These videos are also displayed on the project page.

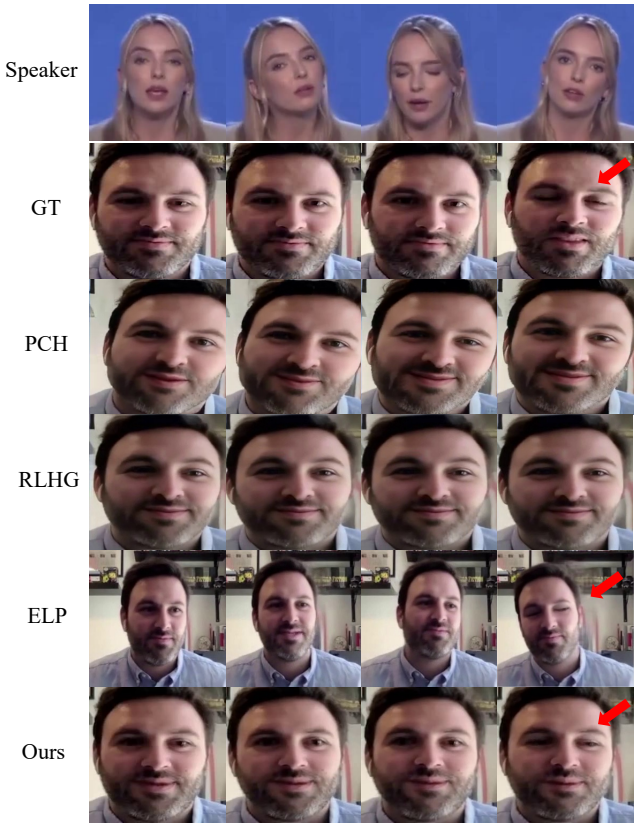


Figure 11. Qualitative comparisons with PCH [15], RLHG [46] and ELP [35].

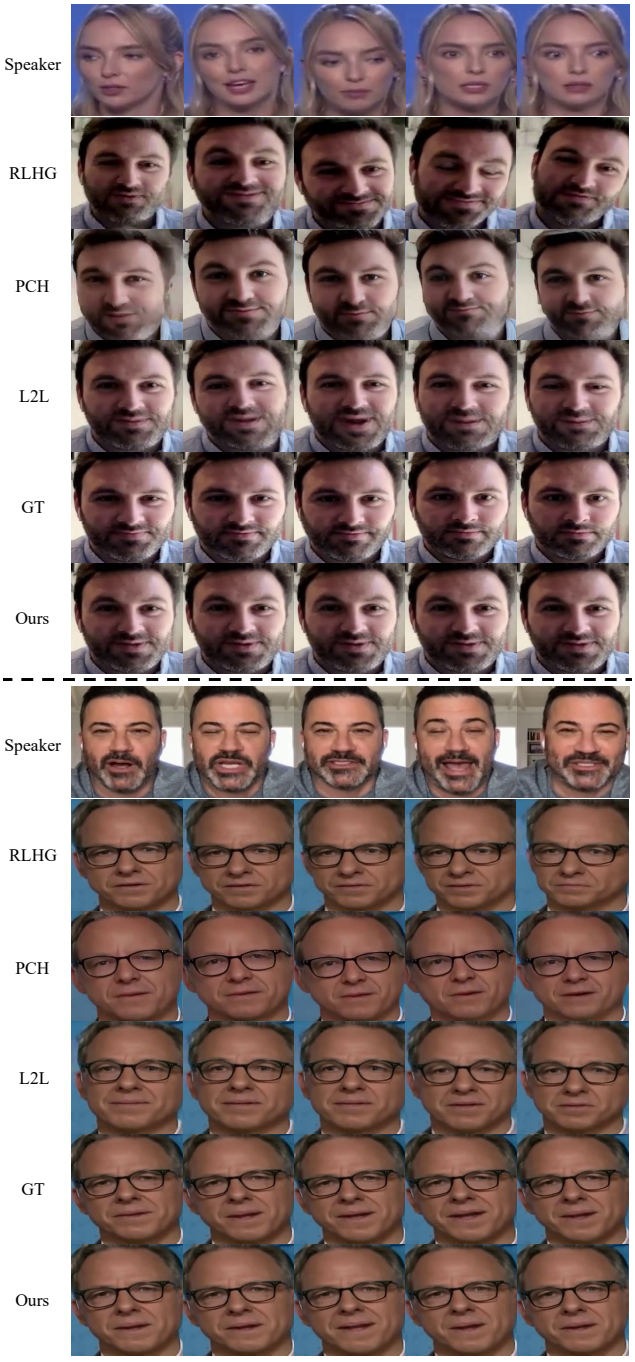


Figure 12. Qualitative comparisons with other methods on ViCo dataset.

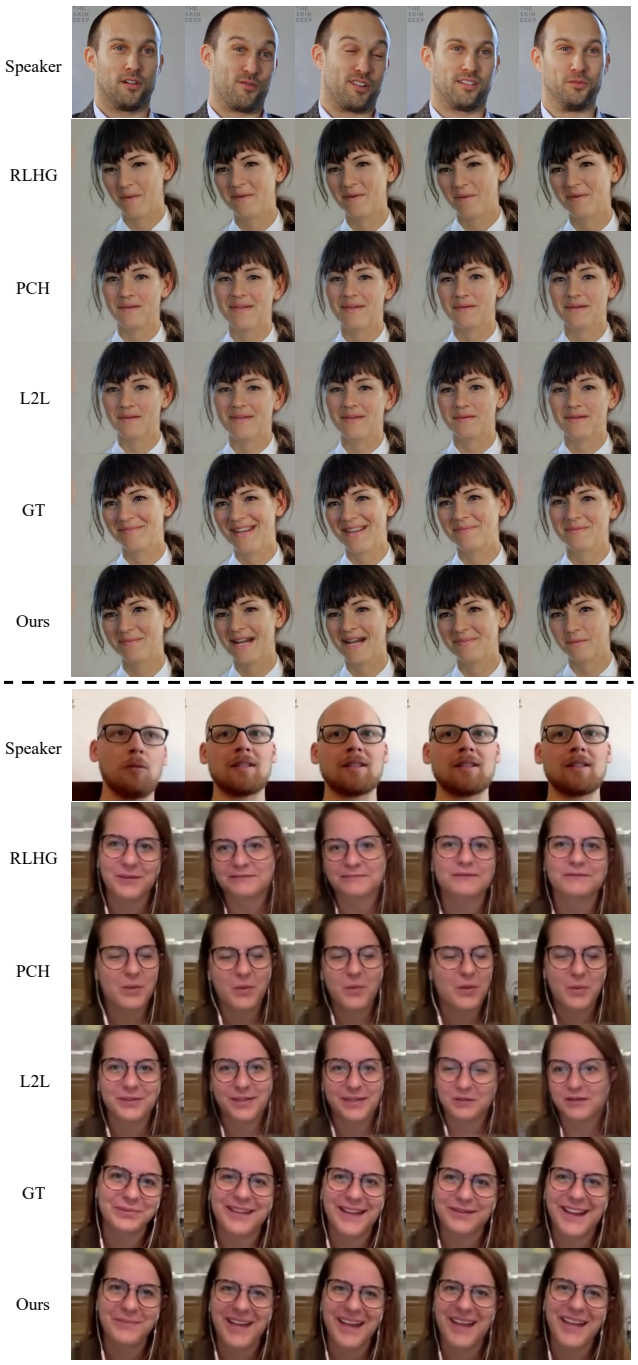


Figure 13. Qualitative comparisons with other methods on RealTalk dataset.


Figure 14. Ablation study of SDP module.