

# DiTaiListener: Controllable High Fidelity Listener Video Generation with Diffusion

Maksim Siniukov\* Di Chang\* Minh Tran  
 Hongkun Gong Ashutosh Chaudhary Mohammad Soleymani  
 University of Southern California  
 Los Angeles, USA  
[haven't-invented.github.io/DiTaiListener](https://haven't-invented.github.io/DiTaiListener)  
 siniukov@usc.edu

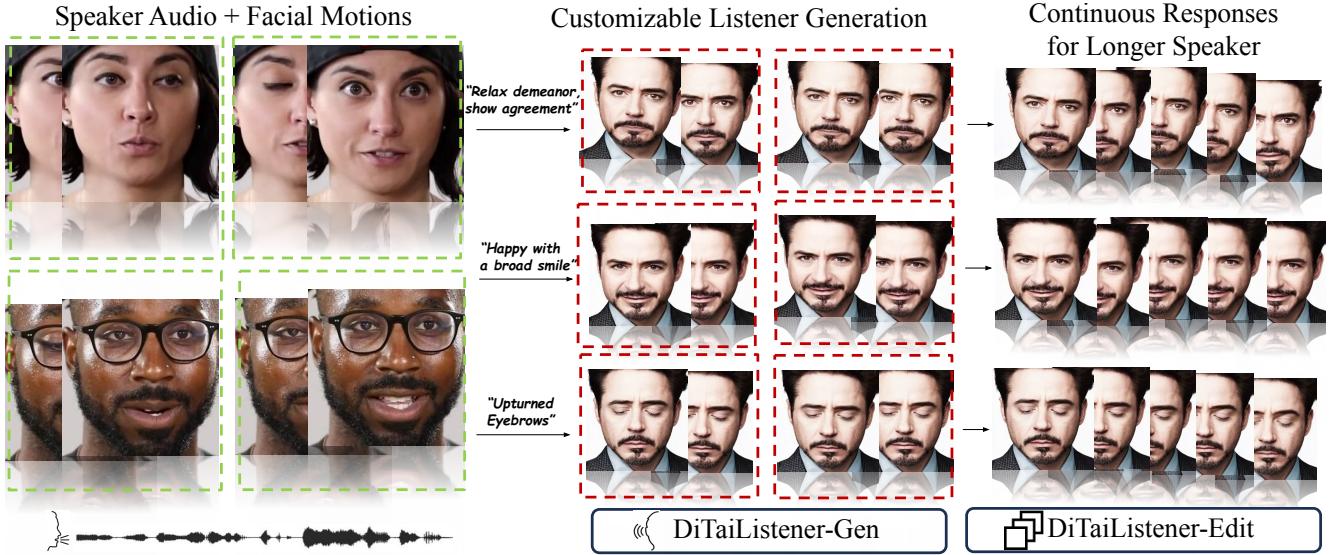


Figure 1. We introduce DiTaiListener, a DiT-based listener generation model that synthesizes high-fidelity human portrait videos of listener behaviors from speaker audio and facial motion inputs in an end-to-end manner. *DiTaiListener-Gen* generates customizable listener responses in short segments, while *DiTaiListener-Edit* ensures seamless transitions between segments, producing continuous and natural listener behaviors. Together, DiTaiListener supports user-friendly customizable listener behavior generation for variable speaker inputs.

## Abstract

Generating naturalistic and nuanced listener motions for extended interactions remains an open problem. Existing methods often rely on low-dimensional motion codes for facial behavior generation followed by photorealistic rendering, limiting both visual fidelity and expressive richness. To address these challenges, we introduce DiTaiListener, powered by a video diffusion model with multimodal conditions. Our approach first generates short segments of listener responses conditioned on the speaker’s speech and facial motions with DiTaiListener-Gen. It then refines the transitional frames via DiTaiListener-Edit for a seamless transition. Specifically, DiTaiListener-Gen adapts a Diffusion Transformer (DiT) for the task of listener head portrait generation by introducing

a Causal Temporal Multimodal Adapter (CTM-Adapter) to process speakers’ auditory and visual cues. CTM-Adapter integrates speakers’ input in a causal manner into the video generation process to ensure temporally coherent listener responses. For long-form video generation, we introduce DiTaiListener-Edit, a transition refinement video-to-video diffusion model. The model fuses video segments into smooth and continuous videos, ensuring temporal consistency in facial expressions and image quality when merging short video segments produced by DiTaiListener-Gen. Quantitatively, DiTaiListener achieves the state-of-the-art performance on benchmark datasets in both photorealism (+73.8% in FID on RealTalk) and motion representation (+6.1% in FD metric on VICO) spaces. User studies confirm the superior performance of DiTaiListener, with the model being the clear preference in terms of feedback, diversity, and smoothness,

\*Equally contributed as first authors

*outperforming competitors by a significant margin.*

## 1. Introduction

Human behavior synthesis is an essential building block for socially intelligent systems, with broad applications from health to entertainment. Hence, audio-driven head motion generation has received considerable attention due to its potential for generating engaging content [21, 36, 39, 40, 48, 53, 55, 59, 66, 77, 80]. A grand majority of work has focused on speaker head generation [8, 35, 41, 51, 53, 58, 65, 66, 70, 73, 75, 77, 81] synthesizing speakers’ facial motions based on speech input, producing vivid lip synchronization yet neglecting listeners’ nonverbal feedback or conversational context. Coordinated and contextual listener behaviors are essential to building a sense of rapport, a crucial element of human interpersonal communication with an impact on interaction quality and trust [17]. Hence, Listening-head generation [32, 36, 39, 40, 48, 49, 55, 76, 79, 82] focuses on coordinated listeners’ responses to speaker behaviors, yet typically restricts them to limited reactive facial movements in a limited temporal context, i.e., responding to immediate speaker behaviors with limited motions. While some recent efforts [36, 55, 59, 80, 82] have begun exploring the dyadic context in generating reactions, most continue to rely on motion tokenization rather than generating high-fidelity portraits with nuanced expressions and facial details.

We propose DiTaiListener, an end-to-end video diffusion-based listener head generation framework. Through extensive experiments, we study the potential of video generation models in learning nuanced and coordinated facial motions that are well-timed, detailed, and controllable. While prior methods generate listener behaviors in a compact latent space, e.g., 3D Morphable Models (3DMMs) or other motion codes, and then render the face, inevitably losing detail and expressiveness, we directly predict the listener’s behavior in a higher-fidelity video space through diffusion, offering more varied and fine-grained outputs. Large Video Diffusion models capture emergent, data-driven behaviors, producing more dynamic and lifelike motions, including natural blinking patterns and subtle facial expressions. Our method also supports flexible controls from users; that is, once emotion guidance or response context is specified through a text prompt, the model can generate listener reactions accordingly.

Reactivity in human interactions is not strictly synchronous; instead, listeners react to events in the temporal context with uncertain lag. For example, smile mimicry occurs almost immediately, whereas a head nod is usually delayed, aligning more with speaker pauses as a non-rigid response. To model this, we use a temporal attention mechanism that learns to adaptively time the reactions. This enables DiTaiListener to learn better response timing. Our design incorporates a causal temporal adapter, which aligns

preceding speaker cues to predicted listener motions, improving temporal coherence, in addition to long-term speaker audiovisual behavior encoders that gather context from up to 10 seconds of speaker video. This architecture allows for the generation of extended video sequences, including those lasting minutes or more, via DiTaiListener-Edit, a temporal frame-smoothing approach. Experimental evaluations show that DiTaiListener achieves state-of-the-art results in listener behavior synthesis according to both quantitative measures and human evaluations.

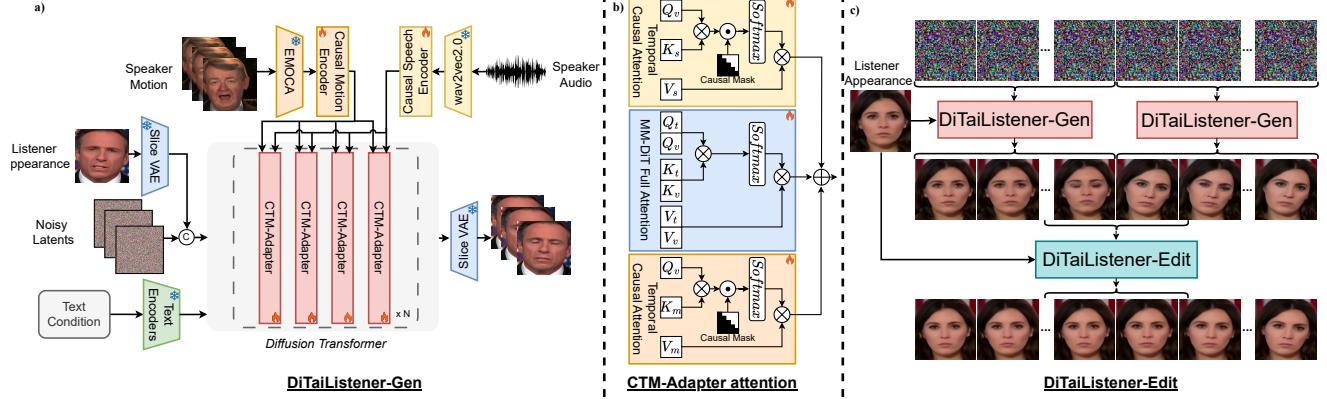
Our main contributions can be summarized as follows.

- We propose an end-to-end listener generation model that can synthesize listener motions in pixel space and demonstrate the potential of video generation methods for socially intelligent, coordinated, and controlled face and head gestures.
- We introduce a causal temporal multimodal adapter (CTM-Adapter) that aligns listener motions with preceding audiovisual speaker behaviors. This improves the temporal consistency between speaker cues and the listener’s responses.
- Our model provides free-form text control to guide the listener’s behaviors. Given emotional guidance and/or dialog context, the model can produce listener reactions according to the provided guidance.
- We develop a long-sequence video generation approach through frame smoothing with DiTaiListener-Edit for variable-length listener videos.

## 2. Related Work

### 2.1. Diffusion-based Portrait Generation

Recent advances in Diffusion Probabilistic Models (DMs) [25, 47, 50] have exhibited extraordinary performance in a variety of generative tasks, such as image synthesis [46], video generation [3], and multi-view rendering [20, 33, 34]. Latent diffusion models (LDMs) [44] build on these successes by conducting the diffusion process within a lower-dimensional latent space, thereby substantially reducing computational overhead. In the realm of portrait animation, pre-trained diffusion models [44, 46] have served as a strong foundation for image-to-video (I2V) generation. A number of previous works [5, 31, 71] have highlighted the efficacy of infusing reference image features into the self-attention layers of LDM UNets, enabling both image editing and video generation while preserving appearance consistency. Additionally, ControlNet [72] expands LDM-based generation by incorporating structural inputs—such as landmarks, segmentations, and dense poses—into the model. Leveraging these strategies, recent research [6, 27, 67] has demonstrated state-of-the-art full-body animation by integrating appearance features, motion control modules, and temporal attention mechanisms [22, 23] within ReferenceNet



**Figure 2. Overview of DiTaiListener.** a) Given the listener’s appearance (reference frame), speaker’s motion, encoded via EMOCA 3DMM coefficients, speech (Wav2Vec2) and an input text control, DiTaiListener learns to generate listener face and head motions in pixel space through a video diffusion model powered by a modified DiT. b) We introduce a Causal Temporal Multimodal Adapter for seamless integration of multimodal speaker input in a temporally causal manner. c) Our long video generation pipeline consists of two video generation models. DiTaiListener-Gen generates video blocks that are fused by the DiTaiListener-Edit model that facilitates the smooth transition between two blocks, improving smoothness and reducing computational cost compared to existing long video generation strategies, e.g., prompt traveling and teacher forcing.

architectures. Subsequent works [10, 37, 61, 65] have extended portrait animation to accommodate a range of facial motion representations and speech signals. Nevertheless, these methods currently remain confined to one-way portrait animation and cannot yet produce convincing two-way conversational interactions between speakers and listeners.

## 2.2. Dyadic Behavior Generation

Unlike human portrait generation, e.g., talking heads, listener behavior generation in dyadic settings focuses on generating a listener’s response to the speaker’s verbal and nonverbal behaviors. Early work relied on rule-based methods and simple machine learning models to generate discrete behaviors for virtual humans, e.g., a head nod or smile [18], and heavily relied on the speaker’s speech. Huang et al. [28] extended the rule-based listener response generation with a conditional random field (CRF) model to generate different kinds of head nods. Bohus and Horvitz [4] studied the effect of facial nonverbal behavior generation for facilitating multiparty turn-taking. Greenwood et al. [19] studied speech-driven synchronized agent motion generation in dyadic settings. Ahuja et al. [1] focus on producing non-verbal full-body motions in dyadic interactions by integrating monadic (within-person) and dyadic (between-person) behaviors to predict socially appropriate body poses. Other techniques aim to generate realistic listener facial movements based on language [9] or speech signals [29, 30]. For example, Song et al. [48] proposed Emotional Listener Portrait (ELP) that learns to generate nonverbal listener motions that can be modulated by emotions. Geng et al. [16] introduced the RealTalk database and proposed a method to retrieve plau-

sible listener expressions by leveraging a large language model to condition motions on listeners’ goals, personalities or backgrounds. Ng et al. [39] proposed Learning2Listen (L2L) that relies on learning a dictionary of listener motions through VQ-VAE [57] and generating listener motions based on the speaker’s motion and speech in an autoregressive manner. Tran et al. [55] proposed Dyadic-Interaction-Modeling (DIM) that employs a similar approach to predict discrete listener head motions but leverages self-supervised pre-training with a large dataset to improve representations and uses PIRenderer [43] for photorealistic synthesis. More recently, CustomListener [36] proposed a text-guided generation strategy that encodes user-defined textual attributes into portrait tokens containing time-dependent cues for speaker-listener coordination and then feeds these tokens into diffusion-based motion priors for controllable video generation. Zhu et al. [82] introduce INFP that takes a two-stage audio-driven approach, learning a low-dimensional motion latent space and mapping dual-track audio to these latent codes via a diffusion-based transformer. Although these diffusion-based methods enable responsive and natural dyadic interactions, they predict the motion latent codes that need to be rendered through a warping mechanism, posing challenges for directly generating highly diverse and detailed results in video space. Unlike previous work, DiTaiListener learns to generate listener motions directly in the pixel space, providing a more realistic and detailed appearance and motion.

## 3. Method

We provide an overview of DiTaiListener in Figure 2. Given the speaker’s audio and facial motions and the listener’s ref-

erence frame (identity), our model synthesizes high-fidelity, photorealistic images of the listener’s expected facial response. Unlike most prior work that first predicts 3DMM features for the listener and then applies photorealistic rendering, DiTaiListener introduces an end-to-end solution that directly synthesizes high-quality listener facial images via a standard diffusion denoising process with a novel diffusion transformer architecture built upon DiT. Our method not only enhances realism but also enables text-based control for customizing listener responses, offering greater flexibility and expressiveness in facial reaction generation. At the core of our approach is the Causal Temporal Multimodal Adapter (CTM-Adapter), a module integrated into the DiT architecture to process multimodal inputs—speaker audio and facial motion features—in a temporally causal manner while enabling text-based control for customized listener responses. We further enhance DiTaiListener by introducing seamless transitions between independently synthesized video segments, allowing for the generation of long (unbounded), coherent listener response videos, a limitation of prior video generation methods. In the following sections, we provide details for each component of DiTaiListener.

### 3.1. Visual and behavioral Descriptors

**Speaker Feature Representations.** For speaker motions, we follow prior work [38, 40, 55] and utilize EMOCA [11] to extract facial representations at a 12 FPS frame rate. For speaker audio, we employ a pre-trained Wav2Vec2-base [2] model to capture rich speech representations. Specifically, we leverage the intermediate features from all 12 layers of Wav2Vec2 to preserve detailed acoustic information. To ensure temporal alignment, we resample the extracted speech representations to match the frame rate of the speaker motion features. The features are passed to transformer-based causal motion encoders and causal speech encoders to extract relative information. We denote the resulting speaker speech representation as  $X_s$ , and the speaker facial motion features as  $X_m$ .

**Listener identity.** We utilize features extracted from Slice-VAE [63] to encode the reference listener image into the diffusion latent space. This approach is commonly used for spatial conditioning in image and video generation through depth maps and edge maps [64]. We extend this approach to provide appearance guidance. We denote the extracted identity representation as  $X_i$ .

**Customized text prompts.** DiTaiListener optionally accepts text prompts to enable controllable listener behavior generation. When a text prompt is provided, we encode it using the T5 [42] and BERT [12] text encoders, allowing for flexible customization of listener responses. We denote the extracted text representation as  $X_t$ .

### 3.2. DiTaiListener

We propose a novel diffusion transformer adapter that seamlessly integrates speaker speech and motion in a temporally causal manner. At the core of DiTaiListener are the CTM-Adapter layers, with our model structured as a hierarchical stack of these modules to effectively capture multimodal dependencies and generate coherent listener responses.

**CTM-Adapter Blocks.** Inspired by IP-Adapter [68] for UNet-based diffusion models [45], we propose CTM adapter – a video diffusion adapter that extends adapters from conventional spatial attention to multimodal temporal attention for better temporally-aligned guidance in a Multimodal Video DiT. Built upon the MM-DiT Full Attention blocks from EasyAnimate [63], our CTM-Adapter blocks introduce two additional Temporal Causal Attention (TCA) modules to process temporally fused speech and motion features. The outputs of the standard MM-DiT Full Attention module are then combined with the speech-guided and motion-guided cross-attention features produced by our TCA modules using a weighted summation, effectively capturing the temporal dependencies between multimodal inputs. Following L2L [39], we apply causal masks to our TCA modules so that current listener motions are only generated with respect to past speaker motions and audio.

Formally, the original MM-DiT Full Attention block takes visual and textual tokens as input, denoted as  $X_v$  and  $X_t$ , respectively. In our case,  $X_v$  represents the generated listener video frames. For appearance control, the first CTM-Adapter block is modified to accept concatenated inputs  $[X_v, X_i]$ , where  $X_v$  is initialized with random noise. The text-video tokens are then concatenated and passed through a multi-layer perceptron (MLP), after which they are split back into text tokens and denoised video tokens. These tokens are subsequently fed into the next layer for further processing. The attention mechanism within the MM-DiT blocks is given by

$$\text{Attn}_{\text{MMDiT}}([X_t X_v]) = \sigma \left( \frac{[Q_t Q_v] [K_t K_v]^\top}{\sqrt{d}} \right) \times [V_t V_v] \quad (1)$$

where  $\sigma$  is a Softmax function,  $[Q_t Q_v]$ ,  $[K_t K_v]$ ,  $[V_t V_v]$  are the concatenation of query, key, and value projections from text  $X_t$  and video  $X_v$  tokens respectively, and  $d$  is the projection dimension.

To adapt MM-DiT for the task of listener behavior generation with additional speech and motion inputs, we leverage the Decoupled Multimodal Attention mechanism [14] to integrate speech-guided and motion-guided information into the attention modules of MM-DiT. Specifically,

$$\text{Attn}_{\text{speech}}([X_t X_v], X_s) = \sigma \left( \frac{M \circ [Q_t Q_v] K_s^\top}{\sqrt{d}} \right) V_s \quad (2)$$

$$\text{Attn}_{\text{motion}}([X_t X_v], X_m) = \sigma \left( \frac{M \circ [Q_t Q_v] K_m^\top}{\sqrt{d}} \right) V_m \quad (3)$$

where  $[K_s, K_m]$  and  $[V_s, V_m]$  are the key and value projections of  $X_s$  and  $X_m$  respectively, and  $M$  correspond to the causal attention masks. Attention is computed along the temporal dimension, while the spatial dimension is combined with the batch dimension. The outputs of the three attention branches are combined via an element-wise summation

$$\begin{aligned} \text{Attn}_{\text{CTM-adapter}}([X_t X_v]) &= \alpha \times \text{Attn}_{\text{speech}}([X_t X_v], X_s) \\ &\quad + \beta \times \text{Attn}_{\text{MMDiT}}([X_t X_v]) \\ &\quad + \gamma \times \text{Attn}_{\text{motion}}([X_t X_v], X_m) \end{aligned} \quad (4)$$

where  $\alpha, \beta, \gamma$  are scaling parameters that control the contribution of each branch during inference.

Overall, DiTaiListener-Gen consists of a stack of  $L$  CTM-Adapter Blocks, trained using the standard diffusion loss similar to [63].

**Causal Long Sequence Generation.** DiTaiListener-Gen is designed to generate segments of  $K$  frames at a time. When synthesizing listener behavior for input sequences longer than  $K$  frames, we partition the inputs into overlapping segments of  $K$  frames and generate each segment independently. However, directly merging these segments often results in abrupt discontinuities in pose and expression due to the lack of temporal consistency across segment boundaries. To address this, we introduce *DiTaiListener-Edit* to produce smooth transitions between consecutive segments.

DiTaiListener-Edit operates by selecting two consecutive segments of length  $K$  generated by DiTaiListener and extracting the last  $K'$  ( $K' \ll K$ ) frames from the first segment and the first  $K'$  frames from the second segment. While these segments should ideally form a seamless motion sequence, independent generation can introduce unnatural transitions. To refine these transitions, we pass the first and the last of concatenated  $2K'$  frames through a standard DiT model, with the objective of reconstructing the corresponding ground-truth frames for these  $2K'$  timesteps. The model is trained as a standard diffusion model, learning to synthesize temporally smooth and coherent frames that effectively bridge discontinuities between independently generated segments.

**Customized text control.** Existing dyadic behavior learning datasets lack annotated descriptions of listener behavior, making them unsuitable for text-guided listener behavior generation. To bridge this gap, we leverage Google Gemini 1.5 [52] to extract affective-related descriptions of listener behavior from existing datasets. These generated textual descriptions are then used to train DiTaiListener, enabling it to learn text-aware behavior generation. During inference,

users can provide custom textual descriptions to guide the model in producing personalized listener responses. Details of our text prompts are provided in the supplementary materials.

Method	Feedback↑	Diversity↑	Smoothness↑	Overall↑
ELP [48]	2.93%	4.04%	5.65%	4.21%
RLHG [79]	10.45%	12.14%	13.00%	11.86%
L2L [38]	4.20%	6.06%	10.95%	7.07%
DIM [55]	10.25%	12.12%	16.11%	12.83%
Ours	<b>72.17%</b>	<b>65.64%</b>	<b>54.29%</b>	<b>64.03%</b>

Table 1. User Study. We ask the participants to choose the best video among all methods for each criterion. The average preference percentage for each method output is provided.

Method	FID↓	FVD↓	LPIPS↓	PSNR↑	SSIM↑
RLFG [38]	52.01	287.48	0.37	17.95	0.61
L2L [38]	56.93	285.24	0.40	18.01	0.64
ELP [48]	67.12	281.95	0.38	17.96	0.61
DIM [55]	49.75	288.27	0.36	18.51	0.62
w/o Text-Control	10.56	57.71	<b>0.28</b>	20.58	0.71
w/o CTM-Adapter	11.56	54.79	<b>0.28</b>	20.77	<b>0.72</b>
DiTaiListener	<b>10.14</b>	<b>53.54</b>	<b>0.28</b>	<b>20.77</b>	<b>0.71</b>

Table 2. Quantitative comparison on ViCo [69] test set. The metrics are reported in photorealistic video space.

Method	FID↓	FVD↓	LPIPS↓	PSNR↑	SSIM↑
RLHG [69]	30.55	217.48	0.41	17.23	0.53
L2L [38]	56.43	522.67	0.46	16.49	0.53
DIM [55]	30.75	204.49	0.40	17.74	0.54
w/o Text-Control	8.05	49.88	0.27	18.70	0.59
w/o CTM-Adapter	8.16	45.62	0.27	18.60	0.55
DiTaiListener	<b>7.99</b>	<b>45.42</b>	<b>0.26</b>	<b>18.73</b>	<b>0.60</b>

Table 3. Quantitative comparison on RealTalk [16] test set. The metrics are reported in photorealistic video space.

## 4. Experiments

### 4.1. Implementation Details

**Datasets** We evaluate DiTaiListener on RealTalk [16] and ViCo [78] datasets. RealTalk [16] is a large dataset containing 692 dyadic conversations with a total duration of 115 hours. To allow comparison with existing methods, we also evaluate our method on ViCo [79], which is a smaller dataset consisting of 483 video sequences featuring 76 unique listeners. We present the details of dataset pre-processing in the supplementary material.

**Model Training and Inference** We used a pre-trained DiT-based video diffusion model as the backbone, with initial-



Figure 3. **Qualitative Comparison on ViCo test set.** Our method generates high-quality, photorealistic facial images with diverse and natural social behaviors, including head movements and blinks, whereas baseline methods often produce less varied and expressive responses.



Figure 4. **Listener generation from DiTaiListener on out-of-domain identities.** Our method can integrate expressions from text conditions and synthesize diverse responses to the speakers.

ization weights from EasyAnimateV5.1-12b [63]. Causal Motion and Speech Encoders are transformers trained from scratch sharing the same architecture with 6 attention blocks, 8 heads, 512 hidden dimensions. For CTM-Adapter we set  $\gamma$ ,  $\beta$ , and  $\alpha$  parameters to be 1.0. We trained our model on RealTalk [16] with 8 NVIDIA H100 GPUs for 20k iterations with a batch size of 8. The training videos were divided into 4-second clips consisting of 144 frames, sampled at a stride of 3 frames. Each frame was resized to  $256 \times 256$  pixels. We used the Adam optimizer with a learning rate of  $2 \times 10^{-5}$ , a weight decay of  $5 \times 10^{-3}$ , and gradient clipping of 0.05. The entire model was trained end-to-end, with text control sequences truncated at 512 tokens per video. Text is encoded and passed to the model via T5 [42] and BERT [12] encoder. For DiTaiListener-Edit we set  $K$  to 49,  $K'$  to 6. To compare with baselines on ViCo [69], we train our DiTaiListener for 5k iterations with the same hyperparameters.

## 4.2. Evaluations and Comparisons

**Baselines** We tested four listener behavior generation methods to benchmark our model against, including:

- **RLHG** [79] is an RNN-based listener prediction model.
- **L2L** [38] uses quantized VQ-VAE to predict listener motions in an autoregressive manner.
- **ELP** [48] is a model that relies on emotion vectors and speaker features to predict listener behavior.
- **DIM** [55] is a recent method that combines two-branch speaker-listener VQ-VAE with contrastive pretraining.
- **INFP** [82] is a recent diffusion-based model which learns a motion latent space trained on in-house data. Currently, the code and data are not open-sourced.

Note that these methods are all multi-stage models, which predict discrete listener motion representation first, then map these predictions into photorealistic videos via PIRenderer [43] or other similar rendering methods [13].

**Evaluation metrics.** Since our method is an end-to-end model that generates photorealistic videos directly, we first evaluate its performance against baseline methods using image-space metrics: Peak Signal-to-Noise Ratio (PSNR), Learned Perceptual Image Patch Similarity (LPIPS) [74], SSIM [60], and Fréchet Inception Distance (FID). PSNR [26] is a pixel-wise image similarity metric. LPIPS and SSIM are perceptual image-based measures that evaluate the similarity between predictions and ground-truth frames. Fréchet Inception Distance [24] is used to measure how different the distributions of the generated frames are from real data. Fréchet Video Distance [56] compares distributions of generated videos with the distribution of ground-truth videos, including temporal information. Following previous methods [38, 55], we also compare the results in the 3DMM space by extracting EMOCA [11] parameters from our generated frames. We use Frechet Distance (FD), Mean Squared Error (MSE), Paired FD (P-FD), Variance (Var), and SI Diversity (SID) as evaluation metrics. Details of these metrics are available in the supplementary material. Note that Var and SID are metrics to evaluate the diversity of data distribution, so we are able to report these for Ground Truth EMOCA parameters as well. FD and P-FD are the primary metrics to evaluate the listener generation quality.

**Quantitative Comparison** We compare our method to the state-of-the-art listener video generation baselines [38, 48, 55, 79, 82]. This work focuses on improving the quality of motions and the photorealism of the generated video. Table 2 and Table 3 present the quantitative analysis of such quality on ViCo and RealTalk datasets. DiTaiListener achieves significant improvements across different baseline models, indicating that the proposed method generates vivid expressiveness of listener head videos. The baseline methods heavily rely on the renderer and are not able to directly generate high-fidelity human portraits. Since previous work [38, 48, 55, 79] widely adopted 3DMM [11, 15] to represent pose and facial expression and provided evaluation in such parameter space, we extract EMOCA [11] parameters from our generated photorealistic videos and presents a quantitative analysis in Tables 4 and 5. Even though our model is not trained to generate 3DMM parameters, it achieves competitive performance with the existing methods, which indicates that DiTaiListener generates high-quality videos with diverse listener motions that provide correct feedback to the speakers.

**Qualitative Comparison** We qualitatively compare the generated listener video from DiTaiListener with previous methods [38, 48, 55, 79] on ViCo in Figure 3. Note that these works used 3DMMs as predictions from their models, we render the photorealistic results by PIRenderer [43] follow-

ing the setting mentioned in [38, 55]. All baselines exhibit limited expressiveness, with most of their generated listener appearing almost static. Please refer to additional video examples provided in the supplementary materials for clearer observations and further comparison. We also provide more visualizations of text-control from DiTaiListener on out-of-domain data in Figure 4. Our method demonstrates user-friendly customization of listener emotions through natural language and strong generalizability to other unseen identities. We provide more video visualizations in the supplementary offline page.

**User Study** We asked 50 participants from **Prolific**, a crowdsourcing platform, to evaluate the quality of the listeners generated with the ViCo test set [79] using L2L [39], RLHG [79], ELP [48], DIM [55] and DiTaiListener. The methods were anonymized and randomly ordered for each question. The average duration of the videos is 14 seconds. For each group of video comparisons, we ask the user to choose the best method according to the following criteria for judgment: 1) The response from the listener provides correct feedback to the speaker. 2) The head motions and facial expressions are diverse. 3) The general video quality is flawless and smooth. The user study demonstrated that most users selected our model output as superior across all scales (see Table 1). We present an example from our user study survey in Figure. 6 of the supplementary material.

### 4.3. Ablation Analysis

In this section, a comprehensive ablation analysis of DiTaiListener is presented. We evaluate the effectiveness of our proposed Causal Multimodal Temporal Adapter and text control on ViCo and RealTalk test sets in Tables 2, 3, 4 and 5. Note that in all these tables, each row represents the individual ablation. **w/o Text-Control** means an empty string replaces the input to text encoders, and **w/o CTM-Adapter** denotes the Causal Temporal Attention is replaced by simple non-causal spatial attention without masks. To further confirm the effectiveness of text control, we show videos of different listener emotions on the supplementary offline page. We also visualize the comparison between our proposed Causal Long Sequence Generation strategy, frame-smoothing **DiTaiListener-Edit**, with widely used prompt travel and teacher forcing [23, 62] in Figure. 5 and the supplementary offline page. Note that **w/o DiTaiListener-Edit** in both figures and offline webpage denotes the video segments are directly concatenated in time dimension without smooth processing with DiTaiListener-Edit. **Teacher Forcing** [62] produces longer videos recurrently relying on its own predictions by replacing the appearance frame with the last generated frame. Teacher forcing suffers from error accumulation and gradual information loss, leading to quality degradation for long sequences. **Prompt Travel** relies on

Method	FD↓		P-FD↓		MSE↓		SID↑		Var↑	
	Exp	Pose	Exp	Pose	Exp	Pose	Exp	Pose	Exp	Pose
ELP [48]	47.17	0.08	47.48	0.08	0.98	<u>0.02</u>	1.76	1.66	1.49	<u>0.02</u>
RLHG [79]	39.02	0.07	40.18	0.07	0.86	<b>0.01</b>	3.62	3.17	1.52	<u>0.02</u>
L2L [39]	33.93	0.06	35.88	<u>0.06</u>	0.93	<b>0.01</b>	2.77	2.66	0.83	<u>0.02</u>
DIM [55]	23.88	0.06	24.39	<u>0.06</u>	0.70	<b>0.01</b>	3.71	2.35	1.53	<u>0.02</u>
INFPT [82]	<u>18.63</u>	0.07	-	-	<b>0.51</b>	<b>0.01</b>	<u>4.78</u>	<b>3.92</b>	<b>2.83</b>	<b>0.18</b>
GT	-	-	-	-	-	-	5.03	4.07	0.93	0.01
w/o Text-Control	18.88	<u>0.05</u>	<u>21.64</u>	<u>0.06</u>	0.85	<u>0.02</u>	<b>4.81</b>	3.59	1.43	<u>0.02</u>
w/o CTM-Adapter	19.54	<u>0.05</u>	22.27	0.06	0.88	<u>0.02</u>	4.68	3.71	1.46	<u>0.02</u>
DiTaiListener	<b>17.49</b>	<b>0.04</b>	<b>20.53</b>	<b>0.05</b>	0.85	<u>0.02</u>	4.73	<u>3.74</u>	1.51	<u>0.02</u>

Table 4. Quantitative comparison on ViCo [79] test set in the 3DMMs (EMOCA [11]) space. † denotes the method did not release code or in-house training data, and the numbers are directly taken from their paper.

Method	FD↓		P-FD↓		MSE↓		SID↑		Var↑	
	Exp	Pose	Exp	Pose	Exp	Pose	Exp	Pose	Exp	Pose
RLHG [79]	69.04	<u>0.05</u>	69.09	<u>0.06</u>	1.37	<b>0.01</b>	0.35	3.23	0.14	<b>0.01</b>
L2L [39]	72.89	0.10	72.94	0.10	1.44	<u>0.02</u>	0.10	2.42	0.07	<b>0.01</b>
DIM [55]	77.97	0.15	78.70	0.15	1.52	<u>0.02</u>	3.49	3.29	0.74	<b>0.01</b>
GT	-	-	-	-	-	-	5.13	3.95	1.36	0.02
w/o Text-Control	15.62	<b>0.02</b>	16.35	<b>0.02</b>	0.67	<b>0.01</b>	5.01	3.94	1.30	<b>0.01</b>
w/o CTM-Adapter	15.10	<b>0.02</b>	<u>15.82</u>	<b>0.02</b>	0.66	<b>0.01</b>	<b>5.11</b>	<b>3.98</b>	1.30	<b>0.01</b>
DiTaiListener	<b>14.28</b>	<b>0.02</b>	<b>15.07</b>	<b>0.02</b>	<b>0.65</b>	<b>0.01</b>	<u>5.09</u>	<u>3.95</u>	<b>1.31</b>	<b>0.01</b>

Table 5. Quantitative comparison on RealTalk [16] test set in the 3DMMs (EMOCA [11]) space.

a sliding window approach for long video generation. The method starts by generating long noise patterns. During each denoising step, a sliding window is taken over overlapping chunks of video; denoising predictions are made for each chunk independently. Predictions from overlapping chunks are averaged and passed to the next denoising step. Prompt Travel requires additional resources proportional to overlap length. While effective against error accumulation, it suffers from blurring artifacts due to averaging and transition artifacts, as the model is not trained on inputs produced via averaging.

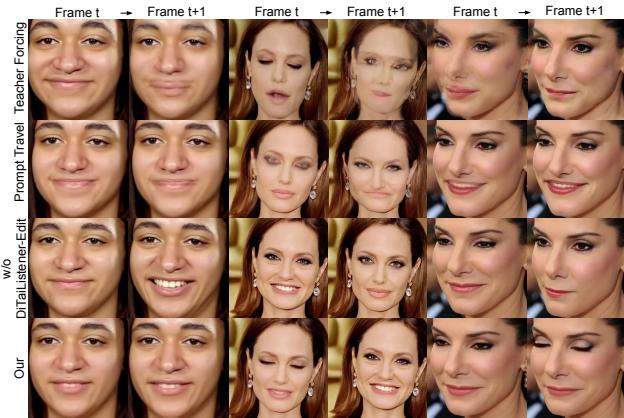


Figure 5. Qualitative comparison of long video generation. Our method generates smoother videos with fewer transition artifacts compared to prompt traveling and teacher forcing methods.

## 5. Conclusion

We introduce DiTaiListener, a DiT-based framework for generating high-fidelity listener response videos from speaker audio and facial motion inputs. Unlike prior methods that rely on intermediate 3DMM representations, DiTaiListener directly synthesizes photorealistic listener portraits in an end-to-end manner while enabling text-controlled customization of listener behaviors. At its core, the Causal Temporal Multimodal Adapter (CTM-Adapter) effectively integrates multimodal inputs in a temporally causal manner. To ensure smooth transitions between generated video segments, we proposed *DiTaiListener-Edit*, which refines segment boundaries for continuous and natural listener behaviors over long sequences. Our model demonstrates state-of-the-art performance in producing expressive, semantically aligned listener responses with high visual fidelity and temporal consistency. This work advances AI-driven human interaction modeling and has broad applications in virtual avatars, human-computer interaction, and social robotics. Future directions include expanding listener behavior diversity, improving real-time inference, and integrating more contextual cues for enhanced responsiveness.

**Limitations and Future Works** Despite its effectiveness in the expressiveness generation of listener videos, our model has certain limitations, particularly in inference efficiency. In the future, we will explore further techniques to accelerate the sampling procedure and provide faster video generation.

**Ethics Statement.** Our work aims to improve human be-

havior generation from a technical perspective and is not intended for malicious use like impersonation. Any future application should implement safeguards to ensure consent from the people whose likeness is being generated, and synthesized videos should clearly indicate their artificial nature.

## **Acknowledgments**

Research was sponsored by the Army Research Office and was accomplished under Cooperative Agreement Number W911NF-25-2-0040. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein

## References

- [1] Chaitanya Ahuja, Shugao Ma, Louis-Philippe Morency, and Yaser Sheikh. To react or not to react: End-to-end visual pose forecasting for personalized avatar during dyadic conversations. In *2019 International conference on multimodal interaction*, pages 74–84, 2019. 3
- [2] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*, 2020. 4
- [3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. 2
- [4] Dan Bohus and Eric Horvitz. Facilitating multiparty dialog with gaze, gesture, and speech. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, pages 1–8, 2010. 3
- [5] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactr: Tuning-free mutual self-attention control for consistent image synthesis and editing. *arXiv preprint arXiv:2304.08465*, 2023. 2
- [6] Di Chang, Yichun Shi, Quankai Gao, Jessica Fu, Hongyi Xu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion, 2024. 2
- [7] Di Chang, Yufeng Yin, Zongjian Li, Minh Tran, and Mohammad Soleymani. Libreface: An open-source toolkit for deep facial expression analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8205–8215, 2024. 1
- [8] Zhiyuan Chen, Jiajiong Cao, Zhiqian Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. *arXiv preprint arXiv:2407.08136*, 2024. 2
- [9] Hang Chu, Daiqing Li, and Sanja Fidler. A face-to-face neural conversation model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7113–7121, 2018. 3
- [10] Jiahao Cui, Hui Li, Yao Yao, Hao Zhu, Hanlin Shang, Kaihui Cheng, Hang Zhou, Siyu Zhu, and Jingdong Wang. Hallo2: Long-duration and high-resolution audio-driven portrait image animation. *arXiv preprint arXiv:2410.07718*, 2024. 3
- [11] Radek Danecek, Michael J. Black, and Timo Bolkart. EMOCA: Emotion driven monocular face capture and animation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20311–20322, 2022. 4, 7, 8, 1
- [12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 4, 6
- [13] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 2663–2671, New York, NY, USA, 2022. Association for Computing Machinery. 7
- [14] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 4
- [15] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 40(8), 2021. 7
- [16] Scott Geng, Revant Teotia, Purva Tendulkar, Sachit Menon, and Carl Vondrick. Affective faces for goal-driven dyadic communication, 2023. 3, 5, 6, 8
- [17] Jonathan Gratch, Anna Okhmatovskaia, Francois Lamothe, Stacy Marsella, Mathieu Morales, Rick J van der Werf, and Louis-Philippe Morency. Virtual rapport. In *International Workshop on Intelligent Virtual Agents*, pages 14–27. Springer, 2006. 2
- [18] Jonathan Gratch, Ning Wang, Jillian Gerten, Edward Fast, and Robin Duffy. Creating rapport with virtual agents. In *Intelligent Virtual Agents (IVA)*, pages 125–138, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. 3
- [19] David Greenwood, Stephen Laycock, and Iain Matthews. Predicting head pose in dyadic conversation. In *International Conference on Intelligent Virtual Agents*, pages 160–169. Springer, 2017. 3
- [20] Yuming Gu, You Xie, Hongyi Xu, Guoxian Song, Yichun Shi, Di Chang, Jing Yang, and Linjie Luo. Diffportrait3d: Controllable diffusion for zero-shot portrait view synthesis, 2023. 2
- [21] Jiazhi Guan, Zhanwang Zhang, Hang Zhou, Tianshu Hu, Kaisiyuan Wang, Dongliang He, Haocheng Feng, Jingtuo Liu, Errui Ding, Ziwei Liu, et al. Stylesync: High-fidelity generalized and personalized lip sync in style-based generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1505–1515, 2023. 2
- [22] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. *arXiv preprint arXiv:2311.16933*, 2023. 2
- [23] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2, 7
- [24] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:1706.08500*, 2017. 7
- [25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 2

- [26] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 7
- [27] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023. 2
- [28] Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. Virtual rapport 2.0. In *International workshop on intelligent virtual agents*, pages 68–79. Springer, 2011. 3
- [29] Patrik Jonell, Taras Kucherenko, Erik Ekstedt, and Jonas Beskow. Learning non-verbal behavior for a social robot from youtube videos. In *ICDL-EpiRob Workshop on Naturalistic Non-Verbal and Affective Human-Robot Interactions, Oslo, Norway, August 19, 2019*, 2019. 3
- [30] Patrik Jonell, Taras Kucherenko, Gustav Eje Henter, and Jonas Beskow. Let’s face it: Probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2020. 3
- [31] Yukang Lin, Haonan Han, Chaoqun Gong, Zunnan Xu, Yachao Zhang, and Xiu Li. Consistent123: One image to highly consistent 3d asset using case-aware diffusion priors, 2023. 2
- [32] Jin Liu, Xi Wang, Xiaomeng Fu, Yesheng Chai, Cai Yu, Jiao Dai, and Jizhong Han. Mfr-net: Multi-faceted responsive listening head generation via denoising diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6734–6743, 2023. 2
- [33] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization, 2023. 2
- [34] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023. 2
- [35] Tao Liu, Feilong Chen, Shuai Fan, Chenpeng Du, Qi Chen, Xie Chen, and Kai Yu. Anitalker: Animate vivid and diverse talking faces through identity-decoupled facial motion encoding, 2024. 2
- [36] Xi Liu, Ying Guo, Cheng Zhen, Tong Li, Yingying Ao, and Pengfei Yan. Customlistener: Text-guided responsive interaction for user-friendly listening head generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2415–2424, 2024. 2, 3
- [37] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024. 3
- [38] Evonne Ng. Learning2listen. <https://evonneng.github.io/learning2listen/>. 4, 5, 6, 7, 1
- [39] Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20395–20405, 2022. 2, 3, 4, 7, 8
- [40] Evonne Ng, Sanjay Subramanian, Dan Klein, Angjoo Kanazawa, Trevor Darrell, and Shiry Ginosar. Can language models learn to listen? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10083–10093, 2023. 2, 4
- [41] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, page 484–492, New York, NY, USA, 2020. Association for Computing Machinery. 2
- [42] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. 4, 6
- [43] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13759–13768, 2021. 3, 7
- [44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2
- [45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10684–10695, 2022. 4
- [46] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 35:36479–36494, 2022. 2
- [47] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2
- [48] Luchuan Song, Guojun Yin, Zhenchao Jin, Xiaoyi Dong, and Chenliang Xu. Emotional listener portrait: Realistic listener motion simulation in conversation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20782–20792. IEEE, 2023. 2, 3, 5, 6, 7, 8
- [49] Siyang Song, Micol Spitale, Cheng Luo, Cristina Palmero, German Barquero, Hengde Zhu, Sergio Escalera, Michel Valstar, Tobias Baur, Fabien Ringeval, Elisabeth Andre, and Hatice Gunes. React 2024: the second multiple appropriate facial reaction generation challenge, 2024. 2
- [50] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2
- [51] Jiaxiang Tang, Kaisiyuan Wang, Hang Zhou, Xiaokang Chen, Dongliang He, Tianshu Hu, Jingtuo Liu, Gang Zeng, and Jingdong Wang. Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. *arXiv preprint arXiv:2211.12368*, 2022. 2
- [52] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. 5, 2

- [53] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions. *arXiv preprint arXiv:2402.17485*, 2024. 2
- [54] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 2
- [55] Minh Tran, Di Chang, Maksim Siniukov, and Mohammad Soleymani. Dim: Dyadic interaction modeling for social behavior generation. In *European Conference on Computer Vision*, pages 484–503. Springer, 2024. 2, 3, 4, 5, 6, 7, 8, 1
- [56] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. 7
- [57] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6309–6318, 2017. 3
- [58] Cong Wang, Kuan Tian, Jun Zhang, Yonghang Guan, Feng Luo, Fei Shen, Zhiwei Jiang, Qing Gu, Xiao Han, and Wei Yang. V-express: Conditional dropout for progressive training of portrait video generation. *arXiv preprint arXiv:2406.02511*, 2024. 2
- [59] Duomin Wang, Bin Dai, Yu Deng, and Baoyuan Wang. Agent-avatar: Disentangling planning, driving and rendering for photorealistic avatar agents. *CoRR*, abs/2311.17465, 2023. 2
- [60] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 7
- [61] Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation, 2024. 3
- [62] Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989. 7
- [63] Jiaqi Xu, Xinyi Zou, Kunzhe Huang, Yunkuo Chen, Bo Liu, MengLi Cheng, Xing Shi, and Jun Huang. Easyanimate: A high-performance long video generation method based on transformer architecture. *arXiv preprint arXiv:2405.18991*, 2024. 4, 5, 6
- [64] Jiaqi Xu, Xinyi Zou, Kunzhe Huang, Yunkuo Chen, Bo Liu, MengLi Cheng, Xing Shi, and Jun Huang. Easyanimate: A high-performance long video generation method based on transformer architecture, 2024. 4
- [65] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024. 2, 3
- [66] Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time. *arXiv preprint arXiv:2404.10667*, 2024. 2
- [67] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1481–1490, 2024. 2
- [68] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. I\_ADAPTER: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 4
- [69] Jun Yu, Shenshen Du, Haoxiang Shi, Yiwei Zhang, Renbin Su, Zhongpeng Cai, and Lei Wang. Responsive listening head synthesis with 3dmm and dual-stream prediction network. In *Proceedings of the 1st International Workshop on Multimedia Content Generation and Evaluation: New Methods and Practice*, pages 137–143, 2023. 5, 6
- [70] Chenxu Zhang, Chao Wang, Jianfeng Zhang, Hongyi Xu, Guoxian Song, You Xie, Linjie Luo, Yapeng Tian, Xiaohu Guo, and Jiashi Feng. Dream-talk: diffusion-based realistic emotional audio-driven method for single image talking face generation. *arXiv preprint arXiv:2312.13578*, 2023. 2
- [71] Lyumin Zhang. [major update] reference-only control · mikubill/sd-webui-controlnet · discussion #1236, 2023. 2
- [72] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Int. Conf. Comput. Vis.*, pages 3836–3847, 2023. 2
- [73] Longhao Zhang, Shuang Liang, Zhipeng Ge, and Tianshu Hu. Personatalk: Bring attention to your persona in visual dubbing. *arXiv preprint arXiv:2409.05379*, 2024. 2
- [74] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7
- [75] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8652–8661, 2023. 2
- [76] Wei Zhao, Peng Xiao, Rongju Zhang, Yijun Wang, and Jianxin Lin. Semantic-aware responsive listener head synthesis. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7065–7069, 2022. 2
- [77] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2
- [78] Mohan Zhou, Yalong Bai, Wei Zhang, Ting Yao, Tiejun Zhao, and Tao Mei. Vico-x: Multimodal conversation dataset. <https://project.mhzhou.com/vico>, 2022. Accessed: 2022-09-30. 5
- [79] Mohan Zhou, Yalong Bai, Wei Zhang, Ting Yao, Tiejun Zhao, and Tao Mei. Responsive listening head generation: a benchmark dataset and baseline. In *European Conference on Computer Vision*, pages 124–142. Springer, 2022. 2, 5, 6, 7, 8
- [80] Mohan Zhou, Yalong Bai, Wei Zhang, Ting Yao, and Tiejun Zhao. Interactive conversational head generation. *arXiv preprint arXiv:2307.02090*, 2023. 2

- [81] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions on Graphics*, 39(6):1–15, 2020. [2](#)
- [82] Yongming Zhu, Longhao Zhang, Zhengkun Rong, Tianshu Hu, Shuang Liang, and Zhipeng Ge. Infp: Audio-driven interactive head generation in dyadic conversations. *arXiv preprint arXiv:2412.04037*, 2024. [2](#), [3](#), [6](#), [7](#), [8](#)

# DiTaiListener: Controllable High Fidelity Listener Video Generation with Diffusion

## Supplementary Material

Please choose the best generated listener video from Method A to Method E, given the speaker video on the left. According to the following criteria:

1. The response from the listener provides correct feedback to the speaker. (Feedback)
2. The head motions and facial expressions are diverse. (Diversity)
3. The general video quality is flawless and smooth. (Smoothness)

The figure shows two examples of a user study survey. Each example consists of a video player interface at the top and a rating matrix below it.

**Example 1 (Top):** Speaker video duration is 0:02 / 0:15. The rating matrix is:

	A	B	C	D	E
Feedback	<input type="radio"/>				
Diversity	<input type="radio"/>				
Smoothness	<input type="radio"/>				

**Example 2 (Bottom):** Speaker video duration is 0:15 / 0:28. The rating matrix is:

	A	B	C	D	E
Feedback	<input type="radio"/>				
Diversity	<input type="radio"/>				
Smoothness	<input type="radio"/>				

Figure 6. A screenshot of user study survey example. The methods are anonymized as A, B, C, D, E, and the order is randomized.

## 6. Detailed User Study

We show an example screenshot of our user study survey in Figure 6. The methods are well-anonymized, and we keep track of the original method order for each comparison offline. The participants are paid \$12/hr for their labor, and we collect the results of each user after careful human review. The incomplete and invalid responses are rejected. After such a review, we have 50 valid responses in total. The result is presented in Table 1.

## 7. Details of Dataset Processing

**RealTalk** database includes videos from The Skin Deep YouTube Podcast and contains 692 dyadic conversations. The total duration of the videos is 115 hours. The database provides bounding boxes of participants' faces, as well as EMOCA 3DMM coefficients [11]. The dataset contains around 50,000 conversational turns of dyadic interactions with specified roles, i.e., speaker vs listener. We noticed that some of the videos contain artifacts that make them hard to use for direct training in the video-generation model; such artifacts include occlusions between the listener and the camera, bounding boxes not covering full faces, shaky

bounding boxes, bounding boxes changing size, and extreme head poses.

Training the model on raw data led to model collapse (floating hands, blurry, distorted faces, face out of frame), necessitating data cleaning. To address this, we applied a multi-step filtering process. First, for each clip, instead of using per-frame bounding boxes, we identified the largest bounding box covering faces in all frames and used it for cropping. This provided a static background and a constant bounding box size and removed frame-to-frame jitter. We then used LibreFace [7] to estimate head pose and only kept frames with frontal-looking faces by applying pitch and yaw thresholds; the acceptable range of yaw and pitch angles was set to  $[-30, 30]$  degrees. In order to remove other frame-level artifacts, we collected annotations in-house. For every one of the 50,000 clips, we randomly selected a frame. Annotators were to label if each frame had artifacts of some category: split screens, text overlays or transition effects, non-hand objects blocking the view, visible hands, and other visual disturbances such as glare. Frames without artifacts were automatically labeled as clean. Initial manual sparse annotations were used to finetune a CLIP image encoder for artifact detection. This method provided 0.95 recall on unseen clips. As a result, 13.6% of all data was marked for removal due to detected artifacts. The resulting model was applied with a high sensitivity threshold to filter artifacts on all the other frames in the dataset.

## 8. Details of 3DMM Space Evaluation Metrics

- Fréchet Distance (FD) measures discrepancies between distributions of real and generated listener's EMOCA 3DMM coefficients. It is separately computed for head poses and facial expressions.
- Paired FD (P-FD) is an extension of FD for synchrony. P-FD is calculated by concatenating listener motions with GT speaker motions and measuring Fréchet Distance of resulting features. It measures how well listener motions are temporally aligned with speaker motions.
- Variance (Var) and SI for Diversity (SID) are used to measure the diversity of generated listener motions, following DIM [55] and L2L [38]. SID is computed by applying k-means clustering to 3DMM features and computing entropy of the histogram of cluster assignments. The higher the SID, the more diverse the motions are. Var is the variance of pose and expression features that indicate how expressive the movements are across temporal dimensions.

## 9. Text control

We add text control inputs to guide the model with text prompts. ViCo and RealTalk datasets do not contain any textual captions. Therefore, we performed captioning in two stages: text extraction and prompt refinement. In the first stage, we used a Large Video-Language-Model (VLM) Google Gemini 1.5 Flash 002 [52] to extract text descriptions of audio-visual emotional cue information from the listener video with the speaker audio. The following prompt was used:

Which emotion is present in the video? Emotion can belong to only one of the following classes - angry, happy, sad, neutral, disgust, fear, surprise. Give your response in JSON format as following: {"emotion": "angry", "reason": "reason for your response"}

Example of model response:

```
{"emotion": "happy", "reason": "The person in the video is expressing feelings of freedom and self-acceptance. The overall tone is positive, suggesting contentment and happiness. He talks about being able to be himself without conforming to external pressures. This indicates a sense of liberation and joy. Phrases such as \"I'm free\" and \"I get to be me\" directly convey positive emotions."}
```

Extracted information has emotional cues from the speech and quotes that are important to describe the context of the conversation. However, those do not directly affect the appearance of the listener; sometimes, the VLM lists what emotions are not present in the video. Therefore, an extra step of text processing was needed to create a text control prompt from extracted information. We use Llama-3.3-70B-Instruct Large Language Model (LLM) [54] with the following prompt:

"""

Instructions:

Given an analytical description of a dyadic conversation with a video of a listener, extract only information about the listener's physical expressions and emotions, conversation emotions. Discard conversation quotes, reasoning thoughts, and any general explanations or definitions of

emotions. Only retain direct observations about the listener's facial expressions, emotions, or other reactions. Give only extracted information in plaintext form in a single sentence.

Analytical description example:  
"The person in the video is shown smiling broadly, with their mouth wide open and eyes crinkled at the corners. This is a classic display of happiness. There are no other indicators of any other emotion present."

Example extracted: Listener looks happy, smiling broadly with mouth open and eyes crinkled.

User:

Analytical description:  
"{"text"}"

Assistant:

Extracted: Listener looks """

The processed text is used as text control. Example of resulting text control: "happy, smiling broadly, showing her teeth".