VividListener: Expressive and Controllable Listener Dynamics Modeling for Multi-Modal Responsive Interaction

Shiying Li[1], Xingqun Qi[2,*], Bingkun Yang[1], Weile Chen[3], Zezhao Tian[1], Muyi Sun[1,*,†], Qifeng Liu[2], Man Zhang[1], Zhenan Sun[4]

[1]Beijing University of Posts and Telecommunications, Beijing, China
[2]The Hong Kong University of Science and Technology, Hong Kong, China
[3]Zhejiang University, Hangzhou, China
[4]Institute of Automation, Chinese Academy of Sciences, Beijing, China

Abstract

Generating responsive listener head dynamics with nuanced emotions and expressive reactions is crucial for dialogue modeling in various virtual avatar animations. Previous studies mainly focus on the direct short-term production of listener behavior. They overlook the fine-grained control over motion variations and emotional intensity, especially in long-sequence modeling. Moreover, the lack of long-term and large-scale paired speaker-listener corpora incorporating head dynamics and fine-grained multi-modality annotations limits the application of dialogue modeling. Therefore, we first newly collect a large-scale multi-turn dataset of 3D dyadic conversation containing more than 1.4M valid frames for multi-modal responsive interaction, dubbed **ListenerX**. Additionally, we propose **VividListener**, a novel framework enabling fine-grained, expressive, and controllable **listener dynamics modeling**. This framework leverages multi-modal conditions as guiding principles for fostering coherent interactions between speakers and listeners. Specifically, we design the **Responsive Interaction Module** (RIM) to adaptively represent the multi-modal interactive embeddings. RIM ensures the listener dynamics achieve fine-grained semantic coordination with textual descriptions and adjustments, while preserving expressive reaction with speaker behavior. Meanwhile, we propose the **Emotional Intensity Tags** (EIT) for emotion intensity editing with multi-modal information integration, applying to both text descriptions and listener motion amplitude. Extensive experiments conducted on our newly collected ListenerX dataset demonstrate that VividListener achieves state-of-the-art performance, realizing expressive and controllable listener dynamics.

Introduction

Listener dynamics modeling aims to generate expressive and emotional head movements which respond to the speech content of the corresponding speakers. These non-verbal interactive behaviors (Volonte et al. 2020; Zhao et al. 2025) like facial expressions, nodding and blinking significantly facilitate message delivery during human daily communications. Meanwhile, modeling listener dynamics displays a wide range of applications in human-machine interaction (Qi et al. 2025, 2024c,b), embodied AI (Duan et al. 2022), robotics (Soori, Arezoo, and Dastres 2023), and virtual avatar animations (Genay, Lécuyer, and Hachet 2021; Qi et al. 2023).

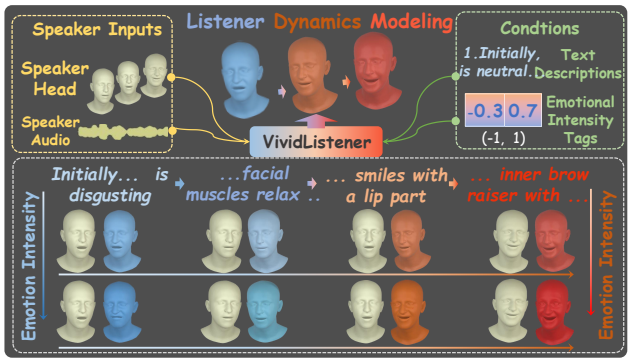

Figure 1: VividListener: listener dynamics modeling for multi-modal responsive interaction. This framework inputs speaker head motions, speaker audios, the conditions of listener textual descriptions, and the emotional intensity tags, which outputs listener head dynamics sequences. The generated listener changes with the expression descriptions (from disgusting to excited), and exhibits varying degrees of facial emotions from light to graydark. (Zoom in for better details.)

Previous studies on dialogue modeling have mostly focused on speaker generation (talking face/ head generation) and achieved satisfactory performances (Chen, Tran-Thien-Y, and Florence 2021; Pei et al. 2024). However, these works overlook the practical significance of interactive scenarios, especially in modeling the response of the listener.

Recently, some works have explored synthesizing the motions of the listener from the given speaker corpus, guided by simplistic emotion labels (Song et al. 2023; Liu et al. 2024b; Tran et al. 2024). Nevertheless, these approaches are typically limited to short-term head motion modeling and analysis, where the semantic and expression are significantly insufficient and incomplete. Furthermore, the coarse perception of listener emotions with few basic categories exacerbates these shortcomings, resulting in monotonous head

*Corresponding Authors.
†Project Leader.

expressions. These methods overlook the fine-grained motion variants and the intensity of emotions within lengthy dialogue sequences, which are more practical in real-world scenes. For example, the listener emotions may fluctuate dynamically during communication, shifting from excitement to frustration. Therefore, in this work, we introduce the new task of **Expressive and Controllable Listener Dynamics Modeling** for multi-modal responsive interaction in long conversational sequences, as illustrated in Figure 1.

There are two main challenges in the listener dynamics modeling task: 1) Datasets containing fine-grained annotations of listener head movements and emotion intensity corresponding to speaker corpus are scarce. 2) Modeling complex and variable emotion shifting of listener behavior is difficult, especially in long dialogue sequences from in-the-wild scenes.

To address the data scarcity problem, we first construct a new dataset of 3D head dyadic conversation which contains 1.4M frames of emotional speaker-listener head movements, dubbed **ListenerX**. Specifically, we extract long-term and continuous dialogue segments from in-the-wild videos, ensuring that both the conversational participant heads remain fully visible throughout the interaction. These segments should also include scenarios where the speaker articulates at least two clear sentences, while the listener reacts dynamically rather than monotonously. To process these segments, we employ an advanced 3D facial avatar estimation method (Daněček and Black 2022) to extract high-quality facial expressions and head poses (FLAME (Li et al. 2017)) for both conversational partners. Meanwhile, to support our insights into multimodal cue-guided modeling, we combine traditional detection models (Chang et al. 2024) with vision-language models (Liu et al. 2024a) to generate fine-grained and emotion-aware facial expression descriptions for continuous head frames. To control emotional intensity, we use facial emotion analysis techniques (Toisoul et al. 2021) to obtain Valence (how positive the emotion is) and Arousal (how calming or exciting it is), providing a continuous representation linked to emotional categories. In this fashion, our ListenerX dataset covers high-quality and long-term listener head movements with corresponding fine-grained expression descriptions and emotion tags, which pave the way for various downstream tasks like human-human interaction analysis (Stergiou and Poppe 2019; Qi et al. 2024a; Zhang et al. 2025), talking face generation (Pei et al. 2024), and responsive interaction modeling (Chen, Tran-Thien-Y, and Florence 2021).

Based on **ListenerX**, we propose **VividListener**, a novel framework that enables expressive and controllable listener head dynamics modeling with Diffusion Transformer (DiT). Our key insight is to build coherent interactions between the speaker and the responsive listener for listener modeling. Hence, vividListener adaptively incorporates the multi-modal conditions as guiding principles via a diffusion-based model. In particular, we propose the Responsive Interaction Module (RIM) to ensure temporal alignment between the generated listener movements and the speaker audio rhythms, while presenting fine-grained coordination *w.r.t.* text guidance. Here, we first model the joint interactive embeddings of the listener emotion descriptions and the integrated conversational cues provided by the given speaker audio and motion prior. In this manner, the discrete text-based descriptions are adaptively transformed into continuous temporal representations. Meanwhile, the speaker prior is seamlessly integrated into the learned joint embeddings. Moreover, we present Emotional Intensity Tags (EIT), a delicate design to achieve fine-grained emotion intensity control for listener dynamics. The combination of EIT and text descriptions (listener motion) serves as the input to learn a dynamic intensity tag through an integrated learning process. Then, we can leverage the intensity tag as a controllable modulation mechanism to balance the dependency of generated facial dynamics. The modulated features are finally fed into the DiT for producing expressive and controllable results.

Overall, our contributions are summarized as follows:

- We introduce a new task of fine-grained controllable listener head dynamics modeling, cooperating with one newly collected large-scale 3D head dyadic conversation dataset, namely ListenerX.
- We propose a novel framework, VividListener, that leverages multi-modal conditions as guiding principles to generate coherent and expressive listener head motions.
- We propose a Responsive Interaction Module to ensure coherent seamless integration of multi-modal conditions and a set of Emotional Intensity Tags to flexibly adjust emotional intensity, encouraging high-fidelity listener motion synthesis with desirable properties.
- Extensive experiments demonstrate that our framework achieves superior performance against various competitors, displaying expressive and interactive coherent listener head dynamics.

Related Work

Listener Dynamics Modeling

Listener dynamics modeling focuses on generating expressive listener head movements in dyadic conversations.Recent advances in 3D facial representation have enhanced the capture of fine facial motions, making it the standard for listener motion generation. Early works (Zhou et al. 2022; Ng et al. 2022; Tran et al. 2024; Ng et al. 2023) explored the correlation between speaker cues and listener motion. RLHG (Zhou et al. 2022) uses an LSTM-based method to produce basic reactions like nodding, while L2L (Ng et al. 2022) and DIM (Tran et al. 2024) employ VQ-VAE to learn motion codebooks. However, these models generate only coarse and uncontrollable motions.

Other works have achieved controllable generation based on input conditions (Song et al. 2023; Liu et al. 2023, 2024b). ELP (Song et al. 2023) and MER-Net (Liu et al. 2023) enable listener motion generation using input emotions as control conditions, but the control over emotions is still limited to simple emotion labels. CustomListener (Liu et al. 2024b) uses simple text input as a control condition, which lacks fine-grained modeling of listener emotions. Meanwhile, all the aforementioned works focus on short-term dimensions, making it challenging to model the com-

Dataset	Scale	Long-term	Modality				Annotation		Acquisition	
			Motion	Audio	Text	Tag.	Coarse	Fine-grained	Sou.	Rep.
L2L Dataset (Ng et al. 2022)	0.75M	✗(2S)	✓	✓	✗	✗	✗	✗	YouTube	FLAME
VICO (Zhou et al. 2022)	0.1M	✗(2S)	✓	✓	✗	✗	✓	✗	YouTube	3DMM
Realtalk (Geng et al. 2023)	-	✗(2S)	✓	✓	✗	✗	✗	✗	YouTube	FLAME
MDS Dataset (Tamon et al. 2024)	3M	✓(6S-11S)	✗	✓	✗	✗	✗	✗	Zoom	Video
ListenerX(Ours 2025)	**1.4M**	✓(8S)	✓	✓	✓	✓	✓	✓	**YouTube**	**FLAME**

Table 1: Statistical comparison of ListenerX against various counterparts. The "Long-term" means continuous and long-duration dialogue clips from in-the-wild videos. The "Coarse" means simple emotional category labels. The abbreviation "Tag." stands for emotional intensity tags. The abbreviations "Sou." and "Rep" refer to the data source and the representation method. The scale represents the number of valid data frames in the dataset, without comparing the duration of the original videos.

plex and dynamic listener reactions embedded with interactive and emotional information. In contrast, we introduce the long-sequence temporal modeling and integrate multi-modal annotations to achieve fine-grained emotional control in listener head dynamics generation.

Multi-Modal Head Generation

Multi-modal head generation aims to create 3D facial motion sequences based on input modalities such as speech audio, textual descriptions or emotional labels. Among these modalities, the audio modality has emerged as a primary focus of research due to the intrinsic correlation between speech and facial motion (Aneja et al. 2024; Cudeiro et al. 2019; Richard et al. 2021).Facetalk (Aneja et al. 2024) and VOCA (Cudeiro et al. 2019) focus on high-fidelity sequence modeling. However, their data acquisition processes are with high thresholds, which limit the practical applications.

Building on single-modality audio-driven approaches, severalmethods (Peng et al. 2023; Daněček et al. 2023; Sun et al. 2024; Thambiraja et al. 2023; Zhao et al. 2024)incorporate multimodal information to enable more precise control over 3D facial animation. EmoTalk (Peng et al. 2023) and EMOTE (Daněček et al. 2023) incorporate emotion labels to facilitate expression control, while DiffPoseTalk (Sun et al. 2024) and Imitator (Thambiraja et al. 2023) leverage reference videos to enhance contextual awareness and generate stylistic facial motions. Media2Face (Zhao et al. 2024) employs a diffusion-based method in the latent space guided by multi-modal inputs, including detailed text and images. However,Most methods focus on speaker modeling while overlooking interactive dynamics in conversations and fail to capture fine-grained conditional information from multi-modal inputs.Inspired by the above, we introduce a multi-model, interactive fine-grained listener head generation task, which enhances nuanced emotional interactions and enables emotional intensity control in dialogues.

ListenerX Dataset

To alleviate the dataset scarcity of the dyadic conversation, we propose **ListenerX**, a newly collected large-scale multi-turn dataset for 3D multi-modal responsive interaction, with listener textual descriptions and emotional intensity tags.

Data Collection

Considering the expensive and labor-consuming cost of 3D scanning and complex motion capture, similar to EMOTE (Daněček and Black 2022), we utilize in-the-wild videos as the data source and extract 3D representations by advanced facial avatar estimators. The original videos are sourced from interview shows and daily conversations, with each scenario accounting for half of the total frames, named as InterviewX and DailyX. InterviewX is characterized by structured contexts and stable emotional expressions, where listener responses primarily involve distinct non-verbal behaviors. In contrast, DailyX contains more casual and emotionally dynamic content, with listener facial movements exhibiting greater diversity and complexity. Constructing a dataset for our task emphasizes the acquisition of i) high-quality long-duration multi-turn interactive head movements between speakers and listeners; ii) fine-grained textual annotations describing motion variations, iii) and authoritative emotional intensity tags for facial expressions.

Multi-Model Annotation Pipeline

3D Dyadic Conversation Reconstruction. Firstly, we utilize face detection and voice source localization techniques (Chung and Zisserman 2016) to produce long-term video segments in which the faces of both participants are simultaneously visible. In these segments, the roles of the speaker and listener remain consistent over continuous periods. Then, we adopt the superior 3D facial estimator EMOCA (Daněček and Black 2022) to acquire the FLAME-based representation for each frame of both participants. Here, each clip is unify cropped as **8 seconds** where the facial expressions and head pose movements are parameterized. In this manner, our dataset supports long-term multi-turn sequence modeling of interactive listener dynamics.

Textual Descriptions. Once we obtain the estimated motion clips, we annotate the varying facial expressions and emotions with fine-grained textual descriptions. We observe that directly adopting the Vision-Language Models (VLM) (Liu et al. 2024a) on facial frames often produces inaccurate or fabricated descriptions, such as describing dimples that are not present in the faces. To this end, we incorporate the facial action unit detector (Chang et al. 2024) to extract high-intensity action units that serve as additional prompts for

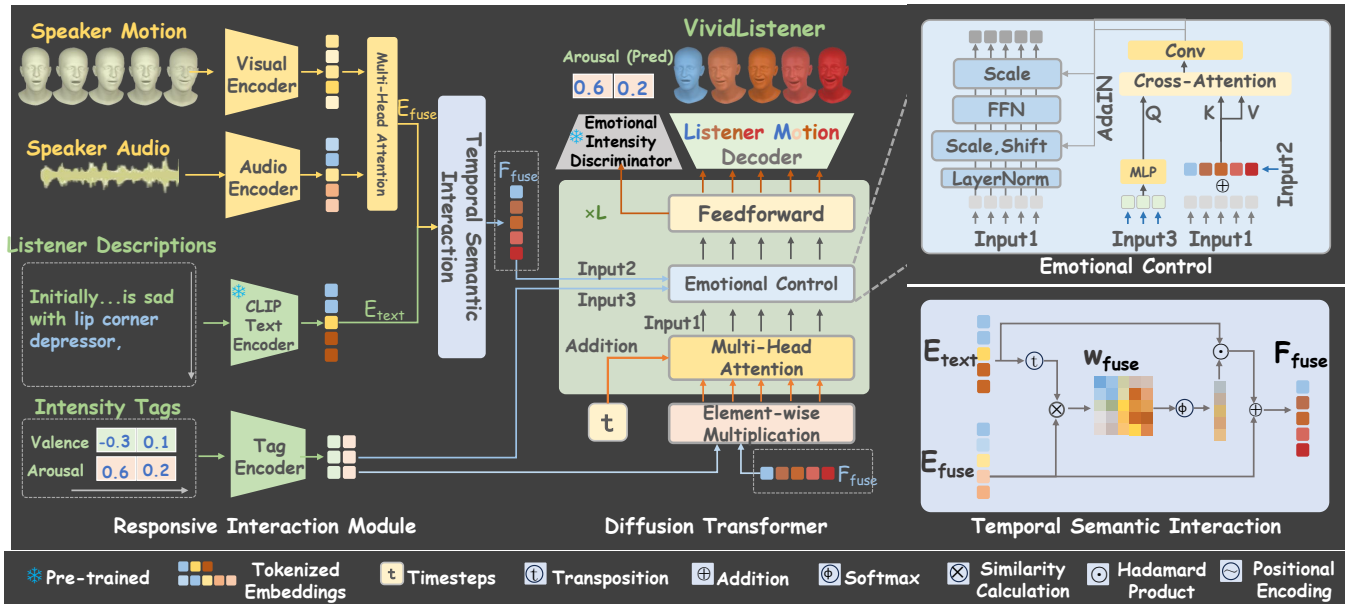

Figure 2: VividListener: listener dynamics modeling framework for multi-modal responsive interaction. This framework first integrates multi-modal inputs of the speaker and the listener through the Responsive Interaction Module. Then, the fused features and conditions are fed into the DiT pipeline for the final listener dynamics modeling, which will change with the text descriptions and intensity tags, from cool tones to warm tones. (Zoom in for better details.)

VLM. In this paradigm, we obtain accurate textual descriptions of the listener expressions. Moreover, we conduct manual revisions to ensure the generated results are coherent and semantically aligned with facial sequences.

Emotional Intensity Tags. To model expressive listener facial dynamics with authoritative emotional intensity tags, we incorporate facial affect analysis (Toisoul et al. 2021), using continuous dimensional representations of emotions: V (Valence, how positive the emotion is) and A (Arousal, how exciting the emotional display looks like) (Russell 1980). Unlike previous methods that use discrete emotion classification, our approach better reflects human emotional expressions in interactive dialogues.

Dataset Analysis

Overall, our ListenerX contains a total of 6,683 videos with 1.4M frames for 3D conversational motion sequences. Each dialogue sequence includes the interactive 3D head movements of both participants, the speaker audio, fine-grained long-text emotion descriptions, and the emotion intensity tags of the listener. Compared with the previous counterparts, ListenerX displays comprehensive modality attributes with fine-grained annotations as reported in Table 1. To the best of our knowledge, our ListenerX dataset is currently the largest for modeling long-term listener dynamics.See supplementary material for details.

VividListener Framework

Problem Formulation

Given the multi-modal inputs of the speaker motion S_m, audio representation S_a, fine-grained text descriptions $text$, and emotional intensity tags σ, the goal of our framework is to generate expressive and controllable listener head motions H. In practice, we leverage the parameterized FLAME model (Li et al. 2017) to represent facial expressions and head poses. We tackle this challenging problem with a specific-designed framework named VividListener. The overall objective is expressed as

$$H = \text{VividListener}(S_m, S_a, text, \sigma) \quad (1)$$

Listener Head Dynamics Modeling

To generate high-fidelity and diverse results, our VividListener adopts the Diffusion Transformer (DiT) (Peebles and Xie 2023) as the backbone, owing to its scalability and superior performance on sequence modeling, displayed in Figure 2. With the text descriptions of the listener movement as prompts, our framework first integrates the audio signals and head motion cues of the conversational speaker to produce fine-grained responsive guidance through the Responsive Interaction Module. Then, incorporated with the multi-modal responsive guidance, we employ the Emotional Intensity Tags to adaptively control the emotion intensity of generated results.

Responsive Interaction Module. To ensure the semantic coherent with text descriptions while preserving the vivid reaction, we propose Responsive Interaction Module (RIM) to model the temporal representations among multi-modal conditions. As illustrated in Figure 2, the speaker features including motion and audio, are firstly fused via a bidirectional Multi-Head Attention (MHA) mechanism. Once we acquire the fused reactive features, we consider taking textual motion scripts as the semantic descriptions to further en-

hance the generation of corresponding listener movements. Here, we utilize a pre-trained CLIP (Radford et al. 2021) as the text encoder to obtain sequence-aware semantic embeddings. The aforementioned reactive features and semantic embeddings are then integrated into a joint representation via Temporal Semantic Interaction for modeling semantic associations and temporal dynamics.

In particular, we model the temporal relevance between the speaker reactive features and semantic text embeddings via a joint similarity matrix, as shown in Figure 2 (bottom right corner). We aim to exploit the temporal interactive correlations to boost the speaker motion cues for drawing listener synthesis. Once we attain this matrix, we conduct temporal-wise adaptive pooling to acquire a set of normalized (norm) learnable weight coefficients. These coefficients represent the activated blending weights aligned with the text embeddings. Then, we obtain the semantic-enhanced sequential speaker reaction features by:

$$\begin{aligned} F_{\text{fused}} &= E_{\text{fused}} + E_{\text{text}} \odot Norm(\text{AdaPool}\left(W_{\text{fuse}}\right)) \\ W_{\text{fuse}} &= E_{\text{fused}} \otimes E_{\text{text}}{}' \end{aligned} \tag{2}$$

Where $W_{\text{fuse}} \in \mathcal{R}^{L \times L}$ is the temporal interactive correlation matrix, and L is the sequence length. AdaPool indicates the adaptive maxpooling operation (LeCun et al. 1998). $\odot$ means Hadamard product, $\otimes$ means matrix multiplication for similarity, and $'$ indicates the transposition operation. The semantic enhanced multi-modal features are further fed into our DiT-based denoiser. In this manner, the temporal interaction authority of the generated listener head is well-preserved.

Emotional Intensity Tags. For achieving fine-grained emotion intensity control in listener dynamics, the Emotional Intensity Tags serve as a key component in the VividListener framework. By integrating multi-modal information (*i.e.*textual descriptions and motion features), EIT are injected into both the input conditions and intermediate layers of the framework. It provides emotional intensity control in a dynamically modulated manner.

Specifically, we first aggregate the emotional intensity tags with semantic-enhanced reactive features via element-wise multiplication to produce the emotion guidance fed into our framework. Different from directly stacking the DiT blocks for listener dynamics generation, we further incorporate the aforementioned intensity tags and fused reactive motion cues into a specific-designed Emotional Control layer to draw a fine-grained facial effect. As depicted in Figure 2 (upper right), the enhanced reactive fused features (*i.e.*Input 2) are integrated with intermediate listener motion embeddings (*i.e.*Input 1). Then we leverage the emotion tag representations as the query Q to match the key features K and value features V via the cross-attention mechanism. Along with this operation, we further obtain the updated emotional guidance embeddings by a convolution layer. Here, we take a consideration leveraging the emotional guidance as the modulated indicators which are exploited to boost the listener motion features via an adaptive instance normalization (AdaIN) layer (Huang and Belongie 2017). By conducting this, we derive the listener dynamics features as:

$$F_{listener} = \text{AdaIN}\left(F_{listener}, conv(F_{emo})\right), \tag{3}$$

where $F_{listener}$ denotes listener motion features, and F_{emo} indicates the emotion guidance produced by cross-attention.

Objective Functions

During the training phase, the entire framework is trained with an end-to-end manner. Given the diffusion timestep t, the current speaker motion S_m and audio S_a, the textual descriptions $text$, emotional intensity tags σ, and the noised listener head motion $H^{(t)}$, the denoiser is designed to generate continuous listener head dynamics. The denoising process is constrained by a simple objective:

$$\mathcal{L}_{\text{s}} = \mathbb{E}_{H,t,\epsilon}\left[\left\| H - \text{VividL}(H^{(t)}, S_m, S_a, text, \sigma, t)\right\|_2^2\right] \tag{4}$$

Where $\epsilon \sim \mathcal{N}(0, I)$ is the added random Gaussian noise, $H^{(t)} = H + \gamma_{(t)}\epsilon$ represents the gradual noise addition process at step t. $\gamma_{(t)} \in (0, 1)$ is a constant hyper-parameter.

Emotional Intensity Predictor. To ensure that the generated listener emotion displays coherent alignment with the input emotional intensity tags, we employ a pre-trained 3D emotional intensity predictor to produce the corresponding emotion tag for our results. The emotional intensity loss is defined as follows:

$$\mathcal{L}_{\text{emotional}} = \|\sigma - \mathcal{P}_{\text{emotional}}(\hat{H})\|_2^2, \tag{5}$$

where $\hat{H}$ is our generated results. $\mathcal{P}_{\text{emotional}}$ denotes emotion intensity predictor.

Moreover, we utilize the velocity loss $\mathcal{L}_{vel}$ to provide supervision on the smoothness (Zhang et al. 2022). Finally, our overall objective function is defined as follows:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{s}}\mathcal{L}_{\text{s}} + \lambda_{\text{emotional}}\mathcal{L}_{\text{emotional}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}}. \tag{6}$$

λ_{simple}, $\lambda_{\text{emotional}}$, and λ_{vel} are the trade-off coefficients.

Experiments

Experimental Settings

Implementation Details. In our experiments, we set the total generated sequence length N = 240 with the normalized fps=30. S_a is initially an audio waveform, converted into 128 × 480 mel-spectrograms using a 512 FFT window and a 160 hop length. The textual descriptions $text$ of the listener are encoded with the CLIP model (Radford et al. 2021). The listener emotional intensity tags σ are obtained at the frequency of 5 Hz. In the training stage, we empirically set λ_{simple}=2, $\lambda_{\text{emotional}}$=0.2, λ_{vel}=0.8. The initial learning rate is set to 1×10^{-5} with AdamW optimizer (Loshchilov 2017).

Evaluation Metrics. To comprehensively evaluate the realism, synchrony, and diversity of the generated listener dynamics, we introduce several metrics, including Fréchet Distance (FD), Paired Fréchet Distance (P-FD), Mean Squared Error (MSE), Shannon Index (SID), Variance (Var), and Residual Pearson Correlation Coefficient (rPCC). Please refer to the supplementary materials for details.

Dataset Scenario	Methods	FD↓		P-FD↓		MSE↓		SID↑		Var↑		rPCC↓	
		Exp	Pose	Exp	Pose	Exp	Pose	Exp	Pose	Exp	Pose	Exp	Pose
InterviewX+ DailyX (ListenerX)	RLHG (Zhou et al. 2022)	5.842	0.021	6.164	0.020	0.133	0.003	4.765	3.797	0.212	0.015	0.033	0.005
	L2L (Ng et al. 2022)	11.510	0.115	11.770	0.113	0.260	0.127	1.930	1.131	0.182	0.014	0.028	0.134
	DIM (Tran et al. 2024)	7.805	0.046	8.291	0.042	0.212	0.006	4.050	3.441	0.216	0.013	0.074	0.011
	Vividlistener (Ours)	**3.792**	**0.036**	**4.177**	**0.038**	**0.120**	**0.006**	**4.803**	**3.830**	**0.210**	**0.015**	**0.040**	**0.037**
DailyX(Train) InterviewX(Test)	RLHG (Zhou et al. 2022)	41.510	0.122	41.961	0.126	0.854	0.028	1.040	2.452	0.882	0.004	0.068	0.171
	L2L (Ng et al. 2022)	44.020	0.053	44.644	0.058	0.920	0.011	1.452	2.630	0.884	0.005	0.043	0.154
	DIM (Tran et al. 2024)	44.643	0.142	44.870	0.147	0.913	0.020	0.910	2.474	0.882	0.005	0.043	0.138
	Vividlistener (Ours)	**16.309**	**0.023**	**16.890**	**0.028**	**0.363**	**0.003**	**1.460**	**1.506**	**0.821**	**0.005**	**0.094**	**0.168**

Table 2: The top table presents a comparison with state-of-the-art methods on the ListenerX dataset, while the bottom table illustrates cross-scenario inference results.↑ means the higher the better, and ↓ indicates the lower the better.

Methods	ListenerX			
	FD↓	SID↑	Var↑	rPCC↓
w/o TSI	6.38	1.82	0.18	0.09
w/o EC	4.00	4.72	0.18	0.05
w/o EIP	5.95	2.07	0.09	0.08
Vividlistener (Full)	**3.79**	**4.80**	**0.21**	**0.04**

Table 3: Ablation study of modules on the ListenerX dataset.

Input Conditions	ListenerX			
	FD↓	SID↑	Var↑	rPCC↓
w/o Text	4.83	4.14	0.18	0.05
w/o Tag	4.79	4.34	0.10	0.04
Vividlistener (Full)	**3.79**	**4.80**	**0.21**	**0.04**

Table 4: Ablation study of conditions on the ListenerX.

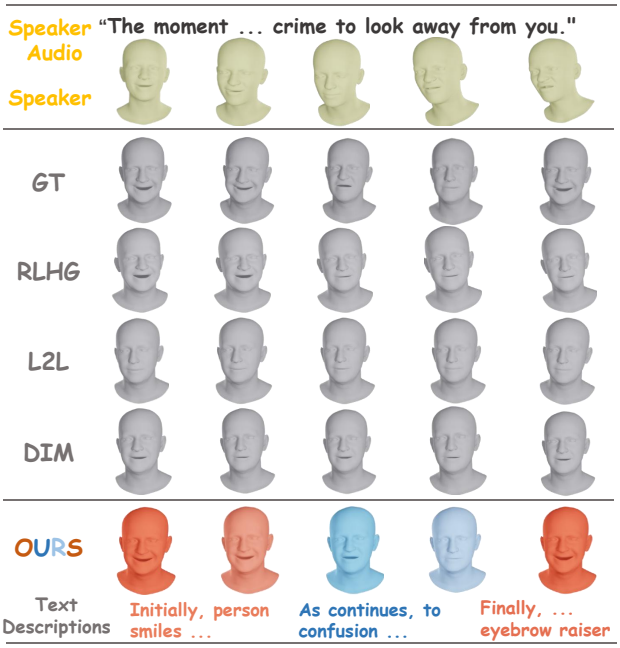

Figure 3: Visual comparisons on ListenerX. We present visualizations of listener motions generated on ListenerX compared to various state-of-the-art methods. Unlike other approaches, our VividListener input incorporates fine-grained textual descriptions (shown in the last row).

Quantitative Analysis

Comparisons with the SOTA counterparts. To verify effectiveness of our method, we compare VividListener with SOTA models for listener modeling: L2L (Ng et al. 2022), RLHG (Zhou et al. 2022), and DIM (Tran et al. 2024). These models are re-trained and tested on our dataset using the source code released by authors. Experiments are conducted in two groups: firstly, on the whole ListenerX dataset to assess performance in complex scenarios, secondly, on DailyX for training and InterviewX for testing to evaluate generalizability. As shown in Table 2, VividListener consistently outperforms competing models. Notably, in second experiment, it achieved a 25.21% reduction in FD compared to sub-optimal models, demonstrating superior generalization.
Ablation Study. To further evaluate the effectiveness of our network design, we conduct the ablation studies on different components and input conditions as variations. **For verifying the effects of the Vividlistener modules,** we conduct ablation studies on the Temporal Semantic Interaction (TSI), Emotional Control (EC), and the Emotional Identity Predictor (EIP). As shown in Table 3, removing TSI weakens semantic association and temporal modeling between text and speaker information, degrading global control and overall performance. When EC is removed, SID and rPCC decrease, indicating its role in enhancing interactive information and control over listener motion. Lastly, removing the EIP results in a decline in the Var metric, indicating its role in maintaining intensity consistency for higher-quality emotional expression. **For verifying the effects of the conditional input guidances,** we conduct ablation studies by removing the text input and the emotion intensity tags, as shown in Table 4. Removing the text disrupts the semantic guidance provided by textual descriptions, leading to semantic alignment loss in the generated listener motion. This re-

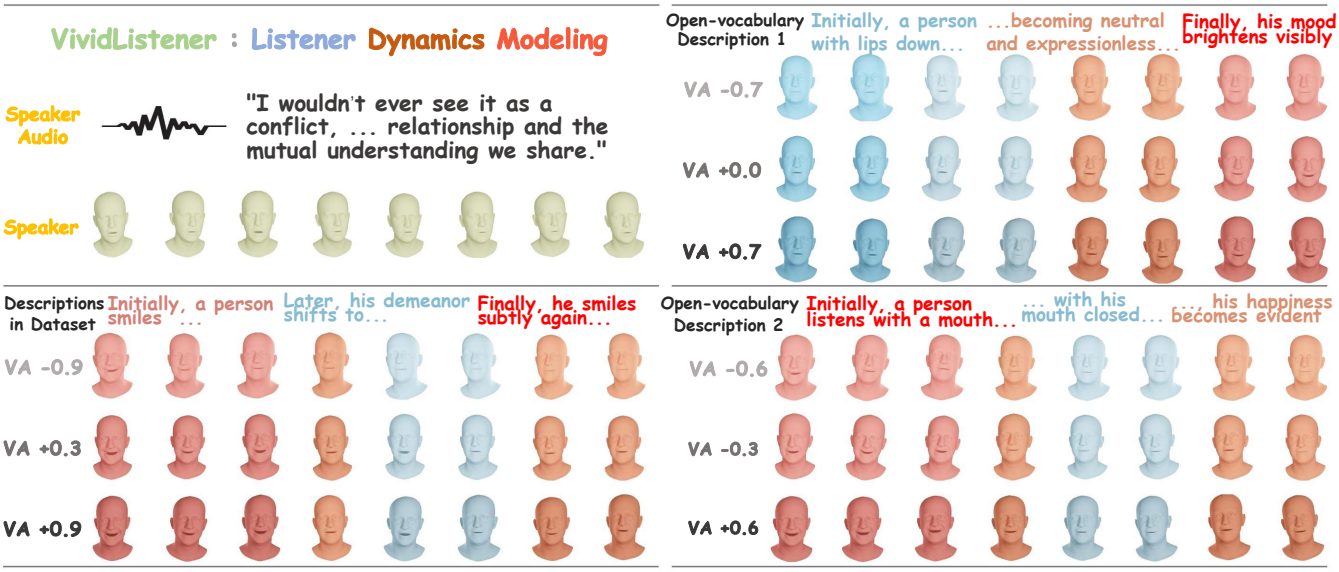

Figure 4: Conditional Control Results. We generate listener head dynamics by inputting the speaker audio and head motion information (shown in the top-left corner) alongside dataset-based emotion descriptions (shown in the bottom-left corner). Additionally, we utilize open-vocabulary text inputs to further enrich the generation process. Each example incorporates three VA values (e.g., -0.9, +0.3, +0.9) to control emotional intensity levels (Please Zoom in for better details.)

Methods	L2L		ViCo	
	FD ↓	SID ↑	FD ↓	SID ↑
RLHG	7.56	4.38	5.37	6.73
L2L	7.21	4.75	6.10	5.95
DIM	6.97	3.76	6.28	6.02
Ours	**5.36**	**6.84**	**5.39**	**7.84**

Table 5: FD and diversity on L2L and ViCo datasets.

sults in noticeable degradation in the rPCC and FD metrics, as the model fails to establish coherent interactions between speaker cues and listener responses. Similarly, removing the emotion intensity tag weakens the model ability to modulate the intensity of the generated motion, reducing the diversity and emotional expressiveness (SID and Var) of the output.

Experiments Conducted on Existing Datasets To validate the effectiveness of our method, we conduct comparative experiments against several state-of-the-art approaches on two representative datasets, L2L and ViCo, both featuring short durations and limited modality coverage. As shown in Table 5, our method consistently surpasses all baselines across key evaluation metrics. These results underscore the advantage of our framework in short-term, low-resource scenarios.

Qualitative Analysis

Comparisons with the SOTA counterparts. To fully demonstrate the superior performance of VividListener, we present visualized keyframes generated by our method compared with various counterparts on ListenerX dataset. As shown in Figure 3 , our method produces more vivid and accurate listener head motions compared to other approaches. Specifically, we observe that L2L tends to generate rigid and monotonous results in long-sequence generation. RLHG achieves higher diversity. however, its facial expressions and head dynamics are limited in terms of expressiveness. DIM generates richer motions, yet results lack synchronization with speaker rhythm, leading to poor interactive dynamics.

Conditional Control Results. Figure 4 illustrates listener head motion generation results under different text-based conditions and emotional intensity controls. By incorporating open-vocabulary textual input conditions, our method surpasses the limitations of fixed-text descriptions in traditional datasets. For instance, Description 1 captures a gradual transition from a "somber expression" to a "brightened mood," while Description 2 show cases richer emotional fluctuations, demonstrating the model's adaptability and capability to generate diverse and nuanced dynamics.More qualitative results in the supplementary materials.

Conclusion

In this paper, we present VividListener, a novel framework for expressive and controllable listener dynamics modeling in multi-modal responsive interaction. To support this task, we introduce ListenerX, a large-scale 3D dyadic conversation dataset with fine-grained multi-modal conditions. We further propose a Responsive Interaction Module for listener-speaker alignment and integration, and design Emotional Intensity Tags for emotion intensity modulation. Experimental results on ListenerX demonstrate that VividListener achieves state-of-the-art performance in generating expressive and controllable listener reactions. In future work, we aim to explore more complex scenarios, such as jointly modeling speakers and listeners in dialogue interactions.

Acknowledgements

This work is supported by the Beijing Natural Science Foundation (Grant No. QY24210) and the National Natural Science Foundation of China (Grant No. 62306309, U23B2054, 62276263).

References

Aneja, S.; Thies, J.; Dai, A.; and Nießner, M. 2024. Facetalk: Audio-driven motion diffusion for neural parametric head models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21263–21273.

Chang, D.; Yin, Y.; Li, Z.; Tran, M.; and Soleymani, M. 2024. LibreFace: An open-source toolkit for deep facial expression analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 8205–8215.

Chen, J.-S.; Tran-Thien-Y, L.; and Florence, D. 2021. Usability and responsiveness of artificial intelligence chatbot on online customer experience in e-retailing. *International Journal of Retail & Distribution Management*, 49(11): 1512–1531.

Chung, J. S.; and Zisserman, A. 2016. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lipreading, ACCV*.

Cudeiro, D.; Bolkart, T.; Laidlaw, C.; Ranjan, A.; and Black, M. J. 2019. Capture, learning, and synthesis of 3D speaking styles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10101–10111.

Daněček, R.; and Black, M. J. 2022. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20311–20322.

Daněček, R.; Chhatre, K.; Tripathi, S.; Wen, Y.; Black, M.; and Bolkart, T. 2023. Emotional speech-driven animation with content-emotion disentanglement. In *SIGGRAPH Asia 2023 Conference Papers*, 1–13.

Duan, J.; Yu, S.; Tan, H. L.; Zhu, H.; and Tan, C. 2022. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2): 230–244.

Genay, A.; Lécuyer, A.; and Hachet, M. 2021. Being an avatar "for real": a survey on virtual embodiment in augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 28(12): 5071–5090.

Geng, S.; Teotia, R.; Tendulkar, P.; Menon, S.; and Vondrick, C. 2023. Affective faces for goal-driven dyadic communication. *arXiv preprint arXiv:2301.10939*.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, 1501–1510.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Li, T.; Bolkart, T.; Black, M. J.; Li, H.; and Romero, J. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.*, 36(6): 194–1.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024a. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, J.; Wang, X.; Fu, X.; Chai, Y.; Yu, C.; Dai, J.; and Han, J. 2023. Mfr-net: Multi-faceted responsive listening head generation via denoising diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6734–6743.

Liu, X.; Guo, Y.; Zhen, C.; Li, T.; Ao, Y.; and Yan, P. 2024b. CustomListener: Text-guided Responsive Interaction for User-friendly Listening Head Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2415–2424.

Loshchilov, I. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ng, E.; Joo, H.; Hu, L.; Li, H.; Darrell, T.; Kanazawa, A.; and Ginosar, S. 2022. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20395–20405.

Ng, E.; Subramanian, S.; Klein, D.; Kanazawa, A.; Darrell, T.; and Ginosar, S. 2023. Can language models learn to listen? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10083–10093.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195–4205.

Pei, G.; Zhang, J.; Hu, M.; Zhang, Z.; Wang, C.; Wu, Y.; Zhai, G.; Yang, J.; Shen, C.; and Tao, D. 2024. Deepfake generation and detection: A benchmark and survey. *arXiv preprint arXiv:2403.17881*.

Peng, Z.; Wu, H.; Song, Z.; Xu, H.; Zhu, X.; He, J.; Liu, H.; and Fan, Z. 2023. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20687–20697.

Qi, X.; Liu, C.; Li, L.; Hou, J.; Xin, H.; and Yu, X. 2024a. Emotiongesture: Audio-driven diverse emotional co-speech 3d gesture generation. *IEEE Transactions on Multimedia*, 26: 10420–10430.

Qi, X.; Liu, C.; Sun, M.; Li, L.; Fan, C.; and Yu, X. 2023. Diverse 3D hand gesture prediction from body dynamics by bilateral hand disentanglement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4616–4626.

Qi, X.; Pan, J.; Li, P.; Yuan, R.; Chi, X.; Li, M.; Luo, W.; Xue, W.; Zhang, S.; Liu, Q.; et al. 2024b. Weakly-Supervised Emotion Transition Learning for Diverse 3D Co-speech Gesture Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10424–10434.

Qi, X.; Wang, Y.; Zhang, H.; Pan, J.; Xue, W.; Zhang, S.; Luo, W.; Liu, Q.; and Guo, Y. 2025. Co3Gesture: Towards Coherent Concurrent Co-speech 3D Gesture Generation with Interactive Diffusion. In *ICLR*.

Qi, X.; Zhang, H.; Wang, Y.; Pan, J.; Liu, C.; Li, P.; Chi, X.; Li, M.; Zhang, Q.; Xue, W.; et al. 2024c. CoCoGesture: Toward Coherent Co-speech 3D Gesture Generation in the Wild. *arXiv preprint arXiv:2405.16874*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Richard, A.; Zollhöfer, M.; Wen, Y.; De la Torre, F.; and Sheikh, Y. 2021. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1173–1182.

Russell, J. A. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6): 1161.

Song, L.; Yin, G.; Jin, Z.; Dong, X.; and Xu, C. 2023. Emotional listener portrait: Neural listener head generation with emotion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20839–20849.

Soori, M.; Arezoo, B.; and Dastres, R. 2023. Artificial intelligence, machine learning and deep learning in advanced robotics, a review. *Cognitive Robotics*, 3: 54–70.

Stergiou, A.; and Poppe, R. 2019. Analyzing human–human interactions: A survey. *Computer Vision and Image Understanding*, 188: 102799.

Sun, Z.; Lv, T.; Ye, S.; Lin, M.; Sheng, J.; Wen, Y.-H.; Yu, M.; and Liu, Y.-j. 2024. Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models. *ACM Transactions on Graphics (TOG)*, 43(4): 1–9.

Tamon, M.; Yasuhisa, F.; Yukoh, W.; Kengo, O.; Ryota, N.; and Norihide, K. 2024. Listening Head Motion Generation for Multimodal Dialog System. In *2024 11th International Conference on Advanced Informatics: Concept, Theory and Application (ICAI)*, 1–6. IEEE.

Thambiraja, B.; Habibie, I.; Aliakbarian, S.; Cosker, D.; Theobalt, C.; and Thies, J. 2023. Imitator: Personalized speech-driven 3d facial animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20621–20631.

Toisoul, A.; Kossaifi, J.; Bulat, A.; Tzimiropoulos, G.; and Pantic, M. 2021. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 3(1): 42–50.

Tran, M.; Chang, D.; Siniukov, M.; and Soleymani, M. 2024. DIM: Dyadic Interaction Modeling for Social Behavior Generation. In *European Conference on Computer Vision*, 484–503. Springer.

Volonte, M.; Hsu, Y.-C.; Liu, K.-Y.; Mazer, J. P.; Wong, S.-K.; and Babu, S. V. 2020. Effects of interacting with a crowd of emotional virtual humans on users' affective and non-verbal behaviors. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 293–302. IEEE.

Zhang, H.; Li, Z.; Qi, X.; Li, M.; Sun, M.; Wang, S.; Zhang, M.; and Han, S. 2025. DanceEditor: Towards Iterative Editable Music-driven Dance Generation with Open-Vocabulary Descriptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12158–12168.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2022. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*.

Zhao, Q.; Long, P.; Zhang, Q.; Qin, D.; Liang, H.; Zhang, L.; Zhang, Y.; Yu, J.; and Xu, L. 2024. Media2face: Co-speech facial animation generation with multi-modality guidance. In *ACM SIGGRAPH 2024 conference papers*, 1–13.

Zhao, Y.; Wang, Y.; Wen, L.; Zhang, H.; and Qi, X. 2025. FreeDance: Towards Harmonic Free-Number Group Dance Generation via a Unified Framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10560–10569.

Zhou, M.; Bai, Y.; Zhang, W.; Yao, T.; Zhao, T.; and Mei, T. 2022. Responsive listening head generation: a benchmark dataset and baseline. In *European Conference on Computer Vision*, 124–142. Springer.