

Real-time Generation of Listener Nodding via Prediction of Kinematic Parameters for Avatar Dialogue Systems

Kazushi Kato
Graduate School of Informatics
Kyoto University
Kyoto, Japan
katou@sap.ist.i.kyoto-u.ac.jp

Koji Inoue
Graduate School of Informatics
Kyoto University
Kyoto, Japan
inoue@sap.ist.i.kyoto-u.ac.jp

Taiga Mori
Graduate School of Informatics
Kyoto University
Kyoto, Japan
mori@sap.ist.i.kyoto-u.ac.jp

Divesh Lala
Graduate School of Informatics
Kyoto University
Kyoto, Japan
lala@sap.ist.i.kyoto-u.ac.jp

Tatsuya Kawahara
Graduate School of Informatics
Kyoto University
Kyoto, Japan
kawahara@i.kyoto-u.ac.jp

Abstract

In human dialogue, we achieve smooth communication by expressing nonverbal cues such as eye contact, nodding, and facial expressions with precise timing. It is expected for conversational avatars to express these cues appropriately to realize natural and human-like interactions. This study focuses on nodding, which is crucial for demonstrating active listening and encouraging further user utterances. We propose a model that predicts both timing and kinematic parameters representing the motion features of listener nodding in real time. The proposed model consists of a timing prediction module and a kinematic parameter prediction module. Each implements a dyadic attention network over the speaker and listener channels based on the technique of Voice Activity Projection (VAP). Unlike conventional models, this approach enables real-time prediction of kinematic parameters based on the specific context of the dialogue rather than just predicting the timing. Furthermore, we demonstrate the effectiveness of fine-tuning the kinematic parameter prediction module initialized from the trained timing prediction module. The proposed model is lightweight and capable of real-time operation, and it has been integrated into an avatar dialogue system. Subjective evaluation experiments shows that our proposed method significantly outperforms both a baseline with stochastic timing and another with fixed-motion nodding. The code and trained models are available at <https://github.com/MaAI-Kyoto/MaAI>.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → *Multi-task learning*.

Keywords

Nodding Prediction; Kinematic Parameters; Human-Robot Interaction; Multimodal Interaction; Spoken Dialogue Systems

ACM Reference Format:

Kazushi Kato, Koji Inoue, Taiga Mori, Divesh Lala, and Tatsuya Kawahara. 2026. Real-time Generation of Listener Nodding via Prediction of Kinematic Parameters for Avatar Dialogue Systems. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '26)*, October 05–09, 2026, Napoli, Italy. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3776574.3831145>

1 Introduction

In human dialogue, we perceive nonverbal cues such as eye contact, nodding, and facial expressions from our interlocutors and express them for smooth communication. It is expected for spoken dialogue systems and conversational avatars to express these cues appropriately to realize natural and human-like interactions. In particular, while the user is speaking, timely nonverbal listener responses such as verbal backchannels and nodding should show that the system is listening and encourage the user to continue speaking.

Recently, research on generating listening head motions for conversational avatars has been conducted [9, 13, 25], including causal real-time models [15, 27] and integrations into dialogue systems [15]. Some approaches generate motions according to the avatar's role, such as listener or speaker, by modeling turn-taking and conversational state in the architecture [5, 12]. However, many of these models do not explicitly model the timing at which the listener can respond (i.e., backchannel opportunity points (BOPs) [4] or backchannel relevance places [22]), or model it only indirectly, posing challenges for generating appropriately timed nodding backchannels.

Nodding is a nonverbal response in which participants shake their heads vertically during dialogue, primarily performed by the listener [14]. It is often performed simultaneously with verbal backchannels such as “um” or “uh-huh”, as a backchannel, and is closely related to turn-taking [14, 17, 24]. Moreover, listeners use various forms of nodding depending on the dialogue context, thereby facilitating smooth communication [1, 18, 20]. Several models have been proposed for predicting nodding in dialogue systems, including models based on sequential probabilistic models [16], models using human-human dialogue data as surrogate data [23], and models that predict three types of nodding continuously in real time [10]. However, most of these models are limited to binary



This work is licensed under a Creative Commons Attribution 4.0 International License. *ICMI '26, Napoli, Italy*
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2318-6/26/10
<https://doi.org/10.1145/3776574.3831145>

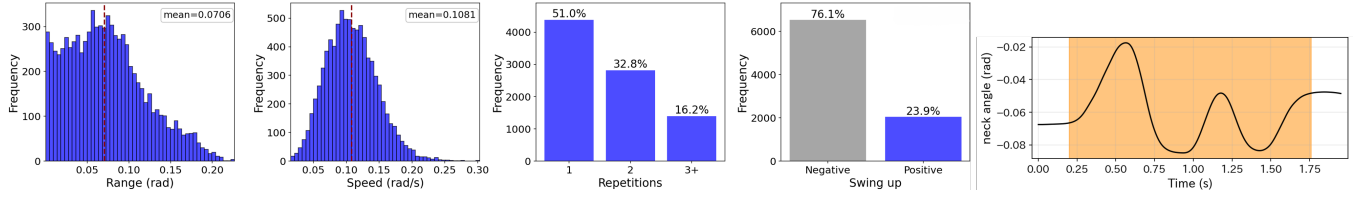


Figure 1: Distributions of kinematic parameters (left) and smoothed motion data with a detected nodding segment (right).

prediction of nodding versus not nodding, or prediction of predefined nodding classes, and systems integrating them can only replay fixed-pattern nodding rather than expressing diverse forms [8].

In this paper, we propose a model that predicts in real time both nod timing and kinematic parameters of listener nodding, such as the number of repetitions and motion range (Figure 1). To model appropriate listener nod timing, the proposed model learns the onset timing of nodding events from annotated dialogue data. In addition, by learning kinematic parameters for each ground-truth interval, the model can predict the form of nodding depending on dialogue context. To the best of our knowledge, no model has been proposed that continuously predicts both timing and kinematic parameters of nodding in real time. The diversity of nodding forms is known to be closely related to the listener’s cognitive state (e.g., understanding, agreement, notice, and surprise) and turn-taking [18, 20]. Therefore, by predicting not only the timing of nodding but also its kinematic parameters and expressing context-appropriate forms of nodding in dialogue systems, it is expected to produce human-like listener responses.

The proposed model consists of a timing prediction module and a parameter prediction module, each implementing a dyadic attention network over the speaker and listener channels based on Voice Activity Projection (VAP) [3], which models turn-taking in dyadic dialogue. Similar VAP-style dyadic architectures have been applied to real-time verbal backchannel prediction and real-time nodding prediction in prior work [6, 10]. Building on these prior studies, we present a model with the following features. First, it takes the speech signals of both the speaker and the listener as input to predict both the timing and kinematic parameters of listener nodding in an end-to-end manner. Second, the timing prediction module is trained with multi-task learning on nodding prediction, turn-taking, and verbal backchannel prediction and detection tasks. We then fine-tune the kinematic parameter module, initialized from the timing-trained module, leveraging turn-taking for appropriate nodding form. Third, instead of classifying nodding into discrete categories and replaying a matched motion, the proposed model directly predicts kinematic parameters and synthesizes diverse and human-like head motion from them. Finally, the proposed model is lightweight and capable of real-time operation, so integrating it into spoken dialogue systems or conversational avatars enables real-time generation of listener nodding.

2 Dataset

We use the attentive listening dataset collected through Wizard-of-Oz experiments with the android ERICA [11]. A trained actor

Table 1: Distribution of nodding segments

	Time (s)	Ratio (%)	Count
Nodding	15477.9	37.0	8581
No nodding	26350.9	63.0	

conducted Japanese attentive-listening dialogues with elderly participants (everyday small talk) and university students (COVID-19-related difficulties). Because the original recordings lacked diverse nodding, we annotated nodding intervals on additionally recorded listener gestures from a prior study [10] and used them for training. The dataset contains 90 dialogues averaging 8 minutes each (about 12 hours in total), and the transcripts include segment-level annotations of listener verbal backchannels.

2.1 Nodding Annotation

We detected nodding from the vertical neck angle (radians) in the motion data. The motion data were resampled to 120 Hz, smoothed using a 15-frame moving average filter and a 4th-order Butterworth low-pass filter, and nodding segments were detected using gradient-based rules. We manually corrected the detected segments by reviewing the motion data (Figure 1, right). Nodding segments account for 37.0% of the total time (Table 1). The large ratio of nodding is attributed to the nature of attentive-listening dialogues.

2.2 Kinematic Parameters

Based on previous studies that analyzed the motion features of nodding [18–20], we define the following four kinematic parameters and extract them from each detected nodding event.

- **Range (rad):** Represents the magnitude of the head movement during a nod. Computed as the absolute difference between the head angle at the segment onset and the angle at the lowest point.
- **Speed (rad/s):** Represents the velocity of head movement during a nod. Computed as the mean absolute angular velocity over the entire nodding segment.
- **Repetitions:** Represents the number of nodding cycles. A single cycle is defined as one consecutive pair of downward and upward head movements [19]. Manually annotated for each nodding event while reviewing the motion data.
- **Swing-up:** Represents whether the nod begins with an upward head movement. Labeled when its motion starts with upward movement.

Figure 1 (left) shows the distributions: mean range 0.0706 rad, mean speed 0.1081 rad/s, single-cycle nods 51.0%, two-cycle nods

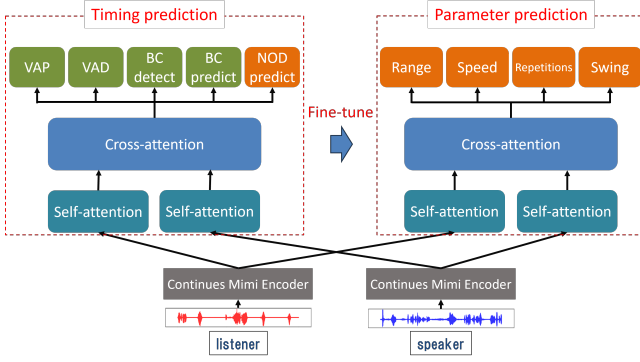


Figure 2: Architecture of the proposed model

32.8%, nods with three or more cycles 16.2%, and swing-up at the start of the motion in 23.9% of events (76.1% without).

3 Proposed Model

This section describes the proposed model for predicting both timing and kinematic parameters of listener nodding in a continuous and real-time manner. Following VAP-style multi-task learning for backchannel and nodding prediction [6, 10] and end-of-utterance prediction [7], our main extension is a dedicated module that directly predicts nodding kinematic parameters beyond timing or fixed classes. We first describe the model architecture, then explain multi-task training for timing prediction, followed by fine-tuning for kinematic parameter prediction.

3.1 Architecture

Figure 2 shows the architecture of our proposed model consisting of a timing prediction module and a kinematic parameter prediction module. The timing prediction task and the kinematic parameter prediction task differ in nature, so we allocate them to separate modules and fine-tune the parameter prediction module from the trained timing module. Each module implements a dyadic attention network over the speaker and listener channels, following VAP [3], which models turn-taking in dyadic dialogue.

We employ a component of the Mimi audio codec as an audio encoder, introduced in Moshi [2]. Specifically, we utilize the continuous features output by Mimi’s neural encoder rather than the codebook-quantized discrete tokens. This preserves causality while allowing predictions to rely on both acoustic and semantic information in the speech signal. We refer to this as the Continuous Mimi Encoder.

In the proposed model, the audio waveforms of the two dialogue participants are each encoded using the Continuous Mimi Encoder, and the output features for each channel are processed by Self-attention Transformers in the respective modules. These representations are then passed through Cross-attention Transformers, enabling reference to each other’s attention states, and finally task-specific linear layers produce the predictions.

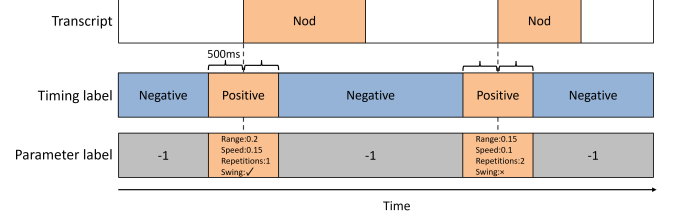


Figure 3: Definition of ground-truth labels

3.2 Multi-task Training for Timing Prediction

The timing prediction module is trained on timing-related multi-task objectives. With the nodding timing prediction loss as the primary loss, training is performed via multi-task learning with the following five tasks.

- **Nodding Prediction:** Predicts the onset timing of listener nodding at each frame. As illustrated in Figure 3, the ground-truth positive interval spans 500 ms before and after each nodding onset (1 s in total). The rationale is threefold: (1) onset timing suffices for the system to start nodding; (2) segment length varies with repetitions and speed, so each nodding event is treated equally; (3) a nod can be produced over an interval, not only at a single instant.
- **Voice Activity Projection (VAP):** Predicts the presence or absence of speech in each of four future intervals within the next 2 seconds for each dialogue participant. The intervals are 0-200 ms, 200-600 ms, 600-1200 ms, and 1200-2000 ms, forming a 2 channels $\times$ 4 intervals grid. The model outputs one of 256 classes representing the joint configuration.
- **Voice Activity Detection (VAD):** Detects the presence or absence of speech for both participants at each frame. It complements the VAP predictions.
- **Backchannel Prediction (BP):** Predicts verbal backchannel of the listener at each frame. Ground-truth labels shift verbal backchannel segments 500 ms earlier. Verbal backchannels often co-occur with nodding, so joint training should help nodding accuracy.
- **Backchannel Detection (BD):** Detects the presence or absence of verbal backchannel of the listener at each frame. Like VAD for VAP, it complements BP by marking backchannel frames.

The loss function is defined as Equation (1).

$$L = L_{nod} + w_{vap}L_{vap} + w_{vad}L_{vad} + w_{bp}L_{bp} + w_{bd}L_{bd} \quad (1)$$

where L_{nod} , L_{vap} , L_{vad} , L_{bp} , and L_{bd} are the frame-level cross-entropy losses for nodding prediction, voice activity projection, voice activity detection, backchannel prediction, and backchannel detection, respectively. The weights w_{vap} , w_{vad} , w_{bp} , and w_{bd} are hyperparameters for adjusting the weighting of loss terms, and all were set to 1 in our experiments. To address the imbalance between positive and negative samples at the frame level for backchannel and nodding, the positive weight was set to 3 during training. Training the timing prediction module also requires ground-truth verbal backchannel annotations in addition to nodding intervals.

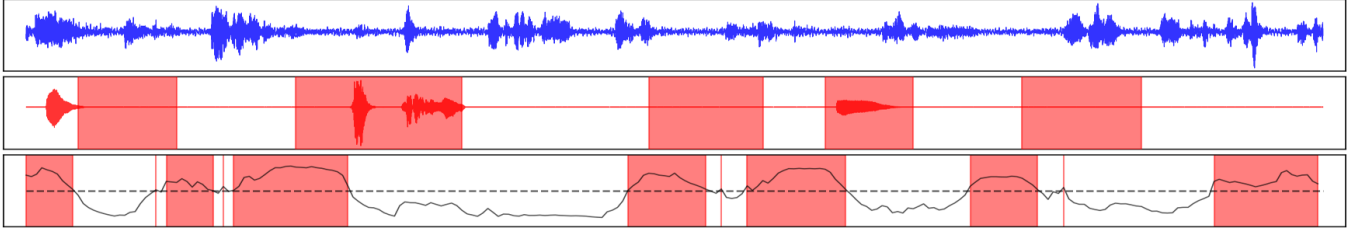


Figure 4: Example outputs of the timing prediction model. First: speaker speech waveform. Second: listener speech waveform with nodding intervals. Third: probability of nodding occurrence with predicted segments.

3.3 Fine-tuning for Kinematic Parameter Prediction

After training the timing prediction module, we fine-tune the kinematic parameter prediction module. Training is performed via multi-task learning with the following four tasks, each corresponding to one kinematic parameter. As illustrated in Figure 3, the loss for each task is computed only over the positive intervals of the nodding timing prediction task. Event-level kinematic parameters (Section 2) are assigned as labels to all frames in each positive interval.

- **Range Prediction:** Predicts the range of the head movement during nodding.
- **Speed Prediction:** Predicts the speed of the head movement during nodding.
- **Repetitions Prediction:** Predicts the number of nodding cycles. Formulated as a 3-class classification (1, 2, 3 or more).
- **Swing-up Prediction:** Predicts the presence or absence of swing-up. Formulated as a 2-class classification.

The loss function is defined as Equation (2).

$$L = w_{rp}L_{rp} + w_{sp}L_{sp} + w_{rep}L_{rep} + w_{sup}L_{sup} \quad (2)$$

where L_{rp} , L_{sp} , L_{rep} , and L_{sup} represent the losses for range prediction, speed prediction, repetitions prediction, and swing-up prediction, respectively. Range prediction and speed prediction use frame-level mean squared error (MSE), while repetitions prediction and swing-up prediction use frame-level cross-entropy. Ground-truth range and speed are z-score normalized using the mean and standard deviation computed on the training data. The weights w_{rp} , w_{sp} , w_{rep} , and w_{sup} are hyperparameters, and all were set to 1 in our experiments.

For this fine-tuning step, the Self-attention and Cross-attention layers of the kinematic parameter prediction module are initialized with the weights learned by the timing prediction module and then updated on the objectives of Equation (2). This improves kinematic parameter prediction by reusing timing features for turn-taking, verbal backchannel, and nodding onset.

4 Experimental Evaluation

We evaluated the proposed model on two frame-level nodding tasks: timing prediction and kinematic parameter prediction. For each task, the contribution of each auxiliary sub-task in multi-task training for timing prediction was examined through an ablation study. The frame rate was 12.5 Hz, the context length 20 seconds, and we used five-fold cross-validation.

Table 2: Results for timing prediction

Method	F1-score (↑)	Precision (↑)	Recall (↑)
Always	32.02±0.00	19.06±0.00	100.00±0.00
Stochastic	23.63±0.04	20.89±0.03	27.20±0.09
CPC	46.14±0.21	36.63±0.38	62.34±1.43
Proposed	52.19±0.40	41.93±0.60	69.11±0.56

The Continuous Mimi Encoder used the weights of Mimi’s Neural Encoder (ConvNet and Transformer layers), pretrained on approximately 7 million hours of speech data. These parameters were frozen during training. We also compared a CPC audio encoder, commonly adopted in the original VAP model [3] and its task-specific extensions [6, 10], with weights pretrained on approximately 60,000 hours of LibriSpeech data and kept frozen during training.

4.1 Timing Prediction

The first task is to predict whether nodding onset occurs at each frame. The evaluation metrics were frame-level precision, recall and F1-score.

The following models were compared.

- (Always) Always predicting nodding in all frames.
- (Stochastic) Predicts nodding onsets by sampling the time between consecutive nod onsets from the interval distribution fit on the training data, with each nodding segment fixed to 1 second.
- (CPC) Using CPC as the audio encoder.
- (Proposed) Using the Continuous Mimi Encoder.

The results are presented in Table 2. The proposed model outperforms both Always, Stochastic and CPC in terms of F1-score. This result indicates that the continuous features from the Mimi Audio Codec, prior to codebook quantization, are effective for predicting the onset timing of nodding.

Figure 4 shows output samples of the proposed model on the timing prediction task, demonstrating that it can effectively predict nodding onset earlier than the actual occurrence.

4.2 Kinematic Parameter Prediction

The next task is to predict the kinematic parameters of nodding. The evaluation metrics are MAE and correlation coefficient (Corr.) for range and speed, macro F1-score for repetitions, and F1-score for swing-up.

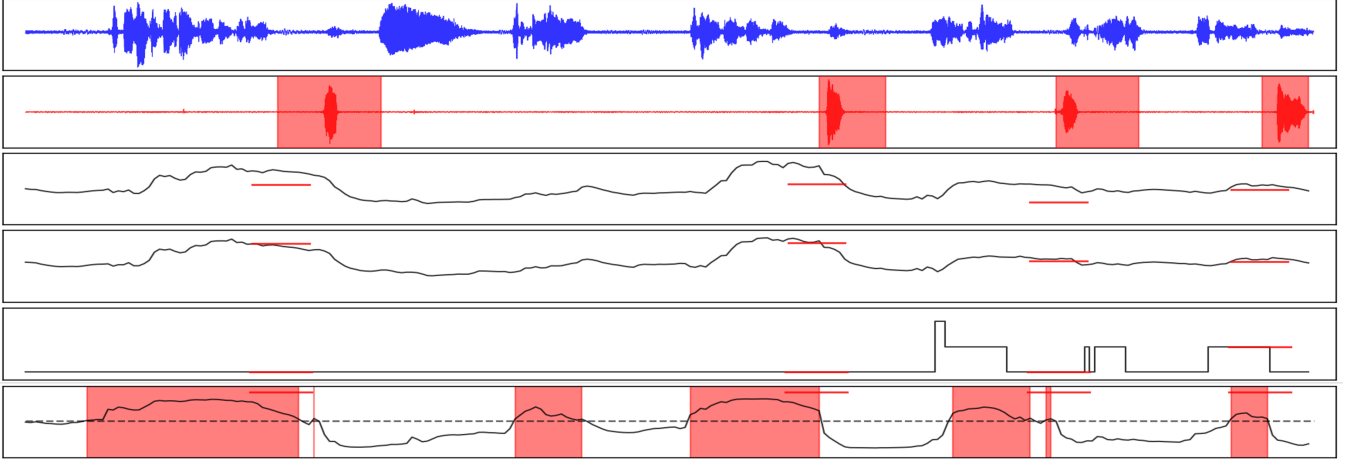


Figure 5: Example outputs of the parameter prediction model. *First row: speaker speech waveform. Second row: listener speech waveform with positive timing intervals shaded. Third–sixth rows: predicted range, speed, repetitions, and swing-up. In rows 3–6, the solid curves show model predictions and the red lines show ground truth (constant within each positive interval).*

Table 3: Results for kinematic parameter prediction

Method	Range (rad)		Speed (rad/s)		Repetitions	Swing-up
	MAE (↓)	Corr. (↑)	MAE (↓)	Corr. (↑)	Macro F1 (↑)	F1-score (↑)
Fixed	0.0395±0.0000	0	0.0605±0.0000	0	22.25±0.00	0
Stochastic	0.0545±0.0003	0.0002±0.0000	0.0848±0.0005	0.0003±0.0000	33.30±0.01	24.10±0.11
SVM	F0+loud. 0.0377±0.0001	0.3314±0.0096	0.0571±0.0002	0.3565±0.0080	36.92±0.75	42.69±0.39
	Mimi 0.0367±0.0002	0.4083±0.0061	0.0559±0.0001	0.4103±0.0072	38.31±0.38	46.00±0.66
CPC (Finetune)	0.0370±0.0004	0.3774±0.0190	0.0568±0.0007	0.3851±0.0157	37.49±0.43	49.47±0.27
Proposed	Finetune 0.0341±0.0003	0.5215±0.0127	0.0525±0.0003	0.5236±0.0111	39.37±0.89	53.45±0.72
	Scratch 0.0350±0.0004	0.4830±0.0224	0.0539±0.0008	0.4835±0.0245	37.22±1.01	50.79±0.34

In this task, we evaluated the proposed model and CPC, as well as support vector machine (SVM) baselines. Support vector classification (SVC) was used for discrete tasks and support vector regression (SVR) for continuous-value tasks, both with the radial basis function (RBF) kernel, trained to predict the kinematic parameters of each nodding segment from the speech features of both speaker and listener over the preceding 2.5 seconds.

The following models were evaluated.

- (Fixed) Predicting the fixed training-data mean for range and speed (0.06946 rad and 0.1046 rad/s); the fixed training-data modes for repetitions and swing-up (one repetition, no swing-up).
- (Stochastic) Sampling kinematic parameters from their joint training distribution.
- SVM
 - (F0+loud.) Using frame-level (100 Hz) F0 and loudness.
 - (Mimi) Using Continuous Mimi Encoder features.
- (CPC) Using CPC as the audio encoder while training the timing prediction model and fine-tuning for parameter prediction.
- Proposed
 - (Finetune) Initializing from the timing-trained module followed by fine-tuning for kinematic parameter prediction.

- (Scratch) Training the kinematic parameter prediction task from scratch.

The results are presented in Table 3. Across all kinematic parameter prediction tasks, the proposed model outperforms Fixed, Stochastic, SVM, and CPC. Furthermore, fine-tuning from the timing-trained module for kinematic parameter prediction substantially improved accuracy compared to training from scratch. This indicates that features related to turn-taking, verbal backchannel, and nodding timing prediction are effective for predicting the form of nodding. Compared with SVM with Mimi features, CPC scored lower on range, speed, and repetitions but higher on swing-up, suggesting that the proposed model’s turn-taking structure contributes more to swing-up prediction accuracy.

Figure 5 shows output samples of the proposed model on the parameter prediction task, demonstrating that it can appropriately predict kinematic parameters continuously.

4.3 Ablation Study

We also examined how much each auxiliary sub-task in multi-task timing training contributes to the two main tasks. Table 4 presents the ablation results. Removing the VAP loss lowered scores across

Table 4: Ablation study: scores for timing and kinematic parameter prediction tasks

	Timing	Range (rad)		Speed (rad/s)		Repetitions	Swing-up
	F1 ($\uparrow$)	MAE ($\downarrow$)	Corr. ($\uparrow$)	MAE ($\downarrow$)	Corr. ($\uparrow$)	Macro F1 ($\uparrow$)	F1 ($\uparrow$)
Original	52.19 $\pm$ 0.40	0.0341 $\pm$ 0.0003	0.5215 $\pm$ 0.0127	0.0525 $\pm$ 0.0003	0.5236 $\pm$ 0.0111	39.37 $\pm$ 0.89	53.45 $\pm$ 0.72
w/o VAD	52.16 $\pm$ 0.21	0.0341 $\pm$ 0.0004	0.5234 $\pm$ 0.0160	0.0526 $\pm$ 0.0003	0.5274 $\pm$ 0.0097	39.19 $\pm$ 0.70	53.28 $\pm$ 0.46
w/o VAP	51.51 $\pm$ 0.12	0.0345 $\pm$ 0.0004	0.5034 $\pm$ 0.0090	0.0529 $\pm$ 0.0007	0.5106 $\pm$ 0.0126	39.30 $\pm$ 0.75	52.17 $\pm$ 0.84
w/o BD	52.25 $\pm$ 0.30	0.0339 $\pm$ 0.0005	0.5278 $\pm$ 0.0176	0.0526 $\pm$ 0.0004	0.5229 $\pm$ 0.0175	39.48 $\pm$ 0.64	52.87 $\pm$ 0.55
w/o BP	51.85 $\pm$ 0.37	0.0338 $\pm$ 0.0002	0.5275 $\pm$ 0.0053	0.0525 $\pm$ 0.0004	0.5228 $\pm$ 0.0080	40.36 $\pm$ 0.56	52.58 $\pm$ 0.58

Table 5: Real-time processing performance

	Avg. inference time (s)	RTF
w/o Mimi ONNX	0.05338	0.6673
w/ Mimi ONNX	0.04568	0.5711

all tasks from the Original model, suggesting that modeling turn-taking structure is critical for both timing and parameter prediction. Relative to other parameters, swing-up prediction showed clearer F1 decreases when either the verbal backchannel prediction (BP) or detection (BD) loss was removed (from 53.45 to 52.58 and 52.87, respectively). A possible reason for the sensitivity of swing-up to backchannel modeling is the close relationship between the timing of the listener’s cognitive shift, indicated by swing-up (e.g., reception of new information, surprise) [20], and the timing of verbal backchannel.

4.4 Real-time Processing Performance

The real-time processing performance of the proposed model was also evaluated. Because the latency of the model is mostly caused by the Continuous Mimi Encoder, we created an ONNX (Open Neural Network Exchange) format (fp32) version of the Continuous Mimi Encoder for real-time processing. Assuming deployment on laptop-class devices without GPUs, we measured the average inference time and the real-time factor (RTF) during inference on a CPU (Intel Core Ultra 9 275HX) over a 2-minute period.

The results are presented in Table 5. The proposed model achieved an RTF of 0.67 without ONNX and 0.57 with ONNX. Therefore, the proposed model is capable of real-time processing and, once integrated into avatar dialogue systems and conversational robots, can generate nodding in real time.

5 Subjective Evaluation

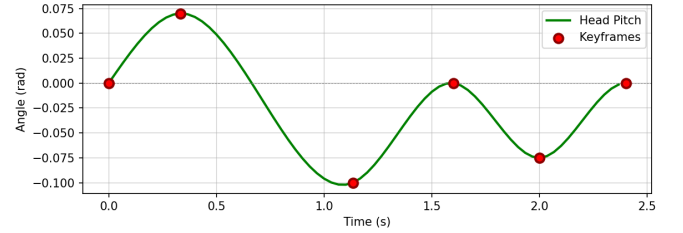
We integrated the proposed model into an avatar dialogue system with a CG agent¹ (Figure 7). In this system, diverse forms of nodding can be generated in real time in response to user utterances. We conducted a subjective evaluation to investigate whether generating nodding with the proposed model gives a better impression to users compared to conventional methods.

5.1 Method

We evaluated the five methods in Table 6; **Stochastic** and **Fixed** follow the same timing and parameter baselines as in Section 4.

Table 6: Methods evaluated in subjective experiment

Method	Timing	Parameters
A) No Nodding	–	–
B) Stochastic + Fixed	Stochastic	Fixed
C) Proposed + Fixed	Proposed	Fixed
D) Proposed + Stochastic	Proposed	Stochastic
E) Proposed + Proposed	Proposed	Proposed

**Figure 6: Nodding motion generation from predicted kinematic parameters via cubic spline interpolation**

In Methods D and E, at runtime (Figure 7), when the nodding probability exceeds a threshold continuously for a fixed number of frames (e.g., $K=5$ frames at 12.5 Hz, equivalent to 400 ms), we average the parameter predictions over those frames to obtain the event-level parameter set. A full nodding trajectory satisfying these parameters is then synthesized by rule-based cubic-spline interpolation (Figure 6) and sent to the avatar. Because the positive interval is defined as ± 500 ms around each onset, the model predicts nods sufficiently earlier than their actual occurrence (Figure 4), so requiring a short run of positive frames yields stable detection without perceptible delay.

We recorded 45 avatar videos by passing nine sessions randomly chosen from held-out test sessions (five-fold CV, Section 4) through each of five variants (9 sessions $\times$ 5 methods).

The nine sessions were divided into 3 groups, and 30 crowd-sourcing workers were recruited per group, totaling 90 participants. Each participant watched 15 videos (3 sessions $\times$ 5 methods) plus one dummy video for attention check (16 total). For this attention check, the last 5 seconds of the dummy video instructed participants to rate all items as 1, and we excluded participants who did not follow the instruction. Each participant evaluated each video on the following 7 metrics.

- (1) How human-like is the avatar’s response? (**human likeness**)

¹CG-CA Gene (c) 2023 by Nagoya Institute of Technology, Moonshot R&D Goal 1 Avatar Symbiotic Society. <https://github.com/mmdagent-ex/gene>

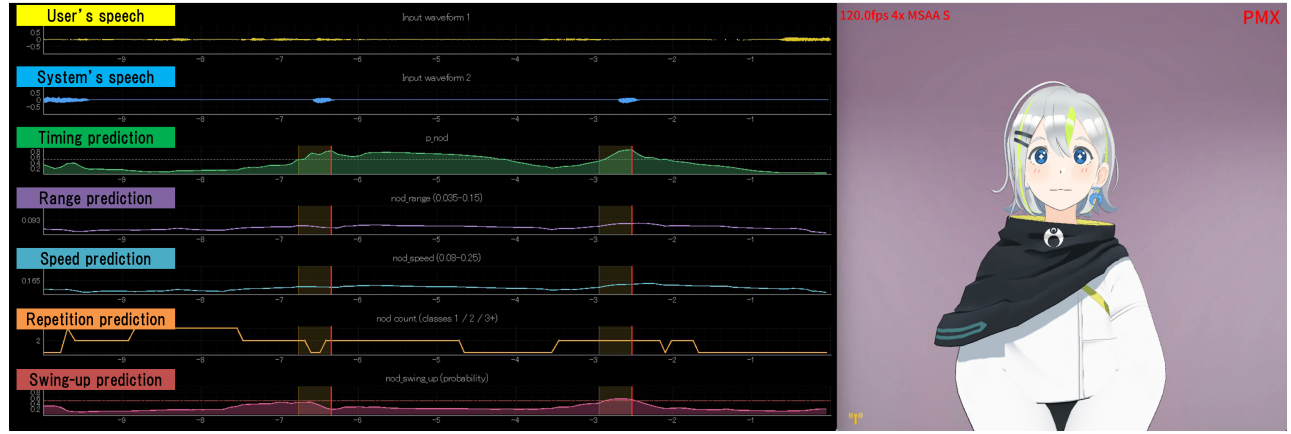


Figure 7: Avatar dialogue system integrated with the proposed model

Table 7: Averaged scores of subjective evaluation ($n=60$). Hum: human likeness; Nat: naturalness; Div: diversity; Att: attentiveness; Fac: facilitation; Und: understanding; Emp: empathy. Method names match Table 6.

Method	Response			Avatar			
	Hum	Nat	Div	Att	Fac	Und	Emp
A) No nodding	(1.64)	(1.59)	(1.09)	1.94	1.13	1.87	1.41
B) Stochastic + Fixed	4.19	4.06	3.11	4.34	3.73	4.23	4.08
C) Proposed + Fixed	4.53	4.52	2.94	4.98	4.26	4.72	4.51
D) Proposed + Stochastic	<u>5.38</u>	<u>5.24</u>	5.08	<u>5.66</u>	5.05	<u>5.43</u>	<u>5.44</u>
E) Proposed + Proposed	<u>5.46</u>	<u>5.42</u>	<u>4.87</u>	<u>5.69</u>	<u>5.02</u>	<u>5.45</u>	<u>5.49</u>

- (2) How natural is the avatar's response? (**naturalness**)
- (3) How varied (non-monotonous) is the response? (**diversity**)
- (4) How well does the avatar appear to be listening to the user? (**attentiveness**)
- (5) How well does the avatar encourage continued speaking? (**facilitation**)
- (6) How well does the avatar appear to understand? (**understanding**)
- (7) How well does the avatar appear to empathize? (**empathy**)

Each metric was scored on a 7-point scale, videos were shown in random order, and the experiment was conducted in Japanese.

5.2 Results and Analysis

Each participant evaluated three videos per method, and the per-method score for each metric was computed by averaging these three ratings. After the attention check, $n=60$ participants remained. Tables 7 and 8 present the mean scores and Bonferroni-corrected post-hoc pairwise comparisons, respectively. A mixed-design ANOVA with group (3 levels) as a between-subjects factor and method (5 levels) as a within-subjects factor revealed a significant main effect of the method for all seven metrics (Greenhouse-Geisser corrected $p < .001$; partial $\eta^2 \approx .78-.85$). The main effect of the group was not significant for any metric ($p > .14$). The group $\times$ method interaction was significant for naturalness ($p = .002$), attentiveness ($p = .042$), and understanding ($p = .043$).

Table 8: Post-hoc pairwise comparison p -values (Bonferroni corrected; $n=60$). Column abbreviations as in Table 7.

Comparison	Hum	Nat	Div	Att	Fac	Und	Emp
B vs. C	.030*	.023*	1.00	<.001**	.002**	.002**	.002**
C vs. E	<.001**	<.001**	<.001**	<.001**	<.001**	<.001**	<.001**
D vs. E	1.00	.205	.742	1.00	1.00	1.00	1.00
						* $p<0.05$	** $p<0.01$

Across all metrics, Methods D and E, which predict nodding timing with the proposed model and generate diverse forms of nodding, tended to achieve the highest scores. Comparing B and C, proposed timing significantly outperformed stochastic timing on all metrics except diversity. In particular, a large difference was observed in attentiveness, indicating that proposed nodding-onset timing prediction enhances the attentiveness of dialogue systems. Comparing C and E, generating diverse forms of nodding yielded significantly higher scores than generating fixed nodding on all metrics. This indicates that a conversational avatar expressing diverse forms of nodding according to context can behave more naturally and human-like, and can better express understanding and empathy. Comparing D and E, diversity was slightly higher for Method D and naturalness slightly higher for Method E descriptively, but no significant differences were found on any metric. These results demonstrate that, by predicting both timing and kinematic parameters of nodding in real time with the proposed model and expressing diverse forms of nodding at appropriate timing, it is possible to enhance the naturalness and human-likeness of the responses and to convey attentive listening, understanding, and empathy.

5.3 Qualitative Analysis

We conducted a qualitative analysis of nodding behavior under Methods D and E, for which the subjective evaluation found no significant differences on any metric. In the example in Figure 8, the nodding sequences generated by the proposed method (Method E) exhibit the following tendencies:

- When the speaker's talk is likely to continue, a single small nod is produced as a continuer.

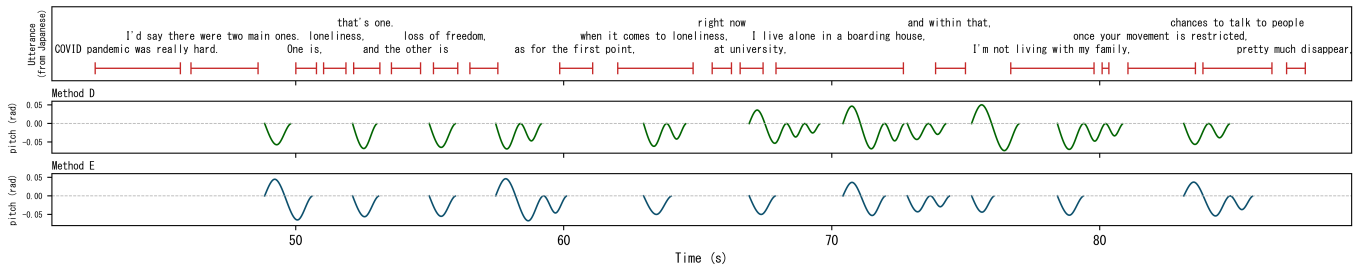


Figure 8: Comparison of nodding motions generated by Method D (stochastic parameters) and Method E (proposed): utterance timeline and head pitch (rad) over time

- When the talk appears to be ending, or when the user presents new information, swing-up tends to accompany the nod, and the range is larger.

Specifically, in Method E, while the speaker describes two main difficulties during the pandemic (about 52-55 s), small single nods continued as continuers. After both “loneliness” and the second difficulty “loss of freedom” had been introduced, a relatively large two-cycle nod with swing-up occurred around 57 s in response to that new information. During the subsequent explanation of loneliness (about 63-68 s), small single nods continued, and around 70 s a swing-up nod appeared when the speaker disclosed living alone. Around 82 s, a swing-up nod followed talk of loneliness causes (e.g., “I’m not living with my family”, “once your movement is restricted.”). This can be viewed as partial uptake or alignment while the speaker explains why loneliness arises. These patterns align with prior analyses that listener swing-up nods often accompany understanding of the speaker’s intent, receipt of new information, or a shift in cognitive state [20, 21, 26]. By contrast, under Method D, large swing-up nods occurred around 67 and 75 s at moments when the topic still seemed likely to continue, which appears less appropriate for a listener in context.

Such characteristics of the proposed model were not clearly reflected in the subjective evaluation results. This is likely because participants watched pre-recorded videos as third-party observers rather than interacting with the avatar directly. In such a format, subtle qualities may be difficult to assess, in particular whether the avatar appears to understand the speaker and respond with empathy. Additionally, since five methods were evaluated simultaneously, the overall evaluation may have been dominated by factors that differ substantially across methods, such as how diverse nodding motions are in each condition and whether timing is appropriate or not. To more clearly demonstrate that the proposed model produces nodding forms that match the conversational content, further experiments such as paired comparisons in face-to-face interactions are needed.

6 Conclusion

In this paper, we have proposed a model that predicts in real time both the timing and kinematic parameters of nodding, an important nonverbal listener response, from the speech signals of both the speaker and the listener. Each of the timing and parameter modules employs a dyadic attention network over the speaker and listener channels. We fine-tune the kinematic parameter prediction

module from the trained timing prediction module. As a result, we achieved an F1 score of 52.19 on timing prediction and, on kinematic parameter prediction, correlations of 0.5215 and 0.5236 for range and speed, macro F1 39.37 for repetitions, and F1 53.45 for swing-up. Furthermore, the effectiveness of this fine-tuning procedure was demonstrated. In addition, the ablation study showed that multi-task learning with the VAP task is important for timing prediction, and that auxiliary tasks related to verbal backchannels also influence swing-up prediction accuracy. We showed that the proposed model is capable of real-time processing and integrated it into avatar dialogue systems and conversational robots. Subjective evaluation experiments demonstrated that the proposed method significantly outperformed both the stochastic timing baseline and the fixed-motion nodding baseline.

Future work includes advancing from predicting nodding kinematic parameters to generating natural nodding motions in a context-aware manner from kinematic parameters, and building models that predict listener nonverbal responses using not only speech but also the speaker’s visual nonverbal cues (e.g., eye gaze, nodding) during dialogue.

Safe and Responsible Innovation Statement

Our models estimate listener nodding from conversational speech to drive attentive-listening interfaces. We expect a relatively small misuse footprint compared with unconstrained generative systems, but interfaces must not suggest that a live human steers nodding when it is model-generated, including claims that an avatar or robot is entirely human-operated. Speech materials and derived features are handled under local access controls. The dialogue corpus was collected with informed consent for research use and sharing, and released recordings exclude directly identifying items.

Acknowledgments

This work was supported by JST Moonshot R&D JPMJPS2011 and JST PRESTO JPMJPR24I4. K. Kato was a scholarship recipient of the Iwaware Scholarship Foundation in fiscal year 2025.

References

- [1] Peter Blomsma, Julija Vaitonytė, Gabriel Skantze, and Marc Swerts. 2024. Backchannel behavior is idiosyncratic. *Language and Cognition* 16, 4 (2024), 1158–1181. doi:10.1017/langcog.2024.1
- [2] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037* (2024).

- [3] Erik Ekstedt and Gabriel Skantze. 2022. Voice Activity Projection: Self-supervised Learning of Turn-taking Events. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*. 5190–5194.
- [4] Jonathan Gratch, Anna Okhmatovskaia, François Lamothe, Stacy Marsella, Mathieu Morales, R. J. van der Werf, and Louis-Philippe Morency. 2006. Virtual Rapport. In *Intelligent Virtual Agents (Lecture Notes in Computer Science, Vol. 4133)*. Springer, Berlin, Heidelberg, 14–27. doi:10.1007/11821830_2
- [5] Ying Guo, Xi Liu, Cheng Zhen, Pengfei Yan, and Xiaoming Wei. 2025. ARIG: Autoregressive Interactive Head Generation for Real-Time Conversations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 12956–12965.
- [6] Koji Inoue, Divesh Lala, Gabriel Skantze, and Tatsuya Kawahara. 2024. Yeah, Un, Oh: Continuous and Real-time Backchannel Prediction with Fine-tuning of Voice Activity Projection. *arXiv preprint arXiv:2410.15929* (2024).
- [7] Ryo Ishii, Shin-ichiro Eitoku, Ryota Yokoyama, and Junichi Sawase. 2025. Predicting End-of-turn and Backchannel Based on Multimodal Voice Activity Prediction Model. In *Proceedings of the 27th ACM International Conference on Multimodal Interaction*. ACM, Canberra, ACT, Australia, 446–455. doi:10.1145/3716553.3750781
- [8] Vidit Jain, Maitree Leekha, Rajiv Ratn Shah, and Jainendra Shukla. 2021. Exploring Semi-Supervised Learning for Predicting Listener Backchannels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, USA, Article 395, 12 pages. doi:10.1145/3411764.3445449
- [9] Siyeol Jung and Taehwan Kim. 2025. DiffListener: Discrete Diffusion Model for Listener Generation. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5.
- [10] Kazushi Kato, Koji Inoue, Divesh Lala, Keiko Ochi, and Tatsuya Kawahara. 2025. Real-time Generation of Various Types of Nodding for Avatar Attentive Listening System. In *Proceedings of the 27th ACM International Conference on Multimodal Interaction*. ACM, Canberra, ACT, Australia, 209–217. doi:10.1145/3716553.3750764
- [11] Tatsuya Kawahara. 2019. Spoken dialogue system for a human-like conversational robot ERICA. In *9th International Workshop on Spoken Dialogue System Technology (IWSDS)*. 65–75.
- [12] Taekyung Ki, Sangwon Jang, Jaehyeon Jo, Jaehong Yoon, and Sung Ju Hwang. 2026. Avatar Forcing: Real-Time Interactive Head Avatar Generation for Natural Conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [13] Miao Liu, Jing Wang, Xinyuan Qian, and Haizhou Li. 2024. ListenFormer: Responsive Listening Head Generation with Non-autoregressive Transformers. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM Multimedia)*. 7094–7103.
- [14] Senko K. Maynard. 1987. Interactional functions of a nonverbal sign: Head movement in Japanese dyadic casual conversation. *Journal of Pragmatics* 11, 5 (1987), 589–606. doi:10.1016/0378-2166(87)90181-0
- [15] Tamon Mikawa, Yasuhisa Fujii, Yukoh Wakabayashi, Kengo Ohta, Ryota Nishimura, and Norihide Kitaoka. 2024. Listening Head Motion Generation for Multimodal Dialog System. In *11th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*. 1–6.
- [16] Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. 2010. A Probabilistic Multimodal Approach for Predicting Listener Backchannels. *Autonomous Agents and Multi-Agent Systems* 20, 1 (2010), 70–84. doi:10.1007/s10458-009-9092-y
- [17] Taiga Mori and Yasuharu Den. 2020. Co-occurrence Relations between Forms of Response Tokens and Nods. *Special Interest Group on Spoken Language Understanding and Dialogue Processing (SIG-SLUD)* 90 (2020), 140–145. (in Japanese).
- [18] Taiga Mori and Yasuharu Den. 2021. Relations between Forms of Response Tokens and Physical Features of Nods. *Special Interest Group on Spoken Language Understanding and Dialogue Processing (SIG-SLUD)* 91 (2021), 62–67. (in Japanese).
- [19] Taiga Mori, Yasuharu Den, and Kristiina Jokinen. 2025. Structure of nods in conversation. *PLOS ONE* 20, 5 (2025), e0323448. doi:10.1371/journal.pone.0323448
- [20] Taiga Mori, Kristiina Jokinen, and Yasuharu Den. 2022. Cognitive States and Types of Nods. In *Proceedings of the 2nd Workshop on People in Vision, Language, and the Mind*. European Language Resources Association, Marseille, France, 17–25. <https://aclanthology.org/2022.pvlam-1.4/>
- [21] Costanza Navarretta, Elisabeth Ahlsén, Jens Allwood, Kristiina Jokinen, and Patrizia Paggio. 2012. Feedback in Nordic first-encounters: A comparative study. In *LREC*. 2494–2499.
- [22] Hiroaki Noguchi, Yasuhiro Katagiri, and Yasuharu Den. 2001. *A Proposal of Building a BRP-Likelihood Corpus*. Technical Report SIG-SLUD-A101. Japanese Society for Artificial Intelligence. 25–32 pages. (in Japanese).
- [23] Shohei Noguchi, Yutaka Nakamura, and Yuya Okadome. 2024. Development of an Attentive Listening Robot Using the Motion Prediction Based on Surrogate Data. In *HCI International 2024 Posters: 26th International Conference on Human-Computer Interaction (HCII 2024) (Communications in Computer and Information Science, Vol. 2114)*. Springer Nature, Cham, 387–394. doi:10.1007/978-3-031-61932-8_44 Conference held in Washington, DC, USA..
- [24] Kazuhiro Otsuka and Masahiro Tsumori. 2020. Analyzing Multifunctionality of Head Movements in Face-to-Face Conversations Using Deep Convolutional Neural Networks. *IEEE Access* 8 (2020), 217169–217195. doi:10.1109/ACCESS.2020.3041672
- [25] Maksim Siniukov, Di Chang, Minh Tran, Hongkun Gong, Ashutosh Chaubey, and Mohammad Soleymani. 2025. DiTaiListener: Controllable High Fidelity Listener Video Generation with Diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 11991–12001.
- [26] Kevin A Whitehead. 2011. Some uses of head nods in “third position” in talk-in-interaction. *Gesture* 11, 2 (2011), 103–122.
- [27] Sizhe Zhao, Chenyang Wang, Weiyu Zhao, Zonglin Li, Ming Li, and Shengping Zhang. 2025. REA-Listener: Real-Time Listening Head Generation with Dynamic Emotion Modeling and Flexible Modality Adaptation. In *Proceedings of the 33rd ACM International Conference on Multimedia*. ACM, New York, NY, USA, 9733–9742.