International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02300-7

Beyond Talking – Generating Holistic 3D Human Dyadic Motion for Communication

Mingze Sun[1] · Chao Xu[2] · Xinyu Jiang[1] · Yang Liu[2] · Baigui Sun[2] · Ruqi Huang[1]

Received: 29 March 2024 / Accepted: 3 November 2024

Abstract
In this paper, we introduce an innovative task focused on human communication, aiming to generate 3D holistic human motions for both speakers and listeners. Central to our approach is the incorporation of factorization to decouple audio features and the combination of textual semantic information, thereby facilitating the creation of more realistic and coordinated movements. We separately train VQ-VAEs with respect to the holistic motions of both speaker and listener. We consider the real-time mutual influence between the speaker and the listener and propose a novel chain-like transformer-based auto-regressive model specifically designed to characterize real-world communication scenarios effectively which can generate the motions of both the speaker and the listener simultaneously. These designs ensure that the results we generate are both coordinated and diverse. Our approach demonstrates state-of-the-art performance on two benchmark datasets. Furthermore, we introduce the HoCo holistic communication dataset, which is a valuable resource for future research. Our HoCo dataset and code will be released for research purposes upon acceptance.

Keywords Dyadic motion · Holistic human mesh · Communication

1 Introduction

Boosted by large-scale video data of human talking, recent approaches have made significant progress in the task of speech-to-motion, namely, generating non-verbal signals within a talk, such as human facial expressions, body poses, and hand gestures, from verbal cues (e.g., audio clips or transcript). Such progress can help artificial intelligence agents understand human behavior within a speech, and therefore contributes to practical applications in healthcare, virtual reality (VR), and human-robot interaction (HRI), to name a few.

However, the prior works all concentrate on playing a *single role*, either speaker or listener, in a talk. For the former, there have been works on the alignment of facial and lip movements with audio (Ye et al., 2023; Zhang et al., 2023c; Guo et al., 2021; Xu et al., 2023) as shown in Fig. 1a, realistic gesture generation (Ao et al., 2022; Zhi et al., 2023; Ahuja et al., 2023) in Fig. 1b, as well as the synthesis of holistic 3D human motion for speakers (Habibie et al., 2021; Yi et al., 2023) in Fig. 1c; For the latter, the recent advances have been primarily devoted to generating facial expression reactions (Zhou et al., 2022; Song et al., 2023a; Liu et al., 2023) in Fig. 1d and the body poses of the listeners (Tuyen et al., 2022; Tuyen and Celiktutan, 2023) in Fig. 1e. To sum up, these works cannot directly generate holistic 3D mesh motions for *both speaker and listener simultaneously* based on audio alone.

Communicated by Shengfeng He.

Mingze Sun, Chao Xu and Xinyu Jiang contributed equally to this work.

✉ Ruqi Huang
ruqihuang@sz.tsinghua.edu.cn

Mingze Sun
smz22@tsinghua.edu.cn

Chao Xu
xc264362@alibaba-inc.com

Xinyu Jiang
xy-jiang23@tsinghua.edu.cn

Yang Liu
ly261666@alibaba-inc.com

Baigui Sun
baigui.sbg@alibaba-inc.com

[1] Tsinghua Shenzhen International Graduate School, Shenzhen, China

[2] Alibaba Group, Hangzhou, China

Published online: 17 December 2024

On the other hand, it has long been recognized that communication is of significance in social interaction between humans. Early research (Jocelyn Scheirer, 1999; Birdwhistell, 1952) even argues that over 80% of human communication is encoded in facial expressions and body movements. In fact, participants of a conversation can simultaneously convey and perceive information throughout, no matter whether he/she is talking or not. For instance, when a speaker says something confusing, a listener may have a furrowed brow (facial expression) and/or rub head with fingers (body movement). The speaker can then adjust his/her expression by noticing the listener's reaction. It is then evident that the ability of *simultaneously* generating motions for speaker and listener is critical in building interactive agents in practice. To be more concrete, an agent fully devoted to playing speaker or listener can not adapt to dynamic changes of role in an interactive conversation with human users. Meanwhile, switching between two independent models (w.r.t speaker and listener) can pose difficulty in maintaining character consistency and communication continuity. From this point of view, the prior works, which all generate non-verbal signals of a fixed role, fall short of taking into consideration the instantaneous mutual influences among participants. As a consequence, the agent trained in this single-role manner lacks the potential of *real-time interacting* with other agents, let alone humans in real life.

In this paper, we present a significant advancement towards the generation of holistic motions for *both* speaker and listener. As shown in Fig. 1, our approach significantly differs from the single-role based generative methods. More specifically, given a talk between two participants, we take consideration of the interaction in between and aim to simultaneously generate holistic 3D mesh sequences including facial expressions, body poses, and hand gestures, for both speaker and listener. Figure 2 shows an example that our method is capable of generating diverse and coordinated communication during a conversation between two participants switching roles.

To achieve this goal, we first propose HoCo dataset, which provides Holistic description of Communication within talks, including verbal (e.g., audio and transcript) and non-verbal signals (e.g., holistic body language) of both speakers and listeners. More concretely, we collect 22, 913 pieces of video clips, whose total duration is 45 hours. We further make our effort to provide comprehensive annotations for both speakers and listeners within the video clips including audio, text aligned with the speech, and the pseudo-labels for SMPL-X.

On the front of modeling, our novel task requires effective modeling of not only each role but also their mutual influence. For the former, we propose to enhance our model by introducing proper factorization of the verbal input, which in turn reduces the complexity of the latent space and allows for more fine-grained control over generation. The existing methods (Kucherenko et al., 2020; Bhattacharya et al., 2021) encode information from both audio and corresponding text. However, the features have not been decoupled based on the influence of facial expressions and body motions. For instance, the shape of the speaker's mouth should be related to the content and style of the input audio, while body motions should consider more of the style of the audio and the semantic information of the corresponding text. In contrast to prior methods, we resort to the recent advance in pre-training audio feature factorization (Li et al., 2022), which encodes input audio into features consisting of three components in correspondence to attributes including energy, pitch, and style. Beyond the audio rhythm and style, since human motions are related to the semantic information of the text and our dataset provides accompanied texts with respect to the verbal input, we further enrich the encoding by appending the respective features extracted by the pre-trained NLP model of (Devlin et al., 2018).

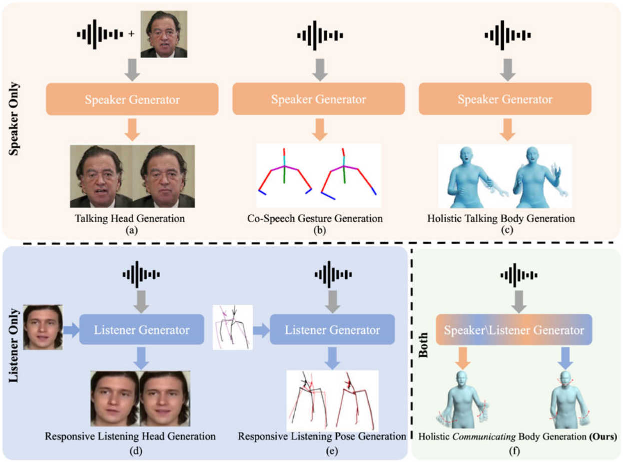

Speaker Only
Speaker Generator
Speaker Generator
Speaker Generator
Talking Head Generation
(a)
Co-Speech Gesture Generation
(b)
Holistic Talking Body Generation
(c)
Listener Only
Listener Generator
Listener Generator
Responsive Listening Head Generation
(d)
Responsive Listening Pose Generation
(e)
Both
Speaker\Listener Generator
Holistic *Communicating* Body Generation **(Ours)**
(f)

Fig. 1 Illustrations of two related tasks and our proposed holistic communicating body generation. Top row: Generation of the head (Zhang et al., 2023c), gesture (Zhi et al., 2023), or holistic body (Yi et al., 2023) for the speaker from speech; Bottom left: Responsive listening head synthesizes videos in responding to the speaker video stream (Zhou et al., 2022) and Responsive listening pose generation based on predicted pose history (Ahuja et al., 2019); Bottom right: Our holistic communicating body generation from speech. We generate 3D holistic motions for both speakers and listeners simultaneously

For the latter, while it seems plausible to combine the existing works on speakers and that on listeners to generate non-verbal interaction, we argue that this naive approach can not fully exploit the interaction presented in the data and leads to sub-optimal results in the end. That is because a) works on speaker (Yi et al., 2023; Yoon et al., 2020; Liu et al., 2022a; Guo et al., 2021) typically ignores listener; b) works on listener (Zhou et al., 2022; Song et al., 2023a; Liu et al., 2023) most require the ground-truth annotations of speakers during inference, which is inconvenient but also

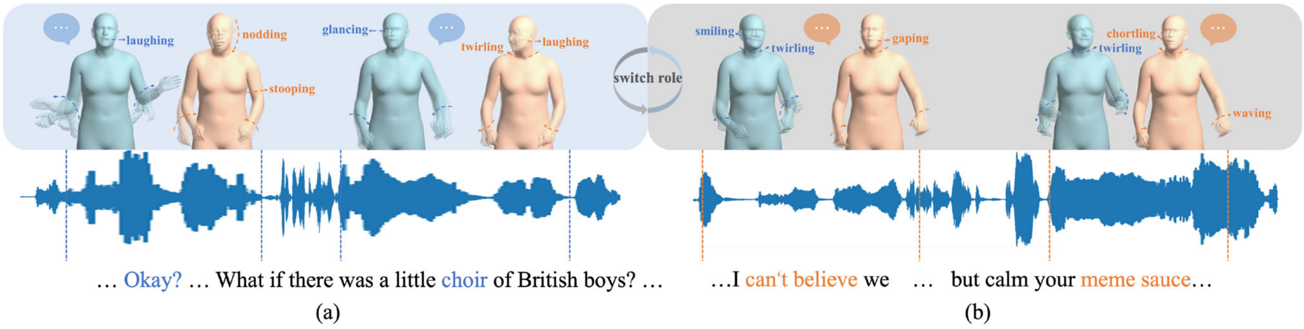

laughing
nodding
stooping
glancing
twirling
laughing
switch role
smiling
twirling
gaping
chortling
twirling
waving
… Okay? … What if there was a little choir of British boys? …
…I can't believe we … but calm your meme sauce…
(a)
(b)

Fig. 2 Holistic communicating body generation example. Two characters with fixed spatial relationships are identified with different colors. Their roles, speaker/listener, are exchanged within a conversation, where the speaker is indicated by the speech balloon regarding color. Given the conversation audio, our method can generate coordinated and diverse communication. **a** The left character is speaking. The listener laughs in response to the speaker's joke, accompanied by changes in body posture. **b** The left character switches to act as a listener. As the speaker narrates an unusual event, the listener expresses surprise with raised gestures and facial expressions in sync with the speaker

favors speakers' impact over listeners much more than the opposite.

In contrast, our method considers the real-time mutual influence between the speaker and the listener. We first separately train VQ-VAEs (Van Den Oord et al., 2017) with respect to the body language of both speaker and listener. Noting that, typically the speaker plays more of a leading role in a conversation, as the listener needs to first perceive the verbal and non-verbal message from the speaker and then react. To better model such a subtle difference, we propose an auto-regressive transformer model with a tailored chain-like structure design. At each moment, we first predict the motion of the speaker and then the listener, forming a chain of mutual interactions over time. More concretely, we condition the generation of the body language of the speaker at frame $T + 1$ on the factorized features and the generated frames from frame 1 to frame T of both speaker and listener. The counterpart for the listener at frame $T + 1$ is similarly conditioned but with an extra input of the speaker at frame $T + 1$.

We demonstrate both quantitative and qualitative experimental comparisons with respect to the prior state-of-the-art. Remarkably, our model generates more coordinated and diverse single-role body language, and we achieve a 27.6% improvement in Frechet Gesture Distance (FGD) and a 46.2% improvement in Variation compared to (Yi et al., 2023). We also significantly outperform the baselines in the more challenging communication generation between two participants, which is benchmarked on our proposed dataset.

To summarize, our contributions are as follows:

- We present HoCo dataset, which comprises high-definition RGB videos of communication within human interaction. The dataset spans 45 h and includes 22,913 video clips, accompanied by multi-modal corresponding information. We have also generated pseudo-labels for SMPL-X in generating full-body movements for both speakers and listeners.
- We design a new audio decoupling method that incorporates text information as conditions for generating both speaker and listener motions. The decoupled features include content, style, and semantic information, which correspond to the control of generating expressions and body motions. This results in motions that exhibit stronger consistency with the audio phonologically and semantically.
- We devise a framework for the chained generation of speaker and listener real-time interactions, allowing for the simultaneous generation of full-body movements for both speakers and listeners based on a given raw audio during inference. The generated speaker and listener motions exhibit a more coordinated and cohesive interaction.

2 Related Work

2.1 Audio-Based Motion Generation

Speaker-Centric Generation from Speech has attracted considerable research interest in recent years. The initial efforts have been made towards talking head generation within 2D (Wang et al., 2021; Zhang et al., 2023c; Doukas et al., 2021) and 3D (Cudeiro et al., 2019; Richard et al., 2021; Fan et al., 2022) domain. Another line of work addresses the generation of body poses of speakers based on audio, which can be divided into rule-based (Kipp, 2005; Kopp et al., 2006) and learning-based methods (Huang and Mutlu, 2014; Levine et al., 2010; Sargin et al., 2008), of which we focus on the latter. 2D or 3D skeletons have been popular representations for body pose generation (Yoon et al., 2020; Zhi et al., 2023; Ao et al., 2022; Zhu et al., 2023), while being efficient, such representations can only encode body movements in a coarse and abstract manner. To this end, full-body motion generalization consisting of body movements and hand gestures (Habibie et al., 2021; Yi et al., 2023) have been proposed. Overall, it is evident that a more and more comprehensive understanding of non-verbal signals has been achieved along this line of work. Indeed, our holistic description of non-verbal communication greatly benefits from the recent SOTA (Yi et al., 2023). The specific task differences can be referenced from the upper part of Fig. 1. Nevertheless, these approaches completely ignore the perspective of listeners, which plays an equally important role in communication.

Listener-Centric Generation from Speech primarily focuses on the non-verbal reactions of listeners with respect to speakers' output. There has been some progress in recent years on the work concerning the facial reactions of listeners. Their task can be summarized as shown in Fig. 1d. Data-driven generation of listeners has been first introduced by (Gillies et al., 2008), which is purely conditioned on the input audio clips. More recent breakthrough (Zhou et al., 2022) leverages deep neural networks and considers the generation of listener's faces conditioned on both audio and speaker's faces. Follow-up works consider more general encoding including the head motion of listeners, and have further improved the performance by utilizing advances on generative models, such as VQ-VAE model (Ng et al., 2022) and diffusion model (Liu et al., 2023). Song et al. (2023a) also introduce extra emotion information to generate realistic listener head motion. The recent REACT challenge (Song et al., 2023b) considers online multiple-listener facial reaction generation based on the speaker's real-time behavior. Similar to the aforementioned works on speakers, the current efforts in listener generation are also biased. Moreover, they all focus on generating facial expressions, which fail to capture the full-body movements. A series of works also focus on interlocutor-aware listener body reactions, as shown in Fig. 1e. Early effort along this line focuses on predicting the listener's response based on the speaker's behavior (Joo et al., 2019; Jonell et al., 2020). Based on these, (Ahuja et al., 2019) further adds the previous listener's pose to predict the listener's next pose. More recently, several works utilize popular generative models like GANs to generate body poses of listeners (Tuyen et al., 2022; Tuyen and Celiktutan, 2023). The GENEA challenge (Kucherenko et al., 2023) considers interlocutor-aware holistic motions, and the DYAD challenge (Palmero et al., 2022) considers behavior forecasting for the upper body, faces, and head of the two interlocutors simultaneously. However, these works all use skeletons to represent 3D poses, which lack detailed expression compared to meshes and cannot be directly used for rendering. Additionally, previous works only consider generating listener behavior and ignore the mutual influence between the speaker and the listener.

We advocate considering a new task, as shown in Fig. 1f, that generates 3D holistic motions for *both speakers and listeners simultaneously* based on the input audio while considering the *mutual influence between the speaker and the listener*.

Audio encoding is a key step in all types of co-speech motion generation tasks. It has been a common practice to leverage MFCC (Davis and Mermelstein, 1980) to encode raw audio into features. A commonly shared treatment is to first encode audio input with MFCC (Zheng et al., 2001). One line of works (Bhattacharya et al., 2021; Qian et al., 2021; Yoon et al., 2020; Kucherenko et al., 2020) directly leverage MFCC features as a whole during follow-up training phase. A more advanced treatment (Yi et al., 2023; Zhao et al., 2024; Yoon et al., 2022; Yin et al., 2023) refines MFCC features with pre-trained wav2vec (Baevski et al., 2020), to improve the feature embedding quality. Yet, the output of wav2vec is not decoupled.

On the other hand, it is gradually recognized that decoupling audio into different components can help align high-level and low-level information with the corresponding motions. For instance, hierarchical approaches (Ao et al., 2022; Liu et al., 2022a; Chang et al., 2023) consider the feature extracted from text via BERT as high-level semantics, and use contrastive learning to extract the components (from MFCC encoding) aligning best to the former as semantic audio feature. While the above approach improves the co-speech generation, their decoupling still lacks interpretability and flexibility in further manipulation.

In contrast, we look outside of the domain of interest, and introduce StyleTTS (Li et al., 2022), a pre-trained model from Text-To-Speech (TTS) field. Thanks to its strong performance and interpretability in decoupling not only boosts the co-speech generation model, but also allows flexible manipulation. We refer readers to the Appendix for more details.

2.2 Speech-to-Motion Dataset Construction

Speech-to-Motion Datasets are the cornerstone of the generation task. We follow the above clustering criterion to categorize them into speaker-based and listener-based. Speech-to-motion based on speaker has been extensively studied, resulting in an abundant variety of datasets in this field, which can be further classified into two types according to the main target: facial expressions and body movements. There exists a set of large-scale datasets for talking heads, including 2D datasets (Nagrani et al., 2017; Lu et al., 2021; Guo et al., 2021) and a 3D one (Cudeiro et al., 2019). Regarding body movements, related datasets have been proposed in Yoon et al. (2019), Yoon et al. (2020), Ginosar et al. (2019). Among them, Ginosar et al. (2019) contains 1766 videos sourced from online TED speech videos and gives 3D poses as pseudo-GroundTruth (p-GT). Noting that the aforementioned datasets only consider facial expressions or body movements during speech, without taking into account full-body actions, Habibie et al. (2021) introduces a dataset comprising 33 h of information involving the holistic movements of speakers. Building upon this, Yi et al. (2023) performed data cleaning and provided the pseudo labels with respect to the well-known SMPL-X model (Pavlakos et al., 2019).

On the other hand, the construction of datasets focusing on listeners is relatively lagged. The ViCo dataset (Zhou et al., 2022) is the first one to contain rich samples of different listener identities. Another concurrent dataset, Learning2Listen (Ng et al., 2022), consists of videos of a total duration of 72 h, collected in the wild, which comes from YouTube with six identities.

There are relatively few open-source datasets available for human-human interaction. (Lee et al., 2019) presents Talking With Hands, which is a dataset of two-person face-to-face spontaneous conversations but fails to capture features beyond body and fingers. Palmero et al. (2021) introduces UDIVA, a non-acted dataset of face-to-face dyadic interactions, where interlocutors perform competitive and collaborative tasks with different behavior elicitation and cognitive workloads. However, this dataset only focuses on the upper body movements of individuals. Some datasets focusing on gesture generation in specified dyadic interactions are also available. For instance, the JESTKOD (Bozkurt et al., 2017) dataset focuses on agreement and disagreement scenarios; LISI-HHI dataset (Tuyen et al., 2023) focuses on different interaction scenarios from wayfinding to tangram games. These datasets are all conducted in controlled experimental environments, which significantly limit the diversity of facial expressions and movements as they are tailored to specific tasks. To this end, we seek *in-the-wild RGB videos* that also include clear facial, hand, and body movements, allowing for regression to corresponding meshes. We, therefore, introduce a large-scale in-the-wild interactive communication dataset, HoCo, which comprises video clips, corresponding audio, text, reconstructed SMPL-X p-GT, and annotations indicating the speaker's position and the listener's emotions. For specific dataset comparisons, please refer to Table 1.

Parametric Models play a role in our dataset construction. We benefit from the fact that, under such models, human (or human parts) shapes can be represented by finite-dimension latent codes, giving rise to an efficient representation in both optimizations and also a strong constraint on the plausibility of the generation. In the task of talking head synthesis, 3D Morphable Models (3DMM) (Paysan et al., 2009; Blanz and Vetter, 2023) is widely used to assist in learning accurate mouth shapes and facial expressions (Suwajanakorn et al., 2017; Liu et al., 2022b; Zhang et al., 2023c). SMPL (Loper et al., 2015) is a widely used 3D body model that represents human bodies in a deformable and articulated manner. SMPL-X Pavlakos et al. (2019) is an important extension of SMPL, which allows for extra encoding and control over hands and face. In order to achieve a holistic generation of non-verbal communication, we utilize SMPL-X in our dataset construction, as well as model formulation. In particular, we adapt the state-of-the-art method for SMPL-X parameter regression AiOS Sun et al. (2024) to compute p-GT with respect to SMPL-X parametric model in our HoCo dataset.

3 Methodology

Figure 3 shows the schematic illustration of our whole pipeline. In the following, we first introduce in Sect. 3.1 how the verbal input is encoded by a mixture of hand-craft feature extractor and pre-trained deep neural networks (Fig. 3a). Then we present details on the VQ-VAE module for generating the human motion of both speaker and listener in Sect. 3.2. (Fig. 3b). Finally, we propose our auto-regressive transformer for in Sect. 3.3 (Fig. 3c).

3.1 Feature Extraction

The input of our pipeline is assumed to be audio clips as well as the accompanied text. In particular, we follow the same procedure in Ao et al. (2022) to align the text with audio clips in the SHOW dataset. We assume to be given a audio clip of T frames $A = [a_1, a_2, \cdots, a_T] \in \mathbb{R}^{1\times T}$. The extracted and aligned text is denoted by $W = [w_1, w_2, \cdots, w_T] \in \mathbb{R}^{1\times T}$.
Identity Encoding Following the previous works (Yi et al., 2023; Ao et al., 2022), we perform one-hot encoding of the speakers' ID as $F^{\mathtt{I}} \in \{0, 1\}^{N_I}$, where N_I represents the number of speakers in the dataset.
Audio Encoding We first encode the input audio clip A with MFCC (Sahidullah and Saha, 2012), yielding an initial set of per-frame features $F^M = [F^M_1, F^M_2, \cdots, F^M_T] \in \mathbb{R}^{80\times T}$.

In order to decouple semantic components of A, we adapt the pre-trained model of StyleTTS (Li et al., 2022), which also takes hand-crafted features from MFCC as input. As a by-product, the features extracted by StyleTTS naturally consist of three components, yielding

$$F^A = [F^A_1, F^A_2, \cdots, F^A_T] \in \mathbb{R}^{130\times T},$$
$$\text{where } F^A_i = [e_i; p_i; y_i],$$

where e_i, $p_i \in \mathbb{R}^1$, $y_i \in \mathbb{R}^{128}$ correspond the features regarding energy, pitch, and style of the $i-$frame of A, respectively. The former two reflect the intensity and pitch of the sound,

Table 1 Comparison of different speech-to-motion datasets

Dataset	Environment	Interaction	Holistic body connection	Body	Head	Hand	Annotation	Length
Multiface (Wuu et al., 2022)	Lab	✗	✗	✗	3D mesh	✗	Multi-camera	–
VOCASET (Cudeiro et al., 2019)	Lab	✗	✗	✗	3D mesh	✗	4D-scan	–
Takeuchi et al. (Takeuchi et al., 2017)	Lab	✗	✗	3D keypoint	✗	✗	MoCap	5 h
Trinity (Ferstl and McDonnell, 2018)	Lab	✗	✗	3D keypoint	✗	✗	MoCap	4 h
Yoon et al. (Yoon et al., 2019, 2020)	Wild	✗	✗	3D keypoint	✗	✗	p-GT	52 h
Speech2Gesture (Ginosar et al., 2019)	Wild	✗	✗	2D keypoint	✗	2D keypoint	p-GT	144 h
Habibie et al. (Habibie et al., 2021)	Wild	✗	✗	3D keypoint	3D mesh	3D keypoint	p-GT	33 h
SHOW (Yi et al., 2023)	Wild	✗	✓	3D mesh	3D mesh	3D mesh	p-GT	27 h
UDIVA (Palmero et al., 2021)	Lab	✓	✗	RGB video	RGB video	✗	Multi-camera	90.5h
Talking With Hands (Lee et al., 2019)	Lab	✓	✗	3D keypoint	✗	3D keypoint	Multi-camera	50 h
JESTKOD (Bozkurt et al., 2017)	Lab	✓	✗	3D keypoint	✗	✗	Multi-camera	4.3h
LISI-HHI (Tuyen et al., 2023)	Lab	✓	✗	3D keypoint	✗	✗	Multi-camera	8.3h
ViCo (Zhou et al., 2022)	Wild	✓	✗	✗	RGB video	✗	p-GT	1.5h
Learning2Listen (Ng et al., 2022)	Wild	✓	✗	✗	3D mesh	✗	p-GT	72 h
Ours	Wild	✓	✓	3D mesh	3D mesh	3D mesh	p-GT	45 h

The horizontal line dividers represent, respectively: non-interactive lab scenes, non-interactive wild scenes, interactive scenes and our provided HoCo dataset

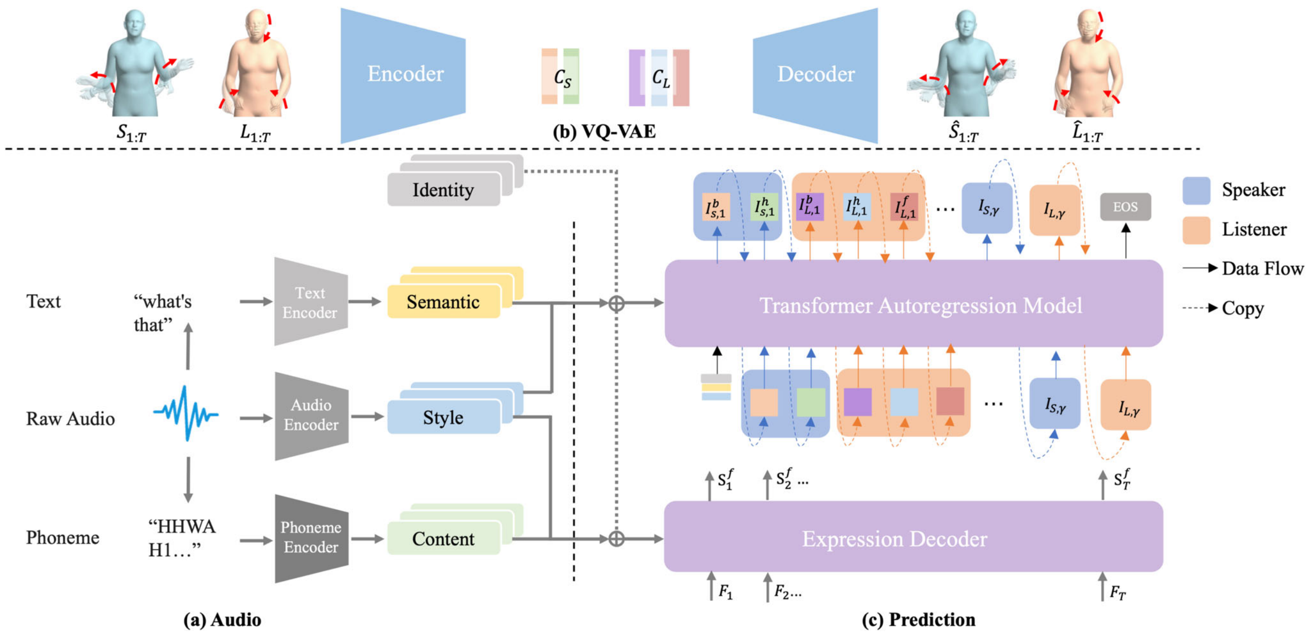


Fig. 3 Overview of the proposed framework for holistic communicating generation: **a** Audio feature extraction; **b** VQ-VAE model for the generation of speaker and listener motion; **c** Transformer-based autoregression model for simultaneously generating the motion of both speaker and listener in a chain-like manner

while the latter encodes the emotional aspects of the audio. The extracted features have a strong correlation with the body and hand motions. As will be shown in Sect. 5.6, such feature decomposition facilitates fine-grained control over the generated body language.

Text Encoding Transcription is a crucial form of speech representation that conveys detailed linguistic information in a concise format. It is usually presented as a sequence of words, and the rate at which words are spoken can vary with the speech tempo. To address this variability, we follow (Ao et al., 2022) to align words with the corresponding speech and convert the text into frame-level features. Subsequently, we take the aligned text as input and use the pre-trained model of BERT (Devlin et al., 2018) to encode W, obtaining $F^W = [F^W_1, F^W_2, \cdots, F^W_T] \in \mathbb{R}^{768\times T}$. We obtain the aligned phoneme in a similar manner which reflects the content information of the audio and input it into the text encoder of StyleTTS to obtain the phoneme feature $F^P = [F^P_1, F^P_2, \cdots, F^P_T] \in \mathbb{R}^{512\times T}$.

In the end, we denote by $F^{\mathtt{m}}$ the extracted features from the input verbal cues A and W, and $F^{\mathtt{f}}$ from A and P:

$$F^{\mathtt{m}} = [F^A; F^W] \in \mathbb{R}^{898\times T},$$
$$F^{\mathtt{f}} = [F^A; F^P] \in \mathbb{R}^{642\times T}. \tag{1}$$

3.2 VQ-VAE Model

We first denote the SMPL-X p-GT label for the speaker and listener over a time duration T by $S_{1:T} = [s_1, ..., s_T]$ and $L_{1:T} = [l_1, ..., l_T]$, respectively. For each s_i (l_j), we indicate the label regarding facial expressions, body movements, and hand gestures by superscripts $\mathtt{f}$, $\mathtt{b}$, and $\mathtt{h}$, respectively. That is $s_i = [s^{\mathtt{f}}_i; s^{\mathtt{b}}_i; s^{\mathtt{h}}_i]$ (similar definition for l_j). For the sake of simplicity, we also stack the label along each attribute, for example, $S^{\mathtt{f}} = [s^{\mathtt{f}}_1, s^{\mathtt{f}}_2, \cdots, s^{\mathtt{f}}_T]$.

Given a raw audio clip A, we first obtain $F^{\mathtt{m}}$ and $F^{\mathtt{f}}$ via Eq.(1). The speaker's lip shape and expression are generally related to the phonemes, rhythm, and other information in the audio (Wu et al., 2023). For the generation of facial expressions for the speaker, we follow (Yi et al., 2023) and model it as a regression task. More specifically, we train a network, $\mathcal{N}^{\mathtt{f}}_S$, to recover $\hat{S^{\mathtt{f}}}$, namely

$$\hat{S^{\mathtt{f}}} = \mathcal{N}^{\mathtt{f}}_S(F^{\mathtt{f}}, F^{\mathtt{I}}) = [\hat{s_1}^{\mathtt{f}}, \hat{s_2}^{\mathtt{f}}, \cdots, \hat{s_T}^{\mathtt{f}}]. \tag{2}$$

We use CNN (Yi et al., 2023) as the backbone of $\mathcal{N}_S$ and Mean Squared Error (MSE) loss for training.

On the other hand, we aim to generate body and hand motions for speakers and full-body movements for listeners based on verbal input. Apart from maximizing diversity as in general generation tasks, we further emphasize the harmony between the generated body and hand motions. To this end, we utilize a VQ-VAE (Van Den Oord et al.,

2017) network and employ an autoencoder to discretely encode the regarding attributes. We take the generation of the speaker's body motion as an example. Given a sequence $S^{\mathrm{b}} = [s_1^{\mathrm{b}}, s_2^{\mathrm{b}}, \cdots, s_T^{\mathrm{b}}]$, we train an autoencoder and a finite-dimensional codebook $C_S^{\mathrm{b}} = [c_{S,1}^{b}, ..., c_{S,K}^{b}]$, which contains K codes.

We denote the encoder and decoder of the VQ-VAE as $\mathcal{E}_S^{\mathrm{b}}$ and $\mathcal{D}_S^{\mathrm{b}}$ respectively. The former takes as input S^{b}, and output latent code $Z_S^{\mathrm{b}} = [z_{S,1}^{\mathrm{b}}, ..., z_{S,\gamma}^{\mathrm{b}}]$, where $\gamma = T/w$ and w is the temporal window size. Then we can quantize the i-th embedding $z_{S,i}^{b}$ by comparing it with the codebook, C_S^{b}, to find the closest code:

$$\hat{z}_{S,i}^{\mathrm{b}} = \underset{c_{S,k}^{\mathrm{b}} \in C_S^{\mathrm{b}}}{\arg\min} \|z_{S,i}^{\mathrm{b}} - c_{S,k}^{\mathrm{b}}\|. \tag{3}$$

Then, we reconstruct the body motion with the decoder, with respect to the latent code in Eq. (3): $\hat{S}^{\mathrm{b}} = \mathcal{D}_S^{\mathrm{b}}(\hat{Z}_S^{\mathrm{b}})$.

Finally, we train the autoencoder and the codebook using the following loss function (Van Den Oord et al., 2017):

$$\begin{aligned} L_{\mathrm{vq}} = L_{\mathrm{rec}}(\hat{S}^{\mathrm{b}}, S^{\mathrm{b}}) + \|sg[Z_S^{\mathrm{b}}] - \hat{Z}_S^{\mathrm{b}}\| \\ + \beta \|Z_S^{\mathrm{b}} - sg[\hat{Z}_S^{\mathrm{b}}]\|, \end{aligned} \tag{4}$$

where L_{rec} denotes the MSE loss, $sg\,[\cdot]$ is the stop-gradient operator and β is a hyper-parameter for the commitment loss. These components ensure that the encoder commits to specific codes and that the codebook is utilized optimally. We independently train the autoencoder and the codebook for each of S^{h}, L^{f}, L^{b}, L^{h}, which are denoted in the same manner as above in the following.

3.3 Architecture of Speaker-Listener Generator

With the learned VQ-VAEs in Sec 3.2, the motion of the speaker $S_{1:T}$ and listener $L_{1:T}$ can be encoded as a sequence of indices $I_{S,1:\gamma}$ and $I_{L,1:\gamma}$ obtained via Eq. (3), with respect to the corresponding encoder and codebook. We can project these indices back to their corresponding codebook entries and decode the obtained latent code to reconstruct the motion. Therefore, we model motion generation as an auto-regressive next-index prediction from audio input. In particular, we not only consider the coordination between the facial expressions and body movements of the speaker and the listener individually but also account for the mutual influence between them. To this end, we need to exploit the interaction between the speaker and the listener. More specifically, we formulate the generation of the speaker's and listener's motion as

$$p(I_S|F^{\mathrm{m}}, F^{\mathrm{I}}) = \prod_{i=1}^{|I_S|} p(I_{S,i}|F^{\mathrm{m}}, F^{\mathrm{I}}, I_{S,1:i-1}, I_{L,1:i-1}), \tag{5}$$

$$p(I_L|F^{\mathrm{m}}, F^{\mathrm{I}}) = \prod_{i=1}^{|I_L|} p(I_{L,i}|F^{\mathrm{m}}, F^{\mathrm{I}}, I_{S,1:i}, I_{L,1:i-1}), \tag{6}$$

where I is assembled by I^{b}, I^{h}, I^{f}, representing the motion parameters of the body, hand, and face for either the speaker or the listener. This process is depicted in Fig. 3c. Specifically, the dataflow starts from fine-grained audio features and ends with an End-of-Sequence (EOS) token. During each index prediction phase, the speaker is predicted before the listener, as the speaker takes a leading role in the conversation process. The entire processing procedure is chain-like, which allows for sufficient mutual interaction between the speaker and the listener during conversation. In this process, the motion of the speaker involves fine-grained movements of the body and hands, while the listener involves movements of the body, hands, and face. The final updated speaker and listener indices are processed through their respective trained VQ-VAE models to obtain the final reconstruction results.

We highlight the difference between our chain-like design and that of (Yi et al., 2023). The latter only considers the generation task for a speaker, which solely enforces the harmony among the generation of the speaker's body and hand. On the other hand, our design further allows for the simultaneous output of both the speaker and listener and takes into consideration the mutual influence between them while the conversation is going on.

We train an auto-regressor with a transformer backbone (Zhang et al., 2023b) in this generation network, where we maximize the log-likelihood of the data distribution:

$$L_{\mathrm{reg}} = E_{I \sim p(I)}[-\log p(I|F^{\mathrm{m}}, F^{\mathrm{I}})]. \tag{7}$$

4 Dataset Construction

In this section, we introduce our new dataset, named HoCo, which is designed for holistic non-verbal communication. The HoCo dataset comprises video clips sourced from natural settings. It encompasses both verbal elements (audio clips and their transcripts) and 3D annotations of non-verbal signals like facial expressions, body postures, and hand gestures. Distinctively, our dataset provides data for both speakers and listeners, unlike previous works that focus on one or the other. Additionally, HoCo includes annotations of listeners' emotions. Although not used in our experiments, this feature adds more depth compared to other leading datasets such as SHOW (Yi et al., 2023).

Table 1 presents a clear comparison between the existing benchmarks and ours. We emphasize that our dataset enjoys several features including being captured in a real-world scenario, carrying rich annotations, and involving the most comprehensive information on speaker-listener inter-

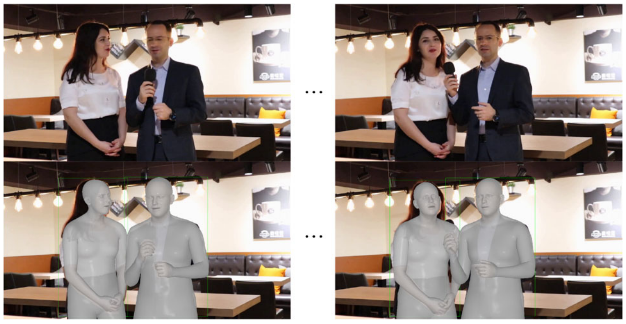


Fig. 4 In HoCo, we provide high-definition videos of two-person communication (top), as well as the corresponding p-GT estimated by SMPL-X (bottom)

action among all. In Fig. 4, we select one challenging scene (occlusion between each other) from our dataset to showcase, which includes the RGB images of the videos and the corresponding visualizations of SMPL-X reconstructions. In the following, we provide details on our dataset construction.

4.1 Raw Data Collection and Screening

We start by collecting conversational video clips from YouTube[1] involving two participants. The scenes include dual hosts, two-person blog programs, and talk show programs. We have curated a total of 45 hours of video data, which includes 22, 913 clips involving 26 speaker IDs. The selected clips pass our screen process based on the following criteria: 1) The video frames should feature two individuals, of which one is speaking, and the other responds accordingly; 2) Both individuals' upper bodies are clear and unobstructed, with arms fully visible during their communication, and their facial expressions are clear; 3) There is clear and noticeable interaction between the participants. The clips with listeners showing significant reactions such as nodding, smiling, shrugging, and rich facial expressions such as happiness or surprise are preferred.

4.2 Annotation Procedure

Unlike existing datasets such as SHOW (Yi et al., 2023), we provide multi-modal data including audio clips, aligned transcripts, and the pseudo-GroundTruth (p-GT) labels with respect to SMPL-X (Pavlakos et al., 2019) model.
Audio and Transcript Processing Based on high-resolution RGB videos, we initially extract audio using the **moviepy**[2] library in Python. Subsequently, based on the audio, we use the **whisper**[3] library to extract the corresponding transcripts. To align each transcript with the corresponding time frames, we employ Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) to detect the start and end times of each word.
3D Body Language Annotation Our dataset also provides 3D whole-body meshes as pseudo-labels. We choose SMPL-X Pavlakos et al. (2019) as the p-GT for the dataset because it contains rich facial, hand, and body details. In particular, we leverage the recent advance on SMPL-X estimation from video clips, AiOS Sun et al. (2024), for automatically generating p-GT. We control the order of the detection boxes to assign p-GTs to the speaker and listener respectively. The p-GT comprises parameters of a shared body shape $\beta \in \mathbb{R}^{10}$, poses $\theta_h \in \mathbb{R}^{156\times T}$, a shared camera pose $\theta_c \in \mathbb{R}^3$ and translation $\epsilon \in \mathbb{R}^3$, and facial expressions $\phi \in \mathbb{R}^{10\times T}$. The experimental results in the main text are all based on HoCo.

5 Experiments

5.1 Datasets

We start by detailing the datasets and evaluation metrics used in our experiments. Since our goal is to generate full body language, apart from our proposed dataset, HoCo, we consider the SHOW dataset for benchmarking our design on the single-role pipeline:
Show (Yi et al., 2023) This dataset is a filtered version based on (Habibie et al., 2021), resulting 26.9 hours of high-quality videos. The dataset comprises 4 speaker IDs. Additionally, p-GT for SMPL-X parameters is annotated. On this basis, we generate text aligned with the number of frames and filter out data that cannot be automatically aligned.
HoCo: We evaluate our method and the baselines with our HoCo dataset. More precisely, HoCo dataset is built on video clips collected *in the wild*, which provides holistic information of talks in the following two perspectives: 1) It includes both verbal inputs– audio clips and accompanied transcripts) and 3D annotations for the non-verbal signals (facial expressions, body poses, and hand gestures); 2) Unlike the prior works biased towards either speakers or listeners, the input and annotation above are constructed for *both* speakers and listeners within the collected talks.

Moreover, we also provide annotations of the listeners' emotions. Though this attribute is not utilized in our experiments, we highlight that HoCo contains more modalities than the state-of-the-art datasets (e.g., SHOW (Yi et al., 2023)).

[1] https://www.youtube.com.

[2] https://pypi.org/project/moviepy/.

[3] https://github.com/openai/whisper.

5.2 Implementation Details

The training is composed of two stages. For the VQ-VAE model, the codebook size is set to 256 × 256. The temporal window size w is 4 and we set the number of codes to be $2,048$. The weight, β, of the commitment loss term in Equation(4) in the main submission is set to be 0.25. For the Speaker-Listener Generator, we choose a transformer model comprising 9 layers and 16 heads. We set the block size to be 376. We adapt Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a learning rate of 0.0001 as the optimizer. For all experiments, the batch size is set to 128, and training is conducted on a single Tesla V100-32 G GPU for 100 epochs with all video clips cropped into 88 frames.

5.3 Evaluation Metrics

We roughly categorize the evaluation metrics involved into deterministic and non-deterministic.

The former is especially applied to the task of generating the speaker's facial expressions, which is performed as a deterministic task. In particular, we use **L2** and Landmark Velocity Difference (LVD) to measure the authenticity and accuracy of generated expressions. The **L2** distance is to measure the difference between the generated facial expression and the ground truth label. LVD quantifies velocities of corresponding facial expressions between the ground truth and the generated expressions. This measurement helps to evaluate how well a model captures the dynamic aspects of facial movements.

Regarding the latter, we follow (Zhi et al., 2023) to assess the body and hand motions of both speaker and listener. The evaluation metrics include Frechet Gesture Distance (FGD) (Yoon et al., 2020), Beat Consistency Score (BC) (Liu et al., 2022a), Variation (Yi et al., 2023), Concordance Correlation Coefficient (CCC) (Song et al., 2023b), and Time Lagged Cross Correlation (TLCC) (Boker et al., 2002; Ng et al., 2022). FGD provides a quantitative measure of the dissimilarity between the generated and ground truth motions which uses a pre-trained autoencoder to project the motion into latent space. BC is used to assess the consistency of motion, particularly in the context of body and hand movements associated with rhythmic beats or gestures. We utilize both FGD and BC together to evaluate the rationality of the generated motion. Variation is calculated by the variance across the time series sequence of the motions.

Meanwhile, there is no golden standard for evaluating whether a listener's reaction is appropriate in response to a speaker's behavior, while manually labeling would be labor-intensive and subjective. Therefore, within the evaluation metrics for the generation of the listener's full-body motions, we incorporate CCC followed REACT23 challenge (Song et al., 2023b), to calculate the correlation between the generated full-body motions and its most similar appropriate motion. Additionally, we also employ TLCC to measure the synchrony between the generated speaker and listener for dyadic motion assessment.

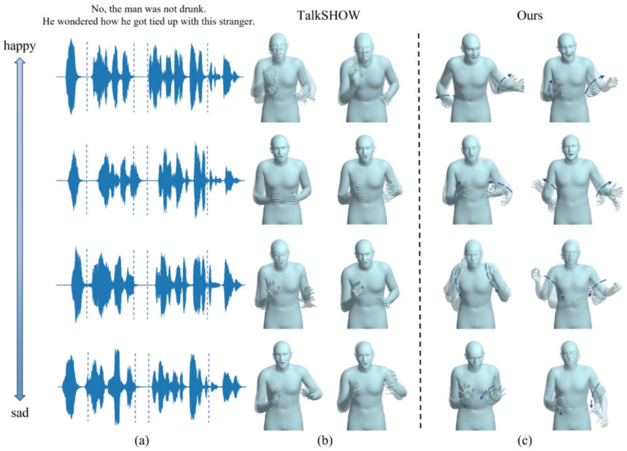

No, the man was not drunk.
He wondered how he got tied up with this stranger.
TalkSHOW
Ours
happy
sad
(a)
(b)
(c)

Fig. 5 **a** shows a piece of transcript and the corresponding audio signals with varying pitches and emotions. **b** displays the inference results from TalkSHOW (Yi et al., 2023), where the generated motions have low sensitivity to changes in the audio signals. **c** presents our inference results, demonstrating high consistency between the generated motions and the audio

5.4 Baselines

Regarding generating speakers' motion from speech, we compared our method with the SOTA methods Habibie et al. Habibie et al. (2021) and TalkSHOW Yi et al. (2023) in the speech-to-motion task. Simultaneously, we construct some baselines to validate the effectiveness of our approach. In the following, Audio + CVAE refers to the network structure of (Petrovich et al., 2021a), using audio as a condition to generate corresponding motion through a VAE autoencoder structure. TalkSHOW + ours feature replaces the audio processing method of TalkSHOW with the decoupled feature approach we proposed.

For our novel task of non-verbal communication generation with the HoCo dataset, as there exists no available baseline, we follow the responsive listening head generation task proposed in Zhou et al. (2022), which uses the audio and speaker's information as conditions for generating the listener. We compare with our formulated variants instead. Namely, for a baseline **X**, we first train the model to generate the speaker with **X** and use the SMPL-X p-GT corresponding to the speaker as the condition to generate the listener. We refer to the results of this approach as **X**-split.

Table 2 Comparison to baselines on SHOW dataset (Yi et al., 2023)

Method	Face		
	L2 ↓		LVD ↓
Habibie et al. (2021)	0.237		0.036
TalkSHOW (Yi et al., 2023)	0.215		0.032
Ours	**0.209**		**0.031**
	Body&Hands		
	FGD ↓	BC ↑	Variation ↑
Habibie et al. (2021)	3.198	0.452	0.051
Audio+CVAE (Petrovich et al., 2021b)	2.561	0.665	0.134
TalkSHOW (Yi et al., 2023)	2.049	0.847	0.332
TalkSHOW+our features	1.902	0.849	0.328
Ours	**1.483**	**0.854**	**0.617**

↑ indicates higher is better and ↓ indicates lower is better

5.5 Decoupling Feature Results

In Section 3.2, we introduce an audio decoupling method, and now we provide a more detailed qualitative validation of its effects. If the input contains only textual information, the corresponding generated results should be similar. However, if the same sentence is spoken by a speaker with different pitches or emotions, there should be distinct motion expressions. Based on this, we select the dataset provided by Wang et al. (2020), where the same text includes two speaking emotions, happy and sad, each with two intensity levels from low to high. We use the model trained on the SHOW dataset (Yi et al., 2023) to directly infer these unseen audios. We compare our method with TalkSHOW (Yi et al., 2023), and the results are shown in Fig. 5.

Our method, built upon the foundation of textual semantic information, explicitly considers audio pitch and style elements, where style represents the emotional content of the audio. Our results show significant variation in gestures generated for different audio emotions. For instance, gestures associated with happiness are predominantly upward, while those for sadness tend toward downward movements. With the same emotion at different pitches, a higher pitch variation corresponds to a more pronounced motion. In contrast, results generated by TalkSHOW (Yi et al., 2023) for the same text under different audio conditions exhibit minimal variation in motion, indicating poorer consistency with the audio.

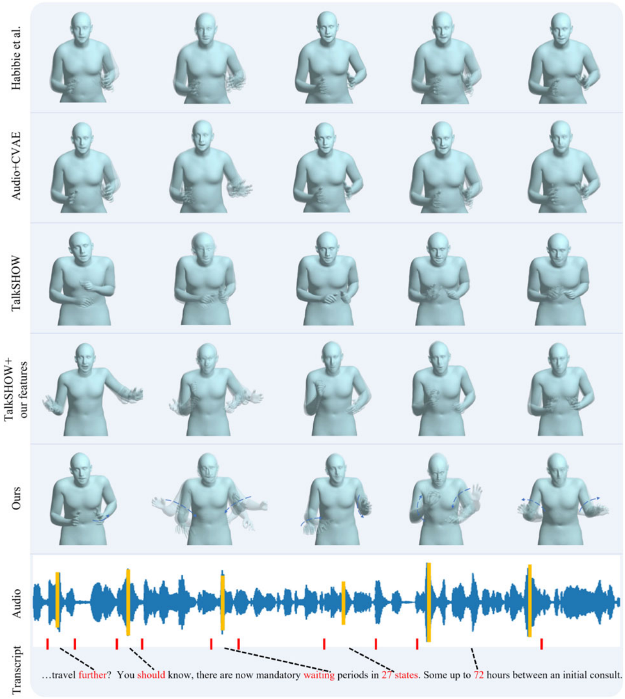

Habibie et al.
Audio+CVAE
TalkSHOW
TalkSHOW+ our features
Ours
Audio
Transcript
…travel further? You should know, there are now mandatory waiting periods in 27 states. Some up to 72 hours between an initial consult.

Fig. 6 Visual comparisons with the various baselines on SHOW dataset. Our method generates diverse motions consistent with the rhythm of the input audio. In this audio clip with progressively emotional content, our generated results exhibit more diverse motions

5.6 Speaker-Centric Generation

Quantitative Analysis Table 2 shows the quantitative results on the SHOW dataset. We first validate the effectiveness of our proposed feature decoupling. Using the model framework from TalkSHOW (Yi et al., 2023) as a baseline, we replace the input features with our decoupled features. In the generation of facial expressions, compared to directly extracting features from a pre-trained wav2vec (Baevski et al., 2020) model, our decoupled features result in more accurate expressions. Moreover, our features assist in generating body and hand movements that are more similar to the ground truth. In comparison to baselines, our generated results exhibit an improvement both in audio consistency and motion diversity.

Table 3 Comparison to baselines on HoCo dataset

Method	Speaker			Listener				Speaker&Listener
	FGD ↓	BC ↑	Variation ↑	FGD ↓	BC ↑	Variation ↑	CCC ↑	TLCC ↓
Habibie et al. split	1.524	0.523	0.121	1.007	0.579	0.040	0.109	23.995
Audio+CVAE split	1.165	0.801	0.163	0.601	0.652	0.055	0.117	21.003
TalkSHOW split	1.112	0.829	0.397	0.633	0.875	0.139	0.145	19.136
TalkSHOW+our features split	0.995	0.827	0.432	0.512	0.871	0.161	0.150	18.895
Ours split	1.297	0.894	0.473	0.501	**0.888**	0.208	0.179	17.248
Ours	**0.959**	**0.897**	**0.504**	**0.375**	0.873	**0.265**	**0.208**	**13.897**

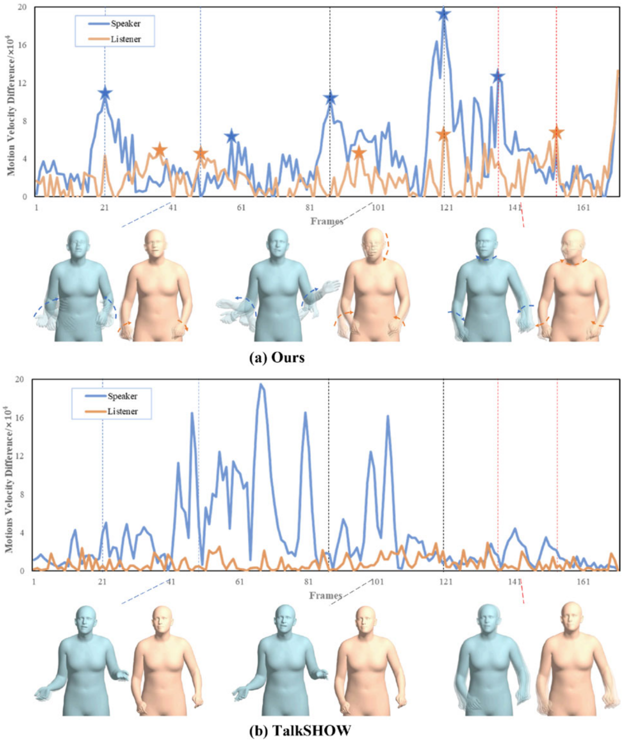
↑ indicates higher is better and ↓ indicates lower is better

Qualitative Analysis Figure 6 shows a comparison between the motion generated by us and those of various baselines for a speech clip with progressive emotions. The sentence contains some emphasized words, including "should" "waiting" etc., accompanied by intonation changes. Our generated results exhibit more noticeable body movement changes in these areas. For emotionally progressive words like "27" and "72", our results introduce gestures such as shrugging. In contrast, results generated by TalkSHOW (Yi et al., 2023) lack diversity, with only small, regular hand movements. Our results are more vivid in comparison. We also provide a demo video as a supplement.

5.7 Speaker and Listener Generation

Quantitative Analysis For the speaker and listener generation framework, we specifically design baselines to validate the effectiveness of our proposed framework. The quantitative results are shown in Table 3. Firstly, we observe the improvement of TalkSHOW + our features-split upon TalkSHOW-split, which confirms again the advantage of our factorized feature. It is also consistent with the results reported in Table 2. Secondly, the performance gap between Ours-split with Ours validates the effectiveness of our chain-design auto-regressor in improving the diversity of generated results. Finally, compared to our proposed variants on top of external baselines, including TalkSHOW-split, TalkSHOW + our features-split and Ours-split. Ours achieves a 3.6% improvement in FGD for the speaker and a 25.1% for the listener. The diversity of our results has also seen improvement. In the generation of listener actions, we consider CCC to assess the appropriateness of the generated motions. Our method achieves a 13.9% improvement, which indicates that our approach can generate not only diverse listener's motions but also actions that are more in line with reality. At the same time, taking into account the structure of both the speaker and the listener results in an 19.4% increase in synchrony within the generated outcomes.

Qualitative Analysis We show an example from the speaker and listener generation experiment to demonstrate the coherence of our generated speaker and listener actions. This is a playful scenario where, in the latter part of the audio, both the speaker and the listener are laughing. Figure 7a above illustrates the variation in SMPL-X parameters generated by our approach for the speaker and listener, demonstrating a consistent change in parameters, as indicated by the stars. We also visualize corresponding significant changes for both in Fig. 7a below, where the listener responds with laughter and a bending posture following the speaker's movements, and the speaker, in turn, reacts with laughter. In contrast, Fig. 7b shows the corresponding results from TalkSHOW, where the listener's reactions are notably lacking.



(a) Ours

(b) TalkSHOW

Fig. 7 Visual comparisons with the TalkSHOW baseline on HoCo dataset. Our approach can generate semantically meaningful movements such as laughing and stooping

Table 4 The percentage of the user's favorite methods in terms of natural, diversity, synchrony, and consistency

Methods	Speaker			Listener			Speaker&Listener
	Natural	Diversity	Synchrony	Natural	Diversity	Synchrony	Consistency
TalkSHOW-split	31.37%	34.31%	40.20%	27.45%	15.69%	19.61%	19.80%
Ours	**68.63%**	**65.69%**	**59.80%**	**72.55%**	**84.31%**	**80.39%**	**80.20%**

Table 5 Ablation study on `HoCo` dataset

Ablation	Speaker			Listener				Speaker&Listener
	FGD ↓	BC ↑	Variation ↑	FGD ↓	BC ↑	Variation ↑	CCC ↑	TLCC ↓
Ours w/o text feature	1.263	0.897	0.499	0.402	0.853	0.245	0.203	19.007
Ours w/o style feature	1.384	0.889	0.409	0.731	0.872	**0.279**	0.202	17.239
Ours w/o chain-design	1.512	0.876	0.490	0.670	0.870	0.241	0.156	21.845
Ours w/o s-l chain-design	1.297	0.894	0.473	0.501	**0.888**	0.208	0.179	17.248
Ours	**0.959**	**0.897**	**0.504**	**0.375**	0.873	0.265	**0.208**	**13.897**

↑ indicates higher is better and ↓ indicates lower is better

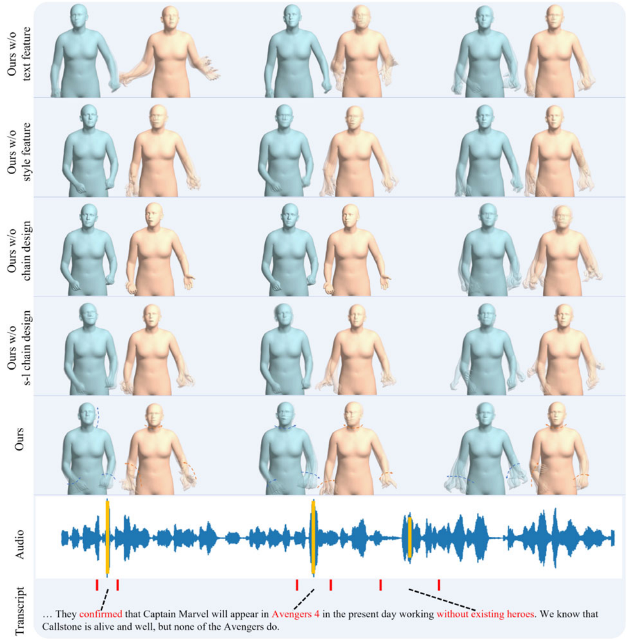

Ours w/o text feature
Ours w/o style feature
Ours w/o chain design
Ours w/o s-l chain design
Ours
Audio
Transcript
… They confirmed that Captain Marvel will appear in Avengers 4 in the present day working without existing heroes. We know that Callstone is alive and well, but none of the Avengers do.

Fig. 8 Visualization about the ablation experiments. Our generated results exhibit more harmonious and diverse motions in this audio clip with progressively emotional content

5.8 User Study

We conduct a user study experiment to demonstrate the differences between our method and the SOTA method (TalkSHOW (Yi et al., 2023)). We interview 102 volunteers and present them with 20 different cases of generated dual-person communication. The respondents evaluated the speaker and listener on three metrics: Natural, Diversity, and Synchrony, as well as the consistency of responses between the speaker and listener. We then calculate the percentage scores for each task across these metrics. As shown in Table 4, our method is more favored by participants in terms of the aforementioned criteria.

5.9 Ablation Study

We conduct ablation experiments primarily focusing on the effects of decoupled features and different chaining designs in our proposed framework. The experiments are validated on the `HoCo` dataset, and the results are presented in Table 5. The first part of the table shows results when not using text and style-related features. The absence of the feature leads to a decrease in the coordination between generated motions and the ground truth. Simultaneously, using all features may result in a slight sacrifice in diversity. The second part validates the rationality of our chaining design. Initially, when we do not consider self-interactions or interactions between the speaker and the listener, as shown in Ours w/o chain-design, the FGD for the speaker increases 34.6% and the FGD for the listener increases 44.0%. Also, the CCC for the listener increases 25.0%. On this basis, when we consider the interaction of their actions, as shown in Ours w/o s-l chain-design, compared to our design, there is an increase in FGD for both 26.1% and 25.1% and there is an increase in CCC for 13.9%. At the same time, our design can enhance the diversity of both speaker and listener generations. The synchrony between the generated speaker and listener has significantly improved. We also provide a visualization of the ablation experiments, as shown in Fig. 8. It can be observed that the motions generated by ours are more harmonious and diverse than the ablated variants, which are consistent with the quantitative results in Table 5.

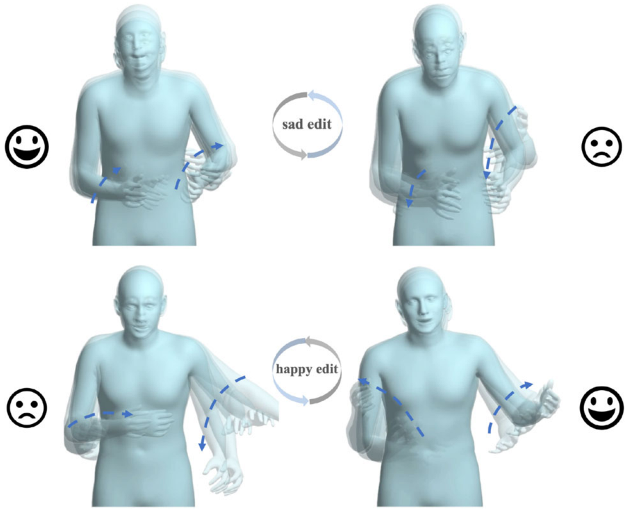


Fig. 9 Left: Given a happy and a sad audio segment, we generate the regarding holistic motions, where the emotion is reflected by the distinctive arm movement directions; Right: We perform emotion editing by interchanging the style feature in our audio encoding, without touching the content of the regarding audio segments. Top-right panel shows the result of injecting sadness into the original happy audio segment, and bottom-right shows the opposite edit. The change of arm movement directions validates our editing effect. See more details in the text

6 Conclusion

In this paper, we take into consideration communication within human interaction and present a novel task that generates 3D holistic human motions for both speakers and listeners. To achieve this goal, we contribute to both dataset and model design. For the former, we provide the `HoCo` communication dataset for future exploration along this task. For the latter, we propose a model tailored for our task, which consists of novel designs including 1) factorization for decoupling audio features which enhances the generation of more realistic and coordinated movements; 2) a chain-like auto-regressive model for characterizing non-verbal communication. Additionally, we achieve state-of-the-art performance on two benchmarks.

Limitations There is still room for improvement in our current research. Our data can only generate p-GT corresponding to the video data through algorithms. Also, we only consider the parametric human model of SMPL-X and do not provide the final realistic rendering results. Additionally, our work only considers situations where the positions of the speaker and the listener remain unchanged, and it cannot simulate scenarios where the positions of the speaker and listener change, which are also quite common in reality. In future work, we will consider a wider range of actions and verbal interactions between people, and we will look into combining different settings to generate realistic dyadic interactions.

Appendix

Emotion Editing via Our Audio Decoupling

To verify the effectiveness of our audio decoupling method, we design the editing experiment shown in Fig. 9. First, we select two segments of audio from the MEAD dataset, containing happy and sad emotions respectively. The texts are "*The revolution now underway in materials handling makes it much easier*" and "*No, the man was not drunk, he wondered how he got up tied up with a stranger*". The generated motion is then shown in the top-left and bottom-left of Fig. 9. Note the opposite directions of arm movement in the left two panels, reflecting the difference in emotion.

Thanks to our audio decoupling design, our generation framework allows for emotion-based editing. We first compute the style feature F^A_{happy}, F^A_{sad} of the input audio clips. Then we switch the style feature between the two audio clips. Namely, following Eq. 1, we construct:

$$F^m_{happy2sad} = [F^A_{sad}; F^W_{happy}],\ F^f_{happy2sad} = [F^A_{sad}; F^P_{happy}], \tag{8}$$

as the input of our trained model. Similarly, we can edit the sad audio clip to a happy one by reversing the above construction.

We remark that, in the above editing, the audio texts are untouched. Top-right of Fig. 9 shows the editing result of making the happy audio sad, and the bottom-right of Fig. 9 shows the opposite. Again, we check the arm movement directions, which confirms the editing effect.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11263-024-02300-7.

Acknowledgements This work was supported by the National Natural Science Foundation of China under contract No. 62171256 and Meituan.

Data availability We claim to release the dataset and code upon acceptance. The datasets generated and analyzed during the current study will be available in our open-source repository.

References

Ahuja, C., Ma, S., Morency, LP., & Sheikh, Y. (2019). 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* (pp. 74–84).

Ahuja, C., Joshi, P., Ishii, R., & Morency, L.P. (2023). Continual learning for personalized co-speech gesture generation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 20893–20903).

Ao, T., Gao, Q., Lou, Y., Chen, B., & Liu, L. (2022). Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM Transactions on Graphics (TOG), 41*(6), 1–19.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems, 33*, 12449–12460.

Bhattacharya, U., Childs, E., Rewkowski, N., & Manocha, D. (2021). Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning. In: Proceedings of the 29th ACM International Conference on Multimedia, pp 2027–2036.

Birdwhistell, R. (1952). Introduction to Kenesics.

Blanz, V., & Vetter, T. (2023). A morphable model for the synthesis of 3d faces. *Seminal Graphics Papers: Pushing the Boundaries, 2*, 157–164.

Boker, S. M., Rotondo, J. L., Xu, M., & King, K. (2002). Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychology Methods, 7*(3), 338.

Bozkurt, E., Khaki, H., Keçeci, S., Türker, B. B., Yemez, Y., & Erzin, E. (2017). The jestkod database: An affective multimodal database of dyadic interactions. *Language Resources and Evaluation, 51*, 857–872.

Chang, Z., Hu, W., Yang, Q., & Zheng, S. (2023). Hierarchical semantic perceptual listener head video generation: A high-performance pipeline. In *Proceedings of the 31st ACM International Conference on Multimedia* (pp. 9581–9585).

Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., & 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* (pp. 10101–10111).

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 28*(4), 357–366.

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Doukas, M.C., Zafeiriou, S., & Sharmanska, V. (2021). Headgan: One-shot neural head synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 14398–14407).

Fan, Y., Lin, Z., Saito, J., Wang, W., & Komura, T. (2022). Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18770–18780).

Ferstl, Y., & McDonnell, R. (2018). Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the 18th international conference on intelligent virtual agents* (pp. 93–98).

Gillies, M., Pan, X., Slater, M., & Shawe-Taylor, J. (2008). Responsive listening behavior. *Computer Animation and Virtual Worlds, 19*(5), 579–589.

Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., & Malik, J. (2019). Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3497–3506).

Guo, Y., Chen, K., Liang, S., Liu, Y.J., Bao, H., & Zhang, J. (2021). Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 5784–5794).

Habibie, I., Xu, W., Mehta, D., Liu, L., Seidel, HP., Pons-Moll, G., Elgharib, M., & Theobalt, C. (2021). Learning speech-driven 3d conversational gestures from video. In *Proceedings of the 21st ACM international conference on intelligent virtual agents* (pp. 101–108).

Huang, CM., & Mutlu, B. (2014). Learning-based modeling of multimodal behaviors for humanlike robots. In *Proceedings of the 2014 ACM/IEEE international conference on human-robot interaction* (pp. 57–64).

Jocelyn Scheirer RWP. (1999). *Affective objects*. MIT Media Laboratory: Tech. rep.

Jonell, P., Kucherenko, T., Henter, GE., & Beskow, J. (2020). 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* (pp. 1–8).

Joo, H., Simon, T., Cikara, M., & Sheikh, Y. (2019). Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10873–10883).

Kipp, M. (2005). Gesture generation by imitation: From human behavior to computer character animation. Universal-Publishers.

Kopp, S., Krenn, B., Marsella, S., Marshall, AN., Pelachaud, C., Pirker, H., Thórisson, KR., & Vilhjálmsson, H. (2006). Towards a common framework for multimodal generation: The behavior markup language. In *Intelligent virtual agents: 6th international conference, IVA 2006, Marina Del Rey, CA, USA, August 21–23*, 2006. Proceedings 6, Springer (pp. 205–217).

Kucherenko, T., Jonell, P., Van Waveren, S., Henter, GE., Alexandersson, S., Leite, I., & Kjellström, H. (2020). Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the 2020 international conference on multimodal interaction* (pp. 242–250).

Kucherenko, T., Nagy, R., Yoon, Y., Woo, J., Nikolov, T., Tsakov, M., & Henter, GE. (2023). The genea challenge 2023: A large-scale evaluation of gesture generation models in monadic and dyadic settings. In *Proceedings of the 25th international conference on multimodal interaction* (pp. 792–801).

Lee, G., Deng, Z., Ma, S., Shiratori, T., Srinivasa, SS., & Sheikh, Y. (2019). Talking with hands 16.2 m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 763–772).

Levine, S., Krähenbühl, P., Thrun, S., & Koltun, V. (2010). Gesture controllers. In Acm siggraph 2010 papers (pp. 1–11).

Li, YA., Han, C., & Mesgarani ,N. (2022). Styletts: A style-based generative model for natural and diverse text-to-speech synthesis. arXiv preprint arXiv:2205.15439.

Liu, J., Wang, X., Fu, X., Chai, Y., Yu, C., Dai, J., & 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* (pp. 6734–6743).

Liu, X., Wu, Q., Zhou, H., Xu, Y., Qian, R., Lin, X., Zhou, X., Wu, W., Dai, B., & Zhou, B. (2022a). Learning hierarchical cross-modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10462–10472).

Liu, X., Xu, Y., Wu, Q., Zhou, H., Wu, W., & Zhou, B. (2022b). Semantic-aware implicit neural audio-driven video portrait generation. In *European conference on computer vision* (Springer, pp. 106–125).

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., & Black, M. J. (2015). SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics, (Proc SIGGRAPH Asia), 34*(6), 248:1-248:16.

Lu, Y., Chai, J., & Cao, X. (2021). Live speech portraits: Real-time photorealistic talking-head animation. *ACM Transactions on Graphics (TOG), 40*(6), 1–17.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. *Interspeech, 2017*, 498–502.

Nagrani, A., Chung, J.S., & Zisserman, A. (2017). Voxceleb: A large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612.

Ng, E., Joo, H., Hu, L., Li, H., Darrell, T., Kanazawa, A., & 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* (pp. 20395–20405).

Palmero, C., Selva, J., Smeureanu, S., Junior, J., Jacques, C., Clapés, A., Moseguí, A., Zhang, Z., Gallardo, D., Guilera, G., et al. (2021). Context-aware personality inference in dyadic scenarios: Introducing the udiva dataset. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 1–12).

Palmero, C., Barquero, G., Junior, JCJ., Clapés, A., Núnez, J., Curto, D., Smeureanu, S., Selva, J., Zhang, Z., Saeteros, D., et al. (2022) Chalearn lap challenges on self-reported personality recognition and non-verbal behavior forecasting during social dyadic interactions: Dataset, design, and results. In *Understanding social behavior in dyadic and small group interactions, PMLR* (pp. 4–52).

Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, AAA., Tzionas, D., & Black, M.J. (2019) Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE conference on computer vision and pattern recognition (CVPR)*.

Paysan, P., Knothe, R., Amberg, B., Romdhani, S., & Vetter, T. (2009). A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance* (Ieee, pp. 296–301).

Petrovich, M., Black, MJ., & Varol, G. (2021a). Action-conditioned 3D human motion synthesis with transformer VAE. In *International Conference on Computer Vision (ICCV)*.

Petrovich, M., Black, M.J., & Varol, G. (2021b). Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10985–10995).

Qian, S., Tu, Z., Zhi, Y., Liu, W., & Gao, S. (2021). Speech drives templates: Co-speech gesture synthesis with learned templates. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 11077–11086).

Richard, A., Zollhöfer, M., Wen, Y., De la Torre, F., & Sheikh, Y. (2021). Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1173–1182).

Sahidullah, M., & Saha, G. (2012). Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition. *Speech Communication, 54*(4), 543–565.

Sargin, M. E., Yemez, Y., Erzin, E., & Tekalp, A. M. (2008). Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30*(8), 1330–1345.

Song, L., Yin, G., Jin, Z., Dong, X., & Xu, C. (2023a). Emotional listener portrait: Realistic listener motion simulation in conversation. arXiv preprint arXiv:2310.00068.

Song, S., Spitale, M., Luo, C., Barquero, G., Palmero, C., Escalera, S., Valstar, M., Baur, T., Ringeval, F., Andre, E., et al. (2023b). React2023: the first multi-modal multiple appropriate facial reaction generation challenge. arXiv preprint arXiv:2306.06583.

Sun, Q., Wang, Y., Zeng, A., Yin, W., Wei, C., Wang, W., Mei, H., Leung, CS., Liu, Z., Yang, L., et al. (2024). Aios: All-in-one-stage expressive human pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1834–1843).

Suwajanakorn, S., Seitz, S. M., & Kemelmacher-Shlizerman, I. (2017). Synthesizing obama: Learning lip sync from audio. *ACM Transactions on Graphics (ToG), 36*(4), 1–13.

Takeuchi, K., Kubota, S., Suzuki, K., Hasegawa, D., & Sakuta, H. (2017). Creating a gesture-speech dataset for speech-based automatic gesture generation. In: HCI International 2017–Posters' Extended Abstracts: 19th International Conference, HCI International 2017, Vancouver, BC, Canada, July 9–14, 2017, Proceedings, Part I 19, Springer, pp 198–202.

Tuyen, NTV., & Celiktutan, O. (2022). Agree or disagreeÆ' generating body gestures from affective contextual cues during dyadic interactions. In *2022 31st IEEE international conference on robot and human interactive communication (RO-MAN)* (IEEE, pp. 1542–1547).

Tuyen, N. T. V., & Celiktutan, O. (2023). It takes two, not one: Context-aware nonverbal behaviour generation in dyadic interactions. *Advanced Robotics, 37*(24), 1552–1565.

Tuyen, N.T.V., Georgescu, A.L., Di Giulio, I., & Celiktutan, O. (2023). A multimodal dataset for robot learning to imitate social human-human interaction. In *Companion of the 2023 ACM/IEEE international conference on human-robot interaction* (pp. 238–242).

Van Den Oord, A., Vinyals, O., et al. (2017). Neural discrete representation learning. *Advances in Neural Information Processing Systems 30*.

Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., He, R., Qiao, Y., & Loy, CC. (2020). Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*.

Wang, T.C., Mallya, A., Liu, M.Y. (2021). One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10039–10049).

Wu, X., Hu, P., Wu, Y., Lyu, X., Cao, Y.P., Shan, Y., Yang, W., Sun, Z., & Qi, X. (2023). Speech2lip: High-fidelity speech to lip generation by learning from a short video. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 22168–22177).

Wuu, Ch., Zheng, N., Ardisson, S., Bali, R., Belko, D., Brockmeyer, E., Evans, L., Godisart, T., Ha, H., Huang, X., et al. (2022). Multiface: A dataset for neural face rendering. arXiv preprint arXiv:2207.11243.

Xu, C., Zhu, J., Zhang, J., Han, Y., Chu, W., Tai, Y., Wang, C., Xie, Z., & Liu, Y. (2023). High-fidelity generalized emotional talking face generation with multi-modal emotion space learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6609–6619).

Ye, Z., Jiang, Z., Ren, Y., Liu, J., He, J., & Zhao, Z. (2023). Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. arXiv preprint arXiv:2301.13430.

Yi, H., Liang, H., Liu, Y., Cao, Q., Wen, Y., Bolkart, T., Tao, D., & Black, M.J. (2023). Generating holistic 3d human motion from speech. In *CVPR*.

Yin, L., Wang, Y., He, T., Liu, J., Zhao, W., Li, B., Jin, X., & Lin, J. (2023). Emog: Synthesizing emotive co-speech 3d gesture with diffusion model. arXiv preprint arXiv:2306.11496.

Yoon, Y., Ko, WR., Jang, M., Lee, J., Kim, J., & Lee, G. (2019). Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 international conference on robotics and automation (ICRA)* (IEEE, pp. 4303–4309).

Yoon, Y., Cha, B., Lee, J. H., Jang, M., Lee, J., Kim, J., & Lee, G. (2020). Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG), 39*(6), 1–16.

Yoon, Y., Wolfert, P., Kucherenko, T., Viegas, C., Nikolov, T., Tsakov, M., & Henter, GE. (2022). The genea challenge 2022: A large evaluation of data-driven co-speech gesture generation. In *Proceedings of the 2022 international conference on multimodal interaction* (pp. 736–747).

Zhang, H., Tian, Y., Zhang, Y., Li, M., An, L., Sun, Z., & Liu, Y. (2023a). Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., & Shen, X. (2023b). T2m-gpt: Generating human motion from textual descriptions with discrete representations. arXiv preprint arXiv:2301.06052.

Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., & Wang, F. (2023c). 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* (pp. 8652–8661).

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

Zheng, F., Zhang, G., & Song, Z. (2001). Comparison of different implementations of mfcc. *Journal of Computer Science and Technology, 16*, 582–589.

Zhi, Y., Cun, X., Chen, X., Shen, X., Guo, W., Huang, S., & Gao, S. (2023). Livelyspeaker: Towards semantic-aware co-speech gesture generation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 20807–20817).

Zhou, M., Bai, Y., Zhang, W., Yao, T., Zhao, T., & Mei, T. (2022). Responsive listening head generation: a benchmark dataset and baseline. In *European conference on computer vision* (Springer, pp. 124–142).

Zhu, L., Liu, X., Liu, X., Qian, R., Liu, Z., & Yu, L. (2023). Taming diffusion models for audio-driven co-speech gesture generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10544–10553).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.