MimicTalker: A Multimodal Interactive and Memory-Enhanced Framework for Real-Time Dyadic 3D Head Generation

Yinuo Wang[1,*,‡] Yanbo Fan[2,*,†] Xuan Wang[1] Boyao Zhou[3] Yu Guo[1,†] Yujun Shen[3] Fei Wang[1]
[1]State Key Laboratory of Human-Machine Hybrid Augmented Intelligence,
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
[2] Nanjing University
[3]Ant Group
nuo1wang.github.io/MimicTalker

Abstract

Dyadic interactive head generation aims to synthesize realistic head motions that respond both verbally and non-verbally to an interlocutor in real-time conversation. The existing works often focus on offline scenarios, and struggle with a shallow understanding of the multimodal conversational context while also lacking long-term coherence. To address these limitations, we propose MimicTalker, a novel method for producing real-time, contextually-aware, and long-term consistent interactive head motions. To this end, we propose a Multimodal Interactive Context Extraction (MICE) module to capture both instantaneous and long-term multimodal interactive information from the interlocutor. To enhance in-depth conversational understanding, we propose a Semantic-enhanced Dynamic Interaction (SDI) module to integrate the intentions and topics of the conversation, which are automatically extracted through an LLM-based analyzer. Further, we propose a semantic-guided Motion Style Memory (MSM) mechanism, enabling the long-term motion consistency throughout the conversation. We conduct experiments on both short conversational segments (25 seconds) and extended dialogues (6 minutes), and the comprehensive experiments demonstrate that our method significantly outperforms existing approaches.

1. Introduction

Real-time dyadic interactive 3D head generation aims to synthesize realistic head motions that respond to an interlocutor both verbally and non-verbally, mimicking real-world face-to-face conversations. It is vital for fostering effective communication between humans and conversational agents, and has promising applications in areas such as educational platforms, online customer service, human-computer interaction, and entertainment.

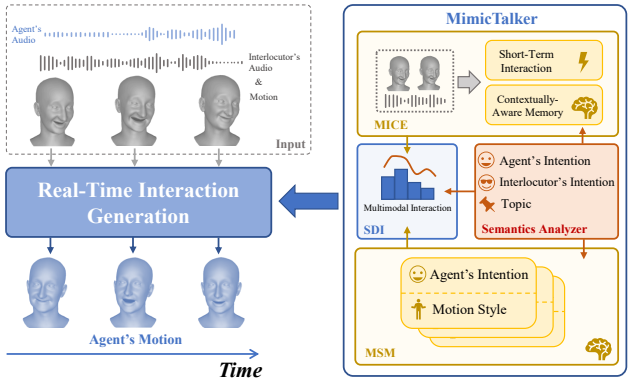

Figure 1. MimicTalker dynamically analyzes and utilizes the multimodal context of a real-time conversation. It consists of MICE to capture both short-term and contextually related long-term information of the interlocutor, SDI to dynamically integrate multimodal context features, and MSM to efficiently maintain long-term motion style consistency. As a result, MimicTalker produces natural, consistent, and seamless reactions to the interlocutor's multimodal input in real time.

In recent years, many studies have been conducted on talking head generation [7, 11, 28, 29, 36, 40] and listening head generation [20, 21, 25, 33, 39, 45]. Talking head generation synchronizes the generated head motion with the input audio but completely ignores the response to the interlocutor. Listening head generation, while focusing on the interaction between two interlocutors, only generates non-verbal responses such as nodding or smiling. These limitations restrict usability in real-world conversations, and therefore, attention has been increasingly drawn to dyadic interactive head generation [6, 14, 23, 26, 34, 37, 46].

*Equal contribution
†Corresponding authors
‡This work was done during Yinuo Wang's internship at Ant Group

Despite being a more practical and universal way of enabling seamless human-agent interaction, real-time dyadic interactive head generation is very challenging and should meet the following criteria. Firstly, it should be able to integrate the interlocutor's multimodal conversational signals (including audio, motions, and semantics) and generate the responsive head motions in real time, *e.g.*, if the interlocutor bursts into laughter, the interactive head should perceive this in real time and give its response. Secondly, it should be aligned with the in-depth conversational semantics, including the conversational topic and intentions of the agent. Thirdly, it should be able to generate long and consistent motion sequences, *e.g.*, a real conversation can be lengthy, and the interactive head should maintain its personal style as long as the conversation lasts. However, most of the existing dyadic interactive head generation methods [26, 30, 34, 37, 46] are designed for offline generation. These methods generate interactive motions in clips, which require all the information in the clip as a prerequisite and thus do not support real-time interaction. As for those designed for real-time interaction, they either underestimate the complex interaction between interlocutors [6, 23] or overlook the in-depth understanding of the conversation [14], resulting in interactions that are less realistic or natural in motion. Moreover, existing methods also oversimplify the generation of longer sequences, which often compromises consistency over extended durations.

To tackle these challenges, we propose MimicTalker, a novel framework for 3D real-time dyadic interactive head generation, as illustrated in Fig. 1. MimicTalker leverages specially designed semantic and memory modules to deeply understand conversational dynamics and generate realistic, style-consistent responses during extended interactions. Specifically, to enable real-time interaction, the framework incorporates the Multimodal Interactive Context Extraction (MICE) module, which processes the interlocutor's multimodal input in a causal manner. The module captures both frame-level instantaneous interactions and contextually related long-term historical interactions, ensuring that the generated reactions are both timely and contextually appropriate. In order to further enhance the framework with in-depth understanding of the conversation context, we introduce the Semantics-enhanced Dynamic Interaction (SDI) module, which integrates various types of semantic information, such as intentions and conversation topic, while continuously interacting with the interlocutor's features in real time. A large language model (LLM) based Automatic Semantics Analyzer is employed to automatically analyze both historical and ongoing conversations, providing detailed semantic insights, including each interlocutor's intentions and the conversation topic. These semantic features are carefully integrated into the corresponding parts of the network, based on their characteristics. To ensure long-term motion consistency, we propose the semantic-guided Motion Style Memory (MSM), which maintains the interactive head's historical intentions and their associated motion styles throughout the conversation. The MSM enables efficient retrieval of the motion style that best corresponds to the current intention and uses them as guidance, allowing the interactive head to maintain a consistent style for the same intention. Consequently, MimicTalker is capable of generating long, contextually appropriate, and style-consistent motion sequences that respond to the interlocutor in real time with natural and coherent reactions.

In brief, our main contributions are as follows:

- We propose a novel framework for dyadic interactive 3D head generation in real-time conversation, producing realistic, coherent, and contextually-aware head motions.
- We design a MICE module and a SDI module to effectively extract and integrate the multimodal interaction information for both interlocutors.
- We introduce a novel semantic-guided Motion Style Memory to efficiently maintain long-term consistency of the generated head motion.
- We conduct comprehensive experiments and demonstrate significant improvements in dyadic interactive head generation compared with existing methods.

2. Related Works

2.1. Talking Head Generation

Talking head generation is an important research area in computer vision, and numerous studies have explored this topic. Some works focus on 2D talking head generation, where the goal is to synthesize realistic facial movements synchronized with audio [7, 16, 19, 36, 41, 43]. Benefiting from the recent advances in diffusion models [32, 42], state-of-the-art methods can generate photorealistic talking head videos [9, 10], but at the cost of long generation time.

Another line of work focuses on generating 3D talking head by animating a 3D face model based on audio. Works such as FaceFormer [11], CodeTalker [40], and SelfTalk [28] are capable of generating accurate lip-synchronized animations. EmoTalk [29] enhances the expressiveness of the animation by generating emotional facial expressions. ARTalk [8] increases the generation speed of talking head models while maintaining realistic and natural head motion. Lee *et al*. [4] enhances lip synchronization through audio-visual representation learning.

However, these methods are designed for one-sided communication, and are mainly built for offline scenarios.

2.2. Listening Head Generation

In contrast to talking head generation, listening head generation is set in a dyadic conversation scenarios and generates non-verbal feedback for the interlocutor. Earlier work

adopts rule-based methods [2, 3, 13], and is limited to predefined reaction motion sequences. Recent studies have turned to learning-based methods through data-driven approaches to automatically learn to generate appropriate reactions based on the speaker's conditions, improving motion diversity and expressiveness [5, 18, 20, 21, 25, 33, 39, 44, 45]. Even though some can generate vivid and realistic responses, they are for non-verbal response only, and are limited to generating short sequences in offline settings.

A simplistic dyadic head generation paradigm can be obtained by combining existing talking head generation and listening head generation. However, this approach fails to model dynamic interactions between interlocutors or enable smooth and natural speaker-listener role switching.

2.3. Dyadic Interactive Head Generation

As a unified framework that seamlessly integrates talking head generation and listening head generation, dyadic interactive head generation has drawn increasing interest from the research community. Some existing works [26, 30, 34, 37, 46] focus on the offline generation of the interlocutors' motions while exploring the interactions between them. Among them, [34, 37] explicitly assign the role of speaker and listener, which reduces usability and leads to stiff transitions. While [26, 30, 46] can freely switch between listening and speaking states, they require the full sequence of both interlocutors and generate head motion video segments in bulk. As a result, they are difficult to implement in real-world interactions where real-time feedback is required. Real-time interaction scenarios are less explored in existing research. AV-Flow [6] generates both audio and head motion from text transcription, but does not effectively capture in-depth interactions between interlocutors. OmniResponse [23] fine-tunes an LLM to generate audio and head-motion responses based on the interlocutor's visual and audio input. However, due to the modality gap, it struggles to generate head motion responses synchronized with dialogue audio, as evidenced by the demo videos on their project page. ARIG [14] is the most relevant to our work. It uses a lightweight diffusion network to generate interlocutor's head motion, given the two interlocutors audio and the other interlocutor's head motion. ARIG's limitation is that it neglects the in-depth understanding of the semantics of the conversation, and similar to other methods, the long-term consistency of the generated sequence. We demonstrate both of those are key to generating realistic and coherent head motions.

3. Method

3.1. Overview

In this section, we introduce MimicTalker, a 3D talking head generation framework for real-time dyadic interactions. The network takes Speaker A's audio ($\mathbf{A}_A$), Speaker B's audio ($\mathbf{A}_B$) and head motion ($\mathbf{M}_B$) as input, and generates Speaker A's head motion ($\hat{\mathbf{M}}_A$) that synchronizes with A's audio as well as reacts to B's verbal cues (audio) and non-verbal cues (head motion) in real time. The process can be formally defined as follows,

$$\hat{\mathbf{M}}_A = \mathrm{f}(\mathbf{A}_B, \mathbf{M}_B, \mathbf{A}_A). \tag{1}$$

The network f models the dynamic interactions between Speaker A and Speaker B with a causal structure to enable A's real-time response to B. Enhanced with a deep understanding of the conversation semantics and specially designed memory modules, the network can model each interlocutor's motion based on the flow of the conversation and the interlocutor's intentions, resulting in better consistency and realism over the long term. As shown in Fig. 2, the network consists of four main components: Multimodal Interactive Context Extraction, Semantics-Enhanced Dynamic Interaction, Semantic-Guided Motion Style Memory, and Automatic Semantics Analyzer.

3.2. Multimodal Interactive Context Extraction

MICE module enables the network to perceive Speaker B's multimodal features in real time. The existing methods process the interlocutor's features in bulk within a time window, which introduces an innate time delay that cannot be avoided due to the dependency on future context. In contrast, our method processes Speaker B's features frame by frame in real time, eliminating the innate latency and enabling responsive interactions. Moreover, enhanced with the Contextually-Aware Memory, the module can model historical interactions that are closely connected to Speaker B's intention.

Instantaneous Frame-Level Perception. Given Speaker B's audio $\mathbf{A}_B$ and head motion $\mathbf{M}_B$, we first project them into a shared feature space using MLPs,

$$\mathbf{H}_B = \mathrm{MLP}(\mathrm{MFCC}(\mathbf{A}_B)), \mathbf{M}'_B = \mathrm{MLP}(\mathbf{M}_B), \tag{2}$$

where MFCC stands for MFCC feature extraction. Both Speaker B's audio and visual features are extracted frame-wise to avoid the dependency on the future, thus eliminating the innate time delay that hinders real-time interaction. Then, Speaker B's features are merged using several cross attention layers and self attention layers, enabling the network to perceive short-term dependencies between audio-visual modalities, better capturing their correlations. Causal masks $\mathcal{M}$ are applied to force the causality,

$$\mathbf{Z}_B = \mathrm{SelfAttn}(\mathrm{CrossAttn}(\mathbf{H}_B, \mathbf{M}'_B, \mathcal{M}), \mathcal{M}). \tag{3}$$

where $\mathbf{Z}_B$ is the instantaneous frame-level feature of Speaker B that aligns the visual representation with the acoustic cues dynamically.

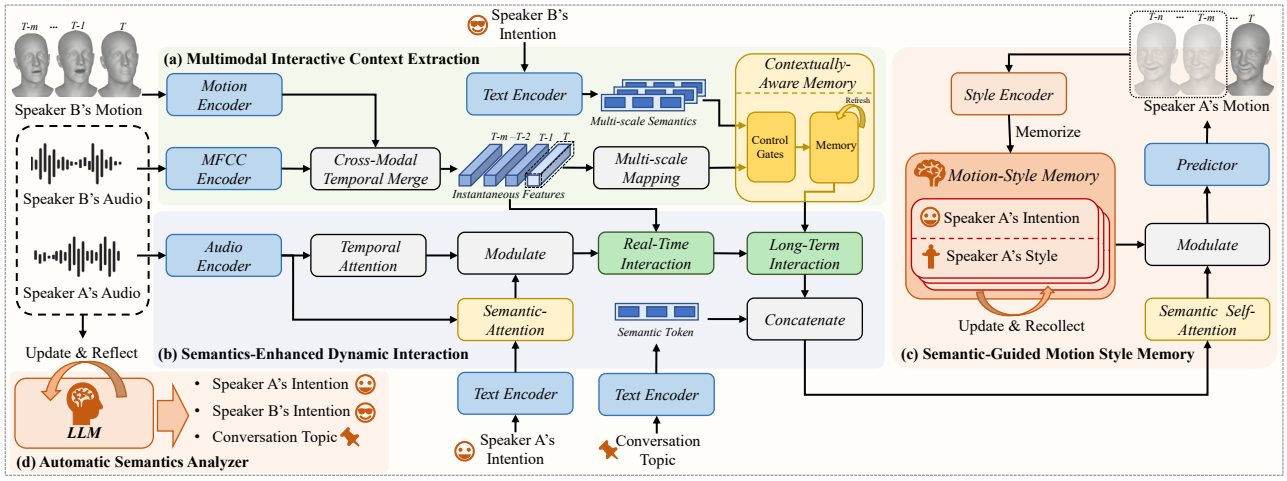

Figure 2. Overview of MimicTalker. Incorporating in-depth semantics of the conversation, the framework dynamically interacts with interlocutor's multimodal features and generates realistic and vivid head motions in real time. The key components are (a) Multimodal Interactive Context Extraction, (b) Semantics-Enhanced Dynamic Interaction, (c) Semantic-Guided Motion Style Memory, and (d) Automatic Semantics Analyzer.

Contextually-Aware Memory. The instantaneous nature of $\mathbf{Z}_B$ constrains that it only contains information of Speaker B within a very short period. To successfully model the dynamics of interaction, the long-term information is crucial. We design a multi-scale memory module that stores useful information from instantaneous feature $\mathbf{Z}_B$ into a memory space. At the same time, semantic features from Speaker B's intention $\mathbf{I}_B$ guides the memory update to be closely connected with the current conversation flow. The memory is updated as follows,

$$\begin{aligned}
&\mathbf{z}_B^{(t)} = \mathrm{MLP}(\mathbf{Z}_B^{(t)}),\ \mathbf{i}_B = \mathrm{MLP}((\mathrm{Pool}(\mathbf{I}_B)),\\
&\mathbf{h}^{(t)} = \mathrm{Concat}(\mathbf{i}_B, \mathbf{z}_B^{(t)}, \mathbf{m}^{(\mathbf{t})}),\\
&\mathbf{g}_i^{(t)} = \mathrm{MLP}(\mathbf{h}^{(t)}),\ \mathbf{g}_f^{(t)} = \mathrm{MLP}(\mathbf{h}^{(t)}),\\
&\mathbf{m}^{(t+1)} = \mathbf{g}_i^{(t)}\,\mathrm{MLP}(\mathbf{h}^{(t)}) + \mathbf{g}_f^{(t)}\mathbf{m}^{(t)},
\end{aligned} \tag{4}$$

where $\mathbf{m}^{(t)} \in \mathbb{R}^{K\times d}$ is the memory feature at frame time t with a dimension of d. $\mathbf{g}_i^{(t)}$ and $\mathbf{g}_f^{(t)}$ are control gates at frame time t that decide which information is memorized and which is forgotten. We use a memory with a size of K to capture multi-scale information from different representation subspaces [38]. The multi-scale features of $\mathbf{Z}_B^{(t)}$ and $\mathbf{I}_B^{(t)}$ are extracted by MLPs. The memory is updated using historical information with semantic guidance, allowing the network to retain and aggregate relevant contextual details over time across multiple scales. As a result, the interaction becomes more realistic and contextually aware.

3.3. Semantics-Enhanced Dynamic Interaction

SDI module captures the temporal correlation between Speaker A's multimodal features, and simultaneously interacts with Speaker B's features in real time. By integrating Speaker A's intention and the conversation topic, we enhance the framework with in-depth understanding of the conversation, enabling it to generate more socially appropriate head motions.

Given Speaker A's raw audio signals $\mathbf{A}_A$, we first extract its high dimensional feature representation $\mathbf{H}_A$ using a pretrained Whisper encoder [31]. Next, convolutional layers and self-attention layers align the audio features to the frame level and refine the temporal dynamics,

$$\mathbf{Z}_A = \mathrm{SelfAttn}(\mathrm{Conv}(\mathbf{H}_A), \mathcal{M}), \tag{5}$$

where $\mathbf{Z}_A \in \mathbb{R}^{N\times d}$ is the Speaker A's audio feature aligned frame-wise with Speaker B's head motion, $\mathcal{M}$ is a causal mask.

The network cannot use Speaker A's intention effectively without knowing when it should be expressed through head motion. By modeling the correlation between Speaker A's audio and intention, we enable the network to learn the head motion suitable for the intention, and more importantly, the appropriate time to express or emphasize the intention through motion. With this motivation, we first use cross attention mechanism to obtain the frame-wise attention features between $\mathbf{Z}_A$ and Speaker A's intention feature $\mathbf{I}_A$,

$$\mathbf{C}_A = \mathrm{CrossAttn}(\mathbf{H}_A, \mathbf{I}_A), \tag{6}$$

where $\mathbf{H}_A$ is used as query and $\mathbf{I}_A$ as key and value. The yielding $\mathbf{C}_A$ contains the frame-wise correlation information between intention and audio. A higher correlation at a frame indicates that it is more likely the intention feature will affect the head motion corresponding to that frame. $\mathbf{C}_A$

is then integrated into Speaker A's audio feature $\mathbf{H}_A$ with adaptive layer norm (adaLN) blocks [27],

$$(\gamma_A, \beta_A) = \text{MLP}(\mathbf{C}_A), \mathbf{F}_A = (1 + \gamma_A)\mathbf{H}_A + \beta_A, \quad (7)$$

where $\mathbf{F}_A \in \mathbb{R}^{N \times d}$ is Speaker A's audio feature enhanced dynamically with semantics.

Then, we use Interaction Blocks that consists of cross attention and adaLN, to integrate Speaker B's instantaneous feature $\mathbf{Z}_B$ and long-term feature $\mathbf{m}^{(t)}$ with Speaker A's semantics enhanced feature $\mathbf{F}_A$ frame-wise, and obtain the interaction feature $\mathbf{F}_{AB}$. This enables the dynamic interaction of features between the two interlocutors in both short-term and long-term contexts.

We view the conversation topic as general guidance for determining the conversational atmosphere and use it to ensure the generated motion remains contextually relevant. Since the topic feature provides an overarching context rather than precise, frame-level guidance like the intention feature, it is treated as a global condition. As a result, we concatenate the conversation topic with the interaction feature $\mathbf{F}_{AB}$ and use causal self-attention to enable the network to attend to the topic feature,

$$\mathbf{F} = \text{SelfAttn}(\text{Concat}(\mathbf{T}, \mathbf{F}_{AB}), \mathcal{M}), \quad (8)$$

where $\mathbf{T}$ is the topic feature, $\mathcal{M}$ is a causal mask.

3.4. Semantic-guided Motion Style Memory

The existing methods for generating dyadic head motion either process at the segment level without connections between segments or condition the current segment on the previous one. In the first case, long-term dependencies beyond the segment length cannot be captured. In the second, while slightly longer-term dependencies are modeled, historical information still fades over time, leading to inconsistent motion styles over extended periods. To address this, we propose MSM, which utilizes an external memory bank to store all of Speaker A's past motion styles, and any stored motion style can then be queried and used as guidance when generating a new segment. With this design, our framework can efficiently maintain a consistent motion style unlimited by sequence length. Compared with the methods that compromise long-term consistency to reduce computational complexity, our method stores information that is highly compressed and representative, and allows the network to efficiently attend to any historical information in storage. This ensures both efficiency and effectiveness in maintaining consistent motion over long sequences.

Specifically, we extract motion style for every segment with a length of w, using a style encoder [24] trained with contrastive loss [35], and construct an external memory bank for the conversation session. For each segment, we extract Speaker A's intention as well as its corresponding motion style, and store them to the memory, where the intention serves as the key, and the motion style as the value. During inference, we use Speaker A's current intention as the query, search the memory bank for the most similar historical intention, and use the corresponding motion style as guidance for Speaker A's head motion generation. If there is no similar intention in the memory bank, *e.g.* when the conversation starts with an empty memory bank, or when all stored intentions and the query intention have a similarity score that falls below a threshold, we feed a segment of all-zero motion to the style encoder to obtain the default style code.

Next, the motion style feature $\mathbf{s}$ obtained by searching memory acts as guidance via adaLN blocks,

$$(\gamma_s, \beta_s) = \text{MLP}(\mathbf{s}), \mathbf{F}' = (1 + \gamma_s)\mathbf{F} + \beta_s. \quad (9)$$

Finally, we feed $\mathbf{F}'$ to an MLP to predict Speaker A's head motion $\hat{\mathbf{M}}_A$.

3.5. Automatic Semantics Analyzer

One major limitation of the existing works is that they only use low-level information from the conversation to model the dyadic conversation, which over-simplifies the complexity of the conversation and often leads to an emotion discrepancy between the generated motion and the input audio. To address this, we leverage an LLM to automatically analyze the conversation at a higher level, yielding the intention of each interlocutor and the topic of the conversation. The interlocutors' intentions help the network better understand social interactions on a semantic level, enabling it to capture the subtle nuances of interpersonal dynamics. Meanwhile, the conversation topic provides general guidance, ensuring the generated motion remains contextually relevant and thematically consistent.

First, we transcribe both Speaker A's and Speaker B's audio with an off-the-shelf tool [31] to obtain the context of the conversation. Then, for every k rounds of conversation, an LLM analyzes the dialogue and infers the two interlocutors' intentions and the conversation's topic. If any of the contextual information cannot be inferred, *e.g.*, at the start of a conversation when data is insufficient, the corresponding information is set to a default sentence (*e.g.*, *the intention is currently undetermined*). Finally, we use Roberta [22] to extract features from the contextual information, yielding semantics features $\mathbf{I}_A$, $\mathbf{I}_B$ and $\mathbf{T}$ for Speaker A's intention, Speaker B's intention, and conversation topic respectively. These features are integrated in the three aforementioned main components of the network to boost the alignment between the generated motion as well as the semantic and emotional context of the interaction.

Methods	FD↓			P-FD↓			MSE↓			SID↑			rPCC↓		
	EXP	JAW	POSE	EXP	JAW	POSE	EXP	JAW	POSE	EXP	JAW	POSE	EXP	JAW	POSE
		$\times 10^3$	$\times 10^2$		$\times 10^3$	$\times 10^2$	$\times 10^1$	$\times 10^3$	$\times 10^2$				$\times 10^2$	$\times 10^1$	$\times 10^1$
FaceFormer [11]	34.90	5.40	8.00	34.90	5.40	8.00	6.97	1.80	2.67	0.54	0.36	0.50	13.05	2.41	5.27
CodeTalker [40]	48.57	6.89	10.74	48.57	6.89	10.74	9.71	2.29	3.58	0	0	0	11.06	2.33	5.11
EmoTalk [29]	29.86	4.33	7.54	30.20	4.36	7.58	6.88	1.76	2.59	2.86	1.72	0.98	9.89	2.19	4.94
SelfTalk [28]	35.77	5.49	8.14	35.77	5.49	8.14	7.15	1.83	2.71	2.49	1.30	1.28	12.25	2.39	4.70
L2L [25]	24.61	3.69	7.08	24.99	3.74	7.13	5.68	1.48	2.49	2.86	1.89	1.19	8.52	2.06	4.11
Dualtalk [30]	11.14	1.90	3.83	11.88	1.97	3.97	3.59	1.04	1.72	3.48	2.23	1.72	4.73	1.37	2.38
MimicTalker	**7.12**	**1.32**	**2.71**	**8.20**	**1.41**	**2.93**	**3.09**	**0.93**	**1.67**	**3.63**	**2.40**	**1.90**	**4.16**	**1.18**	**2.03**
FaceFormer [11]	35.92	5.39	8.60	35.93	5.39	8.60	7.18	1.80	2.87	0.54	0.40	0.51	11.71	2.16	5.73
CodeTalker [40]	50.05	6.95	11.66	50.05	6.95	11.66	10.01	2.32	3.88	0	0	0	10.24	2.18	5.76
EmoTalk [29]	34.12	4.17	8.59	34.44	4.21	8.62	7.73	1.71	2.94	2.89	1.79	0.94	9.44	1.96	5.54
SelfTalk [28]	36.23	5.36	8.89	36.23	5.36	8.89	7.24	1.79	2.96	2.61	1.36	1.08	11.26	2.13	5.67
L2L [25]	30.49	3.82	8.56	30.87	3.86	8.61	6.87	1.54	2.98	2.76	1.91	1.11	9.02	1.94	4.99
Dualtalk [30]	21.71	3.15	5.89	22.56	3.22	6.06	5.97	1.50	2.48	2.98	1.94	1.38	6.86	1.60	3.28
MimicTalker	**15.89**	**2.27**	**4.15**	**17.25**	**2.38**	**4.42**	**5.28**	**1.36**	**2.29**	**3.04**	**2.11**	**1.57**	**6.65**	**1.48**	**2.59**

Table 1. Quantitative results on DualTalk dataset. The ↑ indicates higher is better for the corresponding metric while ↓ indicates lower is better. The top half is the result on the test set, and the bottom half is the result on the OOD set. The best and the second best results are in bold and underlined respectively.

Methods	FD↓			P-FD↓			MSE↓			SID↑			rPCC↓		
	EXP	JAW	POSE	EXP	JAW	POSE	EXP	JAW	POSE	EXP	JAW	POSE	EXP	JAW	POSE
		$\times 10^3$	$\times 10^2$		$\times 10^3$	$\times 10^2$	$\times 10^1$	$\times 10^3$	$\times 10^2$				$\times 10^2$	$\times 10^1$	$\times 10^1$
DualTalk	30.13	4.72	9.43	30.90	4.77	9.58	7.94	2.16	3.78	**2.66**	1.80	1.11	6.77	3.06	4.27
Baseline	27.28	4.87	7.56	28.91	5.05	7.84	8.50	2.51	3.99	2.34	1.75	1.36	5.75	3.75	2.80
with MICE	24.75	4.73	7.49	26.37	4.90	7.78	8.08	2.49	4.13	2.55	1.77	1.39	5.51	3.69	**2.48**
with SDI	26.10	4.36	7.57	27.73	4.54	7.86	8.31	2.37	4.12	2.46	1.79	1.43	5.63	3.57	2.56
with MSM	25.58	3.59	6.87	26.98	3.67	7.14	7.51	1.80	3.45	2.49	1.91	1.44	**5.05**	3.03	3.13
MimicTalker	**24.52**	**3.23**	**6.56**	**25.92**	**3.31**	**6.84**	**7.28**	**1.70**	**3.35**	2.54	**2.03**	**1.46**	5.12	**2.95**	3.02

Table 2. Quantitative results on Seamless Interaction dataset.

4. Experiments

4.1. Experimental Setup

Dataset. We conduct experiments on the mulit-turn dyadic interaction dataset DualTalk [30]. The dataset contains 50 hours of real life dyadic conversation videos of 1,052 unique speakers, with 4,935 clips in the training set, 539 clips in the test set and 384 clips in the out-of-distribution (OOD) set. The maximum and average clip lengths in the DualTalk dataset are 40 seconds and 25 seconds, respectively. Additionally, we randomly sample 20 sequences that are over 5 minutes long from a large-scale dyadic interaction dataset Seamless Interaction [1] to test the performance on longer conversations. The motion sequences used to train and test our model are extracted by Spectre [12].

Compared Methods and Evaluation Metrics. We compare MimicTalker with several state-of-the-art methods for talking head and dyadic interaction generation, including FaceFormer [11], CodeTalker [40], EmoTalk [29], SelfTalk [28], L2L [25] and DualTalk [30]. We retrain these methods on Dualtalk dataset. Following DualTalk, we use evaluation metrics including Fréchet Distance (FD), Paired Fréchet Distance (P-FD), Mean Squared Error (MSE), SI for Diversity (SID), and Residual Pearson Correlation Coefficient (rPCC), to assess motion realism, expression accuracy, motion diversity and synchronization.

Implementation Details. All methods, including MimicTalker are trained on the DualTalk dataset. For our method, we use the Adam [17] optimizer with a batch size of 32. Intentions and conversation topics are extracted automatically using GPT-4o [15]. The segment length w for motion style update is 20 seconds, and the conversation rounds k for intention and topic update are set to 3. Since the clips in the DualTalk dataset are short (25 seconds in average), the topic and intentions within each clip remain constant, and insufficient for constructing a motion style memory bank for each clip. Therefore, during training, we feed the transcript of the whole clip to the LLM and obtain the topic and intentions. The motion styles are extracted from all other clips in the training set associated with the same speaker identity, excluding the current clip. To ensure robust performance when no valid motion style is retrieved from MSM, we randomly set the style motion sequence to all-zeros at a probability of 0.1 during training. Similarly, we assign default sentences to the intention and topic independently with a probability of 0.1.

4.2. Quantitative Results

Compared Methods on DualTalk Dataset. The experimental results of MimicTalker and the compared methods

Methods	FD↓			P-FD↓			MSE↓			SID↑			rPCC↓		
	EXP	JAW $\times 10^3$	POSE $\times 10^2$	EXP	JAW $\times 10^3$	POSE $\times 10^2$	EXP $\times 10^1$	JAW $\times 10^3$	POSE $\times 10^2$	EXP	JAW	POSE	EXP $\times 10^2$	JAW $\times 10^1$	POSE $\times 10^1$
Baseline	7.95	1.47	2.85	9.08	1.56	3.09	3.23	0.98	1.69	3.51	2.36	1.86	4.58	1.22	2.16
with MICE	7.45	1.34	2.75	8.58	1.44	2.99	3.21	0.96	1.71	3.58	2.39	1.89	4.24	1.18	2.08
with SDI	7.69	1.41	2.83	8.79	1.49	3.07	3.19	0.97	1.71	3.56	2.36	1.86	4.43	1.23	2.14
with MSM	7.39	1.38	2.83	8.49	1.46	3.06	3.15	0.95	1.71	3.59	2.38	1.86	4.22	1.20	2.15
MimicTalker	**7.12**	**1.32**	**2.71**	**8.20**	**1.41**	**2.93**	**3.09**	**0.93**	**1.67**	**3.63**	**2.40**	**1.90**	**4.16**	**1.18**	**2.03**
Baseline	19.39	2.54	4.56	20.81	2.62	4.82	5.96	1.46	2.40	2.88	2.07	1.50	7.16	1.53	2.84
with MICE	18.40	2.37	4.45	19.85	2.48	4.73	5.91	1.45	2.46	2.95	2.09	1.54	**6.63**	1.54	2.75
with SDI	18.62	2.46	4.43	20.06	2.57	4.71	5.89	1.47	2.43	2.91	2.07	1.56	6.87	1.55	2.68
with MSM	16.17	2.29	4.33	17.54	2.40	4.59	5.33	1.36	2.35	3.03	2.09	1.55	6.63	**1.46**	2.68
MimicTalker	**15.89**	**2.27**	**4.15**	**17.25**	**2.38**	**4.42**	**5.28**	**1.36**	**2.29**	**3.04**	**2.11**	**1.57**	6.65	1.48	**2.59**

Table 3. Ablation study on DualTalk dataset. The top half is the result on the test set, and the bottom half is the result on the OOD set.

on DualTalk dataset are in Tab. 1. It can be seen that our method outperforms others with respect to all tested metrics on both the test and the OOD set. For the motion realism metric FD and P-FD, the relative improvement for expression and head pose is over 30% on both test and the OOD set, proving that our method can generate more realistic and natural head motions with more convincing interactions between interlocutors. Our method achieves the lowest MSE, with an overall 20% improvement in expression, which demonstrates that our method generates more accurate expressions compared to the other methods. The highest SID and the lowest rPCC with over 10% improvement indicate that our method can generate more diverse head motions while maintaining consistent and active interactions between two interlocutors. Moreover, better result on the OOD set demonstrates the generalization capability of our method. These results verify that our method can significantly improve the head motion quality, enabling realistic and vivid dyadic interaction head generation.

Long Sequence Generation. We test MimicTalker on the Seamless Interaction dataset with real-world settings to evaluate its performance in long conversations. The test samples in the Seamless Interaction dataset are 5-11 minutes long, which is 10 to 20 times longer in duration than the longest clips in the DualTalk dataset. This allows us to strictly maintain a real-world scenario in which the topic and intentions are updated recursively and the style reference sequences are obtained from the previously generated motion. More details can be found in the supplementary material. As shown in Tab. 2, our proposed method achieves the overall best result, with significantly improved realism, accuracy, and synchronization. This demonstrates better generalization capabilities and versatility of our framework.

Real-time Capability. MimicTalker achieves a generation speed of over 300 frames per second (fps) on a single NVIDIA RTX 5000 GPU, far surpassing the 30 fps minimum typically required for real-time applications. For half-minute-long conversation segments, the Semantics Analyzer completes transcription, intention and topic analysis, and semantic feature extraction in under 3 seconds. As the Semantics Analyzer operates in parallel with the main network, it does not hinder the real-time response of the framework. Furthermore, most of the existing methods require a segment of the interlocutor's motion and audio input to respond, resulting in a longer response time equivalent to the segment's length. In contrast, our method processes the interlocutor's input frame by frame in real time, enabling better real-world interactions.

4.3. Qualitative Results

Alongside quantitative metrics, we conduct qualitative evaluations to compare MimicTalker with baseline methods *e.g.* DualTalk, EmoTalk, CodeTalker, and FaceFormer. As illustrated in Fig. 3, our method surpasses the baseline methods in terms of motion accuracy, style consistency, and emotion responsiveness. To be specific, 1) in Fig. 3 (a) and (b), our method generates correct reactions. In (a), our method captures the subtle smiling face of the interlocutor and the encouraging tone of the conversation, responding with a smile. In (b), our method reacts to the interlocutor's serious topic with a solemn expression while preserving the style of the interactive head. The result is a calm and contemplative response. 2) In Fig. 3 (c), our method successfully captures the smiling face of the silent interlocutor, demonstrating responsiveness and effective use of multimodal features. 3) In Fig. 3 (d), (e), and (f), our method generates more accurate lip movements, showcasing its proficiency in the speaking state. 4) In Fig. 3 (e), our method reacts to the interlocutor's smile with a responsive smile while accurately synchronizing lip movements during speech, highlighting its ability to deliver timely and appropriate responses in both speaking and listening states.

4.4. Ablation Study

We conduct an ablation study to demonstrate the necessity of each key module of MimicTalker. The effect of each key module is assessed by adding them to the baseline—a version of our model where all semantic- and memory-related components are removed. As shown in Tab. 3, and Tab. 2,

Figure 3. Visualization of the compared methods. Our method exhibits accurate motion, consistent style, vivid facial expressions, and coherent reactions. The top half is the result of the interlocutor leading the conversation, and the bottom half is the result of the interactive head leading the conversation, where the phonemes corresponding to the displayed frames are marked in red.

the overall performance improves significantly when each key module is added to the baseline.

Adding the MICE module (*i.e*. applying Contextually-Aware Memory to leverage Speaker B's semantics-enhanced historical features, as described in Sec. 3.2) allows the framework to generate more realistic and accurate head motions. This improvement is evident through the significantly better rPCC, emphasizing the importance of long-term historical interactions and the module's ability to effectively utilize them.

Incorporating the SDI module (*i.e*. leveraging Speaker A's intention to enhance their audio features and using conversation topic feature as a global guidance, as described in Sec. 3.3) results in more accurate and realistic facial expressions while maintaining synchronization with the interlocutor. The performance gains in FD, MSE, and rPCC demonstrate the module's ability to dynamically embed semantic features into Speaker A's features, leading to more contextually appropriate head motions.

Introducing MSM (*i.e*. using the motion style corresponding to the intention most similar to the current one as guidance, as described in Sec. 3.4) significantly improves performance on overall head motions. This improvement stems from the module's effectiveness in preserving style consistency over time.

Supplementary Materials. Due to space limitation, we provide network details, training objectives, detailed inference settings, discussion of ethics and limitations, and more experimental results in the supplementary materials.

5. Conclusion

In this paper, we propose a novel real-time interactive head generation framework that utilizes the interlocutor's multimodal input and responds both verbally and non-verbally with realistic, natural, and coherent head motions in real time. The framework perceives both instantaneous and long-term features of the interlocutor, enabling the capture of complex interaction dynamics. By integrating rich semantics of the conversation, the framework is enhanced with a deep understanding of the interaction, generating more contextually relevant head motions. With an efficient long-term style memory, the framework can maintain a consistent motion style over extended periods. Comprehensive experiments demonstrate the superior performance of the framework compared with state-of-the-art methods and the effectiveness of its key modules.

Acknowledgment

This research was funded by Key R&D Program of Shaanxi Province under Grant 2024GX-YBXM-141. This work was also supported by Ant Group Research Intern Program.

References

[1] Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, Julia Buffalini, Fabio Maria Carlucci, Joy Chen, Junming Chen, Zhang Chen, Shiyang Cheng, et al. Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset. *arXiv preprint arXiv:2506.22554*, 2025. 6

[2] Justine Cassell and Kristinn R. Thorisson. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 13(4-5):519–538, 1999. 3

[3] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*, page 413–420, New York, NY, USA, 1994. Association for Computing Machinery. 3

[4] Lee Chae-Yeon, Oh Hyun-Bin, Han EunGi, Kim Sung-Bin, Suekyeong Nam, and Tae-Hyun Oh. Perceptually accurate 3d talking head generation: New definitions, speech-mesh representation, and evaluation metrics. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21065–21074, 2025. 2

[5] Zhigang Chang, Weitai Hu, Qing Yang, and Shibao Zheng. Hierarchical semantic perceptual listener head video generation: A high-performance pipeline. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 9581–9585, New York, NY, USA, 2023. Association for Computing Machinery. 3

[6] Aggelina Chatziagapi, Louis-Philippe Morency, Hongyu Gong, Michael Zollhöfer, Dimitris Samaras, and Alexander Richard. Av-flow: Transforming text to audio-visual human-like interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14270–14282, 2025. 1, 2, 3

[7] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2403–2410, 2025. 1, 2

[8] Xuangeng Chu, Nabarun Goswami, Ziteng Cui, Hanqin Wang, and Tatsuya Harada. Artalk: Speech-driven 3d head animation via autoregressive model. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–9, 2025. 2

[9] Jiahao Cui, Yan Chen, Mingwang Xu, Hanlin Shang, Yuxuan Chen, Yun Zhan, Zilong Dong, Yao Yao, Jingdong Wang, and Siyu Zhu. Hallo4: High-fidelity dynamic portrait animation via direct preference optimization and temporal motion modulation. *arXiv preprint arXiv:2505.23525*, 2025. 2

[10] Yikang Ding, Jiwen Liu, Wenyuan Zhang, Zekun Wang, Wentao Hu, Liyuan Cui, Mingming Lao, Yingchao Shao, Hui Liu, Xiaohan Li, et al. Kling-avatar: Grounding multimodal instructions for cascaded long-duration avatar animation synthesis. *arXiv preprint arXiv:2509.09595*, 2025. 2

[11] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18770–18780, 2022. 1, 2, 6

[12] Panagiotis P Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. Spectre: Visual speech-informed perceptual 3d facial expression reconstruction from videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5755, 2023. 6

[13] Jonathan Gratch, Anna Okhmatovskaia, Francois Lamothe, Stacy Marsella, Mathieu Morales, R. J. van der Werf, and Louis-Philippe Morency. Virtual rapport. In *Intelligent Virtual Agents*, pages 14–27, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. 3

[14] Ying Guo, Xi Liu, Cheng Zhen, Pengfei Yan, and Xiaoming Wei. Arig: Autoregressive interactive head generation for real-time conversations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12956–12965, 2025. 1, 2, 3

[15] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 6

[16] Xiaozhong Ji, Xiaobin Hu, Zhihong Xu, Junwei Zhu, Chuming Lin, Qingdong He, Jiangning Zhang, Donghao Luo, Yi Chen, Qin Lin, et al. Sonic: Shifting focus to global audio perception in portrait animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 193–203, 2025. 2

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 6

[18] Shiying Li, Xingqun Qi, Bingkun Yang, Chen Weile, Zezhao Tian, Muyi Sun, Qifeng Liu, Man Zhang, and Zhenan Sun. Vividlistener: Expressive and controllable listener dynamics modeling for multi-modal responsive interaction. *arXiv preprint arXiv:2504.21718*, 2025. 3

[19] Tianqi Li, Ruobing Zheng, Minghui Yang, Jingdong Chen, and Ming Yang. Ditto: Motion-space diffusion for controllable realtime talking head synthesis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, page 9704–9713, New York, NY, USA, 2025. Association for Computing Machinery. 2

[20] Jin Liu, Xi Wang, Xiaomeng Fu, Yesheng Chai, Cai Yu, Jiao Dai, and Jizhong Han. Mfr-net: Multi-faceted responsive listening head generation via denoising diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 6734–6743, New York, NY, USA, 2023. Association for Computing Machinery. 1, 3

[21] X. Liu, Y. Guo, C. Zhen, T. Li, Y. Ao, and P. Yan. Customlistener: Text-guided responsive interaction for user-friendly listening head generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2415–2424, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 1, 3

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 5

[23] Cheng Luo, Jianghui Wang, Bing Li, Siyang Song, and Bernard Ghanem. Omniresponse: Online multimodal conversational response generation in dyadic interactions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1, 2, 3

[24] Yifeng Ma, Suzhen Wang, Zhipeng Hu, Changjie Fan, Tangjie Lv, Yu Ding, Zhidong Deng, and Xin Yu. Styletalk: One-shot talking head generation with controllable speaking styles. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1896–1904, 2023. 5, 1

[25] Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20395–20405, 2022. 1, 3, 6, 2

[26] Evonne Ng, Javier Romero, Timur Bagautdinov, Shaojie Bai, Trevor Darrell, Angjoo Kanazawa, and Alexander Richard. From audio to photoreal embodiment: Synthesizing humans in conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1001–1010, 2024. 1, 2, 3

[27] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 5

[28] Ziqiao Peng, Yihao Luo, Yue Shi, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5292–5301, 2023. 1, 2, 6

[29] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20687–20697, 2023. 1, 2, 6, 3

[30] Ziqiao Peng, Yanbo Fan, Haoyu Wu, Xuan Wang, Hongyan Liu, Jun He, and Zhaoxin Fan. Dualtalk: Dual-speaker interaction for 3d talking head conversations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21055–21064, 2025. 2, 3, 6

[31] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023. 4, 5, 1

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[33] L. Song, G. Yin, Z. Jin, X. Dong, and C. Xu. Emotional listener portrait: Realistic listener motion simulation in conversation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20782–20792, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 1, 3

[34] Mingze Sun, Chao Xu, Xinyu Jiang, Yang Liu, Baigui Sun, and Ruqi Huang. Beyond talking – generating holistic 3d human dyadic motion for communication. *Int. J. Comput. Vision*, 133(5):2910–2926, 2024. 1, 2, 3

[35] Zhiyao Sun, Tian Lv, Sheng Ye, Matthieu Lin, Jenny Sheng, Yu-Hui Wen, Minjing Yu, and Yong-jin Liu. Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models. *ACM Transactions on Graphics (TOG)*, 43(4):1–9, 2024. 5, 1

[36] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*, pages 244–260. Springer, 2024. 1, 2

[37] Minh Tran, Di Chang, Maksim Siniukov, and Mohammad Soleymani. Dim: Dyadic interaction modeling for social behavior generation. In *European Conference on Computer Vision*, pages 484–503. Springer, 2024. 1, 2, 3

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4, 1

[39] Yinuo Wang, Yanbo Fan, Xuan Wang, Guo Yu, and Fei Wang. Diffusion-based realistic listening head generation via hybrid motion modeling. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15885–15895, 2025. 1, 3

[40] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12780–12790, 2023. 1, 2, 6

[41] Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. Vasa-1: lifelike audio-driven talking faces generated in real time. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2025. Curran Associates Inc. 2

[42] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2

[43] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings*

of the IEEE/CVF conference on computer vision and pattern recognition, pages 8652–8661, 2023. 2

[44] Wei Zhao, Peng Xiao, Rongju Zhang, Yijun Wang, and Jianxin Lin. Semantic-aware responsive listener head synthesis. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 7065–7069, New York, NY, USA, 2022. Association for Computing Machinery. 3

[45] Mohan Zhou, Yalong Bai, Wei Zhang, Tiejun Zhao, and Tao Mei. Responsive listening head generation: A benchmark dataset and baseline. In *European Conference on Computer Vision*, 2021. 1, 3

[46] Yongming Zhu, Longhao Zhang, Zhengkun Rong, Tianshu Hu, Shuang Liang, and Zhipeng Ge. Infp: Audio-driven interactive head generation in dyadic conversations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10667–10677, 2025. 1, 2, 3