Conversational Human Audio-visual Talking Dialogue Generation

Junhao Song[1], Lluis Guasch[2], Xilin He[3], Zhongyu Yang[4], Yingfang Yuan[5], Weicheng Xie[6], Linlin Shen[7,8], Haijun Lin[9], Shizhe Liu[10], Wei Pang[4], and Siyang Song[†11]

[1] Department of Computing, Imperial College London, UK
[2] Department of Earth Science & Engineering, Imperial College London, UK
[3] Mohamed bin Zayed University of Artificial Intelligence, UAE
[4] Department of Computer Science, Heriot-Watt University, UK
[5] School of Computer Science, Northumbria University, UK
[6] College of Computer Science & Software Engineering, Shenzhen University, China
[7] School of Artificial Intelligence, Shenzhen University, China
[8] Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen University, China
[9] School of Engineering and Design, Hunan Normal University, China
[10] Department of Computer Science, University of Oxford, UK
[11] Department of Computer Science, University of Exeter, UK
junhao.song23@imperial.ac.uk and s.song@exeter.ac.uk
† Corresponding author

Abstract. Large-scale dyadic interactive audio-visual dialogue (DIAD) datasets provide fundamental data resources for developing humanoid interactive virtual agents and digital humans. However, collecting such data is time-consuming, expensive, and ethically sensitive. To address this, we propose CHAT, a new dyadic interactive audio-visual dialogue generation (DIADG) framework that generates diverse, paired, and mutually responsive speech-face dialogue clips from a single textual prompt. CHAT unifies large language models and talking face models with interactive audio and facial behaviour refinement modules, enabling the generation of aligned dyadic dialogue clips with diverse contents and facial identities. Experiments show that CHAT outperforms existing related methods designed for similar tasks under both objective and subjective evaluations. Moreover, our synthesised CHAT-AVD-50k dataset serves as effective pre-training data for downstream interactive head generation, consistently improving PerFRDiff and ReactDiff on REACT 2024. CHAT offers a scalable alternative to the costly and ethically sensitive collection of real dyadic interaction data.

Keywords: Audio-visual Human Behaviour Synthesis · Human Behavioural Response Generation · Diffusion Model · Reaction Generation

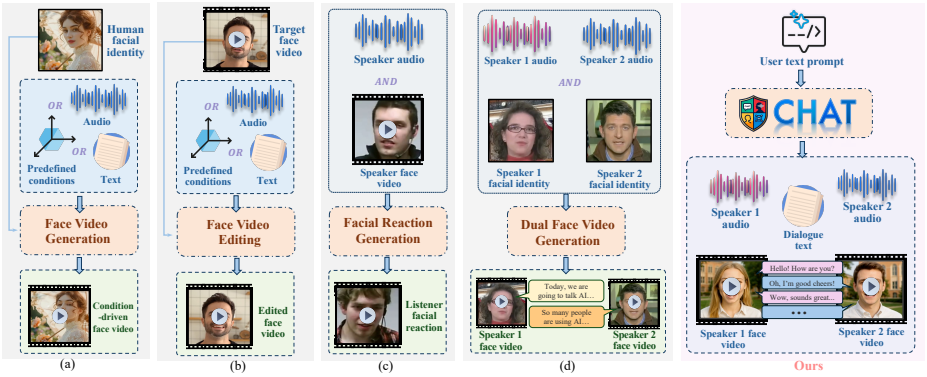

Fig. 1: Existing methods: (a) Condition-driven face video generation synthesises a *non-interactive face video* based on pre-defined conditions (e.g., text/audio) and a facial identity; **(b) Face video editing** utilises pre-defined conditions to modify the given *non-interactive face video*; **(c) Facial reaction generation** takes human audio-visual behaviours to generate the conversational partner's *interactive non-verbal* facial reaction video; and **(d) Interactive talking face video pair generation** takes a pair of pre-defined conditions to accordingly generate *a pair of verbally interactive face videos*. **Our CHAT** enables the generation of diverse, *verbally and non-verbally interactive dyadic audio-visual conversations* directly from a single text prompt, without requiring pre-defined video or audio inputs.

1 Introduction

Dyadic **I**nteractive **A**udio-visual **D**ialogue (DIAD) refers to paired audio-visual recordings capturing the speech, facial behaviours, and mutual responsiveness of two individuals engaged in natural dyadic interactions. Large-scale DIAD datasets are essential for training and evaluating intelligent human-computer interaction (HCI) systems, including virtual agents [65] and digital companions [31], which must model not only verbal content but also non-verbal cues and interpersonal dynamics. However, collecting large-scale DIAD datasets remains challenging due to privacy and ethical constraints, demographic imbalance, annotation cost, and the complexity of synchronising multi-modal signals in natural dyadic settings [28]. As a result, existing DIAD datasets possess limited size, interaction scenario diversity, and demographic coverage. Motivated by this, we argue that scalable synthesis of large-scale, demographically diverse DIAD datasets containing paired interactive audio-visual-text dialogues from anonymised textual prompts is an effective alternative solution.

While the expected DIAD clip pair is to contain two audio-visual clips containing interactive speeches and facial behaviours expressed by a pair of individuals, a large part of existing facial behaviour synthesis methods [6, 10, 68] can only generate independent and non-interactive face videos from a textual/audio sentence [8, 9, 26] or modify a pre-defined face video [12, 34, 46] (**Limitation 1**). Although some advanced approaches [5, 43, 70, 78] can already generate paired

interactive audio-visual clips, they are still driven by pre-defined conversational speeches/transcripts, i.e., they are not only unable to generate interactive verbal behaviours on their own but also unable to account for how one individual's behaviours are influenced by the other's non-verbal vocal and facial behaviours. Although a few recently proposed facial reaction generation (FRG) methods [39,55,56,69,77] can generate facial behaviour videos that are responsive to the input conversational partner's non-verbal behaviours, they are still unable to produce paired human-human interaction audio-visual clips (**Limitation 2**).

Alternatively, recent large language models (LLMs) [14, 16, 18, 51] enable the generation of diverse human-human dialogues. Although most of them are limited to outputting only diverse appropriate interactive textual dialogues from a user-defined textual prompt or chat contents [36], a few advanced LLMs [29,38] can synthesise audio-visual clips. However, to the best of our knowledge, none of the existing LLMs can directly generate diverse, realistic, synchronised, and mutually responsive human-human DIAD clips (**Limitation 3**).

Given the importance of DIADs in developing intelligent HCI systems, this paper introduces a new task, called **Dyadic Interactive Audio-visual Dialogue Generation (DIADG)**, which requires the developed model to automatically generate a set of diverse but contextually appropriate audio-visual dialogues from a single user-defined anonymous textual prompt, where each dialogue contains interactive speech and facial behaviours expressed by a pair of individuals engaged in a human-human dyadic interaction. In this sense, we propose a DIADG framework, called **C**onversational **H**uman **A**udio-visual **T**alking (**CHAT**) Dialogue Generation that consists of three key modules: (i) a **Textual Dialogue Generation (TDG)** module for creating a set of diverse and fine-grained textual dyadic dialogues; (ii) a **Dyadic Audio Dialogue Generation (DADG)** module for generating paired, emotion-aware speeches aligned with the corresponding textual dialogues; and (iii) an **Interactive Facial Behaviour Generation (IFBG)** module for generating mutually responsive and temporally coherent face video pairs aligned with the corresponding interactive speech and textual dialogues. Fig. 1 compares our CHAT with existing related methods discussed above. The main contributions of this paper are summarised as follows:

- We formally define a new human-centred AI task called DIADG, and propose a DIADG framework called CHAT which can generate a set of diverse audio-visual DIAD clips from a single customisable textual prompt, facilitating scalable, demographically diverse DIAD dataset creation without manual data collection.
- We introduce an Interactive Audio Refinement (IAR) block together with a novel trainable Interactive Facial Behaviour Refinement (IFBR) block, transforming the non-interactive, unemotional speeches and face videos produced by pre-trained models into emotional, synchronised and mutually responsive DIAD clips.
- Experiments show that CHAT not only outperforms related methods in synchronisation and interactivity while achieving competitive visual qual-

ity, but also provides effective pre-training data for downstream Interactive Head Generation (IHG), consistently improving models over real-data-trained counterparts on the REACT benchmark [58].

2 Related Work

LLM-based dialogue generation. Recent advancements in LLMs, such as GPT [1], Llama [14], DeepSeek [18] and Gemini [16] series, enable the generation of contextually relevant and coherent human-style dialogues from user-defined textual prompts. These LLMs can be categorised into three groups: text-only models [18, 25] that interact with users via texts only, text-audio models [15, 53] supporting both textual and speech interactions, and multi-modal models that can understand and output text-audio [13, 17], text-image [16, 71] and text-video [38, 73] modalities. However, to the best of our knowledge, none of the existing models can directly synthesise diverse audio-visual interactive and responsive clip pairs from a single user-defined prompt.

Face video synthesis. Existing face video generation approaches (Fig. 1) can be roughly categorised into: (1) Condition-driven face video generation, (2) Face video editing, (3) Facial reaction generation, and (4) Interactive face video pair generation. Specifically, **condition-driven face video generation** methods aim to produce coherent facial animations from a face image [67] or 3D face data [59] conditioned solely on texts [20, 66], audio [8, 26] or other pre-defined attributes (e.g., action units [21, 39] or driving motion fields [19, 76]). Alternatively, **face video editing** methods modify the given face videos through pre-defined or learned facial attributes, enabling expressive and realistic changes in identity [3] or emotions [41]. For example, talking face methods frequently utilise pre-defined speech [60, 74] or textual transcripts [9] to semantically modify face videos with synchronised lip behaviours [34], facial expressions [60] or enhanced hair-face consistency [12]. However, these methods can only output independent, non-interactive face videos defined by the given conditions. Alternatively, an increasing number of methods [24, 37, 40, 42, 44, 56, 57, 61, 64, 77] have been developed to generate diverse but appropriate interactive **facial reaction videos** conditioned on the input human audio-visual behaviours, although they still fail to synthesise paired interactive audio-visual clips. Besides, recent **interactive talking face generation** methods such as DialogueNeRF [70], DualTalk [48], Audio2Photoreal [43], INFP [78], AV-Flow [5] and ChatAnyone [50], enable direct generation of paired verbally interactive face videos. However, these interactive face videos are generated from pre-defined interactive verbal/language behaviours without semantically considering non-verbal facial and vocal behaviours in triggering/responding to their conversational partners' behaviours, and thus these methods fail to generate both verbally and non-verbally interactive outputs. Concurrent works TAVID [27] and JAM-Flow [33] explore related joint generation tasks but require pre-defined scripts or audio inputs. Compared with these methods, CHAT targets the setting of generating complete, mutually re-

sponsive DIAD clips from a **single anonymous scenario prompt** without any pre-defined conditions.

3 Task Definition

The DIADG task aims to develop a machine learning model $\mathcal{H}$ that can generate multiple different DIAD clips $D_1, D_2, \ldots, D_N$ from a textual prompt $\mathcal{P}$. Here, $\mathcal{P}$ provides a de-identified summary of a dyadic interaction scenario where two subjects chat with each other. This DIADG task can be formally formulated as:

$$D_1, D_2, \cdots, D_N = \mathcal{H}(\mathcal{P}), \tag{1}$$

where each DIAD clip $D_n = \{(V_n^i, A_n^i), (V_n^j, A_n^j)\}, n \in \{1, \cdots, N\}$ contains realistic, temporally coherent, synchronised and mutually responsive speeches and facial behaviours expressed by two arbitrary subjects S^i and S^j engaged in a natural dyadic dialogue.

4 Methodology

4.1 CHAT Framework

Our **C**onversational **H**uman **A**udio-visual **T**alking (**CHAT**) Dialogue Generation is a DIADG solution that can generate diverse, contextually appropriate, temporally coherent, and mutually responsive dyadic interactive audio-visual dialogues (DIADs) $D_1, D_2, \cdots, D_N$ from a single user-defined textual prompt $\mathcal{P}$. Throughout Sec. 4, $\bar{x}$ denotes initially generated values, $\hat{x}$ denotes LLM-refined values, and $\tilde{x}$ denotes diffusion-refined values. As illustrated in Fig. 2, it comprises three core modules: Textual Dialogue Generation (**TDG**), Dyadic Audio Dialogue Generation (**DADG**), and Interactive Facial Behaviour Generation (**IFBG**). This modular design allows individual components (e.g., LLM_{text}) to be replaced or updated independently without retraining the full pipeline, while a two-stage scheme, which first pre-trains the individual modules and then jointly optimises the trainable blocks end-to-end with the LLMs and text-to-speech (TTS) models frozen, ensures stable optimisation. Separating dialogue, audio and video into dedicated modules lets each stage build on a model that is already strong in its own domain, and keeps the trainable part of the pipeline small, which matters given the cost of video diffusion.

Textual Dialogue Generation (TDG). Our CHAT starts with the TDG module responsible for generating multi-turn dyadic dialogues and identity descriptions from $\mathcal{P}$. Given a textual prompt $\mathcal{P}$ describing the scenario and style of the dialogue, a large language model (LLM) LLM_{text} is employed to produce N diverse, contextually relevant dialogue scripts $\bar{\mathbf{DG}} = \{\bar{\text{dg}}_1, \bar{\text{dg}}_2, \cdots, \bar{\text{dg}}_N\}$, with each dialogue also paired with two generated textual descriptions ID-txt^i and ID-txt^j representing the identities of two individuals S^i and S^j engaged in the conversation (e.g., S^i is a thoughtful gentleman and S^j is a vibrant lady).

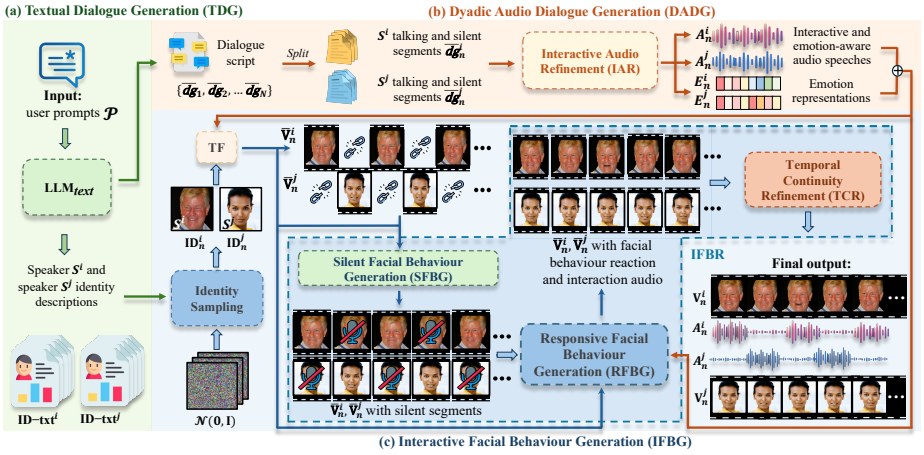

Fig. 2: The CHAT pipeline. (a) TDG generates textual dialogue pairs and identity descriptors from a single user prompt $\mathcal{P}$; **(b)** DADG converts each dialogue into interactive, emotion-aware audio pairs via the IAR block (Sec. 4.2); **(c)** IFBG synthesises mutually responsive face video pairs using a talking face model and the IFBR block (Sec. 4.3). All DIAD clips are assembled from DADG and IFBG outputs.

To ensure robustness and avoid the performance degradation of the LLM in generating long multi-turn conversational dialogues, our TDG generates each dyadic conversation via the LLM conditioned on the speakers' identities, dialogue texts and sentiment/emotion through a user prompt. Here, each conversation is constrained to 5–10 turns, a range where LLMs maintain optimal coherence without degradation [30, 35]. Here, each turn may contain multiple sentences expressed from one speaker.

Dyadic Audio Dialogue Generation (DADG). Given each textual dialogue $\bar{\mathrm{dg}}_n = \{\bar{\mathrm{dg}}_n(1), \cdots, \bar{\mathrm{dg}}_n(K)\}$ containing K sentences, the DADG module first organises them as two sets $\bar{\mathrm{dg}}_n^i$ and $\bar{\mathrm{dg}}_n^j$ based on their orders: odd-numbered sentences $\bar{\mathrm{dg}}_n^i = \{\bar{\mathrm{dg}}_n(1), \mathrm{Si}_n^i(2), \bar{\mathrm{dg}}_n(3), \mathrm{Si}_n^i(4), \ldots, \bar{\mathrm{dg}}_n(K-1), \mathrm{Si}_n^i(K)\}$ expressed by S^i, even-numbered sentences $\bar{\mathrm{dg}}_n^j = \{\mathrm{Si}_n^j(1), \bar{\mathrm{dg}}_n(2), \mathrm{Si}_n^j(3), \bar{\mathrm{dg}}_n(4), \ldots, \mathrm{Si}_n^j(K-1), \bar{\mathrm{dg}}_n(K)\}$ expressed by S^j. Here, $\mathrm{Si}_n^i(k)$ denotes a silent segment with the same temporal duration as $\bar{\mathrm{dg}}_n(k)$ (i.e., S^i keeps silent when S^j expresses the k-th sentence). While directly converting each textual sentence to an audio speech disregards the interactive nature of the human-human interaction, our DADG involves an **Interactive Audio Refinement (IAR)** block (please refer to Sec. 4.2 for details) that encodes each organised dialogue $\bar{\mathrm{dg}}_n^i, \bar{\mathrm{dg}}_n^j$ as an interactive and emotion-aware audio dialogue $\mathcal{A}_n = \{A_n^i, A_n^j\}$ conditioned on a user-defined prompt $\mathcal{P}_{\mathrm{ref}}$ as:

$$A_n^i, A_n^j, E_n^i, E_n^j = \mathrm{IAR}(\bar{\mathrm{dg}}_n^i, \bar{\mathrm{dg}}_n^j, \mathcal{P}_{\mathrm{ref}}), \tag{2}$$

where each speech sentence in A_n^i/A_n^j is expressed with a specific emotion (defined by E_n^i, E_n^j) suitable for the context. In natural human-human conversation,

although turn-taking is the fundamental structure and simultaneous speaking is a statistically rare edge case in most interactive scenarios, our IAR still occasionally inserts short-term interactive speech responses (e.g., "Hmm" and "OK") to silent segments to express the silent individual's verbal emotional responses when the conversational partner is speaking.

Interactive Facial Behaviour Generation (IFBG). Finally, the IFBG module generates a pair of mutually responsive, speech-aligned face videos (V_n^i and V_n^j) from A_n^i and A_n^j. It first extends a talking face model TF [2] to generate the initial talking facial behaviour segments (i.e., $\bar{V}_n^i = \{\bar{v}_n^i(1), \bar{v}_n^i(3), \cdots, \bar{v}_n^i(K-1)\}$ and $\bar{V}_n^j = \{\bar{v}_n^j(2), \bar{v}_n^j(4), \cdots, \bar{v}_n^j(K)\}$) corresponding to the given speech sentences, where the speech features are additionally concatenated with their corresponding emotion representations obtained by IAR to form a joint emotion-aware speech condition, which is injected into the talking face model to guide both accurate lip synchronisation and emotionally expressive facial animations. Here, an identity generation/sampling block [47] converts the textual identity descriptors into sampled synthetic face images ID_n^i and ID_n^j from the texts ID-txt^i and ID-txt^j for defining the facial identities of S^i and S^j, i.e., supporting diverse photorealistic and stylised identities. Based on both face identities, we propose an **Interactive Facial Behaviour Refinement (IFBR)** block to: (i) synthesise facial behaviours of silent segments; (ii) ensure all talking and silent facial behaviours are responsive to their conversational partner's audio-visual behaviours; and (iii) make the final refined facial behaviours V_n^i and V_n^j temporally coherent at the segment boundaries as:

$$V_n^i, V_n^j = \text{IFBR}(\bar{V}_n^i, \bar{V}_n^j, A_n^i, A_n^j). \tag{3}$$

Please refer to Sec. 4.3 for more details.

CHAT-AVD-50k dataset. CHAT-AVD-50k, a large-scale synthetic dataset for DIADG research, contains 50,000 diverse dyadic textual-audio-visual dialogue pairs (100,000 clips, 1388.9 hours) across 100,000 synthesised facial identities, with turn-level emotion, scenario, and speaker identity metadata. It covers casual, professional, emotional, academic, and social interaction scenarios.

4.2 Interactive Audio Refinement

The **IAR block** of the DADG module first applies an LLM ($\text{LLM}_{\text{audio}}$) to refine the obtained textual dialogue $\bar{\text{dg}}_n^i$ and $\bar{\text{dg}}_n^j$ using an audio refine prompt $\mathcal{P}_{\text{ref}}$. This not only refines the contents of $\bar{\text{dg}}_n^i$ and $\bar{\text{dg}}_n^j$ as $\hat{\text{dg}}_n^i$ and $\hat{\text{dg}}_n^j$ by making extensive use of the second-person pronouns and conversational sentences, but also outputs:

(i) a set of interactive words (e.g., Hmm and OK) $\text{IW}_n = \{\text{iw}_n(1), \cdots, \text{iw}_n(K)\}$ allowing each individual to express short but contextually appropriate responses when the conversational partner speaks, making the dialogue more natural and interactive as:

$$\text{iw}_n(k) = \begin{cases} \text{random}(\mathcal{W}_{\text{inter}}) & \text{random}() < P_{\text{inter}} \\ \text{silence} & \text{otherwise} \end{cases}, \tag{4}$$

where $\mathcal{W}_{\text{inter}} = \{\text{“uh-huh”}, \text{“hmm”}, \text{“okay”}, \text{“yeah”}\}$ and P_{inter} denotes a pre-defined probability.
(ii) a pair of emotion representations $E_n^i = \{e_n^i(1), \cdots, e_n^i(K)\}$ and $E_n^j = \{e_n^j(1), \cdots, e_n^j(K)\}$, where each $e_n^i(k)$ is a structured descriptor from $\text{LLM}_{\text{audio}}$ encoding a discrete emotion category and continuous prosody attributes (rate, pitch, energy, pauses), embedded by the emotion encoder described below;
(iii) time stamps $\mathcal{T}_n^i$ and $\mathcal{T}_n^j$ for labelling the start and end time stamps of all sentences and interactive words expressed by S^i and S^j;
(iv) the sound environment SE_n (e.g., studio or open space) of the dialogue. Then, we insert all interactive words into their corresponding silent segments in $\hat{\text{dg}}_n^i$ and $\hat{\text{dg}}_n^j$, resulting in the final interactive dialogues dg_n^i and dg_n^j (written without a hat to denote the interactive-word-inserted sequences):

$$\begin{aligned} \text{dg}_n^i &= \big\{\text{dg}_n(1), \text{iw}_n(2), \text{dg}_n(3), \cdots, \text{iw}_n(K)\big\}, \\ \text{dg}_n^j &= \big\{\text{iw}_n(1), \text{dg}_n(2), \text{iw}_n(3), \cdots, \text{dg}_n(K)\big\}. \end{aligned} \tag{5}$$

As a result, the final obtained dg_n^i, dg_n^j, E_n^i, E_n^j, $\mathcal{T}_n^i$, $\mathcal{T}_n^j$ and SE_n jointly represent an interactive human speech dialogue.

To generate interactive and emotion-aware audio speeches from the refined textual dialogues, each sentence/interactive word $\text{dg}_n^i(k)/\text{iw}_n(k)$ and its corresponding emotion representation $e_n^i(k)$ are first encoded by a content Transformer encoder (TFM-C) and an emotion Transformer encoder (TFM-E) into a pair of latent representations of the same dimension. An element-wise sum then processes these representations to generate an emotion-aware dialogue representation $\mathbf{z}_n^i(k)$ as:

$$\mathbf{z}_n^i(k) = \text{Sum}(\text{TFM-C}(\text{dg}_n^i(k)), \text{TFM-E}(e_n^i(k))). \tag{6}$$

Then, the obtained $\{\mathbf{z}_n^i(k)\}_{k=1}^K$ corresponding to all sentences/interactive words expressed by S^i are used, together with the time stamps $\mathcal{T}_n^i$ and the sound environment SE_n, to condition an emotional speech synthesiser that produces the interactive, emotion-aware audio waveform segments $A_n^i(k)_{k=1}^K$.

4.3 Interactive Facial Behaviour Refinement

The **IFBR block** refines the independently generated *non-interactive* talking facial behaviour segments $\bar{V}_n^i$ and $\bar{V}_n^j$ as a pair of temporally coherent and mutually responsive face videos V_n^i and V_n^j via the following blocks.

Silent Facial Behaviour Generation (SFBG). This block generates each initial silent facial behaviour segment $\bar{v}_n^i(k)/\bar{v}_n^j(k+1)$ when S^i/S^j are not speaking, reflecting temporally and semantically coherent facial behaviours in the context of its preceding and succeeding talking facial behaviour segments $\bar{v}_n^i(k-1)$ and $\bar{v}_n^i(k+1)$. Specifically, the initial silent segment $\bar{v}_n^i(k)$ containing realistic facial behaviours (e.g., nods and blinks) is generated conditioned on S^i's facial

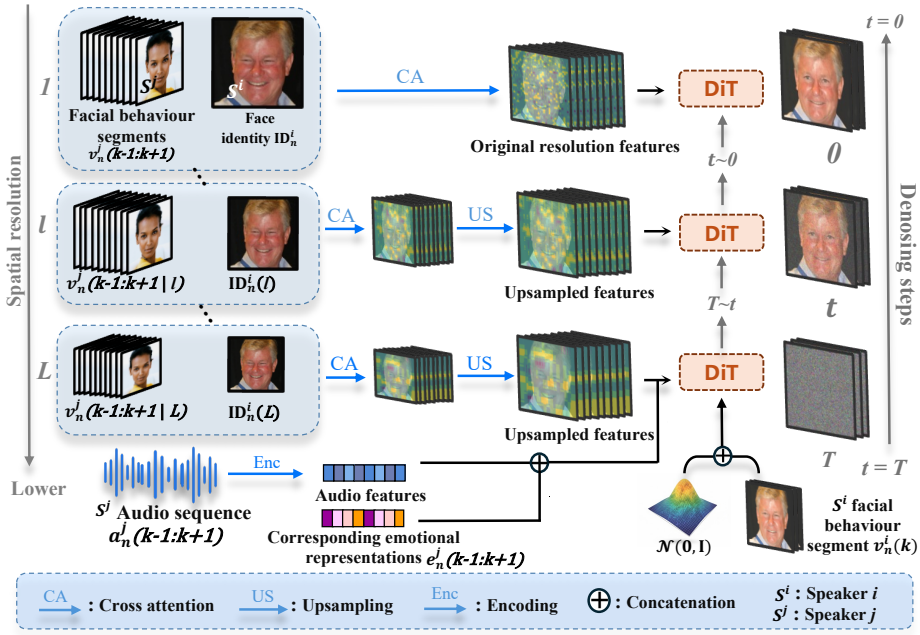

Fig. 3: Illustration of the proposed RFBG sub-block.

identity image ID_n^i as well as $\bar{v}_n^i(k-1)$ and $\bar{v}_n^i(k+1)$ via a video diffusion model SilentDiff as:

$$\bar{v}_n^i(k) = \mathrm{SilentDiff}(\bar{v}_n^i(k-1), \bar{v}_n^i(k+1), \mathrm{ID}_n^i). \tag{7}$$

SilentDiff inherits the same architecture from [10] but is end-to-end optimised with other modules/blocks within our framework. This also maintains rough temporal continuity and semantic coherence between $\bar{v}_n^i(k)$ and its adjacent talking facial segments $\bar{v}_n^i(k-1)/\bar{v}_n^i(k+1)$.

Responsive Facial Behaviour Generation (RFBG). This block employs a video diffusion transformer to refine the initial silent and talking facial behaviour segments $\bar{v}_n^i(k)/\bar{v}_n^i(k-1)$ as facial behaviour segments $\tilde{v}_n^i(k)/\tilde{v}_n^i(k-1)$ that are responsive to the conversational partner's audio-visual behaviours expressed at the same period. Given S^i's facial behaviour $\bar{v}_n^i(k)$, the conditions used in RFBG include S^j's preceding, current and succeeding audio-visual behaviours $\bar{v}_n^j(k-1:k+1)$ and $a_n^j(k-1:k+1)$ expressed from turn $k-1$ to $k+1$, which are injected into the diffusion transformer at multiple denoising steps. As illustrated in Fig. 3, the face identity ID_n^i and S^j's facial behaviours $\bar{v}_n^j(k-1:k+1)$ are spatially downsampled to L scales as $\mathrm{ID}_n^i(l)_{l=1}^L$ and $\bar{v}_n^j(k-1:k+1|l)_{l=1}^L$ (i.e., higher l corresponds to lower spatial resolution), which are then fused via learnable cross-attention (CA) blocks at L scales as:

$$\mathcal{C}_n^j(k-1{:}k+1|l) = \mathrm{CA}(\bar{v}_n^j(k-1:k+1|l), \mathrm{ID}_n^i(l)). \tag{8}$$

This feature extraction enables the capture of both multi-scale coarse-level (e.g., low-resolution conditions generally capture shape and global facial motions) and fine-grained (e.g., high-resolution conditions typically capture facial details such as micro-expressions and local muscle movements) facial behaviours expressed by S^j, while forcing the diffusion transformer to generate facial behaviours of identity ID_n^i. Then, the obtained L conditions $\mathcal{C}_n^j(k{-}1{:}k{+}1|l)_{l=1}^L$ (i.e., each is a 2D feature map sequence) are upsampled to the same spatial resolution as the original $\bar{v}_n^j(k{-}1:k{+}1)/\mathrm{ID}_n^i$. We then gradually integrate conditions obtained from lower spatial resolutions into early denoising steps and conditions of higher spatial resolutions into later denoising steps, as the early steps of the diffusion process typically model overall facial behaviour outlines, while later steps refine facial details [22, 63] as:

$$\tilde{v}_n^i(k) = \mathrm{Diff}\big(\bar{v}_n^i(k) \mid \Phi_k^{i,j}(t), a_n^j(k{-}1:k{+}1), e_n^j(k{-}1:k{+}1)\big), \tag{9}$$

where $\Phi_k^{i,j}(t) = \bigcup_{l=l(t)}^L \mathcal{C}_n^j(k{-}1{:}k{+}1|l)$. Particularly, $l(t) = \max(1, \lceil L\,t/T \rceil)$ denotes that at denoising step t, the facial behaviour conditions $\mathcal{C}_n^j(k{-}1{:}k{+}1|l)$ from scale $l(t)$ to L are injected into the diffusion process. This ensures that the conditions capturing S^j's global facial motions guide early denoising steps to generate responsive facial behaviour outlines of $\tilde{v}_n^i(k)$, whilst the fine-grained conditions are progressively integrated into later denoising steps to help complete details of $\tilde{v}_n^i(k)$. At the first denoising step, we also inject the audio representation learned from $a_n^j(k{-}1:k{+}1)$ and its emotion representation $e_n^j(k{-}1:k{+}1)$ obtained by IAR via cross-attention as a condition.

Temporal Continuity Refinement (TCR). Although the SFBG block maintains rough temporal continuity between adjacent facial segments, segment-level processing conducted by the RFBG block may still result in visual discontinuities at facial segment boundaries. To address this, we utilise a Gaussian-weighted boundary blending strategy [45] to ensure: (1) smooth and perceptually imperceptible facial behaviour transition between facial segment boundaries; and (2) impact on only a narrow temporal window within $2W$ frames centred at each boundary (e.g., empirically set to only $W = 10$ frames, whereas each facial behaviour segment contains 100 frames), as the Gaussian weighting function rapidly decays towards zero outside the boundary region. Importantly, this block conducts **an identical transform to both S^i and S^j facial behaviours** at the same time window to preserve their established mutual responsiveness. Given a pair of adjacent facial behaviour segments $\tilde{v}_n^i(k-1)$ and $\tilde{v}_n^i(k)$, this block applies temporal blending to the last W frames of $\tilde{v}_n^i(k-1)$ and the first W frames of $\tilde{v}_n^i(k)$ as:

$$\begin{aligned} v_n^i(k-1,\, F_{k-1}-\tau+1) &= [1-\alpha(\tau)]\,\tilde{v}_n^i(k-1,\, F_{k-1}-\tau+1) \\ &\quad + \alpha(\tau)\,\tilde{v}_n^i(k,\, \tau), \\ v_n^i(k,\, \tau) &= \alpha(\tau)\,\tilde{v}_n^i(k-1,\, F_{k-1}-\tau+1) \\ &\quad + [1-\alpha(\tau)]\,\tilde{v}_n^i(k,\, \tau), \end{aligned} \tag{10}$$

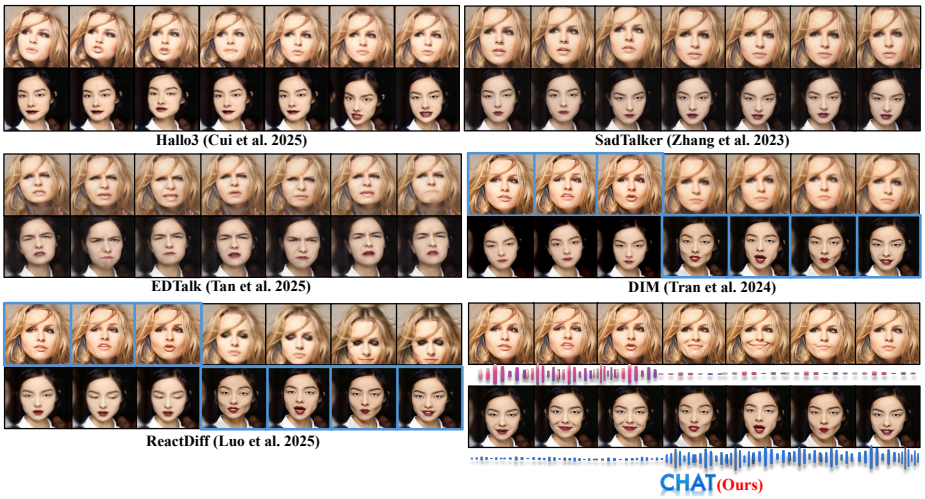

Fig. 4: Qualitative comparison. Talking-face methods (Hallo3, EDTalk, SadTalker) used twice yield non-responsive pairs; FRG methods (DIM, ReactDiff) take user-provided speaker video (blue frames). CHAT generates mutually responsive audio-visual dialogue pairs.

where $1 \leq \tau \leq W$ and F_{k-1} is the total number of frames in segment $k-1$, and $\alpha(\tau) = \frac{1}{2}\exp(-(\tau-1)^2/2\sigma^2)$ is a half-Gaussian blending weight, equal to 0.5 at the boundary and decaying outwards, with $\sigma = W/3$ following the 3-sigma rule. Since the transforms on both segments share the same $\alpha(\tau)$, this symmetric design preserves the established mutual responsiveness between S^i and S^j facial behaviours.

5 Experiments

5.1 Experimental Settings

Implementation details. The LLMs used in TDG and DADG modules are the Gemini models [16], while the emotional audio generation model in IAR is a pre-trained TTS [4]. We pre-train SFBG and RFBG blocks on the HDTF [75] dataset and REACT 2024 dataset [58] using AdamW optimiser respectively, where the diffusion model trained for RFBG is set to have 50 denoising steps. Then, the talking face model TF in IFBG as well as SFBG and RFBG blocks are end-to-end trained, while other blocks (LLMs and TTS models) are frozen. We compare our CHAT with competitors by utilising them to generate 1000 audio-visual dialogues, with each containing 5–10 conversational turns covering diverse topics (e.g., casual conversations and professional discussions).
Evaluation metrics. We assess four aspects: **Visual quality** via FID [23] and FVD [62]; **AV synchronisation** via LSE-C [49] and LSE-D [34]; **Interactive responsiveness** via FRCorr and FRDiv [57] (appropriateness and diversity of

Table 1: Quantitative results. Comparison of CHAT and baselines on evaluation set (mean of 1000 samples). The best result is in **bold** and the second best in underline.

Method	Visual Quality		AV Sync		Reaction ($\times 10^{-2}$)		ID Preservation	
	FID ↓	FVD ↓	LSE-C ↑	LSE-D ↓	FRCorr ↑	FRDiv ↑	CSIM ↑	LPIPS ↓
Hallo3 [10]	20.56	**362.23**	4.61	10.41	N/A	N/A	0.79	0.55
SadTalker [74]	22.53	385.51	6.85	8.17	N/A	N/A	0.81	0.48
DIM [61]	36.52	460.35	6.82	8.89	31.02	12.55	0.83	0.52
EDTalk [60]	18.74	619.92	5.62	9.61	N/A	N/A	0.78	**0.36**
ReactDiff [40]	21.36	386.22	5.23	9.02	48.05	15.31	0.77	0.58
CHAT (Ours)	**17.33**	365.03	**6.89**	**8.07**	**50.02**	**15.34**	**0.85**	0.38

Table 2: Qualitative user study results. Mean opinion scores for each method and evaluation dimension. The best result is in **bold** and the second best in underline.

Method	Visual	Audio	Sync	Expr.	Interact.
Hallo3 [10]	4.5	3.5	3.9	3.8	2.5
SadTalker [74]	2.5	3.7	1.2	2.0	1.3
DIM [61]	3.7	3.1	3.4	2.8	3.6
EDTalk [60]	2.2	3.6	2.3	2.8	2.7
ReactDiff [40]	4.2	3.5	3.8	4.1	3.2
CHAT (Ours)	**4.6**	**4.8**	**4.6**	**4.6**	**4.8**

generated behaviours); **Identity preservation** via CSIM (ArcFace [11]) and LPIPS [72]. Ablation studies additionally report MCD [32], Emo-Acc [52], and ViSQOL [7]. A 60-participant user study provides subjective quality assessment on a 5-point Likert scale.

5.2 Quantitative Results

Since no existing method directly addresses DIADG, we compare CHAT with recent methods for related tasks: **Hallo3** [10] and **SadTalker** [74] for audio-driven talking face generation; **DIM** [61] and **ReactDiff** [40] for facial reaction generation; and **EDTalk** [60] for emotional talking head synthesis. Audio-visual dialogue generation methods [5, 27, 33, 50, 54, 70, 78] are excluded because code is unavailable. For fair comparison, all methods receive the same prompts/audio; since competitors are limited to 12-second clips while CHAT can generate substantially longer conversations, we standardise evaluation to the first ten seconds. Audio-driven baselines (Hallo3, SadTalker, EDTalk) receive CHAT's generated audio A_n and identity ID_n, while FRG baselines (DIM, ReactDiff) receive CHAT's generated speaker video. Table 1 shows that CHAT achieves the best audio-visual synchronisation, interactiveness, FID and CSIM, while ranking second in FVD and LPIPS. Although these baselines were developed for related rather than identical tasks, the results still indicate that CHAT can produce synchronised, identity-consistent, and interactive audio-visual clip pairs, providing a competitive reference for this newly introduced task. FRCorr and FRDiv

Table 3: Ablation quantitative analysis results. Performance comparison of CHAT variants (mean of 1000 samples). All variants include three base modules: TDG, DADG, and IFBG. ✓ indicates enabled, and ✗ indicates disabled. (i) without entire IAR (but keep time stamps); (ii) without interactive words; (iii) without emotion representations; (iv) without sound environment; (v) without entire IFBR; (vi) without SFBG (replaced by interpolation); (vii) without RFBG; and (viii) without TCR. A dash marks an audio metric that is not applicable, as the ablated module leaves the audio path unchanged.

Variant	IAR Components			IFBR Components			Audio Quality			Visual Quality		AV Sync		Reaction ($\times 10^{-2}$)		ID Preservation	
	IW	Emo	SE	SFBG	RFBG	TCR	MCD↓	Emo-Acc↑	ViSQOL↑	FID↓	FVD↓	LSE-C↑	LSE-D↓	FRCorr↑	FRDiv↑	CSIM↑	LPIPS↓
(i) w/o IAR	✗	✗	✗	✓	✓	✓	6.23	41.2%	2.97	21.34	456.78	6.03	9.45	32.45	7.89	0.74	0.53
(ii) w/o IW	✗	✓	✓	✓	✓	✓	5.12	67.8%	3.89	18.67	398.23	6.77	8.67	47.12	9.34	0.81	0.41
(iii) w/o emotion	✓	✗	✓	✓	✓	✓	9.12	38.3%	2.67	23.45	423.56	6.12	9.23	28.34	5.67	0.69	0.59
(iv) w/o SE	✓	✓	✗	✓	✓	✓	4.78	76.2%	3.78	18.45	385.67	6.72	8.34	48.23	14.12	0.84	0.40
(v) w/o IFBR	✓	✓	✓	✗	✗	✗	–	–	–	38.67	623.45	5.72	10.23	28.56	6.78	0.62	0.71
(vi) w/o SFBG	✓	✓	✓	✗	✓	✓	–	–	–	20.23	375.89	6.78	8.94	49.45	10.23	0.80	0.44
(vii) w/o RFBG	✓	✓	✓	✓	✗	✓	–	–	–	20.34	412.56	6.45	8.88	25.78	4.56	0.76	0.42
(viii) w/o TCR	✓	✓	✓	✓	✓	✗	–	–	–	18.56	421.34	6.67	9.12	48.67	14.23	0.82	0.43
Full CHAT	✓	✓	✓	✓	✓	✓	**4.23**	**78.4%**	**3.98**	**17.33**	**365.03**	**6.89**	**8.07**	**50.02**	**15.34**	**0.85**	**0.38**

measure facial reaction appropriateness and diversity, and do not fully reflect turn-taking or semantic coherence at the dialogue level.

5.3 Qualitative Results

We conducted a user study with 60 participants (aged 20–55, from academia and industry across the USA, the UK, Singapore and China), who evaluated randomly ordered video pairs on five aspects (visual realism, audio quality, AV sync, expressiveness, interactivity; 1–5 Likert scale) for both the main comparison (Table 2) and eight ablation variants (Table 4). Participants rated CHAT highest on all five aspects. Fig. 4 also visually compares a dialogue generated by our CHAT and competitors. Whilst competitors exhibit visual artefacts around the mouth and generate strained expressions, our CHAT produces natural smiling facial reactions with higher fidelity and expressiveness. This is because CHAT is specifically designed for the DIADG task, while competitors were only designed for similar but different tasks. CHAT generates the two clips as a responsive pair, so the listening face reacts to the speaker within the same clip. The talking-face baselines animate each face in isolation, and the reaction baselines produce facial motion without the accompanying speech. These differences are consistent with the higher interactivity and synchronisation scores of CHAT in Table 1 and Table 2.

5.4 Ablation Studies

We conducted a series of ablation studies (Table 3) to evaluate our CHAT.
IAR ablation. Removing IAR degrades audio quality consistently; eliminating emotion-aware processing causes the largest drop. IAR also underpins the mutual responsiveness of the generated facial reactions.
IFBR ablation. Removing IFBR eliminates both listener silence behaviours (SFBG) and partner-responsive refinement (RFBG), causing severe degradation:

Table 4: Ablation user study results. Mean opinion scores (1–5); base modules (TDG, DADG, IFBG) always included. Best in **bold**, second best underlined.

Variant	IAR Components			IFBR Components			User Study Scores (1–5)				
	IW	Emo	SE	SFBG	RFBG	TCR	Vis.	Aud.	Sync	Expr.	Inter.
w/o IAR (all)	✗	✗	✗	✓	✓	✓	3.2	2.8	3.1	2.3	2.5
w/o IW	✗	✓	✓	✓	✓	✓	3.6	3.2	3.4	3.5	3.2
w/o Emotion	✓	✗	✓	✓	✓	✓	2.9	3.1	2.8	1.7	2.8
w/o SE	✓	✓	✗	✓	✓	✓	3.6	3.6	3.5	3.6	4.1
w/o IFBR (all)	✓	✓	✓	✗	✗	✗	2.4	3.1	2.6	2.5	2.3
w/o SFBG	✓	✓	✓	✗	✓	✓	3.7	3.4	3.6	3.7	4.3
w/o RFBG	✓	✓	✓	✓	✗	✓	3.4	3.8	3.2	3.4	2.1
w/o TCR	✓	✓	✓	✓	✓	✗	3.6	3.5	3.5	3.8	4.1
Full CHAT	✓	✓	✓	✓	✓	✓	**4.6**	**4.8**	**4.6**	**4.6**	**4.8**

FID increases by 123.1% and FVD by 70.8%, confirming that facial behaviour refinement is fundamental. Removing RFBG alone leads to large interactivity loss, demonstrating its necessity for mutually responsive synthesis. Because these variants leave the DADG audio path unchanged, we omit their audio metrics and report only the visual, synchronisation, responsiveness, and identity results.

User study on ablation variants. Table 4 shows the same pattern. Removing IAR lowers the audio quality score by 2.0 and interactivity by 2.3, whereas removing IFBR lowers visual realism by 2.2 and expressiveness by 2.1. The two blocks therefore play complementary roles.

5.5 Downstream Interactive Head Generation

To assess CHAT-AVD-50k's utility for downstream dyadic interaction tasks, we evaluate it as pre-training data for **Interactive Head Generation (IHG)**, where a model generates appropriate facial reactions conditioned on a conversational partner's audio-visual behaviours. We pre-train PerFRDiff [77] and ReactDiff [40] on CHAT-AVD-50k before fine-tuning on the REACT 2024 [58] training split, then evaluate on its test split under the official protocol. As baselines, we report each model trained on REACT 2024 alone. Since CHAT's IFBR block is itself pre-trained on REACT 2024 (Sec. 5), CHAT-AVD-50k should be regarded as a large-scale diverse augmentation over a related interaction distribution rather than as a fully out-of-domain source.

Table 5 shows consistent improvements across both reported metrics. Specifically, CHAT-AVD-50k pre-training increases FRCorr from 37.21 to 40.11 for PerFRDiff and from 24.19 to 26.12 for ReactDiff, while reducing FRDist (a dynamic time warping distance to the ground-truth reaction, lower is better) from 94.72 to 89.45 and from 86.70 to 83.87, respectively. We therefore interpret these gains as evidence of useful scale and diversity augmentation within a related distribution, rather than as strictly independent transfer. The consistent gains for two different reaction models indicate that the synthetic dialogues carry interaction patterns that transfer beyond a single architecture.

Table 5: Downstream IHG. Pre-training PerFRDiff and ReactDiff on CHAT-AVD-50k before fine-tuning on REACT 2024 [58]. FRCorr is reported as $\times 10^{-2}$. Best per model in **bold**.

Pre-training	**PerFRDiff** [77]		**ReactDiff** [40]	
	FRCorr↑	FRDist↓	FRCorr↑	FRDist↓
None (REACT 2024 only)	37.21	94.72	24.19	86.70
CHAT-AVD-50k	**40.11**	**89.45**	**26.12**	**83.87**

6 Conclusion

This paper formally defines the DIADG task and proposes CHAT, a framework for generating diverse human-human DIAD clips from a single textual prompt. Experiments show that CHAT outperforms existing related methods on DIAD clip generation, and that CHAT-AVD-50k provides effective pre-training data for downstream IHG, enabling both PerFRDiff and ReactDiff to outperform their real-data-trained baselines on the REACT benchmark. We view CHAT as an initial step towards scalable DIAD synthesis, and we hope that the DIADG task and the CHAT-AVD-50k dataset support further work on large-scale, demographically diverse dyadic interaction data. We will release CHAT-AVD-50k with provenance metadata and bias auditing for responsible use.

7 Limitations and Future Work

Inference cost. The main limitation of CHAT is its generation cost. Each dialogue clip runs a large language model, a text-to-speech model and several video diffusion passes, so synthesis at the scale of CHAT-AVD-50k is slow. Here scalable refers to demographic coverage rather than a low per-clip cost, and distillation or faster samplers are a clear next step.
Evaluation and metrics. FRCorr and FRDiv, taken from facial reaction generation, score the appropriateness and diversity of facial behaviour, not turn-taking timing or dialogue-level semantics. A metric for dialogue-level interaction quality remains open.
Scope. CHAT currently generates English dialogues, and the downstream study uses REACT 2024, on which the IFBR block is pre-trained. Extending it to more languages and to independent corpora such as IEMOCAP or NoXi is a natural next step.

Acknowledgements

Thanks to the Bio-inspired Computing and Machine Learning (BCML) Lab at Heriot-Watt University for early research support. Junhao Song is a first-year PhD student in the Department of Computing at Imperial College London, fully funded by the Hitachi-Imperial Centre, a collaboration between Hitachi Ltd, Hitachi Europe and Imperial College London.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Aneja, S., Thies, J., Dai, A., Nießner, M.: Facetalk: Audio-driven motion diffusion for neural parametric head models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21263–21273 (2024)
3. Bai, Z., Chen, P., Peng, X., Liu, L., Yao, N., Chen, H.: Bring your own character: A holistic solution for automatic facial animation generation of customized characters. In: 2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR). pp. 429–438. IEEE (2024)
4. Casanova, E., Davis, K., Gölge, E., Göknar, G., Gulea, I., Hart, L., Aljafari, A., Meyer, J., Morais, R., Olayemi, S., et al.: XTTS: a massively multilingual zero-shot text-to-speech model. arXiv preprint arXiv:2406.04904 (2024)
5. Chatziagapi, A., Morency, L.P., Gong, H., Zollhoefer, M., Samaras, D., Richard, A.: AV-Flow: Transforming text to audio-visual human-like interactions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2025)
6. Chen, K., Li, Z., Cui, F., Ni, M., Wang, S., Che, J., Liu, F., Qi, Y., Zhang, F., Liu, J., et al.: FastTalker: Real-time audio-driven talking face generation with 3d gaussian. Image and Vision Computing p. 105573 (2025)
7. Chinen, M., Lim, F.S., Skoglund, J., Gureev, N., O'Gorman, F., Hines, A.: ViSQOL v3: An open source production ready objective speech and audio metric. In: 2020 twelfth international conference on quality of multimedia experience (QoMEX). pp. 1–6. IEEE (2020)
8. Choi, J., Kim, M., Park, S.J., Ro, Y.M.: Text-driven talking face synthesis by reprogramming audio-driven models. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 8065–8069. IEEE (2024)
9. Cui, J., Chen, B., Xu, M., Shang, H., Chen, Y., Su, Q., Dong, Z., Yao, Y., Wang, J., Zhu, S.: High-fidelity dynamic portrait animation via direct preference optimization and temporal motion modulation. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–10 (2025)
10. Cui, J., Li, H., Zhan, Y., Shang, H., Cheng, K., Ma, Y., Mu, S., Zhou, H., Wang, J., Zhu, S.: Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21086–21095 (2025)
11. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
12. Deng, K., Zheng, D., Xie, J., Wang, J., Xie, W., Shen, L., Song, S.: DEGSTalk: Decomposed per-embedding gaussian fields for hair-preserving talking face synthesis. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
13. Dihan, Q.A., Nihalani, B.R., Tooley, A.A., Elhusseiny, A.M.: Eyes on google's notebooklm: using generative ai to create ophthalmology podcasts with a single click. Nature Eye **39**(2), 215–216 (2025)
14. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The LLaMA 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)

15. Fang, Q., Zhou, Y., Guo, S., Zhang, S., Feng, Y.: LLaMA-Omni2: Llm-based real-time spoken chatbot with autoregressive streaming speech synthesis. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. pp. 18617–18629. ACL (2025)
16. Gemini Team: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
17. Ghosh, S., Kumar, S., Seth, A., et al.: Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 6288–6313 (2024)
18. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al.: Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. Nature **645**(8081), 633–638 (2025)
19. Guo, Y., Yang, C., Rao, A., Agrawala, M., Lin, D., Dai, B.: Sparsectrl: Adding sparse controls to text-to-video diffusion models. In: European Conference on Computer Vision. p. 330–348. Berlin, Heidelberg (2024)
20. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. International Conference on Learning Representations (2024)
21. He, X., Luo, C., Xian, X., Li, B., Khan, M.H., Ge, Z., Xie, W., Song, S., Shen, L., Ghanem, B., et al.: SynFER: Towards boosting facial expression recognition with synthetic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10184–10195 (2025)
22. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
23. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. Advances in Neural Information Processing Systems **30** (2017)
24. Huang, J., Yu, Z.: Multiple appropriate facial reaction generation based on multi-view transformation of speaker video. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 13992–13996 (2025)
25. Inception Lab: Mercury: Ultra-fast language models based on diffusion. arXiv preprint arXiv:2506.17298 (2025)
26. Jang, Y., Kim, J.H., Ahn, J., Kwak, D., Yang, H.S., Ju, Y.C., Kim, I.H., Kim, B.Y., Chung, J.S.: Faces that speak: Jointly synthesising talking face and speech from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8818–8828 (2024)
27. Kim, J.H., Ahn, J., Kwak, D., Chung, J.S., Watanabe, S.: TAVID: Text-driven audio-visual interactive dialogue generation. arXiv preprint arXiv:2512.20296 (2025)
28. Kim, M., Liu, F., Jain, A., Liu, X.: DCFace: Synthetic face generation with dual condition diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12715–12725 (2023)
29. Kondratyuk, D., Yu, L., Gu, X., Lezama, J., Huang, J., Schindler, G., Hornung, R., Birodkar, V., Yan, J., Chiu, M.C., Somandepalli, K., Akbari, H., Alon, Y., Cheng, Y., Dillon, J.V., Gupta, A., Hahn, M., Hauth, A., Hendon, D., Martinez, A., Minnen, D., Sirotenko, M., Sohn, K., Yang, X., Adam, H., Yang, M.H., Essa, I.,

Wang, H., Ross, D.A., Seybold, B., Jiang, L.: VideoPoet: A large language model for zero-shot video generation. In: Proceedings of the 41st International Conference on Machine Learning. pp. 25105–25124 (2024)
30. Koudounas, A., La Quatra, M., Baralis, E.: DeepDialogue: A multi-turn emotionally-rich spoken dialogue dataset. arXiv preprint arXiv:2505.19978 (2025)
31. Kovačević, N., Holz, C., Gross, M., Wampfler, R.: The personality dimensions GPT-3 expresses during human-chatbot interactions. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **8**(2), 1–36 (2024)
32. Kubichek, R.: Mel-cepstral distance measure for objective speech quality assessment. In: Proceedings of IEEE pacific rim conference on communications computers and signal processing. vol. 1, pp. 125–128. IEEE (1993)
33. Kwon, M., Shin, J., Jung, J., Park, J., Uh, Y.: JAM-Flow: Joint audio-motion synthesis with flow matching. arXiv preprint arXiv:2506.23552 (2025)
34. Li, C., Zhang, C., Xu, W., Lin, J., Xie, J., Feng, W., Peng, B., Chen, C., Xing, W.: LatentSync: Taming audio-conditioned latent diffusion models for lip sync with syncnet supervision. arXiv preprint arXiv:2412.09262 (2024)
35. Li, J.N., Tu, Q., Mao, C., Yu, Z., Wen, J.R., Yan, R.: StreamingDialogue: Prolonged dialogue learning via long context compression with minimal losses. In: Advances in Neural Information Processing Systems. vol. 37, pp. 86074–86101 (2024)
36. Liang, Y., Meng, F., Zhang, Y., Chen, Y., Xu, J., Zhou, J.: Emotional conversation generation with heterogeneous graph neural network. Artificial Intelligence **308**, 103714 (2022)
37. Liu, Z., Liang, C., Wang, J., Zhang, H., Liu, Y., Zhang, C., Gui, J., Wang, S.: One-to-many appropriate reaction mapping modeling with discrete latent variable. In: 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–5. IEEE (2024)
38. Liu, Z., Dong, Y., Wang, J., Liu, Z., Hu, W., Lu, J., Rao, Y.: Ola: Pushing the frontiers of omni-modal language model with progressive modality alignment. arXiv preprint arXiv:2502.04328 (2025)
39. Luo, C., Song, S., Xie, W., Spitale, M., Ge, Z., Shen, L., Gunes, H.: ReactFace: Online multiple appropriate facial reaction generation in dyadic interactions. IEEE Transactions on Visualization and Computer Graphics (2024)
40. Luo, C., Song, S., Yan, S., Yu, Z., Ge, Z.: ReactDiff: Fundamental multiple appropriate facial reaction diffusion model. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 5607–5616 (2025)
41. Ma, Y., Liu, H., Wang, H., Pan, H., He, Y., Yuan, J., Zeng, A., Cai, C., Shum, H.Y., Liu, W., et al.: Follow-Your-Emoji: Fine-controllable and expressive freestyle portrait animation. arXiv preprint arXiv:2406.01900 (2024)
42. Mao, Q., Wu, Q., Liu, N., Ding, Y., Gao, L.: Scattering-conditioned diffusion models for multiple appropriate facial reaction generation. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 13985–13991 (2025)
43. Ng, E., Romero, J., Bagautdinov, T., Bai, S., Darrell, T., Kanazawa, A., Richard, A.: From audio to photoreal embodiment: Synthesizing humans in conversations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1001–1010 (2024)
44. Nguyen, M.D., Yang, H.J., Ho, N.H., Kim, S.H., Kim, S., Shin, J.E.: Vector quantized diffusion models for multiple appropriate reactions generation. In: 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–5. IEEE (2024)

45. Niklaus, S., Liu, F.: Softmax splatting for video frame interpolation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5437–5446 (2020)
46. Pang, Y., Zhang, Y., Quan, W., Fan, Y., Cun, X., Shan, Y., Yan, D.m.: DPE: Disentanglement of pose and expression for general video portrait editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 427–436 (2023)
47. Papantoniou, F.P., Lattas, A., Moschoglou, S., Deng, J., Kainz, B., Zafeiriou, S.: Arc2Face: A foundation model for id-consistent human faces. In: European Conference on Computer Vision. pp. 241–261. Springer (2024)
48. Peng, Z., Fan, Y., Wu, H., Wang, X., Liu, H., He, J., Fan, Z.: DualTalk: Dual-speaker interaction for 3d talking head conversations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21055–21064 (2025)
49. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A lip sync expert is all you need for speech to lip generation in the wild. In: Proceedings of the 28th ACM international conference on multimedia. pp. 484–492 (2020)
50. Qi, J., Ji, C., Xu, S., Zhang, P., Zhang, B., Bo, L.: ChatAnyone: Stylized real-time portrait video generation with hierarchical motion diffusion model. arXiv preprint arXiv:2503.21144 (2025)
51. Qwen Team: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
52. Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., et al.: SpeechBrain: A general-purpose speech toolkit. arXiv preprint arXiv:2106.04624 (2021)
53. Sharma, D.V., Ekbote, V., Gupta, A.: IndicSynth: A large-scale multilingual synthetic speech dataset for low-resource indian languages. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. pp. 22037–22060 (2025)
54. Siniukov, M., Chang, D., Tran, M., Gong, H., Chaubey, A., Soleymani, M.: DiTaiListener: Controllable high fidelity listener video generation with diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (October 2025)
55. Song, S., Shao, Z., Jaiswal, S., Shen, L., Valstar, M., Gunes, H.: Learning person-specific cognition from facial reactions for automatic personality recognition. IEEE Transactions on Affective Computing **14**(4), 3048–3065 (2022)
56. Song, S., Spitale, M., Kong, X., Zhu, H., Luo, C., Palmero, C., Barquero, G., Escalera, S., Valstar, M., Daoudi, M., et al.: REACT 2025: the third multiple appropriate facial reaction generation challenge. ACM Multimedia 2025 (2025)
57. Song, S., Spitale, M., Luo, C., Barquero, G., Palmero, C., Escalera, S., Valstar, M., Baur, T., Ringeval, F., André, E., et al.: REACT2023: The first multiple appropriate facial reaction generation challenge. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 9620–9624 (2023)
58. Song, S., Spitale, M., Luo, C., Palmero, C., Barquero, G., Zhu, H., Escalera, S., Valstar, M., Baur, T., Ringeval, F., et al.: React 2024: the second multiple appropriate facial reaction generation challenge. In: 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–5. IEEE (2024)
59. Song, W., Wang, X., Zheng, S., Li, S., Hao, A., Hou, X.: TalkingStyle: personalized speech-driven 3d facial animation with style preservation. IEEE Transactions on Visualization and Computer Graphics (2024)

60. Tan, S., Ji, B., Bi, M., Pan, Y.: EDTalk: Efficient disentanglement for emotional talking head synthesis. In: European Conference on Computer Vision. pp. 398–416. Springer (2024)
61. Tran, M., Chang, D., Siniukov, M., Soleymani, M.: DIM: Dyadic interaction modeling for social behavior generation. In: European Conference on Computer Vision. pp. 484–503. Springer (2024)
62. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
63. Wang, B., Vastola, J.J.: Diffusion models generate images like painters: an analytical theory of outline first, details later. arXiv preprint arXiv:2303.02490 (2023)
64. Wang, P., Xue, P., Liu, X., Ji, T.: Explaining listener reactions: Personality-guided facial response generation with cross-modal attention. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 13997–14003 (2025)
65. Wang, R.E., Ribeiro, A.T., Robinson, C.D., Loeb, S., Demszky, D.: Tutor CoPilot: A human-ai approach for scaling real-time expertise. arXiv preprint arXiv:2410.03017 (2024)
66. Wang, Z., Zhang, P., Qi, J., Xu, S., Zhang, B., et al.: OmniTalker: One-shot real-time text-driven talking audio-video generation with multimodal style mimicking. In: Advances in Neural Information Processing Systems. vol. 38, pp. 21834–21862 (2025)
67. Xiang, J., Guo, Y., Hu, L., Guo, B., Yuan, Y., Zhang, J.: Expressive talking human from single-image with imperfect priors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10398–10409 (2025)
68. Xu, S., Chen, G., Guo, Y.X., Yang, J., Li, C., Zang, Z., Zhang, Y., Tong, X., Guo, B.: VASA-1: Lifelike audio-driven talking faces generated in real time. arXiv preprint arXiv:2404.10667 (2024)
69. Xu, T., Spitale, M., Tang, H., Liu, L., Gunes, H., Song, S.: Reversible graph neural network-based reaction distribution learning for multiple appropriate facial reactions generation. IEEE Transactions on Affective Computing (2026)
70. Yan, Y., Zhou, Z., Wang, Z., Gao, J., Yang, X.: DialogueNerf: Towards realistic avatar face-to-face conversation video generation. Visual Intelligence **2**(1), 24 (2024)
71. Zeng, A., Xu, B., Wang, B., Zhang, C., Yin, D., Zhang, D., Rojas, D., Feng, G., Zhao, H., et al.: ChatGLM: A family of large language models from glm-130b to glm-4 all tools. arXiv preprint arXiv:2406.12793 (2024)
72. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
73. Zhang, S., Fang, Q., Yang, Z., Feng, Y.: LLaVA-Mini: Efficient image and video large multimodal models with one vision token. In: International Conference on Learning Representations (2025)
74. Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., Wang, F.: 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 (2023)
75. Zhang, Z., Li, L., Ding, Y., Fan, C.: Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3661–3670 (2021)

76. Zhao, S., Hong, F.T., Huang, X., Xu, D.: Synergizing motion and appearance: Multi-scale compensatory codebooks for talking head video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26232–26241 (2025)
77. Zhu, H., Kong, X., Xie, W., Huang, X., Shen, L., Liu, L., Gunes, H., Song, S.: PerFRDiff: Personalised weight editing for multiple appropriate facial reaction generation. In: Proceedings of the 32nd ACM International Conference on Multimedia (2024)
78. Zhu, Y., Zhang, L., Rong, Z., Hu, T., Liang, S., Ge, Z.: INFP: Audio-driven interactive head generation in dyadic conversations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10667–10677 (2025)