Agents through the ages: from early computational personas to generative social intelligence

Daksitha Senel Withanage Don, Elisabeth André, Michael Dietz, Silvan Mertes, Florian Lingenfelser

Angaben zur Veröffentlichung / Publication details:

Withanage Don, Daksitha Senel, Elisabeth André, Michael Dietz, Silvan Mertes, and Florian Lingenfelser. 2026. "Agents through the ages: from early computational personas to generative social intelligence." *KI - Künstliche Intelligenz*. https://doi.org/10.1007/s13218-026-00918-y.

KI - Künstliche Intelligenz
https://doi.org/10.1007/s13218-026-00918-y

TECHNICAL CONTRIBUTION

Agents Through the Ages

From Early Computational Personas to Generative Social Intelligence

Daksitha Senel Withanage Don[1] · Elisabeth André[1] · Michael Dietz[1] · Silvan Mertes[2] · Florian Lingenfelser[1]

Received: 16 December 2025 / Accepted: 10 July 2026

Abstract

Over several decades, research on socially interactive agents has transformed human–machine interaction, evolving from early script-based systems to today's generative, multimodal agents. Tracing this development through the ages, this article reflects on major conceptual and technical shifts that shaped the field, drawing on advances across multiple areas of AI, and discusses how progress in natural language processing and multimodal behavior analysis and generation has redefined AI as a social interface. Rather than presenting this trajectory as a narrative of continuous progress, we also examine persistent limitations, including weak grounding, difficult evaluation, opaque decision processes, safety risks in sensitive contexts, and the gap between convincing social simulation and genuine social understanding.

1 Introduction

Human communication is inherently multimodal: speech is interwoven with gestures, facial expressions, posture shifts, and subtle affective cues. Motivated by the goal of enabling more natural and intuitive communication with machines, an increasing body of research has focused on simulating these behaviors in socially interactive agents.

Recent years have seen major breakthroughs in photorealistic avatar technologies, such as Meta's Codec Avatars, which reconstruct animatable head models from phone scans [57], and Relightable Gaussian Codec Avatars, which use 3D Gaussian–based geometry to capture sub-millimeter detail, including pores and individual hair strands, and support physically accurate relighting.[1] These advances demonstrate that visual appearance continues to evolve rapidly. In parallel, neural text-to-speech (TTS) systems have reached a level of expressiveness that enables voices with nuanced prosody, speaking style, and emotion [88], further raising expectations for coherence between an agent's visual and vocal identity.

While such advances in appearance and voice provide increasingly compelling embodiments, recent progress in multimodal representation learning, large language models, and multimodal generative models has transformed the behavioral capabilities of virtual agents. Systems that once relied on handcrafted rules can now generate behavior aligned with linguistic content, conversational rhythm, and interaction context, while adapting to user signals and expressing style, personality, or affect [49]. Recent state-of-the-art systems illustrate this convergence across modular generative architectures, speech, and multimodal behavior generation, including MeLaX for LLM-based agent orchestration [33], VoiceX for personalized agent voices [105], and holistic gesture models such as EMAGE [60]. A timeline of research on socially interactive agents is shown in Fig. 1.

This survey synthesizes contemporary modeling paradigms and identifies the challenges that must be addressed to achieve robust, real-time social interaction with artificial agents. While prior work, including the Handbook on Socially Interactive Agents [62, 63], provides broad coverage of research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics, we focus specifically on recent generative approaches that unify verbal and nonverbal behavior, and on the social-interaction mechanisms that both motivate and constrain these models.

[1] https://nvidianews.nvidia.com/news/digital-humans-ace-generative-ai-microservices

✉ Elisabeth André
andre@informatik.uni-augsburg.de

[1] University of Augsburg, Universitätsstraße 6a, 86159 Augsburg, Germany

[2] Technical University of Applied Sciences Augsburg, An d. Hochschule 1, 86161 Augsburg, Germany

Published online: 10 August 2026

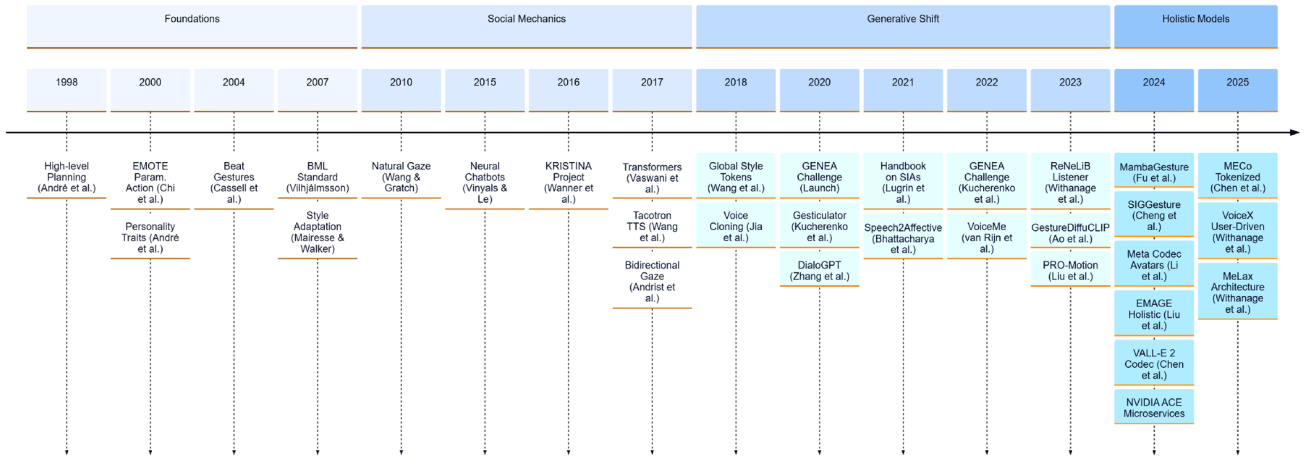

Fig. 1 Timeline of research on socially interactive agents

The paper is structured as follows. We begin by outlining social foundations of interaction (Sect. 2), followed by applications of virtual agents (Sect. 3). We then review core principles (Sect. 4) and architectural approaches for building socially interactive agents (Sect. 5). Next, we survey advances in language, speech technologies (Sect. 6) as well as multimodal behavior generation (Sects. 7 and 8), before discussing challenges in evaluation (Sect. 9). The paper concludes with reflections and future work (Sect. 10).

2 Social Foundations of Interaction

To understand how multimodal behavior supports meaningful interaction, we first consider the social processes that govern how humans interpret and respond to artificial agents. A recent systematic review by Etienne et al. [35] how users attribute emotions, personality traits, and social competence to virtual agents on the basis of multimodal cues such as gaze patterns, facial action units, gesture kinematics, and vocal prosody, and identifies design variables that reliably shift these attributions.

A central goal of designing these behaviors is to elicit a positive social response from the human user, often measured through *social presence*—the feeling of being with another social entity [75]. This sense of co-presence can be influenced by an agent's realism and gaze control [39], and multimodal behavioral cues can serve as predictors for its emergence [73]. Achieving social presence is a gateway to establishing higher-level social connections, such as *rapport* and *trust*. Rapport, or a sense of mutual understanding, is critical for agents in roles that require a strong interpersonal connection between interlocutors, such as health counselling [68], and can be modeled by tracking rapport dynamics over time [19]. Trust is similarly essential, especially for tasks involving self-disclosure [108] or sensitive information [84]. Trust can be influenced, among other things, by familiarity [76], the agent's appearance-voice consistency [3], and can be measured, among other things, by the frequency of advice-seeking [58].

These social outcomes are not automatic; they are built through the nuanced mechanics of social interaction. *Gaze* is perhaps the most-studied social cue. Early work focused on automating "attending behaviors" [26], but researchers quickly found that constant mutual gaze is unnatural and can decrease rapport [98]. This led to more sophisticated, controllable models of gaze aversion for both virtual agents [7] and robots [8] that signal cognitive effort or manage conversational flow. Bidirectional models, in which the agent's gaze is coordinated with the user's, have been shown to improve collaborative task performance [6]. Similarly, *proxemics*, or the use of social space, is a key non-verbal behavior. Studies show that users maintain a comfortable personal space when approaching virtual agents in AR [47] and that an agent's emotional expression (e.g., anger) can increase this preferred distance [15].

Conversational mechanics are also crucial. Agents must manage *turn-taking* [48], appropriately handle user *interruptions* [41], and use *conversational fillers* (e.g., "uhm") to manage response delays, which can make the agent feel more natural [16, 72]. Even subtle social behaviors like *mimicry* [45, 97] and adherence to *politeness* strategies [86] can significantly impact user perception of socially-interactive agents and the outcome of an interaction.

To support credible socio-affective behavior, virtual agents are often equipped with computational models of affect and personality [40, 80]. These models draw on established psychological theories of emotion [74] and personality [69], enabling agents to exhibit coherent internal states rather than merely executing scripted actions. Computational models such as ALMA [40], EMA [67], FAtiMA [31], and WASABI [13] share a common conceptual orientation:

Fig. 2 An older person interacting with the *KRISTINA* multilingual conversational agent, designed to provide companionship, cultural mediation, and health-related support within a smart home environment

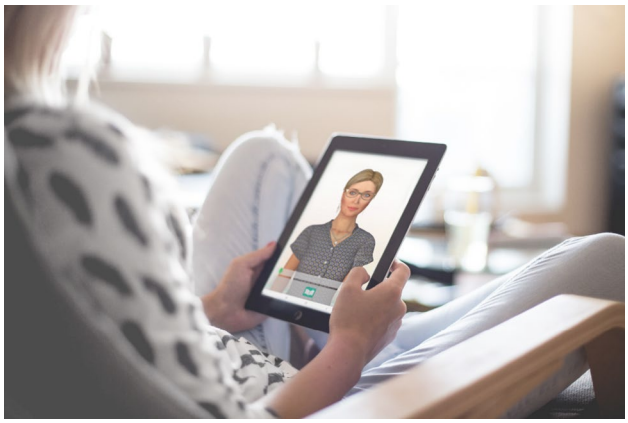

Fig. 3 A patient interacts with the UBIDENZ virtual agent embedded in a mobile application. During their conversation, the sensor data gets used to adapt the avatar's behavior

they ground affective behavior in internal processes rather than surface-level display. Most of these systems build on cognitive appraisal theory; particularly the OCC model [74], which derives emotion types from an agent's evaluation of events, actions, and objects relative to its goals, standards, and attitudes, while others, such as WASABI, additionally incorporate sub-symbolic dynamics inspired by embodied affect.

A vast body of research focuses on the socio-affective perception of an agent's multimodal behavior. A comprehensive review of the socio-affective perception of agents' multimodal behavior [35] shows how users infer emotions, personality, and social capabilities from cues such as gaze, facial expression, gesture, and prosody, offering guidelines for designing socio-affective agents. Studies have shown that an agent's perceived awareness [1] and non-verbal expressions of empathy [66] are key to positive user perception.

3 Applications of Virtual Agents

Healthcare and mental well-being: The healthcare sector presents significant opportunities for socially competent virtual agents, especially in contexts where service demand exceeds capacity. Areas such as eldercare and mental health are particularly in need of technological augmentation. Demographic shifts and workforce shortages contribute to gaps in accessibility and quality. Likewise, mental health services are burdened by long wait times and poor system integration. Virtual agents capable of understanding emotional and social cues can meaningfully enhance both sectors through companionship, monitoring, and adaptive support. An example of intelligent agent technologies addressing eldercare and mental health is the EU-funded project KRISTINA [101], which developed a multilingual conversational companion for older adults. Integrated into a smart home environment (Fig. 2), the agent monitors sleep, reminds users to take medication, and detects emergencies, thereby supporting safety and well-being. The embodied agent is equipped with social and cultural competencies and applying social signal analysis to interpret facial, gestural, and multilingual verbal cues and provide culturally appropriate emotional support that may help reduce loneliness and depression.

The nationally funded UBIDENZ project [32] applied similar technologies to address the limited availability of outpatient mental health care in Germany. It developed an empathic virtual agent that combined automated condition assessment with therapy-specific responses to support individuals after clinical treatment, prevent relapses, and enhance digital therapeutic assistance (see Fig. 3). The dialog content was based on the ABCZ Model [83] and the BDI-II [12] self-assessment questionnaire. The system used a rule-based approach in which avatar responses were carefully selected and modeled by psychological experts due to safety concerns. However, attempts are being made to automatically learn and provide appropriate responses leveraging advances of large language models (LLMs; see Sect. 6).

Education and training: Soft skills training is a very promising domain for the application of socially competent virtual agents. The virtual agents enable the simulation of uncomfortable situations such as job interviews, negotiations, or conflict resolution scenarios in a controlled environment. This allows individuals to safely practice communicative competencies and strategies without the risks associated with real-world interactions.

An illustrative example is the EU project TARDIS, which uses role-play with a virtual agent to help adolescents prepare for job interviews (see Fig. 4). A study showed that adolescents trained with TARDIS performed significantly better in a subsequent interview with a human trainer

Fig. 4 A student engages in role-play with a virtual character serving as a job interviewer

Fig. 5 A trainee teacher engaging with the MITHOS virtual reality training system. The trainee interacts with simulated students, whose behavior is driven by conversational agents responding dynamically to non-verbal cues

than those in a control group [29], appearing less nervous and maintaining more eye contact, indicating increased self-confidence.

Another example is the German MITHOS project [20], which developed a virtual-reality training scenario for student teachers (see Fig. 5). In this setting, conversational agents take on the role of pupils and react dynamically to the trainee's behavior, with their responses informed by analyses of classroom interaction.

4 Core Principles of Socially Interactive Agents

The social phenomena discussed in the previous sections translate into several computational requirements for virtual agents.

A defining challenge of multimodal behavior generation is *coordination*: human communicative behaviors unfold across multiple channels that must align temporally and semantically. Current work still builds on classical principles of embodied conversational agents [18, 92], but must now satisfy them under substantially different constraints: neural generators replace "hand-authored" rules, real-time sensing replaces offline annotation, and evaluation must account for the stochastic and context-dependent nature of learned behavior [71]. As a result, coordination, responsiveness, and context sensitivity are no longer design specifications to be implemented but emergent properties to be verified. Gestures synchronize with prosodic accents; facial expressions modulate affective tone; gaze regulates turn-taking and attention; and posture conveys stance, engagement, or openness. These signals operate on different timescales, yet they must be integrated into a coherent whole. Failure to achieve such coordination leads to a breakdown of naturalness and can disrupt rapport or trust, as users are acutely sensitive to mismatches between speech and nonverbal behavior [18, 87].

Closely related is the principle of *responsiveness*. Social interaction is inherently dynamic and contingent: appropriate behavior depends not only on internal planning but also on the agent's ability to react promptly and meaningfully to unfolding user actions. Responsive agents continuously update their behavior in response to the partner's verbal and nonverbal signals, such as interruptions, changes in affect, gaze shifts, or turn-taking cues. Achieving responsiveness poses substantial computational challenges, as it requires low-latency perception, prediction of interlocutor intent, and real-time adaptation of multimodal output. At the same time, responsiveness entails a trade-off: tightly scripted systems can guarantee safe and predictable reactions, but they often fail when users deviate from expected paths; open-ended generative systems can react more flexibly, but they may introduce latency, inconsistency, or inappropriate responses in socially sensitive situations.

At the same time, artificial agents must generate behavior that is *sensitive to context*. Human communicative behavior varies depending on physical environment, interpersonal distance, emotional state, cultural background, conversational role, and interlocutor behavior. Generative models, therefore, require conditioning signals that capture linguistic content, prosody, semantics, affect, user behavior, or stylistic intent. Modern neural architectures integrate these cues directly: speech, text, affective embeddings, or interlocutor signals influence the latent representations from which gestures, facial expressions, or full-body motion are synthesized. This stands in contrast to earlier rule-based systems, in which mappings from linguistic features to gestures were manually specified and often brittle [53, 92].

Another core concept is *expressive diversity*. Human behavior is inherently variable, no two gestures are exactly

the same, even when conveying similar communicative intentions. Successful generative models must therefore avoid deterministic outputs and instead capture the stochastic nature of human motion. Neural architectures achieve this through probabilistic decoders, latent diffusion processes, tokenized motion representations, or style embeddings, enabling the models to generate multiple plausible variations (see Sect. 7). This diversity supports more natural interaction and enables agents to convey style, personality, or emotional nuance (see Sect. 8).

Finally, multimodal behavior generation is inherently *holistic*. While early systems treated different modalities in separate modules, modern approaches increasingly model these channels jointly. Full-body generative models capture dependencies across the limbs, torso, facial musculature, and gaze, producing more fluid and integrated behaviors. This holistic perspective reflects the underlying reality of human communication: behaviors across channels are not independent but are mutually reinforcing components of a single communicative act.

Taken together, these concepts—coordination, responsiveness, context sensitivity, expressive diversity, and holistic integration—define the computational challenges and opportunities of multimodal behavior generation. They frame the landscape in which contemporary technical approaches operate, from neural sequence models and diffusion-based synthesis to tokenized motion representations and LLM-conditioned behavior planners. They also expose a persistent tension in the field: rule-based architectures offer interpretability, explicit control, and verifiable safety guarantees, but they require extensive manual authoring and degrade quickly outside their designed interaction scope. Neural models, by contrast, learn flexible and variable behaviors from data, but their outputs are difficult to constrain to domain-appropriate ranges, hard to evaluate against communicative intent, and largely opaque when failures occur—a combination that is especially problematic in safety-critical applications such as healthcare or counselling.

The following sections examine these approaches in detail, highlighting how different architectures incorporate linguistic, affective, and contextual cues to produce coherent and socially appropriate expressive behavior.

5 Building Socially Interactive Agents

The development of socially interactive agents encompasses the computational processes that enable artificial characters to produce coordinated verbal and nonverbal signals, such as speech, gestures, facial expressions, gaze, posture, and affective cues, that together shape how the agent is perceived and how effectively it communicates. Traditional pipelines for socially interactive agents separated perception, dialogue, affect modeling, and animation into distinct modules (see Fig. 6). Classical systems such as SAIBA/BML and SmartBody established this modular view by defining explicit markup layers that separate communicative intent from behavior realization [87, 92]. Recent platforms retain modular separation but replace hand-authored intent planners and animation schedulers with data-driven generative components, as illustrated by the MeLaX architecture described below [33]. Although this architectural separation has become blurred in modern neural systems, the underlying idea remains: behavior generation must be driven by representations of what the agent intends to communicate and how it should express itself in the current interactional context.

MeLaX [33] is designed to integrate distinct generative technologies into a single real-time platform, unifying speech recognition, dialogue management, and behavior synthesis through a zero-code graphical interface. The system relies on an event-driven pipeline where data flows from audio classification and transcription to response generation. To ensure the interaction remains coherent and fluid, a Conversation Orchestrator manages contextual flow together with a memory retrieval mechanism such as Retrieval Augmented Generation (RAG). The orchestrator utilizes a traditional state machine to track conversational turns, maintaining a record of the agent's status—whether it is currently speaking, listening, or processing. This explicit state management enables handling real-time conversational dynamics, including barge-in mechanisms that allow users to interrupt the agent and back-channeling behavior while the agent is listening. While this architectural framework provides the necessary structure for real-time control, the communicative power of the individual modules is increasingly derived from the rapid advances in generative models discussed in the following section.

6 Socially Interactive Agents Empowered by Modern Language and Speech Technologies

Language and speech technologies form the backbone of socially interactive agents, and their development has been shaped by major advances in natural language processing and neural generative modeling. large language models (LLMs) now enable fluid dialogue, while modern neural text-to-speech (TTS) systems provide highly expressive voices. These advances fundamentally reshape how agents generate language, speak, and coordinate multimodal behavior. At the same time, these capabilities expose concrete unsolved problems: LLM-generated dialogue can be

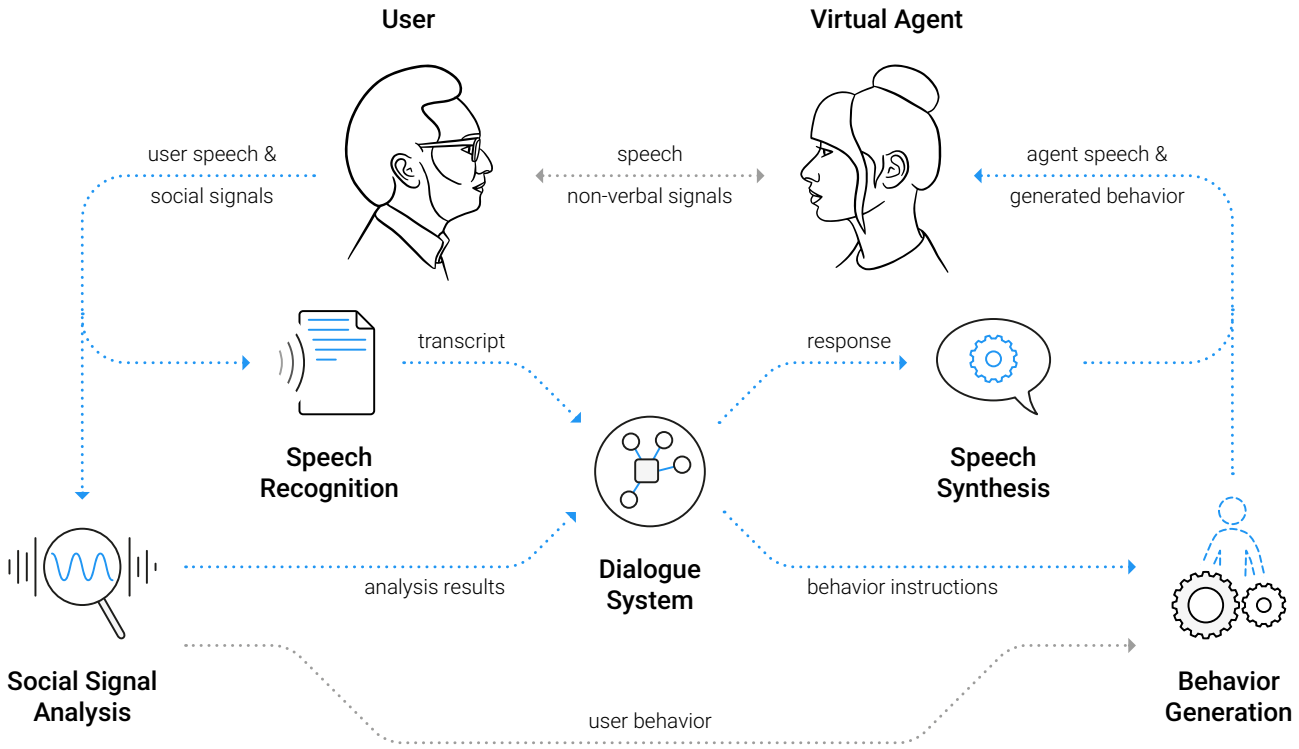

Fig. 6 High-level architecture of an emotionally intelligent virtual agent. The user produces speech and non-verbal social signals that are processed through speech recognition and social signal analysis pipelines. These feed into a central dialogue system, which generates an appropriate multimodal response. Speech synthesis and behavior generation modules then transform the system's decisions into verbal and non-verbal behaviors displayed by the virtual agent, completing the interaction loop

fluent yet factually ungrounded, producing plausible but incorrect or fabricated claims; the models' decision processes are largely opaque, making it difficult to audit why a particular response was chosen; and their apparent social competence can mislead users into attributing understanding or empathy that the system does not possess. The current state of the art is therefore best understood as a continuation of earlier dialogue-system research, from rule-based chatbots [28, 102] to neural response generation [93, 110] and LLM-based emotional support [111], with each generation trading one set of limitations for another.

6.1 Advances Through LLMs

Early conversational agents, or chatbots, were mostly built on handcrafted rules for dialogue management, as illustrated by systems such as ELIZA [102] and PARRY [28]. As such, they were confined to specific domains, scripts, or interaction patterns. Recent advances in large language models (LLMs) have fundamentally changed this landscape. Trained on massive text corpora, LLMs can now generate dialogue that often appears remarkably fluent and human-like and can flexibly handle a broad range of topics, tasks, and conversational situations. Leveraging these capabilities, researchers and developers are building increasingly sophisticated virtual agents. LLMs can also be embedded into agent architectures with memory, planning, retrieval, and tool-use components. However, each additional module introduces new failure modes—retrieval can surface irrelevant or outdated context, planning can commit to goals the user did not intend, and tool use can trigger real-world actions on the basis of misunderstood instructions—making it harder to trace responsibility for errors and increasing the need for explicit audit trails, fallback mechanisms, and human-in-the-loop oversight. However, these models did not emerge overnight; they evolved from predecessors over many iterations (see Fig. 7).

An early fully data-driven approach that employed neural networks to learn conversational patterns was introduced by Vinyals and Le [93]. Their model was one of the first attempts to develop an end-to-end trainable chatbot: it would simply predict the next sentence given the previous conversation. While groundbreaking, these early neural agents suffered from well-documented issues: generic replies, inconsistent persona, and limited conversational grounding [56, 79, 109, 110].

Subsequent work began addressing these shortcomings. Notably, Zhang et al. [109] proposed conditioning neural

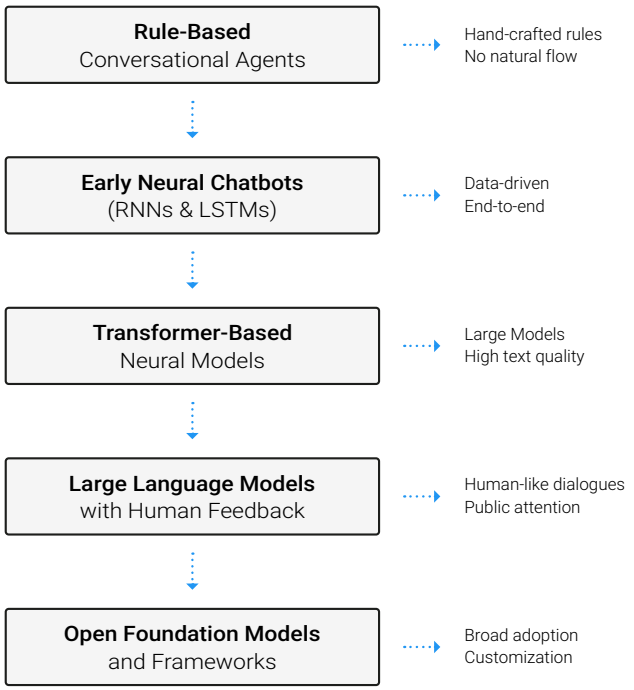

Fig. 7 The evolution of dialogue systems and LLMs

dialogue models on fictional personas, allowing the system to adopt stable speaking styles and preferences. This idea—injecting an identity, backstory, or style into a dialogue agent—has persisted and is now implemented through prompting strategies in modern LLM-based systems.

Early neural dialogue systems based on recurrent networks were limited by small datasets, but the introduction of the Transformer architecture [91] and large-scale pretraining fundamentally reshaped the field. Models such as GPT [77] and DialoGPT [110] showed that fine-tuning large pretrained transformers enables far more fluent and contextually appropriate dialogue than training models from scratch. This progression culminated in ChatGPT, trained with Reinforcement Learning from Human Feedback (RLHF), which brought open-domain conversational agents into mainstream use and demonstrated their viability in everyday interaction. Its impact spurred rapid developments, including Google's LaMDA/PaLM-2, Anthropic's Claude, and Meta's LLaMA, and the emergence of toolkits that simplify LLM integration. Today, LLM-based dialogue generation underpins most modern conversational agents.

A recent challenge in the area of socially interactive agents concerns the use of LLMs to build emotionally supportive virtual agents. Although LLMs can be fine-tuned or strategically prompted to deliver empathetic, psychologically grounded responses, research shows that ensuring consistency, safety, and genuine emotional appropriateness remains an open problem (see the survey by Zheng et al. [111]). Emotional support is a complex, high-stakes domain: effective responses must balance validation, guidance, and warmth while avoiding harmful suggestions, overgeneralization, or inappropriate self-disclosure. To address these challenges, recent work has begun to explore the use of LLMs not only as generators but also as evaluators of emotional support behavior; see, for example, the approach by Madani and Srihari [64].

6.2 Advances in Speech Synthesis Technology

Speech is a primary channel of human communication and not only conveys linguistic content but also affective and social signals that shape how messages are perceived. In the context of virtual agents, the ability to generate speech that captures emotional nuances is important for building trust and rapport, and supporting natural interaction. Neural text-to-speech (TTS) technologies have substantially expanded the expressive capabilities of such virtual agents. State-of-the-art TTS systems allow fine-grained, continuous, and interpretable control of vocal emotion and speaking style. Recent surveys of expressive speech synthesis emphasize that controllability, naturalness, emotional nuance, and speaker identity preservation are now central criteria for agent voices [11, 88]. One milestone of unsupervised neural style modeling was achieved by Wang et al. with their *Global Style Tokens* (GST) [100]. They introduced a bank of latent embeddings within Google's Tacotron [99] architecture. The model was trained to learn embeddings representing diverse prosodic styles, without requiring labeled data. Their approach enabled control over speaking style and supported the transfer of stylistic characteristics from existing speech to newly generated samples. Their work was further extended by Stanton et al. [82] who modified the architecture to infer the style embeddings from text only. As a result, the system no longer depended on reference audio samples. While the use of GST, in theory, allowed for controlling prosody and thus certain aspects of affect, the GST embeddings themselves were not interpretable by human users. Consequently, subsequent research focused on more comprehensible alternatives to steer the speech synthesis process. For instance, Ren et al. [78] introduced *FastSpeech 2*, which enabled direct manipulation of prosodic features such as duration, pitch, and energy (Fig. 8).

Another line of research that has become highly relevant for virtual agents is *voice cloning*. Voice cloning is the ability to generate speech in the voice of a specific target speaker, using only a limited amount of reference data from that speaker. Here, early advances combined a speaker encoder with Tacotron-based synthesizers and neural vocoders. For instance, Jia et al. [50] demonstrated that with just a few seconds of audio, a model could synthesize speech in a new voice while preserving intelligibility and

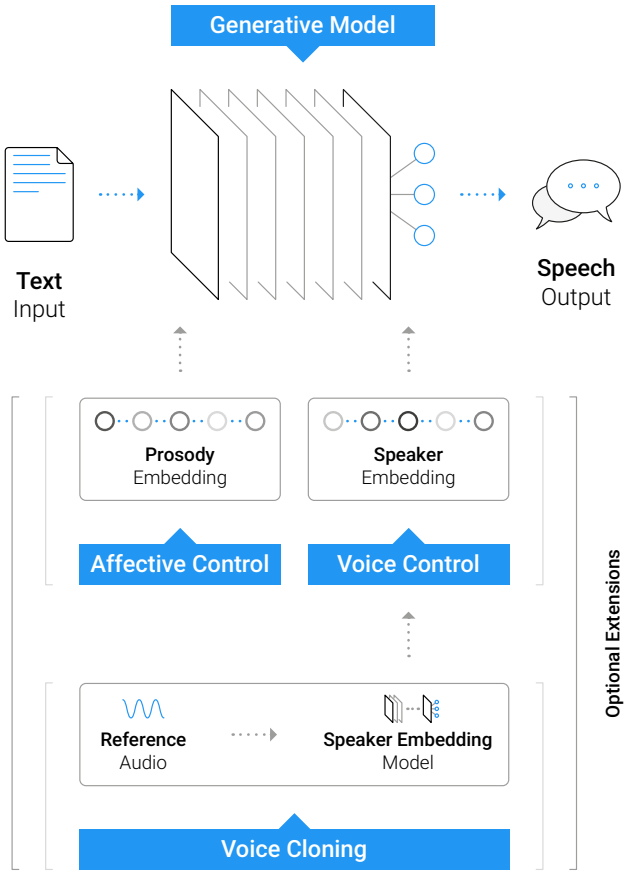

Fig. 8 Data flow in speech synthesis systems with optional affective control, voice control, and voice cloning

voice timbre. This approach laid the foundation for *few-shot* and *zero-shot* speaker adaptation, providing developers with a means to build virtual agents with distinct and personalized voices at low data cost. Subsequent work by Qian et al. [58] disentangled speaker identity from prosodic content, allowing emotional tone to be transferred from one speaker to another while maintaining identity. More recent models use *neural codec language modeling*, i.e., treating speech as a sequence of discrete audio tokens. For instance, VALL-E and VALL-E 2 [23, 96] enable zero-shot cloning from a short prompt and preserve not only emotional coloration, but also more subtle characteristics like recording conditions from reference audio. In parallel, multilingual frameworks, e.g., YourTTS [17], aim to extend zero-shot voice cloning across different languages, thereby maintaining non-verbal voice characteristics, even when switching between languages.

Beyond cloning existing voices, researchers have increasingly explored how to personalize and customize voices for virtual agents in ways that reflect user preferences and character design. For instance, van Rijn et al. introduced *VoiceMe*, an approach that allows users to explore the latent speaker space of a TTS model and create entirely new synthetic voices that perceptually align with certain characters [89]. They also showed that participants could navigate the high-dimensional latent style space of GST-Tacotron models to identify clusters corresponding to emotional prototypes, which could then be transferred to new utterances [90]. More recently, *VoiceX* [105] introduced a framework that combines human-in-the-loop evolutionary search with a user-friendly interface, allowing users to iteratively converge on personalized voices through simple pairwise comparisons. Such methods shift personalized speech synthesis from being purely data-driven to being *user-driven*; that is, end-users are empowered to design voices that not only carry emotional nuance but also align with personal or narrative expectations. For virtual agents, this means that affective TTS is no longer limited to reproducing existing voices or generic emotional categories. Instead, agents can be equipped with unique, customizable identities whose expressivity is tuned for specific contexts of interaction.

7 Advances in Multimodal Behavior Generation

Multimodal behavior generation has undergone rapid change over the past decade, driven by advances in deep learning and large-scale multimodal datasets. Approaches that once relied on handcrafted symbolic AI models have been largely replaced by neural architectures that learn the semantics, expressive, and stylistic nuances of human behavior, as well as their temporal sequence, directly from the data.

As shown in Fig. 9, modern frameworks now orchestrate these complexities through hierarchical layers ranging from social perception to continuous motion synthesis. It is important to note that the synchronization required in these frameworks is not limited to the agent's own modalities (*intrapersonal*), but is also fundamentally *interpersonal*. Agents must not only coordinate their own gestures and speech, but also dynamically adapt their behavior to the timing and actions of users and other agents. This section presents the main families of generative models for synthesizing gestures, facial expressions, and whole-body movements, focusing on how they incorporate linguistic, affective, and contextual cues into this complex multimodal output. The state of the art is characterized by a move from early speech-driven sequence models [2, 54], which often produced averaged motion with limited semantic control, toward three complementary paradigms: adversarial models that push generated motion toward perceptual realism by training against human-motion discriminators [107], diffusion-based models that generate diverse, high-fidelity gesture variants through iterative denoising [9, 24], and

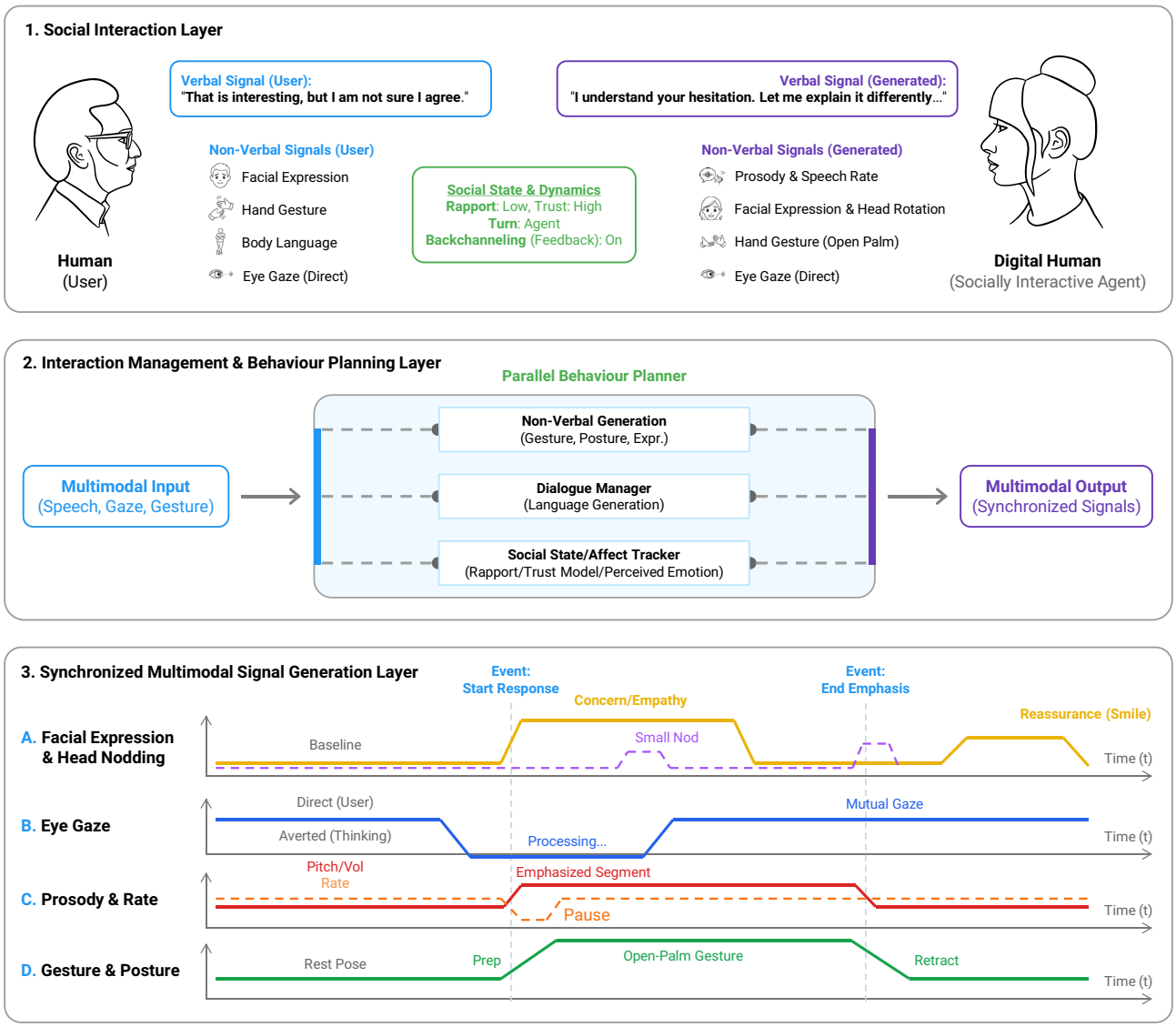

Fig. 9 A multi-layered architecture for modeling complex multimodal behaviors: This diagram details the hierarchical dependencies required to synthesize lifelike interaction. The process is divided into three functional stages: layer 1 (social signal perception) encodes dynamic user inputs, such as gaze and prosody, to establish interactional context and layer 2 (intent and state modeling) serves as the cognitive core, integrating these signals with linguistic goals to derive high-level behavioral strategies and layer 3 (continuous behavior synthesis) transforms these abstract states into synchronized, frame-by-frame time-series data for full-body motion generation

tokenized motion models that discretize movement into reusable units, enabling flexible recombination and LLM-compatible conditioning [21, 60](Table 1).

7.1 Procedural Generation

Before the rise of data-driven neural methods, virtual agent behaviors were synthesized using predetermined scripts or hand-engineered symbolic AI models. Typically, a high-level planner would decide what the agent should do or say to achieve goals, and then an animation model would determine how those actions are performed stylistically [4].

A classic strategy for coupling speech and gestures was to trigger simple beat gestures on prosodic peaks, e.g., a hand movement timed to a stressed syllable [18]. More complex iconic or deictic gestures were usually selected from pre-authored libraries based on keywords in the utterance or dialogue context. As an important extension of this paradigm, Kopp et al. [53] introduced a procedural microplanning approach that derived the form of iconic gestures directly from communicative goals, rather than selecting them from a fixed gesture lexicon, enabling the on-the-fly construction of novel, semantically grounded gestures.

Early work showed that tweaking parameters for personality could systematically alter multiple aspects of an agent's

Table 1 Strengths and limitations of major computational paradigms for socially interactive agents

Paradigm	Strengths	Limitations
Rule-based and procedural models	High interpretability, explicit control, and predictable behavior in constrained domains	Require substantial authoring effort, are brittle under open-ended interaction, and scale poorly to diverse contexts
Affective and cognitive architectures	Make internal states, appraisal variables, and personality assumptions explicit	Translating cognitive theories into computational models often requires simplifying assumptions and hand-designed mappings from internal states to behavior
Neural sequence and adversarial models	Learn temporal regularities and motion style directly from data	May produce averaged or unstable motion and often provide limited semantic control
Diffusion and tokenized motion models	Support diverse, high-fidelity, and flexibly conditioned behavior generation	Can be computationally demanding and do not automatically provide causal, real-time social adaptation
LLM-based and modular generative systems	Enable fluent dialogue, flexible personas, memory integration, and rapid prototyping of complete agents	Remain vulnerable to hallucination, weak grounding, opaque decision processes, and error propagation across modules

behavior [5]. In practical terms, a rule-based persona might specify that an "extraverted" virtual presenter uses broader gestures and a closer speaking distance, whereas an "agreeable" character nods frequently and speaks in a warmer tone. Several groups incorporated psychological models, such as the Big Five traits, to drive consistent behavioral variation in agents. For example, Chi et al. [25] introduced the EMOTE model, which parameterizes gesture "effort" and "amplitude", allowing character motion to be systematically modulated to reflect different personality types. Similarly, Walker et al. [65] proposed a method for adapting an agent's linguistic style based on its personality traits. These handcrafted mappings ensured control over verbal and nonverbal cues, but required substantial expert effort to design.

In classical architectures for earlier socially interactive agents, behavioral control was realized through structured pipelines, such as the SAIBA framework, where communicative intent was transformed into Behavior Markup Language (BML, [92]) specifications describing gesture types, gaze shifts, facial expressions, and their timings. Around the same time, platforms like SmartBody [87] acted as a general behavior realizer converting high-level BML commands into continuous, blended animation in real time.

7.2 Early Neural Sequence Models

Neural approaches marked a shift from explicit rules toward data-driven learning of motion regularities. Early neural models [46] demonstrated that realistic human motion could be synthesized from high-level control parameters by learning compact latent representations of human movement from motion-capture data. Feedforward networks mapped abstract inputs, such as locomotion paths or target locations, into smooth, natural motion trajectories. Subsequent work [2] shifted focus to conversational behavior, where motion is no longer driven by explicit control parameters but must instead be aligned with speech. Neural sequence models, as exemplified by Gesticulator [54], provided evidence that both the timing and expressive qualities of gestures can be learned directly from the temporal structure of speech. These early approaches used recurrent or autoregressive architectures to predict gesture trajectories from prosodic and semantic features. Complementary analytical work [37] showed that some gesture parameters, such as path length, are strongly predictable from speech, whereas others, such as velocity, are less so. This line of work was extended beyond hand gestures to include head and eye movements in multiparty settings [51], where the dynamic directions of the eyes and heads of all interlocutors were predicted from speech signals.

Efficient sequence models and selective state spaces. A promising recent development for multimodal behavior generation is the use of structured and selective state space models. Earlier structured state space models showed that long sequences can be modeled efficiently without relying exclusively on recurrent or attention-based architectures [43]. Mamba extends this line of work with input-dependent selection mechanisms and linear-time sequence processing [42], while Mamba-2 further connects state space models and attention through structured state space duality [30]. For socially interactive agents, these models are relevant because they can support long temporal context over speech, gesture, gaze, facial expression, and user behavior while keeping inference efficient enough for real-time interaction. This makes them especially promising for co-speech gesture generation, attentive listener behavior, turn-taking, and multimodal adaptation. However, selective state space models should be understood as improving temporal modeling and scalability; they do not by themselves solve grounding, safety, or social reasoning.

7.3 Adversarial Networks

Adversarial training further improved realism by encouraging generated motion to match the statistical properties of human gestures. In generative adversarial frameworks,

a generator produces samples, and a discriminator evaluates their quality, pushing the generator toward more realistic virtual-human behaviors. An example includes GestureGAN, presented by Tang et al. [85]. While Tang et al. focused on the realism of the generated gestures, Ferstl et al. [36] presented multi-objective adversarial models that decomposed gesture quality into sub-objectives (plausible gesture dynamics, smoothness, and realistic joint configuration) and monitored them by separate adversaries. Related work by Yoon et al. [107] used adversarial training to generate gestures that convey the style of a particular speaker while remaining aligned with speech content and rhythm. While these models significantly improved motion fidelity, control remained largely implicit, with conditioning signals often entangled in latent representations rather than explicitly modeled. Thus, adversarial learning is powerful for visual plausibility, but adversarial models can be unstable to train and may optimize realism more strongly than communicative appropriateness.

7.4 Diffusion-Based Models

Diffusion models have become a dominant paradigm for motion synthesis due to their robustness, high fidelity, and controllability. By iteratively refining noise into coherent motion trajectories, they naturally capture the variability and fluidity of human gestures. Crucially, diffusion-based models support flexible conditioning on diverse control signals, including speech, text, prosody, affective state, style, and interlocutor behavior (see Sect. 8).

MambaGesture [38] builds on this paradigm by generating gesture trajectories through iterative noising and denoising, conditioned on audio and multimodal cues, including text, emotion, and style. It illustrates how selective state space models can be combined with generative behavior models to improve temporal coherence and multimodal fusion. GestureDiffuCLIP [9] adds style-aware conditioning to a latent diffusion framework: gestures are generated in a compact latent space, with style embeddings from text, exemplar motion, or video modulating the denoising process to produce semantically aligned yet stylistically consistent behavior. This is enabled by CLIP [77], a representation learning model trained with a contrastive objective to align multiple modalities in a shared embedding space.

Beyond single-stage diffusion, some approaches also adopt multi-stage generation pipelines. For instance, SIGGesture [24] adopts a two-stage design: A large-scale rhythmic gesture generator is first trained, and its generated motion sequences are subsequently used as prompts to guide a foundational diffusion model for gesture synthesis. While the rhythmic generator captures the temporal alignment and motion energy of gestures with speech, the subsequent synthesis stage is required to determine the semantic form of the gesture, including hand shape, spatial trajectory, and coordination with other modalities. Diffusion models are particularly attractive because they generate diverse alternatives rather than a single average motion, but their iterative sampling can be computationally costly and can complicate low-latency interaction.

7.5 Tokenized and Discrete Motion Models

An alternative to continuous motion prediction is to tokenize motion sequences using vector-quantized variational autoencoders (VQ-VAEs). These approaches discretize gesture or full-body motion into sequences of codebook entries, i.e. learned motion units analogous to linguistic tokens, enabling generative models to reuse powerful techniques from NLP. Tokenized models are particularly effective for style transfer, stochastic variation, and the generation of multiple plausible outputs conditioned on the same input, as diversity arises from sampling and recombining discrete motion units.

ReNeLiB [106] adopts this approach to learn nuanced listener behaviors by training on real-life therapy interactions, focusing on modeling expert-level attentive responses. The system encodes the variability of listeners' facial motion into a VQ-VAE codebook, capturing rich, reusable motion tokens that reflect empathetic, context-sensitive reactions. Recent work has extended this listener-oriented perspective toward real-time nodding behavior: Kato et al. [52] predict both the timing and type of nods for an avatar attentive-listening system, illustrating that listener generation is increasingly moving from offline synthesis toward online interaction control.

To synthesize these behaviors, ReNeLiB uses a transformer architecture that autoregressively predicts future facial animation sequences based on the speaker's behavior and the encoded style representation of the listener (see Fig. 10).

At the same time, Voss and Kopp [95] applied a similar methodology to the field of communicative gestures made by speakers with their AQ-GT framework (Aligned and Quantized GRU-Transformer). Like ReNeLiB, AQ-GT use a discrete latent space (specifically, a VQ-VAE-2 trained with adversarial loss) to ensure high-resolution movements. However, instead of modeling the reactive behavior of listeners, AQ-GT bridges the gap between semantic intent and neural synthesis. It integrates a GRU-Transformer to map multimodal contexts, including text and audio, directly onto these discrete motion tokens, demonstrating that such representations can successfully capture the subtle semantic dependencies required for speech-driven generation.

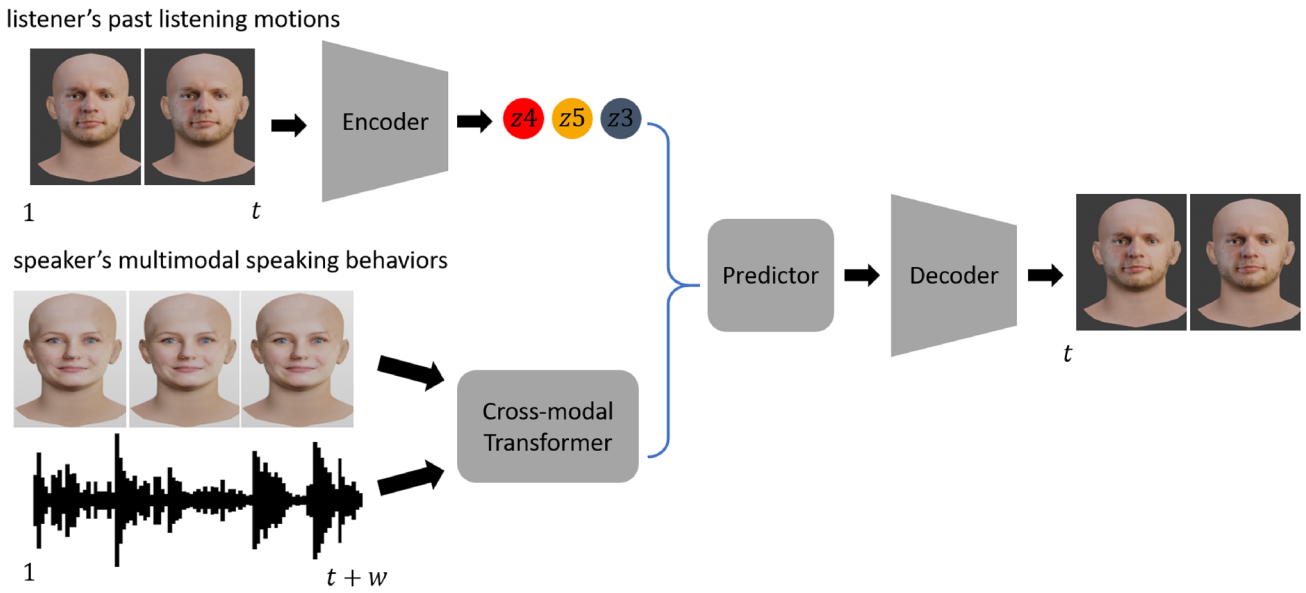

Fig. 10 Predicting appropriate listener behavior in ReNeLiB

More recent work integrates tokenization with LLMs. The MECo framework [21] uses tokenizers to convert both speech and motion exemplars into discrete sequences, which are then jointly processed by an LLM to generate stylistically consistent co-speech gestures. This approach removes the need for pseudo-labels and allows fine-grained control over motion style, body parts, and multimodal inputs (motion clips, static poses, videos, or textual prompts). Holistic tokenized models, such as EMAGE [60], extend these ideas to full-body motion including hands, body, head, and facial gestures using four compositional VQ-VAEs (for the face, upper body, hands, and lower body), and masked-gesture modeling. The advantage of tokenization is that it makes motion easier to model and combine with language-like representations; a limitation is that compact codebooks may discard subtle motion details and may encode dataset-specific biases.

7.6 Summary

Recent progress in multimodal behavior generation reflects a shift from hand-engineered mappings toward data-driven, highly expressive generative architectures. Across these architectures, control mechanisms play a central role in shaping expressive behavior. The next section reviews how linguistic, affective, stylistic, and interpersonal cues condition these generative processes.

8 Control and Conditioning of Multimodal Behavior

Human communicative behavior varies with linguistic content, prosodic rhythm, affective tone, the speaker's personality, and the interlocutor's behavior. Therefore, generative models must incorporate mechanisms that shape the form and timing of gestures, facial expressions, and posture according to these contextual factors. This section examines the conditioning signals that guide contemporary models and the strategies used to control the expressive qualities of generated behavior. Recent state-of-the-art work increasingly treats conditioning as multimodal and interactional: language models provide semantic structure [24, 104], affective and stylistic signals shape expressivity [14, 38], and interlocutor-aware models generate contingent listener responses [52, 59, 106].

8.1 Acoustic and Linguistic Conditioning

Earlier data-driven gesture generation models relied primarily on acoustic features to condition motion, which proved particularly effective for capturing rhythmic aspects such as timing and beat gestures (see, for example, the speech-driven approach by Alexanderson et al. [2]). However, paralinguistic signals, such as speech acoustics, provide only limited access to the semantic and discourse-level information required to generate meaningful iconic or deictic gestures. To address this gap, recent approaches have increasingly leveraged large language models, which offer rich semantic representations and world knowledge that can be exploited for generating nonverbal behavior. Rather than

replacing specialized motion models, LLMs are typically used to provide high-level semantic structure that guides downstream generative processes.

In SIGGesture [24] (see Sect. 7), LLMs propose semantic gesture candidates, such as "left" or "right", for given transcripts. The semantically matched gestures are then used to condition a diffusion-based foundation model for generating rhythmic gestures. In PRO-Motion [61], LLMs act as semantic motion planners that convert open-world text prompts, such as "jump on one foot" or "experience a profound sense of joy", into compositional posture scripts, which then condition diffusion- and transformer-based generative modules for posture synthesis and motion generation, combining LLM commonsense with specialized motion models. The Language of Motion [22] introduces a multimodal language modeling approach with modality-specific tokenizers, enabling flexible multimodal generation and editing across text, speech, and motion.

Together, these approaches illustrate that LLMs encode rich semantic structure that can be effectively translated into controllable motion cues. Notably, text-based embeddings may even outperform audio-only baselines for both beat and semantic gestures, as demonstrated by LLAniMAtion [104], which generates gestures directly from LLM-derived embeddings (LLaMA 2).

8.2 Affective Conditioning

Another key dimension is affect. Human gestures and facial expressions are shaped by emotional state, and generative systems increasingly learn affective conditioning through sentiment embeddings, acoustic correlates of emotion, or multimodal encoders that integrate both. Models such as Speech2AffectiveGestures [14] incorporate affect during both generation and discrimination, producing gestures that reflect emotional tone while preserving linguistic congruence. Affective conditioning also supports empathic interaction: agents can adapt their expressive behavior to signal understanding, warmth, or concern, which is essential for applications in healthcare, mentorship, or social coaching. Related diffusion-based approaches, such as MambaGesture [38], support affective conditioning alongside linguistic and stylistic cues, allowing emotional signals to modulate expressive behavior (see Sect. 7).

8.3 Stylistic and Personality Conditioning

A further dimension of conditioning concerns style and personality, which shape the characteristic manner in which individuals gesture. Modeling such stable differences enhances the consistency and believability of virtual agents. Style can be specified through exemplar clips that are encoded into latent style embeddings or through explicit controls that modulate gesture manner.

Tokenized approaches such as MECo [21] and EMAGE [60] represent motion as sequences of discrete gesture units (see Sect. 7), enabling models to reuse stylistic patterns and to modulate specific body regions or expression channels. Because behavior is composed of token sequences, these methods integrate naturally with style-conditioning modules and support applications such as style variation and personality-aware gesture generation.

Yoon et al. [107] use an adversarial approach to generate gestures that convey the style of a particular speaker while matching speech content and rhythm.

GestureDiffuCLIP [9] provides a complementary mechanism for style control within a latent diffusion model (see Sect. 7). Style embeddings derived from text, reference motion, or video modulate the denoising process, allowing gestures to remain semantically aligned with speech while exhibiting consistent stylistic traits.

Together, these approaches illustrate how modern systems combine generative modeling with flexible style representations to produce individualized and expressive behavior.

8.4 Interlocutor-Aware Conditioning

Increasingly, generative models also rely on interlocutor-aware conditioning. Social interaction is inherently bidirectional: the agent's behavior must correspond to the user's behavior. Interlocutor-aware models incorporate signals from the conversational partner to generate facial reactions, nods, head movements, or adaptive gaze shifts that align with the user's timing and affect. These methods move beyond presenter-driven systems toward models capable of regulating engagement, signaling attentiveness, and producing contingent reactions that reflect conversational rhythm. This trend is strengthened by the shift toward holistic models that integrate facial, bodily, and gaze behavior within a single generative framework, enabling coordinated adaptation to the user. Examples include the previously mentioned ReNeLiB system [106] (see Sect. 7), which learns a listener's facial motions from the listener's and the speaker's previous audiovisual behaviors, and the approach presented by Lin et al. [59], which predict turn-taking and backchannel behaviors from linguistic, acoustic visual signals of the speaker. These systems are promising because they address contingency directly, but they also carry specific risks: user-facing agents must continuously process gaze, facial expression, and vocal affect, which constitutes sensitive biometric data; incorrectly inferred states—for example, misclassifying fatigue as disengagement or nervousness as hostility—can cause the agent to respond with inappropriate reassurance,

unsolicited emotional probing, or premature topic changes, any of which may erode rather than build trust.

8.5 Summary

Overall, control and conditioning have become central design principles for generating multimodal behavior. By integrating linguistic, prosodic, affective, stylistic, and interpersonal cues, modern generative models move beyond animation toward social communication. Future research must therefore examine not only whether generated behavior appears natural, but also whether the conditioning variables—affect labels, personality attributions, inferred engagement levels—are valid and appropriate for a given application, whether users are aware of and can contest what is being inferred about them, and whether models provide sufficient audit trails to diagnose failures when socially consequential behavior goes wrong. The next section examines how this progress is evaluated in practice. It also introduces current benchmarking efforts—particularly the GENEA [55] and REACT Challenges [81]—that highlight the need for robust, interactive, and standardized evaluation paradigms.

9 Evaluation Challenges for Multimodal Behavior Generation

Despite remarkable progress in generative models for gesture, facial expression, and full-body motion, evaluating multimodal behavior generation remains one of the most difficult and least standardized aspects of the field. Recent state-of-the-art evaluation work therefore combines benchmark challenges for gesture and reactive behavior [55, 81] with broader methodological reflections on speech synthesis and animated gesture evaluation [11, 71]. Unlike traditional benchmarks in computer vision or natural language processing, the quality of expressive behavior cannot be fully captured through accuracy or likelihood scores. Instead, evaluation must handle the inherently subjective, context-dependent, and interactive nature of human communication. See Bailly et al. [11] for a discussion of challenges in the evaluation of speech synthesis, and Nagy et al. [71] for an overview of modern practices in animated gesture evaluation.

Current metrics only partially reflect these dimensions, and even the most widely used ones have important limitations. Objective metrics quantify aspects of motion statistics or rhythmic congruence, but they capture neither communicative appropriateness nor user perception in real interaction. Subjective rating scales focus on naturalness or appropriateness but rarely account for how gestures coordinate across modalities or support communicative intent. Moreover, evaluations are typically based on isolated clips rather than interactive scenarios, limiting ecological validity.

A key open problem is the evaluation of gestures within a broader communicative system, where behavior is shaped by language, prosody, affect, personality, and the interlocutor's responses. Assessing generated motion in isolation cannot capture these dependencies. More comprehensive methodologies combining objective measures, perceptual studies, and interactive evaluations are required to judge whether behaviors meaningfully support communication and social engagement.

The GENEA Challenge [55] and the REACT Challenge series [81] have made influential contributions by providing shared datasets, standardized training splits, and controlled evaluation protocols. Together, these challenges cover two complementary aspects of multimodal behavior generation: co-speech gesture synthesis in GENEA and reactive listener behavior in REACT. GENEA focuses on generating gestures for predetermined speech clips, whereas REACT focuses on producing appropriate facial reactions in dyadic social settings. Although both challenges still emphasize offline generation and passive observation, they provide robust benchmarks and pave the way for future evaluations in interactive, adaptive, and conversational contexts.

10 Conclusions and Future Directions

Over the past several decades, research on virtual agents has advanced from early attempts to endow computer characters with simple rule-based behaviors to the development of socially aware, multimodal artificial partners capable of fluid, context-sensitive interaction. These systems now integrate techniques from human–computer interaction, graphics and animation, speech and language processing, and affective computing. As a result, the boundary between technically convincing social behavior and cognitively grounded social intelligence has become increasingly important to examine critically. The evolution of virtual agents reflects a broader trajectory in human–agent interaction, one that progresses from reactive to adaptive, from unimodal to multimodal, and from handcrafted scripts toward generative, data-driven competence.

Ethical and societal risks Equipping agents with human-like social characteristics also means they can trigger human-like social biases. Current state-of-the-art systems therefore have to be assessed not only in terms of naturalness and task performance, but also in terms of fairness, privacy, safety, and the social consequences of simulated empathy and authority. Research has identified a *virtual backlash* effect, in which dominant female agents are liked less than

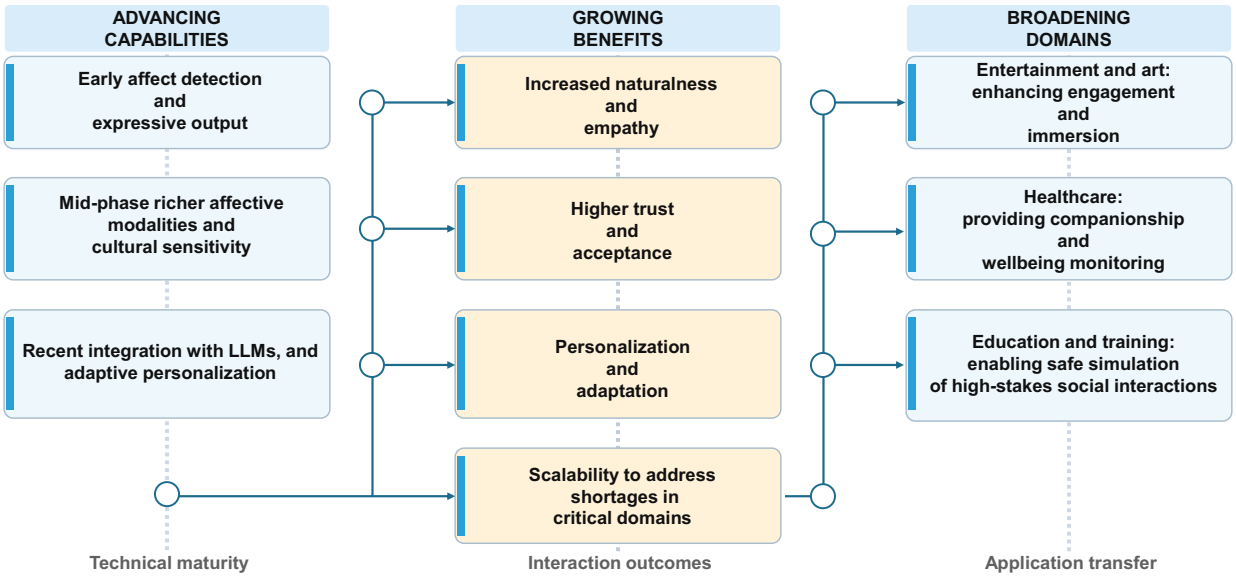

Fig. 11 Conceptual overview of the evolution of socially and emotionally competent virtual agents. Advances in multimodal sensing, generative behavior, and affective intelligence have led to increasing benefits–such as naturalness, empathy, and trust–and to a broadening of applications across domains including entertainment, healthcare, and education

dominant male agents, mirroring real-world gender penalties [103]. Gender stereotypes are prevalent in agent design [10, 70], and embodiment can activate these biases, negatively impacting negotiations for women [44]. Furthermore, stereotypes can influence how users perceive an agent's emotional expressions across different perceived races and genders [34]. These findings underscore that socially expressive behavior is not value-neutral: multimodal expressivity interacts with users' expectations, norms, and prior beliefs.

Generative AI is furthermore prone to *hallucinations*, i.e., producing outcomes that are correct in form but incorrect, implausible, or unfitting in substance. In critical areas of human–agent interaction, where the agent is given some form of implicit authority, such as healthcare, mental-health support, or education, this is not merely a technical issue but a consequential safety and responsibility problem [94]. Such domains may also require empathy and a deeper understanding of the user's socio-emotional needs, which may be simulated by generative approaches but cannot be assumed. Agents may respond with generic validation, overconfident reassurance, or a mismatched emotional tone. Such responses can be experienced as shallow, patronizing, or harmful in situations that require careful support. Against these backdrops, contemporary technical approaches must be designed with awareness of social dynamics, application context, user vulnerability, and potential risks. This requires transparency about system limitations, careful domain restrictions, human oversight in high-stakes settings, and evaluation protocols that assess not only naturalness and usability but also bias, safety, trust calibration, and possible harm.

Revisiting the evolution of agent capabilities helps contextualize both recent breakthroughs and their associated risks. With these challenges in mind, this article has traced the progression of virtual agents through the ages, highlighting advances in capabilities, the diversification of application domains, and the expanding practical and societal relevance of socially and emotionally intelligent virtual agents (see Fig. 11).

Advancing capabilities Early agent systems focused on generating basic conversational behaviors, often using symbolic planners, handcrafted gesture rules, or simple mappings between prosody and motion. These systems demonstrated that even minimal cues could evoke personality impressions, but they were limited in variability, coordination, and context awareness. The following decades introduced psychologically grounded models of affect and personality, enabling agents to express consistent internal states and more coherent multimodal behavior. Frameworks such as BML provided the first unified pipelines for scheduling gaze, speech, gesture, and facial expressions along a shared timeline.

Recent advances, driven by deep learning and large multimodal models, have dramatically expanded the expressive bandwidth of virtual agents. Neural TTS offers fine-grained control over speaking style and affect and speech-driven gesture and motion synthesis achieve rhythmic and semantic alignment. At the same time, increasingly sophisticated models of social cues allow agents not merely to display affect but to adapt their behavior dynamically in response to user's social signals. As a result, virtual agents have evolved from script-following characters to systems capable of

maintaining rapport, expressing personality, and responding with convincing social competence.

Broadening of domains While early virtual agents were primarily developed for entertainment, storytelling, and experimental HCI installations, their improved social and affective capabilities have enabled deployment in a growing number of application areas. In entertainment and media, agents contribute to narrative immersion and character-driven experiences by combining expressive animation, speech, and behavior. In healthcare, socially competent agents support mental health counselling, cognitive assessment, coaching, and adherence to treatment plans, functioning as both companions and assistants. In education and training, agents simulate complex interpersonal scenarios, such as interviews, negotiations, cultural encounters, or classroom dynamics, offering safe, repeatable environments for practicing social, emotional, or communicative skills. With the rise of Extended Reality, virtual agents can now cohabit the user's environment, blending physical and digital contexts and enabling novel forms of embodied interaction.

Growing benefits Across these domains, the maturation of social and affective capabilities has led to measurable improvements in naturalness, engagement and user trust. Users increasingly perceive agents not merely as interfaces but as credible social actors capable of empathy, alignment, and mutual understanding. Generative models enable personalization of voice, gesture style, conversational tone, or affective stance. This supports long-term interaction and emotional attunement. Moreover, scalable virtual agents offer promising avenues for addressing staffing shortages in areas such as healthcare, therapy, and professional training.

In sum, virtual agents have progressed far beyond the early question of *how to animate* toward the deeper challenge of designing interaction behaviors that appear appropriate, responsive, and trustworthy. Advances in perception, multimodal behavior generation, and adaptive interaction modeling have enabled agents to move from scripted displays toward increasingly context-sensitive forms of interaction. However, it is important to distinguish between the convincing simulation of socially appropriate behavior and genuine comprehension of social meaning. Contemporary data-driven systems, particularly those based on large neural models, primarily learn patterns from interaction data rather than explicitly reasoning about intentions, norms, or social consequences. While these approaches can produce plausible and emotionally convincing behaviors, their underlying decision processes often remain opaque.

This distinction between make-believe and actual reasoning is particularly relevant as virtual agents are increasingly deployed in socially sensitive domains such as healthcare, education, and counseling. An agent's ability to appear empathic or socially aware does not necessarily imply actual comprehension, robust reasoning, or emotional understanding. At its core, virtual-agent research faces a dual challenge: improving the technical capabilities of virtual agents while developing computational models that offer greater explainability and more explicit representations of social and mental processes [27]. Virtual-agent research thus remains both an engineering effort to make interaction systems more effective and a scientific effort to better understand the mechanisms underlying human social behavior and communication.

Several future research directions follow from this review. First, virtual agents require stronger grounding in interaction context, task constraints, and user state, rather than relying primarily on pattern completion from large-scale data. Second, multimodal models need more transparent mechanisms for representing affect, intention, uncertainty, and safety-relevant assumptions. Third, evaluation should move beyond isolated naturalness ratings toward longitudinal, interactive, and application-specific studies that measure trust calibration, user well-being, bias, robustness, and potential harm. Fourth, real-time systems must balance expressive richness with latency, controllability, privacy protection, and human oversight. Finally, the field should more explicitly connect data-driven engineering with cognitive and social theory, so that virtual agents become not only more convincing interfaces but also better instruments for studying human communication.

Funding Open Access funding enabled and organized by Projekt DEAL. The work conducted at the University of Augsburg was partially funded by the BMFTR: REGINA, Award Number: 16SV9492 and CONFIDENCE, Award Number: 16SV9383.

References

1. Ahmmed A, Butts E, Naeiji K, Thiamwong L, Ancis J, Daher S (2024) Aware intelligent virtual agent's effect on social presence and empathy of caregivers. In: Proceedings of the 24th ACM international conference on intelligent virtual agents, IVA '24. ACM, New York, NY, USA

2. Alexanderson S, Henter GE, Kucherenko T, Beskow J (2020) Style-controllable speech-driven gesture synthesis using normalising flows. Comput Graph Forum 39(2):487–496
3. Alimardani M, de Roode R, Vaitonyte J, Louwerse MM (2024) Effect of a virtual agent's appearance and voice on uncanny valley and trust in human-agent collaboration. In: Proceedings of the 24th ACM international conference on intelligent virtual agents, IVA '24, New York, NY, USA. ACM
4. André E, Rist T, Müller J (1998) Integrating Reactive and Scripted Behaviors in a Life-Like Presentation Agent. In: Sycara KP, Wooldridge, MJ (eds) Proceedings of the second international conference on autonomous agents, AGENTS 1998, St. Paul, Minnepolis, USA, May 9-13, 1998, pp 261–268. ACM
5. André E, Rist T, van Mulken S, Klesen M, Baldes S (2000) The automated design of believable dialogue. MIT Press, Cambridge, MA, USA, pp 220–255
6. Andrist S, Gleicher M, Mutlu B (2017) Looking coordinated: bidirectional gaze mechanisms for collaborative interaction with virtual characters. In: Proceedings of the 2017 CHI conference on human factors in computing systems, CHI '17, New York, NY, USA. ACM, pp 2571–2582
7. Andrist S, Pejsa T, Mutlu B, Gleicher M (2012) Designing Effective Gaze Mechanisms for Virtual Agents. In: Konstan JA, Chi EdH, Höök K (eds) CHI conference on human factors in computing systems, CHI '12, Austin, TX, USA - May 05 - 10, 2012. ACM, pp 705–714
8. Andrist S, Tan XZ, Gleicher M, Mutlu B (2014) Conversational Gaze Aversion for Humanlike Robots. In: Proceedings of the 2014 ACM/IEEE international conference on human-robot interaction, HRI '14. ACM, New York, NY, USA, pp 25–32
9. Ao T, Zhang Z, Liu L (2023) GestureDiffuCLIP: gesture diffusion model with CLIP latents. ACM Trans Graph 42(4):1–18
10. Araujo V, Schaffer D, Costa AB, Musse SR, (2022) Towards virtual humans without gender stereotyped visual features. In: SIGGRAPH Asia, (2022) Technical Communications, SA '22, New York, NY. ACM, USA
11. Bailly G, André E, Cooper E, Klabbers E, Cowan B, Edlund J, Harte N, King S, Maguer L, Moore S, Roger K (2025) Hot topics in speech synthesis evaluation. In: Proc 13th ISCA speech synthesis workshop (SSW 2025). pp 1–7
12. Beck AT, Steer RA, Brown GK (1996) Manual for the beck depression inventory-II, 2nd edn. Pearson, San Antonio
13. Becker-Asano C, Wachsmuth I (2010) Affective computing with primary and secondary emotions in a virtual human. Auton Agent Multi-Agent Syst 20(1):32–49
14. Bhattacharya U, Childs E, Rewkowski N, Manocha D (2021) Speech2AffectiveGestures: synthesizing co-speech gestures with generative adversarial affective expression learning. In: Proceedings of the 29th ACM international conference on multimedia, MM '21. ACM, New York, NY, USA, pp 2027–2036
15. Bönsch A, Radke S, Ehret J, Habel U, Kuhlen TW (2020) The impact of a virtual agent's non-verbal emotional expression on a user's personal space preferences. In: Marsella S, Jack R, Vilhjálmsson HH, Sequeira P, Cross ES (eds) IVA '20: ACM international conference on intelligent virtual agents, virtual event, Scotland, UK, October 20–22, 2020. ACM, pp 12:1–12:8
16. Boukaram H-A, Ziadee M, Sakr MF (2021) Mitigating the effects of delayed virtual agent response time using conversational fillers. In: Proceedings of the 9th international conference on human-agent interaction, HAI '21. ACM, New York, NY, USA, pp 130–138
17. Casanova E, Weber J, Shulby CD, Júnior AC, Gölge E, Ponti MA (2022) Yourtts: towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In: Chaudhuri K, Jegelka S, Song L, Szepesvári C, Niu G, Sabato S (eds) international conference on machine learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, volume 162 of Proceedings of Machine Learning Research. PMLR, pp 2709–2720
18. Cassell J, Vilhjálmsson HH, Bickmore TW (2004) BEAT: the behavior expression animation toolkit. In: Prendinger H, Ishizuka M (eds) Life-like characters - tools, affective functions, and applications, Cognitive Technologies. Springer, pp 163–186
19. Cerekovic A, Aran O, Gatica-Perez D (2017) Rapport with virtual agents: What do human social cues and personality explain? IEEE Trans Affect Comput 8(3):382–395
20. Chehayeb L, Bhuvaneshwara C, Anglet M, Hilpert B, Meyer A-K, Tsovaltzi D, Gebhard P, Biermann A, Auchtor S, Lauinger N, Knopf J, Kaiser A, Kersting F, Mehlmann G, Lingenfelser F, André E (2024) MITHOS: interactive mixed reality training to support professional socio-emotional interactions at schools. arXiv:2409.12968
21. Chen B, Li Y, Zheng Y, Ding Y-X, Zhou K (2025) Motion-example-controlled co-speech gesture generation leveraging large language models. In: Proceedings of the special interest group on computer graphics and interactive techniques conference conference papers, SIGGRAPH conference papers '25. ACM, New York, NY, USA,
22. Chen C, Zhang J, Lakshmikanth SK, Fang Y, Shao R, Wetzstein G, Fei-Fei L, Adeli E (2024) The language of motion: unifying verbal and non-verbal language of 3D human motion
23. Chen S, Liu S, Zhou L, Liu Y, Tan X, Li J, Zhao S, Qian Y, Wei F (2024) VALL-E 2: neural codec language models are human parity zero-shot text to speech synthesizers. arXiv:2406.05370
24. Cheng Q, Li X, Fu X (2024) SIGGesture: generalized co-speech gesture synthesis via semantic injection with large-scale pre-training diffusion models. SIGGRAPH Asia, (2024) Conference Papers, SA '24. ACM, New York, NY, USA
25. Chi D, Costa M, Zhao L, Badler N (2000) The emote model for effort and shape. In: Proceedings of the 27th annual conference on computer graphics and interactive techniques, SIGGRAPH '00. ACM Press/Addison-Wesley Publishing Co, USA, pp 173–182
26. Chopra-Khullar S, Badler NI (1999) Where to look? automating attending behaviors of virtual human characters. In: Proceedings of the third annual conference on autonomous agents, AGENTS '99. ACM, New York, NY, USA, pp 16–23
27. Cohen PR, Galescu L, Shvo M (2025) A framework for building explainable collaborative multimodal dialogue systems using a theory of mind. Auton Agents Multi Agent Syst 39(2):44
28. Kenneth Mark Colby (1975) Artificial paranoia: a computer simulation of paranoid processes, vol 49. Pergamon Press, Oxford, Pergamon General Psychology Series
29. Damian I, Baur T, Lugrin B, Gebhard P, Mehlmann G, André E (2015) Games are better than books: In-situ comparison of an interactive job interview game with conventional training. In: Conati C, Heffernan NT, Mitrovic A, Verdejo MF (eds) Artificial intelligence in education - 17th international conference, AIED 2015, Madrid, Spain, June 22-26, 2015. Proceedings, volume 9112 of Lecture Notes in Computer Science, pp 84–94. Springer
30. Dao T, Gu A (2024) Transformers are SSMs: generalized models and efficient algorithms through structured state space duality. arXiv:2405.21060
31. Dias J, Mascarenhas S, Paiva A (2014) FAtiMA modular: towards an agent architecture with a generic appraisal framework. Towards pragmatic computational models of affective processes, In: Emotion Modeling, pp 44–56
32. Dietz M (2025) Assistive augmentation of cognitive processes using mobile signal processing. doctoralthesis. University of Augsburg
33. Don DSW, Kiderle T, Mertes S, Schiller D, Ritschel H, André E (2025) Melax: conversations with generative AI in socially interactive agents. In: Li T, Paternò F, Väänänen K, Leiva L, Spano LD, Verbert K (eds) Companion proceedings of the 30th

international conference on intelligent user interfaces, IUI 2025, Cagliari, Italy, March 24-27, 2025. ACM, pp 163–166
34. Durupinar F, Kim J (2022) Facial emotion recognition of virtual humans with different genders, races, and ages. In: ACM symposium on applied perception, (2022) SAP '22. ACM, New York, NY, USA
35. Etienne E, Ristorcelli M, Saufnay S, Quilez A, Casanova R, Schyns M, Ochs M (2024) A systematic review on the socio-affective perception of ivas' multi-modal behaviour. In: Proceedings of the 24th ACM international conference on intelligent virtual agents, IVA '24. ACM, New York, NY, USA,
36. Ferstl Y, Neff M, McDonnell R (2019) Multi-objective adversarial gesture generation. In Proceedings of the 12th ACM SIGGRAPH conference on motion, interaction and games, MIG '19, New York, NY, USA. ACM
37. Ferstl Y, Neff M, McDonnell R (2020) Understanding the predictability of gesture parameters from speech and their perceptual importance. In Proceedings of the 20th ACM international conference on intelligent virtual agents, IVA '20, New York, NY, USA. ACM
38. Fu C, Wang Y, Zhang J, Jiang Z, Mao X, Wu J, Cao W, Wang C, Ge Y, Liu Y (2024) MambaGesture: enhancing co-speech gesture generation with mamba and disentangled multi-modality fusion. In: Proceedings of the 32nd ACM international conference on multimedia, MM '24, pp 10794–10803, New York, NY, USA, ACM
39. Garau M, Slater M, Vinayagamoorthy V, Brogni A, Steed A, Sasse MA (2003) The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. In: Proceedings of the SIGCHI conference on human factors in computing systems, CHI '03. ACM, New York, NY, USA, pp 529–536
40. Gebhard P (2005) ALMA: a layered model of affect. In: Dignum F, Dignum V, Koenig S, Kraus S, Singh MP, Wooldridge MJ (eds) 4th international joint conference on autonomous agents and multiagent systems (AAMAS 2005), July 25–29, 2005, Utrecht, The Netherlands. ACM, pp 29–36
41. Gebhard P, Schneeberger T, Mehlmann G, Baur T, André E (2019) Designing the impression of social agents' real-time interruption handling. In: Proceedings of the 19th ACM international conference on intelligent virtual agents, IVA '19. ACM, New York, NY, USA, pp 19–21
42. Gu A, Dao T (2023) Mamba: linear-time sequence modeling with selective state spaces. arXiv: 2312.00752
43. Gu A, Goel K, Ré C (2022) Efficiently modeling long sequences with structured state spaces. In: International conference on learning representations
44. Hale J, Schweitzer L, Gratch J (2024) Pitfalls of embodiment in human-agent experiment design. In: Proceedings of the 24th ACM international conference on intelligent virtual agents, IVA '24. ACM, New York, NY, USA
45. Hoegen R, van der Schalk J, Lucas G, Gratch J (2018) The impact of agent facial mimicry on social behavior in a prisoner's dilemma. In: Proceedings of the 18th international conference on intelligent virtual agents, IVA '18, New York, NY, USA. ACM, pp 275–280
46. Holden D, Saito J, Komura T (2016) A deep learning framework for character motion synthesis and editing. ACM Trans Graph 35(4):138:1-138:11
47. Huang A, Knierim P, Chiossi F, Lewis L, Chuang R, Welsch (2022) Proxemics for human-agent interaction in augmented reality. In: Proceedings of the 2022 CHI conference on human factors in computing systems, CHI '22. ACM, New York, NY, USA
48. Janowski K, André E (2019) What if i speak now? a decision-theoretic approach to personality-based turn-taking. In: Proceedings of the 18th international conference on autonomous agents and multiagent systems, AAMAS '19. Richland, SC, International Foundation for Autonomous Agents and Multiagent Systems, pp 1051–1059
49. Janowski K, Ritschel H, André E (2022) Adaptive artificial personalities. In: Lugrin B, Pelachaud C, Traum D (eds) The Handbook on socially interactive agents:20 years of research on embodied conversational agents, intelligent virtual agents, and social robotics Volume 2: Interactivity, Platforms, Application, volume 48 of ACM Books. ACM / Morgan & Claypool, pp 155–194
50. Jia Y, Zhang Y, Weiss RJ, Wang Q, Shen J, Ren F, Chen Z, Nguyen P, Pang R, López-Moreno I, Wu Y (2018) Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In: Bengio S, Wallach HM, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems 31: annual conference on neural information processing systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, pp 4485–4495
51. Jin A, Deng Q, Zhang Y, Deng Z (2019) A deep learning-based model for head and eye motion generation in three-party conversations. Proc ACM Comput Graph Interact Tech 2(2):9:1-9:19
52. Kato K, Inoue K, Lala D, Ochi K, Kawahara T (2025) Real-time generation of various types of nodding for avatar attentive listening system. In: Proceedings of the 27th international conference on multimodal interaction, ICMI '25. ACM, New York, NY, USA, pp 209–217
53. Kopp S, Tepper P, Cassell J (2004) Towards integrated microplanning of language and iconic gesture for multimodal output. In: Sharma R, Darrell T, Harper MP, Lazzari G, Turk M (eds) Proceedings of the 6th international conference on multimodal interfaces, ICMI 2004, State College, PA, USA, October 13-15, 2004, ACM, pp 97–104
54. Kucherenko T, Jonell P, van Waveren S, Henter GE, Alexandersson S, Leite I, Kjellström H (2020) Gesticulator: a framework for semantically-aware speech-driven gesture generation. In: Proceedings of the 2020 international conference on multimodal interaction, ICMI '20, New York, NY, USA. ACM, pp 242–250
55. Kucherenko T, Wolfert P, Yoon Y, Viegas C, Nikolov T, Tsakov M, Henter GE (2024) Evaluating gesture generation in a large-scale open challenge: the GENEA challenge 2022. ACM Trans Graph 43(3):32:1-32:28
56. Li J, Galley M, Brockett C, Gao J, Dolan B (2016) A diversity-promoting objective function for neural conversation models. In: Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: human language technologies. Association for Computational Linguistics, San Diego, California, pp 110–119
57. Li J, Cao C, Schwartz G, Khirodkar R, Richardt C, Simon T, Sheikh Y, Saito S, (2024) URAvatar: Universal Relightable Gaussian Codec Avatars. In: SIGGRAPH Asia, (2024) conference papers, SA '24, New York, NY. ACM, USA
58. Lin J, Cronjé J, Käthner I, Pauli P, Latoschik ME (2023) The virtual maze: a tool for measuring trust towards virtual humans. In: Proceedings of the 23rd ACM international conference on intelligent virtual agents, IVA '23. ACM, New York, NY, USA
59. Lin Y, Zheng Y, Zeng M, Shi W (2025) Predicting turn-taking and backchannel in human-machine conversations using linguistic, acoustic, and visual signals. In: Che W, Nabende J, Shutova E, Pilehvar MT (eds) Proceedings of the 63rd annual meeting of the ACL (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025. ACL, pp 15310–15322
60. Liu H, Zhu Z, Becherini G, Peng Y, Su M, Zhou Y, Zhe X, Iwamoto N, Zheng B, Black MJ (2024) EMAGE: towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling

61. Liu J, Dai W, Wang C, Cheng Y, Tang Y, Tong X (2023) Plan. Towards open-world text-to-motion generation, posture and go
62. Lugrin B, Pelachaud C, Traum D (2022) editors. The handbook on socially interactive agents:20 years of research on embodied conversational agents, intelligent virtual agents, and social robotics volume 2: interactivity, platforms, application, volume 48 of ACM Books. ACM / Morgan & Claypool
63. Lugrin B, Pelachaud C, Traum DR (2021) editors. The handbook on socially interactive agents: 20 years of research on embodied conversational agents, intelligent virtual agents, and social robotics volume 1: methods, behavior, cognition, volume 37 of ACM Books. ACM / Morgan & Claypool
64. Madani N, Srihari RK (2025) Esc-judge: A framework for comparing emotional support conversational agents. arXiv:2505.12531
65. Mairesse F, Marilyn A (2007) Walker. PERSONAGE: personality generation for dialogue. In: Carroll J, van den Bosch A, Zaenen A (eds) ACL 2007, Proceedings of the 45th annual meeting of the ACL, June 23-30, 2007, Prague, Czech Republic. ACL
66. Marcoux A, Tessier M-H, Jackson PL (2023) Nonverbal markers of empathy in virtual healthcare professionals. In: Proceedings of the 23rd ACM international conference on intelligent virtual agents, IVA '23. ACM, New York, NY, USA
67. Marsella SC, Gratch J (2009) EMA: a process model of appraisal dynamics. Cogn Syst Res 10(1):70–90
68. Maxim A, Zalake M, Lok B (2023) The impact of virtual human vocal personality on establishing rapport: A study on promoting mental wellness through extroversion and vocalics. In: Proceedings of the 23rd ACM international conference on intelligent virtual agents, IVA '23. ACM, New York, NY, USA
69. McCrae RR, John OP (1992) An introduction to the five-factor model and its applications. J Pers 60(2):175
70. Nag P, Nilay Yalçın Ö (2020) Gender stereotypes in virtual agents. In: Proceedings of the 20th ACM international conference on intelligent virtual agents, IVA '20. ACM, New York, NY, USA
71. Nagy R, Voss H, Hoang-Minh T, Tsakov M, Nikolov T, Zhang Z, Ao T, Yang S, Huang S, Cheng Y, Mughal MH, Dabral R, Chhatre K, Theobalt C, Liu L, Kopp S, McDonnell R, Neff M, Kucherenko T, Yoon Y, Henter GE (2025) Gesture Generation (Still) Needs improved human evaluation practices: insights from a community-driven state-of-the-art benchmark. arXiv:abs/2511.01233
72. Obremski D, Friedrich P, Wienrich C (2025) Please let me think: the influence of conversational fillers on transparency and perception of waiting time when interacting with a conversational AI in virtual reality. In: Proceedings of the 27th international conference on multimodal interaction, ICMI '25, New York, NY, USA. ACM, pp 496–505
73. Ochs M, Bousquet J, Blache P (2019) Multimodal cues of the sense of presence and co-presence in human-virtual agent interaction. In: Proceedings of the 19th ACM international conference on intelligent virtual agents, IVA '19, New York, NY, USA. ACM pp 218–220
74. Ortony A, Clore GL, Collins A (1988) The cognitive structure of emotions. Cambridge University Press, Cambridge
75. Pereira A, Prada R, Paiva A (2014) Improving social presence in human-agent interaction. In: Proceedings of the SIGCHI conference on human factors in computing systems, CHI '14, New York, NY, USA. ACM pp 1449–1458
76. Poletaykina A, Kruse L, Steinicke F (2025) Knowing is trusting! the influence of familiarity on the perception of intelligent virtual agents. In: Proceedings of the 25th ACM international conference on intelligent virtual agents, IVA '25. ACM, New York, NY, USA
77. Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training. Technical report, OpenAI
78. Ren Y, Hu C, Tan X, Qin T, Zhao S, Zhao Z, Liu T-Y (2021). Fastspeech 2: Fast and high-quality end-to-end text to speech. In: 9th international conference on learning representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net
79. Serban IV, Sordoni A, Bengio Y, Courville AC, Pineau J (2016) Building end-to-end dialogue systems using generative hierarchical neural network models. In: Proceedings of the thirtieth AAAI conference on artificial intelligence, AAAI'16. AAAI Press, pp 3776–3784
80. Shvo M, Buhmann J, Kapadia M (2019) An interdependent model of personality, motivation, emotion, and mood for intelligent virtual agents. In: Proceedings of the 19th ACM international conference on intelligent virtual agents, IVA '19, New York, NY, USA. ACM pp 65–72
81. Song S, Spitale M, Luo C, Palmero C, Barquero G, Zhu H, Escalera S, Valstar MF, Baur T, Ringeval F, André E, Gunes H (2024) REACT 2024: the second multiple appropriate facial reaction generation challenge. In: 18th IEEE international conference on automatic face and gesture recognition, FG 2024, Istanbul, Turkey, May 27-31, 2024. IEEE, pp 1–5
82. Stanton D, Wang Y, Skerry-Ryan RJ (2018) Predicting Expressive Speaking Style from Text in End-To-End Speech Synthesis. In: 2018 IEEE spoken language technology workshop, SLT 2018, Athens, Greece, December 18-21, 2018. IEEE, pp 595–602
83. Stavemann HH (2015) Sokratische Gesprächsführung in Therapie und Beratung: Eine Anleitung für Psychotherapeuten, 3rd edn. Berater und Seelsorger, Beltz, Weinheim and Basel
84. Sundar SS, Kim J (2019) Machine heuristic: When we trust computers more than humans with our personal information. In: Proceedings of the 2019 CHI conference on human factors in computing systems, CHI '19, New York, NY, USA, 2019. ACM, pp 1–9
85. Tang H, Wang W, Xu D, Yan Y, Sebe N (2018) Gesturegan for hand gesture-to-gesture translation in the wild. In: Boll S, Lee KM, Luo J, Zhu W, Byun H, Chen CW, Lienhart R, Mei T (eds) 2018 ACM multimedia conference on multimedia conference, MM 2018, Seoul, October 22–26, 2018. ACM, pp 774–782
86. Terada K, Okazoe M, Gratch J (2021) Effect of politeness strategies in dialogue on negotiation outcomes. In: Proceedings of the 21st ACM international conference on intelligent virtual agents, IVA '21, pp 195–202New York, NY, USA. ACM
87. Thiebaux M, Marsella S, Marshall AN, Kallmann M (2008) Smartbody: behavior realization for embodied conversational agents. In: Proceedings of the 7th international joint conference on autonomous agents and multiagent systems - Volume 1, AAMAS '08, Richland, SC. International foundation for autonomous agents and multiagent systems, pp 151–158
88. Triantafyllopoulos A, Björn W, Schuller, Iymen G, Sezgin TM, He X, Yang Z, Tzirakis P, Liu S, Mertes S, André E, Fu R, Tao J (2023) An overview of affective speech synthesis and conversion in the deep learning era. In: Proc IEEE 111(10):1355–1381
89. van Rijn P, Mertes S, Schiller D, Dura P, Siuzdak H, Harrison PMC, André E, Jacoby N (2022) Voiceme: Personalized voice generation in TTS. arXiv:2203.15379
90. van Rijn P, Mertes S, Schiller D, Harrison PMC, Larrouy-Maestri P, André E, Jacoby N (2021) Exploring emotional prototypes in a high dimensional TTS latent space (2021) In: Hermansky H. In: Cernocký H, Burget L, Lamel L, Scharenborg O, Motlícek P (eds) 22nd annual conference of the ISCA, Interspeech 2021, Brno, Czechia, August 30 - September 3. ISCA, pp 3870–3874
91. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, Garnett R (eds) Advances in neural information processing systems 30: annual conference on neural information

processing systems 2017, December 4-9, 2017, Long Beach, CA, USA, pp 5998–6008
92. Vilhjálmsson H, Cantelmo N, Cassell J, Chafai NE, Kipp M, Kopp S, Mancini M, Marsella S, Marshall AN, Pelachaud C, Ruttkay Z, Thórisson KR, van Welbergen H, van der Werf RJ (2007) The behavior markup language: recent developments and challenges. In: Pelachaud C, Martin J-C, André E, Chollet G, Karpouzis K, Pelé D (eds) Intelligent Virtual Agents, Berlin, Heidelberg. Springer Berlin Heidelberg, pp 99–111
93. Vinyals O, Le QV (2015) A neural conversational model. arXiv:1506.05869
94. Volpato R, DeBruine L, Stumpf S (2025) Trusting emotional support from generative artificial intelligence: a conceptual review. Comput Hum Behav Artif Hum 5:100195
95. Voß H, Kopp S (2023) AQ-GT: a Temporally aligned and quantized GRU-transformer for co-speech gesture synthesis. In: Proceedings of the 25th international conference on multimodal interaction, ICMI '23, New York, NY, USA. ACM, pp 60–69
96. Wang C, Chen S, Wu Y, Zhang Z, Zhou L, Liu S, Chen Z, Liu Y, Wang H, Li J, He L, Zhao S, Wei F (2023) Neural codec language models are zero-shot text to speech synthesizers. arXiv:2301.02111
97. Wang I, Calvo R, Wang H, Ruiz J (2023) Stop copying me: evaluating nonverbal mimicry in embodied motivational agents. In: Proceedings of the 23rd ACM international conference on intelligent virtual agents, IVA '23, New York, NY, USA. ACM
98. Wang N, Gratch J (2010) Don't just stare at me! In: Proceedings of the SIGCHI conference on human factors in computing systems, CHI '10, New York, NY, USA. ACM, pp 1241–1250
99. Wang Y, Skerry-Ryan RJ, Stanton D, Wu Y, Weiss RJ, Jaitly N, Yang Z, Xiao Y, Chen Z, Bengio S, Le QV, Agiomyrgiannakis Y, Clark R, Rif A (2017) Saurous. Tacotron: Towards end-to-end speech synthesis. In: Lacerda F (ed) 18th annual conference of the ISCA, Interspeech 2017, Stockholm, Sweden, August 20-24, 2017. ISCA, pp 4006–4010
100. Wang Y, Stanton D, Zhang Y, Skerry-Ryan RJ, Battenberg E, Shor J, Xiao Y, Jia Y, Ren F, Rif A (2018) Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In: Dy JG, Krause A (eds) Proceedings of the 35th international conference on machine learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, volume 80 of proceedings of machine learning research. PMLR, pp 5167–5176
101. Wanner L, André E, Blat J, Dasiopoulou S, Domínguez M, Llorach G, Mille S, Sukno F, Kamateri E, Vrochidis S, Kompatsiaris I, Lingenfelser F, Mehlmann G, Stam A, Stellingwerff L, Vieru B, Lamel L, Minker W, Pragst L, Ultes S (2016) Towards a multimedia knowledge-based agent with social competence and human interaction capabilities. In: Vrochidis S, Wanner L, André E, Elzer Schwartz S (eds) Proceedings of the 1st international workshop on multimedia analysis and retrieval for multimodal interaction (MARMI '16), New York, NY, USA. ACM Press, pp 21–26
102. Weizenbaum J (1966) ELIZA-A computer program for the study of natural language communication between man and machine. Commun ACM 9(1):36–45
103. Wessler J, Schneeberger T, Christidis L, Gebhard P (2022) Virtual backlash: nonverbal expression of dominance leads to less liking of dominant female versus male agents. In: Proceedings of the 22nd ACM international conference on intelligent virtual agents, IVA '22. ACM, New York, NY, USA
104. Windle J, Matthews I, Taylor S (2024) Llanimation: Llama driven gesture animation. In: Proceedings of the ACM SIGGRAPH/Eurographics symposium on computer animation, SCA '24, Goslar, DEU. Eurographics Association, pp 1–10
105. Withanage D, Lingenfelser F, Kuch JM, Grothe O, Schlagowski R, André E, Mertes S (2025) VoiceX as a design tool for virtual agents' voices. In: Gebhard P, Schneeberger T, Biancardi B, Sabouret N, Schmitz M, Yumak Z, (eds) Proceedings of the 25th ACM international conference on intelligent virtual agents, IVA 2025, Berlin, Germany, September 16-19, 2025. ACM, pp 35:1–35:4
106. Withanage Don DS, Müller P, Nunnari F, André E, Gebhard P (2023) ReNeLiB: Real-time neural listening behavior generation for socially interactive agents. In: Proceedings of the 25th international conference on multimodal interaction, ICMI '23, New York, NY, USA. ACM, pp 507–516
107. Yoon Y, Cha B, Lee J-H, Jang M, Lee J, Kim J, Lee G (2020) Speech gesture generation from the trimodal context of text, audio, and speaker identity. ACM Trans Graph 39(6):222:1-222:16
108. You C, Ghosh R, Maxim A, Stuart J, Cooks E, Lok B (2022) How does a virtual human earn your trust? Guidelines to improve willingness to self-disclose to intelligent virtual agents. In: Proceedings of the 22nd ACM international conference on intelligent virtual agents, IVA '22. ACM, New York, NY, USA
109. Zhang S, Dinan E, Urbanek J, Szlam A, Kiela D, Weston J (2018) Personalizing dialogue agents: I have a dog, do you have pets too? In: Gurevych I, Miyao Y (eds) Proceedings of the 56th annual meeting of the ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers. ACL, pp 2204–2213
110. Zhang Y, Sun S, Galley M, Chen Y-C, Brockett C, Gao X, Gao J, Liu J, Dolan B (2020) DIALOGPT: Large-scale generative pretraining for conversational response generation. In: Celikyilmaz A, Wen T-H (eds) Proceedings of the 58th annual meeting of the ACL: system demonstrations, Online, July 5-10, 2020. ACL, pp 270–278
111. Zheng Z, Liao L, Deng Y, Nie L (2023) Building emotional support chatbots in the era of LLMs. arXiv:2308.11584