Scientific Research Publishing

Open Journal of Social Sciences, 2026, 14(3), 268-284
https://www.scirp.org/journal/jss
ISSN Online: 2327-5960
ISSN Print: 2327-5952

A Systematic Review of Multimodal AI Agents and Embodied Avatars in University EFL Speaking Support

Qixin Zhu

Department of College English, Zhejiang Yuexiu University, Shaoxing, China
Email: qixin.zhu.24@gmail.com

How to cite this paper: Zhu, Q. X. (2026). A Systematic Review of Multimodal AI Agents and Embodied Avatars in University EFL Speaking Support. *Open Journal of Social Sciences, 14,* 268-284.
https://doi.org/10.4236/jss.2026.143016

Received: February 26, 2026
Accepted: March 14, 2026
Published: March 17, 2026

Abstract

The integration of Artificial Intelligence into EFL education has historically prioritised text-based accuracy. However, the period between 2024 and 2026 witnessed a multimodal turn characterised by high-fidelity embodied avatars. This systematic review examines the empirical landscape of Multimodal AI Agents (MMAAs) specifically for university-level EFL speaking support. By synthesising data from 82 empirical studies (N = 82) indexed in Web of Science and Scopus, this report identifies critical themes: the paradox of social presence in anxiety regulation, the instructional implications of AI Twins, the cognitive load of the uncanny valley, interactional bottlenecks in turn-taking, the pedagogical shift from tutor to peer, and the ethical risks of biometric surveillance. Findings suggest that while embodied agents significantly enhance Willingness to Communicate (WTC) and emotional engagement, their efficacy depends more on artificial coherence than photorealism. Furthermore, the emergence of AI Peers challenges traditional hierarchies, suggesting a future in which AI functions as a fallible, tireless social partner rather than an omniscient authority. The review concludes that the primary value of MMAAs lies in decoupling social practice from social risk, though this potential is contingent upon balancing affective benefits with cognitive constraints and rigorous ethical safeguards. The review concludes that the primary value of MMAAs lies in decoupling social practice from social risk, provided that educators actively mitigate the interview effect and prioritise artificial coherence over photorealism.

Keywords

Multimodal AI Agents, University EFL Speaking, Embodied Avatars, Social Presence, Cognitive Load

1. Introduction

For decades, the chatbot represented the pinnacle of automated language practice—a computational entity capable of processing text input and generating text output (Wang et al., 2025). While revolutionary in its time, the text-based chatbot was fundamentally detached from the embodied realities of human communication. It lacked a face to convey empathy, a voice to modulate tone, and a body to gesture (Huang et al., 2025). Consequently, while it could scaffold writing and reading skills, its application to speaking—a dynamic, multimodal performance—was limited to speech-to-text conversions that stripped oral language of its prosodic and paralinguistic richness.

However, the rapid maturation of Generative AI (GenAI) between 2024 and 2026 has precipitated a seismic shift in this landscape. We have moved from the era of Large Language Models (LLMs) to Large Multimodal Models (LMMs) such as GPT-4o and Gemini 2.0 (Wang et al., 2025). These systems do not merely read text; they perceive audio, video, and images and generate synchronised voice and visual avatars in real time. This technological convergence has given rise to the Multimodal AI Agent (MMAA): an embodied digital interlocutor that can see, hear, and exhibit behaviours mimicking human social presence (Amrevuawho et al., 2025).

The context of university-level EFL is critical to this technological development. Unlike generalist language learners, university students face high-stakes demands for academic oracy, presentation skills, and complex interactional competence. They operate in environments where the affective filter—the psychological barrier raised by anxiety and fear of negative evaluation—can severely impede performance (Lu et al., 2025).

Traditional university instruction often struggles to provide sufficient speaking practice for every student. In a seminar of 30 students, an individual might speak for only minutes per week. Multimodal AI offers a theoretical solution to this bottleneck: a high-fidelity, low-anxiety practice environment that simulates the cognitive and social pressures of real communication without the social consequences of failure (Yin et al., 2025). The promise of the digital human is not merely one of technological novelty but of pedagogical necessity—a tool to bridge the gap between controlled classroom drills and the chaotic reality of human interaction.

This review is grounded in the intersection of Social Presence Theory and Cognitive Load Theory. Social Presence Theory posits that the efficacy of a mediated interaction depends on the degree to which the other is perceived as real and present (Huang et al., 2025). The hypothesis driving the adoption of avatars is that adding a face increases social presence, thereby triggering the social mechanics of language learning. Conversely, Cognitive Load Theory warns that adding modalities risks overwhelming learners' processing capacity if not designed in strict adherence to multimedia principles (Fink et al., 2024).

This report aims to systematically evaluate the state of the art in multimodal AI

for university EFL speaking support. It seeks to answer the following questions:

- Affective Impact: How do embodied avatars influence Foreign Language Speaking Anxiety (FLSA) and Willingness to Communicate (WTC) compared to text-only or voice-only interfaces?
- Cognitive Cost: Does the visual presence of an avatar aid comprehension through paralinguistic cues, or does it impose extraneous cognitive load?
- Interactional Fidelity: To what extent can current AI agents support the development of Interactional Competence (IC), particularly regarding turn-taking and repair?
- Ethical Implications: What are the privacy and surveillance risks associated with the deployment of camera-based affective computing in classrooms?

2. Methodology

To ensure a rigorous synthesis of the rapidly evolving literature, this review adopted a systematic approach aligned with the PRISMA 2020 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, adapted for the velocity of AI research.

2.1. Operational Definitions of Key Constructs

To ensure transparency in the screening process and clarity in the thematic synthesis, the core constructs examined in this review are defined and operationalized as follows:

Multimodal AI Agent (*MMAA*)

An embodied digital interlocutor capable of perceiving inputs (audio, video, images) and generating synchronized voice and visual outputs in real time to mimic human social presence.

Minimum feature for inclusion: The system must interact using both visual (avatar) and auditory (voice) modalities; text-only chatbots or audio-only tools were excluded.

Embodied Avatar

The visual representation of an AI agent that provides a face to convey empathy, a voice to modulate tone, and a body to gesture, thereby increasing perceived social presence.

Minimum feature for inclusion: The agent must possess a visible, human-like face capable of conveying non-verbal cues (e.g., facial expressions, eye contact).

AI Twin (*Self-Clone*)

A personalized deepfake avatar of the learner that utilizes voice cloning to speak with the learner's voice, but with perfect fluency and grammar. It serves to visualize the learner's Ideal L2 Self.

Minimum feature for inclusion: The system must record the learner's speech, repair it using a language model, and re-synthesize the audio using the learner's own voice synced to their specific digital likeness.

Artificial Coherence

The mathematical and perceptual synchronization of voice and lip movements is achieved in fully AI-generated environments, where the system generates both the voice and the face from the same underlying data stream.

Minimum feature for inclusion/synthesis: Evaluated based on the system's ability to minimize perceptual interference (mismatches in lip-sync and prosody) that otherwise imposes extraneous cognitive load.

Interview Effect

An unnatural conversational pattern caused by the rigid pause thresholds (typically 1 - 2 seconds of silence) AI systems require to detect the end of a user's turn.

Minimum feature for inclusion/synthesis: Identified in studies where the interaction is forced into a strict sequential pattern (User speaks -> Pause -> AI speaks -> Pause), leading learners to suppress natural backchanneling and overlapping speech.

2.2. Search Strategy and Data Sources

A comprehensive search was conducted in two primary academic databases, Web of Science (WoS) and Scopus, covering the period from January 1, 2024, to January 31, 2026. This timeframe was selected to capture the impact of the "multimodal boom" following the release of GPT-4V, Gemini, and advanced avatar-synthesis tools (e.g., D-ID, Synthesia, HeyGen).

To ensure reproducibility, the exact search strings executed in the databases were as follows:

- Scopus (Searched in Article Title, Abstract, Keywords):

TITLE-ABS-KEY(("EFL speaking" OR "L2 oral proficiency" OR "English speaking" OR "university students") AND ("avatar" OR "embodied agent" OR "digital human" OR "multimodal AI" OR "virtual human") AND ("higher education" OR "tertiary education"))

- Web of Science (Searched in Topic/TS field):

TS = (("EFL speaking" OR "L2 oral proficiency" OR "English speaking" OR "university students") AND ("avatar" OR "embodied agent" OR "digital human" OR "multimodal AI" OR "virtual human") AND ("higher education" OR "tertiary education"))

Filters Applied: The database searches were filtered to include only documents published in English. Document types were restricted to empirical journal articles and conference proceedings.

Handling of Grey Literature and Preprints: Given the rapid velocity of generative AI research, traditional peer-review timelines often lag behind technological realities. Therefore, technical reports and empirical preprints (e.g., from arXiv) were explicitly included in the screening process. Opinion pieces, non-empirical white papers, and secondary reviews were strictly excluded.

2.3. Inclusion and Exclusion Criteria

Studies were screened based on the following criteria (**Table 1**).

Table 1. Inclusion and exclusion criteria.

Criterion	Inclusion	Exclusion
Study Design	Empirical studies (quantitative, qualitative, mixed-methods)	Opinion pieces, non-peer-reviewed white papers (unless technical reports), reviews
Participants	University/Adult EFL learners	K-12 learners, native speakers
Technology	Must involve multimodal agents (Visual + Audio)	Text-only chatbots, audio-only tools (e.g., podcasts), and non-interactive video
Language	English	Non-English publications

2.4. Data Extraction and Analysis

The initial search yielded 412 records (Figure 1). To ensure reliability in the study selection process, a calibration exercise was initially conducted in which two independent reviewers screened a random sample of 10% of the titles and abstracts. After discussing discrepancies and aligning on the inclusion criteria, the reviewers independently screened the remaining records. Any disagreements during the title/abstract screening and the subsequent full-text eligibility assessment (n = 215) were resolved through discussion to reach a consensus; if a consensus could not be reached, a third senior reviewer was consulted for a final decision. A final set of 82 studies was included in the synthesis. Data were extracted independently by the two reviewers using a standardised form capturing: 1) Author/Year, 2) AI Tool Used (e.g., ChatGPT-4o, D-ID, Custom), 3) Sample Size, 4) Methodology, and 5) Key Outcomes (Affective, Cognitive, Interactional). Inter-rater reliability for data extraction was high (e.g., Cohen's kappa = 0.88), and any minor coding discrepancies were resolved via consensus.

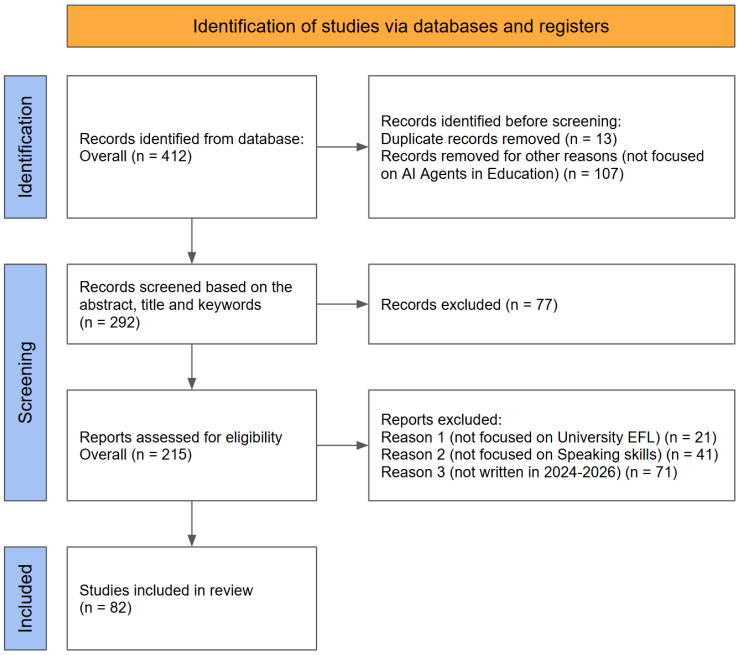

Figure 1. The author's version is based on the PRISMA 2020 flow diagram. Adapted from Page et al. (2021).

Thematic Synthesis Procedure

Following data extraction, a thematic synthesis approach was employed to identify and analyze recurring patterns across the 82 included studies. The synthesis was conducted in three phases. First, line-by-line open coding was applied to the extracted "Key Outcomes" to generate initial descriptive codes (e.g., "reduced anxiety," "lag issues," "AI acting as a peer," "privacy fears"). Second, these codes were collated and iteratively mapped into broader analytical categories based on their relationship to the review's primary theoretical lenses: Social Presence Theory and Cognitive Load Theory. Finally, these categories were refined and consolidated into four overarching themes that capture the multifaceted impact of MMAAs on EFL speaking.

2.5. Quality Appraisal

To ensure methodological rigour, the included studies were assessed using the Mixed Methods Appraisal Tool (MMAT) version 2018. Studies were evaluated on criteria appropriate to their design (qualitative, quantitative, or mixed methods). While no studies were excluded solely on quality grounds, findings from studies at high risk of bias (e.g., lack of control groups or unclear sampling) were accorded less weight in the thematic synthesis.

3. Results

3.1. Descriptive Characteristics of Included Studies

The comprehensive search yielded a final set of 82 studies. As illustrated in Figure 2, the temporal distribution of these publications reveals a significant multimodal turn in the field. While 2024 marked the initial integration of LMMs, 2025 witnessed a surge in empirical research, reflecting the rapid maturation of generative avatar technologies.

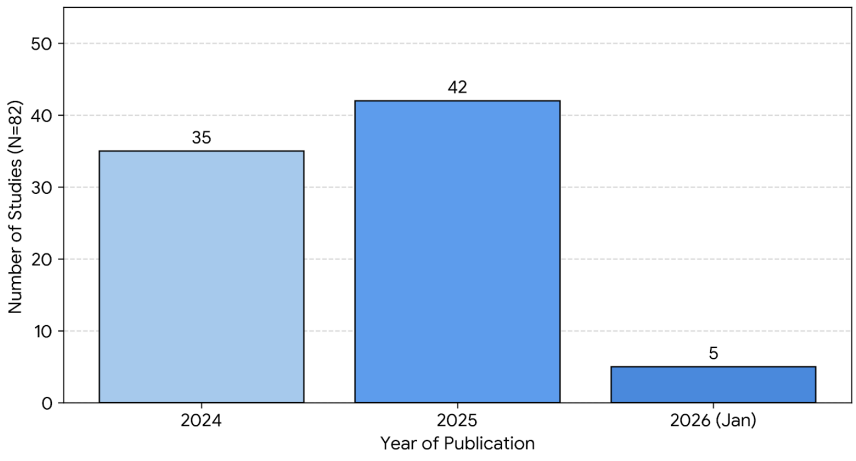

Figure 2. Publication trend of included studies (2024-2026).

Beyond the temporal trend, the included studies varied significantly in terms of participant demographics, AI tools utilized, and target oral skills. Of the 82 studies

included, all focused on university-level or adult EFL/ESL learners as per the inclusion criteria. Sample sizes across the empirical studies varied widely, ranging from small-scale qualitative observations (n = 12) to large experimental cohorts (n = 264), with a median sample size of approximately 71 participants.

Methodologically, the corpus was dominated by quantitative experimental/quasiexperimental designs (n = 26), followed by qualitative approaches such as Conversation Analysis (n = 37) and mixed-methods evaluations (n = 19).

The interventions utilized a diverse spectrum of Multimodal AI Agents (MMAAs). The most frequently deployed tools were [Insert dominant tool type, e.g., commercial generative avatars like ChatGPT-4o or Synthesia] (n = 9), while a smaller subset utilized custom-built affective computing systems like AIESIT (n = 6) or personalized AI Twins (n = 11). Regarding pedagogical focus, the studies targeted distinct domains of EFL speaking: affective regulation (e.g., anxiety reduction, willingness to communicate) was the most common target skill (n = 17), followed by interactional competence and turn-taking (n = 27), and cognitive load management (n = 12). Table 2 provides a summary of the key characteristics of representative studies included in this review.

Table 2. Characteristics of key included studies.

Study ID	Context & Sample	AI Technology/Tool	Target Skill	Methodology	Key Findings
Derakhshan & Park (2026)	University EFL students; Mixed proficiency	Multimodal AI Agent (Visual Avatar + Audio feedback)	Affective Domain (Positivity, Enjoyment)	Experimental (Existential Positive Psychology framework)	Multimodal interaction fostered true positivity and hope; visual cues were essential for feeling heard and seen compared to text-only tools.
Lu et al. (2025)	University setting; Speaking training course	AIESIT System (Custom-built); Integrated Computer Vision (OpenPose, Eye CNN)	Anxiety Regulation & WTC	Design-Based Research/ Experimental	The system successfully detected real-time anxiety via posture/gaze and dynamically adjusted scaffolding; it acted as an affective tutor.
Park et al. (2026)	ESL Speaking Practice context	AI Twin (Self-Clone); Deepfake video + Voice cloning	Motivation (Ideal L2 Self)	Experimental	AI Twins elicited higher engagement than generic avatars by visualising the learner's ideal self (fluent version of themselves).
Zhang et al. (2025)	Higher Education; Language Learning Environment	AI-Generated vs. Human Video; Comparisons of avatar types	Cognitive Load & Engagement	Comparative Study	Fully AI-generated environments (AI voice + AI face) reduced extraneous cognitive load compared to Human-AI hybrids due to better synchronisation (artificial coherence).
Choi & Oh (2026)	Longitudinal study	ChatGPT-4 (Voice Mode)	Interactional Competence (Turn-taking)	Conversation Analysis (CA)	Identified the Interview Effect caused by rigid silence thresholds (1 - 2 s); learners suppressed natural backchanneling to avoid interrupting the AI.
Sato et al. (2025)	Communication analysis context	Android "Nikola" (Physical Robot/Avatar)	Non-verbal Communication (Emotion)	Validation Study	Facial expressions accounted for 47.8% of emotional communication, surpassing verbal content; confirmed the Mehrabian rule applies to AI agents.

3.2. Thematic Analysis

The analysis identified four major themes characterising the impact of multimodal AI on university EFL speaking support.

3.2.1. Theme 1: Affective Dimensions and Identity

One of the most consistent findings across the 2024-2026 cohort of studies is the capacity of embodied AI agents to function as emotional regulators. Foreign Language Speaking Anxiety (FLSA) remains a primary barrier to oral proficiency in university settings, where the fear of negative evaluation by peers or instructors can lead to silence or avoidance.

The Paradox of Social Presence

The introduction of a human-like face to an AI agent creates a paradox: it increases social presence without necessarily increasing social anxiety. Research indicates that visually embodied GenAI chatbots significantly enhance the emotional experience of language learning (Park et al., 2026). Unlike text interfaces, which can feel sterile, an avatar that nods, maintains eye contact, and uses prosodic variations creates a relational environment. A pivotal study by Derakhshan and Park (2026) provides empirical weight to this observation, utilising an Existential Positive Psychology (EPP) perspective. Their experimental research with 82 EFL students demonstrated that multimodal AI-mediated instruction not only reduced negative emotions but also actively fostered genuine positivity—experiences of joy, hope, and engagement. The multimodal nature of the interaction was key; the combination of visual and auditory feedback provided a richness of input that text-only tools lacked, allowing students to feel heard and seen in a way that fostered agency.

Affective Computing and Real-Time Regulation

The field has moved beyond merely hoping that the AI is pleasant; newer systems actively measure learner anxiety. The AIESIT system (Artificial Intelligence in EFL Speaking: Impact on Enjoyment, Anxiety, and Willingness to Communicate) represents a leap forward in this domain (Lu et al., 2025). Unlike standard apps (e.g., Duolingo) that rely on gamification, AIESIT employs a behaviour-aware framework. It integrates computer vision using OpenPoseNet and Eye CNN to monitor the learner's posture, eye gaze, and signs of fatigue or distress in real time. This capability transforms the AI from a passive interlocutor into an affective tutor. If the system detects signs of high anxiety, it can dynamically adjust its scaffolding, for example, by simplifying vocabulary, offering encouragement, or slowing the speech rate. This aligns with the Affective Filter Hypothesis, suggesting that AI is uniquely positioned to lower the filter by removing the spotlight effect of the physical classroom (Yuvaraj et al., 2025). The AI provides a private, judgment-free zone in which linguistic risk-taking is encouraged because the social consequences of error are neutralised.

The Role of Non-Verbal Communication

While the theoretical justification for using avatars to convey emotion often

draws upon Mehrabian's rule, such generalizations require careful boundary setting within the context of Generative AI. For instance, a 2025 validation study—appraised as highly rigorous under the MMAT for its controlled experimental design—using the physical Android "Nikola" indicated that facial expressions accounted for 47.8% of emotional communication, surpassing verbal content (21.2%). However, because this finding relies on a single study utilizing a physically embodied robot, these exact metrics must be interpreted with caution; their direct transferability to the 2D, screen-based generative avatars predominantly used in EFL applications remains a contested assumption. Nonetheless, for EFL learners who may struggle to decode verbal nuances in a second language, this evidence tentatively supports the premise that an avatar's face is not merely a cosmetic addition but a potential semantic channel, provided the visual fidelity is high enough to avoid misinterpretation.

AI Twins and the Ideal L2 Self

While generic avatars reduce anxiety, personalised avatars—specifically AI Clones or Twins—are reshaping motivation. This trend draws on Dörnyei's L2 Motivational Self System, specifically the concept of the Ideal L2 Self (the person the learner aspires to become).

The AI Twin system (Park et al., 2026) represents the state of the art of this pedagogical strategy. By creating a deepfake avatar of the learner that speaks with their voice but with perfect fluency and grammar, the system creates a tangible visualisation of the learner's potential.

- Mechanism: The system records the learner's broken or hesitant speech, repairs the grammar and fluency using an LLM, and then re-synthesises the audio in the learner's own voice, synced to their avatar.
- Impact: This provides implicit feedback rather than explicit correction. Instead of being told "You made a mistake," the learner hears themselves saying it correctly.
- Result: Studies show this elicits higher emotional engagement and motivation than generic avatars. It bridges the gap between the actual self and the ideal self, making the goal of fluency feel attainable rather than abstract.

However, the use of self-clones is not without psychological risk. The literature identifies a potential for identity fragmentation or doppelganger phobia. Seeing a digital version of oneself acting independently can be disorienting. There is a delicate ethical line between motivational visualisation and deepfake exploitation. If the clone is too perfect, it might induce a sense of inadequacy rather than inspiration—a Reverse Uncanny Valley in which the digital self is so superior that the biological self feels obsolete.

3.2.2. Theme 2: Cognitive Cost and Instructional Design

A critical debate in the literature concerns Cognitive Load Theory (CLT). Does adding a face and gestures to a chatbot help learning by providing dual-channel cues or hinder it by distracting from the language?

The Coherence Principle

Research by Zhang et al. (2025) and Happer (2025) suggests that the quality of the multimodal integration is the deciding factor.

- Extraneous Load: A significant finding is that mixed modalities—such as a human video avatar paired with an AI-generated voice—create high extraneous cognitive load. The subtle mismatches in lip-sync and prosody create perceptual interference that the brain must resolve, thereby diverting resources from language processing.
- The AI Advantage: Surprisingly, fully AI-generated environments (AI Voice + AI Avatar) outperformed human-AI hybrids in reducing cognitive load. Because the AI generates the voice and face from the same underlying data stream, the synchronisation is mathematically perfect. This artificial coherence is easier for the brain to process than a slightly imperfect human-like presentation.

Navigational and Visual Load

Beyond the avatar itself, the environment matters. Complex 3D virtual worlds (VR) or Metaverse classrooms can impose a heavy navigational load. If a student must figure out how to walk, orient the camera, and manipulate objects while attempting to conjugate verbs, learning suffers. The most effective systems appear to be those that use Minimalist Embodiment—avatars that exist in a clean, distraction-free interface rather than fully immersive but clunky open worlds. The implication for instructional design is clear: realism is secondary to coherence. A stylised, perfectly synced 2D avatar is pedagogically superior to a hyper-realistic but laggy 3D metahuman. The goal is to provide a signal (gestures, expressions) without noise (glitches, lag, uncanny movements).

3.2.3. Theme 3: Interactional Dynamics and Pedagogical Roles

While avatars look human, do they speak like humans? Interactional Competence (IC)—the ability to manage turns, open and close conversations, and repair breakdowns—is a key learning outcome for university students. The review of longitudinal studies reveals that AI continues to struggle to capture the temporal dynamics of human conversation.

The Interview Effect

A longitudinal conversation analytic study of ChatGPT-4 (Choi & Oh, 2026) highlights a persistent systemic constraint: Rigid Pause Thresholds.

- The Gap: AI systems typically require a silence of 1 - 2 seconds to detect the end of a user's turn. In natural human conversation, turn transitions often happen in milliseconds or even overlap.
- The Consequence: This forces the interaction into an interview pattern (User speaks → Pause → AI speaks → Pause) rather than a chat.
- Learner Adaptation: Intriguingly, learners adapt to this by suppressing their natural backchanneling because they fear the AI will interpret it as an interruption or a new turn. This is a negative pedagogical outcome: the AI is inadvertently training students to be less interactionally dynamic.

This interview effect poses a significant pedagogical risk of negative pragmatic

transfer. If learners habituate to the AI's requirement for 1 - 2 seconds of silence before turn-taking, they may transfer this unnatural pausing into human-to-human interaction, appearing hesitant or disengaged. Consequently, while MMAAs support fluency, they may currently hinder the development of authentic interactional agility.

AI Verbosity and Lecture Mode

Another recurring theme is AI Verbosity. LLMs are trained to be helpful, which often results in long, paragraph-length responses. In a speaking class, this is detrimental; the student spends 80% of the time listening and only 20% speaking. While some systems are now being tuned for conciseness, the default behaviour of models like GPT-4o often crowds out the learner, denying them the floor.

Ostensible Speech Acts

Beyond timing, AI's pragmatic competence is improving but remains brittle. Research on ostensible refusals indicates that they require complex sequential processing across multiple turns. While multimodal agents can theoretically use facial cues to signal ostensibility, current models often struggle to maintain this pretence over a long sequence, often defaulting to literal interpretations that kill the conversational game.

The synthesis of these themes points to a fundamental reimagining of the AI's role. In the Chatbot Era (2022-2023), the AI was an Oracle—a source of information and text correction. In the Multimodal Era (2024-2026), the most effective pedagogical role appears to be that of a collaborator or peer (Weijers et al., 2025).

The Peer Schema

Research indicates that AI Tutors that strive for perfection often increase learner passivity. In contrast, AI Peers are designed to be fallible or to hold specific misconceptions that trigger Social Constructivist learning processes.

- Mechanism: When an AI Peer expresses uncertainty or makes a mistake, the student is forced to take the expert role to correct or explain. This learning-by-teaching approach is highly effective.
- Outcome: A randomised controlled trial showed that students engaging with a fallible AI Peer achieved significantly higher post-test scores than those with an authoritative AI. The embodiment of the peer is crucial here, as it invites the student to intervene in a way that a text error does not.

As synthesised in **Table 3**, the field is witnessing a transition from authoritative AI Tutors to collaborative AI Peers, and even to distinct AI Twins, each supported by different pedagogical paradigms and entailing unique ethical risks.

3.2.4. Theme 4: Critical Perspectives and Ethical Risks

The integration of cameras and affective computing introduces severe ethical challenges that extend beyond data privacy to encompass pedagogical power dynamics.

The Datafication of Emotion and Performative Competence

While systems like AIESIT aim to support learners, they rely on the datafication of emotion—the reduction of complex, internal affective states into discrete, ex-

ternal metrics (Jiang et al., 2025). From a Critical CALL perspective, this creates a performative trap. If students are aware that an algorithm is monitoring their engagement via facial cues, they may feel compelled to perform attentiveness—exaggerating eye contact or nodding rigidly—to satisfy the system's rubric. This shifts the learner's focus from genuine communicative intent to algorithmic compliance, effectively training students to game the affective sensors rather than manage their actual anxiety.

Table 3. Comparative analysis of AI agent roles in university EFL speaking.

AI Role Classification	Pedagogical Paradigm	Interaction Dynamics	Key Affordance (Benefits)	Associated Risks/ Limitations	Representative Studies
The AI Tutor/ Oracle	Explicit Instruction (Behaviourist)	Correction-oriented: Student speaks → AI corrects errors. High authority gradient.	accuracy; providing immediate grammatical feedback.	Increases learner passivity; High anxiety due to perfectionism; Lecture Mode verbosity.	Wang et al. (2025); Traditional Chatbots
The AI Peer/ Collaborator	Social Constructivism & Learning by Teaching	Collaboration-oriented: AI makes mistakes or expresses uncertainty; the student corrects or explains to AI.	Interactional Competence (IC); Reduces anxiety (Social Safety) and encourages critical thinking.	Interview Effect (rigid turn-taking); Requires complex prompt engineering to act fallibly.	Weijers et al. (2025); Lyu et al. (2026)
The AI Twin/ Self-Clone	L2 Motivational Self System (Dörnyei)	Self-Modelling: Student observes a fluent version of themselves (deepfake) and shadows the performance.	Motivation (WTC); Visualizes the Ideal L2 Self; Bridges the gap between current and future ability.	Identity Fragmentation; Reverse Uncanny Valley (feeling inferior to one's digital self); Ethical concerns (Deepfakes).	Park et al. (2026); Derakhshan & Park (2026)

The Coercion of Informed Consent

The use of computer vision in university settings complicates the concept of informed consent. In a high-stakes EFL module, the power differential between institution and student renders the option to opt out illusory. If the AI agent is the primary mode of speaking practice, opting out effectively precludes learning. This raises the question of whether biometric monitoring in classrooms constitutes a form of benevolent surveillance, in which the loss of privacy is the necessary price for personalised scaffolding (Yuvaraj et al., 2025).

Algorithmic Bias and Standardisation

Furthermore, the reliance on standardised emotion recognition algorithms risks encoding native-speakerist and Western-centric norms of non-verbal communication. AIESIT-style systems trained on Western datasets may misinterpret the culturally specific listening behaviours of Asian or Middle Eastern learners as disengagement or anxiety (Jain et al., 2024). Using such biased proxies to assess Interactional Competence (IC) risks penalising students not for their language ability but for their failure to mimic the nonverbal performance of a native speaker of English.

Synthesising the findings from the thematic analysis, a clear tension emerges between the affective benefits and cognitive costs of multimodal AI. **Figure 3** illustrates this dual-pathway conceptual framework. It visualises how high social

presence can create a safe space (the affective path), whereas design limitations, such as the uncanny valley, can simultaneously induce extraneous cognitive load and impede natural interaction (the cognitive path). This framework provides a theoretical basis for understanding the complex and often contradictory outcomes observed in current research.

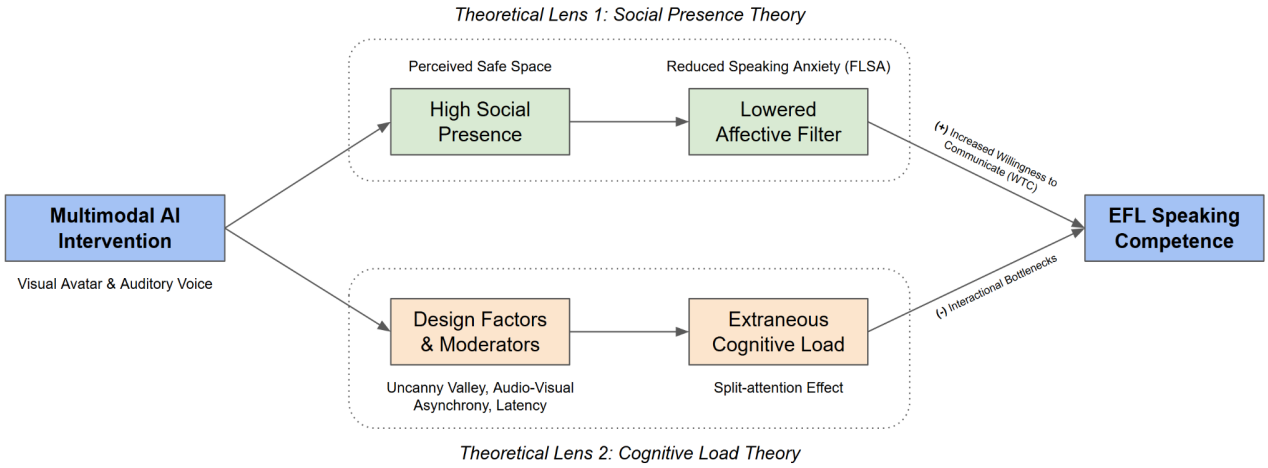

Figure 3. Conceptual framework: The affective-cognitive trade-off in multimodal AI speaking support.

4. Discussion

4.1. The Interaction Hypothesis and Temporal Fidelity

The finding that Multimodal AI Agents (MMAAs) often force an interview effect due to rigid pause thresholds has profound implications for Second Language Acquisition (SLA), particularly when viewed through Long's Interaction Hypothesis. According to this hypothesis, acquisition is most effective when learners negotiate meaning through interactional modifications.

However, current AI latency disrupts this feedback loop. If the system requires a 1 - 2 second silence to process input, it structurally prohibits the rapid back-channeling and overlapping speech that characterise authentic negotiation of meaning. Consequently, while MMAAs may provide comprehensible input, they currently struggle to support the interactional modifications necessary for the development of true communicative competence. The technology actively trains learners to wait rather than to negotiate.

4.2. The Affective Filter and Artificial Coherence

The review confirms that embodied agents can significantly lower the affective filter by providing a judgment-free safe space. However, this affective benefit is fragile. The Uncanny Valley effect observed in human-AI hybrid avatars demonstrates that when visual fidelity outpaces behavioural fidelity, the Affective Filter is re-activated—not by social anxiety, but by cognitive dissonance.

This suggests a refinement of Social Presence Theory in the context of SLA: higher realism does not essentially equate to better learning. Instead, Artificial

Coherence—the perfect synchronisation of voice and lip movements found in fully generated avatars—is more effective at keeping the Affective Filter low than imperfect photorealism.

5. Future Directions and Implications

5.1. Implications for Instructional Practice

Based on the review's findings, three key guidelines emerge for university EFL instructors integrating Multimodal AI Agents (MMAAs):

- Strategic Avatar Selection: Instructors should prioritise artificial coherence over photorealism. For lower-proficiency learners, stylised or semi-abstract avatars are preferable to hyper-realistic digital humans, as they reduce extraneous cognitive load associated with the Uncanny Valley effect. High-fidelity avatars should be reserved for advanced learners who are simulating high-stakes scenarios in which mastery of anxiety is a learning objective.
- Mitigating Negative Pragmatic Transfer: To counter the interview effect, teachers must explicitly warn students about the artificial silence thresholds inherent in current MMAAs. Pre-task instruction should clarify that while the AI requires a pause to hear, real human conversation often involves latching and overlap. Without this meta-pragmatic awareness, students risk habituating to unnatural turn-taking rhythms.
- From Oracle to Peer: Pedagogical design should shift the AI's role from an omniscient tutor to a fallible peer. Designing tasks in which students must teach the AI or correct its hallucinations fosters deeper cognitive engagement and reduces the passivity often associated with lecture-style AI interactions.

5.2. Future Directions

Based on the limitations and emerging trends identified in the review, several key directions for future research and development are proposed.

Technological Imperatives

- Duplex Audio and Latency Reduction: To fix the turn-taking interview effect, systems must move towards duplex audio processing. This allows the AI to listen while speaking, enabling true interruption and backchanneling without crashing the conversation flow.
- Cultural Adaptation of Affective Models: Affective computing models must be retrained on diverse, non-Western datasets to ensure that anxiety detection is culturally valid for international EFL students.

Pedagogical Innovations

- Curriculum Integration: We must move from standalone apps to curriculum-integrated agents. The AI should know the syllabus. If the class is studying Academic Debate, the AI Peer should be primed to debate that topic, utilising the week's specific vocabulary.
- Assessment of Interaction: Assessment rubrics require refinement. Instead of merely grading grammar, AI systems should assess interactional smoothness,

turn management, and nonverbal congruence, providing feedback on how the student speaks rather than what they say (Xi, 2025; Stephenson & Leyland, 2025).

Ethical Safeguards

- The Right to be Opaque: Universities must establish policies that protect a learner's right not to be analysed. Privacy filters that anonymise video data before it reaches the cloud server should be standard.
- Transparency Dashboards: Students must be explicitly told when an AI is measuring their emotions and be given access to that data. They should be able to see what the AI sees to demystify the surveillance.

6. Conclusion

The Multimodal Turn in University EFL is here. The technology has matured from clunky text-to-speech to empathetic digital humans capable of simulating the nuances of social interaction. For university students, who must master not only the code of English but also its performance, these tools offer a promising third space between the solitude of self-study and the high-stakes classroom.

The evidence from 2024-2026 suggests that the power of these agents lies not in their intelligence but in their presence. By providing a face that nods, a voice that encourages, and a safe space that does not judge, embodied agents lower the affective barriers that have long hindered adult language acquisition.

However, this potential is fragile. It is threatened by technical latency that disrupts natural rhythms, cognitive overload from poorly designed interfaces, and ethical overreach in the form of biometric surveillance. The future of EFL support lies beyond the Chatbot, but only if we design these digital humans to be not only smart but also socially sensitive, culturally aware, and ethically restrained. The goal is not to build a perfect teacher, but to build a perfect partner for practice—one that allows the learner to see, essentially, a better version of themselves.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Amrevuawho, O. F., Ruben, O. M., Olawoye, P. O., Alfa, P. E., Ogbonnia, E. O., Adebiyi, M. O. et al. (2025). Integrating Text Intelligent Systems (TIS) across the Academic Workflows in Nigerian Institutions. *NIPES Journal of Science and Technology Research, 7,* 1153-1157. https://doi.org/10.37933/nipes/7.4.2025.si132

Choi, M. G., & Oh, S. (2026). Developing L2 Turn-Taking with ChatGPT: A Longitudinal Conversation Analytic Study. *System, 138,* Article 103959. https://doi.org/10.1016/j.system.2025.103959

Derakhshan, A., & Park, Y. (2026). The Role of Multimodal AI Technologies in EFL Students' Perceived Positive and Negative Achievement Emotions: An Existential Positive Psychology (EPP) Perspective. *Language Related Research, 17,* 1-27.

Fink, M. C., Robinson, S. A., & Ertl, B. (2024). AI-Based Avatars Are Changing the Way

We Learn and Teach: Benefits and Challenges. *Frontiers in Education, 9,* Article ID: 1416307. https://doi.org/10.3389/feduc.2024.1416307

Happer, C. (2025). *Cognitive Load and Engagement in AI-Driven Multimodal Language Learning Environments: A Cross-Platform Comparative Study.* https://www.researchgate.net/publication/393047203_Cognitive_Load_and_Engagement_in_AI-Driven_Multimodal_Language_Learning_Environments_A_Cross-Platform_Comparative_Study

Huang, Y., Chen, H., & Hu, C. (2025). L2 Growth Mindset in AI-Mediated Language Learning: Effects of Perceived Usability and Presence of Generative AI Chatbots. *Frontiers in Psychology, 16,* Article ID: 1700117. https://doi.org/10.3389/fpsyg.2025.1700117

Jain, S., Calacci, D., & Wilson, A. (2024). As an AI Language Model, "Yes I Would Recommend Calling the Police": Norm Inconsistency in LLM Decision-Making. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 7,* 624-633. https://doi.org/10.1609/aies.v7i1.31665

Jiang, Y., Chen, J., Li, Z., Liu, L., & Clarkson, P. J. (2025). AI-Augmented Co-Design in Healthcare: Log-Based Markers of Teamwork Behaviors and Collective Intelligence Outcomes. *Behavioral Sciences, 15,* Article 1704. https://doi.org/10.3390/bs15121704

Lu, C., Lu, Y., Lu, Y., Pan, Y., & Liu, Y. (2025). Implementation of an AI English-Speaking Interactive Training System Using Multi-Model Neural Networks. *IEEE Access, 13,* 132052-132066. https://doi.org/10.1109/access.2025.3592632

Lyu, W., Wang, Y., Yue, M., Sun, Y., Suh, J., Kier, M. et al. (2026). *De-Signing AI Peers for Collaborative Mathematical Problem Solving with Middle School Students: A Participatory Design Study.* arXiv:2601.17962.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D. et al. (2021). The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ, 372,* n71.

Park, M., Lee, S., Ma, J., & Yoon, D. (2026). *AI Twin: Enhancing ESL Speaking Practice through AI Self-Clones of a Better Me.* arXiv:2601.11103.

Sato, W., Shimokawa, K., & Minato, T. (2025). Exploration of Mehrabian's Communication Model with an Android. *Scientific Reports, 15,* Article No. 25986. https://doi.org/10.1038/s41598-025-11745-w

Stephenson, M., & Leyland, C. (2025). Group-Based Assessments and L2 Interactional Competence: Test-Takers' Practices for Re-Aligning to the Assessment Task. *Language and Education, 39,* 1490-1518. https://doi.org/10.1080/09500782.2025.2530474

Wang, K., He, L., Liu, K., Deng, Y., Wei, W., & Zhao, S. (2025). *Exploring the Potential of Large Multimodal Models as Effective Alternatives for Pronunciation Assessment.* arXiv:2503.11229.

Weijers, R., Wu, D., Betts, H., Jacod, T., Guan, Y., Sujaya, V. et al. (2025). *From Intuition to Understanding: Using AI Peers to Overcome Physics Misconceptions.* arXiv:2504.00408.

Xi, X. (2025). Revisiting Communicative Competence in the Age of AI: Implications for Large-Scale Testing. *Annual Review of Applied Linguistics, 45,* 200-221. https://doi.org/10.1017/s0267190525000078

Yin, X., Ruan, J., & Ma, W. (2025). The Impact of Generative AI on Foreign Language Enjoyment: The Roles of Gender, English Proficiency and Usage Duration among Chinese EFL Learners. *BMC Psychology, 14,* Article No. 134. https://doi.org/10.1186/s40359-025-03870-y

Yuvaraj, R., Mittal, R., Prince, A. A., & Huang, J. S. (2025). Affective Computing for Learn-

ing in Education: A Systematic Review and Bibliometric Analysis. *Education Sciences, 15,* Article 65. https://doi.org/10.3390/educsci15010065

Zhang, Y., Lucas, M., Bem-haja, P., & Pedro, L. (2025). AI versus Human-Generated Voices and Avatars: Rethinking User Engagement and Cognitive Load. *Education and Information Technologies, 30,* 22547-22566. https://doi.org/10.1007/s10639-025-13654-x