International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 15, Issue 01, January 2026)

Multimodal AI Tutors Redefining Learner Autonomy in ELT

Dr. Abdul Majeed C.T
Majlis College of Arts and Science, Puramannur, Malappuram, Kerala, India.

***Abstract*-- The rapid advancement of multimodal artificial intelligence (AI)—which integrates text, speech, images, and increasingly, gesture and facial expression recognition—marks a significant turning point for English Language Teaching (ELT). Multimodal AI tutors can provide contextualised, multimodal input and feedback, support tasks across different modalities, personalise pacing and content, and maintain learner motivation. These features promise to shift the focus of control toward learners, enhancing their autonomy in ways that earlier computer-assisted language learning (CALL) systems did not fully achieve. This article synthesises theoretical foundations, including multimodality, autonomy, and self-regulated learning. It surveys recent empirical research and prototypes, examines pedagogical advantages and practical design principles, and critically discusses risks, equity concerns, and future research and classroom implementation. The target audience includes ELT researchers, teacher educators, and practitioners designing AI-enhanced curricula. Additionally, this piece aims to serve as a draft-length article for submission or a long-form report.**

***Keywords*-- Multimodal Artificial Intelligence; Learner Autonomy; English Language Teaching(ELT); AI Tutors; Self-Regulated Learning; Technology-Enhanced Language Learning (TELL)**

I. INTRODUCTION

Learner autonomy, the ability and willingness of learners to take charge of their own education, has long been a goal in language teaching. Traditional classroom limitations, such as fixed schedules, having one teacher for many students, and standardised materials, restrict opportunities for personalised control over task selection, feedback timing, and modes of practice. Although digital technologies have promised to alleviate some of these constraints, many earlier computer-assisted language learning (CALL) systems were limited in their responsiveness and operated primarily in a single mode (such as text or basic audio).Multimodal AI tutors (referred to as "MM-AI tutors") transform the educational landscape. By integrating language models, speech recognition and synthesis, computer vision, and multimodal reasoning, these systems offer simultaneous, context-rich input and feedback. For example, they can combine spoken prompts, on-screen images, and corrective gestures or visual annotations.

This approach tailors interactions to the learners' goals, emotional states, and pace. As a result, students benefit from more diverse practice opportunities and may experience significant changes in how learning activities are directed. Learners can choose different modes of interaction, request clarifications in various formats, and receive adaptive support that fosters self-regulation and encourages gradual independence..

This article explains how MM-AI tutors can effectively support learner autonomy in English Language Teaching (ELT), grounding its claims in recent research and prototypes, and offering practical guidance for design and classroom adoption — balanced with an appraisal of ethical and equity implications.

II. THEORETICAL FOUNDATIONS

2.1Multimodality and Language Learning

Multimodality acknowledges that communication and learning happen through various channels, including spoken language, written text, images, gestures, and spatial layout. In language education, multimodal tasks—such as videos, presentations, and role-plays that utilise visual props—have been shown to enhance motivation, reduce anxiety during oral production, and facilitate meaning-making by distributing cognitive load across different modalities. These multimodal resources help learners connect form and meaning through complementary channels; for instance, gestures can support phonology, while images can aid in lexical recall. Recent reviews emphasise that multimodal pedagogy not only promotes linguistic competence but also enhances strategic and emotional aspects of learning. Multimodality and Language Learning

2.2 Learner autonomy and self-regulation

Autonomy in language learning involves multiple dimensions: cognitive (strategy use), metacognitive (planning, monitoring, evaluating), motivational (ownership, persistence), and social (seeking resources, negotiating with peers). Technologies that support these capacities can enhance autonomy by providing choice architecture, transparent feedback, and tools for self-assessment.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 15, Issue 01, January 2026)

Traditional strategies for fostering autonomy include negotiated syllabi, self-assessment rubrics for learners, and structured reflection prompts. Multimodal AI expands this toolkit by offering continuous support for reflection, monitoring, and adaptive scaffolding across various modes.

2.3 AI tutors as cognitive and metacognitive scaffolds

Intelligent Tutoring Systems (ITS) and conversational tutors emulate key aspects of human tutoring, such as diagnosing errors, prompting reflection, and providing tailored explanations. When integrated with multimodal inputs—like a learner's speech, their written drafts, or referenced images—multimodal AI tutors can develop more comprehensive diagnostic models and offer deeper support. For example, they can identify pronunciation errors in connected speech while simultaneously presenting visual diagrams of articulation and practice videos. These multimodal supports can encourage metacognitive strategies, such as advising learners to "listen again and note three words you mispronounced," which fosters independent self-regulation. Recent prototypes of ITS and multimodal tutors have successfully showcased these capabilities in English Language Teaching (ELT) contexts.

2.4 What makes an AI tutor "multimodal"?

A multimodal AI tutor incorporates at least two different input/output modes (commonly text, speech, and visuals) and analyses the relationships between them. Core components include:

2.4.1Multimodal input processing: Automatic speech recognition (ASR) for learner speech; optical character recognition (OCR) and image analysis for submitted learner images or writing; gesture, pose, or facial-affect recognition when camera data is used.

2.4.2 Multimodal generative models: Engines capable of producing synchronised outputs across various modes—such as a spoken prompt accompanied by animated mouth shapes, or a written correction shown alongside an explanatory image or a brief explainer video.

2.4.3 Dialogue and pedagogical management: A conversational manager oversees turn-taking, scaffolding depth, and learning pathways; it integrates learner models, including knowledge, emotions, and preferences.

2.4.4 Learner modelling and analytics: Continuous assessment algorithms evaluate proficiency, engagement, and strategy usage based on multimodal signals, adapting content and feedback accordingly.

2.4.5 Authoring and teacher dashboards: Teachers have interfaces that allow them to customise tasks, review analytics, and intervene as needed. Practical multimodal AI tutors vary in aspects such as the richness of their modalities, interpretability, latency (whether feedback is real-time or delayed), and whether they are designed for controlled tasks (like pronunciation drills) or open-ended tasks (such as conversational practice). Recent prototype systems—ranging from multimodal dialogic tutors to avatar-based conversational AI—demonstrate how these components can be effectively integrated within English Language Teaching (ELT) contexts.

III. How Multimodal AI Tutors Support Learner Autonomy

In the following sections, I will explore the specific ways in which MM-AI tutors support autonomy, providing illustrative scenarios to clarify these concepts.

3.1 Expanded choice architecture (what, when, how learners practice)

MM-AI tutors enable learners to select not only topics but also modes of learning. For example, learners can choose to practice vocabulary through image labelling, spoken recall, or micro-videos. They can also adjust the timing of feedback—opting for immediate corrections or delayed summaries—as well as the intensity of support, choosing between gentle prompts and explicit corrections. This freedom of choice fosters a sense of ownership in the learning process. Research indicates that perceived choice enhances motivation and persistence among learners. For language learners, selecting modalities that align with their preferred strategies, such as visual or auditory methods, promotes self-directed engagement.

Scenario: A learner preparing for an oral presentation selects the following sequence of modes: (a) automatic pronunciation analysis of rehearsal audio, (b) on-demand visual feedback that shows lip and tongue placement, and (c) a simulated Q&A conversation with a tutor avatar. The system maintains a log of the learner's choices and suggests a pathway with reduced support as their competence improves.

3.2 Immediate, multimodal formative feedback that learners can act upon

MM-AI tutors offer corrections in various modes to effectively communicate issues. Pronunciation errors can be illustrated using spectrograms and animated articulatory diagrams.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 15, Issue 01, January 2026)

Grammatical mistakes can be highlighted in text, accompanied by brief spoken explanations and example sentences. Additionally, learners can practice discourse-level pragmatics through role-play video simulations. This multimodal approach makes feedback actionable and helps learners create multi-sensory memory traces that support independent practice.

Empirical note: Comparative studies of multimodal feedback versus text-only feedback in writing and speaking show greater improvements in willingness to communicate and sometimes in measurable performance, indicating that multimodal feedback has significant pedagogical value.

3.3 Scaffolding of metacognitive strategies

Multi-modal AI systems can incorporate prompts to guide both planning and reflection across different formats. Before a task, a tutor can encourage the learner to set goals through text input, prompt them to record a plan using audio or video, and later ask for a reflective portfolio artifact. By structuring the reflection process—often supported by checklists or visual comparisons of progress—the tutor helps learners develop monitoring habits, which are essential for fostering autonomy.

Scenario: After a series of conversational practice sessions, the tutor creates a brief multimodal dashboard. This dashboard includes a timeline of speaking turns, a heatmap highlighting common errors, and a 60-second compilation video showcasing the learner's best utterances. The tutor then asks the learner to select two specific targets for focused practice in the upcoming week.

3.4 Adaptive fading and gradually increasing independence

Effective tutors modify the level of support they provide. They begin with high scaffolding, which includes hints, examples, and guided practice, and gradually reduce this support as the learner's skills improve. Multimodal AI tutors can implement this gradual reduction of assistance across different channels. For instance, they might decrease explicit visual cues while increasing expectations for self-correction in audio-only interactions. This gradual withdrawal of support encourages learners to take more responsibility for their own Learning encourages learners to take greater responsibility for their own education.

Social and affective support is essential for enabling risk-taking. Feeling safe to experiment is crucial during speaking practice. MM-AI tutors create low-stakes environments that many learners find less intimidating than situations involving peers or teachers.

These tutors also incorporate affect-aware features that can detect anxiety or disengagement through voice or facial cues. As a result, they can adjust the difficulty of tasks and provide motivational support, fostering persistence and self-initiated learning. Studies involving chatbots and AI conversational tutors have shown positive effects on motivation, meeting autonomy-related needs (such as competence and relatedness), and increasing practice opportunities.

3.5 Evidence and prototypes: what recent work shows

Research conducted between 2023 and 2025 provides consistent evidence that AI-mediated, multimodal interventions can enhance autonomy-related outcomes such as motivation, willingness to communicate, and self-regulation, along with measurable language improvements. However, results may differ based on the task, context, and system design.

3.6 Conversational and multimodal tutor evaluations:

Evaluations of AI conversational tutors using mixed methods show improvements in speaking fluency and increased learner confidence. Qualitative reports also indicate more autonomous practice outside the classroom. These tutors often use real-time ASR and LLM-driven scaffolding.

3.7 Multimodal dialogic tutors and prototypes:

Industrial and academic prototypes, such as multimodal dialogic tutors showcased at ACL and ACM venues, demonstrate how synchronised speech, gestures, and visuals can simulate natural interactions and tailor responses. These systems illustrate the feasibility and pedagogical potential across diverse learner populations..

3.8 Systematic reviews & meta-analyses:

Recent systematic reviews of generative AI in language learning indicate that adaptive feedback and scaffolding features are linked to better learning outcomes. Additionally, multimodal designs are showing promise as a direction for future development. However, more rigorous randomised controlled trials (RCTs) and longitudinal studies are needed to identify the causal mechanisms involved.

3.9 CALL-based multimodal pedagogy studies:

Research on computer-assisted multimodal pedagogy indicates improvements in motivation and communication willingness; these emotional shifts likely serve as mediators for increased autonomous practice.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 15, Issue 01, January 2026)

IV. Contemporary Educational Technology

While encouraging, the literature also emphasises the variability in effect sizes and highlights the need for context-aware design—taking into account language level, cultural expectations regarding autonomy, and access to devices.

4.1 Design principles for MM-AI tutors that promote autonomy

To transform technological possibilities into effective pedagogy, designers should make principled design choices grounded in autonomy theory and language education

4.2. Learner control & transparent choices

Give learners clear options (modes, feedback types, timing). Make the consequences of choices transparent (e.g., "Choosing audio-only practice will reduce visual prompts; you can restore them anytime").

4.3 Scaffolded metacognitive routines

Embed short, actionable reflection prompts after tasks (e.g., "Rate your confidence 1–5; list one strategy to practice this week"). Provide templates and multimodal artefacts to support externalised reflection.

4.4 Adaptive, interpretable feedback

Ensure feedback is not only accurate but pedagogically interpretable. Combine corrective signals with brief explanatory content and examples. Allow learners to query "why" and receive multi-level answers (short explanation, expanded example, practice item).

4.5 Gradual fading & mastery paths

Design learning trajectories that gradually reduce external support and nudge learners towards independent practice, making mastery criteria explicit and joint planning visible.

4.6 Portfolios &artefacts for agency

Encourage learners to curate multimodal portfolios (audio clips, videos, annotated texts). Portfolios provide evidence of progress and support self-assessment and goal setting.

4.7 Human-in-the-loop and teacher orchestration

Provide teacher dashboards that make system inferences visible and allow teachers to override, supplement, or redirect learning paths. Teachers remain crucial for higher-order corrective feedback and socio-cultural mediation.

4.8 Privacy-by-design and accessibility

Offer modes that do not require camera/audio streams to respect privacy; include offline or low-bandwidth variants; ensure multimodal content is accessible (captions, transcripts, alternative descriptions). Accessibility provisions are integral to autonomy — learners must be able to choose modalities that suit their context and comfort.Applying these principles increases the likelihood that MM-AI tutors will foster genuine learner autonomy rather than merely automating teacher functions.

V. Practical Classroom Models For Adoption

MM-AI tutors can be integrated into ELT through several practical configurations:

5.1 Flipped-plus-AI model

Assign multimodal practice (micro-conversations, video prompts) for homework where the MM-AI tutor provides formative feedback. Class time becomes for human-facilitated communicative tasks using artefacts generated during AI practice.

5.2 Station rotation

Use MM-AI tutor stations for independent guided practice while the teacher focuses on small-group tasks. The AI station offers multimodal scaffolding, while other stations include peer interaction and teacher-led feedback.

5.3 Portfolio-based assessment

Let learners compile AI-scaffolded multimodal artefacts into a graded portfolio, with rubrics emphasising self-regulation and reflection as much as linguistic accuracy.

5.4 Supplementary informal learning

Encourage autonomous, incidental learning by granting learners access to MM-AI tutors for self-directed study (vocabulary games, pronunciation drills), with gamified progress markers that respect educational priorities.Each model requires clear instructions, teacher orientation to the tutor's affordances and limits, and policies about data and privacy.

VI. Challenges, Limitations, And Ethical Considerations

Despite potential, MM-AI tutors raise significant concerns that must be navigated carefully.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 15, Issue 01, January 2026)

6.1 Validity and bias in multimodal interpretation

AI interpretation of speech, facial affect, or gestures can be biased by accents, dialects, skin tone, and cultural display rules. Incorrect inferences about affect or competence can mislead learners and teachers or undermine confidence. Robust, diverse training data and evaluation across populations are essential.

6.2 Over-reliance and deskilling risks

If AI provides too much scaffolding without effective fading, learners may become dependent on external cues and fail to develop independent monitoring strategies. Designs must intentionally fade support and promote metacognitive practice.

6.3 Data privacy, surveillance, and consent

Multimodal systems often require audio, video, and textual data. Clear, consent-driven policies, local data storage options, and privacy-preserving modes (e.g., on-device processing, anonymisation) are non-negotiable. Teachers and institutions must ensure informed consent and alternatives for learners uncomfortable with camera/microphone use.

6.4 Equity and access

MM-AI tutors require hardware and bandwidth. Without careful deployment, benefits may accrue to well-resourced learners, widening gaps. Low-bandwidth and offline-first designs, plus institutional provisioning strategies, are necessary to ensure inclusive autonomy support.

6.5 Pedagogical alignment and teacher role erosion fears

Some teachers fear replacement. The productive framing is augmentation: AI should take on routine, labour-intensive tasks (e.g., low-level correction, providing controlled practice) so teachers can focus on high-impact activities (curriculum design, affective support). Teacher training is critical for productive integration.

VII. Evaluation And Research Agenda

To move beyond promising prototypes to reliable classroom tools, the field needs a coordinated research agenda:

7.1 Longitudinal, mixed-methods studies

Examine whether MM-AI tutors produce durable autonomy (sustained self-directed practice weeks/months after intervention) and transfer to classrooms without AI support.

7.2 Mechanisms of change

Identify which multimodal features (e.g., visual articulatory cues, multimodal feedback combinations) drive gains in autonomy and language outcomes.

7.3 Cross-cultural perspectives on autonomy

Autonomy has culturally mediated meanings. Research should probe how multimodal choice architectures map onto diverse learner expectations and institutional constraints.

7.4 Bias, fairness, and robustness testing

Evaluate interpretive components (ASR, affect detection) across linguistic varieties, ages, and demographics to surface and mitigate biases.

7.5 Design-based comparative trials

Compare MM-AI tutor configurations (real-time avatar vs. delayed multi-modal report vs. audio-only) to establish task-match heuristics.Existing reviews and systematic work provide initial evidence but call for more rigorous trials and transparency about system limitations.

VIII. Policy And Institutional Considerations

To scale ethically and sustainably, institutions must craft supportive policies:

Data governance frameworks enabling transparency, student control over artefacts, and options to opt out without penalty. Device and connectivity strategies to mitigate digital divides (lend devices, on-campus lab access, low-bandwidth alternatives).Teacher professional development programs that train instructors to interpret analytics, set meaningful learning goals with students, and use AI artefacts in formative assessment. Curriculum alignment to ensure AI practice maps to learning outcomes, assessment standards, and assessment validity. Institutions should pilot MM-AI tutors in controlled phases, gather stakeholder feedback, and iterate on policies before broad rollout.

IX. A Balanced Conclusion: Promise Tempered By Caution

Multimodal AI tutors represent a meaningful opportunity to operationalise learner autonomy in ELT at scale. Their strength lies in multimodal feedback, flexible choice structures, continuous assessment, and capacity for scaffolding and fading — core mechanisms that support cognitive, metacognitive, motivational, and social dimensions of autonomy.

Recent research and prototypes provide promising early evidence of improved motivation, willingness to communicate, and targeted language gains.

However, the promise will be realised only when developers, researchers, teachers, and policymakers collaborate to ensure (1) pedagogically principled design aligned with autonomy theory, (2) robust mitigation of bias and privacy risk, (3) accessible deployment strategies for diverse learners, and (4) rigorous long-term evaluation of learning and autonomy outcomes. Without these safeguards, MM-AI tutors risk widening inequities, fostering dependence, or producing opaque analytics that neither teachers nor learners can meaningfully interpret.

For ELT professionals, the immediate opportunities are practical and accessible: adopt multimodal practice modules for homework, use AI-generated portfolios as reflective instruments, and experiment with AI stations that free teachers for higher-level facilitation. For researchers, the task is to unpack causal mechanisms and cross-cultural dynamics; for technologists, it is to build interpretable, privacy-preserving multimodal models. For learners, the hope is an expanded capacity to choose, monitor, and drive their English learning — an autonomy supported by intelligent, multimodal scaffolds rather than constrained by them.

X. Practical Checklist For Teachers And Program Leaders

10.1 Start small — pilot one multimodal AI activity (e.g., AI-assisted pronunciation practice) with a volunteer group.

10.2 Clarify learning goals — map AI tasks to course outcomes and rubrics.

10.3 Prioritize consent &alternatives — always provide opt-outs for camera/audio; offer equivalent text-based tracks.

10.4 Embed reflection — require a short multimodal reflection artefact after AI practice.

10.5 Use portfolios — collect AI artefacts for formative assessment and to jointly set goals with learners.

10.6 Train teachers — short PD modules on reading dashboards, interpreting AI feedback, and scaffolding fade.

10.7 Monitor equity — track who uses the system and who benefits; address gaps with provisioning and adapted tasks.

10.8 Limitations-- Expand the limitations section with specifics about dataset biases, ASR error rates for target populations, and consent logistics.

REFERENCES:

[1] Szabó, F. (2024). How generative AI promotes autonomy and inclusivity in ELT. ELT Journal.

[2] Crompton, H. (2024). AI and English language teaching: affordances and challenges. British Educational Research Journal.

[3] Liu, Z., et al. (2025). SingaKids: A Multilingual Multimodal Dialogic Tutor for ... (ACL Industry 2025).

[4] Avouris, N. (2025). AI Conversational Tutors in Foreign Language Learning: A Mixed-Methods Evaluation Study. arXiv:2508.05156 — evaluation of conversational AI tutors for language learning.

[5] Ekizer, F. N. (2025). Exploring the impact of artificial intelligence on English language teaching: A meta-analysis. Acta Psychologica, 260, 105649 — meta-analysis of AI in EFL/ELT contexts.

[6] Huang, X., & Derakhshan, A. (2025). Predicting learner autonomy through AI-supported self-regulated learning. Learning and Individual Differences (online) — AI's impact on autonomy via self-regulation.

[7] Anonymous. (2025). AI-driven autonomous interactive English learning tutoring system. Journal of Language & Education Technology — autonomous interactive AI learning platforms.

[8] Peña-Acuña, B., & Durão, R. C. F. (2024). Learning English as a second language with artificial intelligence for prospective teachers: A systematic review. Frontiers in Education, 9.

[9] Jiang, R. (2022). How does artificial intelligence empower EFL teaching and learning nowadays? Frontiers in Psychology, 13. — overview of AI empowering EFL, including autonomy.

[10] Illés, É. (2012). Learner autonomy revisited. ELT Journal, 66(4), 505–513 — foundational theory of learner autonomy relevant to AI contexts.

[11] Benson, P. (2001). Teaching and Researching Autonomy in Language Learning. Longman — seminal work on learner autonomy theory.

[12] Vokhidova, T. (2025). Artificial intelligence in ESL pedagogy: Advancing learner autonomy and personalised instruction. International Journal of Artificial Intelligence, 5(07), 533–536 — ELT focuses on AI and autonomy.

[13] "Learner autonomy, learner engagement and learner satisfaction in multimodal CMC environments". Education and Information Technologies, 28, 14283–14323 — multimodal learning and autonomy.

[14] Ng Kok Wah, J. (2025). Artificial Intelligence in Language Learning: A Systematic Review of Personalisation and Learner Engagement. Forum for Linguistic Studies, 7(9), 327–341 — systematic review of AI in language learning.

[15] Integration of AI in language teaching (2025). EuroGlobal Journal of Linguistics and Language Education, 2(2), 89–98 — qualitative review of AI fostering autonomy.

[16] Lee, G., et al. (2023). Multimodality of AI for Education: Towards Artificial General Intelligence. arXiv:2312.06037 — multimodal AI in education framework.

[17] Chae, H., Kim, M., Kim, C., Jeong, W., Kim, H., Lee, J., & Yeo, J. (2023). TUTORING: Instruction-Grounded Conversational Agent for Language Learners. arXiv:2302.12623 — Intelligent conversational agents in language education.

[18] Advancing Education through Tutoring Systems: A Systematic Literature Review (2025). Liu, V., Latif, E., & Zhai, X. — Review of AI tutoring systems & multimodal interactions.

[19] Szabó, & Csépes (2023). Generative AI in ELT: autonomy, digital literacy and critical thinking (in ELT Journal) — discusses AI, autonomy, and pedagogy.

[20] Frontiers | The human touch in AI: self-determination theory and teacher scaffolding (2025). Frontiers in Psychology — AI personalisation, autonomy, and teacher roles

[21] (Optional foundational) D. Benson & P. Voller (Eds.) (1997). Autonomy and Independence in Language Learning. Longman — established literature on autonomy theory supporting AI contexts.