SOFTWARE ARCHITECTURE AND PEDAGOGICAL PRINCIPLES OF APPLYING MULTIMODAL ARTIFICIAL INTELLIGENCE TECHNOLOGIES IN THE TEACHING AND LEARNING PROCESS

Giyosova Gulnoza Alisher qizi
e-mail: giyasovagulnoza09@gmail.com
DOI:https://doi.org/10.5281/zenodo.21947200

Abstract. The paper substantiates the software and pedagogical foundations of introducing multimodal artificial intelligence — models that jointly process text, speech, images and video — into the educational process. A four-layer integration architecture is proposed, and the didactic principles that make such systems instructionally effective are derived from the cognitive theory of multimedia learning and the UNESCO AI competency framework for teachers. Risks of hallucination, bias, data protection and skill erosion are analysed.

Keywords: multimodal artificial intelligence, generative AI, educational software architecture, multimodal learning analytics, AI literacy, digital didactics, instructional design.

Relevance of the topic

The applied use of artificial intelligence in education has shifted from unimodal text chatbots to large multimodal foundation models able to interpret spoken language, handwriting, diagrams and video within a single dialogue. S. Küchemann and co-authors emphasise that this multimodality opens qualitatively new didactic possibilities — real-time feedback on handwritten mathematics, interpretation of graphs, barrier-free speech interaction and simulation of classroom situations for teacher training [1, p. 3].

The direction is supported institutionally: the Strategy for the Development of Artificial Intelligence Technologies until 2030, approved by Resolution of the President of the Republic of Uzbekistan No. PQ-358 of 14.10.2024, defines the development of human-resource potential and the AI competences of the population as a priority task [7]. In practice, however, AI is used spontaneously — through consumer chat interfaces, outside the learning management system and without instructional design. The result is a gap between the technological capability of the tools and their didactic effectiveness. The aim of the paper is to substantiate an approach in which the software architecture of a multimodal solution and the pedagogical model of its application are designed jointly rather than sequentially.

Psychological and pedagogical basis of multimodality

The value of multimodality is explained by the cognitive theory of multimedia learning: information is processed through two partially independent channels, verbal and visual, each of limited capacity, and meaningful learning requires active selection, organisation and integration of material [3]. Adding a modality is therefore productive only when it distributes the load between channels, whereas unjustified duplication of the same content increases extraneous cognitive load [4]. The leading design rule is that the modality is chosen by the didactic task, not by the technical availability of the function.

The second basis is multimodal learning analytics. A systematic review by M. Mohammadi and co-authors shows that combining digital traces with audio, video and behavioural data models collaboration, engagement and self-regulation far more fully than log data alone, while noting the weak methodological transparency of such studies [2].

Software foundations: a four-layer architecture

Layer 1 — modality capture. Speech, camera image, digital-pen strokes and screen content are captured and normalised through speech, optical and handwriting recognition. Streams resembling biometric data should be processed on the device where possible, and a latency budget of about one second maintained, otherwise the dialogue loses its scaffolding function.

Layer 2 — model and knowledge. A multimodal foundation model is combined with retrieval-augmented generation over an approved corpus of curricula, textbooks and assessment standards. This binds the answer to the national curriculum instead of the model's general statistical priors and makes each statement traceable to a source. For languages with limited digital resources, including Uzbek, localised corpora and fine-tuning are required, otherwise quality degrades unevenly across learner groups.

Layer 3 — pedagogical orchestration. This layer converts model capability into instruction: a learner model, prompt templates encoding teaching strategies (Socratic questioning, worked example with fading guidance, error analysis), adaptation of task difficulty, formative feedback and a teacher dashboard. Interoperability is ensured by educational standards — LTI 1.3, xAPI, Caliper and QTI — which prevents the solution from becoming an isolated application.

Layer 4 — governance and security. Role-based access, pseudonymisation of learner data, auditable logging of outputs and mandatory human confirmation for consequential decisions such as grading or placement. The regulatory context is already restrictive: Regulation (EU) 2024/1689 classifies AI used in education and vocational training as high-risk and directly prohibits the inference of emotions in educational institutions [6], which makes attention- or emotion-tracking unacceptable as a design goal even where technically feasible. Fault tolerance, a low-bandwidth mode and predictable cost per learner determine scalability to rural schools no less than model quality.

Pedagogical foundations of application

Six principles follow from the analysis: didactic primacy — the tool is selected after the learning outcome is defined, in accordance with constructive alignment; controlled multimodality — each additional channel is justified by the cognitive task; preservation of productive struggle — the system offers scaffolds rather than finished solutions, since over-reliance erodes independent problem-solving [1, p. 5]; transparency — learners are informed when they interact with AI and generated fragments are marked; a culture of verification — cross-checking model output is treated as a learning objective forming critical AI literacy; and teacher agency, in which the teacher remains designer, moderator and final authority.

For the last principle the UNESCO AI competency framework for teachers (2024) provides a ready structure — fifteen competencies across five dimensions (human-centred mindset, ethics of AI, AI foundations and applications, AI pedagogy, professional learning) at three levels: Acquire, Deepen, Create [5]. The principles also entail a redesign of assessment: where a model can reproduce a finished product, the object of assessment shifts to the process — drafts, oral defence, portfolios and in-class tasks with documented reasoning. Multimodal analytics must serve formative feedback and not surveillance. Professional-development programmes should therefore extend the TPACK model with an AI component covering the probabilistic nature of generative models and the limits of their applicability.

Conclusion

The main risks are documented: plausible but factually incorrect output, reproduction of biases contained in training data, leakage of learners' personal data, a widening digital divide and skill erosion through over-delegation [1, p. 4]. A separate limitation is the evaluation gap: controlled studies measuring learning gains from multimodal AI remain few compared with the volume of descriptive publications [2], which makes pilot testing with pre- and post-assessment a mandatory element of any roll-out.

Thus the effectiveness of multimodal artificial intelligence in education is determined not by the capability of the model but by the coherence of two designs — the software architecture and the didactic model. The proposed architecture localises the technical decisions and makes them verifiable, while the didactic principles define the conditions under which multimodality yields a learning gain rather than additional cognitive load. Implementation presupposes an institutional AI policy, retrieval binding of the system to the approved curriculum, a teacher competency programme aligned with the UNESCO framework, and a pilot stage with measurement of learning outcomes.

References

1. Küchemann S., Avila K.E., Dinc Y. et al. On opportunities and challenges of large multimodal foundation models in education // npj Science of Learning. – 2025. – Vol. 10. – Art. 11.
2. Mohammadi M., Tajik E., Martínez-Maldonado R. et al. Artificial intelligence in multimodal learning analytics: a systematic literature review // Computers and Education: Artificial Intelligence. – 2025. – Vol. 8. – Art. 100426.
3. Mayer R.E. Multimedia Learning. – 3rd ed. – Cambridge: Cambridge University Press, 2021. – 400 p.
4. Sweller J., van Merriënboer J.J.G., Paas F. Cognitive architecture and instructional design: 20 years later // Educational Psychology Review. – 2019. – Vol. 31, No. 2. – P. 261–292.
5. UNESCO. AI competency framework for teachers. – Paris: UNESCO, 2024.
6. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) // Official Journal of the EU. – 2024. – Art. 5, Annex III.
7. Resolution of the President of the Republic of Uzbekistan No. PQ-358 of 14.10.2024 "On approval of the Strategy for the Development of AI Technologies until 2030". – Tashkent, 2024.