

Learning with Multimodal Artificial Intelligence Tools

Introduction to the Basic Concepts of Artificial Intelligence

Author: Kacsur Annamária, ITStudy

Reviewed by: Hartyányi Mária, ITStudy

M1-T1-MLC6

Learning objective

The latest artificial intelligence solutions have moved far beyond text-based communication. Large language models (ChatGPT, Gemini, Claude) which originally were developed to generate text, can now not only generate increasingly high-quality images, video and audio, but also process digital information from any external source, such as sensors. Models with these expanded capabilities are collectively known as **multimodal artificial intelligence** systems.

This learning material explores the opportunities that multimodal artificial intelligence offers to education and how it goes beyond the multimedia learning materials that have become widespread in recent decades.

How do we learn?

Scientists have long studied the nature of learning, but we still only partly understand how young children learn and how the varied experiences they gain while exploring their surroundings will develop into coherent knowledge. When a young child plays with a ball, the sight, touch, movement and the associated words come together as a shared experience. Understanding emerges not from summarizing these impressions, but from recognising the connections between them.



How do babies learn? (Source: ChatGPT)

Learning takes place in a similar way in older children and adults. We do not perceive the world through separate “channels”: we see, hear, act, speak and read, interpreting these experiences by connecting them with one another and with our prior knowledge. This is consistent with the constructivist learning theory, which holds that learners do not simply receive information conveyed to them. Instead, they interpret it in the light of their own experiences and prior knowledge, establish connections and make sense of it for themselves.

In the 1990s, researchers increasingly explored how different modes of expression and representation support understanding. This gave rise to the theory of **multimodality**¹, and it was during this period that the concepts of **multimodal learning** and **multimodal pedagogy** first appeared in education, emphasising that knowledge can also be constructed by recognising and developing connections between different representations.

From multimedia towards multimodality

Language has long been the main medium of teaching, particularly the written text and speech: the teacher explains, while learners read, take notes and then demonstrate their knowledge orally or in writing.



Traditional education: the teacher explains; students read, take notes and then report orally or in writing (Source: ChatGPT)

As digital technologies have developed, **multimedia**—the combined use of text, images, audio, video and animation—has become increasingly accessible. A “**mode**” or “**modality**” refers to the form in which information is presented or the type of data entering a system. Examples include text,

¹ In the word multimodality, “multi” is a prefix of Latin origin meaning many or multiple; “modality” derives from the Latin *modus* (mode) and, in this context, refers to a mode of representation, communication or expression.

images, audio, video and data from external sensors, such as the daily average temperature recorded by a weather station.

Digital systems can work with:

- text;
- images and graphics;
- speech, music and other sounds;
- video;
- data describing movements and gestures;
- sensor data, such as measurements of position, temperature or speed.

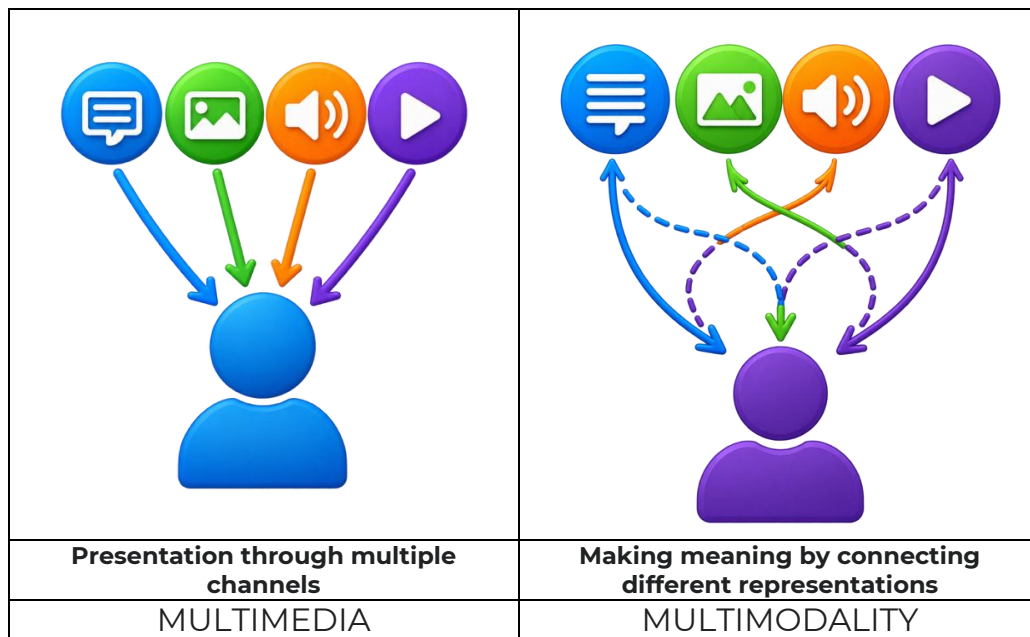
This does not mean that a modality is the same as a file format. JPEG and PNG, for example, are different image formats, but both belong to the same **visual modality**.



Multimedia in teaching: using several media together—information is presented side by side.

Experience over recent decades has shown that interactive educational CDs and colourful presentations enriched with images, animations and sounds support learning, but do not bring a qualitative improvement in understanding.

Most lessons today still focus primarily on text and speech (textbooks, whole-class explanations and notebooks).



If learners still mainly receive ready-made information, and connecting the meanings conveyed by different media in an explanation depends entirely on the teacher's initiative, the media elements communicate the same message without adding to understanding.

In a multimodal approach to teaching, learners not only receive information in several forms, but also compare, connect, transform and create different representations themselves. The emphasis thus shifts from conveying information through multimedia to learners' multimodal meaning-making.



Multimodal knowledge construction (Source: ChatGPT)

A function's formula, its graph and a story describing a real movement are not simply three different "formats" of the same learning material. Learners need to recognise how the changes described in the story appear on the graph and how they can be expressed through mathematical relationships. Understanding develops as learners recognise the connections between representations.

Multimodal artificial intelligence

Multimodal artificial intelligence accepts different data types (text, images, audio and video) as input and generates different types of output on request.



Multimodal artificial intelligence (Source: ChatGPT)

Multimodal artificial intelligence² can process and interpret several types of data—text, images, audio, video and more—within a single model. The term “multimodal” refers to the variety of input and output data that the model can handle.

It not only combines different media formats, but also “learns” the relationships between them

Areas of application

Healthcare

Artificial intelligence can analyse X-rays or MRI scans together with a patient’s written clinical notes, identify patterns by interpreting these sources together and use them to formulate a possible diagnosis.



Multimodal AI in healthcare (Source: ChatGPT)

Customer service and accessibility

Multimodal AI allows users with visual or mobility impairments to ask questions orally and request answers in the mode that suits them.

Content creation

² **Modality encoding:** each type of input data is first processed by a specialised encoder. The image encoder converts picture elements (pixels) into numerical data, the language encoder turns text into tokens, while the audio encoder extracts important features from speech or sound. These encoders capture the essential features of each modality while filtering out irrelevant details.

Multimodal integration (fusion): the model combines information from several modalities to form a unified representation of the input data. Multimodal foundation models usually achieve this integration through shared transformer layers and attention mechanisms, allowing information from one modality to influence the interpretation of another. For example, by interpreting a person’s facial expression together with their tone of voice and spoken words, the model can infer emotions or intent.

Models such as ChatGPT and Gemini use combined text and image input to modify an image or create new creative content.

Robotics and autonomous systems

Self-driving vehicles and robotic systems use multimodal artificial intelligence to integrate camera images and spatial data, build a picture of their surroundings and make navigation decisions in real-world conditions.

Document analysis

When given a document containing charts, diagrams and embedded images, multimodal artificial intelligence processes all the visual elements alongside the text and answers questions by interpreting them together.

Examples from vocational education and training

In an environmental project, learners upload a video showing the condition of a river and a document containing measurement results to an AI system. The AI helps establish the relationship between the phenomena visible in the video and the measurements, and the conclusions that can be drawn by examining the two sources together.



Multimodal AI in an environmental project (Source: ChatGPT)

In a history lesson, students can ask AI to answer their questions by analysing a photograph, a poster and a written historical source from the period together.



Multimodal AI in a history lesson (Source: ChatGPT)

In physics, learners can examine a video of a movement alongside the corresponding measurements and graph: does the graph match what we see in the recording, and how can the relationship be explained using the laws of physics?



Multimodal AI in a physics lesson (Source: ChatGPT)

Students in vocational education and training upload a photograph or video of a faulty component together with its technical documentation and measurement data. By analysing the sources together, AI can answer students' questions about the possible causes of the fault.



Multimodal AI in the classroom (Source: ChatGPT)

Takeaways for teachers

Today's artificial intelligence applications enable teachers to create striking images, videos, presentations and infographics quickly and easily. Developing learning materials becomes easier, their rapid obsolescence becomes less of a problem, and adapting them to learners' individual needs can become a reality.

Multimodal artificial intelligence can bring about a deeper change if we encourage learners to compare content they have created themselves or generated with AI, have it analysed, ask for feedback on their own ideas and request reasons for the answers they receive.

Multimodal AI opens up new possibilities for active learning. With teacher guidance, it can help learners become constructive participants in their own learning. The teacher's role is to design and create a learning environment in which multimodal AI prompts learners to think, interpret, compare and construct their own meaning. Today, these technologies can reduce the dominance of text-based information processing and bring education closer to the true nature of learning.

References

- Simamkele Matuntuta (2026), How AI is Transforming Vocational Education and Training, an article on cloudassess.com (last visited: 3/7/2026)
- Ayush Chauhan (2025): Multimodal AI: Beyond Text and Images, article on <https://www.techaheadcorp.com>,
- Kamalmeet Singh (2026): How Multimodal AI Works?, blog article on <https://medium.com>, last visited on 8/7/2026
- Amol Yadhav (2024): Multimodal AI: The Next Frontier in Personalized Education, blog article on <https://medium.com>, last visited on 8/7/2026

- Peraschi Selvan Subramanian (2025): Multimodal AI: The future of integrated intelligence, article published in World Journal of Advanced Engineering Technology and Sciences, last visited on 9/7/2026

Links

Multimodal AI <https://www.llamaindex.ai/glossary/multimodal-ai>

How do Multimodal AI models work? A YouTube video:

<https://www.youtube.com/watch?v=WkoytIA3MoQ>

Copyright © 2026 VITA Consortium

This work may be reused and adapted under the terms of the Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).



Funded by the European Union. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the National Agency.