EdSutra

Teaching guide

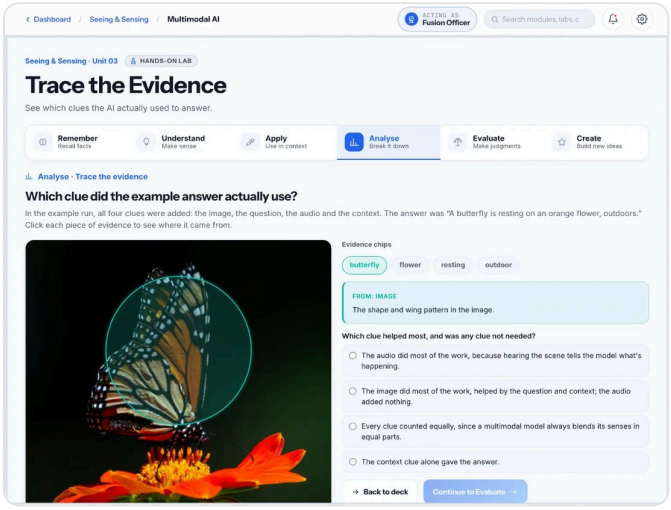
Module 02 · Seeing & Sensing · Unit 03 · Stage 2, Capabilities

Multimodal AI

The Fusion Deck

Grades 8–12 | 45 minutes | Role: **Fusion Officer**

Grade 8: About age 13–14 · Year 9 in the UK · Class 8 in India



The big idea

A multimodal AI combines different kinds of clue, such as a picture, a question, a sound and some context, into one answer. More clues are not always better. A careful user traces which clue each part of an answer came from, notices clues that added nothing, and judges whether the task needs several senses or just one.

The learner's job

Learners are the **Fusion Officer** at Cape Marrow Station · Fusion Deck. The station's AI answers from a picture, a question, a recording and a note. Nobody has checked which of them it actually uses. Their task: put the signals together. Say which one each part of the answer came from.

By the end, learners can say

- I can name the kinds of input a multimodal AI can combine: image, text, audio, video and context
- I can combine clues in the Fusion Deck and see which ones an answer needs
- I can trace each part of an answer back to the clue it came from, and spot a clue that added nothing
- I can judge when combining signals helps and when one signal is enough for the task
- I can design a multimodal pipeline: its senses, its goal, and what each sense could get wrong

Before class

- One computer or tablet for each learner or pair, with EdSutra open (this Module is part of the full EdSutra studio).
- Open the studio once on each device before class, so the Unit still runs if the internet drops.
- Play Remember to Apply yourself, and run the Fusion Deck once with the Context clue and once without it.
- Tell learners the Fusion Deck is scripted: no AI model runs, so the lesson is in tracing the clues.
- Best after Seeing & Sensing and Uni-modal AI.
- Units open in order, and Module 1 comes first. To start a class here, switch on Open every Unit in Settings → Teachers & parents.

Words to introduce

Multimodal: Combining more than one kind of input, such as an image plus text.

Fusion: Putting several clues together to reach one answer.

Evidence: The part of an input that a piece of the answer came from.

Context: Background information about a situation, such as "outdoor scene".

Visual question answering: Answering a question about a picture.

Confidence score: A model's own estimate of how sure it is, not a guarantee.

Pipeline: The planned path from an AI's senses to its goal.

Running the lesson 45 minutes

Open 5 min

"An AI gets a photo, a typed question, a voice recording and a short note. Does it really use all four to answer?"

Before anyone clicks, ask: which of the four clues will matter most for the answer, and will any clue add nothing? Tally the votes on the board and come back to them in Analyze.

Step	What learners do	Ask the class	Listen for
1. Remember 4 min	Tap the five clue cards (image, text, audio, video and context) to see an example of each kind of input a multimodal AI can mix.	• Which of these five kinds of clue does your phone take in? • What counts as context?	Context is background information about the situation, such as "outdoor scene" or "taken at night".
2. Understand 5 min	Click through how AI grew from one kind of input to many: mono-modality, unimodal, multimodal, generative multimodal. Then read how today's tools turn text, images, audio and video into answers.	• What is the difference between unimodal and multimodal? • Name a tool that takes a picture and a question together.	Multimodal means different kinds of information working together, not just more of the same kind.
3. Apply 8 min	In the Fusion Deck, drag clues into the AI Fusion Core: an image and a question are required, audio and context optional. Press Run Fusion, then read the answer and its evidence.	• What changed in the answer when you added or removed the Context clue? • Did adding the audio change anything?	"Outdoors" appears only with the Context clue. Adding the audio changes nothing in the answer.
4. Analyze 7 min	Click each evidence chip (butterfly, flower, resting, outdoor) to see which clue it came from and where in the photo. Then say which clue helped most and whether any clue was not needed.	• Where did "resting" come from? • Look back at our votes on the board. Which clue really mattered most?	The image did most of the work, with the question and context helping. The audio added nothing.
5. Evaluate 8 min	Compare two example answers: Answer A uses only the image ("An animal."), Answer B uses image, text and context. Choose a verdict and the reason that backs it. The numbers are illustrative.	• When would Answer A be good enough? • Does a confidence of 0.93 mean it is 93% sure to be true?	It depends on the task. A confidence score is the model's own estimate; models can be confidently wrong.
6. Create 8 min	Design a multimodal pipeline: pick two or more senses (image, text, audio, video, context, sensors), a goal, and at least one caution that fits the senses chosen. Then submit it.	• Which caution did you pick, and which sense is it about? • Why is "with several senses it can't be wrong" refused?	Each sense can mislead in its own way, and two senses can disagree, so every pipeline needs cautions.

Steps follow Bloom's revised taxonomy, in the order the Unit presents them.

Deepen the thinking

Critical thinking	An AI answers a question about a photo and sounds very sure. How could you check which clue it used before you trust it?
AI thinking	For a robot that helps in a kitchen, which senses would you combine, and which one could you leave out? Why?
Ethics and fairness	An app describes photos aloud for people who are blind. What harm could a confident wrong description do, and who should explain its limits?
Connect to life	When did you work something out by putting two clues together, such as a sound and something you saw?

Misconceptions to listen for

"More signals always give a better answer."

In the Fusion Deck the audio added nothing. Evaluate refuses "more clues always make a better answer": it depends on the task.

"The AI reads the picture like a person does."

Analyze shows the answer built from specific evidence: regions of the image, words in the question, and the context chip.

"A high confidence number means the answer is true."

A confidence score is the model's own estimate, not a guarantee. Evaluate refuses reading 0.93 as 93% certain to be true.

In real life the card learners see at the end of the Unit

- Some AI helpers take a photo plus your question, like "What plant is this?", and answer from both.
- TV weather forecasts blend satellite pictures with readings from weather stations: different kinds of clues, fused into one answer.
- A maps app combines your typed search with context, like where you are and the time, to pick results.

Try it this week. Describe a photo to someone using only words, then show it. Which details did the picture add, and which did your words add?

Watch out for. More clues don't always mean a better answer. Ask which clue each part of an answer came from, and whether a clue added anything at all.

Talk about it. When you're unsure what's happening, which do you check first: what you see, hear or read?

Exit ticket

1. What makes an AI multimodal?

   _____

2. In the Fusion Deck, which clue added nothing to the answer?

   _____

3. Name one task where a single clue is enough.

   _____

Going further

The Computer Vision Lab has both camera and microphone blocks (camera optional; a demo mode runs without one). Learners open an example such as Gesture Reader, change its blocks and wires, and name the sense each part of the result depends on.