DAY 19 · Part 2 · How We Got Here

How do photos and sounds get to stand next to words? — AI that sees and hears

Five-Minute AI Lessons · Sep 25, 2026

Chop photos and sounds into pieces and line them up, and they can stand right where words do.

In a nutshell

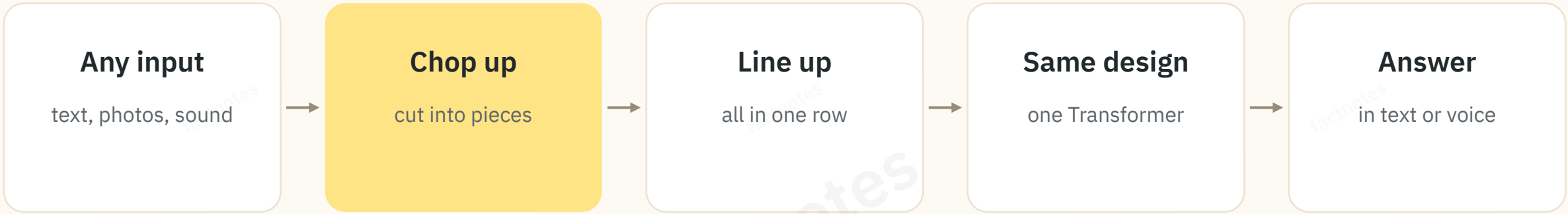
1. The secret to AI seeing photos and hearing sounds wasn't giving it eyes and ears. Photos and sounds got **cut into small pieces** too, and lined up in the same row as word pieces.

2. Text gets cut into word pieces, photos into square tiles on a grid, and sound into **very short slices of time**. Once they're cut, all three are just 'pieces in a row.'

3. That means **there's no need to build a new design**. You can feed them straight into the Transformer, which lets pieces look at each other. Because it handles several senses at once, it's called multimodal.

In one picture

Whatever goes in takes the same road



A model that learned all three together from the start skips the step of turning photos into text, so it's faster and more accurate.

Key idea: it didn't get eyes and ears — it turned everything into pieces and lined them up in the same row

Split text, photos and sound into pieces

Put in text

Cut into word pieces

School | **lunch** | is | **quesad** | **illas** | !

Text gets cut into words or parts of words. 'Quesadillas' doesn't stay one chunk — it can be cut into two or th...

Put in a photo

Cut into grid squares

sky | sky | **cloud** | roof | **window** | wall | road | **person** | road

A photo gets cut into a grid, like a checkerboard. Give each piece a number for which square it is, and you can...

Put in sound

Cut into tiny time slices

0.0s | 0.1s | **0.2s** | **0.3s** | 0.4s | **0.5s**

Sound gets chopped into very short slices along time. At this point all three look the same — text, photos and...

Tap the buttons to switch what goes in. If you can see that whatever you put in ends up as pieces standing in a...

Try it yourself in the web lesson · next-token.net/en/learn/day-19/

Quick check

Just three questions

Q1

What actually happens when you put a photo into an AI?

① The photo gets translated into text first
② It's turned black and white
③ It's cut into a grid and the pieces lined up
④ It's made twice as big

Q2

Why can text, photos and sound all go through one design?

① The model was given eyes and ears
② Photos were text all along
③ Sound has no meaning
④ All three become pieces in the same row

Q3

How does sound get split up?

① Short time slices
② One letter at a time
③ By color
④ Sorted by length

Know the why, too

Q1

③

It's cut up like a checkerboard, and each piece gets a number for which square it is. Then it looks just like word pieces.

Q2

④

Once cut, all three are 'pieces in a row.' So the Transformer, which lets pieces look at each other, works as is.

Q3

①

Sound flows along in time, so it gets chopped into very short slices of time. Each slice is one piece.

Chop photos and sounds into pieces and line them up, and they can stand right where words do.



Tomorrow: Everyone's model race — how the places making these models went from a handful to dozens.