



ARTIFICIAL INTELLIGENCE AND NEURAL NETWORKS

Lecture 9b – Vision-Language Models and Multimodal Reasoning



Chizhi Chris ZHANG

zhangchizhi@ciomp.ac.cn

Advanced Computing and Digital Technology Research Center

University of Chinese Academy of Sciences

Spring 2026

Today's Question

与数字工程研究中心

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

What we are trying to answer

How do neural systems connect what they see with what they say so that an agent can read screens, understand images, follow visual instructions, and act with context?

Why this lecture matters

AI is no longer only about text. Many useful systems must read slides, charts, forms, maps, interfaces, and photos before they can reason well.

What changes from AI9

AI9 focused on the workflow loop of an agent. NN9 focuses on the neural machinery that lets such an agent perceive the world instead of only reading text.

From AI9 to NN9

计算与数字工程研究中心

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Last time

We asked how agents plan, use tools, remember results, and continue across several steps.

Today

We ask what happens when the task includes screenshots, diagrams, documents, photos, charts, and interface state that cannot be reduced to plain text in advance.

One sentence

AI9 was about acting across steps. NN9 is about seeing before acting.



Why Text Alone Is Not Enough

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Text-only systems miss visual evidence

They cannot directly inspect a chart trend, check whether a button is visible, compare two diagrams, or notice that a warning icon appears on the screen.

Practical consequence

If the task depends on layout, appearance, or spatial relations, text alone is often an incomplete description of the real problem.

This is why many modern workflows feel artificial when they rely on manually transcribing visual content back into words.



Why Vision Alone Is Not Enough

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Seeing is not the same as understanding

A model may detect objects or text correctly and still fail to answer what matters, because the real task usually includes goals, rules, and language-based constraints.

Typical gap

Recognizing a chart is easier than answering, “Does this figure support the claim in the report, and why?”

Vision tells us what is there. Language reasoning helps decide what it means in context.



A Campus Example

数字工程研究中心

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A realistic student task

“Look at this course registration screenshot, tell me which requirement is missing, and explain what I should do next.”

What the system must combine

Interface reading, text recognition, layout understanding, course-rule interpretation, and a clear answer in ordinary language.

Why this is a good example

It is not enough to recognize words on the screen. The model has to connect what it sees with what the user is actually trying to solve.

Why Multimodal Models Matter

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

What they make possible

- image and document question answering
- chart and diagram explanation
- interface reading for agents
- grounded visual instructions

The larger shift

Once models can combine visual evidence with language, they move closer to how people actually solve tasks in the real world.

Shared Representation Intuition

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Core idea

Map images and text into a shared representation space so that related image-text pairs land near each other and unrelated pairs stay farther apart.

What this gives us

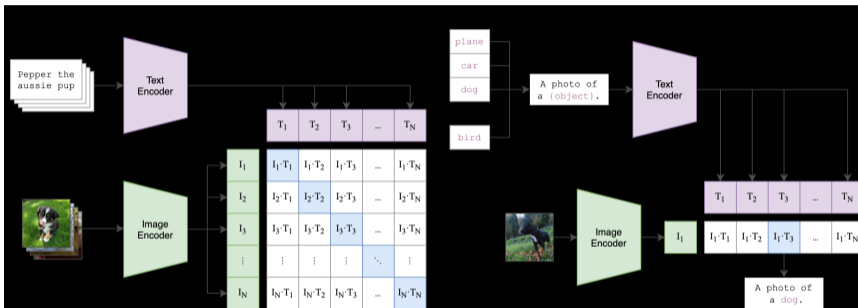
Retrieval, zero-shot recognition, matching, captioning, and later multimodal instruction following all become easier once the two modalities can talk in the same internal language.

This is one of the main conceptual steps that turned vision-language modeling from a collection of separate tricks into a more unified direction.



A CLIP Snapshot

先进计算与数字工程研究中心
ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER



How to read the picture

An image encoder and a text encoder are trained together so matching pairs become similar in one shared space.

Contrastive Intuition

数字工程研究中心
ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Simple form

$$\mathcal{L}_{\text{contrast}} = -\log \frac{\exp(\text{sim}(\mathbf{v}_i, \mathbf{t}_i)/\tau)}{\sum_j \exp(\text{sim}(\mathbf{v}_i, \mathbf{t}_j)/\tau)}$$

Read in words

Make the correct image-text pair more similar than the incorrect pairs in the batch.

The formula matters only because it captures the teaching signal: pull matches together, push mismatches apart.



From CNN to ViT

先进计算与数字工程研究中心

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

CNN strength

Convolutions bring a strong local bias and work very well when nearby pixels matter most.

ViT strength

Patch tokens plus attention make it easier to model longer-range relations across the whole image.

The shift to transformers in vision helped unify visual and language processing around a token-based view.



Patch Embedding Intuition

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Pipeline

Image \rightarrow patches \rightarrow linear projection \rightarrow token sequence.

Why this is convenient

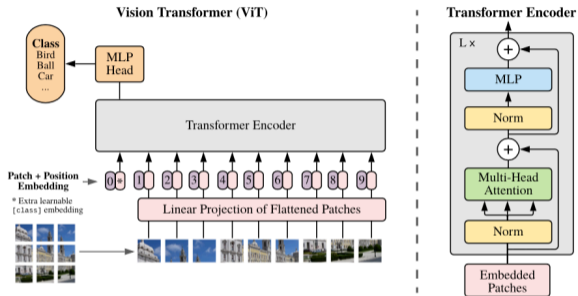
Once an image becomes a token sequence, many transformer tools developed for language can be adapted for vision as well.

This is not saying images and sentences are the same. It says a common token interface makes joint modeling much easier.



A ViT Architecture View

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER



What the diagram is emphasizing

The model chops the image into patches, turns those patches into embeddings, and lets attention model their relationships.

From Patches to Visual Tokens

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

What the encoder produces

Not a sentence, but a set of visual tokens that summarize local content and spatial relations across the image.

Why token count matters

Higher visual resolution usually means more tokens, and more tokens usually mean more memory and compute cost.

This is one reason multimodal systems often face the same scaling tradeoff that language models face with long context.



Three common patterns

- early fusion: mix modalities sooner
- late fusion: process first, combine later
- cross attention: let one modality guide the other

What changes across designs

Where you fuse the information affects cost, flexibility, and how strongly language can stay grounded in visual evidence.

Vision Tokens Meet Language Tokens

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A common design

Visual encoder outputs are projected into a space the language model can consume, so the decoder can answer in normal language while still using visual evidence.

Why projection quality matters

If this bridge is weak, the model may sound fluent while relying only loosely on what the image actually contains.

Many grounding failures are really failures of this connection between seeing and speaking.



A Modern VLM Example

工程研究中心
ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

LLaVA (llama 2)



What is unusual about this image? Can you explain this to a 5-year-old kid?

In the image, there is a man standing on top of a yellow taxi, which is unusual because people typically do not stand on top of cars, especially when they are driving. It is also unusual to see a person having clothes on top of a car, as having is usually done indoors or in a designated area. This scene is quite unique and might be considered humorous or intriguing to a 5-year-old kid. It is important to explain that this is not a typical situation and that people should not attempt to stand on top of cars or wear clothes in such a manner.

Can you explain this in Chinese?

In the image, there is a man standing on top of a yellow taxi, which is quite unusual because people usually do not stand on top of cars, especially when they are driving. Additionally, the man is having clothes on top of the car, which is not a common activity. This scene is quite unique and might be considered humorous or intriguing to a 5-year-old kid. It is important to explain that this is not a typical situation and that people should not attempt to stand on top of cars or wear clothes in such a manner.

LLaVA (llama 1)



What is unusual about this image? Can you explain this to a 5-year-old kid?

In this image, a man is sitting on top of a washing machine, which is attached to the back of a yellow taxi. This is unusual because it is not common to see people sitting on top of washing machines, especially while they are being transported on the back of a taxi. It is a funny and unexpected scene that might catch the attention of a 5-year-old kid.

Can you explain this in Chinese?

在这张图片中，一个人正坐在一台洗衣机的顶部，这台洗衣机被固定在一辆黄色出租车的后部。这是一个不寻常的场景，因为人们通常不会坐在洗衣机的顶部，尤其是在它们被运输的过程中。这是一个有趣且令人意外的场景，可能会吸引5岁孩子的注意力。



Why systems like this matter

Instruction tuning makes multimodal outputs much more useful for real users because the model learns not only to perceive, but also to answer in a usable way.

Multimodal Prompting

数字工程研究中心
ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Typical ingredients

- visual input
- language instruction
- output format or constraint

A practical lesson

Even strong models perform better when the question is clear about what part of the image matters and what form the answer should take.

Good prompting still matters because multimodal models do not magically know which visual detail the user cares about most.



Reading Interfaces and Screens

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Agent value

Many useful agent tasks start with a screen: a form, a menu, a dashboard, an error page, or a configuration panel.

What the model must do

Read visible elements, understand layout, connect them to the user goal, and then suggest or execute the next step carefully.

This is one of the clearest reasons multimodal models matter for practical automation.



Document and Chart Understanding

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Typical task

Read a report, slide, invoice, table, or chart, then answer a question that depends on both visible structure and language meaning.

Why this is harder than OCR

It is not only about extracting words. The model must also understand layout, numerical relations, titles, axes, and the task behind the question.

That is why many business and public-service use cases need multimodal reasoning instead of plain text pipelines.



Retrieval Helps Multimodal Tasks Too

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Why retrieval still matters

An image alone is often not enough. Manuals, policy documents, prior cases, and diagrams can provide the missing context that turns visual recognition into useful reasoning.

A concrete example

The model sees a machine warning light, but the correct action depends on the maintenance manual, the recent repair history, and the current operating rule.

A Multimodal Agent Stack

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Perception

Read pixels, layout, icons, and visible state.

Reasoning

Connect what is visible to the task, the policy, and the next action.

System wrapper

Add memory, permissions, logging, and recovery around the model.

Where NN9 fits

The vision-language model supplies perception and grounding, but the full system still needs planning and control around it.

Why Screens Are Harder Than Photos

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A screenshot is a mixed object

It may contain icons, tiny text, menus, layout cues, hidden state, and interface conventions that only make sense when the model also understands the task.

Why agents care

To use software safely, an agent often has to answer not just “What is visible?” but also “Which visible element matters right now, and what would a wrong click do?”

This is one reason multimodal agents are harder than generic image captioning. They need task-aware perception, not only description.



A Service Counter Example

Scenario

A citizen uploads a damaged photo of a document and asks whether it is sufficient for an application.

What the system should do

Inspect the image, read visible fields, compare them with policy requirements, explain the problem clearly, and decide whether the case should be escalated.

Why this is a serious example

An answer that sounds confident but misreads one field can cause wasted time, unfair rejection, or loss of trust.

Grounding Failure

与数字工程研究中心

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Typical failure

The model describes objects, text, or relations that are not actually present in the image.

Typical fixes

Better supervision, grounded retrieval, confidence-aware answers, and designs that force the model to point to the evidence it is using.

This is the multimodal version of hallucination, and it matters even more once actions depend on visual claims.



Why Evaluation Is Hard

先进计算与数字工程研究中心
ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A score can hide a lot

One benchmark number may combine perception, reading, reasoning, and format compliance, which makes it hard to see what actually failed.

What good evaluation asks

Did the model see the right thing, interpret it correctly, answer the right question, and behave safely in context?

If we do not separate these questions, improvement becomes guesswork.



Failure Analysis Matrix

先进计算与数字工程研究中心

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Three common failure types

- perception error: wrong object or text extraction
- reasoning error: correct observation but wrong conclusion
- action error: good answer but unsafe or wrong next step

Why this framing helps

Different failures need different fixes. More data might help one case, while better guardrails or human review may help another.

Safety Before Action

The key principle

Perception confidence and action permission should not be treated as the same thing.

A better habit

Even when the model gives a plausible answer, the system should still ask whether the task is high-stakes, irreversible, or sensitive before allowing action.

Why this matters

A visually grounded mistake can still trigger a harmful real-world decision if the system is allowed to act too quickly.

When Humans Must Stay in the Loop

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Typical triggers

- low visual confidence
- ambiguous document state
- medical, legal, or financial stakes
- irreversible actions

Goal

Let the model help with perception and explanation, but keep accountability with people when the consequences are large.

What to Watch When Building

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Three design questions

- What visual evidence is actually needed for the task?
- How will the system show uncertainty and ask for help?
- What should be checked before any downstream action is allowed?

Good engineering instinct

Use the simplest multimodal design that solves the task reliably, and spend the saved complexity on evaluation and guardrails.

Why NN9 Leads to AI10

先进计算与数字工程研究中心

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

What NN9 added

We now have models that can combine language with visual evidence and support more realistic agent behavior.

What AI10 asks next

If systems can read documents, inspect screens, and help make decisions, how should society govern them and where should responsibility stay human?

Once models can see as well as speak, the technical question naturally expands into a social question.



Summary

先进计算与数字工程研究中心

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

- Multimodal models matter because many real tasks depend on visual evidence, not only text.
- Shared representation learning helped connect images and language in one usable framework.
- Vision transformers and fusion mechanisms make it possible to combine visual tokens with language reasoning.
- Practical value appears in interfaces, documents, charts, and grounded assistance for agents.
- Reliability still depends on grounding, evaluation, permission design, and human escalation when stakes are high.



Where the course is going

We have now seen agents that can plan and models that can see.

Next lecture

We will widen the lens and ask what frontier AI means for trust, governance, institutions, work, and public life once these systems become more capable and more widely used.



Thank You

