上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Lecture 28: Vision-Language Models

Tao Huang

John Hopcroft Center, School of Computer Science, Shanghai Jiao Tong University

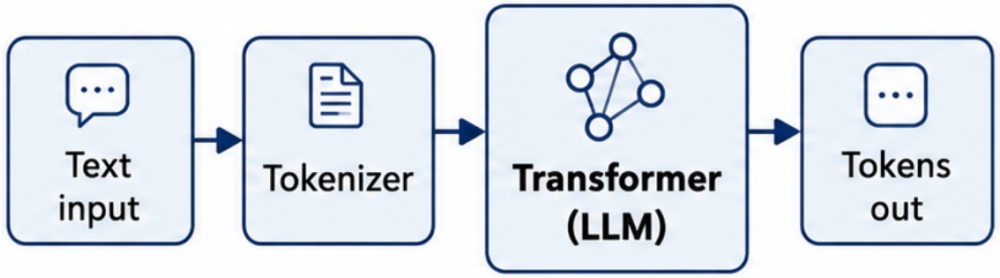
https://taohuang.info/cs3317

https://oc.sjtu.edu.cn/courses/89538

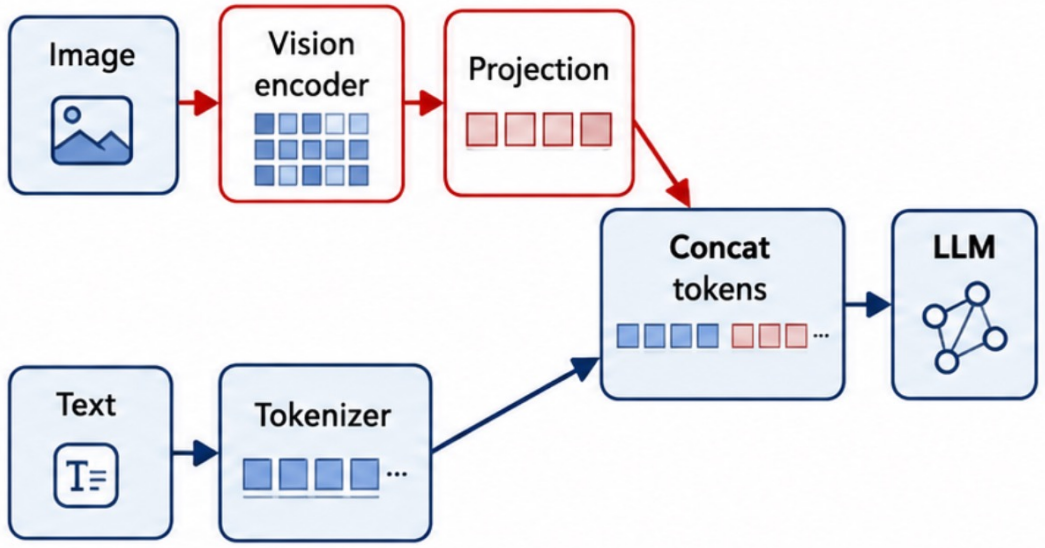
AI tools assisted in generating some figures in these slides. All such content has been reviewed, and the instructor is responsible for its accuracy.

From Reading to Seeing

- L27 made a model that can read.
- This lecture: a model that can see.



One modality. One loss.



Two modalities. New interface problem.

The Core Question

- *Transformers are token-native. Images aren't tokens.*
- *What is the right interface between pixels and a sequence model?*

Today we trace three answers — each forced by a different constraint:

1. Aligned embeddings

(CLIP)

for retrieval

& zero-shot classification

2. Frozen modules + a learned bridge

(BLIP-2, Flamingo, LLaVA)

for instruction-following

VLMs

3. Native multimodal

(Qwen-VL, GPT-4V, Gemini)

when retraining

from scratch is affordable

Objectives

By the end of this lecture, you should be able to:

- Explain why image–text alignment requires a different objective from supervised ImageNet training.
- Compare the three VLM architecture eras (aligned-embedding, frozen-bridge, native) and identify which constraint forced each.
- Sketch the LLaVA architecture and the role of the projection layer.
- Estimate the visual-token budget for a 224×224 image through ViT-L/14, and explain why token compression matters.
- Diagnose a VLM hallucination as a pretraining, alignment, or visual-grounding failure.

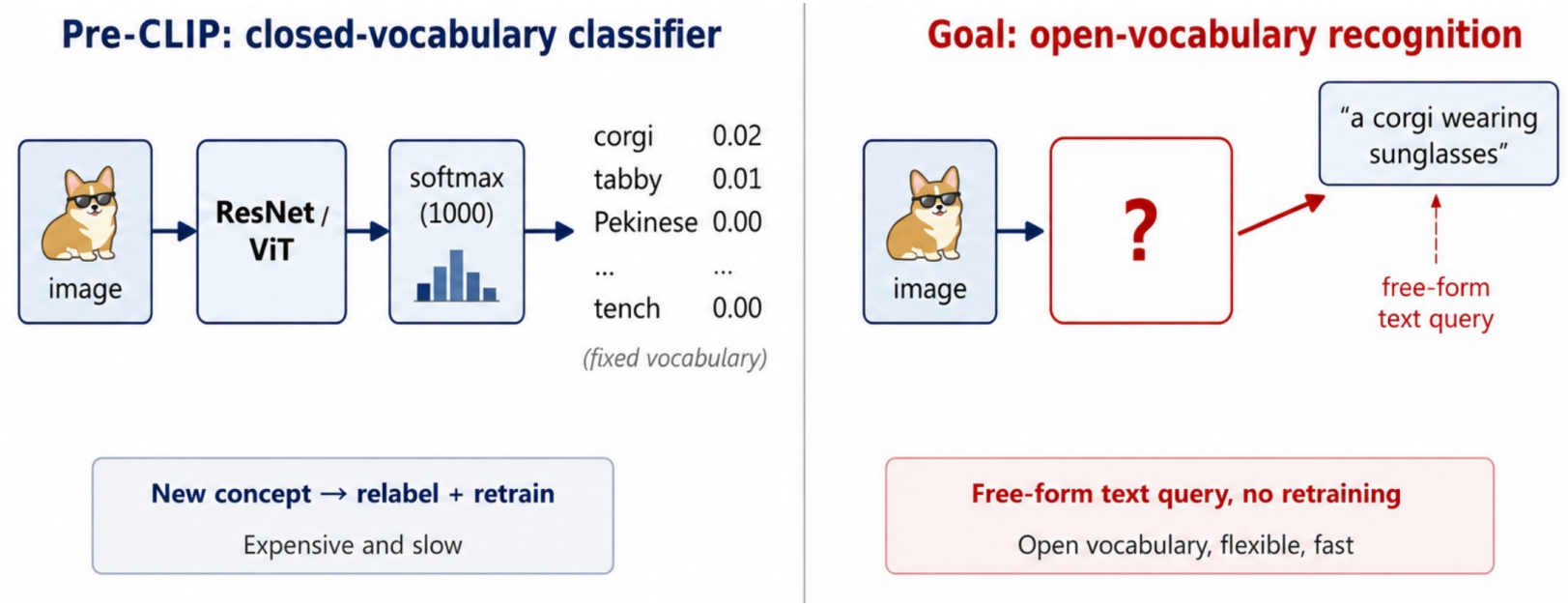
1. Aligning Pixels and Text

Situation: Pre-CLIP Computer Vision

Situation: Through ~2020, vision models were trained as *closed-vocabulary classifiers* — ImageNet → 1000 fixed classes.

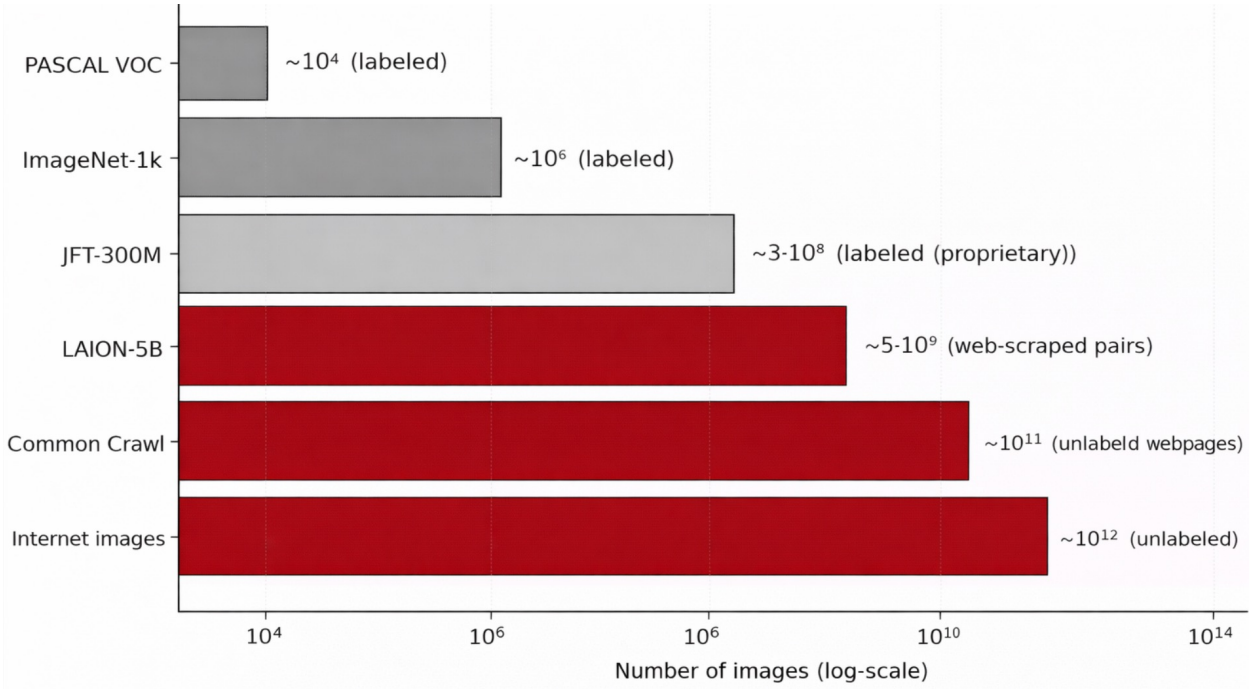
Complication: To recognize a new concept (*"a corgi wearing sunglasses"*), you must collect labels and retrain.

How do we get a vision model that knows concepts it was never explicitly trained on?



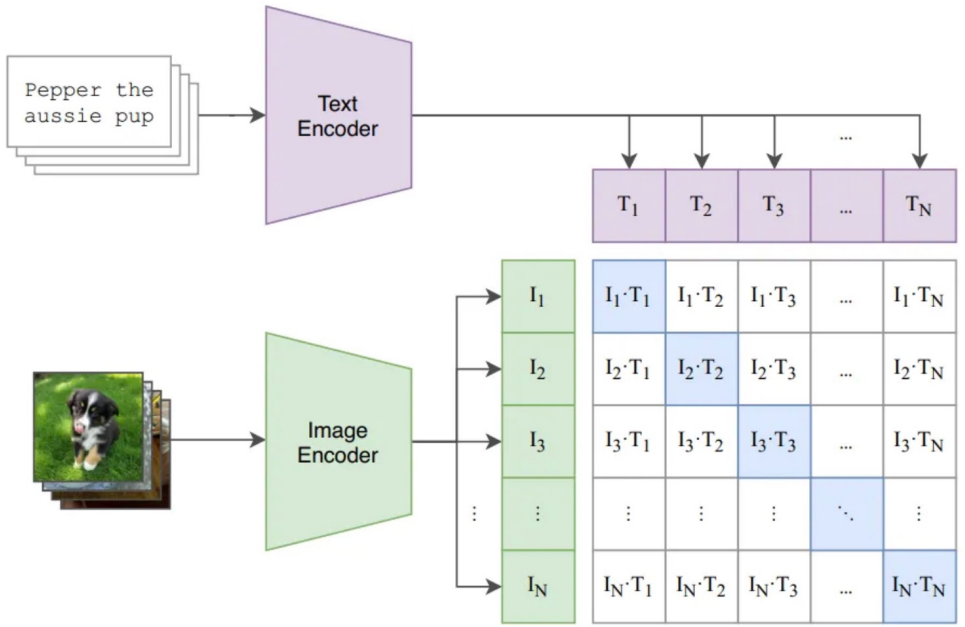
Complication: Labels Are Expensive, Captions Are Free

- ImageNet: 14M images, hand-labeled, ~1000 classes — years of human work.
- The internet: billions of (image, caption) pairs, free. WIT-400M (Radford 2021).
- Captions describe images in open-ended natural language — no fixed vocabulary.



Insight: if we train on (image, caption) pairs directly, we inherit the open vocabulary of language.

CLIP Recap (from L19)



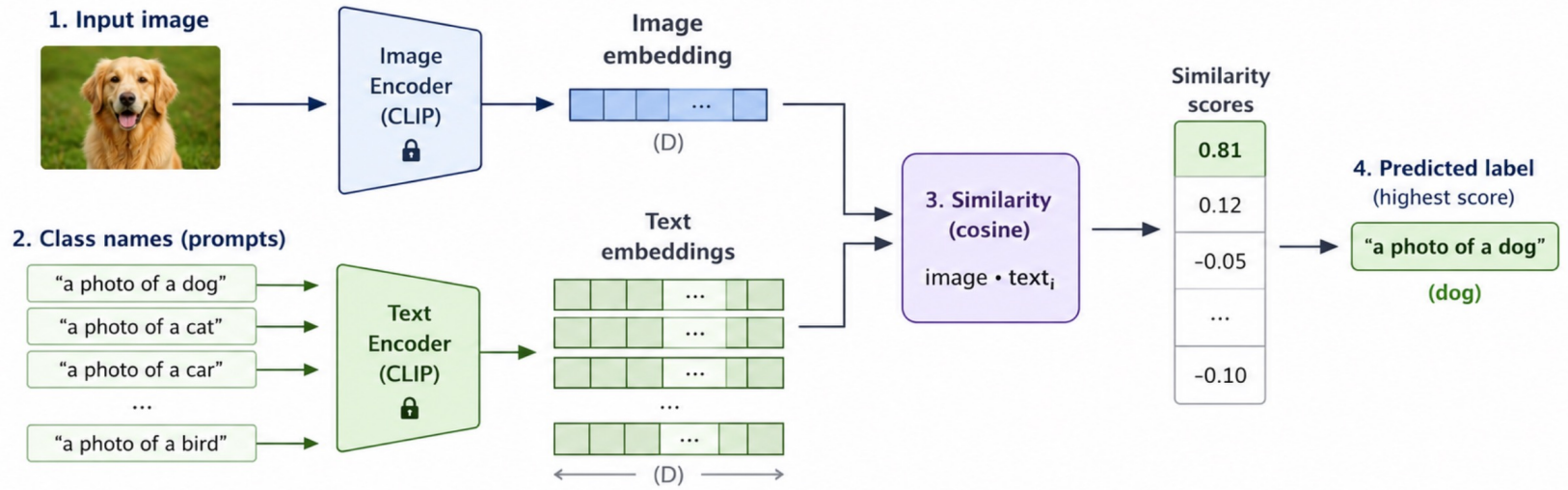
- Two encoders: image (ViT) and text (Transformer) — produce embeddings in a shared space.
- Objective: symmetric InfoNCE — pull paired (image, caption) close, push unpaired apart.

What CLIP Buys: Zero-Shot

Answer: no retraining for new classes — write them as text.

- For each candidate class y, compute similarity (image, “a photo of a {y}”).
- Predicted class = argmax over y.

One pretrained model handles ImageNet, CIFAR, dog breeds, satellite imagery, medical images — by changing the prompt.



	aYahoo	ImageNet	SUN
Visual N-Grams	72.4	11.5	23.0
CLIP	**98.4**	**76.2**	**58.5**

What CLIP Can't Do

- CLIP gives you alignment, not reasoning. Three things it can't do:

*For VQA and instruction-following, we need a model that doesn't just align — it **talks**.*

2. From Alignment to Conversation

Situation: We Already Have Two Strong Models

- L17: vision encoders (ViT, CLIP-ViT) — encode an image into a sequence of patch features.
- L27: pretrained LLMs — read tokens, write tokens, follow instructions.

Naive idea: stitch them together. image → vision encoder → ??? → LLM → text response.

What goes in the ???

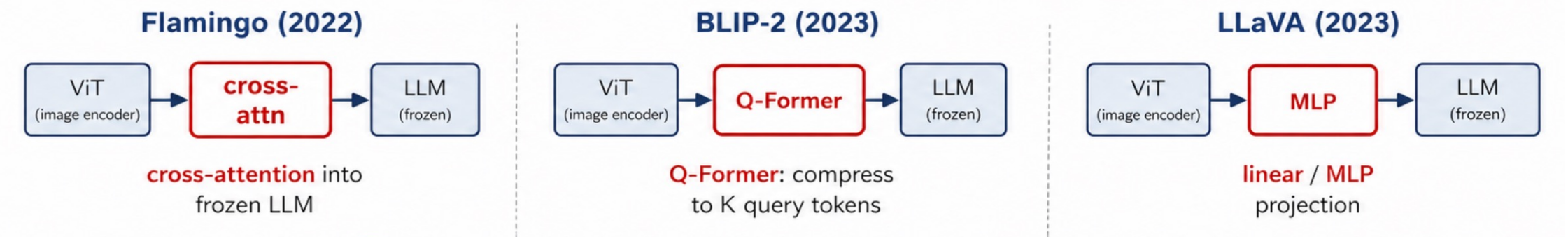
Complication: Different Languages

- ViT-L/14 outputs: 256 patch features × 1024-dim each.
- LLM input: token embeddings × 4096-dim each.

Two mismatches: (1) dimensionality 1024 ≠ 4096 (2) distribution — vision features live on a different manifold

The Three Bridge Designs

Three answers proposed in 2022–2023:

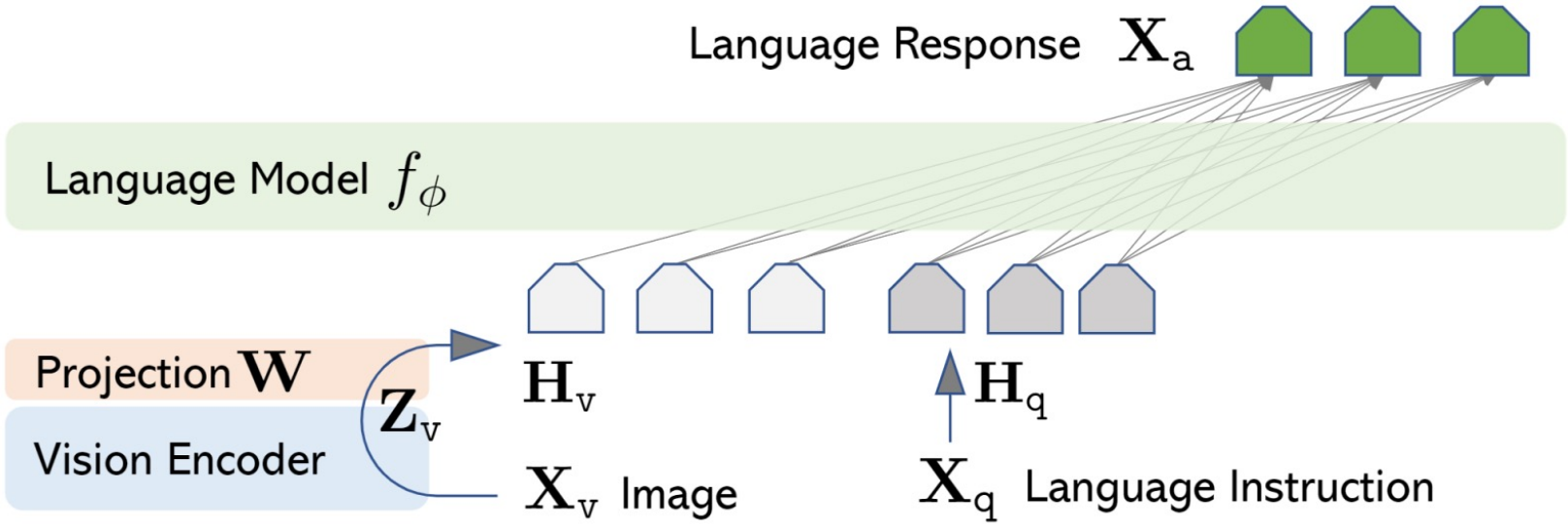


- **Flamingo** (Alayrac 2022) — gated cross-attention layers inserted into a frozen LLM.
- **BLIP-2** (Li 2023) — Q-Former compresses to a fixed K query tokens, then projects.
- **LLaVA** (Liu 2023) — a 1- or 2-layer MLP. The simplest design — and it wins.

LLaVA: The Surprising Winner

Claim: a 2-layer MLP projection is enough — if you train it right.

- CLIP-ViT-L/14 (frozen) → MLP projector (trained) → LLM (Vicuna, fine-tuned).
- Visual features become "tokens" — concatenated to text as a prefix.
- The LLM treats them as a foreign language it must learn to read.



LLaVA Training Recipe — Why It Works

Stage 1 — Feature alignment.

- Freeze ViT and LLM. Train only the MLP. ~600K image-caption pairs.
- Goal: teach the projector to map ViT features into the LLM's input distribution.

Stage 2 — Instruction tuning.

- Unfreeze the LLM. Train on ~150K GPT-4-generated multimodal instruction examples
  ("describe this image", "what's wrong here?", "explain step-by-step").

Why it works: the LLM was already a strong reasoner. Once vision is mapped into its input space, all of L27's instruction-following capability transfers automatically.

Think: How Many Visual Tokens?

Computing the visual-token budget.

- A 224×224 image, ViT-L/14 (patch size 14):

$$\#\text{tokens} = \left(\frac{H}{P}\right) \times \left(\frac{W}{P}\right) = \frac{224}{14} \times \frac{224}{14} = 16 \times 16 = 256$$

- LLaVA-1.5 supports 336×336 → 24×24 = 576 visual tokens per image.

Two questions:

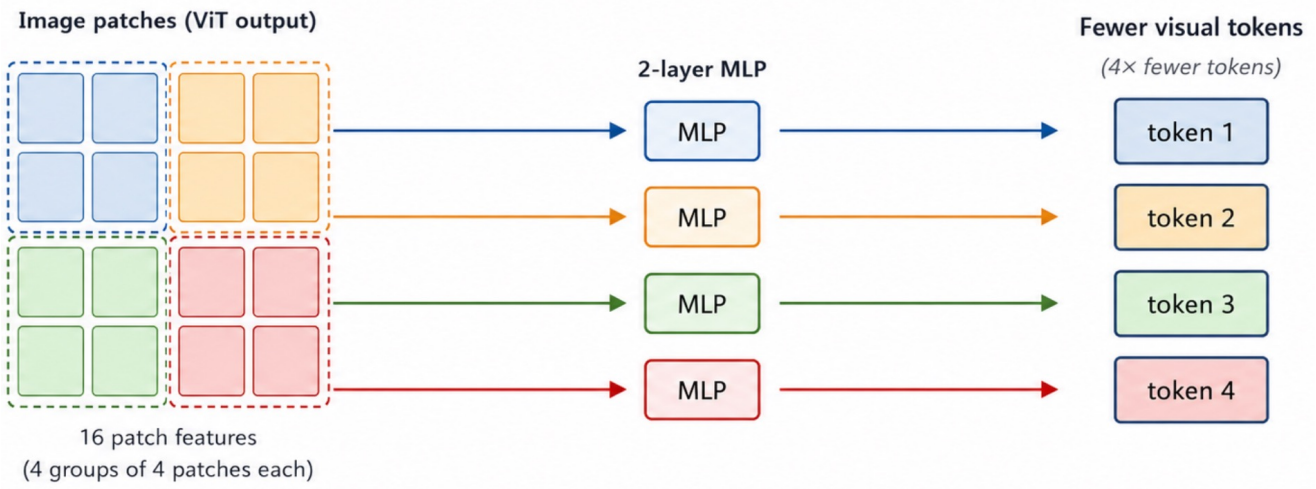
1. If your LLM has a 4K context window, how many images fit before you're out of room for the user's text? 4000 / 576 ≈ 7
2. For a 1-minute video at 1 fps — how many tokens? 576 × 60 = 34560

Token Compression

Two strategies in current systems:

- **Pixel-shuffle / pooling** (Qwen2-VL, InternVL): reduce 4 patches → 1 token. 4× fewer tokens, modest quality drop.
- **Q-Former-style learned compression** (BLIP-2): K query tokens, regardless of image size.

Trade-off: tokens vs. fine-grained spatial detail.



Group neighboring 2×2 patches and map each group to a single token via an MLP.

Where VLMs Fail: Hallucination

VLMs hallucinate more than text-only LLMs. Three diagnosable failure modes:

1. Visual-grounding failure

- Describes an object that isn't there. Often: model defaults to plausible-sounding completions when the image is ambiguous.

2. Counting / spatial-reasoning failure

- "How many people?" answered with a plausible round number, not the actual count. ViT features encode *what's there*, not *how many*.

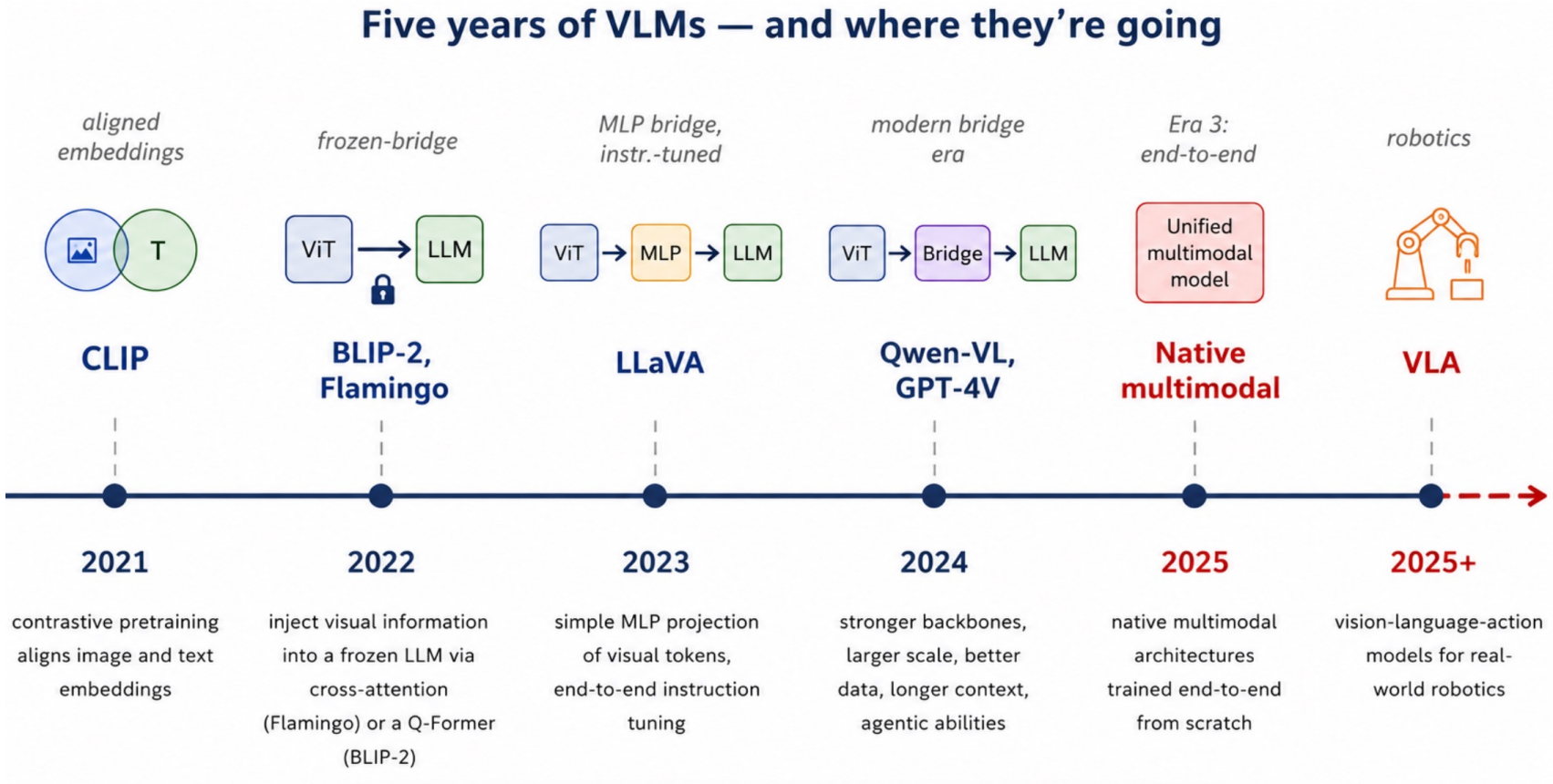
3. OCR / resolution failure

- Small text in image is below patch resolution; model invents text it can't read.

The Frontier

Era 3 — Native multimodal.

- Don’t bridge — train a single transformer on image + text + audio tokens from scratch (or from an LLM init, with vision unfrozen end-to-end).



Summary

- Image–text alignment (CLIP-style contrastive) gives open-vocabulary recognition by replacing labels with captions.
- The interface problem — pixels into a token-native model — drove every VLM design.
- LLaVA's MLP projection works because the LLM is already a strong reasoner; we just need to make vision features legible.
- Token budget is the practical bottleneck for image and especially video VLMs.
- Hallucination is the dominant failure mode — pretraining, alignment, or visual-resolution.

Summary of Module 5

Module 5 (L23–L28) was about modeling distributions over different domains:

- discrete tokens (autoregressive, LLMs)
- continuous latents (VAE)
- implicit distributions (GAN)
- continuous data via score (diffusion)
- joint multimodal distributions (VLMs — today)