

The Evolution of MLLMs

From Text to Unified Multimodal Intelligence (2018–2026)
(Optional)

AI with Deep Learning
EE4016

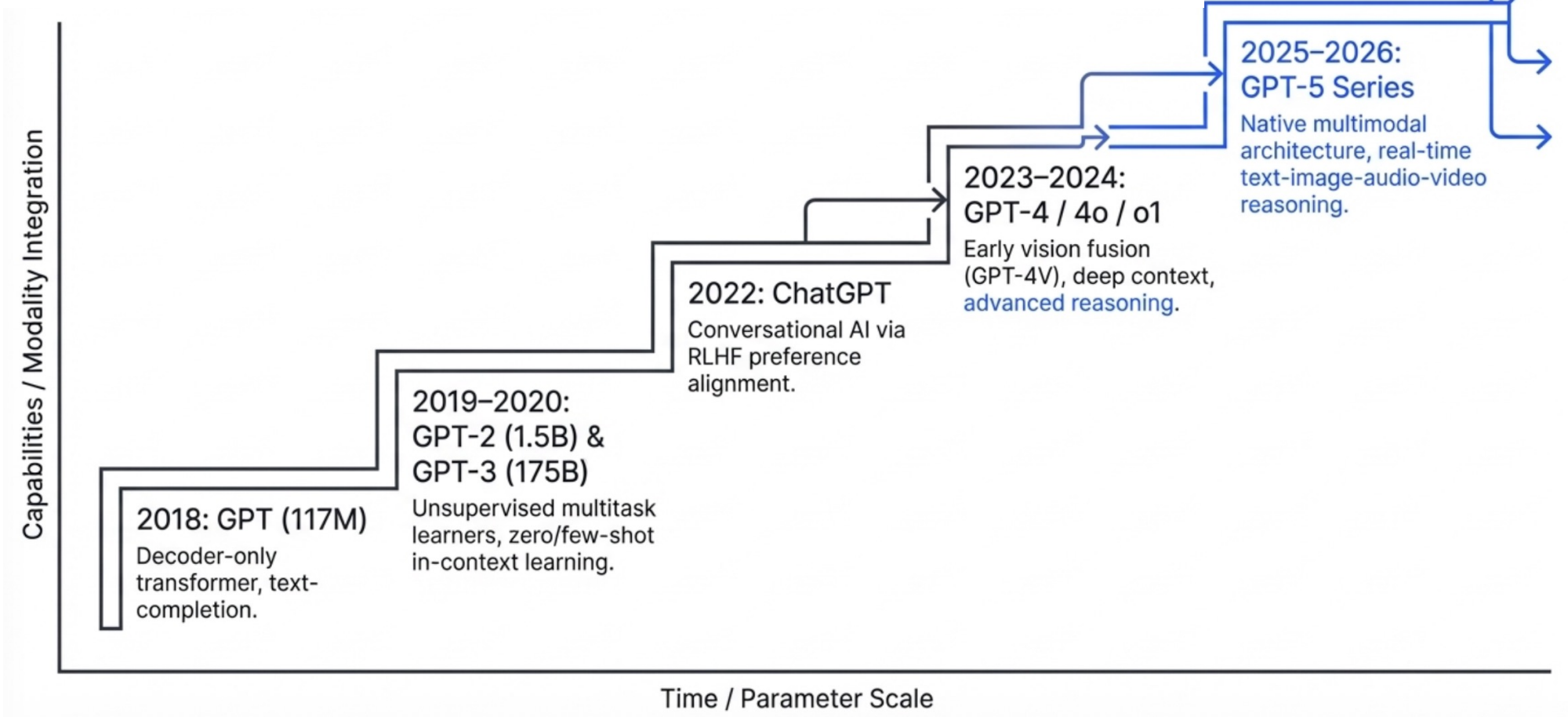
Prof. PO Lai-Man

Department of Electrical Engineering
City University of Hong Kong

What We'll Cover

- **Foundational Era (2018–2020):** Dual Encoders
- **Bridging Era (2021–2022):** Shared Latent Spaces
- **LLaVA Paradigm (2023):** Visual Instruction Tuning
- **Native Multimodality (2024):** Scaling & Integration
- **Unified Intelligence (2025–2026):** Any-to-Any Generation
- **Specialized Applications & Real-World Impact**
- **Challenges & Future Directions**

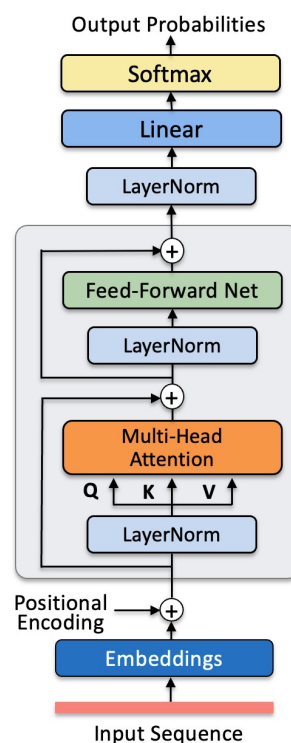
The GPT Lineage: Foundation to Frontier (2018-2026)



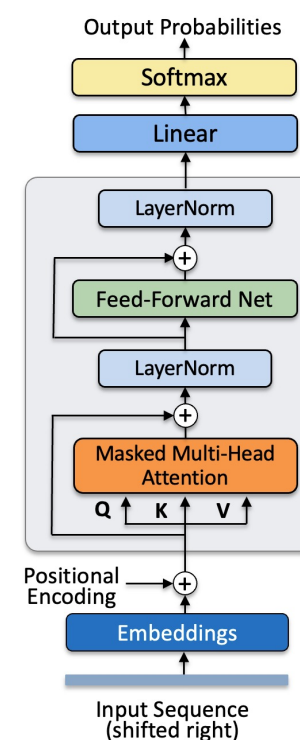
The Text Paradigm: BERT & Early GPT Variants

- **Key Achievements:**
 - Established transformer as dominant architecture
 - Demonstrated emergent reasoning through scaling
 - **✗** Remained "blind" to non-textual data
- **Impact:**
 - Set the stage for multimodal extension but highlighted the need for vision integration

BERT (2018)



GPT-1,2,3 (2018-20)



Foundational Era (2018–2020)

The Dual-Encoder Foundation for Text and Image

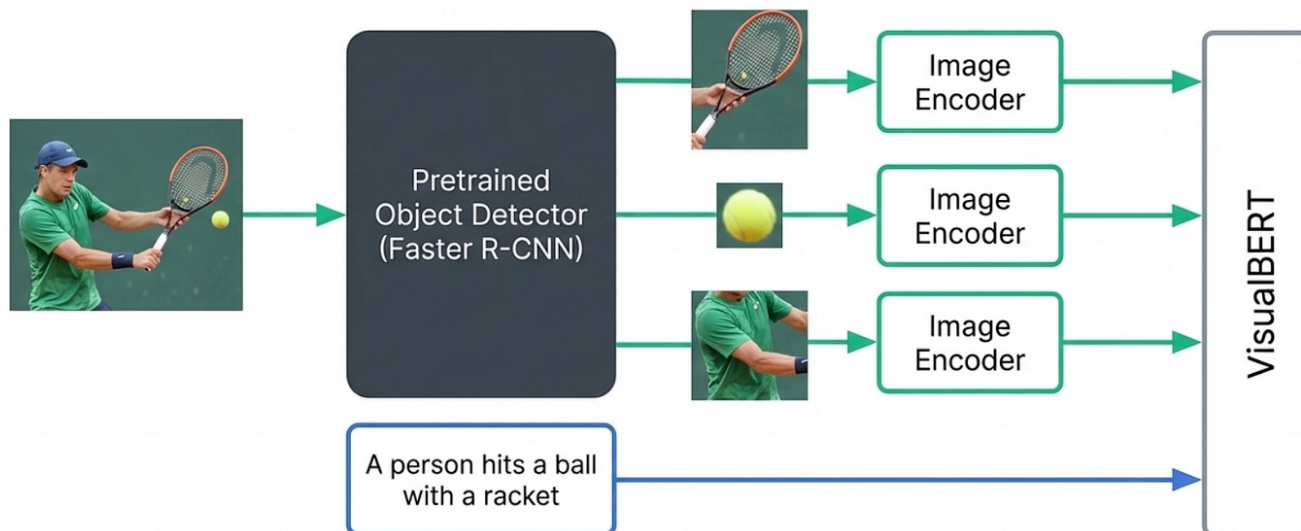
What exactly is a Dual-Encoder (or Dual-Stream) architecture?

- A dual-encoder model has two completely separate neural networks (“encoders”) that process each modality independently:
 1. Vision Encoder (“the eyes”)
 - Takes raw pixels or object regions (usually extracted by Faster R-CNN or similar detector).
 - Outputs a sequence of visual feature vectors/embeddings.
 - Early models used CNNs; later ones started experimenting with transformers.
 2. Text Encoder (“the brain for words”)
 - Usually a BERT-style transformer that turns text into embeddings.

Early Fusion Models: VisualBERT, LXMERT, UNITER (2019-2020)

Architecture:

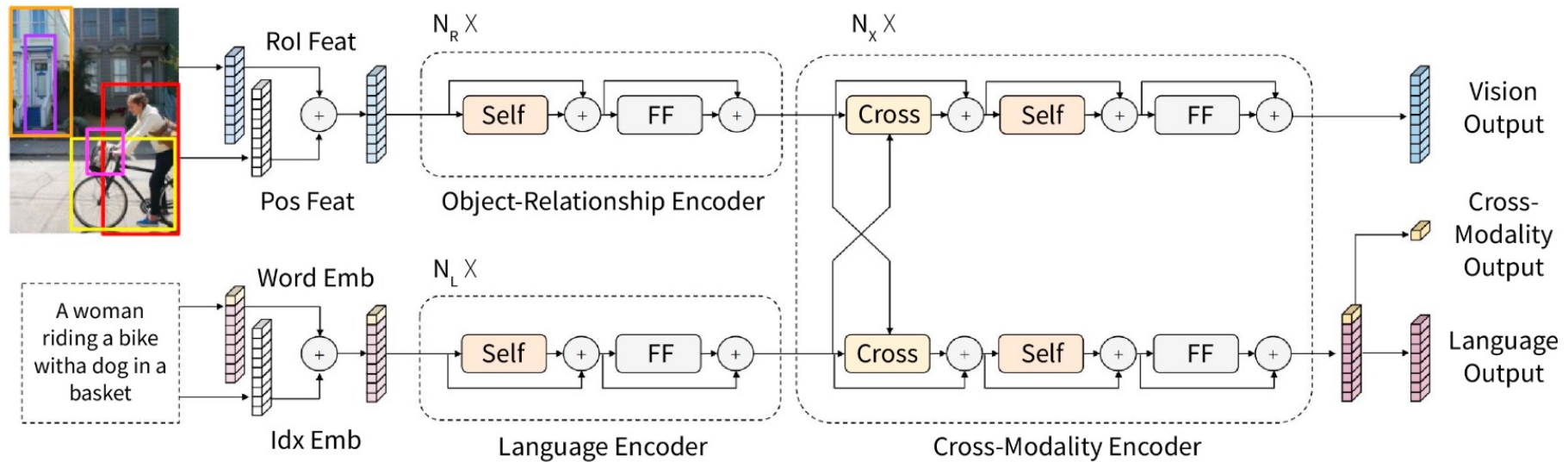
- Dual-stream or cross-attention encoders
- Region-based features from object detectors
- Limited parameter scale



Example: VisualBERT (2019)

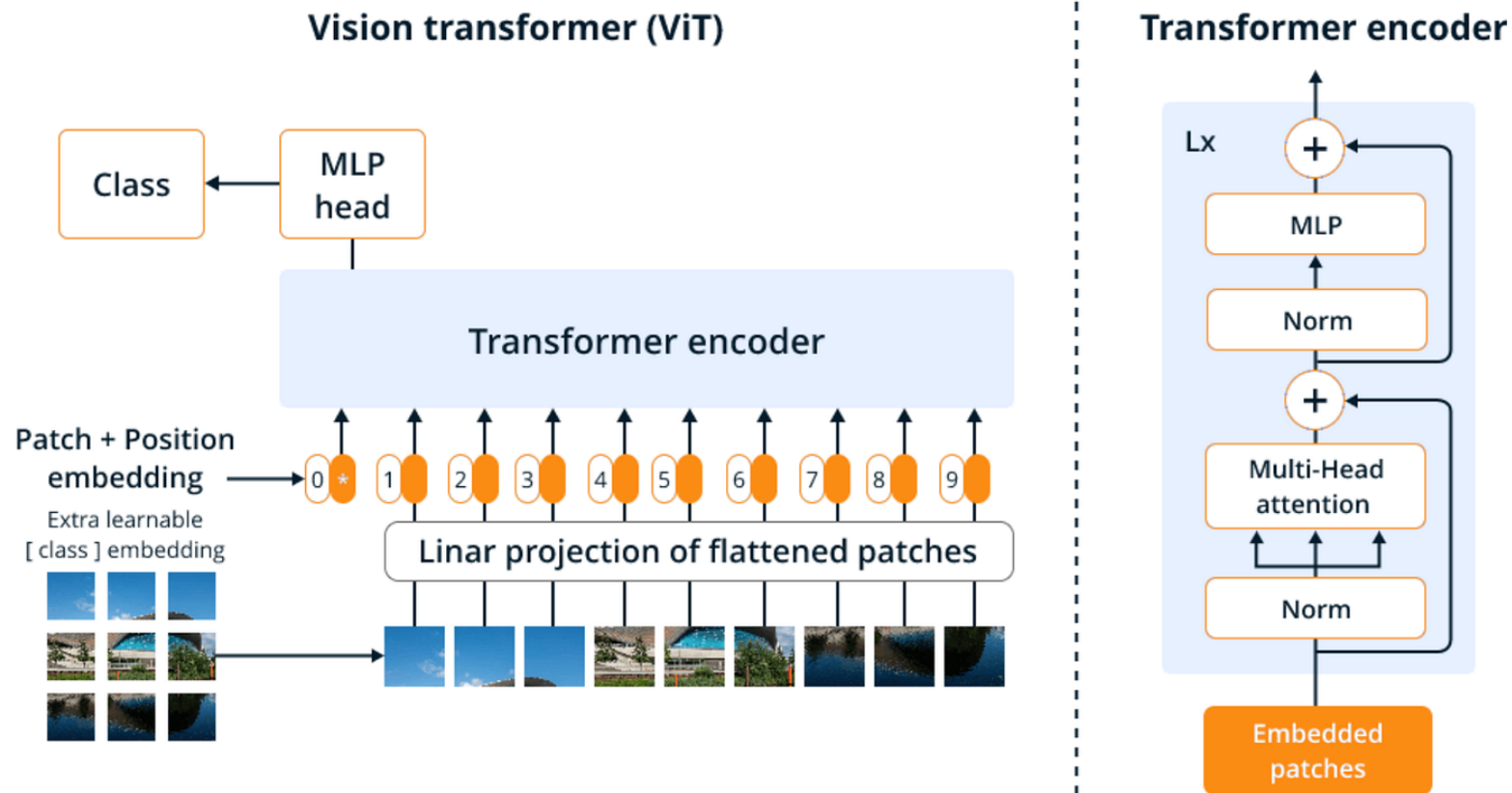
- Feeds text + image regions into BERT
- Discovers internal alignment via self-attention
- **Limitation:** No general instruction-following

LXMERT-style dual-stream architecture



You can clearly see the two separate encoders on the left (Object-Relation Encoder for vision + Language Encoder for text), then a Cross-Modality Encoder in the middle that lets them exchange information via cross-attention. This is exactly the kind of model the article highlights (LXMERT 2019, UNITER 2020).

Vision Transformer (ViT, 2020)







ViT: The Architectural Bridge

- **How ViT Works:**

1. Decomposes image into fixed-size patches (16×16 pixels)
2. Linearly projects each patch into vector embedding
3. Adds positional encodings
4. Processes through standard transformer encoder

- **Why It Mattered:**

-  Architectural Uniformity with LLMs
-  Global Context via self-attention
-  Scalability following LLM scaling laws
-  Patch-Based Representation maps to LLM tokens

Bridging Era (2021–2022): Shared Latent Spaces

- **The Breakthrough:**
 - Development of the "Rosetta Stone" for multimodal AI
 - Align different data types into shared mathematical space
 - Enable cross-modal understanding
- **Key Models:**
 - CLIP (2021)
 - Flamingo (2022)
 - BLIP-2 (2022)

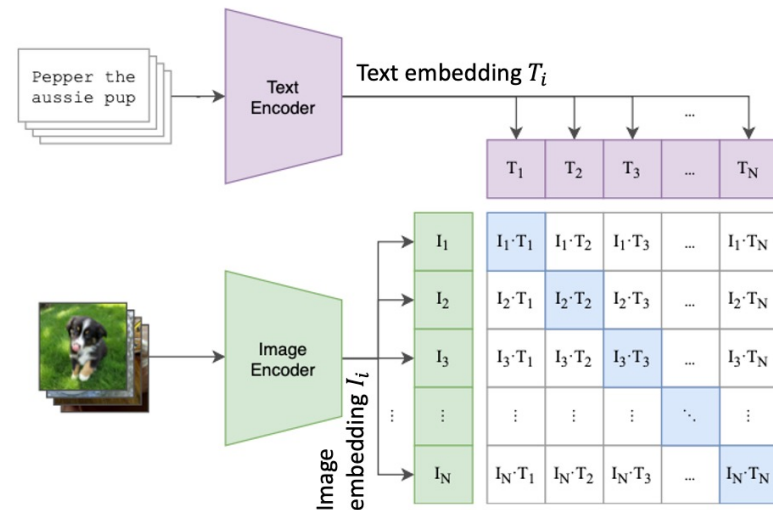
CLIP (2021): Shared Latent Spaces

Concept Overview: In 2021, OpenAI's CLIP solved cross-modal alignment by mapping different data types into the exact same mathematical space.

The Mechanism: Contrastive Language-Image Pretraining (CLIP) jointly trains an image and text encoder.

The InfoNCE Loss: The model learns to predict the correct pairings of (image, text) batches, **pulling true pairs together** and **pushing false pairs apart**.

Result: CLIP became the de facto visual backbone for nearly all subsequent multimodal language

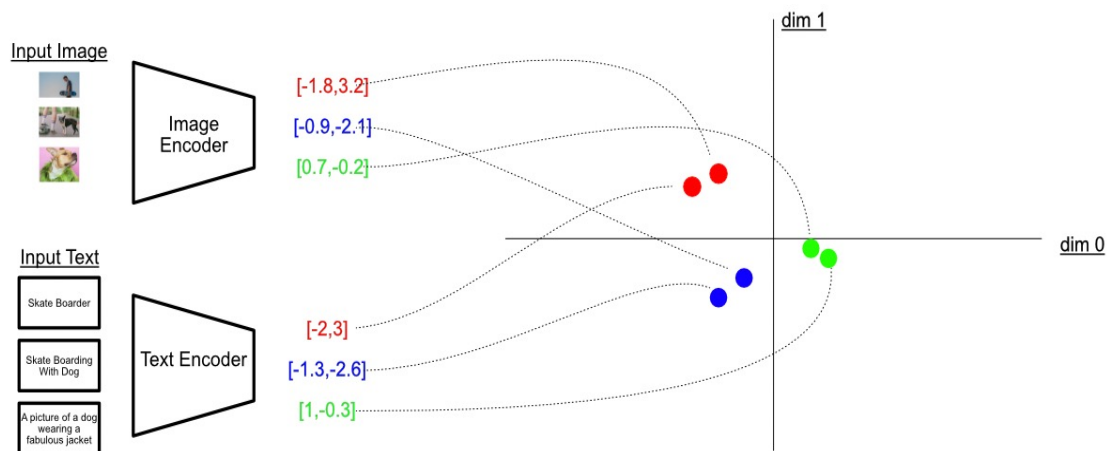


Training Objective

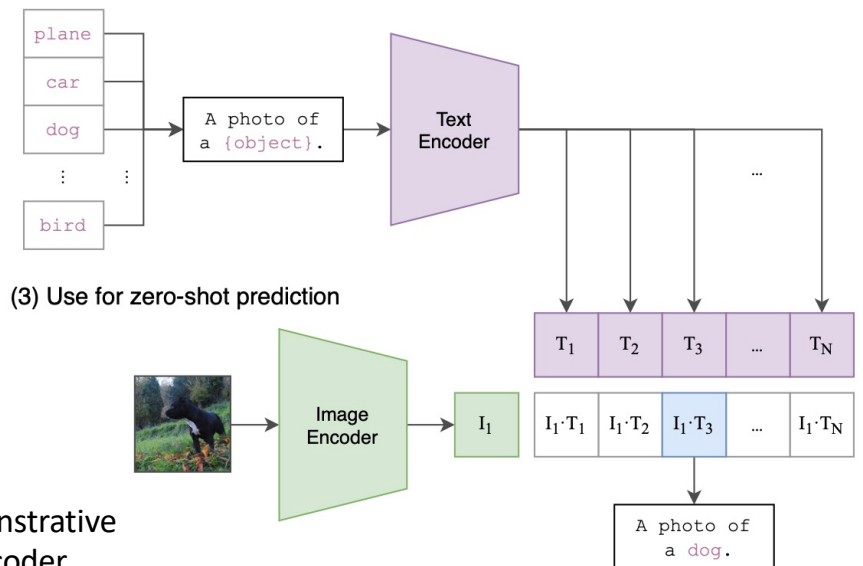
$$\min \left(\sum_{i=1}^N \sum_{j=1}^N (I_i \cdot T_j)_{i \neq j} - \sum_{i=1}^N (I_i \cdot T_i) \right)$$

Dataset and Contrastive Pretraining

WebImageText (WIT) Dataset: CLIP was trained using a freshly assembled dataset comprising **400 million pairs of images and corresponding text** sourced from the Internet.



(2) Create dataset classifier from label text



- An example of CLIP after it's trained, with 2-dimensional embeddings for demonstrative purposes. The image encoder summarizes images into vectors, and the text encoder summarizes text into a vectors.
- Notice how, once the encoders are trained, positive pairs end up close together.

CLIP: Innovation and Impacts

- **Innovation:**
 - Aligned visual features and textual concepts
 - Used InfoNCE loss function
 - Trained on 400M image-text pairs
- **Impact:**
 - Became the **de facto visual backbone** for nearly all subsequent MLLMs
 - Enabled zero-shot transfer to new tasks
 - Created shared embedding space for text and images

Flamingo (2022): First True MLLM-Scale Model

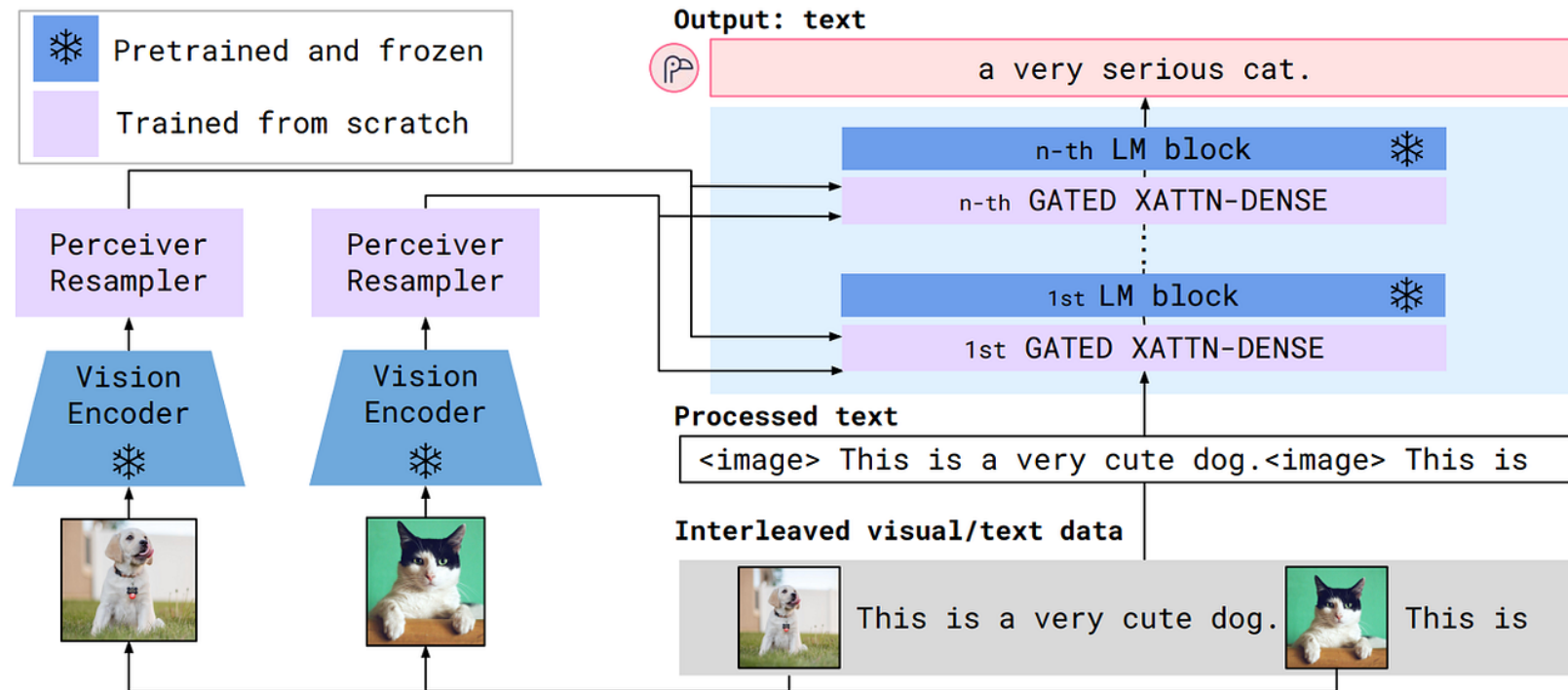


Figure 3: **Flamingo architecture overview.** Flamingo is a family of visual language models (VLMs) that take as input visual data interleaved with text and produce free-form text as output.

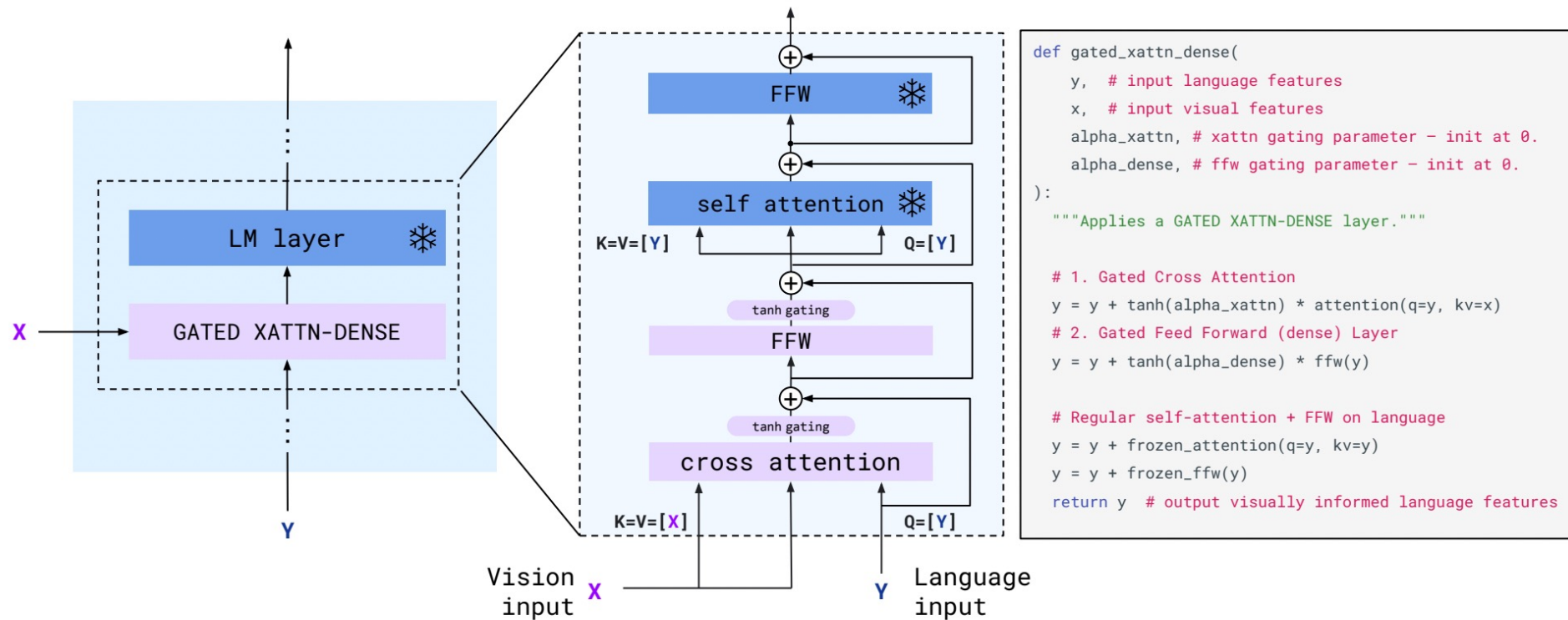


Figure 4: GATED XATTN-DENSE layers. To condition the LM on visual inputs, we insert new cross-attention layers between existing pretrained and frozen LM layers. The keys and values in these layers are obtained from the vision features while the queries are derived from the language inputs. They are followed by dense feed-forward layers. These layers are *gated* so that the LM is kept intact at initialization for improved stability and performance.

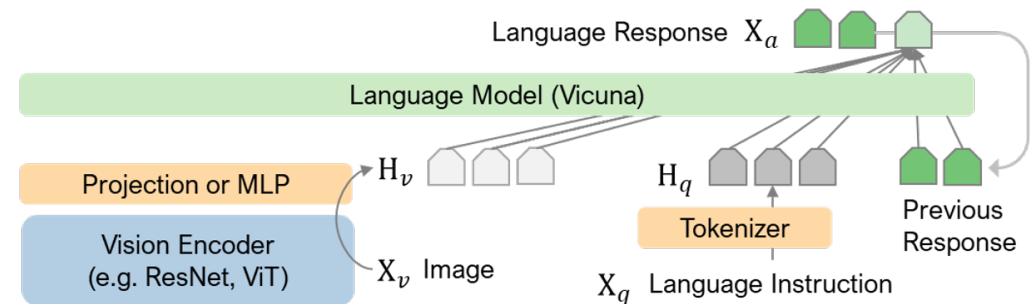
Flamingo (2022): First True MLLM-Scale Model

- **Specifications:**
 - ~80B parameters
 - Gated cross-attention layers
 - Perceiver resampler
- **Key Capability:**
 - Interweaved visual data into text prompts
 - Enabled **few-shot in-context learning**
 - Trained on long multi-image documents

LLaVA Paradigm: Visual Instruction Tuning

- **Minimalist Architecture**

1. Vision Encoder (CLIP ViT)
2. Language Decoder (Vicuna/LLaMA)
3. Simple MLP Projector

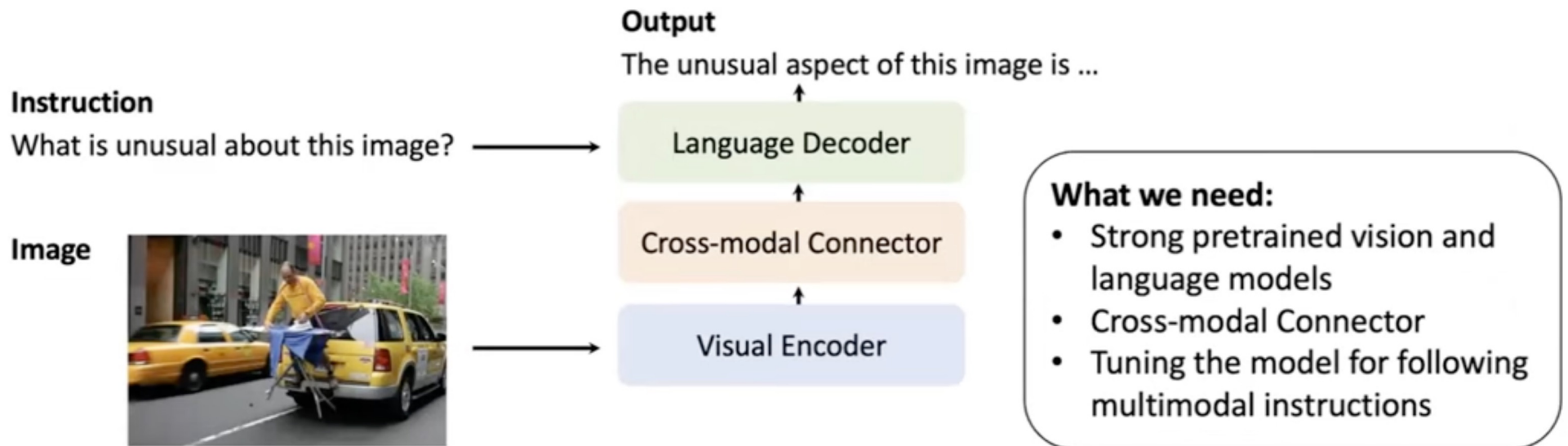


- **Two-Stage Training**

- **Stage 1: Feature alignment (projector only)**
- **Stage 2: Instruction tuning on GPT-4 synthetic data (conversation, description, complex reasoning)**

LLaVA: Visual Instruction Tuning

- Haotian Liu*, Chunyuan Li*, Qingyang Wu, Yong Jae Lee
 - NeurIPS 2023 (oral Presentation), *Equal Contribution



GPT-Assisted Visual Instruction Data Generation

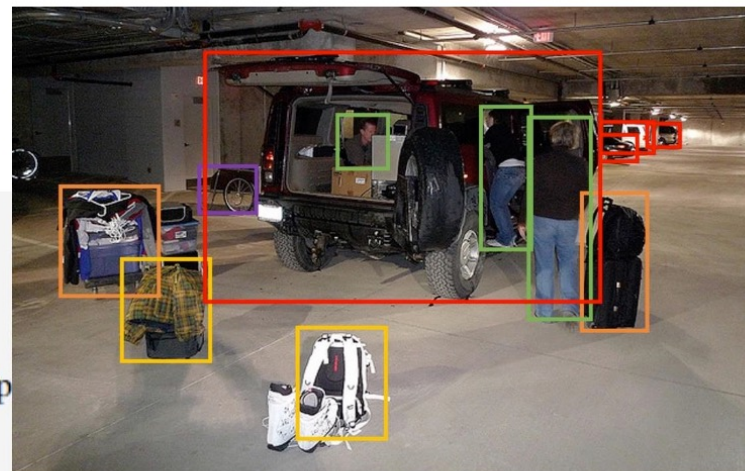
- Rich Symbolic Representations of Images
- In-context-learning with a few manual examples
 - Text-only GPT-4

Context type 1: Captions

A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.

Context type 2: Boxes

person: [0.681, 0.242, 0.774, 0.694], person: [0.63, 0.222, 0.686, 0.516], person: [0.444, 0.233, 0.487, 0.34], backpack: [0.384, 0.696, 0.485, 0.914], backpack: [0.755, 0.413, 0.846, 0.692], suitcase: [0.758, 0.413, 0.845, 0.69], suitcase: [0.1, 0.497, 0.173, 0.579], bicycle: [0.282, 0.363, 0.327, 0.442], car: [0.786, 0.25, 0.848, 0.322], car: [0.783, 0.27, 0.827, 0.335], car: [0.86, 0.254, 0.891, 0.3], car: [0.261, 0.101, 0.787, 0.626]



GPT-Assisted Visual Instruction Data Generation

- Three type of instruction-following responses

Response type 1: conversation

Question: What type of vehicle is featured in the image?

Answer: The image features a black sport utility vehicle (SUV).

Question: Where is the vehicle parked?

Answer: The vehicle is parked in an underground parking area, likely in a public garage.

Question: What are the people in the image doing?

Answer: The people are trying to fit all of their luggage into the SUV, likely preparing for a trip.

Response type 2: detailed description

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip.

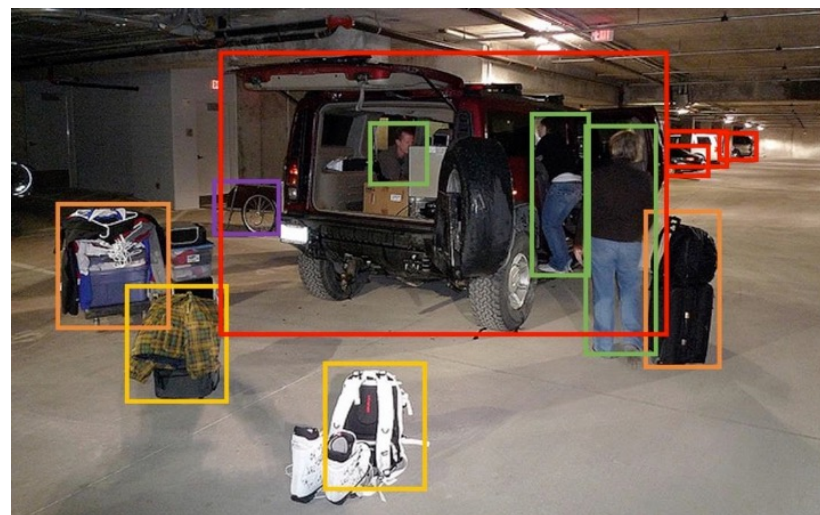
In the parking area, various luggage items are scattered around the vehicle. There are two backpacks, one located near the left rear wheel and the other closer to the right side of the vehicle. Additionally, there are two suitcases, one on the right side of the car and another further away near the center of the parking area. A bicycle can also be seen on the left side of the vehicle.

Other cars are parked around the main SUV, with one car positioned behind it and slightly to the left, another behind and slightly to the right, and the third car further behind on the right side.

Response type 3: complex reasoning

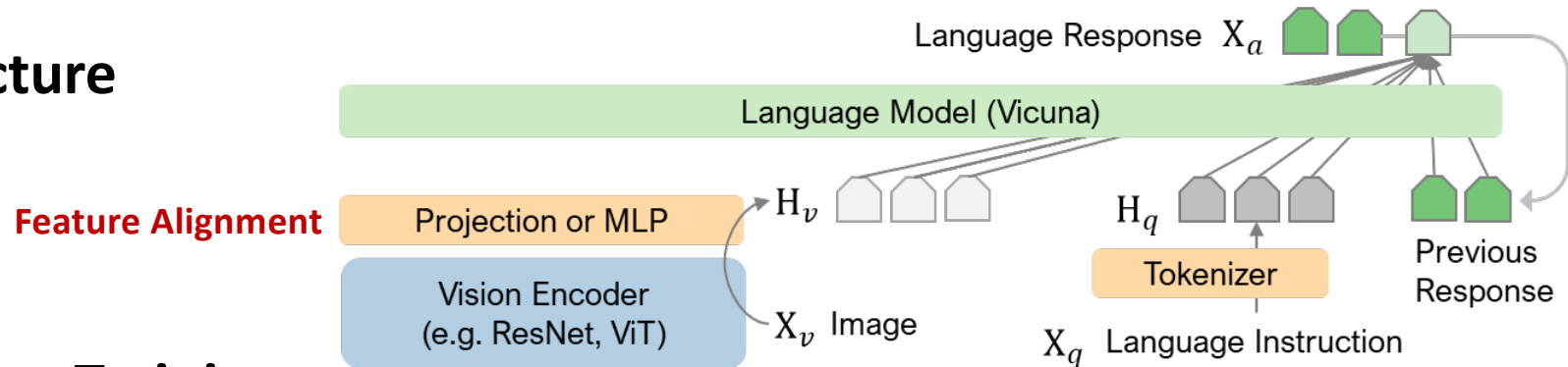
Question: What challenges do these people face?

Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings to accommodate. They might have to strategize and arrange the luggage efficiently to ensure that everything fits properly into the vehicle. Additionally, they need to consider the comfort of the passengers and visibility while driving, so the placement of the luggage must not obstruct the driver's view or make the passengers uncomfortable during the trip.



LLaVA: Architecture and Training

- **Architecture**



- **Two-stage Training**

- **Stage 1: Pre-training for **Feature Alignment****

- Only the projection matrix is updated, based on a subset of CC3M.

- **Stage 2: Fine-tuning End-to-End**

- Both the projection matrix and LLM are updated
- Visual Chat: The generated multimodal instruction data for daily user-oriented applications.
- Science QA: Multimodal reasoning dataset for the science domain.



LLaVA: Large Language and Vision Assistant (2023)

llava-v1.5-13b


Image

Drop Image Here
- or -
Click to Upload

Examples

	What is unusual about this image?
	What are the things I should be cautious about when I visit here?

LLaVA Chatbot



What is unusual about this image?

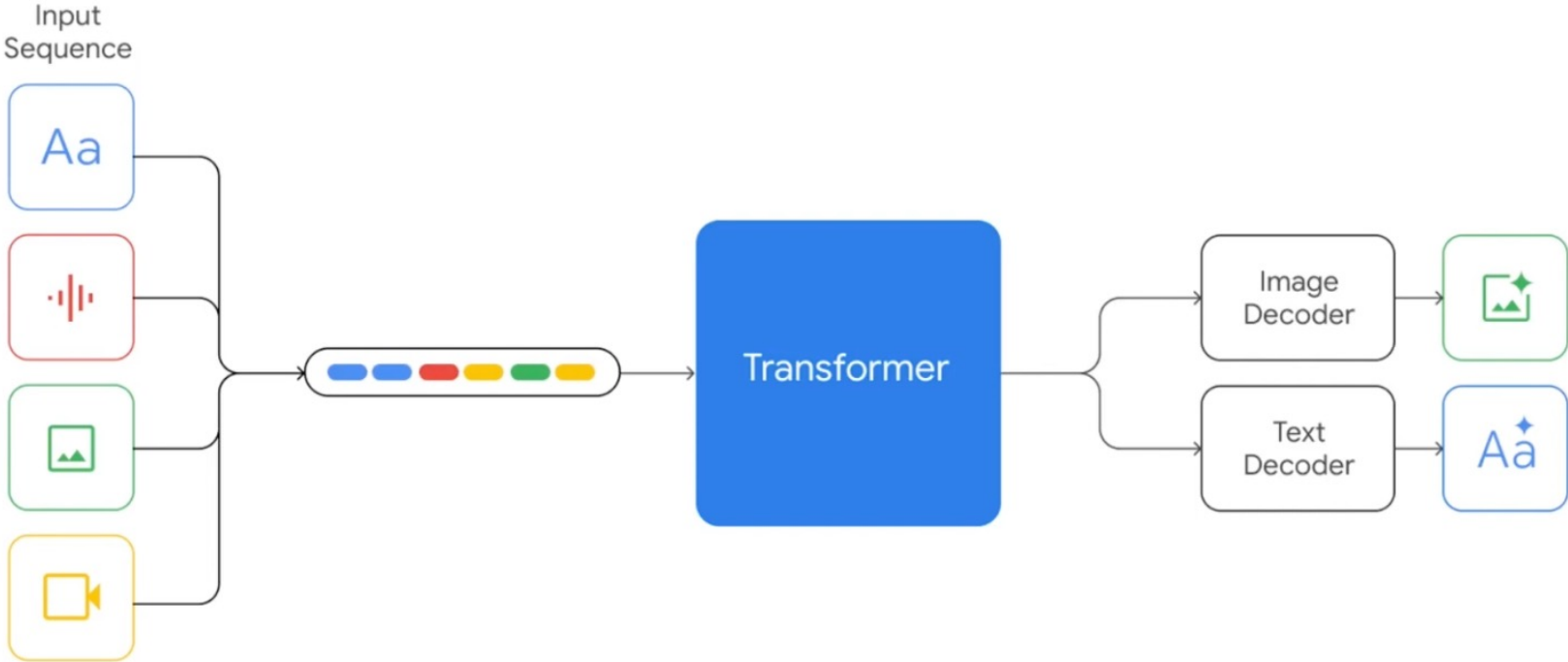
The unusual aspect of this image is that a man is ironing clothes while standing on the back of a moving car. This is not a typical scene, as ironing clothes is usually done indoors, in a more controlled environment, and not while the car is in motion. The man's ability to balance and perform this task on the back of a car is quite impressive and unconventional.

Enter text and press ENTER

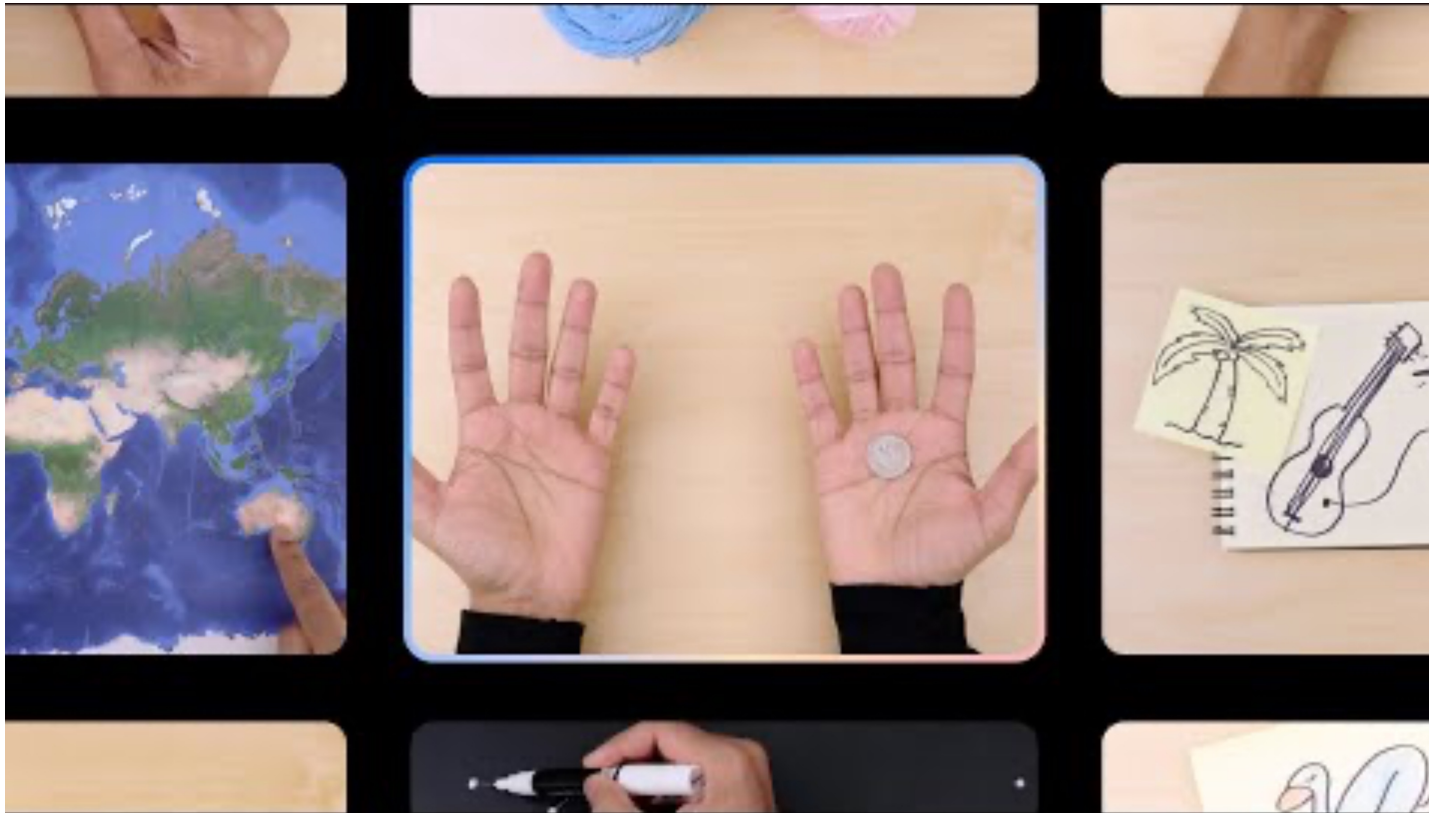
Send

LLaVA Demo: <https://llava.hliu.cc/> GitHub: <https://github.com/haotian-liu/LLaVA>

Google Gemini (2023)



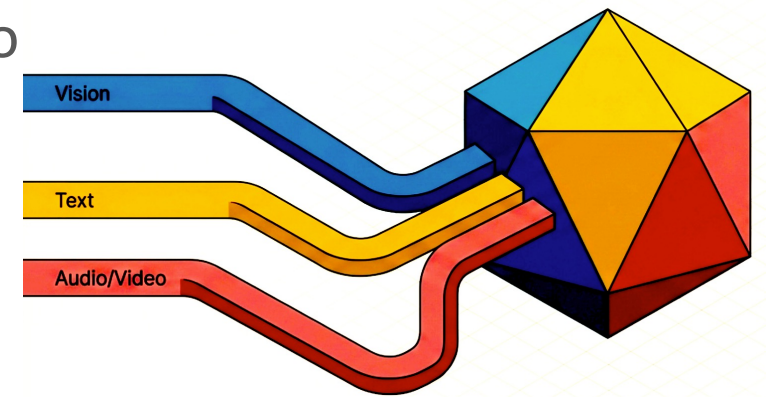
Gemini Demo



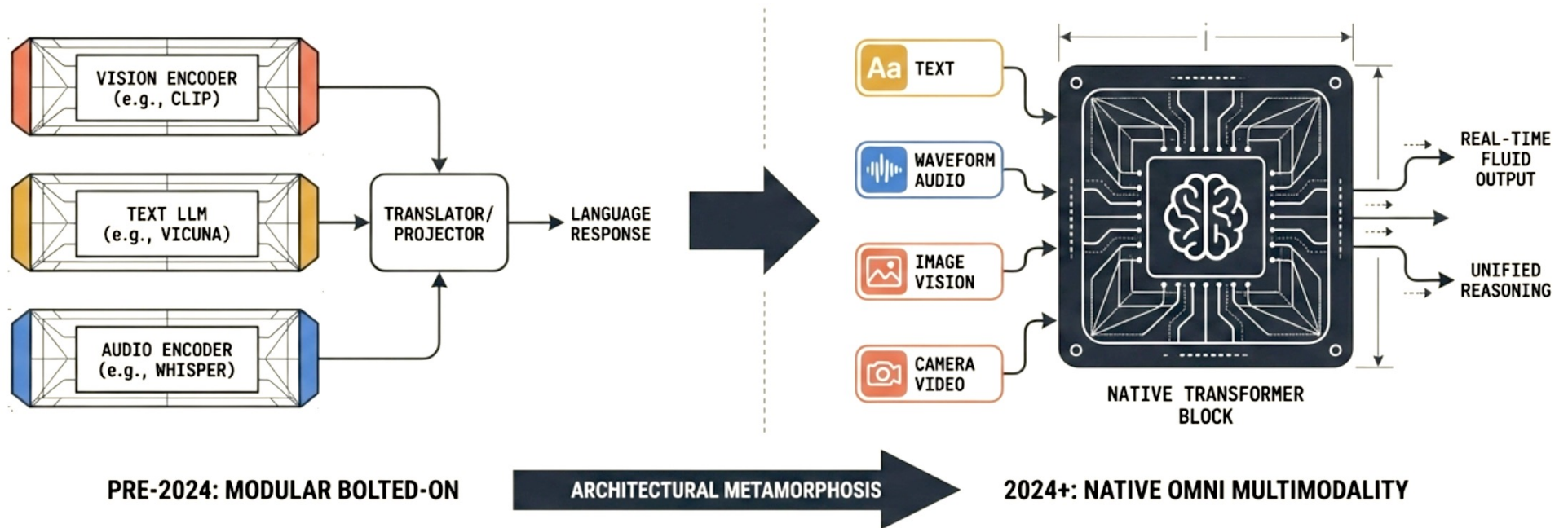
<https://www.youtube.com/watch?v=UIZAiXYceBI>

The Shift to Native Omni Multimodality (2024)

- **The End of Bolted-On:** By 2024, models transitioned to architectures where multiple modalities were integrated from the ground up.
- **GPT-4o Integration:** Text, vision, and audio built entirely natively within a single system.
- **Massive Context:** Gemini 1.5 and Claude 3 establish native vision pipelines with context windows expanding to hour-long video comprehension.



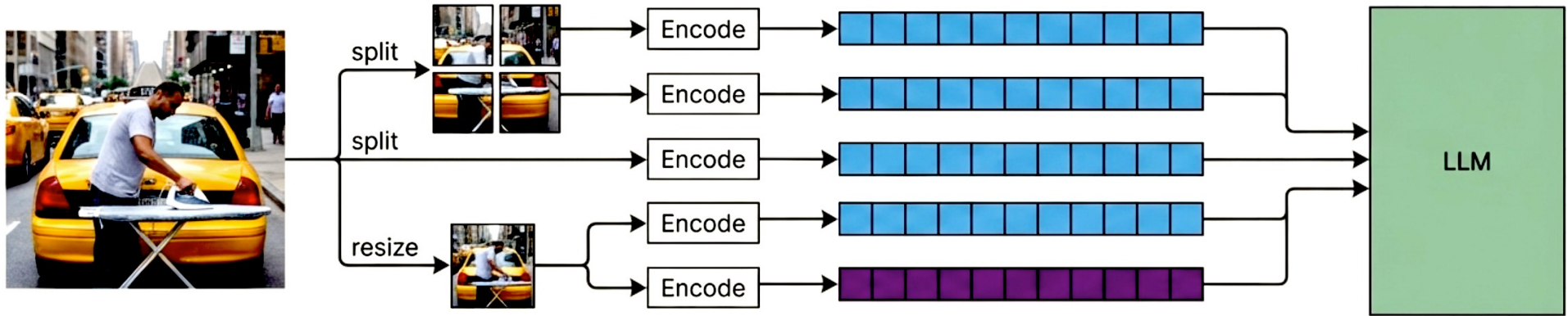
2024: Native Multimodality Architecture



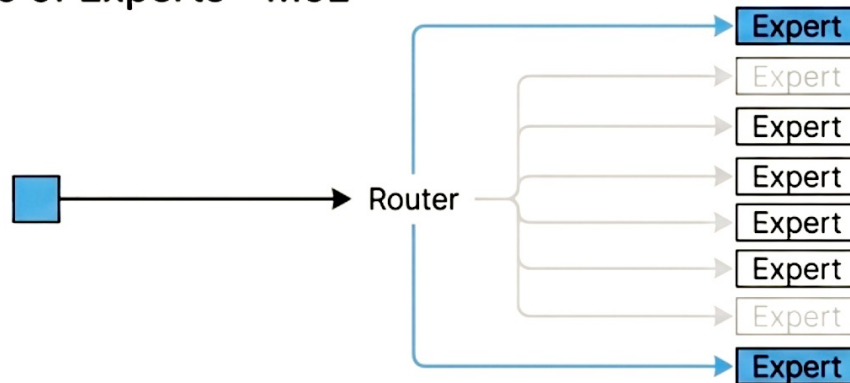
The Core Benefit: Eliminated the latency of translation layers, enabling real-time, fluid voice and visual interactions.

2024: Scaling & Efficiency

Dynamic Slicing - LLaVA-NeXT



Mixture of Experts - MoE



Efficiency: Mixture of Experts (MM1, MoE-LLaVA) routes visual tokens only to specific expert nodes out of a massive cluster, maintaining performance while drastically slashing computational energy.

GPT-4o Realtime Translation Demo



<https://www.youtube.com/watch?v=c2DFg53Zhvw>

GPT-4o Math Tutoring Demo



https://www.youtube.com/watch?v=lvXZCocyU_M

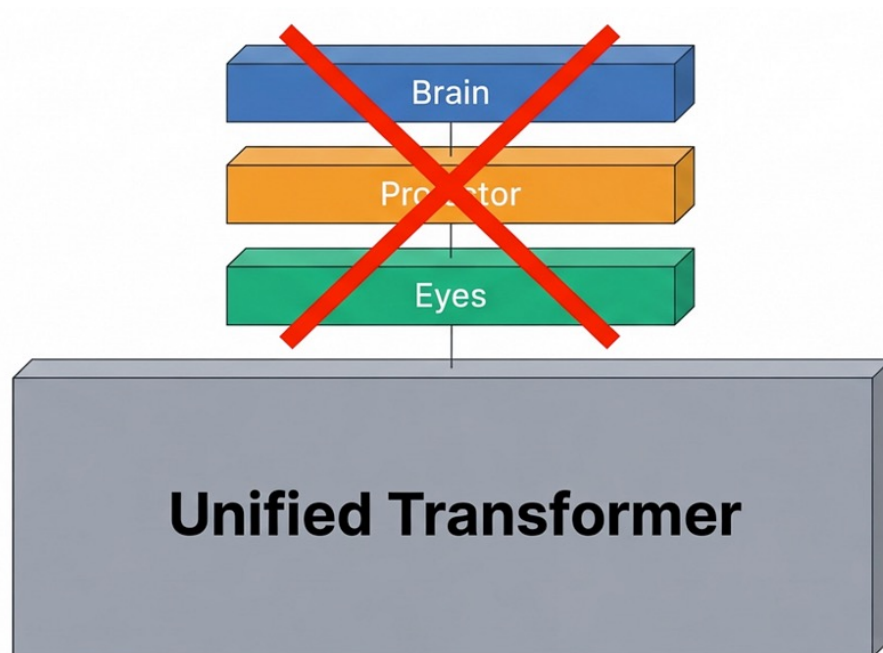
GPT-4o Introduction Full Video



<https://www.youtube.com/watch?v=DQacCB9tDaw>

The Unified Era: Everything is a Token (2025-2026)

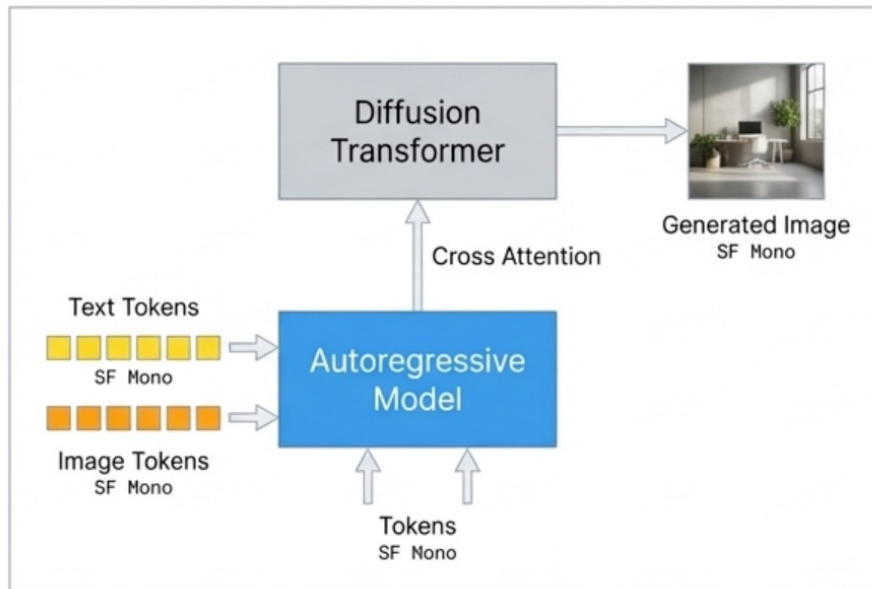
The New Paradigm: Any-to-Any intelligence via pure next-token prediction.



- By 2025, the distinction between understanding and generation blurs completely.
- If images, audio, and text are all converted into the exact same discrete tokens, **complex translators are obsolete.**
- Pure next-token prediction (no separate diffusion models needed)

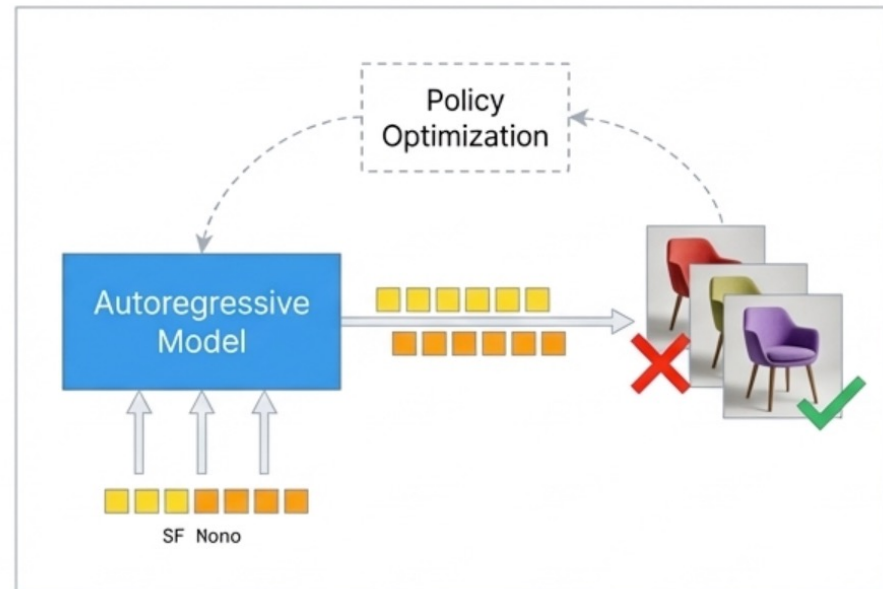
Autoregressive Dominance over Diffusion

The Past: Hybrid / Bolted-on Generation



The Old Paradigm: Generating images required maintaining separate, specialized diffusion models triggered by text cross-attention.

The Present/Future: Unified Generation

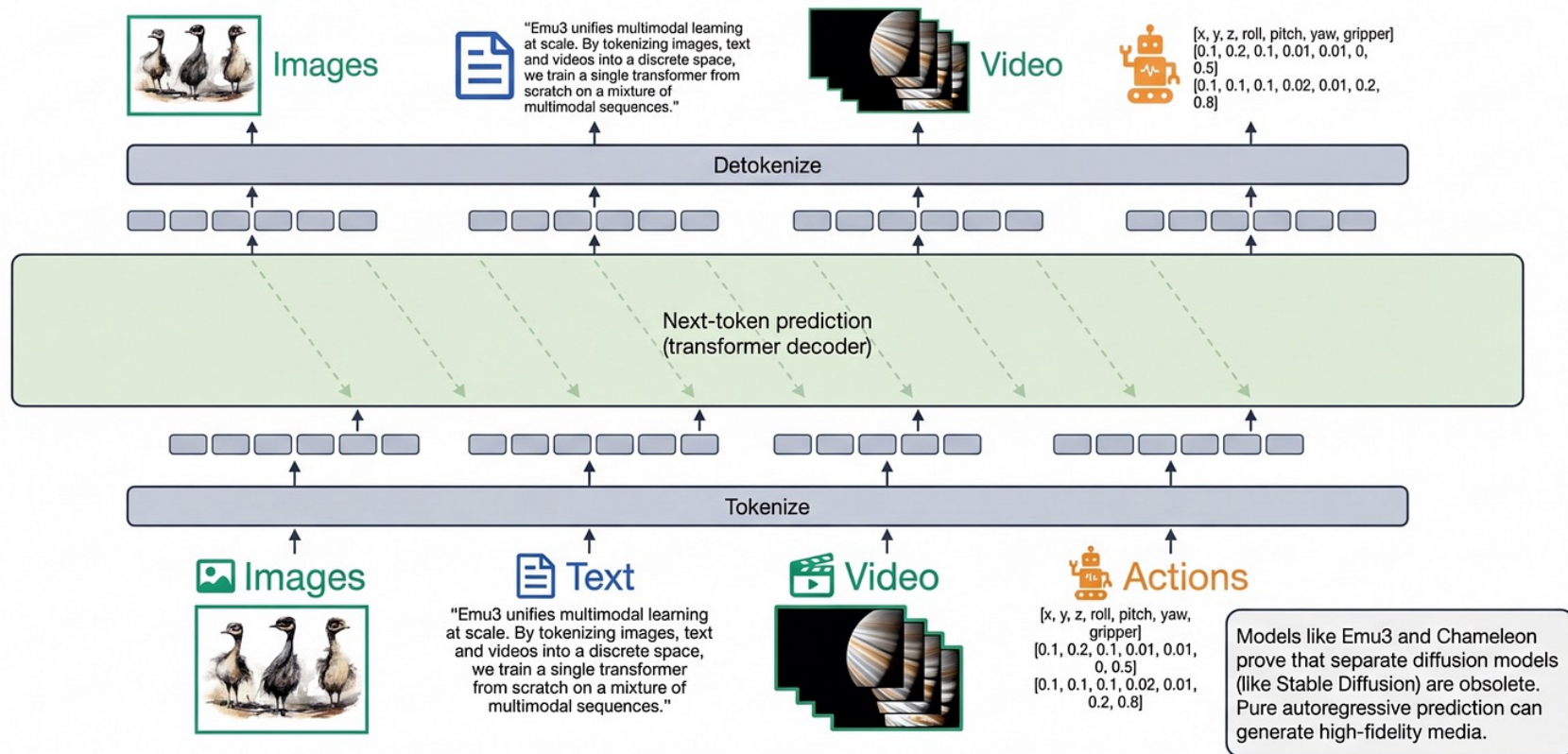


The New Paradigm: By discretizing visual data, MLLMs use pure autoregressive prediction (guessing the next pixel patch) to generate high-fidelity media natively.

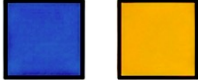


RLHF for Vision: Autoregressive models allow for Reinforcement Learning (Policy Optimization) directly on image generation, allowing models to learn visual rules and physics through feedback.

The Universal Mint

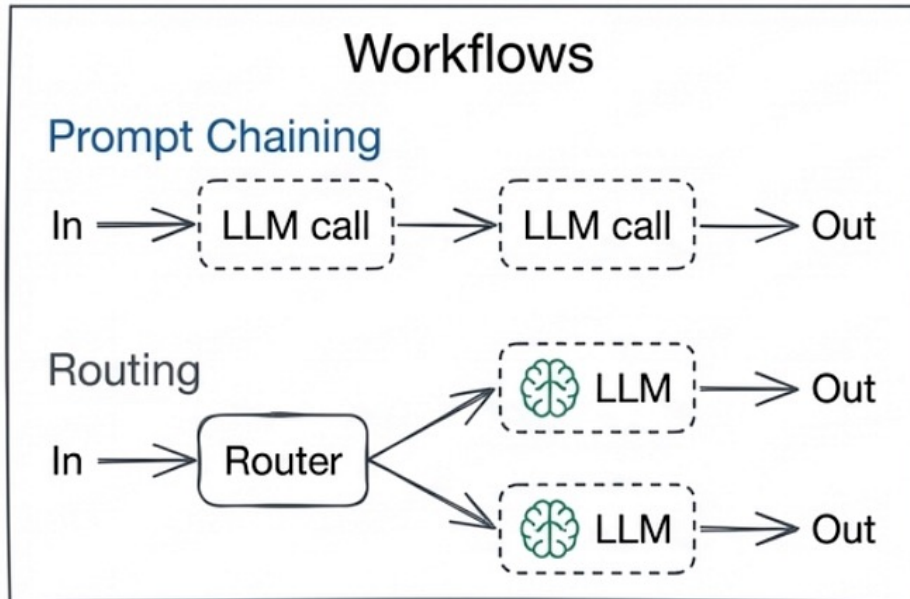
- **Pure Autoregression:** Models like Emu3 and Chameleon prove that specialized diffusion architectures are no longer required.
- **Any-to-Any Space:** By tokenizing images and audio into a discrete space, a single transformer trained from scratch can generate high-fidelity media interchangeably.



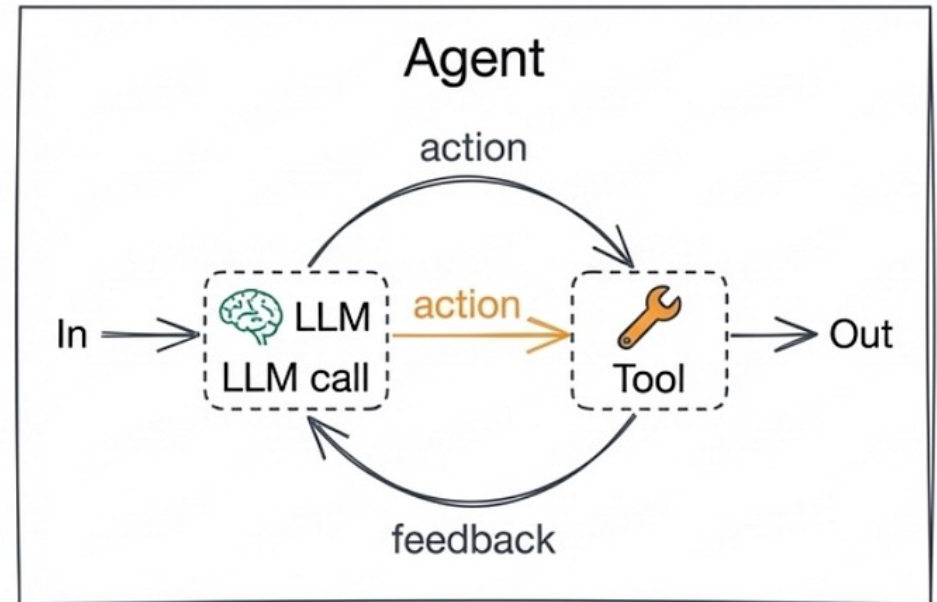
Era-Defining Architecture Matrix

	Early Fusion (2019)	The Projector Era (2023)	Unified Tokenization (2026)
Core Philosophy	Bolted-on & Fragmented	Translated & Bridged	Native & Identical
Key Components	CNN + BERT	ViT + MLP + LLM	Pure Autoregressive Transformer
Handling of Modalities	Separate continuous spaces 	Aligned continuous spaces 	Single discrete space 
Landmark Model	VisualBERT / LXMERT	LLaVA / BLIP-2	Emu3 / Chameleon

Application Shift: Workflows vs. Agents



The LLM is embedded within predefined code paths directed by human logic.



The LLM directs its own actions based on environmental feedback, actively calling tools and interpreting continuous multimodal data.

From Model → Agent

- Prompt chaining → Autonomous agents (Qwen3-Omni, Gemini 3.1 Pro)
- Tool use, continuous video streams, UI navigation

Specialized Applications and Impact

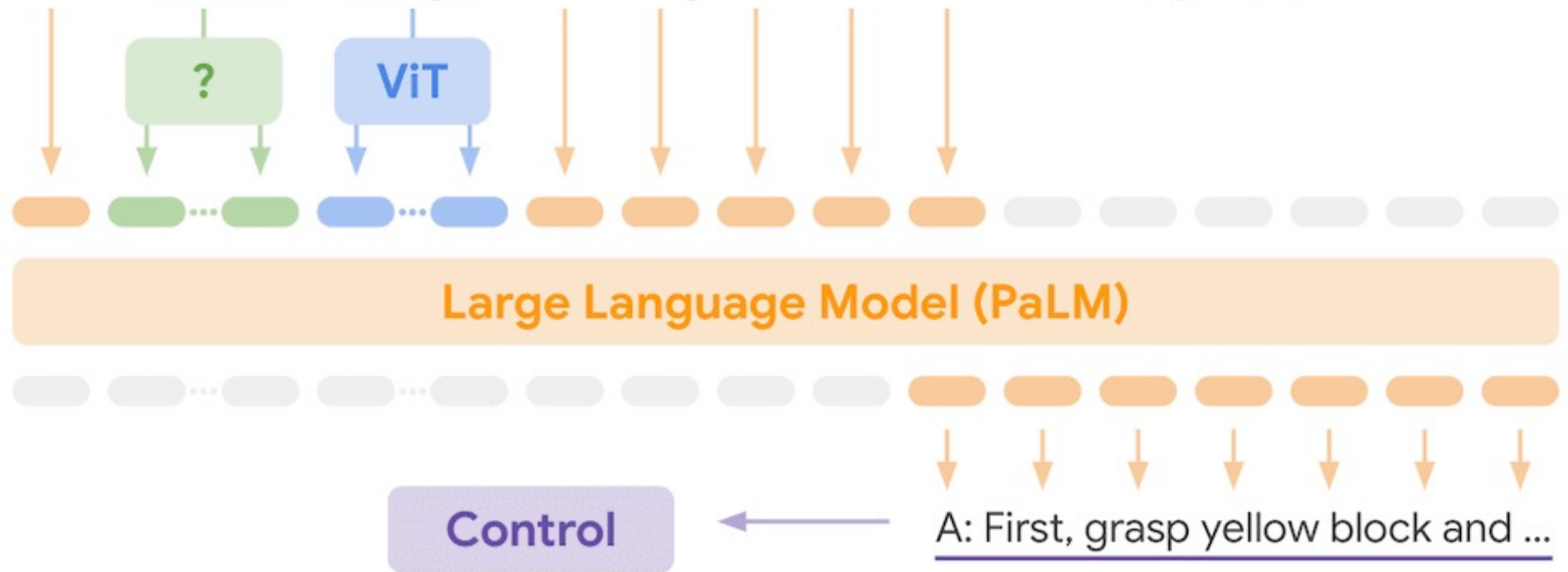


- **Specialized Applications**

- Robotics: PaLM-E (embodied planning)
- Medical: LLaVA-Med (curriculum learning for radiology)
- Creative: LLaVA-Plus (skill repository + external tools)

PaLM-E: An Embodied Multimodal Language Model

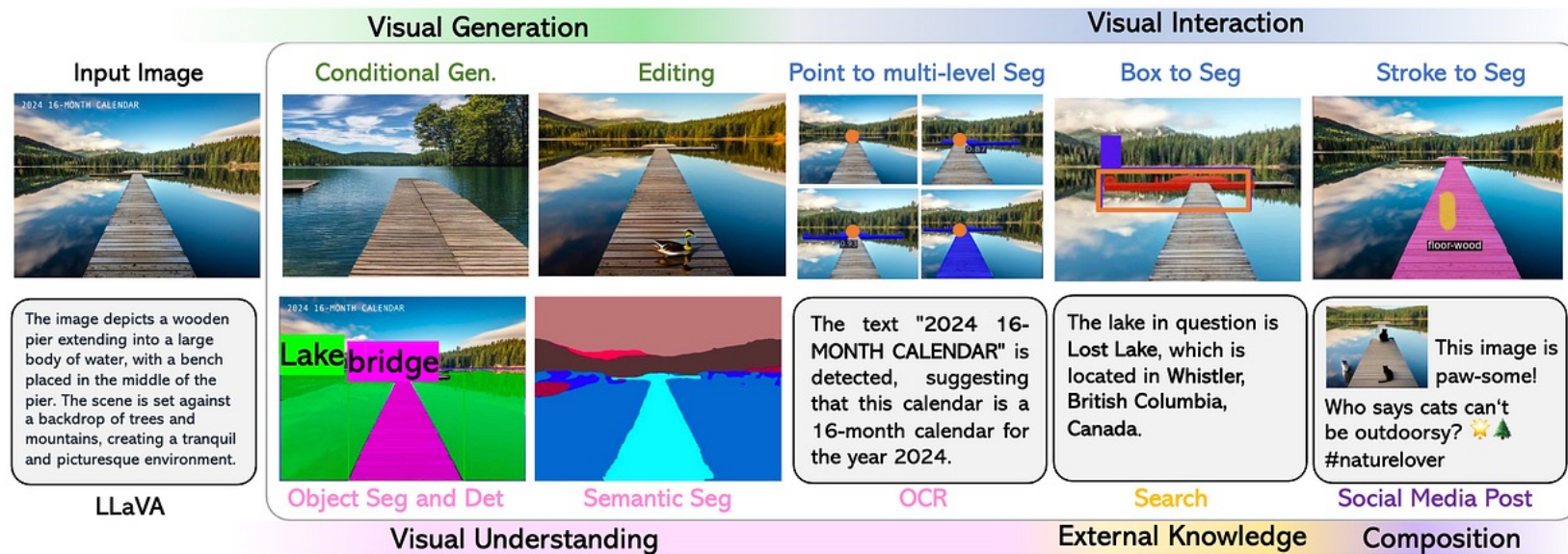
Given **<emb>** ... **** Q: How to grasp blue block? A: First, grasp yellow block



Embodied agents that use MLLMs as backbones for planning and environmental interaction.





Agentic Tools (LLaVA-Plus)

Incorporates a “Skill Repository,” allowing the model to trigger external tools like search engines or segmentation models (Grounding DINO) when internal knowledge is insufficient.



<https://llava-vl.github.io/llava-plus/>

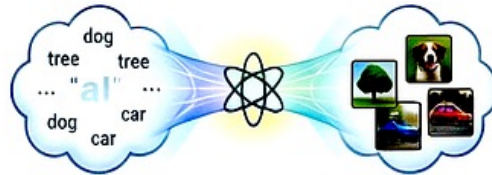
Persistent Challenges

<p>Hallucination </p> <p>Visual grounding failures. The model asserts it 'sees' elements not actually present in the source pixels.</p>	<p>Modality Imbalance </p> <p>Textual reasoning capabilities consistently overpower and dominate visual comprehension logic.</p>
<p>Inference Costs </p> <p>Processing high-resolution video or massive spatial contexts demands unsustainable computational overhead.</p>	<p>Safety & Robustness </p> <p>High vulnerability to multimodal jailbreaks and adversarial, pixel-manipulated image inputs.</p>

Post-2026 Outlook & Future Directions

- Deeper unification → 3D & world models
- Test-time scaling → multimodal Chain-of-Thought
- Reliability → Reinforcement Learning with Verifiable Rewards (RLvR)
- Democratization: Smaller open-source models stay competitive

2018–2022: The Bridging Era



2018–2020

The Text-to-Text Foundation (2018–2020)

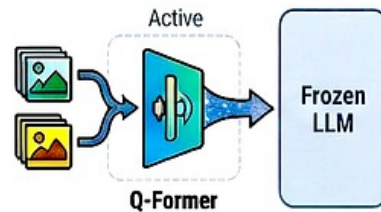
BERT and early GPT models established the transformer architecture but remained “blind” to non-textual data.



2021

CLIP: The Shared Latent Space
CLIP aligned visual and textual concepts into a shared mathematical space, becoming the industry’s visual backbone.

2022

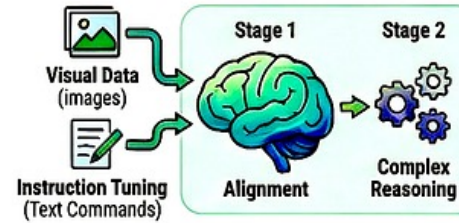


2022

Efficient Bridging via Q-Former
BLIP-2 introduced lightweight modules to extract visual features for frozen LLMs, reducing computational alignment costs.

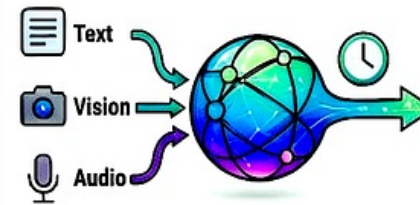
Architecture Type	Primary Function	Notable Example
Encoder-Only	Understanding	BERT (2018)
Decoder-Only	Generation	GPT (2018)
Encoder-Decoder	Translation	T5 (2019)

2023–2026: The Unified Era



2023

LLaVA & Visual Instruction Tuning
Models were taught to reason about images using a two-stage alignment and complex reasoning pipeline.



2024

Native “Omni” Multimodality
Systems like GPT-4o integrated text, vision, and audio natively to enable real-time, low-latency interactions.



2024

Native “Omni” Multimodality
Systems like GPT-4o integrated text, vision, and audio natively to enable real-time, low-latency interactions.



2025–2026

Any-to-Any Unified Intelligence
Unified tokenization allows models like Emu3 to generate and understand all media types interchangeably.