

Quiz 3

15 minutes

CS388: Natural Language Processing

Lecture 18: Vision-Language Models and Vision-Language- Action Models



TEXAS

The University of Texas at Austin

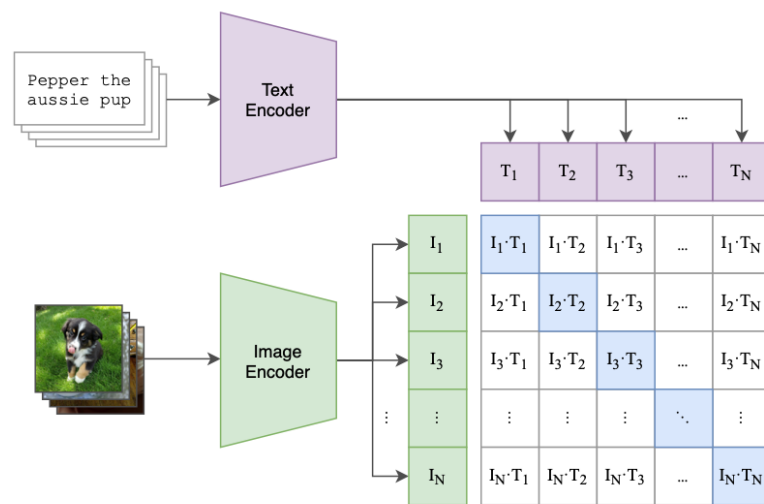
Today

- Problems with CLIP
- What are VLMs?
- VLM history
 - Unifying Vision-and-Language Tasks via Text Generation
 - Flamingo
 - BLIP and the Q-former
 - Llava
- VLMs now, VLM applications
- VLAs
 - Problem framing
 - Early VLAs
 - Diffusion policies
 - VLAs now

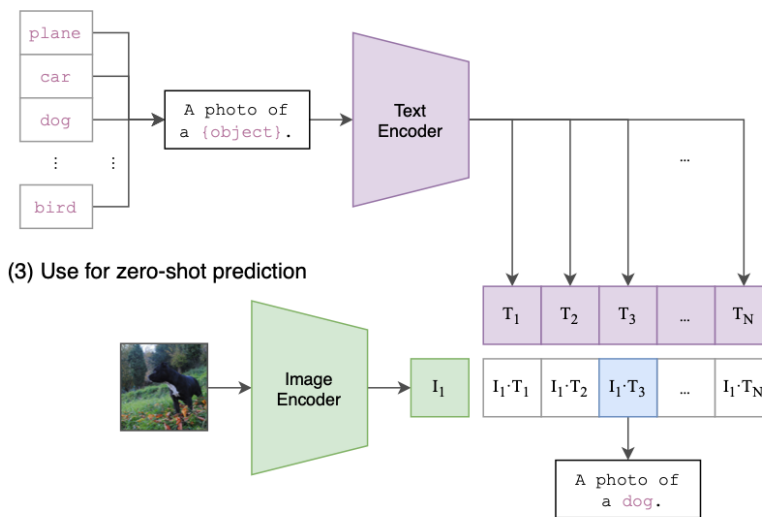
Recap

- **Grounding:** tie words to perception and action
- **CLIP:** contrastive pretraining on 400M image–text pairs
- **Zero-shot:** $\hat{c} = \arg \max_c \cos(g(I), f(t_c))$, $t_c = \text{“a photo of a } \{c\}\text{”}$

(1) Contrastive pre-training



(2) Create dataset classifier from label text



(3) Use for zero-shot prediction

Problems with CLIP

- Can't generate text
 - Only scores (image, text) pairs
- Limited in how it can be used
 - Needs a candidate set: classification, retrieval
 - No VQA, captioning, or dialogue
- Weak compositionality
 - Behaves like a bag of words: word order, relations, counting



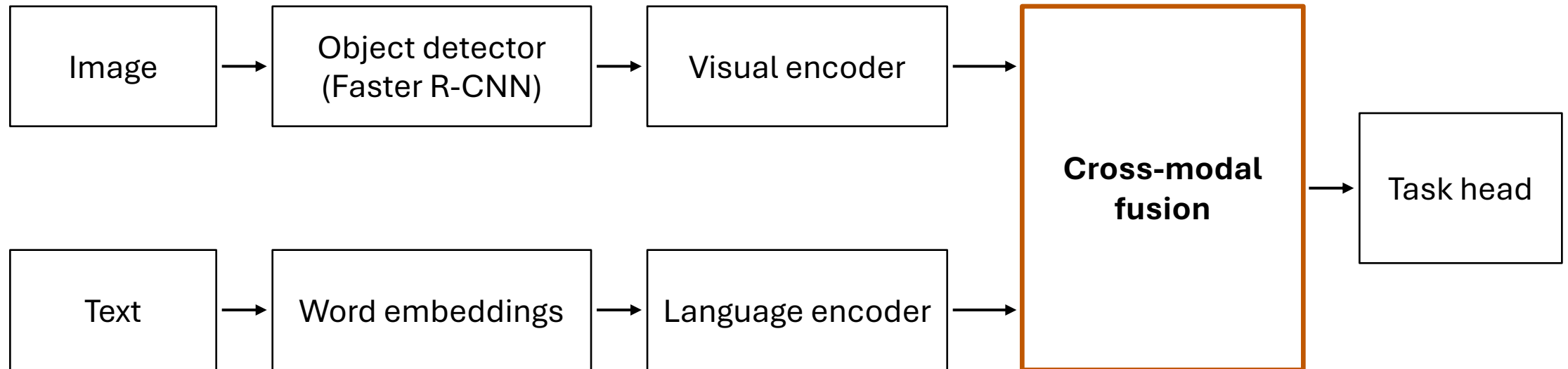
(a) some plants
surrounding a
lightbulb



(b) a lightbulb surrounding some plants

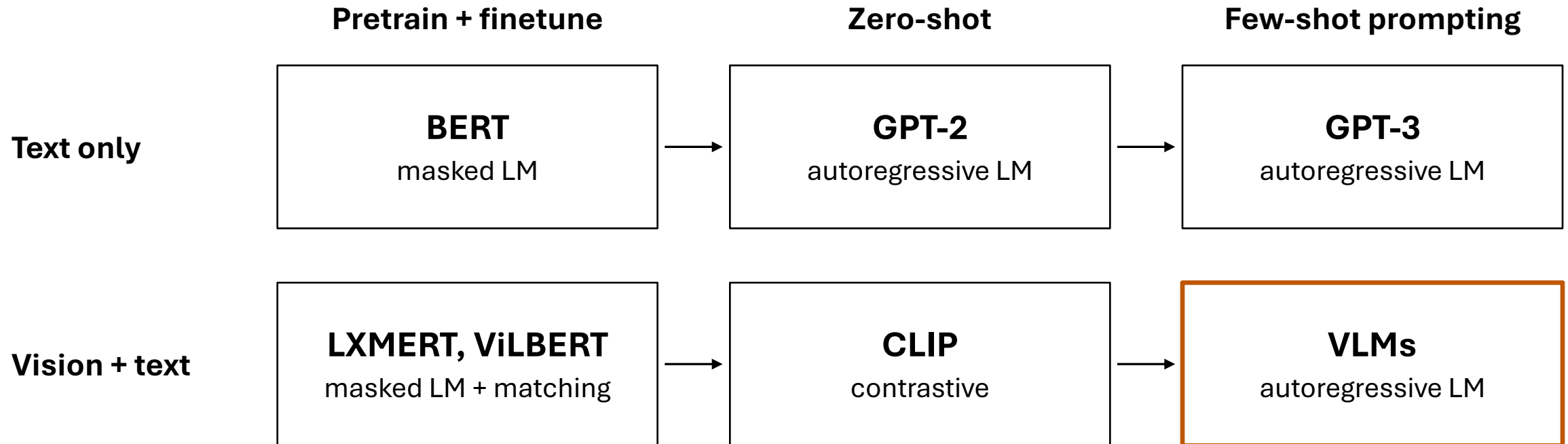
Recall dual-stream architectures

- LXMERT, ViLBERT
- Pretraining+finetuning paradigm

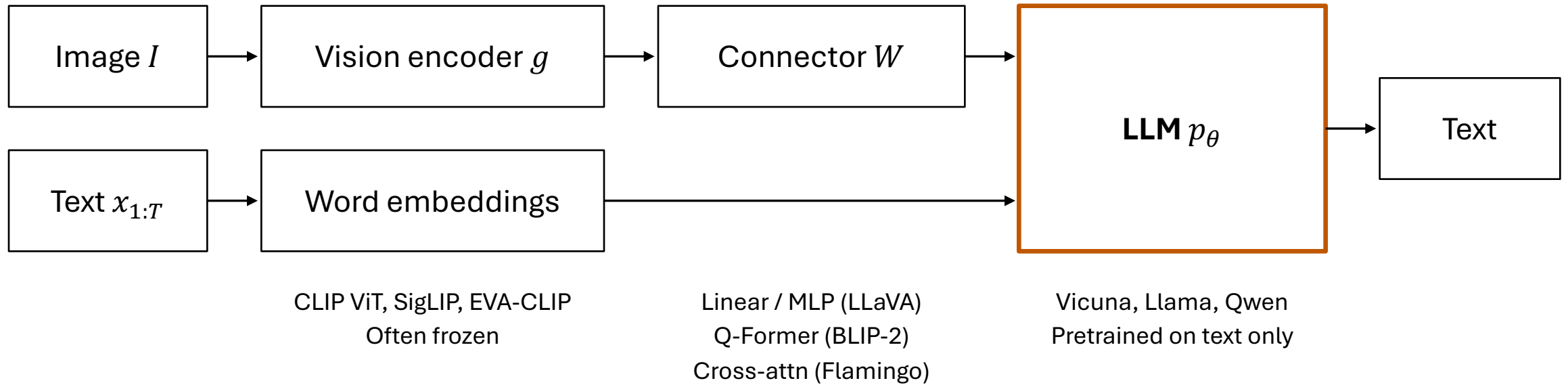


VLMs

- Vision + language followed ~the same path as text-only NLP



VLM typical architecture

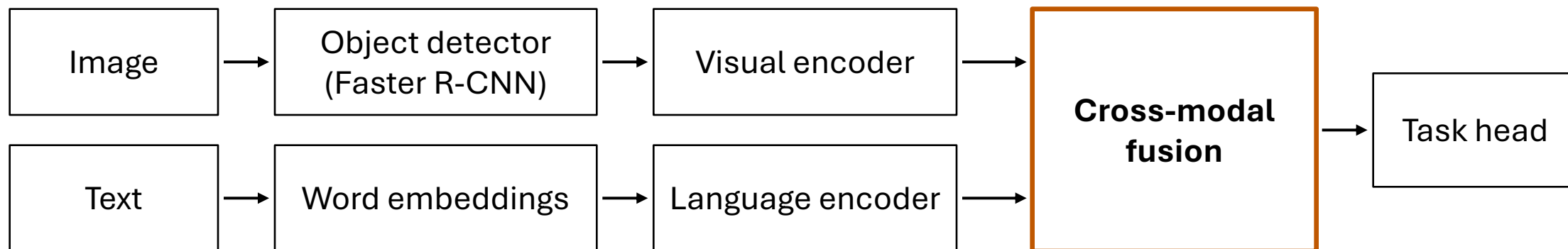


Visual tokens: $H_v = W g(I)$

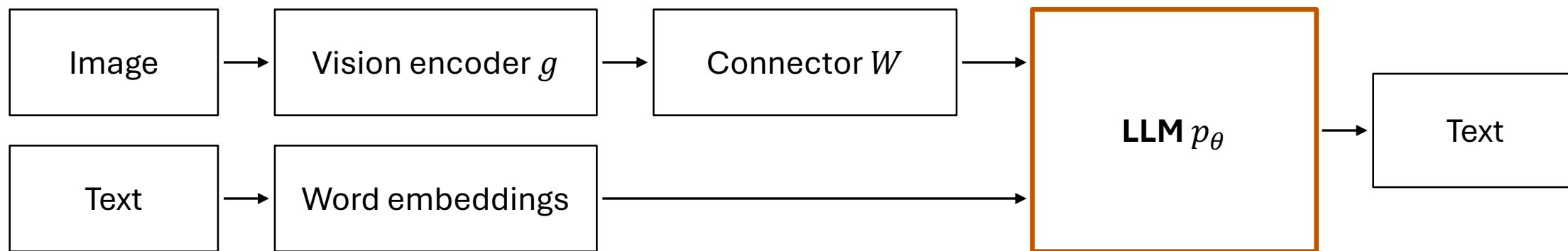
Train with the usual LM loss: $\mathcal{L} = -\sum_{t=1}^T \log p_{\theta}(x_t | H_v, x_{<t})$

VLMs vs LXMERT-style

LXMERT-style



VLM



Unifying Vision-and-Language Tasks via Text Generation

- One T5/BART model for all tasks
- Input: region features + task prefix
 - “vqa: what is the man jumping over?”
- Output is always text
 - Even grounding: “<vis_3>”
- Same LM loss, no task-specific heads

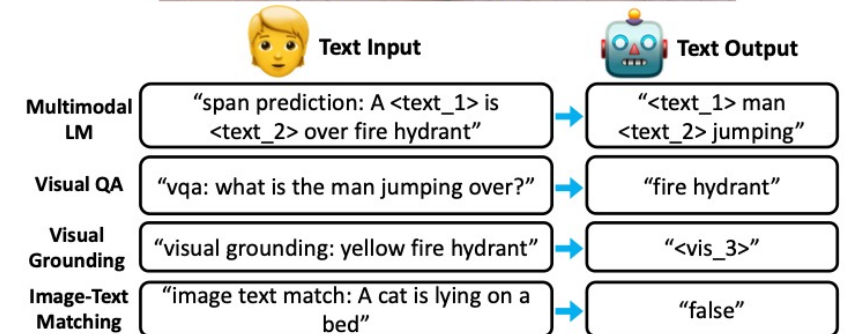
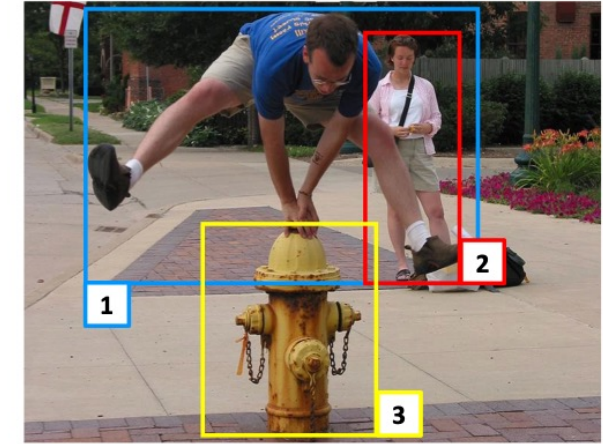


Figure 1. Our unified framework for learning vision-and-language tasks. While existing methods require designing task-specific architectures for different tasks, our framework unifies them together as generating text labels conditioned on multimodal inputs.

Flamingo

- **Few-shot learning:** image–text examples in the prompt, like GPT-3
- No finetuning: with 32 shots, beats finetuned SOTA on 6 of 16 tasks



Flamingo

- **Frozen:** vision encoder and LLM (Chinchilla 70B)
- **Perceiver Resampler:** image features → 64 visual tokens
- **Gated cross-attention** layers inserted between LM blocks
- Data: 43M interleaved webpages + image/video–text pairs

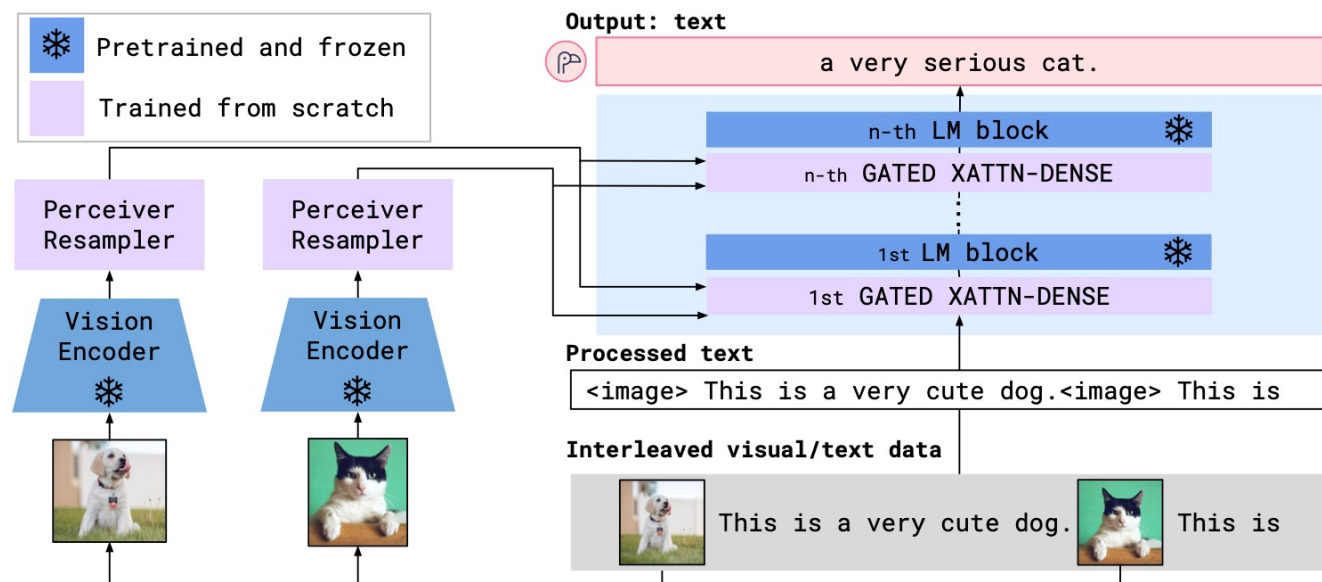
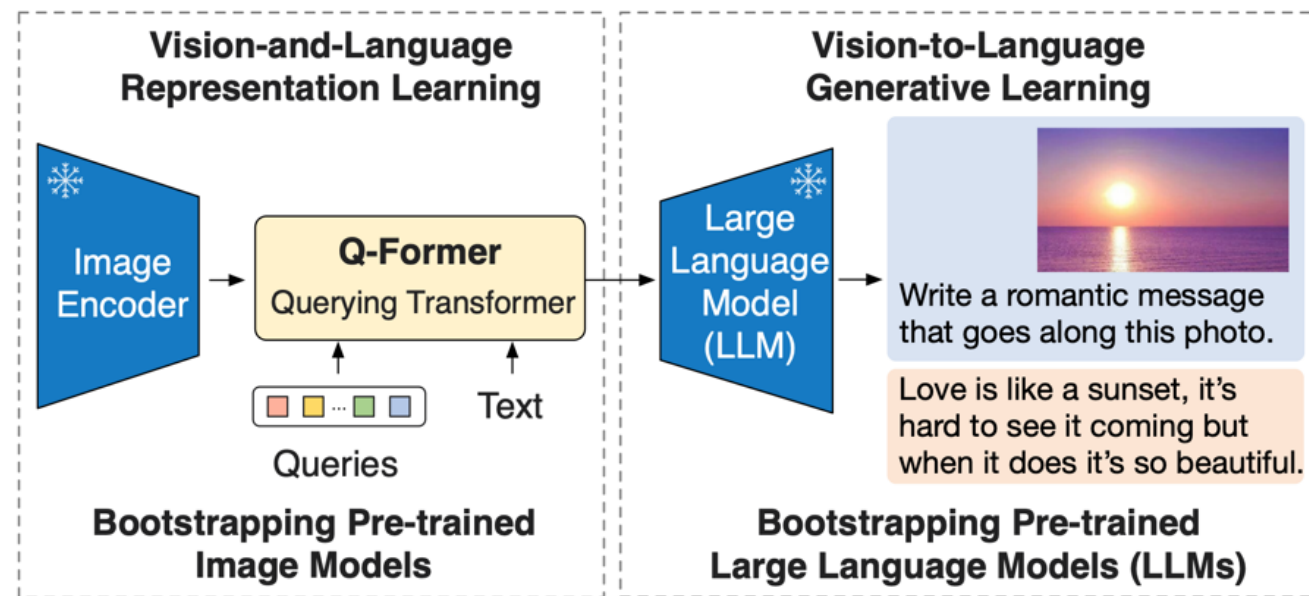


Figure 3: **Flamingo architecture overview.** Flamingo is a family of visual language models (VLMs) that take as input visual data interleaved with text and produce free-form text as output.

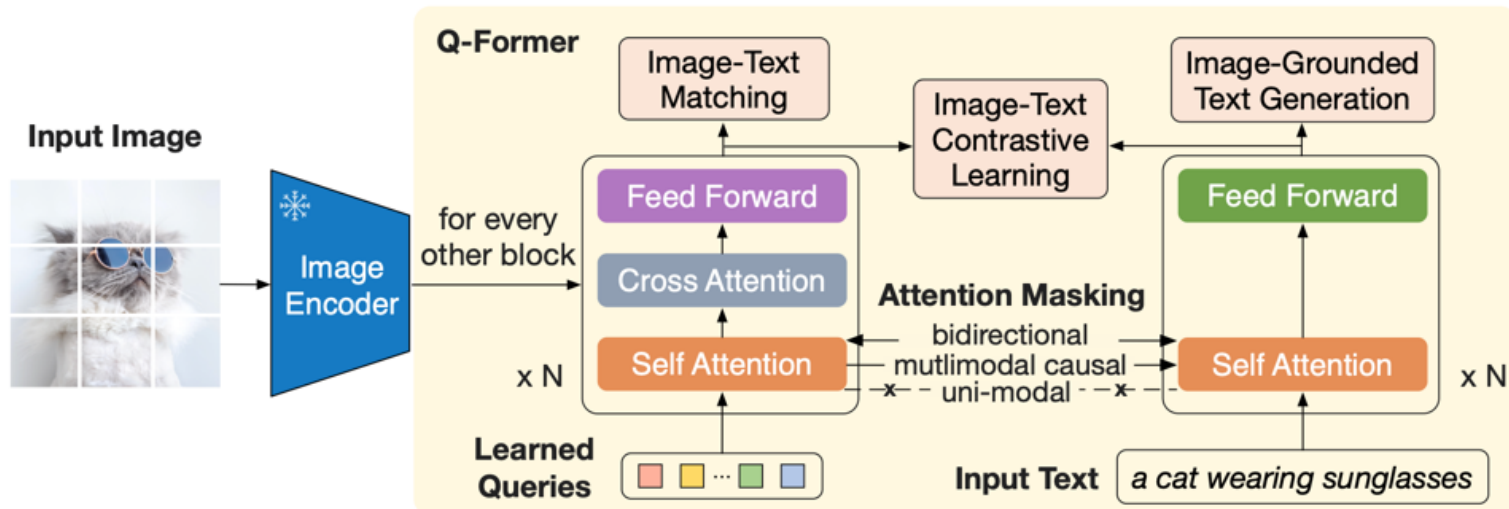
BLIP-2 and Q-former

- **Freeze** both big models
 - Image encoder (ViT-g) and LLM (OPT, FlanT5)
- Train only the **Q-Former** (188M)
 - 32 learned queries → 32 visual tokens
- Two training stages
 - 1: learn visual features that match text
 - 2: feed queries to the LLM as a soft prompt



BLIP-2 and Q-former

- Learned queries self-attend, then cross-attend to frozen image features
- Shares self-attention layers with a text transformer (init. from BERT-base)
- Output: 32×768 query embeddings, projected into H_v for the LLM
- Fixed 32 tokens, regardless of image resolution



BLIP-2 objectives

- Stage 1: three losses, same model, different attention masks
- **ITM**: bidirectional; classify match vs. no match
- **ITG**: causal; generate the caption from the queries
- **ITC**: unimodal; contrastive, like CLIP
- Stage 2: queries → frozen LLM, standard LM loss

Q: query token positions; **T**: text token positions.

■ masked □ unmasked

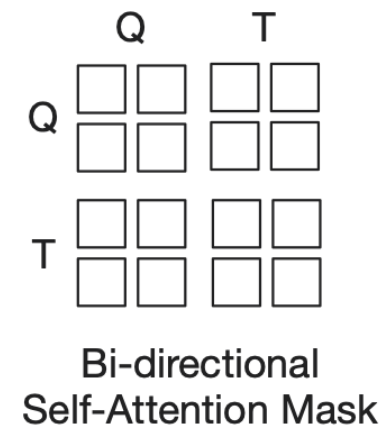


Image-Text
Matching

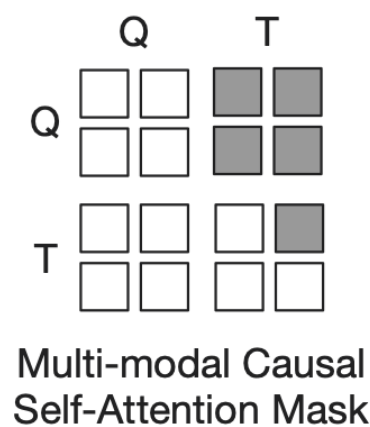


Image-Grounded
Text Generation

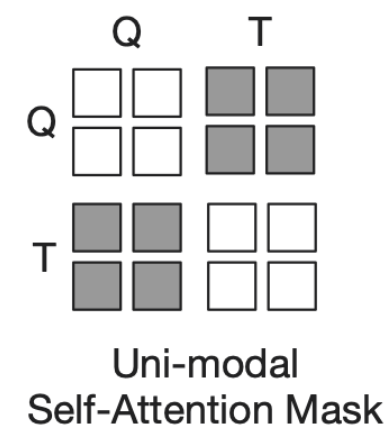


Image-Text
Contrastive Learning

BLIP-2 results

- Free-form text output, zero-shot
- Zero-shot VQAv2: **65.0** vs. 56.3 for Flamingo-80B
- **54× fewer** trainable parameters
- Follows image + instruction prompts



Explain the advantages of this product.

The audi e-tron quattro concept is a plug-in hybrid electric sports car that has a range of 310 miles and can accelerate from 0 to 60 mph in just 4 seconds.



Models	#Trainable Params	Open-sourced?	Visual Question Answering	Image Captioning		Image-Text Retrieval	
			VQAv2 (test-dev) VQA acc.	NoCaps (val) CIDEr	SPICE	Flickr (test) TR@1	IR@1
BLIP (Li et al., 2022)	583M	✓	-	113.2	14.8	96.7	86.7
SimVLM (Wang et al., 2021b)	1.4B	✗	-	112.2	-	-	-
BEIT-3 (Wang et al., 2022b)	1.9B	✗	-	-	-	94.9	81.5
Flamingo (Alayrac et al., 2022)	10.2B	✗	56.3	-	-	-	-
BLIP-2	188M	✓	65.0	121.6	15.8	97.6	89.7

Table 1. Overview of BLIP-2 results on various **zero-shot** vision-language tasks. Compared with previous state-of-the-art models. BLIP-2 achieves the highest zero-shot performance while requiring the least number of trainable parameters during vision-language pre-training.

LLaVA

- **Connector: a single linear layer**

$$H_v = W Z_v, \quad Z_v = g(I)$$

- Vision encoder: CLIP ViT-L/14, frozen
- LLM: Vicuna (Llama tuned on chats)
- Visual tokens + instruction $\rightarrow$ response

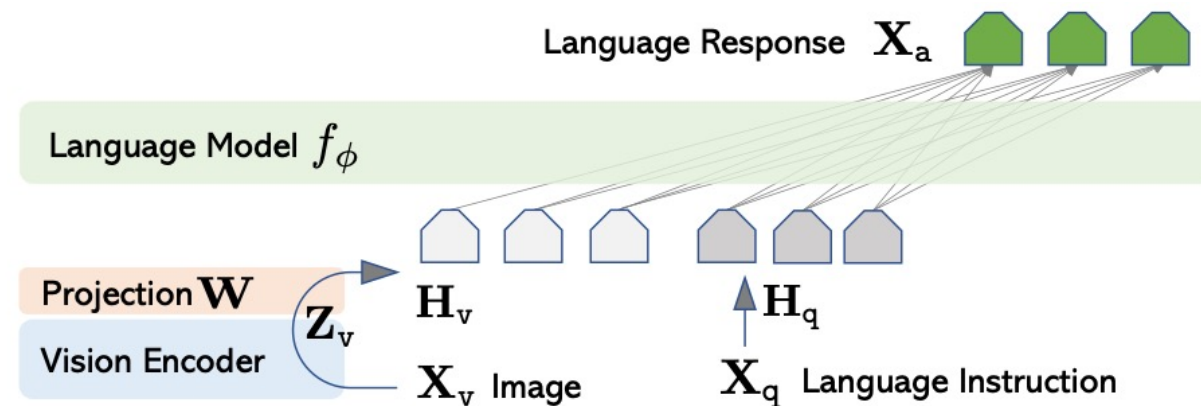


Figure 1: LLaVA network architecture.

LLaVA: visual instruction data

- Few image–instruction datasets existed
- Synthesized data via GPT-4
- Done without the image
 - Gets COCO captions + bounding boxes as text
- Generates 158K examples
 - 58K conversations
 - 23K detailed descriptions
 - 77K complex reasoning
- What are some issues with this?

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], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: conversation

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

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

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. ...<omitted>

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 ...<omitted>

LLaVA training

Stage	Data	Trained
1. Feature alignment	595K image–caption pairs (CC3M)	W only
2. Instruction tuning	158K instruction data	W + LLM

- **LLaVA-1.5:** MLP connector, 336px images, academic VQA data
- Simple recipe became the default open VLM design

VLMs now

- **Higher resolution:** image tiles or native dynamic resolution
- **Video** and long multimodal context
- **Pointing and GUI control:** output coordinates, not just text
- **Reasoning:** RL-trained “thinking” VLMs
- **Native multimodal pretraining** instead of a bolted-on encoder

Open weights

Qwen3-VL, InternVL3.5, GLM-4.6V, Gemma, Molmo (open data)

Proprietary

GPT-5, Gemini 3, Claude

VLM applications

Application	Input	Output
Document and chart QA	Scans, PDFs, plots	Answers, extracted fields
GUI agents	Screenshots	Clicks, typing
Accessibility	Phone camera	Scene descriptions
Medical imaging	X-rays, pathology slides	Draft reports
Video understanding	Long videos	Summaries, timestamps
Robotics	Robot cameras	Plans, actions (VLAs)

VLA's

- **Goal:** one model that sees, reads an instruction, and acts
- A VLM whose outputs are robot actions: $\pi_{\theta}(a_t \mid o_t, \ell)$
- Reuse web-scale knowledge for robot control



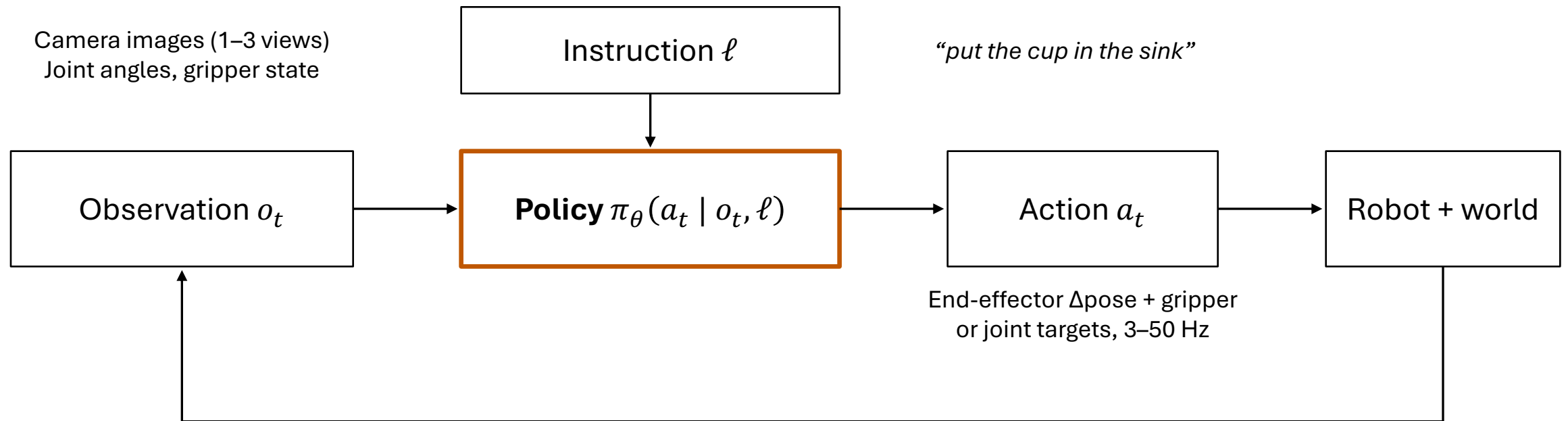
<https://www.pi.website/blog/pi0>



<https://robotics-transformer2.github.io>

Robotics overview

- **Goal:** map observations + instruction to actions, in a closed loop (recall RL)

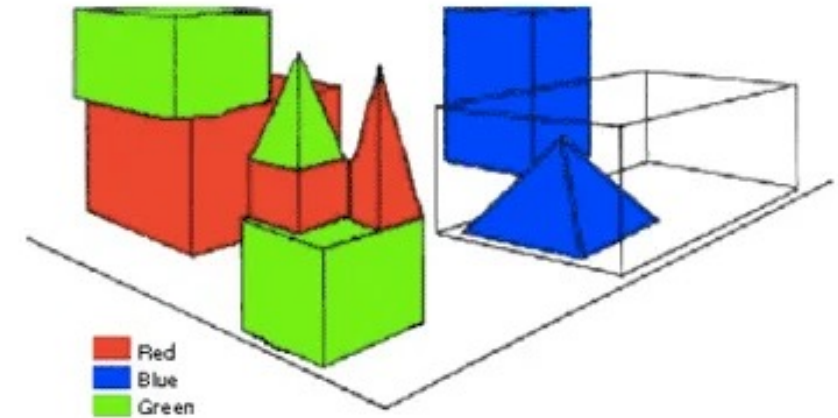


Robotics: constraints

- **Perception:** partial views, occlusion, clutter
- **Physical feasibility:** reach, collisions, stable grasps
- **Real time:** the loop runs at 3–50 Hz
- **Errors compound** and happen in the physical world
- **Data:** no internet-scale corpus of robot actions

SHRDLU

- **Long-standing goal:** talk to a robot in natural language
- Simulated arm in a blocks world
- Parsed commands, answered questions, planned moves
 - *“Find a block which is taller than the one you are holding and put it into the box.”*
- Hand-written grammar and world model; didn't scale



Person: Pick up a big red block.

Computer: OK.

Person: Grasp the pyramid.

Computer: I don't understand which pyramid you mean.

Early VLAs

LLM as
planner

SayCan
2022

ProgPrompt
2022

Learned
policy

RT-1
2022

VIMA
2022

RT-2
2023

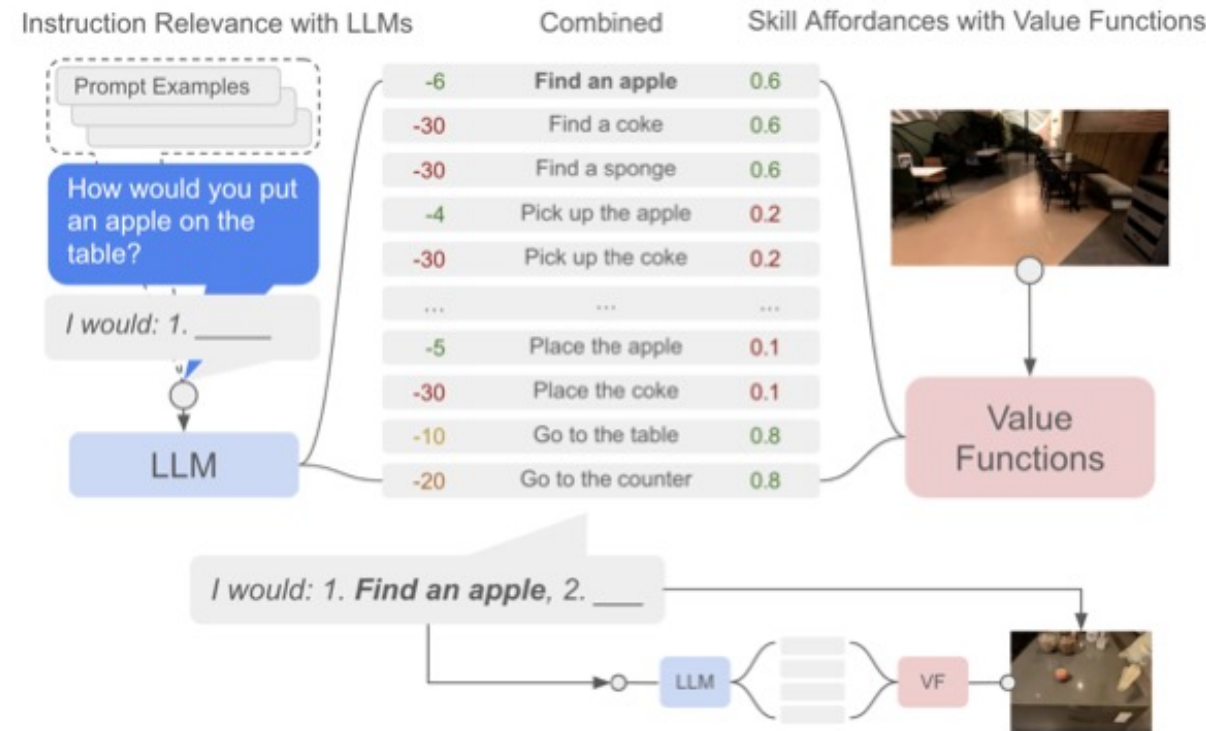


SayCan

- LLM proposes a skill; value function checks feasibility

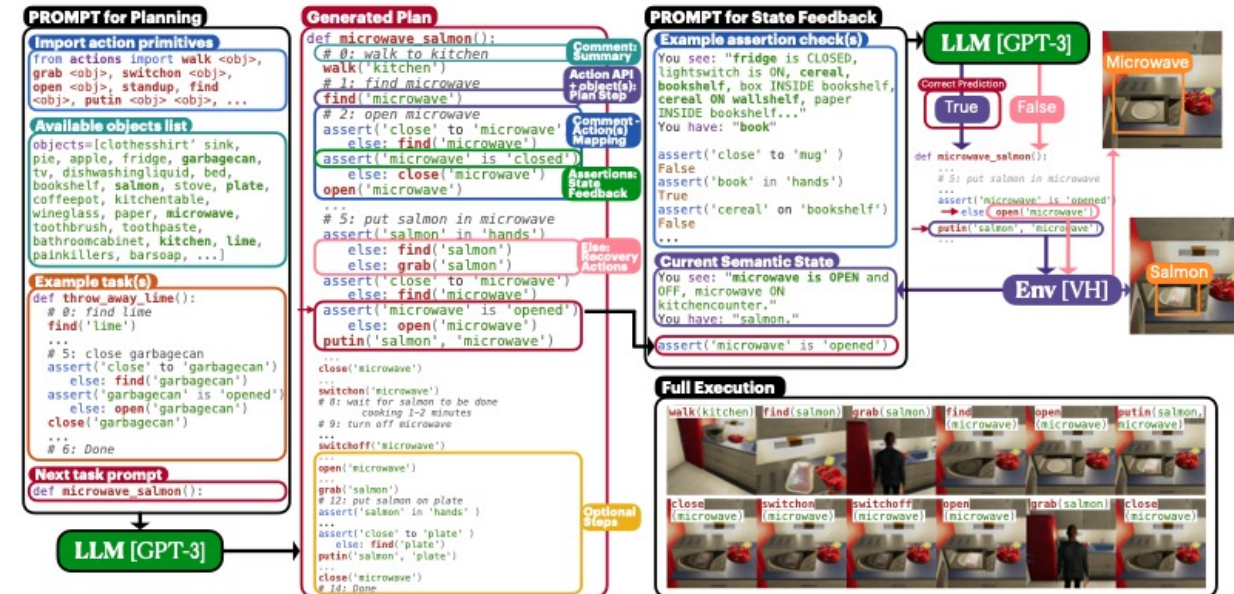
$$k^* = \arg \max_k p_{\text{LLM}}(k \mid \ell) \cdot V(o_t, k)$$

- Say** = useful for the task; **Can** = feasible now
- 551 pretrained skills, 101 kitchen tasks
- Affordances nearly double success



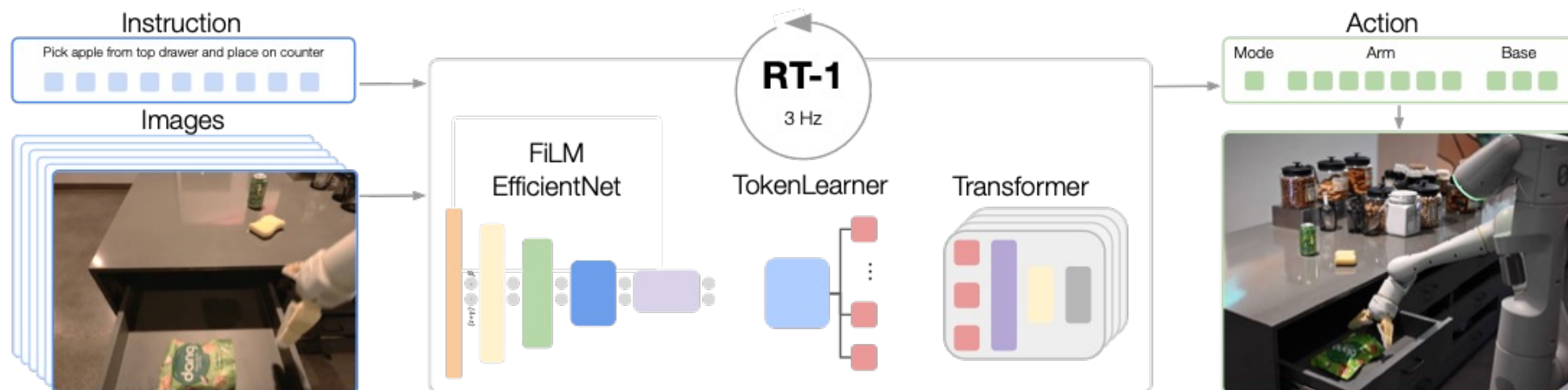
ProgPrompt

- Prompt the LLM with code
 - Available actions as imports, objects as a list
 - A few example programs
- LLM writes a Python plan for the new task
 - Assertions check state and trigger recovery
- Plans use only real actions and objects



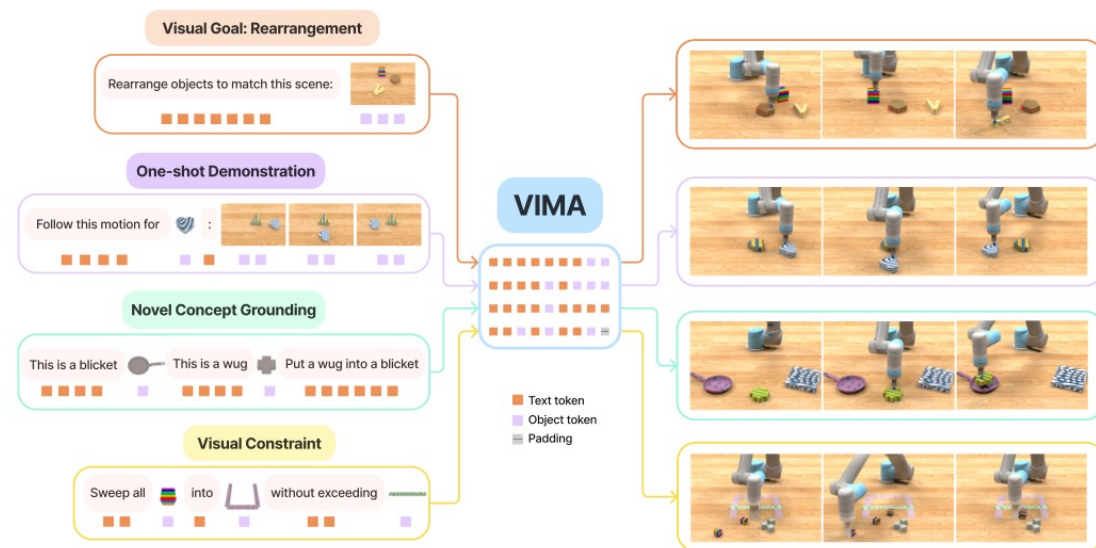
RT-1

- End-to-end: images + instruction → actions
 - EfficientNet with FiLM → TokenLearner → Transformer
- 130K demos, 700+ tasks, 97% success
- 35M params; no web pretraining



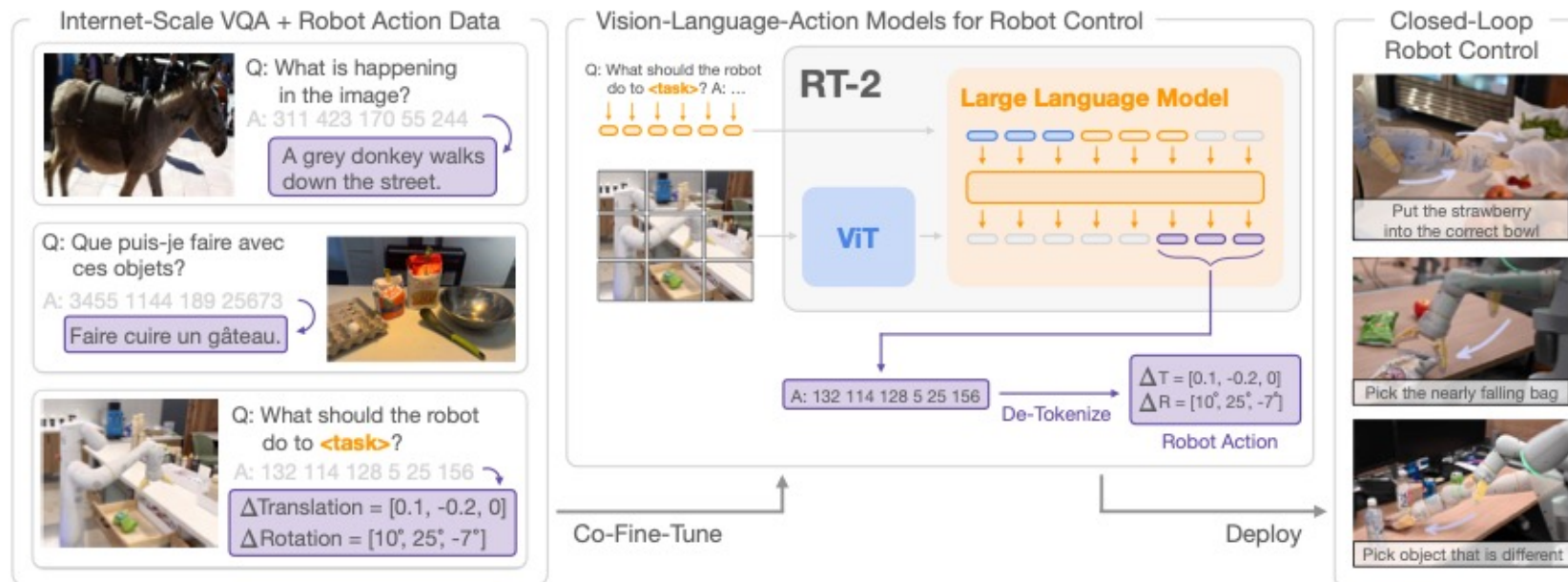
VIMA

- Tasks given as multimodal prompts
 - “Rearrange objects to match this scene: [image]”
- Object tokens + T5 prompt encoder
- Decoder outputs pick-and-place poses
- 600K+ simulated demos
- Up to 2.9× better in zero-shot generalization
 - Language unlocks zero-shot abilities



RT-2: the first “VLA”

- Start from a VLM (PaLI-X 55B, PaLM-E 12B)
- Write actions as text tokens
 - “1 128 91 241 5 101 127 217”
- Co-finetune on web VQA + robot demos
- Emergent skills from web data
 - “Pick up the extinct animal”
- Slow: 1–3 Hz for the 55B model



VLAAs now

Model	VLM backbone	Action output	Weights
RT-2 (2023)	PaLI-X, PaLM-E	Discrete tokens	Closed
Octo (2024)	None (small transformer)	Diffusion head	Open
OpenVLA (2024)	Llama 2 7B	Discrete tokens	Open
π 0 (2024)	PaliGemma 3B	Flow matching	Open
π 0.5 (2025)	PaliGemma	Tokens + flow	Open
GR00T N1 (2025)	Eagle-2	Flow-matching DiT	Open
Gemini Robotics 1.5 (2025)	Gemini	Not disclosed	Closed

- Shared data: **Open X-Embodiment**, 1M+ trajectories from 22 robots

Takeaways

- **VLM** = vision encoder + connector + pretrained LLM
- Trained with the same LM loss as text-only models
- Instruction data turns a VLM into an assistant
- **VLA** = VLM whose outputs are robot actions
- Actions as tokens (RT-2, OpenVLA) or flows (π_0)
- Open problems: robot data, speed, safety