Foundation Models for Vision, Language, and Robotics

Sangwoo Mo

Assistant Professor
POSTECH IME

한국통계학회 2025년 동계학술대회에 대해서 알려줘

Show thinking

한국통계학회 2025년 동계학술대회는 **바로 내일인 12월 19일(금)부터** 서울대학교에서 개최됩니다.

현재 사전등록 기간은 마감되었으므로, 참석을 원하시면 **현장 등록**을 하셔야 합니다. 주요 일정과 정보를 정리해 드립니다.

1. 개최 개요

- **일시:** 2025년 12월 19일(금) ~ 12월 20일(토) (2일간)
- **장소:** 서울대학교 관악캠퍼스
- **주최:** (사)한국통계학회
- **주관:** (사)한국통계학회, 서울대학교 통계학과, 서울대학교 과학데이터혁신연구소

2. 등록 안내 (현장 등록)

사전등록은 11월 26일부로 마감되었으며, 행사 당일 현장에서 등록이 가능합니다.

- **일반(회원):** 130,000원
- **학생(회원):** 80,000원
  - *비회원 등록비는 상이할 수 있으니 현장에서 확인이 필요합니다.*
  - *리셉션(30,000원)은 사전등록자에 한해 운영되므로 현장 등록이 불가할 수 있습니다.*

3. 주요 프로그램 및 일정

이번 학술대회는 정기총회를 겸하여 진행되며, 다양한 세션이 마련되어 있습니다.

- **논문 발표:** 구두 발표 및 포스터 발표 (포스터 파일 제출은 12월 7일 마감됨)
- **주요 행사:**
  - 튜토리얼 및 초청 강연 (상세 연사 정보는 학회 홈페이지 일정표 참조)
  - 정기총회
  - 분과별 세션 (통계학 이론, 응용, 데이터 사이언스 등)

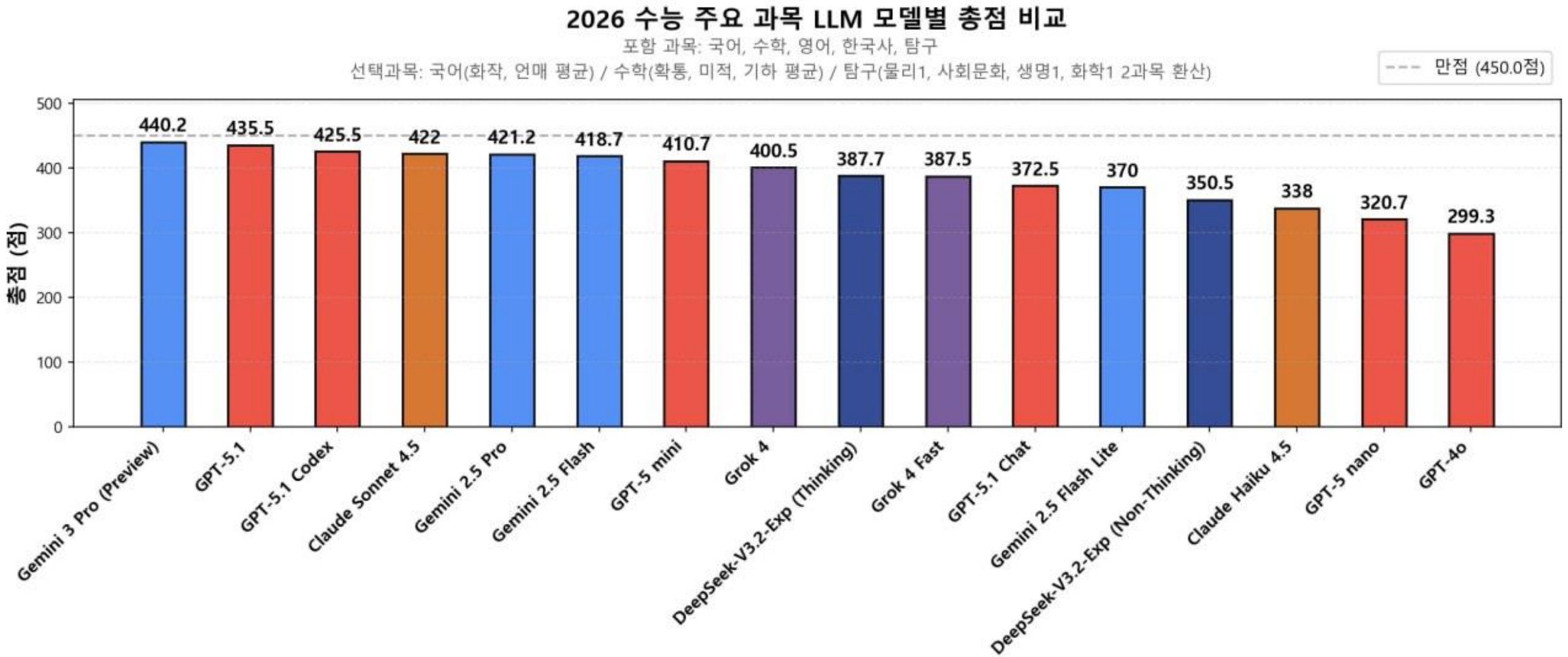
Ask Gemini

Tools

Pro

“2025년 한국통계학회 동계학술대회 현장 풍경을 그려줘”




2026 수능 주요 과목 LLM 모델별 총점 비교
포함 과목: 국어, 수학, 영어, 한국사, 탐구
선택과목: 국어(화작, 언매 평균) / 수학(확통, 미적, 기하 평균) / 탐구(물리1, 사회문화, 생명1, 화학1 2과목 환산)
만점 (450.0점)
총점 (점)
500
400
300
200
100
0
440.2
435.5
425.5
422
421.2
418.7
410.7
400.5
387.7
387.5
372.5
370
350.5
338
320.7
299.3
Gemini 3 Pro (Preview)
GPT-5.1
GPT-5.1 Codex
Claude Sonnet 4.5
Gemini 2.5 Pro
Gemini 2.5 Flash
GPT-5 mini
Grok 4
DeepSeek-V3.2-Exp (Thinking)
Grok 4 Fast
GPT-5.1 Chat
Gemini 2.5 Flash Lite
DeepSeek-V3.2-Exp (Non-Thinking)
Claude Haiku 4.5
GPT-5 nano
GPT-4o

Problem Statement

Let C denote the largest constant for which one has

$$\max_{-1/2\le t\le 1/2}\int_{\mathbb{R}} f(t-x)f(x)dx \ge C\left(\int_{-1/4}^{1/4} f(x)dx\right)^2$$

for all non-negative $f:\mathbb{R}\to\mathbb{R}$. What is C?

Comments

Upper bounds for C

Bound	Source	Date
$\pi/2 = 1.57059$	Schinzel-Schmidt	2002
1.50992	Matolcsi-Vinuesa	8 Jul 2009
1.5032	Georgiev-Gómez-Serrano-Tao-Wagner [Colab 1] [Colab 2]	14 May 2025

Lower bounds for C

Bound	Source	Date
1	Trivial	
1.182778	Martin-O'Bryant	30 Sep 2004
1.262	Martin-O'Bryant	1 Apr 2009
1.2748	Matolcsi-Vinuesa	8 Jul 2009

See also Damek Davis's meta-analysis of this problem.

Resources

Explore in Google Colab

See Section 6.2 in "Mathematical Exploration and Discovery at Scale" paper




autonomous, 2x speed

FM Landscape

Perception	Generation	Action
		

FM Landscape

	Perception	Generation	Action
Vision			
Language			







Disclaimer

This talk focuses on “what FMs are”,

not on “how they work”

Visual Perception

	Perception	Generation	Action
Vision			
Language			

Visual Perception

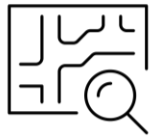
Abstract Perception

“what” (high-level, conceptual)



Dense Perception

“where” (low-level, spatial)



Abstract Perception

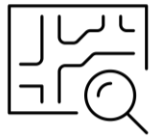
Abstract Perception

“what” (high-level, conceptual)

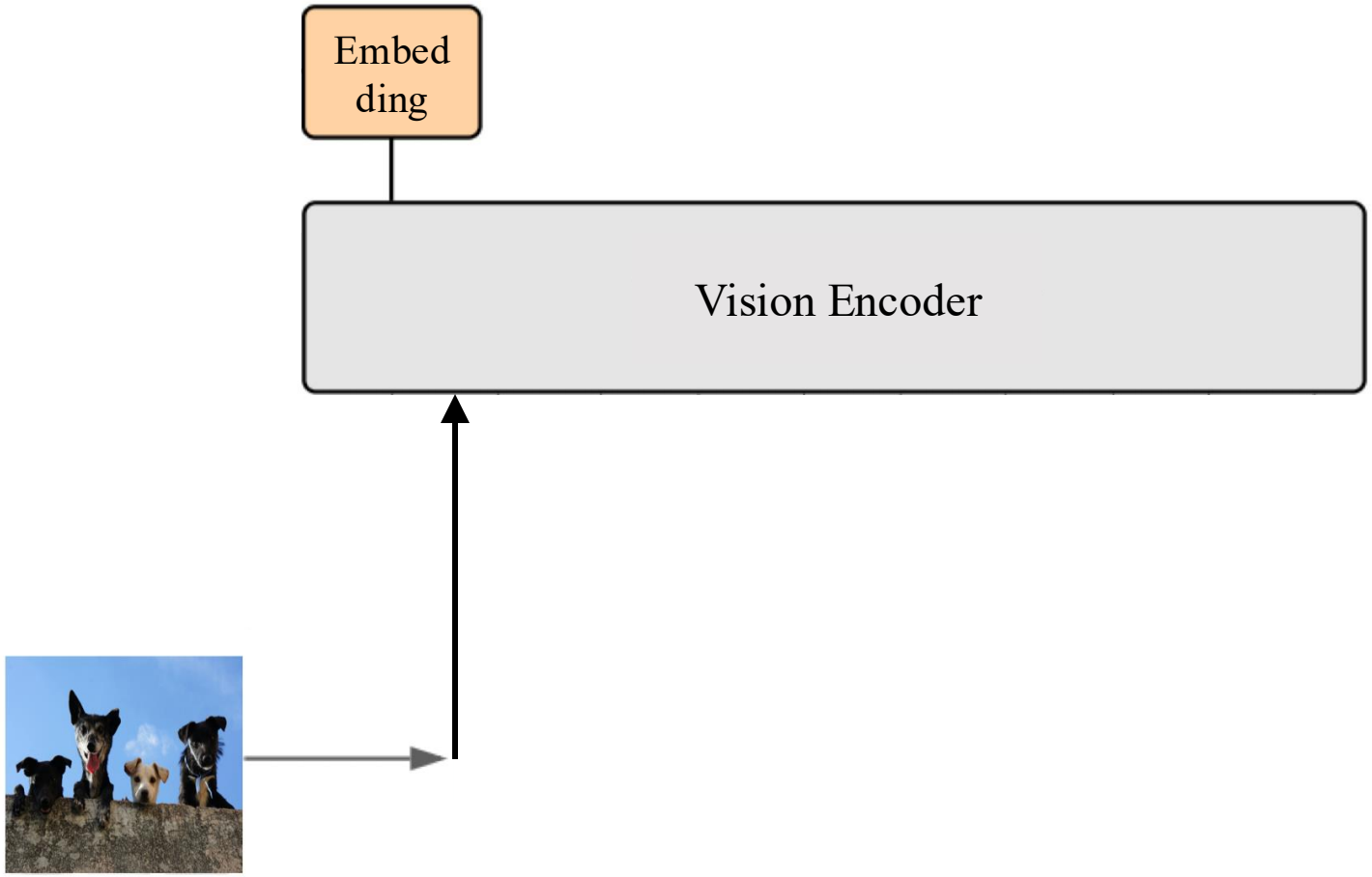


Dense Perception

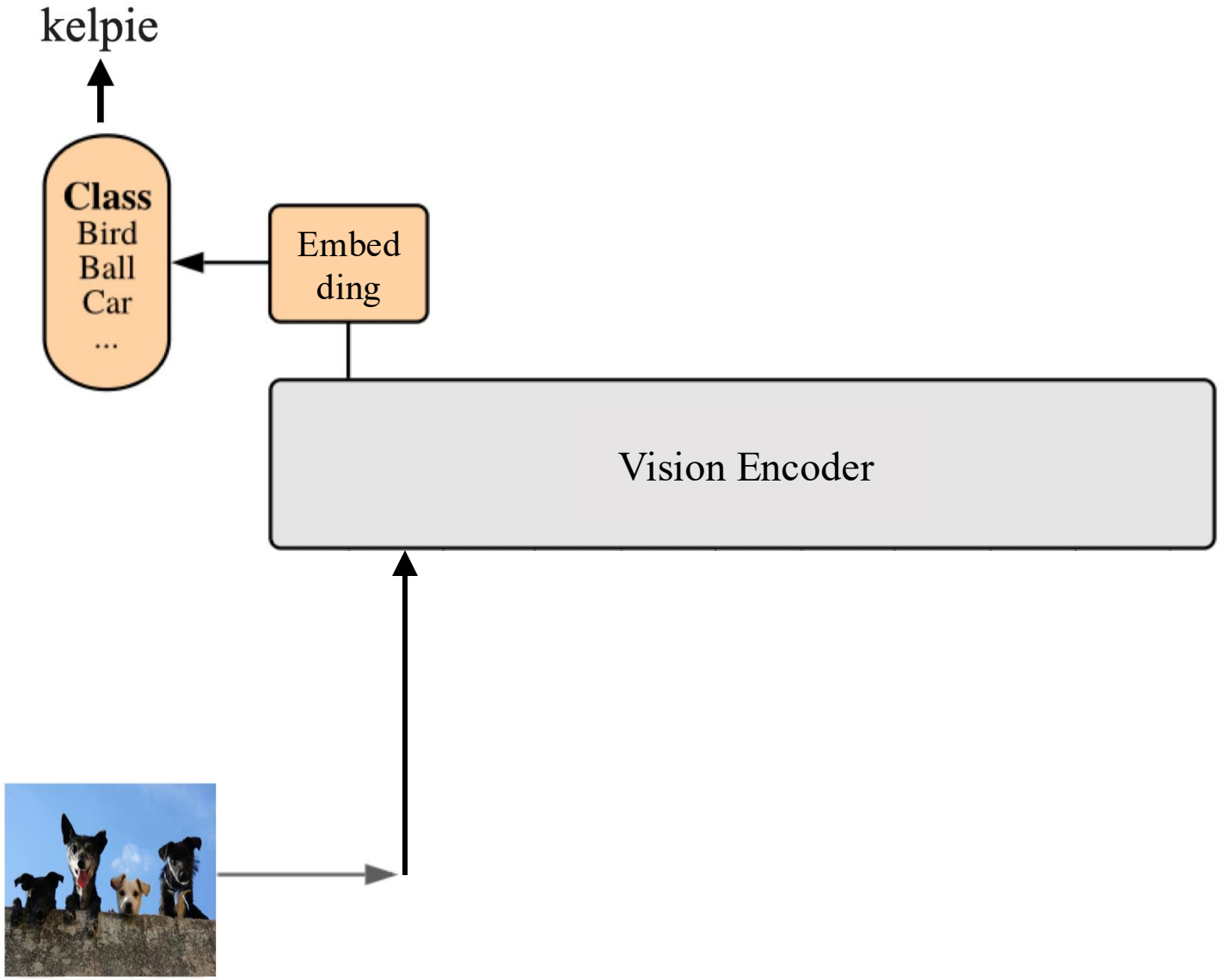
“where” (low-level, spatial)



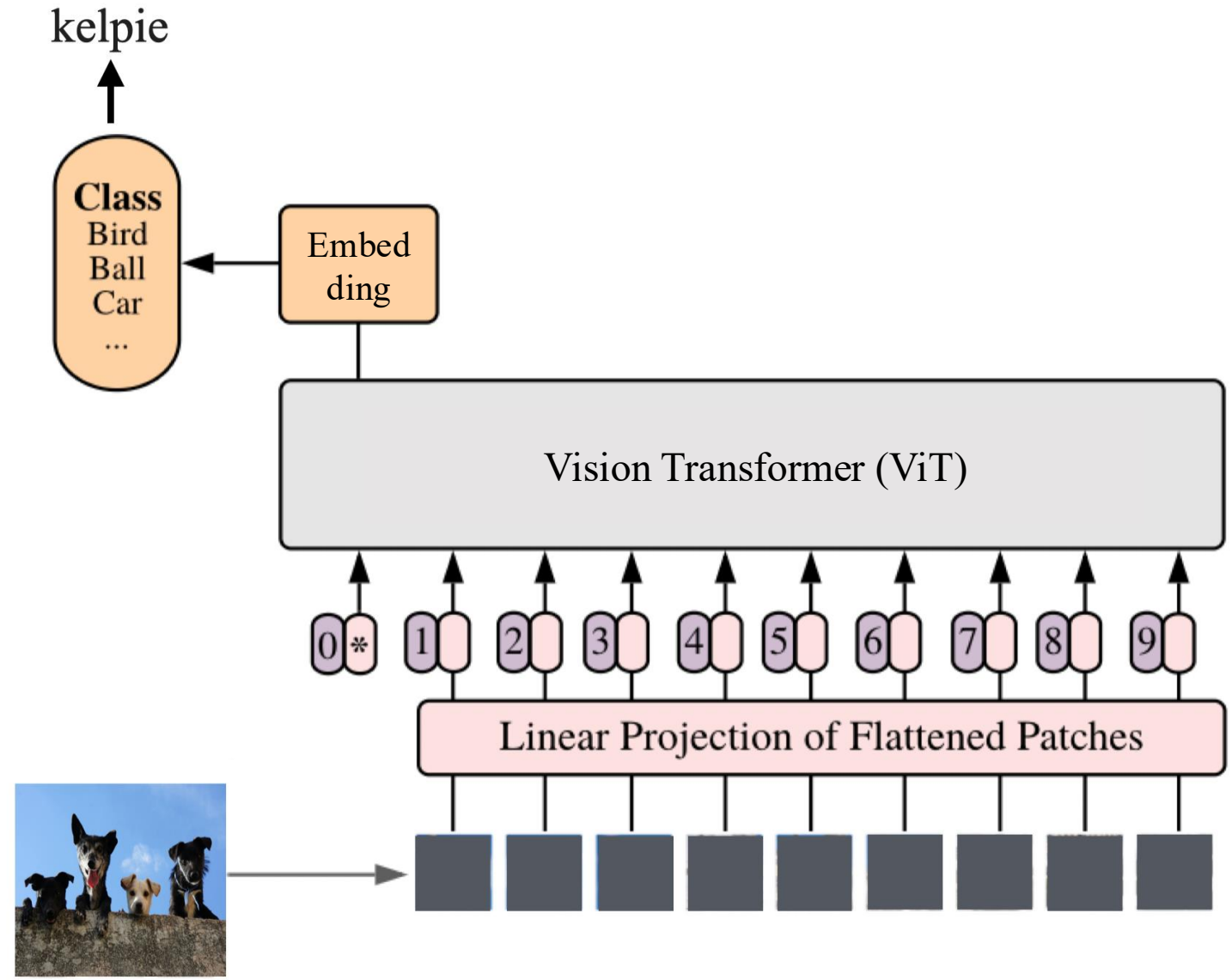
Abstract Perception



Abstract Perception



Abstract Perception



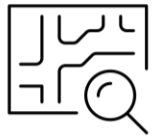
Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021.

Abstract Perception

Abstract Perception
“what” (high-level, conceptual)



Dense Perception
“where” (low-level, spatial)



Visual

Semantic

Abstract Perception – Visual Cue

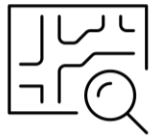
Abstract Perception

“what” (high-level, conceptual)



Dense Perception

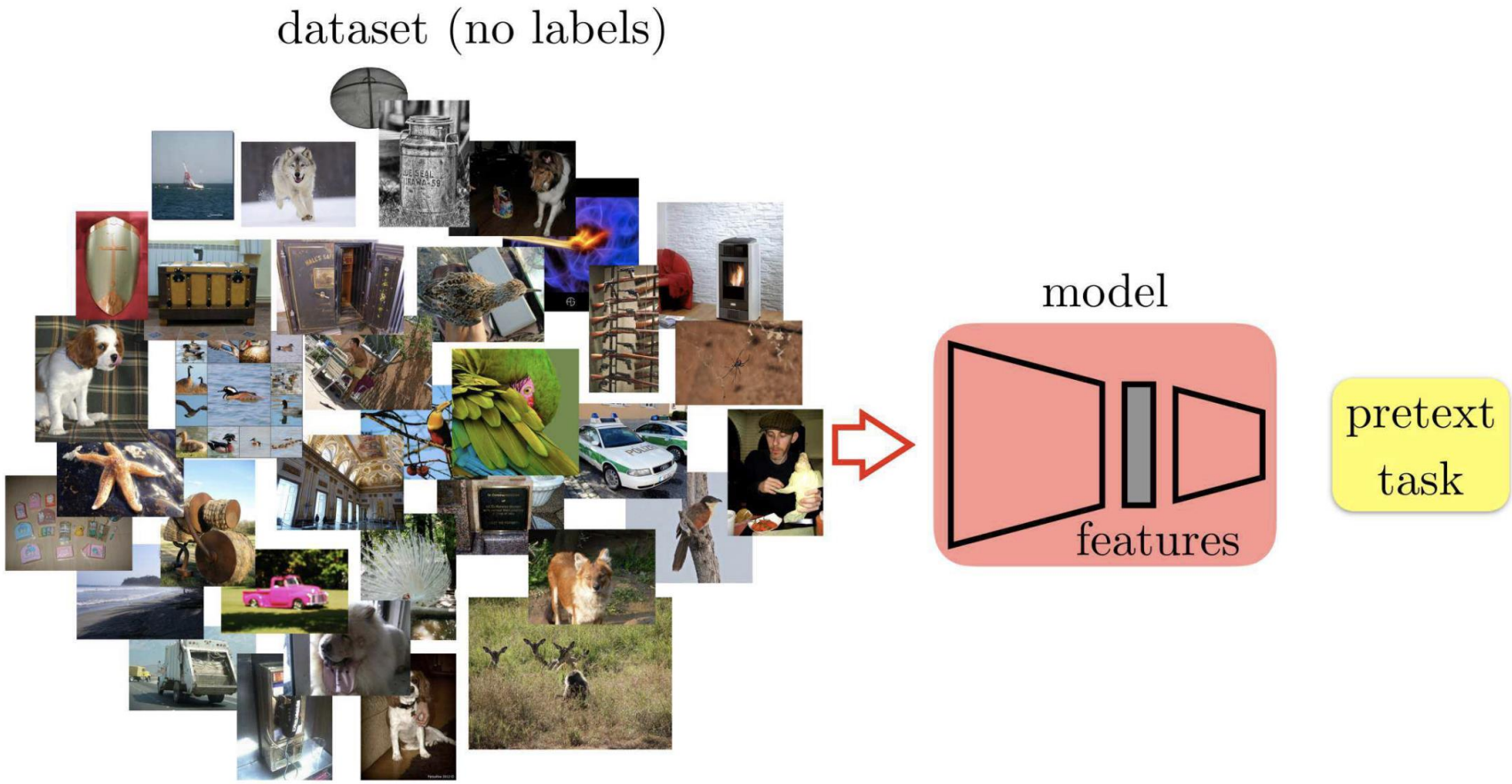
“where” (low-level, spatial)



Visual

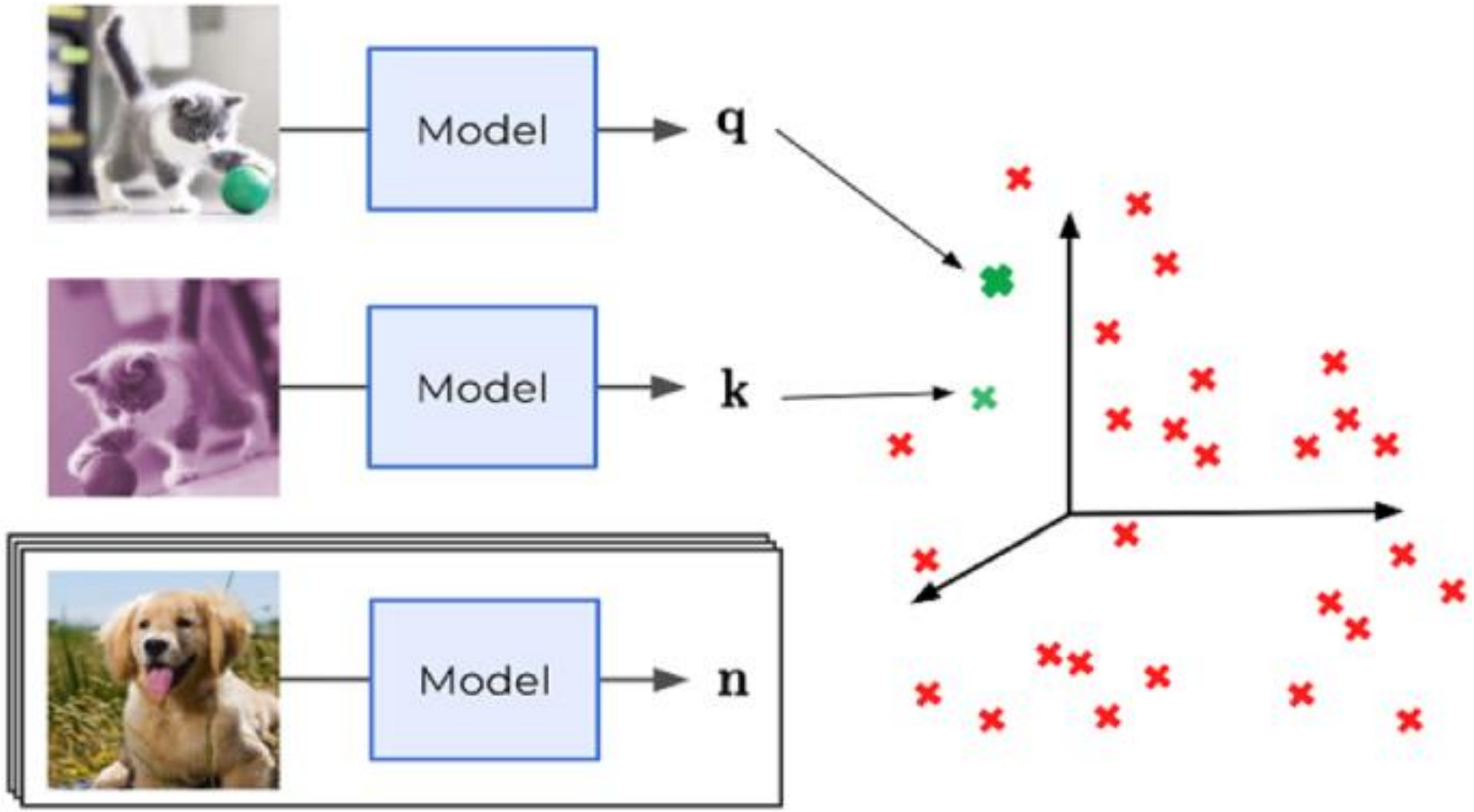
Semantic

Visual Cue: Self-Supervised Learning (e.g., DINO)



Visual Cue: Self-Supervised Learning (e.g., DINO)

Standard for Image Encoder: Contrastive Learning



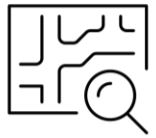
Siméoni et al. DINOv3. arXiv 2025.

Abstract Perception – Semantic Cue

Abstract Perception
“what” (high-level, conceptual)



Dense Perception
“where” (low-level, spatial)

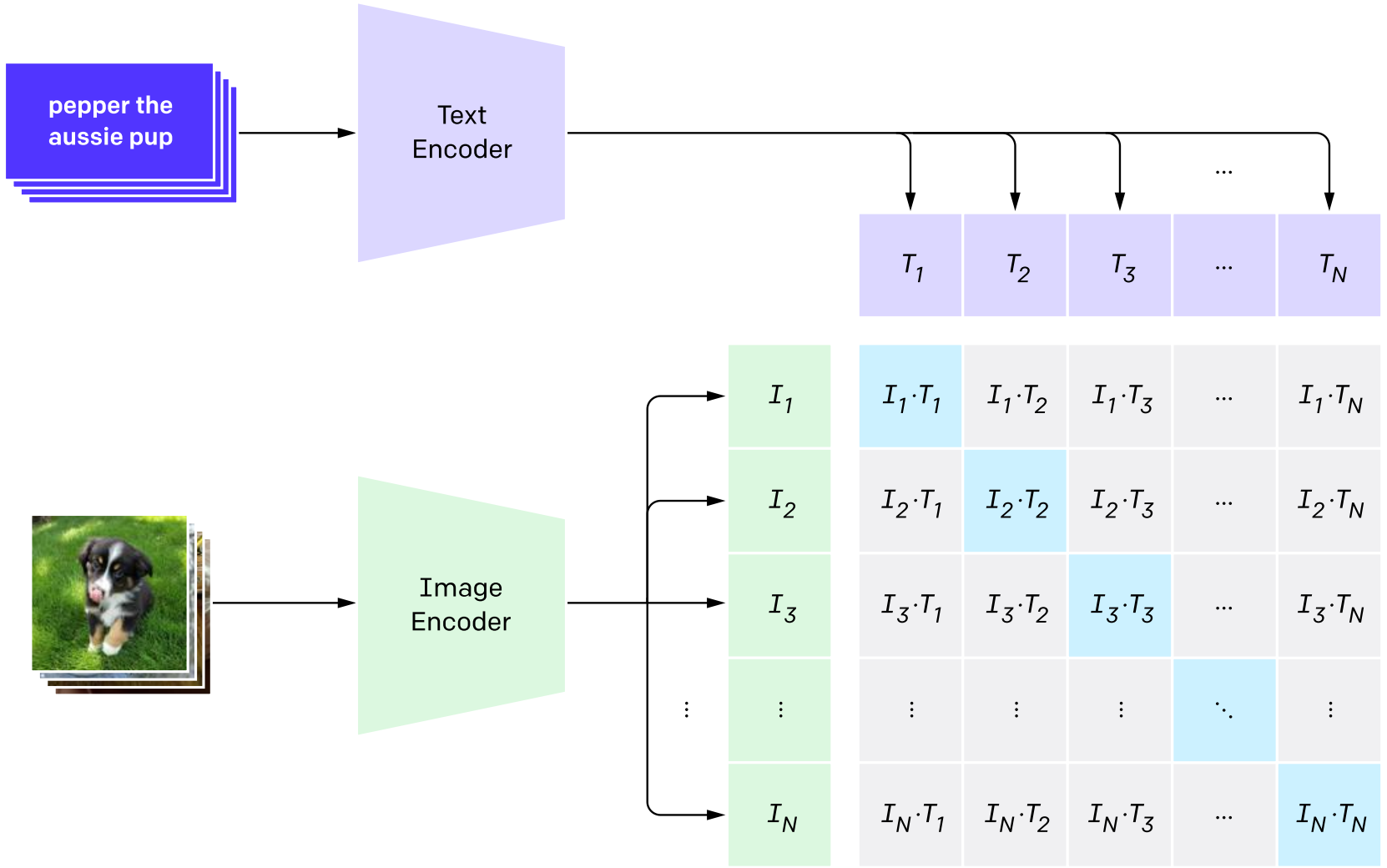


Visual

Semantic

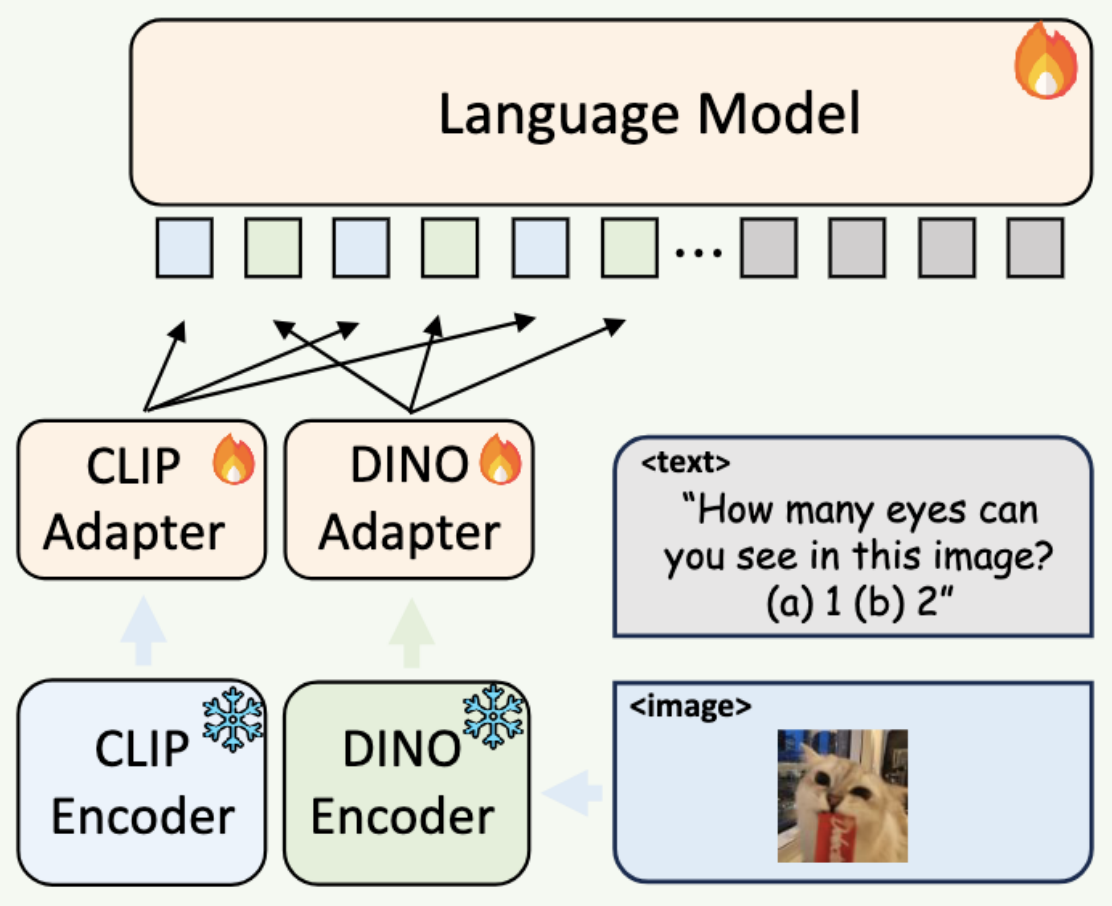
Self-Supervised Learning (e.g., DINO)

Semantic Cue: Vision-Language Model (e.g., CLIP)



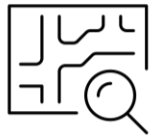
Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.

Complementarity of Visual and Semantic Cues



Tong et al. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. CVPR 2024.

Dense Perception

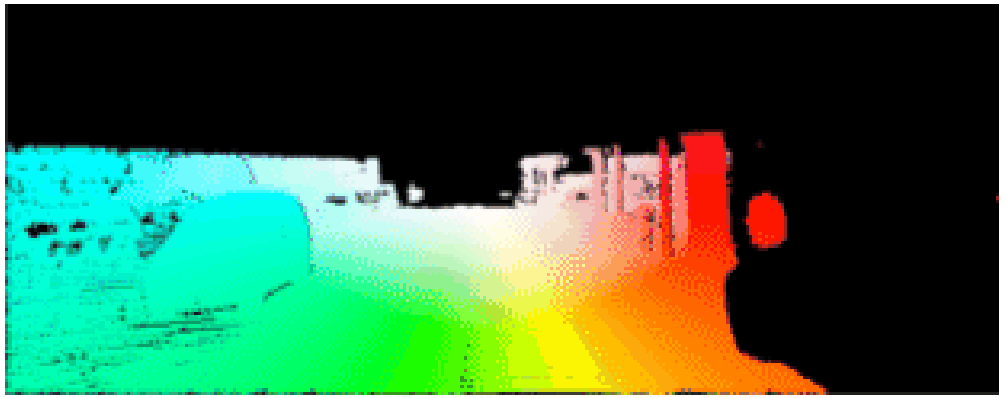
	Abstract Perception “what” (high-level, conceptual)	**Dense Perception** “where” (low-level, spatial)
		
Visual	Self-Supervised Learning (e.g., DINO)	
Semantic	Vision-Language Model (e.g., CLIP)	

Dense Perception

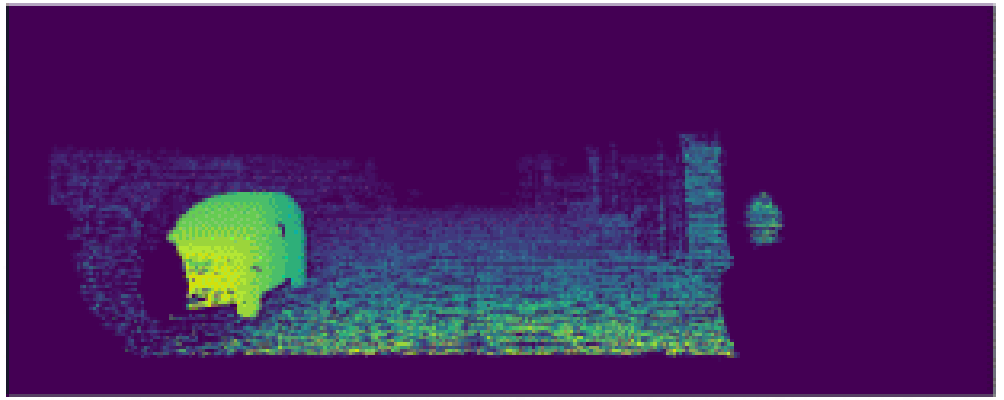
segmentation







optical flow



depth

1. Spatialization of Abstract Perception

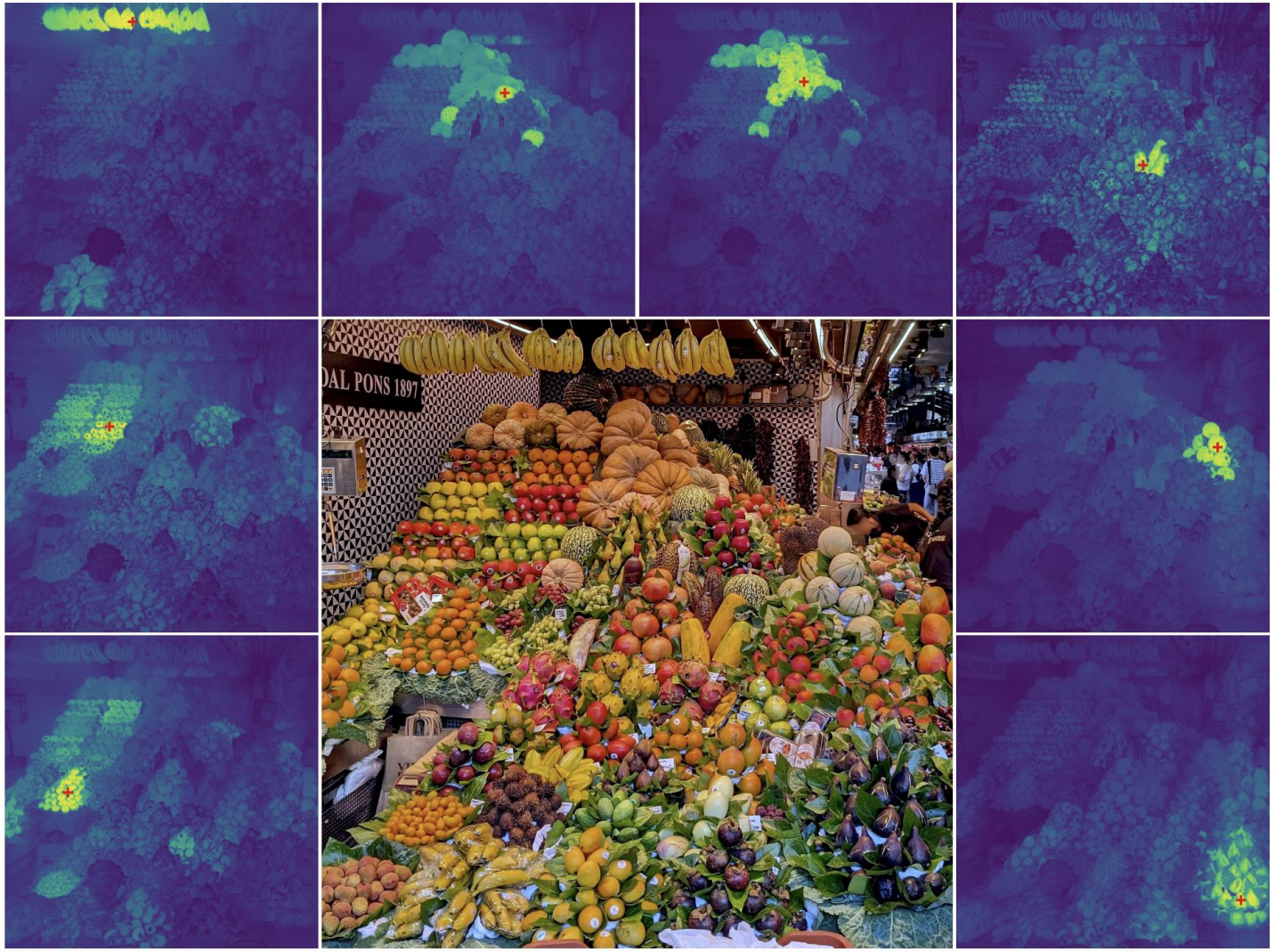


Figure 3: High-resolution dense features. We visualize the cosine similarity maps obtained with DINOv3 output features between the patches marked with a red cross and all other patches. Input image at 4096×4096. *Please zoom in*, do you agree with DINOv3?

2. Dense Supervision (e.g., SAM)



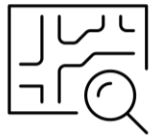
Carion et al. SAM 3: Segment Anything with Concepts. arXiv 2025.

2. Dense Supervision (e.g., SAM)



Lin et al. Depth Anything 3: Recovering the Visual Space from Any Views. arXiv 2025.

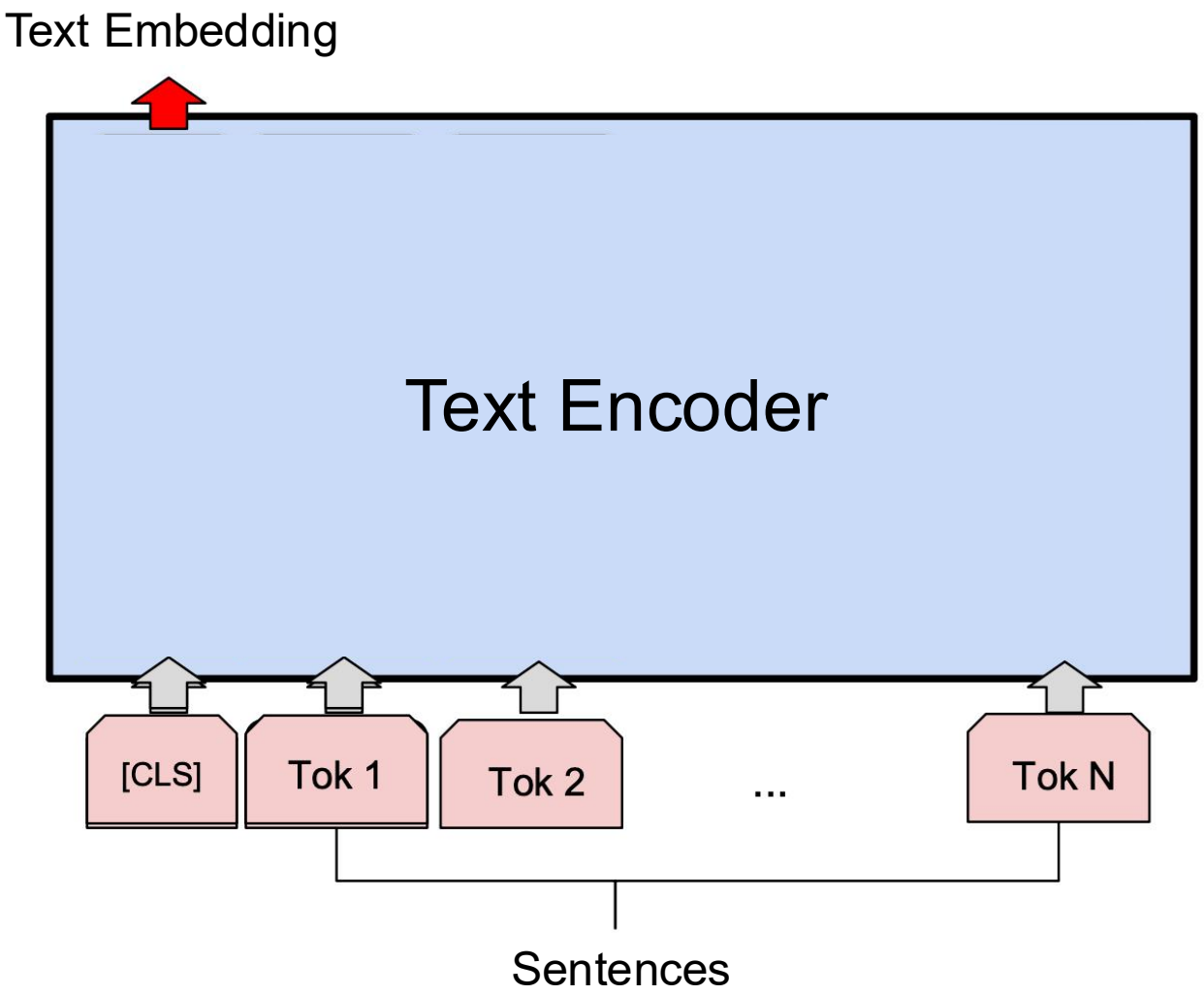
Summary of Part I: Visual Perception

	Abstract Perception “what” (high-level, conceptual) 	**Dense Perception** “where” (low-level, spatial) 
Visual	Self-Supervised Learning (e.g., DINO)	1. Spatialization of Abstract Perception 2. Dense Supervision (e.g., SAM)
Semantic	Vision-Language Model (e.g., CLIP)	

Language Understanding

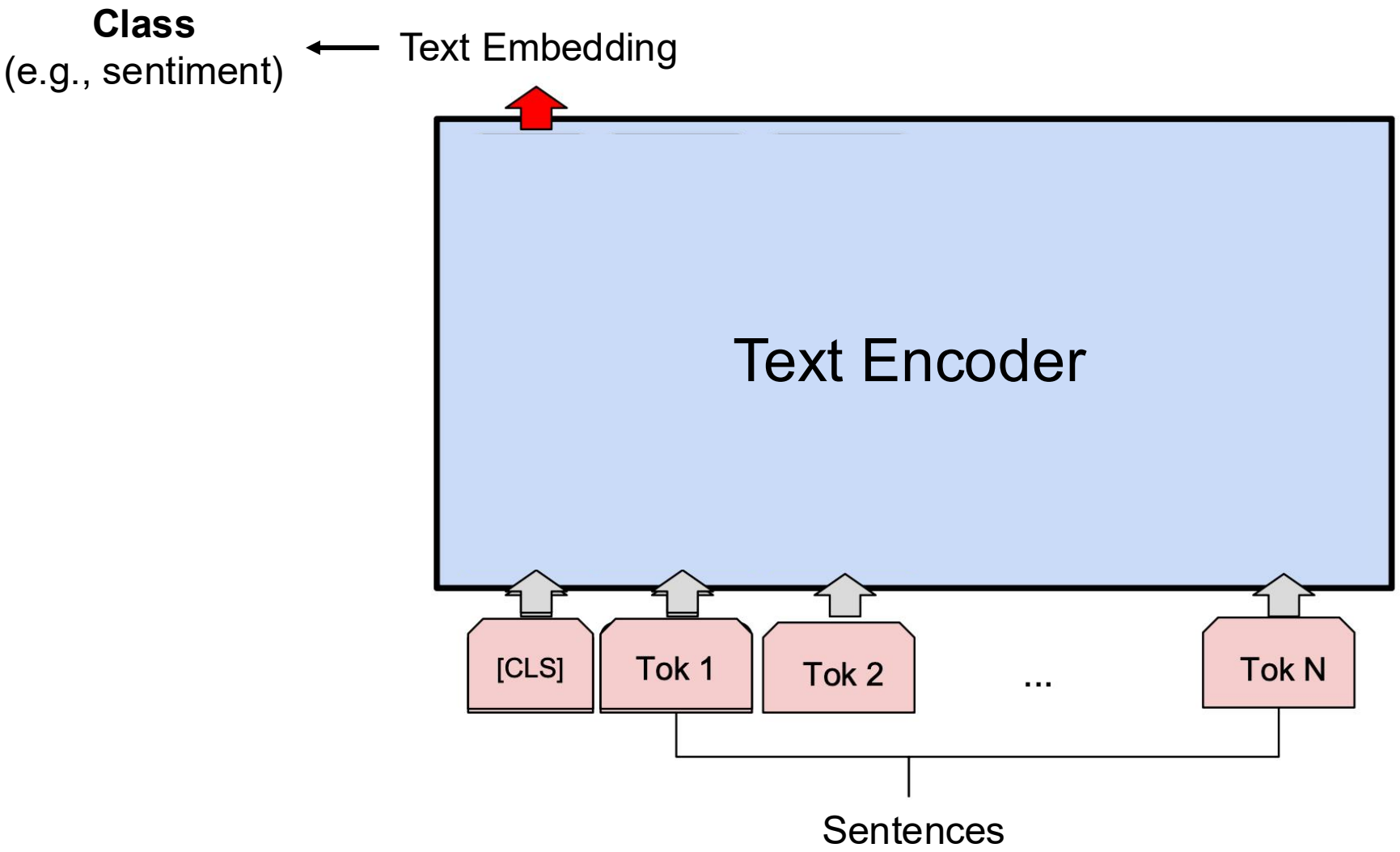
	Perception 	Generation 	Action 
Vision	DINO, CLIP, SAM, etc.		
Language	BERT		

Language Understanding



Delvin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2019.

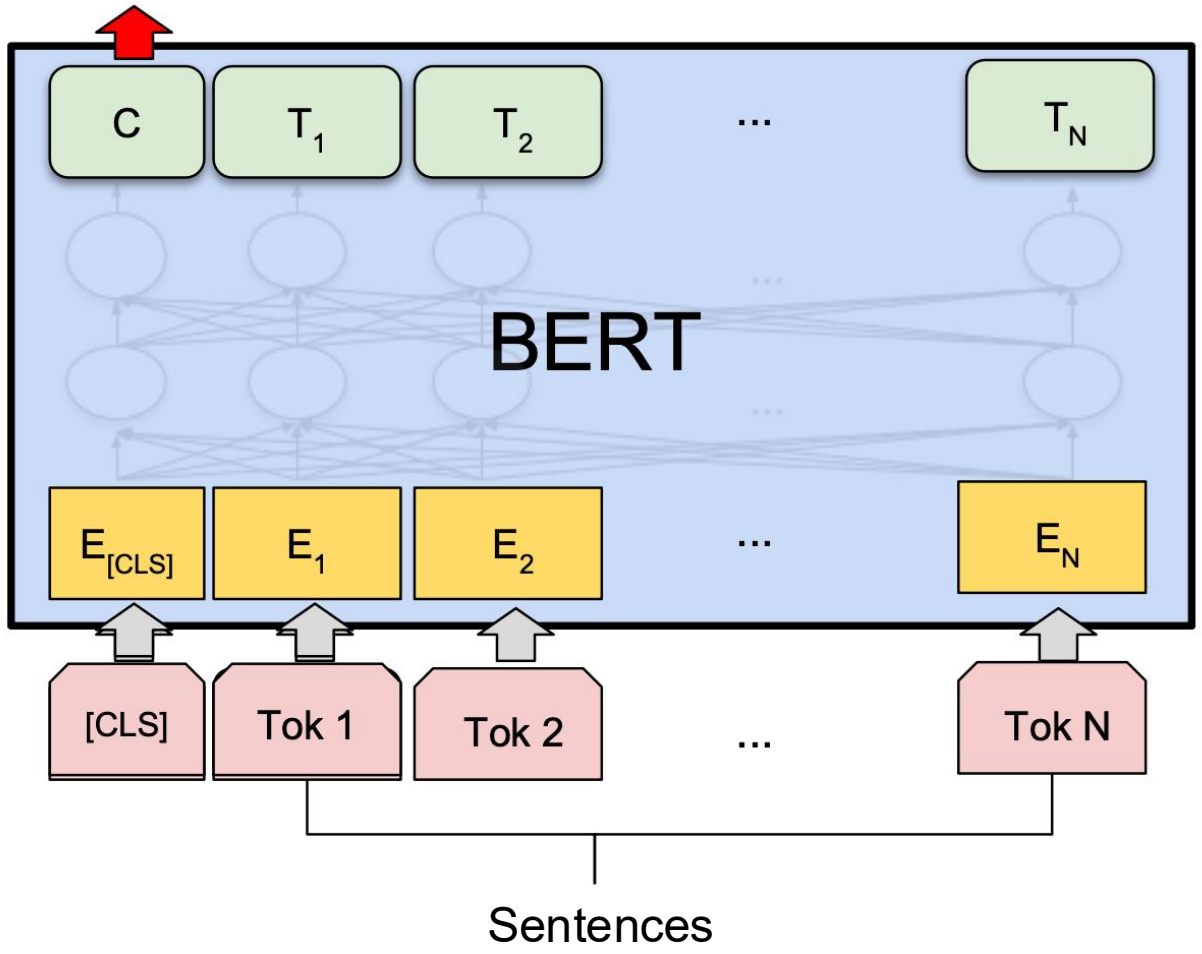
Language Understanding



Delvin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2019.

Language Understanding

Self-Supervised Learning (e.g., BERT)



Delvin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2019.

Language Generation

	Perception	Generation	Action
			
Vision	DINO, CLIP, SAM, etc.		
Language	BERT	LLM	

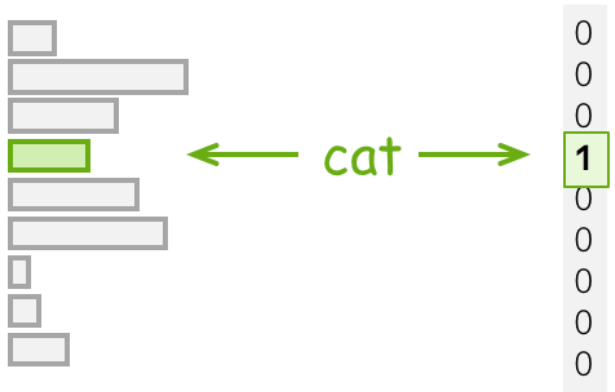
Language Generation

= Repetition of Next Token Prediction

we want the model
to predict this

Training example: I saw a cat on a mat <eos>

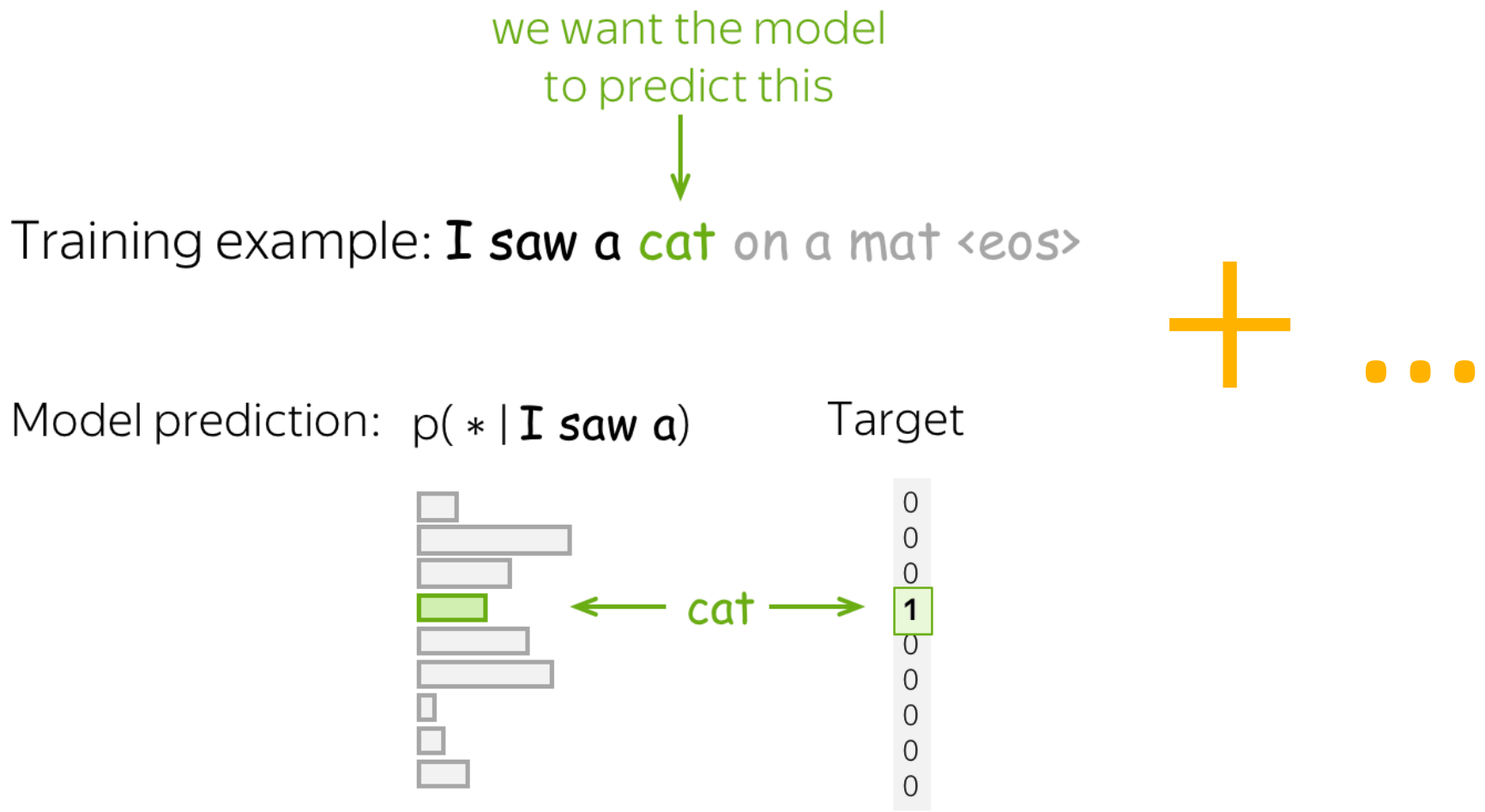
Model prediction: $p(* \mid \text{I saw a})$ Target



Brown et al. Language Models are Few-Shot Learners. NeurIPS 2020.

Language Generation

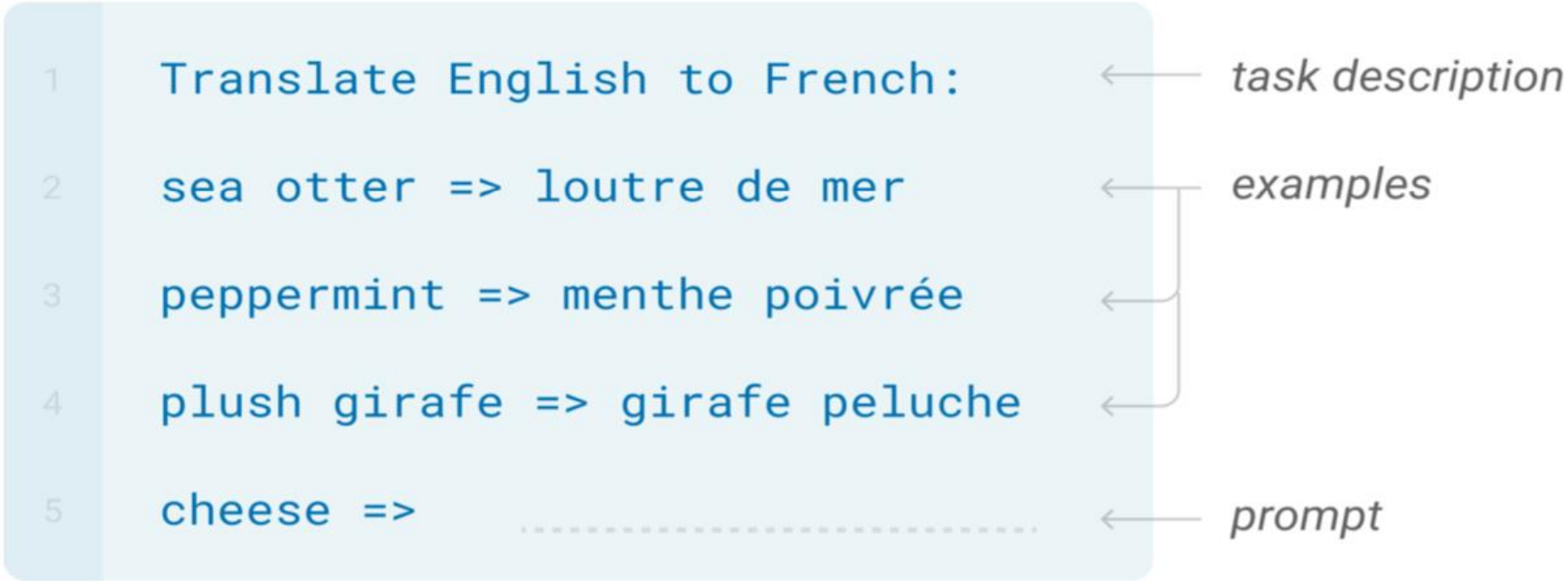
In practice, Pre-train + SFT + RL (alignment / reasoning)



Brown et al. Language Models are Few-Shot Learners. NeurIPS 2020.

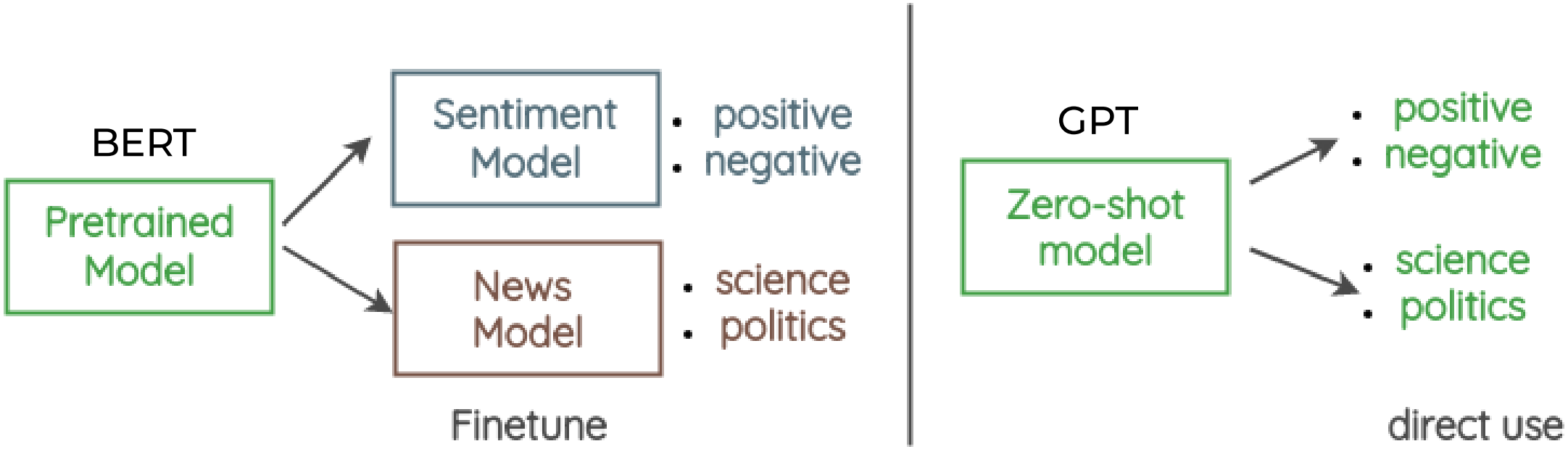
In-Context Learning (ICL): New Paradigm of FM

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



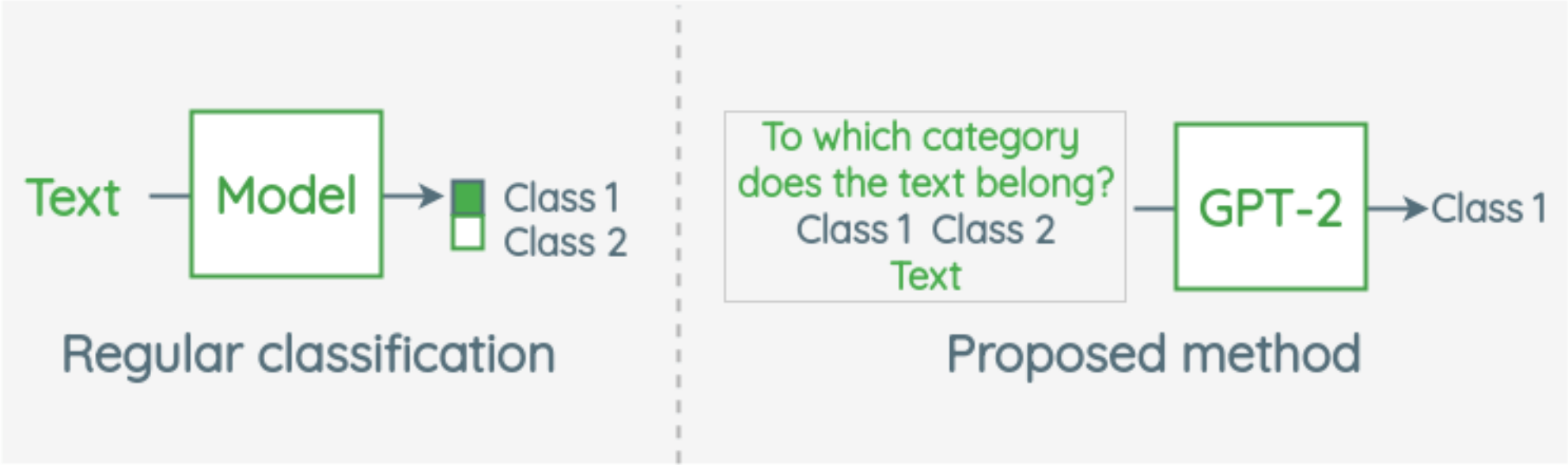
Brown et al. Language Models are Few-Shot Learners. NeurIPS 2020.

In-Context Learning (ICL): New Paradigm of FM

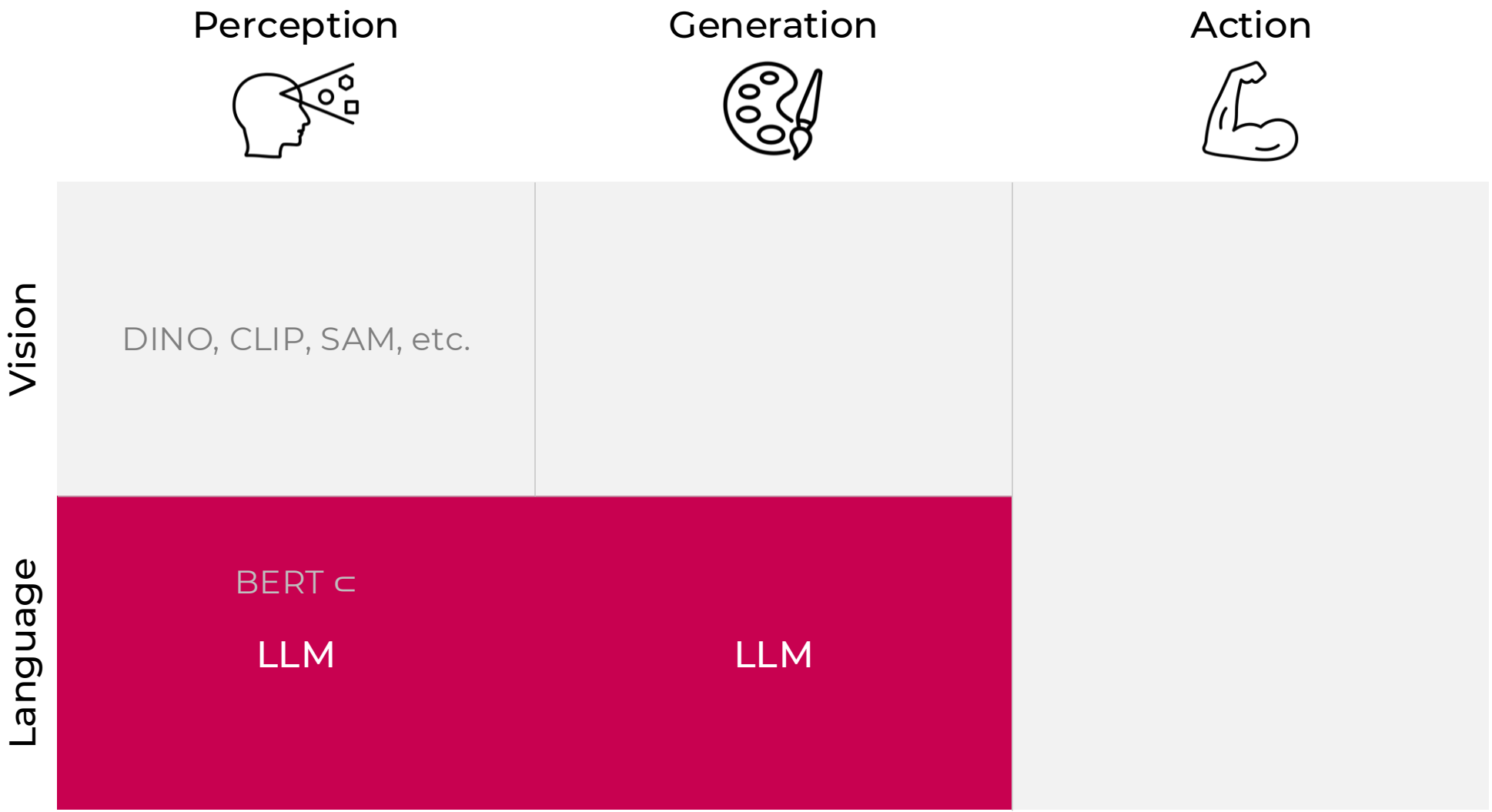


Brown et al. Language Models are Few-Shot Learners. NeurIPS 2020.

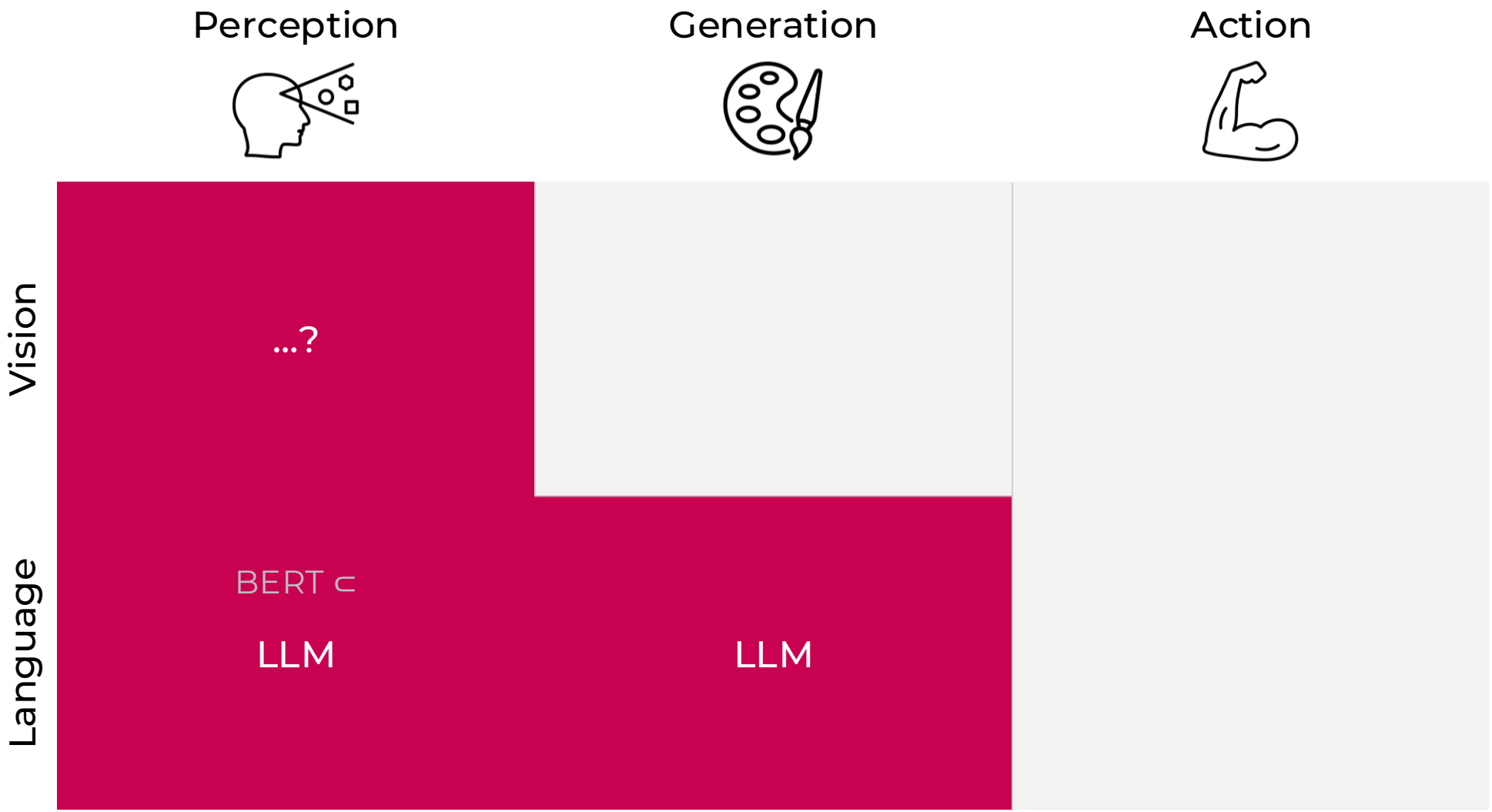
Understanding ⊂ Generation



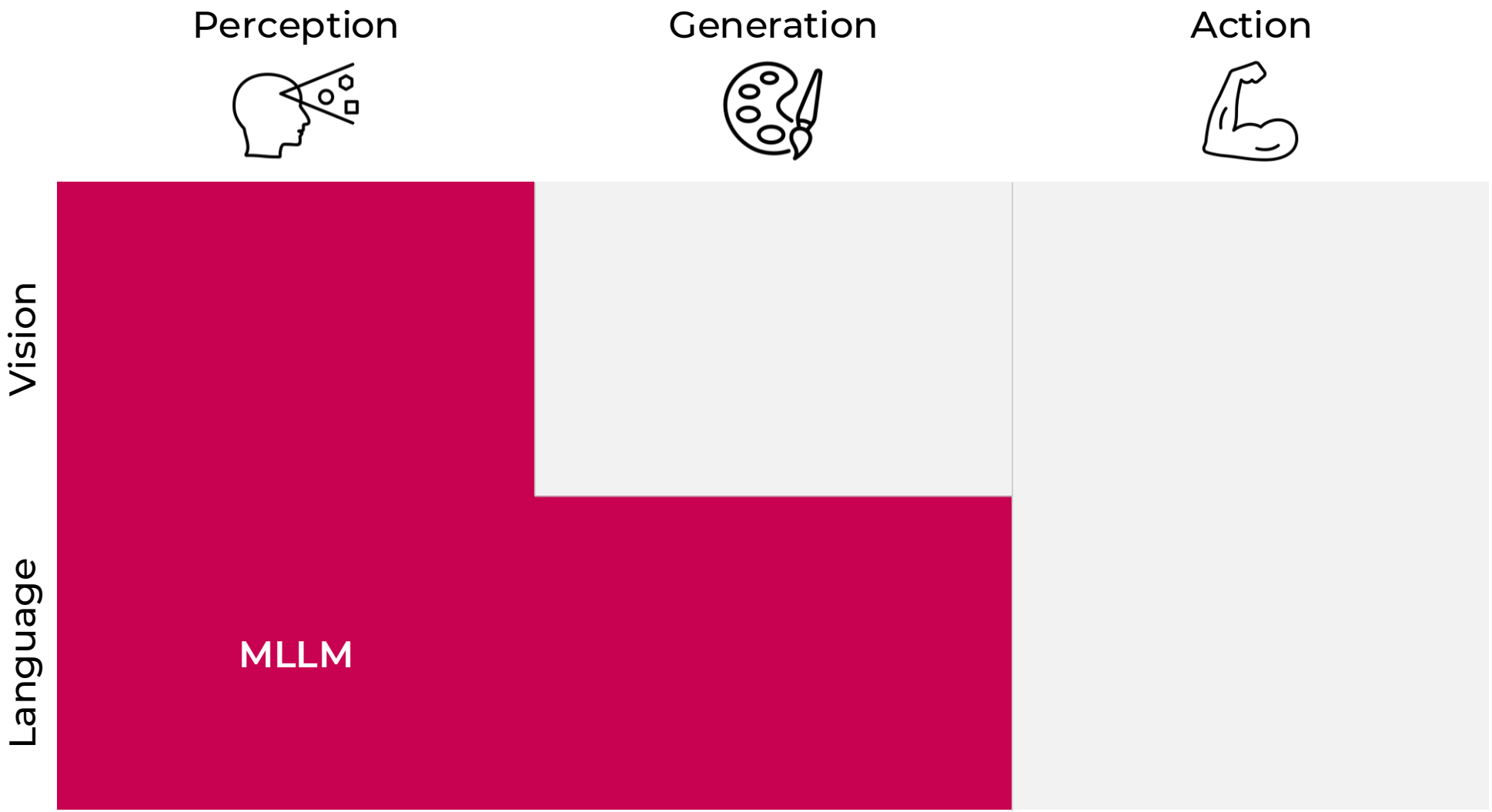
Understanding ⊂ Generation



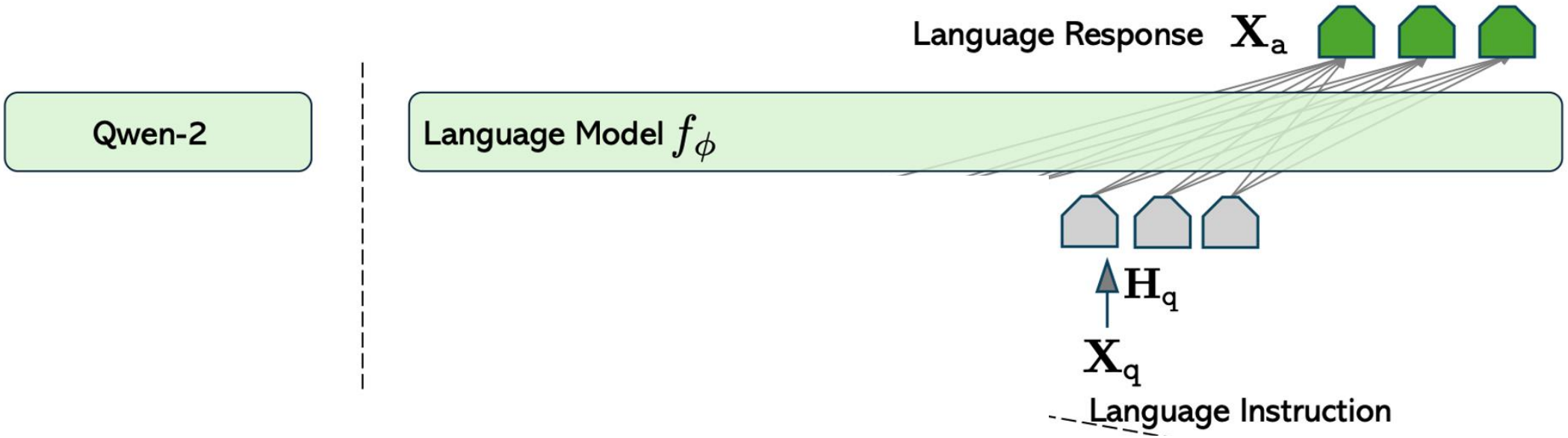
LLM + Visual Perception?



Multimodal LLM

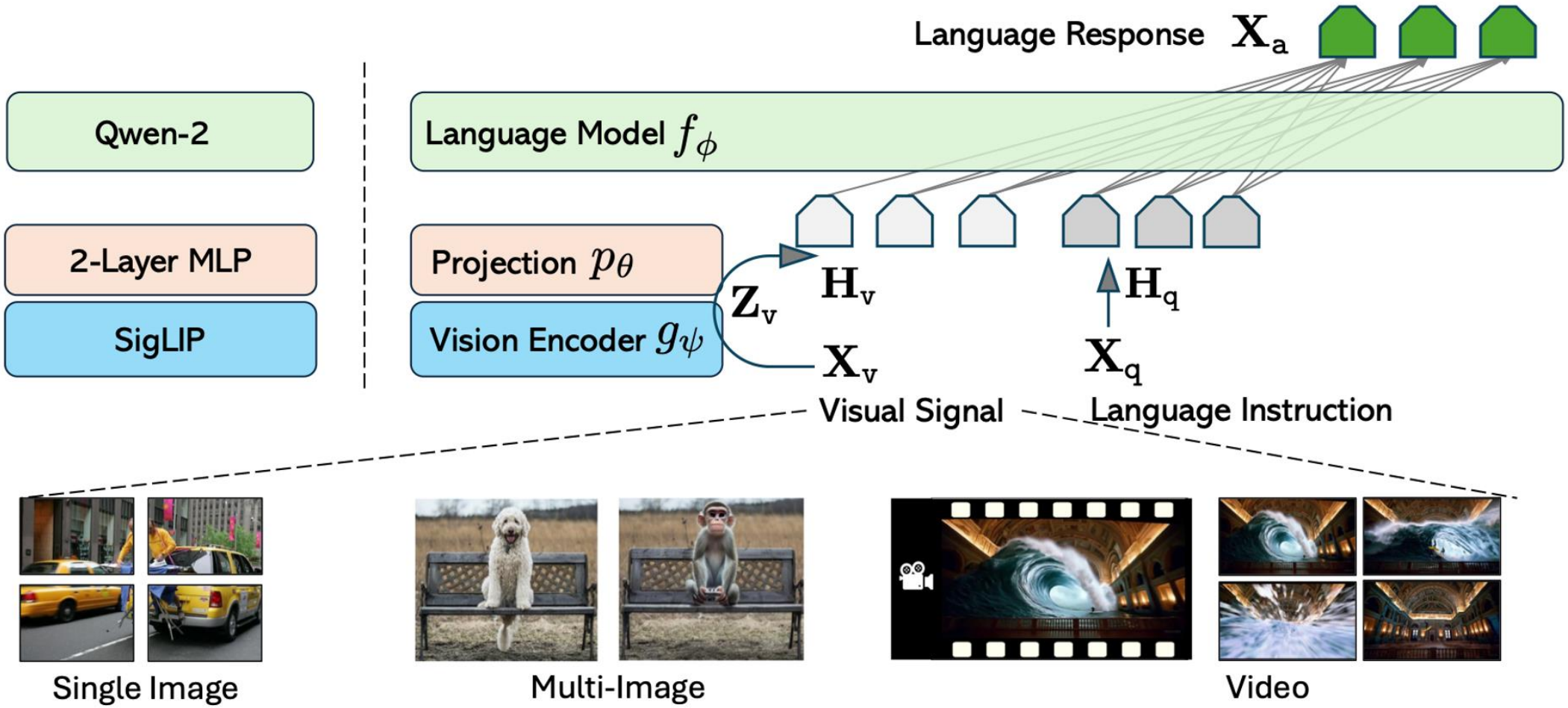


Multimodal **LLM**



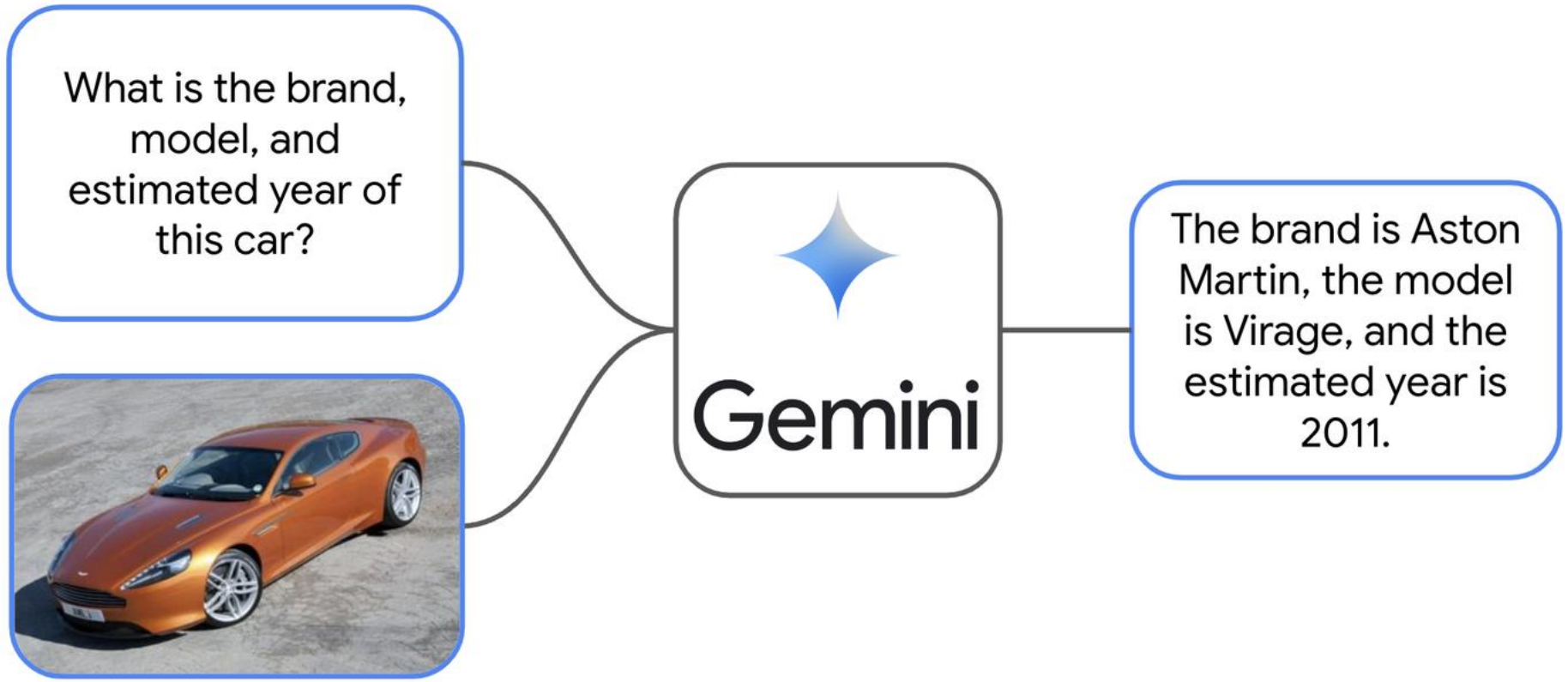
Liu et al. Visual Instruction Tuning. NeurIPS 2023.

Multimodal LLM



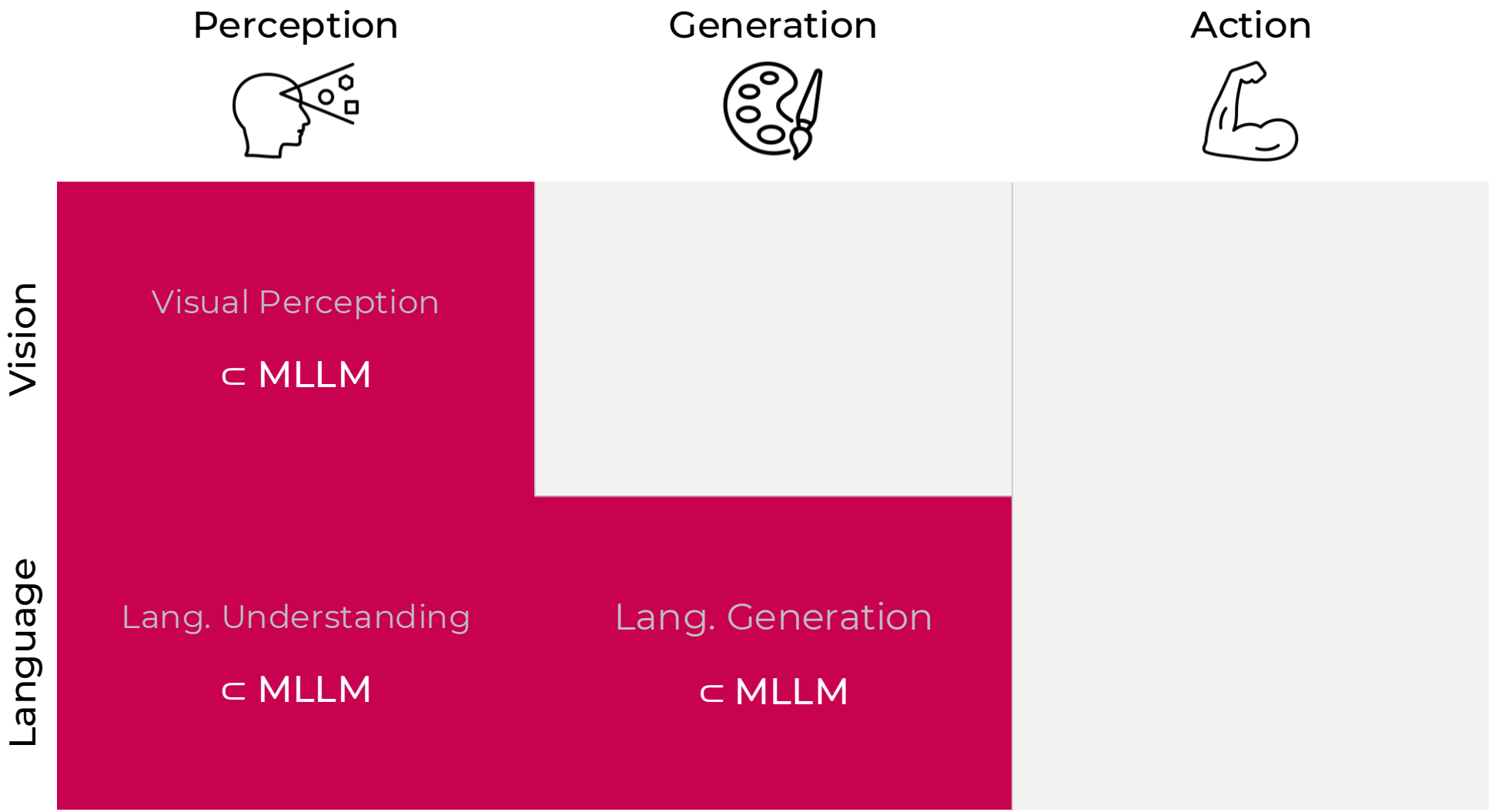
Liu et al. Visual Instruction Tuning. NeurIPS 2023.

Multimodal LLM: Visual Reasoning



Liu et al. Visual Instruction Tuning. NeurIPS 2023.

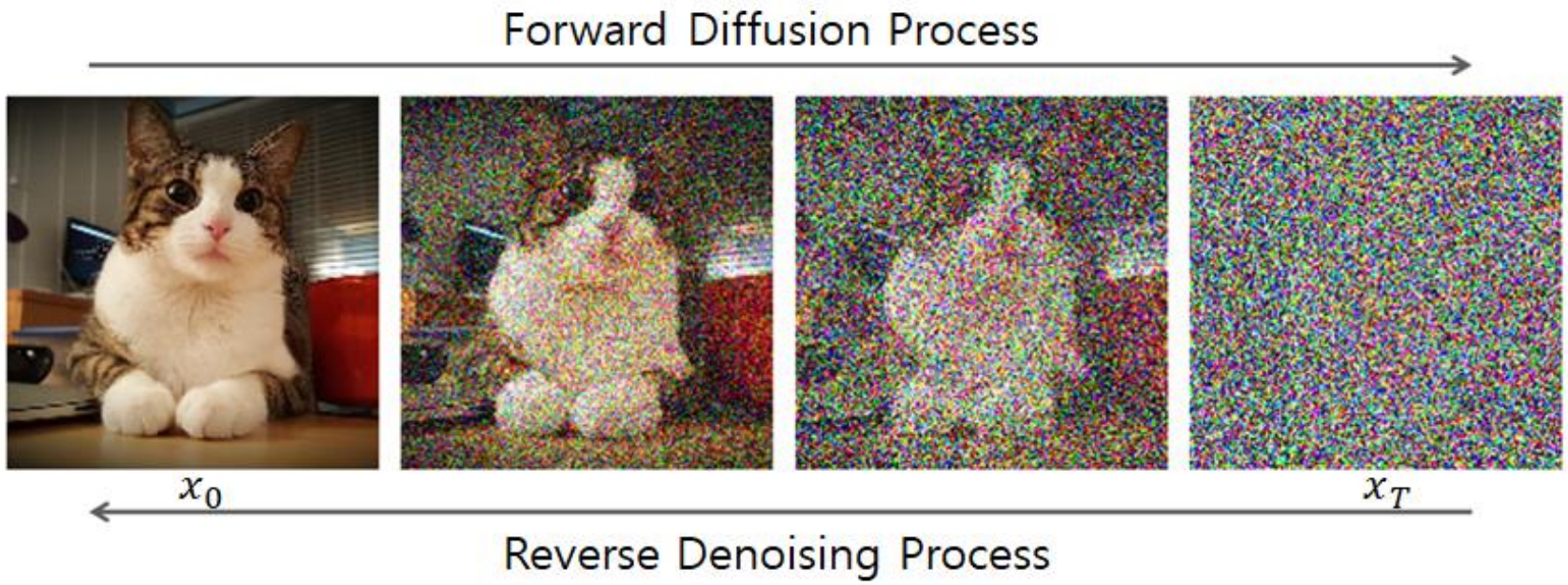
Summary of Part II: Multimodal LLM



Visual Generation

	Perception 	Generation 	Action 
Vision	MLLM		
Language	MLLM	MLLM	

Current Standard: Diffusion / Flow Matching



Esser et al. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. arXiv 2024.

… + 1-Step Distillation



Geng et al. Mean Flows for One-step Generative Modeling. NeurIPS 2025.

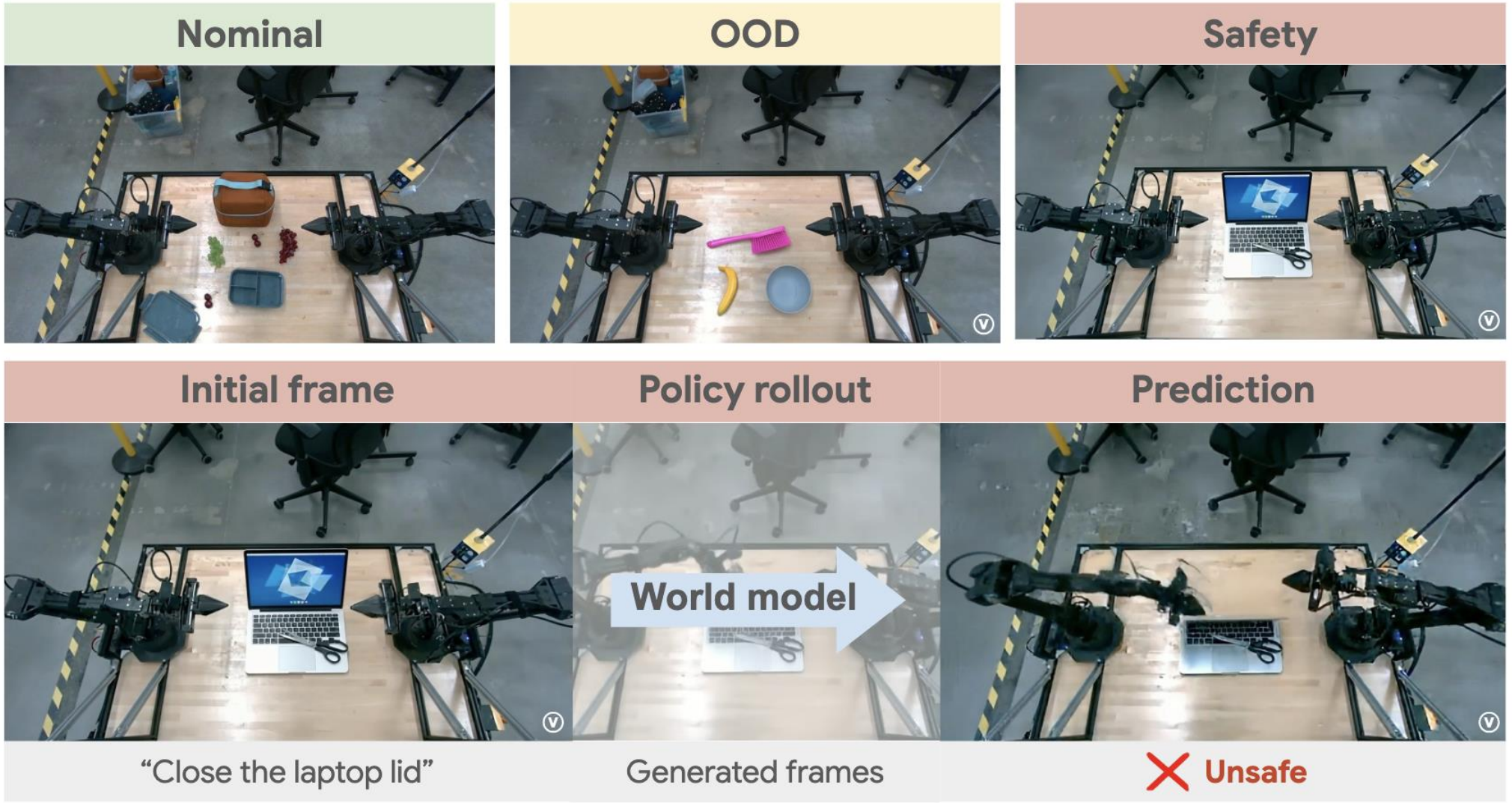
Pretty Image / Video / 3D Generation



More Than... Pretty Image / Video / 3D Generation

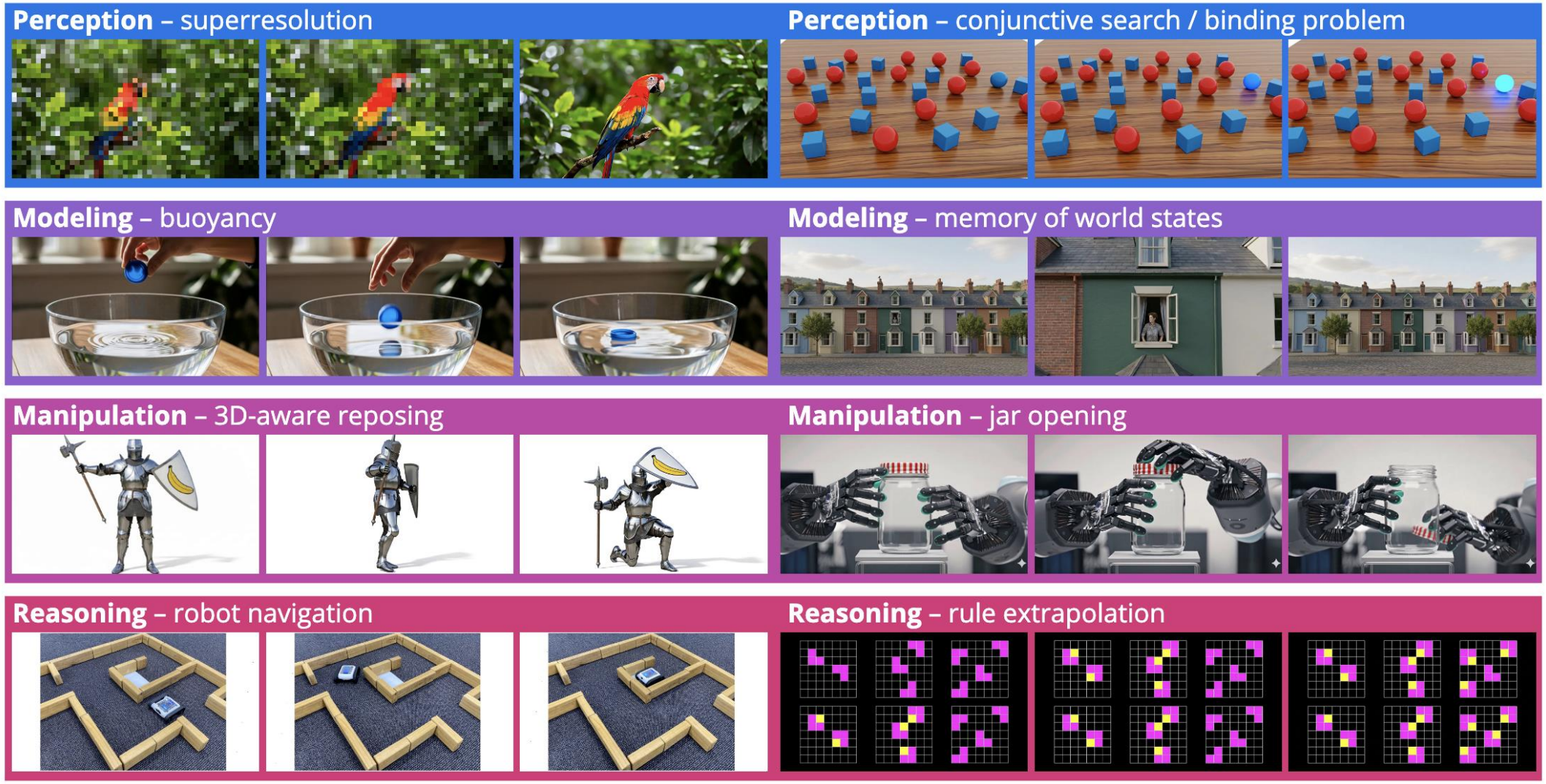


1. Physics Emerges from Visual Generation



Devin et al. Evaluating Gemini Robotics Policies in a Veo World Simulator. arXiv 2025.

2. ICL Emerges from Visual Generation



Wiedemer et al. Video models are zero-shot learners and reasoners. arXiv 2025.

3. Reasoning Emerges from Visual Generation

Physics 222, Modern Physics, Exam 2

NAME_____

You are graded on your work, with partial credit where it is deserved.
Please be clear and well-organized in all your steps.

1. (15) Show that

$$\Psi(x,t) = e^{-ipx/\hbar} e^{-iEt/\hbar}$$

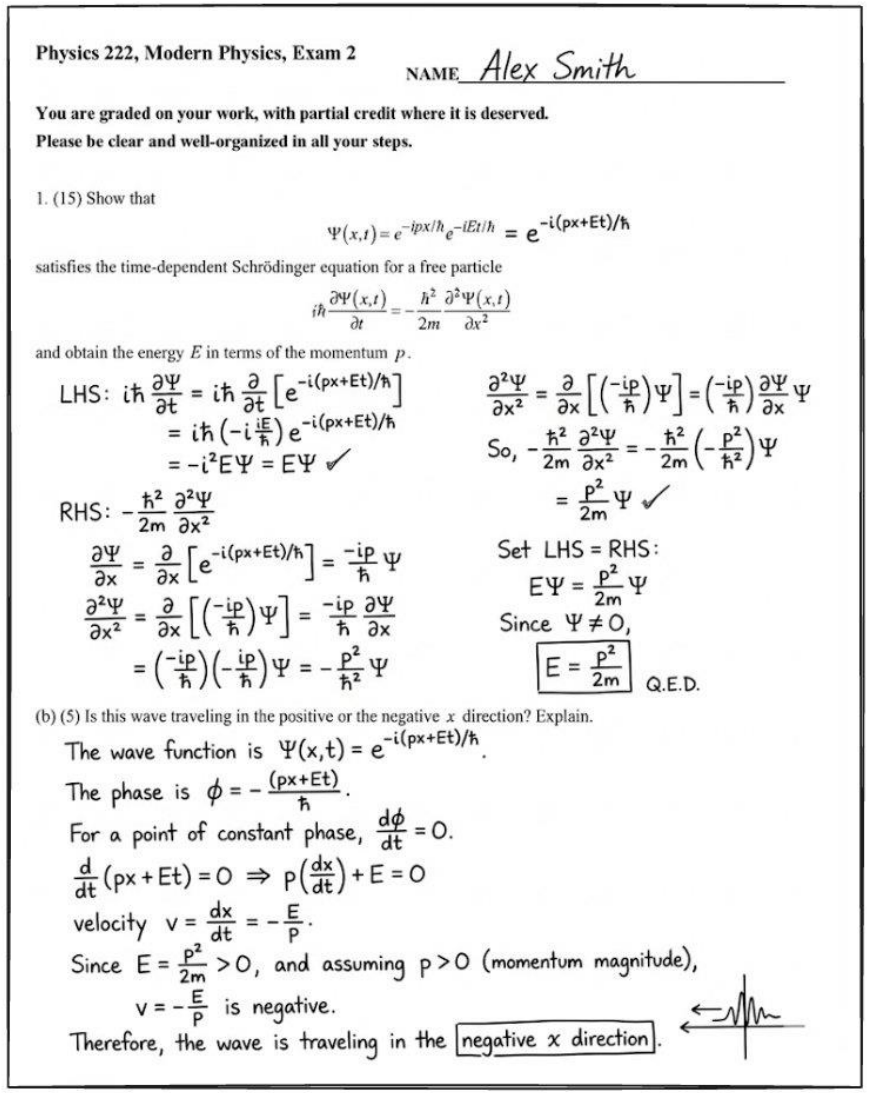
satisfies the time-dependent Schrödinger equation for a free particle

$$i\hbar \frac{\partial \Psi(x,t)}{\partial t} = -\frac{\hbar^2}{2m} \frac{\partial^2 \Psi(x,t)}{\partial x^2}$$

and obtain the energy E in terms of the momentum p.

(b) (5) Is this wave traveling in the positive or the negative x direction? Explain.

Gemini Nano Banana Pro

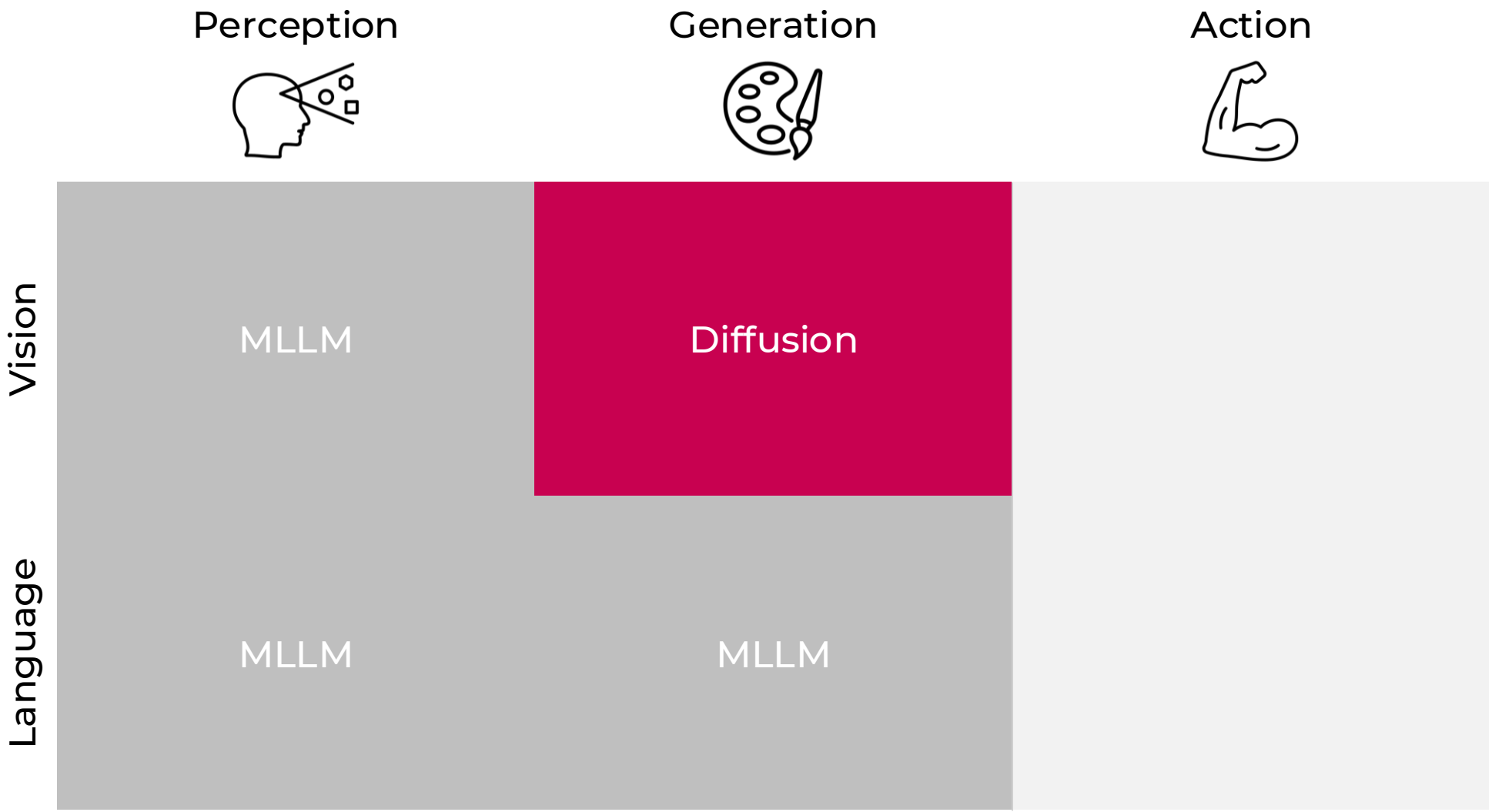


https://x.com/karpathy/status/1992655330002817095/photo/1

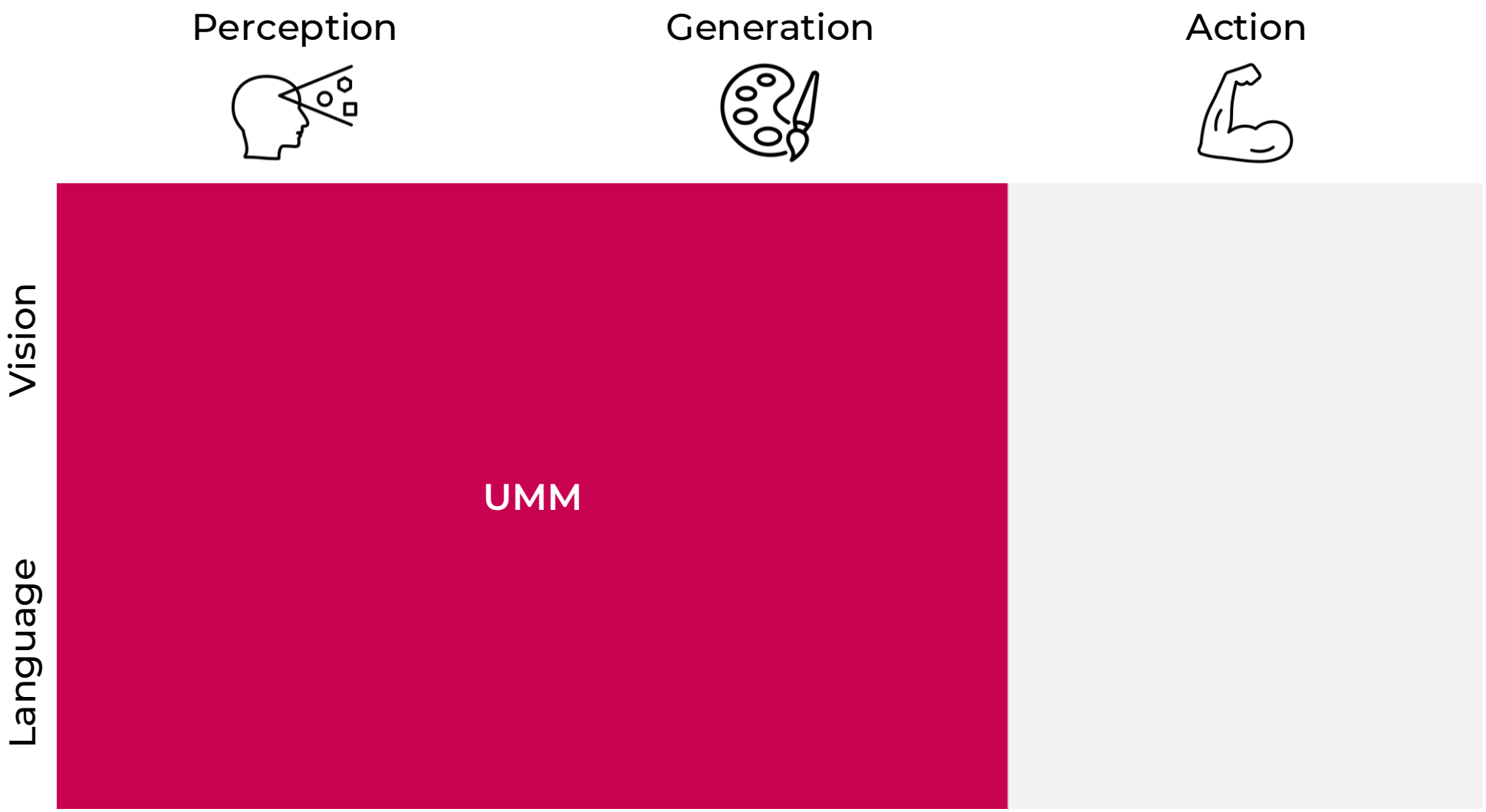
Visual Generation

	Perception	Generation	Action
			
Vision	MLLM	Diffusion	
Language	MLLM	MLLM	

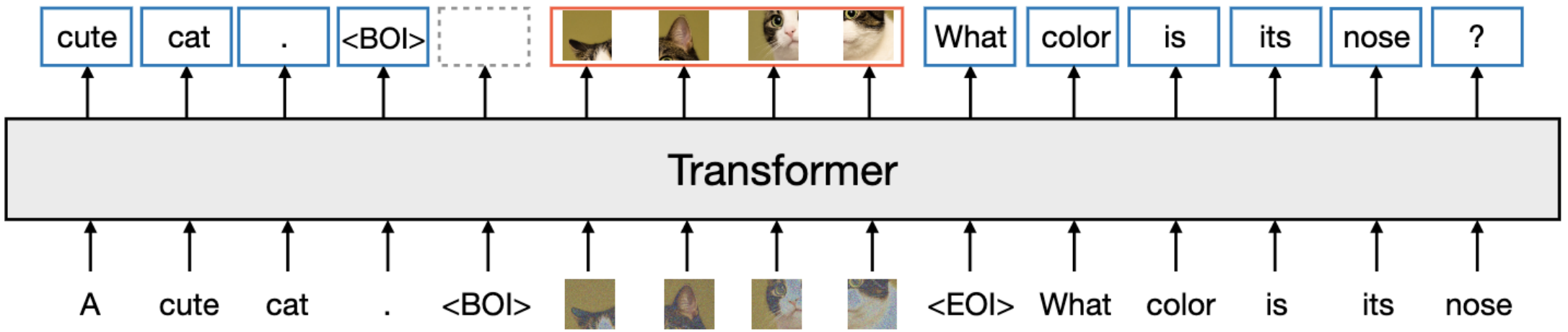
MLLM + Visual Generation?



Unified Multimodal Models

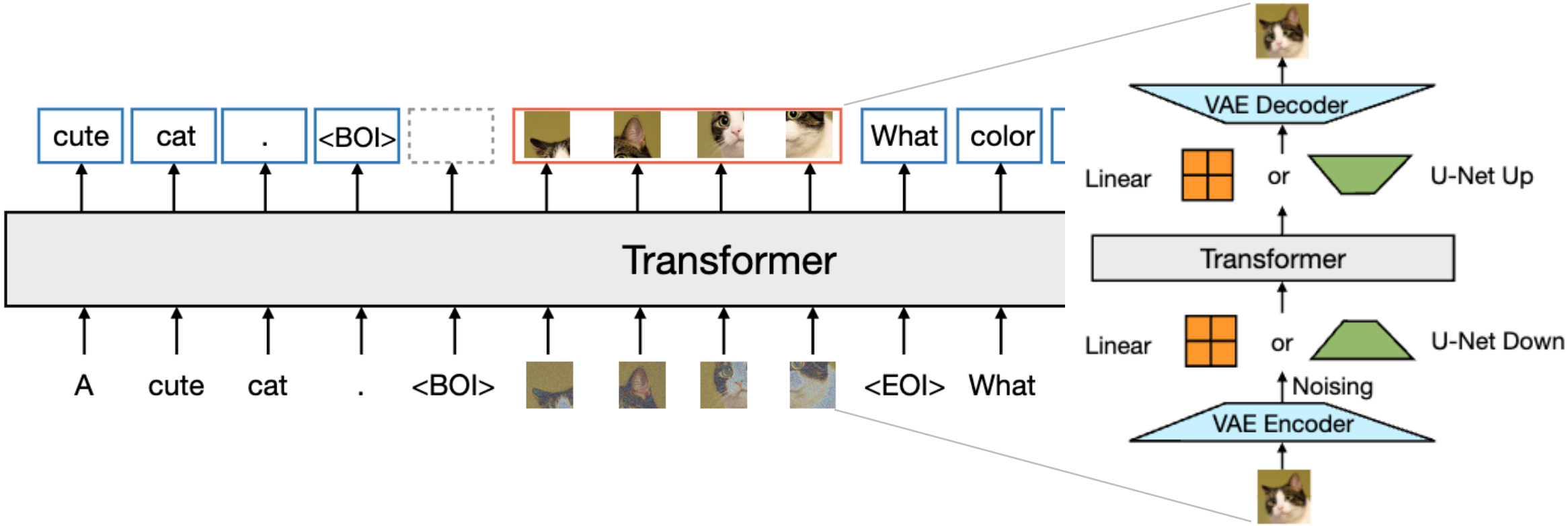


Unified Multimodal Models



Zhou et al. Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model. ICLR 2025.

Unified Multimodal Models

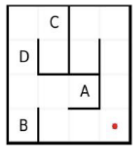


Zhou et al. Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model. ICLR 2025.

Unified Model: Visual Chain-of-Thought (CoT)

Input

Maze



Action Sequence:

Go left. Go left. Go up. Go left. Go up.

MiniBehavior



Action Sequence:

Go right. Go right. Pick up. Go left. Go left. Go up. Drop.

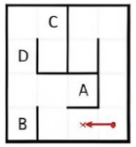
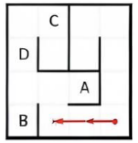
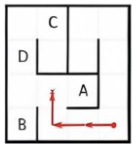
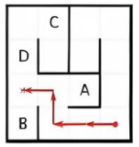
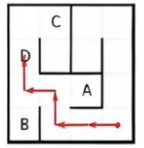
FrozenLake

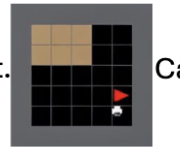
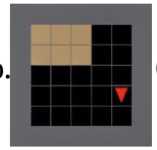
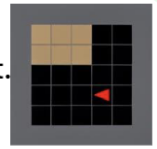


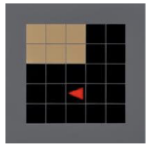
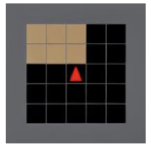
Action Sequence:

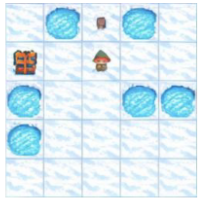
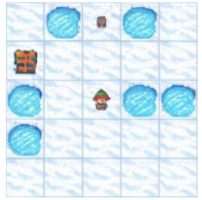
Go down. Go down. Go right. Go right. Go up. Go left. Go left.

Generated Response

Go left.  Go left.  Go up.  Go left.  Go up.  **The answer is D.**

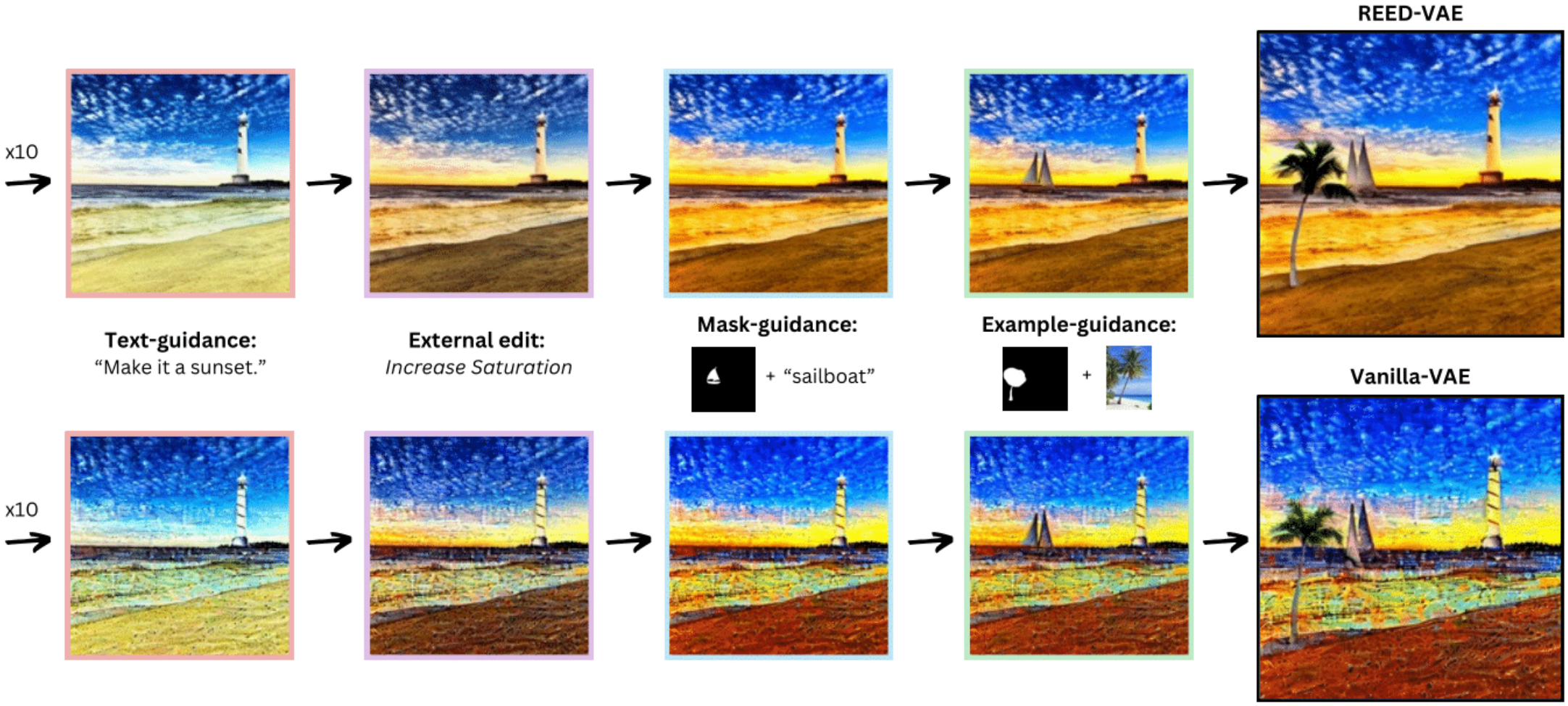
Go right.  Carrying: None. Go right.  Carrying: None. Pick up.  Carrying: printer. Go left. 

Carrying: printer. Go left.  Carrying: printer. Go up.  Carrying: printer. Drop.  Carrying: None. **Action Success.**

Go down.  Go down.  Go right.  **Action Failed: Fall into the Hole.**

Li et al. Imagine while Reasoning in Space: Multimodal Visualization-of-Thought. arXiv 2025.

Unified Model: Iterative Image Editing

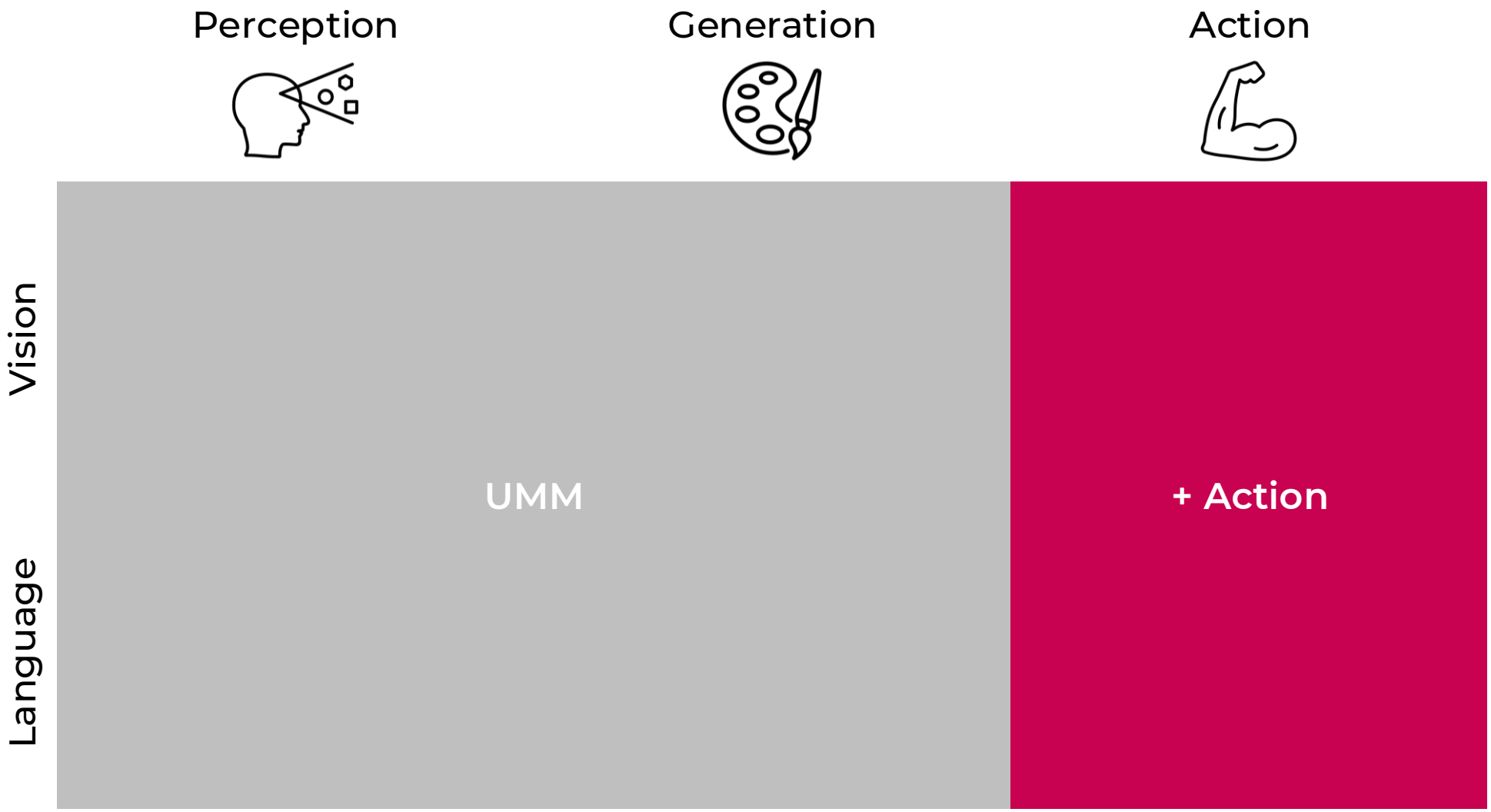


Amog et al. REED-VAE: RE-Encode Decode Training for Iterative Image Editing with Diffusion Models. arXiv 2025.

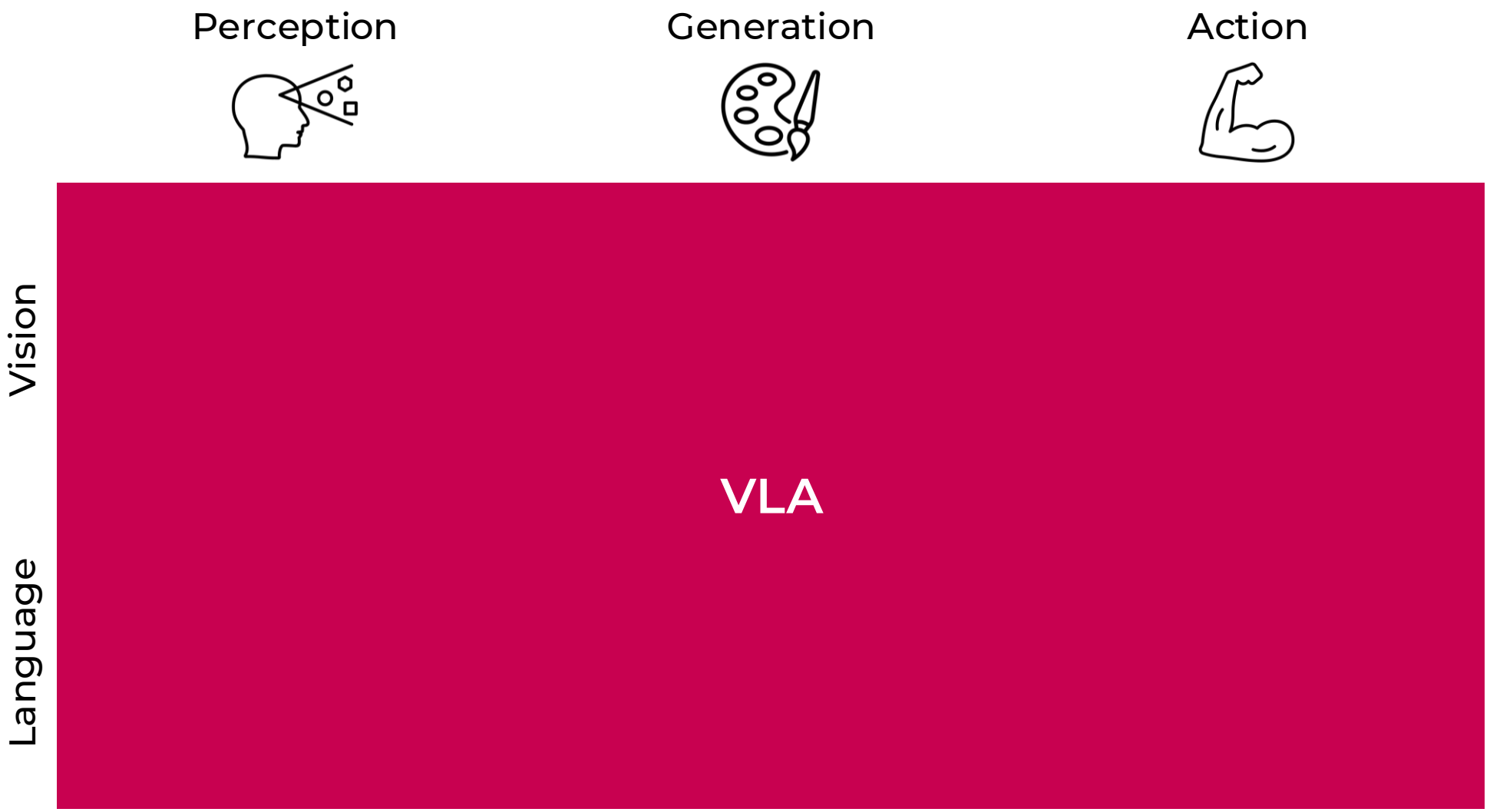
Summary of Part III: Unified Models

	Perception	Generation	Action
			
Vision	Visual Perception ⊂ UMM	Visual Generation ⊂ UMM	
Language	Lang. Understanding ⊂ UMM	Lang. Generation ⊂ UMM	

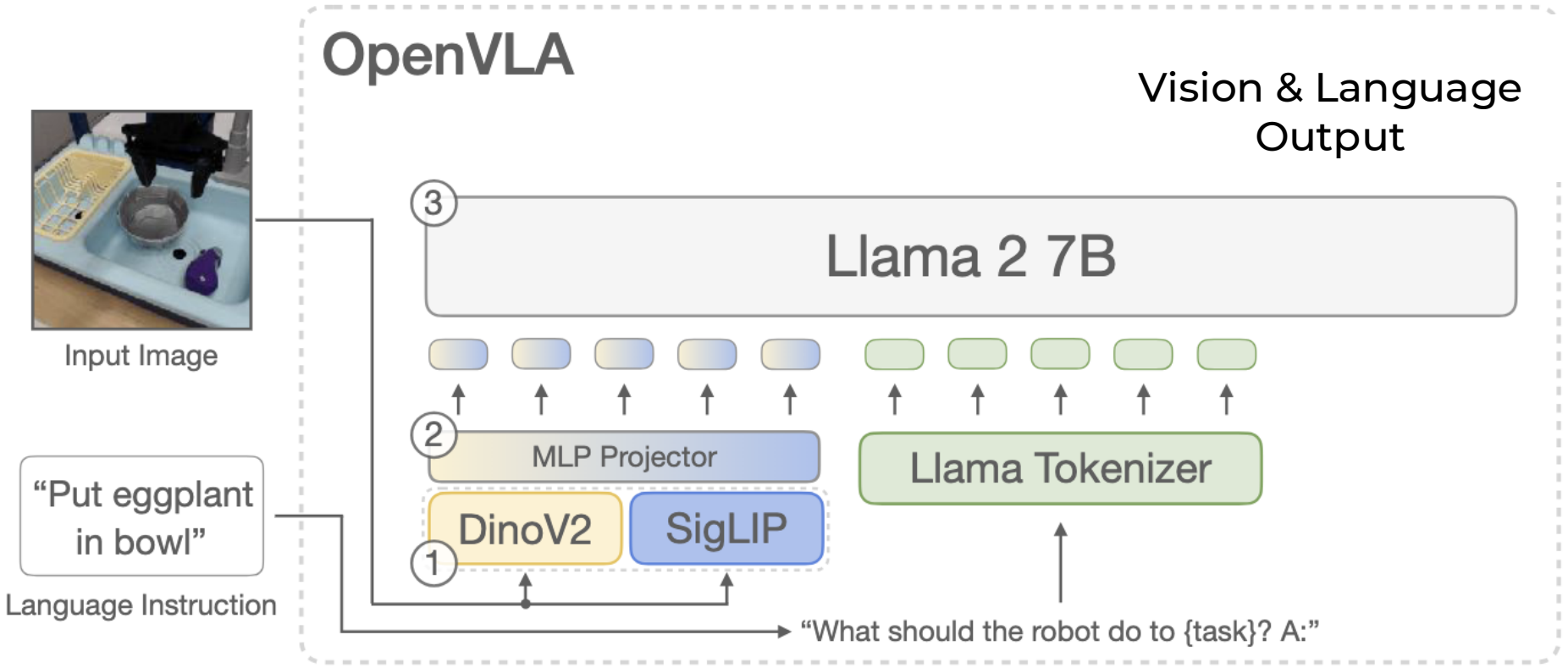
Multimodal Models + Action



Vision-Language-Action (VLA)

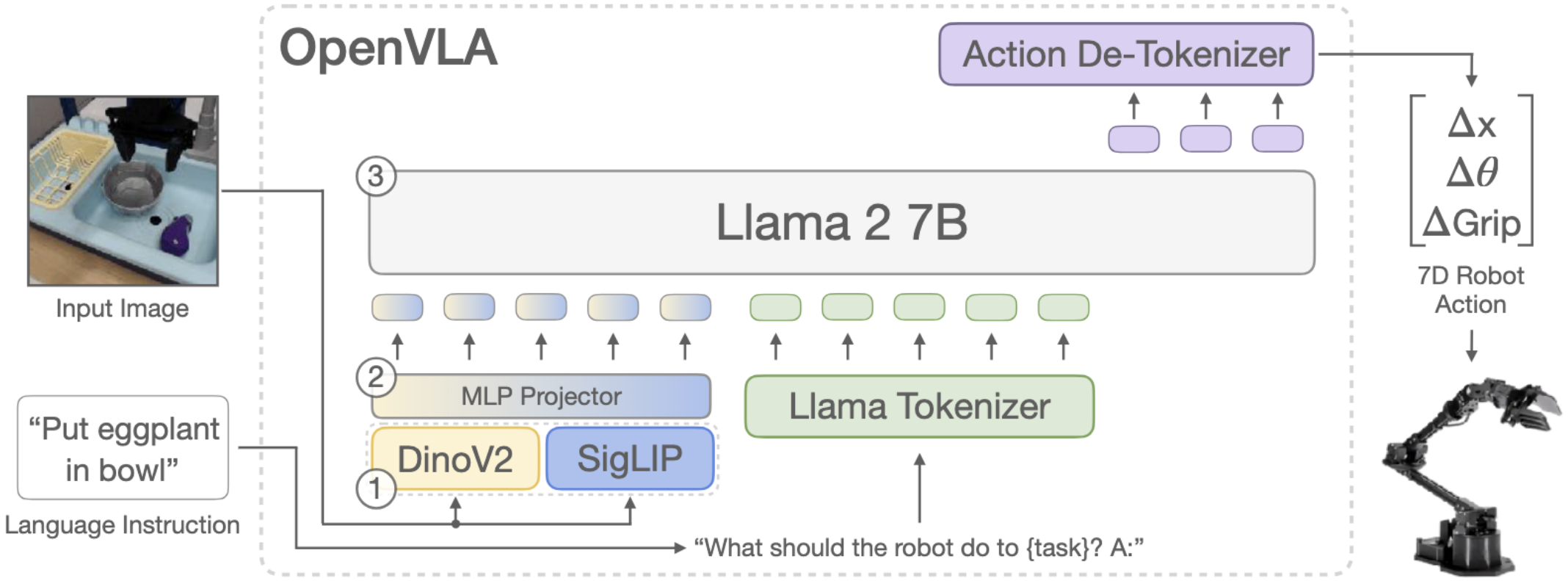


Vision-Language-Action (VLA)



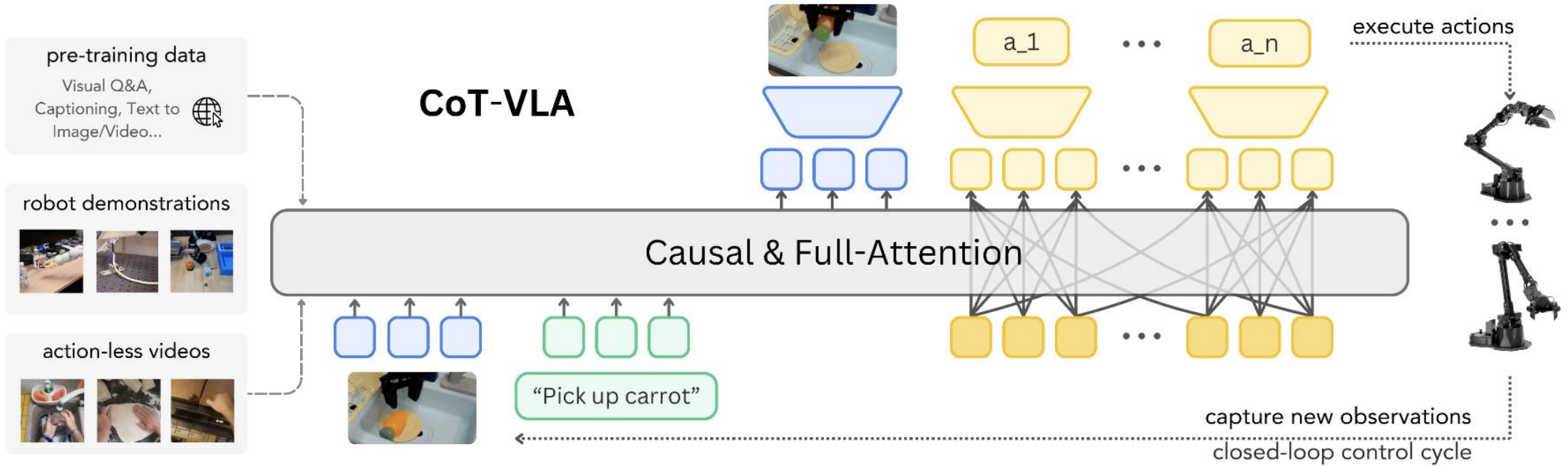
Kim et al. OpenVLA: An Open-Source Vision-Language-Action Model. arXiv 2024.

Vision-Language-Action (VLA)



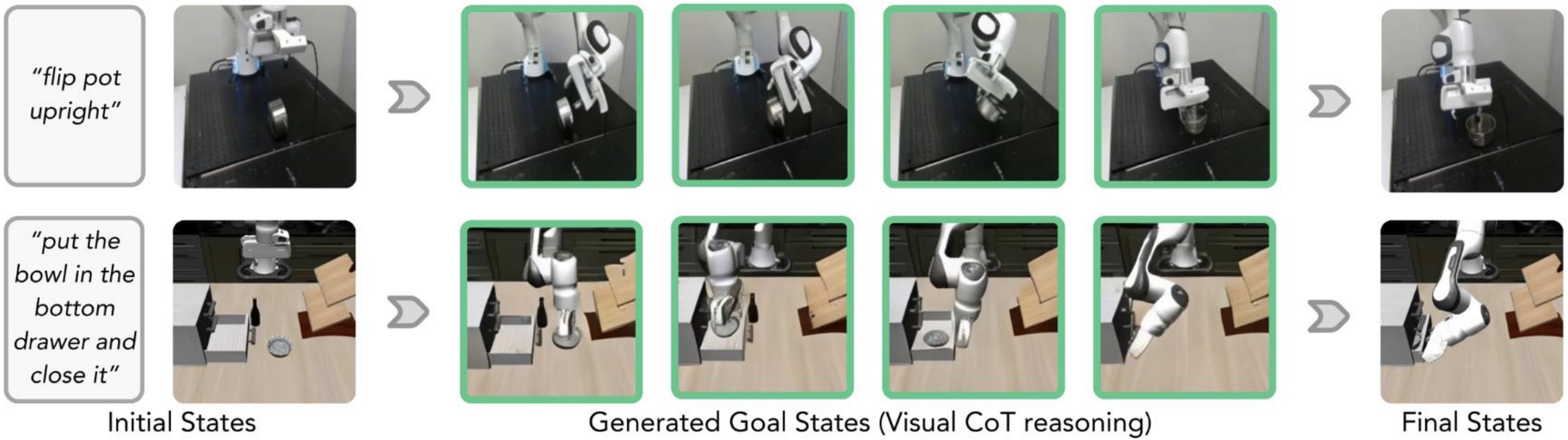
Kim et al. OpenVLA: An Open-Source Vision-Language-Action Model. arXiv 2024.

Unified Model + VLA → Visual CoT



Zhao et al. CoT-VLA: Visual Chain-of-Thought Reasoning for Vision-Language-Action Models. CVPR 2025.

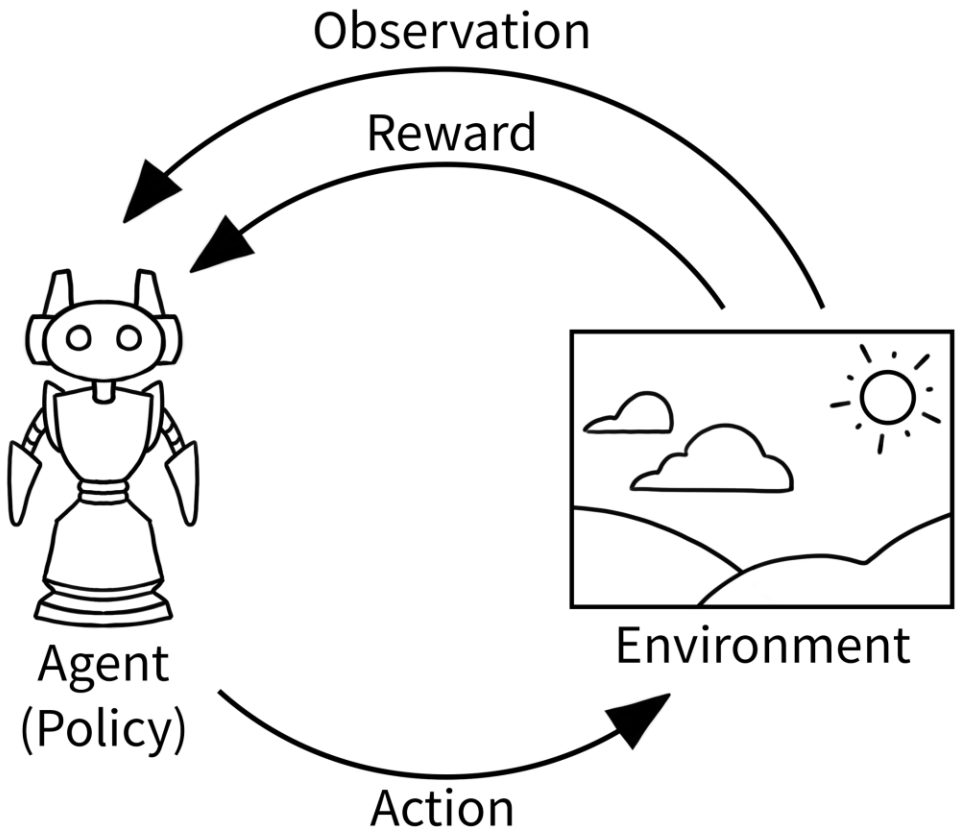
Unified Model + VLA → Visual CoT



Zhao et al. CoT-VLA: Visual Chain-of-Thought Reasoning for Vision-Language-Action Models. CVPR 2025.

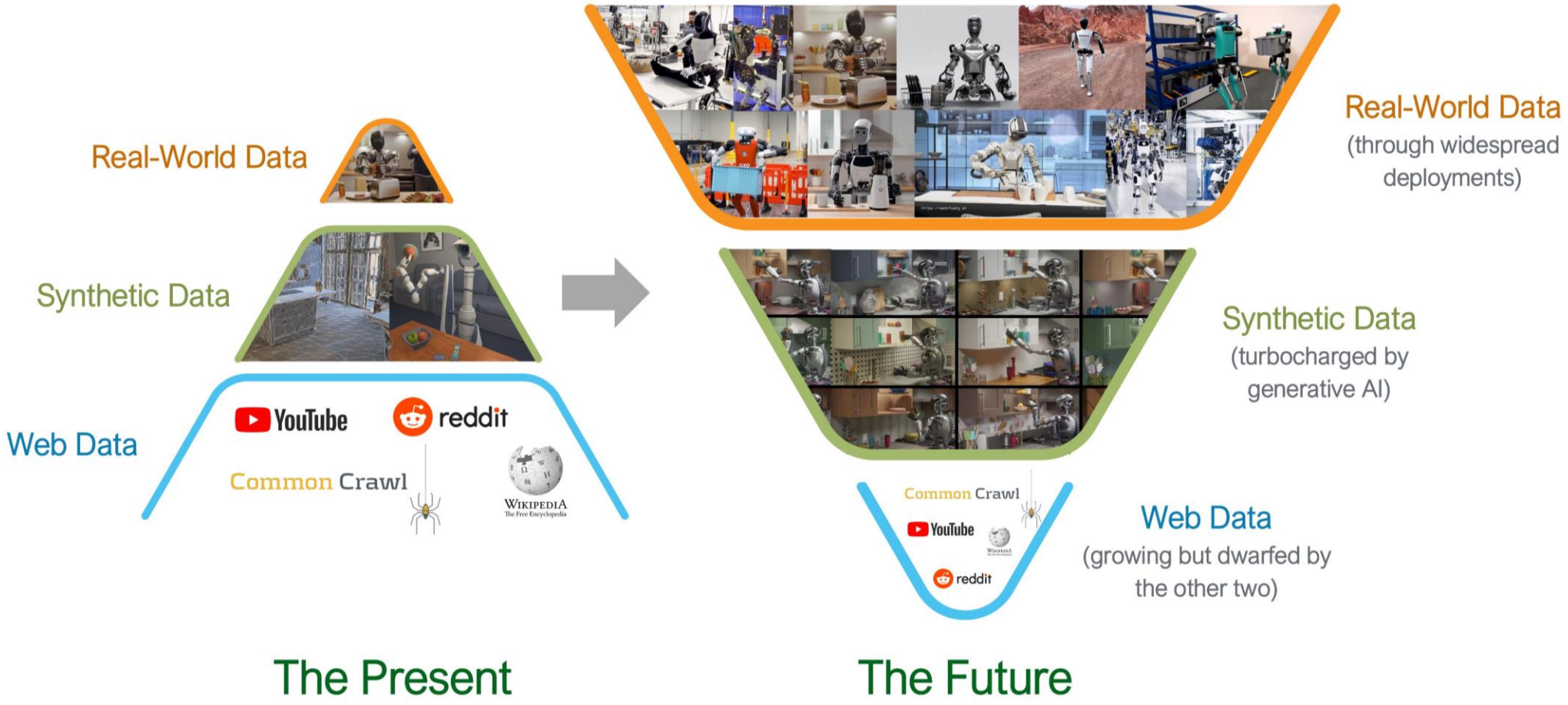
VLA Is More Than Just Adding an Action Decoder

Interactive Learning from Dynamic Environment



VLA Is More Than Just Adding an Action Decoder

Interactive Learning from Dynamic Environment



Yuke Zhu. Data Pyramid and Data Flywheel for Robotic Foundation Models. Talk 2024.

Summary of Part IV: Unified VLA

	Perception	Generation	Action
			
Vision	Visual Perception → VLA	Visual Generation → VLA	Action (from Interactive and Dynamic Env.) → VLA
Language	Lang. Understanding → VLA	Lang. Generation → VLA	