北京航空航天大學
BEIHANG UNIVERSITY

生成式AI与大模型第5讲

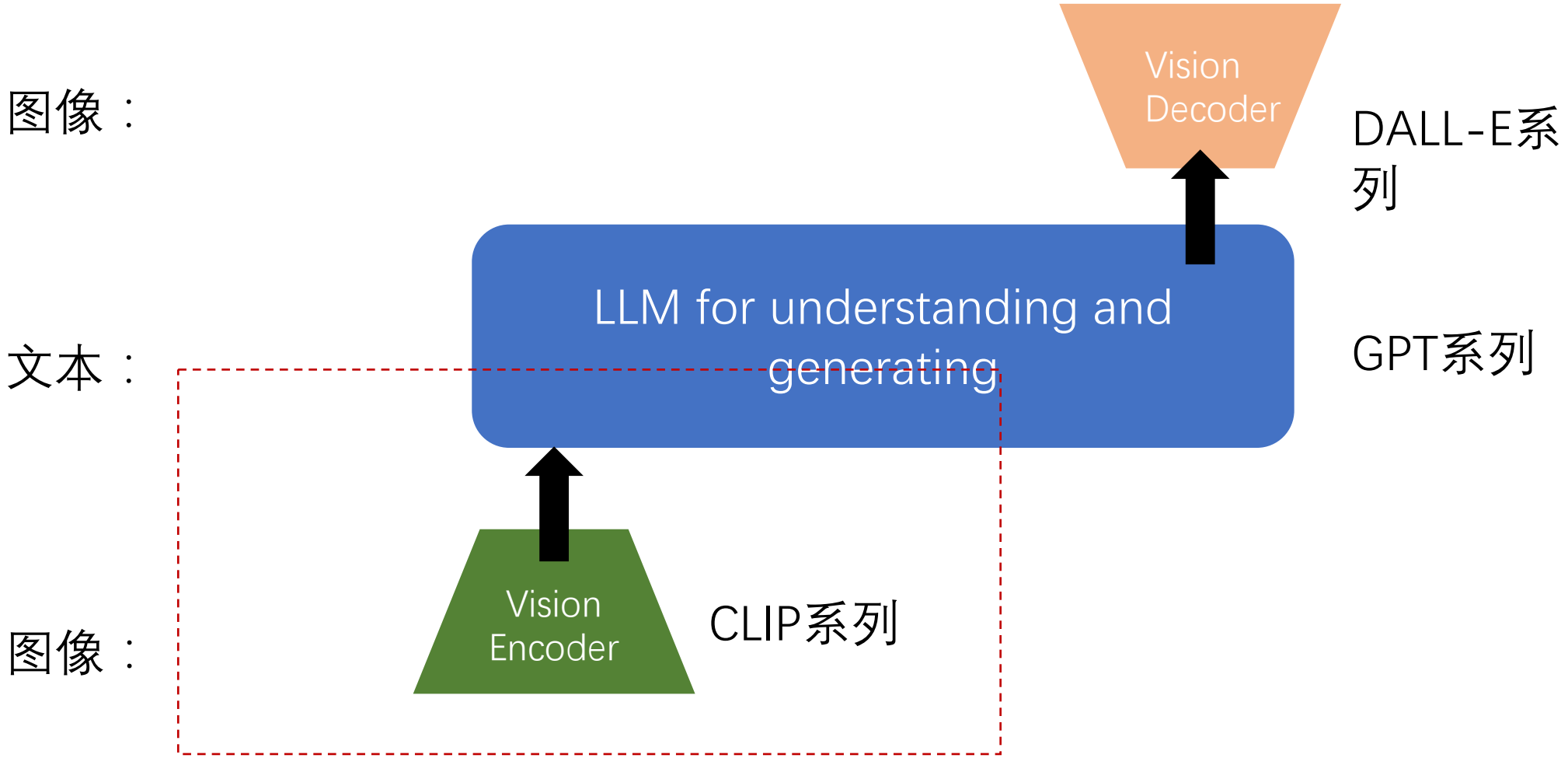
多模态对齐

Beihang University

人工智能研究院
黄雷

01 模态对齐

多模态和视觉大模型

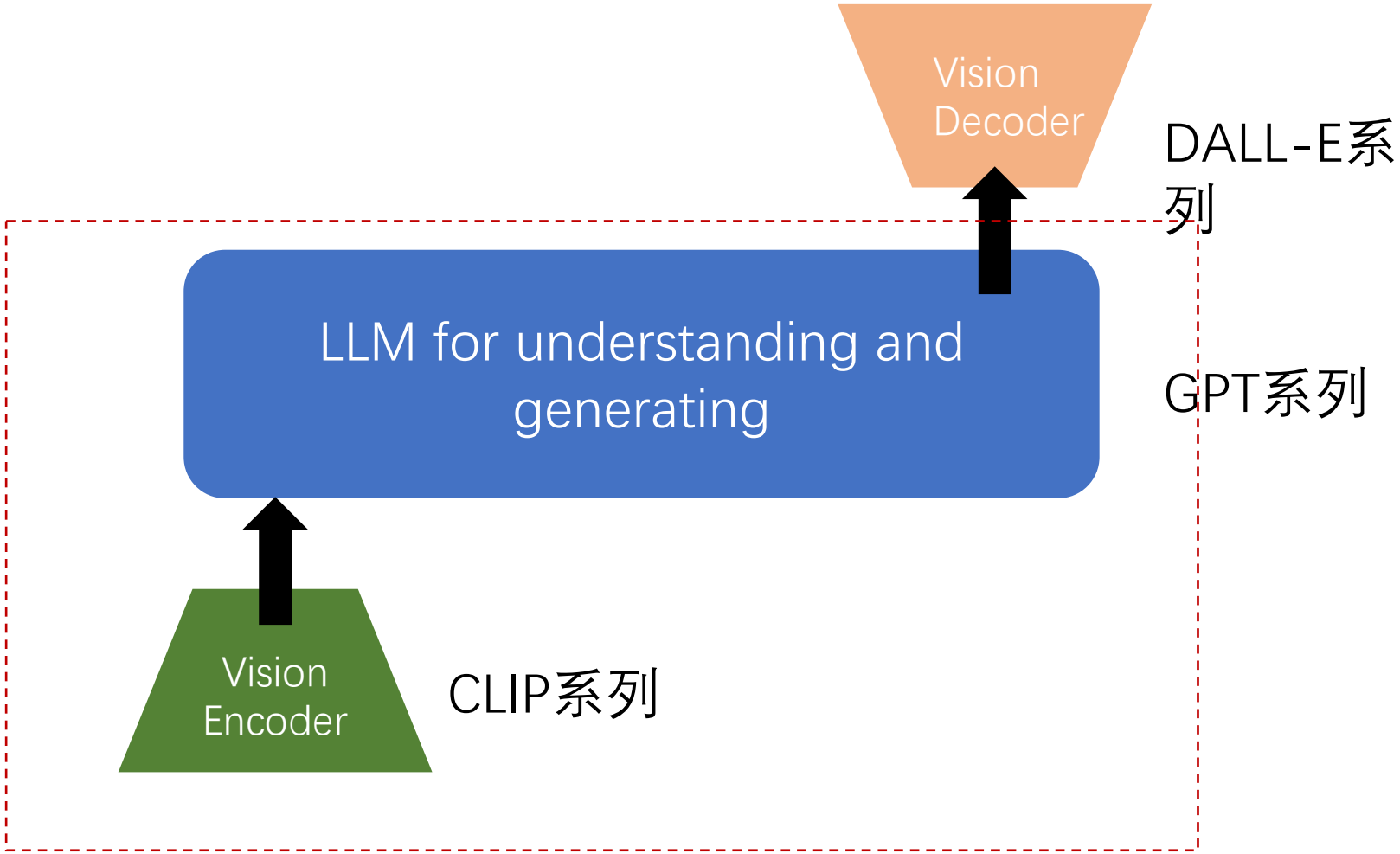

图像：
文本：
图像：
Vision Decoder
DALL-E系列
LLM for understanding and generating
GPT系列
Vision Encoder
CLIP系列

多模态和视觉大模型

图像：

文本：

图像：

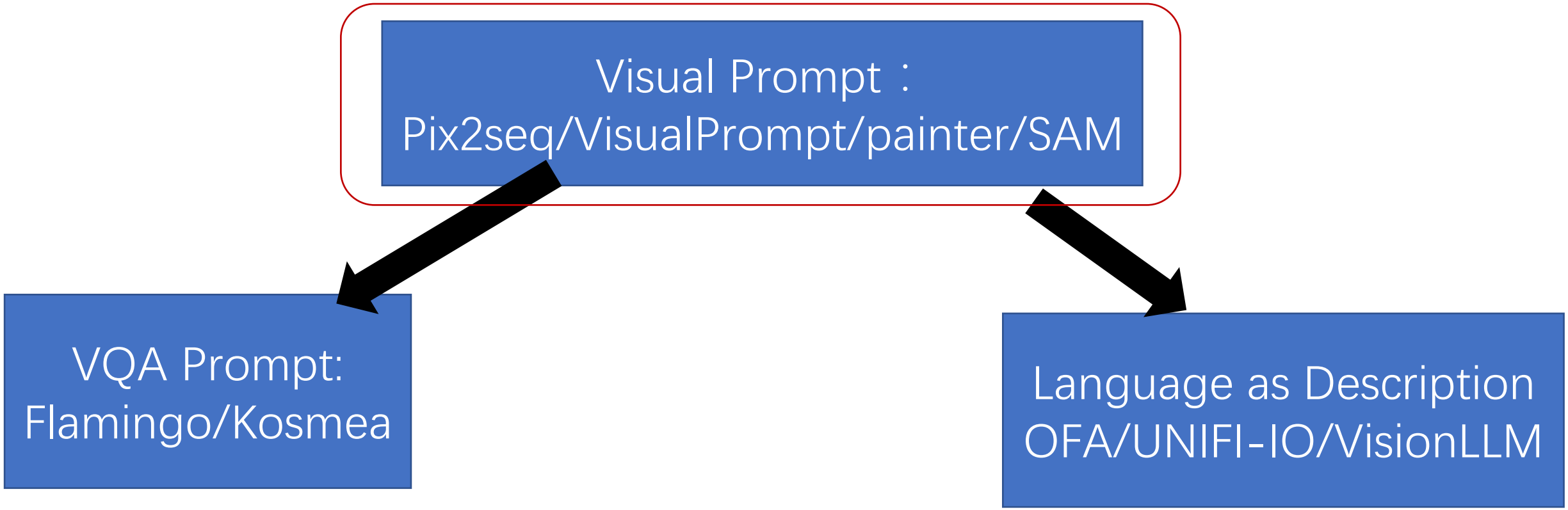


DALL-E系列

GPT系列

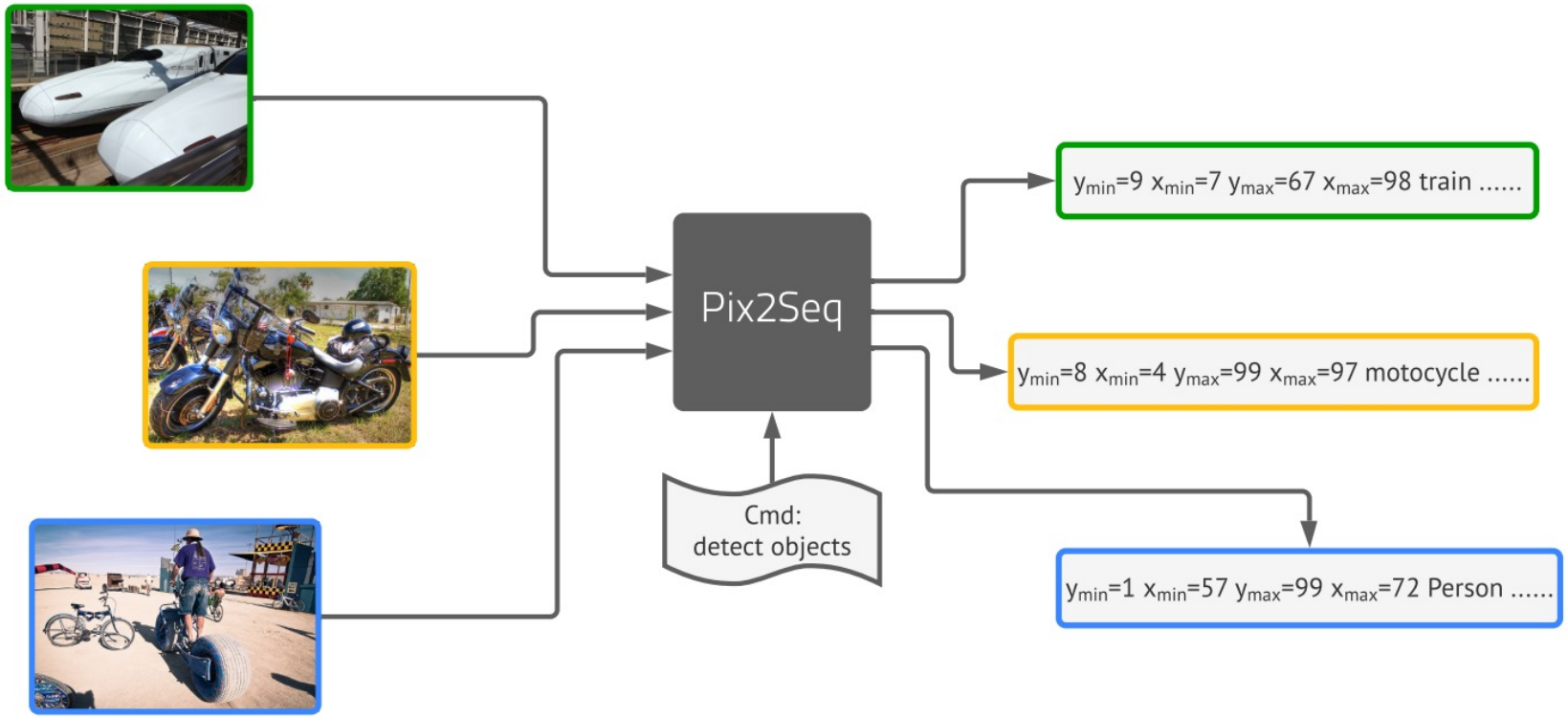
CLIP系列

多模态和视觉大模型（Prompt-like)



Visual Prompt

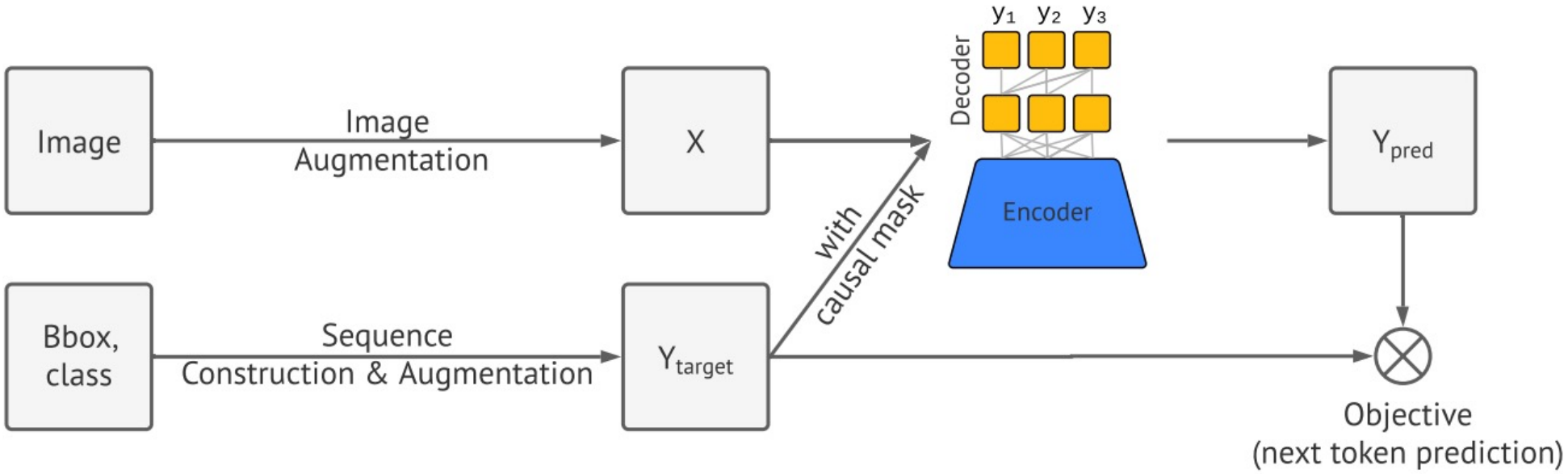
- Pix2Seq (Output as token sequences)



PIX2SEQ: A LANGUAGE MODELING FRAMEWORK FOR OBJECT DETECTION. ICLR 2022

Visual Prompt

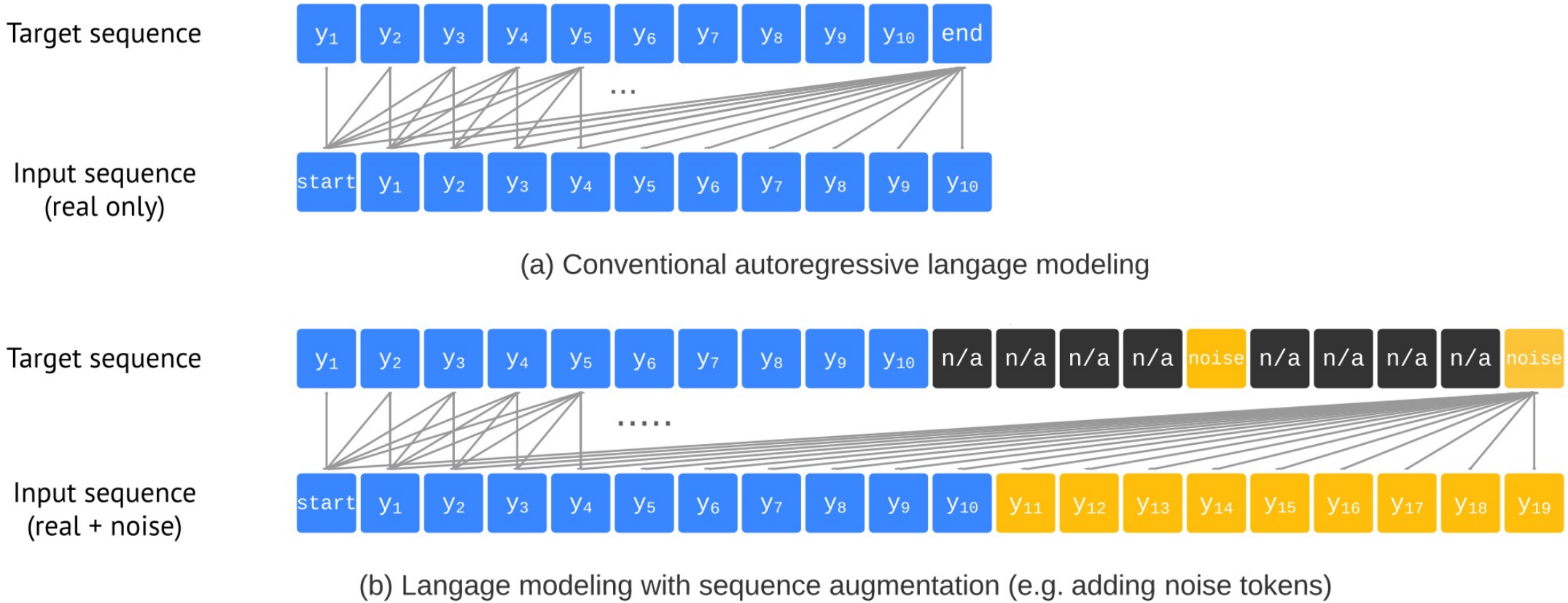
- Pix2Seq (Output as token sequences)



PIX2SEQ: A LANGUAGE MODELING FRAMEWORK FOR OBJECT DETECTION. ICLR 2022

Visual Prompt

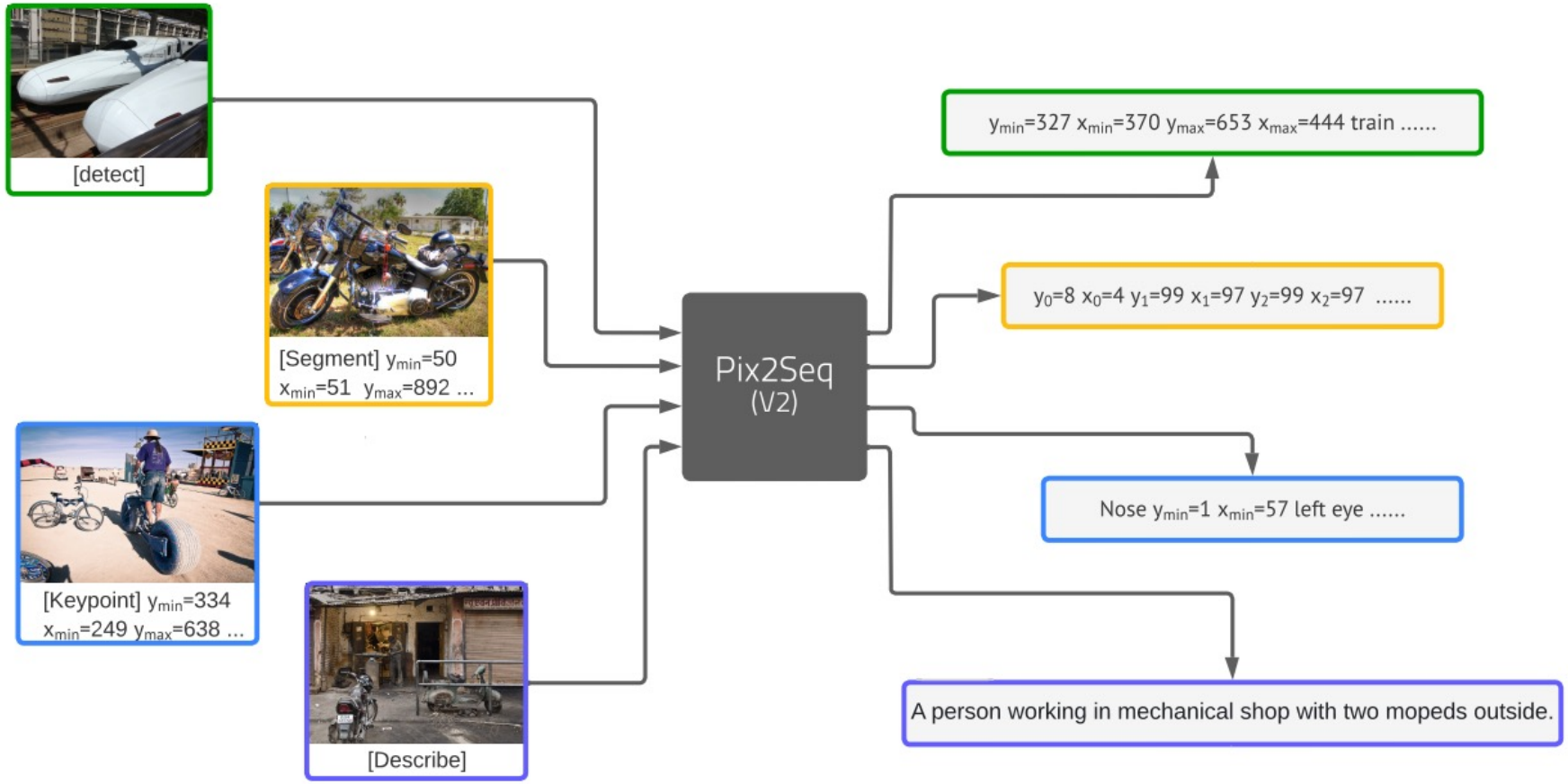
- Pix2Seq (Output as token sequences)



PIX2SEQ: A LANGUAGE MODELING FRAMEWORK FOR OBJECT DETECTION. ICLR 2022

Visual Prompt

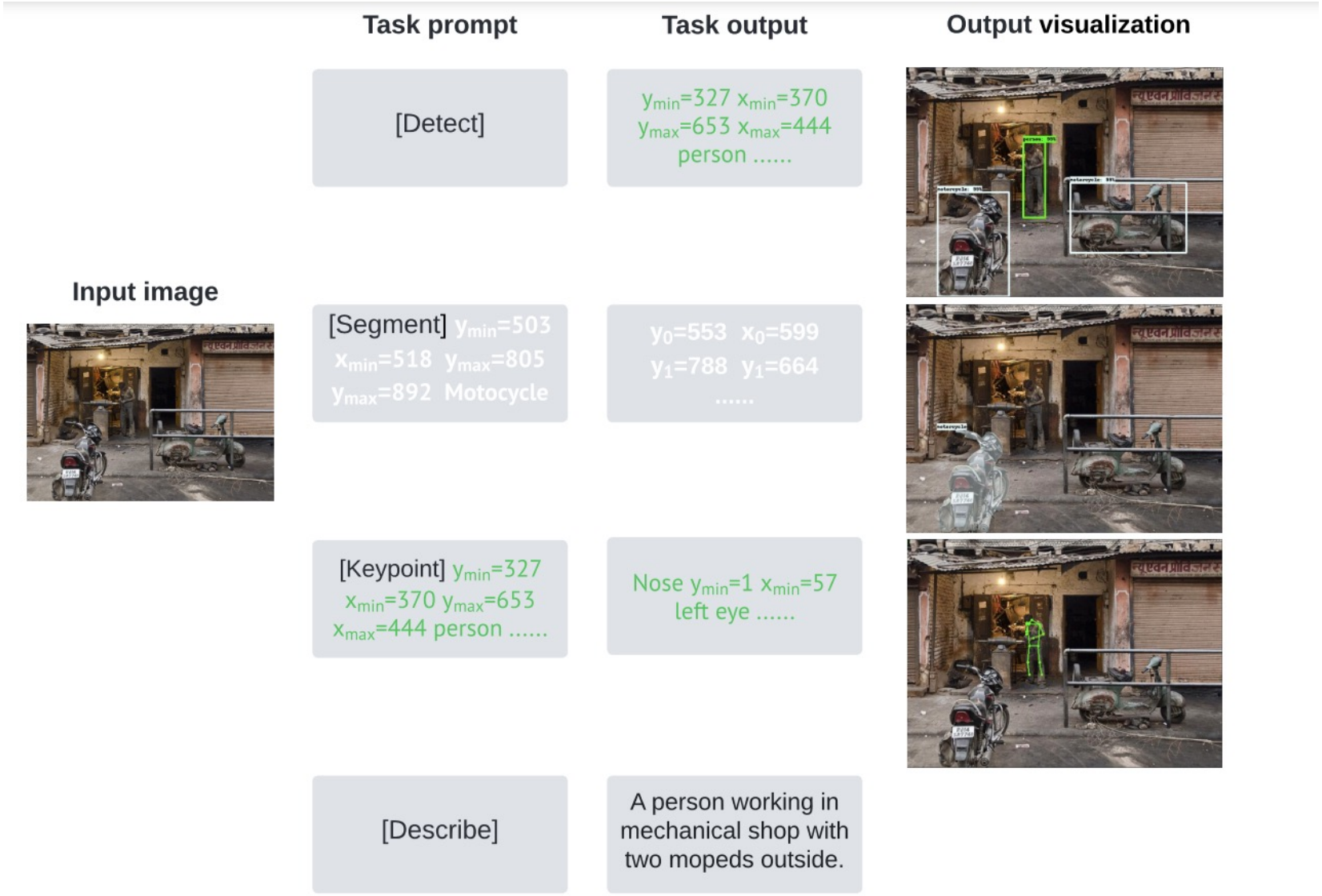
- Pix2Seq-v2 (Detection/Segmentation)



A Unified Sequence Interface for Vision Tasks. NeurIPS 2022

Visual Prompt

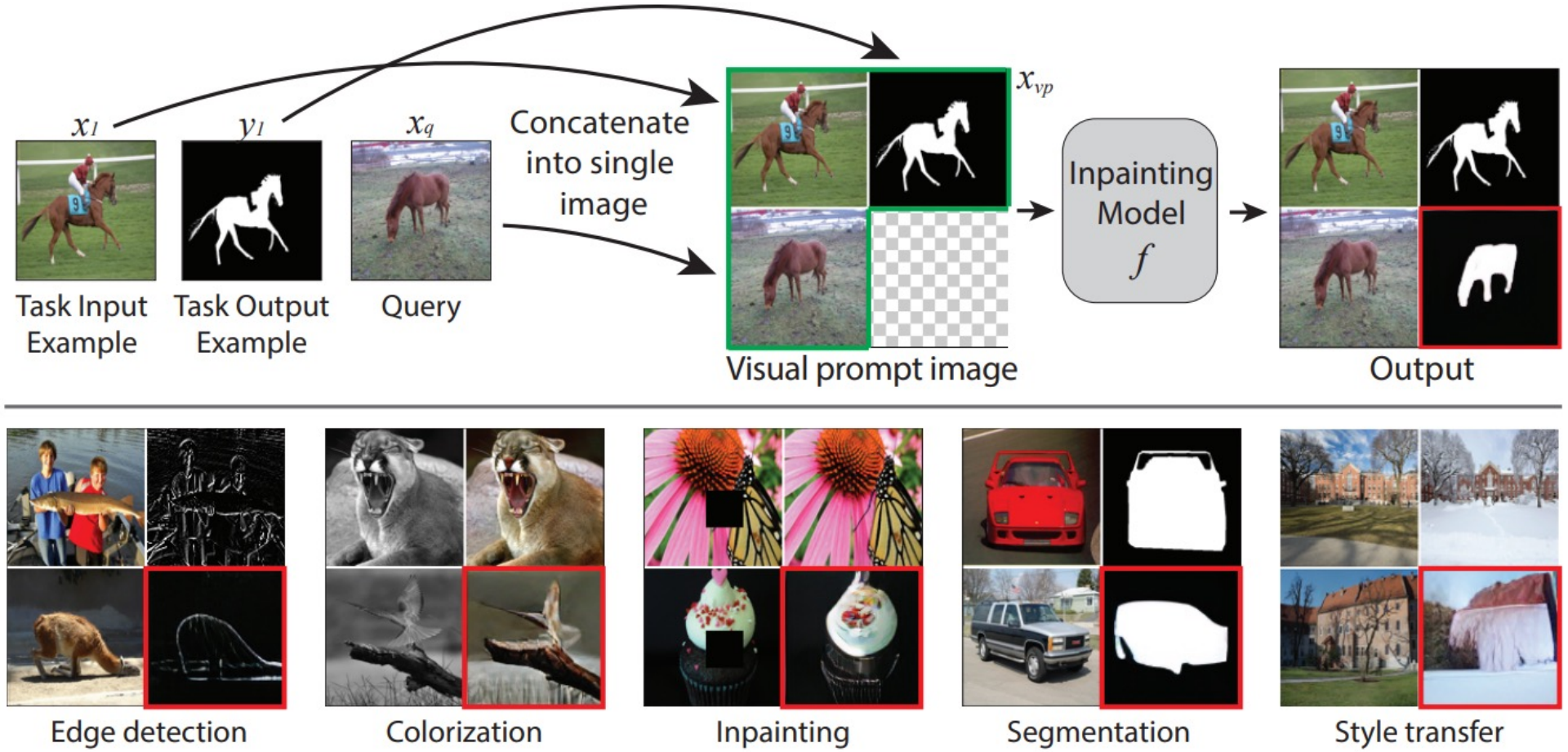
- Pix2Seq-v2



A Unified Sequence Interface for Vision Tasks. NeurIPS 2022

Visual Prompt

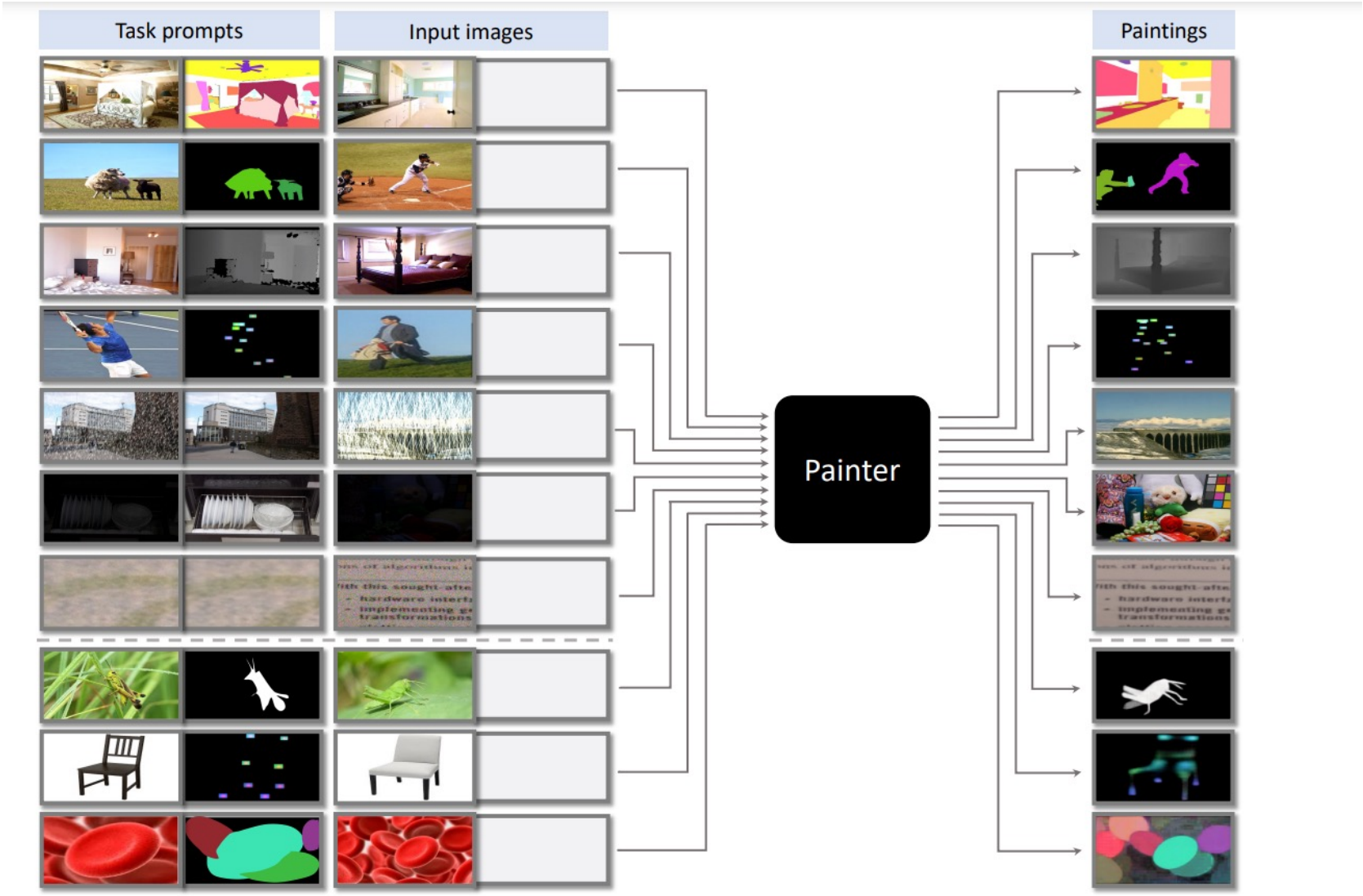
- Vision Prompter



Visual Prompting via Image Inpainting. NeurIPS 2022

Visual Prompt

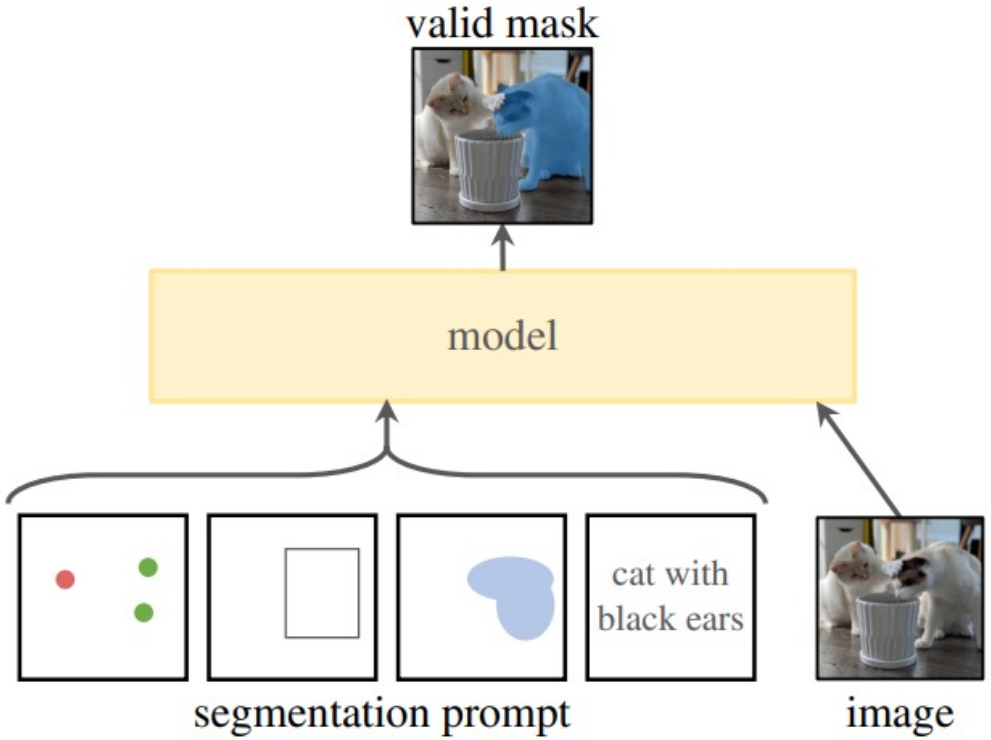
- Painter



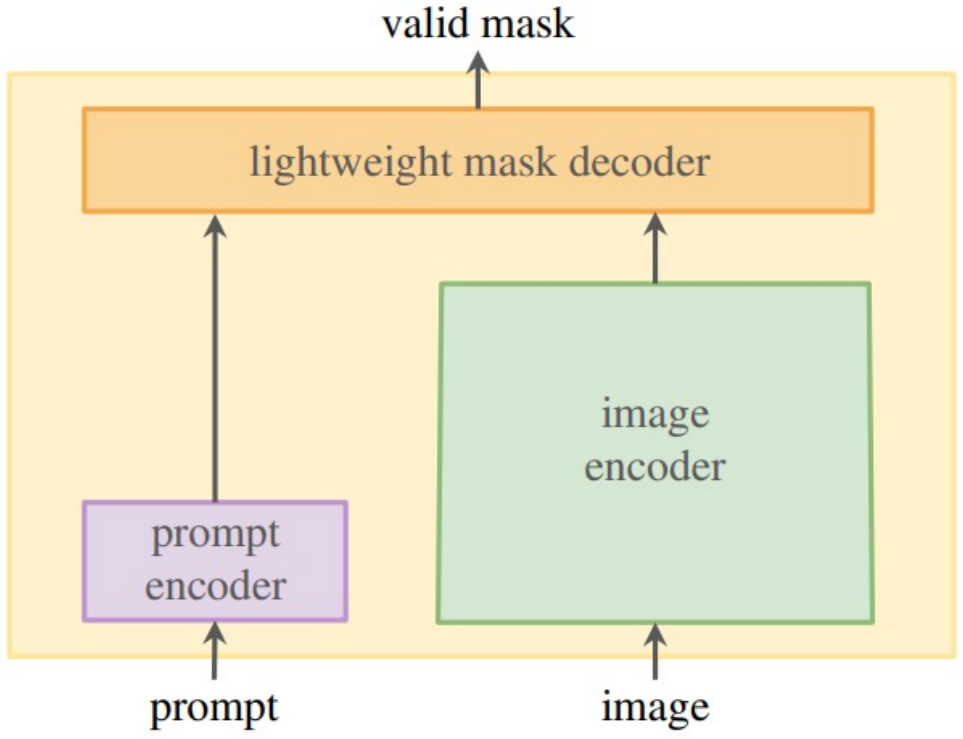
Images Speak in Images: A Generalist Painter for In-Context Visual Learning. CVPR 2023

Visual Prompt

- SAM (Segment Anything Model)



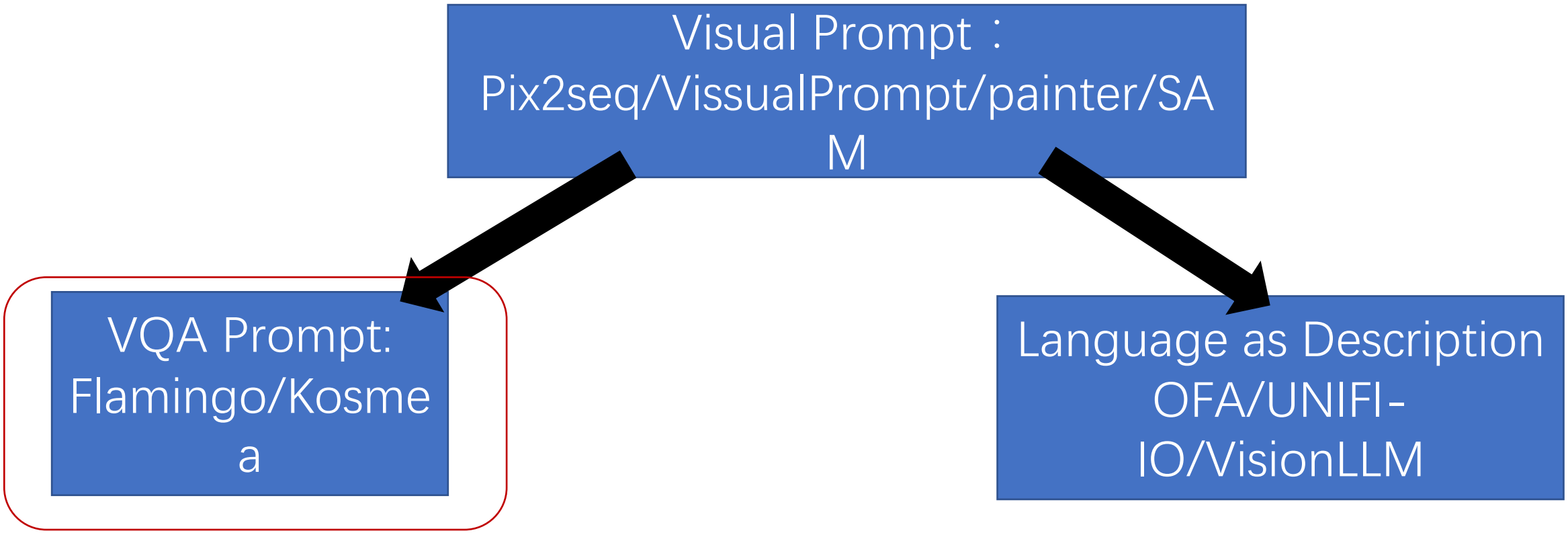
(a) **Task**: promptable segmentation



(b) **Model**: Segment Anything Model (**SAM**)

Segment Anything. Meta. 2023

多模态大模型（Prompt-like)



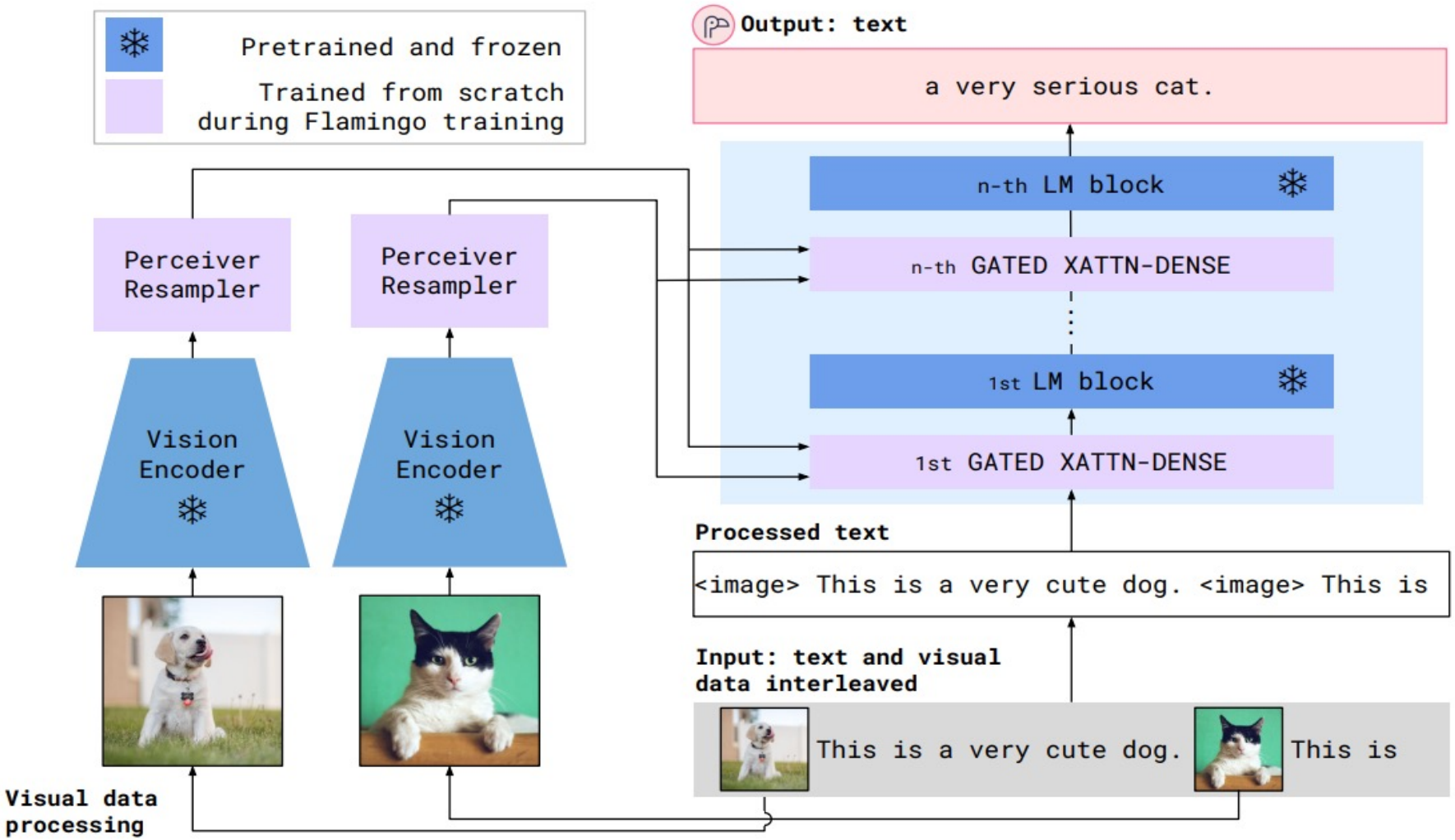
VQA Prompt

- Flamingo

Flamingo: a Visual Language Model for Few-Shot Learning, Deep Mind, NeurIPS 2022

VQA Prompt

- Flamingo

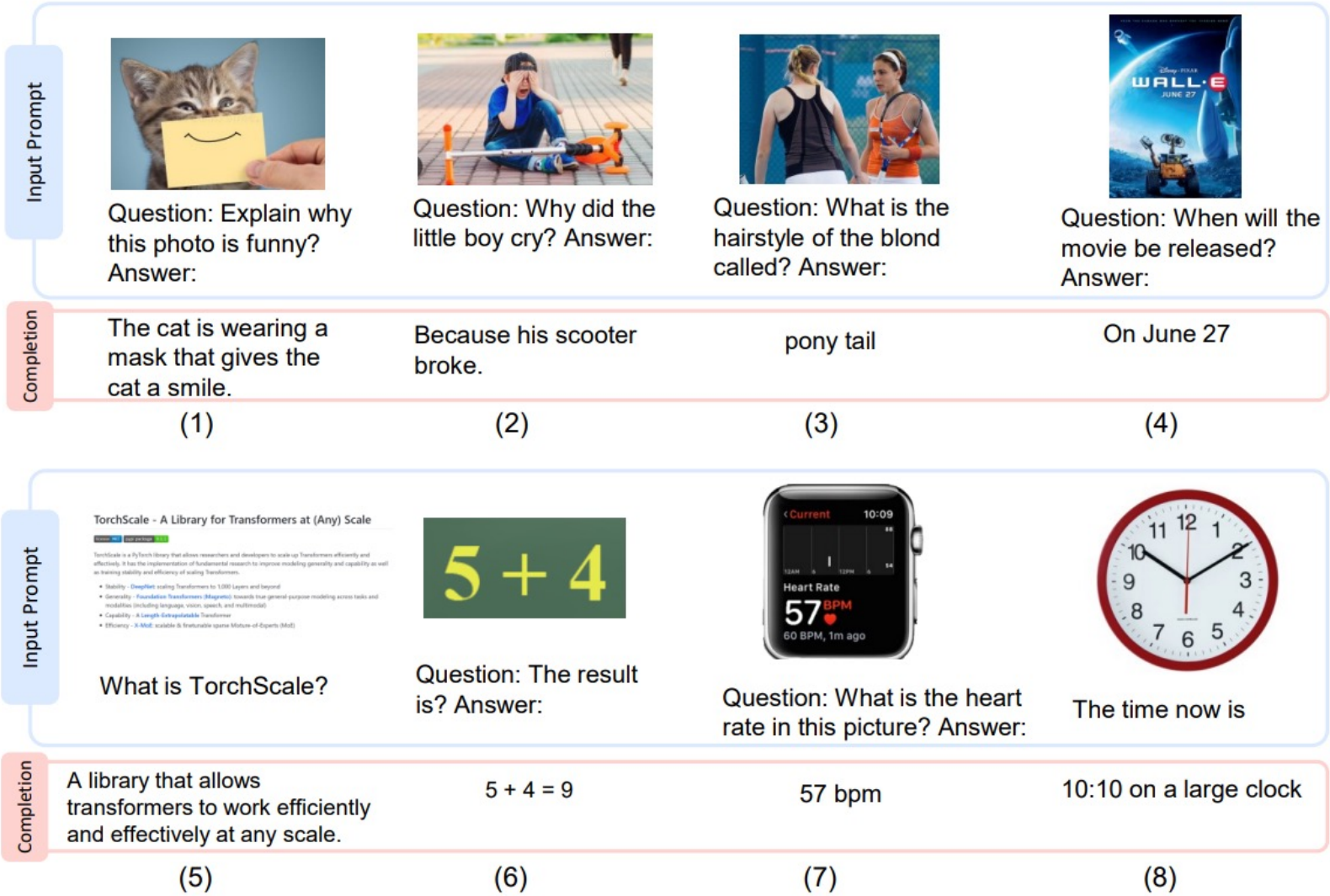


Flamingo: a Visual Language Model for Few-Shot Learning, Deep Mind, NeurIPS 2022

VQA Prompt

- KOSMOS-1

Zero-shot



Language Is Not All You Need: Aligning Perception with Language Models. Microsoft. 2023

VQA Prompt

- LLaVA系列
  - LLaVA, TinyLLaVA等

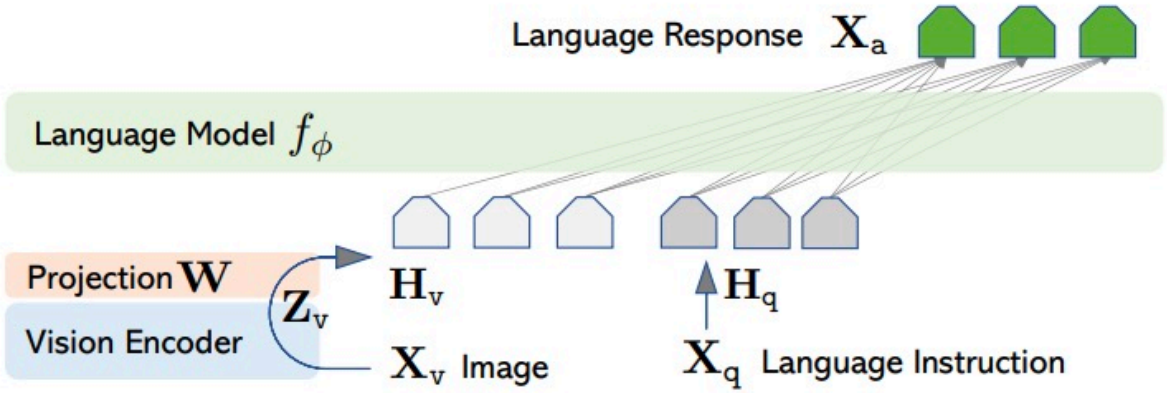
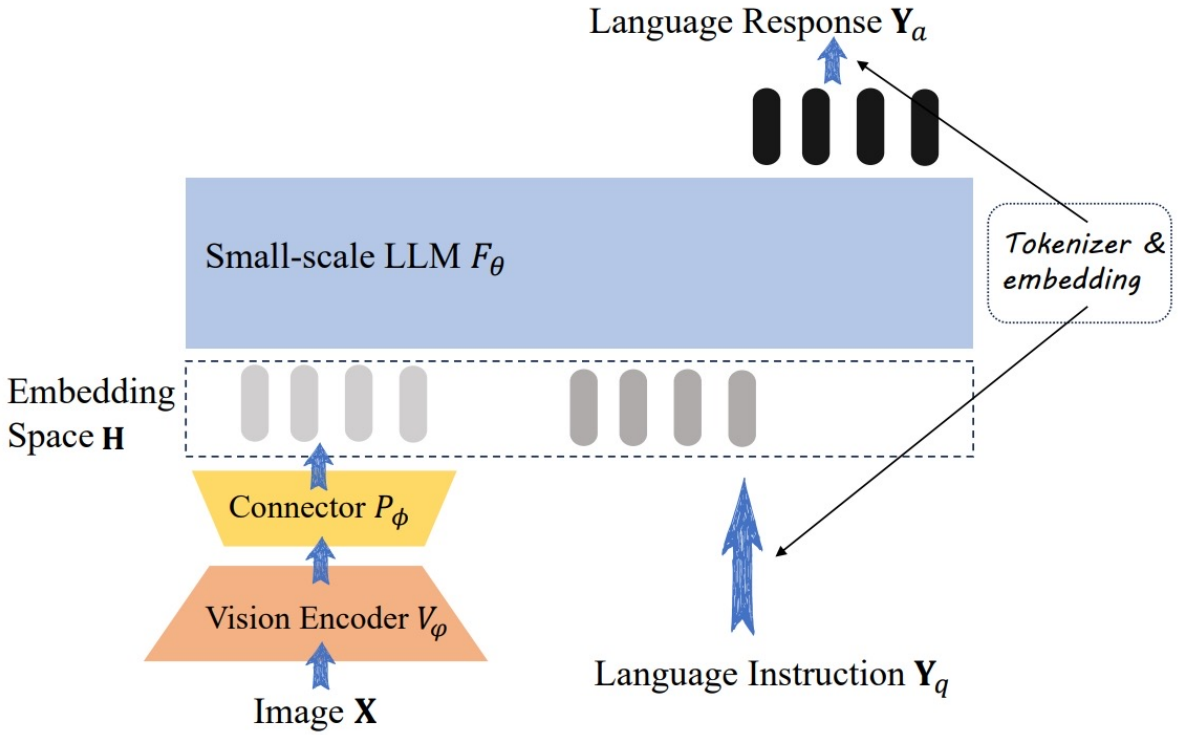


Figure 1: LLaVA network architecture.

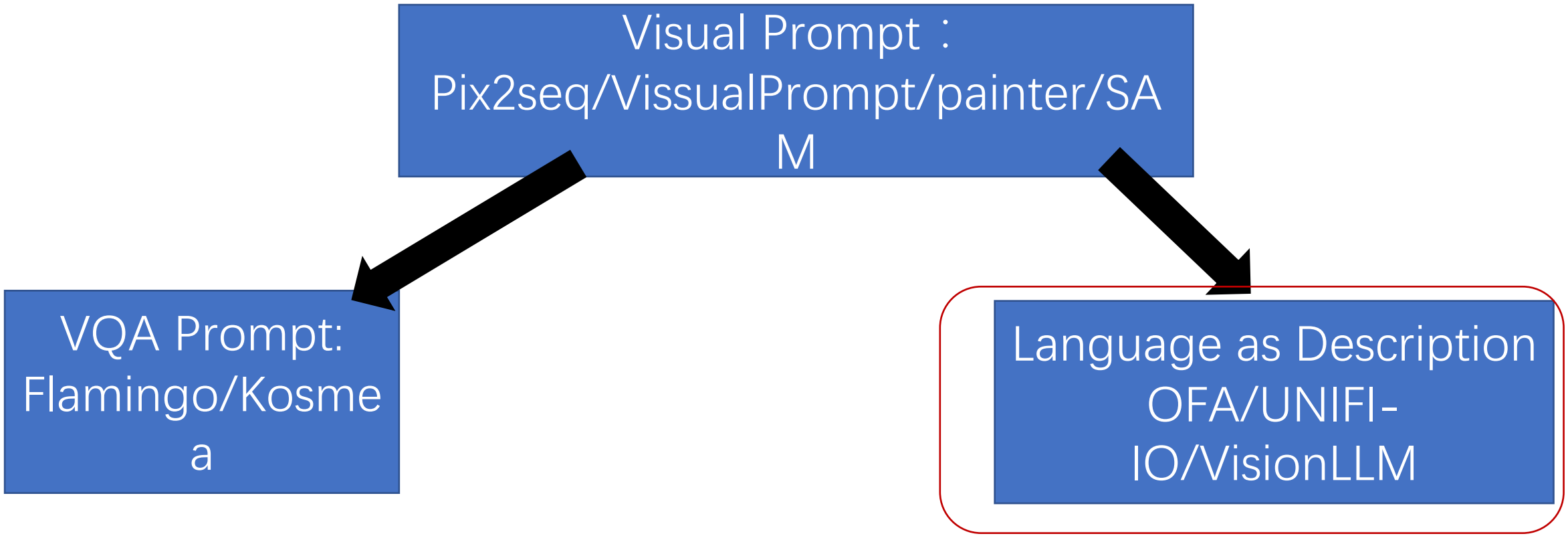
LLaVA



TinyLLaVA

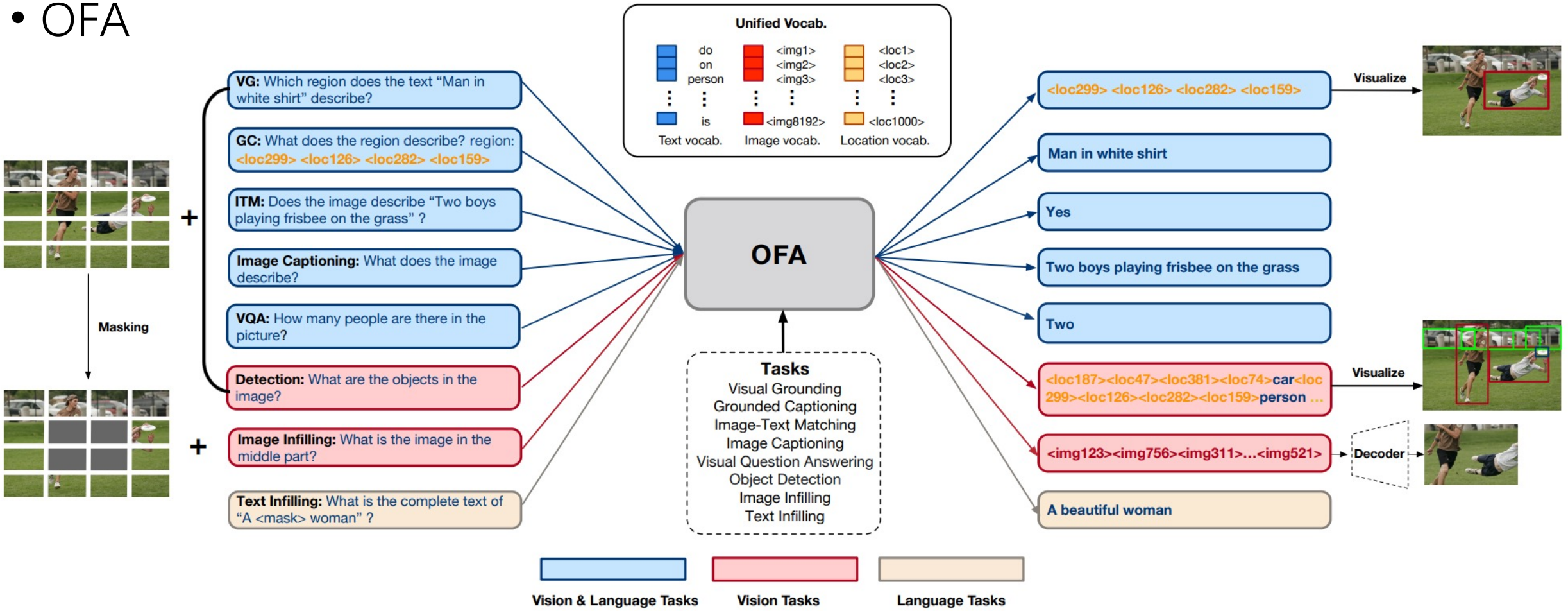
Zhou et al. TinyLLaVA: A Framework of Small-scale Large Multimodal Models. arXiv:2402.14289

多模态大模型（Prompt-like)



Language as Description

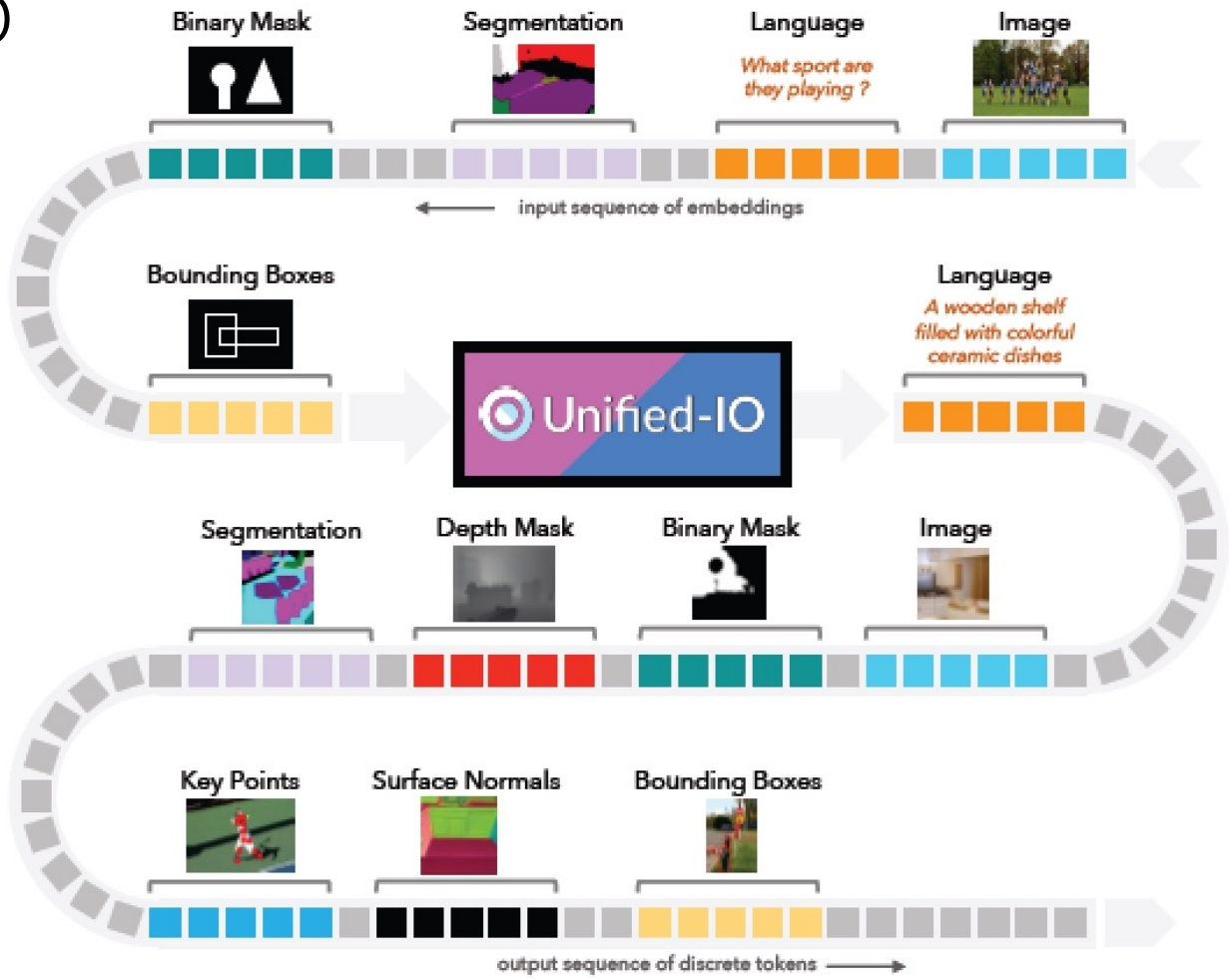
- OFA



OFA: UNIFYING ARCHITECTURES, TASKS, AND MODALITIES THROUGH A SIMPLE SEQUENCE-TO-SEQUENCE LEARNING FRAMEWORK. ICML 2022

Language as Description

- Unified-IO



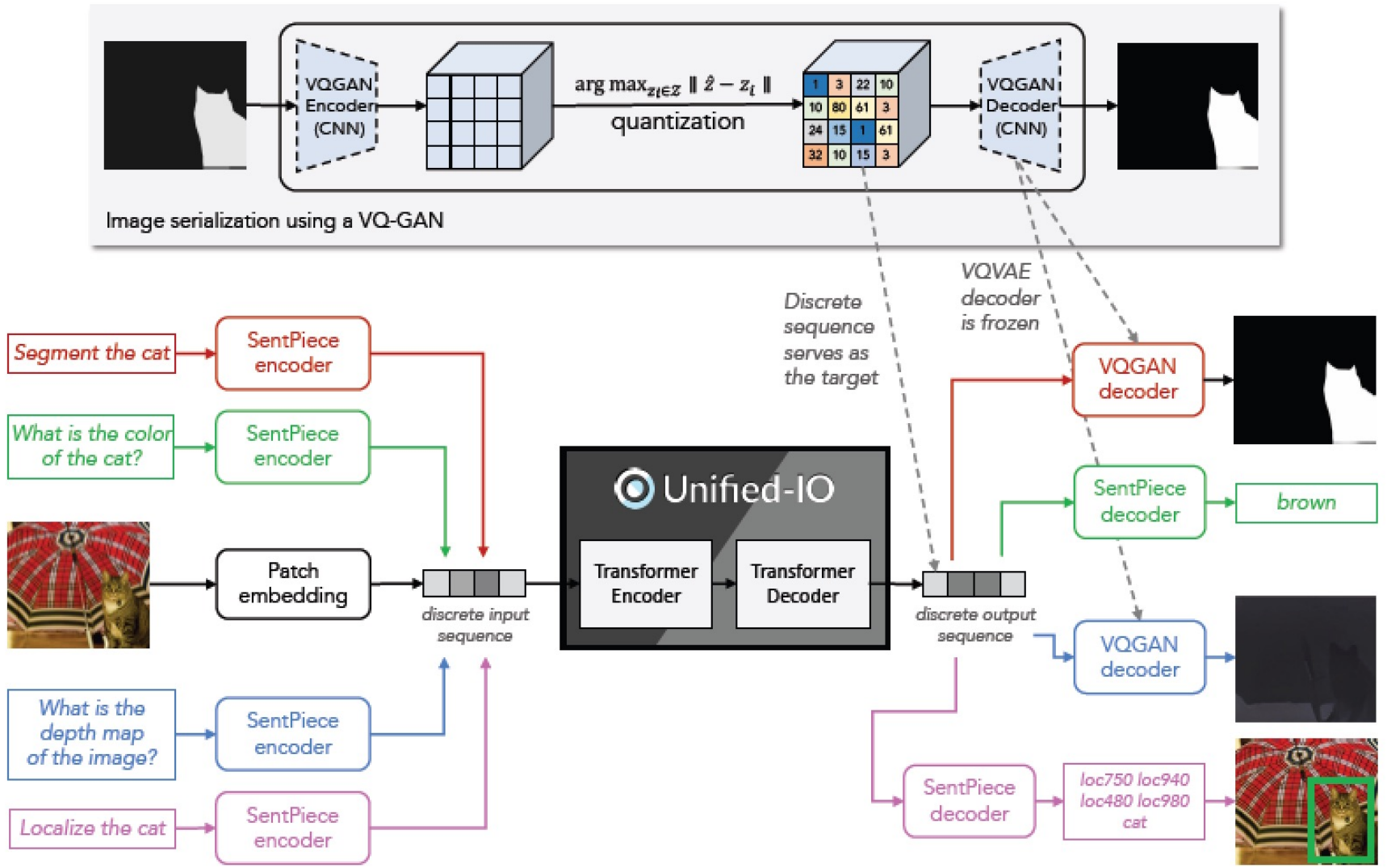
Tasks

Image Classification
Object Detection
Semantic Segmentation
Depth Estimation
Surface Normal Estimation
Segment-based Image Generation
Image Inpainting
Pose Estimation
Relationship Detection
Image Captioning
Visual QA
Referring Expressions
Situation Recognition
Text-based Image Generation
Visual Commonsense
Classification in context
Region Captioning
GLUE Benchmark tasks
Reading comprehension
Natural Language Inference
Grounded Commonsense Inference

UNIFIED-IO: A UNIFIED MODEL FOR VISION, LANGUAGE, AND MULTI-MODAL TASKS. ICLR 2023

Language as Description

- Unified-IO



UNIFIED-IO: A UNIFIED MODEL FOR VISION, LANGUAGE, AND MULTI-MODAL TASKS. ICLR 2023

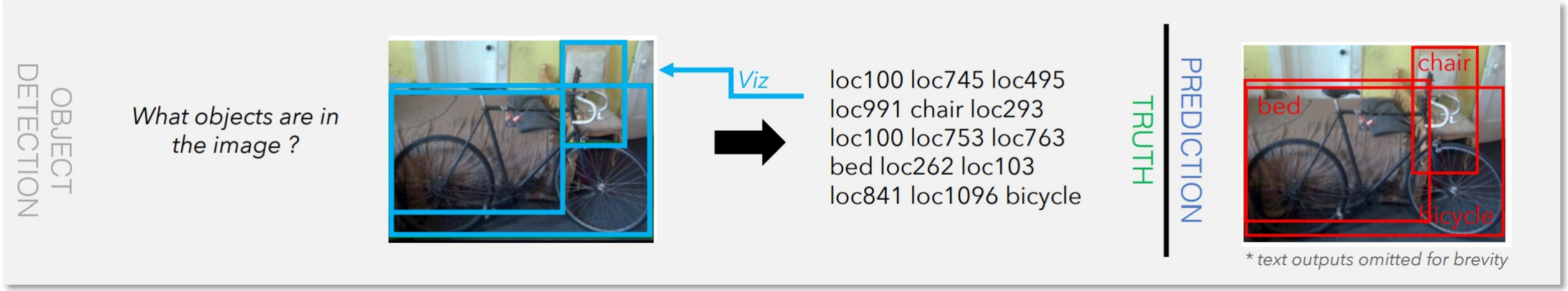
Language as Description

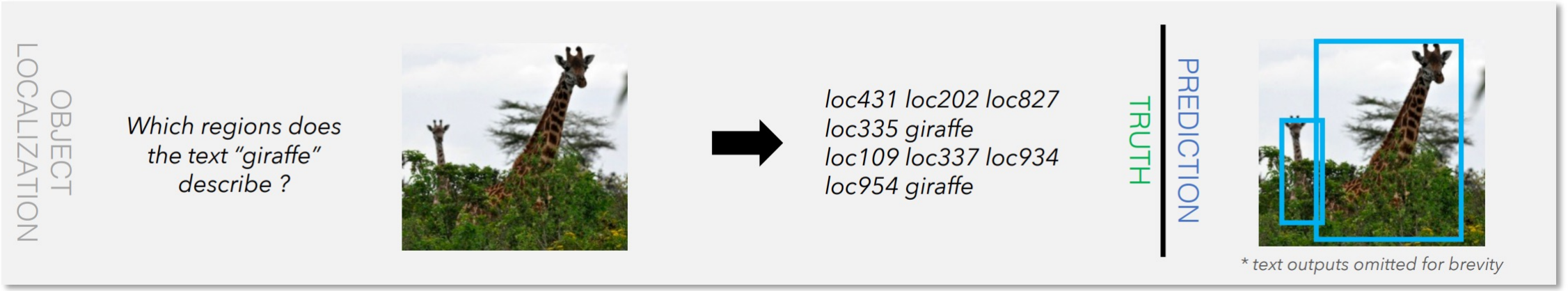
- Unified-IO: Image synthesis

UNIFIED-IO: A UNIFIED MODEL FOR VISION, LANGUAGE, AND MULTI-MODAL TASKS. ICLR 2023

Language as Description

- Unified-IO: Sparse label





UNIFIED-IO: A UNIFIED MODEL FOR VISION, LANGUAGE, AND MULTI-MODAL TASKS. ICLR 2023

Language as Description

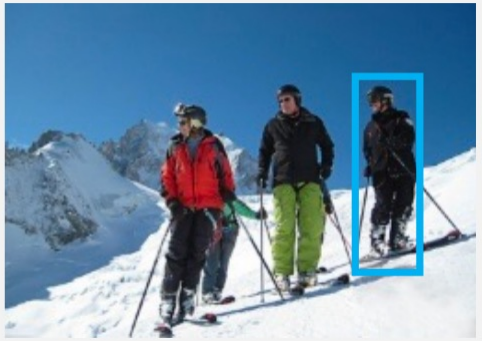
- Unified-IO: Sparse label

REFERRING EXPRESSIONS

Which region does the text "man in solid black" describe ?

loc270 loc856 loc640 loc982

TRUTH | PREDICTION



KEYPOINT ESTMATION

Find the human joints in the region "loc260 loc375 loc726 loc545".

loc307 loc487 loc299 loc499 loc295 loc481 loc315 loc507 loc305 loc455 loc369 loc517 loc359 loc413 loc457 loc527 loc429 loc387 loc423 loc529 ….

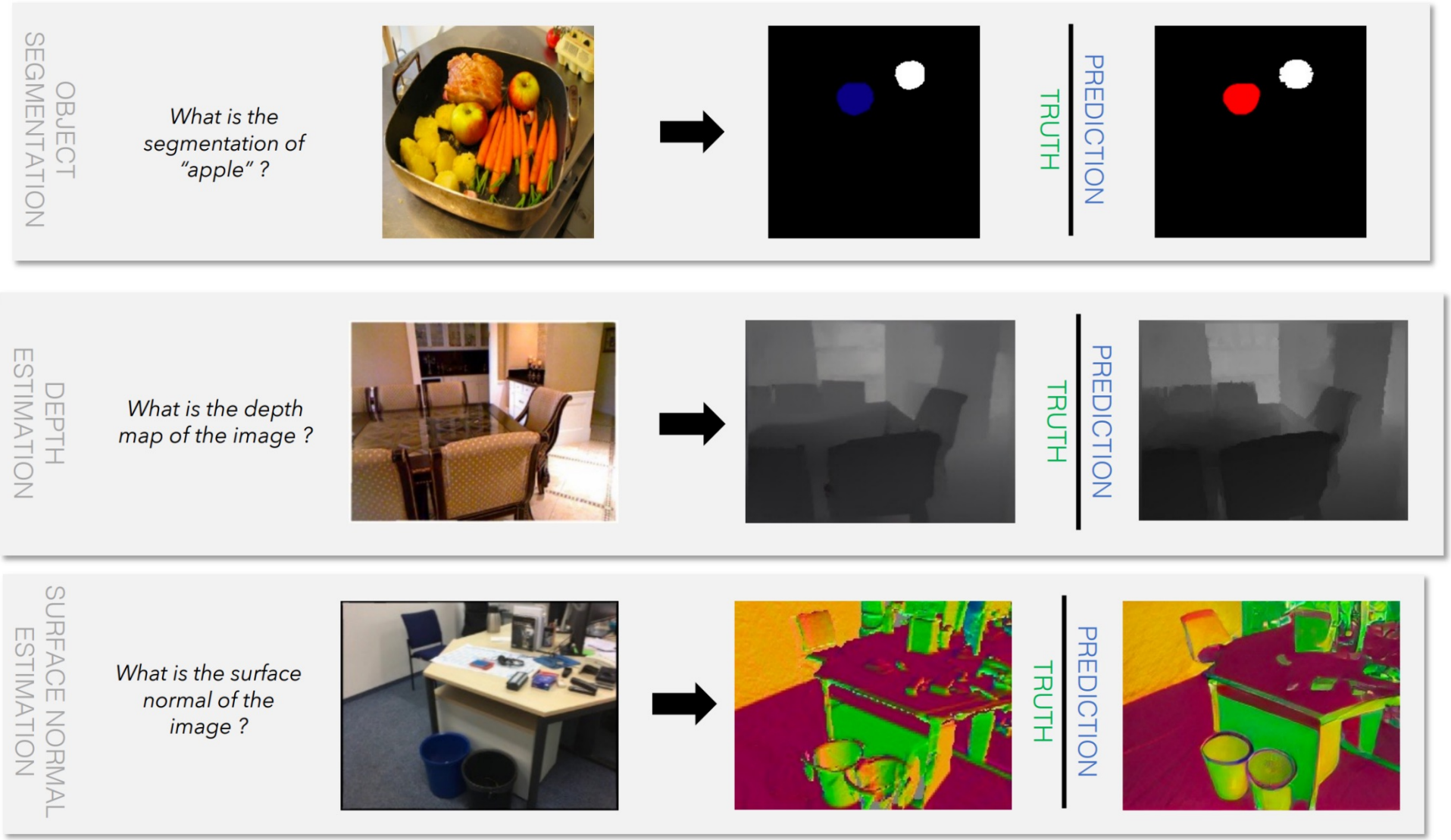
TRUTH | PREDICTION

** text outputs omitted for brevity*

UNIFIED-IO: A UNIFIED MODEL FOR VISION, LANGUAGE, AND MULTI-MODAL TASKS. ICLR 2023

Language as Description

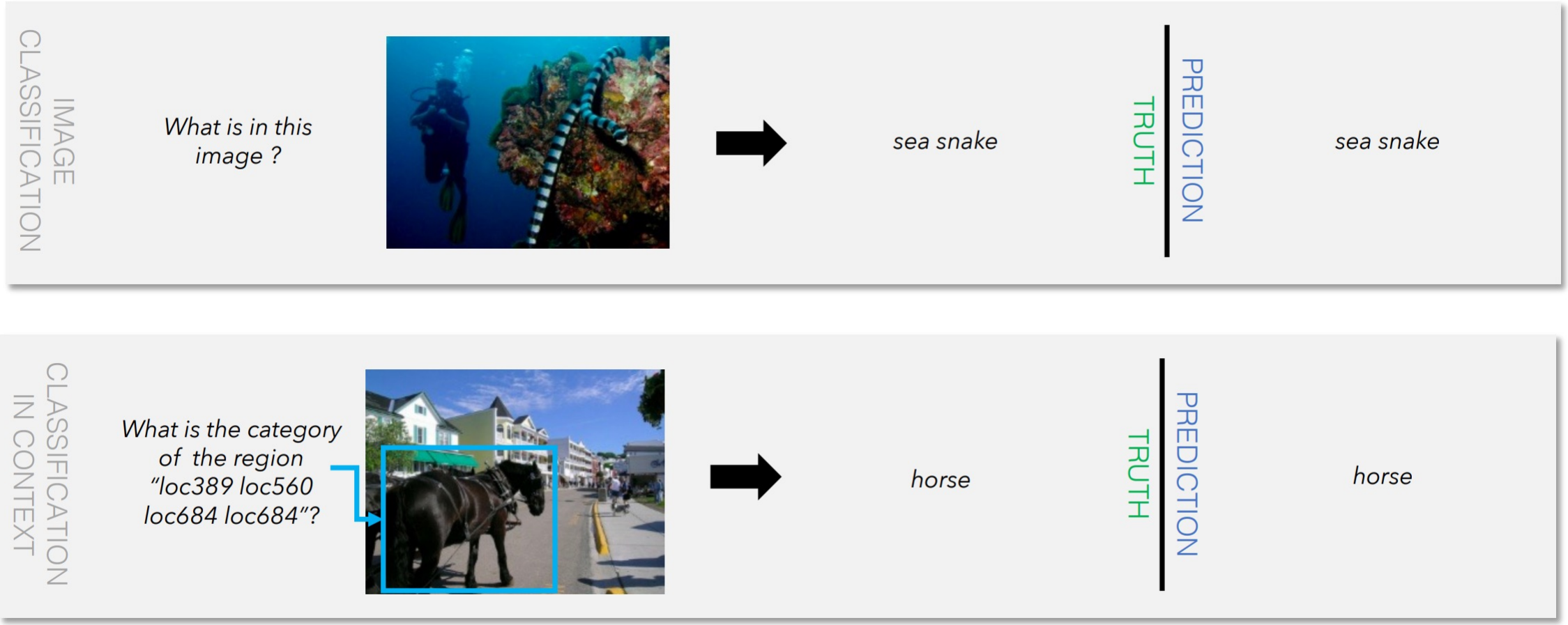
- Unified-IO
  - Dense Label



UNIFIED-IO: A UNIFIED MODEL FOR VISION, LANGUAGE, AND MULTI-MODAL TASKS. ICLR 2023

Language as Description

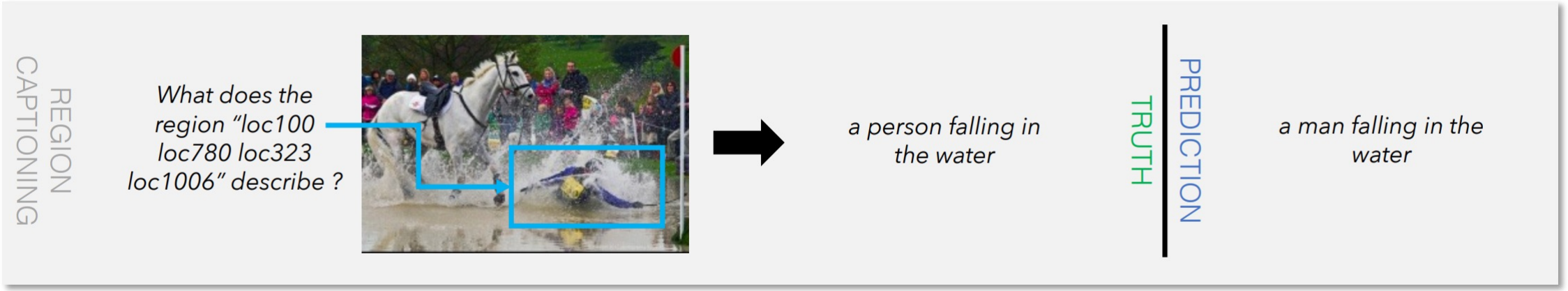
- Unified-IO: Image Classification



UNIFIED-IO: A UNIFIED MODEL FOR VISION, LANGUAGE, AND MULTI-MODAL TASKS. ICLR 2023

Language as Description

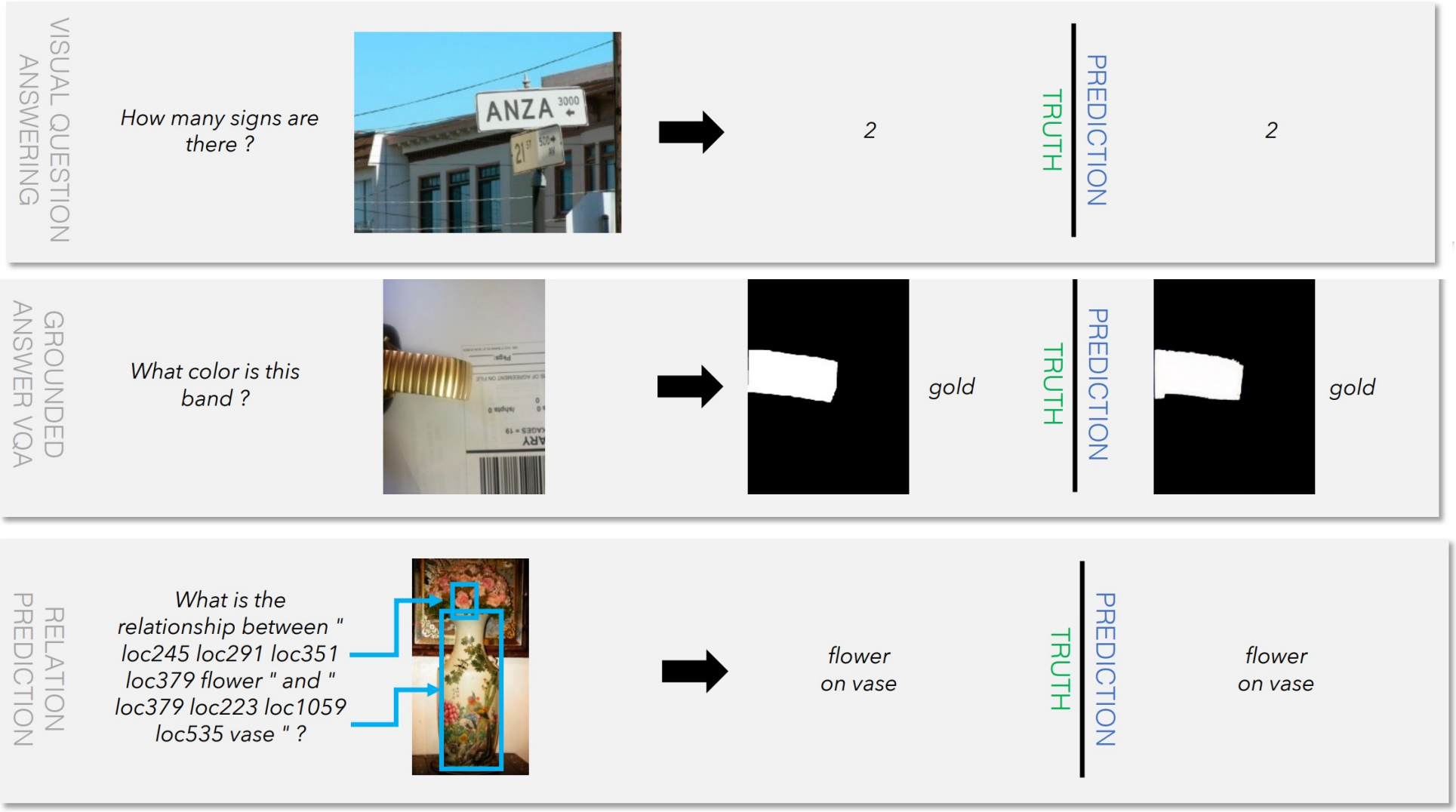
- Unified-IO: Image Captioning Tasks



UNIFIED-IO: A UNIFIED MODEL FOR VISION, LANGUAGE, AND MULTI-MODAL TASKS. ICLR 2023

Language as Description

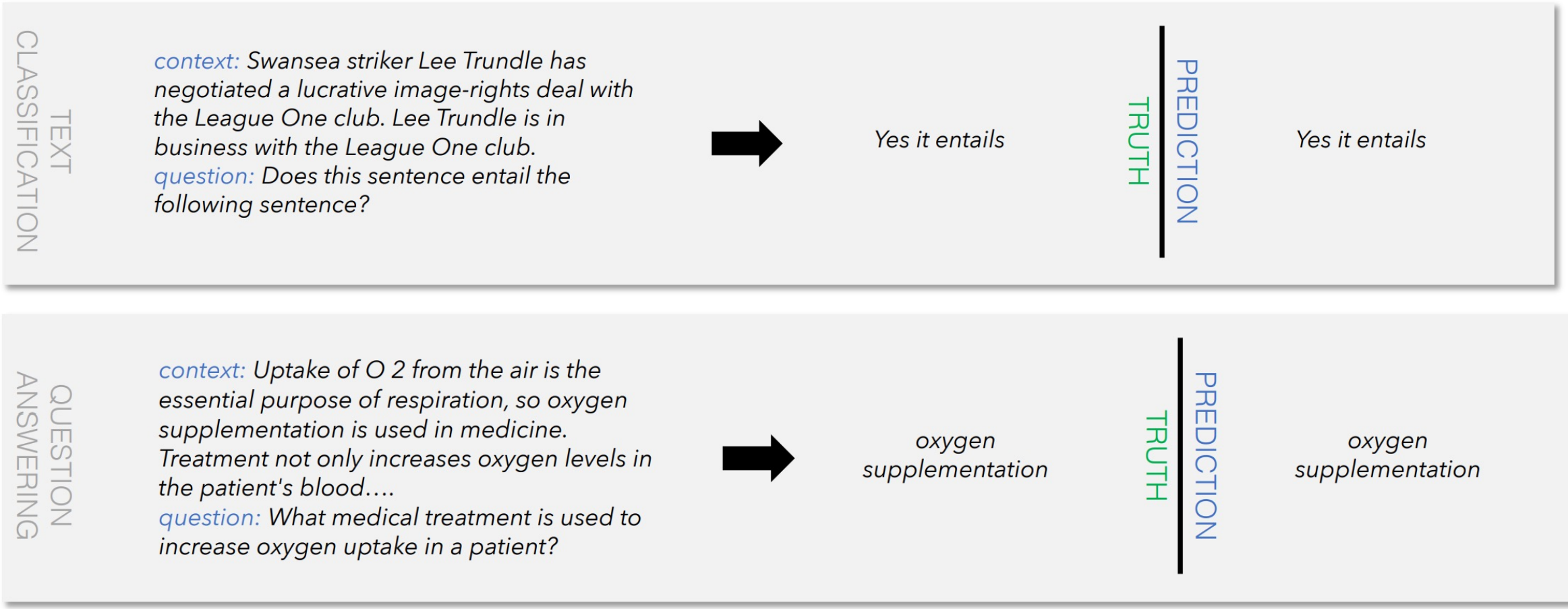
- Unified-IO
  - V&L Tasks



UNIFIED-IO: A UNIFIED MODEL FOR VISION, LANGUAGE, AND MULTI-MODAL TASKS. ICLR 2023

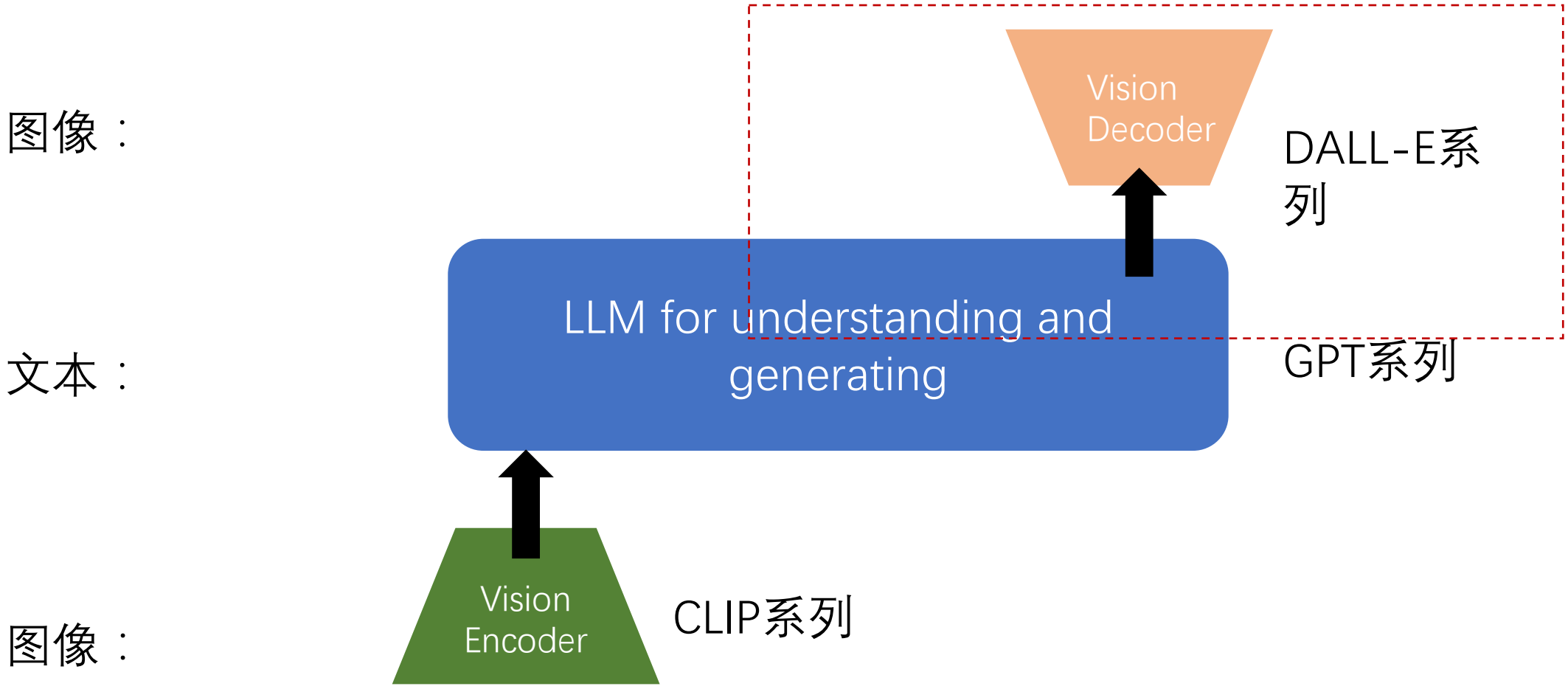
Language as Description

- Unified-IO：NLP Tasks



UNIFIED-IO: A UNIFIED MODEL FOR VISION, LANGUAGE, AND MULTI-MODAL TASKS. ICLR 2023

多模态和视觉大模型


图像：
文本：
图像：
Vision Decoder
DALL-E系列
LLM for understanding and generating
GPT系列
Vision Encoder
CLIP系列

谢谢！