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

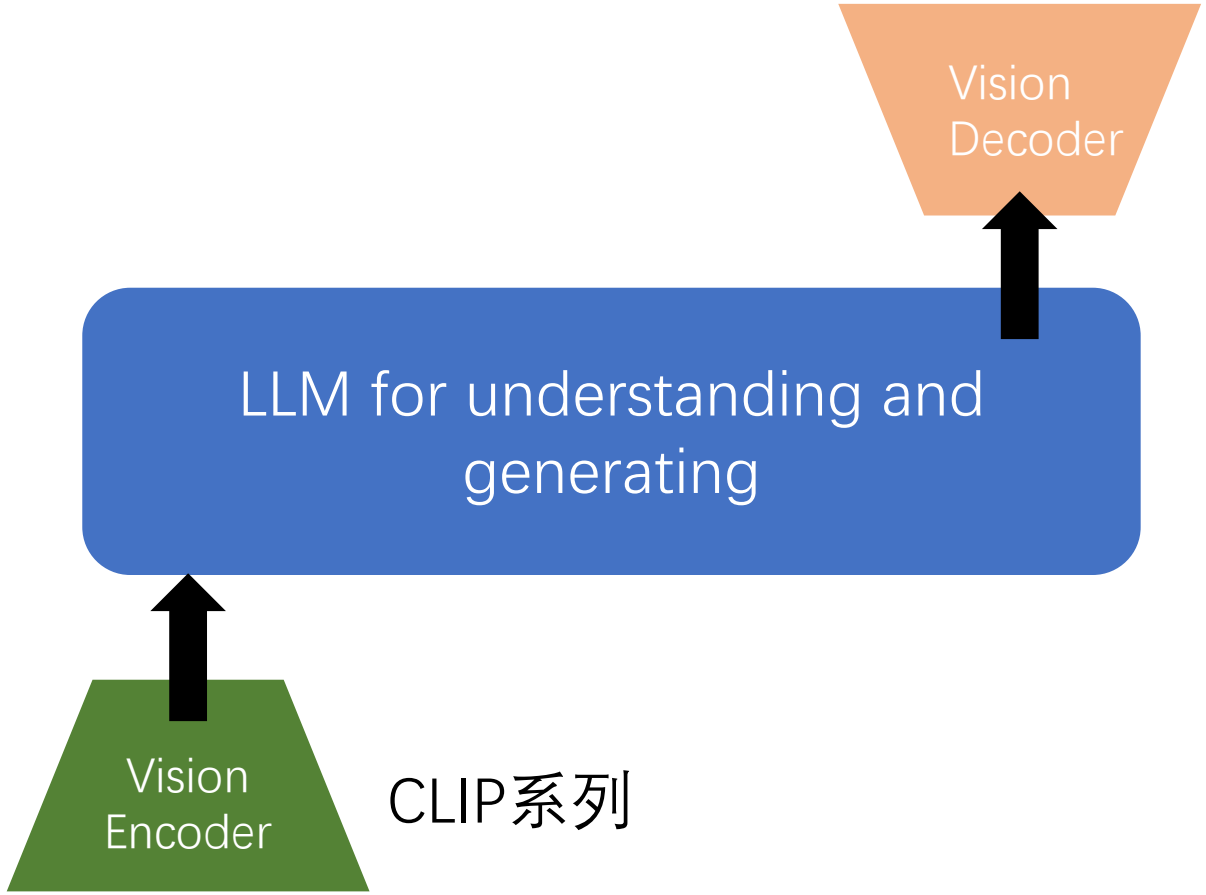
多模态和视觉大模型

Open-AI路线

图像：

文本：

图像：



DALL-E系列

GPT系列

CLIP系列

00

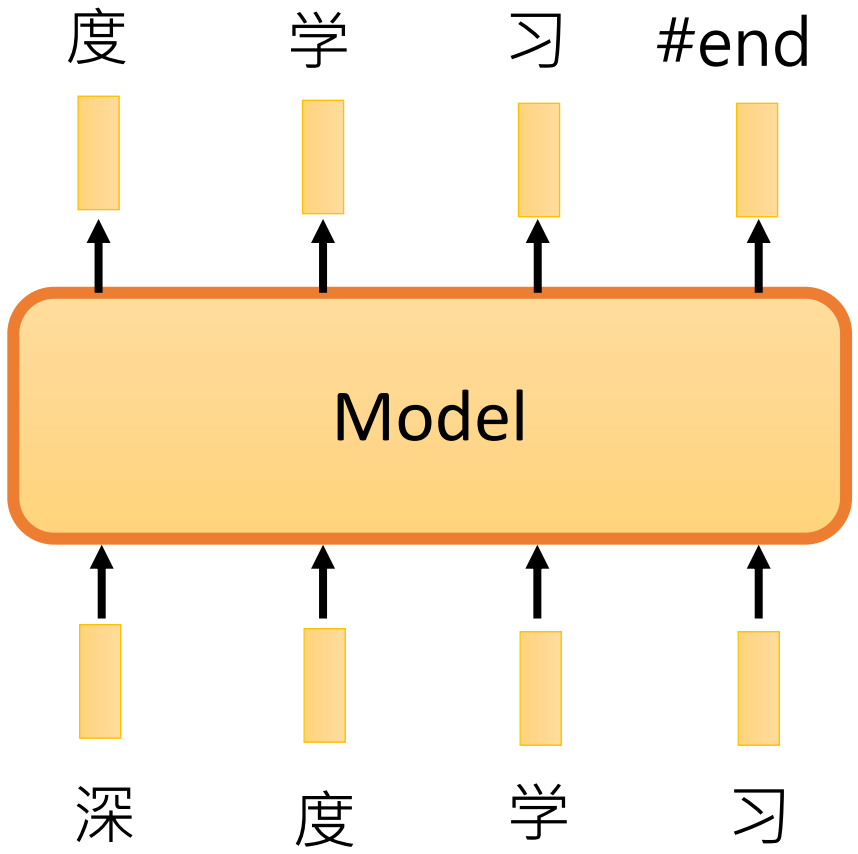
语言自回归模型（回顾）

Language Model

- Auto-Regressive (自回归)
  - GPTs

深	0.1
度	0.7
学	0.1
习	0.05
#end	0.05

$$P(w_t|w_1, w_2, \ldots, w_{t-1})$$



(all characters)

One-hot Encoding

深 = [1 0 0 0 0]

度 = [0 1 0 0 0]

学 = [0 0 1 0 0]

习 = [0 0 0 1 0]

#end = [0 0 0 0 1]

Vision Model

Explicit density model

$$p(x) = p(x_1, x_2, \ldots, x_n)$$

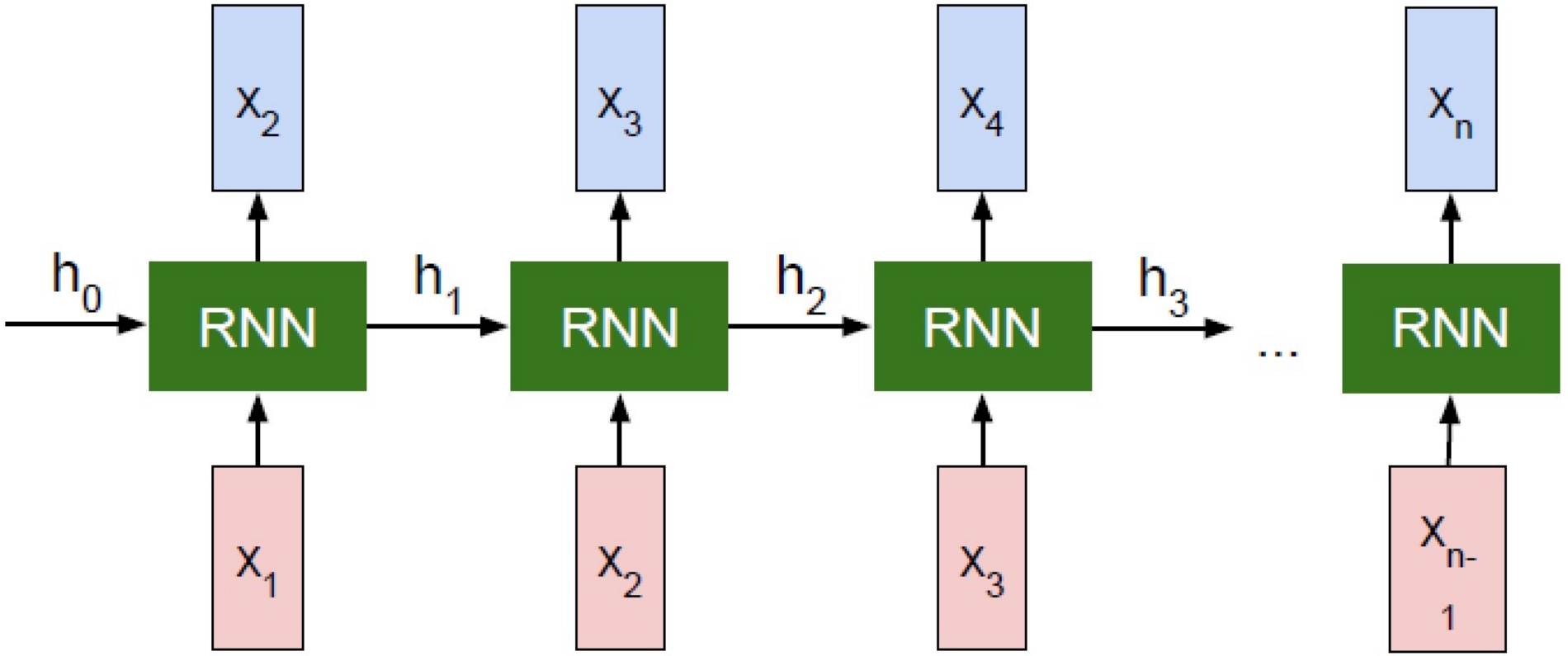
Likelihood of image x

Joint likelihood of each pixel in the image

Fully visible belief network(FVBN)

Incontext Modelling

Recurrent Neural Network(RNN)

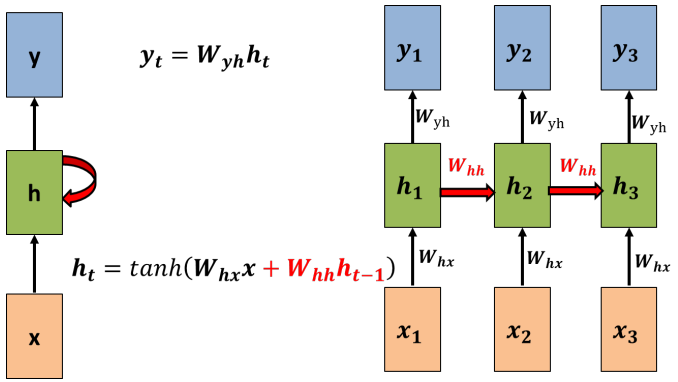


$$p(x_i|x_1, \ldots, x_{i-1})$$

Incontext Modelling

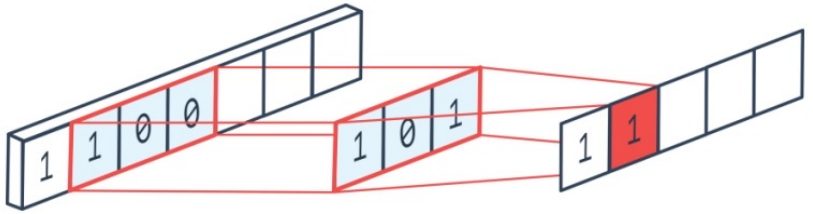
$X \in \mathcal{R}^{N \times \mathrm{d}}$ $Y \in B^{N \times c} / Y \in R^{N \times c}$

- RNN



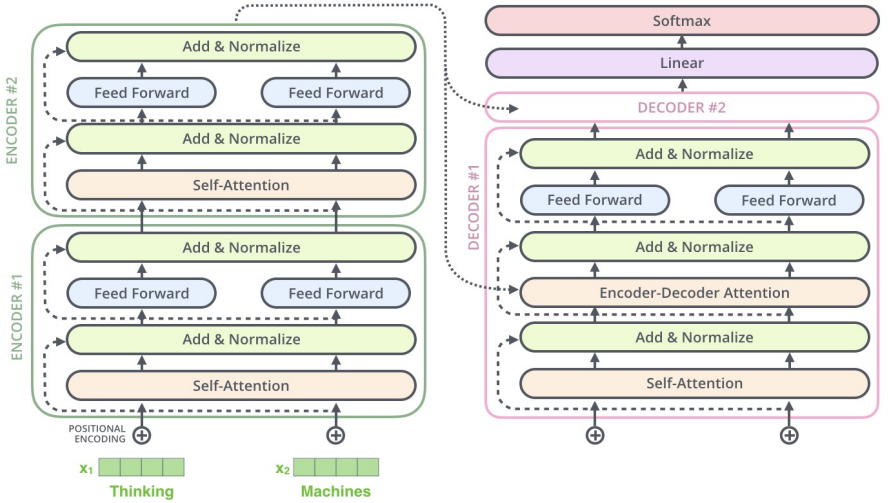
- Theoretically with infinite context window
- Practically suffer from vanishing/exploding gradient
- Training: $O(N)$, not parallelizable
- Inference: constant time for each token.

- CNN



- Finite context window (depending on kernel size)
- Need to materialize the kernel before using it.
- Training and inference depend on kernel size.
- Easily parallelizable

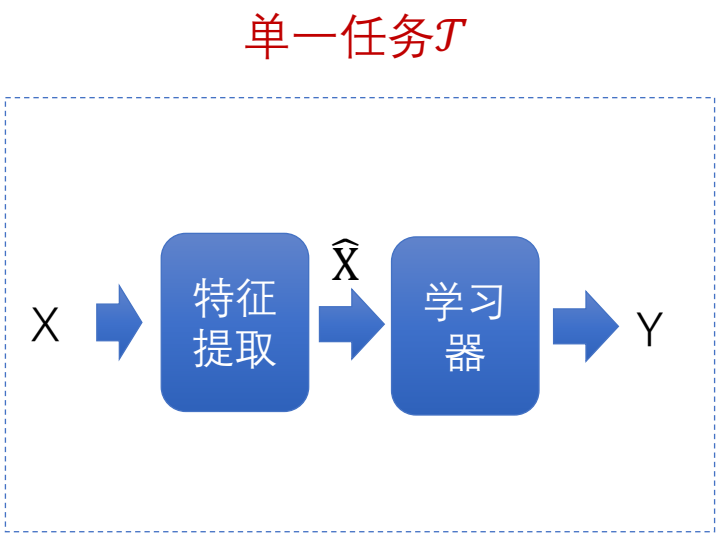
- Transformer



- Finite context window
- Training: $O(N^2)$
- easily parallelizable
- Inference:$O(N^2)$; may $O(N)$ when using KV-Cache, for each token.

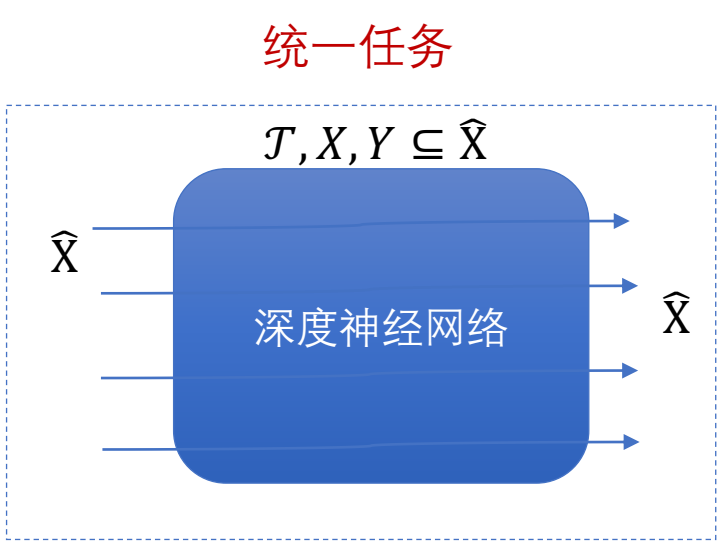
大语言模型

- 大语言模型的核心功能和特点
  - 统一任务
  - 自回归模型(Auto-regressive), 自监督学习（Self-supervised Learning）



特征工程时代

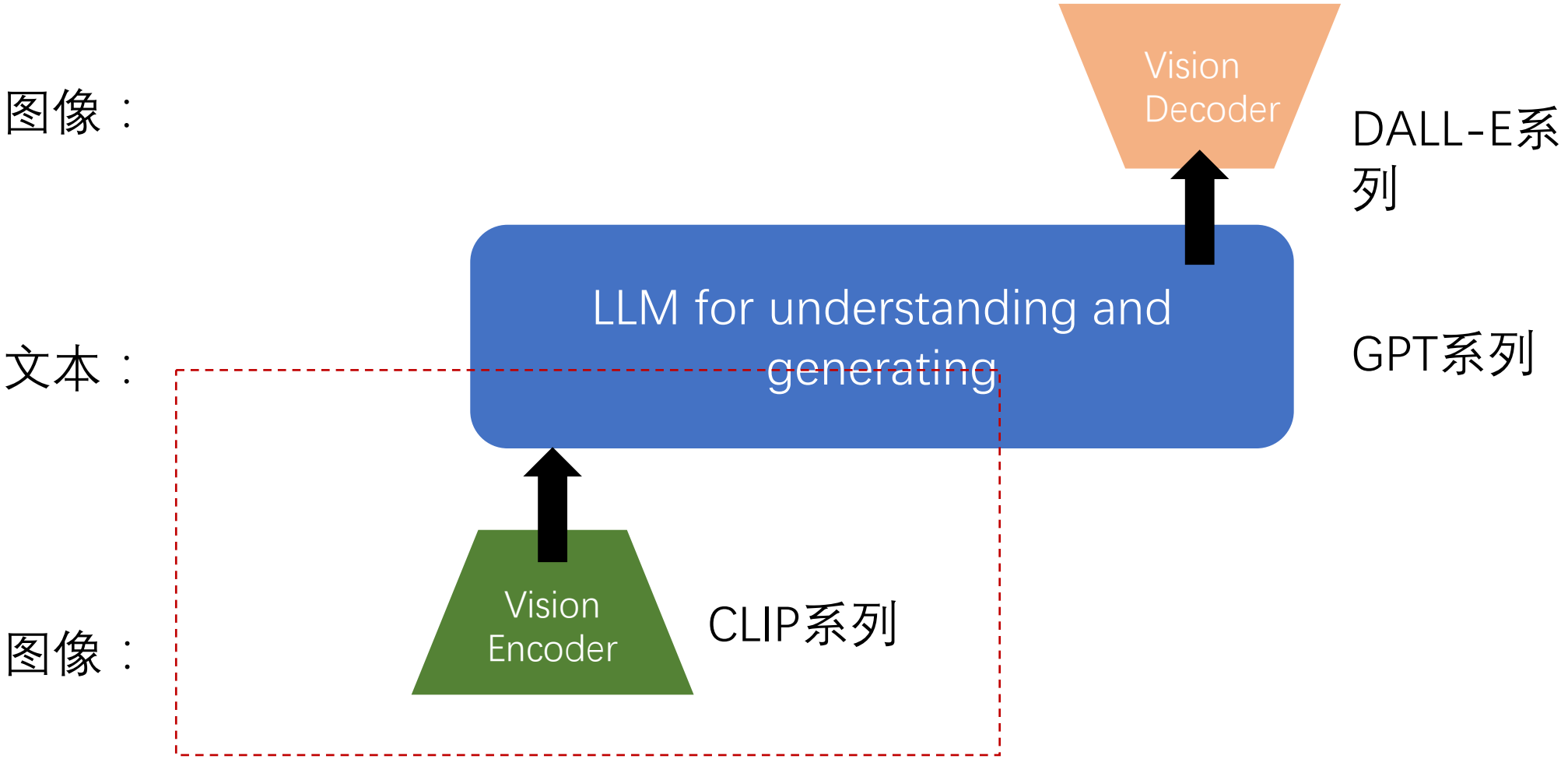
深度学习时代



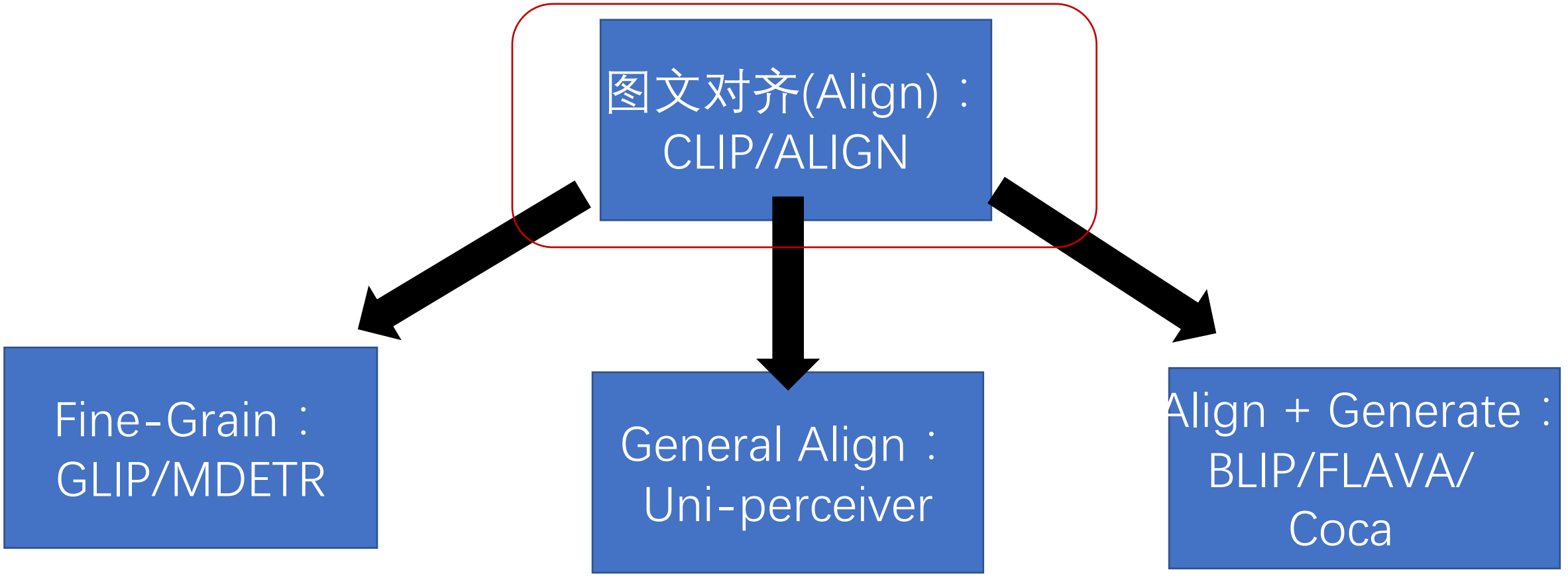
大模型时代

01 模态对齐

多模态和视觉大模型

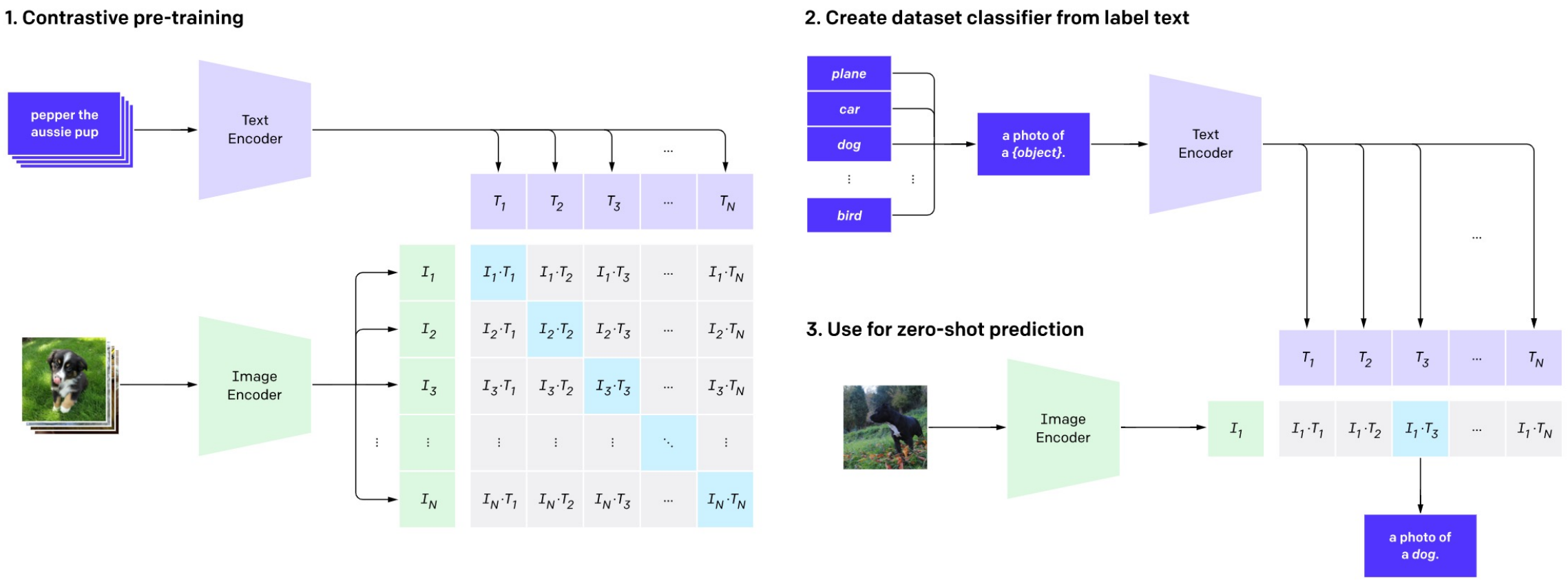

图像：
文本：
图像：
Vision Decoder
DALL-E系列
LLM for understanding and generating
GPT系列
Vision Encoder
CLIP系列

多模态大模型 (BERT-like)



Align the text and vision

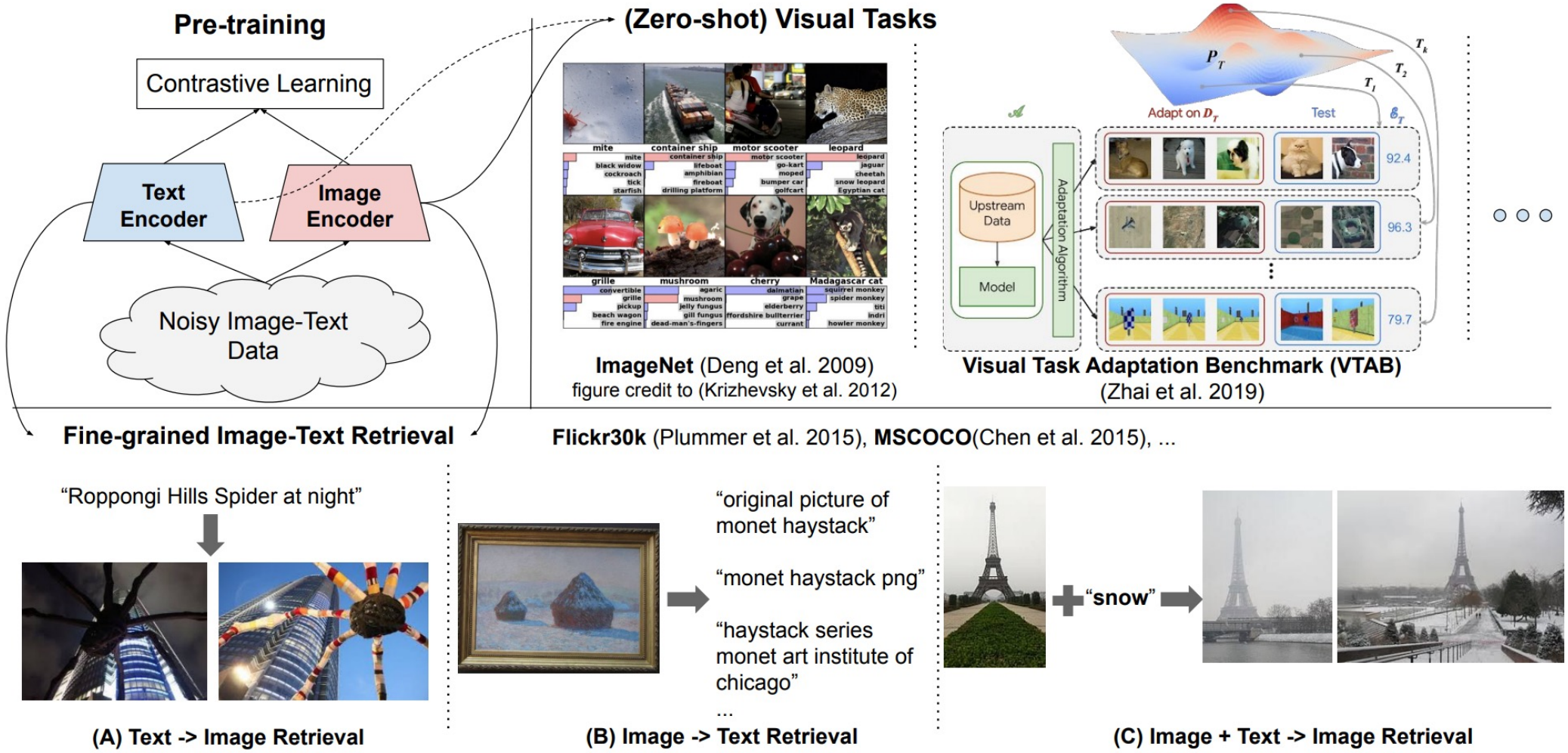
- CLIP: Contrastive Language-Image Pre-training



Learning Transferable Visual Models From Natural Language Supervision, ICML 2021

Align the text and vision

- ALIGN: A Large-scale ImaGe and Noisy-text embedding

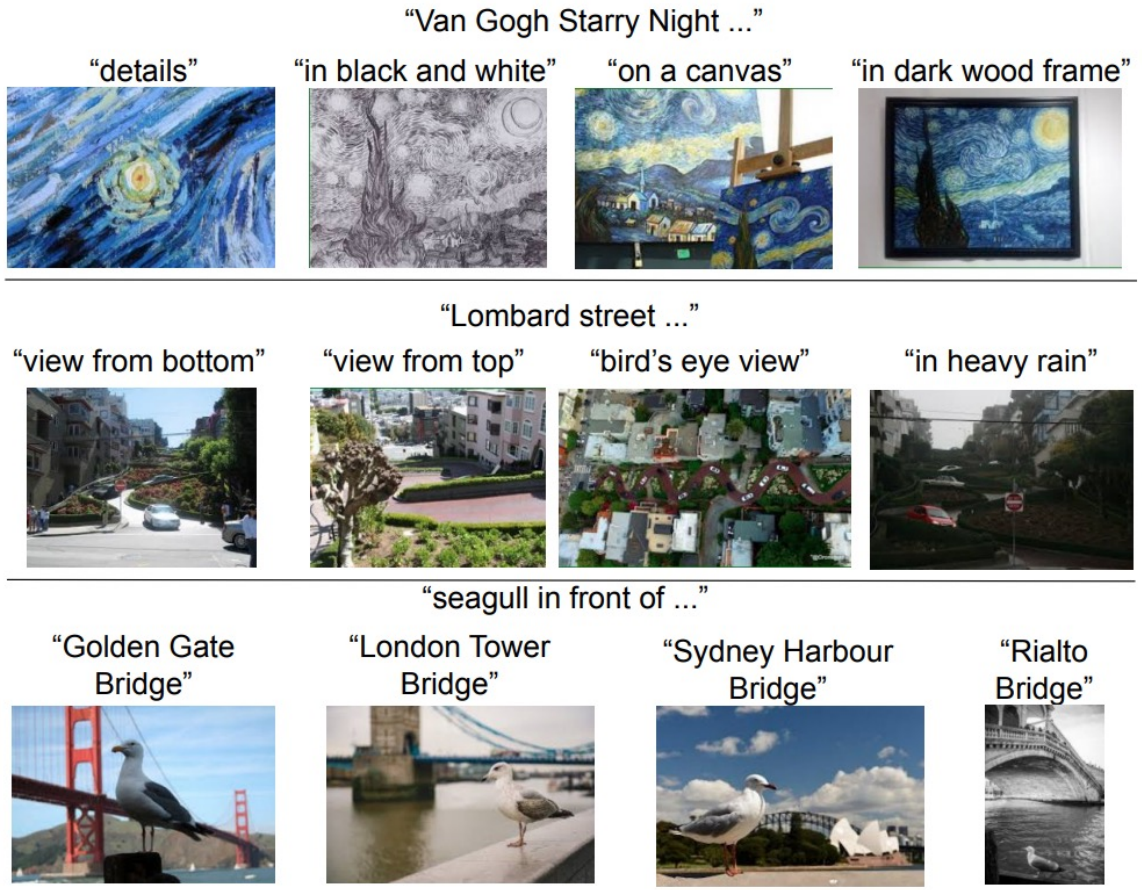


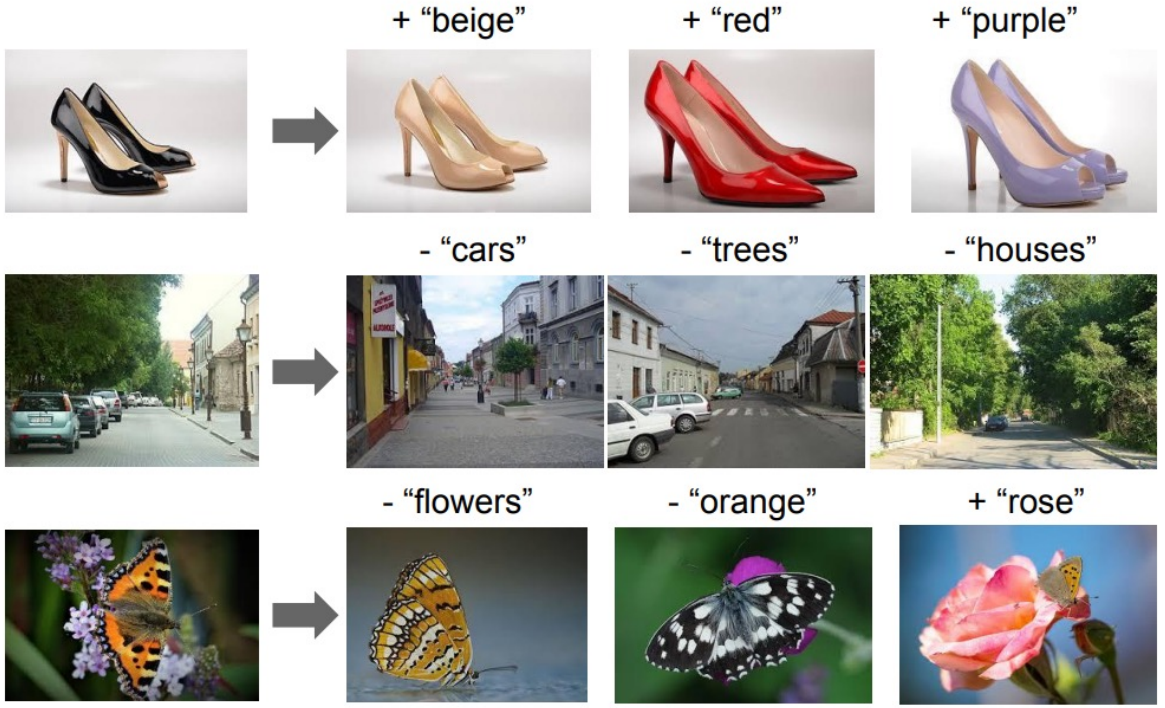
Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. ICML 2021

Align the text and vision

- ALIGN

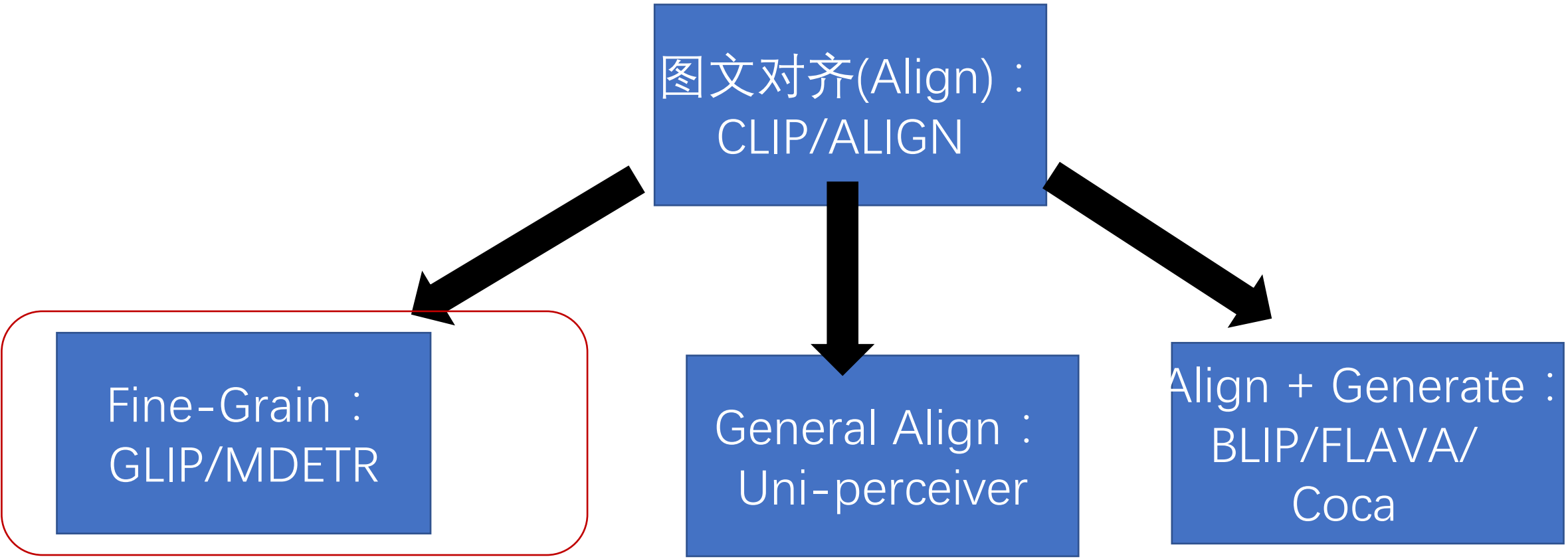
Retrieval (检索):





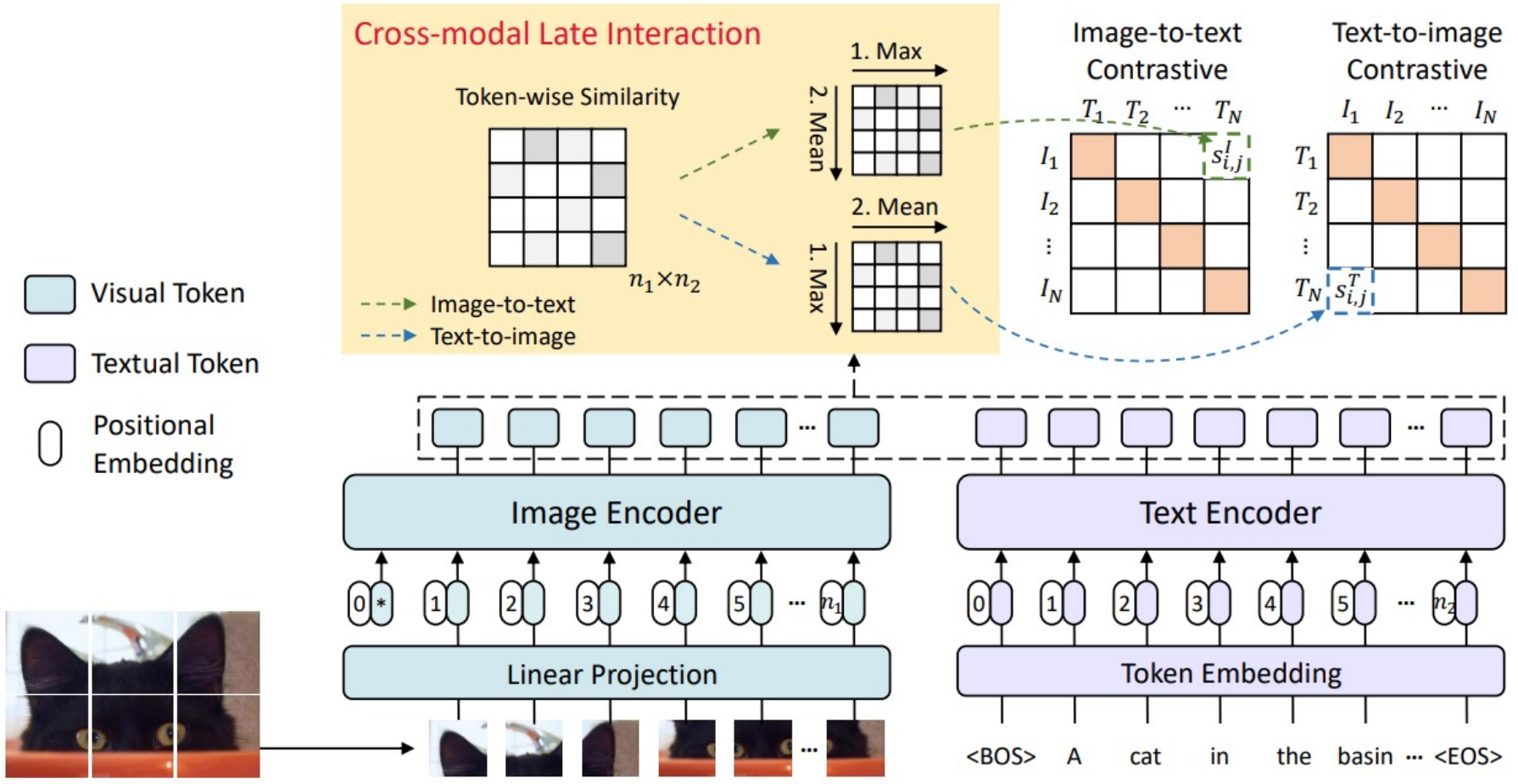
Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. ICML 2021

多模态大模型 (BERT-like)



Fine-grained Align the text and vision

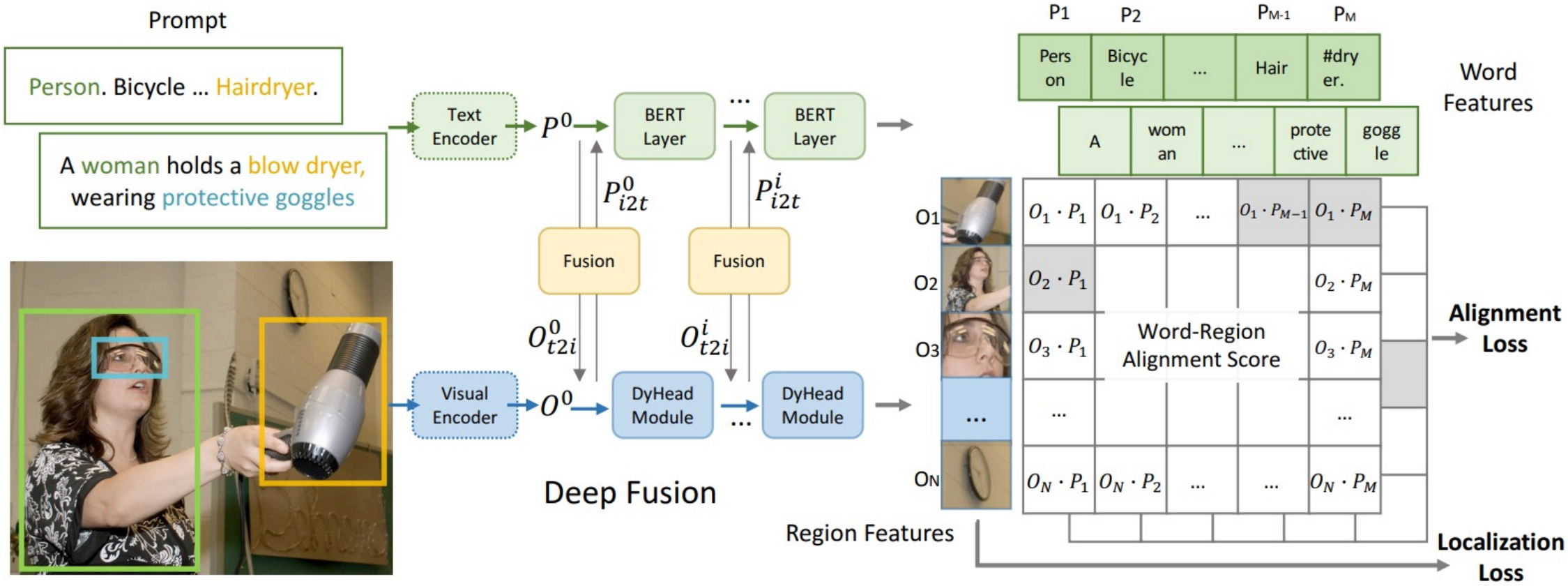
- FILIP: Fine-grained Interactive Language-Image Pre-training



FILIP: FINE-GRAINED INTERACTIVE LANGUAGEIMAGE PRE-TRAINING. ICLR 2022

Fine-grained Align the text and visual

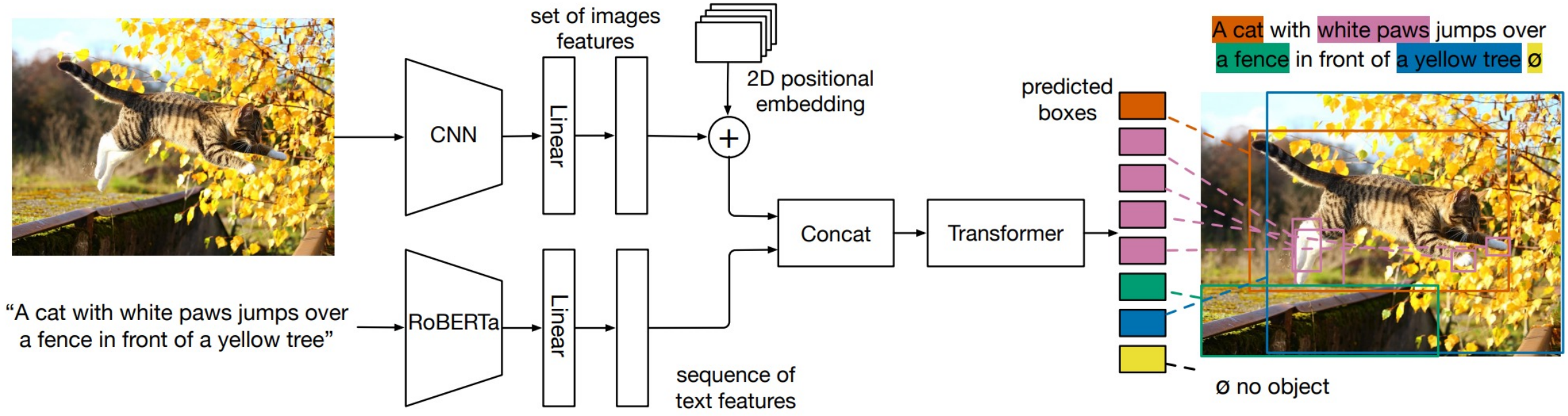
- GLIP: grounded language-image pretraining



Grounded Language-Image Pre-training. CVPR 2022

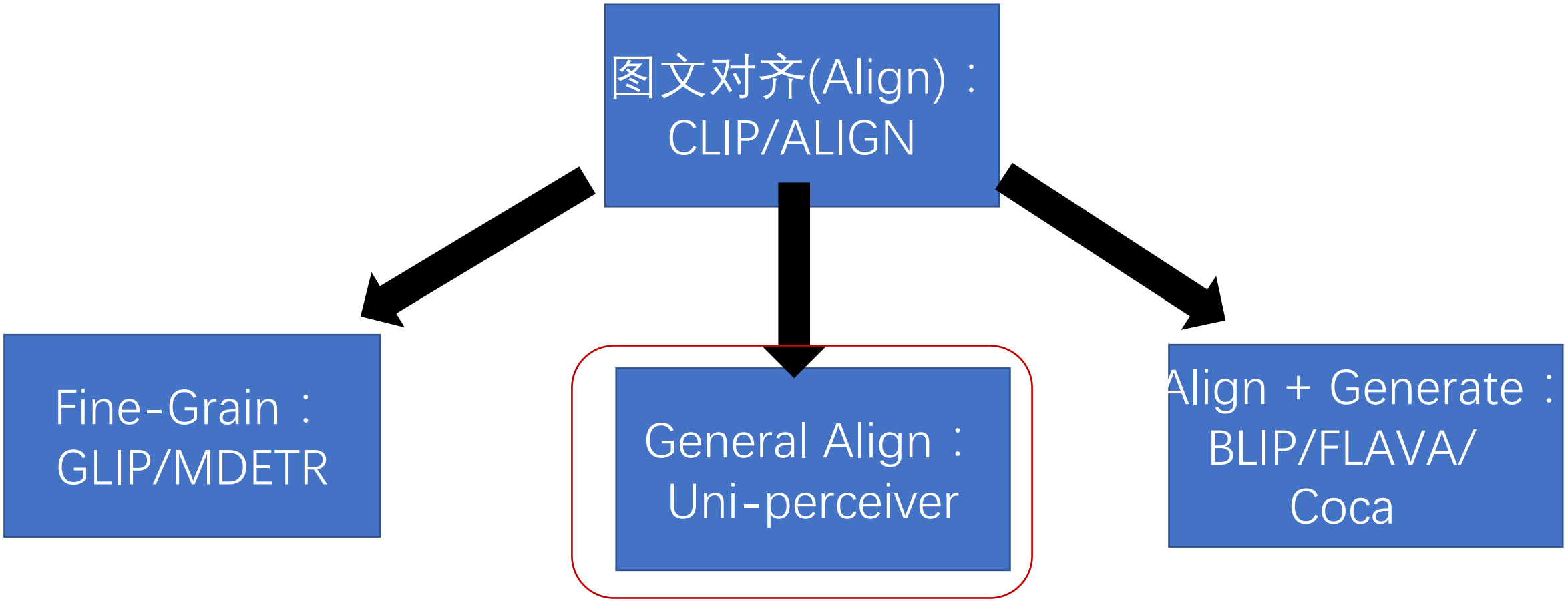
Fine-grained Align the text and visual

- MDETR: modulated detector based on DETR



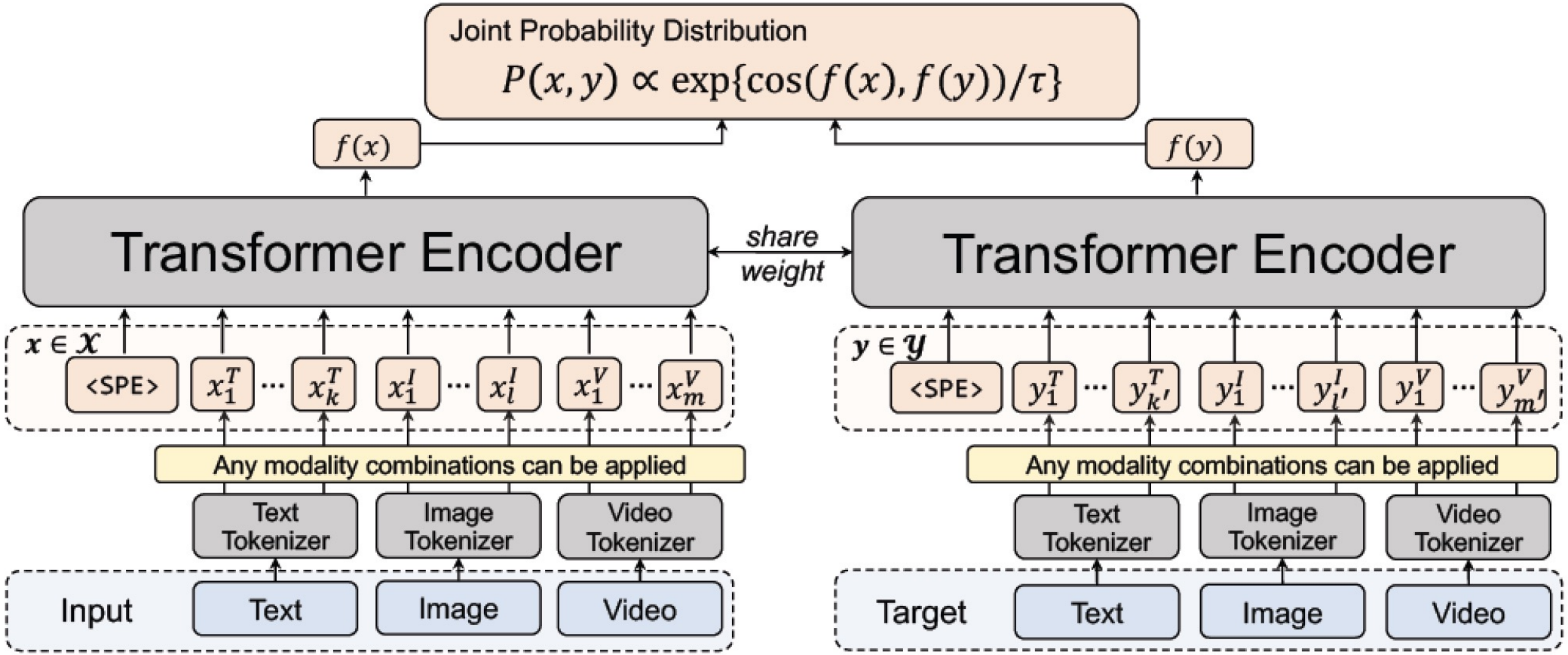
MDETR - Modulated Detection for End-to-End Multi-Modal Understanding. CVPR 2022

多模态大模型 (BERT-like)



General Align

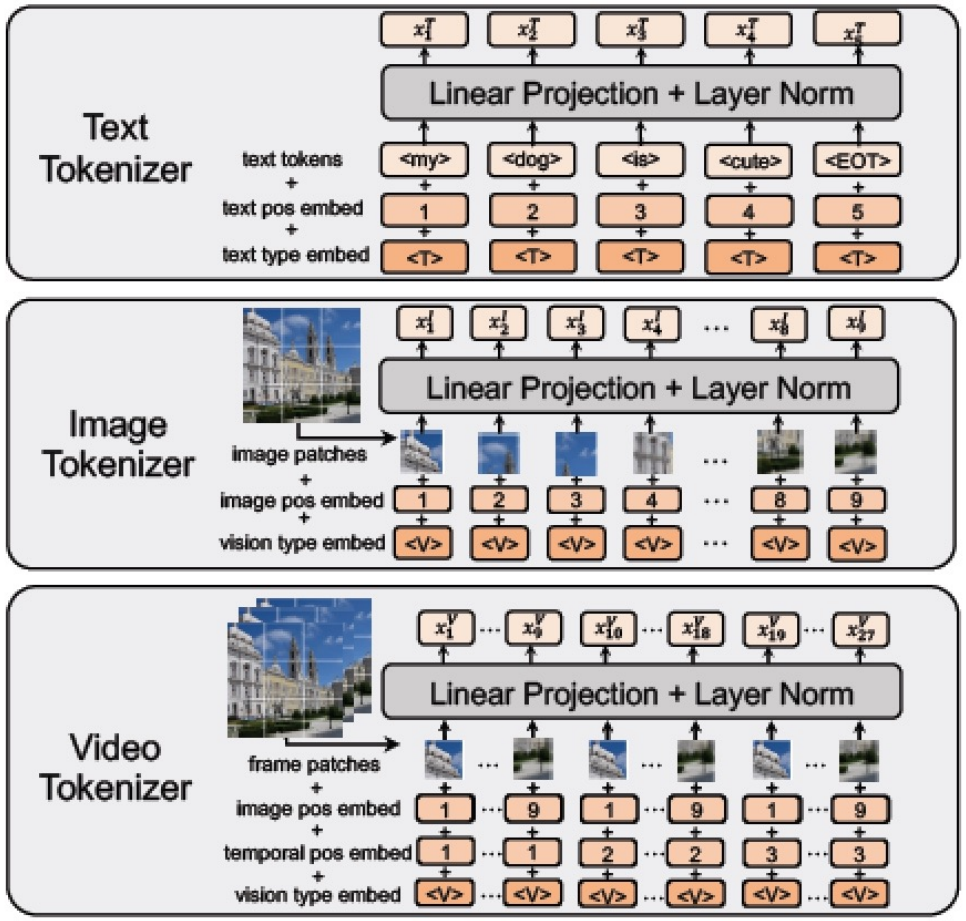
- Uni-Perceiver
  - No specific design for tasks: focus on the input and output process

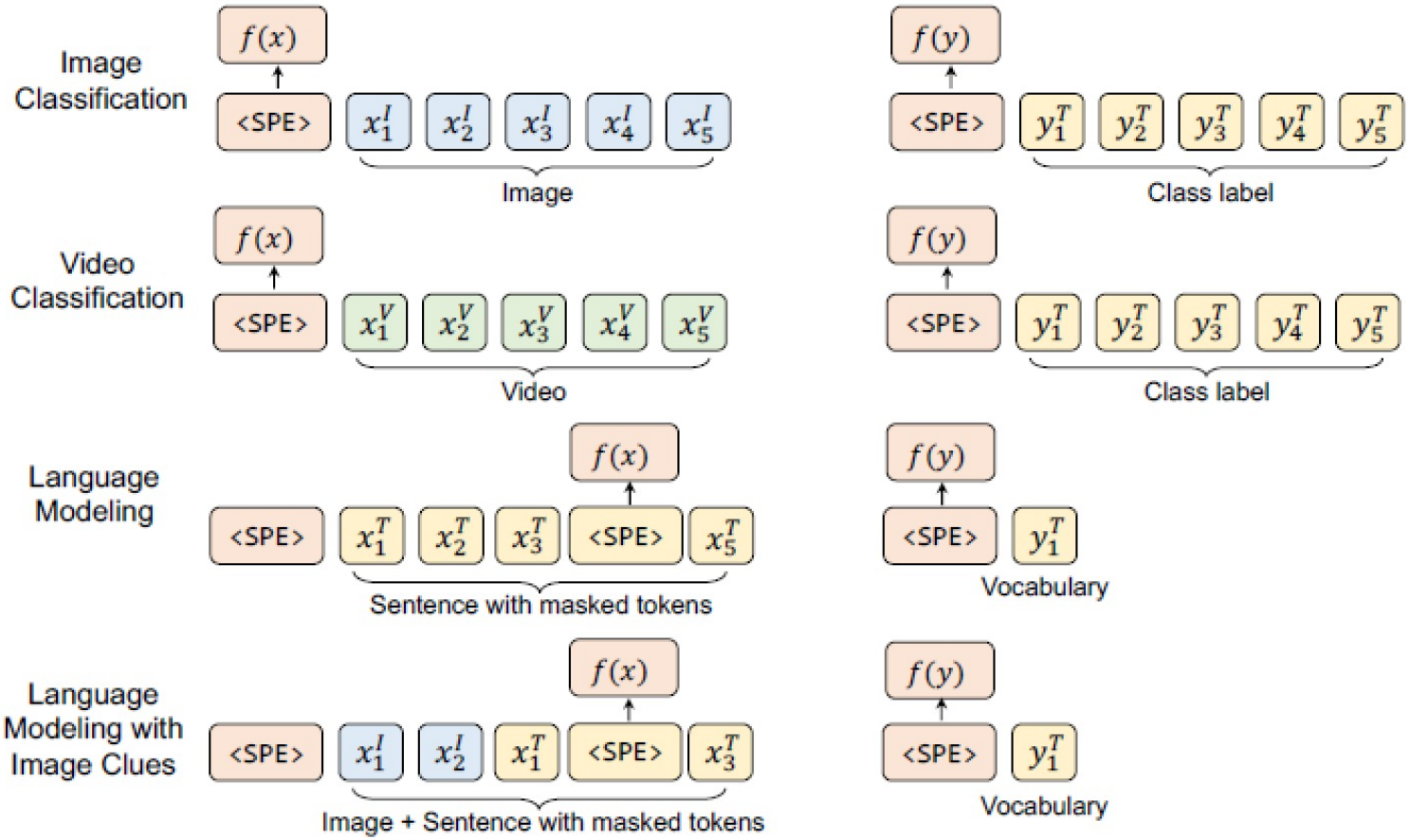


Uni-Perceiver: Pre-training Unified Architecture for Generic Perception for Zero-shot and Few-shot Tasks. CVPR 2022

General Align

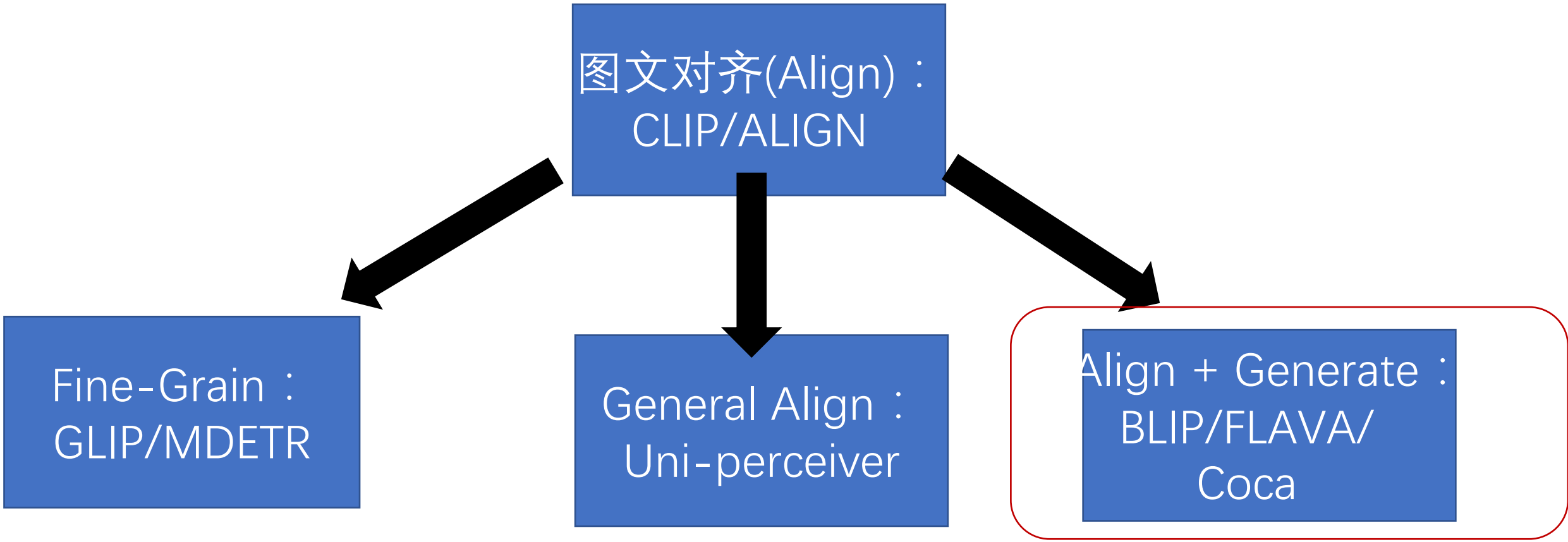
- Uni-Perceiver





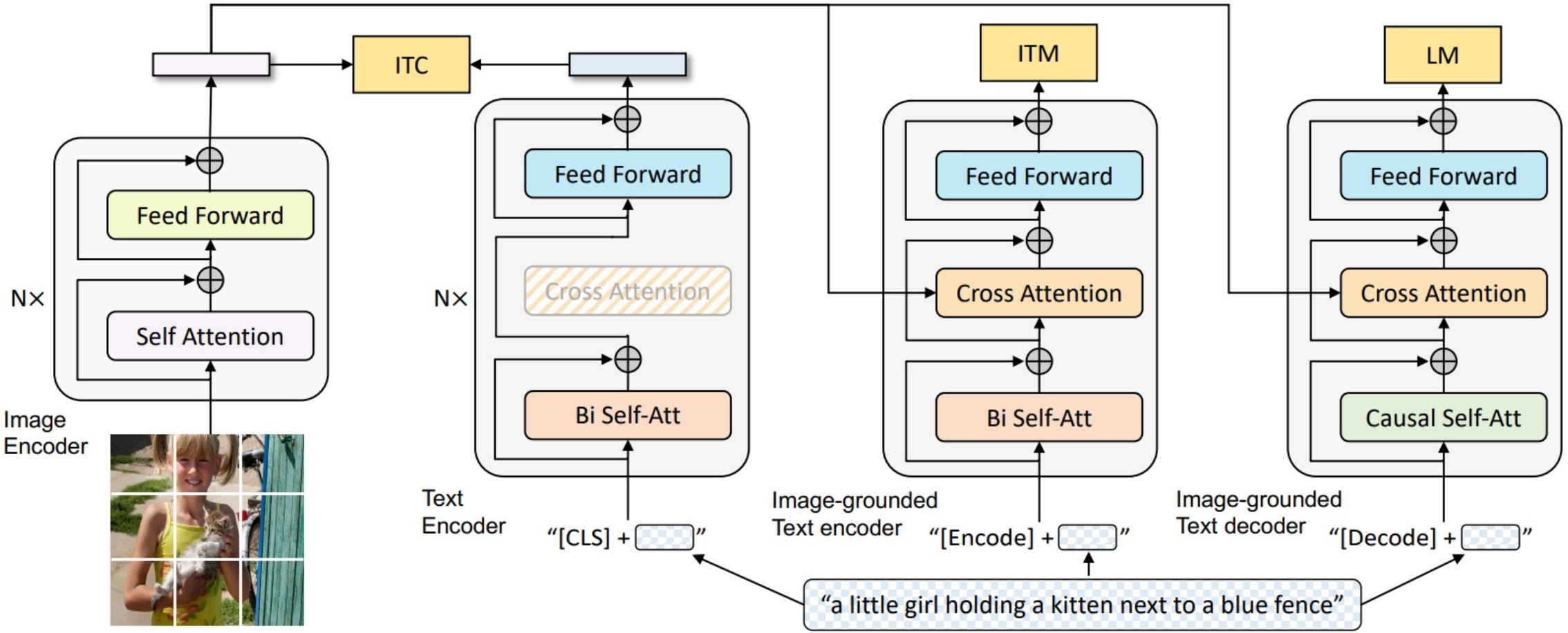
Uni-Perceiver: Pre-training Unified Architecture for Generic Perception for Zero-shot and Few-shot Tasks. CVPR 2022

多模态大模型 (BERT-like)



Align + Generate

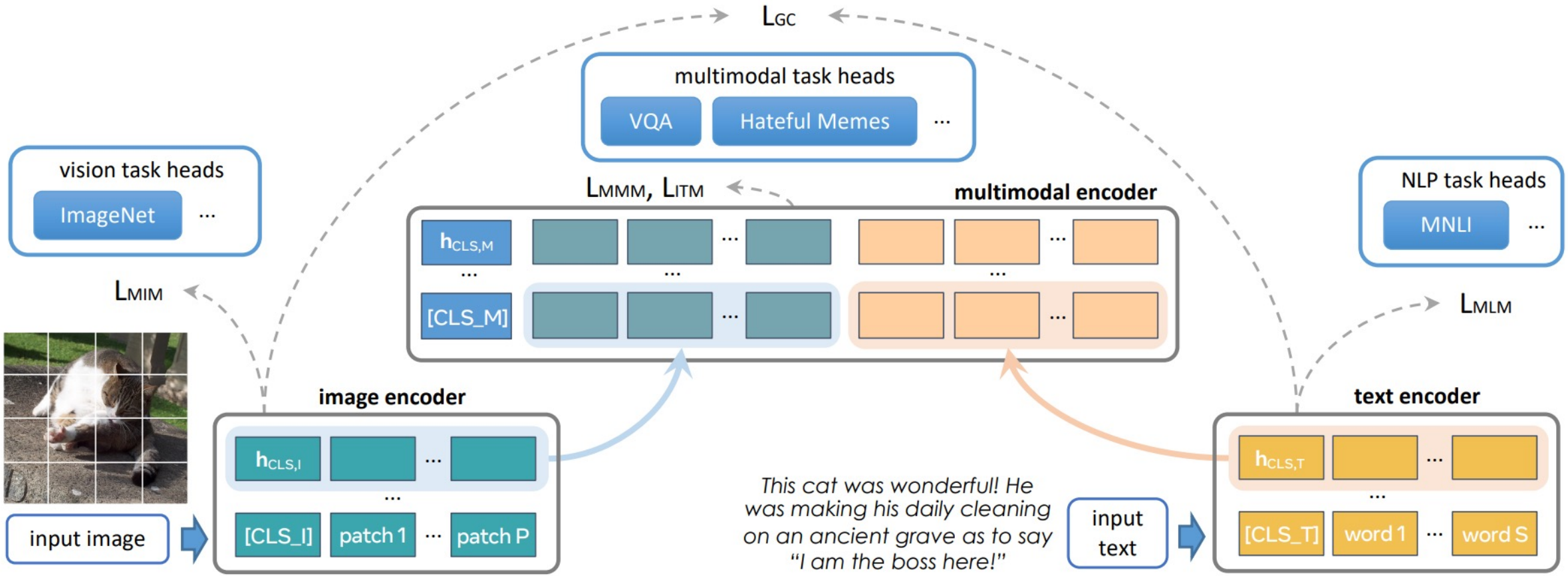
- BLIP: Bootstrapping Language-Image Pre-training



BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. CVPR 2022

Align + Generate

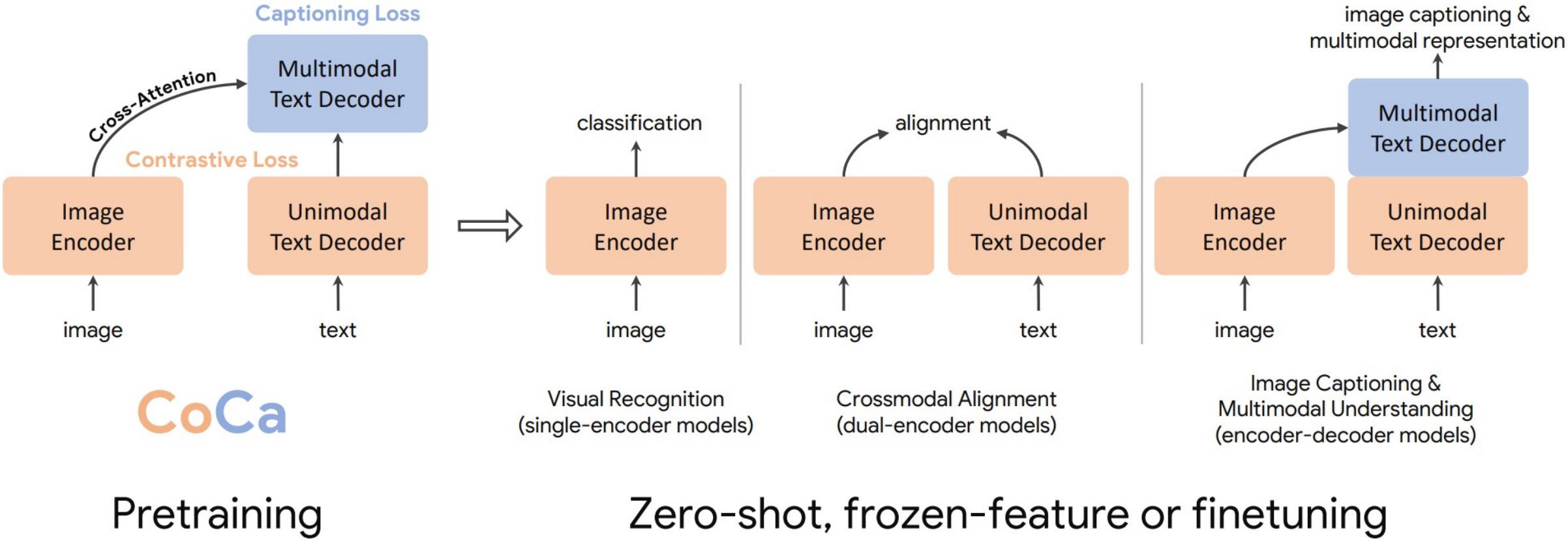
- FLAVA: Foundational LAnguage and Vision Alignment model



FLAVA: A Foundational Language And Vision Alignment Model. CVPR 2022

Align + Generate

- Coca: Contrastive Captioner



CoCa: Contrastive Captioners are Image-Text Foundation Models. Google. 2022

谢谢！