4. Transformers for Computer Vision

Overview of the techniques

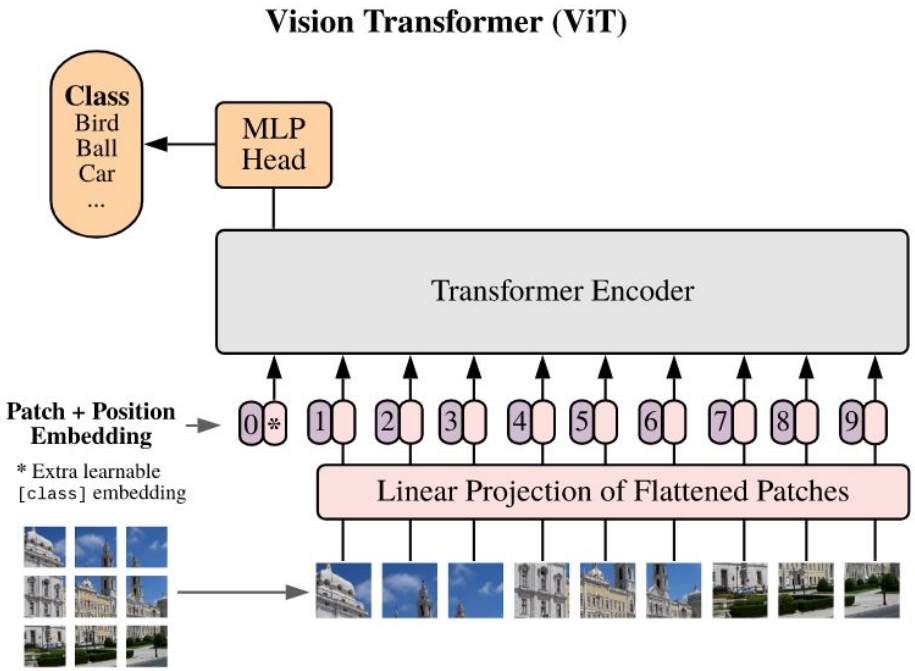
For all of them, we need to **tokenize images**.

- **Encoder-only**: ViT, Swin
- **Encoder-decoder**: DETR, DINO, SAM, TrOCR, DocParser, *DAN,* PaliGemma
- **Decoder-only**: Idefics2-3, ChatGPT-vision, DeepSeek-VL
- (marginal) cross-attention towers

(another way: disjoint networks with contrastive training: CLIP)

4.1. Encoder-only

Encodeurs pour les images



L'image est **découpée en fragments disjoints**, et on y ajoute une information de **position spatiale** en 2D.

Il devient possible de **capturer des dépendances éloignées** entre les indices visuels.

Exemple de l'architecture ViT [Dosovitskiy et al. 2020]

Swin: multi-scale vision transformer

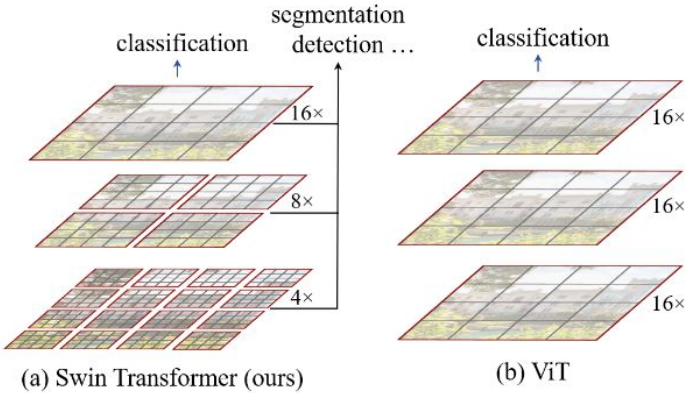


Figure 1. (a) The proposed Swin Transformer builds hierarchical feature maps by merging image patches (shown in gray) in deeper layers and has linear computation complexity to input image size due to computation of self-attention only within each local window (shown in red). It can thus serve as a general-purpose backbone for both image classification and dense recognition tasks. (b) In contrast, previous vision Transformers [19] produce feature maps of a single low resolution and have quadratic computation complexity to input image size due to computation of self-attention globally.

Better image description thanks to **multi-scale encoding.**

Reduce computational cost by **limiting the scope of self-attention.**

Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," ICCV 2021.

Uses for encoder-only vision transformers

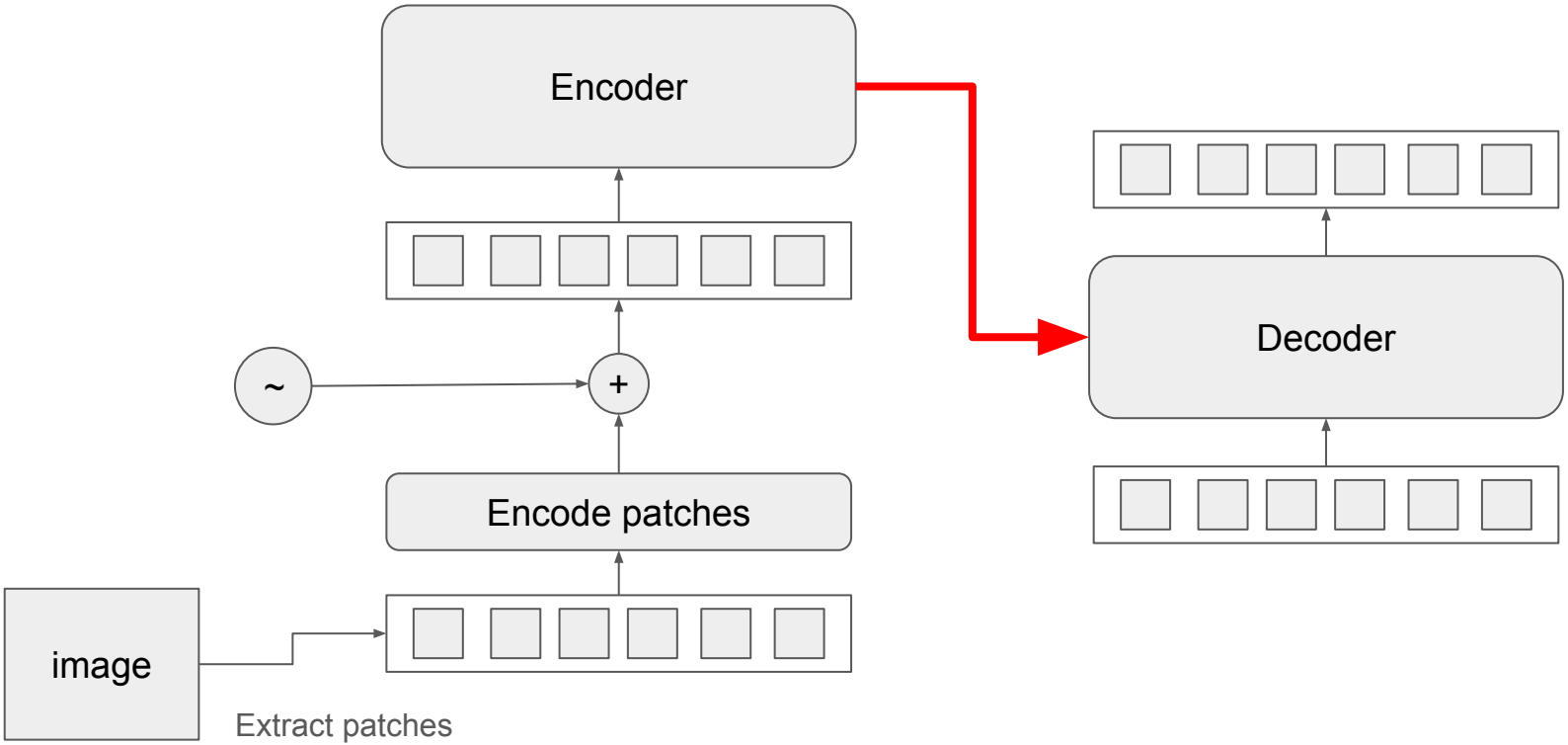
Embedding

As an encoder

Classification

4.2. Encoder-decoder

Cross-attention is the key here



4.2.a. Document analysis systems

TrOCR

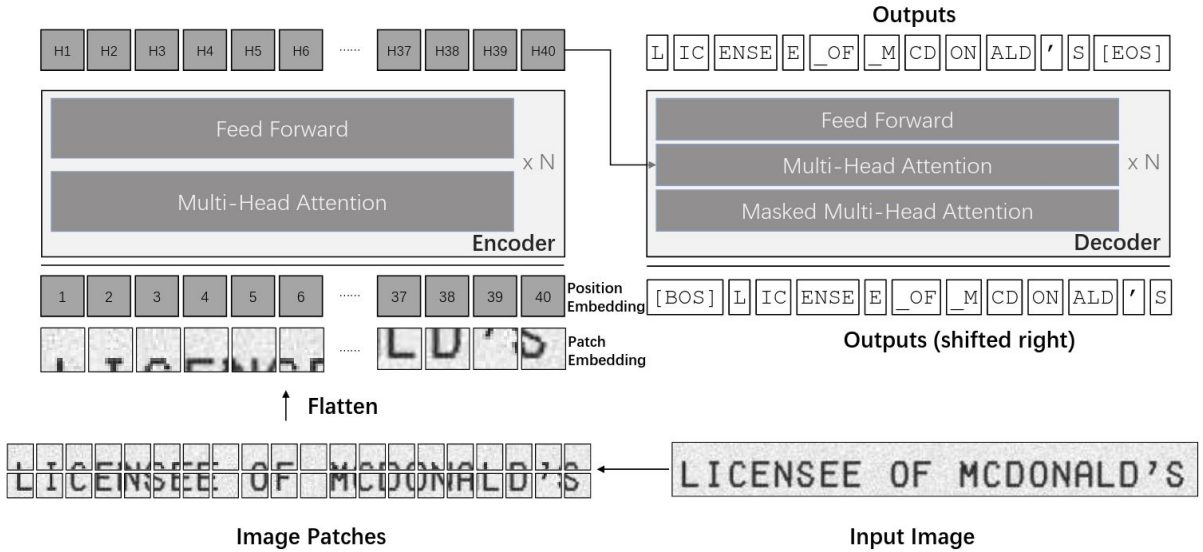


Figure 1: The architecture of TrOCR, where an encoder-decoder model is designed with a pre-trained image Transformer as the encoder and a pre-trained text Transformer as the decoder.

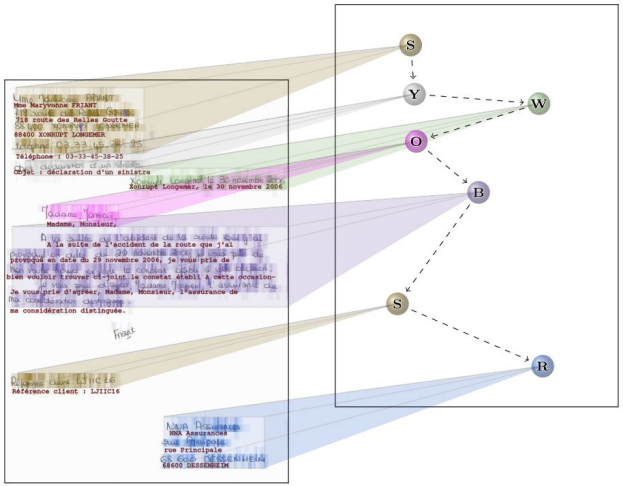
- ViT-like encoder.
- Uses BERT's pretrained weights to initialize decoder!
- Pretraining on millions of synthetic text lines.

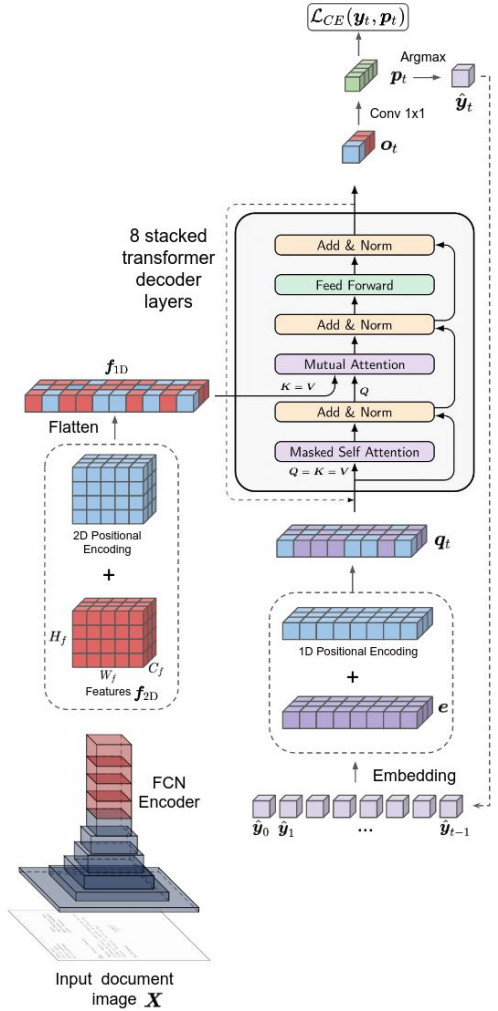
M. Li *et al. (2022)*, "TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models."

DAN: a CNN as encoder!

Full-page image to XML system.

Character-level predictions (improved since).





D. Coquenet, C. Chatelain, and T. Paquet (2022), "DAN: a Segmentation-free Document Attention Network for Handwritten Document Recognition."

4.2.b. Detectors

DETR: End-to-End Object Detection with Transformers

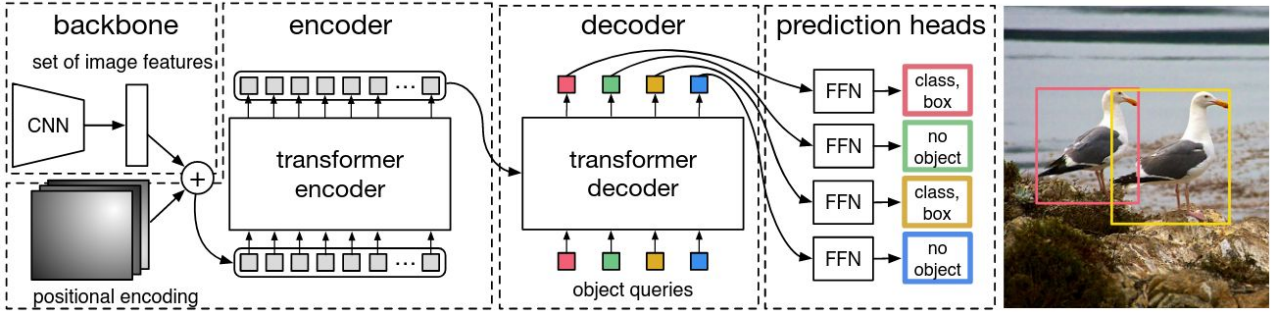


Fig. 2: DETR uses a conventional CNN backbone to learn a 2D representation of an input image. The model flattens it and supplements it with a positional encoding before passing it into a transformer encoder. A transformer decoder then takes as input a small fixed number of learned positional embeddings, which we call *object queries*, and additionally attends to the encoder output. We pass each output embedding of the decoder to a shared feed forward network (FFN) that predicts either a detection (class and bounding box) or a "`no object`" class.

⚠️ not an autoregressive decoder, but a "BERT-like" decoder.
⚠️ must choose the number of queries carefully.

N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," *ECCV 2020*.

Segment Anything: promptable detector

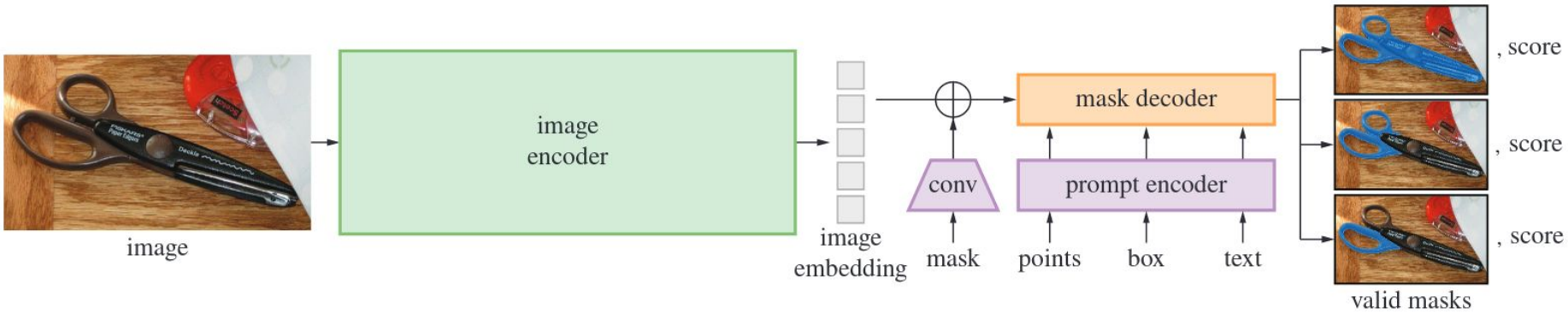


Figure 4: Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks at amortized real-time speed. For ambiguous prompts corresponding to more than one object, SAM can output multiple valid masks and associated confidence scores.

👍 Many super-cool use-cases with text, point or region prompting!

A. Kirillov *et al.*, “Segment Anything,” Apr. 05, 2023, *arXiv*: arXiv:2304.02643. doi: 10.48550/arXiv.2304.02643.

4.3. Decoder-only

General-Purpose Interfaces

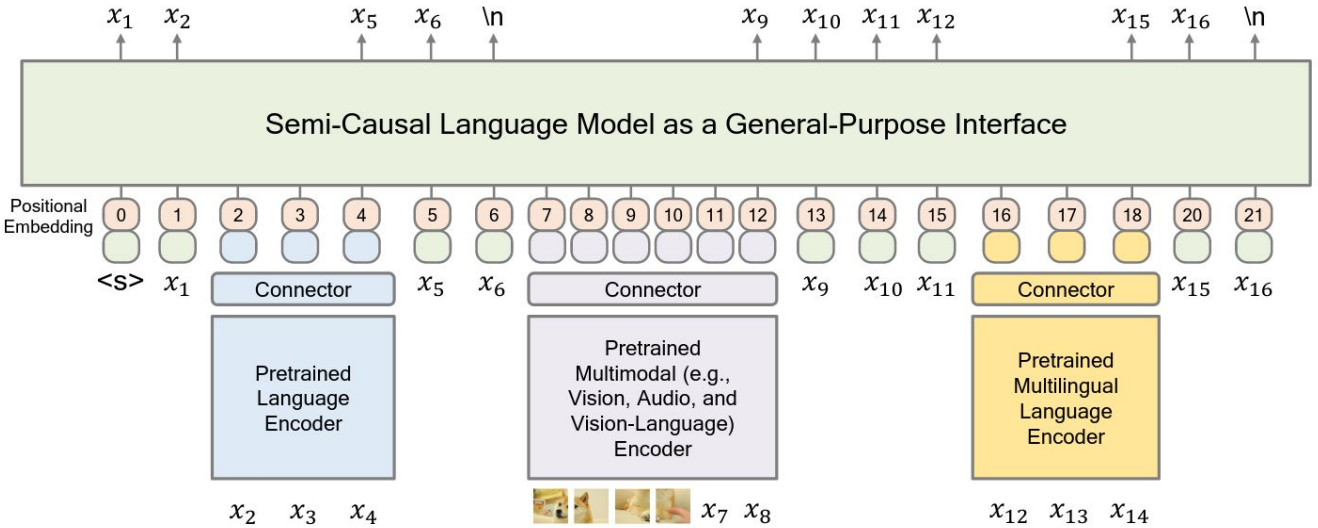


Figure 2: Overview of METALM. The semi-causal language model serves as a general-purpose interface and supports interactions with various foundation models.

Y. Hao *et al.*, "Language Models are General-Purpose Interfaces," Jun. 13, 2022, 10.48550/arXiv.2206.06336 ← Microsoft Research Asia

Deepseek strategy: how to train adapters

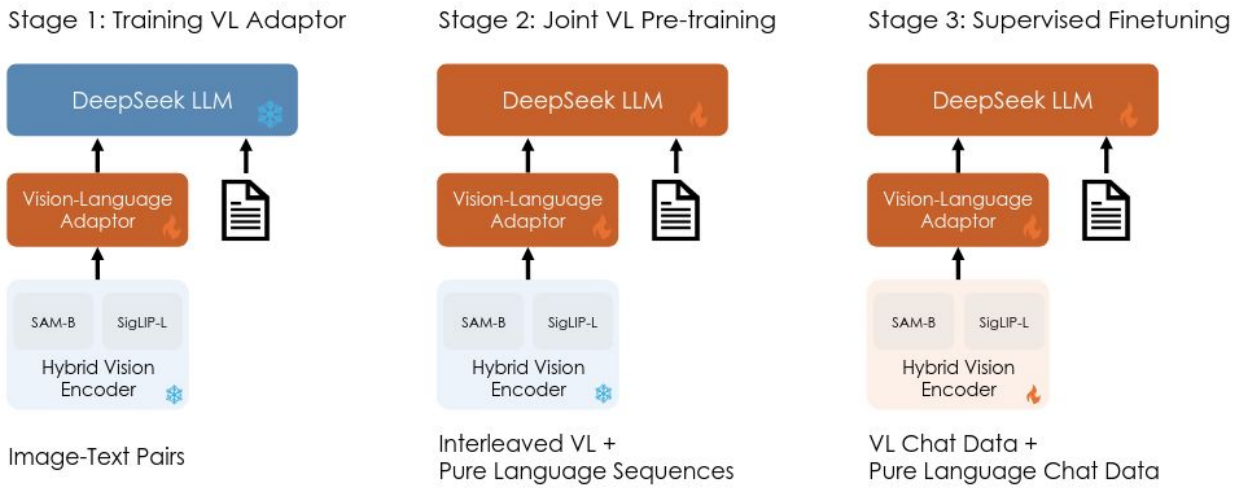
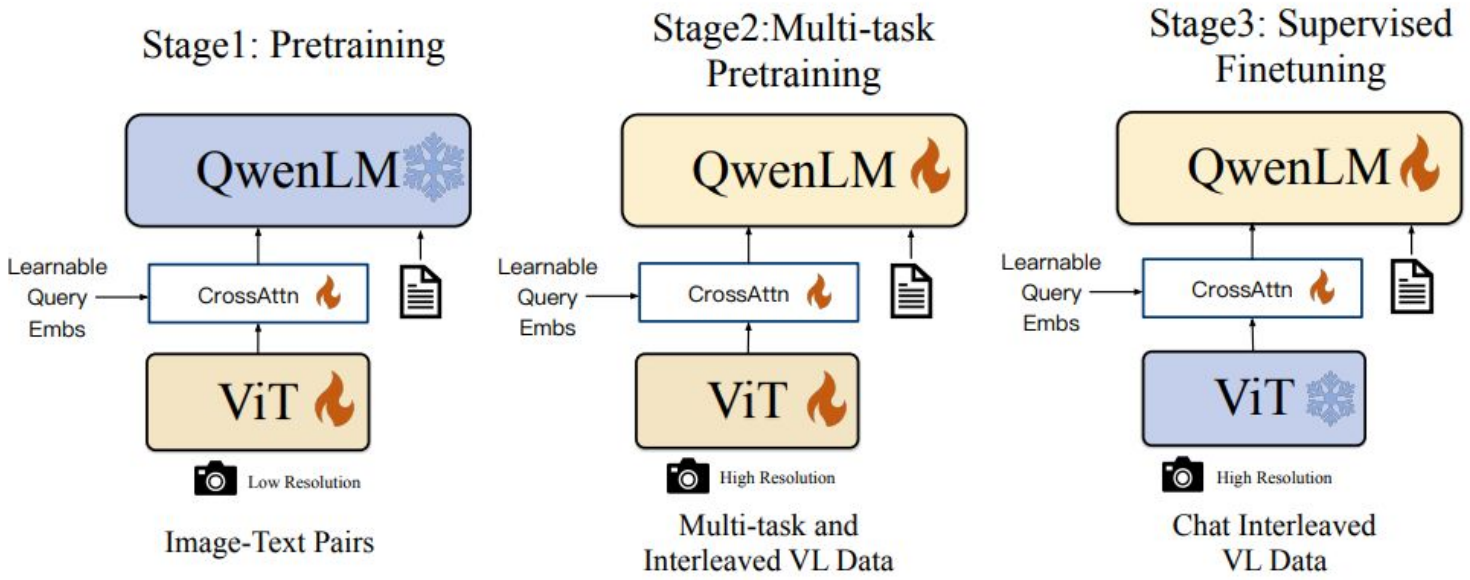


Figure 3 | Our training pipelines consist of three stages. Stage 1 involves training the Vision-Language (VL) adaptor while keeping the hybrid vision encoder and language model fixed. Stage 2 is the crucial part of the joint vision and language pretraining, where both VL adaptor and language model are trainable. Stage 3 is the supervised fine-tuning phase, during which the low-resolution vision encoder SigLIP-L, VL adaptor, and language model will be trained.

H. Lu *et al. (2024)*, "DeepSeek-VL: Towards Real-World Vision-Language Understanding."

Same for Qwen-VL



Y. Hao *et al.*, “Language Models are General-Purpose Interfaces,” Jun. 13, 2022, 10.48550/arXiv.2206.06336 ← Microsoft Research Asia

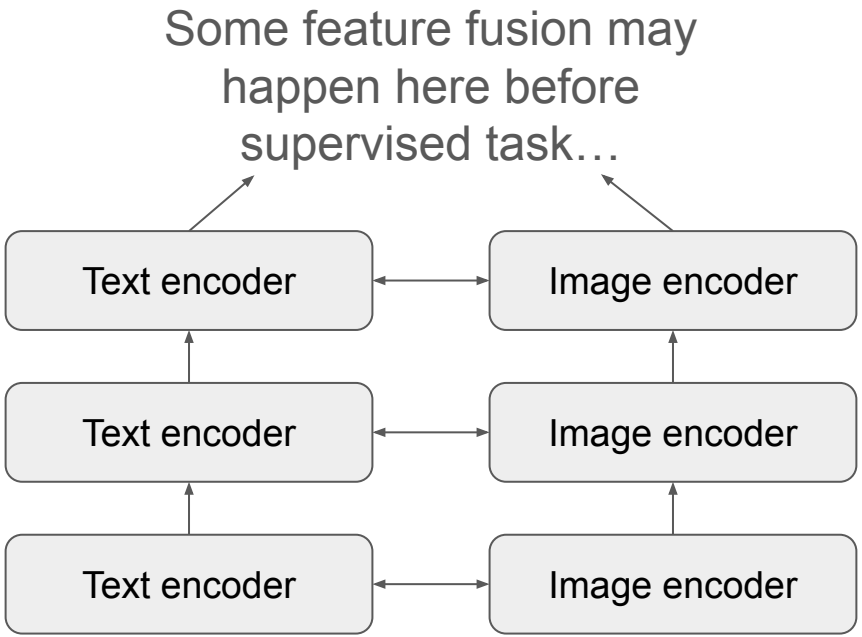
Segmentation capabilities: Idefics2 example

(still looking for the illustration…)

Predict special **<class>** and **<loc> tokens** to point places in the image!

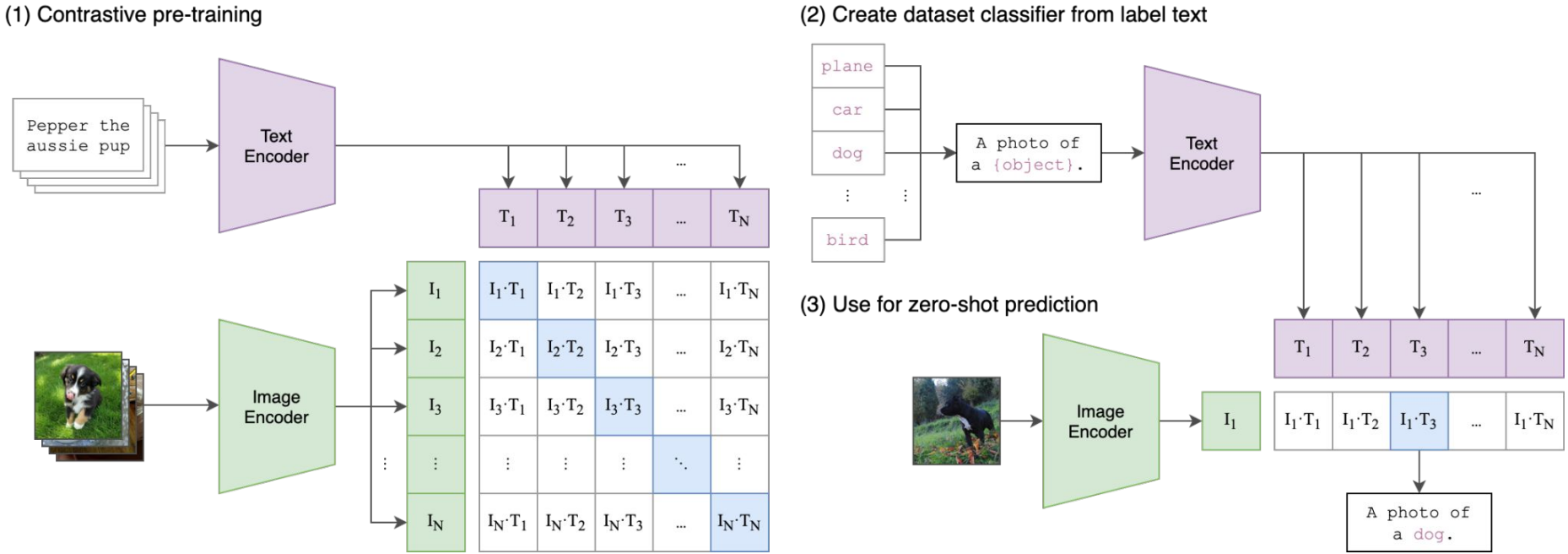
4.4. Cross-attention towers

Complicated beasts… Very specific uses.



4.5. Other

CLIP: powerful image semantic embedding (see SigLIP too)



A. Radford *et al. (2021)*, "Learning Transferable Visual Models From Natural Language Supervision."