

VLM ARCHITECTURE



1

CONTRASTIVE LEARNING

Aligns image-text pairs

2

PREFIX LANGUAGE MODELING (PREFIXLM)

Uses images as text prefixes

3

FROZEN PREFIXLM

Freezes language model parameters

4

MULTIMODAL FUSION WITH CROSS-ATTENTION

Integrates modalities via cross-attention

5

MASKED LANGUAGE MODELING & IMAGE-TEXT MATCHING

Predicts masked text & matches pairs

6

NO-TRAINING

Leverages pre-trained embeddings directly

7

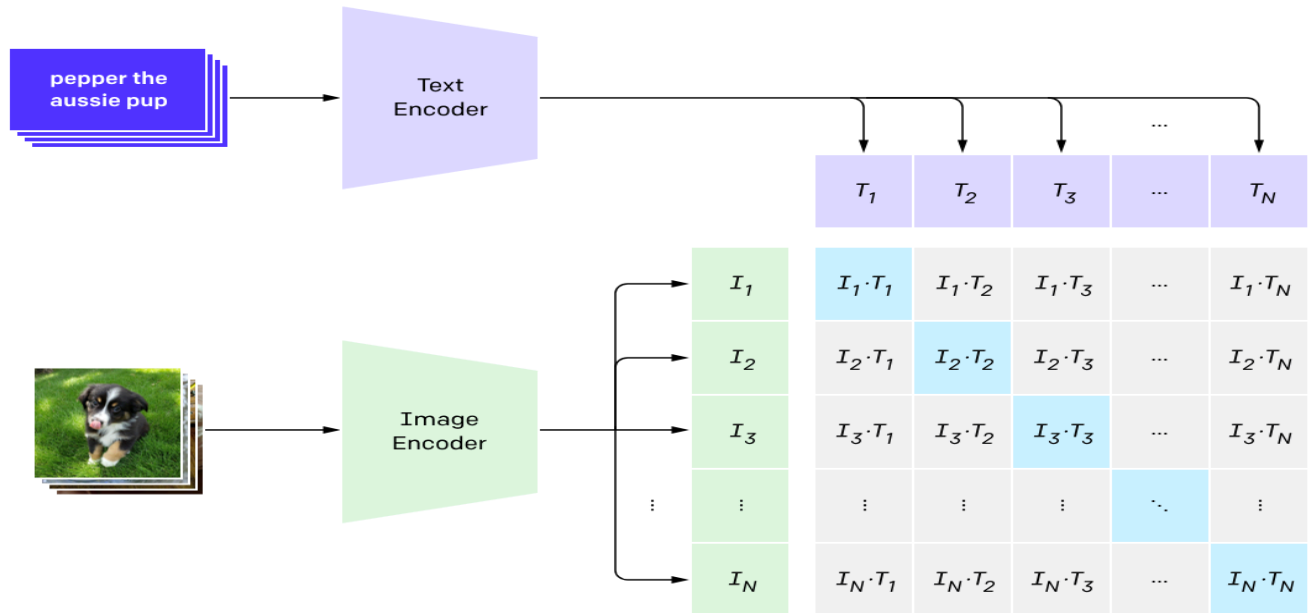
KNOWLEDGE DISTILLATION

Transfers knowledge to smaller models

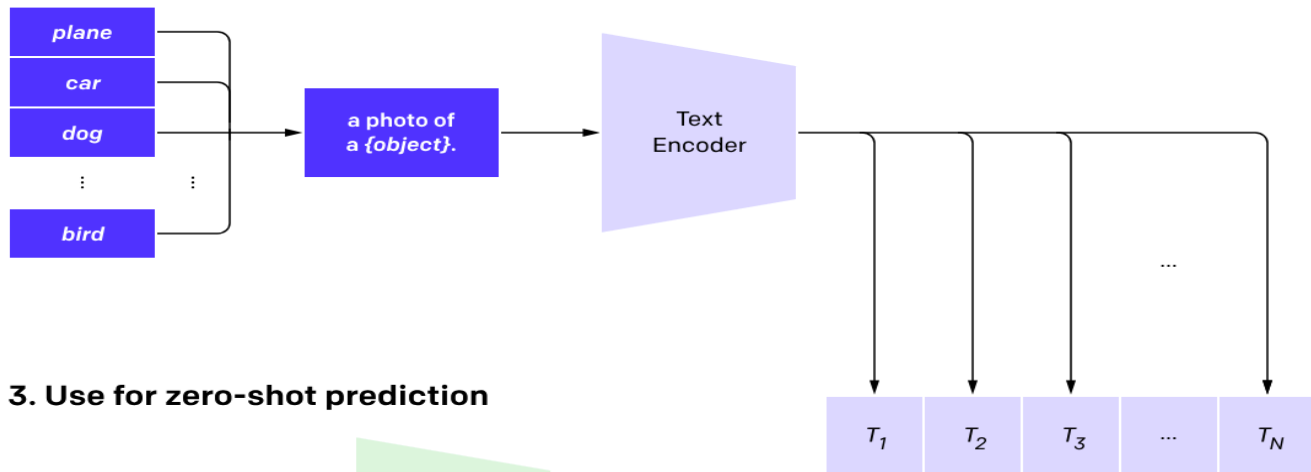
Contrastive Learning

CLIP Architecture

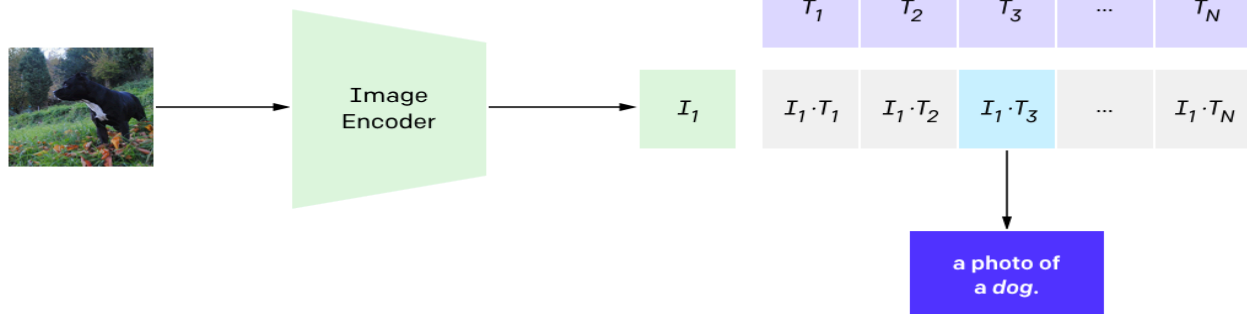
1. Contrastive pre-training



2. Create dataset classifier from label text

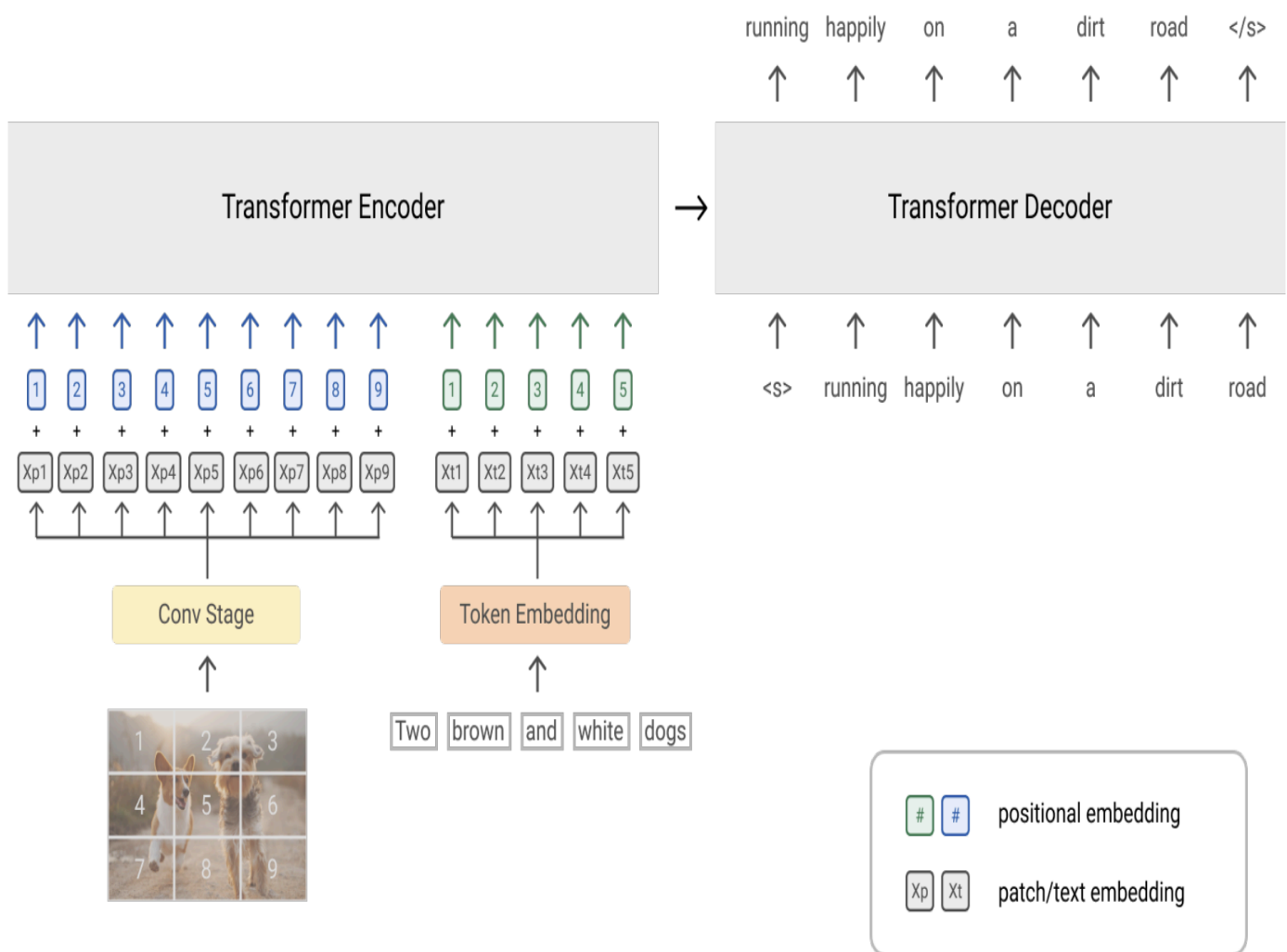


3. Use for zero-shot prediction



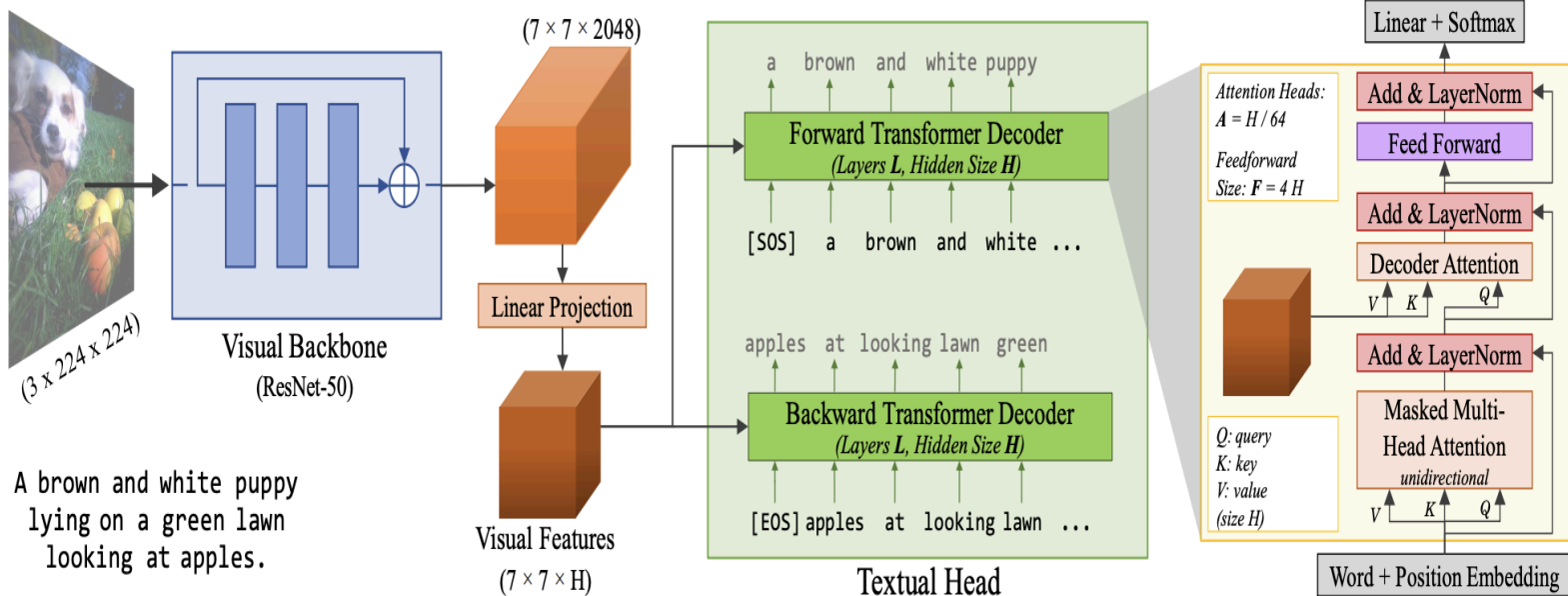
PrefixLM

SimVLM Architecture



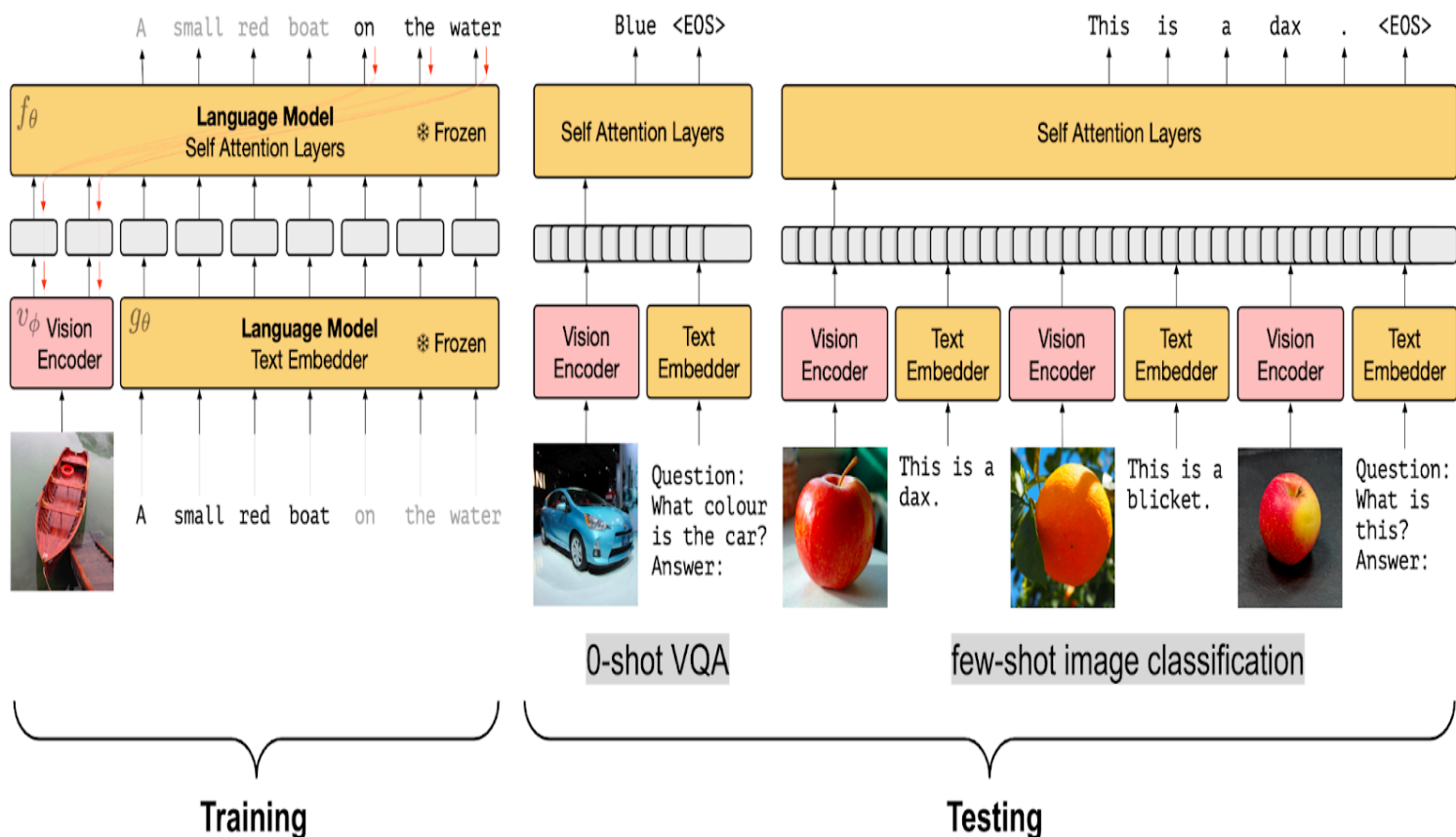
PrefixLM

VirTex Architecture



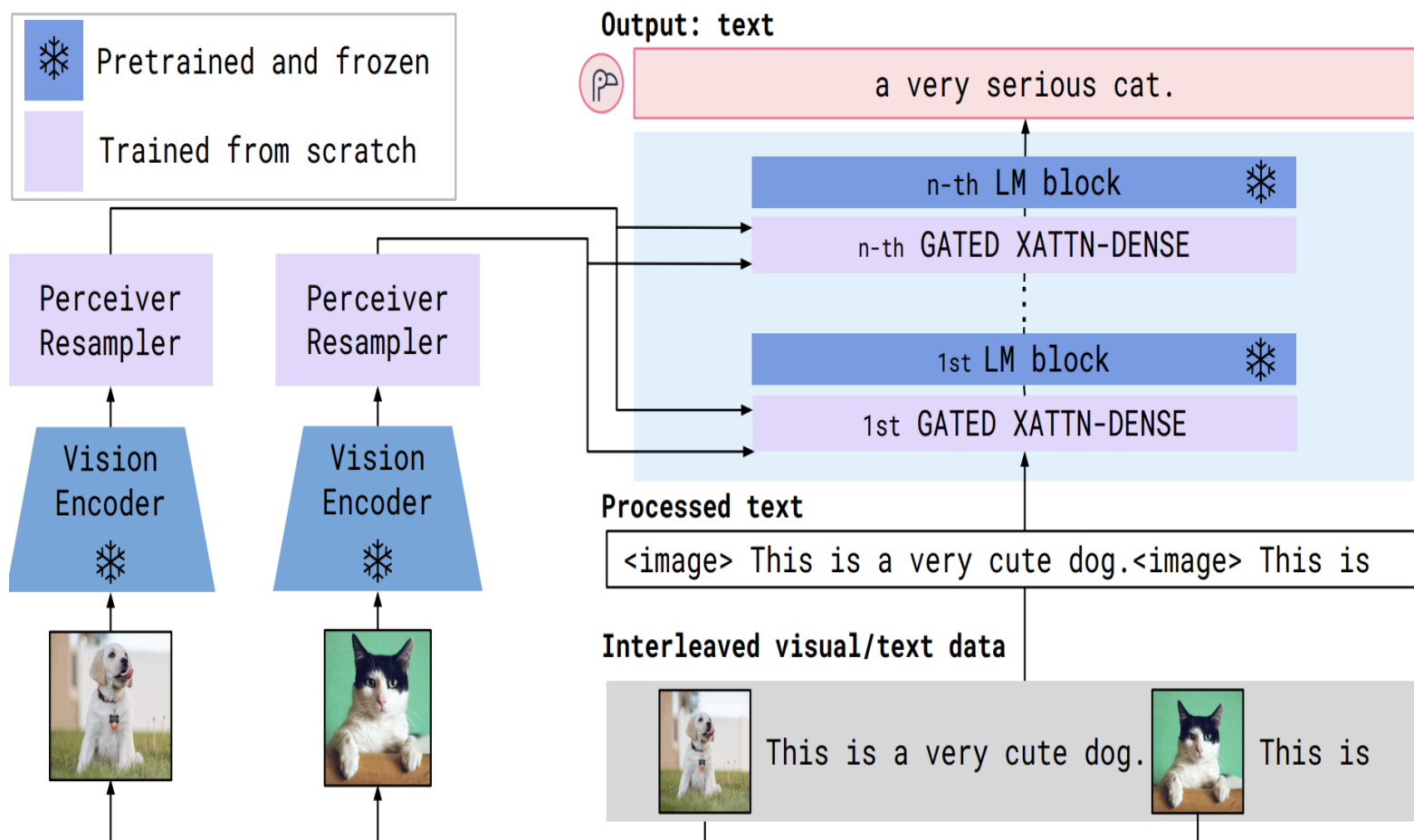
Frozen PrefixLM

Frozen Architecture



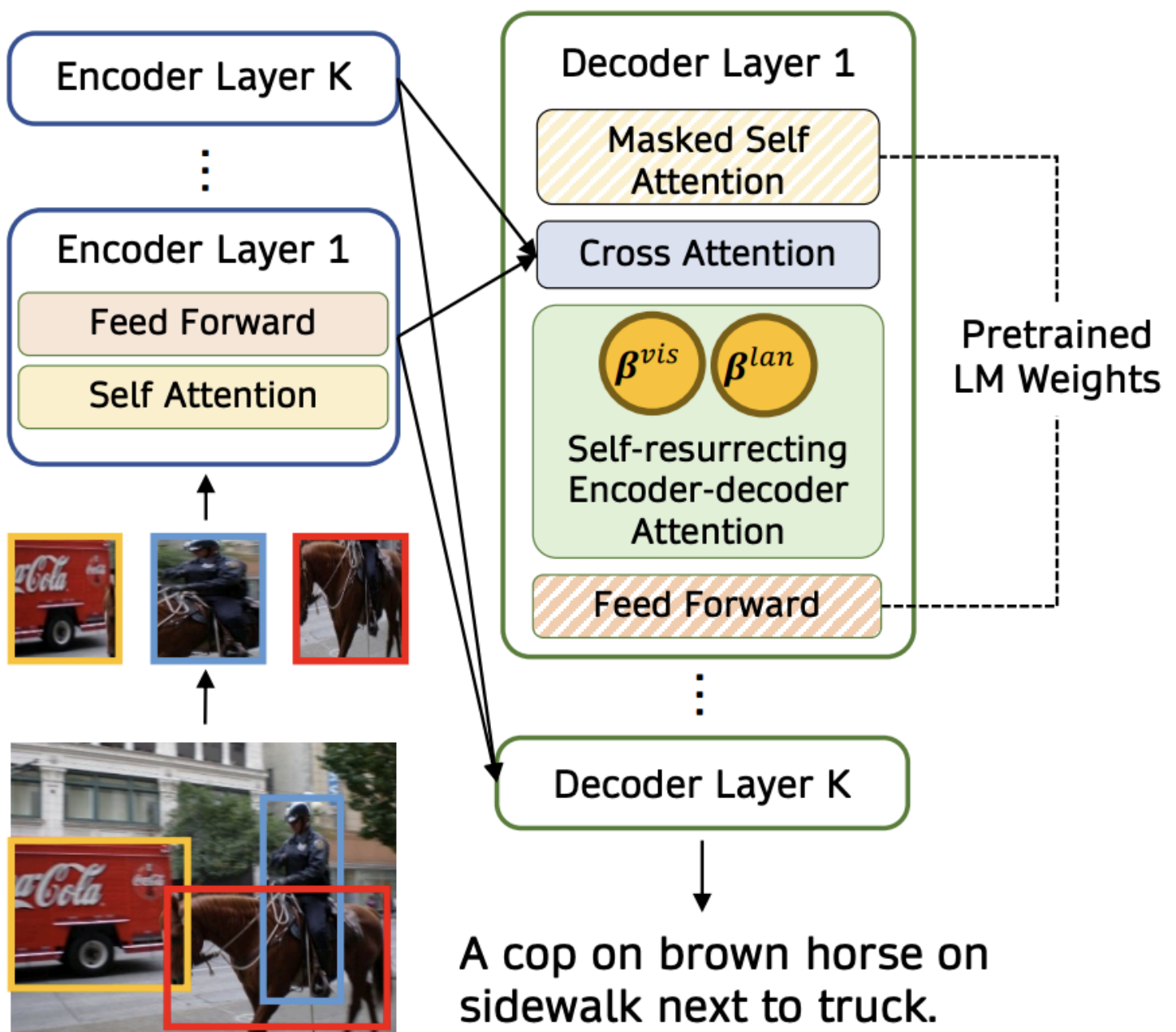
Frozen PrefixLM

Flamingo Architecture



Multimodal Fusing with Cross-Attention

VisualGPT Architecture

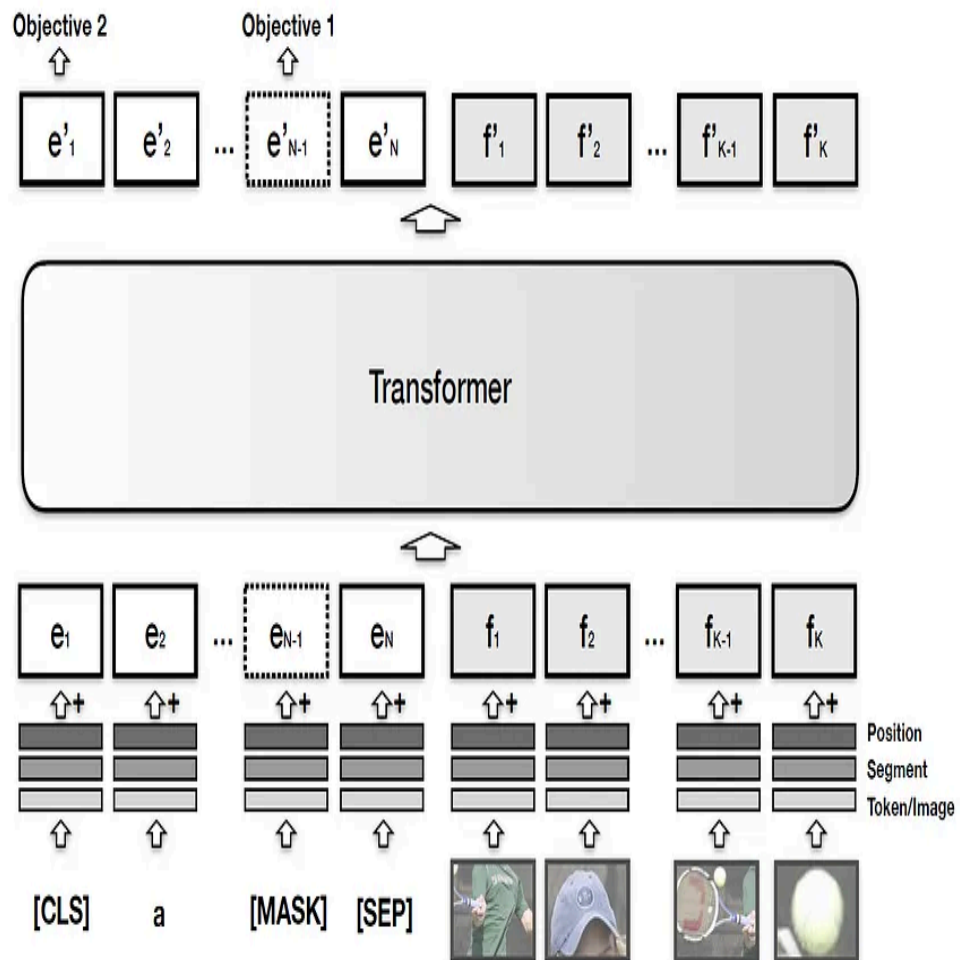


Masked-language Modeling (MLM) and Image-Text Matching (ITM)

VisualBERT Architecture

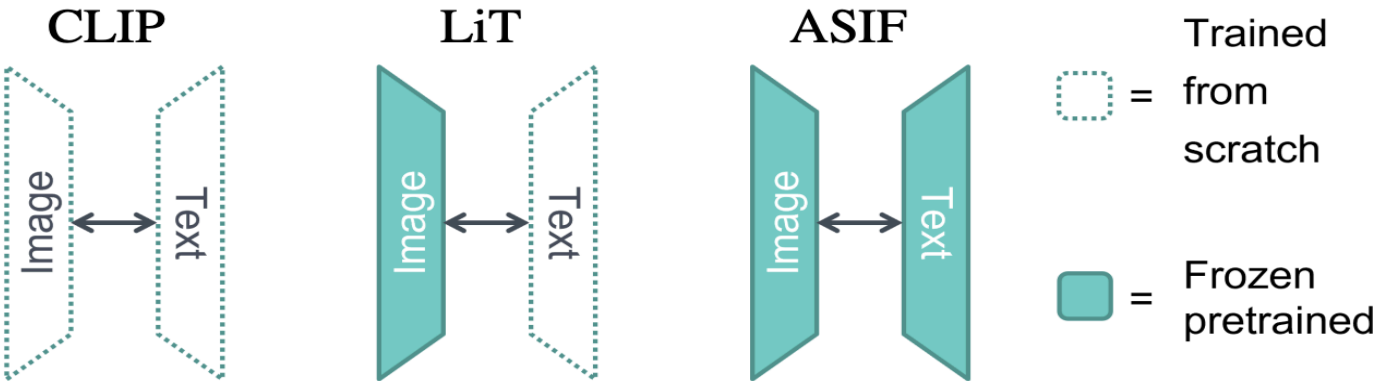


A person hits a ball with a tennis racket



No Training

ASIF Prediction Strategy

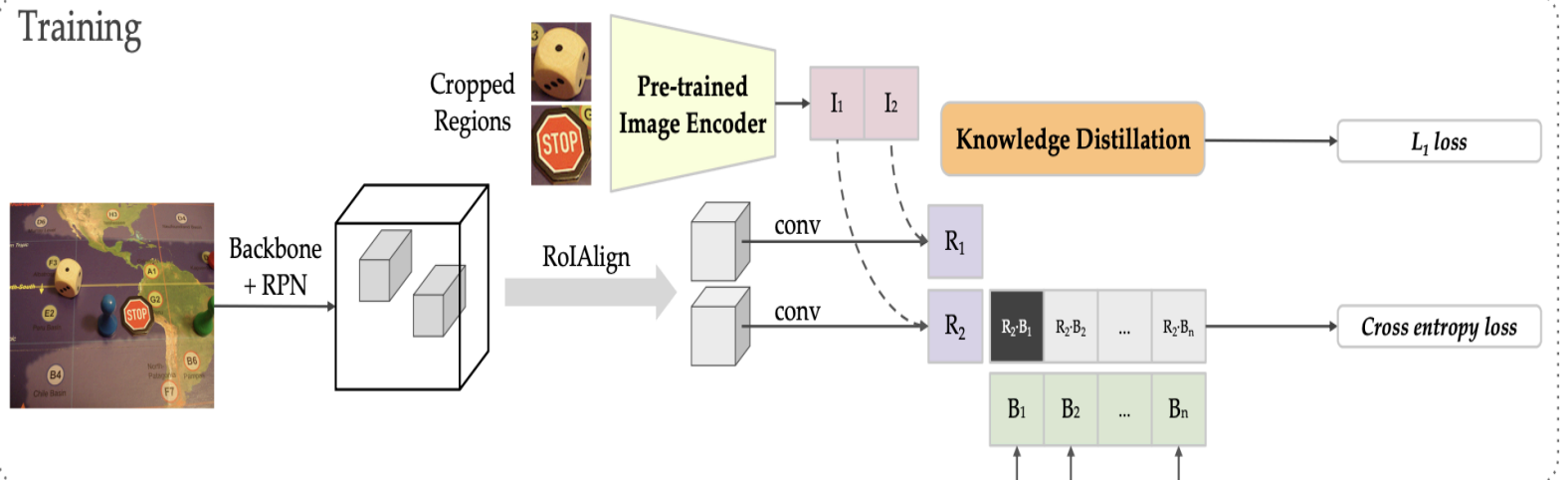


Most similar images in the multimodal dataset					Query image x^*														

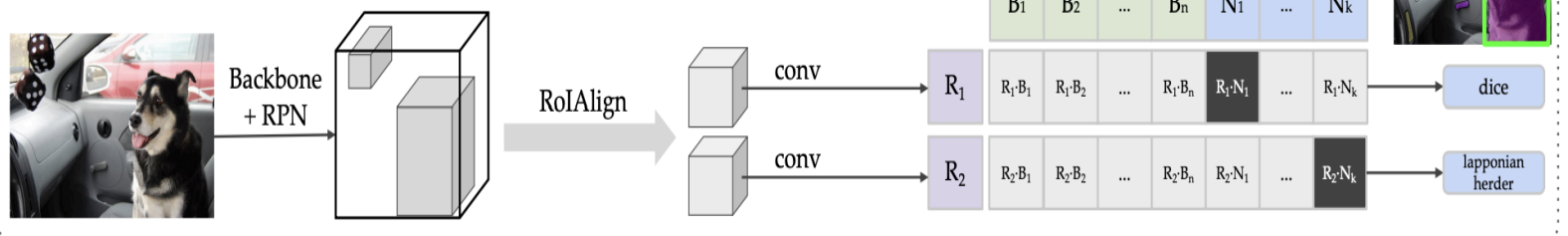
Knowledge Distillation

ViLD Architecture

Training



Inference





Download Complete Guide 

<https://github.com/Abonia1/VLM-Architecture>