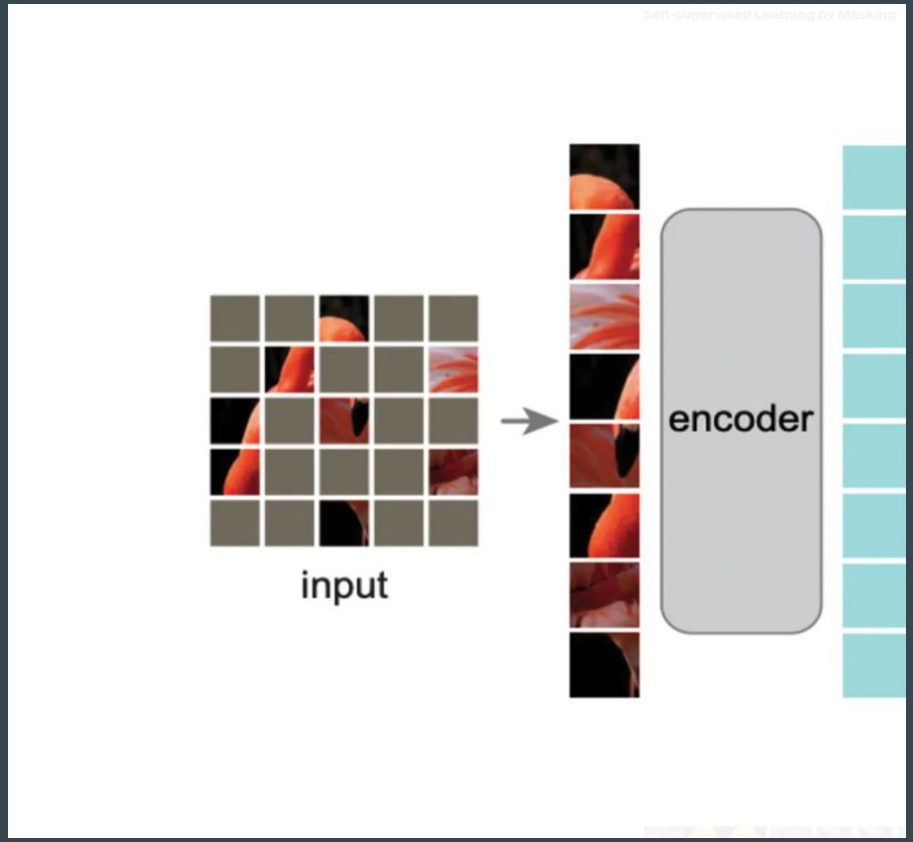
From Diagrams to Dialogue Understanding Miro Boards with VLMs

What is the Million Dollar Idea $:

The Rough Idea is to Extract Information Visual Board
Either it is Miro , Mural or Lucid etc

Vision Encoder:



(High Level VIT)CLIP ,SigLIP etc

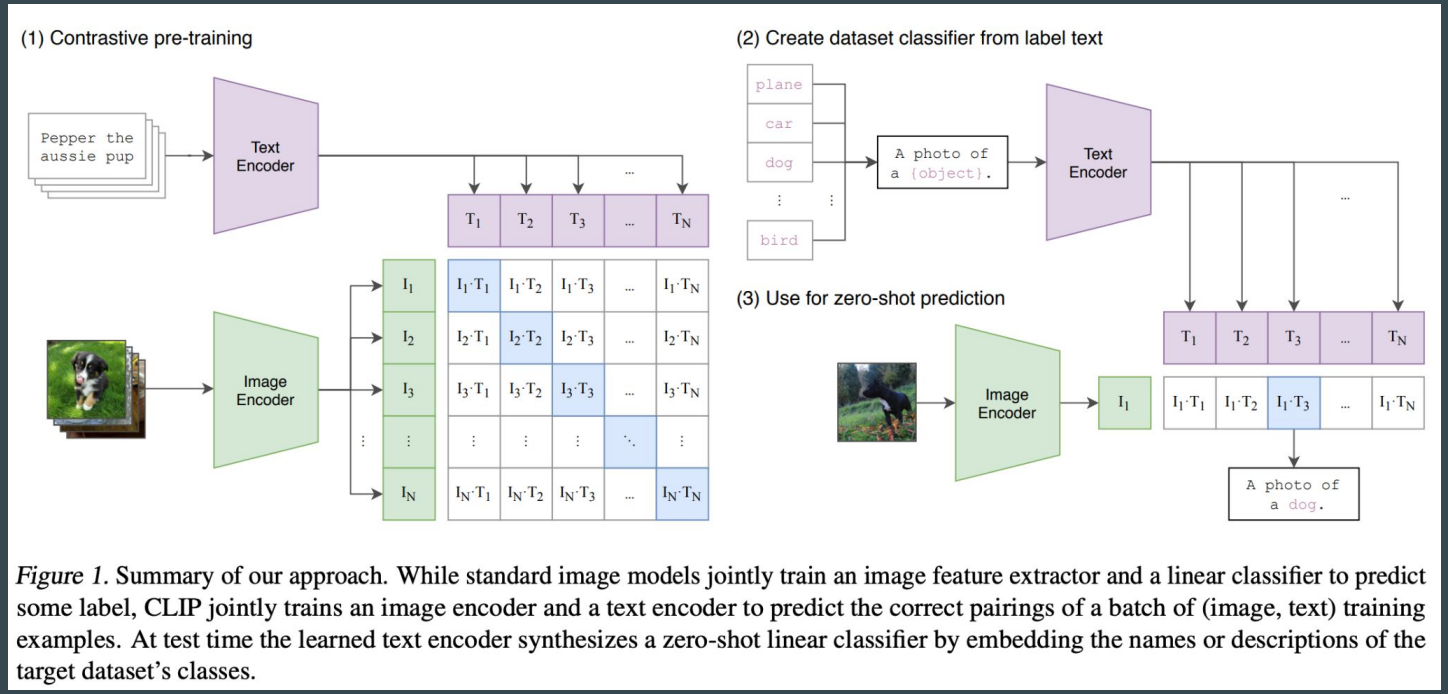
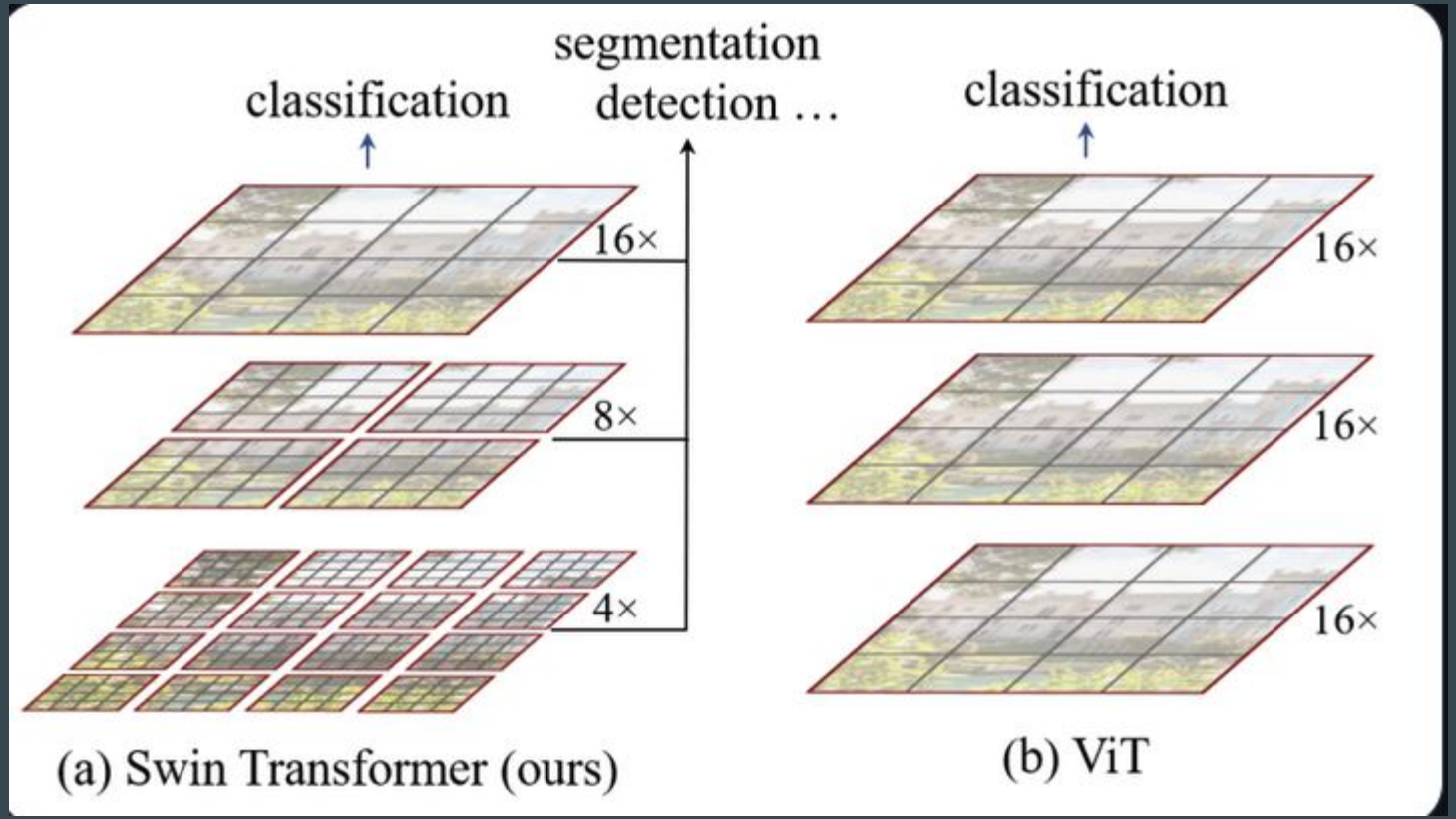


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

Swim Transformer:



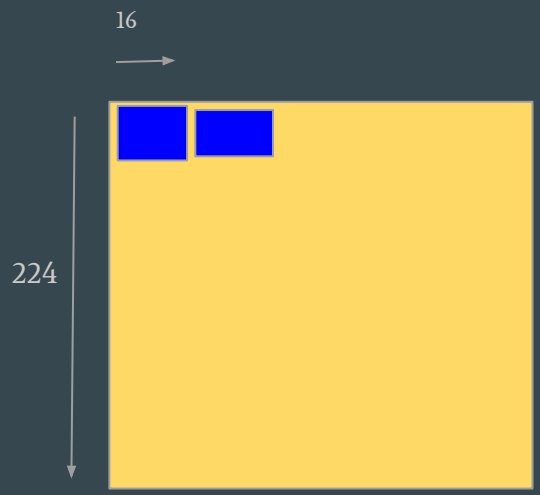
Simple Math Behind It

VIT (Vision Transformation)
Patches per row = 224 ÷ 16 = 14
Patches per column = 224 ÷ 16 = 14

Total patches = 14 × 14 = 196 patches

Each patch = 16 × 16 × 3 = 768 dimensions

All patches = 196 patches × 768 dimensions
Attention Matrix = 196 × 196 = 38,416 computations
Every patch attends to ALL other patches
Complexity = $O(N^2)$ where N = 196



Simple Math Behind It

Initial patches = 4×4 pixels (smaller than ViT)
Patches per row = 224 ÷ 4 = 56
Patches per column = 224 ÷ 4 = 56
Total patches = 56 × 56 = 3,136 patches

Windowing
Window size = 7×7 patches
Patches per window = 7 × 7 = 49 patches
Total windows = 3,136 ÷ 49 = 64 windows

Attention per window = 49 × 49 = 2,401 computations
Total attention = 64 windows × 2,401 = 153,664 computations

Simple Math Behind It

Swin Transform Operation > VIT Operation
153,664 38,416

But Swin Transform Operation = Complexity = O(N) - Linear!

For 448×448 image (2x larger):
ViT:
- Patches = 28 × 28 = 784
- Attention = 784^2 = 614,656 operations

Swin:
- Windows still 7×7 = 49 patches per window
- More windows, but attention per window stays constant

Baby Step:

Figuring out the Layout in the Board

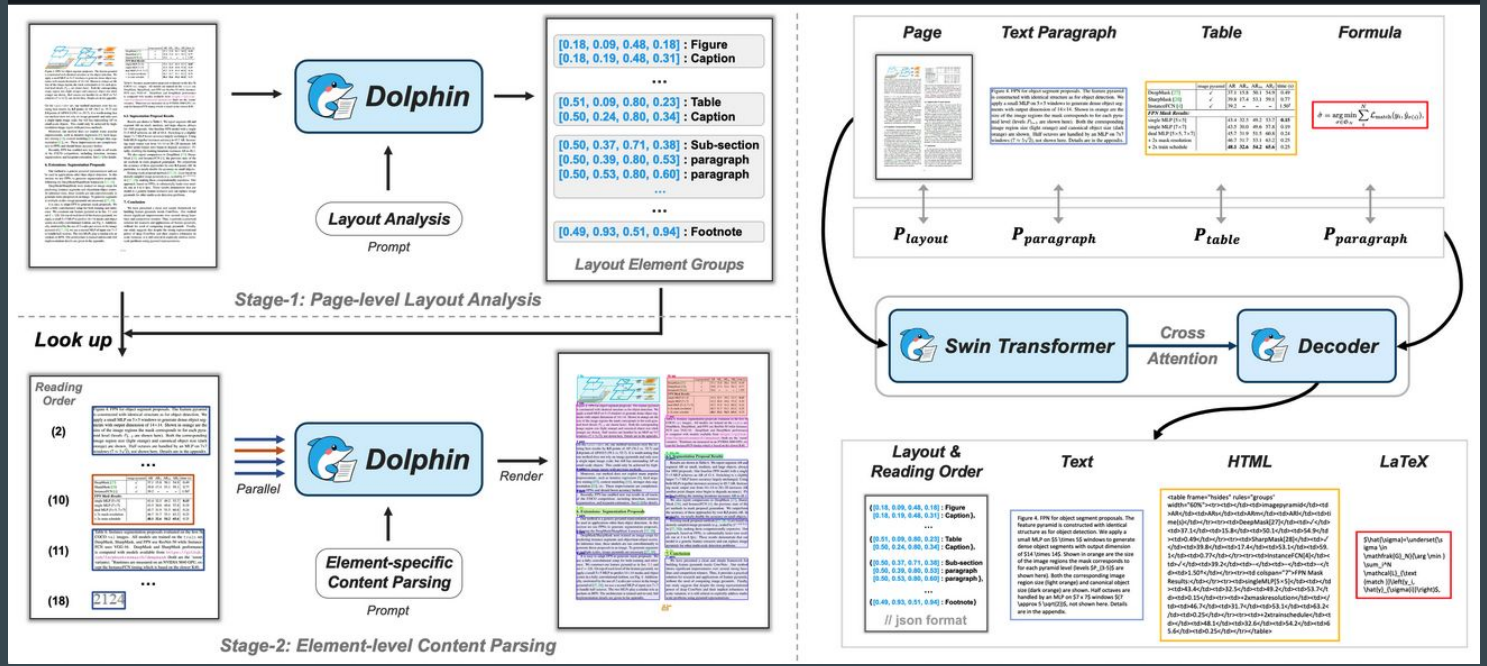
ScreenAI:
https://research.google/blog/screenai-a-visual-language-model-for-ui-and-visually-situated-language-understanding/

Simple Idea:

First Figure out the Layout and different Components in the Board

Pass down the Split images to Multi Model LLM (Preferable InternVL my personal Favorite for Diagram Analysis)

Dolphin Model (Document Image Parsing via Heterogeneous Anchor Prompting)



Repo

Profile