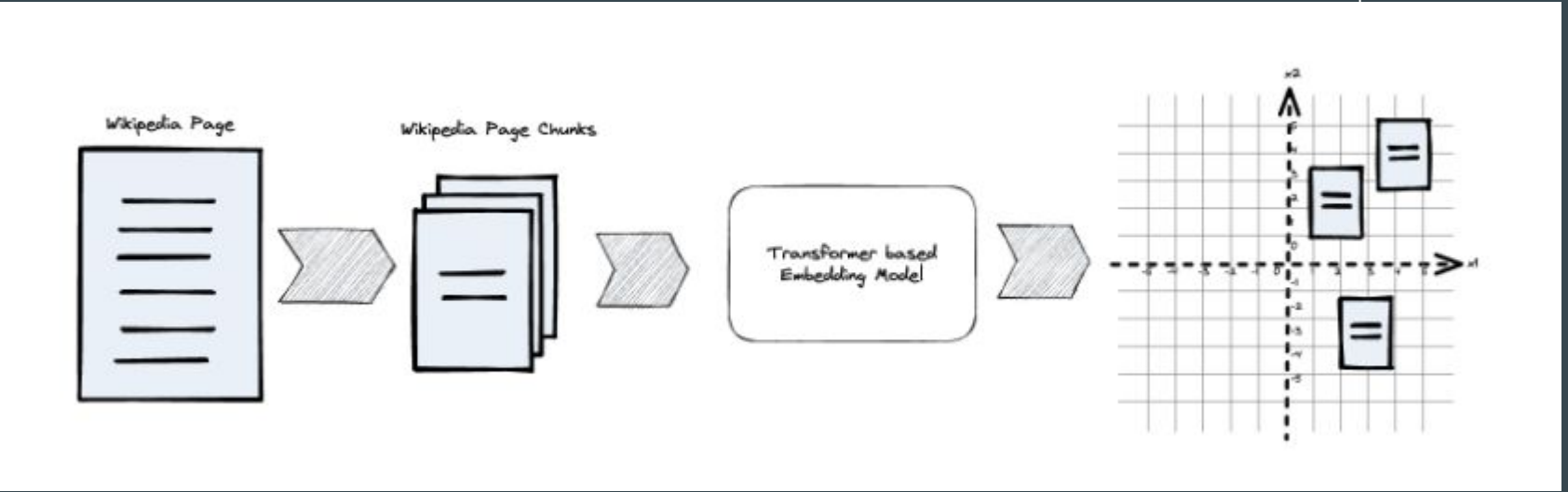
# OCR Multimodal Retrieval using ColiPali Model

# Traditional Retrival Document Method



# ColiPali = Colbert + PaliGemma

# Vision Models
# CLIP ,SigLIP etc

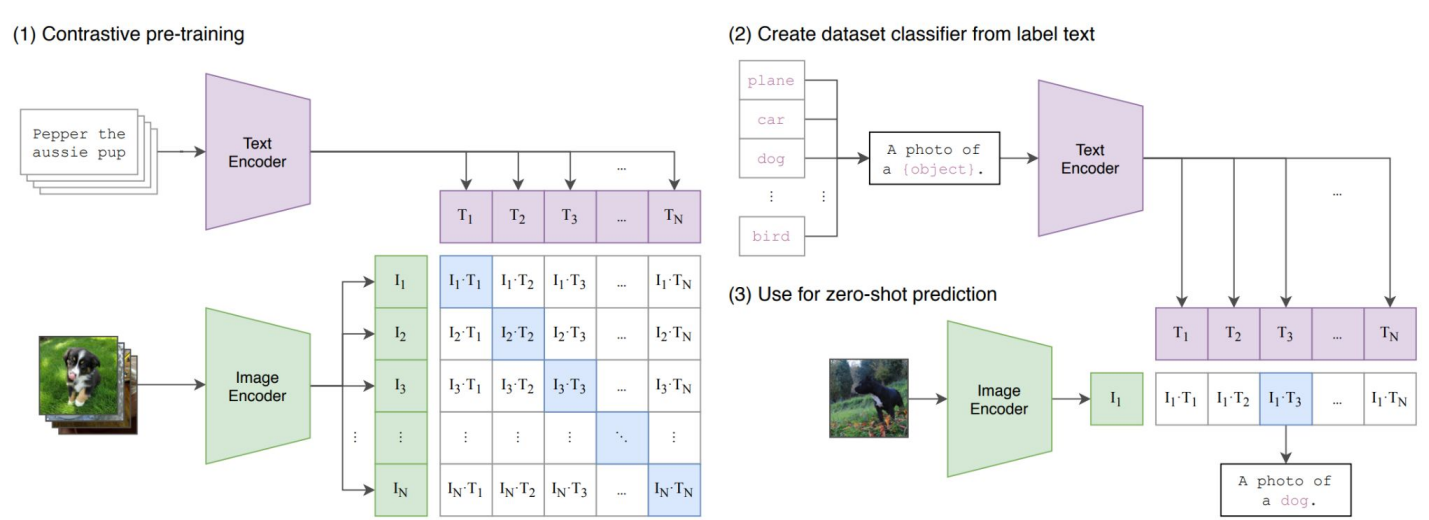
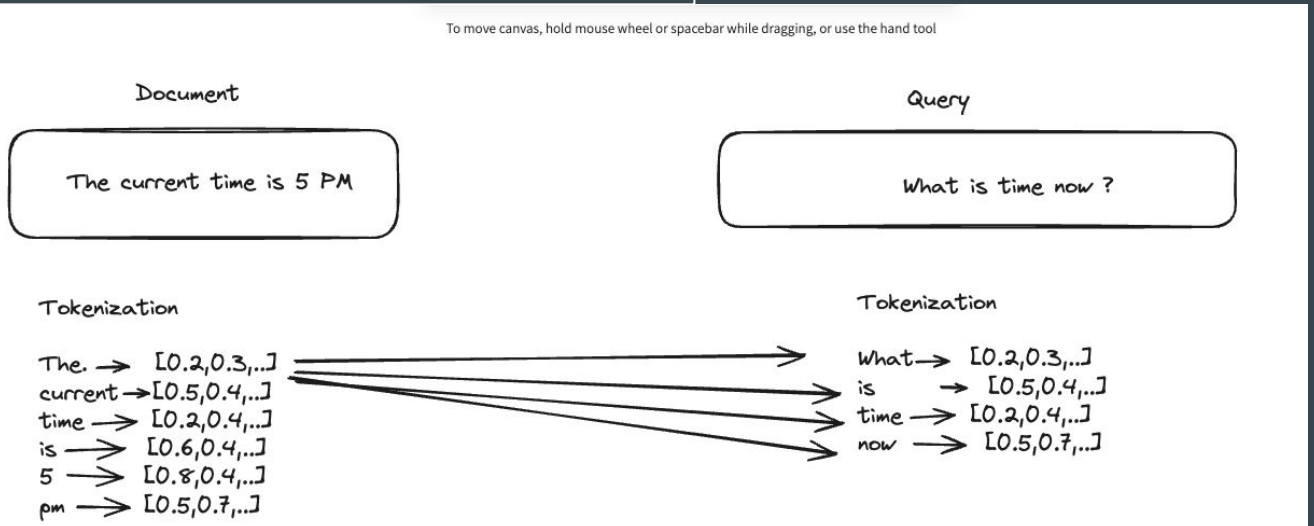



*Figure 1.* Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time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

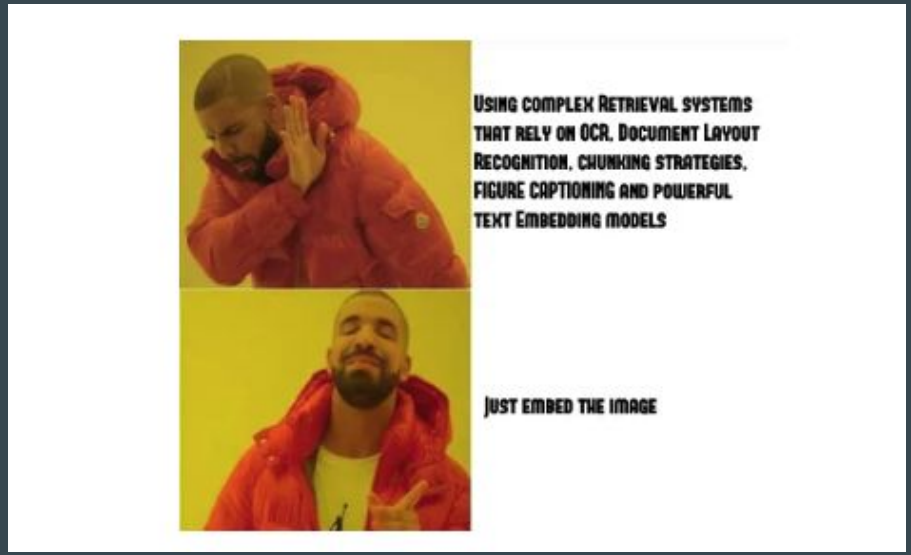
# ColBert

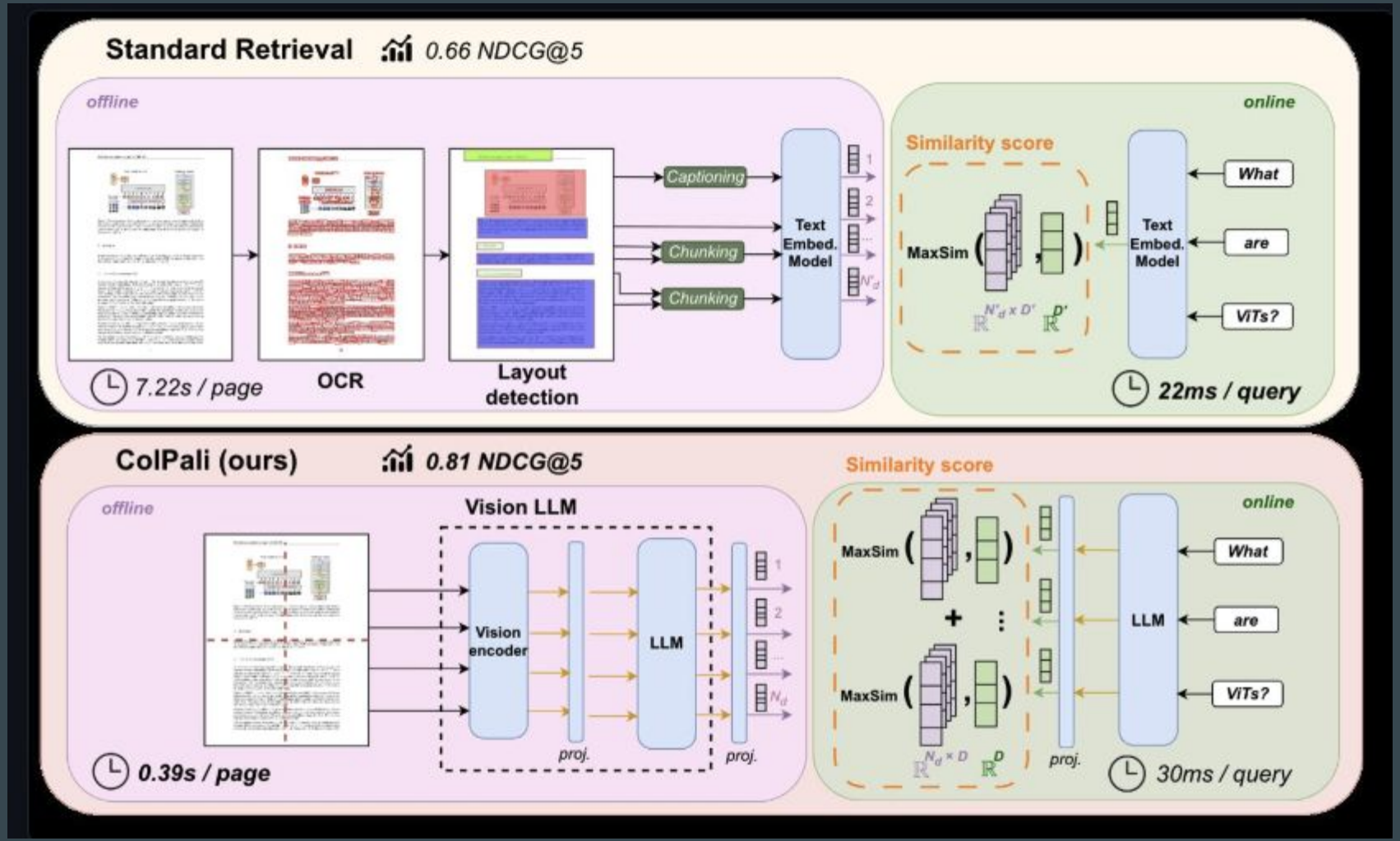



Now document token and Query Token is compared and then similarity score is calculated.
So, then every token contributes to the Similarity Score

This is known as Contextual Late Interaction

# Colipali


Using complex Retrieval systems that rely on OCR, Document Layout Recognition, chunking strategies, figure captioning and powerful text Embedding models
Just embed the image


Standard Retrieval
0.66 NDCG@5
offline
Captioning
Chunking
Chunking
Text Embed. Model
7.22s / page
OCR
Layout detection
online
Similarity score
MaxSim
Text Embed. Model
What
are
ViTs?
22ms / query
ColPali (ours)
0.81 NDCG@5
offline
Vision LLM
Vision encoder
LLM
proj.
proj.
0.39s / page
Similarity score
MaxSim
MaxSim
online
What
are
ViTs?
LLM
proj.
30ms / query