☑ What is CLIP?

CLIP stands for **Contrastive Language–Image Pretraining**. It was introduced by **OpenAI** in 2021.

🧠 Core Idea:

CLIP learns to **match images and text** by training on **millions of (image, caption)** pairs. It learns a **shared space** where related image-text pairs are **close**, and unrelated ones are **far**.

How it works:

- An **image encoder** (e.g., ViT or ResNet).
- A **text encoder** (e.g., Transformer or BERT).
- It learns by **contrastive loss**:
  - Pushes **matching** pairs together.
  - Pushes **non-matching** apart.

☑ Why Use CLIP?

CLIP is **versatile** and powerful because:

Reason	Benefit
🔍 Zero/Few-shot Learning	No need for large labeled datasets
🧠 Multimodal Understanding	Understands **text and image** together
⚡ Transferable	Works well even on **unseen tasks**
📉 Low annotation cost	No need to label 1000s of medical images

In medical AI, this is huge because labeling X-rays is **expensive** and **requires radiologists**.

☑ Types / Variants of CLIP

Variant	Description	Use-case
Original CLIP	Trained on internet images + captions	Generic image-text matching
MedCLIP	Trained on medical image–text (MIMIC-CXR, etc.)	Chest X-ray reports, radiology
BioViL	Biomedical Vision-Language	Medical images + text reports
GLoRIA	Aligns global + local regions	Better for localized findings
CheXzero	Zero-shot classification of CXR	Best for **zero-shot** disease detection
MedCLIP-2D, **BioCLIP**, etc.	Variants of MedCLIP with better tuning	Higher performance on small datasets

☑ Official CLIP Models by OpenAI

These are trained on general web image-text pairs:

Model Name	Description
`ViT-B/32`	Vision Transformer (Base) with 32×32 patches — **Fastest, common default.**
`ViT-B/16`	Higher resolution (16×16 patches) — **Better performance.**
`ViT-L/14`	Larger model, better accuracy — **Slower but more powerful.**
`ViT- L/14@336px`	Larger input size (336px) — **Highest accuracy.**
`RN50`, `RN101`	ResNet backbones — **Older, less popular than ViT.**

☑ Which CLIP Model to Use for Medical Datasets?

Dataset	Recommended Model	Why
MIMIC-CXR	MedCLIP / BioViL / GLoRIA	Trained on or designed for chest X-rays
CheXpert	CheXzero	Trained/tested for zero-shot on CheXpert
NIH ChestX-ray14	MedCLIP / BioViL	Large scale, MedCLIP handles this well
TB Datasets (like yours)	MedCLIP or CheXzero	Both generalize well, even to unseen diseases
General Medical	BioViL / GLoRIA	Broader coverage of anatomy and disease types

Which Model Should we use and Why ??

☑ PART 1: What Are the Official CLIP Models by OpenAI?

These are **general-purpose CLIP models** trained by OpenAI on **400M image–text pairs from the internet (non-medical)**. They're great for general vision tasks, but **not specialized for medical images**.

Model Name	Architecture	Patch Size / Details	Strengths	Weaknesses
ViT-B/32	Vision Transformer (Base)	32×32 patches	Fast, lightweight	Lower resolution → lower accuracy
ViT-B/16	Vision Transformer (Base)	16×16 patches	Better resolution & accuracy than B/32	Slower than B/32
ViT-L/14	Vision Transformer (Large)	14×14 patches	Strong performance	Bigger model → more compute
ViT-L/14@336px	Larger input image size (336px)	Higher detail level	Best accuracy	Slowest, highest GPU need
RN50 / RN101	ResNet-50 / ResNet-101	CNN backbones	Familiar for classic CV users	Lower performance vs ViTs

🟡 **Used for**: General images — dogs, buildings, cars, memes, etc.
🔴 **Not trained** on X-rays, CTs, MRIs → worse performance on medical data.

☑ PART 2: Why Do We Use MedCLIP, CheXzero, BioViL, etc. for Medical Data?

These models are built **on top of CLIP concepts** but trained with **medical-specific data**:

Model	Based On	Trained With	Best For
MedCLIP	CLIP-style ViT	MIMIC-CXR (chest X-rays)	TB, NIH, MIMIC, general CXRs
CheXzero	CLIP + T5-style text encoder	CheXpert + radiology text	Zero-shot on CheXpert, TB
BioViL	CLIP-like ViT + BERT	Biomedical images + text	General biomedical vision-language tasks
GLoRIA	ViT + Local–Global attention	Biomedical data	Fine-grained disease localization + general use

These models still follow the **CLIP architecture style**:

- Visual encoder (ViT or CNN)
- Text encoder (BERT, T5, etc.)
- Contrastive loss: Match image ↔ text

But they're trained **specifically for**:

- Radiology reports
- Medical image datasets (MIMIC-CXR, CheXpert, etc.)
- Better terminology understanding ("pneumothorax", "opacity", "nodule", etc.)

🟢 So Which Should *You* Use?

- For **TB Chest X-rays** or similar:

Need	Use
Zero-shot classification	☑ **CheXzero** (no retraining needed)
Few-shot with own classifier	☑ **MedCLIP** (feature extraction + Logistic Regression)
General biomedical tasks	☑ **BioViL** or **GLoRIA**

☑ Summary Table

Model Type	Purpose	Good For Medical?	Notes
OpenAI CLIP (ViT-B/32, etc.)	General images	❌ No	Great for dogs/cars, weak on X-rays
MedCLIP, CheXzero, BioViL	Medical data	☑ Yes	Trained on chest X-rays, radiology reports