Self-Supervised CLIP Fine-Tuning with Medical Image-Text Data

Gabriel Bülow, Abbas Alubeid, Karl Duckert Karlsson,
Gor Durgaryan, Johan von Axelson, Lukas Ingemarsson

Introduction

Introduction to the project

- A project about finetuning CLIP on a specific domain
- Finetuning on medical images (specializing in the domain)
- Image-text pairs

Reason for finetuning on medical images

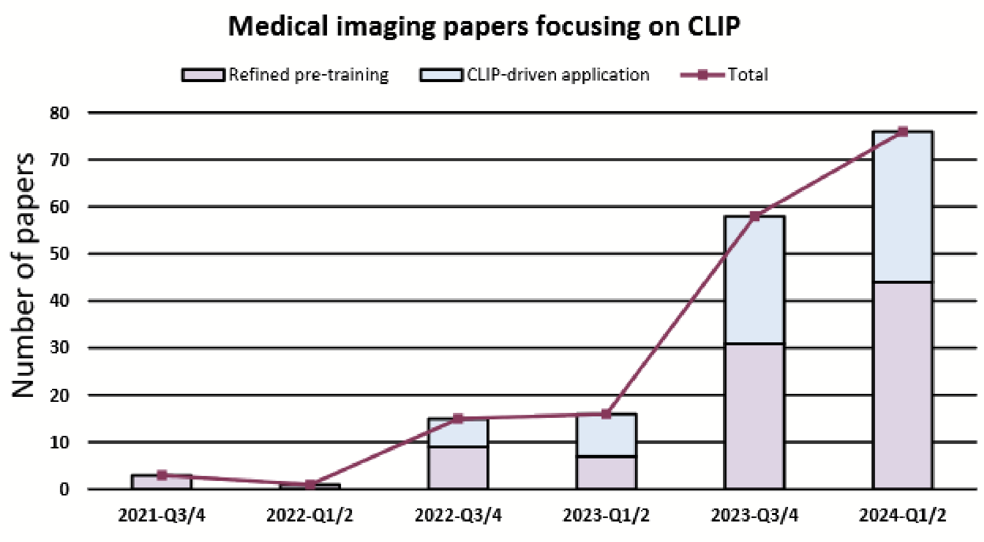


Fig. 2. Rapid increase of the number of medical imaging papers focusing on CLIP. Refined pre-training and CLIP-driven application are the two main taxonomy categories introduced in this survey.

Picture from "*CLIP in medical imaging: A comprehensive survey*" (Zhao et al., 2024)

CLIP has gained a lot of interest in the medical imaging domain in recent years

CLIP (Contrastive Language-Image Pre-Training)

- Learns the relation between embedded images and text
  - Uses separate encoders for text and images to transform input into numerical vector representations (embeddings)
- Pretrained on a dataset with 400M image-text pairs (Radford et al., 2021)

Zero-shot learning

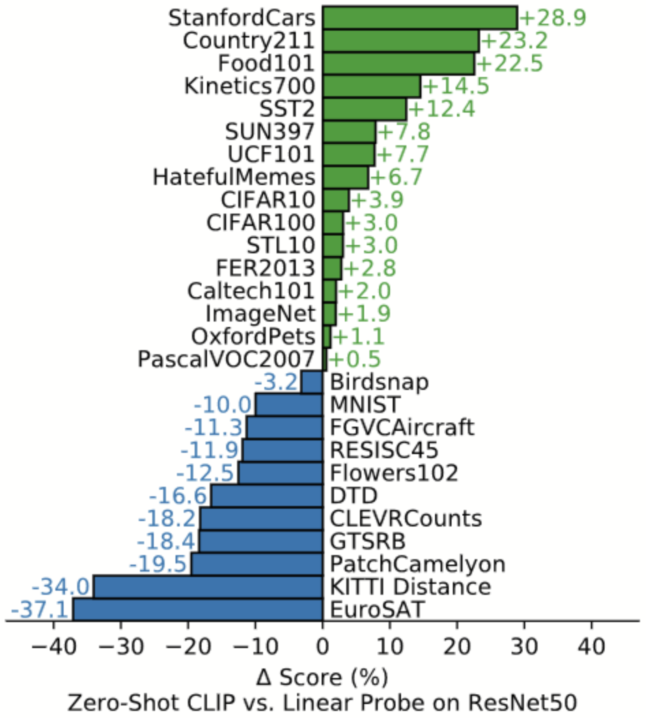


Figure 5. **Zero-shot CLIP is competitive with a fully supervised baseline.** Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet-50 features on 16 datasets, including ImageNet.

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

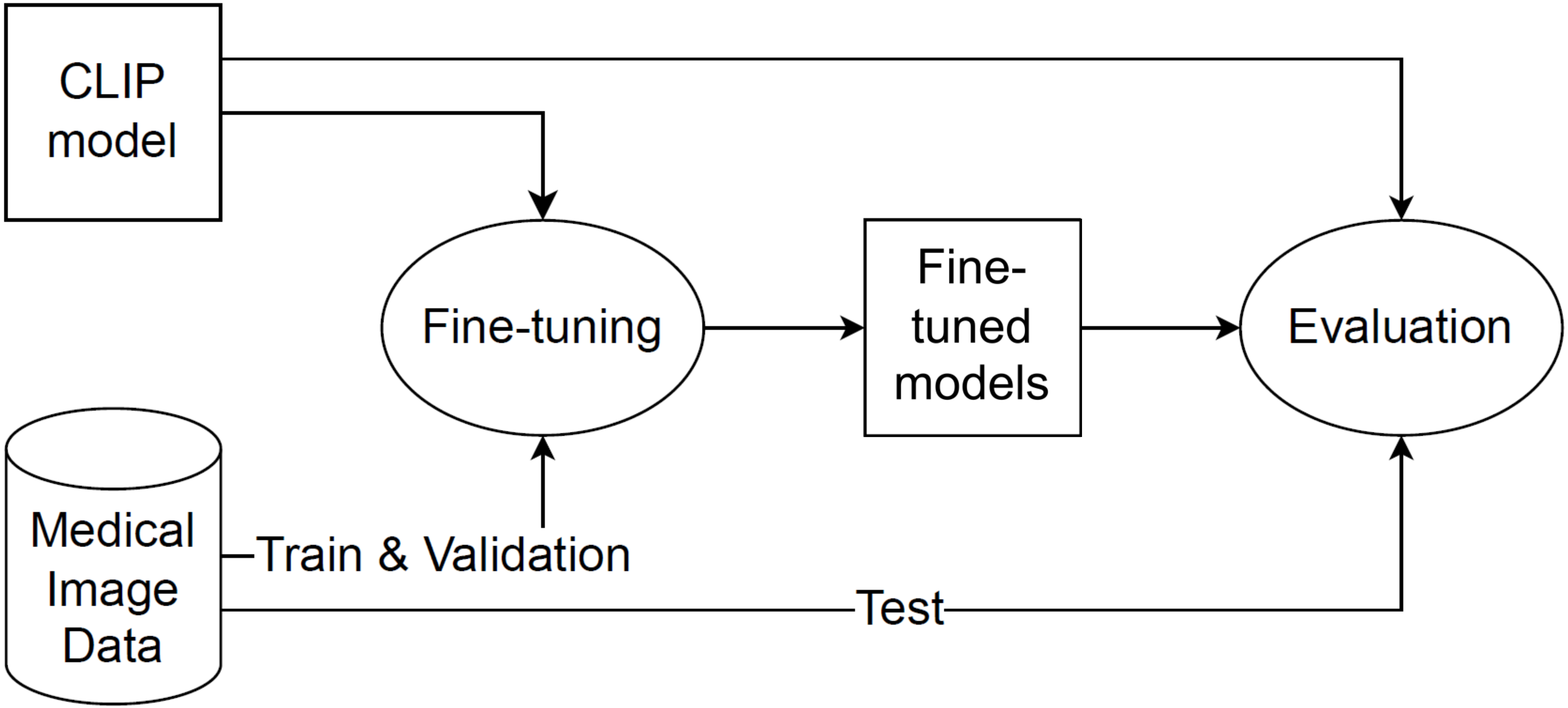
- CLIP is known for its zero-shot learning abilities
  - Classify images in categories it was never explicitly trained on.

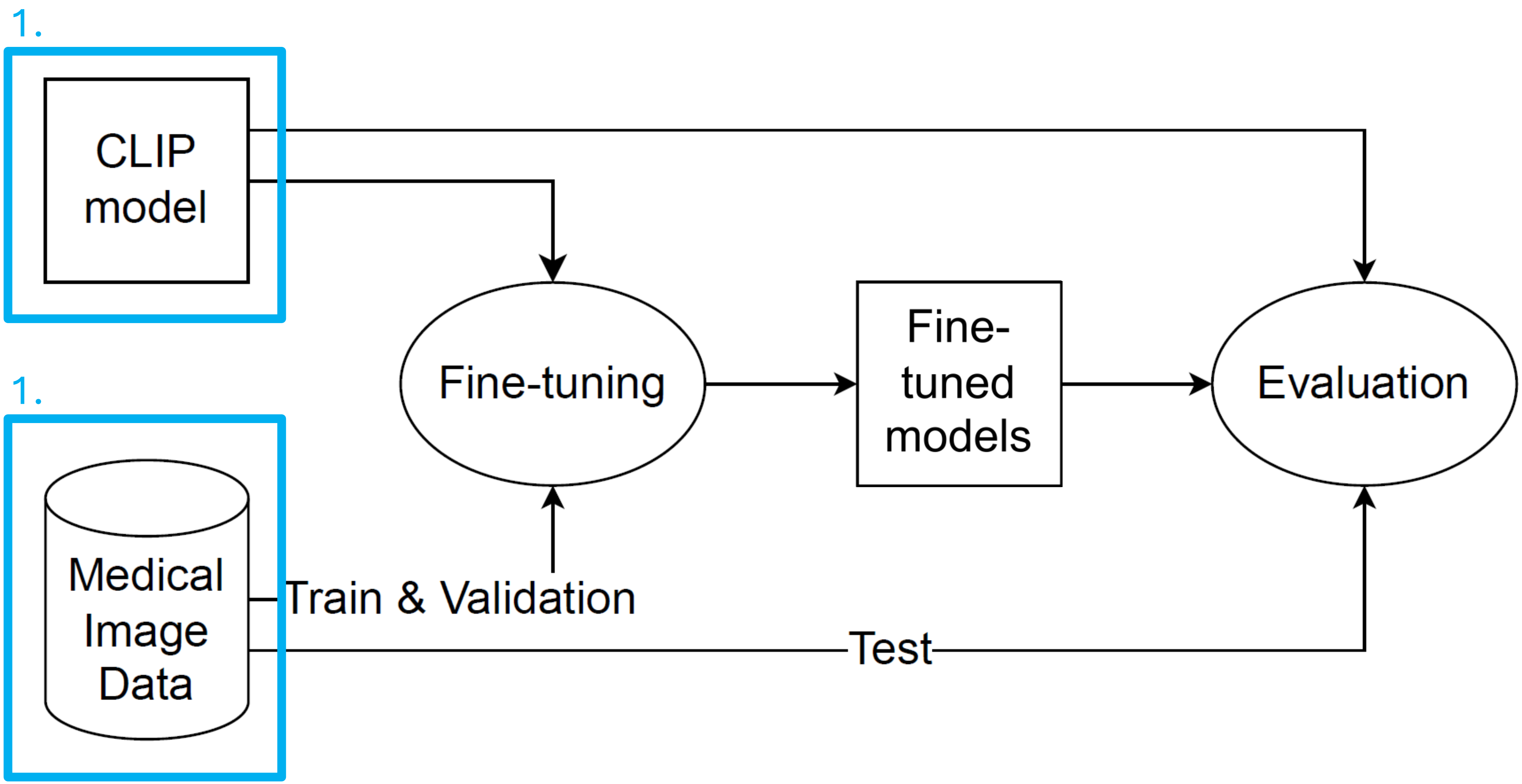
Main question for the project

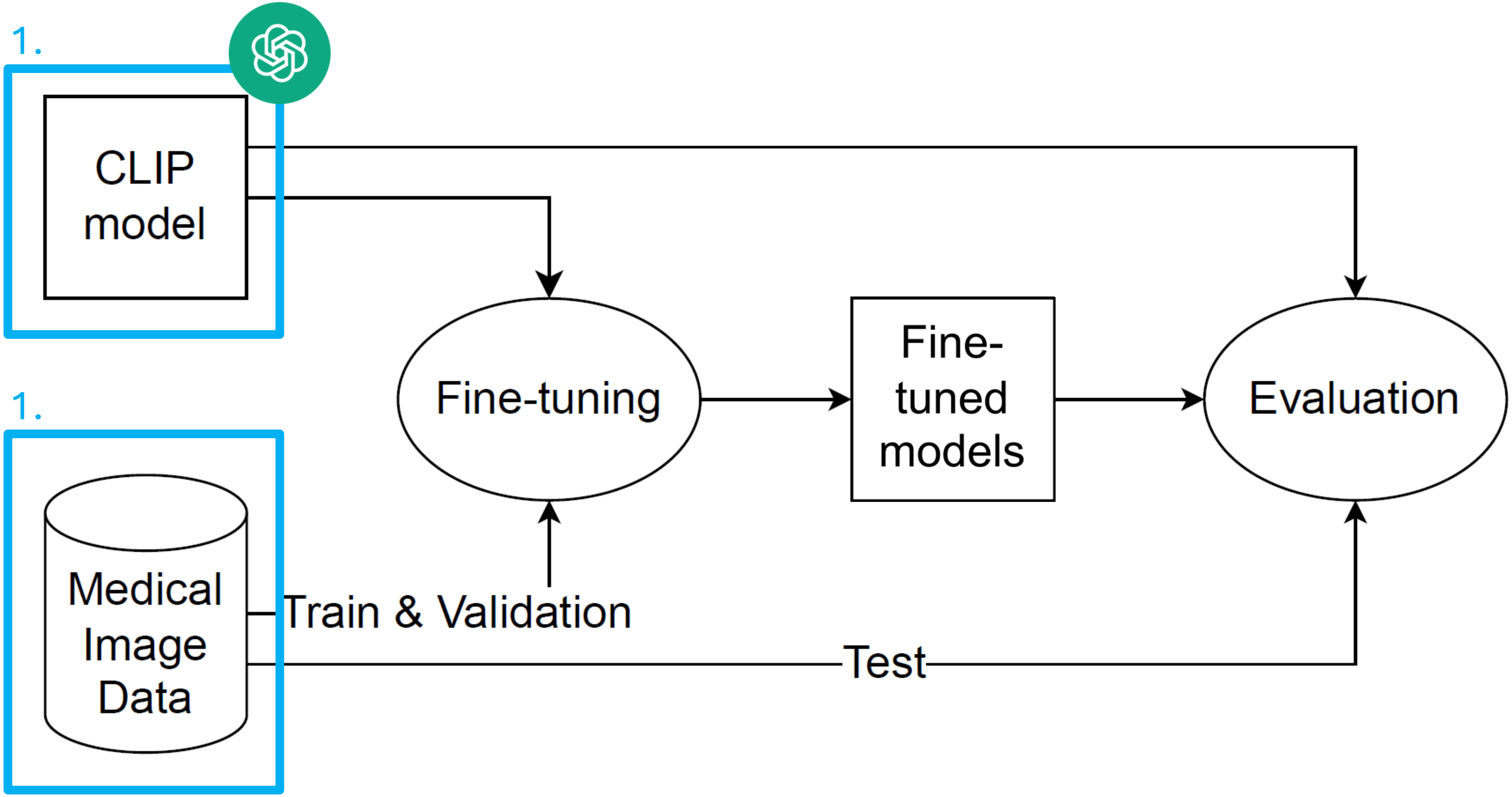
Can we get useful finetuning and increase the accuracy of CLIP in a specific domain with just image and text data?

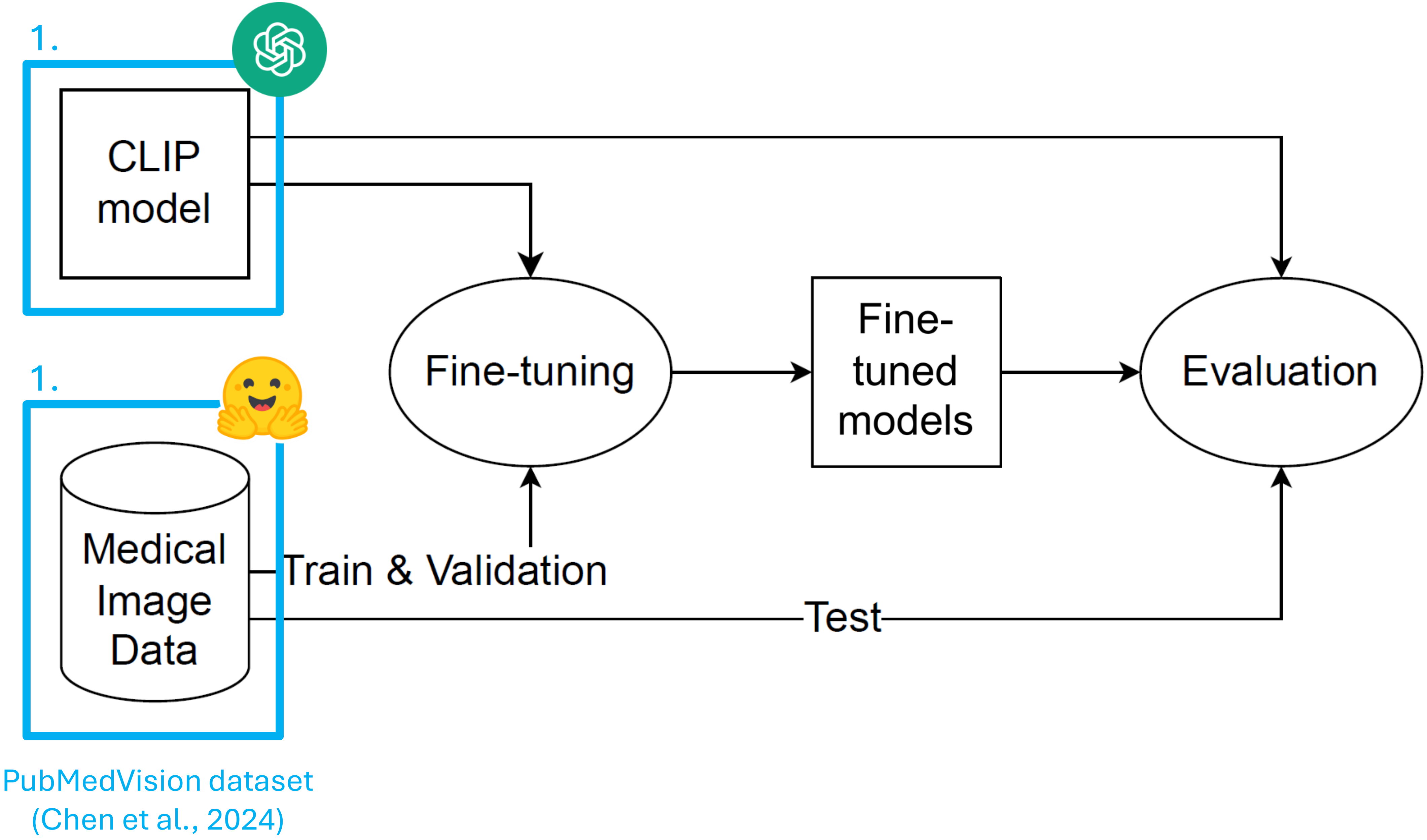
Method

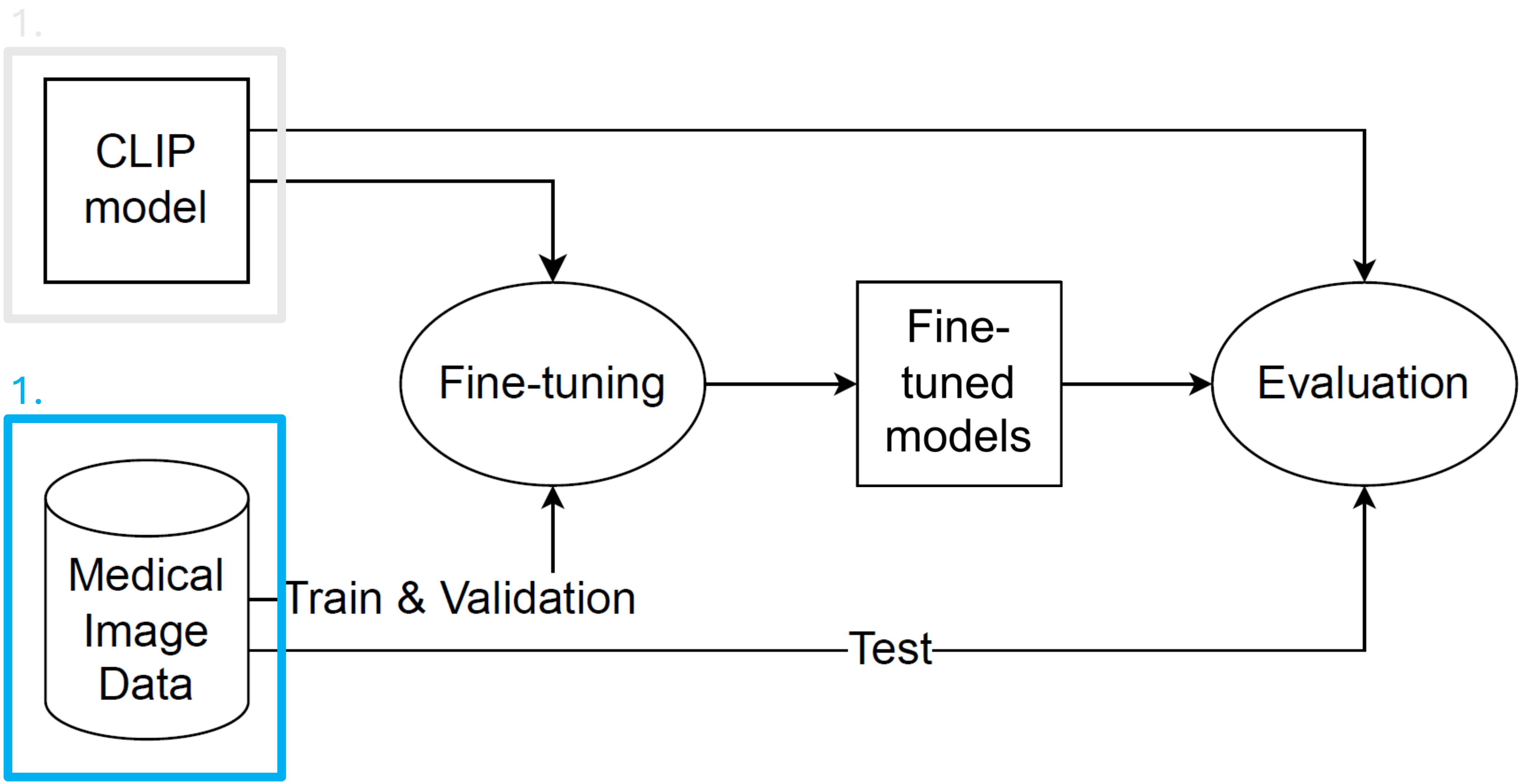
Fine-tuning Pipeline



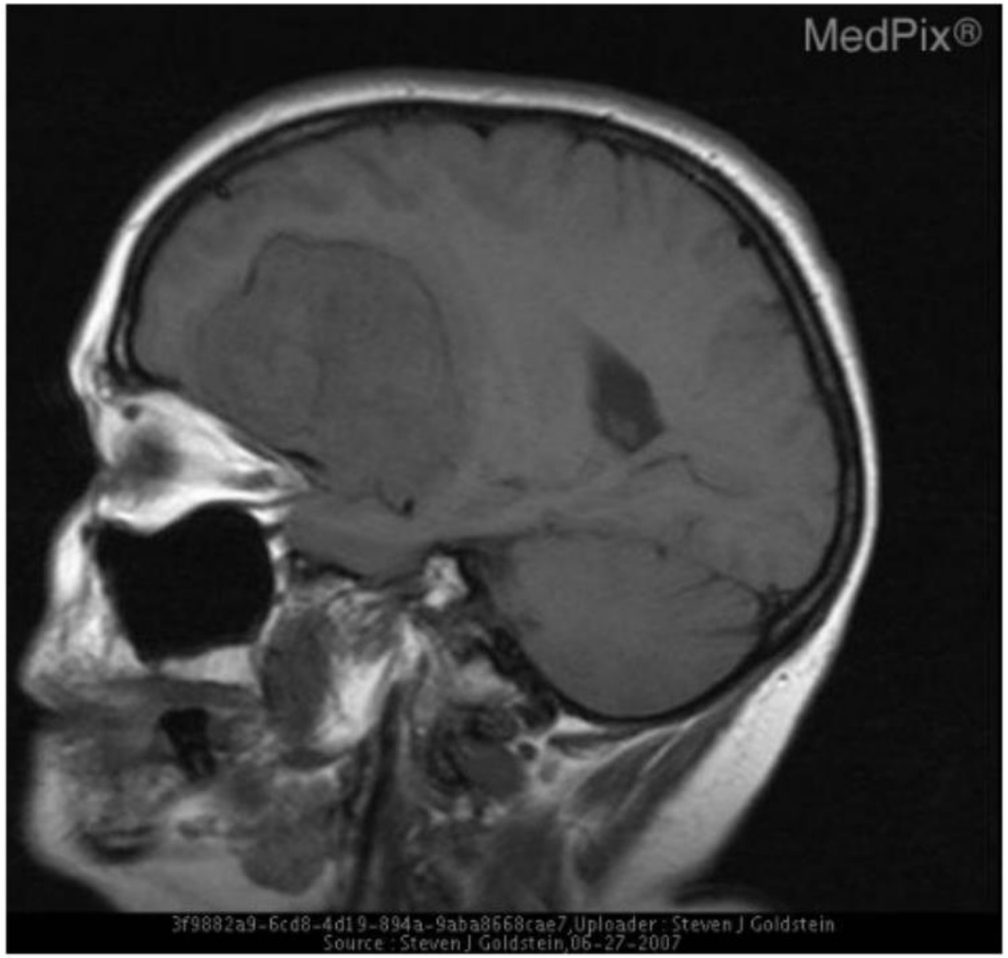

1.
CLIP model
1.
Medical Image Data
Fine-tuning
Fine-tuned models
Evaluation
Train & Validation
Test


1.
CLIP model
1.
Medical Image Data
Fine-tuning
Fine-tuned models
Evaluation
Train & Validation
Test


1.
CLIP model
1.
Medical Image Data
Fine-tuning
Fine-tuned models
Evaluation
Train & Validation
Test
PubMedVision dataset
(Chen et al., 2024)


1.
CLIP model
1.
Medical Image Data
Fine-tuning
Fine-tuned models
Evaluation
Train & Validation
Test

Image

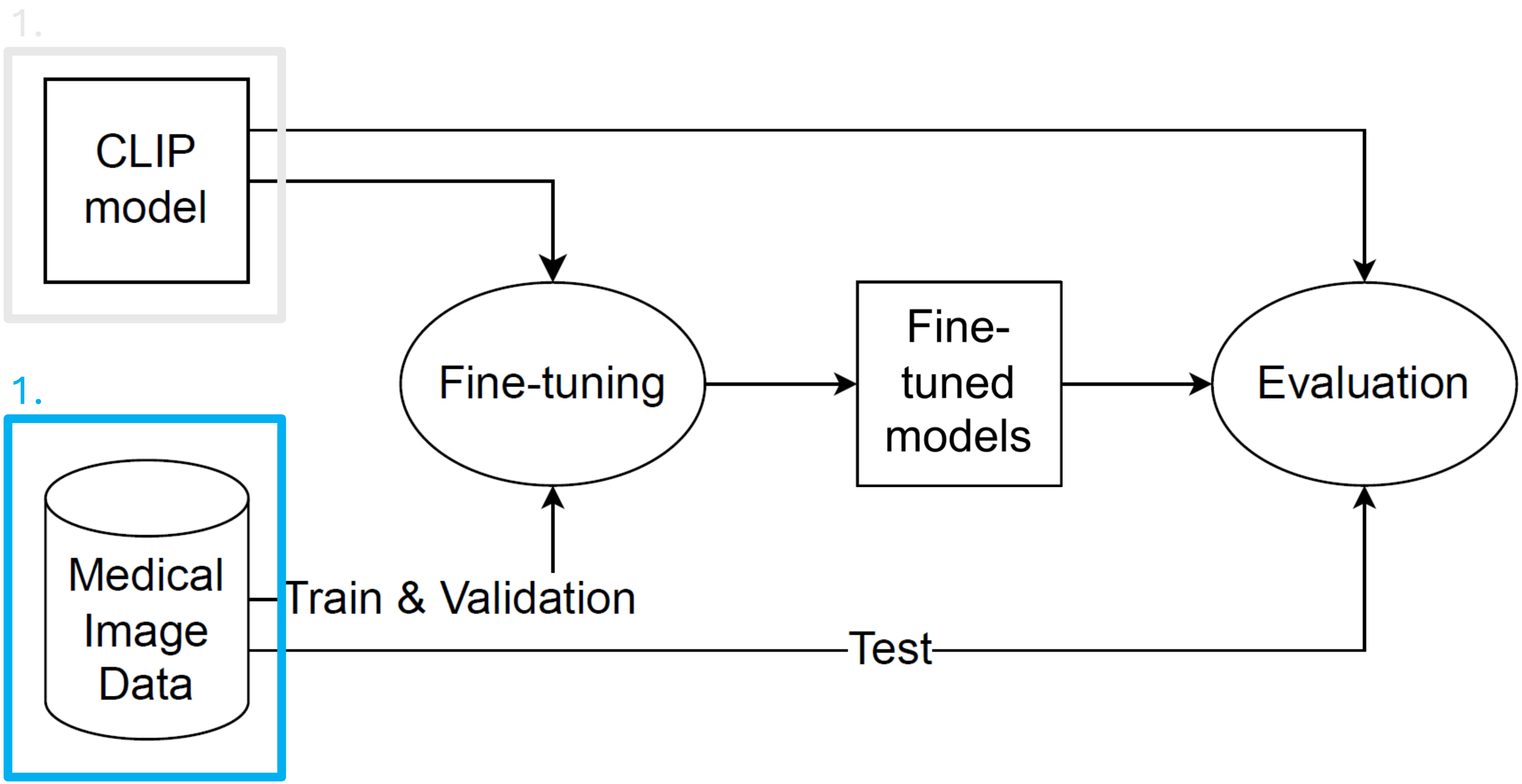


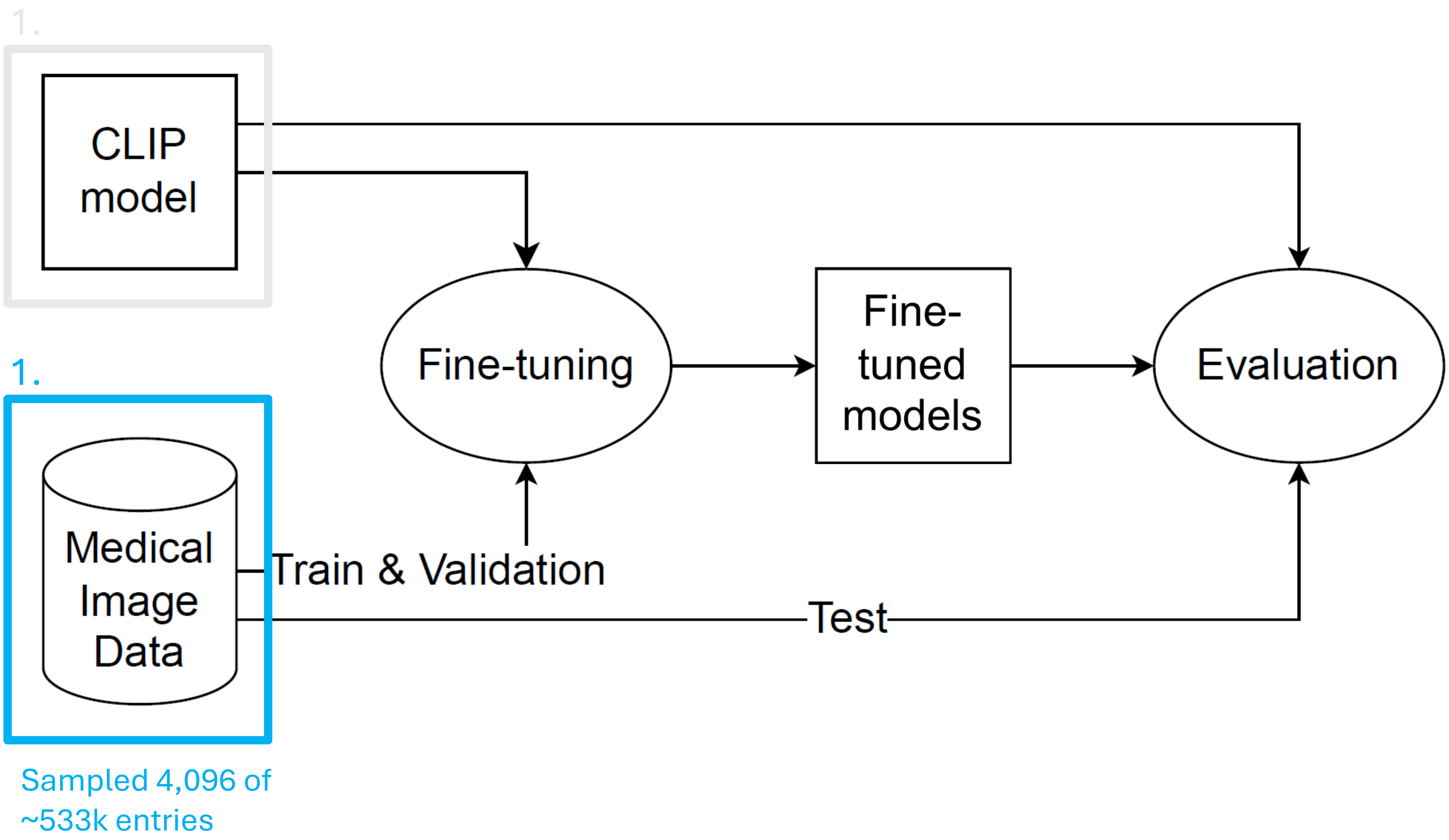
Text

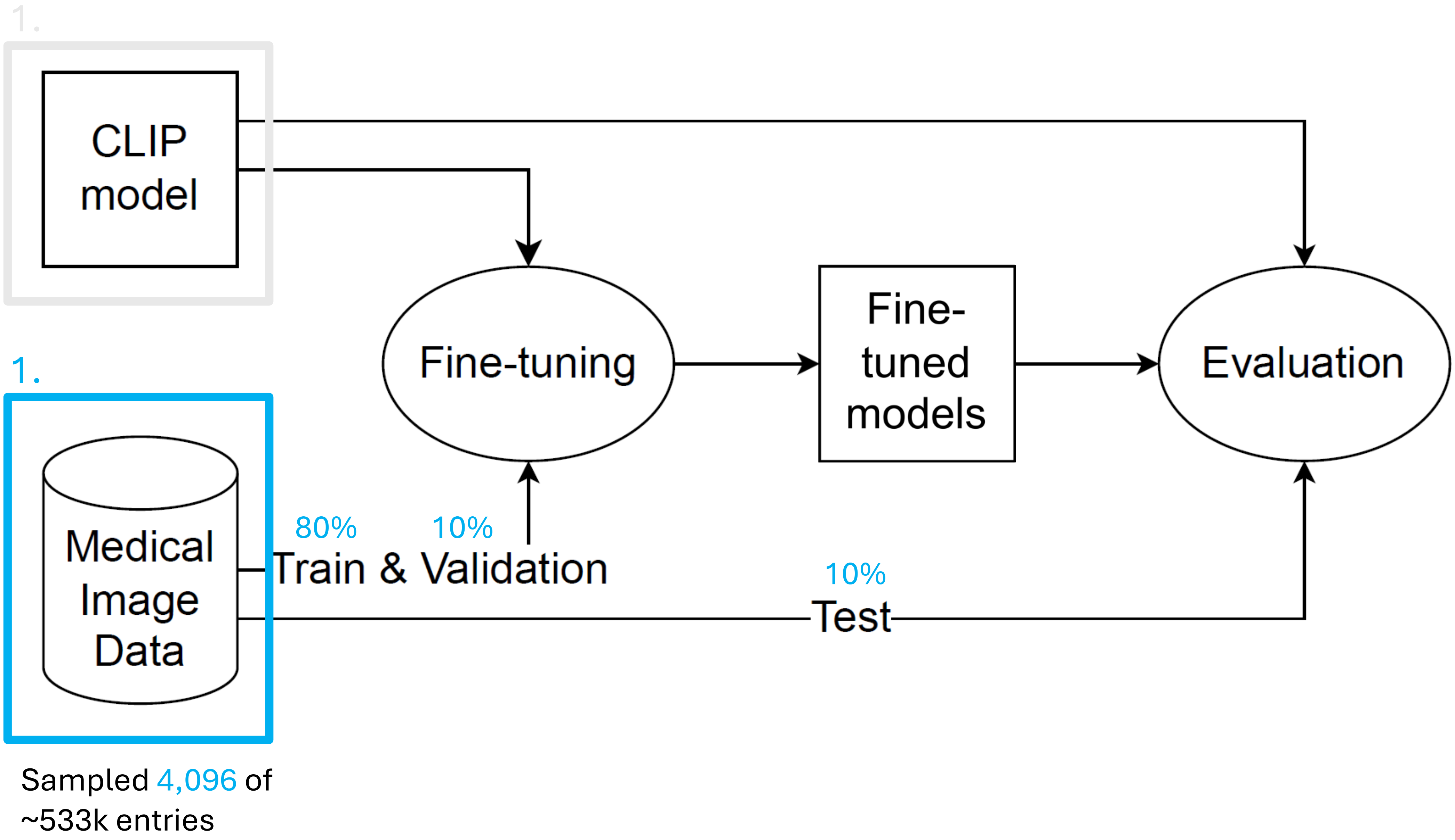
"The provided medical image is a sagittal MRI scan of the brain, showing a prominent lesion that is hyperintense on this T1-weighted image. The lesion is located extracranially but impinges upon the cerebral tissue, specifically around the right sphenoid ridge. It exhibits characteristics suggestive of a significant calcification, as indicated by its brightness. This lesion also demonstrates ...”

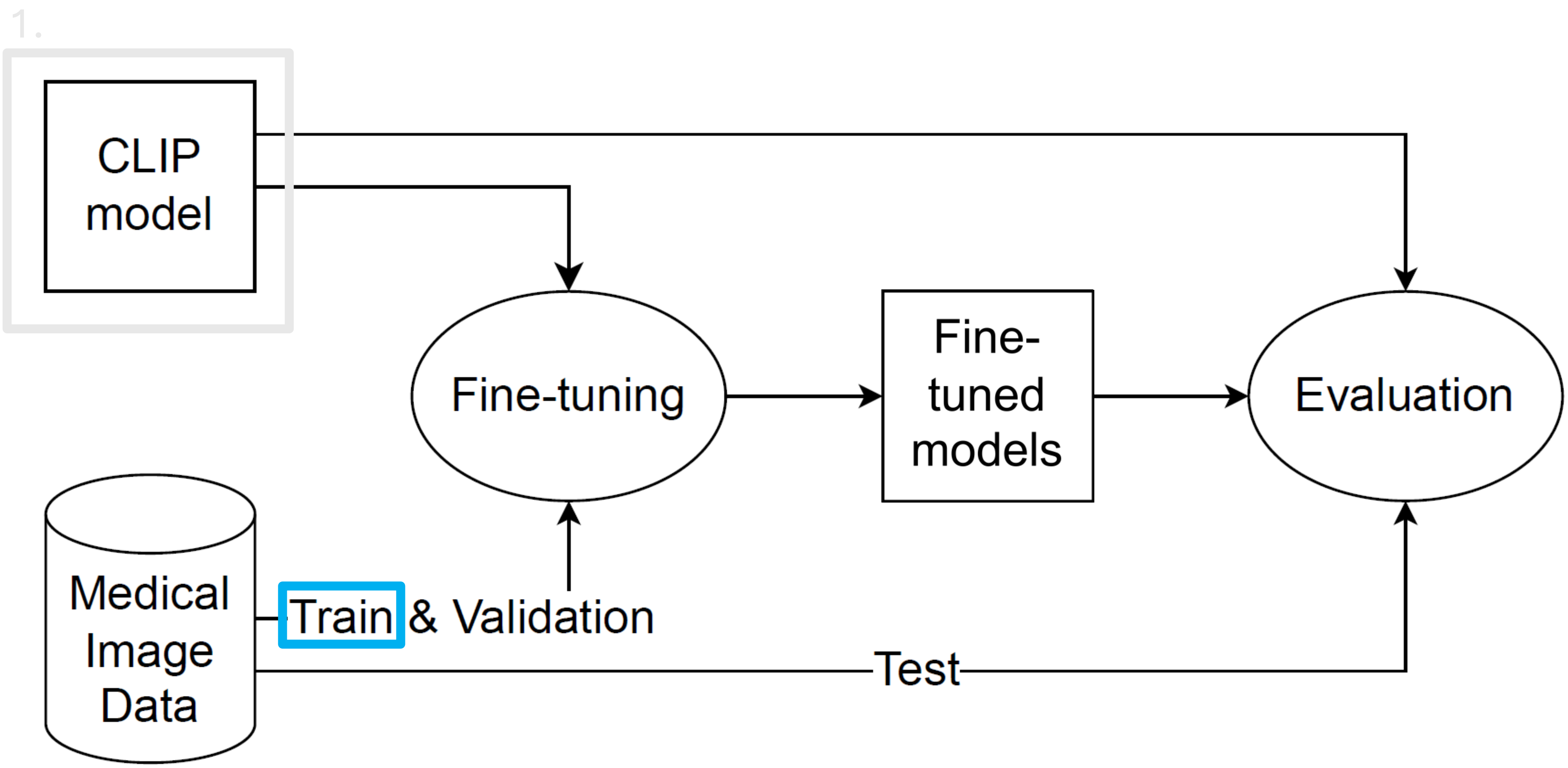
Label

Brain

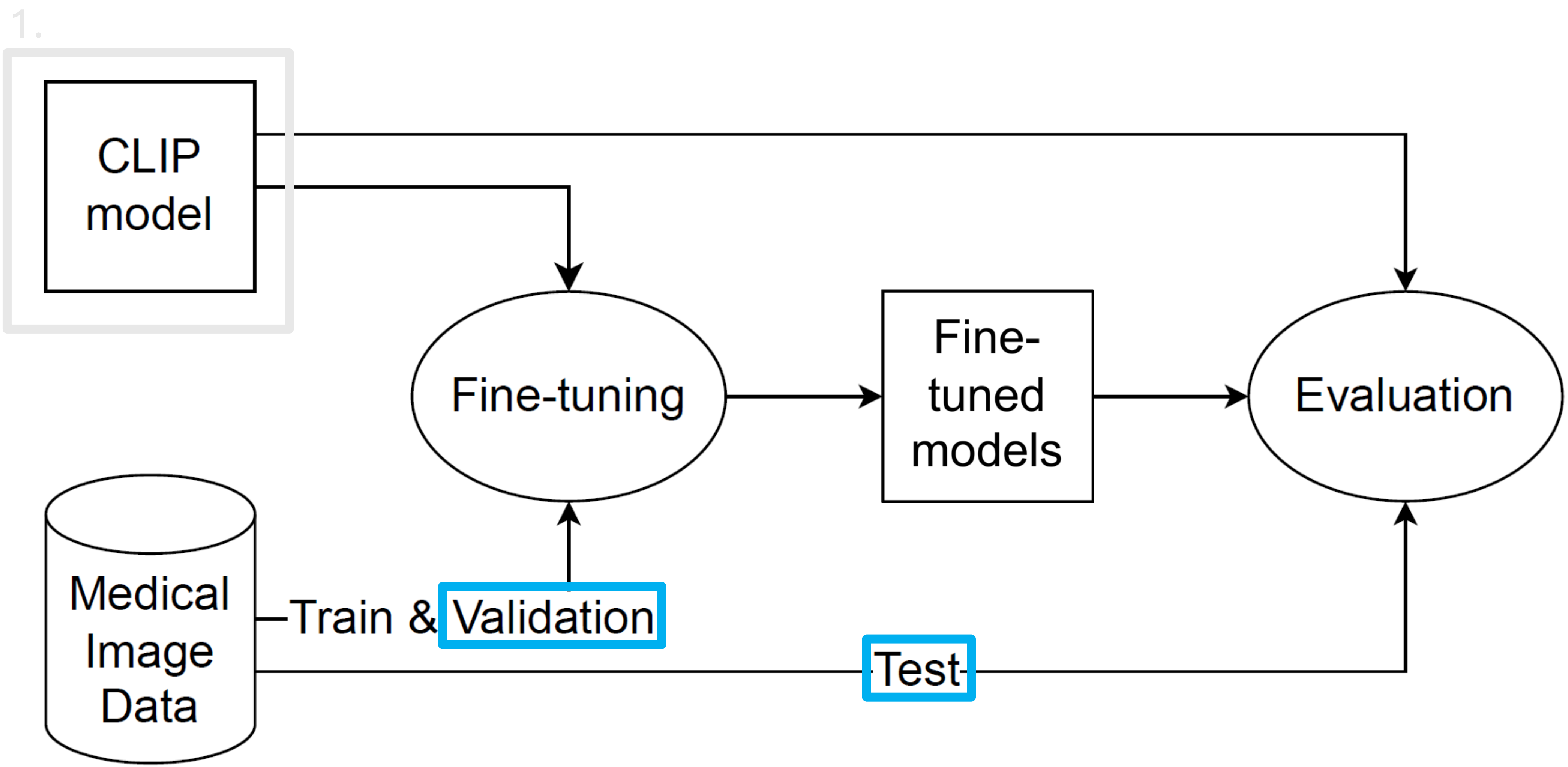

1.
CLIP model
1.
Medical Image Data
Fine-tuning
Fine-tuned models
Evaluation
Train & Validation
Test


1.
CLIP model
Fine-tuning
Fine-tuned models
Evaluation
1.
Medical Image Data
Train & Validation
Test
Sampled 4,096 of ~533k entries

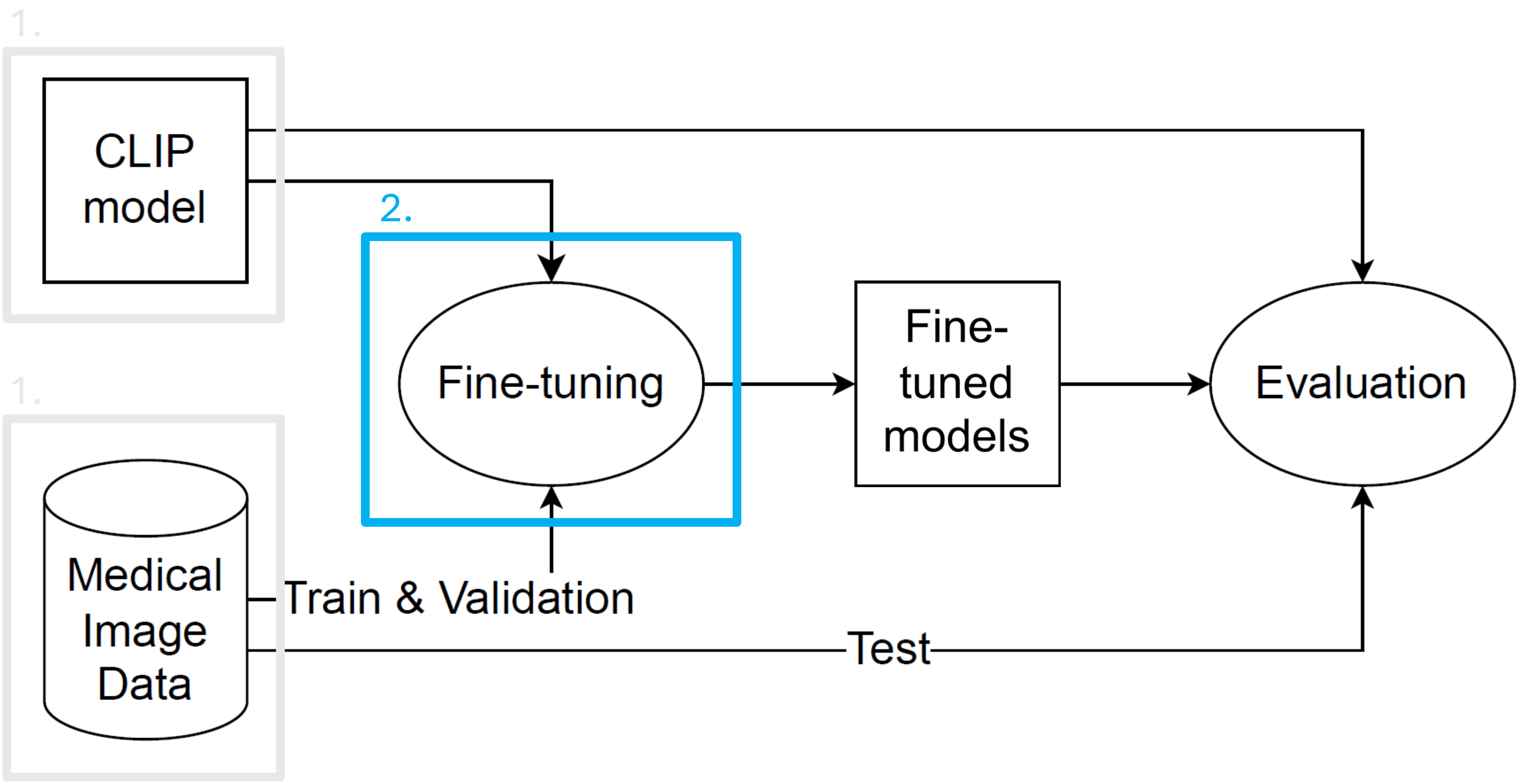

1.
CLIP model
1.
Medical Image Data
Fine-tuning
Fine-tuned models
Evaluation
80%
10%
Train & Validation
10%
Test
Sampled 4,096 of
~533k entries

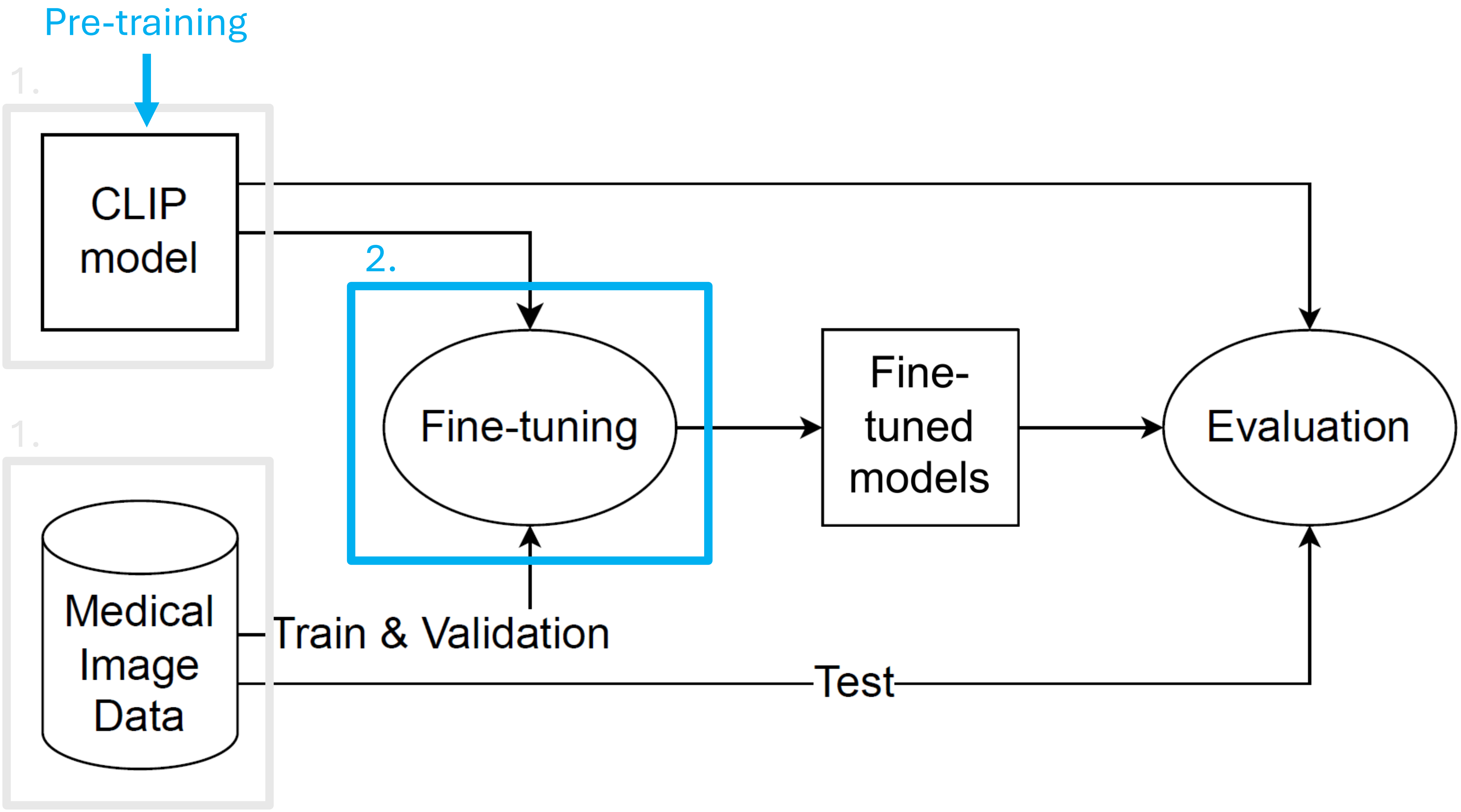


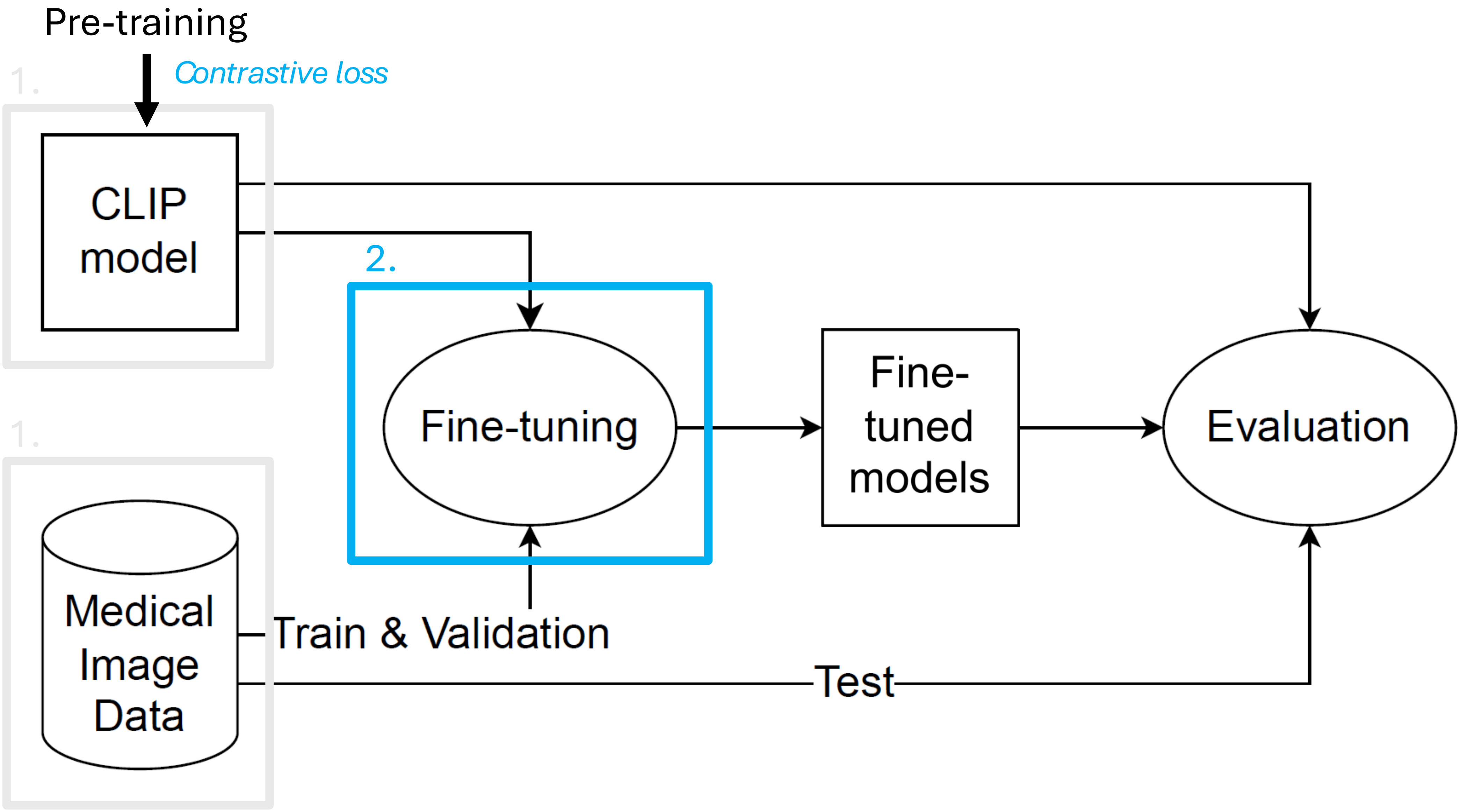
Sampled 4,096 of
~533k entries



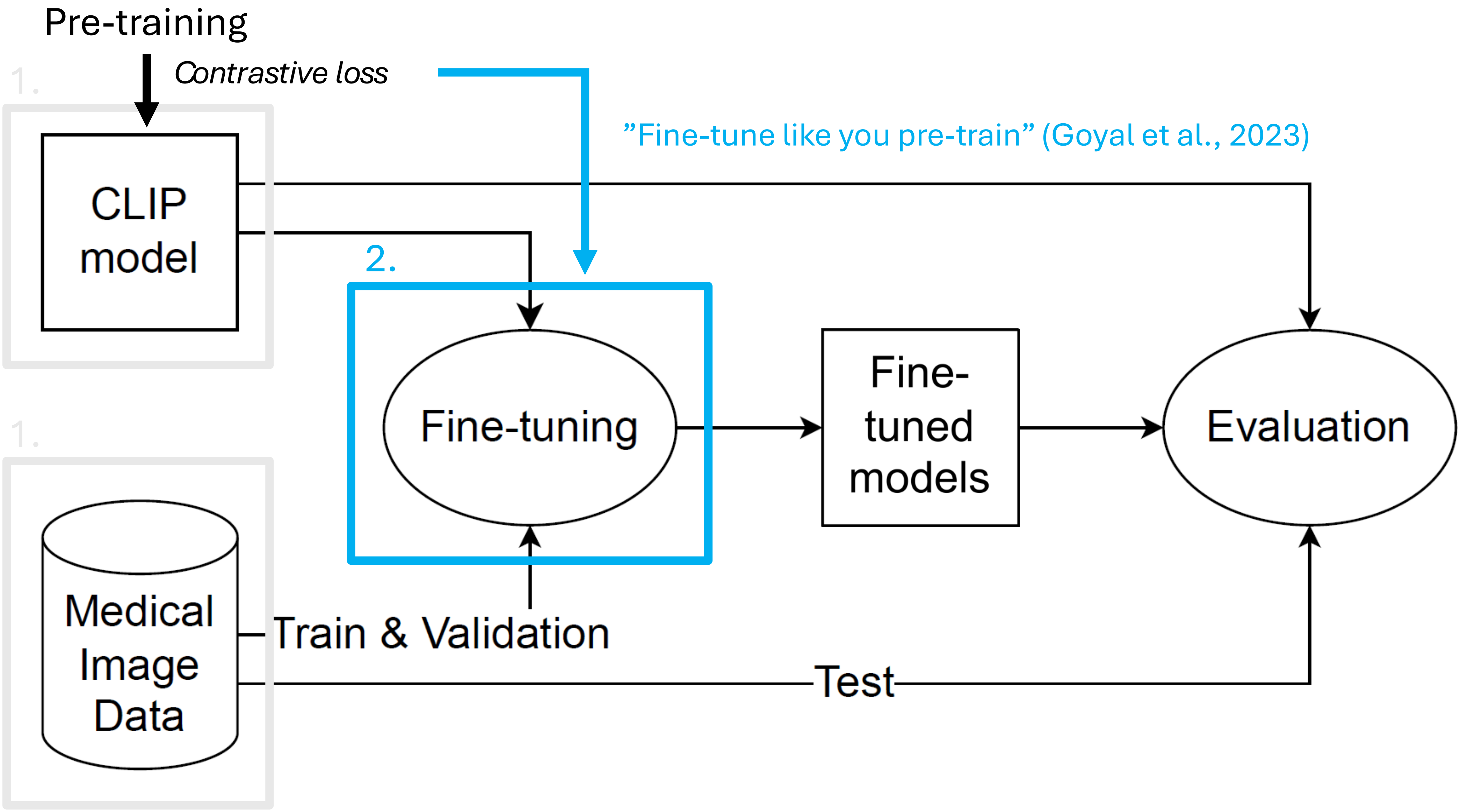
Sampled 4,096 of
~533k entries


1.
CLIP model
1.
Medical Image Data
2.
Fine-tuning
Fine-tuned models
Evaluation
Train & Validation
Test


Pre-training
1.
CLIP model
1.
Medical Image Data
2.
Fine-tuning
Fine-tuned models
Evaluation
Train & Validation
Test



Pre-training
Contrastive loss
1.
CLIP model
2.
Fine-tuning
Fine-tuned models
Evaluation
1.
Medical Image Data
Train & Validation
Test


Pre-training
Contrastive loss
1.
CLIP model
”Fine-tune like you pre-train” (Goyal et al., 2023)
2.
Fine-tuning
Fine-tuned models
Evaluation
1.
Medical Image Data
Train & Validation
Test

Contrastive Loss

- Maximize similarity between GT pairs, while minimizing to the rest

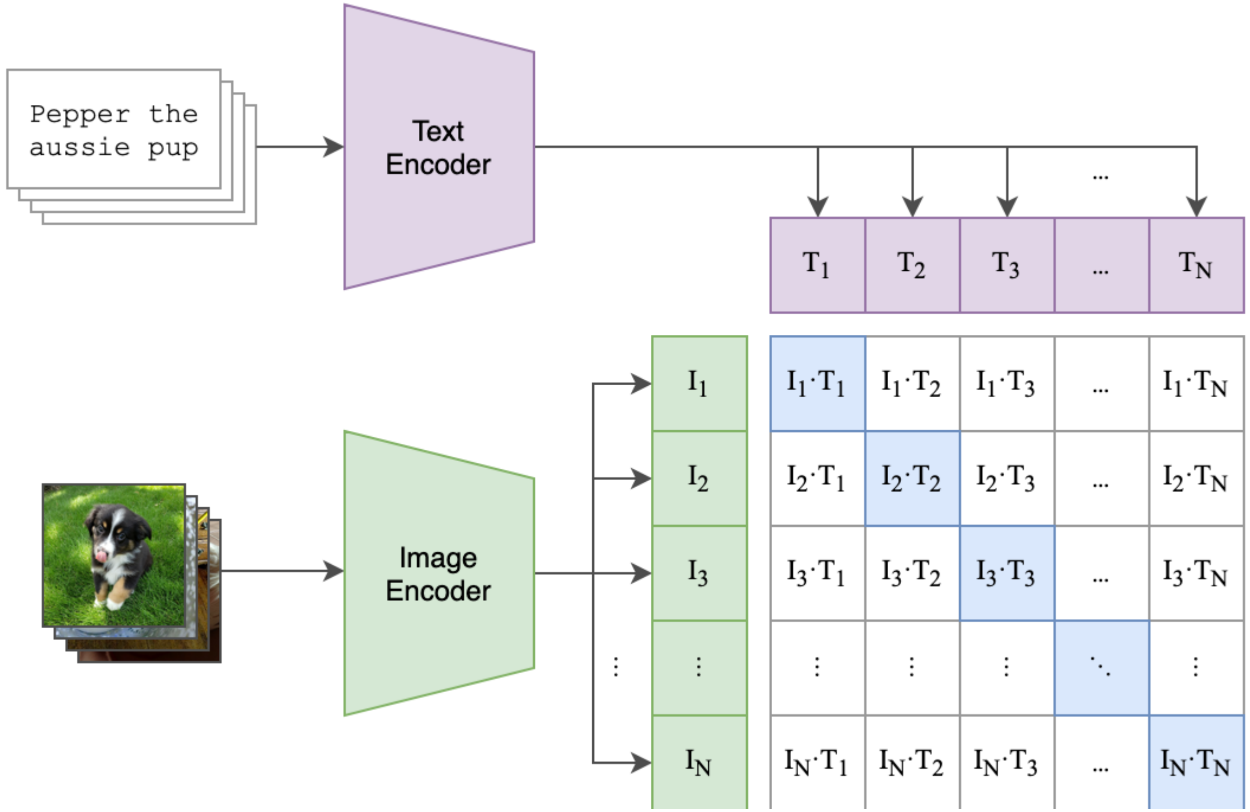


Image from: https://github.com/openai/CLIP

Contrastive Loss

- Maximize similarity between GT pairs, while minimizing to the rest

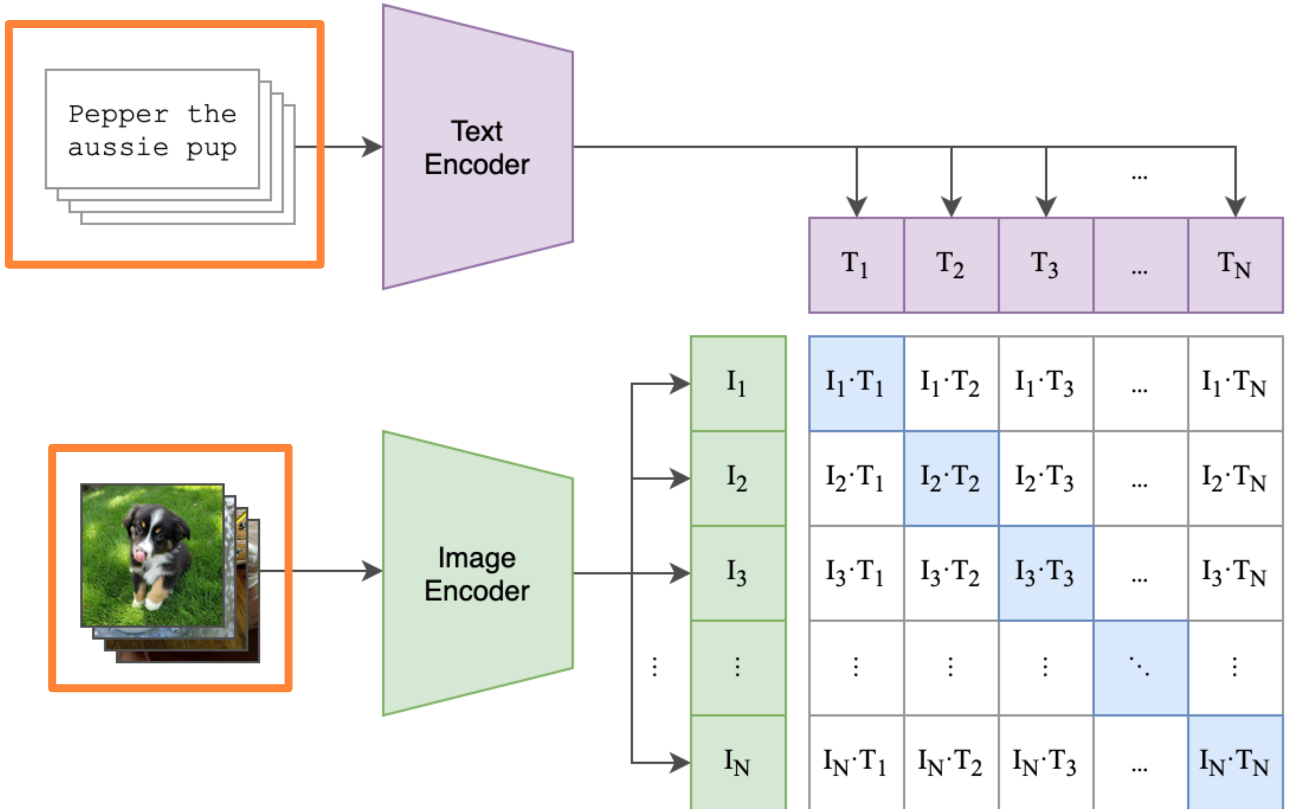


Image from: https://github.com/openai/CLIP

Contrastive Loss

- Maximize similarity between GT pairs, while minimizing to the rest

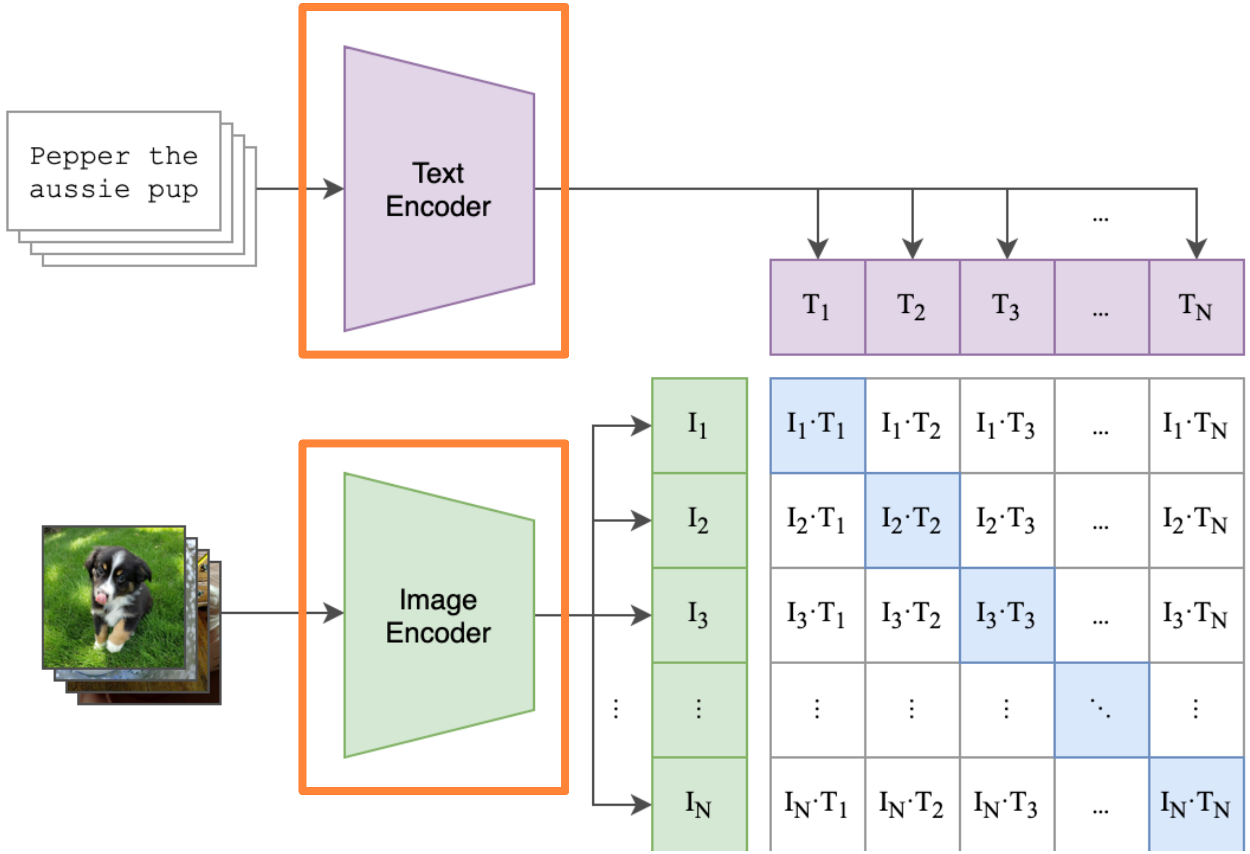


Image from: https://github.com/openai/CLIP

Contrastive Loss

- Maximize similarity between GT pairs, while minimizing to the rest

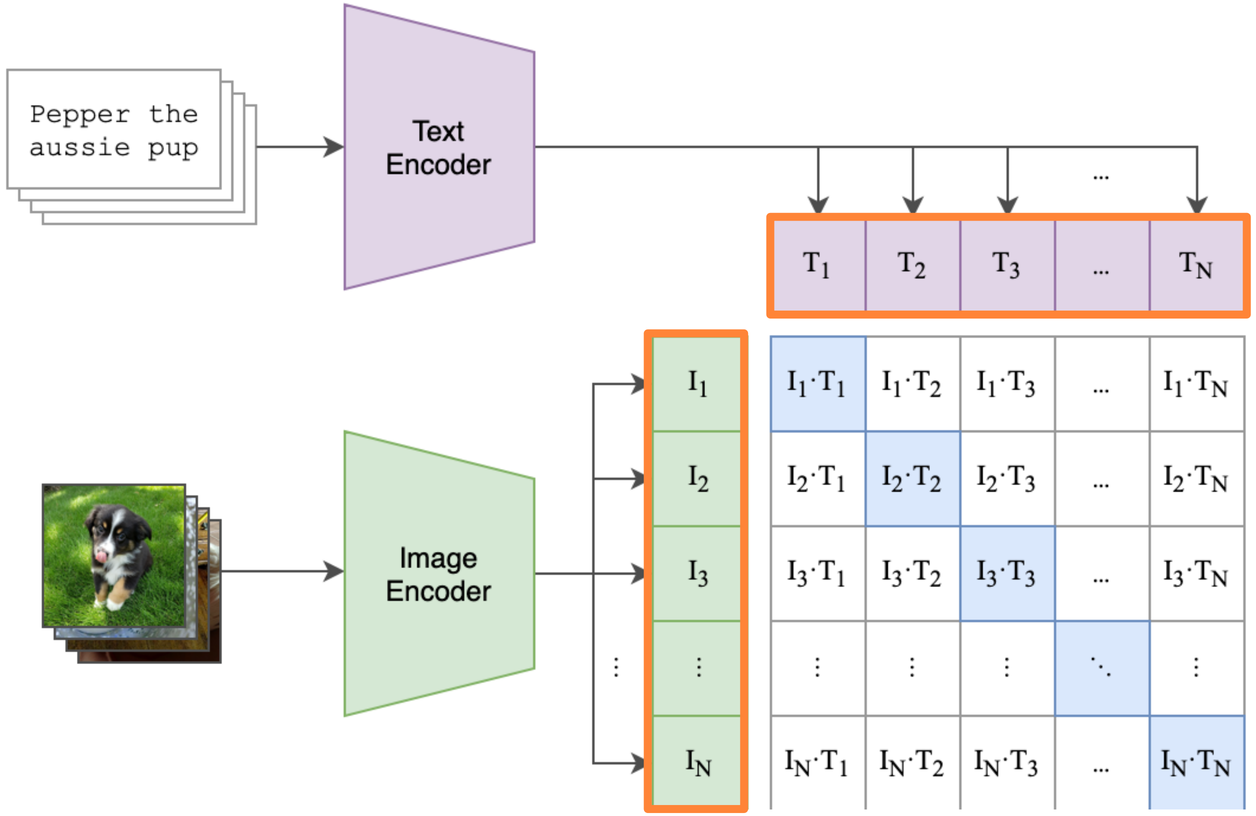


Image from: https://github.com/openai/CLIP

Contrastive Loss

- Maximize similarity between GT pairs, while minimizing to the rest

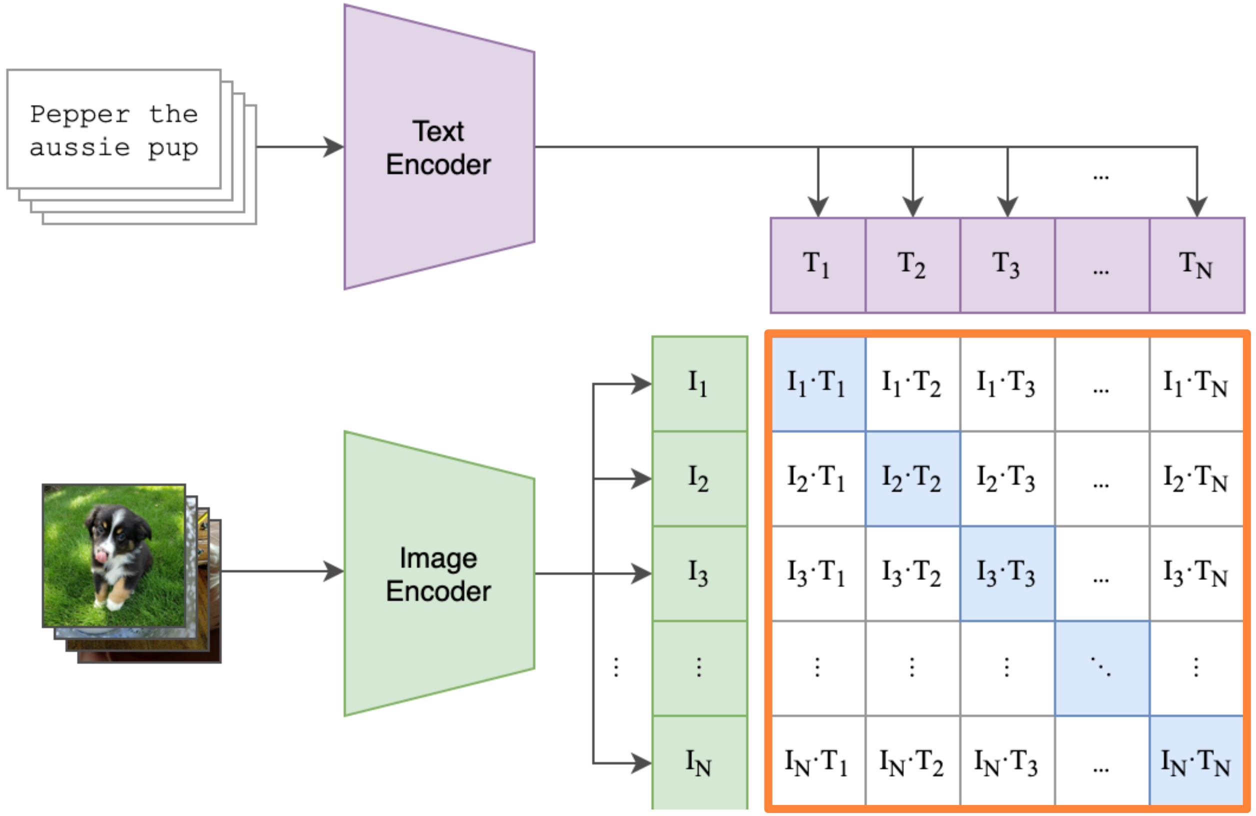


Image from: https://github.com/openai/CLIP

Contrastive Loss

- Maximize similarity between GT pairs, while minimizing to the rest

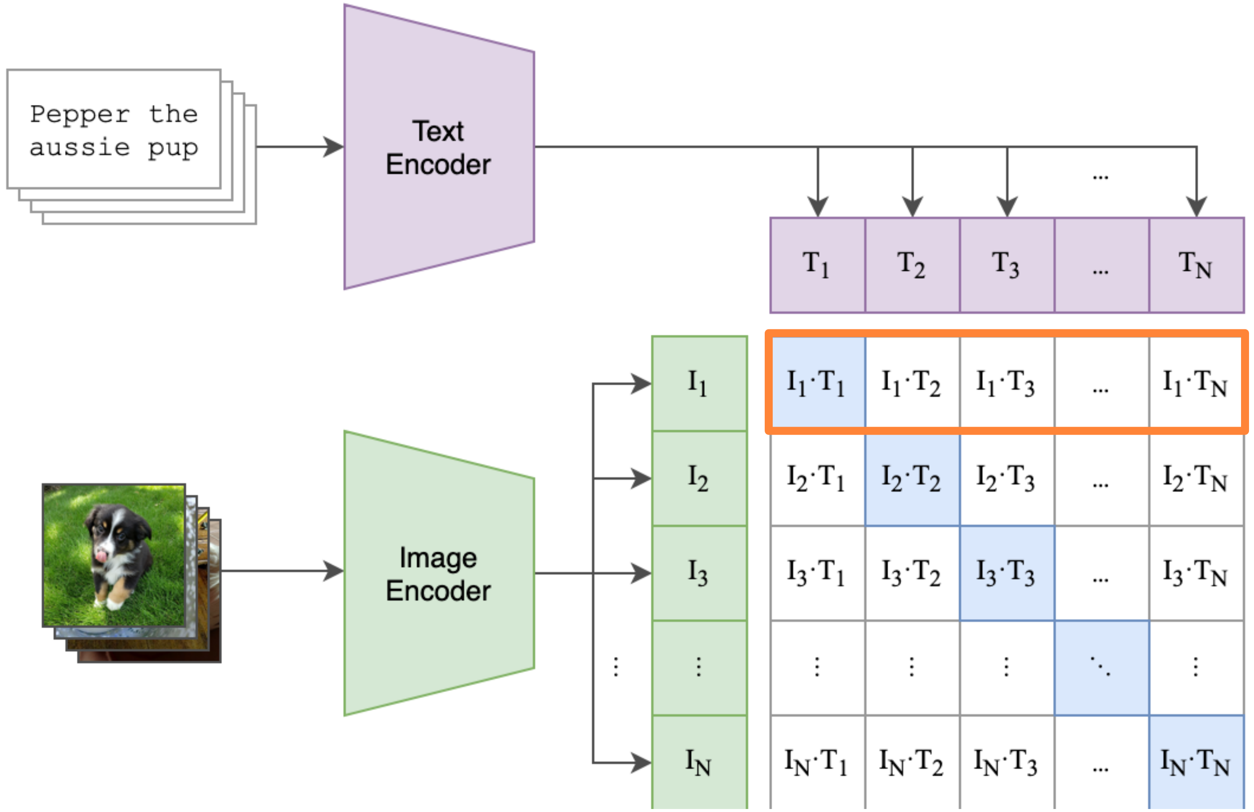


Image from: https://github.com/openai/CLIP

Contrastive Loss

- Maximize similarity between GT pairs, while minimizing to the rest

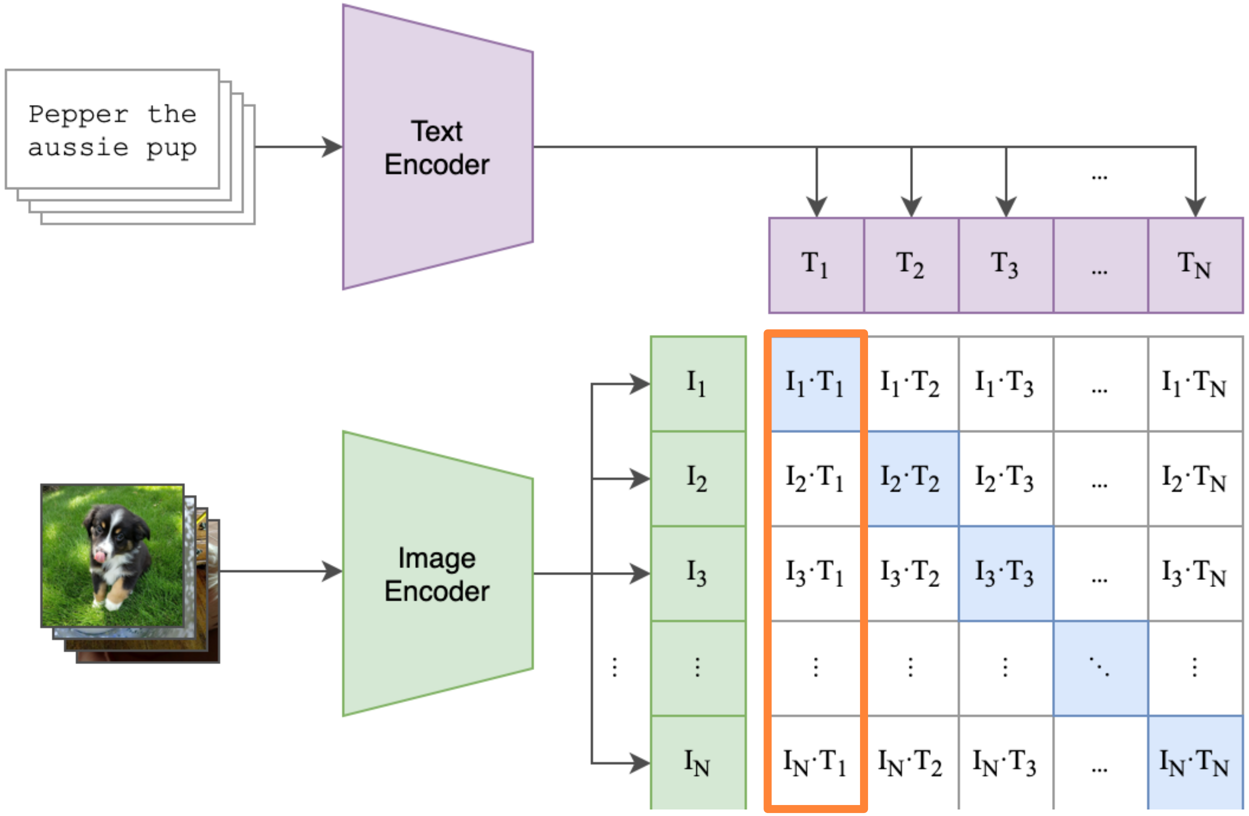


Image from: https://github.com/openai/CLIP

Contrastive Loss

- Maximize similarity between GT pairs, while minimizing to the rest

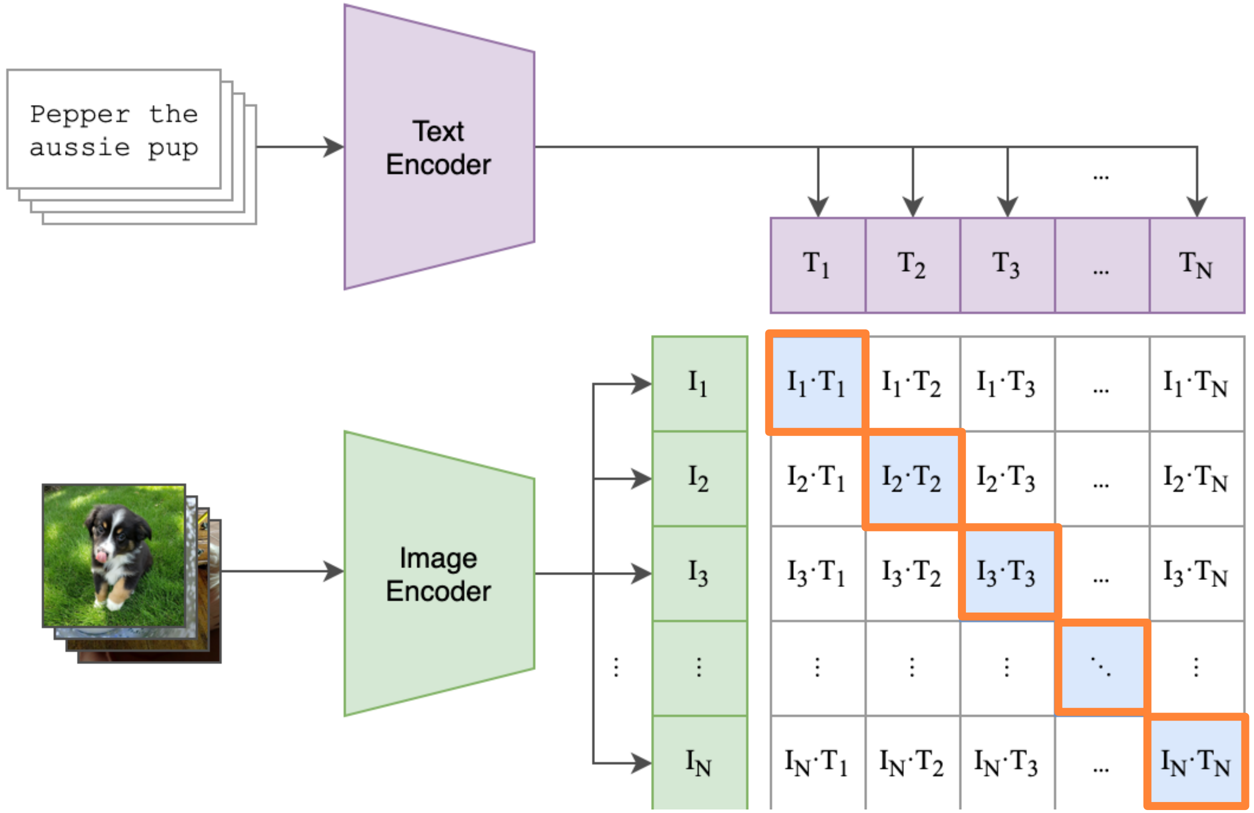


Image from: https://github.com/openai/CLIP

Contrastive Loss

- `total_loss = (image_to_text_loss + text_to_image_loss) / 2`

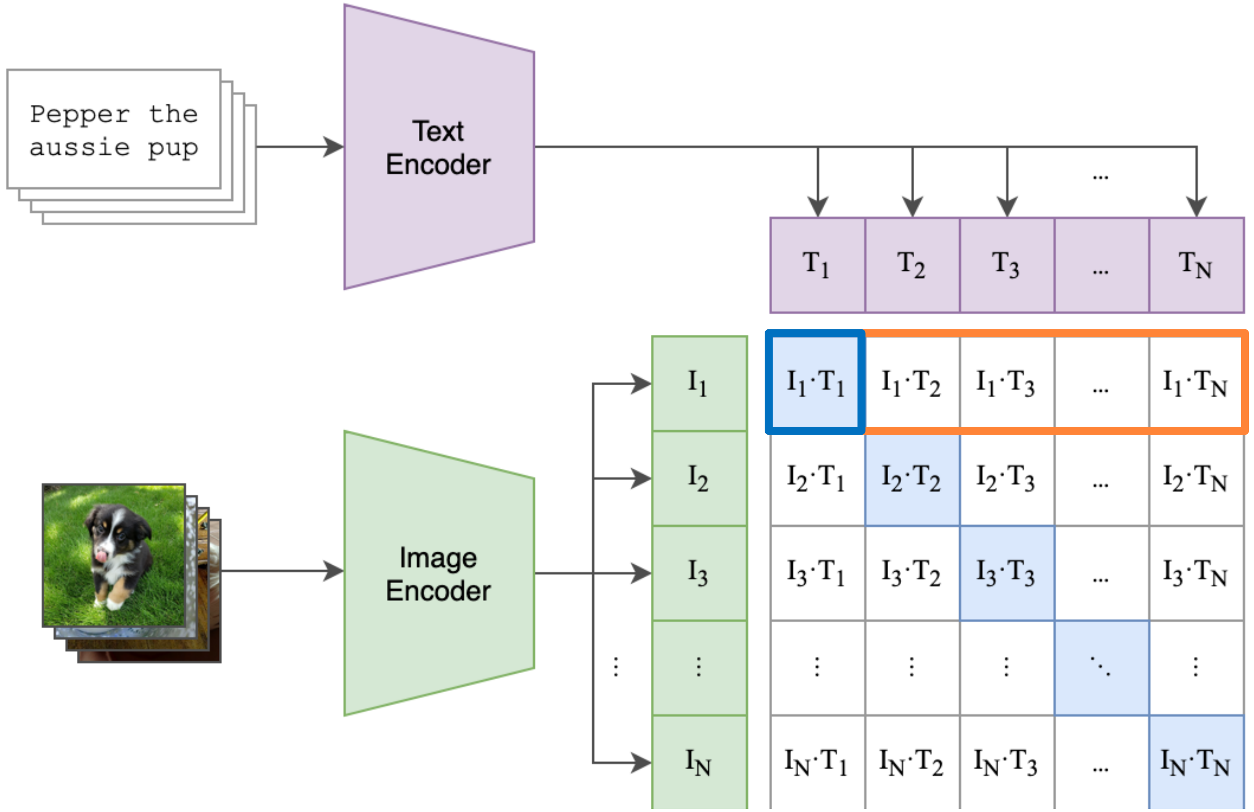


Image from: https://github.com/openai/CLIP

Contrastive Loss

- `total_loss = (image_to_text_loss + text_to_image_loss) / 2`

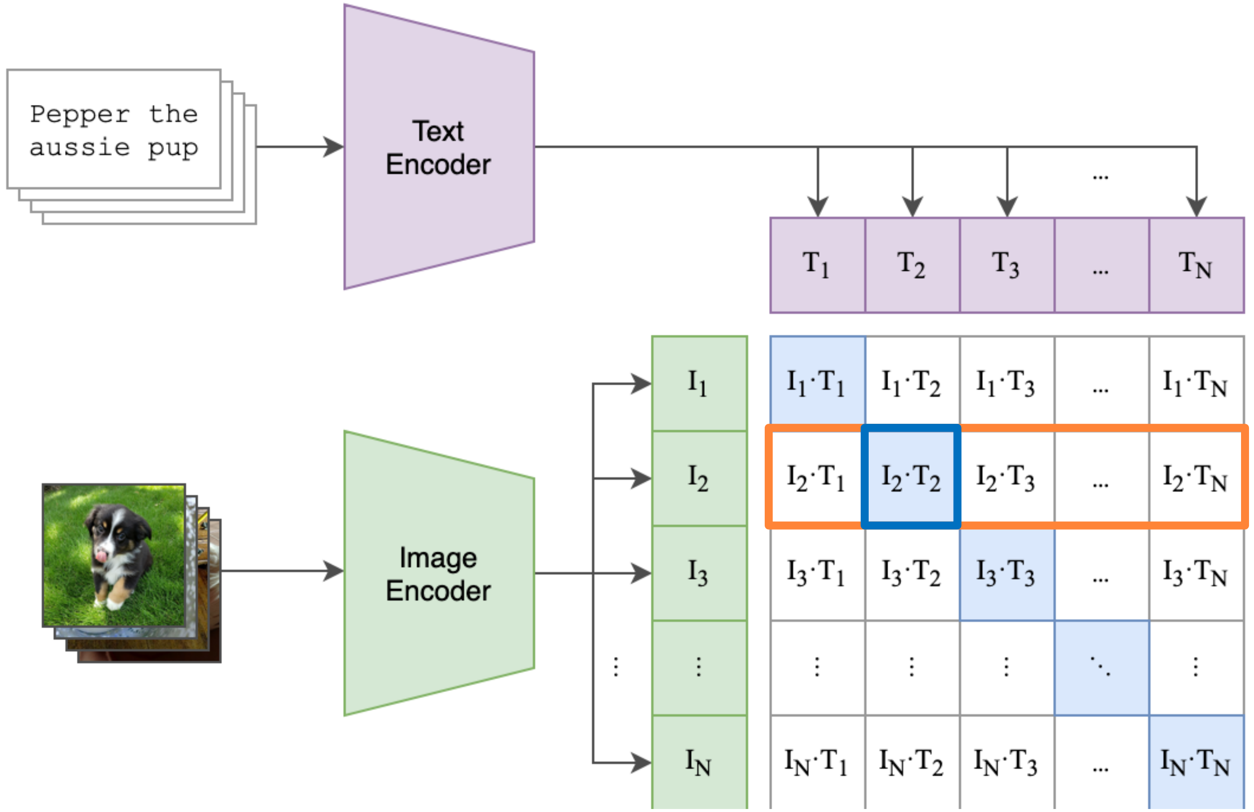


Image from: https://github.com/openai/CLIP

Contrastive Loss

- `total_loss = (image_to_text_loss + text_to_image_loss) / 2`

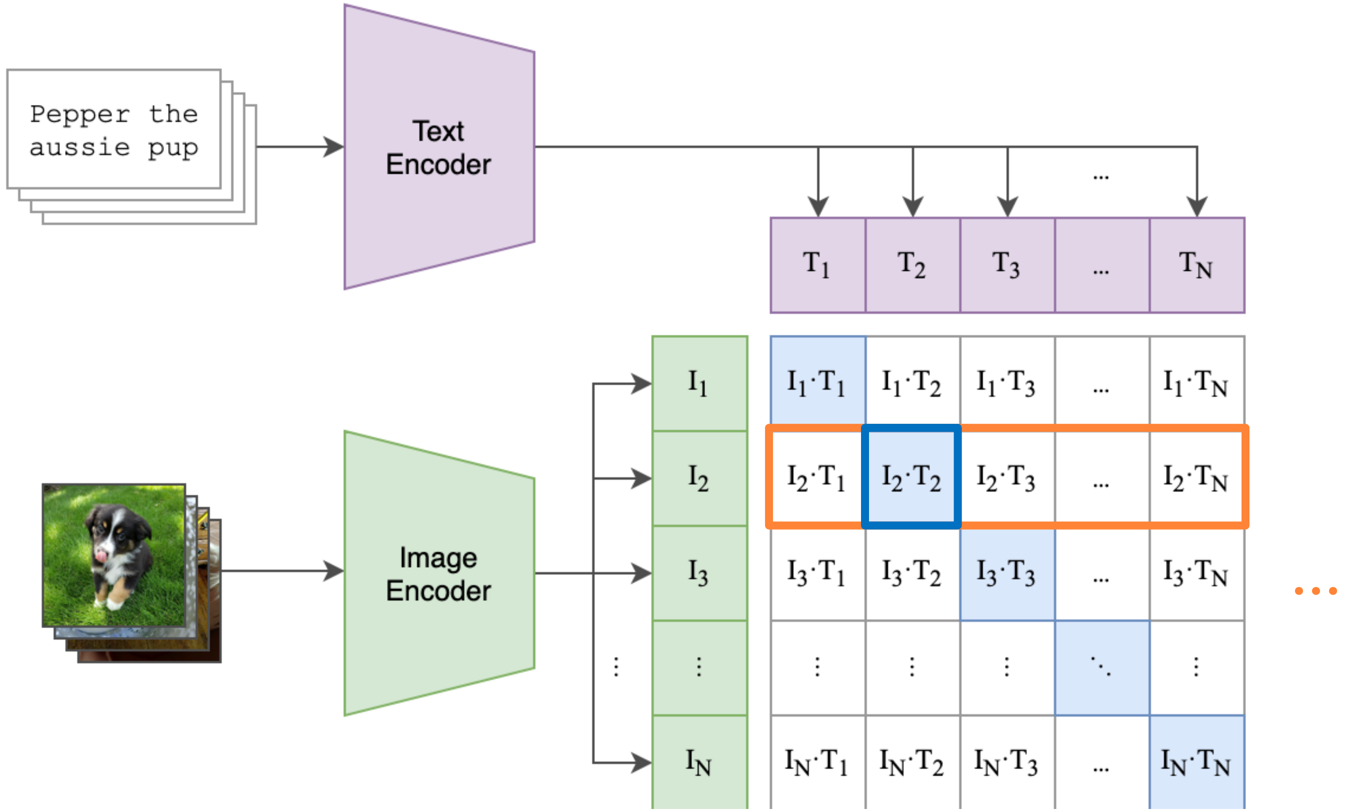


Image from: https://github.com/openai/CLIP

Contrastive Loss

- `total_loss = (image_to_text_loss + text_to_image_loss) / 2`

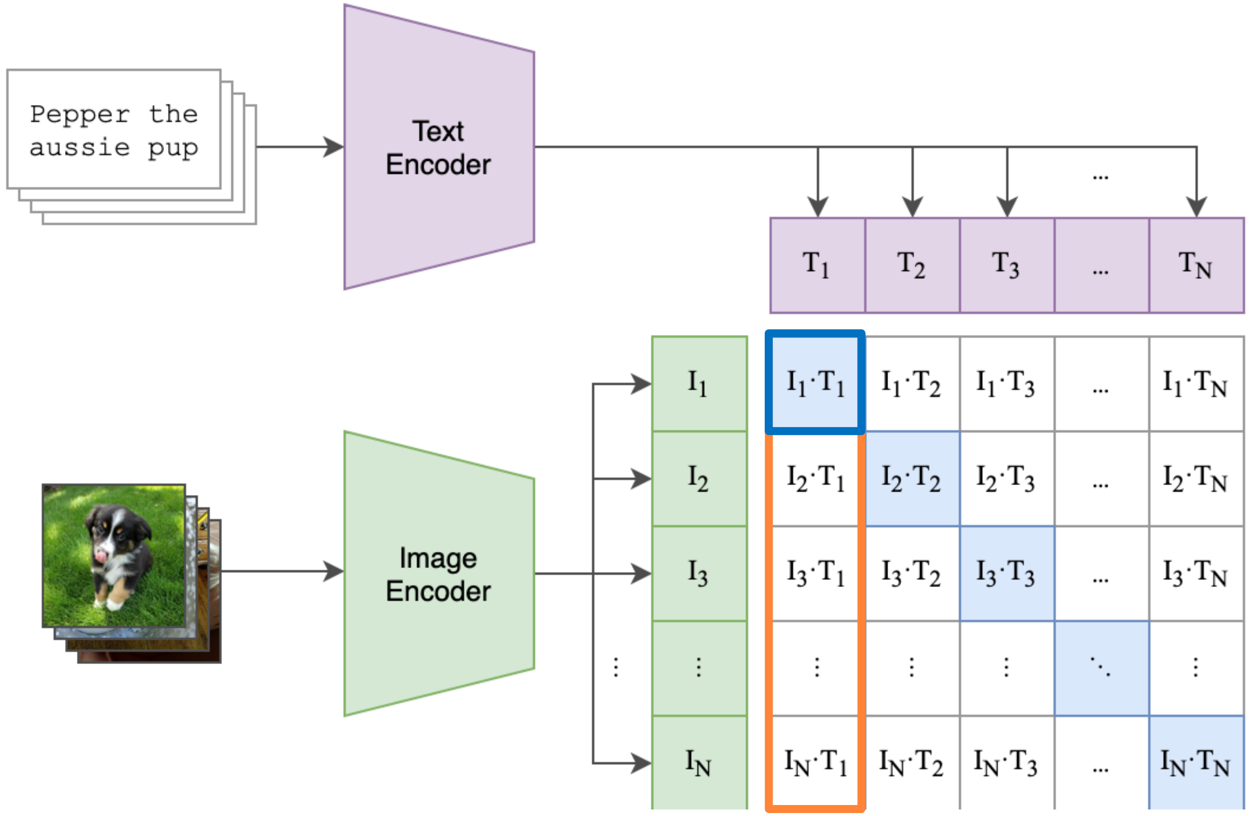


Image from: https://github.com/openai/CLIP

Contrastive Loss

- `total_loss = (image_to_text_loss + text_to_image_loss) / 2`

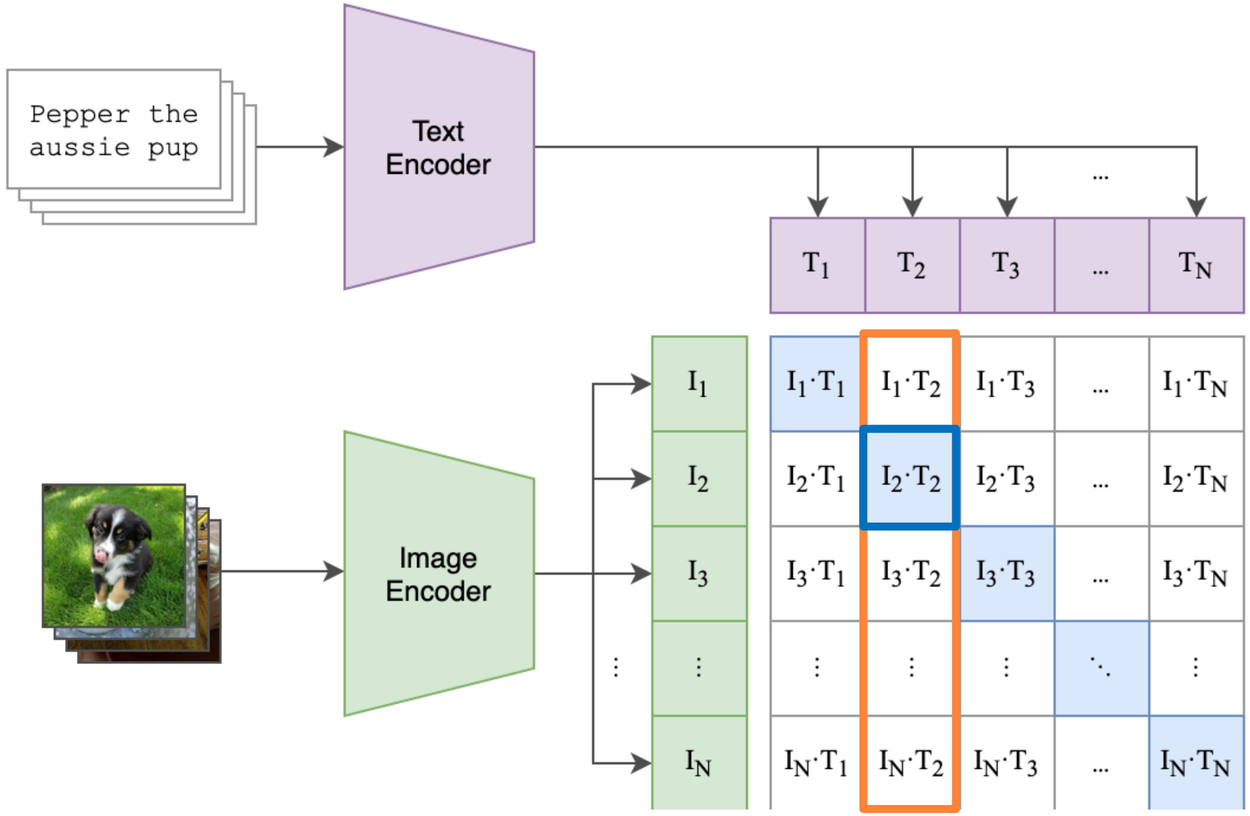


Image from: https://github.com/openai/CLIP

Contrastive Loss

- `total_loss = (image_to_text_loss + text_to_image_loss) / 2`

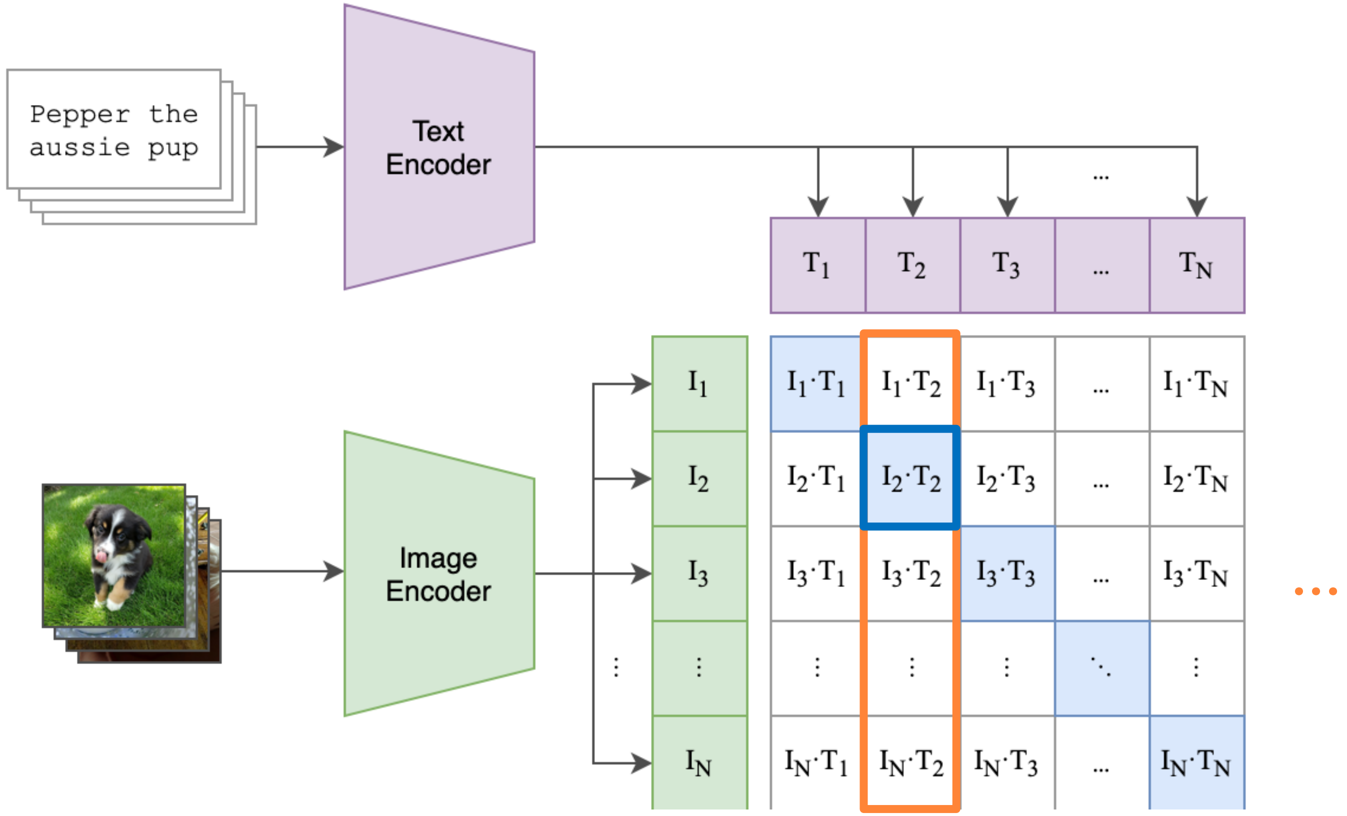


Image from: https://github.com/openai/CLIP

Contrastive Loss

- `total_loss = (image_to_text_loss + text_to_image_loss) / 2`

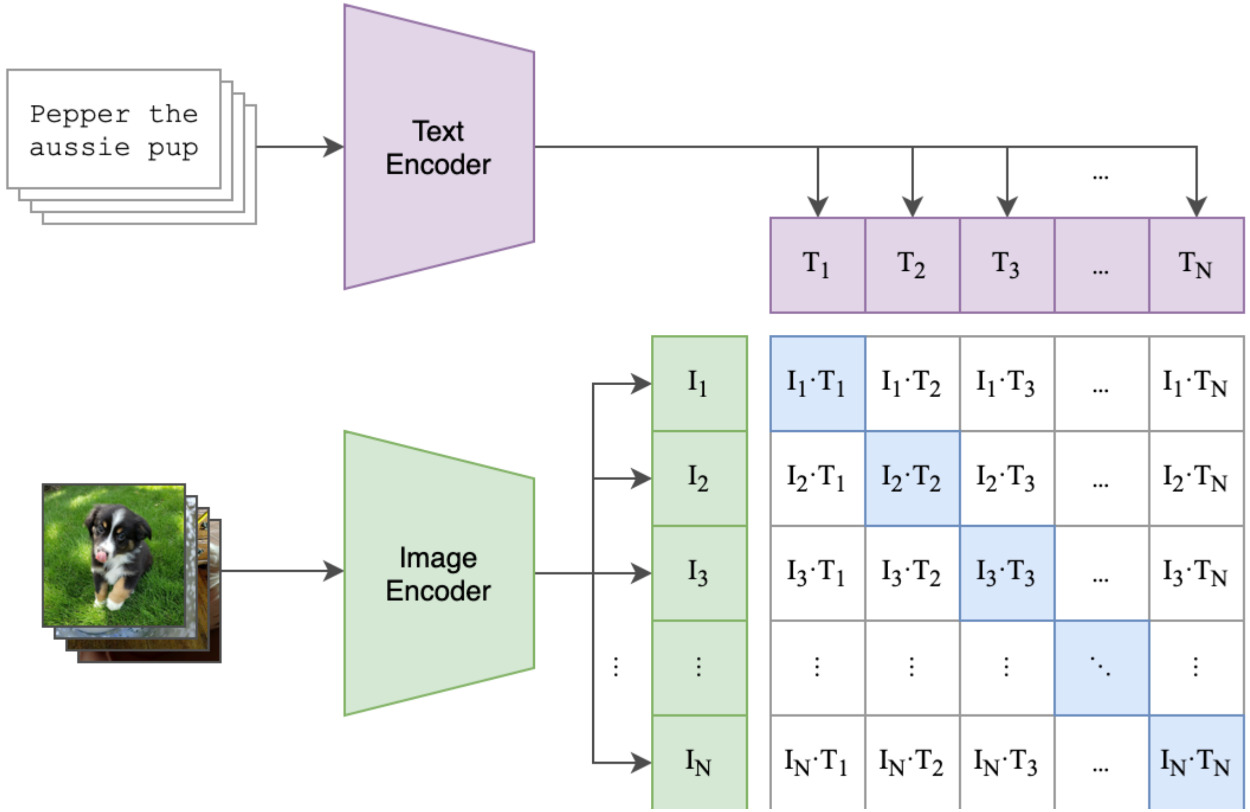
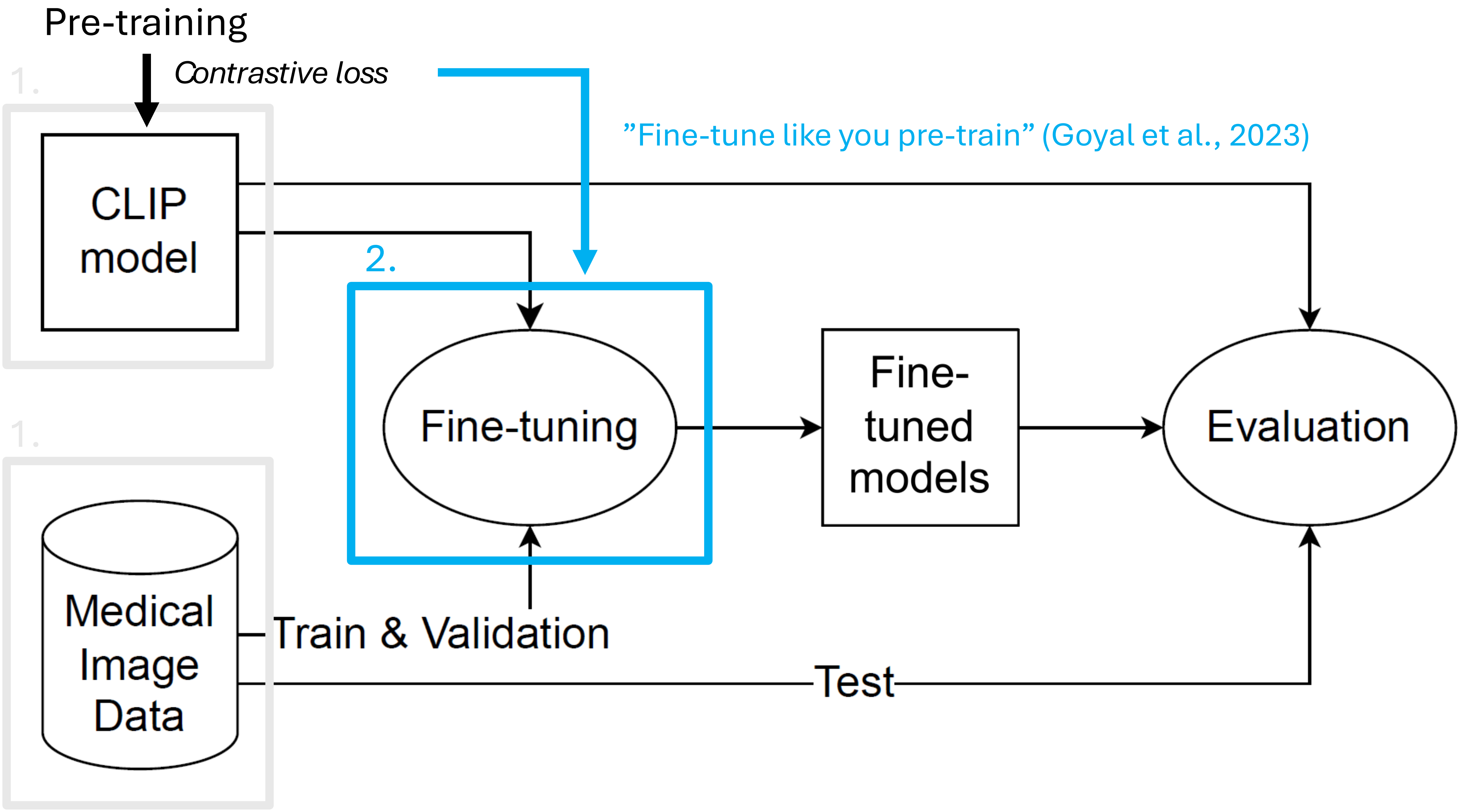
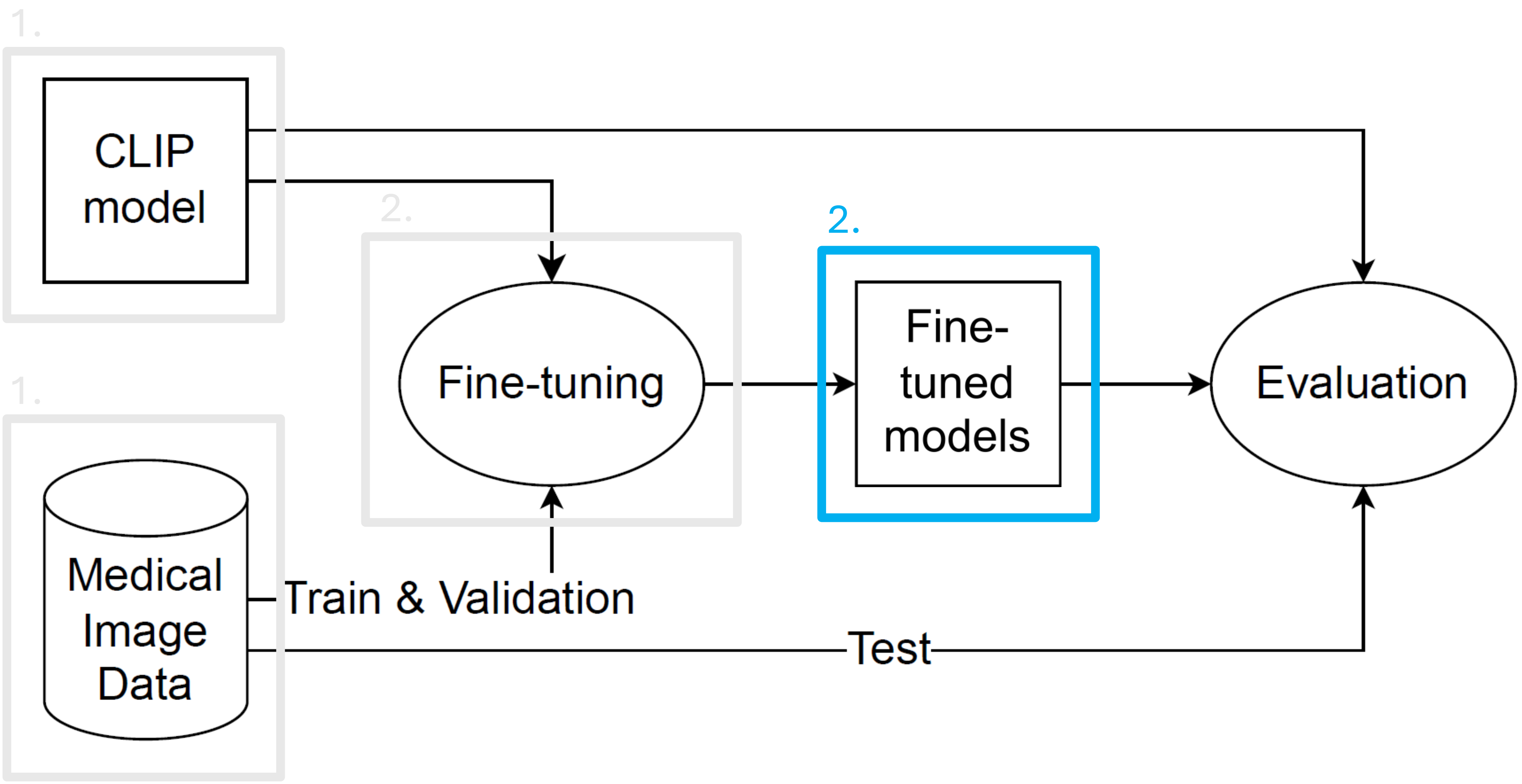
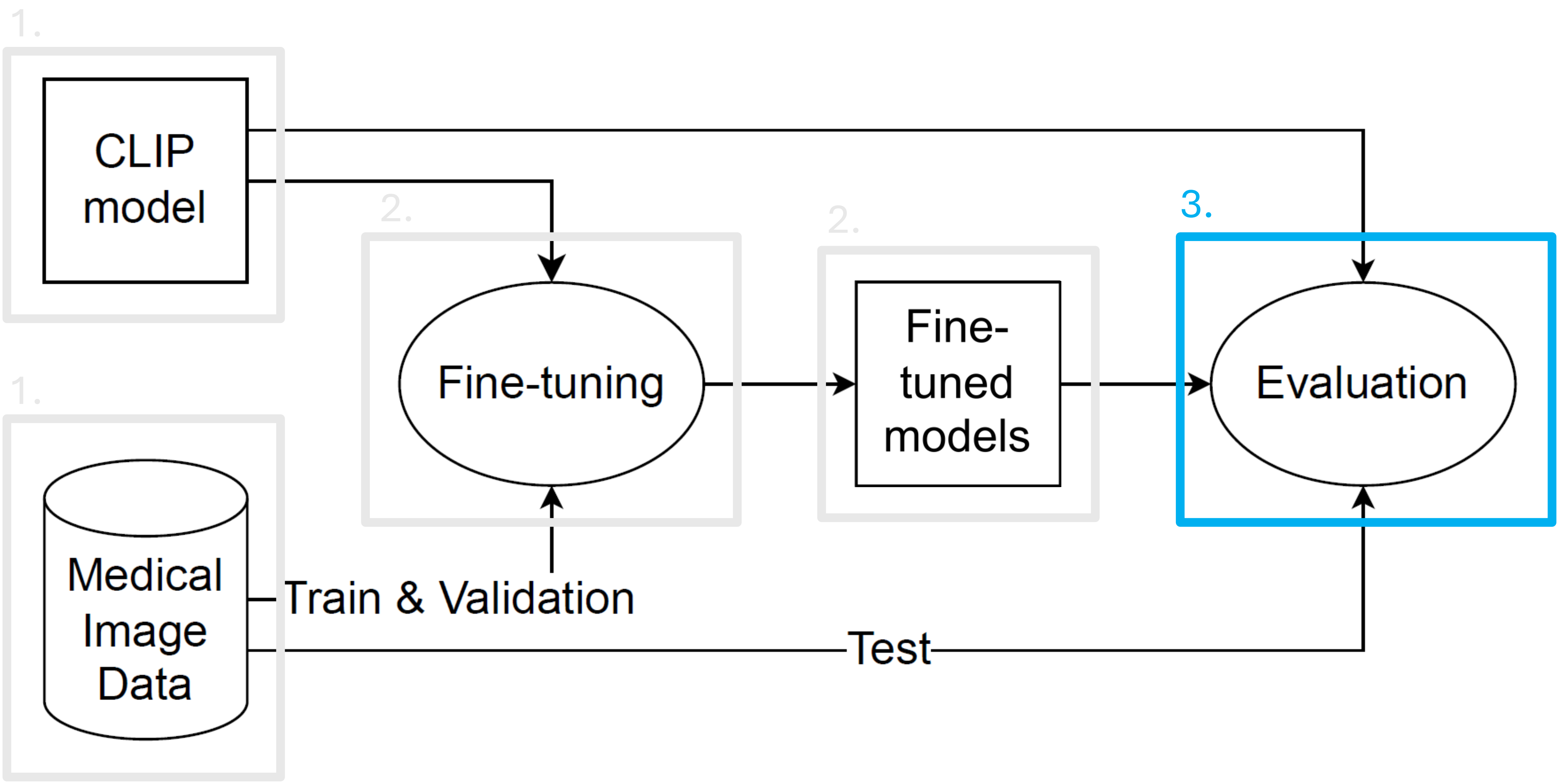


Image from: https://github.com/openai/CLIP


Pre-training
1.
Contrastive loss
”Fine-tune like you pre-train” (Goyal et al., 2023)
CLIP model
2.
Fine-tuning
Fine-tuned models
Evaluation
1.
Medical Image Data
Train & Validation
Test

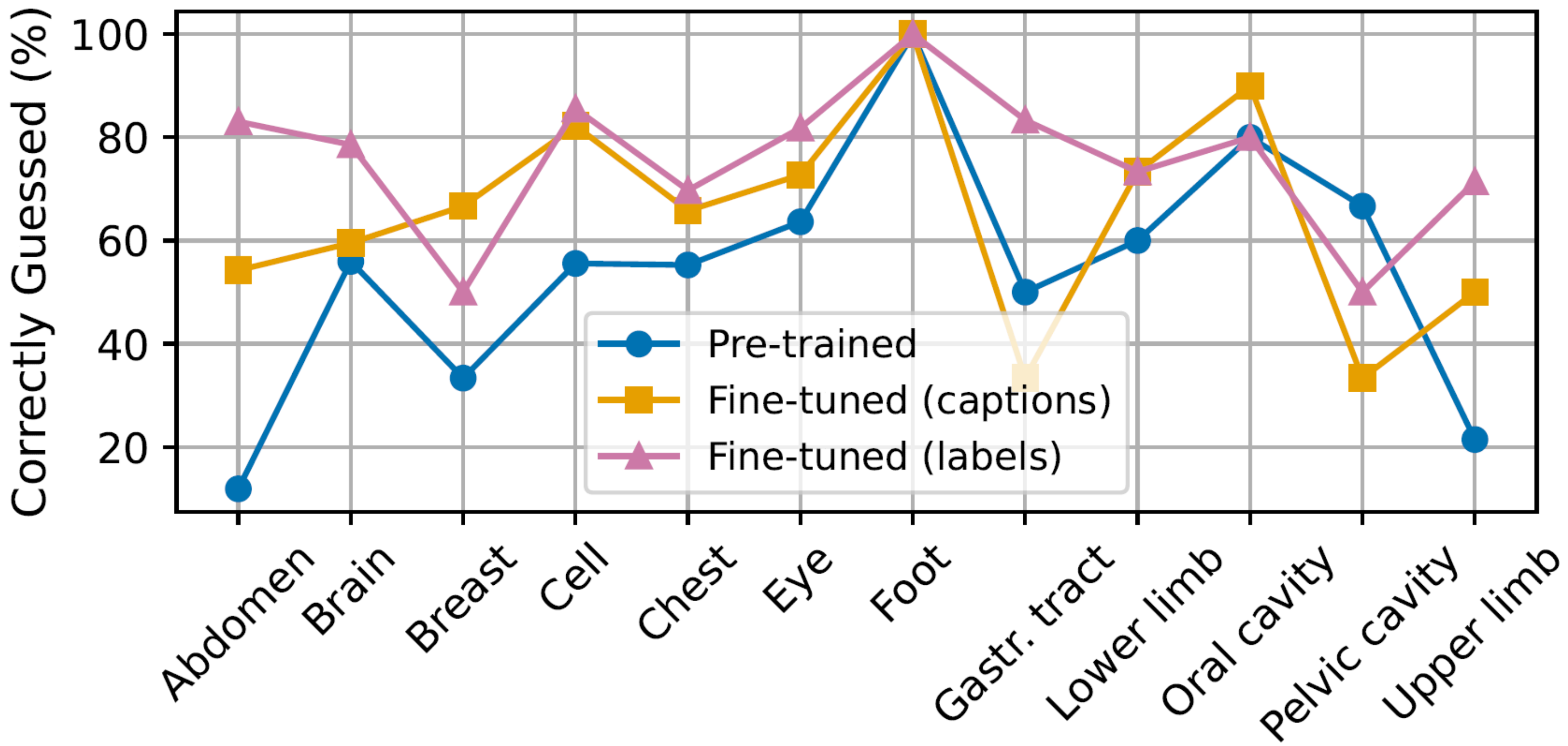

1.
CLIP model
1.
Medical Image Data
2.
Fine-tuning
2.
Fine-tuned models
Evaluation
Train & Validation
Test


1.
CLIP model
1.
Medical Image Data
2.
Fine-tuning
2.
Fine-tuned models
3.
Evaluation
Train & Validation
Test

Results & Analysis

Model	**Top-1** (%)	**Top-3** (%)
Pre-trained	49.4	78.9
Fine-tuned (captions)	64.3	83.1
Fine-tuned (labels)	76.9	92.1

Accuracy scores



Prediction accuracy per body part of each model during evaluation

Questions?