Strategy for Training

To implement your hybrid model on the CUBIC cluster, we are combining three specific frameworks: the **DINOv3 ViT-B** (Vision), the **CheXzero/OpenAI CLIP** (Text), and the **PLIP** (Training Strategy).

Here are the definitive parameters for our implementation:

1. Architectural Parameters

Because the vision and text encoders come from different families, the "bridge" between them is the most critical parameter.

Parameter	**Value**	**Description**
Vision Backbone	DINOv3 ViT-B/16	Pretrained ViT-Base with 16x16 patch size.
Vision Output Dim	**768**	The raw embedding size from DINOv3 ViT-B.
Text Backbone	CLIP-ViT-B/32	The text encoder used in the CheXzero/OpenAI repo.
Text Output Dim	**512**	The latent space dimension used by CLIP/CheXzero.
Projection Head	**Linear Layer**	A layer mapping the Vision **768** ->**512** to match the Text side.
Input Resolution	**224 x 224**	Standard input for both encoders; uses Bicubic interpolation.

2. Training Strategy Parameters (PLIP-Style)

We are shifting from the CheXzero "Epoch" approach to the PLIP "Step" approach.

Parameter	Value	Description
Total Training Steps	**25,000**	The total iteration count derived from the PLIP paper.
Batch Size	**64**	Strictly following the CheXzero batch size for memory stability.
Eval Interval	**500 steps**	Frequency of zero-shot validation on the external dataset.
Loss Function	**InfoNCE**	Symmetrical contrastive loss with learnable temperature.
Temperature	**0.07**	Initial logit scaling (often learnable in CLIP-based models).

3. Optimizer Hyperparameters (CheXzero-Style)

We are retaining the original CheXzero optimization settings to ensure the learning dynamics remain stable for medical images.

Parameter	Value	Description
Optimizer	**SGD**	Stochastic Gradient Descent.
Learning Rate	**10^{-4}**	The baseline rate from the CheXzero paper.

Momentum	**0.9**	Standard momentum for medical image convergence.
Weight Decay	**0**	CheXzero does not typically use weight decay in the final recipe.
Text Context Length	**512 tokens**	Uses the distilled context length for long medical reports.

4. Data Preprocessing & Augmentation

Parameter	**Value**	**Description**
Resize Strategy	**320 -> 224**	Images stored at 320px, then cropped/resized to 224px.
Normalization Mean	[0.481, 0.457, 0.408]	Specific to OpenAI CLIP weights.
Normalization Std	[0.268, 0.261, 0.275]	Specific to OpenAI CLIP weights.
Augmentations	Random Crop, Flip	Standard PLIP/CheXzero visual augmentations.