– Supplementary Material –

Foundation X: Integrating Classification, Localization, and Segmentation through Lock-Release Pretraining Strategy for Chest X-ray Analysis

A. Experiment details

Here, we discuss the setup of the training process for Foundation X. The backbone is Swin-B, initialized with Ark-6 [22] pretrained weights. For the classification task, linear layers serve as classification heads. For localization, we integrate the DINO localization approach [38], modify to handle multiple datasets with one localization encoder and multiple localization decoders. For segmentation, we use UperNet [36], modify to include multiple segmentation heads. We pretrain Foundation X on all 11 datasets (see Table 1) using a single A100 GPU, employing the Cyclic and Lock-Release pretraining strategies. We also employ the Student-Teacher learning paradigm, where the teacher model is an exact copy of the student model at the start. The teacher model is updated after each epoch using an exponential moving average (EMA) [32] with a momentum of 0.80. The configuration for Foundation X is detailed in Table 7.

Backbone	Swin-B† [20]
Classification Branch	Linear Layer‡
Localization Branch	DINO‡ [38]
Segmentation Branch	UperNet‡ [36]
Input Resolution	224 x 224
Optimizer	AdamW
Batch Size	24
Number of Workers	12
Backbone Learning Rate	1e-5
Localization Learning Rate	1e-4
Segmentation Learning Rate	1e-4
Learning Rate Scheduler	Step-decay
Evaluation Metric for Classification	AUC
Evaluation Metric for Localization	mAP40
Evaluation Metric for Segmentation	Dice

† Initialized with Ark-6 [22] pretrained weights.
‡ Initialized with random weights.

Table 7. Experiment settings for Foundation X.

B. Model Parameters

The Foundation X model consists of several key components, contributing to a total of approximately 173 million parameters (See Table 9). The backbone, responsible for feature extraction, accounts for 86.8 million parameters. The localization encoder adds 7.7 million parameters, while the localization decoders total 57.6 million, with each decoder contributing 9.6 million. The segmentation decoder comprises 20.9 million parameters. Although adding a dedicated localization decoder for each dataset increases the model size, only the relevant decoder is active during training, with the others, along with the classification and segmentation heads, remaining frozen. This approach keeps the computational load manageable and ensures efficient GPU utilization.

C. Lock-Release pretraining strategy

Foundation X effectively handles classification, localization, and segmentation tasks. The model leverages the Student-Teacher learning paradigm along with Cyclic and Lock-Release pretraining strategies, ensuring it retains general knowledge for all tasks while avoiding overfitting to any single task. In Table 8, we illustrate the Lock-Release pretraining strategy using the VinDr-CXR organ dataset for organ localization and segmentation. For this demonstration, we treat localization and segmentation of the heart, left lung, and right lung as separate tasks.

D. Cross-Dataset and Cross-Task learning analysis

The full figure illustrating the Cross-Dataset and Cross-Task learning analysis for all six datasets is included in this supplementary material (see Figure 4). This figure highlights the performance trends of Foundation X across various datasets under both focused and unfocused training scenarios, showcasing its ability to generalize and retain knowledge effectively through the Cyclic and Lock-Release pretraining strategies. We include plots for these six datasets because they contain multiple tasks, including classification, localization, and segmentation.

E. Ablation study

The ablation studies demonstrate the effectiveness of the Student-Teacher learning paradigm, Cyclic and Lock-Release pretraining strategies across various tasks. Foundation X-L (see Table 11), trained on six disease localization tasks, generally outperforms the baseline model Swin-B + DINO. Similarly, Foundation X-S (see Table 12), trained on three disease segmentation datasets, consistently surpasses the baseline model Swin-B + UperNet. Additionally, Foundation X-CL (see Table 13), which handles both classification and localization tasks, and Foundation X-LS (see Table 14), which integrates localization and segmentation tasks, both show superior performance compared to their

Algorithm 1: A round of Foundation X's Cyclic Lock-Release pretraining

Data: Datasets: $\mathcal{D} = \{D_1, D_2, ..., D_n\}$; Sample: image-label pair $(x, y) \in \mathcal{D}_i$
Functions: Data augmentation: ε; Dataset/task-specific losses: $\{\mathcal{L}D1(\cdot,\cdot), \mathcal{L}D2(\cdot,\cdot), ..., \mathcal{L}Dn(\cdot,\cdot)\}$; Consistency loss: $\{\mathcal{L}const(\cdot,\cdot)\}$; Loss update by AdamW optimizer: $Update_{adamw}(\cdot,\cdot)$
Trainable Parameters: Student's encoder, localization encoder, segmentation decoder: e_s, $LocEnc_s$, $SegDec_s$; Classification heads $\mathcal{C} = \{C_1, C_2, ..., C_n\}$; Localization decoders $\mathcal{L} = \{L_1, L_2, ..., L_n\}$; Segmentation heads $\mathcal{S} = \{S_1, S_2, ..., S_n\}$;
Stop Gradient: Teacher's encoder, localization encoder, segmentation decoder: e_t, $LocEnc_t$, $SegDec_t$;
Hyperparameters: Momentum: λ

$\{e_t, LocEnc_t, SegDec_t\} \leftarrow \{e_s, LocEnc_s, SegDec_s\}$ `// initialize teacher with student's parameters`
for D_i ***in*** $D_1, D_2, ..., D_n$ **do**
  `/* train student for one epoch */`
  for (x, y) ***in*** D_i **do**
    $x' = \varepsilon(x)$
    if D_i *has Classification Annotation* **then**
      for $j \leftarrow 1$ **to** *2* **do**
        if j = *1* **then**
          $Freeze\ \{e_s\}$ `// Lock mode on, using a random half of the dataset`
        else
          $Unfreeze\ \{e_s\}$ `// Release mode on, using full dataset`
        $emb_t, emb_s = e_t(x'), e_s(x')$
        $pred = C_i(emb_s)$
        $Loss = \mathcal{L}Di(pred, y) + \mathcal{L}const1(emb_t, emb_s)$
        $Update(\{e_s, p_s, C_i\}, Loss)$
    if D_i *has Localization Annotation* **then**
      for $j \leftarrow 1$ **to** *2* **do**
        if j = *1* **then**
          $Freeze\ \{e_s, LocEnc_s\}$ `// Lock mode on, using a random half of the dataset`
        else
          $Unfreeze\ \{e_s, LocEnc_s\}$ `// Release mode on, using full dataset`
        $emb_t, emb_s = e_t(x'), e_s(x')$
        $embLocEnc_s, embLocEnc_t = LocEnc_s(emb_s), LocEnc_t(emb_t)$
        $pred = L_i(embLocEnc_s)$
        $Loss = \mathcal{L}Di(pred, y) + \mathcal{L}const(emb_t, emb_s) + \mathcal{L}const(embLocEnc_t, embLocEnc_s)$
        $Update(\{e_s, LocEnc_s, L_i\}, Loss)$
    if D_i *has Segmentation Annotation* **then**
      for $j \leftarrow 1$ **to** *2* **do**
        if j = *1* **then**
          $Freeze\ \{e_s, SegDec_s\}$ `// Lock mode on, using a random half of the dataset`
        else
          $Unfreeze\ \{e_s, SegDec_s\}$ `// Release mode on, using full dataset`
        $emb_t, emb_s = e_t(x'), e_s(x')$
        $embSegDec_s, embSegDec_t = SegDec_s(emb_s), SegDec_t(emb_t)$
        $pred = S_i(embSegDec_s)$
        $Loss = \mathcal{L}Di(pred, y) + \mathcal{L}const(emb_t, emb_s) + \mathcal{L}const(embSegDec_t, embSegDec_s)$
        $Update(\{e_s, SegDec_s, S_i\}, Loss)$
  `/* Update teacher by student's parameters via epoch-wise EMA */`
  $\{e_t, LocEnc_t, SegDec_t\} \leftarrow \lambda\{e_t, LocEnc_t, SegDec_t\} + (1-\lambda)\{e_s, LocEnc_s, SegDec_s\}$

	Epoch #	Data Size	Backbone	Loc.Enc	Loc.Dec	Seg.Dec	Seg.Head	Mode	Training Task
Cycle 1	1	Half	F	F	T	-	-	Lock	Localization of Heart
	2	Full	T	T	T	-	-	Release	Localization of Heart
	3	Half	F	F	T	-	-	Lock	Localization of Left Lung
	4	Full	T	T	T	-	-	Release	Localization of Left Lung
	5	Half	F	F	T	-	-	Lock	Localization of Right Lung
	6	Full	T	T	T	-	-	Release	Localization of Right Lung
	7	Half	F	-	-	F	T	Lock	Segmentation of Heart
	8	Full	T	-	-	T	T	Release	Segmentation of Heart
	9	Half	F	-	-	F	T	Lock	Segmentation of Left Lung
	10	Full	T	-	-	T	T	Release	Segmentation of Left Lung
	11	Half	F	-	-	F	T	Lock	Segmentation of Right Lung
	12	Full	T	-	-	T	T	Release	Segmentation of Right Lung

Table 8. Demonstrating the Lock-Release pretraining strategy for organ localization and segmentation using the VinDr-CXR dataset. The model completes a single cycle when it goes through all tasks once (from epoch #1 to #12). 'F' denotes a frozen component, and 'T' denotes a trainable component. In Lock mode, the model is trained using half of the dataset, while in Release mode, it is trained using the full dataset. After each epoch in Release mode, the model is tested on the localization and segmentation of the heart, left lung, and right lung.

Component	**Parameters**
Backbone	86,751,673 [+]
Classification Heads	70,725 [+]
Localization Encoder	7,693,056 [+]
Localization Decoders	57,653,868 [+]
Each Localization Decoder	*9,608,978*
Segmentation Decoder	20,894,464 [+]
Segmentation Heads	20,754 [+]
Total	**173,084,540**

Table 9. Parameter distribution across the key components of the Foundation X model, trained on 11 datasets and 20 tasks.

respective baseline methods in most cases.

The ablation study presented in Table 10 highlights the impact of incorporating the Lock-Release pretraining strategy and the Student-Teacher learning paradigm on the performance of the Foundation X model. The results demonstrate that when both components are enabled, the model achieves the best performance across all evaluated VinDr-CXR organs (Heart, Left Lung, and Right Lung) localization. Specifically, the combination of Lock-Release and Student-Teacher results in the highest mAP, with scores of 88.39% for Heart, 95.78% for Left Lung, and 96.78% for Right Lung. These findings suggest that each component complements the other, with the Lock-Release strategy preventing task-specific overfitting and the Student-Teacher paradigm ensuring stable learning by reducing drastic model shifts. Together, these strategies create a synergistic effect that enhances the model's generalization and overall performance, outperforming configurations where one or both components are disabled. This highlights the importance of integrating both the Lock-Release strategy and the Student-Teacher paradigm to maximize the effectiveness of our approach.

Lock-Release	Student-Teacher	Heart	Left Lung	Right Lung
✗	✗	85.45	93.63	94.47
✗	✓	86.41	94.39	94.95
✓	✗	87.50	95.37	96.44
✓	✓	**88.39**	**95.78**	**96.78**

Table 10. Ablation study is conducted on the VinDr-CXR organ localization dataset. We evaluate the model with and without the Lock-Release pretraining strategy, as well as with and without the Student-Teacher model. The results demonstrate that the Foundation X model achieves comparatively better performance when both the Lock-Release pretraining strategy and the Student-Teacher learning paradigm are employed.

F. Acknowledgements

This research was partially supported by ASU and Mayo Clinic through a Seed Grant and an Innovation Grant, as well as by the NIH under Award Number R01HL128785. The authors are solely responsible for the content, which does not necessarily reflect the official views of the NIH. This work also utilized GPUs provided by ASU Research Computing, Bridges-2 at the Pittsburgh Supercomputing Center (allocated under BCS190015), and Anvil at Purdue University (allocated under MED220025). These resources are supported by the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, funded by the National Science Foundation under grants #2138259, #2138286, #2138307, #2137603, and #2138296. We also extend our gratitude to Anirudh Kaniyar Narayana Iyengar for his contributions to collecting localization data, preparing bounding boxes in COCO format, and developing some of the dataloaders. Finally, the content of this paper is covered by patents pending.

Dataset	Baseline Loc.† [mAP40%]	Foundation X-L [mAP40%]
TBX11K	**78.10**	77.77 ↓ 0.33
NODE21	37.50	**42.79** ↑ 5.29
CANDID-PTX	50.90	**53.75** ↑ 2.85
RSNA Pneumonia	21.70	**29.37** ↑ 7.67
ChestX-Det	38.00	**40.13** ↑ 2.13
SIIM-ACR	28.00	**36.20** ↑ 8.20

† Swin-B version of DINO where the backbone is initialized with Ark-6 pretrained weights.

Table 11. We train Foundation X-L on six disease localization tasks utilizing Cyclic and Lock-Release pretraining strategies and compare its performance with the baseline model, DINO [38]. In most cases, Foundation X-L outperforms the baseline across the datasets during pretraining.

Dataset	Baseline Seg.† [Dice%]	Foundation X-S [Dice%]
CANDID-PTX	86.36	**89.58** ↑ 3.23
ChestX-Det	79.33	**83.46** ↑ 4.13
SIIM-ACR	81.92	**83.83** ↑ 1.91

† Swin-B version of UperNet where the backbone is initialized with Ark-6 pretrained weights.

Table 12. We train Foundation X-S on three disease segmentation datasets using the Cyclic and Lock-Release pretraining strategies and compare its performance with the baseline model, UperNet [36]. In all cases, Foundation X-S outperforms the baseline across the datasets during pretraining.

Dataset	Baseline Cls. [AUC%]	Baseline Loc. [mAP40%]	Foundation X-CL [AUC%]	Foundation X-CL [mAP40%]
TBX11K	99.89±0.06	**78.10**	**99.96** ↑ 0.07	72.56 ↓ 5.54
NODE21	99.35±0.45	37.50	**99.68** ↑ 0.33	**47.54** ↑ 10.04
CANDID-PTX	72.61±0.57	50.90	**74.00** ↑ 1.39	**51.61** ↑ 0.71
RSNA Pneumonia	88.87±0.21	21.70	**96.57** ↑ 7.70	**26.08** ↑ 4.38
ChestX-Det	**88.17**±0.33	**38.00**	81.82 ↓ 6.35	37.03 ↓ 0.97
SIIM-ACR	95.01±0.16	28.00	**95.19** ↑ 0.18	**34.98** ↑ 6.98

Table 13. We train Foundation X-CL on six disease datasets, each containing both classification and localization annotations, using our Cyclic and Lock-Release pretraining strategies. The table demonstrates that, in most cases, Foundation X-CL outperforms the baseline methods during pretraining.

Dataset	Baseline Loc. [mAP40%]	Baseline Seg. [Dice%]	Foundation X-LS [mAP40%]	Foundation X-LS [Dice%]
TBX11K	**78.10**	-	73.03 ↓ 5.07	-
NODE21	37.50	-	**46.09** ↑ 8.59	-
CANDID-PTX	50.90	86.36	**53.01** ↑ 2.11	**89.47** ↑ 3.11
RSNA Pneumonia	21.70	-	**27.80** ↑ 6.10	-
ChestX-Det	38.00	**79.33**	**39.22** ↑ 1.22	70.90 ↓ 8.43
SIIM-ACR	28.00	81.92	**36.63** ↑ 8.63	**84.25** ↑ 2.33

Table 14. We train Foundation X-LS on six disease localization and three disease segmentation datasets, using our Cyclic and Lock-Release pretraining strategies. The table demonstrates that, in most cases, Foundation X-LS outperforms the baseline methods during pretraining.

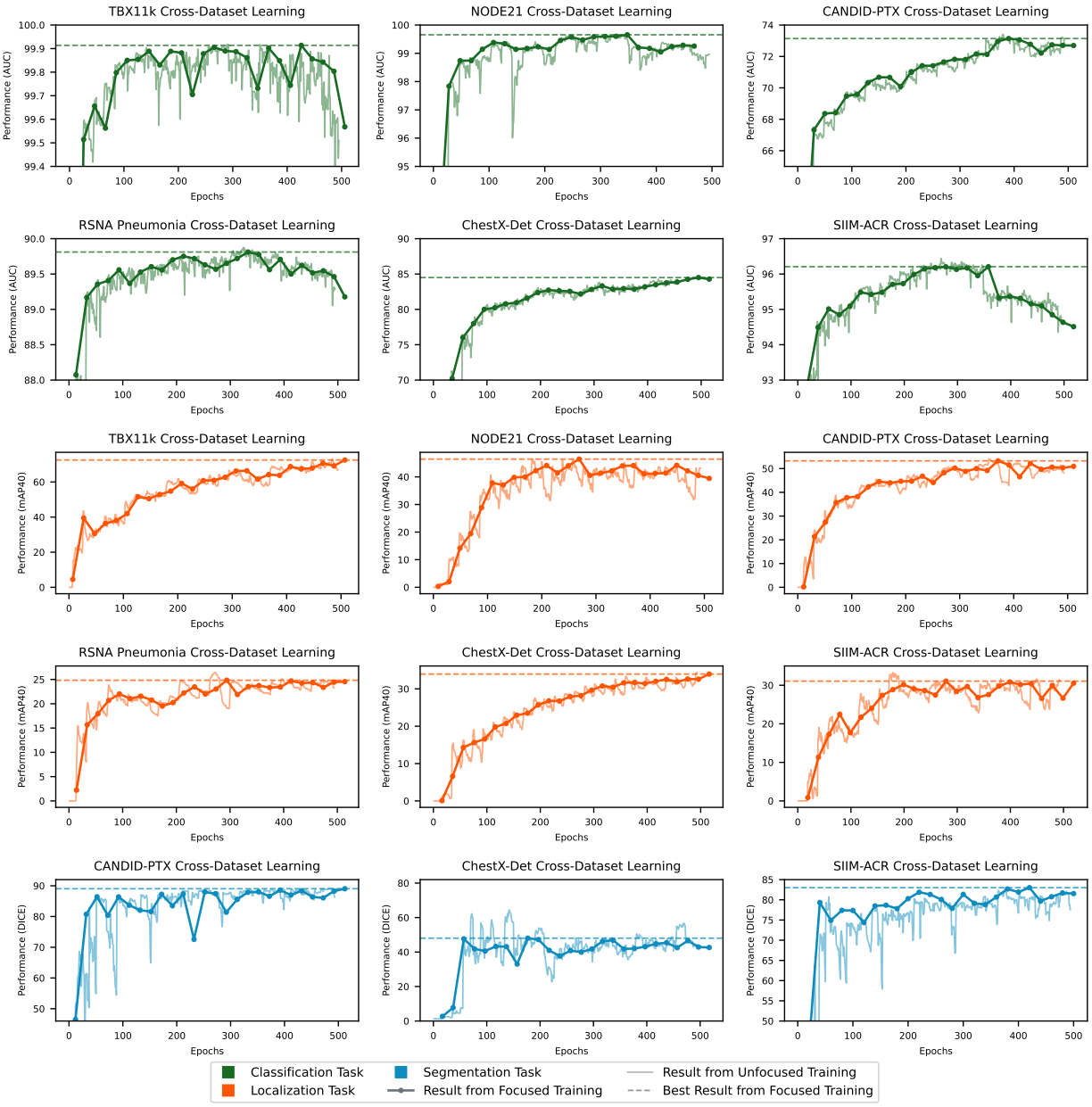

Figure 4. Cross-Dataset & Cross-Task Learning Analysis. The figure demonstrates the performance trends of Foundation X across multiple datasets for both focused and unfocused training scenarios. Focused training refers to scenarios where the model is explicitly trained on the specific dataset being evaluated, while unfocused training refers to scenarios where the model is trained on other datasets and not directly on the dataset being evaluated. The green, orange, and blue lines represent the classification, localization, and segmentation tasks, respectively. Dark-colored lines indicate the testing results during focused training, where the model is explicitly trained on the specific dataset. Light-colored lines show the testing results during unfocused training, where the model is trained on other datasets but tested on the specific dataset. Dashed lines represent the best testing results achieved from focused training for each specific dataset. The results indicate that, during unfocused training, initial performance dips are common as the model is not explicitly trained on the specific dataset. However, performance improves over time, demonstrating the model's ability to generalize effectively and retain knowledge due to the Cyclic and Lock-Release pretraining strategies. In all cases, the unfocused training results do not drift away from the task, highlighting the model's efficient generalization and knowledge retention. Additionally, in some instances, unfocused training achieves even better performance than focused training, showcasing the advantages of cross-task and cross-dataset learning in enhancing the overall capabilities of Foundation X.

Task	Dataset	Official Split	Train Split	Val Split	Test Split	Expert Labels
CLS	CheXpert [9]	✓	223415	234	-	No finding, Enlarged Cardiomediastinum, Cardiomegaly, Lung Opacity, Lung Lesion, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture, Support Devices
	NIH ChestX-ray14 [35]	✓	75312	11212	25596	Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening, Hernia
	VinDr-CXR [25]	✓	15000	-	3000	PE, Lung Tumor, Pneumonia, Tuberculosis, Other diseases, No finding
	NIH Shenzhen CXR [11]	✗	463	65	134	Tuberculosis
	MIMIC-II [12]	✓	368878	2991	5159	No finding, Enlarged Cardiomediastinum, Cardiomegaly, Lung Opacity, Lung Lesion, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture, Support Devices
CLS, LOC	TBX11k [19]	✓	6600	1800	-	Tuberculosis
	NODE21 [31]	✗	4178	-	1046	Nodule
	RSNA Pneumonia [26]	✓	21295	2680	2709	CLS: No lung opacity/Not normal, Normal, Lung Opacity; LOC: Pneumonia
CLS, LOC, SEG	CANDID-PTX [3]	✓	13748	1964	3928	Pneumothorax
	ChestX-Det [16]	✓	3025	-	553	Atelectasis, Calcification, Cardiomegaly, Consolidation, Diffuse Nodule, Effusion, Emphysema, Fibrosis, Fracture, Mass, Nodule, Pleural Thickening, Pneumothorax
	SIIM-ACR [1]	✗	9607	1068	1372	Pneumothorax
	Total CLS images		**741521**	**22014**	**43997**	
	Total LOC images		**58453**	**7512**	**9608**	
	Total SEG images		**26380**	**3032**	**5853**	
LOC FT	VinDr-CXR [25]	✓	15000	-	3000	Aortic enlargement, Atelectasis, Calcification, Cardiomegaly, Consolidation, ILD, Infiltration, Lung Opacity, Nodule/Mass, Other lesion, Pleural effusion, Pleural thickening, Pneumothorax, Pulmonary fibrosis
SEG FT	CheXmask VinDr-CXR [5]	✓	15000	-	3000	Heart, Left Lung, Right Lung
	VinDr-RibCXR [24]	✓	196	-	49	20 Ribs
	NIH Montgomery [11]	✗	92	15	31	Lung
	JSRT [33]	✗	173	25	49	Heart, Lung, Clavicle

Table 15. Foundation X was pretrained on the above 11 classification datasets, 6 localization datasets, and 3 segmentation datasets. CLS stands for classification task, LOC stands for localization task, SEG stands for segmentation task. "CLS, LOC" denotes the datasets used for classification and localization tasks. "CLS, LOC, SEG" denotes the datasets used for classification, localization, and segmentation tasks. "LOC FT" and "SEG FT" denotes the datasets used only during the finetuning of the localization and segmentation task, respectively.