Detection-Gated Cavity Segmentation on Chest X-rays with Vision–Language Pretraining

Seongeun Lee(✉), Hannah Yun, Taejin Jeong, Mingyeong Jeon, Junhyun Park, Hyunwoong Kim, and Jonggwon Park

Team MultimodalAI, DEEPNOID Inc., Seoul, Republic of Korea
selee@deepnoid.com, hnyun@deepnoid.com, tjjeong@deepnoid.com, christy4526@deepnoid.com, jh_park@deepnoid.com, hwkim@deepnoid.com, jgpark@deepnoid.com

Abstract. We describe our submission to the Cavity Detection and Segmentation task (Task 1) of the TREAT-MMTB-2026 challenge [1], which targets binary pulmonary-cavity segmentation on chest X-ray (CXR) evaluated by a composite of case-level detection (0.7) and region Dice (0.3). Detection carries most of the score, yet only 444 mask-labeled images are available. We therefore train the case-level decision on a much larger, publicly available set of case-labeled CXR and keep it separate at inference: a cavity-presence classifier determines the case label, and the segmentation model produces a mask for predicted positives and an empty mask otherwise. Both encoders are adapted to CXR by vision–language pretraining (VLP) on MIMIC-CXR, whose radiology reports we enrich with cavity- and tuberculosis-related sentences from LLM labeling. Rather than using the pretrained encoder directly for segmentation, we first fine-tune it as a cavity-presence classifier on 13,174 case-labeled images, so that it has learned to detect cavities at scale before ever seeing a mask. Pairing a single unaugmented detector with a four-member segmentation ensemble trained on the 444 masks alone scores a composite of 0.6265 on the previously released test split and 0.6001 on the organizers' held-out external evaluation. Between the two sets, detection transferred well (0.757 → 0.811) while Dice score decreased (0.323 → 0.108). Our model placed third in Task 1.

Keywords: Tuberculosis · Chest X-ray · Cavity segmentation · Vision–language pretraining.

1 Introduction

Cavitary lesions on chest X-ray (CXR) are key indicators of tuberculosis infectiousness, so their automatic detection and localization are clinically useful. The challenge poses this problem as binary cavity segmentation, scored by

$$S = 0.7\, A_{\text{det}} + 0.3\, \overline{\text{Dice}}, \tag{1}$$

where A_{det} is case-level presence accuracy and $\overline{\text{Dice}}$ is the Dice score averaged over cases, at the original resolution excluding cases where both ground truth

Fig. 1. The three channels of the [*eq*, CLAHE, *base*] input stack on a CXR provided by the challenge. The encoder consumes the three renderings as one three-channel image.

and prediction are empty. The released mask-labeled set is limited (444 train / 111 test, of which 58 are cavity-positive and 53 negative), and many cases are cavity-negative. We therefore chose to train detection on a much larger set of case-labeled images and use the limited mask dataset for segmentation.

Since detection carries the larger weight in the composite score, our model design prioritizes case-level classification.

Our contributions are: (i) a vision–language pretraining (VLP) stage on a large report-paired CXR dataset that adapts a general-domain vision backbone without requiring task-specific labels; (ii) a two-model design that delegates the case-level decision to a classifier trained on over 30× more labels than the mask set; (iii) a staged encoder transfer that first fine-tunes the segmentation encoder as a cavity-presence classifier before pairing it with the segmentation decoder.

2 Method

2.1 Input Representation

A three-channel stack of different contrast views is expected to carry more information than the raw grayscale alone and to be more robust to variability across acquisition devices and imaging conditions. To this end, from each min–max-normalized, MONOCHROME1-inverted grayscale image, denoted *base*, we build the three-channel stack [*eq*, CLAHE, *base*], where *eq* is a globally histogram-equalized *base* and CLAHE is contrast-limited adaptive histogram equalization [15] (clip 1.0, 8×8 tiles) (Fig. 1). Images keep their aspect ratio and are zero-padded to 1024×1024. The stack and the geometry are identical in pretraining, fine-tuning, and inference, and both the classification fine-tuning and the segmentation training share a top-left padding convention. Dice is scored after resizing predictions back to native resolution.

2.2 Vision–Language Pretraining

We adapt DINOv3 [3] backbones to CXR with a vision–language objective on MIMIC-CXR [5], following GLINT [2]. Each report is decomposed into finding-

level sentences. Because the target findings are rare in free text, we additionally label each report with an LLM (Qwen3.6-35B-A3B [6]) for cavity and tuberculosis status. We encode sentences with MPNet [7] and align them to image features with a multi-positive InfoNCE loss [8]. We pretrain two vision encoders on frontal studies only, both at 1024: a ViT-L/16 for detection and a ConvNeXt-L [4] for segmentation, both initialized from DINOv3 [3] weights. Both encoders are trained without the dense feature regularization (DFR) branch of GLINT, which distills patch features from a frozen vision model teacher. The ViT-L encoder has one CLS and four register tokens ahead of its patch sequence; we exclude these five tokens and pool over patch tokens only. The ConvNeXt-L encoder has no such special tokens, so its feature map is used directly.

2.3 Case-Level Cavity Detection

Case-level cavity labels exist at a far larger scale than masks, so we train the presence classifier on them. From the NIAID TB Portals program [17] we take the union of three published CXR releases (August 2023, March 2025, March 2026 v2) after deduplication and derive a binary cavity label per study from their sextant-level manual annotations. We remove any study matching a released challenge test image, along with all other studies of those patients. Pooled with the challenge training images and split by patient, this gives 14,902 training cases (6,587 positive / 8,315 negative) and 1,646 held-out cases; sampling one negative per positive leaves 13,174 for training.

The detector consists of the pretrained ViT-L encoder with attentive pooling over its patch tokens and a two-layer MLP head. It is trained end-to-end with binary cross-entropy at 1024 resolution using AdamW [16] (encoder lr 2×10^{-5}, pool/head lr 1×10^{-4}), a cosine schedule with 5% warmup, weight decay 0.05, bfloat16, and an effective batch 64, for at most 10 epochs with early stopping on held-out AUROC (patience 4). The held-out split shares no image or patient with the challenge test set. The submitted detector model reaches AUROC 0.8386 on it.

2.4 Staged Encoder Transfer for Segmentation

With only 444 masks available, the segmentation encoder's pretrained representations matter more than what it can learn from the masks alone. We therefore insert a classification stage between pretraining and segmentation. The ConvNeXt-L encoder is first fine-tuned as a cavity-presence classifier on exactly the data above, with the same recipe as the detector, and its weights are then exported to initialize every segmentation model. The segmentation encoder consequently arrives having seen 13,174 case-labeled cavity images, compared to the 444 masks it is about to be trained on. The detector used at inference remains a separate ViT-L model, ensuring that the case-level decision and the mask are never produced by the same network.

2.5 Segmentation Model

The transferred ConvNeXt-L encoder feeds a ConvNeXt-aware U-Net [9] with four encoder stages aggregated coarse-to-fine by a pyramid-pooling bottleneck [10], all-stage skip connections, decoder widths $[512, 256, 128, 64]$, GroupNorm/GELU blocks, and deep-supervision heads [11,12]. The decoder emits a single cavity logit per pixel plus a global-average-pooled presence logit. Each model is fine-tuned end-to-end on the 444 masks at 1024 input resolution under

$$\mathcal{L} = \mathcal{L}_{\text{Focal}} + \mathcal{L}_{\text{Tversky}} + 0.5\,\mathcal{L}_{\text{pres}} + 0.4\,\mathcal{L}_{\text{ds}}, \tag{2}$$

where $\mathcal{L}_{\text{pres}}$ is a binary cross-entropy on the pooled presence logit and $\mathcal{L}_{\text{ds}}$ is the mean of the same pixel loss over the auxiliary decoder heads, with Focal [13] at $\gamma = 2$ and Tversky [14] at $\alpha = 0.3$, $\beta = 0.7$. Both terms address the foreground/background imbalance, and the Tversky asymmetry penalizes false negatives more than false positives. Training uses AdamW [16] (encoder lr 2×10^{-5}, decoder/heads lr 3×10^{-4}), a cosine schedule with 5% warmup, weight decay 0.05, bfloat16, and an effective batch size of 32 over eight GPUs, for at most 60 epochs with early stopping on the composite (patience 8). The loss is evaluated on a 1024 mask grid, matching the encoder input.

Since only 444 mask-labeled images exist, a single model has high variance. We therefore ensemble *four* members that share this recipe and differ only in random seed, averaging their probability maps at inference.

2.6 Operating Points and Inference

The case-level decision is taken from the detector alone and is never overridden by the segmentation output. Predicted negatives receive an empty mask, following the challenge protocol. A predicted positive with an otherwise empty thresholded mask retains its top 1% most confident pixels, so the case label and the mask remain consistent.

The two thresholds are fitted on different objectives and different data, never jointly on the composite: the mask threshold maximizes ungated mean Dice on the released 111-image test split, a property of the segmentation model alone, while the detection threshold τ maximizes presence accuracy on the detector's own 1,646-case held-out split, which contains none of the images being scored. The submitted operating point is $\tau = 0.70$ with a mask threshold of 0.79 on the member-averaged probability map.

3 Experiments and Results

3.1 Two Evaluations

The challenge provides two evaluation settings. Stage 1 is the released 111-case test split (58 positive / 53 negative) with ground-truth masks. Since no separate validation split is provided, we also use this dataset for epoch selection and mask

Table 1. Stage 1 ungated mean Dice (111 cases), mask threshold fitted per row.

Configuration	Threshold	$\overline{\text{Dice}}$
Submitted configuration (four seeds)		
M1	0.96	0.4086
M2	0.75	0.3475
M3	0.74	0.3951
M4	0.90	0.4060
ensemble of M1–M4	0.74	**0.4391**
Encoder learning rate variants		
enc. 1×10^{-5}	0.57	0.3626
enc. 5×10^{-6}	0.93	0.3650
Other configurations		
S0 (no detection pretraining)	0.86	0.4365
S1 (BCE+Dice loss)	0.74	0.3711

threshold fitting, making it optimistic by construction. Stage 2 is the organizers' held-out external evaluation, never disclosed to participants, and serves as the final challenge result.

3.2 Detection Threshold Selection

Because a predicted positive assigned an empty mask contributes Dice 0 rather than being excluded, the detection threshold τ moves both terms of Eq. 1 in opposite directions. The submitted entry uses $\tau = 0.70$. For predicted positives whose thresholded mask is empty, we retain the top 1% most confident pixels, a fraction chosen to approximate the median foreground pixel count among positive cases in the training set.

3.3 Segmentation Ablation on Stage 1

Table 1 compares segmentation models alone, as ungated mean Dice on 111 cases with the mask threshold fitted per model. The detector's presence gating is excluded so that the numbers measure mask quality alone.

Three observations emerge from the results. First, seed variance is large: the four members of a single configuration span 0.348 to 0.409, motivating the use of ensembling, which consistently improves mean Dice with diminishing returns and reduces variance across member combinations. Second, the submitted learning-rate pairing dominates both alternatives by about 0.045. Third, S0 already reaches 0.4365 alone, within the noise margin of the four-seed ensemble, a fact revisited in Sect. 3.4.

3.4 What Transferred and What Did Not

The submitted system scores a composite of 0.6265 on the 111 released test cases ($A_{\text{det}} = 0.7568$, $\overline{\text{Dice}} = 0.3227$, at $\tau = 0.70$ and mask threshold 0.79)

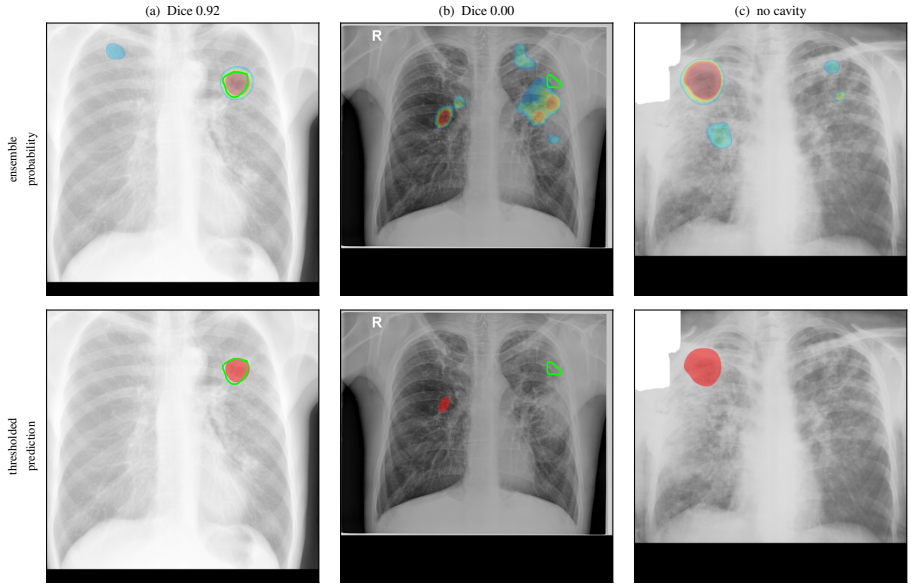

Fig. 2. Qualitative results at mask threshold 0.79, before presence gating. Green: ground truth, Red: prediction. (a) Correctly localized cavity. (b) Mask placed in the wrong region (Dice 0). (c) Cavity-negative case; false positive removed by presence gating.

and 0.6001 on the external test set (0.8112, 0.1077). Detection improved across the two sets: 0.757 → 0.811, while Dice decreased: 0.323 → 0.108. The case-level decision, trained on 13,174 images from a different corpus, transferred. The mask geometry, learned from 444 images with its threshold fitted on Stage 1, did not (Fig. 2).

The ensemble also did not transfer: on released 111 cases, it improved Dice by +0.031 over the best single member (Table 1), but on the external set it scored 0.1077, below the single model at 0.1130 under identical detection accuracy.

4 Limitations

We did not fully ablate the staged encoder transfer against direct initialization from the pretrained encoder. We also explored weak supervision using TB Portals sextant-level annotations, which divide each lung into six zones, to improve localization; however, it did not improve Dice, likely because the zones are too coarse at 1024 resolution to sharpen boundaries. Finally, our ensemble relies on seed-level diversity alone; a single model with stronger intrinsic robustness would be preferable to variance reduction through averaging.

5 Conclusion

The composite score weights case-level detection far more than boundary accuracy, while mask supervision is two orders of magnitude smaller than the publicly available case-level labels. Designing our system around this asymmetry yielded a composite of 0.6001 on the external evaluation, with detection transferring well across test sets and Dice not. The components that generalized are those supervised at scale, suggesting that further improvement would benefit more from acquiring additional mask annotations to improve segmentation robustness. The submission placed third in the final Task 1 ranking.

Acknowledgments. This work was supported by the Technology Innovation Program (RS-2025-02221011, Development of Medical-Specialized Multimodal Hyperscale Generative AI Technology for Global Integration) funded by the Ministry of Trade Industry & Energy (MOTIE, South Korea), and by the "Advanced GPU Utilization Support Program" funded by the Government of the Republic of Korea (Ministry of Science and ICT).

Data were obtained from the TB Portals (https://tbportals.niaid.nih.gov), which is an open-access TB data resource supported by the National Institute of Allergy and Infectious Diseases (NIAID) Office of Cyber Infrastructure and Computational Biology (OCICB) in Bethesda, MD. These data were collected and submitted by members of the TB Portals Consortium (https://tbportals.niaid.nih.gov/Partners). Investigators and other data contributors that originally submitted the data to the TB Portals did not participate in the design or analysis of this study [17].

Disclosure of Interests. All authors are employees of DEEPNOID Inc.

References

1. TREAT-MMTB: Transformative Research and Efficient AI Technologies for Multimodal Management of Tuberculosis 2026. Zenodo (2026). https://doi.org/10.5281/zenodo.19732124
2. Park, J., et al.: GLINT: Sparsely Gated Vision-Language Alignment for Fine-Grained Radiology Representations. arXiv:2606.03180 (2026)
3. Siméoni, O., et al.: DINOv3. arXiv:2508.10104 (2025)
4. Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: CVPR, pp. 11966–11976 (2022)
5. Johnson, A.E.W., et al.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci. Data **6**, 317 (2019)
6. Qwen Team: Qwen3.6-35B-A3B: Agentic Coding Power, Now Open to All (2026). https://qwen.ai/blog?id=qwen3.6-35b-a3b
7. Song, K., Tan, X., Qin, T., Lu, J., Liu, T.-Y.: MPNet: Masked and Permuted Pre-training for Language Understanding. In: NeurIPS, vol. 33, pp. 16857–16867 (2020)
8. Lee, J., Kim, J., Shon, H., Kim, B., Kim, S.H., Lee, H., Kim, J.: UniCLIP: Unified Framework for Contrastive Language-Image Pre-training. In: NeurIPS, vol. 35 (2022)

9. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: MICCAI, LNCS, vol. 9351, pp. 234–241. Springer (2015)
10. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid Scene Parsing Network. In: CVPR, pp. 2881–2890 (2017)
11. Lee, C.-Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-Supervised Nets. In: AISTATS, pp. 562–570 (2015)
12. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
13. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal Loss for Dense Object Detection. In: ICCV, pp. 2980–2988 (2017)
14. Salehi, S.S.M., Erdogmus, D., Gholipour, A.: Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks. In: MLMI, LNCS, vol. 10541, pp. 379–387. Springer (2017)
15. Zuiderveld, K.: Contrast Limited Adaptive Histogram Equalization. In: Heckbert, P.S. (ed.) Graphics Gems IV, pp. 474–485. Academic Press (1994)
16. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. In: ICLR (2019)
17. Rosenthal, A., Gabrielian, A., Engle, E., Hurt, D.E., Alexandru, S., Crudu, V., et al.: The TB Portals: an Open-Access, Web-Based Platform for Global Drug-Resistant-Tuberculosis Data Sharing and Analysis. J. Clin. Microbiol. **55**(11), 3267–3282 (2017)