LocSAM3: Box-Supervised Adaptation of SAM3 for Text-Only Chest X-Ray Segmentation

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Pixel-level segmentation masks are expensive to annotate, particularly in chest X-rays, where anatomical structures overlap and pathological boundaries are often ill-defined. Bounding boxes (BBoxes) provide a lower-burden form of localization supervision and are available in several chest X-ray datasets. We present LocSAM3, a mask-free adaptation of SAM3 for text-only anatomy and pathology segmentation that uses only BBoxes and concept names during training. Rather than learning radiograph-specific masks directly, LocSAM3 adapts SAM3's concept-grounding and localization modules while keeping its pretrained mask decoder and more than 95% of its vision backbone frozen. The fixed decoder then converts the learned in-domain localization into a pixel-level prediction. During inference, the model receives only an image and a text concept. On a 10-class MIMIC-CXR benchmark, LocSAM3 improves text-only segmentation over Medical-SAM3 from 0.48 to 0.69 mIoU for anatomical structures and from 0.09 to 0.53 for pathological findings. These results demonstrate that coarse localization annotations can support effective text-to-pixel adaptation in chest X-rays without requiring dense mask annotations for training or spatial prompts at inference.

Keywords: Medical image segmentation · Text-guided segmentation · Bounding box supervision · Chest radiography · SAM3.

1 Introduction

Dense segmentation annotation is a major bottleneck in medical imaging. Producing a pixel-level mask requires tracing the complete target boundary, which is time-consuming for well-defined anatomy and especially difficult for overlapping structures or diffuse abnormalities [22]. Bounding boxes (BBoxes) provide coarser but less burdensome supervision and are already available in datasets such as Chest ImaGenome [23]; concept names can also be obtained from structured labels or radiology reports. This motivates a practical question: *Can boxes and concept names replace pixel masks when adapting a model for text-to-mask segmentation?* Such models could ground clinical concepts in image regions for downstream tasks such as report grounding and temporal knowledge graphs [13,12,25], but requiring a user-provided point or box for every prediction limits automation. We therefore study a harder setting in which the model receives only a chest X-ray and a text concept and must both localize and segment the target.

Promptable foundation models offer a natural starting point, but chest radiography presents two challenges. First, grayscale appearance, superimposed anatomy, and weak boundaries make localization from appearance alone difficult [16]. Reliable grounding also requires anatomical priors because thoracic structures occupy characteristic locations [1,24]. Second, always supplying boxes during training can create a shortcut, where the model learns to refine a given region rather than localize it from text, leading to poor text-only performance when the box is removed.

We introduce LocSAM3, a box-supervised adaptation of SAM3 [2] for text-only inference. It never uses pixel masks for optimization: we adapt SAM3's concept-grounding and localization modules while keeping the mask decoder frozen, and the box is supplied as an input prompt for only a random subset of training examples so the model learns to localize from text alone.

Our contributions are:

1. We formulate a mask-free setting in which models learn from concept names and BBoxes but produce pixel masks from text alone at inference.
2. We introduce box-prompt dropout, which retains boxes as localization targets while withholding them as input prompts for a fraction of training examples.
3. We adapt SAM3 by updating specific encoders and heads while freezing the mask decoder and over 95% of the vision backbone.
4. On a 10-class MIMIC-CXR benchmark, LocSAM3 significantly improves text-only mIoU over Medical-SAM3 for both anatomy (0.48 to 0.69) and pathology (0.09 to 0.53), proving that partial box prompting is highly effective.

2 Related Work

Before foundation models, medical segmentation relied on task-specific CNNs; for example, a model trained on pulmonary opacities could not segment cardiac boundaries without retraining. The Segment Anything Model (SAM) family removed this restriction [11]: SAM and SAM2 [19] generate masks from spatial prompts, and SAM3 [2] extends this to text-to-mask inference through Promptable Concept Segmentation (PCS). Because these models are trained mainly on natural RGB images, they transfer poorly to low-contrast radiographs. Vision-language localization models face a similar limitation: GroundingDINO [15] enables open-vocabulary localization from text but is limited to BBoxes, and DINOv3 [20] produces strong ViT features but has no mask decoder. Lightweight adapters such as SAM4MLLM [4], which uses an LLM to generate point prompts for a frozen SAM, and SAM3-Adapter [3], which adds trainable bottleneck layers, leave the core perception weights largely unchanged.

In the medical domain, several works have already moved toward text-driven segmentation, but none have closed the text-only gap we study. Grounding DINO-US-SAM [18] performs fully automated text-prompted ultrasound segmentation by pairing a LoRA-tuned Grounding DINO detector with a frozen SAM2 decoder; the box is generated internally from text, but this is a two-stage detect-then-segment pipeline on a different modality that leaves the segmentation model

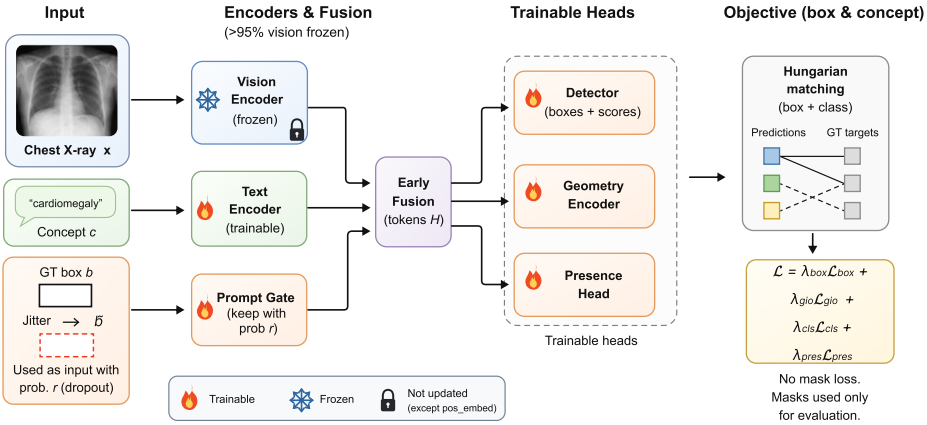

Fig. 1. Training procedure of LocSAM3. The ground-truth box b is jittered and injected as an input prompt only with probability r (box-prompt dropout). Only the text encoder and the Detector, Geometry, and Presence heads are trained, with a minimal `pos_embed` update, while more than 95% of the vision backbone is frozen.

off-the-shelf. MedSAM3 [14] adapts SAM3 to medical PCS by fine-tuning only its detector and is text-promptable, but its text-only mode is reported to be weaker than its text+BBox mode, i.e., the same prompt dependency we remove. Full fine-tuning does not resolve this either: although Medical-SAM3 [9] is fully fine-tuned without bounding box prompts during training, it attains reasonable accuracy only when provided with BBoxes at inference and collapses under text-only inference, especially on dense, overlapping structures such as the heart. In contrast, LocSAM3 targets the text-only setting directly: it withholds the BBox for a fraction of training samples so that a single SAM3 model learns to localize from text alone, without any pixel-mask supervision.

3 Methodology

Let $x \in \mathbb{R}^{H \times W \times 3}$ be a chest X-ray, c a target concept, and $b \in \mathbb{R}^4$ its BBox. A binary mask $y \in \{0,1\}^{H \times W}$ is used only for held-out evaluation and never for optimization. Our goal is to learn text-only segmentation,

$$\hat{y} = f_\theta(x, c),$$

using only concept and box supervision. As illustrated in Fig. 1, the box is always a supervision target but is provided as an input prompt only for a subset of training examples.

3.1 Parameter-Efficient Model Adaptation

We build on SAM3 [2], which comprises vision and text encoders, a Detector, Geometry Encoder, Presence Head, and mask decoder. To reduce overfitting

and catastrophic forgetting on limited medical data [21], we train the text encoder, Detector, Geometry Encoder, and Presence Head while freezing more than 95% of the vision backbone; only its positional embeddings (`pos_embed`) are updated, using a learning rate of 10^{-7}, to permit limited adaptation to chest X-ray geometry [6]. The mask decoder remains frozen and provides the pretrained dense-segmentation prior.

3.2 Mask-Free Box-and-Concept Supervision

Because the detector predicts a fixed set of candidate boxes, predictions are first matched to ground-truth annotations using Hungarian assignment. The model is optimized using box-regression, generalized Intersection over Union (IoU), concept-classification, and presence losses:

$$\mathcal{L} = \lambda_{\text{box}}\mathcal{L}_{\text{box}} + \lambda_{\text{giou}}\mathcal{L}_{\text{giou}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{pres}}\mathcal{L}_{\text{pres}}.$$

The objective contains no mask term: training adapts only concept grounding and localization, and the frozen SAM3 decoder converts the resulting representation into a dense mask.

3.3 Box-Prompt Dropout for Text-Only Inference

BBoxes provide useful spatial guidance in chest X-rays, where structures overlap and pathological boundaries are diffuse, but always supplying a box during training makes the model depend on a prompt that is unavailable at text-only deployment. We therefore separate the box as a *supervision target* from its optional role as an *input prompt*.

The target b is retained for every training example. A perturbed copy $\tilde{b} = \mathcal{J}(b)$, generated using random noise ($\sigma = 0.1$, shift $\leq 20\,\text{px}$), is supplied as input only with probability r:

$$z \sim \text{Bernoulli}(r), \qquad p_b = \begin{cases} \tilde{b}, & z = 1, \\ \varnothing, & z = 0. \end{cases}$$

Training is conditioned as $\hat{y} = f_\theta(x, c, p_b)$, whereas inference always uses $p_b = \varnothing$. Perturbation discourages copying the exact target coordinates, and dropout exposes the model to both box-guided and text-only examples, so r trades geometric guidance against the text-only deployment condition. This mechanism is related to conditioning and modality dropout [8,17].

4 Experiments

4.1 Datasets and Splits

We use MIMIC-CXR-JPG [10] as the image source and combine annotations from three complementary resources. For anatomy, we consider the Heart, the

Left Lung, and the Right Lung. BBoxes are obtained from Chest ImaGenome [23], while CheXmask [7] provides masks used only for evaluation. We retain CheXmask annotations with Reverse Classification Accuracy (RCA) $\geq 90\%$. This subset contains 1,365 training, 186 validation, and 110 test radiographs.

For pathology, we use the predefined splits of MIMIC-CXR-Ext-ILS [5] and include Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Opacity, and Pneumonia. Together, the anatomical and pathological subsets form a 10-class benchmark. All prompts are standardized to short concept names without prompt ensembling. The pathology training set is imbalanced, ranging from 3,919 concept–image instances for Consolidation to 40,243 for Cardiomegaly. We oversample and augment underrepresented classes during training only; the validation and test distributions remain unchanged. Because all annotation resources are linked to the longitudinal MIMIC-CXR collection, we audit splits using patient identifiers. The train, validation, and test partitions are pairwise patient-disjoint, including across the anatomical and pathological subsets: no `subject_id` occurs in more than one partition.

4.2 Baselines

We compare LocSAM3 with Medical-SAM3 [9] under text-only inference, where each model receives an image and concept name but no spatial prompt. We additionally evaluate text+box inference to measure dependence on explicit localization cues. For a fair comparison, all models use the same test instances, concept prompts, input resolution, output-selection rule, and mask post-processing.

4.3 Implementation Details

We train LocSAM3 for 30 epochs on six NVIDIA RTX 3090 GPUs using AdamW, bfloat16 mixed precision, and an effective batch size of six. The model uses $d_{\text{model}} = 256$, 200 queries, and early fusion with one feature level. Learning rates are 5×10^{-6} for the trainable transformer modules, 1×10^{-6} for the text encoder, and 1×10^{-7} for the trainable vision parameters. We use an inverse-square-root schedule, weight decay 0.05, vision layer decay 0.95, and gradient clipping with maximum norm 0.01.

Training assignment uses Hungarian matching with box and class costs, plus a one-to-many auxiliary matcher. The objective contains box, GIoU, concept-classification, and presence losses. Training images are processed at resolutions up to 1008×1008 and augmented with shifts, scaling, rotations up to 7°, brightness and contrast changes, and Gaussian noise or blur. For the smaller anatomical subset, we generate five augmented training variants per image. Unless otherwise stated, results use the final checkpoint and random seed 300.

4.4 Box-Prompt Ablation

Using the three anatomical classes, we vary the box-prompt rate $r \in \{.0,.2,.4,.6,.8,1.0\}$ and evaluate each resulting model under text-only inference, isolating the effect

of training-time box exposure on prompt-free performance. Both endpoints are suboptimal: at $r = 1.0$ the model never trains without a box and transfers poorly to text-only inference, while low rates provide little spatial guidance. Validation performance peaks near $r = 0.6$ (a box-drop probability of 0.4), which we fix before evaluating the 10-class test benchmark.

4.5 Evaluation Protocol

We report IoU and Dice for each positive concept–image test instance. Scores are first averaged within each class. The *Overall* pathology result is then computed as a sample-weighted average because class frequencies differ substantially; it covers 2,217 test instances. The anatomical classes contain equal numbers of evaluated instances, so weighted and unweighted class averages are identical. At inference, we select the predicted mask associated with the highest-scoring matched concept query and binarize it using a threshold selected on the validation set. If the model returns no valid mask, the instance receives an IoU and Dice score of zero.

5 Results and Discussion

Text-only segmentation. Table 1 compares both models under the intended deployment condition: each receives only an image and text concept, with no spatial prompt. LocSAM3 outperforms Medical-SAM3 on all ten classes, improving overall anatomy mIoU by 0.211 (0.691 vs. 0.480) and instance-weighted pathology mIoU by 0.441 (0.526 vs. 0.085). The class-macro pathology mIoU also rises from 0.064 to 0.446, so the gain is not driven only by frequent classes. The largest gains occur for cardiomegaly (+0.759), opacity (+0.420), heart (+0.394), and consolidation (+0.383); right lung is the only near-tie (0.847 vs. 0.841). Effusion and heart remain the weakest targets. Figure 2 illustrates this: LocSAM3 places masks on the correct region across effusion and pneumonia cases, while the baseline segments the wrong lung or the whole field, though our boundaries are not always tight.

These gains are obtained without pixel-level masks during optimization. Because the mask decoder remains frozen, the results are consistent with improved in-domain concept grounding and localization being converted into dense predictions by SAM3's pretrained segmentation prior. However, upstream representations also condition the decoder, so the improvement should not be attributed exclusively to box localization.

Dependence on spatial prompts. Table 2 compares text-only and text+box inference. For Medical-SAM3, removing the box reduces pathology mIoU from 0.400 to 0.085, an absolute decrease of 0.315 and a relative decrease of 78.7%. Its anatomy mIoU decreases from 0.620 to 0.480 (22.6% relative). This indicates that the evaluated baseline is particularly dependent on explicit spatial guidance for pathological targets. LocSAM3 also benefits from a box at inference, but its prompt-removal gap is smaller. Pathology mIoU decreases from 0.679 to 0.526 (22.5% relative), while anatomy decreases from 0.781 to 0.691 (11.5%

Table 1. Text-only segmentation on the 10-class benchmark. Values are mean per-instance IoU and Dice; n is the number of test instances. The pathology macro average weights classes equally, whereas the weighted average weights test instances.

		LocSAM3		**Medical-SAM3**	
Category	n	**mIoU**	**Dice**	**mIoU**	**Dice**
Anatomy					
Left lung	110	**0.826**	**0.890**	0.593	0.635
Right lung	110	**0.847**	**0.912**	0.841	0.901
Heart	110	**0.401**	**0.483**	0.007	0.014
Overall	330	**0.691**	**0.762**	0.480	0.517
Pathology					
Atelectasis	250	**0.375**	**0.500**	0.051	0.080
Cardiomegaly	765	**0.767**	**0.866**	0.007	0.014
Consolidation	74	**0.405**	**0.528**	0.022	0.038
Edema	623	**0.428**	**0.555**	0.242	0.336
Effusion	198	**0.300**	**0.419**	0.033	0.058
Opacity	187	**0.426**	**0.559**	0.006	0.012
Pneumonia	120	**0.418**	**0.549**	0.087	0.129
Macro average	–	**0.446**	**0.568**	0.064	0.095
Instance-weighted average	2,217	**0.526**	**0.643**	0.085	0.123

Table 2. Effect of inference-time box prompting. For each model, the same checkpoint is evaluated with text vs text with bounding box. Pathology results are instance-weighted.

		Anatomy		**Pathology**	
Model	**Input**	**mIoU**	**Dice**	**mIoU**	**Dice**
Medical-SAM3	Text only	0.480	0.517	0.085	0.123
Medical-SAM3	Text + box	0.620	0.726	0.400	0.529
LocSAM3	Text only	**0.691**	**0.762**	**0.526**	**0.643**
LocSAM3	Text + box	**0.781**	**0.869**	**0.679**	**0.796**

relative). Moreover, text-only LocSAM3 exceeds Medical-SAM3 even when the latter receives a box: 0.526 versus 0.400 on pathology and 0.691 versus 0.620 on anatomy. Thus, box-prompt dropout reduces, although it does not eliminate, dependence on inference-time spatial guidance.

Limitations. First, the frozen mask decoder cannot directly adapt to radiograph-specific boundaries; box supervision primarily improves grounding and localization. Second, the weaker heart result under joint training indicates sensitivity to class imbalance or negative transfer. Third, all images originate from MIMIC-CXR, so cross-institutional generalization remains untested. BBoxes are less burdensome than dense masks but may still require expert annotation, particularly for diffuse findings.

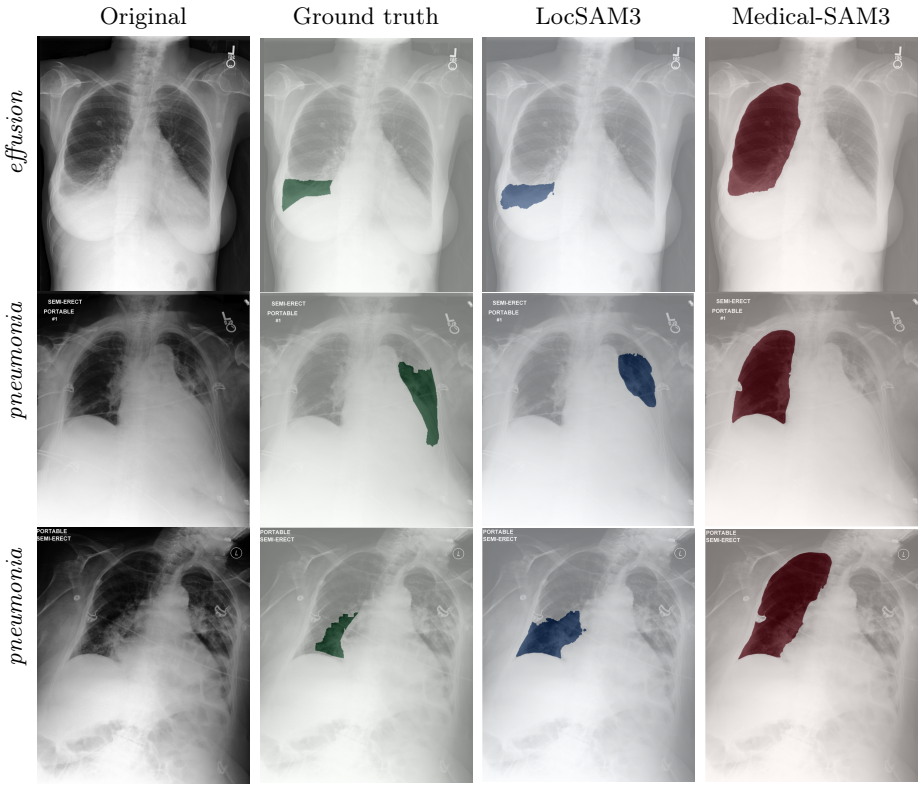

Fig. 2. Text-only predictions on three findings. LocSAM3 (blue) localizes the finding in agreement with the reference (green), whereas Medical-SAM3 (red) mislocalizes or floods the hemithorax. Boundary tightness remains imperfect (row 2).

6 Conclusion & Future Work

We presented LocSAM3, a mask-free adaptation of SAM3 for text-only anatomy and pathology segmentation in chest X-rays. Using only BBox and concept-name supervision, and no pixel masks at any point during training, the model learned stronger in-domain localization which, decoded by the pretrained SAM3 mask head, produced text-only masks that improved over a fully fine-tuned baseline. The central ingredient is a simple box-prompt dropout: by withholding the box for a fraction of training samples, the model learns to localize from text alone at inference. LocSAM3 offers a practical, low-annotation-cost baseline for automated text-to-pixel grounding in limited-resource medical settings. Future work will focus on improving segmentation quality beyond localization, for example through partial adaptation of the pretrained SAM3 mask decoder. We also plan to investigate localization-focused fine-tuning strategies for Medical-SAM3 to better understand the interaction between localization and segmentation performance.

References

1. Ahmad, I.S., et al.: Foundation models for X-ray interpretation: a narrative review of current techniques and future perspectives in diagnostic imaging. Quantitative Imaging in Medicine and Surgery **16**(4) (2026). https://doi.org/10.21037/qims-2025-1-2782
2. Carion, N., et al.: SAM 3: Segment Anything with Concepts. In: The International Conference on Learning Representations (2026), https://openreview.net/forum?id=r35clVtGzw
3. Chen, T., et al.: SAM3-Adapter: Efficient Adaptation of Segment Anything 3 for Camouflage Object Segmentation, Shadow Detection, and Medical Image Segmentation (2025). https://doi.org/10.48550/arXiv.2511.19425
4. Chen, Y.C., et al.: SAM4MLLM: Enhance Multi-Modal Large Language Model for Referring Expression Segmentation. In: European Conference on Computer Vision. Cham (2024). https://doi.org/10.1007/978-3-031-73004-7_19
5. Choi, G., et al.: Instruction-Guided Lesion Segmentation for Chest X-rays with Automatically Generated Large-Scale Dataset. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2026), https://openaccess.thecvf.com/content/CVPR2026/html/Choi_Instruction-Guided_Lesion_Segmentation_for_Chest_X-rays_with_Automatically_Generated_Large-Scale_CVPR_2026_paper.html
6. Dong, X., et al.: CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022). https://doi.org/10.1109/CVPR52688.2022.01181
7. Gaggion, N., et al.: CheXmask: a large-scale dataset of anatomical segmentation masks for multi-center chest x-ray images. Scientific Data **11**(1) (2024). https://doi.org/10.1038/s41597-024-03358-1
8. Ho, J., et al.: Classifier-Free Diffusion Guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021), https://openreview.net/forum?id=qw8AKxfYbI
9. Jiang, C., et al.: Medical SAM3: A Foundation Model for Universal Prompt-Driven Medical Image Segmentation (2026). https://doi.org/10.48550/arxiv.2601.10880
10. Johnson, A.E.W., et al.: MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs (2019). https://doi.org/10.48550/arXiv.1901.07042
11. Kirillov, A., et al.: Segment Anything. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2023). https://doi.org/10.1109/ICCV51070.2023.00371
12. Koleilat, T., et al.: MedCLIP-SAM: Bridging Text and Image Towards Universal Medical Image Segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 643–653. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-72390-2_60
13. Li, D., et al.: CliCARE: Grounding Large Language Models in Clinical Guidelines for Decision Support over Longitudinal Cancer Electronic Health Records. AAAI Conference on Artificial Intelligence **40**(37) (2026). https://doi.org/10.1609/aaai.v40i37.40421
14. Liu, A., et al.: MedSAM3: Delving into Segment Anything with Medical Concepts (2025). https://doi.org/10.48550/arXiv.2511.19046
15. Liu, S., et al.: Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. In: European Conference on Computer Vision. Cham (2024). https://doi.org/10.1007/978-3-031-72970-6_3

16. Mochurad, L.: Approach for enhancing the accuracy of semantic segmentation of chest X-ray images by edge detection and deep learning integration. Frontiers in Artificial Intelligence **8** (2025). https://doi.org/10.3389/frai.2025.1522730
17. Neverova, N., et al.: ModDrop: Adaptive Multi-Modal Gesture Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(8) (2016). https://doi.org/10.1109/TPAMI.2015.2461544
18. Rasaee, H., et al.: Grounding DINO-US-SAM: Text-Prompted Multiorgan Segmentation in Ultrasound With LoRA-Tuned Vision–Language Models. IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control **72**(10) (2025). https://doi.org/10.1109/TUFFC.2025.3605285
19. Ravi, N., et al.: SAM 2: Segment Anything in Images and Videos. In: International Conference on Learning Representations (2024), https://openreview.net/forum?id=Ha6RTeWMd0
20. Siméoni, O., et al.: DINOv3. Transactions on Machine Learning Research (2026), https://openreview.net/forum?id=2NlGyqNjns
21. Teuber, C., et al.: Parameter Efficient Fine-Tuning of Segment Anything Model for Biomedical Imaging. In: Medical Imaging with Deep Learning (2025), https://openreview.net/forum?id=TWPnpixpW2
22. Wang, S., et al.: Annotation-efficient deep learning for automatic medical image segmentation. Nature Communications **12**(1) (2021). https://doi.org/10.1038/s41467-021-26216-9
23. Wu, J.T., et al.: Chest ImaGenome Dataset for Clinical Reasoning. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2021), https://openreview.net/forum?id=H-d5634yVi&noteId=N6xn6-Ec02
24. Yue, Y., et al.: CheXWorld: Exploring Image World Modeling for Radiograph Representation Learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025). https://doi.org/10.1109/CVPR52734.2025.01935
25. Zhao, Z., et al.: Large-vocabulary segmentation for medical images with text prompts. npj Digital Medicine **8**(1), 566 (Sep 2025). https://doi.org/10.1038/s41746-025-01964-w