Diagnosing 3D Volumes from Camera Video: A Semantic-Robust Framework Without Direct DICOM Access

Junlei Wu[1†], Haolin Huang[1†], Jiaming Li[1], and Qian Wang[1,2](✉)

[1] School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai, China
qianwang@shanghaitech.edu.cn
[2] Shanghai Clinical Research and Trial Center, Shanghai, China

Abstract. Medical imaging traditionally assumes that high-fidelity, lossless data are required for reliable computer-aided diagnosis (CAD), an assumption inherited by AI models trained on pristine DICOM volumes. While 2D radiographs have shown tolerance to smartphone recapture, it remains unknown whether 3D volumetric diagnosis can survive the severe degradation introduced when a camera records slice scrolling as a video stream. This question is increasingly relevant as DICOM-level access and PACS integration impose substantial infrastructural barriers, particularly in resource-limited settings. We introduce **ScreenCAD**, a framework designed to test whether a 3D medical volume, transformed into a camera-captured video, still preserves sufficient semantic information for AI diagnosis. ScreenCAD combines cascaded ROI extraction, a frozen medical foundation model with strong semantic invariance, and a Task-Token Mixture-of-Experts aggregator to operate directly on low-fidelity, variable-length video streams. On CT-RATE with real-world screen recordings, ScreenCAD achieves competitive diagnostic performance under video inputs, while conventional 3D baselines degrade sharply. These results provide initial evidence that volumetric diagnosis can remain feasible under substantial information loss, highlighting camera-based acquisition as a promising interface for future medical AI systems when direct PACS integration is infeasible. Our code will be released at GitHub.

Keywords: Camera-Captured Video · Foundation Models · Volumetric Diagnosis · Semantic Robustness.

1 Introduction

For decades, both clinical radiology and computer-aided diagnosis (CAD) have been built upon a foundational assumption that high-fidelity, lossless images are essential for reliable interpretation. This belief has driven continuous improvements in imaging hardware, hospital information systems, and PACS infrastructures, all designed to preserve pixel-level integrity during acquisition,

† These authors contributed equally to this work.

storage, and transmission. Modern AI systems [1–4] have inherited this assumption: nearly all diagnostic models are trained and evaluated on pristine DICOM volumes, implicitly treating image quality as a prerequisite for algorithmic power.

Recent advances in large-scale self-supervised learning challenge this long-standing premise. Medical foundation models [5–7] increasingly exhibit strong semantic invariance, capturing anatomical structure rather than texture-level details. This raises a fundamental scientific question: *must AI rely on high-fidelity images, or can it remain diagnostically reliable even when substantial visual information is lost during transmission?*

In the 2D domain, early evidence suggests that AI can tolerate certain forms of degradation. Studies on camera-captured radiographs [8] have shown that perspective correction and simple enhancement can recover sufficient information for downstream diagnosis. These results imply that, at least for planar images, the requirement for perfect digital fidelity may be weaker than assumed.

However, the central challenges of medical imaging arise not from 2D radiographs but from 3D volumetric modalities such as CT and MRI. Volumetric interpretation requires clinicians to scroll through hundreds of slices, forming a temporally ordered visual stream that reflects the underlying anatomy. When this process is captured by a camera, the 3D volume is effectively transformed into a video sequence containing Moiré patterns, compression artifacts, motion blur, variable scrolling speed, and frame drops. Unlike 2D photos, such video streams cannot be trivially rectified or restored, and no prior work has systematically investigated whether AI can perform volumetric diagnosis from these low-fidelity, camera-recorded representations.

Beyond the scientific question of fidelity, the second challenge concerns global accessibility. Integrating AI systems into hospital IT infrastructure requires substantial financial and technical investment, including PACS interoperability, cybersecurity approval, and workflow modification [9, 10]. Such requirements pose significant barriers for low- and middle-income healthcare institutions [11], where upgrading existing systems is often infeasible. As a result, the benefits of modern AI diagnosis remain unevenly distributed, reinforcing long-standing inequities in global healthcare.

These two considerations motivate a new paradigm: *camera-as-a-universal-interface for volumetric AI diagnosis*. Instead of requiring direct access to DICOM volumes or modification of hospital systems, AI models would accept screen-recorded video as input, enabling deployment in diverse clinical environments with minimal infrastructural demands. If feasible, such a paradigm could dramatically reduce integration costs, broaden global accessibility, and challenge the assumption that high-fidelity imaging is a prerequisite for AI-assisted diagnosis.

In this work, we take the first step toward this vision by introducing **ScreenCAD**, a framework that performs volumetric diagnosis directly from camera-captured video streams. To our knowledge, this is the first study to investigate whether a 3D medical volume, when converted into a video stream during routine slice scrolling, still contains sufficient semantic information for reliable AI diag-

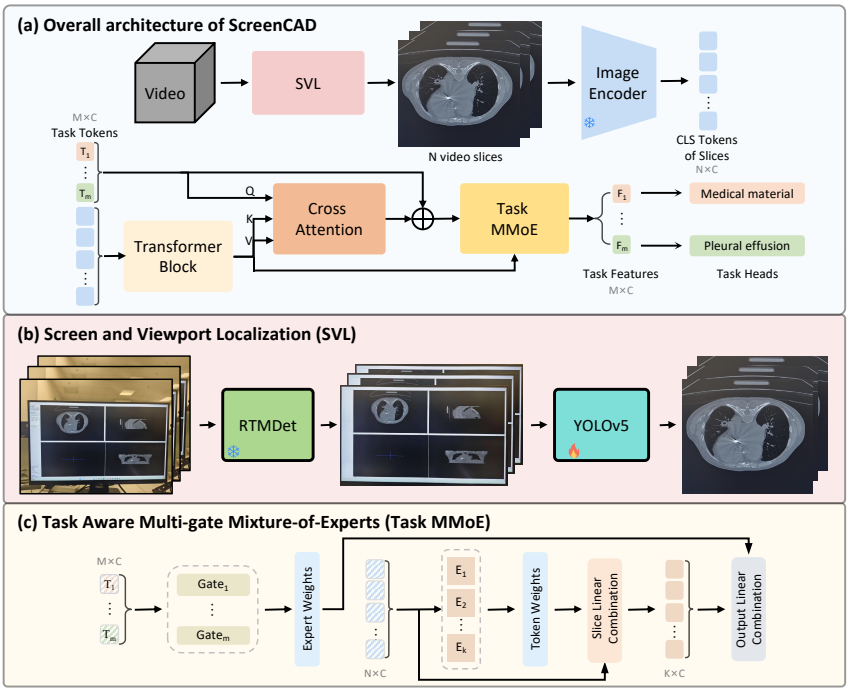

Fig. 1. Overview of proposed ScreenCAD. (a) The overall architecture of ScreenCAD. The pipeline ingests a raw video, extracts valid frames via SVL module, encodes them into slice embeddings, and aggregates them for diagnosis using our Task-Token MMoE Aggregation method. (b) Screen and Viewport Localization (SVL). (c) Task Aware Multi-gate Mixture-of-Experts (Task MMoE).

nosis. Rather than attempting to restore degraded pixels, ScreenCAD leverages the semantic robustness of a frozen medical foundation model and aggregates variable-length slice sequences through a Task-Token Mixture-of-Experts architecture. Evaluated on CT-RATE [12] with real-world screen recordings, ScreenCAD achieves diagnostic performance that closely matches its clean-DICOM counterpart, despite substantial visual degradation in the input. These findings provide the first empirical evidence that 3D volumetric diagnosis may remain feasible under severe information loss, and they highlight camera-based acquisition as a practical and scalable interface for future medical AI systems.

2 Methodology

2.1 Video Acquisition via Agent-Based Slice Browsing

As illustrated in Figure 1a, to systematically study whether a 3D medical volume retains sufficient semantic information when viewed through a camera, we construct a controlled acquisition setup that mimics how clinicians browse volumetric studies. An automated agent interacts with a standard DICOM viewer,

scrolling through the axial slices of each CT volume using mouse-wheel events. The agent follows a human-like browsing pattern: it scrolls forward through the volume at non-uniform speeds, occasionally pauses on diagnostically relevant slices, and intermittently reverses direction to simulate re-examination. Throughout this process, an external camera is fixed in front of the monitor and captures the screen at a constant frame rate.

Because the scrolling speed varies, each anatomical slice in the underlying 3D volume corresponds to a variable number of video frames (typically 1–5). This setup produces a controlled yet variable camera-captured video stream with stochastic scrolling, pauses, and limited reversals. The resulting video serves as the raw input to our system.

2.2 Screen and Viewport Localization

A camera-captured video frame contains not only the medical image but also environmental background, monitor bezels, UI elements, and software toolbars. To isolate the true slice content, we perform a two-stage localization procedure.

First, a pre-trained RTMDet model detects the physical monitor within the scene. We extract the tightest quadrilateral around the detected screen and apply a perspective homography to rectify it into a canonical front-facing view. This step removes camera-angle distortion and standardizes the spatial geometry. Second, a YOLOv5 detector, fine-tuned on annotated screen frames, localizes the medical imaging viewport within the rectified monitor. This removes toolbars, patient lists, and other UI clutter, yielding a clean sequence of cropped images that correspond to the slices viewed during scrolling. These viewport crops form the basis for subsequent temporal segmentation and feature extraction.

2.3 Slice-Level Feature Encoding

Each viewport crop is processed by a frozen MedDINOv3 [5] backbone to obtain a frame-level embedding. We select this foundation model for two key reasons. First, it is built upon DINOv3 [13], which follows DINOv2 [14] to enforce representation invariance across aggressive data augmentations. This mechanism implicitly trains the model to ignore low-level degradations, making it naturally robust to the Moiré patterns and compression artifacts inherent in screen recording. Second, MedDINOv3 performs domain-adaptive pretraining on large-scale CT datasets, ensuring that its features capture semantically meaningful anatomical structures rather than generic natural image patterns.

Let r_i denote the i-th cropped frame and $s_i \in \mathbb{R}^C$ its corresponding embedding. To preserve the full temporal context of the scrolling action, including pauses, reversals, and speed variations, we retain the complete sequence of video frames without downsampling or thresholding. This treats the video analysis as a Multiple Instance Learning (MIL) problem. The setup is analogous to Whole Slide Imaging in pathology, where diagnosis relies on aggregating features from constituent patches [15, 16]. This results in a slice sequence $\{r_1, \ldots, r_N\}$ and

corresponding embeddings $S = \{s_1, \ldots, s_N\}$ that serve as the direct input to our aggregation network.

2.4 Task-Token MMoE Aggregation and Training Objective

Given the slice embeddings $S \in \mathbb{R}^{N \times C}$, ScreenCAD aggregates them into task-specific volumetric representations using a Task-Token Multi-gate Mixture-of-Experts (MMoE) architecture.

First, we process the sequence S with a Transformer Encoder to capture inter-slice dependencies, yielding context-aware slice features $\hat{S}$. To probe this volume for M specific pathologies, we introduce a set of learnable Task Tokens $T = \{T_1, \ldots, T_M\} \in \mathbb{R}^{M \times C}$. A Multi-Head Cross-Attention layer uses T as queries and $\hat{S}$ as keys/values, allowing each task token to selectively gather initial pathology-relevant information from the slice sequence, producing enhanced tokens $T' = \{T'_1, \ldots, T'_M\}$.

To obtain the final volume-level representation, we employ an MMoE layer comprising K Slice-Attention Experts $\{E_1, \ldots, E_K\}$. Each expert E_k independently computes a global feature vector $v_k \in \mathbb{R}^C$ by performing a weighted temporal pooling over $\hat{S}$. Specifically, the attention weights $\alpha_k \in \mathbb{R}^N$ for the k-th expert are computed as:

$$\alpha_k = \text{Softmax}\left(W_k^{(2)}\text{GELU}\left(W_k^{(1)}\hat{S} + b_k^{(1)}\right)\right), \tag{1}$$

where $W_k^{(1)}, b_k^{(1)}$ and $W_k^{(2)}$ are learnable parameters. The expert output is then the weighted sum $v_k = \sum_{j=1}^{N} \alpha_{k,j}\hat{S}_j$.

Simultaneously, for each task m, a gating network G_m takes the corresponding enhanced task token T'_m as input and outputs a softmax probability distribution over the K experts. The final task feature F_m is a weighted sum of the experts' global vectors, fused with the residual task token:

$$F_m = \text{LayerNorm}\left(T'_m + \sum_{k=1}^{K} G_m(T'_m)_k \cdot v_k\right). \tag{2}$$

Each task-specific feature F_m is passed to a linear classifier to produce the diagnostic probability $\hat{y}_m$.

Each diagnostic task m is a binary classification problem. We optimize the entire aggregation network (with the MedDINOv3 backbone frozen) using a filtered binary cross-entropy loss summed over all M tasks:

$$\mathcal{L} = \sum_{m=1}^{M} \left[-y_m \log(\hat{y}_m) - (1 - y_m)\log(1 - \hat{y}_m)\right] \tag{3}$$

where $y_m \in \{0, 1\}$ is the label and $\hat{y}_m = \sigma(Y_m)$ is the predicted probability.

3 Experiments

3.1 Datasets and Experimental Setup

Dataset. We validated on CT-RATE [12], selecting a training set of 5,447 chest CT volumes from patients exhibiting at least one positive label for our target pathologies (2,885 positive for Medical Material, 2,857 for Pleural Effusion), and using the official validation set (1,564 scans) for testing. Medical Material was present in 10.3% and Pleural Effusion in 12.4% of the test set. To evaluate the domain gap, we constructed a parallel video dataset by recording both subsets with a Microsoft Azure Kinect RGB camera (2560×1440, 30fps) filming a Dell D2720DS monitor (2560×1440 resolution) running ITK-SNAP, varying camera angles (±15°) and distances (0.5–0.6 m). The video train/test split mirrors the DICOM split at the patient level. All recordings used de-identified data. While this controlled setup does not capture the full variability of clinical screen-sharing (diverse phones, monitors, PACS viewers, lighting), it provides an initial reproducible testbed for measuring the domain gap.

Implementation Details. For feature extraction, we resize all cropped viewports to 512×512 pixels before feeding them into the frozen MedDINOv3 (ViT-B/16) backbone, which outputs C=768-dimensional CLS embeddings. The downstream aggregator employs a 2-layer Transformer Encoder with 8 heads, M=2 task tokens, and K=4 experts (hidden dimension 256), totaling approximately 14.9M trainable parameters. We train for 50 epochs using the AdamW optimizer (learning rate 5×10^{-4}, weight decay 1×10^{-4}, cosine annealing schedule) with a batch size of 16 on a single NVIDIA A100 GPU. ROI localization utilizes RTMDet (pre-trained on COCO, frozen) for monitor detection and a YOLOv5 (fine-tuned on 10,000 annotated frames) for viewport extraction. Performance is evaluated using Area Under the ROC Curve (AUC), Accuracy (ACC), and F1-score (F1) with 95% confidence intervals reported.

Baselines and Variants. We define three evaluation protocols: Clean (train/test on DICOM), Video ZS (Zero-shot, train on DICOM, test on video), and Video FT (Fine-tuned, train/test on video). We compare against: (1) **CT-Net** [17], a ResNet-18 + 3D convolution architecture for multi-abnormality classification; (2) **CT-CLIP** [12], a zero-shot 3D foundation model using contrastive learning; (3) **Ours (ABMIL)**: replacing our aggregator with Attention-Based MIL [18] over MedDINOv3 embeddings, a strategy recently benchmarked for CT-RATE analysis [19]; and (4) **Ours (Task Token)**: our full method. All methods receive the same SVL-preprocessed viewport crops; CT-Net sequences are resized to 420×420×402 and CT-CLIP sequences are temporally sampled and padded/center-cropped to 480×480×240, while ScreenCAD natively processes variable-length sequences.

Table 1. Quantitative comparison for Medical Material Detection across three evaluation protocols.

Method	Metric	Clean	Video ZS	Video FT
CT-Net [17]	AUC	0.664 [.621, .699]	0.530 [.491, .575]	0.544 [.498, .582]
	ACC	**0.697** [.675, .725]	0.526 [.488, .571]	0.530 [.485, .570]
	F1	0.751 [.718, .779]	0.460 [.417, .498]	0.592 [.556, .637]
CT-CLIP [12]	AUC	0.733 [.691, .769]	0.540 [.493, .582]	0.631 [.590, .670]
	ACC	0.661 [.635, .703]	0.536 [.500, .580]	0.592 [.547, .638]
	F1	0.729 [.718, .763]	0.670 [.628, .707]	0.770 [.736, .799]
Ours (ABMIL)	AUC	0.638 [.583, .702]	0.581 [.541, .619]	0.609 [.570, .643]
	ACC	0.597 [.548, .641]	0.571 [.528, .609]	0.630 [.589, .662]
	F1	0.723 [.694, .748]	0.762 [.733, .789]	0.702 [.663, .736]
Ours (Task Token)	AUC	**0.740** [.695, .780]	**0.722** [.677, .759]	**0.723** [.679, .760]
	ACC	0.654 [.607, .696]	**0.629** [.584, .669]	**0.670** [.626, .709]
	F1	**0.795** [.769, .816]	**0.850** [.818, .877]	**0.781** [.750, .807]

3.2 Resilience to Domain Shift

We evaluated two complementary tasks chosen to probe distinct failure modes: Medical Material detection (metallic implants with high-frequency features) and Pleural Effusion detection (low-contrast soft tissue). As shown in Table 1 and Table 2, conventional baselines perform acceptably on clean DICOMs but suffer substantial degradation when exposed to video inputs. Specifically, CT-Net's AUC for Pleural Effusion drops from 0.772 to 0.532, while the zero-shot CT-CLIP drops from 0.854 to 0.561. These sharp declines indicate that standard 3D architectures trained on pristine voxels struggle to generalize to the temporal and spatial artifacts introduced by screen recording.

While CT-Net and CT-CLIP remain competitive on native DICOM inputs, our primary focus is achieving robustness under the video domain shift. In the video setting, ScreenCAD demonstrates stronger zero-shot robustness. For Medical Material, our method retains an AUC of 0.722 on video (*vs.* 0.740 on clean), outperforming CT-CLIP's 0.540. Crucially, for Pleural Effusion, our zero-shot video AUC (0.883) matches our clean DICOM performance, suggesting that our Task-Token aggregator successfully extracts invariant semantic features that are more robust to the camera capture process. Furthermore, while baselines require fine-tuning simply to regain basic functionality (*e.g.*, CT-Net recovering from 0.532 to 0.733), fine-tuning ScreenCAD pushes performance even higher, reaching 0.896 AUC for Pleural Effusion.

3.3 Ablation Study

Table 3 details the contribution of each component to the Video (zero-shot) performance. Removing the Transformer Block causes the most significant performance drop ($\Delta = -0.045$), bringing the AUC down to 0.761. This confirms

Table 2. Quantitative comparison for Pleural Effusion Detection across three evaluation protocols.

Method	Metric	Clean	Video ZS	Video FT
CT-Net [17]	AUC	0.772 [.741, .794]	0.532 [.477, .582]	0.733 [.694, .766]
	ACC	0.747 [.713, .769]	0.530 [.503, .563]	0.720 [.683, .772]
	F1	0.790 [.766, .809]	0.333 [.285, .386]	0.749 [.712, .779]
CT-CLIP [12]	AUC	0.854 [.829, .896]	0.561 [.516, .603]	0.711 [.669, .746]
	ACC	0.856 [.794, .905]	0.550 [.513, .575]	0.693 [.653, .728]
	F1	**0.842** [.819, .860]	0.513 [.465, .555]	**0.849** [.812, .874]
Ours (ABMIL)	AUC	0.852 [.817, .882]	0.801 [.770, .820]	0.819 [.781, .859]
	ACC	0.756 [.718, .789]	0.759 [.720, .793]	0.750 [.706, .789]
	F1	0.804 [.783, .821]	**0.873** [.849, .892]	0.802 [.771, .828]
Ours (Task Token)	AUC	**0.880** [.856, .902]	**0.883** [.854, .898]	**0.896** [.870, .917]
	ACC	**0.878** [.849, .895]	**0.814** [.774, .849]	**0.817** [.788, .841]
	F1	0.834 [.812, .851]	0.862 [.837, .882]	0.823 [.807, .841]

Table 3. Ablation Study. Average AUC on Video (zero-shot) across both tasks.

Configuration	Avg. AUC	Δ
Full model	**0.806**	–
w/o MMoE (Task Token + Linear probing)	0.788	−0.018
w/o Task Token (Shared query)	0.780	−0.026
w/o Transformer Block	0.761	−0.045
w/o Cross-Attention	0.793	−0.013

that modeling inter-slice dependencies is crucial for handling the variable speeds and fragmentation inherent in screen-recorded video. The Task Token mechanism is the next most critical component; replacing it with a shared query reduces AUC to 0.780 ($\Delta = -0.026$), validating our hypothesis that probing distinct pathologies requires attending to specific temporal subsets. Finally, simplifying the architecture by removing the MMoE or Cross-Attention layers yields smaller but consistent penalties, further highlighting the benefit of specialized experts and context-aware aggregation in noisy environments.

4 Conclusion

In this study, we presented ScreenCAD, a PACS-agnostic framework that enables volumetric diagnosis directly from degraded screen-recorded video by combining cascaded ROI extraction, a frozen medical foundation model, and a Task-Token MMoE aggregator. Our experiments on CT-RATE demonstrate that standard 3D diagnostic models suffer substantial performance degradation when applied to video inputs, confirming that the domain gap introduced by camera capture

is non-trivial. In contrast, ScreenCAD achieves competitive diagnostic performance across both soft-tissue and high-frequency pathologies without requiring image restoration or retraining on video data. These results provide initial empirical evidence that structure-level features from a frozen foundation model can partially bridge the screen-capture domain gap, highlighting camera-based acquisition as a promising interface for future medical AI systems in settings where PACS integration is infeasible.

This first testbed is limited to CT-RATE, two pathologies, one camera/viewer setup, and an agent-controlled browsing protocol. Future validation should cover clinician-recorded videos, additional devices/viewers, larger cohorts, and broader modalities including ultrasound, where HDMI capture offers a degradation-free alternative. Screen-recorded medical data must also be de-identified, consented, institutionally approved, securely stored, and compliant with local regulations.

Acknowledgments. This work was partially supported by the AI4S Initiative and HPC Platform of ShanghaiTech University.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chang, Y.W., Ryu, J.K., An, J.K., Choi, N., Park, Y.M., Ko, K.H., Han, K.: Artificial intelligence for breast cancer screening in mammography (AI-STREAM): preliminary analysis of a prospective multicenter cohort study. Nature Communications **16**(1), 2248 (2025)
2. Dembrower, K.E., Crippa, A., Eklund, M., Strand, F.: Human-AI interaction in the screentrustcad trial: recall proportion and positive predictive value related to screening mammograms flagged by AI CAD versus a human reader. Radiology **314**(3), e242566 (2025)
3. Ma, K., Zheng, M., Chen, W., Qi, Y., Rong, H.: Research progress in computer-aided diagnosis systems for lung cancer. npj Digital Medicine **8**(1), 722 (2025)
4. Huang, H., Shen, Z., Wang, J., Wang, X., Lu, J., Lin, H., Ge, J., Zuo, C., Wang, Q.: MetaAD: Metabolism-aware anomaly detection for parkinson's disease in 3D ^{18}F-FDG PET. In: Linguraru, M.G., Dou, Q., Feragen, A., Giannarou, S., Glocker, B., Lekadir, K., Schnabel, J.A. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 291–301. Springer Nature Switzerland, Cham (2024)
5. Li, Y., Wu, Y., Lai, Y., Hu, M., Yang, X.: Meddinov3: How to adapt vision foundation models for medical image segmentation? (2025), https://arxiv.org/abs/2509.02379
6. Li, Y., Hu, M., Yang, X.: Polyp-sam: transfer sam for polyp segmentation. In: Medical Imaging (2023), https://api.semanticscholar.org/CorpusID:258426990
7. Wang, Y., Xiong, H., Sun, K., Liu, J., Lin, X., Chen, Z., He, Y., Wang, Q., Shen, D.: Unisyn: A Generative Foundation Model for Universal Medical Image Synthesis across MRI, CT and PET . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. vol. LNCS 15962. Springer Nature Switzerland (September 2025)

8. Li, J., Wu, J., Wang, S., Xiong, H., Cai, J., Zhao, Z., Zhu, Y., Yin, Y., Shen, D., Wang, Q.: Visioncad: An integration-free radiology copilot framework (2025), https://arxiv.org/abs/2511.00381
9. Burns, J.L., Gichoya, J.W., Kohli, M.D., Jones, J.F., Purkayastha, S.: Theory of radiologist interaction with instant messaging decision support tools: A sequential-explanatory study. PLOS Digital Health **3** (2023), https://api.semanticscholar.org/CorpusID:259198993
10. Nair, M., Svedberg, P., Larsson, I., Nygren, J.M.: A comprehensive overview of barriers and strategies for ai implementation in healthcare: Mixed-method design. PLOS ONE **19**(8), 1–27 (08 2024). https://doi.org/10.1371/journal.pone.0305949, https://doi.org/10.1371/journal.pone.0305949
11. Bhushan, I., Anandampillai, K., Agarwal, S.: Navigating complexities: agile digital health initiatives in developing countries. Oxford Open Digital Health **2**, oqae032 (08 2024). https://doi.org/10.1093/oodh/oqae032, https://doi.org/10.1093/oodh/oqae032
12. Hamamci, I.E., Er, S., Wang, C., et al.: Generalist foundation models from a multimodal dataset for 3d computed tomography. Nature Biomedical Engineering (2026). https://doi.org/10.1038/s41551-025-01599-y
13. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: Dinov3 (2025), https://arxiv.org/abs/2508.10104
14. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2024), https://arxiv.org/abs/2304.07193
15. Jiang, H., Huang, H., Cai, J., Xu, M., Shen, Z., Fei, M., Wang, X., Zhang, L., Wang, Q.: Multi-task Screening for Cervical Diseases via Feature Routing and Asymmetric Distillation . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. vol. LNCS 15973. Springer Nature Switzerland (September 2025)
16. Cao, M., Fei, M., Cai, J., Liu, L., Zhang, L., Wang, Q.: Detection-free pipeline for cervical cancer screening of whole slide images. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 243–252. Springer Nature Switzerland, Cham (2023)
17. Draelos, R.L., Dov, D., Mazurowski, M.A., Lo, J.Y., Henao, R., Rubin, G.D., Carin, L.: Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. Medical Image Analysis **67**, 101857 (2021). https://doi.org/https://doi.org/10.1016/j.media.2020.101857, https://www.sciencedirect.com/science/article/pii/S1361841520302218
18. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 2127–2136. PMLR (10–15 Jul 2018), https://proceedings.mlr.press/v80/ilse18a.html

19. Liu, C., Chen, Y., Shi, H., Lu, J., Jian, B., Pan, J., Cai, L., Wang, J., Yu, J., Gao, Z., Zhang, X., Bai, L., Zhang, Y., Li, J., Bercea, C.I., Ouyang, C., Chen, C., Xiong, Z., Wiestler, B., Wachinger, C., Duncan, J.S., Rueckert, D., Bai, W., Arcucci, R.: Does dinov3 set a new medical vision standard? benchmarking 2d and 3d classification, segmentation, and registration (2026), https://arxiv.org/abs/2509.06467