JANUS: Anatomy-Conditioned Gating for Robust CT Triage Under Distribution Shift

Lavsen Dahal[1,2 [0000−0002−8991−759X]], Yubraj Bhandari[1,3 [0009−0004−7279−4097]], Geoffrey Rubin[4 [0000−0002−3820−2500]], and Joseph Y. Lo[1,2 [0000−0002−9540−5072]]

[1] Center for Virtual Imaging Trials, RAI Labs, Department of Radiology, Duke University, Durham NC 27708, USA
[2] Electrical and Computer Engineering, Pratt School of Engineering, Duke University, Durham, NC 27708, USA
[3] Department of Mathematics, Trinity College of Arts & Sciences, Duke University, Durham, NC 27708, USA
[4] Department of Radiology and Imaging Sciences, University of Arizona College of Medicine, Tucson, AZ 85004, USA
{lavsen.dahal,joseph.lo}@duke.edu

Abstract. Automated CT triage requires models that are simultaneously accurate across diverse pathologies and reliable under institutional shift. While Vision Transformers provide strong visual representations, many clinically significant findings are defined by quantitative imaging biomarkers rather than appearance alone. We introduce **JANUS**, a physiology-guided dual-stream architecture that conditions visual embeddings on macro-radiomic priors via *Anatomically Guided Gating*. On the MERLIN test set (N=5,082), JANUS attains macro-AUROC 0.88 and AUPRC 0.74, outperforming all reproduced baselines. It generalizes to an external dataset (N=2,000; AUROC 0.87), with the largest gains on findings defined by size and attenuation as well as improved calibration on both datasets. We further quantify prediction suppression using the *Physiological Veto Rate* (PVR), showing that under domain shift JANUS reduces high-confidence false positives substantially more often than true positives. Together, these results are consistent with physically grounded conditioning that improves both discrimination and reliability in CT triage. Code is made publicly available at github repository https://github.com/lavsendahal/janus and model weights are at https://huggingface.co/lavsendahal/janus.

Keywords: Vision Transformers (ViTs) · CT Abdomen · Triage · Classification

1 Introduction

Automated CT triage seeks to flag clinically significant findings, many defined by quantitative thresholds rather than appearance alone: organomegaly by size, aneurysm by vessel caliber, calcific burden by HU density [12, 7, 17, 15]. Missing such measurement-defined findings can have major consequences, e.g., an

undetected aneurysm can rapidly escalate to a life-threatening event. As CT utilization grows and radiologist workloads intensify, reliable automated triage becomes a clinical necessity [13, 22].

CT foundation models have emerged to meet this need, learning rich visual representations via large-scale pretraining [14, 13, 2, 19], yet lack explicit mechanisms for the measurement cues that define many urgent findings [11]. Under distribution shift, this gap becomes consequential: underlying anatomy is protocol-invariant, but visual appearance can shift with scanner and reconstruction settings [8].

Prior attempts to bridge this gap incorporated only coarse organ-level statistics (mean volume and mean HU) as auxiliary inputs to visual embeddings [3], demonstrating that quantitative priors carry signal beyond visual features. *Macro-Radiomics* [5] enables a richer, clinically aligned measurement space, including organ-level descriptors, diameters, densitometry profiles, calcific burden, and fluid volumes. Unlike microscopic texture features, Macro-radiomics operate at the scale of radiological decision-making and are reliably computable for large, well-defined organs via modern segmentation models [21, 4].

Yet how to enforce such measurements as structural constraints remains unaddressed. Additive fusion lacks an explicit suppression mechanism: under distribution shift, visual models could latch onto spurious correlations [6] and concatenation may not encourage selective suppression when predictions conflict with available quantitative priors. Multiplicative gating [16, 10] provides this inductive bias: physical measurements modulate visual evidence through a learned bottleneck, structurally enabling suppression rather than merely shifting predictions.

We introduce **JANUS**, a physiology-guided architecture that operationalizes this principle: quantitative measurements should constrain visual predictions, not merely augment them.

Contributions.

- **Architecture.** JANUS integrates the full macro-radiomic phenotype space with a visual stream via *Anatomically Guided Gating*: a disease-specific multiplicative bottleneck that modulates visual evidence through scalar priors, providing an explicit suppression mechanism.
- **Metric.** We introduce the *Physiological Veto Rate* (PVR) to quantify suppression of high-confidence baseline false positives at fixed thresholds. To assess whether suppression is selective rather than uniform, we additionally report *veto selectivity*, defined as the ratio of false-positive suppression to true-positive suppression. On the external dataset, veto selectivity reaches 10.8×, indicating suppression is strongly concentrated on high-confidence false positives while largely preserving true positives.
- **Empirical.** JANUS achieves the highest macro-AUROC among reproduced baselines on MERLIN and maintains performance on an external dataset, while improving calibration (lower ECE) on both; results support macro-radiomic priors improving discrimination and reliability.

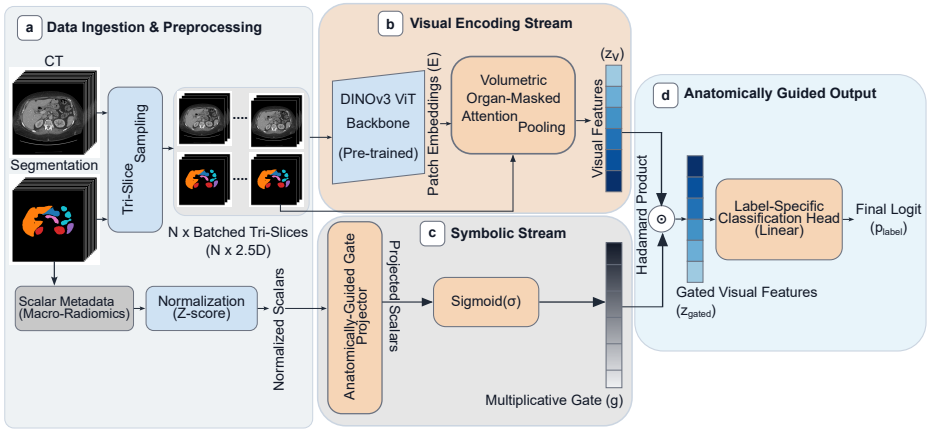

Fig. 1. JANUS Architecture. (a) A 3D CT volume is sampled into N 2.5D tri-slices; segmentation masks yield macro-radiomic scalar priors. **(b)** A DINOv3 backbone extracts patch embeddings condensed into a label-specific visual feature z_v via Organ-Masked Attention Pooling. **(c)** Scalar priors are projected and sigmoid-bounded to form a physiological gate g. **(d)** g modulates z_v via Hadamard product ($\odot$), acting as a *Physiological Veto*; gated features z_{gated} are passed to a label-specific linear head.

2 Methods

2.1 Problem Formulation and the Geometric Gap

We consider automated triage as a multi-label classification problem over a CT volume $X \in R^{D\times H\times W}$ with partially observed targets $\mathbf{y} \in \{0, 1, \emptyset\}^L$, where $\emptyset$ denotes missing labels. Let $\delta_\ell = 1[y_\ell \neq \emptyset]$ indicate label availability. Our goal is to learn a physiologically conditioned predictor $P(\mathbf{y} \mid X, S)$, where S encodes segmentation-derived macro-radiomic priors that provide explicit *quantitative measurement context* spanning anatomical scale and tissue densitometry.

We initialize the visual backbone with DINOv3 weights [20]. Standard attention pooling is invariant to token count by construction: geometric quantities such as organ volume or vessel diameter are encoded in how many tokens an organ occupies, not per-token appearance, and is thus not explicitly represented in the pooled embedding, the *pooling-induced geometric gap*. We address this by conditioning visual evidence on explicit scalar priors S derived from segmentation.

2.2 JANUS Architecture

We introduce **JANUS** (Fig. 1), a dual-stream physiology-guided model that couples (i) a visual stream extracting appearance features from X and (ii) a symbolic stream encoding macro-radiomic priors. JANUS fuses the two streams through a multiplicative gate that modulates the visual embedding using explicit anatomical measurements.

Visual Stream: 2.5D Tri-slice Encoding. Given X, we construct a 2.5D representation by sampling axial slice centers $\{t\}_{t=1}^{T}$ with stride s and forming a 3-channel input at each center by stacking adjacent slices: $[X_{t-1}, X_t, X_{t+1}]$ (boundary slices are replicated). Each tri-slice is resized to a fixed resolution and normalized with ImageNet statistics. Passing all tri-slices through a DINOv3 ViT produces patch token embeddings. We discard the CLS token and DINOv3 register tokens and reshape tokens as $\{u_i\}_{i=1}^{N}$ per tri-slice, yielding $u \in R^{B\times T\times N\times d}$, where B is the batch size, N is the number of spatial patch tokens per tri-slice, and d is the embedding dimension (d=768 for ViT-B).

Dilated ROI-Masked Attention Pooling. For each label ℓ, we define a disease-specific 3D ROI mask M_ℓ by composing segmentation channels relevant to that pathology: a single organ for focal findings (e.g., aorta for Abdominal Aortic Aneurysm), a multi-organ union where appropriate (e.g., liver and spleen for hepatic steatosis), or a localization box where no organ proxy exists (e.g., appendicitis). To capture peri-organ context (e.g., fat stranding), we apply a resolution-adaptive morphological dilation of M_ℓ by a physical radius r_ℓ (mm), using voxel spacing to convert millimeters to per-axis kernel sizes. We align the 3D ROI mask with the tri-slice sampling by extracting a 2D mask at each tri-slice center t and resizing it to the encoder input grid. We then downsample this ROI to the ViT patch grid via nearest-neighbor interpolation to obtain a binary token mask $m_{t,i,\ell} \in \{0, 1\}$ indicating whether spatial position i in tri-slice t lies inside the (dilated) ROI.

Given tokens $\{u_{t,i}\}$ where $t \in \{1, \ldots, T\}$ indexes tri-slices and $i \in \{1, \ldots, N\}$ indexes spatial positions, JANUS pools a label-specific visual embedding $z_{v,\ell}$ using ROI-masked attention:

$$
\begin{aligned}
a_{t,i,\ell} &= \frac{\phi_\ell(u_{t,i}) + \beta_{\text{in},\ell}\, m_{t,i,\ell}}{\tau_\ell}, \quad w_{t,i,\ell} = \text{softmax}_{i:\, m_{t,i,\ell}=1}(a_{t,i,\ell})\,, \\
z_{v,\ell} &= \frac{1}{T}\sum_{t=1}^{T}\sum_{i:\, m_{t,i,\ell}=1} w_{t,i,\ell}\, u_{t,i}.
\end{aligned}
\tag{1}
$$

Here $a_{t,i,\ell}$ denotes the unnormalized attention logit for token i at tri-slice t (label ℓ) prior to softmax normalization; $\phi_\ell(\cdot)$ is a learned linear scorer, $\beta_{\text{in},\ell}$ is a learnable logit bias for ROI tokens, and τ_ℓ is a learnable temperature. The softmax normalizes over spatial positions i within the ROI *independently per slice* t, so $\sum_{i:\, m_{t,i,\ell}=1} w_{t,i,\ell} = 1$ for each t. Outside tokens receive zero weight; if an ROI is empty for a sample, we fall back to uniform weights over all tokens.

Symbolic Stream: Anatomically Guided Gating. From the same anatomical masks used to define M_ℓ, we compute a disease-specific macro-radiomic prior vector $s_\ell \in R^{K_\ell}$. The feature list for each disease is fixed in the released code base and was selected by the scalar-only ML baseline using cross-validated logistic regression, rather than manually chosen for JANUS. Each feature is z-score normalized using training-set statistics.

We map the prior vector to a physiological gate $g_\ell \in [0,1]^d$ and modulate visual evidence by element-wise multiplication:

$$g_\ell = \sigma(W_{g,\ell} s_\ell + b_{g,\ell}), \qquad z_{\text{gated},\ell} = z_{v,\ell} \odot g_\ell, \tag{2}$$

where $W_{g,\ell} \in R^{d \times K_\ell}$ and $b_{g,\ell} \in R^d$ are learned. Because $g_\ell \in [0,1]^d$, the gate implements a bounded, dimension-wise re-weighting of the visual embedding conditioned on quantitative measurements, which can attenuate components of the representation when the priors provide countervailing evidence. We initialize $b_{g,\ell} = 2.0$ so that $g_\ell \approx 0.88$ at initialization, keeping the gate mostly open and ensuring stable gradient flow. This initialization biases the model toward preserving the visual stream unless training data support stronger gating, which can mitigate failure modes when scalar priors are weakly informative.

2.3 Prediction and Training Objective

For each label ℓ, we apply a disease-specific linear head to obtain logits $\hat{y}_\ell = \mathbf{w}_\ell^\top z_{\text{gated},\ell} + b_\ell$ and probabilities $p_\ell = \sigma(\hat{y}_\ell)$. We optimize a curriculum-weighted BCE over partially observed labels $y_\ell \in \{0, 1, \emptyset\}$:

$$\mathcal{L}^{(e)} = \frac{\sum_{\ell=1}^{L} \left[\delta_\ell \cdot \text{BCE}(p_\ell, y_\ell) + (1-\delta_\ell) \cdot w^{(e)} \cdot \text{BCE}(p_\ell, 0)\right]}{\sum_{\ell=1}^{L} \left[\delta_\ell + (1-\delta_\ell) \cdot w^{(e)}\right]} \tag{3}$$

where $w^{(e)}$ controls the contribution of unannotated labels: these labels are ignored for $E_{\text{ignore}} = 10$ epochs and then linearly ramped over $E_{\text{ramp}} = 10$ epochs to a final weight $w_{\max} = 0.3$. Thus, missing labels are not treated as confirmed negatives, but only as weak negatives later in training.

3 Experimental Setup

Datasets. We train and evaluate on MERLIN [2], a retrospective abdominal CT dataset (N=25,275; 30 labels; splits 15,175/5,018/5,082). We additionally evaluate on an external dataset from a different [redacted] U.S. hospital (N=2,000), with labels derived from radiology reports via dual-LLM consensus: a finding is positive only if both Qwen-3 [23, 1] and MedGemma [18] independently agree, and negative only if both confirm absence; abstaining studies are excluded, with explicit contradictions in fewer than 0.5% of cases. Organ masks are obtained via TotalSegmentator [21].

Baselines. We compare against three baselines: (i) *ViT-Baseline*, DINOv3 ViT-B/16 with global average pooling and no scalar integration; (ii) *ORACLE-CT* [3], a recent concurrent anatomy-aware baseline; and (iii) *ORACLE-CT+OSF* [3], which reported the strongest supervised results on the official MERLIN test split and injects scalar priors via projection+concatenation at the pooled embedding (originally mean volume/HU). In our experiments, ORACLE-CT+OSF uses the same macro-radiomic prior bank and backbone as JANUS, isolating prior usage (projection+concatenation vs. prior-modulated features).

Table 1. Global benchmarking. AUROC, AUPRC, and ECE (95% bootstrap CIs; n=1000 resamples) on MERLIN (internal) and Duke-Abdomen2026 (external) datasets (27 shared labels). **JANUS** achieves the strongest overall performance across metrics.

	MERLIN (Internal, $N = 5{,}082$)			Duke-Abdomen (External, $N = 2{,}000$)		
Model	AUC ↑	AUPRC ↑	ECE ↓	AUC ↑	AUPRC ↑	ECE ↓
ViT-Baseline [20]	0.84 (0.83–0.85)	0.66 (0.64–0.68)	0.13 (0.12–0.14)	0.84 (0.82–0.85)	0.67 (0.65–0.70)	0.16 (0.16–0.17)
ORACLE-CT [3]	0.86 (0.85–0.87)	0.69 (0.68–0.72)	0.15 (0.14–0.16)	0.85 (0.83–0.86)	0.69 (0.67–0.72)	0.19 (0.18–0.20)
ORACLE-CT+OSF [3]	0.84 (0.83–0.85)	0.66 (0.65–0.68)	0.17 (0.16–0.18)	0.81 (0.80–0.83)	0.66 (0.64–0.68)	0.22 (0.21–0.23)
JANUS (Ours)	**0.88** (0.87–0.89)	**0.74** (0.72–0.76)	**0.09** (0.09–0.11)	**0.87** (0.85–0.88)	**0.72** (0.71–0.75)	**0.15** (0.15–0.16)

Training. Volumes are clipped to $[-1000, 1000]$ HU, resampled to 1.5×1.5×3.0 mm, and center-cropped to 224×224×160, tri-slice stride (s = 1). Augmentations include 3D affine transforms, gamma jitter, and noise. DINOv3 ViT-B/16 [20] is trained end-to-end for 20 epochs on 4×A6000 GPUs (batch size 10).

Evaluation. All AUROC and AUPRC values are macro-averaged over labels. Calibration is measured via ECE (10 equal-width bins) [9] .

Physiological Veto Rate (PVR). For label ℓ, let $\mathcal{F}_\ell = \{i : \ y_{i\ell} = 0, \ p_{i\ell}^{\mathrm{ViT}} \geq 0.8\}$ denote high-confidence false positives under the ViT baseline. PVR measures the fraction of these baseline errors suppressed by JANUS below a decision threshold:

$$\mathrm{PVR}_\ell = \frac{1}{|\mathcal{F}_\ell|} \sum_{i \in \mathcal{F}_\ell} 1\big[p_{i\ell}^{\mathrm{JANUS}} < 0.5\big] . \tag{4}$$

We report the mean over labels with $|\mathcal{F}_\ell| \geq 5$. We additionally report the true-positive suppression rate $\mathrm{TSR}_\ell = \Pr(p_{i\ell}^{\mathrm{JANUS}} < 0.5 \mid y_{i\ell} = 1)$ and summarize *veto selectivity* as $\mathrm{PVR}_\ell/\mathrm{TSR}_\ell$, where values $\gg 1$ indicate suppression concentrated on baseline false positives rather than uniformly reducing confidence.

4 Results

4.1 Global Performance, Reliability, and Robustness

Overall performance. Table 1 reports global results; unless otherwise stated, ΔAUROC is measured on the external dataset relative to ViT-Baseline. ORACLE-CT shows organ-level context adds signal ($\Delta\mathrm{AUROC} = +0.02$). Additive scalar fusion (+OSF) improves size and density-driven findings but does not improve overall external macro-AUROC ($\Delta\mathrm{AUROC} = -0.03$) and has the highest external ECE. JANUS better leverages macro-radiomic priors via anatomically guided gating, achieving the strongest performance on both datasets (AUROC/AUPRC 0.88/0.74 internal, 0.87/0.72 external) with the lowest ECE on both.

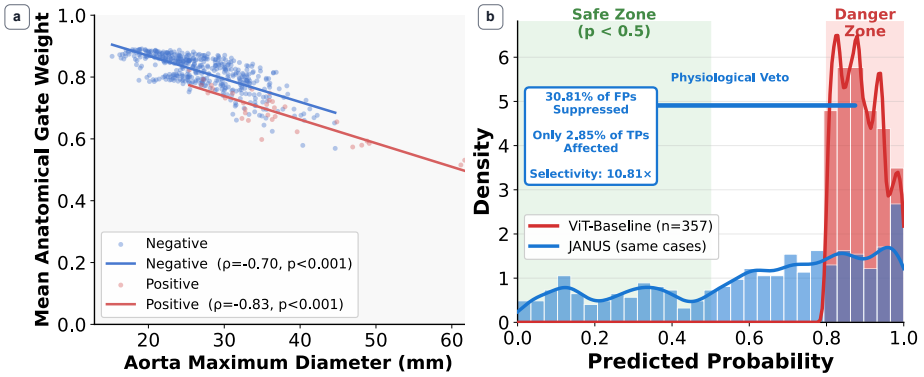

Fig. 2. External Dataset: Gate behavior and physiological veto. (a) Mean anatomical gate weight (mean of $g = \sigma(Ws + b) \in (0, 1)^d$) versus aortic maximum diameter for AAA negatives and positives; lower values indicate stronger down-weighting of visual features. **(b)** PVR on the external dataset: among overconfident ViT false positives ($p_{\text{ViT}} \geq 0.8$, n=357), JANUS reduces $p < 0.5$ more often for negatives than positives, consistent with selective suppression under shift.

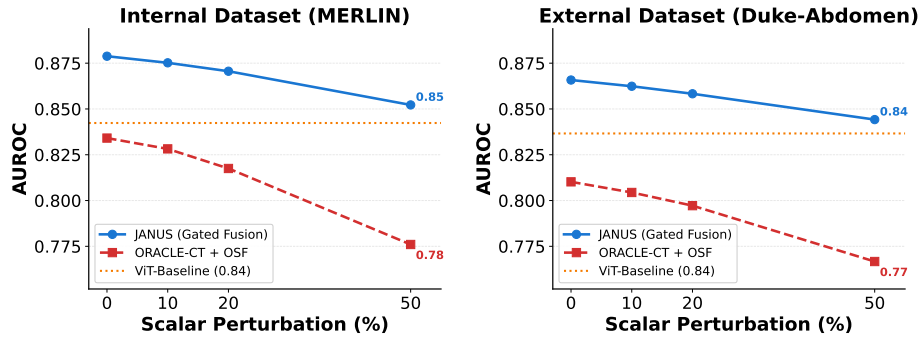

Fig. 3. Robustness to scalar corruption. AUROC under 10%, 20%, and 50% corruption of macro-radiomic priors on internal and external datasets. JANUS degrades gracefully and remains above the ViT-Baseline (dotted) at all corruption levels, including 50%, suggesting the multiplicative gate bounds the influence of corrupted inputs.

Gate Mechanism and Physiological Veto. Figure 2(a) examines how the learned gate varies with a quantitative biomarker on the external dataset. Using Abdominal Aortic Aneurysm (AAA) as a probe, the *mean anatomical gate weight* (mean of $g = \sigma(Ws + b) \in (0, 1)^d$) decreases monotonically with aortic maximum diameter within both label groups, consistent with sensitivity to the scalar measurement rather than simply mirroring disease status. Across all 27 pathologies (Fig. 2b), among baseline overconfident false positives, JANUS reduces 30.8% below threshold while affecting only 2.85% of true positives (10.8×

Table 2. Stratified performance by pathology. AUROC on MERLIN (internal) and external (Duke-Abdomen2026) datasets across 27 shared labels. Group rows report category means; two representative diseases are shown per group. **Gain** (ΔAUROC, JANUS vs. ViT-Baseline, external) is largest where physiology scalars are well-defined (Geometric, Densitometric, Chronic). As expected, degradation is minimal for Focal/Control Group, where scalar priors are inherently uninformative.

	MERLIN				**EXTERNAL**				**Gain**
Pathology	ViT	ORACLE	+OSF	**JANUS**	ViT	ORACLE	+OSF	**JANUS**	(Ext.)
Geometric (7)	0.89	0.92	0.90	**0.94**	0.86	0.89	0.88	**0.91**	+0.05
Prostatomegaly	0.80	0.84	0.88	0.89	0.74	0.83	0.89	0.88	+0.14
Aortic Aneurysm	0.83	0.91	0.94	0.98	0.88	0.87	0.95	0.96	+0.08
Densitometric (5)	0.83	0.84	0.88	**0.90**	0.82	0.83	0.89	**0.88**	+0.06
Hepatic Steatosis	0.79	0.82	0.95	0.96	0.79	0.81	0.88	0.88	+0.09
Gallstones	0.72	0.72	0.80	0.84	0.54	0.62	0.80	0.81	+0.27
Fluid/Global (3)	0.97	0.97	0.93	**0.97**	0.98	0.98	0.95	**0.98**	0.0
Pleural Effusion	0.97	0.97	0.93	0.97	0.98	0.98	0.94	0.98	0.0
Ascites	0.95	0.95	0.89	0.95	0.97	0.97	0.92	0.97	0.0
Chronic (6)	0.85	0.87	0.85	**0.90**	0.82	0.83	0.83	**0.87**	+0.05
Renal Cyst	0.81	0.82	0.82	0.89	0.81	0.75	0.87	0.95	+0.14
Pancreatic Atrophy	0.80	0.87	0.90	0.90	0.72	0.84	0.81	0.83	+0.11
Focal/Control (6)	0.73	**0.75**	0.65	0.73	0.76	**0.75**	0.57	0.74	-0.02
Appendicitis	0.66	0.67	0.59	0.65	0.85	0.86	0.53	0.81	-0.04
Fracture	0.78	0.80	0.68	0.75	0.80	0.84	0.75	0.78	-0.02

selectivity). Because PVR is computed per label, operating points could be set per pathology to trade off false-positive reduction against true-positive retention.

Scalar robustness. To stress-test sensitivity to scalar quality, we corrupt each feature by uniform random noise scaled to $p\%$ of its value at inference ($p \in \{10, 20, 50\}$), spanning mild to deliberately extreme degradation. JANUS degrades more gracefully across all levels on both datasets, remaining above the ViT baseline even at p=50 (external AUROC 0.84) while +OSF approaches it (0.77), consistent with the sigmoid-bounded gate providing a structural bound on corrupted inputs (Figure 3).

4.2 Stratified Analysis by Pathology

Table 2 stratifies performance by diagnostic mechanism, grouping findings by the type of physical signal available to the gate.

Measurement-defined targets. JANUS yields the largest gains on geometric and densitometric findings, where scalar measurements directly encode the diagnostic criterion (group means ΔAUROC = +0.05 and = +0.06 externally). Fluid/Global findings saturate near 0.98 AUROC across all models, leaving no headroom. Chronic targets show consistent improvement (ΔAUROC = +0.05), with Renal Cyst benefiting most (ΔAUROC = +0.14).

Focal targets and non-local priors. Focal pathologies provide a natural stress test for macro-radiomic priors, which summarize organ-level or global measure-

ments and are not designed to capture fine spatial localization. Accordingly, JANUS shows only a small change on these targets (ΔAUROC $= -0.02$). In contrast, +OSF decreases substantially (ΔAUROC $= -0.19$), suggesting that directly fusing non-local scalar features can yield an unfavorable accuracy.

5 Conclusion

We introduced **JANUS**, a physiology-guided framework that integrates a comprehensive macro-radiomic prior bank with organ-localized visual representations via disease-specific multiplicative gating. Under a matched training and evaluation protocol using released baseline code, JANUS achieves the highest macro-AUROC on MERLIN (0.88) and generalizes under distribution shift to an external dataset (AUROC 0.87), with substantially improved calibration on both datasets, suggesting an association between physical grounding and imporved accuracy and reliability.

Limitations. JANUS relies on a sequential pipeline where segmentation precedes gating, limiting guidance for focal findings where organ-level priors provide limited signal. Scalar robustness is evaluated via simulated corruption rather than realistic segmentation errors, and evaluation with CT-pretrained backbones remains future work.

Acknowledgments. This work was funded by the Center for Virtual Imaging Trials, NIH grants P41EB028744, R01EB001838, and R01CA261457.

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

References

1. Agrawal, K.K., Liu, L., Lian, L., Nercessian, M., Harguindeguy, N., Wu, Y., Mikhael, P., Lin, G., Sequist, L.V., Fintelmann, F., et al.: Pillar-0: A new frontier for radiology foundation models. arXiv preprint arXiv:2511.17803 (2025)
2. Blankemeier, L., Cohen, J.P., Kumar, A., Van Veen, D., Gardezi, S.J.S., Paschali, M., Chen, Z., Delbrouck, J.B., Reis, E., Truyts, C., et al.: Merlin: A vision language foundation model for 3d computed tomography. Research Square pp. rs–3 (2024)
3. Dahal, L., Bhandari, Y., Rubin, G.D., Lo, J.Y.: Organ-aware attention improves ct triage and classification. arXiv preprint arXiv:2601.13385 (2026)
4. Dahal, L., Ghojoghnejad, M., Vancoillie, L., Ghosh, D., Bhandari, Y., Kim, D., Ho, F.C., Tushar, F.I., Luo, S., Lafata, K.J., et al.: Xcat 3.0: A comprehensive library of personalized digital twins derived from ct scans. Medical Image Analysis **103**, 103636 (2025)
5. Dahal, L., Lo, J.Y.: Ct-idp: Segmentation-derived quantitative phenotypes for interpretable abdominal ct disease classification (2026), https://arxiv.org/abs/2605.09002
6. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)

7. Gücük, A., Üyetürk, U.: Usefulness of hounsfield unit and density in the assessment and treatment of urinary stones. World journal of nephrology **3**(4), 282 (2014)
8. Guo, B., Lu, D., Szumel, G., Gui, R., Wang, T., Konz, N., Mazurowski, M.A.: The impact of scanner domain shift on deep learning performance in medical imaging: an experimental study. arXiv preprint arXiv:2409.04368 (2024)
9. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)
10. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
11. Li, J., Chen, J., Tang, Y., Wang, C., Landman, B.A., Zhou, S.K.: Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives. Medical image analysis **85**, 102762 (2023)
12. Linguraru, M.G., Sandberg, J.K., Jones, E.C., Petrick, N., Summers, R.M.: Assessing hepatomegaly: automated volumetric analysis of the liver. Academic radiology **19**(5), 588–598 (2012)
13. Momin, E., Cook, T., Gershon, G., Barr, J., De Cecco, C.N., van Assen, M.: Systematic review on the impact of deep learning-driven worklist triage on radiology workflow and clinical outcomes. European radiology **35**(11), 6879–6893 (2025)
14. Pai, S., Hadzic, I., Bontempi, D., Bressem, K., Kann, B.H., Fedorov, A., Mak, R.H., Aerts, H.J.: Vision foundation models for computed tomography. arXiv preprint arXiv:2501.09001 (2025)
15. Paudyal, R., Shah, A.D., Akin, O., Do, R.K., Konar, A.S., Hatzoglou, V., Mahmood, U., Lee, N., Wong, R.J., Banerjee, S., et al.: Artificial intelligence in ct and mr imaging for oncological applications. Cancers **15**(9), 2573 (2023)
16. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
17. Radiopaedia.org: Splenomegaly, https://radiopaedia.org/articles/splenomegaly, accessed: 2026-02-15
18. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
19. Shui, Z., Zhang, J., Cao, W., Wang, S., Guo, R., Lu, L., Yang, L., Ye, X., Liang, T., Zhang, Q., et al.: Large-scale and fine-grained vision-language pre-training for enhanced ct image understanding. arXiv preprint arXiv:2501.14548 (2025)
20. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
21. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5), e230024 (2023)
22. Winder, M., Owczarek, A., Chudek, J., Pilch-Kowalczyk, J., Baron, J.: Are we overdoing it? changes in diagnostic imaging workload during the years 2010-2020 including the impact of the sars-cov-2 pandemic. healthcare (basel) (2021)
23. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)