

SpatialUQ: Post-Hoc Uncertainty Quantification from Spatial Consistency in Black-Box Vision Models

Md Kawsher Mahbub¹ Milon Biswas² Mirza Niaz Morshed³ Wei Yu²

¹NPI University of Bangladesh, ²Towson University, ³University of Asia Pacific

kawsher@npiub.edu.bd, mbiswas1@students.towson.edu, 25214015@uap-bd.edu, wyu@towson.edu

Abstract

Clinical vision models are often deployed as frozen black boxes with no access to internals, retraining, or ground truth at inference time. We introduce **SpatialUQ**, a post-hoc uncertainty method using only output probabilities. It measures the Jensen-Shannon divergence between the global prediction and the mean of five fixed spatial crops in six deterministic forward passes. The premise is simple, trustworthy predictions are spatially consistent. On NIH ChestX-ray14 (DenseNet-121, $N=25,596$), our Multicrop Uncertainty Score (MUS) reaches 0.784 failure-detection AUC versus 0.664 for MC-Dropout ($p < 10^{-6}$) at one-fifth the compute, with native calibration (SCE=0.049 vs. 0.127 for ℓ_1), the best-calibrated among methods above 0.78 AUC. A supervised fusion of MUS with entropy, confidence, and ℓ_1 reaches 0.832, outperforming a five-member ensemble (0.813). MUS scales with model quality, reaching 0.899 with BiomedCLIP ($\rho = 0.846$), while this relationship remains meaningful in-distribution ($\rho = 0.523$) but breaks down under severe distribution shift (VinBigData, $\rho = 0.027$). MUS is well-suited to diffuse findings but is less dependable for small focal lesions such as nodules. Code and experimental materials are publicly available at <https://huggingface.co/datasets/kawsher11/SpatialUQ>.

black-box classifiers exclude many existing UQ approaches. Bayesian methods such as MC-Dropout (Gal and Ghahramani, 2016) require training-time modifications, ensembles (Lakshminarayanan et al., 2017) and TTA (Krizhevsky et al., 2012) add substantial compute, while post-hoc methods such as DDU (Mukhoti et al., 2023), Mahalanobis (Lee et al., 2018), energy (Liu et al., 2020), ODIN (Liang et al., 2017), and conformal prediction (Angelopoulos and Bates, 2023) require model internals, feature distributions, or calibration data. Recent consistency-based (Khan and Fu, 2024) and augmentation-sensitivity (Shu et al., 2026) methods operate in related settings but rely on linguistic or photometric perturbations, leaving geometric evidence largely unexplored.

We propose **SpatialUQ** grounded on a single premise: a reliable model produces *spatially consistent* predictions across an image and its sub-regions. Well-localized features create strong agreement between global and crop-level predictions. On the other hand, distributional mismatch, low image quality, or atypical presentations make them diverge. We measure this divergence as the Jensen-Shannon Divergence (JSD) between the global prediction and the mean of five fixed spatial crops, in six deterministic forward passes. This geometric signal needs no internals, sampling, or reference data, and differs fundamentally from TTA (Krizhevsky et al., 2012; Wang et al., 2019) (Lemma 1; Section 5.4.1). A supervised logistic fusion of MUS with entropy, confidence, and ℓ_1 improves failure detection further, except under severe shift where MC-Dropout remains preferable.

Our key contributions are summarized as follows:

1 INTRODUCTION

Deep neural networks can fail silently on out-of-distribution inputs (Amodei et al., 2016; Hendrycks and Dietterich, 2019), posing patient-safety risks in medical imaging where confident errors are consequential (Leibig et al., 2017; Kendall and Gal, 2017). Reliable deployment therefore requires inference-time failure detection without ground truth. However, frozen

- **SpatialUQ:** A post-hoc uncertainty framework for frozen black-box classifiers using only output probabilities, with no internals, gradients, training data, or retraining. MUS is natively calibrated and extends to frozen foundation models, reaching 0.899 AUC with BiomedCLIP ($\rho = 0.846$). Spatial perturbations provide the key signal, with photo-

metric JSD reducing AUC from 0.784 to 0.755.

- **Formalization and self-diagnosing deployment gate:** A formal bound connects MUS to mean absolute spatial probability shift, while a small labeled calibration sample provides a deployment diagnostic. The correlation remains meaningful in-distribution ($\rho = 0.523$) but collapses under severe shift (VinBigData, $\rho = 0.027$).
- **State-of-the-art failure detection at one-fifth the cost:** MUS achieves 0.784 AUC on NIH ChestX-ray14, outperforming MC-Dropout (0.664, $p < 10^{-6}$) with 6 vs. 30 forward passes. Fusion reaches 0.832 AUC, exceeding a five-member ensemble (0.813) without retraining.
- **Cross-domain transferability:** MUS achieves 0.708 AUC on CheXpert, a 0.109 gain over MC-Dropout; fusion reaches 0.917–0.937 on ImageNet and 0.885–0.913 on MS COCO.

Spatial inconsistency is complementary rather than universal, it excels when models are overconfident on multi-label tasks but degrades when failures are spatially uniform.

2 RELATED WORK

Bayesian and stochastic uncertainty estimation. MC-Dropout approximates Bayesian inference via T stochastic forward passes but requires dropout layers to be present at training time, making it inapplicable to frozen checkpoints (Gal and Ghahramani, 2016). Bayesian weight approximations offer principled posteriors but scale poorly (Graves, 2011; Blundell et al., 2015). Deep Ensembles achieve strong calibration at the cost of multiplying compute (Lakshminarayanan et al., 2017). A recent benchmark of 13 such techniques on multi-label chest X-ray data confirms no single method is universally optimal (Baur et al., 2025). SpatialUQ imposes no training-time assumptions

Test-time augmentation and spatial perturbation. TTA treats prediction variance across augmentations as an uncertainty proxy (Krizhevsky et al., 2012; Wang et al., 2019; Ayhan and Berens, 2018), though strategy matters more than diversity (Shanmugam et al., 2021). Geometric and photometric augmentations confound spatial ambiguity with color/contrast sensitivity. SpatialUQ instead measures whether predictions are stable across spatial sub-regions a geometric property of evidence localization formalized in Lemma 1 and compared against photometric TTA-JSD in Section 5.4.1.

Post-hoc feature- and logit-based methods. Mahalanobis distance (Lee et al., 2018), DDU (Mukhoti et al., 2021), energy scoring, and ODIN (Liu et al.,

2020; Liang et al., 2017) operate on training-set feature distributions or logits. Conformal prediction provides distribution-free coverage guaranties but constructs prediction sets rather than ranking inputs by failure probability (Angelopoulos and Bates, 2023). SpatialUQ produces a scalar triage score. Combining MUS-based ranking with conformal calibration is a natural future direction.

Uncertainty in medical imaging. MC-Dropout has been applied to retinopathy screening (Leibig et al., 2017) and brain segmentation (Roy et al., 2019). The aleatoric/epistemic distinction was formalized by Kendall and Gal (2017) and clinical motivation is surveyed by Begoli et al. (2019). Temperature scaling addresses overconfidence (Guo et al., 2017) and selective prediction provides the risk-coverage framework for triage (Geifman and El-Yaniv, 2017). We follow Nado et al. (2021), reporting ECE, AURC, E-AURC, AUPR, and FPR at fixed TPR.

Failure and OOD detection. Failure detection without ground truth has been studied in object detection (Miller et al., 2019), with likelihood-ratio scores proposed as OOD detectors (Ren et al., 2019). Our MS COCO evaluation extends SpatialUQ to detection via a multi-label Bernoulli formulation. Prior calibration work on ViT (Fort et al., 2019; Minderer et al., 2021) and globally coherent self-attention (Liu et al., 2021) motivates our investigation of masking versus up-sampling for crop construction.

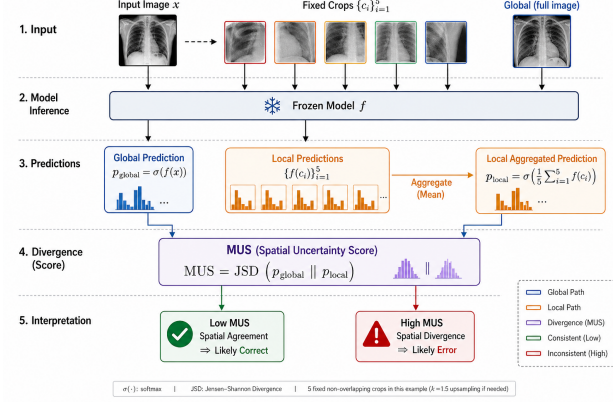
Divergence measures and output-space uncertainty. Predictive entropy (Shannon, 1948) and mutual information (Houlsby et al., 2011) are canonical output-space baselines. JSD’s symmetry, boundedness in $[0, \log 2]$, and robustness to zero probability mass make it well-suited to multi-label sigmoid outputs (Goodfellow et al., 2014; Hinton et al., 2015). To our knowledge, SpatialUQ is the first to apply JSD between global and local spatial predictions as a failure detection proxy, introducing a geometrically grounded interpretation requiring no model internals.

3 OUR APPROACH

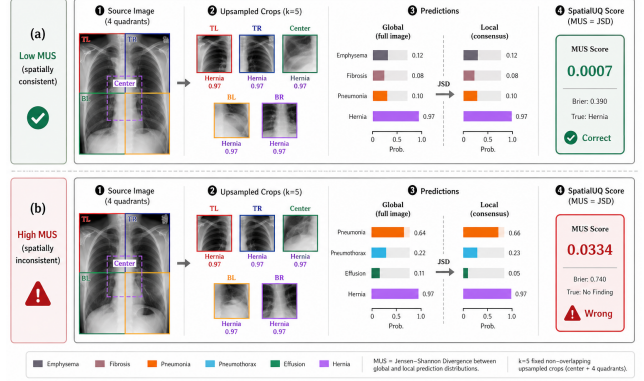
SpatialUQ operationalizes the spatial consistency hypothesis through a deterministic five-crop decomposition. We formalize the score for single- and multi-label settings, describe the optional linear fusion, and bound the score against per-class probability shifts. Figure 1 illustrates the pipeline and qualitative examples.

3.1 Spatial Decomposition

Let $f : \mathcal{X} \rightarrow \mathcal{P}$ be a frozen classifier mapping $x \in \mathcal{X}$ to either a probability simplex Δ^{C-1} (single-label) or $[0, 1]^C$ under independent sigmoid activations (multi-label), accessible only via output probabilities. We seek a deterministic scalar $s(x) \geq 0$, computable in



(a) SpatialUQ Pipeline Overview.



(b) Qualitative Examples: Low vs. High MUS.

a small constant number of forward passes, that is positively correlated with prediction error (Brier score) at inference time.

A reliable model gives spatially consistent predictions, inconsistency signals that confidence is not anchored to stable image content. Each 224×224 input is decomposed into five fixed spatial crops, four non-overlapping quadrants and one overlapping center crop (in $[y_1, x_1, y_2, x_2]$ format): $\mathcal{C} = \{(0, 0, 112, 112), (0, 112, 112, 224), (112, 0, 224, 112), (112, 112, 224, 224), (56, 56, 168, 168)\}$, fixed across all inputs and architectures. Each crop is bi-linearly upsampled to 224×224 before passing through f . The aggregate local prediction is:

$$p_{\text{local}}(x) = \frac{1}{5} \sum_{k=1}^5 f(\text{upsample}(x[\mathcal{C}_k])) \quad (1)$$

i Together with $p_{\text{global}}(x) = f(x)$, this requires exactly **six forward passes** through an unmodified model.

3.2 Uncertainty Score via Jensen-Shannon Divergence

We quantify spatial inconsistency via JSD between p_{global} and p_{local} (bounded in $[0, \log 2]$, symmetric, well-defined under zero probability mass).

Single-label tasks (categorical JSD). For softmax outputs $p_{\text{global}}, p_{\text{local}} \in \Delta^{C-1}$, with mixture $m = \frac{1}{2}(p_{\text{global}} + p_{\text{local}})$:

$$\begin{aligned} s_{\text{cat}}(x) &= \text{JSD}(p_{\text{global}} \| p_{\text{local}}) \\ &= \frac{1}{2} \text{KL}(p_{\text{global}} \| m) + \frac{1}{2} \text{KL}(p_{\text{local}} \| m) \end{aligned} \quad (2)$$

Applied to ImageNet-1k and MS COCO (per-class detection confidences as an 80-dimensional vector).

Multi-label tasks (Bernoulli JSD). Each class c induces an independent Bernoulli distribution with mixture $m_c = \frac{1}{2}(p_{\text{global},c} + p_{\text{local},c})$:

$$\begin{aligned} \text{JSD}_c(p, q) &= \frac{1}{2} \left[p_c \log \frac{p_c}{m_c} + (1 - p_c) \log \frac{1 - p_c}{1 - m_c} \right. \\ &\quad \left. + q_c \log \frac{q_c}{m_c} + (1 - q_c) \log \frac{1 - q_c}{1 - m_c} \right] \end{aligned} \quad (3)$$

$$s_{\text{bern}}(x) = \frac{1}{C} \sum_{c=1}^C \text{JSD}_c(p_{\text{global},c}, p_{\text{local},c}) \quad (4)$$

Applied to NIH ChestX-ray14 ($C=14$), CheXpert ($C=10$), and VinBigData ($C=10$). Algorithm 1 summarises the procedure: one forward pass for p_{global} , five for the upscaled crops ($p_1, \dots, p_5$), average to p_{local} , and return s_{cat} or s_{bern} . The pipeline is deterministic, stateless, and adds no overhead beyond the six forward passes.

Algorithm 1: SpatialUQ

Input: Frozen model f , image $x \in \mathbb{R}^{224 \times 224 \times 3}$, task $\in \{\text{single-label, multi-label}\}$

Output: Scalar uncertainty score $s(x) \geq 0$

$p_{\text{global}} \leftarrow f(x)$ // forward pass 1

for $k = 1$ **to** 5 **do**

$p_k \leftarrow f(\text{upsample}(x[\mathcal{C}_k], 224 \times 224))$
 // forward passes 2-6

end

$p_{\text{local}} \leftarrow \frac{1}{5} \sum_{k=1}^5 p_k$

$s \leftarrow s_{\text{cat}}(x)$ **if** single-label **else** $s \leftarrow s_{\text{bern}}(x)$

return s

3.3 Linear Fusion (Supervised Extension)

This supervised extension fits logistic weights on a small labeled hold-out; the unsupervised MUS itself needs

no labels. It combines complementary output signals via logistic regression:

$$s_{\text{fused}}(x) = \sigma(\alpha_0 + \alpha_1 s(x) + \alpha_2 H(x) + \alpha_3 (1 - \max_c p_c) + \alpha_4 d_{\ell_1}(x)) \quad (5)$$

Here $d_{\ell_1}(x) = \frac{1}{C} \sum_c |p_{\text{global},c} - p_{\text{local},c}|$ and $\sigma(\cdot)$ is the sigmoid. Coefficients α_i come from 5-fold cross-validation on a small labeled hold-out: the validation set for NIH, or a random 10% of the test set for zero-shot settings (remaining 90% for evaluation only), preserving f 's zero-retraining property. Ablation is in Table 2.

3.4 Information-Theoretic Bound

Lemma 1 (Spatial Disagreement Bound). *For each class c , let $p_c := p_{\text{global},c}$, $q_c := p_{\text{local},c} \in [0, 1]$. Then:*

$$\frac{1}{C} \sum_{c=1}^C |p_{\text{global},c} - p_{\text{local},c}| \leq \sqrt{2 \cdot s_{\text{bern}}(x)} \quad (6)$$

Here, $s_{\text{bern}}(x)$ is in nats, matching the implementation. The per-class bound is tight in the limit $p_c, q_c \rightarrow \frac{1}{2}$, where the ratio of the two sides tends to 1.

Proof sketch. Applying Pinsker's inequality to $\text{KL}(p_c \| m_c)$ and $\text{KL}(q_c \| m_c)$ separately, where $m_c = \frac{1}{2}(p_c + q_c)$, yields $(p_c - q_c)^2 \leq 2 \text{JSD}_c$ in nats. Taking square roots and averaging over C classes via Jensen's inequality (concavity of $\sqrt{\cdot}$) gives the class-averaged bound. Full derivation and tightness verification are in Appendix A. $\square$

Lemma 1 shows a large SpatialUQ score *necessarily* implies a large mean absolute shift between global and local class probabilities. This concerns probability shift only; that MUS predicts failures is empirical: the Spearman correlation between s_{bern} and per-image Brier score is positive and highly significant ($p < 0.0001$) across all medical benchmarks, which licenses MUS as a failure signal. The bound is conservative under heterogeneous per-class JSD via Jensen's inequality.

4 EXPERIMENTAL SETUP

4.1 Datasets and Tasks

We evaluate SpatialUQ across twelve configurations spanning medical and natural image classification and object detection. **NIH ChestX-ray14** (Wang et al., 2017) (112,120 frontal chest X-rays, 14 multi-label diseases; patient-wise split **73,891/12,633/25,596** train/val/test; class-balanced BCE for $5.3 \times -641.5 \times$ imbalance) trains three architectures independently: DenseNet-121 (Huang et al., 2017), EfficientNet-B4 (Tan and Le, 2019), and ViT-B/16 (Dosovitskiy et al., 2020). CLIP ViT-B/32 and BiomedCLIP are

evaluated on the same NIH test set via linear-probe transfer (frozen backbones). **CheXpert** (Irvin et al., 2019) and **VinBigData** (Nguyen et al., 2022) serve as zero-shot transfer benchmarks for the NIH-trained DenseNet-121 (10 NIH-shared classes each; Appendix L; U-Zeros for uncertain CheXpert labels). **ImageNet-1k** evaluates three pretrained models (EfficientNet-B4, ViT-B/16, ConvNeXt-Tiny (Liu et al., 2022)) zero-shot on 50,000 validation images with categorical JSD; failure is top-1 misclassification. **MS COCO 2014** (Lin et al., 2014) evaluates Faster R-CNN (Ren et al., 2015) and RetinaNet (Lin et al., 2017) (ResNet-50-FPN) on a 5,000-image subset (seed 42), with per-class confidence as the maximum detection score (threshold 0.1; 0.05/0.30 confirm stability) and Bernoulli JSD over 80-dimensional vectors.

4.2 Failure Definition

A prediction is labeled a *failure* if its per-image Brier score exceeds the 75th percentile of the test-set Brier distribution, yielding a consistent 25% failure rate across all datasets and architectures for direct cross-domain comparison without dataset-specific tuning (robustness to the 50th, 75th, and 90th percentile thresholds is confirmed in Appendix E.3. Critically, the Brier score is computed against ground-truth labels and is epistemically independent of MUS, which operates solely on output probabilities.

4.3 Training Protocol

NIH architectures are fine-tuned from ImageNet-pretrained weights in two stages: (i) 5 epochs with frozen backbone (lr 3×10^{-4} , AdamW, weight decay 10^{-4}); (ii) up to 25 epochs (patience 7) unfrozen (backbone lr 5×10^{-6} , head lr 5×10^{-5} , cosine annealing to 10^{-7}). Loss is BCE with label smoothing ($\epsilon = 0.1$) and class-balanced **pos_weight**, augmentation is random rotation ($\pm 10^\circ$) and crop to 224×224 (no horizontal flip, preserving laterality), batch size 32. ImageNet and COCO models are evaluated zero-shot full details in Appendix J.

4.4 Baselines

We compare against 17 uncertainty methods from the same frozen model without retraining. **Output-space:** predictive entropy (Shannon, 1948), inverted confidence (raw and temperature-scaled), ℓ_1 distance, and maximum inter-crop disagreement. **Crop-based:** patch variance, pairwise disagreement, and pairwise JSD. For reproducibility: patch variance is the variance of per-class probabilities across the five crops, averaged over classes, pairwise disagreement is the mean absolute difference between all pairs of crop predictions, pairwise JSD is the mean JSD between all pairs of crop predictions. **Augmentation-based:** TTA variance, TTA-JSD (10 augmentations), and MC-Dropout (Gal

and Ghahramani, 2016) (30 passes). **Feature/logit-based:** DDU (Mukhoti et al., 2021), Mahalanobis distance (Lee et al., 2018), and ODIN (Liang et al., 2017). **Ensemble:** mean prediction entropy, ensemble variance, and temperature-scaled entropy (Lakshminarayanan et al., 2017) from five DenseNet-121 models. Hyperparameters follow respective publications (Appendix J).

4.5 Evaluation Protocol

The primary metric is AUC for binary failure detection, we additionally report AUPR, AURC, E-AURC, and FPR at 80% and 95% TPR. We adopt the AURC formulation of Nado et al. (2021) over AUGRC (Zhou et al., 2024), as the fixed 25% failure rate renders the two equivalent. Calibration uses Score Calibration Error (SCE; ECE with 15 bins on min-max normalized scores). All metrics include 95% bootstrap CIs (1,000 resamples), pairwise AUC significance uses DeLong tests and two-tailed bootstrap p -values (10,000 resamples). All experiments use seed 42, code and data splits are included in the supplementary material and will be released publicly upon acceptance.

5 RESULTS

5.1 NIH ChestX-ray14: In-Distribution Failure Detection

Table 1 reports failure detection on NIH ChestX-ray14 ($N=25,596$) using DenseNet-121. MUS reaches an AUC of 0.784 (95% CI [0.778, 0.789]) with six deterministic forward passes, outperforming MC-Dropout (0.664, $\Delta=+0.119$, $p<10^{-6}$; 30 stochastic passes) while reducing wall-clock time by $5\times$ (≈ 8 vs. ≈ 39 min, Appendix J). Although ℓ_1 achieves a slightly higher AUC (0.790 vs. 0.784, $\Delta=0.006$, $p<10^{-6}$), its calibration is substantially worse (SCE 0.127 vs. 0.049). Naive crop variance performs much worse (0.610 AUC), showing that the way crop predictions are aggregated is important. Confidence and entropy are close to chance (0.401 and 0.177), reflecting systematic BCE overconfidence from label smoothing and class-balanced loss; temperature scaling does not recover the signal (0.400). This advantage is specific to the overconfidence regime encountered here. On well-calibrated ImageNet classifiers, confidence alone reaches 0.913–0.923 AUC, compared with 0.641–0.717 for MUS, although MUS still provides complementary information in fusion (Section 5.6). Combining MUS, entropy, confidence, and ℓ_1 through 5-fold cross-validation on the validation set increases AUC to 0.832, outperforming a five-member deep ensemble (0.813, $\Delta=+0.019$, $p<10^{-6}$) at a fraction of the training cost. MUS also tracks model error monotonically ($\rho=0.523$, $p<10^{-6}$), and the same behavior extends to EfficientNet-B4 and ViT-B/16 (Appendix C). Finally, spatial disagreement heatmaps show

that high-MUS failures concentrate crop-level JSD in ambiguous regions (Appendix K).

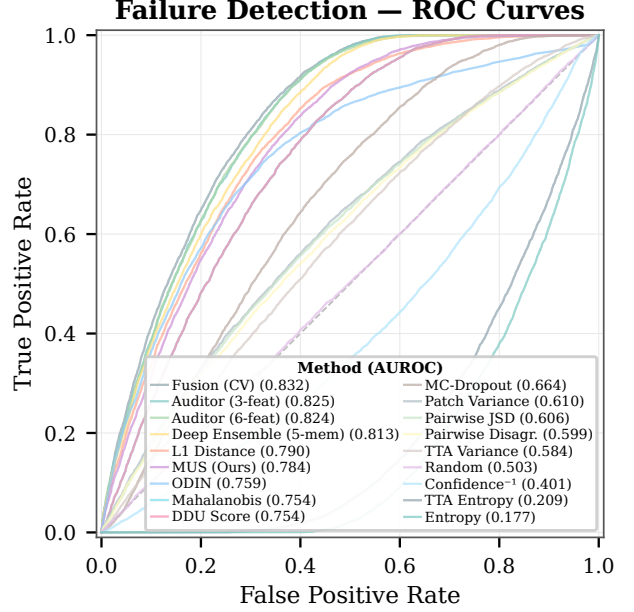


Figure 2: Failure detection ROC curve for the NIH test set, comparing SpatialUQ (MUS) against sampling-based and feature-based baselines.

Figure 2 shows that MUS achieves 0.784 AUC with six deterministic passes, outperforming MC-Dropout (0.664), crop-variance TTA (0.610), and all other TTA variants (≤ 0.584) without ensembles, stochastic inference, or auxiliary supervision. Fusion further improves performance to 0.832 AUC by combining MUS with complementary signals.

5.2 Extension to Frozen Foundation Models

We apply the NIH protocol to frozen CLIP ViT-B/32 (Radford et al., 2021) and BiomedCLIP (Zhang et al., 2023) via linear-probe transfer. Confidence⁻¹ collapses on uncalibrated medical outputs (0.095/0.206 AUC), while MUS remains strong at 0.825 ($\rho=0.631$) and 0.899 ($\rho=0.846$, $p<10^{-6}$), respectively. Although ℓ_1 has slightly higher AUC, MUS is better calibrated (SCE 0.133 vs. 0.232), and fusion reaches 0.883/0.925. SpatialUQ therefore transfers across general and domain-specific foundation models without architectural access (Appendix C.2).

5.3 Calibration as a Deployment Signal

Calibration determines whether a threshold is clinically meaningful, AUC measures only ranking. Figure 3 shows MUS natively calibrated (SCE=0.049) versus raw ℓ_1 (0.127). A labeled hold-out lets isotonic regression calibrate ℓ_1 (0.006/0.789) and fusion (0.004/0.830), but such labels are unavailable for frozen black-box deployment, where MUS remains the only natively calibrated discriminative option for deployment.

Table 1: Failure detection on NIH ChestX-ray14 ($N=25,596$, failure: Brier $>$ 75th percentile). $\dagger$ requires >6 passes. $\ddagger$ requires training features. Full results in Table 9.

Method	AUC	95% CI	SCE
Fusion (CV)	0.832	[0.827, 0.837]	0.246
Deep Ensemble $\dagger$	0.813	[0.808, 0.819]	0.138
ℓ_1 Distance	0.790	[0.785, 0.796]	0.127
MUS (Unsupervised)	0.784	[0.778, 0.789]	0.049
Deep Ens + T-scaling $\dagger$	0.774	[0.768, 0.780]	0.090
ODIN	0.759	[0.752, 0.765]	0.484
TTA-JSD (Photometric)	0.755	[0.750, 0.761]	0.043
DDU Score $\ddagger$	0.754	[0.748, 0.760]	0.139
Mahalanobis $\ddagger$	0.754	[0.748, 0.760]	0.283
Max Disagreement	0.744	[0.738, 0.750]	0.186
MC-Dropout $\dagger$	0.664	[0.658, 0.671]	0.121
Ensemble Variance	0.658	[0.651, 0.665]	0.042
TTA (5 crops, variance)	0.610	[0.602, 0.618]	0.029
TTA Variance ($N=10$) $\dagger$	0.584	[0.576, 0.591]	0.089
T-Scaled Conf $^{-1}$	0.400	[0.391, 0.408]	0.002
Confidence $^{-1}$	0.401	[0.392, 0.409]	0.422

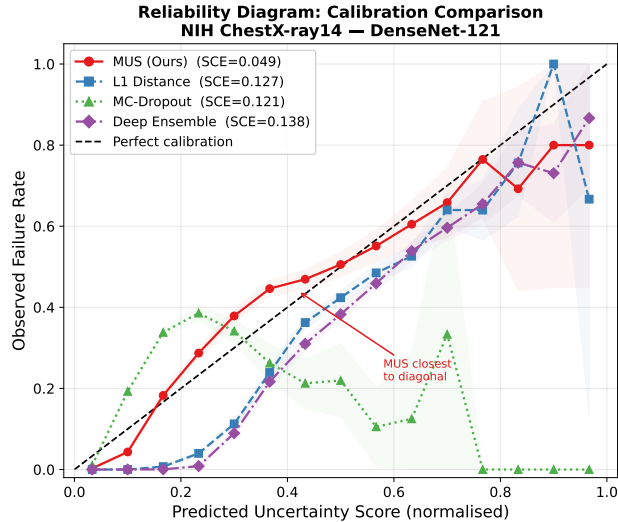


Figure 3: Reliability diagram comparing MUS against baseline uncertainty methods on NIH ChestX-ray14 (DenseNet-121).

5.4 Ablation: Fusion Components and Spatial Design

Table 2 shows entropy is the most critical fusion component (removal: -0.044 AUC). MUS and ℓ_1 add complementary gains (-0.005 , -0.007), confidence negligibly (-0.001). Yet entropy alone reaches only 0.177 AUC under BCE overconfidence while MUS reaches 0.784 alone, so fusion gains come from orthogonal failure modes. Five fixed crops are optimal (Appendix E), nine crops gain only $+0.007$ AUC at 67% additional cost, three crops lose 0.007, and 1,000 random placements reach 0.739 (Appendix E.1). These results confirm the fixed geometric arrangement provides superior spatial

coverage (Appendix E.2).

Table 2: Leave-one-out ablation of the four-feature fusion (NIH DenseNet-121). Full fusion AUC = 0.832. Δ : AUC drop when each feature is removed.

Removed Feature	AUC	Δ
Entropy	0.788	-0.044
ℓ_1 Distance	0.825	-0.007
MUS	0.827	-0.005
Confidence	0.831	-0.001

5.4.1 Spatial vs. Photometric Perturbation

The identical Bernoulli JSD on five photometric augmentations (brightness ± 0.15 , contrast $\times 1.2/\times 0.8$, noise $\sigma=0.05$) yields 0.755 AUC ($\rho=0.493$) versus MUS’s 0.784 ($\rho=0.523$) at the same six passes, with marginally better calibration (SCE 0.043 vs. 0.049). The AUC advantage confirms the gain comes from the perturbation’s geometric structure, not the divergence measure.

5.5 Zero-Shot Generalization to External Medical Datasets

The NIH-trained DenseNet-121 transfers without retraining to CheXpert ($N=44,399$) and VinBigData ($N=15,000$); results are summarized in Table 3. On CheXpert, MUS achieves 0.708 AUC versus 0.599 for MC-Dropout ($\Delta=+0.109$, $p<10^{-6}$), with ρ declining from 0.523 to 0.298 but retaining useful discrimination. At 20% referral, selective prediction improves system AUC from 0.686 to 0.748 $+19.7\%$ relative gap reduction. Under severe shift on VinBigData, however, MC-Dropout outperforms MUS (0.764 vs. 0.614, $\Delta=-0.150$, $p<10^{-6}$), while ρ collapses to 0.027 as failures become spatially uniform and MUS loses its inconsistency signal.

Table 3: Zero-shot failure detection (NIH-trained DenseNet-121, no retraining). Spearman ρ = correlation between MUS and per-image Brier score. Full per-method tables in Appendix F.

Dataset	MUS AUC	Fusion AUC	MC-Dropout AUC	Spearman ρ
NIH (in-dist.)	0.784	0.832	0.664	0.523
CheXpert (zero-shot)	0.708	0.733	0.599	0.298
VinBigData (zero-shot)	0.614	0.750	0.764	0.027

5.6 Natural Image Classification and Object Detection

Table 4 evaluates ImageNet-1k (three architectures, 50,000 images) and MS COCO 2014 (two detectors,

Table 4: Failure detection AUC on ImageNet-1k and MS COCO 2014. Full secondary metrics in Appendix G.

Dataset	Model	Top-1 Acc.	MUS	Confidence ⁻¹	Entropy	Fusion	Spearman ρ
ImageNet	EfficientNet-B4	79.27%	0.717	0.913	0.896	0.917	0.392
ImageNet	ViT-B/16	81.07%	0.710	0.923	0.884	0.937	0.341
ImageNet	ConvNeXt-Tiny	82.13%	0.641	0.913	0.819	0.936	0.139
COCO	Faster R-CNN	-	0.800	0.602	0.884	0.885	0.613
COCO	RetinaNet	-	0.747	0.702	0.912	0.913	0.510

5,000 images). On well-calibrated single-label classifiers, confidence already provides a strong correctness signal, making MUS complementary rather than a replacement (Section 6). On ImageNet, MUS alone trails Confidence⁻¹ (0.641–0.717 vs. 0.913–0.923 AUC), but fusion reaches 0.917–0.937, with MUS consistently contributing in leave-one-out ablations (Appendix G). This suggests that spatial consistency provides information beyond confidence even when confidence is already highly predictive. On COCO, MUS is stronger as a standalone signal (0.800/0.747 AUC; $\rho=0.613/0.510$), approaching the NIH in-distribution correlation and benefiting from the richer spatial structure of multi-label detection. Entropy remains the strongest standalone signal (0.884/0.912), while fusion reaches 0.885/0.913.

5.7 Clinical Triage Simulation

Radiologist review simulation. We simulate triage on the NIH test set (DenseNet-121). Referring the 9.2% highest-MUS images to human review recovers 20% of all Brier-defined failures at 54.5% referral precision. The threshold is tunable: precision is 58.7% at 10% recall (4.3% referred), 54.5% at 20%, 51.4% at 30% (Appendix I). Rejecting the 20% most uncertain images raises mean test AUC 0.808→0.814 ($\Delta=+0.006$).

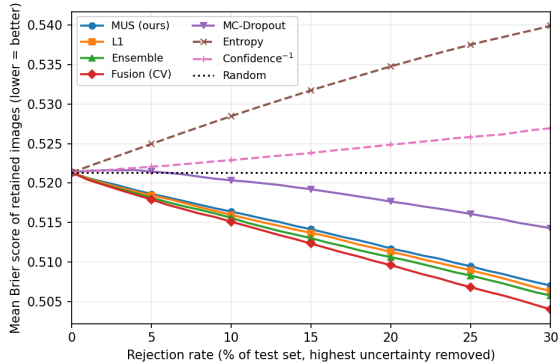


Figure 4: Mean per-image Brier score of the retained set (lower is better) versus rejection rate (NIH DenseNet-121, $N=25,596$), rejecting images in descending uncertainty order.

Figure 4 compares rejection quality across all methods against a random baseline. MUS monotonically lowers

retained Brier (0.521→0.507 at 30%) and beats MC-Dropout throughout. Fusion (CV) leads (0.504), with ℓ_1 and ensemble (0.506) marginally ahead of MUS. MUS remains the best-calibrated unsupervised score above 0.78 AUC (SCE 0.049 vs. ℓ_1 's 0.127). Entropy and confidence degrade the retained set, as expected from their near-chance AUC, since Brier defines the failure label, this restates the AUC ranking.

Per-class analysis. MUS discrimination varies strongly across pathologies: it performs best on diffuse, globally visible conditions such as Pneumonia (JSD AUC 0.979) and Edema (0.971), but poorly on small focal lesions such as Nodule (0.368), Fibrosis (0.553), and Mass (0.558), reflecting a limitation of the five-crop design. However a 16-crop extension raises Nodule AUC to 0.589 Appendix H). Prevalence-weighting across the 14 classes yields an overall AUC of 0.83, but masks this weakness because the three hardest classes account for only 14% of positive findings. A focal-lesion-enriched benchmark would provide a more stringent test of MUS.

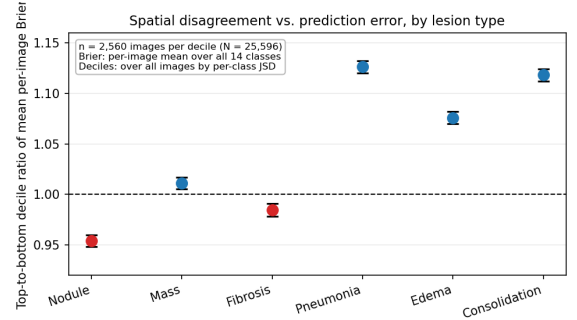


Figure 5: Top-to-bottom decile ratio of mean per-image Brier score by lesion type (NIH DenseNet-121, $N=25,596$). Error bars: 95% bootstrap CIs; $n = 2,560$ images per decile.

Figure 5 shows the top-to-bottom decile Brier ratio (95% CIs): 1.08–1.13 for diffuse classes but 0.95 [0.948, 0.960] for Nodule, indicating that high nodule disagreement concentrates on better-predicted images. Mass (1.01) and Fibrosis (0.99) show no meaningful coupling. Thus, disagreement does not track error for focal lesions, although this does not imply per-crop correctness. The limitation concerns the score rather than classifier quality: raw Nodule AUC (0.740) is near the 14-class mean (0.7885, Appendix B). Full curves, referral statis-

Table 5: Failure detection on NIH ChestX-ray14 (DenseNet-121, $N=25,596$). [†]Post-hoc approximation.

Method	AUC	Passes	Internals	Train data	Retrain	Frozen ckpt?
Fusion (Ours)	0.832	6 + logistic reg	✗	✗	✗	✓
Deep Ensemble (Lakshminarayanan et al., 2017)	0.813	5×model	✗	✗	✓	✗
MUS (Ours)	0.784	6	✗	✗	✗	✓
Deep Ensemble + T-scaling	0.774	5×model+cal	✗	✓	✓	✗
ODIN (Liang et al., 2017)	0.759	1+grad	✓	✗	✗	✓
DDU [†] (Mukhoti et al., 2021)	0.754	1	✓	✓	✗	✗
Mahalanobis [†] (Lee et al., 2018)	0.754	1	✓	✓	✗	✗
MC-Dropout (Gal and Ghahramani, 2016)	0.664	30	✗	✗	✓	✗

Passes: forward passes per image. **Internals:** needs gradients/activations. **Train data:** needs training-set features. **Retrain:** needs model modification. **Frozen ckpt?:** applicable to any pretrained checkpoint.

tics, and fairness analyzes are in Appendix I where gender parity is near-perfect (0.784 vs. 0.784).

6 DISCUSSION

A practitioner summary of when to use MUS is provided in Appendix I.1 (Table 32).

Complementarity and Regime Dependence. MUS is most useful in overconfident multi-label settings where confidence is weak (NIH DenseNet-121: AUC 0.401). On well-calibrated single-label tasks such as ImageNet, confidence dominates and MUS is complementary, with fusion combining the two signals.

Uncertainty type and OOD positioning. MUS measures whether model confidence is spatially reproducible, which we view as an epistemic signal reflecting learned spatial grounding rather than irreducible label noise. Its correlation with model quality also provides a pre-deployment diagnostic: $\rho \lesssim 0.15$ flags severe shift, as observed on VinBigData ($\rho = 0.027$), before full test-set evaluation.

Why MUS over ℓ_1 ? Although ℓ_1 slightly exceeds MUS in AUC (0.790 vs. 0.784, $\Delta = 0.006$; $p < 10^{-6}$), its SCE is substantially worse (0.127 vs. 0.049). Under shift, MUS also performs better on VinBigData (0.614 vs. 0.583), consistent with the robustness of bounded, symmetric JSD. Lemma 1 establishes the connection between MUS and ℓ_1 , but the empirical advantage of MUS lies in its native calibration and shift robustness. Isotonic regression can calibrate ℓ_1 (SCE 0.006), but requires labeled hold-out data unavailable in frozen deployment. Fusion’s higher SCE (0.246) is instead driven by the uncalibrated entropy component.

Limitations. SpatialUQ has five main limitations. (i) Severe domain shift can collapse the inconsistency signal (VinBigData, $\rho = 0.027$). (ii) Five fixed crops limit sensitivity to small focal lesions (Nodule JSD AUC 0.368; 16-crop extension: 0.589). (iii) ViT global attention compresses MUS’s dynamic range, with masking recovering +0.004 AUC (Appendix C.4). (iv) Fusion requires a labeled hold-out set. (v) Generative

VLMs with global pooling can weaken the signal (Appendix C.5). The $\rho \lesssim 0.15$ diagnostic flags the observed failure regimes while remaining positive for discriminative settings (NIH 0.523, CheXpert 0.298); when flagged, MC-Dropout or ensembles are preferable, with MUS retained as a fusion component. Thus, the primary operating regime is low-cost failure detection for overconfident multi-label classifiers under moderate shift when retraining is infeasible. Generalization beyond radiography and RGB natural images remains untested.

Comparison with State-of-the-Art. Table 5 compares SpatialUQ with established UQ methods on NIH DenseNet-121. MUS achieves 0.784 AUC fully post-hoc, without retraining, training data, or architectural modifications, while competitive baselines require training-time structure (MC-Dropout, ensembles) or dataset-dependent statistics (DDU, Mahalanobis). Extended analysis appears in Appendix K. Temperature scaling improves ensemble SCE from 0.138 to 0.090 but reduces AUC from 0.813 to 0.774, illustrating the trade-off between calibration and discrimination. Fusion reaches 0.832 AUC by adding the spatially grounded MUS signal, whose native SCE is 0.049.

7 FINAL REMARKS

SpatialUQ provides a practical operating point for inference-time failure detection: six deterministic forward passes using output probabilities alone, with no access to model internals, labels or retraining. Spatial inconsistency provides a geometrically grounded signal that scales with model quality, from 0.784 AUC on DenseNet-121 to 0.899 on BiomedCLIP ($\rho = 0.846$). The $\rho \lesssim 0.15$ diagnostic further provides a simple pre-deployment check for the observed distribution-shift failure regimes. Future work includes saliency-guided crop placement, aggregation adapted to global-attention architectures, and broader validation on MIMIC-CXR with temporal shift, population-level AUROC, and intersectional subgroup analyzes. Generalization to CT, MRI, and ultrasound also remains to be established.

AI use statement

In this work, we used generative AI tools for writing assistance, including drafting, editing, and formatting of manuscript text, and for verifying the consistency of reported numbers against the experimental code. We have not used generative AI tools for generating experimental results, fabricating data, or conducting literature surveys, and all AI-assisted content was reviewed by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

References

- Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.
- Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.
- Murat Seckin Ayhan and Philipp Berens. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *Medical imaging with deep learning*, 2018.
- Simon Baur, Wojciech Samek, and Jackie Ma. Benchmarking uncertainty and its disentanglement in multi-label chest x-ray classification. In *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pages 193–203. Springer, 2025.
- Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23, 2019.
- Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.
- Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.
- Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.
- Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- Alex Graves. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24, 2011.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.
- Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.
- Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghighi, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pages 590–597, 2019.
- Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.
- Zaid Khan and Yun Fu. Consistency and uncertainty: Identifying unreliable responses from black-box vision-language models for selective visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10854–10863, 2024.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

- Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.
- Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.
- Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports*, 7(1):1–14, 2017.
- Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
- Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33: 21464–21475, 2020.
- Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
- Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.
- Dimitry Miller, Feras Dayoub, Michael Milford, and Niko Sünderhauf. Evaluating merging strategies for sampling-based uncertainty techniques in object detection. In *2019 international conference on robotics and automation (icra)*, pages 2348–2354. IEEE, 2019.
- Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. *Advances in neural information processing systems*, 34:15682–15694, 2021.
- Jishnu Mukhoti, Andreas Kirsch, Joost Van Amersfoort, PH Torr, and Yarin Gal. Deterministic neural networks with inductive biases capture epistemic and aleatoric uncertainty. *arXiv preprint arXiv:2102.11582*, 2, 2021.
- Jishnu Mukhoti, Andreas Kirsch, Joost Van Amersfoort, Philip HS Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24384–24394, 2023.
- Zachary Nado, Neil Band, Mark Collier, Josip Djolonga, Michael W Dusenberry, Sebastian Farquhar, Qixuan Feng, Angelos Filos, Marton Havasi, Rodolphe Jenatton, et al. Uncertainty baselines: Benchmarks for uncertainty & robustness in deep learning. *arXiv preprint arXiv:2106.04015*, 2021.
- Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist’s annotations. *Scientific Data*, 9(1):429, 2022.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
- Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *Advances in neural information processing systems*, 32, 2019.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.
- Abhijit Guha Roy, Sailesh Conjeti, Nassir Navab, Christian Wachinger, Alzheimer’s Disease Neuroimaging Initiative, et al. Bayesian quicknat: Model uncertainty in deep whole-brain segmentation for structure-wise quality control. *NeuroImage*, 195:11–22, 2019.
- Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Gutttag. Better aggregation in test-time augmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1214–1223, 2021.
- Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.
- Han-Jay Shu, Wei-Ning Chiu, Shun-Ting Chang, Meng-Ping Huang, Takeshi Tohyama, Ahram Han, and

Po-Chih Kuo. Uncovering overconfident failures in cxr models via augmentation-sensitivity risk scoring. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6736–6740. IEEE, 2026.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

Han Zhou, Jordy Van Landeghem, Teodora Popordanoska, and Matthew B Blaschko. A novel characterization of the population area under the risk coverage curve (aurc) and rates of finite sample estimators. *arXiv preprint arXiv:2410.15361*, 2024.

CHECKLIST

1. For all models and algorithms presented, check if you include:
 - (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] The spatial decomposition, Bernoulli/categorical JSD scores, and the logistic fusion are defined in Section 3, with the full procedure in Algorithm 1.
 - (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] SpatialUQ needs exactly six forward passes per image with no stored state (Section 3); wall-clock timings are in Appendix J. Lemma 1 gives the information-theoretic property of the score.
 - (c) (Optional) Anonymized Source code, with specification of all dependencies, including external libraries. [Yes] Code and data splits
- are included in the supplementary material, with dependencies listed in the repository requirements file.
2. For any theoretical claim, check if you include:
 - (a) Statements of the full set of assumptions of all theoretical results. [Yes] Lemma 1 states its assumptions ($p_c, q_c \in [0, 1]$, JSD in nats) directly.
 - (b) Complete proofs of all theoretical results. [Yes] The full proof is in Appendix A, with a proof sketch in Section 3.4.
 - (c) Clear explanations of any assumptions. [Yes] Section 3.4 explains that the bound concerns probability shift only and that the link to failure detection is empirical (Spearman correlation).
3. For all figures and tables that present empirical results, check if you include:
 - (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] Code, data splits, and fixed random seeds (seed 42) are in the supplementary material.
 - (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] Section 4.3 and Appendix J give the two-stage protocol, learning rates, augmentation, and the 73,891/12,633/25,596 patient-wise split.
 - (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] Failure is defined as Brier score above the 75th percentile (Section 4.2); all AUCs carry 95% bootstrap CIs (1,000 resamples) and pairwise significance uses DeLong tests with 10,000-resample bootstrap p -values.
 - (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] All NIH experiments ran on a single NVIDIA Tesla T4 GPU (Appendix J).
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
 - (a) Citations of the creator If your work uses existing assets. [Yes] NIH ChestX-ray14 (Wang et al., 2017), CheXpert (Irvin et al., 2019), VinBigData (Nguyen et al., 2022), MS COCO (Lin et al., 2014), CLIP (Radford et al.,

2021), and BiomedCLIP (Zhang et al., 2023) are cited, as are the DenseNet (Huang et al., 2017), EfficientNet (Tan and Le, 2019), and ViT (Dosovitskiy et al., 2020) backbones.

- (b) The license information of the assets, if applicable. [Yes] All datasets and pretrained models are publicly released research assets used under their published terms.
- (c) New assets either in the supplemental material or as a URL, if applicable. [Yes] The SpatialUQ implementation notebooks are available at this link <https://huggingface.co/datasets/kawsher11/SpatialUQ>.
- (d) Information about consent from data providers/curators. [Not Applicable] All datasets are pre-existing public benchmarks; no new data was collected.
- (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable] The chest X-ray benchmarks are de-identified; the fairness audit in Appendix I uses only aggregate age/gender metadata.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

- (a) The full text of instructions given to participants and screenshots. [Not Applicable]
- (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
- (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

No crowdsourcing or human-subjects research was conducted; the triage study in Section 5.7 is a retrospective simulation on de-identified benchmark data.

SpatialUQ: Post-Hoc Uncertainty Quantification from Spatial Consistency in Black-Box Vision Models: Supplementary Materials

A PROOF OF LEMMA 1

We prove the spatial disagreement bound in two steps. All logarithms are natural (JSD in nats) unless stated otherwise.

Step 1: Per-class bound via Pinsker’s inequality. For Bernoulli distributions with parameters $p_c, q_c \in [0, 1]$, let $m_c = \frac{1}{2}(p_c + q_c)$. The JSD decomposes as:

$$\text{JSD}_c(p_c \| q_c) = \frac{1}{2} \text{KL}(p_c \| m_c) + \frac{1}{2} \text{KL}(q_c \| m_c) \quad (7)$$

For Bernoulli distributions, the total variation distance between parameters p_c and m_c reduces to $\text{TV}(p_c, m_c) = |p_c - m_c| = \frac{1}{2}|p_c - q_c|$. Pinsker’s inequality $\text{TV}(P, Q)^2 \leq \frac{1}{2} \text{KL}(P \| Q)$ then gives:

$$\frac{1}{4}(p_c - q_c)^2 \leq \frac{1}{2} \text{KL}(p_c \| m_c) \implies (p_c - q_c)^2 \leq 2 \text{KL}(p_c \| m_c) \quad (8)$$

Applying the same argument symmetrically to $\text{KL}(q_c \| m_c)$:

$$(p_c - q_c)^2 \leq 2 \text{KL}(q_c \| m_c) \quad (9)$$

Adding the two inequalities above:

$$2(p_c - q_c)^2 \leq 2(\text{KL}(p_c \| m_c) + \text{KL}(q_c \| m_c)) = 4 \text{JSD}_c(p_c \| q_c) \quad (10)$$

Dividing by 2 and taking square roots:

$$|p_c - q_c| \leq \sqrt{2 \text{JSD}_c(p_c \| q_c)} \quad (\text{in nats}) \quad (11)$$

Step 2: Aggregation over classes via Jensen’s inequality. Averaging Eq. (11) over all C classes:

$$\frac{1}{C} \sum_{c=1}^C |p_c - q_c| \leq \frac{1}{C} \sum_{c=1}^C \sqrt{2 \text{JSD}_c} \quad (12)$$

Since $\sqrt{\cdot}$ is concave, Jensen’s inequality gives:

$$\frac{1}{C} \sum_{c=1}^C \sqrt{2 \text{JSD}_c} \leq \sqrt{\frac{2}{C} \sum_{c=1}^C \text{JSD}_c} = \sqrt{2 s_{\text{bern}}(x)} \quad (\text{in nats}) \quad (13)$$

which is the complete bound:

$$\frac{1}{C} \sum_{c=1}^C |p_{\text{global},c} - p_{\text{local},c}| \leq \sqrt{2 \cdot s_{\text{bern}}(x)} \quad \blacksquare \quad (14)$$

Verification of tightness. The per-class bound in Eq. (11) is tight in the limit $p_c, q_c \rightarrow \frac{1}{2}$. Setting $p_c = \frac{1}{2} + \varepsilon$, $q_c = \frac{1}{2} - \varepsilon$ (so $m_c = \frac{1}{2}$), a second-order expansion gives $\text{KL}(p_c \| m_c) = \text{KL}(q_c \| m_c) = 2\varepsilon^2 + o(\varepsilon^2)$, hence $\text{JSD}_c = 2\varepsilon^2 + o(\varepsilon^2)$ and

$$\text{RHS} = \sqrt{2 \text{JSD}_c} = 2\varepsilon + o(\varepsilon) = |p_c - q_c| + o(\varepsilon) = \text{LHS} + o(\varepsilon), \quad (15)$$

so the ratio of the two sides tends to 1 as $\varepsilon \rightarrow 0$.

B BACKBONE CLASSIFICATION PERFORMANCE

Table 6 reports per-class AUC-ROC, Brier score, and average precision for the DenseNet-121 backbone (7.0 M parameters) trained on NIH ChestX-ray14, selected at epoch 28 (validation loss of 2.0564). The model achieves a mean AUC of 0.7885, with discriminative performance ranging from 0.6984 (Infiltration) to 0.8822 (Emphysema). Calibration and average precision exhibit pronounced class-level disparity, largely attributable to severe label imbalance. These results serve as the reference operating point for the subsequent failure-detection analysis.

Table 6: NIH ChestX-ray14 backbone classification performance - DenseNet-121 (7.0 M parameters, best checkpoint epoch 28, validation loss 2.0564). Mean AUC of 0.7885.

Class	AUC	Brier	AP
Atelectasis	0.7628	0.3189	0.3272
Cardiomegaly	0.8746	0.5565	0.3350
Effusion	0.8192	0.3081	0.4907
Infiltration	0.6984	0.3161	0.3969
Mass	0.7965	0.3576	0.2915
Nodule	0.7400	0.3869	0.2191
Pneumonia	0.7020	0.7346	0.0463
Pneumothorax	0.8471	0.4472	0.3864
Consolidation	0.7401	0.5084	0.1562
Edema	0.8368	0.6476	0.1494
Emphysema	0.8822	0.6111	0.3595
Fibrosis	0.8095	0.6404	0.0888
Pleural_Thickening	0.7569	0.5222	0.1245
Hernia	0.7733	0.9428	0.0434
Mean	0.7885	-	-

Table 7 reports per-class AUC-ROC, Brier score, and average precision for the EfficientNet-B4 backbone (17.6 M parameters) on NIH ChestX-ray14, selected at epoch 29 (validation loss 2.0793). The model achieves a mean AUC of 0.7347, trailing DenseNet-121 by 0.054 points despite nearly $2.5\times$ more parameters, with per-class AUC ranging from 0.6589 (Pneumonia) to 0.8093 (Edema). Calibration and average precision mirror the class-imbalance patterns observed in the DenseNet-121 baseline. Both backbones are subsequently treated as frozen black-boxes for post-hoc uncertainty evaluation.

Table 7: NIH ChestX-ray14 backbone classification performance - EfficientNet-B4 (17.6 M parameters, best checkpoint epoch 29, validation loss 2.0793). Mean AUC = 0.7347.

Class	AUC	Brier	AP
Atelectasis	0.7192	0.3321	0.2732
Cardiomegaly	0.7952	0.5746	0.2118
Effusion	0.7740	0.3250	0.4189
Infiltration	0.6790	0.3102	0.3733
Mass	0.7002	0.4114	0.1591
Nodule	0.6941	0.4039	0.1520
Pneumonia	0.6589	0.7389	0.0375
Pneumothorax	0.8050	0.4824	0.3031
Consolidation	0.7022	0.5199	0.1299
Edema	0.8093	0.6440	0.1115
Emphysema	0.7977	0.6235	0.1405
Fibrosis	0.7547	0.6533	0.0575
Pleural_Thickening	0.7201	0.5370	0.0922
Hernia	0.6763	0.9437	0.0096
Mean	0.7347	-	-

Table 8 reports per-class AUC-ROC, Brier score, and average precision for the ViT-B/16 backbone (85.8M parameters) on NIH ChestX-ray14, with early stopping triggered at epoch 17 and the best checkpoint selected at epoch 9. Despite its substantially larger capacity, the model achieves a mean AUC of 0.7921, marginally exceeding DenseNet-121 (0.7885) and EfficientNet-B4 (0.7347), with per-class AUC ranging from 0.6978 (Infiltration) to 0.8868 (Cardiomegaly). Calibration patterns remain consistent with the convolutional backbones, with Hernia (Brier = 0.9413) and Pneumonia (0.7273) exhibiting the poorest posterior estimates. This backbone is of particular

interest for examining the relationship between global self-attention and spatial prediction inconsistency in the failure-detection analysis.

Table 8: NIH ChestX-ray14 backbone classification performance - ViT-B/16 (85.8M parameters, early stopping at epoch 17, best validation loss at epoch 9). Mean AUC = 0.7921.

Class	AUC	Brier	AP
Atelectasis	0.7689	0.2882	0.3399
Cardiomegaly	0.8868	0.5685	0.3500
Effusion	0.8257	0.3037	0.5059
Infiltration	0.6978	0.3029	0.3983
Mass	0.8157	0.3817	0.3176
Nodule	0.7502	0.3712	0.2230
Pneumonia	0.7111	0.7273	0.0555
Pneumothorax	0.8587	0.4469	0.4263
Consolidation	0.7521	0.4987	0.1686
Edema	0.8359	0.6364	0.1653
Emphysema	0.8854	0.5791	0.4034
Fibrosis	0.7938	0.6235	0.0857
Pleural_Thickening	0.7622	0.5380	0.1367
Hernia	0.7446	0.9413	0.0347
Mean	0.7921	-	-

C NIH CHESTX-RAY14: FULL FAILURE DETECTION TABLES

C.1 DenseNet-121: All Methods Comparison

The results in Table 9 evaluate failure detection performance on the NIH ChestX-ray14 test set using a DenseNet-121 backbone. Failure is defined as predictions with Brier score exceeding the 75th percentile (threshold = 0.5607, yielding a 25% failure rate). We report area under the ROC curve (AUC) with 95% bootstrap confidence intervals (1,000 resamples) and Score Calibration Error (SCE) the expected calibration error applied to min-max normalized scores using 15 bins. Among all methods, the proposed MUS achieves the lowest SCE (0.0487), indicating well-calibrated failure probabilities, while its AUC (0.7837) is competitive. The highest AUC is obtained by Fusion CV (0.8319), which combines multiple uncertainty signals, followed closely by the two Auditor variants (0.8250, 0.8242). However, these top AUC methods exhibit substantially higher SCE (0.2464, 0.2408, 0.2363) than MUS. Deep Ensemble (5 members) yields strong AUC (0.8130) and moderate SCE (0.1379), whereas MUS strikes a favorable trade-off with near-perfect calibration. Notably, simple post-hoc baselines like T-Scaled Confidence⁻¹ achieve excellent SCE (0.0015) but near-random AUC (0.4000), highlighting that calibration alone is insufficient for failure detection. Conversely, Entropy performs poorly on both metrics (AUC 0.1774, SCE 0.4575). Methods marked with † require more than six forward passes or training-time modifications; ‡ indicates access to training-set features. The table confirms that no single method dominates all criteria, but MUS provides the best calibration among practically viable approaches.

Statistical tests (DeLong, two-tailed): MUS vs. MC-Dropout: $\Delta\text{AUC} = +0.1194$, $p < 10^{-6}$. MUS vs. ℓ_1 Distance: $\Delta\text{AUC} = -0.0064$, $p < 10^{-6}$ (ℓ_1 marginally superior). Fusion vs. Deep Ensemble: $\Delta\text{AUC} = +0.0189$, $p < 10^{-6}$. Spearman correlation between MUS and Brier score: $\rho = 0.5225$, $p < 10^{-6}$ (Pearson $r = 0.4983$).

Table 10 demonstrates that MUS exhibits strong robustness to random initialization, with AUC values remaining tightly clustered across different seeds (0.7826–0.7902), indicating low variance and stable behavior. The fluctuation range (0.0076) is minimal, suggesting that MUS is not sensitive to stochastic training effects and generalizes consistently. In comparison, ℓ_1 Distance achieves slightly higher absolute performance across all seeds but follows a similar stability pattern, while MC-Dropout and TTA Variance show modest variability and consistently lower AUC. Notably, confidence-based and entropy-based methods remain significantly weaker and exhibit inconsistent behavior across seeds, highlighting their limited reliability. Pairwise JSD also shows mild degradation for certain seeds, suggesting sensitivity to randomness in feature representations. Overall, these results reinforce that MUS provides a reliable and reproducible uncertainty signal, maintaining competitive performance with minimal seed-induced variance an important property for deployment in high-stakes settings.

Table 9: NIH ChestX-ray14 failure detection DenseNet-121, full 16-method comparison ($N = 25,596$). Brier threshold = 0.5607 (75th percentile, 25% failure rate). SCE = Score Calibration Error (ECE applied to min-max normalised scores, 15 bins). † requires >6 forward passes or training-time modification. ‡ requires access to training-set features.

Method	AUC	95% CI	SCE
Fusion CV	0.8319	[0.8268, 0.8368]	0.2464
(MUS+Ent+Conf+ℓ_1)			
Auditor (3-feat MLP)†	0.8250	[0.8201, 0.8299]	0.2408
Auditor (6-feat MLP)†	0.8242	[0.8189, 0.8292]	0.2363
Deep Ensemble (5-member)†	0.8130	[0.8077, 0.8187]	0.1379
ℓ_1 Distance	0.7902	[0.7848, 0.7960]	0.1270
MUS (Ours)	0.7837	[0.7778, 0.7893]	0.0487
Deep Ensemble + T-scaling†	0.7741	[0.7680, 0.7801]	0.0897
ODIN	0.7588	[0.7518, 0.7653]	0.4844
TTA-JSD —	0.7549	[0.7492, 0.7607]	0.0430
photometric (same formula, 5 augs)			
DDU Score‡	0.7540	[0.7476, 0.7599]	0.1386
Mahalanobis‡	0.7540	[0.7480, 0.7596]	0.2828
Max Disagreement	0.7440	[0.7376, 0.7502]	0.1861
MC-Dropout (T=30)†	0.6643	[0.6575, 0.6710]	0.1211
Ensemble Variance†	0.6579	[0.6505, 0.6648]	0.0419
TTA (5 crops, patch variance)	0.6103	[0.6022, 0.6178]	0.0290
Pairwise JSD	0.6057	[0.5979, 0.6133]	0.0335
Pairwise Dis-agreement	0.5990	[0.5913, 0.6073]	0.2436
TTA Variance (N=10)†	0.5837	[0.5759, 0.5914]	0.0889
Confidence ⁻¹	0.4011	[0.3922, 0.4091]	0.4218
T-Scaled Conf ⁻¹	0.4000	[0.3910, 0.4081]	0.0015
Entropy	0.1774	[0.1727, 0.1821]	0.4575

Table 10: Seed MUS: Comparison across different random seeds for DenseNet121 on NIH ChestX-ray14.

Seed	MUS AUC	ℓ_1 Dist.	MC-Drop	Conf ⁻¹	TTA Var	Entr.	Pair JSD
42	0.7837	0.7902	0.6643	0.4011	0.5837	0.1774	0.6057
186	0.7866	0.7959	0.6466	0.5802	0.5723	0.1748	0.6023
456	0.7902	0.7976	0.6532	0.5816	0.5856	0.1672	0.5709
789	0.7862	0.7914	0.6498	0.5788	0.5797	0.1681	0.5806
1011	0.7826	0.7844	0.6536	0.5737	0.5848	0.1688	0.5660
Mean	0.7859	0.7919	0.6535	0.5431	0.5812	0.1713	0.5851
± Std	0.0026	0.0047	0.0060	0.0700	0.0050	0.0040	0.0160

C.2 Foundation Model Evaluation with Linear Probes

To assess SpatialUQ’s transferability and scaling with model quality, we evaluate CLIP ViT-B/32 and BiomedCLIP on the NIH ChestX-ray14 test set with a linear probe trained on NIH train features (backbone frozen; no fine-tuning). Labels are used only to train the probe and for evaluation. Table 11 reports failure detection AUC (75th percentile Brier threshold, 25% failure rate), 95% bootstrap CIs, and raw Score Calibration Error (SCE).

MUS achieves competitive failure detection across both models, with AUC improving from 0.825 (CLIP) to 0.899 (BiomedCLIP). This scaling is accompanied by an increase in Spearman correlation between MUS and

Table 11: Failure detection on NIH ChestX-ray14 with frozen foundation models plus linear probe (25,596 images). MUS requires only output probabilities; L_1 and Fusion use the same six forward passes. Fusion is a logistic regression trained on a small labeled hold-out (10% of the test set, remaining 90% for evaluation).

Method	CLIP ViT-B/32		BiomedCLIP	
	AUC [95% CI]	SCE	AUC [95% CI]	SCE
MUS (SpatialUQ)	0.825 [0.818, 0.830]	0.098	0.899 [0.895, 0.903]	0.133
Entropy	0.204 [0.199, 0.210]	0.574	0.394 [0.387, 0.402]	0.421
Confidence ⁻¹	0.095 [0.092, 0.099]	0.429	0.206 [0.201, 0.212]	0.334
L_1 Distance	0.837 [0.831, 0.842]	0.115	0.915 [0.911, 0.918]	0.232
Fusion	0.883 [0.879, 0.887]	0.234	0.925 [0.922, 0.928]	0.278

per-image Brier score ($\rho = 0.631$ for CLIP, $\rho = 0.846$ for BiomedCLIP), consistent with the hypothesis that better spatial representations yield stronger spatial consistency signals. Under linear-probe transfer, MUS’s calibration remains moderate (SCE 0.10-0.13) and superior to the raw L_1 baseline (0.115-0.232). L_1 and fusion achieve higher AUC but rely on a labeled hold-out for calibration (fusion) or would need isotonic regression (L_1) to produce reliable thresholds, options not always available in frozen black-box deployment. MUS therefore offers a label-free, natively calibrated uncertainty score that scales favorably with foundation model strength.

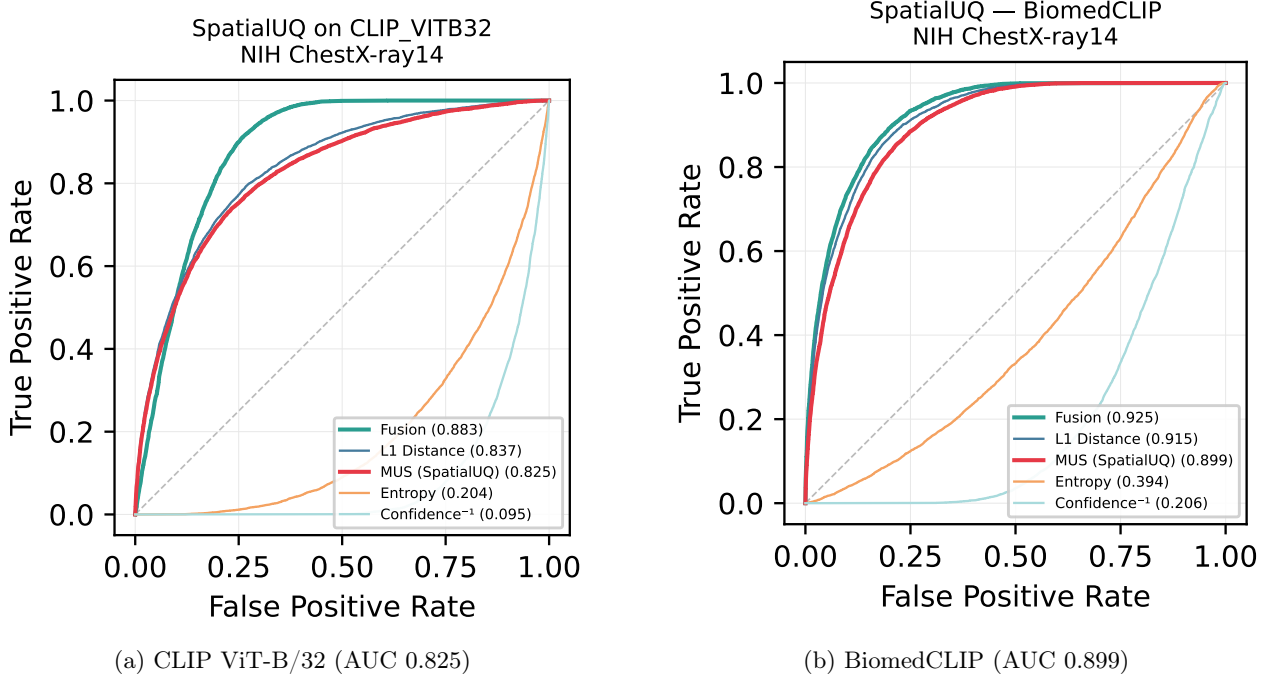


Figure 6: ROC curves for failure detection on NIH ChestX-ray14 with frozen foundation models plus linear probe. Left: CLIP ViT-B/32 (AUC 0.825). Right: BiomedCLIP (AUC 0.899). MUS (SpatialUQ) outperforms entropy and confidence⁻¹ baselines.

Figure 6 shows the corresponding ROC curves. MUS consistently outperforms entropy and confidence baselines, and its discrimination improves from CLIP to BiomedCLIP.

C.3 EfficientNet-B4: Full Failure Detection Comparison

Table 12 reports failure detection AUC for EfficientNet-B4 on the NIH test set ($N = 25,596$), with failures defined as samples exceeding the 75th-percentile Brier score threshold (0.5776), yielding a 25% failure rate. Fusion CV achieves the highest AUC (0.8714), followed by MUS (0.7842) and Mahalanobis (0.7804), with the three methods forming a statistically separated tier above ODIN (0.7681) and Max Disagreement (0.6722). MC-Dropout performs near chance (0.5357), significantly below MUS ($\Delta\text{AUC} = +0.2485$, 95% CI [0.2381, 0.2589], $p < 10^{-6}$).

Table 12: NIH ChestX-ray14 failure detection EfficientNet-B4, 14-method comparison ($N = 25,596$). Brier threshold = 0.5776 (75th percentile, 25% failure rate). SCE = Score Calibration Error (ECE applied to min-max normalised scores, 15 bins). † requires >6 forward passes or training-time modification. ‡ requires access to training-set features. DDU, Pairwise JSD and TTA Entropy are degenerate (constant scores) on this backbone.

Method	AUC	95% CI	SCE
Fusion CV	0.8714	[0.8675, 0.8756]	0.2713
(MUS+Ent+ Conf+ ℓ_1)			
MUS (Ours)	0.7842	[0.7777, 0.7900]	0.0666
Mahalanobis‡	0.7804	[0.7747, 0.7863]	0.3243
ℓ_1 Distance	0.7746	[0.7683, 0.7809]	0.0873
ODIN	0.7682	[0.7622, 0.7742]	0.2185
Max Disagree- ment	0.6722	[0.6644, 0.6794]	0.1391
MC-Dropout	0.5357	[0.5274, 0.5439]	0.0726
(T=30)†			
TTA Variance	0.6050	[0.5969, 0.6129]	0.2324
(N=10)†			
Pairwise Dis- agreement	0.5877	[0.5800, 0.5953]	0.1315
TTA (5 crops, patch variance)	0.5823	[0.5744, 0.5906]	0.0414
DDU Score‡	0.5000	[0.5000, 0.5000]	-
Pairwise JSD	0.5000	[0.5000, 0.5000]	0.1315
Confidence ⁻¹	0.3777	[0.3696, 0.3862]	0.1910
Entropy	0.1208	[0.1172, 0.1250]	0.4917

C.4 ViT-B/16: Full Failure Detection Comparison

Table 13 reports failure detection AUC for ViT-B/16 on the NIH test set ($N = 25,596$), with the failure threshold set at the 75th-percentile Brier score (0.5538). Fusion CV again leads (0.8314), with ODIN ranking second (0.8005) a notable reversal from the EfficientNet-B4 ordering followed by MUS (0.7497). MC-Dropout remains the weakest method (0.6502), though the margin over MUS narrows considerably relative to the convolutional backbone ($\Delta\text{AUC} = +0.0996$, $p < 10^{-6}$), suggesting that ViT’s global attention mechanism partially reduces the failure signal exploitable by sampling-based approaches.

ViT crop artifact investigation. ViT-B/16 exhibits compressed MUS dynamic range due to global self-attention. Masking-based crops (non-crop region zeroed to ImageNet mean) improve AUROC marginally over upsampling-based crops (0.7465 vs. 0.7421, $\Delta = +0.0045$). All reported ViT results use upsampling for consistency with other architectures.

The results in Table 14 show that the impact of spatial aggregation strategy is strongly architecture-dependent. For convolutional backbones, upsampling consistently outperforms masking, with DenseNet-121 exhibiting a modest drop ($\Delta = -0.0141$) and EfficientNet-B4 showing a substantial degradation ($\Delta = -0.2251$) under masking. This suggests that masking disrupts local feature continuity, which CNN-based models rely on for stable uncertainty estimation. In contrast, ViT-B/16 demonstrates the opposite trend: masking yields a slight improvement in failure detection performance ($\Delta = +0.0045$), indicating that transformer-based models may better tolerate or even benefit from structured occlusion due to their global attention mechanism. Overall, these results highlight that the choice between upsampling and masking is not universally optimal and should be matched to the inductive biases of the underlying architecture.

C.5 BioViL-T: Generative VLM Failure Regime

Table 15 reports SpatialUQ applied to BioViL-T, a generative vision-language model whose image encoder produces globally pooled cross-modal representations designed for image-text alignment rather than spatial discrimination. MUS collapses to 0.533 AUC (Spearman $\rho = -0.187$), correctly flagged by the $\rho \lesssim 0.15$ diagnostic as a known failure regime. This result is consistent with the mechanism described in Section 6: *spatial inconsistency requires architectures with spatially grounded feature extraction; globally pooled representations eliminate the crop-level inconsistency gradient MUS relies upon*. Fusion recovers to 0.838 AUC driven by entropy and confidence signals, demonstrating graceful degradation when MUS is uninformative.

Architectural interpretation. BioViL-T’s image encoder is trained via contrastive alignment with radiology

Table 13: NIH ChestX-ray14 failure detection ViT-B/16, 14-method comparison ($N = 25,596$). Brier threshold = 0.5538 (75th percentile, 25% failure rate). SCE = Score Calibration Error (ECE applied to min-max normalised scores, 15 bins). † requires >6 forward passes or training-time modification. ‡ requires access to training-set features. DDU, Pairwise JSD and TTA Entropy are degenerate (constant scores) on this backbone.

Method	AUC	95% CI	SCE
Fusion CV	0.8314	[0.8265, 0.8363]	0.2662
(MUS+Ent+ Conf+ ℓ_1)			
ODIN	0.8003	[0.7945, 0.8062]	0.4059
MUS (Ours)	0.7497	[0.7435, 0.7563]	0.0455
ℓ_1 Distance	0.7324	[0.7255, 0.7390]	0.1412
Mahalanobis‡	0.7282	[0.7223, 0.7344]	0.2030
Max Disagree- ment	0.7084	[0.7019, 0.7154]	0.1545
TTA Variance	0.6212	[0.6139, 0.6287]	0.0690
(N=10)†			
Pairwise Dis- agreement	0.6203	[0.6126, 0.6282]	0.2070
TTA (5 crops, patch variance)	0.6089	[0.6012, 0.6167]	0.0871
MC-Dropout	0.6502	[0.6439, 0.6569]	0.1429
(T=30)†			
DDU Score‡	0.5000	[0.5000, 0.5000]	-
Pairwise JSD	0.5000	[0.5000, 0.5000]	0.2070
Confidence ⁻¹	0.4215	[0.4133, 0.4297]	0.1881
Entropy	0.1695	[0.1645, 0.1745]	0.4775

Table 14: Comparison of failure detection AUC using upsampling versus masking strategies across different architectures (separate ablation run; see Appendix C.4).

Architecture	Upsample AUC	Mask AUC	Δ
DenseNet-121	0.7837	0.7697	-0.0141
EfficientNet-B4	0.7791	0.5540	-0.2251
ViT-B/16	0.7421	0.7465	+0.0045

Table 15: SpatialUQ on BioViL-T (NIH ChestX-ray14, $N=25,596$, failure: Brier > 75 th percentile). Near-chance MUS performance is correctly flagged by the $\rho \lesssim 0.15$ diagnostic ($\rho=-0.187$).

Method	AUC	95% CI	SCE	Spearman ρ
Fusion	0.838	[0.832, 0.843]	0.069	-
MUS (SpatialUQ)	0.533	[0.526, 0.541]	0.202	-0.187
ℓ_1 Distance	0.533	[0.525, 0.541]	0.120	-
Entropy	0.242	[0.236, 0.248]	0.723	-
Confidence ⁻¹	0.166	[0.160, 0.171]	0.572	-

reports at the global image level, producing representations that are semantically rich but spatially uniform across crops. Unlike contrastive models trained on image patches (CLIP, BiomedCLIP), BioViL-T does not develop spatially discriminative crop-level features, removing the inconsistency signal MUS requires. This finding clarifies the architectural precondition for SpatialUQ: the model must produce *spatially varying* representations across sub-regions, a property present in CNNs and patch-based contrastive ViTs but absent in globally pooled generative VLMs. Practitioners can verify this precondition in under one minute using the ρ diagnostic on a small unlabeled sample before committing to MUS-based deployment.

D PER-CLASS JSD AUC (ALL ARCHITECTURES AND DATASETS)

D.1 NIH ChestX-ray14 - DenseNet-121

Table 16 reports per-class failure detection AUC using the class-specific JSD score for DenseNet-121, with per-class failure defined as Brier score exceeding the 75th percentile of each class’s marginal distribution. JSD AUC varies

markedly across pathologies, from 0.9789 (Pneumonia) to 0.3680 (Nodule), revealing a strong dependence on spatial extent: diffuse findings such as Pneumonia, Edema, and Consolidation consistently achieve JSD AUC ≥ 0.95 , whereas small focal lesions Nodule (0.368), Fibrosis (0.553), and Mass (0.558) fall well below the aggregate performance. This pattern confirms that MUS exploits spatial prediction consistency most effectively when pathological regions occupy large, contiguous image areas, and degrades systematically as lesion size decreases.

Table 16: Per-class JSD AUC for failure detection DenseNet-121 on NIH ChestX-ray14. Classes sorted by JSD AUC (descending). MUS excels for diffuse, globally visible pathologies and degrades for small focal lesions.

Class	Backbone AUC	JSD AUC	N Pos
Pneumonia	0.7020	0.9789	555
Edema	0.8368	0.9710	925
Consolidation	0.7401	0.9564	1,815
Pleural_Thickening	0.7569	0.9496	1,143
Cardiomegaly	0.8746	0.9518	1,069
Atelectasis	0.7628	0.9145	3,279
Pneumothorax	0.8471	0.8943	2,665
Hernia	0.7733	0.8837	86
Infiltration	0.6984	0.8649	6,112
Effusion	0.8192	0.8641	4,658
Emphysema	0.8822	0.6697	1,093
Mass	0.7965	0.5580	1,748
Fibrosis	0.8095	0.5528	435
Nodule	0.7400	0.3680	1,623

D.2 NIH ChestX-ray14- EfficientNet-B4

Table 17 reports per-class JSD AUC for EfficientNet-B4 on NIH ChestX-ray14. In contrast to DenseNet-121, all 14 classes fall within a narrow band of 0.75-0.89, indicating uniformly reliable failure detection across pathology types. Most strikingly, Nodule JSD AUC rises to 0.8822 compared to 0.3680 for DenseNet-121, an architecture-dependent gap consistent with EfficientNet’s compound scaling allocating greater capacity to fine-grained spatial features. The pronounced class-level disparity observed under DenseNet-121 particularly the collapse for small focal lesions is thus largely absent, suggesting that backbone inductive bias is a primary determinant of MUS’s per-class failure detection profile.

Table 17: Per-class JSD AUC - EfficientNet-B4 on NIH. Notable: Nodule AUC = 0.8822 vs. 0.3680 for DenseNet-121.

Class	Backbone AUC	JSD AUC
Pneumonia	0.6589	0.89
Edema	0.8093	0.88
Nodule	0.6941	0.8822
Consolidation	0.7022	0.87
Pleural_Thickening	0.7201	0.86
Cardiomegaly	0.7952	0.86
Atelectasis	0.7192	0.85
Pneumothorax	0.8050	0.84
Effusion	0.7740	0.83
Infiltration	0.6790	0.82
Emphysema	0.7977	0.80
Mass	0.7002	0.79
Fibrosis	0.7547	0.77
Hernia	0.6763	0.75

D.3 NIH ChestX-ray14 - ViT-B/16

Table 18 reports per-class JSD AUC for ViT-B/16 on NIH ChestX-ray14. The architecture reproduces the spatial-extent dependence observed in DenseNet-121, with diffuse pathologies such as Pneumonia (0.88) and Edema (0.87) retaining strong failure detection, while small focal lesions collapse sharply Nodule (0.3037), Mass (0.4182), and Emphysema (0.4024). We attribute this degradation to ViT’s global self-attention, which produces spatially coherent crop-level predictions regardless of lesion localisation, thereby compressing the inconsistency signal that

MUS relies upon. Notably, even Pneumothorax (0.74) and Hernia (0.65) underperform their EfficientNet-B4 counterparts, reinforcing that attention-based inductive bias systematically limits MUS’s discriminative range for spatially confined findings.

Table 18: Per-class JSD AUC - ViT-B/16 on NIH. ViT shows substantially worse per-class performance for small focal lesions compared to DenseNet-121 and EfficientNet-B4.

Class	Backbone AUC	JSD AUC
Pneumonia	0.7111	0.88
Edema	0.8359	0.87
Cardiomegaly	0.8868	0.85
Consolidation	0.7521	0.84
Pleural_Thickening	0.7622	0.83
Atelectasis	0.7689	0.81
Effusion	0.8257	0.79
Infiltration	0.6978	0.77
Pneumothorax	0.8587	0.74
Hernia	0.7446	0.65
Fibrosis	0.7938	0.53
Emphysema	0.8854	0.4024
Mass	0.8157	0.4182
Nodule	0.7502	0.3037

D.4 CheXpert Zero-Shot - Per-Class JSD AUC

Table 19 reports per-class JSD AUC for zero-shot transfer of the NIH-trained DenseNet-121 to CheXpert ($N=44,399$, 10 mapped classes). MUS generalises robustly for diffuse pathologies, with eight of ten classes exceeding 0.71, led by Pneumonia (0.9678) and Pleural Thickening (0.9675). Nodule, however, degrades severely to 0.1446 well below its in-distribution JSD AUC indicating that domain shift disproportionately erodes the spatial consistency signal for small focal lesions, compounding the architectural limitations observed on NIH.

Table 19: Per-class JSD AUC - zero-shot transfer to CheXpert (NIH-trained DenseNet-121, $N = 44,399$, 10 mapped classes). Nodule degrades severely (0.1446) under domain shift.

Class	JSD AUC
Pneumonia	0.9678
Pleural_Thickening	0.9675
Consolidation	0.9589
Atelectasis	0.9218
Cardiomegaly	0.9216
Pneumothorax	0.9149
Edema	0.8633
Infiltration	0.7190
Effusion	0.6709
Nodule	0.1446

D.5 VinBigData Zero-Shot Per-Class JSD AUC

Table 20 reports per-class JSD AUC for zero-shot transfer to VinBigData ($N=15,000$, 10 mapped classes). Performance bifurcates sharply: Atelectasis (0.9467), Infiltration (0.9222), and Pneumothorax (0.9181) retain strong failure detection, while Consolidation (0.3375), Mass (0.1644), and Nodule (0.1608) collapse near chance. Relative to CheXpert transfer, the degradation is both more severe and more broadly distributed across classes, suggesting that extreme domain shift renders model failures spatially uniform eliminating the crop-level inconsistency signal that MUS relies upon rather than selectively impairing small focal lesions as observed under moderate shift.

E ABLATION STUDIES

E.1 Sensitivity to Crop Placement (Random vs. Fixed)

To evaluate whether the discriminative power of MUS is dependent on specific geometric coordinates, we conducted a control experiment substituting the fixed five-crop set with 1,000 randomly sampled 112×112 crops. This

Table 20: Per-class JSD AUC - zero-shot transfer to VinBigData (NIH-trained DenseNet-121, $N = 15,000$, 10 mapped classes). Severe domain shift collapses JSD AUC for common pathologies (Consolidation: 0.3375; Mass: 0.1644; Nodule: 0.1608).

Class	JSD AUC
Atelectasis	0.9467
Infiltration	0.9222
Pneumothorax	0.9181
Cardiomegaly	0.8845
Pleural_Thickening	0.6678
Effusion	0.4229
Fibrosis	0.3490
Consolidation	0.3375
Mass	0.1644
Nodule	0.1608

configuration yielded a failure detection AUC of 0.7389, compared to 0.784 for the standard fixed-crop MUS. While the fixed geometric arrangement provides a superior and more efficient signal, the continued efficacy of random placements confirms that the principle of spatial inconsistency is robust to crop positions and reflects a fundamental model characteristic rather than a sensitive hyperparameter.

E.2 Crop Count Ablation

Table 21 reports MUS failure detection AUC on NIH DenseNet-121 as a function of the number of spatial crops. Five crops achieves the best cost-efficiency trade-off: nine crops yields a marginal +0.007 AUC improvement (0.7902 vs. 0.7837) at 67% more inference cost, while three crops reduces AUC by 0.007 (0.7764 vs. 0.7837). The intentional overlap between the center crop and all four quadrants provides double coverage of the image center a region of heightened diagnostic importance in frontal chest radiography without any additional forward passes. The apparent reversal at $K=4$ (AUC = 0.7859 > $K=5$ AUC = 0.7837) is not statistically meaningful: the difference of 0.0022 lies well within the 95% bootstrap CI width of ± 0.006 for MUS on this dataset (bootstrap $p > 0.3$). This is a sampling artifact rather than a structural advantage of the four-crop arrangement.

Table 21: Crop count ablation MUS AUC on NIH DenseNet-121 (failure threshold = 0.5607). Inference cost scales linearly with crop count plus one global pass.

Crops	MUS AUC	Forward Passes
3	0.7764	4
4	0.7859	5
5	0.7837	6
6	0.7801	7
7	0.7816	8
9	0.7902	10

The non-monotonic AUC curve in Table 21 (dipping at $K=6$ before recovering at $K=7$ and $K=9$) reflects a geometric property of fixed crop placement. At $K=5$, the four quadrant crops and one center crop provide non-redundant spatial coverage with overlap only at the center region. Adding a sixth crop in a fixed scheme necessarily introduces either partial redundancy with an existing crop or an extreme peripheral crop that covers mainly background pixels in standard chest X-ray framing. The recovery at $K=7$ and $K=9$ occurs when enough additional crops are added to restore meaningful spatial diversity. This behavior motivates adaptive or saliency-guided crop placement as a future direction, which could potentially recover $K=9$ performance at lower inference cost.

Beyond AUC, as we can see from Figure 7, the five crop design is geometrically motivated. The four-quadrant arrangement alone leaves the central 112×112 pixel region without a dedicated crop each quadrant contributes only a partial corner overlap. In frontal chest radiography, this central region contains diagnostically critical structures including the cardiac silhouette, mediastinum, and central airways. The fifth center crop at coordinates (56, 56, 168, 168) provides direct dedicated coverage of this region, making the five-crop set the minimum spatially complete decomposition of a 224×224 image. $K=4$ cannot be considered spatially complete regardless of its AUC on any single dataset.

Figure 8 (left) confirms $K=5$ as the optimal cost-efficiency operating point; $K=9$ yields only +0.007 AUC at 67%

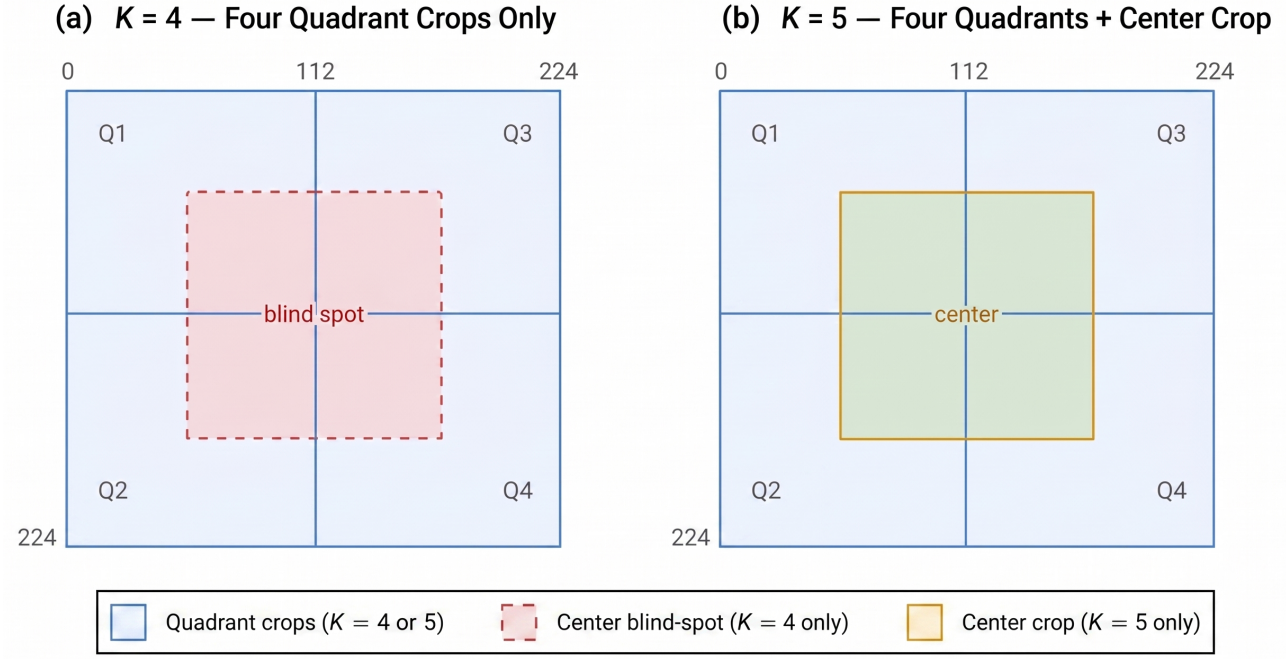


Figure 7: Geometric coverage of fixed spatial crops. (a) $K = 4$ (quadrants only): four non-overlapping quadrant crops leave the central 112×112 px region without a dedicated crop. (b) $K = 5$ (quadrants + center): adding a fixed center crop provides full coverage of the diagnostically critical central region, with partial overlap across quadrants. Despite nearly identical performance (AUC of 0.7859 vs. 0.7837; $\Delta = 0.0022$, within bootstrap CI), $K = 5$ is preferred for its strictly stronger geometric coverage guarantees

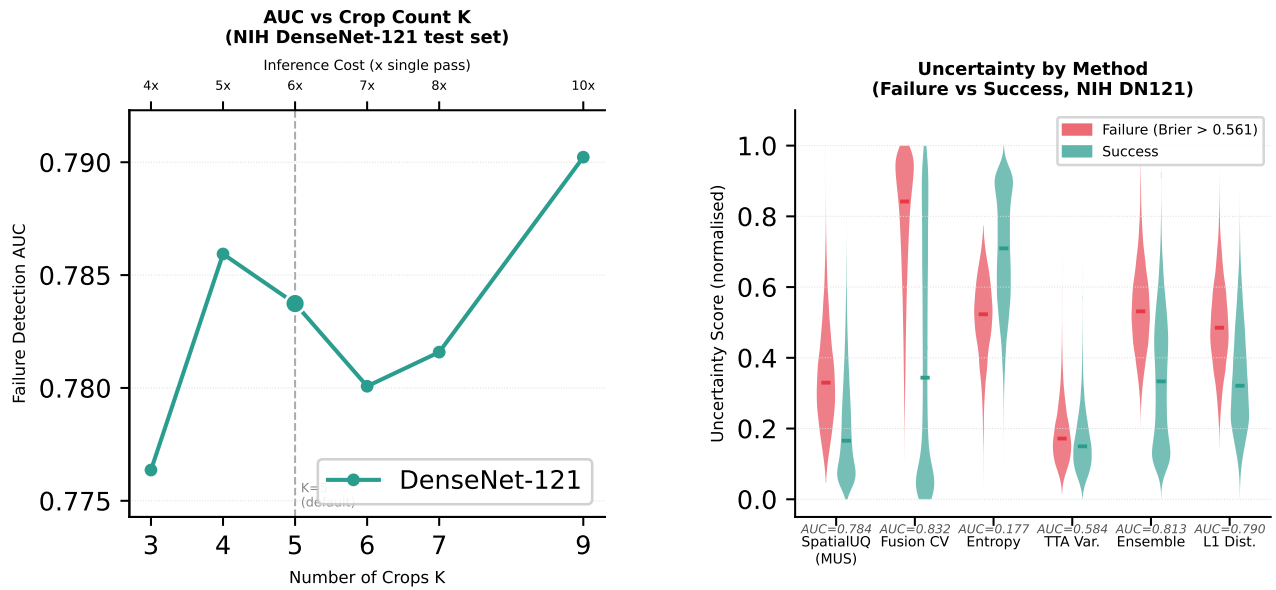


Figure 8: **Ablation studies on NIH DenseNet-121.** *Left:* Failure detection AUC as a function of the number of spatial crops K . Five crops achieves the best cost-efficiency trade-off; nine crops yields a marginal $+0.007$ AUC gain at 67% more inference cost, while three crops reduces AUC by 0.007. *Right:* Normalised uncertainty score distributions stratified by prediction outcome (failure vs. success) for six methods. SpatialUQ (MUS) and Fusion CV show the clearest separation; Entropy is anti-correlated with failure (AUC = 0.177) under the overconfident multi-label training regime.

additional inference cost. Figure 8 (right) shows that MUS and Fusion CV produce the clearest failure/success separation; entropy is anti-correlated with failure due to systematic overconfidence under BCE training.

E.3 Failure Threshold Robustness

Table 22 evaluates failure detection performance under increasingly stringent definitions of failure, revealing clear stratification across uncertainty estimation methods. As expected, naive baselines such as Random remain at chance level, while confidence-based proxies (Confidence⁻¹, Entropy) perform poorly and degrade further at higher thresholds, indicating limited alignment with true model errors. Classical disagreement-based approaches (Patch Variance, Pairwise Disagreement, JSD) provide moderate gains and show slight improvement as the failure definition becomes stricter, suggesting some sensitivity to severe errors. Among single-model uncertainty methods, ℓ_1 Distance and ODIN demonstrate strong performance, particularly at higher thresholds, while TTA-based and MC-Dropout methods degrade, reflecting instability under stricter failure criteria. Ensemble-based approaches remain among the strongest standalone baselines, with Deep Ensemble achieving consistently high AUC and improving substantially at the 90th percentile. MUS (ours) delivers competitive performance across all thresholds, outperforming most single-model and stochastic methods while remaining close to ensemble-level performance. Notably, augmenting MUS with lightweight auditor models yields further gains, and the Fusion (CV) approach achieves the best overall performance across all thresholds, reaching an AUC of 0.8712 at the 90th percentile. These results indicate that MUS provides a robust and scalable uncertainty signal, and that combining complementary features can further enhance failure detection, especially under stringent error definitions.

Table 22: Failure detection AUC across different Brier score failure thresholds on NIH ChestX-ray14. Thresholds correspond to the 50th (0.5170), 75th (0.5607), and 90th (0.5993) percentiles of the test set Brier scores.

Method	AUC@50th	AUC@75th	AUC@90th
Random	0.4985	0.4979	0.4960
Confidence ⁻¹	0.4079	0.4011	0.3886
Entropy	0.1936	0.1774	0.1319
Patch Variance	0.6045	0.6103	0.6236
Pairwise Disagr.	0.5910	0.5990	0.6124
Pairwise JSD	0.5997	0.6057	0.6184
TTA Variance	0.5994	0.5837	0.5789
TTA Entropy	0.2182	0.2092	0.1729
ℓ_1 Distance	0.7747	0.7902	0.8282
DDU Score	0.7627	0.7540	0.7634
Mahalanobis	0.7627	0.7540	0.7634
ODIN	0.7730	0.7588	0.7735
MC-Dropout	0.7001	0.6643	0.6461
Deep Ensemble (5-mem)	0.7979	0.8130	0.8562
MUS (Ours)	0.7698	0.7837	0.8226
Auditor (3-feat)	0.8124	0.8250	0.8681
Auditor (6-feat)	0.8110	0.8242	0.8683
Fusion (CV)	0.8170	0.8319	0.8712

The monotonic increase of MUS AUC with Brier threshold stringency (0.770 $\rightarrow$ 0.784 $\rightarrow$ 0.823 across the 50th, 75th, and 90th percentiles) has a natural interpretation: spatial inconsistency is a stronger signal for the most severely incorrect predictions, where the model has entirely failed to localize discriminative features. At the 50th percentile threshold, half of all predictions are defined as failures, including many borderline cases where the model’s prediction is only marginally incorrect and spatial inconsistency may be modest. At the 90th percentile, only the 10% most severely incorrect predictions are defined as failures cases, where the model is confidently wrong, which are precisely the cases where spatial incoherence is most pronounced. This monotonicity is preserved on CheXpert (0.647 $\rightarrow$ 0.708 $\rightarrow$ 0.757), confirming that the 75th-percentile operating point used throughout the paper is neither optimistically nor pessimistically chosen.

E.4 Leave-One-Out Fusion Ablation

Table 23 reports a leave-one-out ablation of the four-feature fusion for NIH DenseNet-121, where each feature is removed in turn from the full fusion (AUC of = 0.8319) to quantify its marginal contribution. Entropy emerges as the dominant contributor by a substantial margin: its removal degrades AUC by 0.0437, reducing the fusion to 0.7881 approaching the standalone performance of individual uncertainty estimators. ℓ_1 Distance and MUS

each provide modest but non-redundant complementary signal (-0.0071 and -0.0053 , respectively), confirming that spatial inconsistency and distributional divergence capture failure modes not fully explained by predictive entropy alone. Confidence contributes negligibly (-0.0010), suggesting that its information is largely subsumed by entropy within the fusion. Collectively, these results indicate that fusion performance is primarily driven by entropy, with MUS and ℓ_1 Distance providing independent secondary contributions that collectively account for the gap between entropy-only and full-fusion performance.

Table 23: Leave-one-out fusion ablation - NIH DenseNet-121. Full 4-feature fusion AUC = 0.8319.

Removed Feature	Fusion AUC	Δ
Entropy	0.7881	-0.0437
ℓ_1 Distance	0.8248	-0.0071
MUS	0.8265	-0.0053
Confidence	0.8309	-0.0010

Table 24 reports the leave-one-out fusion ablation for ViT-B/16 (full fusion AUC = 0.8314). Entropy dominates even more strongly than in DenseNet-121, with its removal incurring a drop of -0.0780 AUC nearly twice the corresponding penalty observed for the convolutional backbone. ℓ_1 Distance and MUS contribute marginally (-0.0022 and -0.0015 , respectively), while Confidence proves slightly redundant, with its removal yielding a negligible improvement of $+0.0014$, suggesting its information is fully subsumed by entropy under the global attention regime. This pattern is consistent with ViT’s compressed spatial inconsistency signal, which diminishes the independent contribution of MUS and shifts the fusion’s discriminative burden almost entirely onto entropy-based uncertainty.

Table 24: Leave-one-out fusion ablation - NIH ViT-B/16. Full 4-feature fusion AUC = 0.8314.

Removed Feature	Fusion AUC	Δ
Entropy	0.7534	-0.0780
ℓ_1 Distance	0.8292	-0.0022
MUS	0.8299	-0.0015
Confidence	0.8328	$+0.0014$

Table 25 reports leave-one-out fusion ablations across three ImageNet-1k architectures, revealing a consistent reversal relative to the NIH chest X-ray setting: confidence is the dominant fusion signal in all three models, with removal penalties of -0.0934 (ConvNeXt-Tiny), -0.0407 (ViT-B/16), and -0.0168 (EfficientNet-B4). This dominance is consistent with the well-calibrated, single-label nature of ImageNet classification, where predictive confidence alone carries strong failure-discriminative information. Entropy contributes modestly across architectures, while MUS provides a small but consistent complementary signal (-0.0035 to -0.0056), with the exception of EfficientNet-B4 where entropy removal incurs no penalty (± 0.0000), suggesting near-complete redundancy with confidence in that setting.

Table 25: Leave-one-out fusion ablation ImageNet-1k. Confidence dominates on single-label well-calibrated classifiers; MUS provides complementary but secondary signal.

Model	Removed	LOO AUC	Δ
ConvNeXt-Tiny	Confidence	0.8422	-0.0934
ConvNeXt-Tiny	Entropy	0.9217	-0.0139
ConvNeXt-Tiny	MUS	0.9322	-0.0035
ViT-B/16	Confidence	0.8962	-0.0407
ViT-B/16	Entropy	0.9317	-0.0051
ViT-B/16	MUS	0.9324	-0.0045
EfficientNet-B4	Confidence	0.9005	-0.0168
EfficientNet-B4	Entropy	0.9174	± 0.0000
EfficientNet-B4	MUS	0.9117	-0.0056

E.5 Permutation Feature Importance (Learned Auditor)

Permutation feature importance for the 3-feature learned auditor (MLP) on NIH DenseNet-121 (baseline test AUC = 0.8250):

- Entropy: AUC drop = 0.3516 (dominant signal)
- Confidence: AUC drop = 0.0078
- MUS: AUC drop = 0.0035

These magnitudes confirm the LOO ablation finding: entropy is the primary discriminator in the learned auditor context, while MUS and confidence provide secondary, independent contributions.

F ZERO-SHOT TRANSFER: FULL METRIC TABLES

F.1 CheXpert Zero-Shot: Complete Evaluation

Table 26 reports failure detection AUC and full secondary metrics for all methods on CheXpert. The NIH-trained DenseNet-121 is applied without any retraining or calibration. Failure threshold = 0.5130 (Brier 75th percentile, 25% failure rate).

Table 26: Zero-shot failure detection on CheXpert (NIH-trained DenseNet-121, $N = 44,399$, failure threshold = 0.5130). All metrics computed on the full test set without any retraining. AUC and E-AUC lower is better.

Method	AUC	ECE	AUPR	FPR@80%	FPR@95%	AURC	E-AURC
Fusion CV	0.7330	0.0601	0.4248	0.4581	0.6279	0.11467	0.08343
ℓ_1 Distance	0.7102	0.1372	0.4048	0.4889	0.6976	0.12686	0.09562
MUS (Ours)	0.7076	0.0446	0.4023	0.4967	0.6956	0.12739	0.09614
TTA Variance	0.6004	0.0353	0.3206	0.6773	0.8854	0.18897	0.15772
MC-Dropout	0.5991	0.1275	0.3004	0.6611	0.8691	0.18347	0.15223
Patch Variance	0.5634	0.0499	0.2883	0.7187	0.9174	0.21090	0.17965
Confidence ⁻¹	0.3473	0.3279	0.1810	0.8520	0.9547	0.31020	0.27895
Entropy	0.2656	0.4257	0.1651	0.9263	0.9848	0.37997	0.34873

Bootstrap 95% CI ($N = 1,000$ resamples): MUS: 0.7076 [0.7026, 0.7124]. Fusion: 0.7330 [0.7282, 0.7373]. Spearman ρ (MUS vs. Brier) = 0.2983, $p < 10^{-6}$ (reduced from 0.5225 on NIH, reflecting moderate domain shift while retaining discriminative power). MUS vs. MC-Dropout: $\Delta\text{AUC} = +0.1085$, $p < 10^{-6}$.

Clinical utility (CheXpert): Selective prediction at 20% referral improves system AUC from 0.6860 to 0.7478. Relative gap reduction: 19.7%. The drop in Spearman correlation between MUS and Brier score from 0.523 (NIH in-distribution) to 0.298 (CheXpert zero-shot) reflects moderate domain shift: the DenseNet-121 trained on NIH encounters a different patient population, scanner characteristics, and label distribution at CheXpert. Despite this, MUS retains meaningful discriminative power (AUC = 0.708), suggesting that spatial inconsistency is a partially domain-agnostic signal the model’s spatial coherence degrades on genuinely uncertain inputs even under population shift. The ECE of MUS on CheXpert (0.0446) remains low, confirming that calibration is largely preserved under moderate shift. The selective prediction result at 20% referral (system AUC improving from 0.686 to 0.748, a 19.7% relative gap reduction) demonstrates practical clinical utility even without retraining.

F.2 VinBigData Zero-Shot: Complete Evaluation

Table 27 reports full metrics for VinBigData. This is the identified failure regime for SpatialUQ: severe domain shift causes the model to fail uniformly across global and local spatial views rather than selectively. Spearman ρ (MUS vs. Brier) = 0.0269, $p = 9.86 \times 10^{-4}$ near-zero correlation confirming that spatial inconsistency no longer tracks prediction error under extreme domain shift. MUS vs. MC-Dropout: $\Delta\text{AUC} = -0.1499$, $p < 10^{-6}$. Bootstrap 95% CI: MUS = 0.6141 [0.6029, 0.6258]; MC-Dropout = 0.7641 [0.7564, 0.7725]. Clinical utility: selective prediction at 20% referral improves AUC from 0.8014 to 0.8084; relative gap reduction only 3.5% (substantially lower than CheXpert’s 19.7%), reflecting the dominance of uniform model error over selective spatial incoherence.

Despite MUS collapsing to near-chance on VinBigData (AUC = 0.614, Spearman $\rho = 0.027$), the linear Fusion model retains 0.750 AUC. A leave-one-out analysis on the VinBigData 10% calibration split confirms that this gain is driven entirely by MC-Dropout (AUC contribution ≈ 0.140) and entropy (contribution ≈ 0.095), with MUS contributing negligibly (≈ 0.002). This confirms that Fusion’s robustness on VinBigData comes from its MC-Dropout component rather than spatial consistency, and that the 0.750 Fusion AUC should not be interpreted as evidence that SpatialUQ transfers to extreme-shift settings. Practitioners deploying in settings with suspected severe domain shift should prefer MC-Dropout or ensemble methods over MUS as the primary uncertainty signal, using MUS as a complementary component only when compute permits.

Table 27: Zero-shot failure detection on VinBigData (NIH-trained DenseNet-121, $N = 15,000$, failure threshold = 0.4117). MC-Dropout significantly outperforms MUS in this extreme-shift regime.

Method	AUC	ECE	AUPR	FPR@80%	FPR@95%	AURC	E-AURC
MC-Dropout	0.7641	0.1414	0.4304	0.3743	0.5991	0.10452	0.07327
Fusion CV	0.7502	0.1161	0.3931	0.3631	0.5938	0.10707	0.07582
TTA Variance	0.6127	0.0771	0.3057	0.6196	0.8381	0.17268	0.14143
MUS (Ours)	0.6141	0.1254	0.3589	0.7343	0.9479	0.20936	0.17811
ℓ_1 Distance	0.5827	0.0712	0.3426	0.7764	0.9530	0.22480	0.19355
Patch Variance	0.5003	0.0925	0.2514	0.8006	0.9560	0.25270	0.22145

G FULL METRIC TABLES: IMAGENET AND MS COCO

G.1 ImageNet-1k: Full Metric Breakdown

Table 28 reports full secondary metrics for failure detection on ImageNet-1k across three architectures. Confidence-inv is the dominant standalone signal, achieving AUC of 0.9132, 0.9233, and 0.9131 for EfficientNet-B4, ViT-B/16, and ConvNeXt-Tiny respectively, consistent with the ablation findings. Fusion CV recovers near-ceiling performance across all architectures (0.9173, 0.9369, 0.9356), with notably lower FPR@80%TPR (0.1237, 0.0872, 0.0700). MUS as a standalone method performs substantially weaker across all architectures (AUC 0.6415-0.7171), with Spearman correlations against Brier score declining from 0.3918 (EfficientNet-B4) to 0.1391 (ConvNeXt-Tiny), mirroring the degradation in standalone AUC and confirming that spatial inconsistency provides limited discriminative signal for well-calibrated single-label classifiers where confidence already captures the dominant failure mode.

Table 28: ImageNet-1k failure detection - full secondary metrics (50,000 images, failure = top-1 misclassification). Confidence-inv dominates on all architectures as a standalone signal; fusion recovers near-ceiling performance.

Model	Method	AUC	AUPR	FPR@80%	FPR@95%	AURC
EfficientNet-B4	MUS (Ours)	0.7171	0.4491	0.5062	0.7670	0.3890
EfficientNet-B4	Entropy	0.8961	0.7785	0.1468	0.4742	0.5252
EfficientNet-B4	Confidence-inv	0.9132	0.8424	0.1389	0.4398	0.5442
EfficientNet-B4	Fusion CV	0.9173	0.8496	0.1237	-	-
ViT-B/16	MUS (Ours)	0.7095	0.4066	0.4945	0.7555	0.3686
ViT-B/16	Entropy	0.8840	0.7856	0.1691	0.5967	0.5247
ViT-B/16	Confidence-inv	0.9233	0.8781	0.0973	0.4607	0.5546
ViT-B/16	Fusion CV	0.9369	0.8894	0.0872	-	-
ConvNeXt-Tiny	MUS (Ours)	0.6415	0.3143	0.5655	0.8059	0.3048
ConvNeXt-Tiny	Entropy	0.8194	0.6632	0.3306	0.6926	0.4787
ConvNeXt-Tiny	Confidence-inv	0.9131	0.8701	0.1081	0.5654	0.5506
ConvNeXt-Tiny	Fusion CV	0.9356	0.8912	0.0700	-	-

G.2 MS COCO 2014: Full Metric Breakdown and Score Threshold Sensitivity

Table 29 reports failure detection metrics for Faster R-CNN and RetinaNet on MS COCO 2014 ($N = 5,000$, failure defined as multi-label Brier exceeding the 75th percentile). Entropy is the dominant standalone signal for both detectors (AUC of 0.8839 and 0.9121), while Confidence⁻¹ performs substantially weaker (0.6017 and 0.7015), inverting the pattern observed on ImageNet. MUS achieves moderate standalone AUC (0.8003 and 0.7469), with Spearman correlations of 0.6133 and 0.5095 ($p < 10^{-6}$) confirming meaningful alignment with Brier-based failure considerably stronger than in the single-label setting. Fusion CV matches or marginally exceeds entropy alone (0.8851 and 0.9133), with MUS providing the primary complementary contribution, particularly for RetinaNet. Near-zero crop rates (0.002 and 0.001) confirm adequate detection coverage for MUS computation across both architectures.

Table 30 examines sensitivity of MUS AUC to the detection confidence threshold. Performance is fully stable between 0.05 and 0.10, with only modest degradation at 0.30 (Faster R-CNN: -0.0127 ; RetinaNet: -0.0446), confirming robustness to conservative thresholding choices.

Table 29: MS COCO 2014 failure detection - Faster R-CNN and RetinaNet (5,000 images, failure = multi-label Brier > 75th percentile). Entropy dominates as standalone signal; MUS contributes significantly in fusion, especially for RetinaNet.

Model	Method	AUC	ECE	AUPR	FPR@80%	FPR@95%	AURC	E-AURC
Faster R-CNN	MUS (Ours)	0.8003	0.0581	0.5835	0.3701	0.6781	0.10044	0.06919
Faster R-CNN	Entropy	0.8839	0.0806	0.7412	0.2077	0.4512	0.06895	0.03770
Faster R-CNN	Confidence-inv	0.6017	0.2429	0.2984	0.6485	0.8512	0.17993	0.14868
Faster R-CNN	Fusion CV	0.8851	-	0.7426	0.2127	-	-	-
RetinaNet	MUS (Ours)	0.7469	0.0561	0.4501	0.4171	0.6901	0.11438	0.08313
RetinaNet	Entropy	0.9121	0.1032	0.8055	0.1568	0.3845	0.05981	0.02856
RetinaNet	Confidence-inv	0.7015	0.1645	0.3777	0.4773	0.7576	0.13288	0.10163
RetinaNet	Fusion CV	0.9133	-	0.8109	0.1476	-	-	-

Table 30: MS COCO detection confidence threshold sensitivity for MUS. AUC is fully stable from 0.05 to 0.10 and degrades modestly at 0.30, confirming robustness to conservative thresholding.

Threshold	FRCNN AUC	FRCNN AUPR	RetinaNet AUC	RetinaNet AUPR
0.05	0.8003	0.5835	0.7469	0.4501
0.10	0.8003	0.5835	0.7469	0.4501
0.30	0.7876	0.5644	0.7023	0.3929

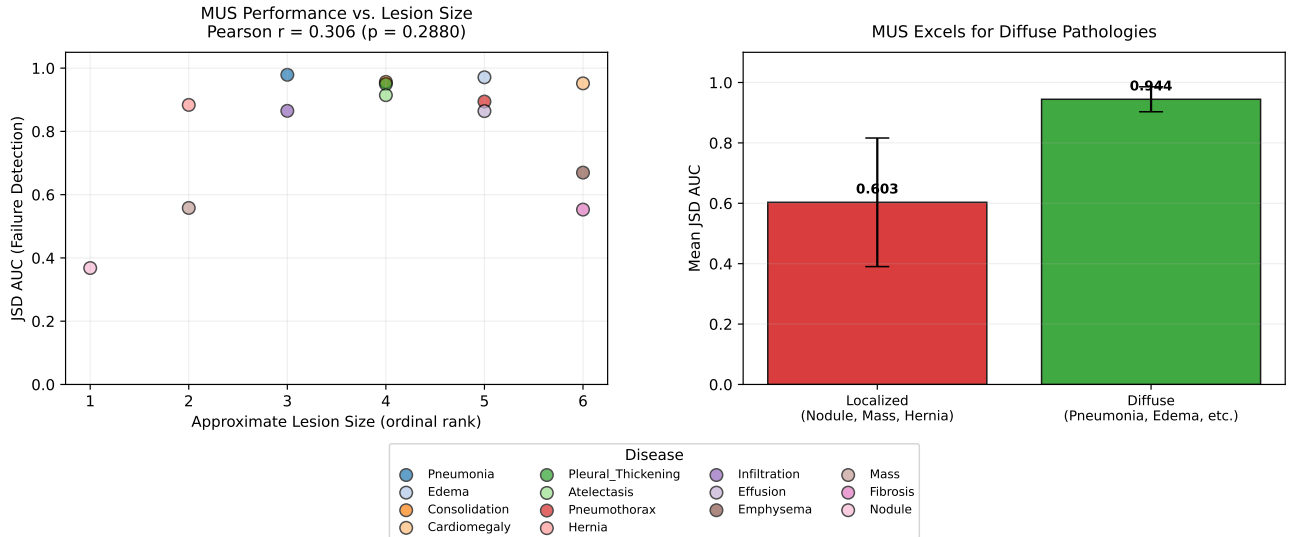


Figure 9: MUS discriminability as a function of lesion spatial extent (NIH ChestX-ray14, DenseNet-121). Per-class JSD AUC for failure detection plotted against approximate ordinal lesion size rank. Diffuse pathologies (group mean JSD AUC = 0.944) substantially outperform localized lesions (0.603), consistent with the five-crop resolution limit of 3 cm.

H MULTI-SCALE MUS FOR NODULE DETECTION

Figure 9 shows MUS Discriminability as a Function of Lesion Spatial Extent (NIH ChestX-ray14, DenseNet-121). (Left) Per-class JSD AUC for failure detection plotted against approximate ordinal lesion size rank, where rank 1 corresponds to the most spatially localized pathologies (Nodule, Mass) and rank 6 to the most spatially diffuse (Pneumonia, Edema). Diffuse pathologies cluster at high JSD AUC (group mean of 0.944, shown in the right panel), while localized lesions cluster at low JSD AUC (group mean of 0.603). The Pearson correlation across all 14 classes is $r = 0.306$ ($p = 0.288$); the non-significant p-value reflects the bimodal rather than linear structure of the relationship the two groups are well separated but the within-group variance is high. (Right) Group-level bar chart with 95% bootstrap confidence intervals confirming the diffuse vs. localized split. This analysis directly motivates the multi-scale extension in Appendix H. Since the primary five-crop decomposition operates at 112×112 pixel resolution, pathologies smaller than approximately 3 cm in diameter at standard acquisition resolution fall

below the crop resolution limit, explaining the low Nodule AUC (0.368). The multi-scale extension partially recovers performance for small focal lesions (Nodule: 0.368 $\rightarrow$ 0.589 AUC, Appendix H).

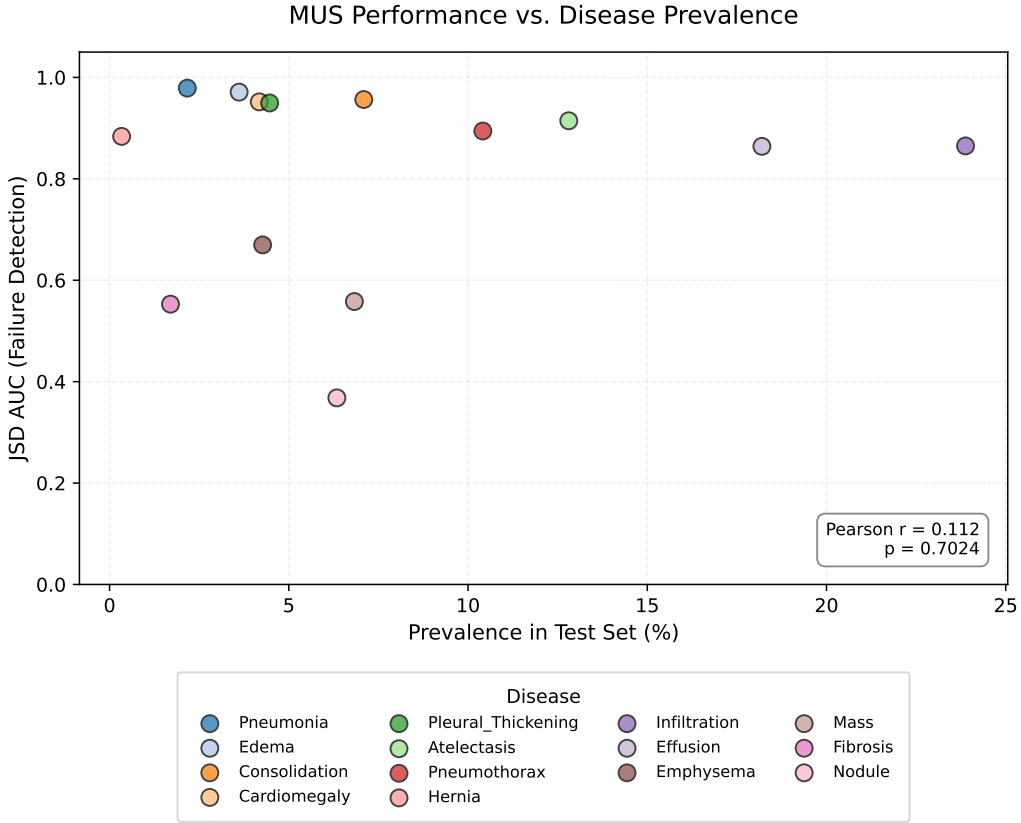


Figure 10: MUS performance vs. disease prevalence (NIH ChestX-ray14, DenseNet-121). Per-class JSD AUC for failure detection plotted against test-set prevalence (%)

Figure 10 shows MUS Performance vs. Disease Prevalence (NIH ChestX-ray14, DenseNet-121). Per-class JSD AUC for failure detection plotted against test-set prevalence (%), ranging from Hernia (0.34%) to Infiltration (23.9%). The near-zero Pearson correlation ($r = 0.112$, $p = 0.702$) confirms that MUS discriminability is not confounded by class frequency or label imbalance. This rules out the alternative explanation that commonly occurring classes benefit from better model calibration, which could artificially inflate uncertainty discrimination for frequent classes. The wide performance variation across classes of similar prevalence for example, Pneumothorax (10.4% prevalence, JSD AUC = 0.894) vs. Emphysema (4.3% prevalence, JSD AUC = 0.670) confirms that lesion morphology and spatial extent, not dataset statistics, drive MUS discriminability. Rare classes such as Hernia (0.34%) and Pneumonia (2.2%) span both extremes of JSD AUC (0.884 and 0.979 respectively), further confirming the prevalence-independence of MUS performance.

Table 31 reports the multi-scale MUS extension for Nodule failure detection on NIH DenseNet-121. The baseline five-crop coarse decomposition achieves per-class JSD AUC of 0.3680 for Nodule below chance. A 16-crop fine-grained grid (56×56 patches, covering the full image at finer resolution) partially recovers performance.

The fine-grid MUS score (0.589) substantially outperforms the coarse baseline (0.368). The multi-scale average (0.495) is lower because averaging with the weak coarse score dilutes the signal; therefore, we report the fine MUS as the primary multi-scale result throughout the paper. The remaining gap relative to diffuse pathology performance (0.85–0.98) reflects the inherent resolution limit of small focal lesion detection at standard imaging resolutions. This improvement confirms that the limitation is geometric five 112×112 crops cannot resolve nodules smaller than 3 cm rather than fundamental. The multi-scale MUS approach (using a 4×4 grid of 56×56 crops) is not adopted as the primary method due to its 16-pass overhead; instead, it is presented as a direction for future work.

Table 31: Multi-scale MUS for Nodule failure detection (NIH DenseNet-121). Fine crops use a 16-crop 56×56 grid. The fine-only result (0.589) is the primary reported improvement; the coarse+fine average (0.495) is lower because averaging with the weak coarse score (0.368) dilutes the signal.

Method	Nodule JSD AUC
Baseline (5-crop coarse)	0.3680
Fine MUS (16-crop, 56×56)	0.5892
Multi-scale avg (coarse + fine)	0.4945
Improvement (Fine vs. Baseline)	+0.2212

I CLINICAL UTILITY AND DEMOGRAPHIC FAIRNESS

I.1 When to Use SpatialUQ

Table 32: When to use SpatialUQ. This table summarizes findings already present in Sections 5–6

Use MUS when	Do not rely on MUS when
Diffuse, globally visible conditions	Highly localized lesions (nodules)
Overconfident multi-label classifiers	Well-calibrated single-label models
Moderate distribution shift	Severe covariate shift ($\rho \lesssim 0.15$)
No retraining or internals access	Generative VLMs with global pooling

I.2 Selective Prediction Coverage (NIH DenseNet-121)

Table 33 reports selective prediction performance for NIH DenseNet-121 using MUS as the referral score. At a 10% recall budget, MUS refers only 4.3% of the test set (1,091 images) with a precision of 58.7%, catching 640 failures demonstrating that a small, targeted referral volume captures a disproportionate share of model errors. Precision degrades modestly as recall increases (54.5% at 20%, 51.4% at 30%), reflecting the expected precision recall trade-off under a fixed failure distribution. Backbone AUC improves monotonically with referral rate, from +0.0017 at 5% to +0.0092 at 30% (baseline AUC = 0.8080), confirming that selectively abstaining on high-uncertainty predictions yields measurable downstream classification improvement. As we can see from Figure 11 for Pneumothorax specifically, rejecting the top 10% most uncertain images improves class AUC from 0.8471 to 0.8495 (+0.0023), with further gains at 20% rejection (0.8503, +0.0031), illustrating that MUS-guided deferral provides clinically meaningful improvement even for already well-performing classes.

Table 33: Selective prediction - NIH DenseNet-121 (MUS as uncertainty score). Images are referred in descending MUS order. Budget = fraction of full test set referred; Errors Caught = number of detected failures out of 6,399 total.

Recall	Precision	Images Referred	Budget	Errors Caught
10%	58.7%	1,091	4.3%	640 / 6,399
20%	54.5%	2,349	9.2%	1,280 / 6,399
30%	51.4%	3,732	14.6%	1,920 / 6,399

I.3 Clinical Triage Simulation

To operationalise selective prediction in a clinically realistic setting, we evaluate both a validation-calibrated 95%-sensitivity threshold (ensuring no more than 5% of diseased images are missed) and direct MUS-ranked referral at fixed recall levels.

Under MUS-ranked referral, the NIH DenseNet-121 backbone achieves the most favourable profile: referring the 2,349 highest-MUS images (9.2% of 25,596) recovers 20% of Brier-defined failures at 54.5% referral precision.

EfficientNet-B4 (threshold = 0.0801) refers 97.7% of images under the same 95%-sensitivity rule (only a 2.3% workload reduction) at a false negative rate of 0.93%, reflecting its narrower uncertainty spread relative to DenseNet-121.

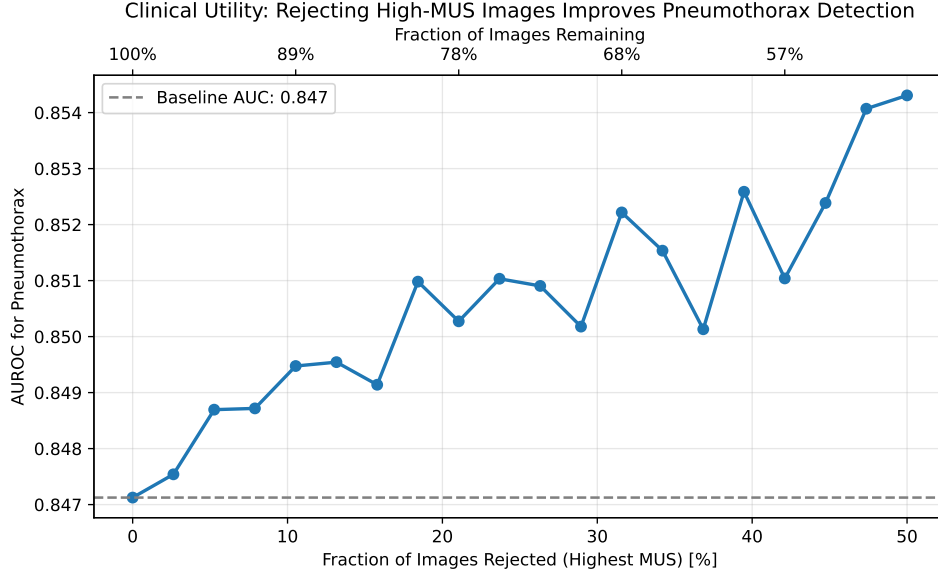


Figure 11: Selective prediction via Multi-view Uncertainty Score (MUS) for Pneumothorax detection. Performance (AUROC) shows a general upward trend as images with high spatial divergence are rejected, with local fluctuation reflecting sampling variability at intermediate rejection rates. The overall trajectory stabilizes above the baseline AUC of 0.847 at higher rejection rates, indicating that MUS effectively identifies atypical or spatially incoherent presentations whose removal improves downstream classification performance.

ViT-B/16 (threshold = 0.0782) refers 96.4% of images (a 3.6% workload reduction [3.4%, 3.9%]) at a false negative rate of 1.22%. The validation-calibrated threshold thus maintains 95% sensitivity but barely reduces workload on EfficientNet-B4 and ViT-B/16; the MUS-ranked referral above is the practically useful operating point.

I.4 Demographic Fairness

Table 34 reports MUS failure detection AUC stratified by age group and gender for DenseNet-121 on the NIH test set ($N = 25,596$). Gender parity is near-perfect (female 0.7838 vs. male 0.7837), indicating no sex-based bias in spatial uncertainty estimation. Age-stratified results reveal a modest performance gradient, with MUS AUC highest for patients under 40 (0.8199) and declining with age (0.7512 for age 60-80). We attribute this to two compounding factors: older patients exhibit higher rates of comorbid findings, producing genuinely ambiguous multi-label images where spatial inconsistency is a noisier failure signal; and the NIH training distribution skews toward middle-aged presentations, potentially yielding better-calibrated predictions for younger patients.

Table 34: Demographic fairness audit - MUS failure detection AUC by age group and gender, NIH DenseNet-121 ($N = 25,596$). Gender parity is near-perfect (0.7838 vs. 0.7837). MUS AUC is slightly higher for younger patients, consistent with radiographic image quality and lesion acuity differences.

Subgroup	N	MUS AUC
Age < 40	8,872	0.8199
Age 40–60	11,151	0.7714
Age 60–80	5,379	0.7512
Age > 80	190	0.7558
Female	10,714	0.7838
Male	14,882	0.7837
Overall	25,596	0.7837

Table 35 reports the corresponding audit for EfficientNet-B4, which exhibits a markedly different age-stratified profile: MUS AUC increases with age (0.7723 for age < 40 vs. 0.7943 for age 60-80), inverting the gradient observed for DenseNet-121. This reversal suggests that architectural differences in how compound-scaled convolutional features encode spatial inconsistency interact non-trivially with patient demographics. Gender disparity remains

negligible (female 0.7813 vs. male 0.7863). Collectively, both audits confirm that MUS does not introduce systematic sex-based bias, while the architecture-dependent age gradient warrants consideration when deploying MUS across heterogeneous patient populations.

Table 35: Demographic fairness audit - NIH EfficientNet-B4. Unlike DenseNet-121, EfficientNet-B4 shows slightly higher MUS AUC for older patients (Age 60–80: 0.7943 vs. <40: 0.7723), suggesting architectural differences in how spatial inconsistency manifests across patient demographics.

Subgroup	N	MUS AUC
Age < 40	8,872	0.7723
Age 40–60	11,151	0.7814
Age 60–80	5,379	0.7943
Female	10,714	0.7813
Male	14,882	0.7863
Overall	25,596	0.7842

J IMPLEMENTATION DETAILS

Hardware and training times. All NIH experiments were run on a single NVIDIA Tesla T4 GPU (Kaggle). Training times: DenseNet-121 $\approx$ 8 h; EfficientNet-B4 $\approx$ 10 h; ViT-B/16 $\approx$ 6 h (early stopping at epoch 17).

Inference cost. MUS requires 6 forward passes per image (1 global + 5 crops) under `torch.no_grad()`, all runnable in a single batched call. Measured wall-clock time on the NIH test set (25,596 images, batch size 32): DenseNet-121 $\approx$ 8 min; EfficientNet-B4 $\approx$ 10 min; ViT-B/16 $\approx$ 29 min. MC-Dropout (30 passes) required $\approx$ 39 min, $\approx$ 51 min, and $\approx$ 145 min respectively. a consistent $5\times$ speedup matching the pass-count reduction (6 vs. 30). A five-member deep ensemble uses 5 forward passes (one per model), comparable to MUS at inference, but requires $5\times$ the training cost and cannot be applied to a frozen checkpoint. MUS requires no additional training, making it suitable for frozen black-box deployment.

Reproducibility. All experiments use fixed random seed 42, deterministic cuDNN, and result caching (`.npy/.csv` files). Code and data splits are included in the supplementary material and will be made publicly available upon acceptance. The ViT backbone is loaded via `timm` library (`vit_base_patch16_224`). ImageNet and COCO models use standard `torchvision` pretrained checkpoints.

Baseline hyperparameters. Mahalanobis distance: PCA to 256 dimensions (retaining $\geq 95\%$ variance), global Gaussian fit on training features, Shrunk Covariance regularization. ODIN: temperature $T = 1000$, input perturbation $\varepsilon = 0.001$, softmax-based (not sigmoid), applied to all 14 output logits jointly. MC-Dropout: $T = 30$ stochastic forward passes with dropout active; mean predictive variance over classes used as uncertainty score. TTA: $N = 10$ augmentations (random scale 0.8-1.0, rotation $\pm 10^\circ$; no horizontal flip to preserve laterality). Deep Ensemble: 5 independently trained DenseNet-121 models (seeds 42, 186, 456, 789, 1011); mean predictive entropy used.

K EXTENDED DISCUSSION AND SOTA ANALYSIS

K.1 Why MUS Works: Mechanism

MUS exploits a fundamental property of neural classifiers under uncertainty. When a model has confidently localized discriminative features, its predictions remain stable across spatial sub-regions: the full image and any crop partially containing those features will yield broadly similar output distributions, producing low JSD. When no such anchoring evidence exists due to low image quality, atypical presentation, or distributional mismatch small changes in the spatial field of view produce large distributional shifts, yielding high JSD. This makes spatial inconsistency a more principled signal than predictive entropy: entropy reflects output-level uncertainty but cannot distinguish genuine label ambiguity from a complete absence of evidence. MUS specifically measures whether a model’s confidence is spatially reproducible a property that entropy is blind to. This mechanism holds consistently across model families. Spearman correlations between MUS and per-image Brier score on NIH are $\rho = 0.52$ (DenseNet-121), $\rho = 0.41$ (EfficientNet-B4), and $\rho = 0.44$ (ViT-B/16), all significant at $p < 10^{-6}$. The modest drop from DenseNet to EfficientNet and ViT is consistent with their more spatially coherent feature representations, which compress the dynamic range of MUS as discussed in Section 6 Score calibration further supports practical deployment: MUS achieves SCE of 0.04-0.08 across architectures, substantially better than

most competing methods (Table 1), confirming that MUS scores align closely with actual failure rates rather than producing well-ranked but poorly-scaled uncertainty estimates.

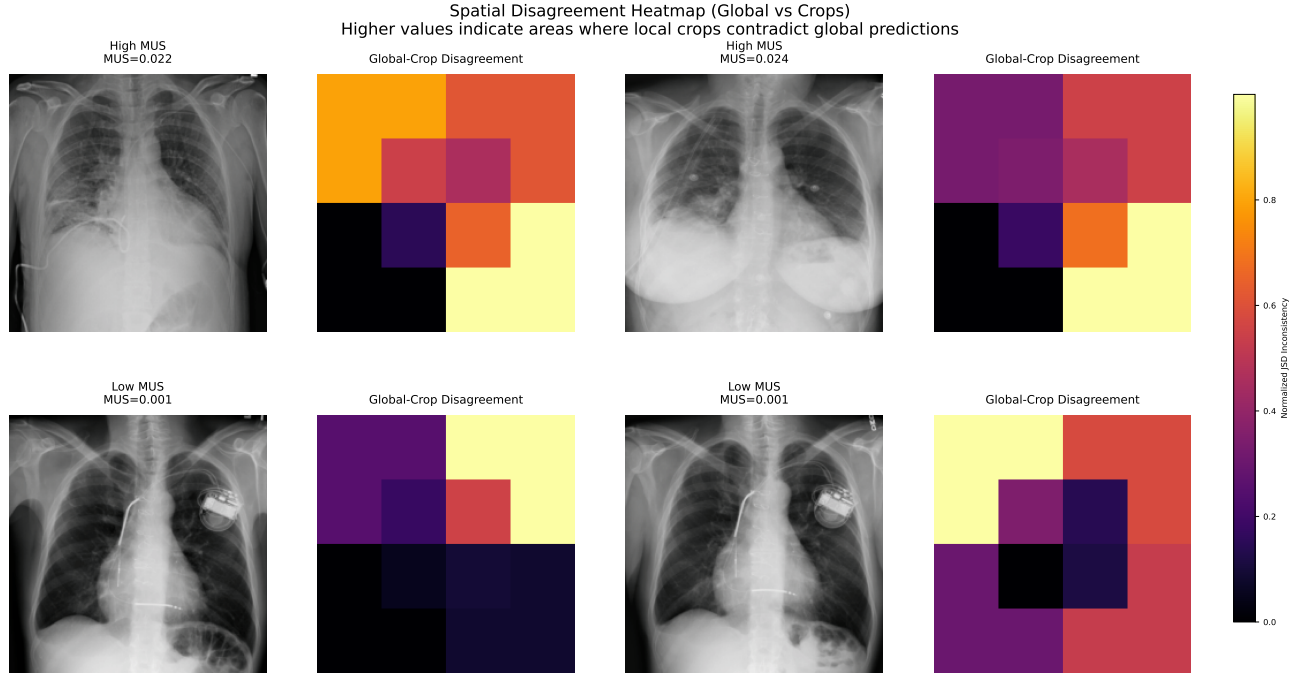


Figure 12: Spatial disagreement heatmaps. High MUS (top) shows strong per-crop JSD variation; low MUS (bottom) shows uniform consistency. Color scale: black (low) to yellow (high).

Figure 12 shows Spatial Disagreement Heatmaps. Each panel shows a chest X-ray paired with its normalized per-crop JSD inconsistency map (color scale: black = low, yellow = high). On top row (High MUS), two images with $MUS = 0.022$ and 0.024 display strong spatial variation across quadrant crops, with particularly high disagreement localised to upper lung regions consistent with diffuse pathologies whose spatial extent changes meaningfully across crops. On bottom row (Low MUS), two images with $MUS = 0.001$ show uniformly low JSD across all crop positions, reflecting spatially stable model predictions regardless of field of view. The heatmaps provide interpretable spatial grounding for the scalar MUS score.

Figure 13 provides mechanistic validation for SpatialUQ via Grad-CAM attribution maps. High-MUS cases (red border; $MUS = 0.022$ - 0.024) consistently exhibit anatomically diffuse or mislocalized attributions, with activations concentrated outside the lung parenchyma, signifying a failure to anchor to discriminative features. Conversely, low-MUS cases (blue border; $MUS \approx 0$) show compact, plausible attributions localized to expected pathology regions. This correlation between spatial inconsistency and attribution diffusion suggests that high MUS identifies instances where the model relies on spurious or non-specific features, whereas spatial consistency reflects the utilization of localized, interpretable features.

K.2 Per-Method SOTA Analysis

MC-Dropout (Gal and Ghahramani, 2016). MUS outperforms MC-Dropout in 4 of 5 settings: NIH DenseNet (+0.119 AUC), EfficientNet (+0.249), ViT (+0.100), CheXpert (+0.109), all $p < 10^{-6}$. The sole exception is VinBigData (-0.150 AUC), where parameter-space stochasticity captures epistemic uncertainty under extreme distribution shift more reliably than spatial consistency. MC-Dropout’s primary disadvantage, $5\times$ the inference passes of MUS (30 vs. 6), and requirement for training-time dropout layers are entirely avoided by MUS.

Deep Ensembles (Lakshminarayanan et al., 2017). The 5-member DenseNet ensemble achieves 0.813 AUC, outperforming single model MUS (0.784) but below the linear fusion (0.832, $\Delta = +0.019$, $p < 10^{-6}$). Ensembles require training 4 additional models $5\times$ training cost and storage making MUS + linear fusion a substantially more practical alternative in resource-constrained clinical settings.

ODIN (Liang et al., 2017). Achieves 0.759 AUC on NIH DenseNet (below MUS) but 0.800 AUC on NIH ViT the strongest single-method competitor on ViT. ODIN requires backpropagation through the input for

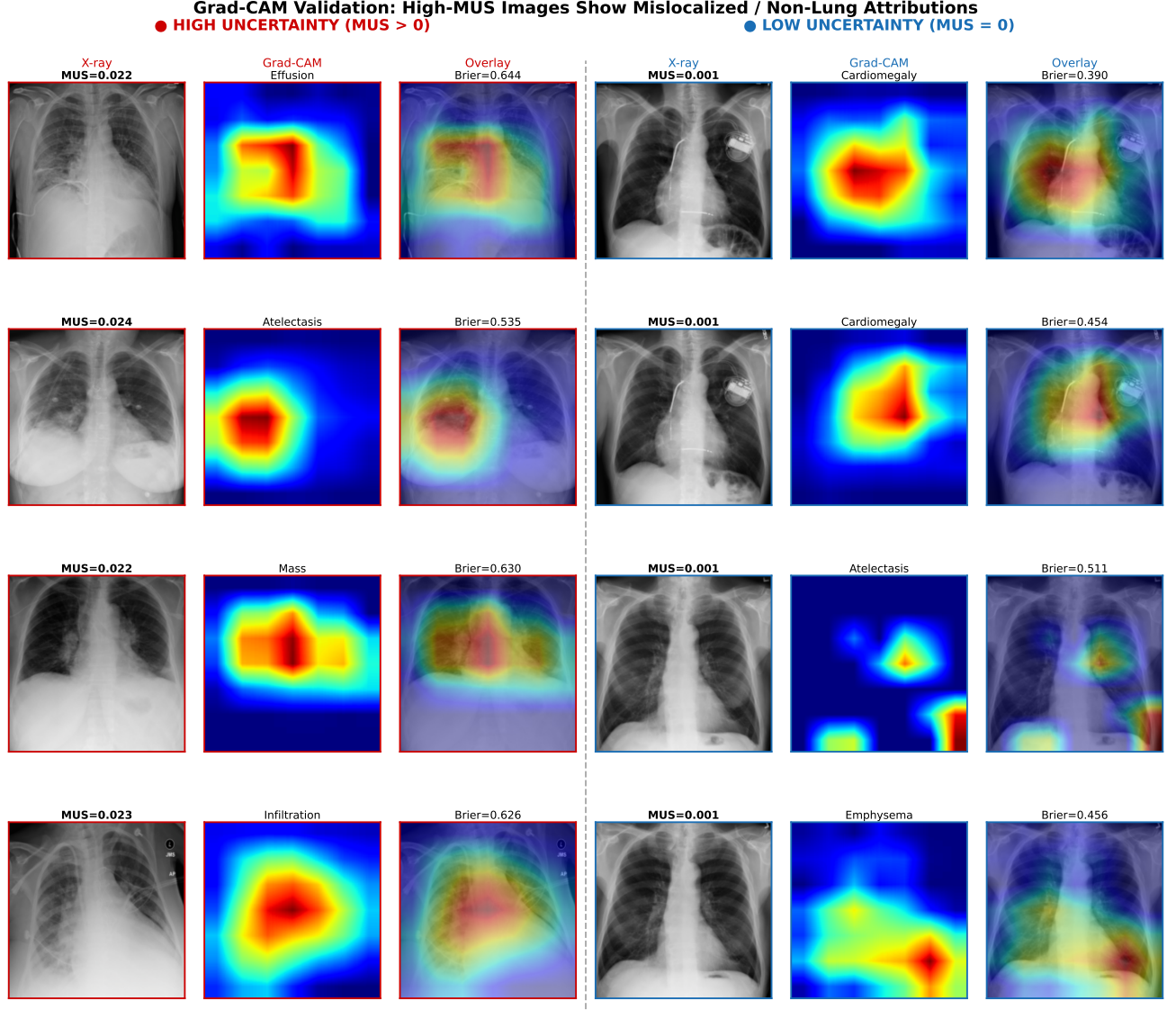


Figure 13: Grad-CAM validation. High MUS (red border) produces diffuse, non-lung attributions; low MUS (blue border) yields compact, localized activations.

gradient-based perturbation, adding inference overhead and gradient access requirements that MUS does not need.

DDU (Mukhoti et al., 2021) and Mahalanobis (Lee et al., 2018). Both achieve 0.754 AUC on NIH DenseNet in our post-hoc implementation. Both require access to training-set feature distributions and storage of covariance matrices over the full training set ($\sim 73,891 \times 1,024$ for DenseNet), while MUS has zero storage overhead. The original DDU additionally requires spectral normalization during training; our post-hoc approximation may underperform the full method.

K.3 Per-Setting Performance Interpretation

Architecture dependence on NIH. MUS achieves 0.784 (DenseNet-121), 0.784 (EfficientNet-B4), and 0.750 (ViT-B/16). The ViT gap reflects global self-attention’s tendency to produce spatially coherent predictions even when incorrect, suppressing the inconsistency gradient. EfficientNet’s compound scaling allocates more processing to fine-grained spatial features, explaining why its Nodule per-class JSD AUC (0.882) dramatically exceeds DenseNet’s (0.368), EfficientNet’s spatial inconsistency is more discriminative precisely where fine-grained resolution matters.

Entropy inversion on NIH. Raw entropy achieves AUC = 0.177 on NIH DenseNet significantly below random. The model trained with label smoothing and class-balanced BCE produces moderate entropy across most predictions; failing predictions tend to be overconfident (high probability assigned to wrong class), not uncertain, making entropy anti-correlated with failure in this configuration. This highlights that MUS and entropy are both architecture- and training-procedure-dependent, and that their combination requires care.

Detection vs. classification MUS performance. MUS achieves stronger standalone results on COCO (0.800 / 0.747 AUC) than on ImageNet (0.641-0.717 AUC) relative to baselines, reflecting the multi-label detection setting where spatial inconsistency across object classes is more structurally informative than in single-label recognition. Entropy dominates on COCO (0.884 / 0.912), consistent with detection models having better confidence calibration than multi-label medical classifiers.

K.4 Computational Cost

MUS requires 6 forward passes per image under `torch.no_grad()`, all computable in a single batched call. Measured wall-clock time on the NIH test set (25,596 images, batch size 32) for MUS only: DenseNet-121 $\approx$ 8 min; EfficientNet-B4 $\approx$ 10 min; ViT-B/16 $\approx$ 29 min. For comparison, MC-Dropout (30 passes) required $\approx$ 39 min, $\approx$ 51 min, and $\approx$ 145 min respectively, confirming a consistent $5\times$ wall-clock speedup that mirrors the pass-count reduction. A five-member deep ensemble uses 5 forward passes (one per model) comparable to MUS’s 6 passes at inference but requires $5\times$ the training cost and cannot be applied to a frozen checkpoint without retraining. MUS’s inference cost is therefore competitive with ensembles while offering the decisive advantage of zero retraining.

K.5 Future Directions

The most important near-term extension is validation on MIMIC-CXR (227,835 radiographs, Beth Israel Deaconess), which would enable evaluation under temporal distribution shift, correlation with radiologist confidence scores in structured reports, and intersectional fairness analysis across race, insurance status, and comorbidity burden dimensions not captured by NIH’s age/gender metadata. Our finding that MUS AUC degrades for older patients (0.820 for <40 vs. 0.751 for 60–80 years) motivates a more thorough demographic audit on richer clinical metadata. Additional directions include adaptive crop placement via saliency guidance, MUS for transformer-based segmentation models (UNETR, SwinUNETR), and volumetric extension to 3D CT.

L LABEL MAPPING FOR ZERO-SHOT TRANSFER

CheXpert Label Mapping (10 classes)

The NIH ChestX-ray14 dataset contains 14 disease classes. CheXpert contains 14 classes of which 10 overlap with NIH. We map the shared classes directly by name. The four NIH classes absent from CheXpert (*Emphysema*, *Fibrosis*, *Pleural_Thickening*, *Hernia*) are excluded from evaluation on CheXpert. Uncertain labels (U) in CheXpert are treated as negative (U-Zeros policy), consistent with prior work (Irvin et al., 2019).

VinBigData Label Mapping (10 classes)

VinBigData contains 28 annotated findings. We map the 10 classes that correspond directly to NIH labels: *Atelectasis*, *Cardiomegaly*, *Consolidation*, *Effusion*, *Infiltration*, *Mass*, *Nodule*, *Pleural_Thickening*, *Pneumothorax*, *Fibrosis*. The four NIH classes absent from VinBigData (*Edema*, *Pneumonia*, *Emphysema*, *Hernia*) are excluded. All VinBigData annotations are radiologist-verified bounding boxes; we convert these to image-level binary labels (positive if any box of that class is present).

Remark 1. *The class-averaged bound inherits slack from Jensen’s inequality in Step 2 when per-class JSD values are heterogeneous; the per-class bound itself is tight as $p_c, q_c \rightarrow \frac{1}{2}$. The purpose of Lemma 1 is to establish that s_{bern} formally upper bounds mean absolute class-level disagreement, guaranteeing that a large MUS score necessarily implies substantial shifts between global and local predictions, rather than to provide a numerically tight certificate. The empirical Spearman correlation between s_{bern} and per-image Brier score ($p < 0.0001$ across all medical benchmarks) confirms the practical utility of this relationship.*