MORE INFORMATION, LESS UNCERTAINTY? MONOTONIC UNCERTAINTY REGULARIZATION UNDER PROGRESSIVE INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Many real-world predictive systems acquire information progressively. In hospitals, for example, clinicians may order a costlier, slower, or more invasive test only when a simpler one is inconclusive. At each step, a clinician must decide whether she has enough evidence or should acquire more, and predictive uncertainty is a natural signal for this choice. In machine learning, however, such uncertainty is usually evaluated and assessed at each information level individually, not along the sequence of progressive acquisition where these decisions must be made based on uncertainty. Through empirical evaluation, we uncover a phenomenon of *aggregate monotonicity illusion*: as inputs exhibit richer and cleaner signal, models become more accurate and less uncertain on average across a whole dataset, yet when considering each sample separately, their uncertainty increases with additional signal more than half of the times. This holds across three datasets, Imagenette, ImageNet-1k, and chest X-ray classification, spanning different scales, domains, tasks, and architectures. We address this issue with **M**onotonic **U**ncertainty **R**egularization (MUR), a family of lightweight fine-tuning objectives that links predictions for the same sample across multiple information levels and encourages progressively sharper predictive distributions. MUR significantly reduces the issue of aggregate monotonicity illusion while maintaining predictive performance, and improves the cost–performance trade-off of progressive acquisition. These results caution against relying on aggregate uncertainty trends for per-sample acquisition decisions and show that lightweight cross-level fine-tuning can make uncertainty a more reliable signal for deciding whether to acquire additional information.

1 INTRODUCTION

Many diagnostic and decision-making processes acquire information progressively (Yeo, 2016; Trapeznikov et al., 2012). A clinician, for instance, may begin with an inexpensive, non-invasive examination such as a chest X-ray and proceed to a more informative test only when the evidence is insufficient (Yeo, 2016; Christensen et al., 2025). Each escalation can improve the decision but adds monetary cost, delay, radiation exposure, or demand on limited clinical resources (Smith-Bindman et al., 2009; Trapeznikov & Saligrama, 2013). Similar trade-offs arise when predictive systems receive progressively higher-resolution inputs, additional views, or higher-quality sensor measurements (Yang et al., 2020; Shim et al., 2018; Trapeznikov et al., 2012). We call this setting *progressive information*: the same sample is observed through an ordered sequence of increasingly informative inputs. We call the resulting decision process *progressive acquisition*: at each level, the system must decide whether current evidence supports a reliable prediction or richer information is worth its cost.

Predictive uncertainty is a natural signal for this stopping decision. Confidence-based methods such as selective prediction, early-exit networks, and model cascades use it to abstain or allocate more computation, but hold the input fixed (Geifman & El-Yaniv, 2017; 2019; Teerapittayanon et al., 2016; Huang et al., 2017; Enomoto & Eda, 2021; Jitkrittum et al., 2023). Acquisition methods instead use confidence to route samples toward richer observations or features (Shim et al., 2018). Although these routing decisions rely on uncertainty along the trajectory of the same sample, standard training optimizes and evaluates predictors at each information level separately. This raises a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

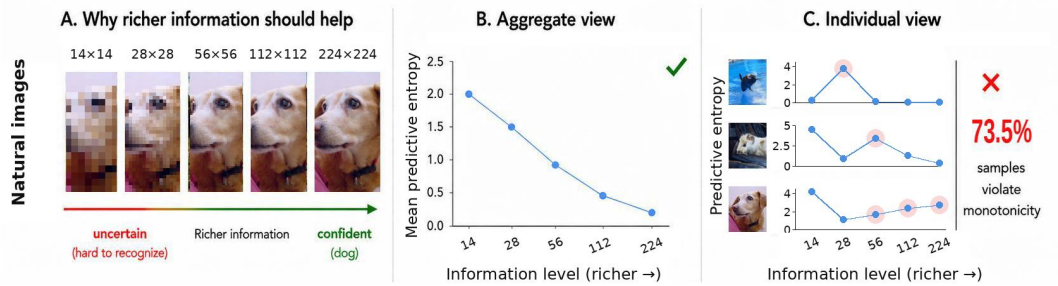

Figure 1: **The Aggregate Monotonicity Illusion under progressive information.** **(A)** Progressively richer observations are expected to reduce predictive uncertainty. **(B)** At the dataset level, mean predictive entropy decreases with increasing information, suggesting monotonic behavior. **(C)** Individual trajectories, however, reveal frequent violations: 73.5% of ImageNet-1k samples become more uncertain at least once. Thus, a favorable aggregate trend can conceal widespread non-monotonicity at the sample level, where progressive acquisition decisions are made.

basic question for progressive acquisition: **Do neural networks actually become more certain as progressively richer information becomes available for the same sample?**

We find that the answer changes markedly when uncertainty is examined at the sample level. Under standard training, models can become more accurate and less uncertain on average as information increases while producing non-monotonic uncertainty trajectories for individual samples (Figure 1). We call this discrepancy the **Aggregate Monotonicity Illusion**. In a controlled five-level resolution hierarchy on Imagenette, 90.4% of samples exhibit at least one entropy-monotonicity violation. The failure persists at ImageNet-1k scale and in multi-label chest X-ray classification under a clinically motivated degraded-to-clean hierarchy, spanning domains, task formulations, architectures, and information axes (Figure 2). This distinction matters because aggregate uncertainty is evaluated over populations, but acquisition decisions are made for individual samples. We do not assume that every increase in uncertainty is undesirable, since richer evidence may legitimately reveal ambiguity. Rather, monotonicity provides an operational notion of cross-level coherence, and we ask whether reducing widespread reversals makes uncertainty more useful for progressive acquisition.

The illusion reflects a mismatch between training and deployment. Standard cross-entropy objectives optimize predictions at each information level but impose no relationship on uncertainty for the same sample across levels. However, deployment relies on exactly this cross-level behavior to decide whether to acquire richer information. To address this mismatch, we introduce **Monotonic Uncertainty Regularization (MUR)**, a family of lightweight second-stage fine-tuning objectives that explicitly couple predictions at adjacent information levels. **MUR-U** directly regularizes scalar uncertainty ordering by penalizing increases in predictive entropy, while **MUR-AU** transfers the richer-level predictive distribution to the poorer level, jointly regularizing class-specific predictive structure and uncertainty through cross-level distribution matching. The two variants expose an important distinction: MUR-U isolates trajectory coherence, whereas MUR-AU additionally preserves the predictive structure needed to identify which samples may benefit from richer observations.

Our contributions are:

1. We identify and formalize the **Aggregate Monotonicity Illusion**: models that appear increasingly accurate and certain in aggregate can exhibit widespread non-monotonic uncertainty trajectories for individual samples. We establish the breadth of this phenomenon across complementary natural and clinical imaging settings.
2. We introduce two complementary forms of **Monotonic Uncertainty Regularization**, applicable as lightweight second-stage fine-tuning to existing predictors. MUR-U regularizes scalar uncertainty ordering, while MUR-AU jointly regularizes predictive structure and uncertainty through cross-level distribution matching. Both sharply reduce sample-level uncertainty violations while preserving performance, repairing incoherent transitions while rarely disrupting coherent ones.
3. We show that MUR more reliably supports progressive acquisition than level-wise calibration or within-level confidence control. MUR-AU consistently improves the cost–performance trade-off across all settings, by up to 18.9%, whereas MUR-U achieves stronger uncertainty ordering in some settings without always improving routing.

2 THE AGGREGATE MONOTONICITY ILLUSION

We study whether *each sample's* predictive uncertainty evolves coherently as progressively richer observations of that sample become available. We now formalize this setting, distinguish aggregate from sample-level monotonicity, and explain why the distinction matters for progressive acquisition.

2.1 PREDICTION UNDER PROGRESSIVE INFORMATION

Let $\mathcal{D} = \{(z_i, y_i)\}_{i=1}^N$ denote a dataset of observation sequences $z_i = (x_i^{(0)}, \ldots, x_i^{(L)})$ and targets y_i. For each sample i, the acquisition process provides an ordered sequence of observations

$$x_i^{(0)} \prec x_i^{(1)} \prec \cdots \prec x_i^{(L)}, \tag{1}$$

where $x_i^{(\ell)}$ is the observation at information level ℓ, and $\prec$ denotes an ordering from less to more informative observations. The ordering is induced by the acquisition process rather than inferred by the model and may represent progressively higher image resolution, improved image quality, additional measurements or increasingly informative diagnostic examinations. Each level is associated with an increasing acquisition cost $c_0 < c_1 < \cdots < c_L$, where cost may represent computation, acquisition time, monetary expense, radiation exposure, or another application-dependent resource. Progressive information therefore creates a natural trade-off: richer observations may support better predictions, but acquiring the richest observation for every sample may be unnecessary or costly.

At level ℓ, a predictor f_ℓ receives $x_i^{(\ell)}$ and produces a predictive distribution $p_\ell(\cdot \mid x_i^{(\ell)})$. We denote the uncertainty of this prediction by $H_\ell(z_i)$: predictive entropy for multiclass prediction and mean binary predictive entropy across labels for multilabel prediction.

A natural coherence property for the uncertainty trajectory of sample i is

$$H_0(z_i) \geq H_1(z_i) \geq \cdots \geq H_L(z_i), \tag{2}$$

so that uncertainty does not increase as richer information about the same sample becomes available. We do not treat this ordering as universal: richer evidence may legitimately reveal ambiguity for particular samples. Rather, we ask how frequently standard predictors violate it and whether reducing such violations makes uncertainty more useful for progressive acquisition.

2.2 AGGREGATE MONOTONICITY CAN HIDE SAMPLE-LEVEL VIOLATIONS

Predictions under progressive information can be evaluated at two distinct levels. At the dataset level, let

$$\overline{H}_\ell = \frac{1}{N} \sum_{i=1}^{N} H_\ell(z_i) \tag{3}$$

denote mean predictive entropy, and let $\mathcal{P}_\ell$ denote aggregate predictive performance at level ℓ, such as accuracy or AUROC. The expected aggregate behavior is $\mathcal{P}_{\ell+1} \geq \mathcal{P}_\ell$ and $\overline{H}_{\ell+1} \leq \overline{H}_\ell$ as richer information becomes available.

Aggregate monotonicity, however, does not imply monotonic uncertainty for an individual sample:

$$\overline{H}_0 \geq \overline{H}_1 \geq \cdots \geq \overline{H}_L \quad \not\Rightarrow \quad H_0(z_i) \geq H_1(z_i) \geq \cdots \geq H_L(z_i). \tag{4}$$

Mean uncertainty can decrease because reductions for some samples outweigh increases for others.

We quantify this discrepancy using the entropy-violation rate

$$\mathrm{VR}_H = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left\{\exists\, \ell < L : H_{\ell+1}(z_i) > H_\ell(z_i)\right\}, \tag{5}$$

the fraction of samples whose uncertainty increases at least once as richer information becomes available. We call the coexistence of favorable aggregate trends and frequent sample-level violations the *Aggregate Monotonicity Illusion*. Section 4 shows that this discrepancy can be substantial across natural and medical imaging settings.

Because VR_H records whether a trajectory contains any violation, its magnitude depends partly on hierarchy depth: longer hierarchies provide more opportunities for a violation. We therefore use it to compare methods within the same progressive-information setting rather than treating values across hierarchies of different lengths as directly comparable measures of severity.

2.3 PROGRESSIVE ACQUISITION

Progressive information naturally induces progressive acquisition: acquire the least costly observation that provides sufficient confidence for prediction. A low-information-first policy begins at level 0 and acquires progressively richer observations only while the current prediction remains insufficiently confident. Let $C_\ell(z_i)$ denote the routing confidence at level ℓ, obtained from the same predictive distribution as $H_\ell(z_i)$, and let $\boldsymbol{\tau} = (\tau_0, \ldots, \tau_{L-1})$ denote a set of stopping thresholds. In our experiments, we use a single threshold shared across levels, $\tau_\ell = \tau$ for all $\ell < L$. The policy exits at the first level whose confidence meets its threshold, or at level L if none does:

$$r_{\boldsymbol{\tau}}(z_i) = \min\left(\{\ell < L : C_\ell(z_i) \geq \tau_\ell\} \cup \{L\}\right). \tag{6}$$

For a routing policy $\boldsymbol{\tau}$, the average acquisition cost is

$$\mathrm{Cost}(\boldsymbol{\tau}) = \frac{1}{N}\sum_{i=1}^{N} c_{r_{\boldsymbol{\tau}}(z_i)}, \tag{7}$$

while the routed predictions determine task performance $\mathcal{P}(\boldsymbol{\tau})$. Varying the shared threshold traces a cost–performance curve, which we summarize using its Pareto frontier and routing hypervolume.

3 MONOTONIC UNCERTAINTY REGULARIZATION

Section 2.2 formalizes the Aggregate Monotonicity Illusion, which we show in Section 4.2 to be widespread under standard training. Standard supervised objectives optimize predictions independently at each information level but leave uncertainty across levels for the same sample unconstrained.

Let $\mathrm{CE}(p_\ell, y)$ denote cross-entropy for multiclass prediction or binary cross-entropy for multilabel prediction at information level ℓ. The standard task objective is

$$\mathcal{L}_{\mathrm{task}} = \frac{1}{L+1}\sum_{\ell=0}^{L} \mathbb{E}_i\left[\mathrm{CE}\left(p_\ell(\cdot \mid x_i^{(\ell)}), y_i\right)\right]. \tag{8}$$

Because this objective imposes no relationship between adjacent information levels, we introduce two complementary forms of **Monotonic Uncertainty Regularization (MUR)**. **MUR-U** directly regularizes scalar uncertainty ordering, whereas **MUR-AU** transfers predictive structure and uncertainty from richer to poorer information. We summarize the methods and their cross-level coupling in Table 1 in the appendix; baseline and experimental details are provided in Section 4.1.

3.1 MUR-U: UNCERTAINTY ORDERING

For levels ℓ and $\ell+1$, we encourage the richer-information prediction to reduce entropy by a margin $m \geq 0$, such that $H_\ell(z_i) - H_{\ell+1}(z_i) \geq m$. The corresponding uncertainty-ordering penalty is

$$\mathcal{R}_{\mathrm{U}} = \sum_{\ell=0}^{L-2} \mathbb{E}_i\left[\max\left(0,\, m - [H_\ell(z_i) - H_{\ell+1}(z_i)]\right)\right] + \mathbb{E}_i\left[\max\left(0,\, m - [H_{L-1}(z_i) - \mathrm{sg}(H_L(z_i))]\right)\right], \tag{9}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. This treats the richest-information prediction as a fixed reference in the final regularization term, so that term updates only the poorer-information predictor. The complete objective is

$$\mathcal{L}_{\text{MUR-U}} = \mathcal{L}_{\mathrm{task}} + \lambda \mathcal{R}_{\mathrm{U}}, \tag{10}$$

where λ controls the strength of cross-level regularization.

When $m = 0$, MUR-U penalizes entropy inversions, $H_{\ell+1}(z_i) > H_\ell(z_i)$; when $m > 0$, it additionally encourages a minimum entropy reduction between adjacent levels. MUR-U is *shape-blind*: predictive distributions with different class-probability structures receive the same regularization signal if they have equal entropy. It therefore regularizes scalar uncertainty ordering without constraining how probability mass is distributed across classes.

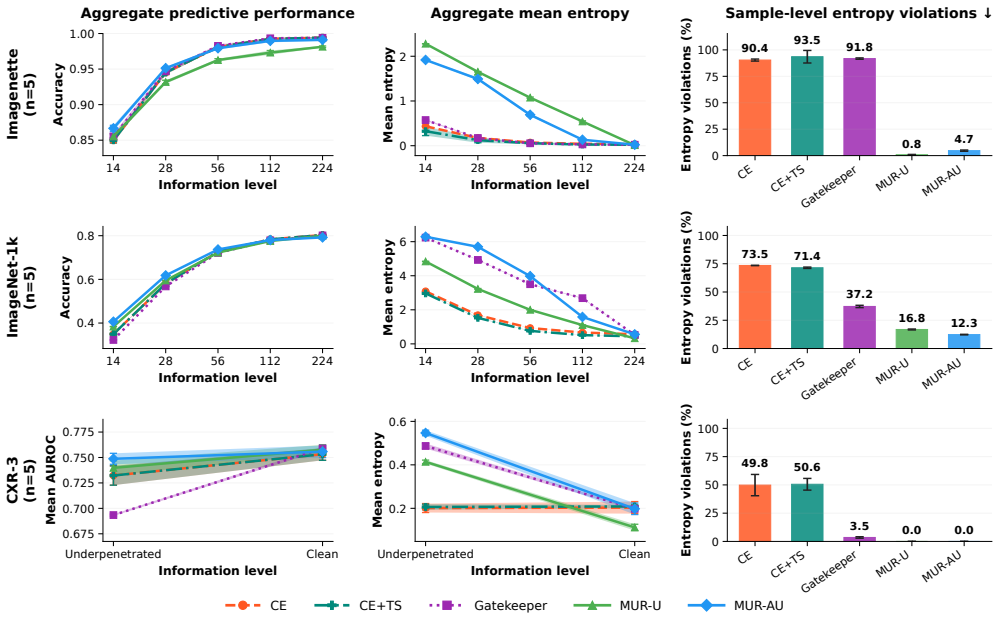

Figure 2: **Cross-level regularization restores sample-level uncertainty coherence.** Rows correspond to Imagenette, ImageNet-1k, and CXR-3; columns show predictive performance, mean entropy, and sample-level entropy-violation rate. CE/BCE and CE+TS exhibit widespread violations despite strong performance at each information level. Gatekeeper leaves violations high on the natural-image datasets, while its low violation rate on CXR-3 comes at the cost of degraded performance on underpenetrated images. In contrast, MUR-U and MUR-AU sharply reduce violations while preserving predictive performance. Curves and bars show mean±std over five seeds.

3.2 MUR-AU: Predictive Structure and Uncertainty

Entropy is only a scalar summary and does not capture which classes remain plausible or their relative probabilities. MUR-AU therefore couples full predictive distributions across adjacent levels through knowledge distillation (Hinton et al., 2015). For each pair $(\ell, \ell+1)$, the richer-information predictor serves as a detached teacher for the poorer-information predictor. We construct the softened teacher distribution

$$q_{\ell+1}^{(\alpha)}(k \mid z_i) = \frac{p_{\ell+1}(k \mid x_i^{(\ell+1)})^{\alpha}}{\sum_{j \in \mathcal{Y}} p_{\ell+1}(j \mid x_i^{(\ell+1)})^{\alpha}}, \qquad 0 < \alpha \leq 1, \tag{11}$$

and minimize the adjacent-level distillation loss

$$\mathcal{R}_{\mathrm{AU}} = \sum_{\ell=0}^{L-1} \mathbb{E}_i \left[- \sum_{k \in \mathcal{Y}} \mathrm{sg}\left(q_{\ell+1}^{(\alpha)}(k \mid z_i) \right) \log p_{\ell}(k \mid x_i^{(\ell)}) \right]. \tag{12}$$

For multilabel prediction, we use the analogous per-label binary distillation objective. The complete objective is

$$\mathcal{L}_{\text{MUR-AU}} = \mathcal{L}_{\mathrm{task}} + \mathcal{R}_{\mathrm{AU}}, \tag{13}$$

where $\mathcal{R}_{\mathrm{AU}}$ has unit weight and α controls the softness of the transferred distribution.

As $\alpha \to 0$, the teacher target approaches a uniform distribution and retains little class-specific information; as $\alpha \to 1$, it approaches the original richer-information prediction. Intermediate values preserve relative class preferences while softening confidence, allowing the poorer-information predictor to inherit predictive structure without reproducing the richer predictor's full certainty. Thus, MUR-U directly regularizes scalar uncertainty ordering, whereas MUR-AU encourages cross-level consistency in both *what* the model predicts and *how certain* it is.

We apply either MUR variant through second-stage fine-tuning. We first train the level-specific predictors using $\mathcal{L}_{\mathrm{task}}$ and then jointly fine-tune them with MUR-U or MUR-AU. MUR therefore reshapes the cross-level behavior of already capable predictors without requiring training from scratch. Section 4.1 provides the initialization, optimization, and hyperparameter-selection details.

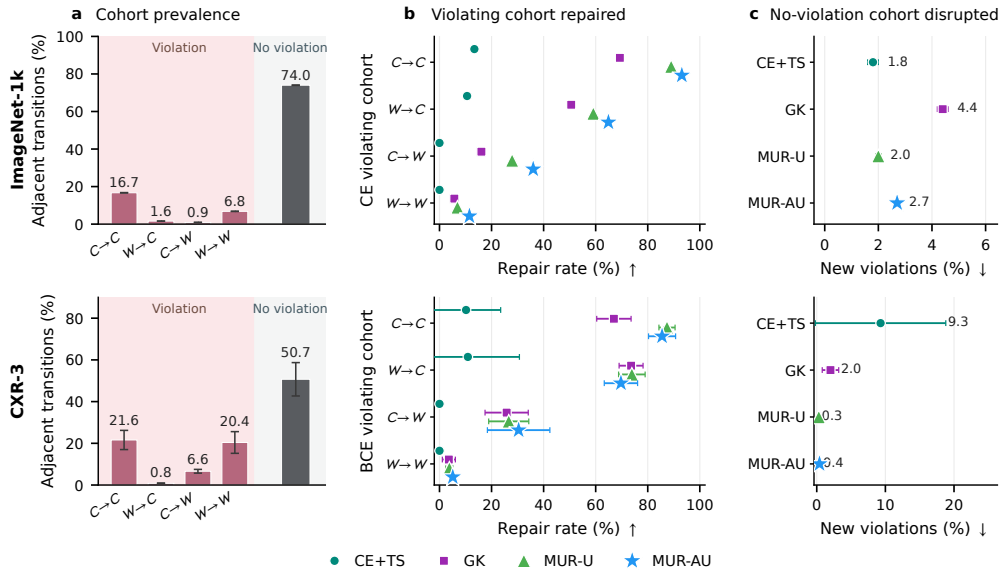

Figure 3: **Which CE baseline failure modes are repaired, and what new violations are introduced?** C and W denote correct and incorrect predictions, respectively, and GK denotes Gatekeeper. **(a)** Adjacent-level CE transitions are partitioned into four entropy-violating cohorts by correctness (pink) and a nonviolating cohort (gray). **(b)** For each baseline violating cohort, the repair rate is the fraction whose entropy increase is eliminated and whose richer-level prediction is correct; for $C \to C$, both levels must remain correct. **(c)** For baseline non-violating transitions, the new-violation rate is the fraction that develop an entropy increase, irrespective of correctness. MUR repairs many violations while preserving $C \to C$ correctness and $W \to C$ corrections, and rarely introduces new violations. Points and error bars show mean±standard deviation over five seeds.

4 Results

We first describe the experimental setup (Section 4.1) and establish the prevalence of the Aggregate Monotonicity Illusion (Section 4.2). We then evaluate whether MUR restores sample-level coherence while preserving predictive quality, examine which violations it repairs and its operating regimes (Section 4.3), and test whether these changes improve progressive acquisition (Section 4.4).

4.1 Experimental Setup

Datasets and information hierarchies. We study progressive resolution on Imagenette and ImageNet-1k using five levels, $14^2 \prec 28^2 \prec 56^2 \prec 112^2 \prec 224^2$. For clinical images, we construct a two-level underpenetrated-to-clean hierarchy on NIH ChestX-ray14. Following radiologists' suggestions for this task, we focus on Mass, Pneumothorax, and Consolidation (CXR-3) in the main paper and extend the analysis to all 14 findings in Appendix C. Radiologists reviewed the degradation on a subset of images for clinical plausibility. Together, the resolution and image-quality hierarchies test cross-level uncertainty across distinct domains, tasks, and information axes. Dataset splits, degradation details, and examples are given in Appendix A.1.

Training and model selection. Each information level uses an independent predictor: an ImageNet-pretrained ResNet-50 for natural images or a TorchXRayVision DenseNet-121 for CXR. We first train the level-specific predictors with CE/BCE, then fine-tune MUR-U, MUR-AU, and Gatekeeper (Rabanser et al., 2025) separately from the same checkpoints, isolating how each method alters cross-level uncertainty. Temperature scaling (Guo et al., 2017) is fitted independently at each level. We select hyperparameters by minimizing the validation entropy-violation rate, subject to richest-level validation performance remaining within five percentage points of CE/BCE. All results report mean and standard deviation over five seeds. Training and selection details are in Appendix A.2; broader parameter sweeps are in Appendix B.4.

Evaluation. We report accuracy for natural images and mean AUROC for CXR. We measure sample-level uncertainty coherence by the entropy-violation rate VR_H (Section 2.2). For progres-

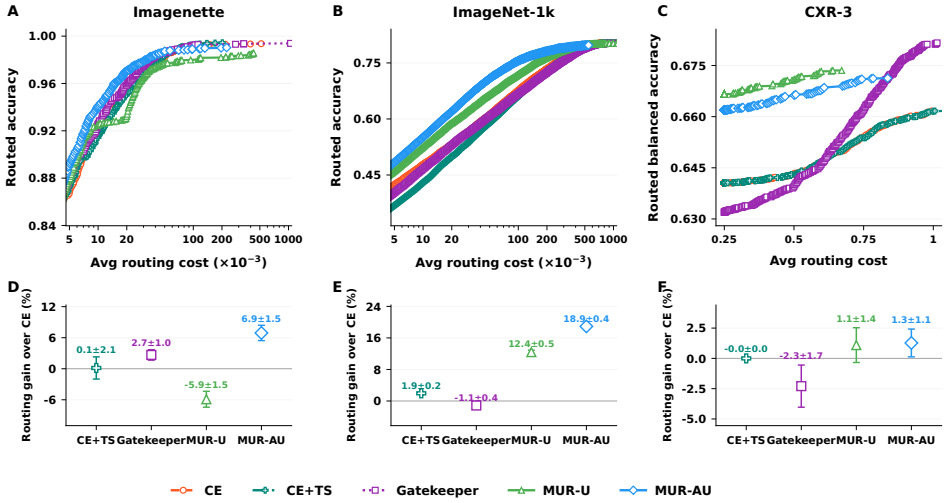

Figure 4: **Cross-level regularization supports more reliable progressive acquisition. (A–C)** Cost–performance frontiers under confidence-threshold routing on Imagenette, ImageNet-1k, and CXR-3. **(D–F)** Routing hypervolume gains over the corresponding CE/BCE baseline. MUR-AU improves routing across all three settings, with its largest gain of 18.9% on ImageNet-1k. MUR-U improves routing on ImageNet-1k and CXR-3 but underperforms CE on Imagenette, whereas Gatekeeper yields inconsistent gains. Error bars denote standard deviation over five seeds.

sive acquisition, we report cost–performance frontiers and routing hypervolume gain over CE/BCE, using routed accuracy for natural images and routed balanced accuracy for CXR. Routing details are in Appendix A.3. In figures, CE denotes BCE for CXR.

4.2 The Aggregate Monotonicity Illusion Is Widespread

Under standard CE/BCE training, aggregate evaluation shows expected behavior across natural and clinical imaging: predictive performance improves and mean entropy decreases as observations progress from lower to higher resolution or from degraded to clean radiographs (Figures 1 and 2).

Individual trajectories reveal widespread violations hidden by these averages. Across the five-level trajectory, 90.4% of Imagenette samples and 73.5% of ImageNet-1k samples exhibit at least one entropy increase, showing that the phenomenon persists from a controlled ten-class benchmark to ImageNet-1k scale. The discrepancy also persists under a shift in domain, architecture, task formulation, and information axis: in the clinically motivated underpenetrated-to-clean CXR setting, 49.8% of CXR-3 studies exhibit an entropy violation under standard BCE training despite a single transition. The same conclusion holds on CXR-14 (Appendix C) and under entropy tolerances (Appendix B.1). Together, these results establish the *Aggregate Monotonicity Illusion*: monotonicity in average uncertainty does not imply monotonicity along individual uncertainty trajectories.

4.3 Cross-Level Regularization Restores Uncertainty Coherence

MUR substantially reduces entropy violations across scale and domain. We next test whether MUR can correct this hidden failure without sacrificing the performance of standard training. Figure 2 jointly compares level-wise predictive performance, mean entropy, and sample-level violation rates. MUR sharply reduces violations while maintaining competitive performance across information levels: MUR-U and MUR-AU reduce the rates from 90.4% to 0.8% and 4.7% on Imagenette, and from 73.5% to 16.8% and 12.3% on ImageNet-1k. Both variants eliminate violations on CXR-3, while MUR-AU retains the strongest low-information performance among the regularized methods. Appendix C shows that these results extend to CXR-14, and Appendix D confirms substantial reductions for individual findings in both CXR settings.

MUR repairs violations while preserving correctness. Under CE/BCE, entropy violations occur even when predictions remain correct at both levels: $C \to C$ accounts for 16.7% of adjacent transitions on ImageNet-1k and 21.6% on CXR-3 (Figure 3a). MUR repairs most of these violations without losing correctness, retains many $W \to C$ corrections, and recovers correct richer-level

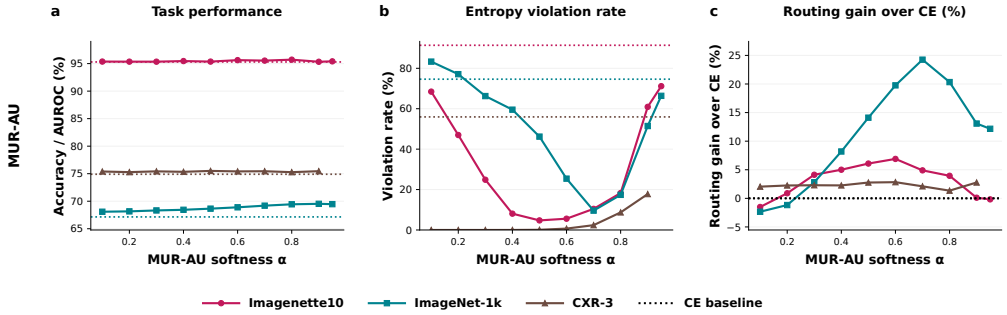

Figure 5: **MUR-AU exhibits broad and interpretable operating regime.** Performance, entropy-violation rate, and routing gain as the softness parameter α varies. Across all three settings, a broad intermediate-softness regime ($\alpha \in [0.5, 0.8]$) substantially reduces violations and yields positive routing gains while preserving baseline performance. Dotted lines denote the CE baselines.

predictions in some $C \to W$ cases (Figure 3b). It rarely introduces violations among previously nonviolating transitions (Figure 3c); the full cohort analysis, including Imagenette and CXR-14, appears in Appendix B.3.

MUR exhibits broad, interpretable operating regimes. As expected from the role of α described in Section 3.2, intermediate MUR-AU softness values $\alpha \in [0.5, 0.8]$ preserve predictive performance, substantially reduce entropy violations relative to CE/BCE, and yield positive routing gains across all three settings (Figure 5). MUR-U likewise preserves performance and reduces violations across a broad range of positive margins and penalty weights. These sweeps identify favorable hyperparameter ranges rather than isolated settings, whereas Gatekeeper shows no regime that jointly achieves low violations, preserved performance, and improved routing (Appendix B.4).

Within-level confidence control is insufficient. Temperature scaling rescales confidence post hoc at each level and leaves violation rates close to standard training (Figure 2), despite low calibration error at the lowest information level (Appendix B.2). Gatekeeper is a more informative control: it shapes confidence during training but never compares predictions for the same sample across levels. Consequently, it fails to reduce violations on Imagenette and reduces them only to 37.2% on ImageNet-1k, more than twice MUR's rate. On CXR-3, Gatekeeper reduces violations to 3.5% but substantially weakens performance on underpenetrated images. In this binary multi-label setting, its correctness-conditioned objective can induce entropy ordering by flattening incorrect predictions toward 0.5 and sharpening correct ones, without coupling the two levels. Thus, low violation rates must be evaluated alongside predictive quality: within-level control does not reliably restore cross-level coherence, whereas MUR directly couples predictions across levels.

4.4 MUR Makes Uncertainty More Reliable for Progressive Acquisition

MUR-AU reaches the same performance with less acquisition. We next test whether MUR's improved cross-level coherence makes uncertainty more useful for deciding when to acquire richer information. Figure 4 shows the cost–performance frontiers and routing hypervolume gains over CE/BCE under the low-information-first policy of Section 2.3. MUR-AU achieves positive routing gains across all three primary settings: 6.9% on Imagenette, 18.9% on ImageNet-1k, and 1.3% on CXR-3. MUR-U improves ImageNet-1k and CXR-3 by 12.4% and 1.1%, respectively, but underperforms CE on Imagenette by 5.9% despite achieving the lowest entropy-violation rate. Temperature scaling yields gains of at most 1.9%, while Gatekeeper improves Imagenette but degrades routing on ImageNet-1k and CXR-3, showing that within-level confidence control does not reliably support progressive acquisition. The broader CXR-14 results reinforce this conclusion: MUR-U and MUR-AU improve routing hypervolume by 1.0% and 1.7%, whereas Gatekeeper reduces it by 8.1% (Appendix C). Together, these results reveal that uncertainty coherence and routing utility are related but not equivalent. MUR-U directly enforces entropy ordering but does not explicitly preserve class-specific predictive structure, explaining why its lowest-violation solution on Imagenette does not yield better routing. MUR-AU couples both properties, preserving the structure needed to distinguish which samples may benefit from richer observations and consistently improving the trade-off between predictive performance and acquisition cost.

5 Related Work

Adaptive computation on a fixed input. Early-exit, anytime, and resolution-adaptive networks allocate computation according to input difficulty, spending more only on hard inputs (Teerapittayanon et al., 2016; Huang et al., 2017; Kaya & Dumitras, 2018; Wołczyk et al., 2021; Yang et al., 2020; Zhu et al., 2021; Wang et al., 2021; 2020; Phuong & Lampert, 2019). Individual predictions can reverse across stages (Kaya & Dumitras, 2018; Wang et al., 2020), and recent work restores monotonicity, nestedness, or calibration across exits post hoc (Jazbec et al., 2023a;b; Meronen et al., 2023). In all of these, the full input is available from the start: later stages extract more from it, and the cost of continuing is computation rather than acquisition.

Learned acquisition policies. When evidence must be acquired at a cost, a separate line of work learns acquisition policies that select which measurement to take next or when to stop, through multi-stage reject classifiers (Trapeznikov et al., 2012; Trapeznikov & Saligrama, 2013), reinforcement learning (Shim et al., 2018; Janisch et al., 2019; Yu et al., 2023), or information-theoretic criteria (Ma et al., 2018; Covert et al., 2023), including in clinical settings (Yu et al., 2023; Xu et al., 2022; Zhang et al., 2019). Many of these policies use the predictor's uncertainty to decide what input to acquire or when to stop. In our work, the order in which inputs are acquired is fixed, and we instead aim to make that uncertainty a more reliable signal for deferral, so MUR can help such policies defer reliably rather than compete with them.

Confidence for acceptance and deferral. The stopping signal itself is studied in selective prediction, learning to defer, and calibration (Chow, 1970; Geifman & El-Yaniv, 2017; 2019; Madras et al., 2017; Mozannar & Sontag, 2020; Hendrycks & Gimpel, 2016; Guo et al., 2017; Cattelan & Silva, 2023), and in model cascades, where calibrated confidence need not yield good deferral (Enomoto & Eda, 2021; Jitkrittum et al., 2023; Rabanser et al., 2025). These methods evaluate or improve confidence at a single information level; none asks whether it stays coherent along the information trajectory of the same sample, which per-level calibration does not ensure (Appendix B.2).

Taken together, prior work varies computation on a fixed input, learns acquisition policies that consume uncertainty without examining it, or shapes confidence within a single stage. To our knowledge, no prior work tests whether uncertainty remains coherent for the same sample as newly acquired observations become richer, trains predictors for that coherence, or measures its effect on acquisition cost. We address this gap by revealing that such coherence often fails, introducing MUR to restore it, and characterizing when restored coherence reduces acquisition cost.

6 Conclusion and Limitations

Conclusion. We identify the Aggregate Monotonicity Illusion: favorable population-level uncertainty trends can conceal widespread inconsistency along the individual trajectories that govern acquisition decisions. By explicitly learning cross-level structure, MUR substantially reduces these inconsistencies while preserving predictive quality across natural and medical imaging. However, coherence alone does not guarantee better routing: MUR-U can enforce uncertainty ordering without always improving acquisition. MUR-AU, which also transfers class-specific predictive structure, improves the cost–performance trade-off across all settings, suggesting that effective acquisition benefits from both. More broadly, uncertainty may be evaluated in aggregate, but it is acted upon per sample. Thus, when uncertainty governs whether to acquire additional information, it should be evaluated, and when necessary learned, along the trajectory on which that decision is made.

Limitations and Future Work. We note three limitations of this work, each of which points to a direction for future research. First, we treat monotonicity as a useful coherence property rather than a universal requirement, since richer evidence may legitimately reveal ambiguity for some samples. Distinguishing such informative increases from incoherent fluctuations that weaken uncertainty-based decisions remains an open problem. Second, our hierarchies are controlled by design. This isolates cross-level behavior cleanly, but it does not capture real acquisition pathways. Evaluating MUR on paired or longitudinal pathways such as CXR→CT, where successive observations differ in modality, diagnostic content, and acquisition consequences, is an important next step. Third, our routing analysis assigns each level a simplified scalar cost, whereas real decisions weigh expense, delay, radiation exposure, procedural risk, and clinical benefit. Developing application-specific utilities with domain experts and evaluating them prospectively would bring uncertainty-guided acquisition closer to deployment in progressive systems.

AI USE STATEMENT

In this work, we did not use generative AI tools for any tasks requiring disclosure. We have not used generative AI tools to generate synthetic datasets; develop theoretical models or conceptual frameworks; formulate mathematical claims; propose or refine hypotheses; design or provide feedback on the research methodology or experiments; implement methods; clean or reformat datasets; or interpret results. Providing ingredients for mathematical proofs, writing proofs, translation, and qualitative or thematic data analysis were not applicable to this work. Additionally, we used generative AI tools to clean existing software code, including plotting code used to prepare scientific figures; improve the presentation of those figures; edit and paraphrase author-provided manuscript text for readability; identify relevant literature; and suggest candidate titles. We reviewed all AI-assisted work: we ran the edited code and checked the figures against the experimental results, reviewed the paraphrased text for accuracy of meaning, claims, and numerical values, and verified retained references against the original papers. We take responsibility for the final content of this work, including text, claims, code, and figures produced with the aid of generative AI.

ETHICS STATEMENT

We use publicly available datasets and collect no new patient data. Radiologists advised on the CXR setting and reviewed examples of the simulated underpenetration. Our experiments do not evaluate a real clinical acquisition pathway, and the resulting models should not be used for patient care without further clinical evaluation.

REPRODUCIBILITY STATEMENT

Sections 2 and 3 define the setting, evaluation metrics, and MUR objectives. Section 4.1 describes the experimental protocol. We use publicly available Imagenette, ImageNet-1k, and NIH ChestX-ray14 datasets. Appendix A provides the dataset splits, image transformations, training settings, hyperparameters, and routing details. Appendices B–D contain additional analyses and results. To assess stability across runs, we report quantitative results over five seeds. Experiments used NVIDIA H100 80GB GPUs: Imagenette and CXR training used one GPU per process, while ImageNet-1k used eight GPUs with distributed training. Code to reproduce the experiments is provided in the supplementary materials.

REFERENCES

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019. URL https://api.semanticscholar.org/CorpusID:196194314.

Luís Felipe P. Cattelan and Danilo Silva. How to fix a broken confidence estimator: Evaluating post-hoc methods for selective classification with deep neural networks. In *Conference on Uncertainty in Artificial Intelligence*, 2023. URL https://api.semanticscholar.org/CorpusID:258888148.

C.-K. Chow. On optimum recognition error and reject tradeoff. *IEEE Trans. Inf. Theory*, 16:41–46, 1970. URL https://api.semanticscholar.org/CorpusID:206730137.

Jared D Christensen, Michael R. Harowicz, Christopher M Walker, Brent P. Little, Kiran Batra, Anupama Gupta Brixey, Melissa B Carroll, Lydia Chelala, Christian W. Cox, M. Bradley Drummond, Nicole Geissen, Jay Halpern, Rachna Madan, Venu Pararath Gopalakrishnan, Girish S. Shroff, Christina S. Thornton, Jennifer Zreloff, and Jonathan H. Chung. Acr appropriateness criteria® chronic dyspnea-noncardiovascular origin: 2024 update. *Journal of the American College of Radiology : JACR*, 22 5S:S163–S176, 2025. URL https://api.semanticscholar.org/CorpusID:278833817.

Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P. Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and

Hadrien Bertrand. Torchxrayvision: A library of chest x-ray datasets and models. In *International Conference on Medical Imaging with Deep Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:240353861.

Ian Covert, Wei Qiu, Mingyu Lu, Nayoon Kim, Nathan White, and Su-In Lee. Learning to maximize mutual information for dynamic feature selection. In *International Conference on Machine Learning*, 2023. URL https://api.semanticscholar.org/CorpusID:255372817.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. URL https://api.semanticscholar.org/CorpusID:57246310.

Shohei Enomoto and Takeharu Eda. Learning to cascade: Confidence calibration for improving the accuracy and computational cost of cascade inference systems. *ArXiv*, abs/2104.09286, 2021. URL https://api.semanticscholar.org/CorpusID:233296030.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *ArXiv*, abs/1705.08500, 2017. URL https://api.semanticscholar.org/CorpusID:491127.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning*, 2019. URL https://api.semanticscholar.org/CorpusID:59316904.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *ArXiv*, abs/1706.04599, 2017. URL https://api.semanticscholar.org/CorpusID:28671436.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015. URL https://api.semanticscholar.org/CorpusID:206594692.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *ArXiv*, abs/1610.02136, 2016. URL https://api.semanticscholar.org/CorpusID:13046179.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015. URL https://api.semanticscholar.org/CorpusID:7200347.

Jeremy Howard. Imagenette: A smaller subset of 10 easily classified classes from imagenet, March 2019. URL https://github.com/fastai/imagenette.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2016. URL https://api.semanticscholar.org/CorpusID:9433631.

Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens van der Maaten, and Kilian Q. Weinberger. Multi-scale dense networks for resource efficient image classification. In *International Conference on Learning Representations*, 2017. URL https://api.semanticscholar.org/CorpusID:3475998.

Jaromír Janisch, Tomáš Pevnỳ, and Viliam Lisỳ. Classification with costly features using deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3959–3966, 2019.

Metod Jazbec, James Urquhart Allingham, Dan Zhang, and Eric T. Nalisnick. Towards anytime classification in early-exit architectures by enforcing conditional monotonicity. *ArXiv*, abs/2306.02652, 2023a. URL https://api.semanticscholar.org/CorpusID:259075688.

Metod Jazbec, Patrick Forré, Stephan Mandt, Dan Zhang, and Eric T. Nalisnick. Early-exit neural networks with nested prediction sets. In *Conference on Uncertainty in Artificial Intelligence*, 2023b. URL https://api.semanticscholar.org/CorpusID:265128872.

Wittawat Jitkrittum, Neha Gupta, Aditya Krishna Menon, Harikrishna Narasimhan, Ankit Singh Rawat, and Surinder Kumar. When does confidence-based cascade deferral suffice? *ArXiv*, abs/2307.02764, 2023. URL https://api.semanticscholar.org/CorpusID:259360997.

Yiğitcan Kaya and Tudor Dumitras. How to stop off-the-shelf deep neural networks from overthinking. *ArXiv*, abs/1810.07052, 2018. URL https://api.semanticscholar.org/CorpusID:53113950.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. URL https://api.semanticscholar.org/CorpusID:53592270.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, José Miguel Hernández-Lobato, Sebastian Nowozin, and Cheng Zhang. Eddi: Efficient dynamic discovery of high-value information with partial vae. *ArXiv*, abs/1809.11142, 2018. URL https://api.semanticscholar.org/CorpusID:52891994.

David Madras, Toniann Pitassi, and Richard S. Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Neural Information Processing Systems*, 2017. URL https://api.semanticscholar.org/CorpusID:52175668.

Lassi Meronen, M. Trapp, Andrea Pilzer, Le Yang, and Arno Solin. Fixing overconfidence in dynamic neural networks. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2668–2678, 2023. URL https://api.semanticscholar.org/CorpusID:256827640.

Hussein Mozannar and David A. Sontag. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, 2020. URL https://api.semanticscholar.org/CorpusID:219260564.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, 2015:2901 – 2907, 2015. URL https://api.semanticscholar.org/CorpusID:6292807.

Mary Phuong and Christoph H. Lampert. Distillation-based training for multi-exit architectures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Stephan Rabanser, Nathalie Rauschmayr, Achin Kulshrestha, Petra Poklukar, Wittawat Jitkrittum, Sean Augenstein, Congchao Wang, and Federico Tombari. Gatekeeper: Improving model cascades through confidence tuning. *Advances in Neural Information Processing Systems 38*, 2025. URL https://api.semanticscholar.org/CorpusID:276617635.

Hajin Shim, Sung Ju Hwang, and Eunho Yang. Joint active feature acquisition and classification with variable-size set encoding. In *Neural Information Processing Systems*, 2018. URL https://api.semanticscholar.org/CorpusID:54025962.

Rebecca Smith-Bindman, Jafi A. Lipson, Ralph Marcus, Kwang Pyo Kim, Mahadevappa Mahesh, Robert G. Gould, Amy Berrington de González, and Diana L. Miglioretti. Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. *Archives of internal medicine*, 169:2078 – 2086, 2009. URL https://api.semanticscholar.org/CorpusID:6194273.

Surat Teerapittayanon, Bradley McDanel, and H. T. Kung. Branchynet: Fast inference via early exiting from deep neural networks. *2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 2464–2469, 2016. URL https://api.semanticscholar.org/CorpusID:2916466.

Kirill Trapeznikov and Venkatesh Saligrama. Supervised sequential classification under budget constraints. In *International Conference on Artificial Intelligence and Statistics*, 2013. URL https://api.semanticscholar.org/CorpusID:3132246.

Kirill Trapeznikov, Venkatesh Saligrama, and David A. Castañón. Multi-stage classifier design. *Machine Learning*, 92:479 – 502, 2012. URL https://api.semanticscholar.org/CorpusID:39491.

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. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3462–3471, 2017. URL https://api.semanticscholar.org/CorpusID:263796294.

Yikai Wang, Fuchun Sun, Duo Li, and Anbang Yao. Resolution switchable networks for runtime efficient image recognition. In *European Conference on Computer Vision*, 2020. URL https://api.semanticscholar.org/CorpusID:220647099.

Yulin Wang, Rui Huang, Shiji Song, Zeyi Huang, and Gao Huang. Not all images are worth 16x16 words: Dynamic transformers for efficient image recognition. In *Neural Information Processing Systems*, 2021. URL https://api.semanticscholar.org/CorpusID:239885997.

Maciej Wołczyk, Bartosz W'ojcik, Klaudia Balazy, Igor T. Podolak, Jacek Tabor, Marek Śmieja, and Tomasz Trzcinski. Zero time waste: Recycling predictions in early exit neural networks. In *Neural Information Processing Systems*, 2021. URL https://api.semanticscholar.org/CorpusID:235391036.

Yanbo Xu, Alind Khare, Glenn Scott Matlin, Monish Ramadoss, Rishikesan Kamaleswaran, Chao Zhang, and Alexey Tumanov. Unfoldml: Cost-aware and uncertainty-based dynamic 2d prediction for multi-stage classification. *ArXiv*, abs/2210.15056, 2022. URL https://api.semanticscholar.org/CorpusID:253157318.

Le Yang, Yizeng Han, Xi Chen, Shiji Song, Jifeng Dai, and Gao Huang. Resolution adaptive networks for efficient inference. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2366–2375, 2020. URL https://api.semanticscholar.org/CorpusID:212725240.

Cha Dong Yeo. Improving diagnosis in health care. *Military medicine*, 181 3:183–5, 2016. URL https://api.semanticscholar.org/CorpusID:38538286.

Zheng Yu, Yikuan Li, Joseph Kim, Kai Huang, Yuan Luo, and Mengdi Wang. Deep reinforcement learning for cost-effective medical diagnosis. *ArXiv*, abs/2302.10261, 2023. URL https://api.semanticscholar.org/CorpusID:257050266.

Zizhao Zhang, Adriana Romero, Matthew Muckley, Pascal Vincent, Lin Yang, and Michal Drozdzal. Reducing uncertainty in undersampled mri reconstruction with active acquisition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2049–2053, 2019. URL https://api.semanticscholar.org/CorpusID:59842971.

Mingjian Zhu, Kai Han, Enhua Wu, Qiulin Zhang, Ying Nie, Zhenzhong Lan, and Yunhe Wang. Dynamic resolution network. In *Neural Information Processing Systems*, 2021. URL https://api.semanticscholar.org/CorpusID:235358802.

APPENDIX

Organization. In the appendix, we provide the experimental settings (Appendix A), additional analyses on the violation tolerance, the temperature-scaling objective, prediction correctness, and operating regimes (Appendix B), additional results on CXR-14 (Appendix C), and per-pathology results on CXR-3 and CXR-14 (Appendices D.1 and D.2).

A EXPERIMENTAL SETTINGS

A.1 DATASETS AND INFORMATION HIERARCHIES

Natural images. We hold out 10% (Imagenette; Howard, 2019) and 5% (ImageNet-1k; Deng et al., 2009) of the official training split for validation and use the official validation split for final evaluation. Each level downsamples the image to $\{14, 28, 56, 112, 224\}^2$ with antialiased bilinear interpolation and resizes it back to 224^2 with nearest-neighbor interpolation, so levels differ in visual information rather than input size. We use standard random resized cropping and horizontal flipping for training, and center cropping for evaluation.

Chest X-rays. We use the official NIH ChestX-ray14 (Wang et al., 2017) train/test split and hold out a patient-disjoint 20% of training patients for validation. Given a clean image with intensities $u \in [0, 1]$, the underpenetrated image is $d = \mathrm{clip}(1.30\,u + 0.22,\ 0,\ 1)$, followed by an intensity-adaptive blur $d \leftarrow \mathrm{clip}\big(d\,G_\sigma(d) + (1 - d)\,d,\ 0,\ 1\big)$ with Gaussian filter G_σ, $\sigma = 3$, which reduces contrast and blurs brighter regions more strongly. Figure 6 shows examples. The transformation was reviewed with radiologists on a subset of images for clinical plausibility. CXR-3 evaluates Mass, Pneumothorax, and Consolidation; CXR-14 evaluates all 14 findings under the same two-level hierarchy.

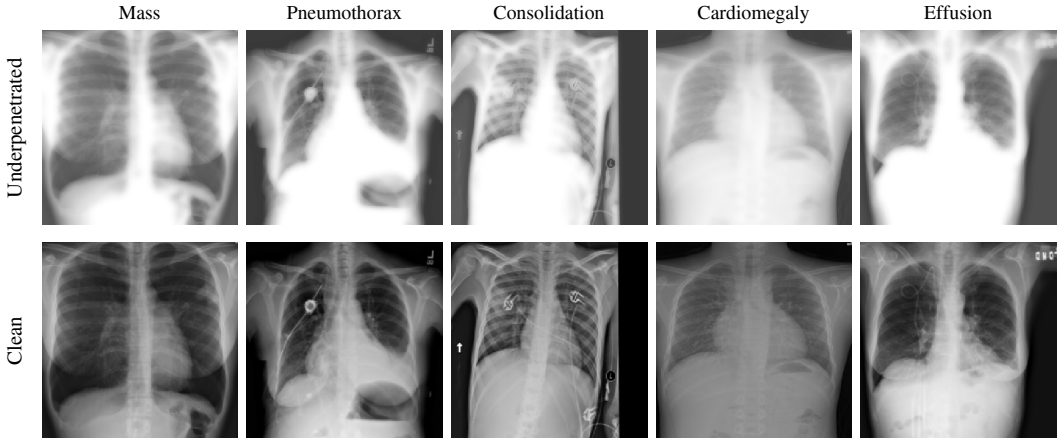

Figure 6: **Examples from the CXR underpenetrated-to-clean hierarchy.** Each column shows the same study at both information levels. Mass, Pneumothorax, and Consolidation form CXR-3; Cardiomegaly and Effusion are additional CXR-14 findings. The degradation reduces contrast and blurs brighter regions more strongly.

A.2 TRAINING AND HYPERPARAMETER SELECTION

We train a separate predictor per level, without parameter sharing, using ImageNet-pretrained ResNet-50 (He et al., 2015) for natural images and TorchXRayVision (Cohen et al., 2021) DenseNet-121 (Huang et al., 2016) for CXR, optimized with AdamW (Loshchilov & Hutter, 2017). We train CE/BCE for 20 (Imagenette), 10 (ImageNet-1k), and 10 (CXR) epochs, then fine-tune MUR-U, MUR-AU, and Gatekeeper from the same checkpoints for 20 (Imagenette), 5 (ImageNet-1k), and 10 (CXR) epochs. Fine-tuning takes roughly half the time of CE/BCE training on ImageNet-1k and CXR. Table 1 summarizes the methods.

We tune all methods with Optuna (TPE sampler) (Akiba et al., 2019). For MUR-U, MUR-AU, and Gatekeeper, we minimize the validation entropy-violation rate subject to validation performance at the richest level remaining within five percentage points of CE/BCE. We search learning rate in $[10^{-5}, 3{\cdot}10^{-4}]$ and weight decay in $[10^{-6}, 10^{-2}]$ (log-uniform), MUR-U $\lambda \in [0.05, 5]$ (log-uniform) and $m \in [0, 1]$, MUR-AU $\alpha \in [0.1, 0.95]$ (natural images) or $[0.2, 0.8]$ (CXR), and Gatekeeper's trade-off weight $\alpha \in [0.1, 0.9]$, which is unrelated to MUR-AU's softness α. We tune hyperparameters once on the validation set and train the selected configuration with five seeds, each retraining CE/BCE and then fine-tuning all methods from that seed's checkpoint. Across seeds, hyperparameters, architectures, and evaluation data are fixed, so the reported standard deviations reflect only training randomness; on ImageNet-1k, the seed also determines the train/validation split. We report mean and standard deviation on the test set. Temperature scaling fits one temperature per level (and per finding for CXR) by minimizing validation NLL, without updating model weights.

Method	**Coupling**	**Target**
CE/BCE	No	Task prediction
CE/BCE+TS	No	Within-level calibration
Gatekeeper	No	Within-level confidence
MUR-U	Yes	Scalar uncertainty
MUR-AU	Yes	Predictive structure and uncertainty

Table 1: **Comparison of methods.** Only MUR explicitly couples adjacent levels.

A.3 Routing

Routing confidence is the maximum softmax probability for natural images and the mean over findings of $\max(p, 1-p)$ for CXR. We use a single threshold shared across levels and sweep it over the observed confidence range. Acquisition costs are geometric: acquiring level m costs $\Delta_m = 4^{-(L-m)}$, and a sample exiting at level ℓ pays for all acquired levels, $c_\ell = \sum_{m \le \ell} \Delta_m$. For natural images, Δ_m is proportional to the pixel count of each resolution; for CXR ($L = 1$), this gives $c_0 = 0.25$ and $c_1 = 1.25$. Routed performance is accuracy for natural images and mean per-pathology balanced accuracy for CXR. We summarize each cost–performance frontier by its hypervolume, computed over log cost, and report the gain relative to CE/BCE.

B Additional Analyses

B.1 Sensitivity to the Violation Tolerance

The violation rate in the main analysis counts any increase in entropy between adjacent information levels. To assess whether the result is driven by small increases, we repeat the analysis with a tolerance ε. We normalize entropy by its maximum value, using $\tilde{H} = H / \log K$ for multiclass prediction and $\tilde{H} = H / \log 2$ for multilabel prediction, where H is the mean binary entropy across labels in the latter case. A sample is counted as violating only if

$$\max_{\ell < L} \big[\tilde{H}_{\ell+1}(z_i) - \tilde{H}_\ell(z_i)\big] > \varepsilon.$$

At $\varepsilon = 0$, this recovers the strict violation rate reported in the main paper.

Figure 7 shows that CE/BCE violation rates decline with increasing tolerance, yet remain above those of MUR-U and MUR-AU at every evaluated tolerance on all four settings. Thus, the difference between CE/BCE and MUR persists even when small entropy increases are ignored.

B.2 Calibration Does Not Imply Cross-Level Coherence

Does calibration resolve violations? Temperature scaling is the standard post-hoc remedy for miscalibration. If calibrating each level were enough to make uncertainty coherent across levels, a dedicated cross-level objective would be unnecessary. Figure 8 shows that it is not: CE/BCE and CE+TS achieve low expected calibration error (ECE) at the lowest information level (at most 0.06)

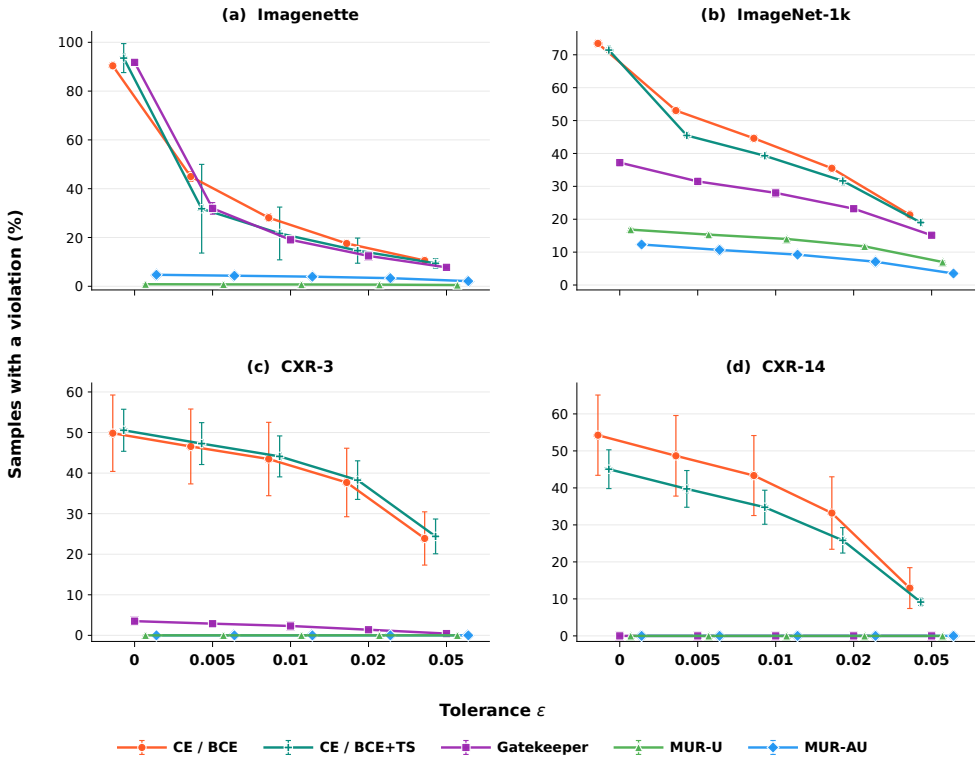

Figure 7: **Sensitivity of sample-level entropy violations to the tolerance ε.** A sample violates when normalized predictive entropy increases by more than ε at any adjacent information level. CE/BCE retains violations across all tolerances, whereas MUR-U and MUR-AU remain lower throughout, so the illusion is not an artifact of negligible entropy fluctuations. Error bars indicate standard deviation over five seeds.

in all four settings, yet 45%–94% of samples still violate entropy monotonicity. Calibration adjusts confidence within each level, whereas progressive acquisition depends on how uncertainty changes across levels for the same sample; MUR targets this cross-level behavior directly.

Sensitivity to the temperature-scaling objective. Standard temperature scaling selects a positive scalar temperature for each information level by minimizing negative log-likelihood (NLL) on the validation set. To test whether our conclusions depend on this choice, we additionally select each temperature by minimizing validation expected calibration error (ECE) (Naeini et al., 2015). Both variants use the same trained predictors, data splits, and evaluation protocol; no model parameters are updated. We report the mean and standard deviation across five seeds. Table 2 compares the resulting sample-level entropy violations and routing hypervolume gains. Accuracy and AUROC are omitted because positive temperature scaling preserves predicted classes and score rankings. ECE-based tuning produces results similar to standard NLL-based tuning across all four settings. It modestly reduces the violation rate on ImageNet-1k, from 71.44% to 66.00%, and increases the routing gain from 1.95% to 3.50%, but the remaining violation rate is still high. Results on Imagenette, CXR-3, and CXR-14 are nearly unchanged. Thus, our conclusion is insensitive to the temperature-selection objective: calibrating predictions independently within each information level does not ensure coherent uncertainty across levels.

B.3 Correctness-Conditioned Violation Analysis

What do the violations represent? Figure 3a partitions adjacent-level CE/BCE transitions by whether the prediction at each level is correct (C) or wrong (W). Violating $C \to C$ transitions

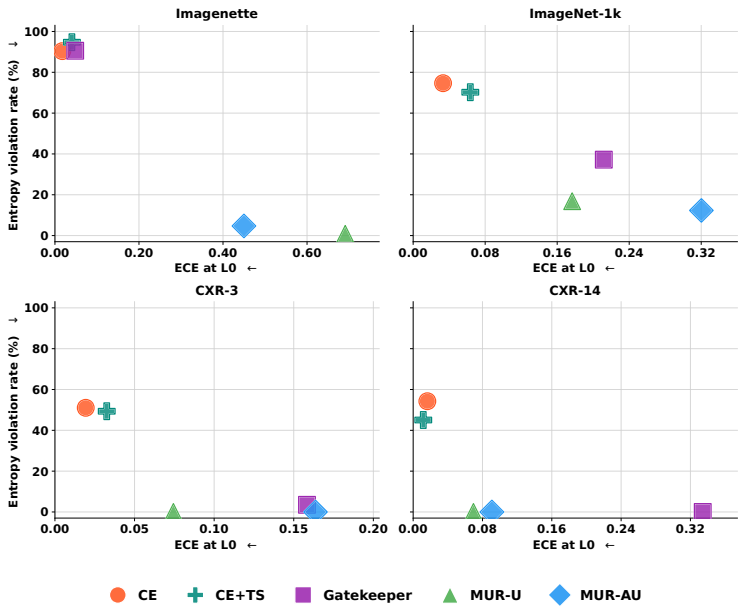

Figure 8: **Calibration does not imply cross-level uncertainty coherence.** ECE at the lowest information level (L0) is plotted against the sample-level entropy-violation rate on Imagenette, ImageNet-1k, CXR-3, and CXR-14; lower is better on both axes. CE/BCE and CE+TS are well calibrated at L0 yet show high violation rates, whereas MUR reduces violations at the cost of higher L0 ECE. Thus, ECE at a single level does not capture whether uncertainty decreases coherently as information increases.

Table 2: **Sensitivity of temperature scaling to the tuning objective.** Temperatures are selected independently at each information level using validation NLL or validation ECE. Results report mean ± standard deviation across five seeds. Both objectives yield similar sample-level entropy violations and routing performance.

Dataset	Method	Entropy violations (%) ↓	Routing HV gain (%) ↑
Imagenette	CE+TS (NLL)	93.53 ± 5.94	0.15 ± 2.15
Imagenette	CE+TS (ECE)	93.40 ± 6.59	1.16 ± 1.98
ImageNet-1k	CE+TS (NLL)	71.44 ± 0.59	1.95 ± 0.23
ImageNet-1k	CE+TS (ECE)	66.00 ± 4.37	3.50 ± 0.57
CXR-3	BCE+TS (NLL)	50.55 ± 5.18	-0.001 ± 0.013
CXR-3	BCE+TS (ECE)	50.56 ± 6.28	0.000 ± 0.012
CXR-14	BCE+TS (NLL)	45.06 ± 5.24	0.008 ± 0.006
CXR-14	BCE+TS (ECE)	45.44 ± 3.90	0.008 ± 0.008

Note. Routing hypervolume gains are measured relative to CE for Imagenette and ImageNet-1k and relative to BCE for CXR-3 and CXR-14. Accuracy and AUROC are omitted because positive temperature scaling preserves predicted classes and score rankings.

account for 16.7% of all transitions on ImageNet-1k and 21.6% on CXR-3: entropy can increase even when both predictions are correct. CXR-3 also has a substantial violating $W \to W$ cohort (20.4%). Figure 9a shows that these patterns extend to Imagenette, where $C \to C$ accounts for 35.5% of transitions, and CXR-14, where $C \to C$ and $W \to W$ account for 23.8% and 19.9%, respectively. Thus, violations arise across different correctness behaviors; they cannot be characterized by level-wise accuracy alone.

Which violations does MUR repair? A baseline violation is counted as repaired when its entropy increase disappears and the richer-level prediction is correct; for $C \to C$, both predictions must remain correct. Figure 3b shows that, on ImageNet-1k and CXR-3, MUR repairs many $C \to C$

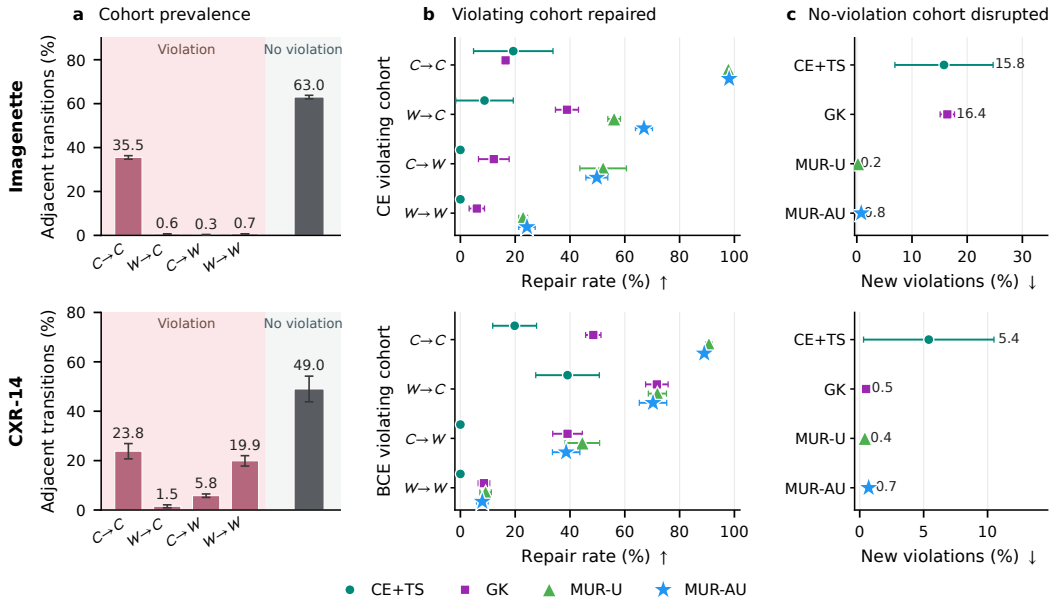

Figure 9: **Correctness-conditioned violation analysis on Imagenette and CXR-14.** Panels follow Figure 3: **(a)** prevalence of violating cohorts (pink) and the nonviolating cohort (gray) under CE/BCE; **(b)** repair rate of each violating cohort; **(c)** new-violation rate among baseline nonviolating transitions. MUR repairs 85.5%–98.1% of $C \to C$ violations without sacrificing correctness while introducing new violations in fewer than 1% of coherent transitions. Error bars show standard deviation over five seeds.

violations without sacrificing correctness, retains many $W \to C$ corrections, and corrects the richer-level prediction in some $C \to W$ cases. Figure 9b shows the same behaviors on Imagenette and CXR-14. Across the four settings, MUR repairs 85.5%–98.1% of violating $C \to C$ transitions, retains and renders coherent 56.1%–73.9% of violating $W \to C$ transitions, and repairs 26.6%–52.1% of violating $C \to W$ transitions by correcting the richer-level prediction. Repair is less frequent for $W \to W$ transitions, where the richer-level prediction must also change from wrong to correct.

Does repair disrupt previously coherent transitions? Figure 3c shows that MUR introduces violations in only 0.3%–2.7% of previously nonviolating transitions on ImageNet-1k and CXR-3, compared with 2.0%–4.4% for Gatekeeper. Figure 9c extends this finding to Imagenette, where MUR's rates are 0.2%–0.8% versus 16.4% for Gatekeeper. On CXR-14, all three methods have low new-violation rates: 0.4%–0.7% for MUR and 0.5% for Gatekeeper. The CXR-14 distinction therefore comes from the repairs and preserved predictive quality, rather than a uniformly lower rate of newly introduced violations.

B.4 Additional Operating-Regime Analyses

The main paper shows that intermediate MUR-AU softness preserves predictive performance, reduces violations, and improves routing across Imagenette, ImageNet-1k, and CXR-3 (Figure 5). Here, we examine the corresponding MUR-U and Gatekeeper sweeps (Figures 10 and 11). Appendix C extends both MUR sweeps to CXR-14.

Positive margins drive MUR-U's coherence gains. In MUR-U, the margin m sets the required entropy decrease between adjacent information levels: $m = 0$ penalizes only entropy increases, whereas $m > 0$ enforces an explicit separation. Positive margins sharply reduce violations over a broad range of penalty weights λ while preserving predictive performance (Figure 10). The settings with the fewest violations do not necessarily provide the greatest routing gain; on Imagenette, routing can remain below the CE baseline despite substantial reductions in violations. Thus, MUR-U offers a broad regime for improving coherence, but coherence and routing utility must be evaluated separately.

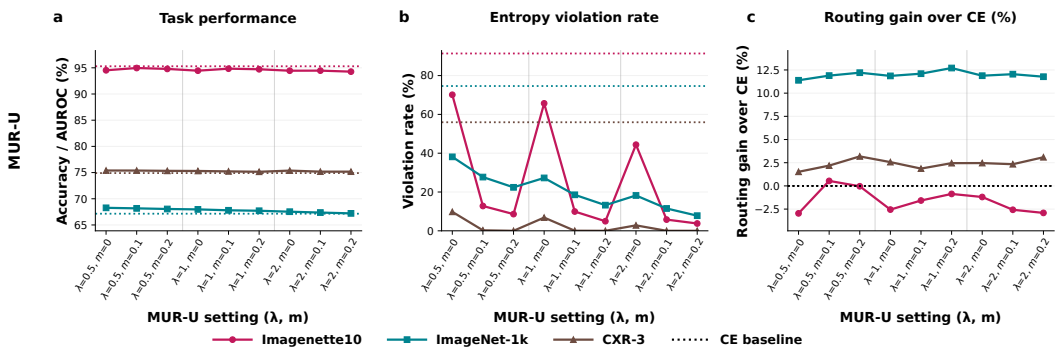

Figure 10: **MUR-U operating regimes.** Predictive performance, entropy-violation rate, and routing gain across margin m and penalty weight λ on Imagenette, ImageNet-1k, and CXR-3. Positive margins are the main driver of coherence, but the lowest-violation settings do not always give the largest routing gains, so coherence and routing utility must be evaluated separately. Dotted lines denote the corresponding CE/BCE baselines.

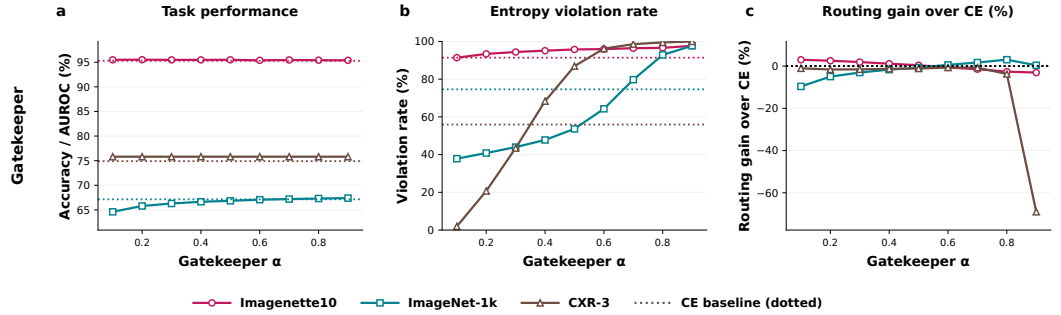

Figure 11: **Gatekeeper operating regimes.** Predictive performance, entropy-violation rate, and routing gain across Gatekeeper settings on Imagenette, ImageNet-1k, and CXR-3. Here, α denotes Gatekeeper's trade-off weight, not MUR-AU softness. No Gatekeeper setting jointly achieves low violations, preserved performance, and improved routing, indicating that shaping confidence within each level does not substitute for coupling levels. Dotted lines denote the corresponding CE/BCE baselines.

Gatekeeper does not exhibit a comparable regime. Gatekeeper's sweep shows no parameter range that jointly provides low violation rates, preserved predictive performance, and improved routing across the three main settings (Figure 11). Violations remain high on Imagenette; on ImageNet-1k, settings that reduce violations have weaker routing. On CXR-3, low violations can coexist with the weaker underpenetrated-image performance shown in Figure 2. These results support the distinction between coupling predictions across information levels and shaping confidence within each level.

C Results on CXR-14

The main paper focuses on CXR-3: Mass, Pneumothorax, and Consolidation, which were selected through radiologist consultation as findings for which underpenetration is clinically meaningful. Here, we apply the same underpenetrated-to-clean hierarchy and evaluation protocol to all 14 ChestX-ray14 findings to test whether our conclusions depend on this selection.

MUR extends to the full CXR-14 label set. Under standard BCE training, 54.2% of studies exhibit an increase in label-averaged predictive entropy from the underpenetrated to the clean image (Figure 12). MUR-U and MUR-AU eliminate these violations while preserving competitive mean AUROC at both information levels. Gatekeeper also eliminates these violations, but substantially

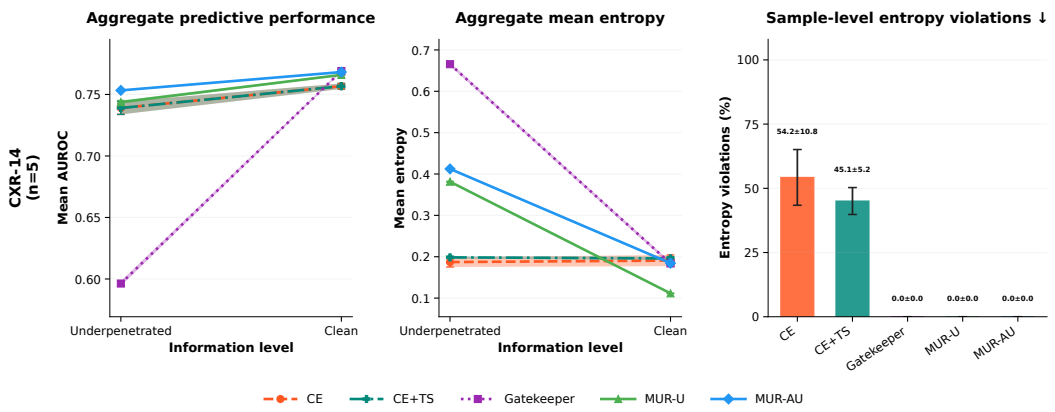

Figure 12: **Uncertainty coherence and predictive performance on CXR-14.** Extends Figure 2 to all 14 findings: mean AUROC, mean entropy, and sample-level entropy-violation rate across the underpenetrated-to-clean hierarchy. MUR-U and MUR-AU eliminate violations while preserving mean AUROC; Gatekeeper also eliminates violations but degrades underpenetrated-image performance. Error bars show standard deviation over five seeds.

reduces performance on underpenetrated images. Thus, low violation rates alone do not establish a useful uncertainty ordering; predictive quality at the lower-information level must also be preserved.

Improved coherence supports progressive acquisition. The distinction is reflected in routing performance (Figure 13). MUR-U and MUR-AU improve routing hypervolume over BCE by 1.0% and 1.7%, respectively, whereas Gatekeeper reduces it by 8.1% despite eliminating violations. MUR therefore improves coherence while retaining the low-information predictions required for effective routing. This result reinforces the main-paper finding that uncertainty coherence and routing utility are related but not equivalent.

The favorable operating regimes persist on CXR-14. Across the MUR-AU softness sweep, mean AUROC remains stable, while intermediate values of α maintain low violation rates and positive routing gains (Figure 14, top). Violations increase at the largest tested α, consistent with the weaker uncertainty separation observed toward that end of the main-paper sweep (Figure 5). The MUR-U sweep similarly contains a broad range of positive-margin settings with low violations, stable predictive performance, and positive routing gains (Figure 14, bottom). These results extend the operating-regime behavior observed on the three main settings rather than relying on a single selected configuration.

D Per-Pathology Results

The CXR results summarize uncertainty using predictive entropy averaged across findings. Monotonicity of this average, however, does not imply monotonicity for each finding. We therefore repeat the analysis separately for each finding on CXR-3 and CXR-14.

D.1 CXR-3

MUR improves coherence for each pathology. Under BCE, per-pathology violation rates range from 43.2% to 54.5% (Figure 15). MUR-U reduces these rates to 0.1%–3.3%, and MUR-AU to 0.2%–3.1%, while preserving competitive AUROC at both information levels. Thus, the zero label-averaged violation rates in the main paper reflect substantial reductions for each pathology, although small pathology-specific violation rates remain. Gatekeeper also reduces violations, but does so with lower AUROC on underpenetrated images, showing again that violation rates must be considered alongside predictive quality.

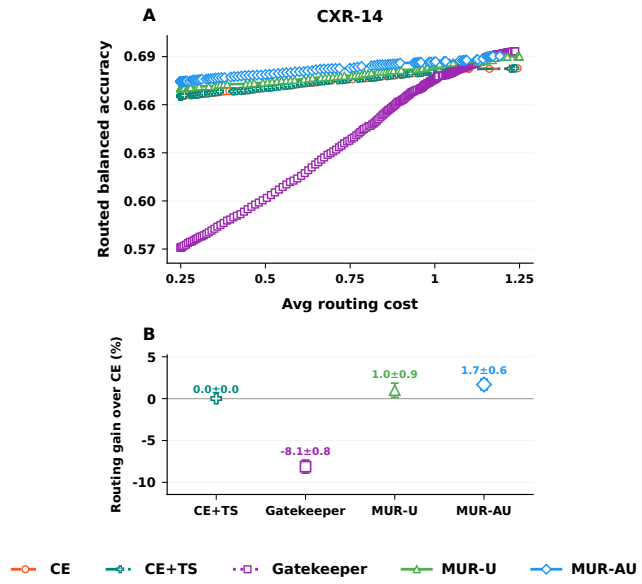

Figure 13: **Progressive acquisition on CXR-14.** Extends Figure 4. **(A)** Cost–performance Pareto frontiers under confidence-threshold routing. **(B)** Routing hypervolume gain relative to BCE. Despite identical zero violation rates, MUR improves routing by up to 1.7% while Gatekeeper degrades it by 8.1%.

The improvements extend to progressive acquisition. Both MUR variants provide stronger low-cost routed balanced accuracy across Mass, Pneumothorax, and Consolidation (Figure 16). Correspondingly, both achieve positive mean routing hypervolume gains over BCE for all three pathologies, whereas Gatekeeper has negative mean gains throughout (Figure 17). The aggregate routing improvement reported in the main paper is therefore supported across the individual findings in the radiologist-informed CXR-3 setting.

D.2 CXR-14

MUR improves coherence for individual findings. Figure 18 shows that MUR-U and MUR-AU substantially reduce violation rates across individual findings while preserving competitive AUROC. Violations remain for some findings, most notably Effusion and Infiltration. Gatekeeper also reduces violations, yet degrades underpenetrated-image performance for several findings.

MUR improves routing across individual findings. The per-pathology routing frontiers show that MUR generally provides stronger low-cost performance across the 14 findings, whereas Gatekeeper often remains below BCE until routing approaches the clean observation (Figure 19). MUR-AU and MUR-U improve routing for most findings, whereas Gatekeeper degrades routing for nearly all findings (Figure 20). The full-label analysis therefore supports the main conclusion beyond the three radiologist-selected findings, with MUR-AU providing the most consistent per-pathology routing improvements.

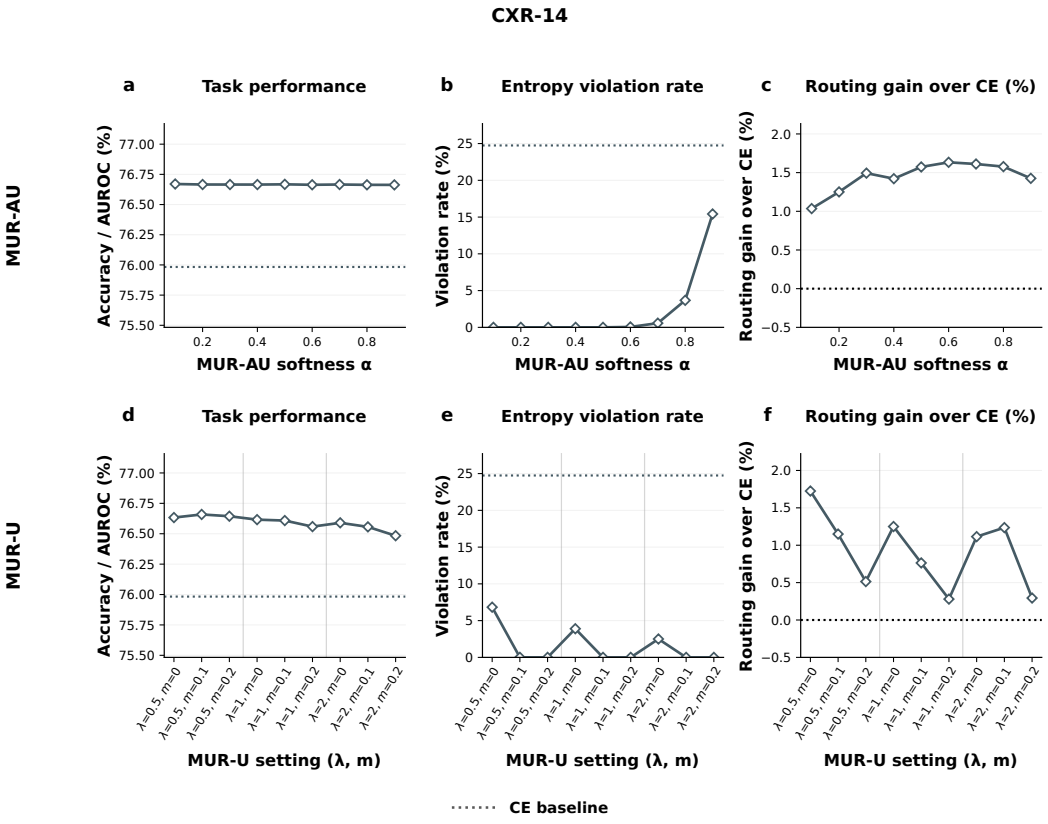

Figure 14: **Operating regimes on CXR-14.** Extends Figures 5 and 10: MUR-AU softness α (top) and MUR-U penalty weight λ and margin m (bottom). The favorable regimes observed on the three main settings transfer to all 14 findings: intermediate α and positive m yield low violations and positive routing gains at stable AUROC. Dotted lines denote the BCE baseline.

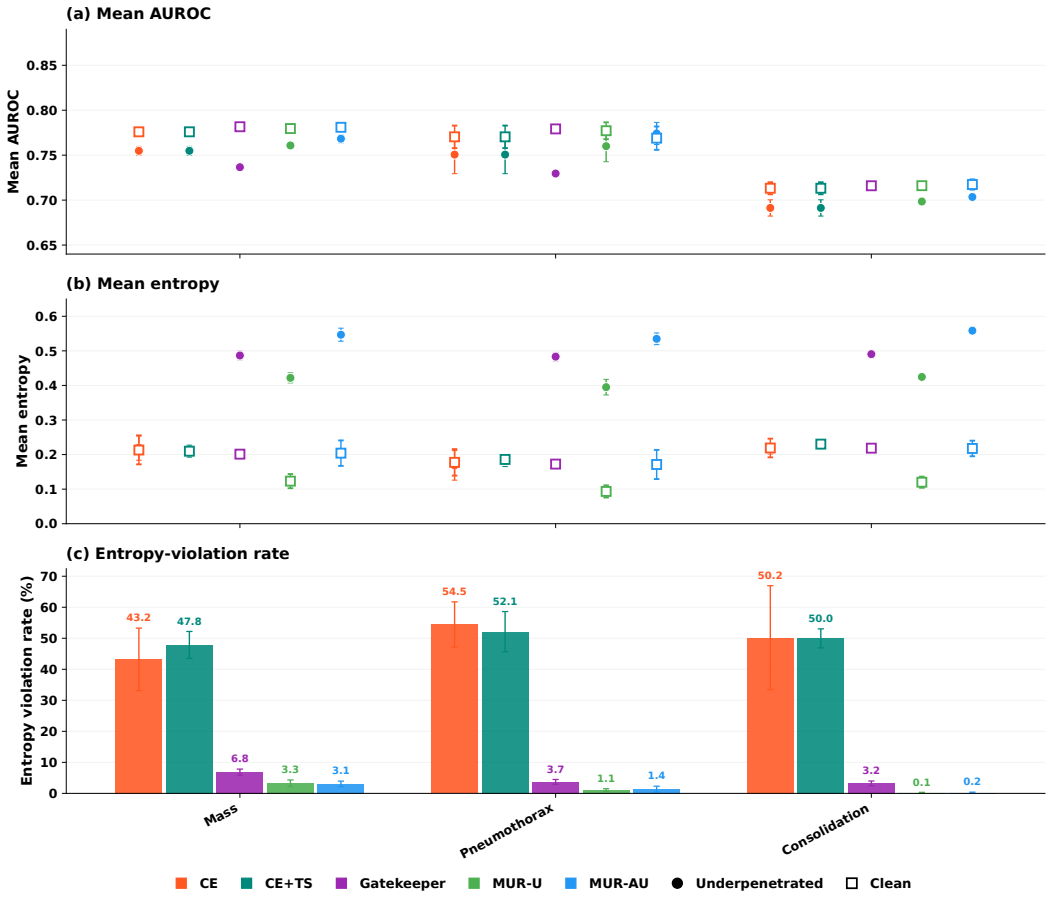

Figure 15: **Per-pathology uncertainty coherence on CXR-3.** **(a)** Mean AUROC and **(b)** mean predictive entropy on underpenetrated (○) and clean (□) images; **(c)** sample-level entropy-violation rate. MUR reduces per-pathology violations from 43.2%–54.5% to at most 3.3% while preserving AUROC, so the label-averaged result reflects coherence for each finding. Error bars show standard deviation over five seeds.

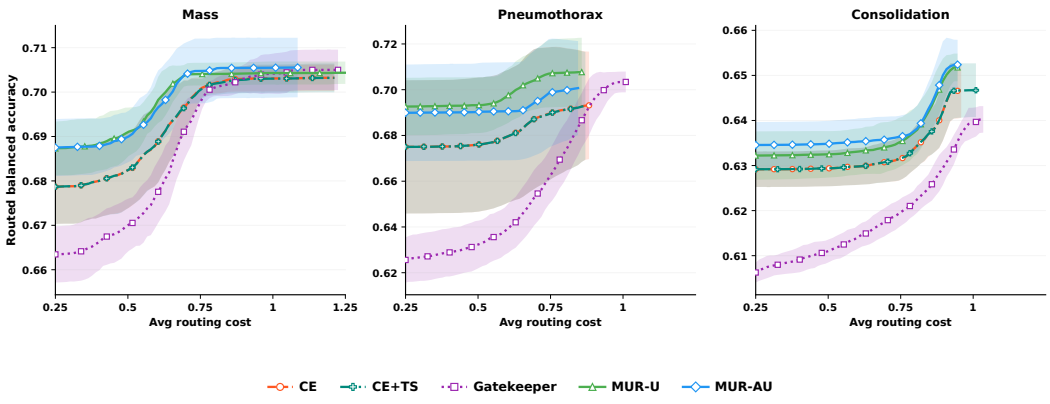

Figure 16: **Per-pathology routing frontiers on CXR-3.** Routed balanced accuracy versus average routing cost for Mass, Pneumothorax, and Consolidation. Shaded regions show standard deviation over five seeds. MUR dominates at low routing cost for all three findings, whereas Gatekeeper trails BCE until most samples are routed to the clean image.

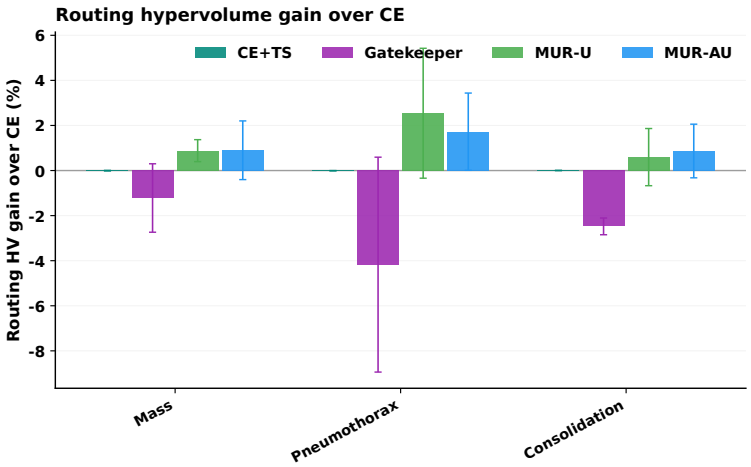

Figure 17: **Per-pathology routing hypervolume gain on CXR-3.** Gain relative to BCE for Mass, Pneumothorax, and Consolidation. Error bars show standard deviation over five seeds. Both MUR variants improve routing for all three findings, whereas Gatekeeper degrades it for all three.

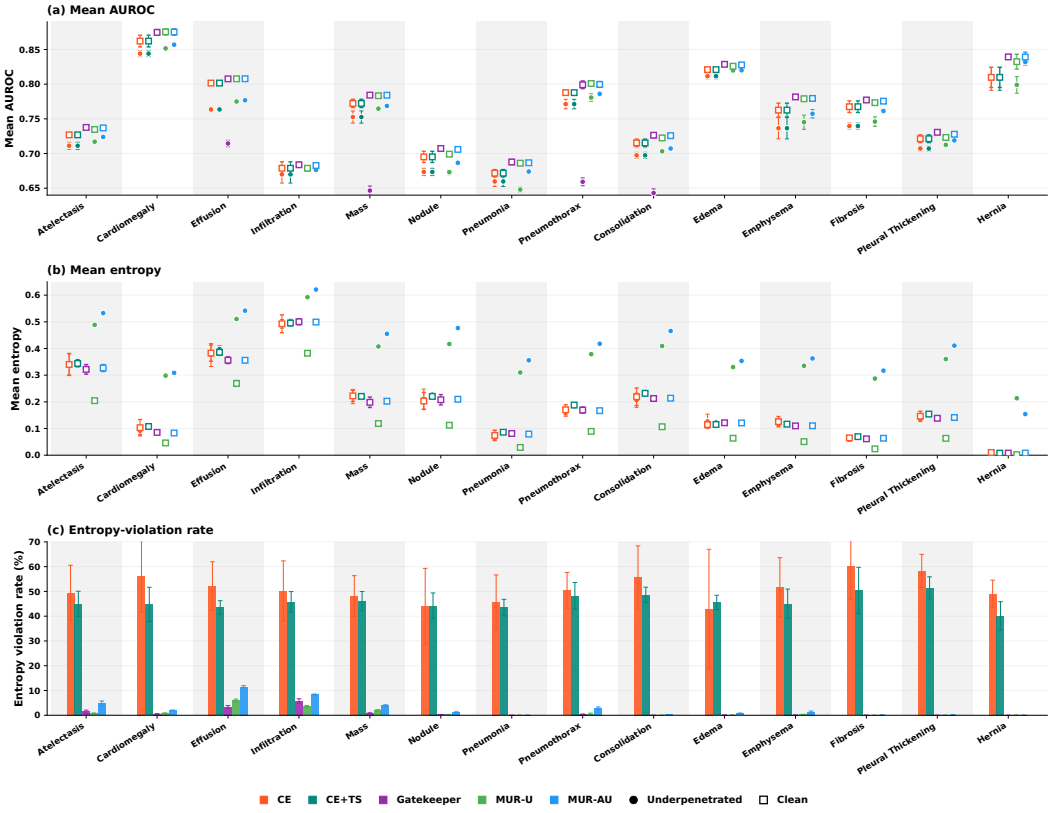

Figure 18: **Per-pathology uncertainty coherence on CXR-14.** **(a)** Mean AUROC and **(b)** mean predictive entropy on underpenetrated (○) and clean (□) images; **(c)** sample-level entropy-violation rate. Error bars show standard deviation over five seeds. MUR reduces violations across findings while preserving AUROC, although violations persist for Effusion and Infiltration.

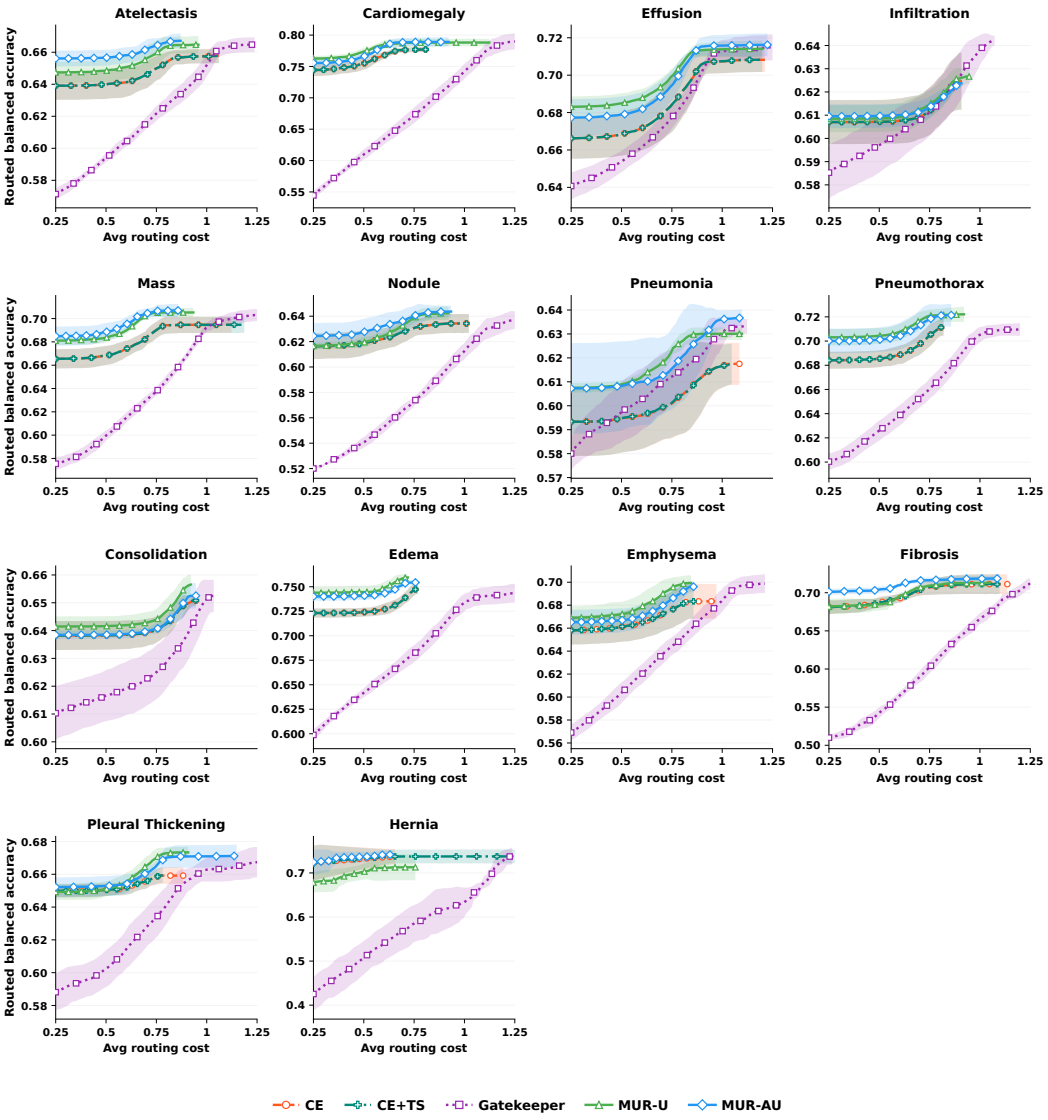

Figure 19: **Per-pathology routing frontiers on CXR-14.** Routed balanced accuracy versus average routing cost for all 14 findings. Shaded regions show standard deviation over five seeds. MUR provides stronger low-cost performance for most findings, whereas Gatekeeper generally trails BCE until routing approaches the clean image.

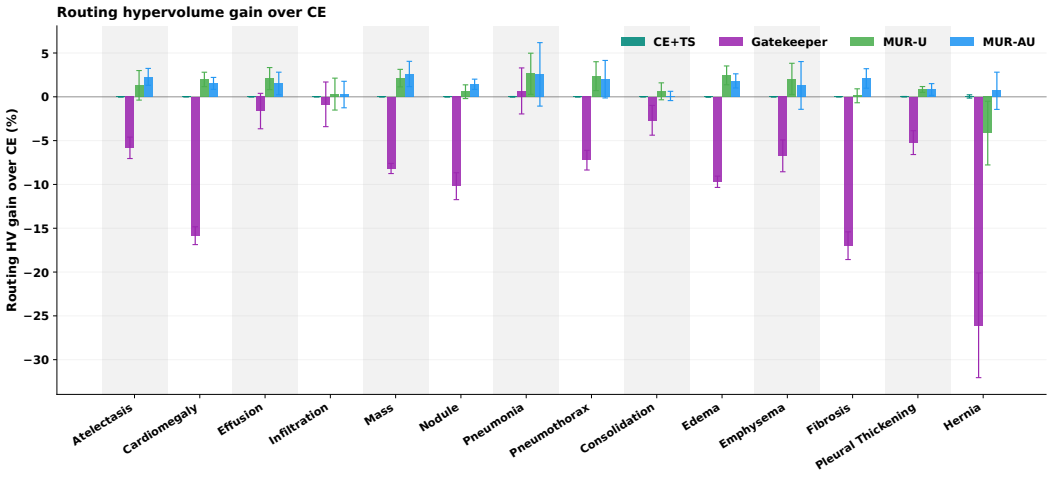

Figure 20: **Per-pathology routing hypervolume gain on CXR-14.** Gain relative to BCE across all 14 findings. Error bars show standard deviation over five seeds. MUR-AU and MUR-U improve routing for most findings, whereas Gatekeeper degrades routing for nearly all findings.