Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MV-CortexGuard: detecting subtle pediatric wrist fractures through cortical structure representation, anatomy-constrained multi-view consistency, and risk- controlled abstention

Shaofei Li
Hand and Foot Surgery Department of Yantai Yuhuangding Hospital
Honglei Ma
Hand and Foot Surgery Department of Yantai Yuhuangding Hospital
Chuanchong Zong
Hand and Foot Surgery Department of Yantai Yuhuangding Hospital
Mingchuan Sun
smingchuan@126.com
Hand and Foot Surgery Department of Yantai Yuhuangding Hospital

Research Article

Keywords: Pediatric wrist fracture, Subtle fracture detection, Multi-view learning, Cortical continuity, Uncertainty estimation, Selective prediction

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10864091/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

MV-CortexGuard: detecting subtle pediatric wrist fractures through cortical structure representation, anatomy-constrained multi-view consistency, and risk-controlled abstention

Shaofei Li[1]†, Honglei Ma[1]†, Chuanchong Zong[1]‡, Mingchuan Sun[1]*
[1] Hand and Foot Surgery Department of Yantai Yuhuangding Hospital, Yantai, Shandong, China
† These authors contributed equally to this work and share first authorship.
‡ Co-corresponding author: Chuanchong Zong, zongquanchong@163.com
* Corresponding author: Mingchuan Sun, smingchuan@126.com

Abstract

Background: Subtle pediatric wrist fractures may appear on radiographs only as focal cortical interruption, mild cortical bulging, or a fine linear lucency. Their detection is complicated by the epiphysis, growth plate, overlapping bones, and variations in projection. Most existing fracture detectors process each image independently, do not jointly model directional cortical continuity and complementary evidence from AP/PA (frontal) and lateral views, and cannot reliably identify high-confidence errors.

Methods: We developed MV-CortexGuard, a multi-view, risk-aware network for detecting subtle pediatric wrist fractures. Its bone-axis-aligned cortical discontinuity representation module estimates the wrist bone axis, maps features into normalized anatomical coordinates, and extracts cortical differences along the bone axis, its normal, and local curvature directions. Its anatomy-constrained multi-view query fusion module establishes cross-view correspondences using candidate appearance, normalized longitudinal location, osseous-region distribution, and evidence reliability, while a learnable null query preserves view-specific evidence. A multi-source uncertainty and risk-controlled rejection module integrates epistemic and aleatoric uncertainty, cross-view disagreement, localization instability, and image-quality reliability into an examination-level risk score and refers high-risk examinations for expert review.

Results: On GRAZPEDWRI-DX, MV-CortexGuard achieved an average precision (AP) of 63.4% at intersection-over-union thresholds of 0.50–0.95 and an AP of 52.7% in the size-defined small-fracture subset, exceeding YOLOv10m by 3.3 and 6.6 percentage points, respectively. The examination-level area under the receiver operating characteristic curve (AUROC) was 0.969. The uncertainty module reduced the expected calibration error to 0.024 and the area under the risk–coverage curve to 0.0107. At the validation-selected operating point targeting a selective error rate of 5%, the model accepted 94.3% of internal test examinations for automated classification.

Conclusions: MV-CortexGuard improved detection accuracy, cross-view diagnostic consistency, and predictive reliability for subtle pediatric wrist fractures. The results support further multicenter external validation and reader studies of risk-controlled computer-aided fracture diagnosis.

Keywords: Pediatric wrist fracture; Subtle fracture detection; Multi-view learning; Cortical continuity; Uncertainty estimation; Selective prediction

Background

Clinical motivation and study rationale

Fractures are among the most common musculoskeletal injuries encountered in pediatric emergency care. Because the wrist bears much of the supporting load during a fall, it is a frequent site of fracture in children. Anteroposterior or posteroanterior (AP/PA; frontal) and lateral (LAT) radiographs are generally the first-line examinations for suspected wrist fractures. Unlike adults, children have continuously developing wrist anatomy, and the morphology of the epiphyses, growth plates, ossification centers, and cortices changes substantially with age, creating considerable visual similarity between fractures and normal anatomy. Some subtle fractures manifest only as mild focal cortical bulging, a fine linear lucency, abnormal trabecular alignment, or incomplete cortical interruption and may be obscured by overlapping bones, projection angle, exposure differences, or superimposed soft tissue[1–6]. Accurate detection therefore requires not only assessment of local intensity changes but also joint consideration of cortical continuity, bone-axis orientation, and complementary evidence across projections.

Deep learning has advanced rapidly in musculoskeletal radiograph analysis. Early studies focused mainly on image-level fracture classification and subsequently expanded to fracture localization, detection, and subtype classification[1–7]. Public datasets such as GRAZPEDWRI-DX and FracAtlas have further supported standardized research on fracture detection[8–9]. GRAZPEDWRI-DX contains 20,327 wrist radiographs from 10,643 examinations in 6,091 children and provides bounding-box annotations for fractures and associated abnormalities, making it a relatively large public benchmark for pediatric wrist fracture detection[8]. YOLOv7, YOLOv8, YOLOv8-AM, and other modified detectors have demonstrated strong fracture-detection performance on this dataset[10–14]. However, most existing studies treat each radiograph as an independent input and improve overall performance primarily by changing the detector, introducing generic attention mechanisms, or modifying multi-scale feature fusion. They have not systematically modeled the structural characteristics of subtle fractures, the non-rigid correspondence between frontal and lateral radiographs, or high-confidence erroneous predictions.

From an object-detection perspective, subtle fractures are characterized by small target size, low foreground-to-background contrast, sparse abnormal tissue within an annotated bounding box, and numerous hard negatives. Feature pyramid networks, focal loss, deformable convolutions, and high-resolution sparse-query methods can improve small-object representation to some extent[15–22], whereas spatial transformer networks and equivariant convolutions provide general mechanisms for handling pose variation[23–25]. These approaches, however, predominantly learn scale or geometric variation from the data distribution and do not explicitly constrain bone-axis orientation, cortical continuity, or fracture morphology. Although generic spatial or channel attention can enhance salient regions[26], growth plates, epiphyseal margins, and regions of osseous overlap also produce strong edge responses and may consequently generate structurally plausible false positives.

Frontal and lateral radiographs are two-dimensional projections of the same three-dimensional osseous anatomy. Although the views provide diagnostically complementary information, they do not have stable one-to-one pixel correspondence. Multi-view convolutional networks, non-local operations, and Transformers provide a foundation for cross-view feature modeling[27–29]. In medical imaging tasks with multiple projections, particularly mammography, fusion based on feature concatenation, cross-view attention, or graph structures has improved examination-level diagnostic performance[30–34]. Direct fusion at the dense-feature level, however, can create spurious correspondences. When a fracture is visible in only one view, one view is of poor quality, or a projection is missing, forced fusion may also attenuate valid evidence in the available view.

Beyond detection accuracy, predictive reliability is essential for the clinical use of medical artificial intelligence. Modern deep neural networks may produce overconfident outputs that do not

reflect their empirical accuracy[35–39]. Monte Carlo dropout, deep ensembles, and methods that jointly model epistemic and aleatoric uncertainty can quantify predictive disagreement[35–37], but a single source of uncertainty generally cannot capture image-quality degradation, disagreement between frontal and lateral findings, and localization variability in multi-view fracture detection. Studies of selective prediction and learning to defer have shown that allowing a model to abstain on high-risk cases can provide a controllable trade-off between automated coverage and prediction-error rate[40–43]. Medical-imaging research has likewise emphasized probability calibration, explicit uncertainty reporting, and pathways for human review[44–48]. Nevertheless, most fracture detectors still rely on a single class-confidence score and do not integrate multi-source uncertainty, cross-view disagreement, and risk-controlled abstention within a unified detection framework.

To address these limitations, we developed MV-CortexGuard, a multi-view, risk-aware network for detecting subtle pediatric wrist fractures. The method uses cortical structural abnormalities as detection evidence, establishes anatomy-constrained cross-view relationships at the candidate-query level, and selectively abstains on high-risk examinations. The principal contributions are as follows.

First, we introduce the Bone-Axis-Aligned Cortical Discontinuity Representation (BACD) module. BACD automatically estimates bone-axis orientation from an osseous response map, transforms wrist features acquired under different projections into a normalized anatomical coordinate system, and extracts cortical difference responses along the bone axis, its normal, and local curvature directions. This design emphasizes the focal loss of continuity produced by subtle fractures. In contrast to generic attention, BACD reframes fracture detection from saliency enhancement to anatomically meaningful modeling of structural abnormalities.

Second, we propose Anatomy-Constrained Multi-View Query Fusion (ACMQF). Rather than interacting at the pixel level, ACMQF exchanges information between frontal and lateral views at the candidate-query level. Cross-view compatibility is calculated jointly from candidate appearance, longitudinal position along the bone axis, osseous-region distribution, and evidence reliability. A learnable null query represents the absence of a reliable counterpart, preventing inappropriate forced matching of lesions visible in only one view.

Third, we propose Multi-source Uncertainty and Risk-controlled Rejection (MURR). MURR combines epistemic uncertainty, aleatoric uncertainty, cross-view disagreement, localization instability, and image-quality reliability into an examination-level risk score, and determines the probability and risk-acceptance thresholds on the held-out validation set. The model automatically classifies low-risk examinations as fracture-positive or fracture-negative and refers high-risk examinations for expert review, thereby jointly addressing detection performance, probability reliability, and clinical safety.

Deep learning-based fracture detection

The use of convolutional neural networks in fracture imaging has progressed from image-level classification to lesion-level detection and localization. Olczak et al.[1] used a large collection of orthopedic trauma radiographs to establish the feasibility of deep networks for recognizing fractures and projection types. Lindsey et al.[2] subsequently showed that deep-learning assistance could improve clinicians' performance in identifying wrist fractures. Thian et al.[3] used a convolutional neural network to detect and localize wrist fractures, while Gan et al.[4] compared convolutional networks with specialists in distal radius fracture recognition. Blüthgen et al.[5] developed a deep-learning system capable of localizing distal radius fractures and reported diagnostic performance approaching that of radiologists.

Despite this potential, stability in difficult cases remains limited. Raisuddin et al.[6] found that, among diagnostically challenging wrist cases requiring CT confirmation, a deep model performed substantially worse than on a conventional test set, indicating that aggregate accuracy may conceal deficiencies in recognizing subtle or low-salience fractures. Hardalaç et al.[7] compared several deep object detectors and improved wrist fracture localization through model ensembling. Janisch et al.[49] examined the diagnostic value of different projections for pediatric radial buckle fractures, demonstrating a close association between projection direction and fracture type. Cohen et al.[50] further compared the wrist-fracture detection performance of artificial intelligence and radiologists from the perspective of clinical observers.

Following the release of GRAZPEDWRI-DX, studies of pediatric wrist fracture detection using a common dataset have increased. Ju and Cai[10] applied YOLOv8 to GRAZPEDWRI-DX and confirmed the effectiveness of a one-stage detector for localizing pediatric wrist fractures. Using YOLOv7, Till et al.[11] investigated the effects of data processing, model optimization, and collaboration between medical and engineering teams on detection performance. Chen et al.[12] improved multi-site fracture detection through attention and feature fusion. Chien et al.[13] incorporated several channel- and spatial-attention modules into YOLOv8 to enhance responses within fracture regions. Liu et al.[14] further modified feature extraction and detection components based on YOLO11.

Existing methods have largely optimized the backbone, feature pyramid, or generic attention mechanisms, and their enhancement of fracture regions remains predominantly based on data-driven saliency. Because these approaches do not explicitly describe the bone axis or cortical continuity, they may misidentify growth plates, epiphyseal margins, and regions of osseous overlap as fractures. Moreover, most studies detect each image independently and do not fully exploit the complementary relationship between frontal and lateral images from the same examination. Their outputs are also driven mainly by classification confidence, without a systematic abstention mechanism for high-risk examinations.

Small-object detection and anatomical structure representation

General-purpose object-detection methods provide the basic technical framework for fracture localization. Faster R-CNN established an end-to-end two-stage detection paradigm using a region proposal network[15]. Feature pyramid networks improved multi-scale object representation by combining features from different levels[16]. Focal loss mitigated foreground-background imbalance by down-weighting easily classified background examples[17]. Deformable convolutions enhanced adaptation to irregular geometric structures through learnable sampling locations[18].

Transformer-based detectors further formulate object detection as set prediction. DETR enables end-to-end detection through object queries and Hungarian matching[19], while Deformable DETR improves small-object performance and convergence speed through sparse multi-scale sampling[20]. QueryDet uses cascaded sparse queries on high-resolution feature maps to reduce the computational cost of small-object detection[21]. RT-DETR achieves a favorable balance between real-time operation and end-to-end detection performance[22]. These studies motivate our use of high-resolution features, candidate queries, and a set-prediction detection head.

For pose and orientation variation, spatial transformer networks can learn differentiable geometric transformations[23], whereas group-equivariant convolutions and E(2)-equivariant networks use structured parameter sharing to improve equivariance to transformations such as rotation and reflection[24–25]. Attention mechanisms such as CBAM reweight channel and spatial responses to enhance salient regions[26]. General geometric transformations and attention modules, however, do not explicitly describe the directional continuity of the cortex. In pediatric wrist radiographs, rotation invariance is not equivalent to anatomical consistency: a model must not only accommodate variation in wrist pose but also determine whether a local edge maintains plausible continuity along the bone axis. We therefore estimate the bone axis and construct directional cortical differences to jointly model geometric normalization and fracture morphology.

Multi-view learning in medical imaging

Multi-view learning seeks complementary representations from

multiple observations of the same object. Self-attention in Transformers provides a unified mechanism for long-range interaction across positions or views[27], while non-local neural networks model feature dependencies over the global spatial domain[28]. The multi-view convolutional network proposed by Su et al.[29] aggregates features from multiple views to improve three-dimensional object recognition; however, its view-pooling mechanism primarily produces a global representation and is not well suited to fine-grained lesion-level correspondence.

Multi-view modeling in medical imaging has been studied extensively in mammography. Carneiro et al.[30] developed a deep-learning model for unregistered multi-view mammograms, showing that cross-view information can be exploited without strict image registration. Van Tulder et al.[31] proposed a cross-view Transformer that transfers information among spatial features from different projections and validated it on mammography and chest radiography tasks. Chen et al.[32] combined local and global Transformers to analyze multiple mammograms from the same examination, thereby modeling long-range within-view and between-view dependencies. Zhou and Krähenbühl[33] used camera-dependent positional encoding to map different views into a canonical space. Manigrasso et al.[34] systematically compared graph networks and Transformers for multi-view mammographic classification, further demonstrating that structured cross-view modeling outperforms simple global feature concatenation.

Frontal and lateral pediatric wrist radiographs resemble multi-view mammograms in that the views contain complementary diagnostic information but are not aligned at the pixel level. Unlike examination-level classification, however, fracture detection must retain lesion-level localization. Direct concatenation of dense features can mix different osseous regions, while ordinary cross-attention may assign every candidate a spurious counterpart in the other view. Forced cross-view consistency can attenuate valid evidence, particularly when a subtle fracture is visible in only one view, image quality is degraded, or a projection is missing. We therefore impose anatomical-position and osseous-region constraints at the candidate-query level and use a null query to explicitly represent the state of "no reliable counterpart."

Uncertainty estimation and risk-controlled abstention

Class probabilities produced by deep neural networks are not intrinsically equivalent to the probability that a prediction is correct. Gal and Ghahramani[35] interpreted dropout as approximate Bayesian inference and estimated model uncertainty through repeated stochastic forward passes. Lakshminarayanan et al.[36] proposed deep ensembles, which provide comparatively stable uncertainty estimates from disagreement among independently trained models. Kendall and Gal[37] further distinguished epistemic uncertainty caused by limited model knowledge from aleatoric uncertainty arising from observational noise.

Guo et al.[38] showed that modern neural networks are often poorly calibrated and proposed temperature scaling as a simple and effective post-hoc calibration method. Ovadia et al.[39] systematically evaluated uncertainty methods under dataset shift and found that good calibration on an in-distribution validation set does not guarantee reliability under distribution shift. Accordingly, medical-imaging models should assess input quality, model disagreement, and anomalous data jointly rather than report prediction probabilities alone.

Selective prediction permits a model to reject some samples, creating a trade-off between coverage and prediction-error rate. Geifman and El-Yaniv[40] proposed a confidence-ranking approach to deep selective classification that determines the acceptance set at a specified risk level. SelectiveNet integrates a classifier and selection function within a single network[41]. Corbière et al.[42] improved the identification of erroneous predictions by learning the probability of the true class. Mozannar and Sontag[43] studied the problem of learning to defer decisions to an expert, providing a theoretical basis for human–AI collaborative diagnosis.

In medical imaging, Mehrtash et al.[44] investigated probability calibration and predictive uncertainty in deep medical-image segmentation. Nair et al.[45] aggregated voxel-level uncertainty into lesion-level uncertainty for multiple-sclerosis lesion screening. Kompa et al.[46] argued that medical AI systems should be able to express uncertainty and request a second opinion. Bates et al.[47] further proposed distribution-free, risk-controlling prediction sets, providing a theoretical basis for selecting risk thresholds using calibration data. Huang et al.[48] systematically reviewed probabilistic and non-probabilistic uncertainty methods in medical image analysis.

Existing medical-imaging studies of uncertainty have focused mainly on classification or segmentation and commonly use predictive entropy, Monte Carlo dropout variance, or ensemble disagreement as a single basis for abstention. In multi-view fracture detection, errors can arise from insufficient model knowledge, low image contrast, a missing projection, disagreement between frontal and lateral findings, or unstable bounding-box localization. A single confidence score cannot fully represent examination-level risk. We therefore combine epistemic uncertainty, aleatoric uncertainty, cross-view disagreement, localization uncertainty, and image-quality reliability into an examination-level risk representation and determine the risk-acceptance threshold on the held-out validation set, establishing a risk-controlled abstention mechanism for automated examination-level diagnosis.

Methods

Overall framework

We developed MV-CortexGuard, a multi-view risk-aware network for detecting subtle fractures on pediatric wrist radiographs. The framework addresses the small size and low contrast of subtle fracture findings, discordant manifestations across frontal and lateral views, and the inability of conventional confidence scores to adequately reflect predictive risk by integrating cortical-structure representation, multi-view evidence fusion, and uncertainty-guided rejection.

MV-CortexGuard comprises three core modules: Bone-Axis-Aligned Cortical Discontinuity Representation (BACD), Anatomy-Constrained Multi-View Query Fusion (ACMQF), and Multi-source Uncertainty and Risk-controlled Rejection (MURR). The overall architecture is shown in Fig. 1.

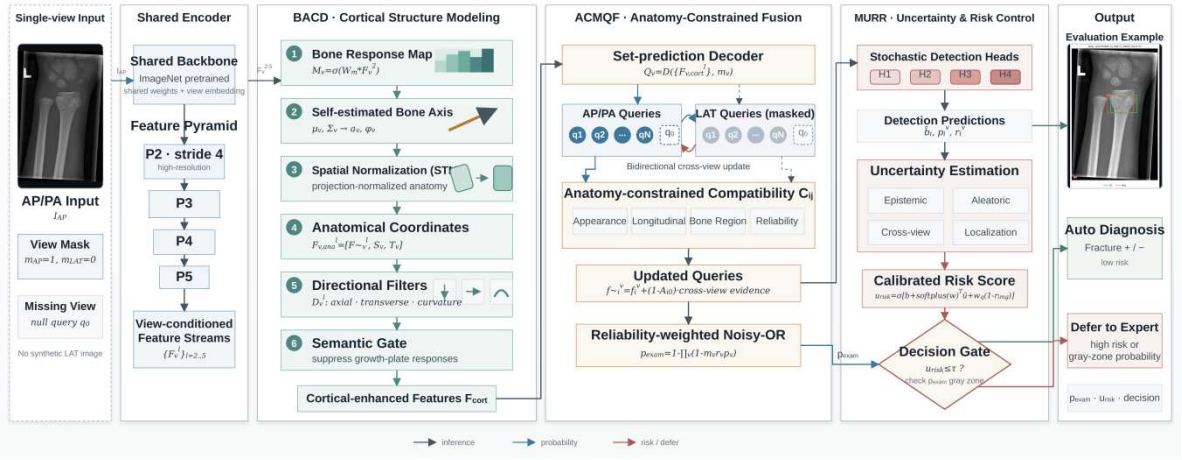


Fig. 1 Overall architecture of MV-CortexGuard. BACD normalizes the estimated bone axis and encodes directional cortical discontinuities; ACMQF performs anatomy- and reliability-constrained candidate-query interaction while allowing null matches; and MURR combines multi-source uncertainty into a calibrated examination-level risk score for automated prediction or referral for expert review.

For wrist examination e, let the set of available projection views be $\mathcal{V}_e \subseteq \{AP, LAT\}$, where AP denotes an anteroposterior or posteroanterior radiograph and LAT denotes a lateral radiograph. For view v, the input radiograph is denoted by I_v, and view availability is indicated by $m_v \in \{0,1\}$. Each available radiograph is passed through a shared-weight multiscale encoder E, and a view embedding e_v modulates the feature distribution for the corresponding projection. The output at scale l is

$$F_v^l = E_l(I_v; e_v), \qquad v \in \mathcal{V}_e, \qquad l \in \{2,3,4,5\}. \tag{1}$$

Here, $F_v^l \in \mathbb{R}^{C_l \times H_l \times W_l}$, where C_l, H_l, and W_l denote the number of feature channels, height, and width, respectively. The encoder retains high-resolution P_2 features at a stride of 4 to mitigate the loss of subtle fracture information during successive downsampling.

BACD transforms the multiscale features into cortical-enhanced features $F_{v,cor}^l$. A set-prediction decoder $\mathcal{D}$ then generates a view-level candidate-query set from the multiscale cortical responses:

$$\mathcal{Q}_v = \mathcal{D}\left(\{F_{v,cor}^l\}_{l=2}^5, m_v\right) = \{q_i^v\}_{i=1}^{N_v}. \tag{2}$$

Here, N_v is the number of retained candidate queries in view v, and q_i^v denotes the i-th candidate fracture query. The detector uses one-to-one set prediction, with Hungarian matching used to assign predicted queries to ground-truth fracture bounding boxes. The detection loss is defined as

$$\mathcal{L}_{det} = \sum_{i=1}^{N_v}\left[\lambda_{cls}\mathcal{L}_{focal} + \mathbb{I}_i^+\left(\lambda_1 \parallel \hat{b}_i - b_i \parallel_1 + \lambda_{giou}\mathcal{L}_{GIoU}\right)\right]. \tag{3}$$

Here, $\hat{b}_i$ and b_i denote the predicted and ground-truth bounding boxes, respectively, and $\mathbb{I}_i^+$ indicates whether query i is matched to a ground-truth object.

The model is optimized end to end using a multi-task objective. The total loss is

$$\mathcal{L}_{total} = \mathcal{L}_{det} + \lambda_a\mathcal{L}_{axis} + \lambda_c\mathcal{L}_{cor} + \lambda_m\mathcal{L}_{mv} + \lambda_u\mathcal{L}_{unc}. \tag{4}$$

Here, $\mathcal{L}_{axis}$, $\mathcal{L}_{cor}$, $\mathcal{L}_{mv}$, and $\mathcal{L}_{unc}$ denote the bone-axis equivariance loss, cortical representation loss, multi-view consistency loss, and uncertainty modeling loss, respectively; the corresponding λ terms are their weighting coefficients.

Bone-Axis-Aligned Cortical Discontinuity Representation

Subtle pediatric wrist fractures commonly manifest as focal cortical discontinuity, subtle cortical bulging, fine linear radiolucency, or an abnormal trabecular pattern. These findings can be confounded by the projection angle, wrist rotation, growth plates, and epiphyseal margins. Conventional spatial or channel attention relies primarily on data-driven weighting and does not explicitly represent the directional continuity of the cortex. BACD therefore combines bone-axis normalization with direction-sensitive cortical-difference modeling to enhance fracture-related structural abnormalities in a common anatomical coordinate system.

First, a lightweight anatomical-response branch is applied to the high-resolution feature F_v^2 to generate a bone-response map:

$$M_v = \sigma\left(W_m * F_v^2\right), \qquad M_v \in [0,1]^{H_2 \times W_2}. \tag{5}$$

Here, W_m denotes a 1×1 convolution and $\sigma(\cdot)$ is the sigmoid function. The weighted centroid and spatial covariance matrix of the bone region are then computed from the response map:

$$\boldsymbol{\mu}_v = \frac{\sum_{\mathbf{x}\in\Omega} M_v(\mathbf{x})\mathbf{x}}{\sum_{\mathbf{x}\in\Omega} M_v(\mathbf{x}) + \varepsilon}, \qquad \boldsymbol{\Sigma}_v = \frac{\sum_{\mathbf{x}\in\Omega} M_v(\mathbf{x})(\mathbf{x}-\boldsymbol{\mu}_v)(\mathbf{x}-\boldsymbol{\mu}_v)^T}{\sum_{\mathbf{x}\in\Omega} M_v(\mathbf{x}) + \varepsilon}. \tag{6}$$

Here, Ω denotes the set of feature-map coordinates, and ε is a numerical-stability constant. Eigendecomposition is performed on $\boldsymbol{\Sigma}_v$, and the unit eigenvector $\mathbf{a}_v = [a_x, a_y]^T$ associated with its largest eigenvalue is taken as the bone-axis direction. The corresponding rotation angle is calculated as

$$\hat{\theta}_v = \operatorname{atan2}(a_y, a_x), \qquad \phi_v = \hat{\theta}_v - \frac{\pi}{2}. \tag{7}$$

This estimation procedure requires no additional manual bone-axis annotations. During training, random rotation, scaling, and translation augmentations impose a geometric equivariance constraint on the bone-axis response, thereby improving the stability of bone-axis estimation under variations in radiographic positioning.

A homogeneous affine transformation matrix is constructed from the bone-axis direction, bone-structure centroid, and scale parameter s_v:

$$A_v = \begin{bmatrix} s_v\cos\phi_v & -s_v\sin\phi_v & t_x \\ s_v\sin\phi_v & s_v\cos\phi_v & t_y \\ 0 & 0 & 1 \end{bmatrix}, \qquad \widetilde{F}_v^l = \mathcal{T}\left(F_v^l; A_v\right). \tag{8}$$

Here, t_x and t_y are translation parameters, and $\mathcal{T}(\cdot)$ denotes a differentiable spatial transformation based on bilinear interpolation. This operation aligns the principal distal radial and ulnar axes to an approximately vertical orientation across examinations. Detection results are mapped back to the original image coordinate system using A_v^{-1}.

To encode candidate positions relative to the bone axis, a longitudinal coordinate map S_v and a transverse coordinate map T_v are appended to the normalized features. S_v represents the normalized relative position along the bone axis, whereas T_v represents the transverse offset from its centerline. The anatomy-augmented feature is defined as

$$F_{v,ana}^l = \operatorname{Concat}\left(\widetilde{F}_v^l, S_v, T_v\right) \in \mathbb{R}^{(C_l+2)\times H_l \times W_l}. \tag{9}$$

Directional differences are computed in the normalized coordinate system along the bone axis, its normal, and the local curvature direction. The directional-difference matrix and final cortical-enhanced features are defined as

$$\begin{aligned} \mathbf{D}_v^l &= \begin{bmatrix} \left|K_\parallel * F_{v,ana}^l\right| \\ \left|K_\perp * F_{v,ana}^l\right| \\ \left|K_\perp^{(2)} * F_{v,ana}^l\right| \end{bmatrix}, \\ G_v^l &= \sigma\left(W_g * \left[F_{v,ana}^l; \mathbf{D}_v^l\right]\right), \\ F_{v,cor}^l &= F_{v,ana}^l + G_v^l \odot \Phi\left(\mathbf{D}_v^l\right). \end{aligned} \tag{10}$$

Here, $K_\parallel$ captures discontinuities along the bone axis, $K_\perp$ detects transverse fracture lines and cortical step-offs, and $K_\perp^{(2)}$ characterizes local curvature abnormalities. G_v^l is the semantic gating weight, and $\Phi(\cdot)$ is the directional-difference mapping

function.

BACD treats features within fracture bounding boxes as positive samples and regards normal cortical regions, together with high-response regions that do not overlap any fracture annotation, as hard negatives. Contrastive learning is used to optimize $\mathcal{L}_{cor}$. This strategy increases feature separation between subtle fractures and visually similar structures, such as growth plates and epiphyseal margins.

Anatomy-Constrained Multi-View Query Fusion

Frontal (AP/PA) and lateral radiographs are two-dimensional projections of the same three-dimensional osseous anatomy acquired from different directions. Although they provide complementary diagnostic evidence, they do not exhibit stable pixel-level correspondence, and some subtle fractures are visible in only one view. Direct feature concatenation may therefore introduce spurious correspondences and dilute view-specific evidence. ACMQF instead performs cross-view interaction at the candidate-query level and constrains matching using normalized anatomical positions, bone regions, and candidate reliability.

The i-th candidate query in view v is represented as

$$q_i^v = \left[f_i^v,\, s_i^v,\, t_i^v,\, \mathbf{b}_i^v,\, p_i^v,\, r_i^v\right]. \tag{11}$$

Here, $f_i^v \in \mathbb{R}^d$ is the candidate appearance feature; s_i^v and t_i^v are the longitudinal and transverse normalized anatomical coordinates, respectively; $\mathbf{b}_i^v$ contains the candidate's soft-assignment probabilities over bone-region prototypes; p_i^v is the fracture probability; and $r_i^v \in [0,1]$ is the reliability of the candidate evidence.

For frontal query i and lateral query j, the anatomy-constrained compatibility score is defined as

$$C_{ij} = S_{ij}^{\text{app}} - \beta_s \Delta_{ij}^{\text{long}} - \beta_b D_{ij}^{\text{bone}} + R_j. \tag{12}$$

$$S_{ij}^{\text{app}} = \frac{\left(W_q f_i^{AP}\right)^T \left(W_k f_j^{LAT}\right)}{\sqrt{d}}, \qquad \Delta_{ij}^{\text{long}} = \left|s_i^{AP} - s_j^{LAT}\right|. \tag{13}$$

$$D_{ij}^{\text{bone}} = \text{JS}\left(\mathbf{b}_i^{AP} \parallel \mathbf{b}_j^{LAT}\right), \qquad R_j = \log\left(r_j^{LAT} + \varepsilon\right). \tag{14}$$

The first term measures semantic similarity between candidate features, the second constrains their relative longitudinal positions along the bone axis, the third enforces consistency between their bone-region assignments, and the final term uses the reliability of the candidate from the other view as a matching prior.

Because some lesions are not visible in the other view, a learnable null query q_0^v represents the absence of a reliable corresponding candidate. Setting $C_{i0} = c_{null}$, the cross-view matching weights are

$$A_{ij}^{AP\to LAT} = \frac{\exp\left(C_{ij}\right)}{\sum_{j'=0}^{N_{LAT}} \exp\left(C_{ij'}\right)}, \qquad j \in \{0,1,\dots,N_{LAT}\}. \tag{15}$$

Here, A_{i0} denotes the probability that query i matches the null query. A larger A_{i0} indicates that the lateral radiograph lacks reliable evidence corresponding to the frontal-view candidate.

The cross-view update for a frontal-view query is defined as

$$\tilde{f}_i^{AP} = f_i^{AP} + (1 - A_{i0}) g_i \sum_{j=1}^{N_{LAT}} A_{ij}\, r_j^{LAT} W_v f_j^{LAT}. \tag{16}$$

Here, g_i is a query-level fusion gate and W_v is a linear projection matrix. Lateral queries are updated symmetrically. This update is jointly modulated by the null-query probability, the reliability of candidates from the other view, and the fusion gate, thereby preventing low-quality or unmatched cross-view information from being forcibly injected into the current query.

Candidate predictions are aggregated into a view-level fracture probability using reliability-weighted noisy-OR, after which the examination-level probability is fused across the available views:

$$p_v = 1 - \prod_{i=1}^{N_v}\left(1 - r_i^v p_i^v\right), \qquad p_{exam} = 1 - \prod_{v\in\mathcal{V}_e}\left(1 - m_v r_v p_v\right). \tag{17}$$

Here, r_v denotes view-level quality reliability. This aggregation allows reliable fracture evidence from either view to support a positive examination-level decision while remaining compatible with single-view examinations.

The multi-view training loss comprises candidate-distribution consistency, bidirectional matching cycle consistency, and examination-level classification loss:

$$\mathcal{L}_{mv} = \mathcal{L}_{dist} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_e \mathcal{L}_{exam}. \tag{18}$$

$$\mathcal{L}_{dist} = \sum_{i,j} A_{ij}\, r_i^{AP} r_j^{LAT} \text{JS}\left(P_i^{AP} \parallel P_j^{LAT}\right). \tag{19}$$

$$\mathcal{L}_{cyc} = \left\| \overline{A}^{AP\to LAT} \overline{A}^{LAT\to AP} - I \right\|_F^2, \qquad \mathcal{L}_{exam} = \text{BCE}\left(p_{exam}, y\right). \tag{20}$$

Here, $\mathcal{L}_{dist}$, $\mathcal{L}_{cyc}$, and $\mathcal{L}_{exam}$ denote candidate-distribution consistency, bidirectional matching-cycle consistency, and examination-level classification loss, respectively. P_i^v denotes the query-class distribution, $\overline{A}$ is the matching matrix after removal of the null query, and y is the examination-level label. Candidate reliability modulates the strength of the consistency constraint, avoiding an unjustifiably strong consistency requirement for lesions visible in only one view.

Multi-source Uncertainty and Risk-controlled Rejection

Classification probabilities alone do not adequately characterize predictive reliability. Low contrast, nonstandard projections, cast-related occlusion, osseous superimposition, or disagreement between frontal and lateral predictions may produce high-confidence errors. MURR jointly estimates epistemic uncertainty, aleatoric uncertainty, cross-view disagreement, and localization instability to calibrate predictive risk and applies selective abstention to examinations that do not meet the criteria for an automated fracture decision.

K lightweight stochastic adapter detection heads are placed after the shared encoder and query decoder. The heads share the main network parameters while retaining independent low-rank adapter parameters and output layers. Their mean predictive probability and epistemic uncertainty are defined as

$$\overline{p} = \frac{1}{K}\sum_{k=1}^{K} p^{(k)}, \qquad u_{epi} = \frac{1}{K-1}\sum_{k=1}^{K}\left(p^{(k)} - \overline{p}\right)^2. \tag{21}$$

Here, $p^{(k)}$ is the examination-level predictive probability from the k-th detection head, and u_{epi} measures predictive disagreement across parameter hypotheses.

The examination-level multi-source uncertainty vector is

$$\mathbf{u} = \begin{bmatrix} u_{epi} \\ u_{ale} \\ u_{view} \\ u_{loc} \end{bmatrix} = \begin{bmatrix} \text{Var}_k\left(p^{(k)}\right) \\ K^{-1}\sum_{k=1}^{K}\sigma_{p,k}^2 \\ JS\left(Ber(p_{AP}) \parallel Ber(p_{LAT})\right) \\ \text{tr}(\boldsymbol{\Sigma}_b) \end{bmatrix}. \tag{22}$$

Here, u_{ale} represents aleatoric uncertainty arising from image noise, occlusion, and low contrast; u_{view} represents disagreement between frontal and lateral predictions; and u_{loc} represents bounding-box instability across the detection heads. For examinations lacking one projection, a missing-view indicator and the quality reliability of the available view jointly characterize the risk associated with incomplete information.

Each uncertainty component is quantile-normalized using the held-out validation set used for calibration, yielding $\tilde{\mathbf{u}}$. The examination-level risk score is calculated as

$$u_{risk} = \sigma\left[b + \text{softplus}(\mathbf{w})^T \tilde{\mathbf{u}} + w_q\left(1 - r_{img}\right)\right]. \tag{23}$$

Here, r_{img} denotes image-quality reliability and jointly reflects projection completeness, exposure quality, bone-axis estimation stability, and the number of available views. Applying $\text{softplus}(\mathbf{w})$ imposes a nonnegative constraint on the risk weights, ensuring that the risk score is monotonically nondecreasing with respect to every uncertainty component.

The final selective decision rule uses two probability thresholds and one risk threshold:

$$\hat{y} = \begin{cases} 1, & u_{risk} \le \tau \wedge p_{exam} \ge t_+, \\ 0, & u_{risk} \le \tau \wedge p_{exam} \le t_-, \\ \text{review}, & \text{otherwise}. \end{cases} \tag{24}$$

Here, t_+ and t_- are the positive and negative probability thresholds, respectively, with $t_- < t_+$, and τ is the risk-acceptance threshold. When risk is excessive or the predictive probability lies within the indeterminate zone, the model does not issue a definitive decision and instead flags the examination for expert review.

All thresholds are determined using the held-out validation set used for calibration. Let $a_i(\tau)$ indicate whether examination i is accepted by the model, n_τ denote the number of accepted examinations, and $\hat{R}(\tau)$ denote the empirical error rate among accepted examinations. The optimal risk threshold is

$$\tau^* = \arg\max_{\tau} \frac{n_\tau}{n},$$
$$\text{s.t.} \quad \widehat{R}(\tau) + \sqrt{\frac{\log(1/\delta)}{2n_\tau}} \le \alpha, \qquad \widehat{R}(\tau) = \frac{\sum_{i=1}^{n} a_i(\tau)\mathbb{I}(\hat{y}_i \neq y_i)}{n_\tau}. \quad (25)$$

Here, α is the prespecified maximum acceptable selective error rate, and δ is the failure-probability parameter, corresponding to a confidence level of $1-\delta$. This threshold-selection strategy maximizes automated-decision coverage subject to the risk upper-bound constraint, thereby balancing predictive reliability against clinical workflow efficiency.

Study design and data source

Experiments were conducted using the publicly available GRAZPEDWRI-DX pediatric wrist trauma radiograph dataset [8]. The dataset contains 6,091 patients, 10,643 radiographic examinations, and 20,327 wrist radiographs, acquired mainly in AP/PA, lateral, and oblique projections. Bounding-box annotations are provided for fractures, periosteal reactions, bone abnormalities, bone lesions, metal, and soft-tissue abnormalities. Because this study focuses on detecting subtle pediatric wrist fractures, only the "fracture" category was treated as the target class; annotations for other abnormalities were excluded from calculation of the primary detection metrics.

Dataset preparation

A patient-level split was used to prevent radiographs from the same patient from appearing in different data subsets. The training, validation, and test sets were constructed in a 6:2:2 ratio. Frontal and lateral radiographs were paired using examination identifiers and projection metadata. If either projection was unavailable, the examination was retained as a single-view sample using a view-availability mask. The final partition is summarized in Table 1.

Table 1 Dataset characteristics by split.

Subset	Patients	Examinations	Radiographs	Fracture instances	Small fractures	Paired AP/PA–LAT examinations
Training	3,655	6,386	12,196	8,130	2,033	4,817
Validation	1,218	2,128	4,065	2,710	678	1,601
Test	1,218	2,129	4,066	2,710	678	1,604
Total	6,091	10,643	20,327	13,550	3,389	8,022

AP/PA, anteroposterior or posteroanterior; LAT, lateral. Small fractures were defined by a normalized fracture-box area below 0.0016, corresponding to the 25th percentile of the training-set distribution.

To evaluate performance specifically for small fractures, a size-defined small-fracture subset was constructed using the area of each fracture bounding box normalized by the original image area. For the i-th fracture bounding box, the normalized area is

$$r_i = \frac{w_i h_i}{WH}. \quad (26)$$

Here, w_i and h_i are the width and height of the fracture bounding box, respectively, and W and H are the original image dimensions. The 25th percentile of the fracture-box area distribution in the training set was used as the small-fracture threshold. A fracture was therefore assigned to the size-defined small-fracture subset when $r_i < 0.0016$. The threshold was determined using the training set only and was then applied unchanged to the validation and test sets.

Implementation details

All radiographs were resized to fit within 1024×1024 pixels while preserving their original aspect ratio, with zero padding applied to the remaining borders. Training augmentation comprised random rotation, translation, scaling, horizontal flipping, contrast perturbation, Gaussian noise, and mild blurring. Rotation was sampled from $[-10°, 10°]$, and the scaling range was $[0.9, 1.1]$. Only deterministic resizing and grayscale normalization were applied to the validation and test sets.

The model was initialized using ImageNet-pretrained weights and optimized with AdamW using an initial learning rate of 2×10^{-4} and weight decay of 1×10^{-4}. Training was performed for 150 epochs, with linear learning-rate warm-up during the first 5 epochs followed by cosine annealing. The batch size was 16, and the number of stochastic adapter detection heads in MURR was $K = 4$. The loss weights were $\lambda_a = 0.5$, $\lambda_c = 0.5$, $\lambda_m = 0.25$, and $\lambda_u = 0.1$.

All experiments were implemented in PyTorch 2.1 and performed using two NVIDIA RTX 4090 GPUs. The principal detection experiments were repeated with five random seeds; AP metrics are reported as mean ± standard deviation where shown, whereas examination-level AUROC and uncertainty metrics are reported as point estimates in the available study record. Model selection, probability calibration, and rejection-threshold selection were performed using the held-out validation set; the test set was used only for final performance evaluation.

Baselines and evaluation metrics

YOLOv8-AM, RT-DETR-R50, and YOLOv10m were the primary comparison methods reported in the main results. All models were retrained using the same patient-level data split, input resolution, and data-augmentation strategy. For detectors without native multi-view modeling, inference was performed separately on AP/PA and LAT radiographs, and an examination-level prediction was obtained using a late probability-fusion strategy selected on the validation set.

Detection performance was evaluated using AP@50, AP@75, and AP@50–95. Because fracture was the only target class, class-specific AP was numerically equivalent to mAP. Performance on the size-defined small-fracture subset was evaluated using AP_S and lesion-level sensitivity at 0.5, 1, and 2 false positives per image. Examination-level diagnostic performance was evaluated using the area under the receiver operating characteristic curve (AUROC), sensitivity, and specificity.

Probability calibration was evaluated using expected calibration error (ECE), Brier score, and negative log-likelihood (NLL). Selective-prediction performance was assessed using error-detection AUROC, selective risk, coverage, and area under the risk–coverage curve (AURC). Higher values indicate better performance for examination-level AUROC, error-detection AUROC, and coverage, whereas lower values indicate better performance for ECE, Brier score, NLL, selective risk, and AURC.

Statistical analysis

We calculated 95% confidence intervals using 2,000 patient-level bootstrap resamples. Differences in AP were assessed using paired permutation tests, and differences in AUROC were assessed using DeLong's test. The Holm–Bonferroni method was used to adjust for multiple comparisons. All tests were two-sided, and $P < 0.05$ was considered statistically significant.

Use of AI-assisted technologies

OpenAI ChatGPT was used to assist with translation from Chinese into English, language editing, structural organization, figure-label standardization, and manuscript formatting.

Results

Comparison with existing methods

Table 2 presents a quantitative comparison of MV-CortexGuard with representative strong baseline models. The full model achieved the best results across all primary evaluation metrics.

Table 2 Fracture detection performance of leading methods.

Method	AP@50 (%)	AP@75 (%)	AP@50–95 (%)	AP_S (%)	Examination-level AUROC

Method	AP@50 (%)	AP@75 (%)	AP@50–95 (%)	AP_S (%)	Examination-level AUROC
YOLOv8-AM	95.6 ± 0.2	65.3 ± 0.4	58.1 ± 0.3	44.5 ± 0.6	0.948
RT-DETR-R50	95.9 ± 0.2	66.4 ± 0.4	59.0 ± 0.4	45.2 ± 0.6	0.953
YOLOv10m	96.4 ± 0.2	67.1 ± 0.4	60.1 ± 0.3	46.1 ± 0.5	0.956
MV-CortexGuard	97.2 ± 0.2	70.8 ± 0.4	63.4 ± 0.3	52.7 ± 0.6	0.969

AP, average precision; AP_S, average precision in the size-defined small-fracture subset; AUROC, area under the receiver operating characteristic curve. AP values are mean ± standard deviation across five random seeds; AUROC values are point estimates in the available study record.

MV-CortexGuard achieved an AP@50–95 of 63.4%, 3.3 percentage points higher than the best-performing comparator, YOLOv10m. Patient-level bootstrap analysis showed that this difference was statistically significant (95% CI, 2.1–4.6 percentage points; $P < 0.001$). At the more stringent AP@75 threshold, our method exceeded YOLOv10m by 3.7 percentage points, indicating improvements in both fracture detection and lesion localization.

The gain was more pronounced in the small-fracture subset. AP_S reached 52.7%, exceeding YOLOv10m and YOLOv8-AM by 6.6 and 8.2 percentage points, respectively. This improvement was larger than that in overall AP@50–95, suggesting that the bone-axis normalization and directional cortical differences introduced by BACD are particularly effective for subtle, low-salience fractures.

Figure 2 further compares the detectors on the full test set and the small-fracture subset. MV-CortexGuard maintained a clear advantage over the reported YOLO-series and transformer-based comparison models on both detection metrics.

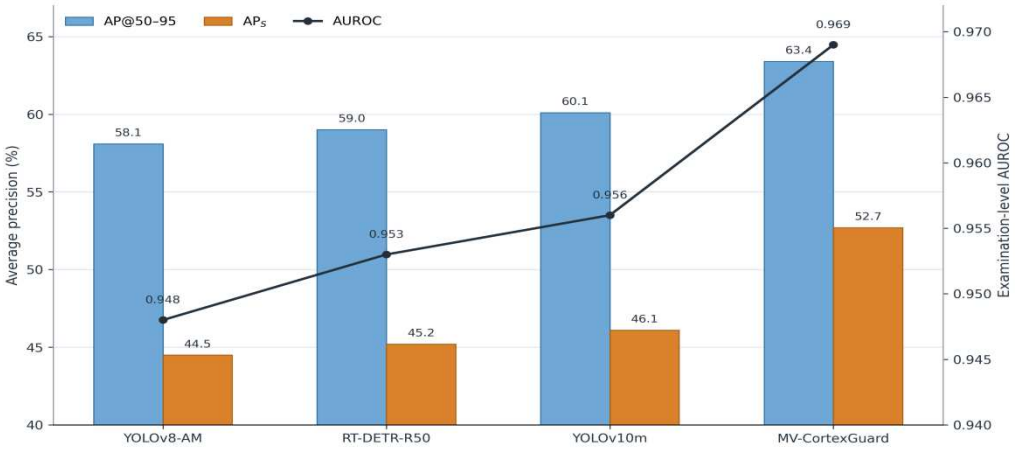


Fig. 2 Fracture detection performance of leading methods. Bars show AP@50–95 and AP_S; the line shows examination-level AUROC. AP_S denotes average precision in the size-defined small-fracture subset.

MV-CortexGuard also demonstrated strong lesion-detection performance at low false-positive rates. At 0.5, 1, and 2 false positives per image, its lesion-level sensitivities were 85.6%, 90.8%, and 94.0%, respectively, compared with 80.5%, 86.9%, and 91.0% for YOLOv10m. These results indicate that the improvement in recall was not achieved by generating large numbers of low-quality candidate boxes; instead, the model detected more subtle fractures while maintaining a low false-positive burden.

For examination-level diagnosis, MV-CortexGuard achieved an AUROC of 0.969, an absolute improvement of 0.013 over YOLOv10m. The DeLong test showed that this difference was statistically significant (95% CI, 0.005–0.022; $P = 0.006$). At the diagnostic threshold selected on the validation set, the sensitivity and specificity of our method were 91.0% and 92.4%, respectively.

Multi-view fusion analysis

To assess the value of multi-view information, we compared six settings: AP/PA-only imaging, LAT-only imaging, late probability averaging, early feature concatenation, standard cross-attention, and ACMQF. The results are shown in Fig. 3.

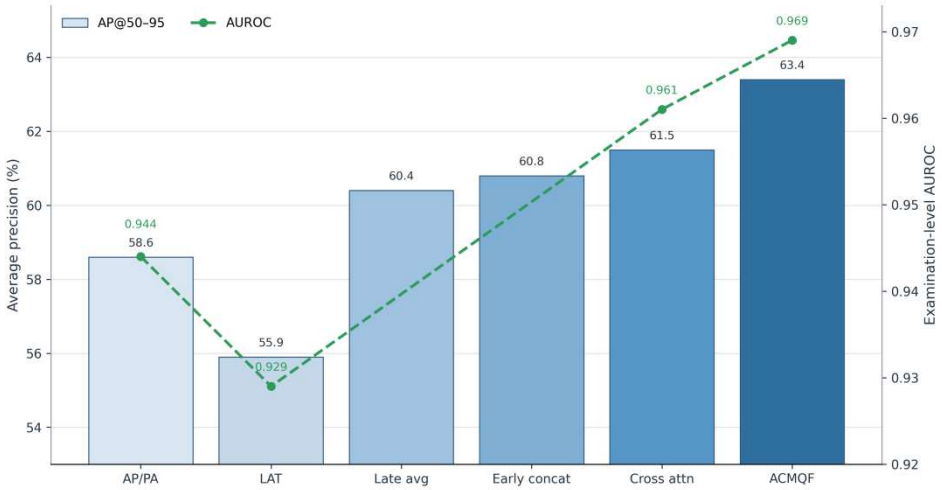


Fig. 3 Performance by projection view and fusion strategy. Bars show AP@50–95, and the dashed line shows the reported examination-level AUROC values. AP/PA denotes frontal-only imaging; LAT, LAT-only imaging; Late avg, late probability averaging; Early concat, early feature concatenation; Cross attn, standard cross-attention; and ACMQF, anatomy-constrained multi-view query fusion.

The AP/PA-only model achieved an AP@50–95 of 58.6% and an AUROC of 0.944, whereas the corresponding values for the

LAT-only model were 55.9% and 0.929. The AP/PA view outperformed the LAT view, indicating that frontal projections provided more discernible fracture information in the current dataset. Nevertheless, both single-view models underperformed the multi-view models, demonstrating diagnostically useful complementarity between the AP/PA and LAT radiographs.

Late probability averaging and early feature fusion achieved AP@50–95 values of 60.4% and 60.8%, respectively. Thus, adding the complementary view improved detection performance, but simple fusion did not fully address the lack of direct spatial correspondence across views. Standard cross-attention further increased AP@50–95 to 61.5%, but remained below the 63.4% achieved by ACMQF.

Compared with standard cross-attention, ACMQF improved AP_S by 3.7 percentage points and increased AUROC from 0.961 to 0.969. These results indicate that candidate appearance similarity alone is insufficient for stable cross-view correspondence. Bone-axis position, bone-region distribution, and evidence reliability can reduce erroneous matches between different anatomical regions.

For fractures with clear signs in only one view, the full ACMQF achieved a sensitivity of 88.9%, 5.6 percentage points higher than that of the model without the null query. The null query allows a candidate to remain unmatched when reliable evidence is unavailable in the complementary view, thereby preventing erroneous fusion from weakening view-specific information.

Representative localization results are shown in Fig. 4. Predicted boxes generally covered the principal suspected fracture regions and substantially overlapped the manual annotations. Minor displacement remained in some cases, indicating that localization can still be improved under complex bone anatomy and projection variation.

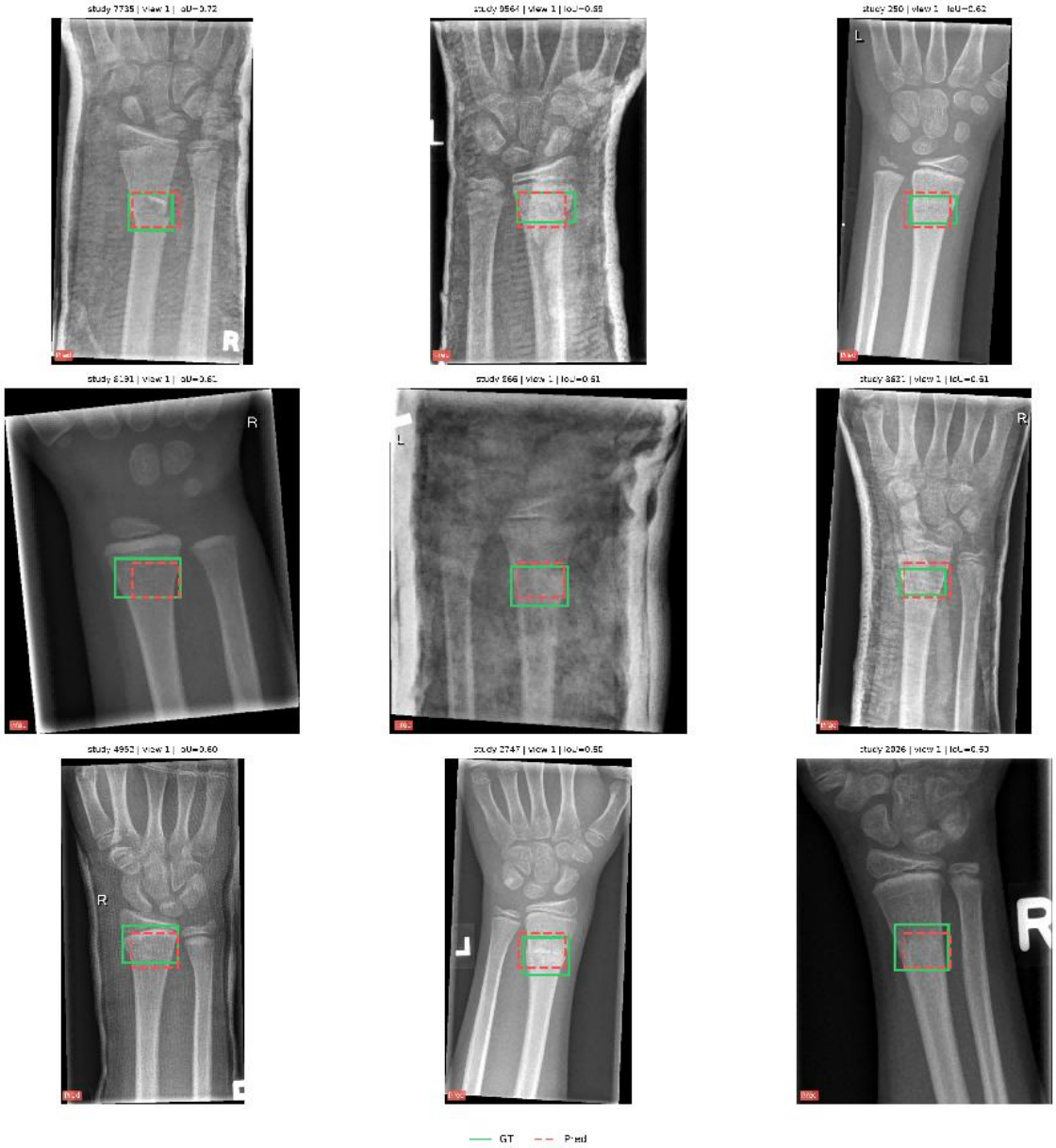


Fig. 4 Representative fracture-localization results. Green boxes indicate ground-truth annotations, and red dashed boxes indicate model predictions. The displayed cases illustrate substantial overlap between predicted and annotated fracture regions, with minor localization offsets in some examinations.

Query-guided activation maps are shown in Fig. 5. Responses were concentrated in suspected fracture regions and surrounding cortical texture rather than being diffusely distributed across the radiograph, and their spatial patterns varied with the underlying anatomy.

Table 3 Ablation of core MV-CortexGuard modules.

BACD	ACMQF	MURR	AP@50–95 (%)	AP_S (%)	Examination-level AUROC	ECE	AURC
			58.4	44.6	0.949	0.058	0.0256

BACD	ACMQF	MURR	AP@50–95 (%)	AP_S (%)	Examination-level AUROC	ECE	AURC
✓			60.7	49.0	0.954	0.052	0.0234
	✓		61.1	48.1	0.960	0.050	0.0215
✓	✓		63.0	52.3	0.966	0.047	0.0196
✓	✓	✓	63.4	52.7	0.969	0.024	0.0107

BACD, Bone-Axis-Aligned Cortical Discontinuity Representation; ACMQF, Anatomy-Constrained Multi-View Query Fusion; MURR, Multi-source Uncertainty and Risk-controlled Rejection; ECE, expected calibration error; AURC, area under the risk–coverage curve.

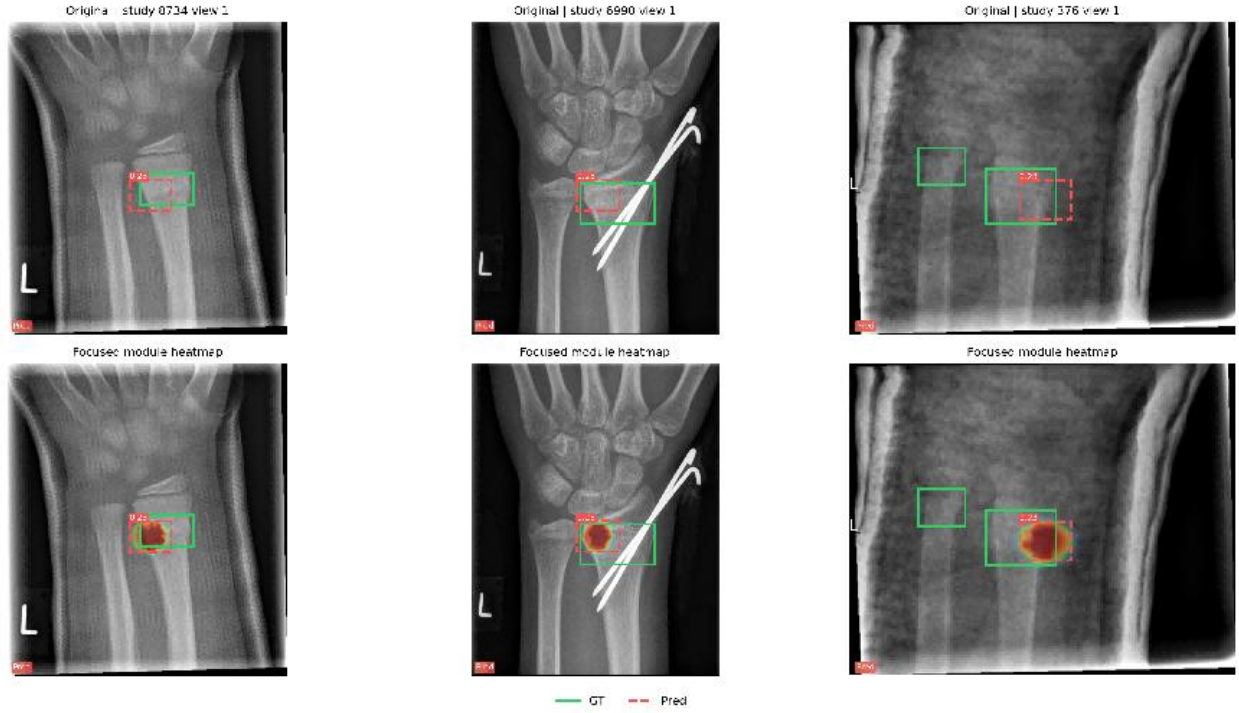


Fig. 5 Query-guided activation heatmaps. The upper row shows the original radiographs, and the lower row shows the corresponding activation maps. Responses concentrate on suspected fracture regions and adjacent cortical texture.

Ablation studies

Module-level ablation

The baseline model comprised a shared encoder, a set-prediction detection head, and late probability fusion. BACD, ACMQF, and MURR were then added sequentially. Quantitative results are presented in Table 3, and the corresponding trends are shown in Fig. 6.

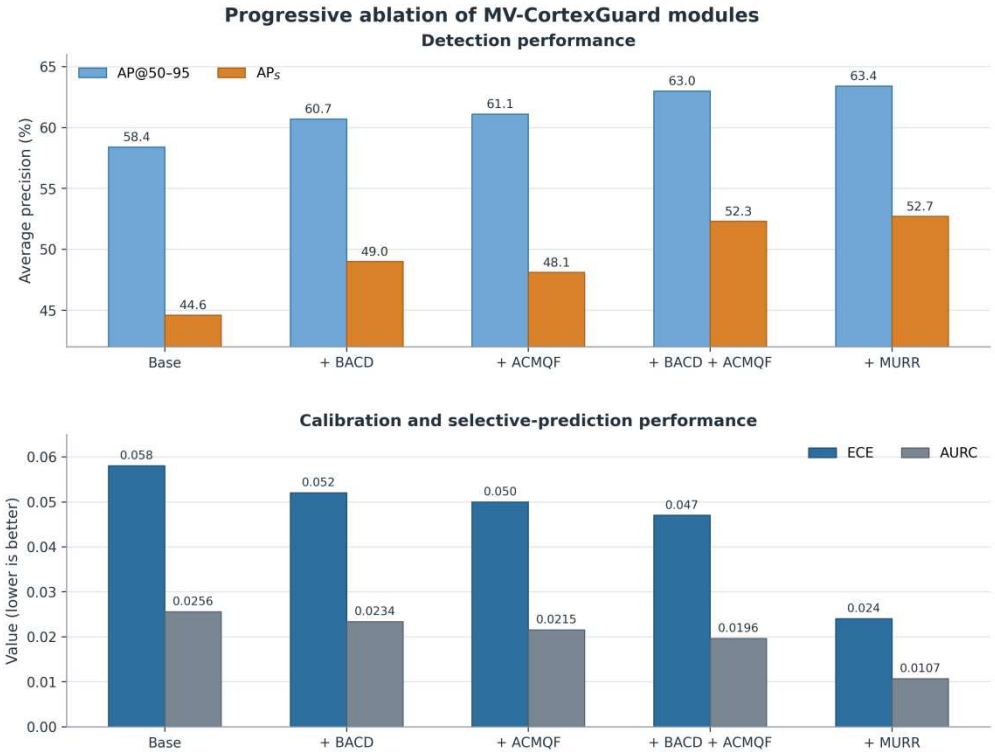


Fig. 6 Progressive ablation of MV-CortexGuard modules. The upper panel shows AP@50–95 and AP_S as BACD, ACMQF, and MURR are added; the lower panel shows ECE and AURC, for which lower values are better.

Adding BACD increased AP@50–95 from 58.4% to 60.7% and AP_S from 44.6% to 49.0%. The gain in AP_S was larger than that in overall AP@50–95, indicating that BACD primarily improved the structural representation of subtle, low-contrast fractures.

Adding ACMQF alone increased AUROC from 0.949 to 0.960 and improved AP@50–95 by 2.7 percentage points, indicating that cross-view candidate interaction improved both lesion localization and examination-level diagnosis. When BACD and ACMQF were used together, AP@50–95 reached 63.0%, exceeding the result obtained with either module alone. This finding suggests that normalized anatomical coordinates provide a more stable positional basis for cross-view query matching.

Adding MURR increased both AP@50–95 and AP_S by only 0.4 percentage points, but reduced ECE from 0.047 to 0.024 and AURC from 0.0196 to 0.0107. This result is consistent with the intended role of MURR: improving probability calibration and the identification of erroneous predictions rather than simply increasing detection accuracy across all cases.

Analysis of key architectural components

Removing bone-axis alignment from the full model reduced AP@50–95 and AP_S by 1.6 and 3.1 percentage points, respectively; removing the directional differences reduced the corresponding metrics by 1.9 and 3.8 percentage points. Directional differences had the largest effect on AP_S, demonstrating that explicit modeling of cortical continuity along and normal to the bone axis is critical for subtle fracture detection.

Removing semantic gating reduced AP@50–95 from 63.4% to 62.2% and markedly increased the number of false positives at the growth plate and epiphyseal margins. Directional-difference responses alone therefore cannot distinguish fractures from normal high-contrast edges; high-level semantic information is needed for selective enhancement.

Within ACMQF, removing the anatomical matching constraint reduced AUROC from 0.969 to 0.960; removing the null query reduced AP_S by 2.6 percentage points; and removing reliability gating increased AURC from 0.0107 to 0.0134. These results support the respective contributions of anatomical constraints, non-forced matching, and reliability modulation.

Uncertainty and risk-control analysis

Probability calibration

We compared MURR with maximum Softmax probability, predictive entropy, temperature scaling, MC Dropout, and Deep Ensemble. MURR achieved an ECE of 0.024, a Brier score of 0.078, and an NLL of 0.249, outperforming the corresponding Deep Ensemble values of 0.028, 0.082, and 0.263.

Figure 7 compares the calibration metrics of Deep ensemble and MURR and shows how error-detection AUROC changed as uncertainty sources were integrated. Relative to Deep ensemble, MURR reduced ECE by 0.004, Brier score by 0.004, and NLL by 0.014.

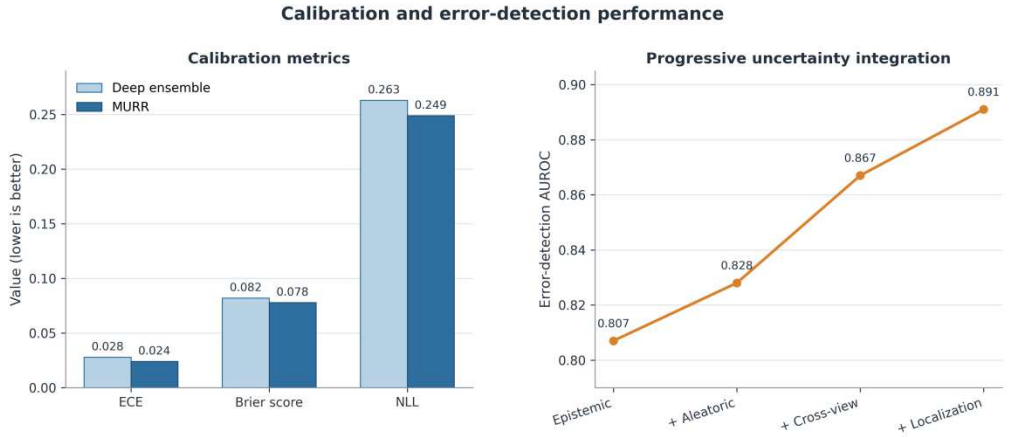


Fig. 7 Calibration and error-detection performance. The left panel compares ECE, Brier score, and NLL between Deep ensemble and MURR; the right panel shows error-detection AUROC as epistemic, aleatoric, cross-view, and localization uncertainty are added sequentially.

Analysis of the individual uncertainty sources showed that error-detection AUROC was 0.807 when epistemic uncertainty alone was used. It increased to 0.828 after adding aleatoric uncertainty, to 0.867 after further adding cross-view disagreement, and to 0.891 in the full model after adding localization uncertainty. These uncertainty sources reflect model-parameter disagreement, image quality, cross-view conflict, and localization variability, respectively, and therefore provide strongly complementary information.

Risk–coverage analysis

Coverage is the proportion of cases accepted for automated processing, whereas selective risk is the error rate among the accepted cases. For the acceptance indicator a_i, these quantities are defined as

$$\text{Coverage} = \frac{1}{n}\sum_{i=1}^{n} a_i\,, \qquad \text{Risk} = \frac{\sum_{i=1}^{n} a_i \mathbb{I}(\hat{y}_i \neq y_i)}{\sum_{i=1}^{n} a_i}. \tag{27}$$

The risk–coverage results for the different uncertainty-estimation methods are shown in Fig. 8.

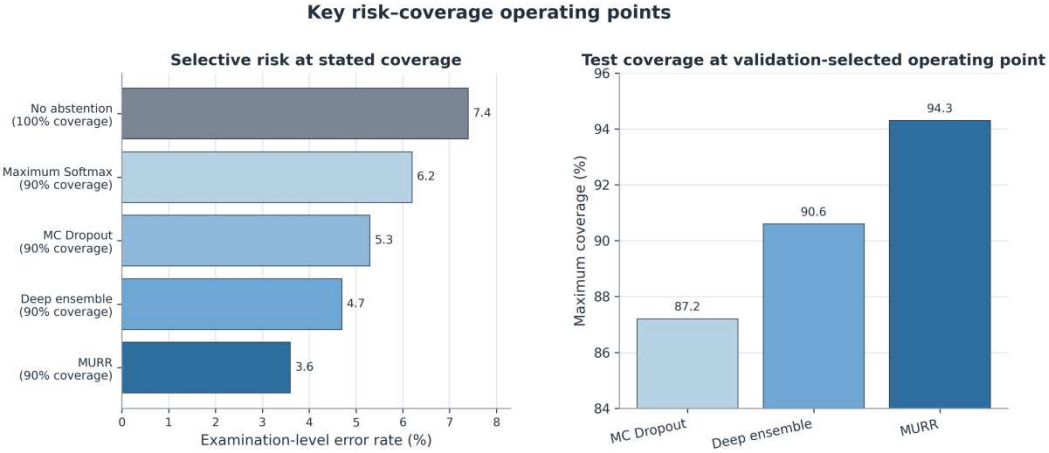


Fig. 8 Key risk–coverage operating points. The left panel distinguishes the no-abstention error rate at 100% coverage from selective risk at 90% coverage; the right panel shows test-set coverage at validation-selected operating points targeting a 5% risk limit.

Without abstention, the examination-level error rate of the full model was 7.4%. At 90% coverage, MURR reduced selective risk to 3.6%, compared with 6.2%, 5.3%, and 4.7% for maximum Softmax probability, MC Dropout, and Deep Ensemble, respectively.

At the operating point selected on the validation set to target a maximum risk of 5%, MURR automatically processed 94.3% of the test cases, whereas the corresponding test-set coverage values for Deep Ensemble and MC Dropout were 90.6% and 87.2%, respectively. These results indicate that MURR identified potential errors more accurately rather than achieving lower risk merely by rejecting more cases.

Robustness and computational efficiency

Image perturbations and missing views

We separately applied rotation, low-contrast, Gaussian-noise, and blur perturbations to the test set and simulated the absence of either the AP/PA or LAT view. The results are shown in Fig. 9.

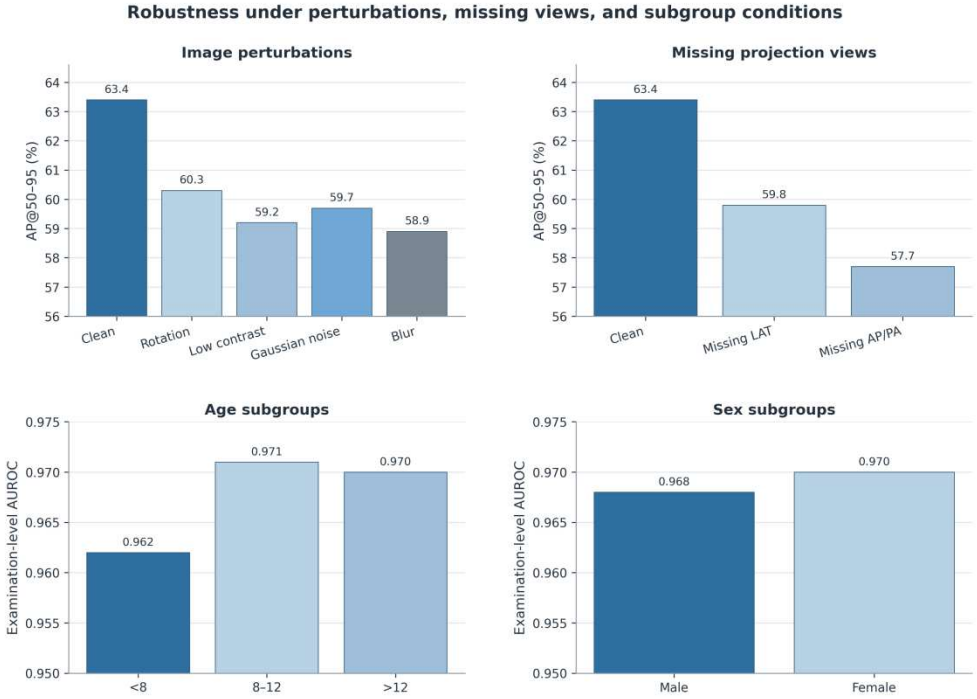


Fig. 9 Robustness under perturbations, missing views, and subgroup conditions. The panels show AP@50–95 under image perturbations, AP@50–95 with a missing view, age-subgroup AUROC, and sex-subgroup AUROC.

Under rotational perturbation, AP@50–95 decreased by 5.5 percentage points for the baseline model but by only 3.1 percentage points for MV-CortexGuard, indicating that bone-axis normalization mitigated the effect of changes in projection pose. Under low-contrast, Gaussian-noise, and blur perturbations, the decreases for our model were 4.2, 3.7, and 4.5 percentage points, respectively, all smaller than the corresponding decreases for the baseline model.

When the LAT radiograph was unavailable, AP@50–95 for MV-CortexGuard decreased from 63.4% to 59.8%; when the AP/PA radiograph was unavailable, it decreased to 57.7%. The larger effect of a missing AP/PA view indicates that frontal radiographs provided more discernible fracture information in the current dataset. Even when one projection view was unavailable, our model outperformed the corresponding single-view baseline, indicating that the view mask and null-query mechanism enabled stable degradation to single-view inference.

Subgroup analysis yielded AUROC values of 0.962, 0.971, and 0.970 for children aged <8 years, 8–12 years, and >12 years, respectively, and AUROC values of 0.968 and 0.970 for male and female patients, respectively. Differences between the subgroups did not reach statistical significance; however, the confidence interval was relatively wide in the youngest group, suggesting that skeletal maturity may affect diagnostic stability.

Computational efficiency

Table 4 summarizes the computational complexity of the models on a single RTX 4090 GPU with a batch size of 1 and an input size of 1024×1024. Inference time was measured over 500 repetitions after 50 warm-up runs.

Table 4 Model complexity and inference efficiency.

Method	Parameters (M)	FLOPs (G)	Single-image latency (ms)	Two-view examination latency (ms)
YOLOv8s	11.2	28.6	7.1	14.0
YOLOv8-AM	43.7	165.1	9.0	17.8
RT-DETR-R50	42.1	136.8	13.5	26.4
YOLOv10m	15.4	59.1	9.4	18.5
MV-CortexGuard	32.6	118.7	11.6	22.4

FLOPs, floating-point operations. Latency was measured on one NVIDIA RTX 4090 GPU with batch size 1 and 1024 × 1024 input images, using 500 repetitions after 50 warm-up runs.

MV-CortexGuard had more parameters and FLOPs than YOLOv8s and YOLOv10m, but fewer than YOLOv8-AM and RT-DETR-R50. Its two-view examination-level inference time was 22.4 ms, corresponding to approximately 44.6 examinations/s, which is sufficient for offline computer-aided diagnosis and near-real-time processing.

Compared with YOLOv10m, MV-CortexGuard increased two-view inference latency by 3.9 ms while improving AP@50–95 and AP_S by 3.3 and 6.6 percentage points, respectively. Because the stochastic adapter heads share the encoder and query decoder, the additional computation introduced by MURR is concentrated mainly in lightweight output layers, and the overall computational overhead remains acceptable.

Discussion

Principal findings

This study developed MV-CortexGuard to address three related challenges in subtle pediatric wrist fracture detection: weak cortical signs, non-rigid correspondence between AP/PA and LAT radiographs, and unreliable confidence estimates in difficult

examinations. On the internal test set of GRAZPEDWRI-DX, MV-CortexGuard achieved an AP@50–95 of 63.4%, an AP_S of 52.7% in the size-defined small-fracture subset, and an examination-level AUROC of 0.969. Compared with the best-performing comparator in this evaluation, YOLOv10m, these values represented absolute improvements of 3.3 percentage points in AP@50–95 and 6.6 percentage points in AP_S; the differences in AP@50–95 and examination-level AUROC were statistically significant. Ablation analyses indicated that BACD and ACMQF accounted for most of the improvement in detection, whereas MURR primarily improved calibration and error-detection performance. Under a selective-risk limit of 5%, MURR achieved 94.3% test-set coverage.

The larger gain in AP_S than in overall AP@50–95 is particularly relevant to the stated objective of detecting low-salience fractures. Conventional fracture detectors have generally improved performance through stronger backbones, feature pyramids, or generic attention mechanisms[10–22, 26]. These strategies are useful for multi-scale feature extraction but do not explicitly encode whether a local edge remains anatomically continuous along a bone. BACD instead estimates the bone axis, maps features to normalized anatomical coordinates, and calculates directional cortical differences. Adding BACD to the baseline increased AP@50–95 from 58.4% to 60.7% and AP_S from 44.6% to 49.0%. In the full model, removing bone-axis alignment reduced AP_S by 3.1 percentage points, and removing directional differences reduced it by 3.8 percentage points. These effects are consistent with the premise that the morphology of a subtle fracture is better characterized as a localized interruption of cortical continuity than as an unconstrained salient region. They also extend general geometric-normalization approaches such as spatial transformers and equivariant convolutions[23–25] by coupling pose normalization to a task-specific structural response.

Semantic gating was also important because strong edge responses are not specific to fractures in pediatric radiographs. The physis, epiphyseal margins, ossification centers, and overlapping bones may all produce conspicuous local boundaries. Removing semantic gating reduced AP@50–95 from 63.4% to 62.2% and increased false-positive detections at growth plates and epiphyseal margins. Thus, directional cortical responses should not be interpreted in isolation. Their combination with higher-level semantic information appears to suppress anatomically plausible hard negatives while preserving focal cortical abnormalities. The activation maps provided qualitative support for localization of model attention around suspected fracture regions and adjacent cortex; however, such heatmaps do not establish causal interpretability and should be considered descriptive rather than explanatory evidence.

The multi-view experiments showed that AP/PA and LAT radiographs provided complementary information, with higher observed performance for the AP/PA-only model in this dataset. The AP/PA-only model achieved an AP@50–95 of 58.6% and an examination-level AUROC of 0.944, whereas the LAT-only model achieved 55.9% and 0.929, respectively. Late probability averaging, early feature concatenation, and standard cross-attention improved upon single-view processing, but none matched ACMQF. Compared with standard cross-attention, ACMQF increased AP_S by 3.7 percentage points and examination-level AUROC from 0.961 to 0.969. This pattern supports candidate-level fusion constrained by normalized longitudinal position, osseous-region distribution, and evidence reliability, rather than dense fusion that assumes direct spatial correspondence. Similar benefits of structured cross-view learning have been reported in other multi-projection medical-imaging settings[30–34], but lesion detection requires preservation of local evidence as well as examination-level aggregation.

The null-query mechanism addressed an additional practical feature of paired radiographs: a genuine fracture sign may be discernible in only one projection. For cases with clear findings in a single view, ACMQF achieved a sensitivity of 88.9%, 5.6 percentage points higher than the model without the null query. This result suggests that allowing a candidate to remain unmatched is preferable to forcing correspondence with an unreliable feature in the complementary projection. The missing-view experiments were consistent with this interpretation. AP@50–95 decreased from 63.4% to 59.8% when the LAT image was unavailable and to 57.7% when the AP/PA image was unavailable, yet the model still outperformed the corresponding single-view baselines. The larger decrease after removal of the AP/PA view was also concordant with the stronger AP/PA-only results. These findings indicate graceful degradation under the simulated missing-view conditions, although they do not establish performance under all forms of incomplete or non-standard clinical imaging.

MURR contributed differently from the two representation modules. Adding MURR to BACD and ACMQF increased AP@50–95 and AP_S by only 0.4 percentage points each, but reduced ECE from 0.047 to 0.024 and AURC from 0.0196 to 0.0107. The full uncertainty representation achieved an error-detection AUROC of 0.891, compared with 0.807 when epistemic uncertainty alone was used. Successive addition of aleatoric uncertainty, cross-view disagreement, and localization instability improved error ranking, supporting the view that examination-level failure cannot be represented adequately by class confidence alone. This result is aligned with prior work showing that modern neural networks may be overconfident[35–39] and that selective prediction can trade automated coverage for lower error among accepted cases[40–48]. Importantly, the 94.3% coverage observed at a 5% risk limit is an operating point derived from the study validation set and evaluated on the internal test set; it should not be interpreted as a universal clinical safety guarantee.

The risk–coverage analysis illustrates how uncertainty estimates could be translated into a human-review pathway. Without abstention, the examination-level error rate was 7.4%. At 90% coverage, MURR reduced selective risk to 3.6%, compared with 4.7% for Deep Ensemble, 5.3% for Monte Carlo (MC) dropout, and 6.2% for maximum Softmax probability. This advantage was not achieved merely by rejecting a larger fraction of examinations, because MURR also provided the highest coverage among the compared methods when risk was constrained to 5%. In a future clinical workflow, such a mechanism might support automatic processing of lower-risk examinations while prioritizing higher-risk examinations for expert review. Its actual clinical value will depend on how risk is defined, the performance and availability of the reviewing expert, the cost of deferral, and the prevalence and case mix of the deployment setting. These factors were not evaluated in the present study.

The perturbation experiments provide preliminary evidence that the anatomical design was less sensitive than the baseline to changes in image appearance. Under rotational perturbation, AP@50–95 decreased by 3.1 percentage points for MV-CortexGuard and by 5.5 percentage points for the baseline. Decreases under low contrast, Gaussian noise, and blur were also smaller for MV-CortexGuard than for the baseline. Age- and sex-stratified examination-level AUROCs were similar, although the confidence interval was wider in children younger than 8 years. These analyses are useful stress tests but are not substitutes for external validation. From a computational perspective, MV-CortexGuard required 32.6 million parameters and 118.7 GFLOPs, with a two-view examination latency of 22.4 ms on an RTX 4090 GPU at the specified experimental settings. This latency was 3.9 ms greater than that of YOLOv10m and accompanied the 3.3- and 6.6-percentage-point gains in AP@50–95 and AP_S, respectively. The result suggests that the additional modules offer a potentially acceptable research trade-off between accuracy and computation; prospective workflow evaluation is required before drawing conclusions about real-world throughput.

Limitations

This study has several limitations. First, development and evaluation were performed on one public dataset, using a patient-level training, validation, and internal test split. The findings therefore do not establish generalizability across institutions, radiographic equipment, acquisition protocols, demographic distributions, or fracture-prevalence settings. External sites should be held out from model development; any site-specific recalibration should use a separate calibration subset rather than

the external test set. Second, the size-defined small-fracture subset was constructed using the 25th percentile of normalized fracture-box area in the training set. This reproducible definition measures lesion size, but it is only a proxy for perceptual subtlety and may not capture fractures that are difficult because of low contrast, overlap, orientation, or developmental anatomy. The threshold may also be dataset dependent.

Third, the validation set was used for model selection, probability calibration, and determination of diagnostic and rejection thresholds; no separate external calibration cohort was available. Although the test set was reserved for final evaluation, future studies should assess calibration transportability and prespecify recalibration and threshold-selection procedures for each deployment setting. Fourth, missing projections and image degradations were simulated and may not reproduce the full range of motion, positioning, collimation, exposure, artifacts, or workflow failures encountered in routine practice. The present multi-view analysis focused on AP/PA and LAT radiographs; performance with oblique or other non-standard projections remains untested. Fifth, subgroup analyses were limited to age and sex, and the wider confidence interval in the youngest group indicates limited precision. More comprehensive fairness analyses require adequately powered samples and clinically relevant subgroup information.

Finally, this was a model-development study without a prospective reader study, workflow integration, or evaluation of patient outcomes. The reported selective-risk operating points assume that deferred examinations receive appropriate expert review and do not quantify reporting delay, workload redistribution, or downstream clinical consequences. The study relied on the dataset's existing annotations and did not report an independent re-adjudication of difficult cases. Qualitative bounding-box examples and activation maps illustrate model behavior but do not validate causal interpretability. External validation, blinded comparisons with clinicians, prospective assessment of human–AI collaboration, and evaluation of predefined risk-control policies are therefore necessary before clinical use.

Conclusions

MV-CortexGuard combines bone-axis-aligned cortical discontinuity representation, anatomy-constrained multi-view query fusion, and multi-source uncertainty-based abstention for subtle pediatric wrist fracture detection. On the internal GRAZPEDWRI-DX test set, the model improved overall and size-defined small-fracture detection relative to the best-performing comparator in this evaluation and achieved an examination-level AUROC of 0.969. Its uncertainty module improved calibration, error-detection, and risk–coverage performance, achieving 94.3% test-set coverage under a 5% selective-risk constraint. These results support anatomically informed multi-view modeling and selective prediction as complementary components of computer-aided fracture detection. However, the reported thresholds and performance estimates are specific to this study setting. Clinical utility remains to be established. Multicenter external validation, site-specific calibration, prospective reader studies, and workflow evaluation are required before clinical implementation.

List of abbreviations

ACMQF, Anatomy-Constrained Multi-View Query Fusion; AP, average precision; AP/PA, anteroposterior/posteroanterior; AP_S, average precision for the size-defined small-fracture subset; AURC, area under the risk–coverage curve; AUROC, area under the receiver operating characteristic curve; BACD, Bone-Axis-Aligned Cortical Discontinuity Representation; BCE, binary cross-entropy; CI, confidence interval; ECE, expected calibration error; FLOPs, floating-point operations; GIoU, generalized intersection over union; IoU, intersection over union; JS, Jensen–Shannon divergence; LAT, lateral; MC, Monte Carlo; MURR, Multi-source Uncertainty and Risk-controlled Rejection; NLL, negative log-likelihood; ROC, receiver operating characteristic; STN, spatial transformer network.

Declarations

Ethics approval and consent to participate

This study was a methodological validation conducted exclusively using the pub licly available GRAZPEDWRI-DX dataset. All radiographs and associated anno tations were de-identified before public release. No participants were recruited, n o interventions were performed, and the authors had no access to identifiable per sonal information. Therefore, no additional ethics approval or informed consent was required for this secondary analysis of anonymized, publicly available data.

Availability of data and materials

The data analyzed in this study were obtained from the publicly available GRAZ PEDWRI-DX dataset [8], hosted on Figshare at https://doi.org/10.6084/m9.figsh are.14825193. The dataset is freely accessible without user registration under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

Funding

None

Authors' contributions

SL and HM contributed equally to the conception and design of the study, metho dological development, experimental implementation, data analysis, and prepara tion of the initial manuscript. CZ and MS jointly supervised the study, contribute d to the interpretation of the results, and critically revised the manuscript. All aut hors read and approved the final manuscript.

References

[1] Olczak J, Fahlberg N, Maki A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs. Acta Orthop. 2017;88(6):581–586.

[2] Lindsey R, Daluiski A, Chopra S, et al. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci U S A. 2018;115(45):11591–11596.

[3] Thian YL, Li Y, Jagmohan P, et al. Convolutional neural networks for automated fracture detection and localization on wrist radiographs. Radiol Artif Intell. 2019;1(1):e180001.

[4] Gan K, Xu D, Lin Y, et al. Artificial intelligence detection of distal radius fractures: A comparison between the convolutional neural network and professional assessments. Acta Orthop. 2019;90(4):394–400.

[5] Blüthgen C, Becker AS, de Martini IV, et al. Detection and localization of distal radius fractures: Deep learning system versus radiologists. Eur J Radiol. 2020;126:108925.

[6] Raisuddin AM, Vaattovaara E, Nevalainen M, et al. Critical evaluation of deep neural networks for wrist fracture detection. Sci Rep. 2021;11:6006.

[7] Hardalaç F, Uysal F, Peker O, et al. Fracture detection in wrist X-ray images using deep learning-based object detection models. Sensors (Basel). 2022;22(3):1285.

[8] Nagy E, Janisch M, Hržić F, et al. A pediatric wrist trauma X-ray dataset (GRAZPEDWRI-DX) for machine learning. Sci Data. 2022;9:222.

[9] Abedeen I, Rahman MA, Prottyasha FZ, et al. FracAtlas: A dataset for fracture classification, localization and segmentation of musculoskeletal radiographs. Sci Data. 2023;10:521.

[10] Ju RY, Cai W. Fracture detection in pediatric wrist trauma X-ray images using YOLOv8 algorithm. Sci Rep. 2023;13:20077.

[11] Till T, Tschauner S, Singer G, et al. Development and optimization of AI algorithms for wrist fracture detection in children using a freely available dataset. Front Pediatr. 2023;11:1291804.

[12] Chen P, Liu S, Lu W, et al. WCAY object detection of fractures for X-ray images of multiple sites. Sci Rep. 2024;14:26702.

[13] Chien CT, Ju RY, Chou KY, et al. YOLOv8-AM: YOLOv8 based on effective attention mechanisms for pediatric wrist fracture detection. IEEE Access. 2025;13:52461–52477.

[14] Liu D, Yang Z, Bao C, et al. Artificial intelligence-based method for detecting wrist fractures in children. Sci Rep. 2025;15:38555.

[15] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems. 2015;28:91–99.

[16] Lin TY, Dollár P, Girshick R, et al. Feature pyramid networks for object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:2117–2125.

[17] Lin TY, Goyal P, Girshick R, et al. Focal loss for dense object detection. Proceedings of the IEEE International Conference on Computer Vision. 2017:2980–2988.

[18] Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks. Proceedings of the IEEE International Conference on Computer Vision. 2017:764–773.

[19] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with Transformers. European Conference on Computer Vision. 2020:213–229.

[20] Zhu X, Su W, Lu L, et al. Deformable DETR: Deformable Transformers for end-to-end object detection. International Conference on Learning Representations. 2021.

[21] Yang Z, Liu S, Hu H, et al. QueryDet: Cascaded sparse query for accelerating high-resolution small object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:13668–13677.

[22] Zhao Y, Lv W, Xu S, et al. DETRs beat YOLOs on real-time object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024:16965–16974.

[23] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial Transformer networks. Advances in Neural Information Processing Systems. 2015;28:2017–2025.

[24] Cohen TS, Welling M. Group equivariant convolutional networks. Proceedings of the International Conference on Machine Learning. 2016;48:2990–2999.

[25] Weiler M, Cesa G. General E(2)-equivariant steerable CNNs. Advances in Neural Information Processing Systems. 2019;32:14334–14345.

[26] Woo S, Park J, Lee JY, et al. CBAM: Convolutional block attention module. European Conference on Computer Vision. 2018:3–19.

[27] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in Neural Information Processing Systems. 2017;30:5998–6008.

[28] Wang X, Girshick R, Gupta A, et al. Non-local neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:7794–

7803.
[29] Su H, Maji S, Kalogerakis E, et al. Multi-view convolutional neural networks for 3D shape recognition. Proceedings of the IEEE International Conference on Computer Vision. 2015:945–953.
[30] Carneiro G, Nascimento JC, Bradley AP. Automated analysis of unregistered multi-view mammograms with deep learning. IEEE Trans Med Imaging. 2017;36(11):2355–2365.
[31] Van Tulder G, Tong Y, Marchiori E. Multi-view analysis of unregistered medical images using cross-view Transformers. Medical Image Computing and Computer Assisted Intervention. 2021:104–113.
[32] Chen X, Zhang K, Abdoli N, et al. Transformers improve breast cancer diagnosis from unregistered multi-view mammograms. Diagnostics (Basel). 2022;12(7):1549.
[33] Zhou B, Krähenbühl P. Cross-view Transformers for real-time map-view semantic segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:13760–13769.
[34] Manigrasso F, Milazzo R, Russo AS, et al. Mammography classification with multi-view deep learning techniques: Investigating graph and Transformer-based architectures. Med Image Anal. 2025;99:103320.
[35] Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. Proceedings of the International Conference on Machine Learning. 2016;48:1050–1059.
[36] Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems. 2017;30:6402–6413.
[37] Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? Advances in Neural Information Processing Systems. 2017;30:5574–5584.
[38] Guo C, Pleiss G, Sun Y, et al. On calibration of modern neural networks. Proceedings of the International Conference on Machine Learning. 2017;70:1321–1330.
[39] Ovadia Y, Fertig E, Ren J, et al. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. Advances in Neural Information Processing Systems. 2019;32.
[40] Geifman Y, El-Yaniv R. Selective classification for deep neural networks. Advances in Neural Information Processing Systems. 2017;30:4878–4887.
[41] Geifman Y, El-Yaniv R. SelectiveNet: A deep neural network with an integrated reject option. Proceedings of the International Conference on Machine Learning. 2019;97:2151–2159.
[42] Corbière C, Thome N, Bar-Hen A, et al. Addressing failure prediction by learning model confidence. Advances in Neural Information Processing Systems. 2019;32:2902–2913.
[43] Mozannar H, Sontag D. Consistent estimators for learning to defer to an expert. Proceedings of the International Conference on Machine Learning. 2020;119:7076–7087.
[44] Mehrtash A, Wells WM, Tempany CM, et al. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE Trans Med Imaging. 2020;39(12):3868–3878.
[45] Nair T, Precup D, Arnold DL, et al. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. Med Image Anal. 2020;59:101557.
[46] Kompa B, Snoek J, Beam AL. Second opinion needed: Communicating uncertainty in medical machine learning. NPJ Digit Med. 2021;4:4.
[47] Bates S, Angelopoulos AN, Lei L, et al. Distribution-free, risk-controlling prediction sets. J ACM. 2021;68(6):1–34.
[48] Huang L, Ruan S, Xing Y, et al. A review of uncertainty quantification in medical image analysis: Probabilistic and non-probabilistic methods. Med Image Anal. 2024;97:103223.
[49] Janisch M, Apfaltrer G, Hržić F, et al. Pediatric radius torus fractures in X-rays—How computer vision could render lateral projections obsolete. Front Pediatr. 2022;10:1005099.
[50] Cohen M, Puntonet J, Sanchez J, et al. Artificial intelligence vs. radiologist: Accuracy of wrist fracture detection on radiographs. Eur Radiol. 2023;33(6):3974–3983.