BEYOND CARTESIAN SCANNING: UNIFYING CONFLICT AND GEOMETRY IN STATE MODELING FOR FINE-GRAINED MEDICAL IMAGE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-grained medical image classification depends on subtle local variations whose diagnostic significance is closely related to surrounding anatomical context. Effective recognition therefore requires refining weak discriminative features during contextual integration and organizing surrounding evidence through meaningful structural relationships. Although Vision Mamba efficiently captures long-range dependencies, global connectivity alone does not explicitly address these requirements. We propose Conflict Recalibration and Geometric Aggregation (CRGA), a framework that enhances pretrained VMamba through directional-response-guided feature refinement and circular–radial context aggregation. Specifically, Conflict-Conditioned Recalibration (CCR) compares spatially aligned horizontal and vertical scan responses before fusion and uses their discrepancies as conditioning information for subsequent representation refinement. Through a one-hop connection, this signal modulates selective state writing and readout in the next stage, enabling contextual integration to account explicitly for relationships between directional responses. To organize surrounding structural context, Geometric Circular–Radial Structured Aggregation (GCRSA) introduces a geometric prior inspired by angular–radial formulations in medical imaging. It organizes existing tokens into closed circumferential and open radial sequences around the feature-grid center, distinguishing contextual interactions within radial bands from those across radial distances. Both components retain the original feature grid and Cartesian scan order, while zero-initialized modulation projections and residual gates preserve the pretrained backbone mapping at initialization. Experiments on a clinically acquired EUS dataset show that CRGA improves VMamba accuracy by 2.2 percentage points, while evaluations on two public medical image benchmarks demonstrate competitive performance. Same-backbone ablations on EUS show accuracy gains from each component, with the combined model achieving the highest accuracy and F1 among the evaluated variants.

1 INTRODUCTION

Fine-grained medical image classification aims to distinguish disease subtypes with highly similar visual appearances but potentially different pathological implications and clinical management strategies Park & Ryu (2024). Unlike conventional image classification, discriminative evidence is often confined to subtle, localized variations in texture, boundaries, or morphology, which can be further obscured by imaging noise, intra-class variability, and lesion heterogeneity. Reliable interpretation therefore requires not only identifying local abnormalities but also evaluating them within the global image context and surrounding anatomical structures Agu et al. (2021). This requirement is particularly important for layered or center-oriented anatomical patterns, where lesion characteristics are closely related to the involved tissue layers, site of origin, and extent of spread into adjacent regions Peng et al. (2019). Consequently, effective fine-grained recognition requires jointly preserving subtle local evidence and modeling long-range, anatomically structured dependencies.

Convolutional neural networks (CNNs) provide strong local inductive biases for extracting texture, boundary, and morphological cues in medical images Ronneberger et al. (2015), but linking these

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.

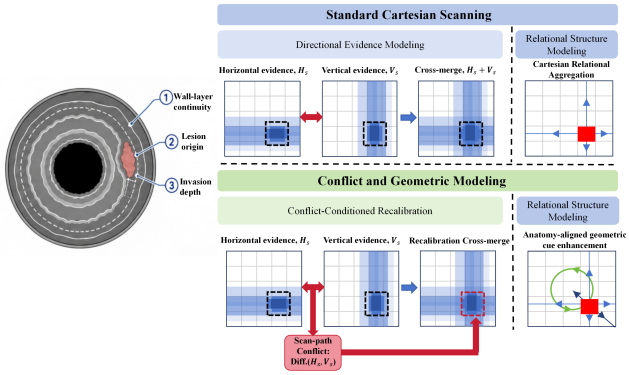


Figure 1: Motivation of CRGA. Standard VMamba does not explicitly exploit cross-axis disagreement or encode circular–radial relationships. CCR uses directional conflict to condition state writing and readout in the next stage, while GCRSA introduces circular–radial aggregation to model circumferential continuity and center-to-periphery context.

cues to distant anatomical structures generally requires deep stacking or enlarged receptive fields Luo et al. (2016); Wang et al. (2018). Vision Transformers enable content-dependent global interactions through self-attention Dosovitskiy et al. (2020); Agu et al. (2021). Beyond global connectivity, fine-grained recognition requires preserving weak class-specific cues during contextual integration and modeling their anatomical relationships. Selective state space models offer input-dependent state writing and readout for long-range contextual modeling Gu & Dao (2023). VMamba extends this formulation to images through multidirectional two-dimensional selective scanning, producing contextual responses associated with different traversal orders Liu et al. (2024b). As illustrated in Fig. 1, two limitations remain. First, standard Cross-Merge combines horizontal and vertical contextual responses into a unified feature representation, while their cross-axis discrepancies are not explicitly carried forward to guide subsequent contextual integration. Prior work in medical segmentation has used inter-rater disagreement to guide spatial feature refinement Ji et al. (2021). Inspired by this use of disagreement, we investigate scan-response discrepancies as conditioning information for subsequent state writing and readout. Second, Cartesian scan orders provide long-range connectivity but do not explicitly separate interactions within radial bands from those across radial distances. Angular–radial organization has been explored in radial ultrasound acquisition and in polar-coordinate image reconstruction for CT and radial MRI Peng et al. (2019); Kim et al. (2013); Rastegar Jooybari et al. (2026). Inspired by these formulations, we abstract this angular–radial organization into two types of contextual relationships on the feature grid: interactions within radial bands and interactions across radial distances. These relationships provide an additional organization for aggregating VMamba features while retaining the backbone's Cartesian selective-scan order. Together, these observations motivate incorporating cross-axis response relationships into selective state computation and distinguishing circular–radial relationships during contextual aggregation.

To address these issues, we propose CRGA, which extends pretrained VMamba through cross-directional conditioning of state computation and circular–radial organization of contextual aggregation. Conflict-Conditioned Recalibration (CCR) characterizes directional conflict as the discrepancy between spatially aligned horizontal and vertical scan responses before Cross-Merge. Through a one-hop connection, the resulting representation conditions the input-to-state mapping B and state-to-output mapping C in the next stage, explicitly connecting directional response comparison with subsequent state writing and readout. Geometric Circular–Radial Structured Aggregation (GCRSA) operates on the Cross-Merged features and organizes existing tokens into closed circumferential and open radial sequences, using the feature-grid center as a fixed geometric reference. The two branches aggregate context within radial bands and across radial distances, respectively, and their outputs are integrated through channel-wise residual gates. This construction separates the geometric relations used for aggregation from the traversal order used by selective scanning. Polar coordinates specify token grouping and ordering without polar resampling or additional visual tokens. Both components retain the original Cartesian scan order, while zero-initialized modulation projections and residual gates preserve the pretrained backbone mapping at initialization.

Our contributions are summarized as follows:

- We propose CRGA for fine-grained medical image classification and develop CCR to connect cross-axis response comparison with subsequent contextual integration. CCR extracts discrepancies before directional fusion and transfers the resulting signal to the next stage to condition selective state writing and readout.
- We introduce GCRSA to organize contextual aggregation through closed circumferential and open radial sequences over existing tokens. The module models within-band and cross-radius relationships directly on the original feature grid, without polar resampling or changes to the backbone's Cartesian scan order.
- We evaluate CRGA on an in-house EUS dataset and two public medical image benchmarks. CRGA improves classification accuracy over VMamba by 2.2 percentage points on EUS. Same-backbone ablations on EUS further assess the contributions of cross-axis conditioning and circular–radial aggregation.

2 METHOD

As illustrated in Fig. 2, CRGA integrates Conflict-Conditioned Recalibration (CCR) and Geometric Circular–Radial Structured Aggregation (GCRSA) into a four-stage pretrained VMamba backbone. CCR guides feature refinement through cross-axis state conditioning, while GCRSA organizes surrounding context through circular–radial aggregation.

2.1 PRELIMINARIES

2.1.1 SELECTIVE STATE SPACE MODEL

State space models (SSMs) encode historical information into latent states, enabling long-range sequence modeling with linear computational complexity. Given a continuous-time input $x(t)$, latent state $h(t)$, and output $y(t)$, a linear SSM is formulated as

$$\dot{h}(t) = Ah(t) + Bx(t), y(t) = Ch(t) + Dx(t), \tag{1}$$

where A governs state evolution, B and C control input writing and state readout, respectively, and D provides a direct input-to-output mapping.

Mamba conditions the discretization step Δ_t and mappings B_t and C_t on the current input x_t, while A and D remain input-independent Gu & Dao (2023). The resulting discrete recurrence is

$$h_t = \overline{A}tht - 1 + \overline{B}_t x_t, \qquad y_t = C_t h_t + Dx_t. \tag{2}$$

Under zero-order-hold discretization, the parameters are given by

$$\overline{A}_t = \exp\left(\Delta_t A\right), \overline{B}_t = \left(\Delta_t A\right)^{-1} \left[\exp\left(\Delta_t A\right) - I\right] \Delta_t B_t. \tag{3}$$

where I denotes the identity matrix. In practice, VMamba uses the first-order approximation $\overline{B}_t \approx \Delta_t B_t$ for the input term Liu et al. (2024b).

2.1.2 2D SELECTIVE SCANNING AND DIRECTIONAL FUSION IN VMAMBA

VMamba extends selective state space modeling to images through 2D Selective Scan (SS2D), comprising Cross-Scan, Selective Scan, and Cross-Merge Liu et al. (2024b). For a VSS Block at stage s, let $X_s \in \mathbb{R}^{B\times H_s\times W_s\times C_s}$ denote the feature map entering Cross-Scan. The transformation $\mathcal{T}_d$ unfolds this map along direction $d \in \mathcal{D} = h^+, h^-, v^+, v^-$, corresponding to horizontal and vertical traversal orders and their reverses:

$$X_s^d = \mathcal{T}_d(X_s), \qquad Y_s^d = \mathrm{SelectiveScan}(X_s^d),\ \widehat{Y}_s^d \quad = \mathcal{T}_d^{-1}(Y_s^d), \qquad d \in \mathcal{D}, \tag{4}$$

where $\mathcal{T}_d^{-1}$ restores each directional output to the original spatial coordinates.

The aligned responses can be grouped by axis, and standard Cross-Merge is expressed as

$$\mathbf{H}_s = \widehat{Y}_s^{h^+} + \widehat{Y}_s^{h^-}, \qquad \mathbf{V}_s = \widehat{Y}_s^{v^+} + \widehat{Y}_s^{v^-},\ F_s^{\mathrm{merge}} \quad = \mathbf{H}_s + \mathbf{V}_s. \tag{5}$$

Standard Cross-Merge aggregates directional context without retaining cross-axis discrepancies as a separate conditioning signal. We use unmodified VMamba with standard SS2D and Cross-Merge as the baseline.

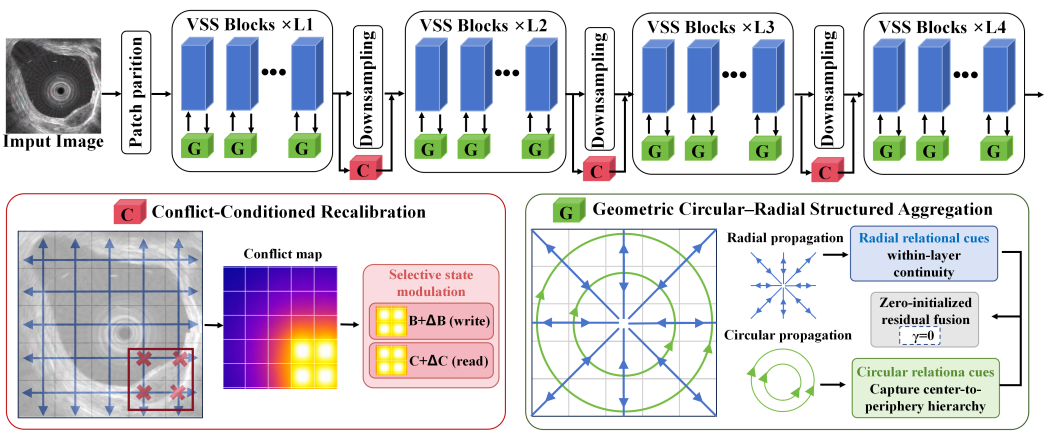


Figure 2: Overall architecture of CRGA on a four-stage VMamba backbone. CCR derives a cross-axis conflict map from spatially aligned directional responses in the last VSS block of stage s and uses it to modulate B and C in the first VSS block of stage $s+1$. GCRSA organizes Cross-Merged tokens into circular and radially ordered groups for aggregation, followed by gated residual fusion. Both modules preserve the original feature grid and the backbone's selective scan order.

2.2 Conflict-Conditioned Recalibration

Directional conflict denotes the relative variation between spatially aligned horizontal and vertical contextual responses. Inspired by the use of inter-rater disagreement for feature refinement in medical segmentation Ji et al. (2021), CCR extracts this signal before Cross-Merge and transfers it through a single-hop connection to modulate B and C in the next stage. This conditions state writing and readout on cross-axis variation to refine subtle local cues within their context. Conflict information enters solely through parameter modulation, preserving the original feature grid, token count, and selective-scan order.

2.2.1 Scan-Path Conflict Evidence Modeling

For the s-th stage, we collect the four directional selective-scan outputs from the last VSS Block before Cross-Merge and restore them to a shared two-dimensional coordinate system. The bidirectional horizontal and vertical evidence are defined as

$$\begin{aligned} H_s &= \widehat{Y}_s^{h^+} + \widehat{Y}_s^{h^-}, V_s = \widehat{Y}_s^{v^+} + \widehat{Y}_s^{v^-}. \\ G_s^{\mathrm{mag}} = \mathrm{Mean}_c(|H_s - V_s|), G_s^{\mathrm{rel}} &= \frac{G_s^{\mathrm{mag}}}{\mathrm{Mean}_c(|H_s|) + \mathrm{Mean}_c(|V_s|) + \varepsilon}, \\ G_s^{\cos} &= 1 - \mathrm{CosSim}_c(H_s, V_s). \end{aligned} \tag{6}$$

The three descriptors capture complementary aspects of variation across directional scan evidence: absolute response discrepancy, scale-normalized discrepancy, and disagreement in channel-wise feature direction. To reduce distributional differences among them, each descriptor is standardized over the spatial dimensions of the corresponding sample. They are then concatenated and mapped to a single-channel conflict response:

$$\begin{aligned} Q_s &= \left[\mathrm{SNorm}(G_s^{\mathrm{mag}}), \mathrm{SNorm}\left(G_s^{\mathrm{rel}}\right), \mathrm{SNorm}(G_s^{\cos})\right], \\ M_s &= \sigma(f_s(Q_s)), M_s \in [0,1]^{B\times 1\times H_s\times W_s}. \end{aligned} \tag{7}$$

Here, $\mathrm{SNorm}(\cdot)$ denotes per-sample spatial standardization, and f_s is a stage-specific point-wise mapping composed of two 1×1 convolutions with a SiLU activation. Because no spatial kernel is introduced, f_s learns only the local combination of the three discrepancy descriptors. The resulting map M_s is generated without pixel-level supervision and serves as conditional information for selective state modulation in the subsequent stage, rather than as a segmentation, saliency, or uncertainty map.

2.2.2 CONFLICT-CONDITIONED SELECTIVE STATE RECALIBRATION

Instead of directly injecting the conflict map into the backbone features, we propagate it to the subsequent stage as conditional information for selective state modulation. The map produced at stage s is first resized to the spatial resolution of stage $s+1$ and projected into two state-dependent modulation maps:

$$\widetilde{M}_s = \mathrm{Resize}\left(M_s; H_{s+1}, W_{s+1}\right), \left[\Delta B^{\mathrm{map}}_{s+1}, \Delta C^{\mathrm{map}}_{s+1}\right] = \mathrm{Split}\left(P_{s+1}(\widetilde{M}_s)\right), \tag{8}$$

where P_{s+1} is a bias-free 1×1 convolution that maps the single-channel conflict response to $2N$ channels, and N denotes the state dimension.

To preserve alignment with the original SS2D sequences, the two modulation maps are unfolded using the corresponding directional Cross-Scan transformation $\mathcal{T}_d$. They are then added to the input-writing and state-readout parameters of the first VSS Block in the subsequent stage:

$$\widehat{B}^d_{s+1} = B^d_{s+1} + \mathcal{T}_d\left(\Delta B^{\mathrm{map}}_{s+1}\right), \widehat{C}^d_{s+1} = C^d_{s+1} + \mathcal{T}_d\left(\Delta C^{\mathrm{map}}_{s+1}\right), d \in \{h^+, h^-, v^+, v^-\}. \tag{9}$$

This design conditions how the current input is written into the hidden state and how the state is read out, while leaving the transition parameter A, discretization step Δ, skip parameter D, scan input, and original scan order unchanged. Each conflict map is consumed only once by the first block of the next stage and is discarded afterward, preventing repeated within-stage amplification. Moreover, P_{s+1} is initialized with zero weights, so the additional branch exactly preserves the pretrained VMamba mapping at initialization.

2.3 GEOMETRIC CIRCULAR–RADIAL STRUCTURED AGGREGATION

GCRSA introduces a circular–radial geometric prior for organizing surrounding context. It distinguishes within-band interactions from cross-radius interactions through closed circumferential and open radial aggregation sequences, respectively.

2.3.1 CIRCULAR–RADIAL RELATIONAL CONSTRUCTION

Let $F_s \in \mathbb{R}^{B\times C_s\times H_s\times W_s}$ denote the Cross-Merged feature map at stage s. We use the feature-grid center as a geometric reference for token grouping and ordering, without requiring localization of an anatomical or acquisition center. Each token i is assigned a radial distance and polar angle:

$$\begin{aligned} r_i &= \sqrt{(v_i - c_x)^2 + (u_i - c_y)^2}, \\ \theta_i &= \mathrm{atan2}(u_i - c_y,\, v_i - c_x), c_x = \frac{W_s - 1}{2}, c_y = \frac{H_s - 1}{2}. \end{aligned} \tag{10}$$

Here, (u_i, v_i) denotes the row and column indices of token i. We set $\theta_i = 0$ when $r_i = 0$. These coordinates are used only to group and order existing tokens, without polar resampling.

Circular relations are constructed within the maximum valid inscribed circle, while radial relations cover the entire feature grid. Tokens are grouped and ordered as

$$\begin{aligned} &\mathcal{R}_{s,k} = \{i \mid r_i \le \rho_s,\ q_s(r_i) = k\}, \qquad \pi^{\mathrm{cir}}_{s,k} = \mathrm{Sort}_\theta(\mathcal{R}_{s,k}), \\ &a_i = \min\left(\left\lfloor K_s \frac{\theta_i + \pi}{2\pi} \right\rfloor, K_s - 1\right), \mathcal{S}_{s,m} = \{i \mid a_i = m\}, \pi^{\mathrm{rad}}_{s,m} = \mathrm{Sort}_r(\mathcal{S}_{s,m}). \end{aligned} \tag{11}$$

Here, $q_s(\cdot)$ discretizes radial distances into approximately concentric groups, ρ_s is the radius of the maximum inscribed circle, and K_s denotes the number of angular groups. Each circular group is ordered by angle and processed with cyclic connectivity. Each angular group is ordered by increasing radius to form an open center-to-periphery sequence.

2.3.2 BIDIRECTIONAL GEOMETRIC AGGREGATION AND GATED FUSION

For a geometric sequence $Z = [z_0, \ldots, z_{N-1}]$, we learn independent channel-shared profiles for the clockwise, counterclockwise, inner-context, and outer-context directions. For $N \ge 2$, their weights

are

$$p_d^q = \frac{\exp(\ell_d^q)}{\sum_{j=1}^{N-1} \exp(\ell_j^q)}, d = 1, \ldots, N-1, q \in \{\mathrm{cw}, \mathrm{ccw}, \mathrm{in}, \mathrm{out}\}. \tag{12}$$

The index d denotes an offset in sequence order. For sequences of different lengths, the corresponding profile is truncated to the available offsets and renormalized.

For a circular sequence of length N_k and a radial sequence of length N_m, the four directional representations are computed as

$$\widetilde{z}_i^{\mathrm{cw}} = \sum_{d=1}^{N_k-1} p_d^{\mathrm{cw}}\, z_{(i-d) \bmod N_k}, \qquad \widetilde{z}_i^{\mathrm{ccw}} = \sum_{d=1}^{N_k-1} p_d^{\mathrm{ccw}}\, z_{(i+d) \bmod N_k},$$

$$\widetilde{z}_i^{\mathrm{in}} = \begin{cases} \dfrac{\sum_{d=1}^{i} p_d^{\mathrm{in}} z_{i-d}}{\sum_{d=1}^{i} p_d^{\mathrm{in}}}, & 1 \le i \le N_m - 1, \\ z_i, & i = 0, \end{cases} \widetilde{z}_i^{\mathrm{out}} = \begin{cases} \dfrac{\sum_{d=1}^{N_m-1-i} p_d^{\mathrm{out}} z_{i+d}}{\sum_{d=1}^{N_m-1-i} p_d^{\mathrm{out}}}, & 0 \le i < N_m - 1, \\ z_i, & i = N_m - 1. \end{cases} \tag{13}$$

The circular expressions apply to $N_k \ge 2$; singleton sequences preserve their original token features, and empty groups are skipped. Circular aggregation uses cyclic indexing to connect the ends of each circumferential sequence. Radial aggregation renormalizes weights over the available inner or outer context, with the original feature retained when that context is empty.

Let $T_s \in \mathbb{R}^{B \times (H_s W_s) \times C_s}$ denote the spatially flattened tokens of F_s. The four directional representations are scattered back to their original token positions as $\widetilde{T}_s^q$ and fused through channel-wise residual gates:

$$R_s^q = \widetilde{T}_s^q - T_s, \gamma_s^q \in \mathbb{R}^{1 \times 1 \times C_s}, T_s^{\mathrm{geo}} = T_s + \sum_q \gamma_s^q \odot R_s^q, \tag{14}$$

where $q \in \{\mathrm{cw}, \mathrm{ccw}, \mathrm{in}, \mathrm{out}\}$. The four gates are independently learned and initialized to zero, ensuring $T_s^{\mathrm{geo}} = T_s$ at initialization. Circular residuals are set to zero outside the inscribed circle, where tokens still participate in radial aggregation and the original backbone computation. GCRSA is inserted after Cross-Merge and before output normalization, retaining the original token count, spatial arrangement, and Cartesian selective-scan paths.

3 Experiments

3.1 Datasets and Experimental Setup

Datasets. We evaluate CRGA on a clinically acquired in-house esophageal endoscopic ultrasonography (EUS) cohort and two public medical image classification benchmarks. The EUS cohort contains 1,658 images from 822 patients, with five categories defined by lesion layer of origin and invasion depth. Given the limited availability of labeled EUS data, this cohort provides annotations from five experienced clinicians for fine-grained recognition of lesion origin and invasion depth. The task requires interpreting local lesion appearance in relation to surrounding tissue layers. We use 1,324 images for development and 334 for held-out testing, with five-fold training–validation partitioning performed within the development set. All splits are constructed at the patient level, preventing patient overlap between development and testing or between training and validation within each fold.

To broaden the evaluation beyond esophageal EUS, we additionally use FETAL_PLANES_DB and OrganAMNIST$_{224}$. FETAL_PLANES_DB comprises six maternal–fetal ultrasound image categories: abdomen, brain, femur, thorax, maternal cervix, and other planes. OrganAMNIST$_{224}$ contains 11 organ categories from axial abdominal CT images. These public benchmarks complement the lesion-layer classification task by evaluating CRGA in additional anatomical settings and imaging modalities. Detailed dataset descriptions and partition statistics are provided in the appendix.

Implementation details. All input images are converted to three channels and resized to 224×224. Random rotation and horizontal and vertical flipping are applied during training across the three datasets. For EUS, these transformations aim to reduce dependence on display orientation

Table 1: Comparing our method with the classic models and recent state-of-the-art (SOTA) FGVC models on In-house Esophageal Dataset(Unit: %, FLOPs unit: G).

Method	Category	Acc.	Recall	F1.	Spe.	AUC	FLOPs
ResNet-18 He et al. (2016)	CNN	68.9	60.7	63.4	91.2	89.0	1.82
ResNet-50 He et al. (2016)	CNN	69.5	63.4	66.2	91.4	89.7	4.13
VGG-16 Simonyan & Zisserman (2014)	CNN	65.6	55.7	58.3	91.2	84.1	16.9
ViT Dosovitskiy et al. (2020)	Transformer	61.6	59.0	61.2	89.4	85.9	16.9
IELT Xu et al. (2023)	ViT-based FGVC	66.7	62.3	65.4	90.6	88.2	19.3
FFVT Wang et al. (2021)	ViT-based FGVC	73.8	73.0	74.5	92.7	92.0	16.4
Cross-X Luo et al. (2019)	CNN-based FGVC	74.2	71.4	72.8	92.7	92.0	4.13
AA-Trans Wang et al. (2023)	ViT-based FGVC	74.3	73.7	75.3	92.8	91.5	25.7
SIM-Trans Sun et al. (2022)	ViT-based FGVC	75.9	72.5	74.9	93.2	91.7	16.9
TransFG He et al. (2022)	ViT-based FGVC	77.9	75.9	78.1	93.8	89.3	17.0
SRRM-ViT Liu et al. (2024a)	ViT-based FGVC	78.7	77.4	79.3	94.0	93.8	32.5
Vim Zhu et al. (2024)	mamba-based	77.1	74.8	77.7	93.4	91.7	10.1
LocalMamba Huang et al. (2024)	mamba-based	78.9	75.4	78.1	93.9	92.6	11.4
VSSD Shi et al. (2025)	mamba-based	79.2	76.0	78.7	94.1	93.2	8.1
VMamba Liu et al. (2024b)	mamba-based	79.9	76.5	79.9	94.2	94.1	8.7
Ours	Proposed method	**82.1**	**80.8**	**82.9**	**95.0**	**94.4**	9.1

Table 2: Comparison with representative methods on two public datasets (Unit: %).

Method	Fetal-Planes-DB		Organ AMNIST$_{224}$	
	ACC	AUC	ACC	AUC
ResNet-50 He et al. (2016)	93.1	99.3	92.8	98.5
ViT-B Dosovitskiy et al. (2020)	84.0	96.5	89.0	98.8
Swin-T Liu et al. (2021)	88.9	98.5	91.0	98.9
MedViT-S Manzari et al. (2023)	91.0	98.6	92.8	99.6
MedMamba-T Yue & Li (2024)	93.9	99.3	94.6	**99.8**
MedKAFormer-T Wang et al. (2025)	93.4	99.4	92.2	99.5
InceptionMamba Huang et al. (2026)	94.6	**99.5**	93.0	99.7
Ours	**95.5**	**99.5**	**96.3**	**99.8**

in classifying lesion origin and invasion depth relative to surrounding tissue layers. Similar augmentation has been used in EUS and fetal-plane classification Liu et al. (2024a). The VMamba baseline and all CRGA variants share the same ImageNet-pretrained VMamba-Small checkpoint, augmentation pipeline, optimization settings, and training schedule. ViT-based baselines use ViT-B/16 pretrained on ImageNet-21K. All EUS comparisons use identical patient-level data partitions. We optimize the models using AdamW with a learning rate of 5×10^{-5} and weight decay of 0.05. The batch size is 32, and each EUS development fold is trained for 500 epochs. Method-specific configurations, pretrained checkpoint sources, and additional implementation details are provided in Appendix. We report the mean of each evaluation metric across the corresponding training runs. For EUS, these comprise five development-fold runs evaluated on the same test set. Public benchmarks are evaluated on their official test splits.

3.2 Results and Discussion

Results on the In-house EUS dataset. We first conduct the main experiments on the clinically acquired in-house EUS dataset to evaluate the ability of different methods to distinguish subtle variations in the lesion layer of origin and invasion depth. As shown in Table 1, conventional CNNs and general-purpose Vision Transformers show relatively limited performance, indicating that local appearance modeling or global self-attention alone may be insufficient to characterize the subtle and localized structural differences in EUS images. Fine-grained recognition methods provide further improvements, but their region-selection mechanisms, originally designed for natural images, may not fully conform to the structural relationships present in ultrasound images. In comparison, the Vision Mamba variants generally achieve stronger results, suggesting that multidirectional selective scanning is better suited to modeling long-range structural dependencies in EUS images.

Built upon the same VMamba backbone, CRGA improves the classification accuracy from 79.9% to 82.1%, while consistently increasing Recall, F1, Specificity, and AUC. The improvement is therefore not restricted to the overall number of correct predictions, but also reflects more balanced recogni-

Table 3: Ablation study on different modules (Unit: %).

Module			Performance			
Baseline	CCR	GCRSA	Acc.	Recall	F1.	AUC
✓			79.9	76.5	79.9	94.1
✓	✓		81.3	**81.5**	82.7	93.5
✓		✓	81.1	81.2	81.6	**95.0**
✓	✓	✓	**82.1**	80.8	**82.9**	94.4

tion across the fine-grained categories. The gains in accuracy, recall, and F1 support enriching VMamba's contextual representations through cross-axis conditioning and circular–radial aggregation. The geometric branch introduces within-band and cross-radius interactions over the existing feature grid, providing an additional organization of surrounding context alongside Cartesian selective scanning. Moreover, CRGA increases the computational cost only from 8.7 G to 9.1 G FLOPs, showing that the performance gain is achieved without a substantial increase in model complexity.

Results on the public datasets. We further evaluate the scalability of CRGA on FETAL_PLANES_DB and OrganAMNIST$_{224}$. FETAL_PLANES_DB contains visually similar standard fetal ultrasound planes and is used to examine whether the proposed method generalizes to another fine-grained ultrasound classification task. OrganAMNIST$_{224}$ consists of abdominal CT images and is used to assess its applicability across imaging modalities and anatomical structures. As reported in Table 2, CRGA achieves competitive ACC and AUC on both benchmarks. Its performance on FETAL_PLANES_DB indicates that the proposed conflict and geometric modeling strategies are not restricted to esophageal EUS, but can be extended to fetal ultrasound images with different fields of view and spatial organizations. The results on OrganAMNIST$_{224}$ further demonstrate that the proposed relational modeling remains effective for CT-based organ classification. Overall, the consistent results across datasets and imaging modalities support the scalability of CRGA rather than dataset-specific optimization.

3.3 Ablation Studies

Module-level ablation. Under the same VMamba backbone and training protocol, CCR and GCRSA individually improve accuracy from 79.9% to 81.3% and 81.1%, respectively (Table 3). CCR yields larger gains in macro-averaged recall and F1 through cross-axis-conditioned state refinement, while GCRSA supplements Cartesian scanning with within-band and cross-radius aggregation and achieves the highest AUC of 95.0%. Combining them yields the best accuracy and F1 among the ablated configurations, at 82.1% and 82.9%. The accuracy gains of 0.8 and 1.0 percentage points over CCR and GCRSA alone support their joint use, consistent with CRGA's leading accuracy in the main EUS comparison (Table 1).

CCR nevertheless introduces an AUC trade-off: adding it decreases AUC by 0.6 percentage points both without GCRSA (94.1% to 93.5%) and with GCRSA (95.0% to 94.4%), despite improving accuracy and F1 in both settings. This reflects differing effects on final class assignments and the class-wise score rankings evaluated by one-vs-rest AUC. GCRSA increases AUC by 0.9 percentage points in both settings, leaving the complete model above the baseline in AUC. The combination therefore offers the strongest accuracy and F1, while GCRSA alone provides the best AUC.

Effectiveness validation. Table 4 examines the conditioning selection in CCR and the relational organization in GCRSA. For CCR, the sparse variant uses the same conflict-map generation and state-recalibration procedure but retains only the eight lowest-scoring spatial positions during both training and testing. Its accuracy improves from the baseline's 79.9% to 80.5%, showing that this restricted conditioning remains useful despite retaining only low-scoring positions. Using the full map further improves accuracy, recall, and F1 to 81.3%, 81.5%, and 82.7%, exceeding the sparse variant by 0.8, 2.1, and 2.1 percentage points, respectively, while AUC is slightly lower (93.5% versus 93.7%). These results support retaining spatially distributed cross-axis response information for subsequent state recalibration, with the full conflict map achieving higher accuracy, recall, and F1 than conditioning on only eight low-scoring positions.

Table 4: Effectiveness validation of the conflict and geometric components on the in-house EUS dataset (Unit: %).

Configuration	ACC	Recall	F1.	AUC
Conflict-Conditioned Recalibration				
Baseline	79.9	76.5	79.9	94.1
Full conflict map	81.3	81.5	82.7	93.5
8 lowest-conflict positions	80.5	79.4	80.6	93.7
Geometric Circular-Radial Aggregation				
Baseline	79.9	76.5	79.9	94.1
Circular only	80.4	79.2	81.5	94.7
Radial only	80.7	78.2	80.7	93.2
Horizontal-vertical	78.4	77.3	79.6	93.4
Circular-radial	81.1	81.2	81.6	95.0

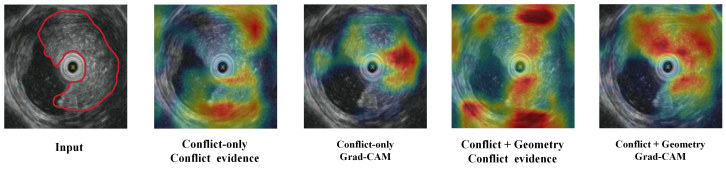


Figure 3: Visualization of class-discriminative activation and cross-axis conflict evidence.

For GCRSA, circular and radial aggregation individually improve accuracy to 80.4% and 80.7%, respectively. Circular aggregation yields higher F1 and AUC, whereas radial aggregation achieves higher accuracy. Their combination reaches 81.1% accuracy and 95.0% AUC, supporting the joint use of within-band and cross-radius contextual interactions. Replacing circular–radial organization with horizontal–vertical aggregation under a comparable aggregation form reduces accuracy to 78.4%, with F1 and AUC also below the baseline. Thus, adding an aggregation branch does not by itself guarantee improved recognition. The results favor the circular–radial organization among the tested alternatives and highlight the importance of how contextual relationships are constructed over the feature grid.

Figure 3 illustrates how CCR and GCRSA jointly affect feature responses on an infiltration-type EUS image. With CCR alone, Grad-CAM mainly emphasizes the right-sided portion of the lesion, while conflict responses are distributed across the lesion and surrounding tissue. After incorporating GCRSA, Grad-CAM covers a broader portion of the lesion, particularly its upper and right-sided extent, together with adjacent tissue regions. The accompanying conflict map retains spatially distributed responses, providing cross-axis contextual variation for state recalibration. This qualitative comparison suggests that the combined model captures a broader range of lesion-related and perilesional features, supporting the joint use of conflict-conditioned feature refinement and circular–radial context aggregation to interpret local abnormalities within their surrounding tissue context.

4 Conclusion

We presented CRGA, a VMamba-based framework that combines feature refinement with circular–radial context aggregation for fine-grained medical image classification. CCR uses discrepancies between spatially aligned directional responses to condition selective state writing and readout in the next stage, explicitly connecting response comparison with subsequent contextual refinement. GCRSA introduces a geometric prior that organizes surrounding context through closed circumferential and open radial sequences over existing tokens. Both components retain the original feature grid and Cartesian scan order, while zero-initialized modulation projections and residual gates preserve the pretrained backbone mapping at initialization. CRGA improves VMamba accuracy by 2.2 percentage points on the in-house EUS dataset and achieves competitive performance on FETAL_PLANES_DB and OrganAMNIST$_{224}$. Same-backbone ablations on EUS further demonstrate accuracy gains from each component. Future work will extend evaluation to larger multicenter cohorts and investigate the application of CRGA to three-dimensional medical imaging tasks.

5 AI ASSISTANCE STATEMENT

Generative AI tools were used for language editing and limited assistance with non-core auxiliary scripting. All experimental procedures, results, and scientific conclusions were independently verified by the authors, who take full responsibility for the submitted work.

6 ETHICS STATEMENT

The retrospective study involving the in-house EUS dataset was approved by the institutional ethics committee of the participating hospital. The approval number has been withheld during the double-blind review to avoid revealing the authors' institutional identity and will be provided in the camera-ready version if the paper is accepted.

7 REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of the proposed architecture, training configurations, dataset partitions, and evaluation protocols in the main paper and supplementary materials. The VMamba baseline and CRGA variants share the same ImageNet-pretrained VMamba-Small backbone and training protocol. The in-house EUS dataset is partitioned at the patient level, with five-fold cross-validation performed on the development set and evaluation on an independent test set. The public benchmarks are evaluated using their specified dataset partitions. Implementation details, including software versions and computing infrastructure, are provided in the supplementary materials. The implementation code will be made publicly available upon publication. The in-house clinical dataset is not publicly available.

REFERENCES

Nkechinyere N Agu, Joy T Wu, Hanqing Chao, Ismini Lourentzou, Arjun Sharma, Mehdi Moradi, Pingkun Yan, and James Hendler. Anaxnet: Anatomy aware multi-label finding classification in chest x-ray. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 804–813. Springer, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, and Changhu Wang. Transfg: A transformer architecture for fine-grained recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 852–860, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Bingquan Huang, Yue Liu, Bin Tang, and Gang Fang. Inceptionmamba: A lightweight and effective model for medical image classification revealing mamba's low-frequency bias. *Neural Processing Letters*, 58(1):15, 2026.

Tao Huang, Xiaohuan Pei, Shan You, Fei Wang, Chen Qian, and Chang Xu. Localmamba: Visual state space model with windowed selective scan. In *European conference on computer vision*, pp. 12–22. Springer, 2024.

Wei Ji, Shuang Yu, Junde Wu, Kai Ma, Cheng Bian, Qi Bi, Jingjing Li, Hanruo Liu, Li Cheng, and Yefeng Zheng. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12336–12346. IEEE, 2021.

Joshua Kim, Dan Ionascu, and Tiezhi Zhang. Algebraic iterative image reconstruction using a cylindrical image grid for tetrahedron beam computed tomography. *Medical physics*, 40(8):081909, 2013.

Tianyi Liu, Shuaishuai Zhuang, Jiacheng Nie, Geng Chen, Yusheng Guo, Guangquan Zhou, Jean-Louis Coatrieux, and Yang Chen. A rotation-invariant texture vit for fine-grained recognition of esophageal cancer endoscopic ultrasound images. In *European Conference on Computer Vision*, pp. 360–377. Springer, 2024a.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024b.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Wei Luo, Xitong Yang, Xianjie Mo, Yuheng Lu, Larry S Davis, Jun Li, Jian Yang, and Ser-Nam Lim. Cross-x learning for fine-grained visual categorization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8242–8251, 2019.

Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems*, 29, 2016.

Omid Nejati Manzari, Hamid Ahmadabadi, Hossein Kashiani, Shahriar B Shokouhi, and Ahmad Ayatollahi. Medvit: a robust vision transformer for generalized medical image classification. *Computers in biology and medicine*, 157:106791, 2023.

Wongi Park and Jongbin Ryu. Fine-grained self-supervised learning with jigsaw puzzles for medical image classification. *Computers in Biology and Medicine*, 174:108460, 2024.

Jue Peng, Xiaozhen Li, Hu Tang, Lucai Ma, Ting Zhang, Yunfei Li, and Siping Chen. Miniaturized high-resolution integrated 360° electronic radial ultrasound endoscope for digestive tract imaging. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 66(5):975–983, 2019.

Fatemeh Rastegar Jooybari, Ali Aghaeifar, Elham Mohammadi, Klaus Scheffler, and Abbas Nasiraei-Moghaddam. Polar fourier transform in practice: its efficiency and characteristics in reconstructing radially acquired mri images. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 39(1):21–36, 2026.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Yuheng Shi, Mingjia Li, Minjing Dong, and Chang Xu. Vssd: Vision mamba with non-causal state space duality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10819–10829, 2025.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Hongbo Sun, Xiangteng He, and Yuxin Peng. Sim-trans: Structure information modeling transformer for fine-grained visual categorization. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 5853–5861, 2022.

Guoli Wang, Qikui Zhu, Chaoda Song, Benzheng Wei, and Shuo Li. Medkaformer: When kolmogorov-arnold theorem meets vision transformer for medical image representation. *IEEE Journal of Biomedical and Health Informatics*, 2025.

Jun Wang, Xiaohan Yu, and Yongsheng Gao. Feature fusion vision transformer for fine-grained visual categorization. *arXiv preprint arXiv:2107.02341*, 2021.

Qi Wang, JianJun Wang, Hongyu Deng, Xue Wu, Yazhou Wang, and Gefei Hao. Aa-trans: Core attention aggregating transformer with information entropy selector for fine-grained visual classification. *Pattern Recognition*, 140:109547, 2023.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794–7803, 2018.

Qin Xu, Jiahui Wang, Bo Jiang, and Bin Luo. Fine-grained visual classification via internal ensemble learning transformer. *IEEE Transactions on Multimedia*, 2023.

Yubiao Yue and Zhenzhang Li. Medmamba: Vision mamba for medical image classification. *arXiv preprint arXiv:2403.03849*, 2024.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.