Decoupled Ordinal Refinement with Geometric Alignment for Chest X-ray Severity Scoring

Geon Jeong[1], Jung Soo Kim[2], and Hyun Gyu Lee[3(✉)]

[1] Department of Statistics, Inha University, Republic of Korea
[2] Inha University Hospital, Republic of Korea
[3] College of Medicine, Inha University, Republic of Korea
12191910@inha.edu, acecloer31@gmail.com, hglee@inha.ac.kr

Abstract. Multi-region severity scoring on chest X-rays (CXR) faces three structural challenges: extreme class imbalance across regions of interest (ROIs), ordinal label discontinuity, and inter-rater subjectivity. We argue that performance degradation in sparse high-severity grades stems from ordinal geometry collapse—the compounding interaction of Neural Collapse [1] and Minority Collapse [2] in the representation space—rather than from data scarcity alone.
To address this, we propose the Decoupled Ordinal Refinement with Geometric Alignment (DORGA), which operates within an explicitly designed Ordinal-Aligned Latent Space. Our approach is threefold: (i) learnable anchors constrained by polar fixing and ordinal separation pin the severity coordinate system on the hypersphere a priori; (ii) ROI embeddings are orthogonally decoupled into severity and structure components, preventing severity similarity from distorting inter-ROI attention; and (iii) pattern-conditional Bayesian propagation refines ambiguous severity embeddings via confidence-gated spherical interpolation, guided by oracle-supervised structure-aware attention during training. DORGA achieves an average MAE of 0.301 on the Brixia consensus test set, compared to 0.441 (BS-Net) and 0.350 (PAFE). Without fine-tuning, it attains QWK 0.797 with the senior radiologist on the external Cohen dataset, matching inter-rater agreement (0.790).

Keywords: Ordinal Regression · Representation Geometry · Chest X-ray · Severity Scoring · Imbalanced Learning

1 Introduction

The Brixia Score [3] assigns ordinal grades 0–3 to six lung ROIs for CXR severity assessment. Automating this scoring is hindered by sharp clinical imbalances: while Grade 3 is rare in the upper lobes, Grade 0 shows a notably low prevalence in the lower lobes. Standard Cross-Entropy (CE) training biases decision boundaries toward majority classes, while non-uniform ordinal spacing and inter-rater variability act as compounding label noise.

We argue that these failures are rooted in ordinal geometry collapse: under class imbalance, Neural Collapse drives features toward a symmetric Simplex

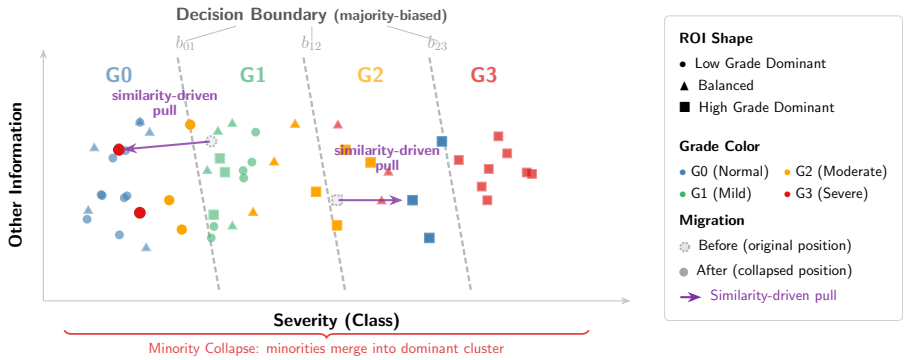

Fig. 1. Conceptual illustration of Minority Collapse under CE: per-ROI minority grades lose independent structure and merge into each ROI's dominant cluster, biasing decision boundaries toward majority classes. Direct mechanistic evidence is provided in Fig. 3.

ETF that ignores ordinal proximity, while Minority Collapse shrinks sparse-class representations toward majority-class directions. Their interaction in the multi-region ordinal setting destroys the coordinate system that severity grading requires.

We propose the Decoupled Ordinal Refinement with Geometric Alignment (DORGA), whose components share a single geometric substrate—the Ordinal-Aligned Latent Space on S^{d-1}—and address complementary failure modes of ordinal geometry collapse: **(1) Coordinate Initialization**: learnable anchors, constrained by polar fixing and ordinal separation, establish the severity axis before class imbalance can distort it. **(2) Axis Protection**: orthogonal u/v projections decouple severity from structure, ensuring inter-ROI attention reflects anatomical similarity rather than grade proximity [4,5]. **(3) Embedding Refinement**: pattern-conditional Bayesian propagation refines ambiguous severity embeddings by propagating probabilistic beliefs through the preserved geometry, with oracle supervision applied only during training. Ablation (Sec. 4.2) confirms that removing any one stage degrades the others, evidencing non-trivial interdependence rather than additive gains.

2 Related Work

Neural Collapse & Minority Collapse. Neural Collapse [1] and Minority Collapse [2,6] are well-established phenomena in imbalanced learning. Our contribution is not identifying these mechanisms but analyzing how their interaction in the multi-region ordinal setting specifically distorts inter-ROI attention, and designing a geometrically principled framework to counteract ordinal geometry collapse.

Ordinal & Contrastive Approaches. Contrastive methods—SupCon [7], AdaCon [8], SCOL [9], HCOR [10]—enforce ordinal margins but entangle severity and anatomical identity in a shared embedding, distorting GNN attention toward grade similarity rather than anatomical relationships [4,5]. DORGA avoids contrastive losses entirely, achieving ordinal structure through anchor alignment and blocking attention bias via explicit decoupling.

Feature Decoupling. Bi et al. [11] introduced channel-wise decorrelation for domain-generalizable medical image segmentation via deep whitening transforms. We adopt this principle at the embedding level, enforcing orthogonality against a geometrically meaningful severity axis w—directly constraining representation geometry before learning begins, rather than modifying loss functions or contrastive objectives.

3 Method

3.1 Stage 1: ROI Feature Extraction

We employ ViT-B/16 initialized with MRM [12] pre-trained weights as the backbone, producing patch tokens $\mathbf{X} \in \mathbb{R}^{P\times d_{\text{vit}}}$ and a CLS token $\mathbf{x}_{\text{cls}}$. Per-patch importance scores w_p are computed via a lightweight depthwise–pointwise convolution with learnable positional bias and temperature-scaled softmax. Each ROI feature $z_r \in \mathbb{R}^{d_{\text{vit}}}$ is obtained by weighted pooling, combining w_p with segmentation mask membership $m_{rp} \in [0, 1]$ across R=6 Brixia regions.

3.2 Stage 2: Ordinal Geometry Construction

Class Anchors. Learnable anchors $A = \{a_0, \ldots, a_{C-1}\} \subset \mathbb{S}^{d-1}$ are initialized as equally spaced points on a great arc. A polar loss $\mathcal{L}_{\text{polar}} = (1 + a_0 \cdot a_{C-1})^2$ enforces antipodality between extreme grades, and an ordinal separation loss $\mathcal{L}_{\text{ord}} = \text{MSE}(a_k \cdot a_{k+1},\ \cos\frac{\pi}{C-1})$ enforces equal angular spacing between adjacent anchors. Together they define the severity axis $w = \frac{a_{C-1}-a_0}{\|a_{C-1}-a_0\|}$.

Severity–Structure Decoupling. To prevent the severity-driven attention bias described in Sec. 2, we decouple each ROI into two orthogonal components: a shared projector ϕ maps z_r to severity embedding $u_r = \phi(z_r) \in \mathbb{S}^{d-1}$, while region-specific projectors $\{\phi_r\}_{r=1}^{R}$ produce structure embeddings $v_r = \phi_r(z_r) \in \mathbb{S}^{d-1}$, constrained to be orthogonal to the severity axis w. Both projectors consist of a two-layer MLP followed by ℓ_2-normalization.

Stage 2 Losses. *Severity alignment* pulls u_r toward a_{y_r} via: $\mathcal{L}_{\text{align}} = 1 - u_r \cdot a_{y_r}$ (orbital drift), $\mathcal{L}_{\text{1D}} = |\arccos(u_r \cdot w) - \arccos(a_{y_r} \cdot w)|$ (polar gradient vanishing), and $\mathcal{L}_{\text{obtuse}}$ with linear cost when $u_r \cdot a_{y_r} < 0$ (wrong-hemisphere correction).

Structure orthogonality prevents severity leakage into v_r: $\mathcal{L}_{\text{orth}} = |\arccos(v_r \cdot w) - \pi/2|$.

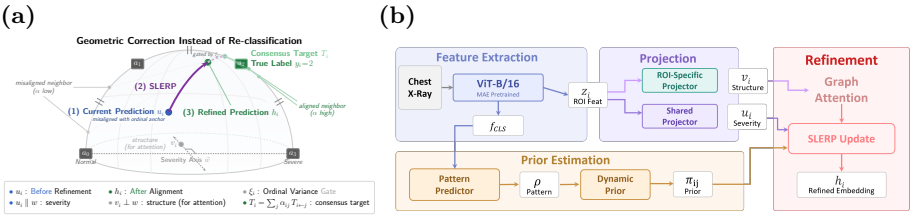

Fig. 2. Ordinal-aligned geometry and SLERP-based refinement in DORGA. (a) On the hypersphere $\mathbb{S}^{d-1}$: the severity embedding u_i is rotated toward a consensus target T_i via confidence-gated SLERP, preserving ordinal geometry. (b) End-to-end pipeline illustrating geometry construction, decoupling, and refinement stages.

The angular classifier computes logits as negative angular distances scaled by learnable temperature τ_{S_2}, and the total Stage 2 loss combines all terms:

$$\mathcal{L}_{S2} = \underbrace{\mathcal{L}_{\text{CE}}}_{\text{boundary}} + \underbrace{\mathcal{L}_{\text{align}} + \mathcal{L}_{\text{1D}} + \mathcal{L}_{\text{obtuse}}}_{\mathcal{L}_{\text{severity}}} + \underbrace{\mathcal{L}_{\text{orth}}}_{\text{structure}} + \underbrace{\mathcal{L}_{\text{polar}} + \mathcal{L}_{\text{ord}}}_{\mathcal{L}_{\text{anchor}}}$$

3.3 Stage 3: Inter-ROI Refinement via Graph Attention

Stage 3 exploits inter-ROI relationships: structure embeddings v guide neighbor selection via attention, while probabilistic message passing refines severity beliefs.

Belief Estimation & Dynamic Prior. Per-ROI belief is extracted from the Stage 2 severity embedding: $p_i = \text{onehot}(\arg\max_c \text{CosSim}_{i,c})$, with stop-gradient on u_i. The relationship between neighboring grades depends on the global disease pattern. We derive K discrete patterns by applying K-means to the six-region score vectors of the training set; K=7 is selected empirically as the smallest value preserving distinct transition patterns. We model this via K pattern-specific conditional matrices $\{\pi^{(k)}\}_{k=1}^{K} \in \mathbb{R}^{r\times r\times c\times c}$, composed into a conditional prior:$\pi_{ij} = \sum_k \rho_k \cdot \pi_{ij}^{(k)}$, where $\pi_{ij,cd} = P(y_i{=}c \mid y_j{=}d)$ and $\rho = \text{softmax}(\text{MLP}([\text{sg}[\mathbf{x}_{\text{cls}}]; \text{sg}[\text{logits}_{S2}]_{\text{flat}}]))$ is the pattern distribution.

Ordinal Message Passing & Structure-Aware Attention. Given neighbor j's belief p_j and the transition matrix π_{ij}, we compute $q_{i\leftarrow j}(c) = \sum_d \pi_{ij,cd} \cdot p_j(d)$. Crucially, $q_{i\leftarrow j}$ is *not* a forwarded belief from j; it is *neighbor j's expected distribution over ROI i's severity*, computed via the conditional prior π_{ij}. This operation corresponds to one iteration of loopy belief propagation on a fully connected factor graph, where π_{ij} acts as the pairwise potential and structure-based attention α_{ij} modulates message aggregation, yielding the consensus $\bar{q}_i(c) = \sum_j \alpha_{ij} \cdot q_{i\leftarrow j}(c)$.

Oracle Attention. To guide the learned attention α toward weighting neighbors that can effectively correct predictions, we construct an oracle target α^*. For each neighbor $j \neq i$, the opinion $q_{i \leftarrow j}$ is computed from j's ground-truth label; for the self-connection ($j=i$), the Stage 2 model prediction is used instead, so that misclassified ROIs naturally receive lower self-attention in the oracle. We measure each opinion's expected ordinal distance from ROI i's ground truth: $d_{ij} = |\sum_c c \cdot q_{i \leftarrow j}(c) - y_i|$. The oracle assigns high weight to neighbors whose opinion closely predicts i's true grade: $\alpha^*_{ij} \propto \exp(\frac{-d_{ij}}{\tau_o})$. Additionally, a per-node confidence c_i restricts oracle supervision to misclassified ROIs: $c_i = \mathbb{1}[\hat{y}_i \neq y_i] \cdot \exp(-\min_j d_{ij})$, where $\hat{y}_i$ is the Stage 2 prediction. The attention loss is: $\mathcal{L}_{\text{att}} = \frac{\sum_i c_i \cdot \mathrm{KL}(\alpha_i^* \| \alpha_i)}{\sum_i c_i}$. We emphasize that oracle targets are constructed from ground-truth labels and are used exclusively as a training signal for $\mathcal{L}_{\text{att}}$; at inference, no ground-truth information is available, and attention weights are computed solely from the learned parameters. Through $\mathcal{L}_{\text{att}}$, the oracle implicitly teaches the structure projectors to identify neighbors whose transition-based opinions are most corrective, without exposing severity information to the attention mechanism.

Confidence-Gated Spherical Refinement. The $\bar{q}_i$ may be unimodal or bimodal/dispersed. We quantify this via ordinal variance $\sigma_i^2 = \sum_c \bar{q}_i(c)(c - \mu_i)^2$ where $\mu_i = \sum_c c \cdot \bar{q}_i(c)$, yielding confidence: $\xi_i = \text{clamp}(1 - \sigma_i^2 / V_{\max},\ 0,\ 1)$ where $V_{\max} = (C-1)^2/4$. The consensus is back-projected to a target direction $T_i = \frac{\sum_c \bar{q}_i(c)\, a_c}{\left\| \sum_c \bar{q}_i(c)\, a_c \right\|}$, and the embedding is refined on the hypersphere: $h_i = \text{SLERP}(u_i,\ T_i,\ t_{\text{eff},i}), t_{\text{eff},i} = \sigma(s) \cdot \xi_i$, where $\sigma(s)$ is a learnable global gate that controls the maximum refinement magnitude.

3.4 Stage 4: Angular Classification

Both u (Stage 2) and h (Stage 3) pass through the angular classifier with independent learnable temperatures τ_{S_2}, τ_{S_3}, where the CE loss applies inverse-frequency class weights. The total objective is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pat}} + \mathcal{L}_{S2} + \mathcal{L}_{S3} + \mathcal{L}_{\text{att}}$, where $\mathcal{L}_{\text{pat}}$ is CE over K pattern classes obtained by K-means clustering of the six-region label vectors in the training set, and $\mathcal{L}_{S3} = \mathcal{L}_{\text{CE}}(\text{logits}_{S3}, y)$. At inference, only logits_{S3} are used and oracle serves purely as a training-time curriculum that teaches the attention that structurally good neighbors are informative; once learned, this behavior generalizes without privileged information.

4 Experiments

4.1 Comparison with State-of-the-Art

Dataset & Setup. All experiments use the Brixia COVID-19 dataset [3] (4,695 CXRs, six-region severity 0–3). We evaluate on (1) the **Brixia Consensus**

Table 1. Per-ROI grade distribution on the Brixia training set (N=4,695). R/L denote right/left lungs; T/M/B denote top, middle, and bottom zones. The distribution exhibits severe per-ROI imbalance: Grade 0 dominates the upper zones (ROI 1, ROI 4), while Grade 3 concentrates in the lower zones (ROI 3, ROI 6).

ROI	Grade 0	Grade 1	Grade 2	Grade 3
ROI 1 (RT)	2,121	1,509	721	344
ROI 2 (RM)	953	1,228	1,512	1,002
ROI 3 (RB)	543	1,032	1,738	1,382
ROI 4 (LT)	2,572	1,389	535	199
ROI 5 (LM)	1,010	1,200	1,499	986
ROI 6 (LB)	568	998	1,621	1,508

Test Set (150 images, 5-radiologist majority vote) and (2) the **Cohen Public Dataset** (192 images, independent senior/junior annotations). Training uses AdamW (per-group lr 10^{-5}–3×10^{-4}, cosine decay, 100 epochs) with mixed-precision and gradient accumulation (effective batch 64). The number of severity patterns is K=7, projection dimension d=768, and attention heads H=4.

Table 2 compares DORGA against BS-Net [3] and PAFE [13]. On the blind test set (Part A), DORGA reduces average MAE from 0.441 to **0.426** over BS-Net, with average CC improving from 0.713 to **0.744**. The improvement is moderate because both models are evaluated against the same single-annotator labels (R_0), whose inter-rater noise limits the measurable gap.

On the consensus gold standard (Part B), where the reference is a 5 radiologist majority vote, the gap widens sharply: DORGA achieves MAE **0.301** versus BS-Net's 0.424 and PAFE's 0.350, with accuracy **71.8%**. Quantitatively, DORGA's Blind→Consensus improvement (Δ_{MAE}=0.125) is substantially larger than BS-Net's (Δ_{MAE}=0.017), consistent with the geometry-denoising hypothesis. This asymmetry supports the ordinal geometry hypothesis: single-annotator noise masks the benefit of a stable severity axis, but consensus references reveal it.

4.2 Ablation Study

We evaluate three ablation variants on the consensus test set (Table 2, Part B): **w/o decoupling**, **w/o oracle**, and **S2-only** (definitions in table footnotes).

Quantitative Impact. Removing decoupling degrades MAE from 0.301 to 0.329. This reintroduces severity-driven attention bias: u serving both roles distorts neighbor selection toward grade similarity, undermining Stage 2's geometric separation.

Removing oracle supervision is worse (0.337) despite retaining decoupled projectors: v embeddings carry structural similarity but, without oracle targets, lack supervised signal to learn which structurally similar neighbors actually improve predictions; attention thus degenerates toward uniform weighting.

Table 2. Quantitative comparison on the Brixia dataset. **A**: Blind test set (DORGA N=453; BS-Net N=447 [3]). **B**: Consensus test set (N=150, 5-radiologist majority vote).

Model	Metric	ROI 1	ROI 2	ROI 3	ROI 4	ROI 5	ROI 6	Avg	Global
Part A: Blind Test Set									
BS-Net [3]	MEr	0.169	−0.038	−0.056	0.125	−0.045	−0.192	−0.006	−0.036
	MAE	0.459	0.448	0.412	0.374	0.459	0.494	0.441	1.728
	SD	0.604	0.524	0.540	0.541	0.574	0.609	0.565	1.429
	CC	0.675	0.779	0.731	0.682	0.737	0.672	0.713	0.862
Ours	ACC	0.658	0.651	0.572	0.649	0.598	0.530	0.610	0.073*
	MEr	0.071	−0.113	−0.091	0.104	−0.071	−0.075	−0.029	−0.174
	MAE	**0.375**	**0.382**	0.466	0.382	**0.442**	0.512	**0.426**	1.892
	SD	**0.548**	0.550	0.573	0.550	**0.571**	**0.582**	**0.565**	1.544
	CC	**0.750**	**0.798**	**0.723**	**0.724**	**0.768**	**0.702**	**0.744**	**0.858**
Part B: Consensus Test Set (N=150)									
BS-Net‡ [3]	MAE	—	—	—	—	—	—	0.424	1.787
PAFE [13]	ACC	–	–	–	–	–	–	67.1	–
	MAE	0.30†	0.38†	0.31†	0.32†	0.37†	0.43†	0.350	–
Ours	ACC	**0.767**	**0.760**	**0.680**	**0.700**	**0.733**	**0.667**	**0.718**	0.200*
	MEr	0.053	0.020	0.027	0.160	0.060	0.073	0.066	0.393
	MAE	**0.267**	**0.247**	**0.333**	**0.333**	**0.273**	**0.353**	**0.301**	**1.233**
	SD	**0.525**	**0.446**	**0.499**	**0.537**	**0.460**	**0.518**	**0.500**	**1.146**
	CC	**0.817**	**0.893**	**0.853**	**0.716**	**0.860**	**0.832**	**0.829**	**0.946**
w/o decoupl.§	MAE	0.280	0.293	0.327	0.347	0.307	0.420	0.329	1.293
	CC	0.815	0.874	0.858	0.713	0.838	0.791	0.815	0.941
w/o oracle‖	MAE	0.340	0.267	0.273	0.407	0.320	0.413	0.337	1.287
	CC	0.783	0.884	0.878	0.708	0.843	0.801	0.816	0.939
S2-only#	MAE	0.307	0.260	0.327	0.467	0.300	0.420	0.347	1.293
	CC	0.805	0.891	0.867	0.707	0.836	0.797	0.817	0.940

*Global ACC = exact match of six-score sum (0–18); ACC values in %
§w/o decoupl.: u for attention+messages (no decoupling, no $\mathcal{L}_{\text{att}}$).
‖w/o oracle: v for attention, u for messages, no $\mathcal{L}_{\text{att}}$.
#S2-only: Stage 2 classifier only (no GNN).
‡ BS-Net CbGS from Table 6 of [3], †PAFE MAE from Fig. 3(a) of [13].

S2-only (0.347) already matches PAFE (0.350), yet full DORGA (0.301) achieves the largest gain—Stage 2 prevents geometric collapse while Stage 3 resolves residual inter-ROI ambiguity, confirming both components are jointly essential.

Attention Distribution Analysis. Fig. 3 provides the **primary mechanistic evidence** for the ordinal-geometry hypothesis at the representation level: we average α_{ij} conditioned on grade pairs (g_i, g_j) over all test samples. With

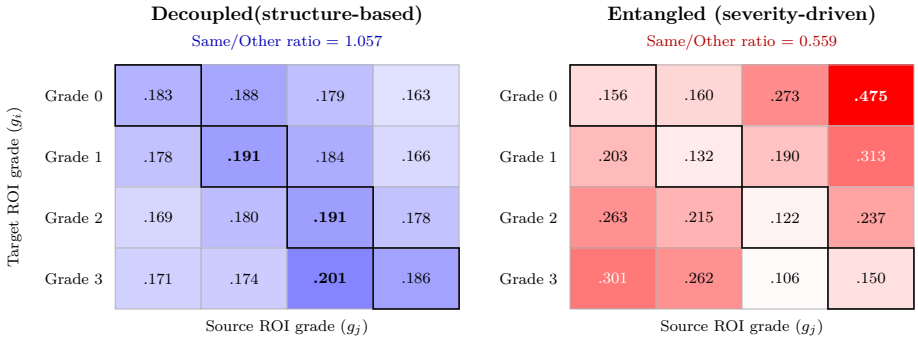

Fig. 3. Effect of severity–structure decoupling on inter-ROI attention. Decoupled representations produce structure-based attention with balanced cross-grade interactions, while entangled representations show severity-driven collapse toward similar grades. The Same/Other ratio measures attention balance (≈ 1 indicates stability).

Table 3. Inter-rater agreement on the Cohen dataset (N=192, R_s as reference).

(a) Full dataset (N=1,152)						**(b) Disagreement subset** (N=422)					
								MAE		**ACC**		**Vote**
Comp.	κ	**QWK**	**ICC**	**MAE**	**ACC**	**Subset**	N	→Sr	→Jr	→Sr	→Jr	Sr:Jr
Ours vs. Sr	**.503**	**.797**	**.798**	**.400**	**63.3**	All	422	**.595**	.711	**46.7**	39.1	**219**:184
Ours vs. Jr	.453	.774	.773	.443	60.5	$\lvert\Delta\rvert$=1	377	**.578**	.634	**47.2**	42.7	**199**:178
Sr vs. Jr	.497	.790	.790	.407	63.4	$\lvert\Delta\rvert$=2	43	**.698**	1.35	**44.2**	9.3	**19**:5
BS-Net† vs. Sr	—	—	—	.518	—	$\lvert\Delta\rvert$=3	2	1.50	1.50	0.0	0.0	1:1

†BS-Net-HA (trained on Brixia, no fine-tuning on Cohen) from Table 9 of [3]

decoupling, the same/other attention ratio is 1.057 (values near 1.0 indicate severity-independent, balanced attention), confirming that v excludes severity information from neighbor selection. Without decoupling, the ratio drops to 0.559, revealing strong anti-diagonal bias: W_Q, W_K learn to prefer severity-dissimilar neighbors, since same-grade opinions echo ROI i's own belief through π.

4.3 Inter-Rater Agreement on External Dataset

We evaluate on the Cohen public CXR dataset (N=192, 1,152 ROI ratings) with independent senior (R_s, 22 yr) and junior (R_j, 2nd-year resident) annotations [3]. DORGA achieves QWK **0.797** and MAE **0.400** with R_s (Table 3a), outperforming BS-Net-HA at MAE 0.518 [3] and exceeding senior–junior agreement (QWK 0.790, MAE 0.407). Both DORGA and BS-Net-HA are trained on the Brixia dataset and evaluated on Cohen without fine-tuning. On the 422 ratings (36.6%) where R_s and R_j disagree (Table 3b), DORGA consistently aligns with R_s (MAE .595 vs. .711; vote 219:184), and the effect strengthens with annotator distance: at $|\Delta|$=2, ACC to Senior is 44.2% vs. 9.3%.

5 Conclusion

We presented DORGA, which addresses ordinal geometry collapse—the compounding interaction of Neural Collapse and Minority Collapse that degrades multi-region CXR severity scoring under class imbalance. The consistent asymmetry between moderate blind-test gains and sharp consensus improvements confirms that the ordinal geometry acts as a structural denoiser, converging toward expert agreement despite noisy single-annotator training. Ablation confirms that no single component compensates for the absence of another: without coordinate initialization, the axis never forms; without axis protection, attention distorts it; without boundary correction, residual ambiguity persists. This mutual dependence is the defining characteristic of the framework.

Limitation. The Brixia score offers a rare setting for multi-region ordinal CXR severity scoring, with both consensus annotations and independent external readings. This protocol aligns with our hypothesis of convergence from noisy single-annotator training toward expert consensus; evaluation on other ordinal scoring tasks remains an important direction for future validation.

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF) (RS-2026-25473885) and the Institute of Information & Communications Technology Planning & Evaluation (IITP) (RS-2022-II220641) grants funded by the Ministry of Science and ICT (MSIT), Republic of Korea.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Papyan, V., Han, X.Y., Donoho, D.L.: Prevalence of neural collapse during the terminal phase of deep learning training. PNAS **117**(40), 24652–24663 (2020)
2. Fang, C., He, H., Long, Q., Su, W.J.: Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. PNAS **118**(43), e2103091118 (2021)
3. Signoroni, A., Savardi, M., Benini, S., et al.: BS-Net: Learning COVID-19 pneumonia severity on a large chest X-ray dataset. Med. Image Anal. **71**, 102046 (2021)
4. Li, M., et al.: Disentangled contrastive learning on graphs. In: NeurIPS, vol. 34 (2021)
5. Robinson, J., et al.: Can contrastive learning avoid shortcut solutions? In: NeurIPS, vol. 34 (2021)
6. Hong, W., Ling, S.: Neural collapse for unconstrained feature model under cross-entropy loss with imbalanced data. JMLR **25**(192), 1–48 (2024)
7. Khosla, P., et al.: Supervised contrastive learning. In: NeurIPS, vol. 33 (2020)
8. Dai, W., et al.: Adaptive contrast for image regression in computer-aided disease assessment. IEEE TMI **41**(5), 1255–1268 (2022)
9. Saleem, A., et al.: SCOL: Supervised contrastive ordinal loss for abdominal aortic calcification scoring on VFA scans. In: MICCAI 2023. LNCS, vol. 14225, pp. 273–283. Springer (2023)

10. Saleem, A., Lewis, J.R., Gilani, S.Z.: A hybrid contrastive ordinal regression method for advancing disease severity assessment in imbalanced medical datasets. In: MICCAI 2025. pp. 14–23. Springer (2025)
11. Bi, Q., et al.: Learning generalized medical image representation by decoupled feature queries. IEEE TPAMI (2025). DOI: 10.1109/TPAMI.2025.3597364
12. Zhou, H.Y., et al.: Advancing radiograph representation learning with masked record modeling. In: ICLR 2023 (2023)
13. Lee, J.B., Kim, J.S., Lee, H.G.: COVID19 to pneumonia: Multi region lung severity classification using CNN transformer position-aware feature encoding network. In: MICCAI 2024. LNCS, vol. 15001, pp. 469–479. Springer (2024)