Clinically-Grounded Hierarchical Classification for Consistent Chest X-ray Interpretation

Jong Hak Moon*[0000−0002−6708−3918], Minjun Kim[0009−0000−3615−1220], and Minjun Kim[0009−0009−8429−5240]

Yeji X, Seoul, South Korea
jh.moon@yejix.ai, jayden.kim@yejix.ai, mj.kim@yejix.ai

Abstract. Accurate chest X-ray interpretation is inherently hierarchical. Clinical decisions depend not only on what abnormality is present but where it is situated, requiring reasoning from broad anatomical systems down to specific pathological findings. Yet existing automated systems largely treat this as a flat classification problem, failing to capture inter-level dependencies or enforce coherence between coarse and fine predictions. We propose CHASE (**C**lassification with **H**ierarchical **A**nalysis and **S**tructured **E**nforcement), a unified single-stage framework that mirrors radiologists' coarse-to-fine reasoning through a clinically-driven three-level taxonomy of 9 anatomical regions, 17 sub-regions, and 28 pathological findings. CHASE jointly optimizes multi-level supervision, cross-level probability alignment, and a hierarchy-violation penalty within a shared Vision Transformer backbone. This ensures that fine-grained findings are anatomically supported by their coarser-level context rather than predicted in isolation. Experiments demonstrate that CHASE outperforms flat and hierarchical baselines across all levels while achieving superior probabilistic hierarchy consistency, with level-wise attention maps confirming anatomically grounded predictions. Code is available at: https://github.com/yejix-ai/CHASE.

Keywords: Hierarchical Multi-label Chest X-ray Classification · Clinically-Grounded Taxonomy · Cross-Level Probabilistic Coherence · Chest X-ray

1 Introduction

Chest X-ray (CXR) interpretation is an inherently hierarchical, coarse-to-fine process [3, 4, 6, 10, 14]. Radiologists first localize abnormalities within broad anatomical systems before refining observations into specific sub-regions and findings [11, 13, 21]—a flow that is clinically critical, as visually similar densities (e.g., pneumonia vs. pleural effusion) require fundamentally different management based on anatomical context [8, 9, 17].

Recent anatomy-aware approaches [1, 12, 23] capture spatial dependencies but remain detection-centric two-stage pipelines that classify findings independently, without enforcing structural consistency across levels. General hierarchical methods [5, 15, 19] enforce label coherence via regularization, but are

* Corresponding author

ill-suited to the multi-label, multi-granularity nature of medical imaging where a single radiograph may exhibit findings across multiple anatomical systems simultaneously. CXR-specific efforts [4, 16] address this partially via conditional probability training, yet remain confined to relatively shallow, disease-centric taxonomies with a limited set of 14 abnormalities. Consequently, current systems frequently produce coarse-to-fine contradictions that undermine clinical reliability [2, 14].

We argue these failures reflect the difficulty of learning reliable coarse-to-fine dependencies without explicit cross-level coherence [4, 10]—evidenced in our ablation, where removing hierarchy-enforcing regularizers reduces probabilistic consistency by over 9 points (Section 3.3). To address this, we propose CHASE (**C**lassification with **H**ierarchical **A**nalysis and **S**tructured **E**nforcement), a unified single-stage framework grounded in a three-level taxonomy (9 regions $\rightarrow$ 17 sub-regions $\rightarrow$ 28 findings) mirroring the coarse-to-fine clinical flow. Within a shared Vision Transformer backbone, CHASE jointly optimizes multi-level supervision, cross-level probability alignment, and a hierarchy-violation penalty, structurally enforcing that fine-grained predictions remain anatomically grounded in their coarser-level context. Our contributions are as follows:

- **Clinically-grounded taxonomy:** A three-level hierarchy (9 regions $\rightarrow$ 17 sub-regions $\rightarrow$ 28 findings) mirroring radiologists' coarse-to-fine reading order, which induces 197 anatomically valid Top–Mid–Leaf paths.
- **Unified training objective:** A joint loss combining multi-level supervision, Global KL Alignment, and a Hierarchy-Violation Penalty for cross-level probabilistic consistency.
- **Empirical validation:** CHASE outperforms baselines across all levels in AUC, achieves the highest probabilistic consistency (74.26%), and is validated through localization analysis confirming anatomical grounding.

2 Method

We present CHASE, a single-stage framework for fine-grained, clinically grounded multi-level classification from chest radiographs, allowing multiple labels to be predicted at each hierarchy level. As illustrated in Figure 1, CHASE is built on (i) a clinically grounded three-level hierarchy and (ii) a unified training objective that enforces cross-level coherence via three complementary loss components.

2.1 Clinically Grounded Three-Level Hierarchy

We structure the label space into a three-level hierarchy that aligns with radiologists' coarse-to-fine localization process (Table 1). Level 1 (**Top: Region**) comprises $C_1 = 9$ broad thoracic systems, establishing a global anatomical framework. Level 2 (**Mid: Sub-region & Laterality**) refines these into $C_2 = 17$ precise zones, differentiating spatially adjacent structures and incorporating laterality. Level 3 (**Leaf: Finding & Disease**) encompasses $C_3 = 28$ pathological labels, each linked to anatomically feasible parent locations.

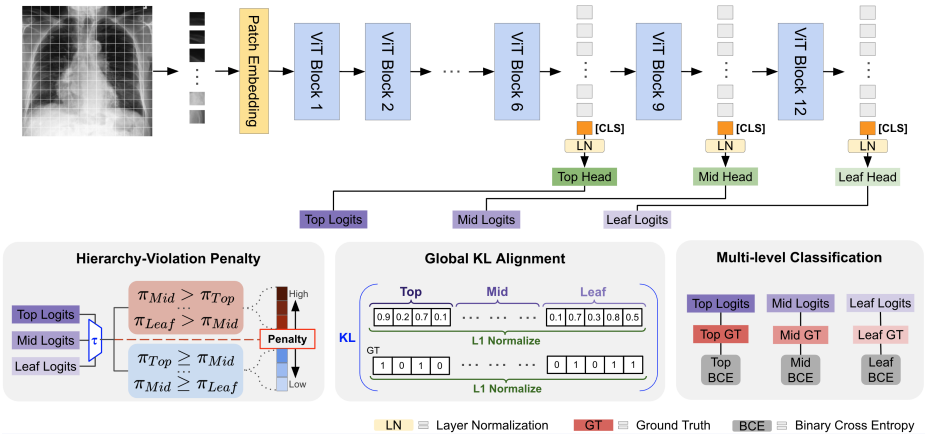

Fig. 1. Overview of the proposed hierarchical classification framework. The model (top) adopts a ViT backbone with level-specific heads for *Top*, *Mid*, and *Leaf* attached to progressively deeper transformer blocks. Training uses a unified objective (bottom) comprising: (left) a **Hierarchy-Violation** penalty that suppresses parent–child contradictions; (center) a **Global KL Alignment** that couples all hierarchy levels via L1-normalized probability distributions; and (right) a **Multi-level Classification** loss that applies Binary Cross Entropy supervision at each granularity.

Valid hierarchical relations. Let $\mathcal{E}_{12} \subseteq \{1,\dots,C_1\} \times \{1,\dots,C_2\}$ represent the valid Region–Sub-region relations, $\mathcal{E}_{23} \subseteq \{1,\dots,C_2\} \times \{1,\dots,C_3\}$ denote the admissible Sub-region–Finding relations, and $\mathcal{E}_{13} \subseteq \{1,\dots,C_1\} \times \{1,\dots,C_3\}$ denote the induced Region–Finding relations along valid coarse-to-fine paths. Together, they define the set of anatomically valid hierarchical relations that serve as structural priors for cross-level coherence during training:

$$\mathcal{E} = \mathcal{E}_{12} \cup \mathcal{E}_{23} \cup \mathcal{E}_{13}. \tag{1}$$

The 9 Region, 17 Sub-region, and 28 Finding labels induce 197 anatomically valid *Top–Mid–Leaf* paths.

2.2 CHASE Architecture

CHASE adopts a Vision Transformer as its shared backbone (Figure 1). Given an input radiograph x, level-specific linear heads are attached to `[CLS]` embeddings extracted from blocks $k \in \{6, 9, 12\}$, corresponding to the Top, Mid, and Leaf levels respectively:

$$\mathbf{z}_\ell = f_\ell\Big(\mathrm{LN}(\mathbf{h}_{\texttt{cls}}^{(k_\ell)})\Big), \quad \ell \in \{1,2,3\}, \tag{2}$$

where f_ℓ is a linear projection and LN denotes Layer Normalization. Unlike CNNs, ViTs exhibit uniform representations across layers with global information accessible from early blocks [18], meaning a coarse-to-fine hierarchy does

Table 1. Three-level label hierarchy used in our study. Top captures coarse anatomical systems, Mid refines to sub-region and laterality, and Leaf defines specific findings.

Level	Label vocabulary
Top	`No Finding`, `lung`, `heart`, `mediastinum`, `trachea`, `spine`, `abdomen`, `clavicle`, `aorta`
Mid	`abdomen_bilateral_general`, `aorta_bilateral_general`, `clavicle_left_general`, `clavicle_right_general`, `heart_bilateral_general`, `lung_left_general`, `lung_left_lower`, `lung_left_mid`, `lung_left_upper`, `lung_right_general`, `lung_right_lower`, `lung_right_mid`, `lung_right_upper`, `mediastinum_bilateral_general`, `mediastinum_bilateral_upper`, `spine_bilateral_general`, `trachea_bilateral_general`
Leaf	`alveolar hemorrhage`, `aspiration`, `copd/emphysema`, `fluid overload/heart failure`, `goiter`, `granulomatous disease`, `interstitial lung disease`, `lung cancer`, `pericardial effusion`, `pneumonia`, `aorta changes`, `atelectasis`, `bone changes`, `bronchiectasis`, `cardiac changes`, `consolidation`, `cyst/bullae`, `hernia`, `hyperaeration`, `ild pattern`, `lung opacity`, `mass/nodule`, `mediastinal changes`, `pleural effusion`, `pleural scarring`, `pneumothorax`, `pulmonary edema`, `vascular changes`

not emerge naturally with depth—making explicit multi-depth supervision a necessity. We attach heads at blocks 6, 9, and 12 (lower, middle, and final thirds of the 12-block backbone) to balance discriminative semantics and inter-level separation. Shallower representations, such as blocks 1/2/12, may better preserve level-specific separation but are less semantically discriminative, whereas later representations, such as blocks 5/11/12, provide stronger semantics but can make hierarchy-level predictions less distinct. This progressive placement provides granularity-specific supervision for regions, subregions, and findings within a shared encoder, while the core novelty of CHASE lies in the joint training objective described below.

Training Objective The model is trained to achieve per-level accuracy while maintaining cross-level coherence by minimizing:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \alpha\,\mathcal{L}_{\text{KL}} + \beta\,\mathcal{L}_{\text{HVP}}, \tag{3}$$

where $\mathcal{L}_{\text{cls}}$ provides per-level supervision, $\mathcal{L}_{\text{KL}}$ couples all levels by global distribution alignment, and $\mathcal{L}_{\text{HVP}}$ enforces hierarchy consistency under the valid hierarchical relations $\mathcal{E}$. The weights $(\alpha, \beta) = (0.1, 0.3)$ are selected via ablation (Section 3.3).

(i) Multi-level Classification Let $\mathbf{t}_\ell \in \{0,1\}^{C_\ell}$ represent the multi-hot ground-truth labels for level ℓ. We impose independent binary cross-entropy (BCE) over each label dimension at every hierarchy level:

$$\mathcal{L}_{\text{cls}} = \sum_{\ell=1}^{3} \sum_{j=1}^{C_\ell} \left[-t_{\ell,j} \log \hat{p}_{\ell,j} - (1 - t_{\ell,j}) \log(1 - \hat{p}_{\ell,j}) \right], \tag{4}$$

where $\hat{p}_{\ell,j}$ is the predicted probability for the j-th label at level ℓ. This ensures that all three levels develop discriminative representations directly from their own supervision signals.

(ii) Global KL Alignment While $\mathcal{L}_{\text{cls}}$ optimizes each level independently, cross-level coherence requires that the joint predicted distribution remains compatible with the ground-truth active label set. We enforce this by aligning the concatenated prediction distribution with the ground-truth via KL divergence. Concatenating logits and targets yields $\mathbf{z} = [\mathbf{z}_1; \mathbf{z}_2; \mathbf{z}_3]$ and $\mathbf{t} = [\mathbf{t}_1; \mathbf{t}_2; \mathbf{t}_3]$. Probabilities are derived via temperature-scaled sigmoid followed by L1 normalization:

$$\mathbf{p} = \frac{\sigma(\mathbf{z}/\tau)}{\|\sigma(\mathbf{z}/\tau)\|_1 + \varepsilon}, \qquad \mathbf{q} = \frac{\mathbf{t}}{\|\mathbf{t}\|_1 + \varepsilon}, \tag{5}$$

with temperature $\tau = 1$ fixed across all experiments. The KL term is applied to samples with at least one positive label:

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}(\mathbf{q} \,\|\, \mathbf{p}). \tag{6}$$

This regularizer discourages probability mass from concentrating on fine-grained findings that lack support from coarser anatomical contexts.[1]

(iii) Hierarchy-Violation Penalty To suppress parent-child contradictions, we introduce a differentiable regularizer that enforces probabilistic monotonicity across hierarchy levels (Figure 1). Let $\boldsymbol{\pi}_\ell = \sigma(\mathbf{z}_\ell)$ denote the predicted probabilities at level ℓ. Clinically, a finding should not be predicted with higher confidence than its anatomical parent—i.e., valid predictions satisfy $\pi_{\text{Top}} \geq \pi_{\text{Mid}} \geq \pi_{\text{Leaf}}$ along any coarse-to-fine path. For each admissible relation $(p, c) \in \mathcal{E}$, we penalize violations of this monotonicity:

$$\mathcal{L}_{\text{HVP}} = \sum_{(p,c)\in\mathcal{E}} \pi_c \,(1 - \pi_p). \tag{7}$$

The penalty is *asymmetric* by design: it targets the clinically critical error of activating a fine-grained finding without broader anatomical support ($\pi_c > \pi_p$), while leaving the reverse unpunished—reflecting that a radiologist may flag a broad region as abnormal before committing to a specific finding.

3 Results

3.1 Experimental Setup

Dataset and implementation details. We derive a three-level hierarchy (Top/Mid /Leaf) from Chest ImaGenome [20] using MIMIC-CXR [7]. Following the official splits and keeping all frontal views, we obtain 213,361/1,733/3,041 (train/val/test) studies, corresponding to 237,070/1,959/3,403 frontal images after retaining all available frontal images per study. All models are evaluated on a unified 54-label

[1] For normal studies (*No Finding* only), $\mathbf{q}$ degenerates to a coarsest-level one-hot vector, naturally suppressing all fine-grained activations—the clinically desired behaviour for normal radiographs.

Table 2. Main performance comparison on hierarchical classification. **Single-level** models are trained and evaluated on specific hierarchy levels, while **Multi-level** models handle all levels simultaneously. Best results are highlighted in **bold**.

Model	# of Classes	AUC (↑)			ACC (↑)			Consis. (↑)	
		Top	Mid	Leaf	Top	Mid	Leaf	Det.	Prob.
Single-level Evaluation (Top / Mid / Leaf)									
Flat-ViT (Top)	9	0.847±.009	–	–	91.81±.184	–	–	–	–
Flat-ViT (Mid)	17	–	0.805±.007	–	–	87.63±.164	–	–	–
Flat-ViT (Leaf)	28	–	–	0.847±.004	–	–	91.96±.094	–	–
Multi-level Evaluation (Top & Mid & Leaf)									
Flat-ViT	54	0.732±.012	0.702±.009	0.770±.005	91.51±.180	86.61±.174	91.59±.093	92.77±.047	70.16±.065
Hier-ViT	54	0.799±.010	0.764±.007	0.774±.005	91.60±.183	87.25±.166	91.64±.090	93.46±.055	71.12±.101
H-CAST	54	0.785±.011	0.755±.007	0.792±.005	90.81±.191	87.27±.165	91.64±.093	93.42±.055	72.10±.115
Ours (Full)	54	**0.826**±.009	**0.803**±.007	**0.848**±.004	**91.63**±.183	**87.77**±.162	**91.83**±.096	**94.96**±.057	**74.26**±.115

space, with metrics reported at the image level ($N = 3,403$). To ensure performance gains reflect our proposed designs rather than latent representation quality, all models share an identical EVA-X [22] backbone, providing domain-specific visual features via a ViT-based foundation model. We fine-tune end-to-end for 20 epochs with $(\alpha, \beta) = (0.1, 0.3)$[2].

Metrics. We report per-level macro-averaged AUC and accuracy (ACC). To quantify cross-level coherence beyond independent level-wise metrics, we introduce hierarchy consistency (Consis.) metric. As defined in Sec. 2.1, the valid relations $\mathcal{E}$ specify anatomically admissible Top→Mid→Leaf paths. Consis. is the fraction of samples whose predictions are monotone along every such path. Let $\pi_{\ell,j}^{(i)}$ be the predicted probability of label j at level ℓ for sample i, and set $g_{\ell,j}^{(i)} = \mathbf{1}[\pi_{\ell,j}^{(i)} \geq \tau]$ (Consis.-Det.) or $g_{\ell,j}^{(i)} = \pi_{\ell,j}^{(i)}$ (Consis.-Prob.). We define $\text{Consis.}(\tau) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\Big[g_{\text{Top},j}^{(i)} \geq g_{\text{Mid},k}^{(i)} \geq g_{\text{Leaf},l}^{(i)},\ \forall (j,k) \in \mathcal{E}_{12},\ (k,l) \in \mathcal{E}_{23}\Big]$ and report it at $\tau = 0.5$. Consis.-Prob. is a strictly more demanding criterion, as it requires probabilistic monotonicity to hold continuously rather than only after hard thresholding.

Baselines. Prior hierarchical CXR methods [1, 4, 12, 16] rely on mutually incompatible label spaces, making cross-study AUC comparison statistically ill-defined. We instead establish a reproducible 54-label protocol derived from Chest ImaGenome [20], with all baselines evaluated under identical conditions. We compare CHASE against *Flat-ViT* (all 54 labels, no hierarchical bias) and single-level *Flat-ViT (Top/Mid/Leaf)* as oracle performance ceilings—the key measure being how much CHASE narrows the gap relative to multi-level baselines. We further include *Hier-ViT* and *H-CAST* [15], adapted to our label space by replacing classification heads with level-specific projections while preserving all hierarchical inductive biases. Critically, both follow a *fine-to-coarse* aggregation

[2] Hyperparameters: batch size 512, AdamW (LR=1×10^{-5}, WD=0.05), 2-epoch warmup. Results are reported as mean±std over 1,000 bootstrap seeds.

Table 3. Weight ablation of Global KL (α) and Hierarchy-Violation Penalty (β). *Consis.-Det.* and *Consis.-Prob.* denote consistency at τ=0.5 and probabilistic monotonicity, respectively. Best and second-best results are in **bold** and underlined.

Configuration	Weights		AUC (↑)			ACC (↑)			Consis. (↑)	
	α	β	Top	Mid	Leaf	Top	Mid	Leaf	Det.	Prob.
Anchor configurations										
No reg.	0.0	0.0	0.763±.012	0.752±.008	0.791±.007	91.43±.184	87.33±.165	91.95±.094	91.96±.078	64.89±.119
KL only	1.0	0.0	0.819±.009	0.803±.007	0.848±.004	91.01±.191	87.70±.162	**92.10**±.092	91.88±.088	65.26±.162
HVP only	0.0	1.0	0.825±.009	0.801±.007	0.837±.004	91.53±.188	87.50±.162	90.40±.106	**97.57**±.048	**78.72**±.108
Both (1.0, 1.0)	1.0	1.0	0.801±.011	0.800±.007	0.843±.004	90.99±.184	87.71±.165	91.69±.098	95.64±.051	76.63±.096
Both (0.5, 0.5)	0.5	0.5	0.809±.010	0.801±.007	0.846±.004	91.32±.181	87.78±.164	91.82±.096	95.13±.054	75.67±.098
$\alpha\downarrow\beta\uparrow$ *progression*										
$\alpha\downarrow\beta\uparrow$-1	0.5	0.1	0.822±.009	0.804±.007	0.848±.004	91.37±.184	87.79±.164	92.09±.092	93.00±.066	70.45±.136
$\alpha\downarrow\beta\uparrow$-2	0.3	0.1	0.825±.008	**0.805**±.007	**0.849**±.004	91.45±.185	**87.81**±.161	92.09±.093	93.12±.065	70.49±.135
CHASE ($\alpha\downarrow\beta\uparrow$-3)	**0.1**	**0.3**	**0.826**±.009	0.803±.007	0.848±.004	**91.63**±.183	87.77±.162	91.83±.096	94.96±.057	74.26±.115
$\alpha\downarrow\beta\uparrow$-4	0.1	0.5	0.822±.009	0.802±.007	0.846±.004	**91.63**±.184	87.65±.163	91.42±.098	95.95±.055	75.30±.112

strategy—the reverse of our top-down approach—providing a direct ablative contrast for the benefit of coarse-to-fine alignment.

3.2 Hierarchical Classification Performance

Table 2 compares CHASE against both single-level and multi-level models. CHASE achieves the strongest performance across all levels (Top/Mid/Leaf AUC: 0.826 /0.803/0.848), matching or exceeding the oracle single-level ceilings demonstrating that hierarchical training benefits rather than compromises fine-grained classification. The multi-level Flat-ViT shows a large AUC drop (0.847→0.732 at Top), confirming that naïve joint training is insufficient; Hier-ViT and H-CAST recover some performance but remain below CHASE, particularly at Mid and Leaf levels where fine-to-coarse error propagation is most pronounced. Beyond level-wise AUC, CHASE achieves the highest probabilistic consistency (Consis.-Prob.: 74.26%) over the 197 anatomically valid *Top–Mid–Leaf* paths, confirming improved cross-level coherence across the full hierarchy.

3.3 Ablation Studies

Table 3 examines the individual and joint effects of the Global KL weight (α) and Hierarchy-Violation Penalty (β). Without regularization (α=β=0), the model fails to maintain structural logic, yielding low consistency (64.89). KL alignment serves as a calibration signal: aligning predicted distributions with labels improves per-level AUC (e.g., Leaf AUC: 0.791 → 0.848) but yields negligible consistency gains (65.26). In contrast, HVP directly enforces monotonicity; HVP-only (β=1.0) maximizes consistency (78.72) but degrades Leaf AUC (0.837) by over-constraining fine-grained predictions. Our final setting (α=0.1, β=0.3) achieves the most favorable balance, yielding the highest Top AUC (0.826) and strong consistency (74.26) without accuracy degradation. Further increasing β (e.g., 0.5) leads to diminishing returns, where accuracy drops without commensurate gains in coherence.

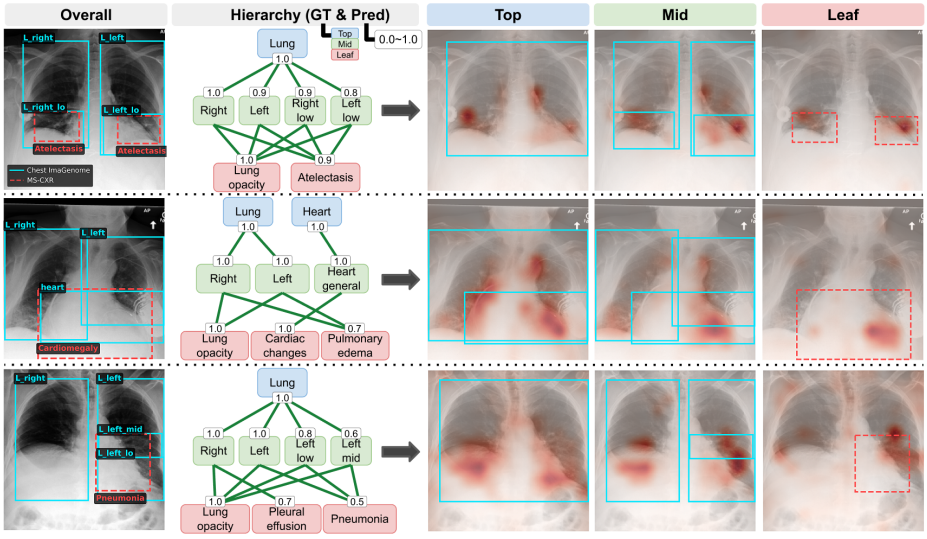

Fig. 2. Hierarchy-consistent localization via attention maps. For each example, we show (left) the input image with reference boxes from Chest ImaGenome (cyan) and MS-CXR (red dashed), (middle) the hierarchy graph with ground-truth nodes and predicted probabilities, and (right) level-wise attention maps derived from ViT self-attention (`[CLS]`→patch) at Blocks 6/9/12 for Top/Mid/Leaf, respectively.

3.4 Hierarchy-Consistent Localization Results

Figure 2 shows how CHASE grounds its predictions in anatomically meaningful evidence via attention maps from Blocks 6, 9, and 12. Leaf-level attention is highly localized, closely matching disease-specific regions and reference bounding boxes. Moving to Mid and Top levels, attention expands to cover predicted subregions and broader anatomical systems. This progression suggests that coarse labels are supported by localized abnormal areas rather than diffuse "whole-organ" cues. The hierarchy panel confirms this probabilistic consistency: Top-level attention highlights the correct anatomical system, Mid-level distributes across corresponding sub-regions, and Leaf-level concentrates on the final finding. This path-level evidence suggests practical clinical utility by making predictions anatomically auditable: radiologists can inspect whether a fine-grained finding is supported by the expected sub-region and broader anatomical system. Collectively, these results confirm that CHASE produces hierarchy-consistent predictions that follow an anatomically coherent coarse-to-fine reasoning process.

4 Conclusions

We introduced CHASE, a single-stage hierarchical multi-label classifier that aligns chest X-ray interpretation with clinical coarse-to-fine reasoning. By at-

taching level-specific heads at multiple ViT depths, CHASE imposes explicit hierarchical supervision that aligns each block toward its assigned anatomical granularity level. To ensure cross-level coherence, we optimized a unified objective combining multi-level supervision with Global KL Alignment and a Hierarchy-Violation Penalty. Empirically, CHASE achieves strong performance across all levels while significantly improving probabilistic consistency. Attention maps further confirm that predictions progress from broad anatomical coverage to localized disease evidence, supporting the model's interpretability. These results demonstrate that clinically aligned hierarchical supervision improves the reliability of fine-grained CXR classification by promoting path-level anatomical coherence. Future work will focus on extending the taxonomy and evaluating robustness across diverse clinical settings.

Acknowledgements This work was supported by the TIPS (Tech Incubator Program for Startup) Program of the Ministry of SMEs and Startups and Korea Startup Foundation under Grant No. RS-2025-25467010.

Disclosure of Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Agu, N.N., Wu, J.T., Chao, H., Lourentzou, I., Sharma, A., Moradi, M., Yan, P., Hendler, J.: 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)
2. Asadi, M., Sodoké, K., Gerard, I.J., Kersten-Oertel, M.: Clinically-inspired hierarchical multi-label classification of chest x-rays with a penalty-based loss function. arXiv preprint arXiv:2502.03591 (2025)
3. Chen, H., Miao, S., Xu, D., Hager, G.D., Harrison, A.P.: Deep hierarchical multi-label classification of chest x-ray images. In: International conference on medical imaging with deep learning. pp. 109–120. PMLR (2019)
4. Chen, H., Miao, S., Xu, D., Hager, G.D., Harrison, A.P.: Deep hiearchical multi-label classification applied to chest x-ray abnormality taxonomies. Medical image analysis **66**, 101811 (2020)
5. Giunchiglia, E., Lukasiewicz, T.: Coherent hierarchical multi-label classification networks. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 33, pp. 9662–9673 (2020)
6. Hanif, M.S., Bilal, M., Alsaggaf, A.H., Al-Saggaf, U.M.: Enhancing multi-label chest x-ray classification using an improved ranking loss. Bioengineering **12**(6), 593 (2025)
7. Johnson, A., Lungren, M., Peng, Y., Lu, Z., Mark, R., Berkowitz, S., Horng, S.: Mimic-cxr-jpg-chest radiographs with structured labels. PhysioNet **101**(215-220), 1 (2019)

8. Kitazono, M.T., Lau, C.T., Parada, A.N., Renjen, P., Miller Jr, W.T.: Differentiation of pleural effusions from parenchymal opacities: accuracy of bedside chest radiography. American Journal of Roentgenology **194**(2), 407–412 (2010)
9. Klein, J.S., Rosado-de Christenson, M.L.: A systematic approach to chest radiographic analysis. Diseases of the Chest, Breast, Heart and Vessels 2019-2022: Diagnostic and Interventional Imaging pp. 1–16 (2019)
10. Levatić, J., Kocev, D., Džeroski, S.: The importance of the label hierarchy in hierarchical multi-label classification. Journal of Intelligent Information Systems **45**(2), 247–271 (2015)
11. Moon, J.H., Choi, G., Rabaey, P., Kim, M.G., Hong, H.G., Lee, J.O., Yoon, H., Doe, E.W., Kim, J., Sharma, H., et al.: Lunguage: A benchmark for structured and sequential chest x-ray interpretation. arXiv preprint arXiv:2505.21190 (2025)
12. Müller, P., Ehlers, G., Ezhov, I., Braren, R., Rueckert, D.: Anatomy-driven pathology detection on chest x-rays. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2023: 26th International Conference, Vancouver, BC, Canada, October 8–12, 2023, Proceedings, Part X. pp. 250–260. Springer (2023)
13. Nobel, J.M., van Geel, K., Robben, S.G.: Structured reporting in radiology: a systematic review to explore its potential. European radiology **32**(4), 2837–2854 (2022)
14. Noor, K.T., Robles-Kelly, A., Zhang, L.Y., Bouadjenek, M.R., Luo, W.: A consistency-aware deep capsule network for hierarchical multi-label image classification. Neurocomputing **604**, 128376 (2024)
15. Park, S., Zhang, Y., Yu, S.X., Beery, S., Huang, J.: Visually consistent hierarchical image classification. arXiv preprint arXiv:2406.11608 (2024)
16. Pham, H.H., Le, T.T., Tran, D.Q., Ngo, D.T., Nguyen, H.Q.: Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels. Neurocomputing **437**, 186–194 (2021)
17. Rabaey, P., Moon, J.H., Lee, J.O., Kim, M.G., Yoon, H., Demeester, T., Choi, E.: Modeling clinical uncertainty in radiology reports: from explicit uncertainty markers to implicit reasoning pathways. arXiv preprint arXiv:2511.04506 (2025)
18. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision transformers see like convolutional neural networks? Advances in neural information processing systems **34**, 12116–12128 (2021)
19. Wehrmann, J., Cerri, R., Barros, R.: Hierarchical multi-label classification networks. In: International conference on machine learning. pp. 5075–5084. PMLR (2018)
20. Wu, J., Agu, N., Lourentzou, I., Sharma, A., Paguio, J., Yao, J.S., Dee, E.C., Mitchell, W., Kashyap, S., Giovannini, A., et al.: Chest imagenome dataset. Physio Net (2021)
21. Yan, S., Cheung, W.K., Tsang, I.W., Chiu, K., Tong, T.M., Cheung, K.C., See, S.: Ahive: anatomy-aware hierarchical vision encoding for interactive radiology report retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14324–14333 (2024)

22. Yao, J., Wang, X., Song, Y., Zhao, H., Ma, J., Chen, Y., Liu, W., Wang, B.: Eva-x: a foundation model for general chest x-ray analysis with self-supervised learning. npj Digital Medicine **8**(1), 678 (2025)
23. Yu, K., Ghosh, S., Liu, Z., Deible, C., Batmanghelich, K.: Anatomy-guided weakly-supervised abnormality localization in chest x-rays. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII. pp. 658–668. Springer (2022)