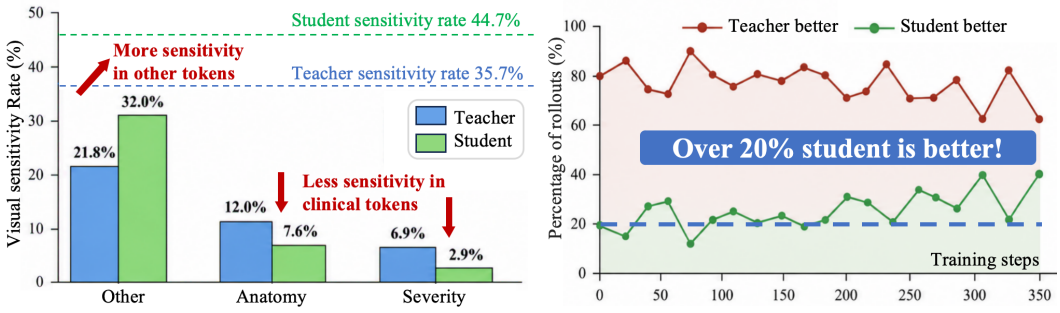
VITAL: VISUAL AND TASK ADVANTAGE LEARNING FOR ON-POLICY DISTILLATION IN MEDICAL DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical vision-language models (VLMs) have shown strong diagnostic performance. On-policy distillation (OPD) transfers these capabilities to smaller VLMs for faster inference and easier deployment. However, existing OPD methods often overlook what the student actually lacks. We uncover two counter-intuitive forms of student deficit: (1) Visual Deficit: misaligned rather than simply weaker visual dependence reduces sensitivity to clinical findings, causing diagnostic errors. (2) Task Deficit, where student rollouts are not consistently inferior to the teacher, and indiscriminate imitation overwrites useful student behaviors and impair generalization. Motivated by these observations, we propose **VITAL**, a **VI**sual and **T**ask-**A**dvantage **L**earning framework for on-policy distillation in medical diagnosis. VITAL improves distillation from two complementary perspectives. Specifically, Visual Advantage Policy optimization (VAP) estimates the teacher's token-level visual dependence through image perturbations and scales distillation accordingly to improve the student's visual dependence. Task Advantage Distillation (TAD) estimates the teacher's task advantage over each student rollout and rejects rollouts without positive teacher advantage, avoiding redundant or potentially harmful imitation. Experiments across four adult and neonatal datasets show VITAL outperforms prior OPD methods in-domain, while boosting out-of-domain F1 by 4.60% (CheXpert Plus) and 2.79% (NeoCXR-EV) over the strongest baseline.

1 INTRODUCTION



(a) Visual deficit (b) Task deficit

Figure 1: Motivations for VITAL in medical diagnosis. (a) Student shows higher visual sensitivity but lower clinical-token dependence. (b) 20% of student's diagnostic rollouts outperform teacher.

Medical disease diagnosis aims to automatically translate medical images into clinically meaningful textual descriptions, with the potential to improve radiology workflow efficiency and reduce reporting burden. Medical vision-language models (VLMs) have achieved strong performance on this task Zhang et al. (2024); Li et al. (2023); Chen et al. (2024); Xu et al. (2025a), but large VLMs incur substantial computational and memory costs that hinder clinical deployment. Compact VLMs are more practical, yet remain considerably weaker. Efficiently transferring capabilities from large VLMs to compact models is therefore important for efficient medical disease diagnosis.

On-policy distillation (OPD) offers a promising approach to this transfer by applying teacher supervision to responses generated by the current student policy Agarwal et al. (2024). By training

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.

on the state distribution induced by the student itself, OPD reduces the mismatch between training and inference in autoregressive generation. Existing OPD methods mainly focus on improving how teacher knowledge is transferred, for example through better distillation objectives, sampling strategies, and teacher–student distribution alignment Gu et al. (2024); Ko et al. (2024); Xu et al. (2025b). Recent extensions to vision-language models further improve multimodal knowledge transfer by introducing more informative visual views or enhancing the teacher's visual sensitivity Bousselham et al. (2026); Yuan et al. (2026); Venkatraman et al. (2026); Liu et al. (2026a). Despite these advances, such methods largely assume that stronger or more effective teacher supervision can better compensate for student weaknesses, while paying less attention to a more fundamental question: what does the student actually lack in disease diagnosis?

Our analysis on MIMIC-CXR dataset Johnson et al. (2019) reveals two counter-intuitive observations. **First, the student's visual limitation is not simply insufficient visual sensitivity, but misaligned visual dependence.** As shown in Fig. 1(a), under controlled image perturbations, the student exhibits response variations comparable to, or even stronger than, those of the teacher. Further analysis shows that such sensitivity is not always allocated to clinically relevant content: generic or templated expressions (other tokens) can be highly responsive to image changes, whereas clinically meaningful findings (anatomy and severity tokens) remain insufficiently grounded in visual evidence. We refer to this phenomenon as **Visual Deficit**. This observation suggests that simply strengthening visual supervision does not address the underlying problem; instead, the key is to correct the student's visual dependence so that it is concentrated on content that should genuinely be determined by the image. **Second, for the same prompt, student-generated responses are not consistently inferior to the teacher.** On-policy sampling produces rollouts with substantially different task quality, and some student responses already match or even outperform the teacher, as shown in Fig. 1(b). Thus, teacher superiority does not hold uniformly across student rollouts. Treating all rollouts equally disperses optimization effort over responses that may not require correction, thereby weakening improvement on genuinely poor responses. More importantly, forcing already strong student responses to imitate the teacher may overwrite useful student behaviors and impair generalization. We refer to this rollout-dependent teacher–student performance gap as **Task Deficit**. This observation suggests that teacher supervision should not be applied indiscriminately, but should depend on whether the teacher retains a genuine advantage over the current student response.

Motivated by these observations, we propose **VITAL**, a novel **VI**sual and **T**ask-**A**dvantage **L**earning framework for on-policy distillation in medical diagnosis. VITAL comprises two complementary components that address the visual and task deficits. Specifically, Visual Advantage Policy Optimization (VAP) uses controlled image perturbations to estimate the teacher's token-level visual dependence and scales the corresponding OPD supervision accordingly, guiding the student toward more appropriate reliance on visual evidence. Task Advantage Distillation (TAD) estimates the teacher's task advantage over each student rollout and filters out rollouts without positive teacher advantage. This preserves useful student behaviors while concentrating distillation on responses that still require improvement. In essence, VAP guides where the student should rely on visual evidence, whereas TAD guides when the student should still learn from the teacher. Together, they shift OPD from indiscriminate teacher imitation toward selective distillation tailored to the student's actual deficits. We evaluate VITAL on two adult and two neonatal diagnosis benchmarks in both in-domain (ID) and out-of-domain (OOD) settings. VITAL achieves competitive ID performance and strong OOD generalization while successfully shifting the student's visual sensitivity from generic, templated language to clinically meaningful findings.

Our main contributions are summarized as follows:

- We identify two student-specific deficits in OPD for medical diagnosis: 1) a visual deficit, where the student's visual dependence is misaligned with clinically relevant details, and 2) a task deficit, where the teacher's advantage varies across student rollouts.
- We propose VITAL, which integrates VAP to improve students' misaligned visual dependence through adaptive token-level OPD supervision, and TAD to selectively distill rollouts with positive teacher task advantage to improve generalization.
- Experiments on four medical diagnosis benchmarks show that VITAL achieves the best overall performance on MIMIC-CXR and NeoCXR, while demonstrating strong generalization with F1 gains of 4.60% and 2.79% on CheXpert Plus and NeoCXR-EV.

2 Related Work

Medical Vision-Language Models. Vision-language models have recently emerged as a promising paradigm for medical image understanding, supporting applications such as medical disease diagnosis, visual question answering, and multimodal clinical reasoning. General-purpose medical VLMs, including BiomedGPT Zhang et al. (2024), LLaVA-Med Li et al. (2023), HuatuoGPT-Vision Chen et al. (2024), and Lingshu Xu et al. (2025a), typically combine large-scale multimodal pretraining with instruction tuning to improve vision–language alignment and medical reasoning. Beyond instruction tuning, Med-Flamingo Moor et al. (2023) explores multimodal few-shot learning in medicine, while Med-PaLM M Tu et al. (2024) integrates clinical language, imaging, and genomics within a unified model. Med-Gemini Yang et al. (2024) further extends these capabilities to chest X-ray and 3D CT report generation, demonstrating the potential of medical multimodal models across diverse imaging settings. In parallel, models tailored to specific imaging scenarios further strengthen domain-specific perception and generation capabilities. For example, LLaVA-Ultra Guo et al. (2024) targets ultrasound understanding, LLaVA-Rad Zambrano Chaves et al. (2025) focuses on clinically accessible chest X-ray interpretation, and RadFM Wu et al. (2025) extends medical multimodal modeling to large-scale 2D and 3D radiological data. These advances substantially improve the ability of VLMs to produce clinically coherent descriptions from medical images. However, existing medical VLMs typically incur substantial computational costs. Their large model size and generation latency hinder practical deployment in resource-constrained environments.

On-policy distillation. On-policy distillation (OPD) transfers knowledge from a larger teacher to a smaller student through feedback on student-generated responses, reducing distribution mismatch while improving task performance at lower inference. Agarwal et al. Agarwal et al. (2024) train students with teacher feedback on self-generated responses. MiniLLM Gu et al. (2024) adopts reverse KL divergence, while DistiLLM Ko et al. (2024) combines skew KL objectives with adaptive off-policy sampling. SKD Xu et al. (2025b) interleaves teacher and student sampling, and DistiLLM-2 Ko et al. (2025) introduces contrastive distillation. However, these methods largely treat teacher supervision as uniformly beneficial across samples, overlooking the substantial quality variation across student rollouts and thus failing to concentrate supervision on the generations that most need correction. More recently, OPD has been extended to multimodal models. VOLD Bousselham et al. (2026) transfers reasoning capabilities from language models to vision-language models through on-policy supervision. CVPD Venkatraman et al. (2026) constructs dense supervision from zoomed and region-removed views. MedOPD Qian et al. (2026) adapts the visual-advantage principle to medical report generation through evidence-aware supervision. However, prior methods primarily strengthen visual perception or grounding by emphasizing visually sensitive tokens. In contrast, VITAL targets misaligned visual dependence, correcting where visual sensitivity is allocated toward clinically relevant content, while selectively distilling rollouts based on the teacher's task advantage.

3 Preliminaries: on-policy distillation

Let π_{S} and π_{T} denote the student and teacher policies, respectively. Given an input $x \sim \mathcal{D}$, the student autoregressively samples a response $y = (y_1, \ldots, y_T)$ according to π_{S}. At each decoding step t, the input and the previously generated tokens form the context $c_t = (x, y_{<t})$. The student predicts the next token using $\pi_{\mathrm{S}}(\cdot \mid c_t)$, while the teacher evaluates the same context to provide a reference distribution $\pi_{\mathrm{T}}(\cdot \mid c_t)$. Thus, teacher feedback is obtained along the student's own generation trajectory. The discrepancy between the student's and teacher's next-token distributions at each step is quantified by the reverse Kullback–Leibler (KL) divergence:

$$D_{\mathrm{KL}}\left(\pi_{\mathrm{S}}(\cdot \mid c_t) \| \pi_{\mathrm{T}}(\cdot \mid c_t)\right) = \sum_{v \in \mathcal{V}} \pi_{\mathrm{S}}(v \mid c_t) \log \frac{\pi_{\mathrm{S}}(v \mid c_t)}{\pi_{\mathrm{T}}(v \mid c_t)}, \tag{1}$$

where $\mathcal{V}$ denotes the vocabulary.

OPD optimizes the student policy by minimizing the reverse KL divergence averaged over decoding steps, in expectation over training inputs and student rollouts:

$$\mathcal{L}_{\mathrm{OPD}} = \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y \sim \pi_{\mathrm{S}}(\cdot|x)} \left[\frac{1}{T} \sum_{t=1}^{T} D_{\mathrm{KL}}\left(\pi_{\mathrm{S}}(\cdot \mid c_t) \| \pi_{\mathrm{T}}(\cdot \mid c_t)\right) \right], \tag{2}$$

where $\mathbb{E}$ denotes expectation over inputs $x \sim \mathcal{D}$ and student rollouts $y \sim \pi_{\mathrm{S}}(\cdot \mid x)$.

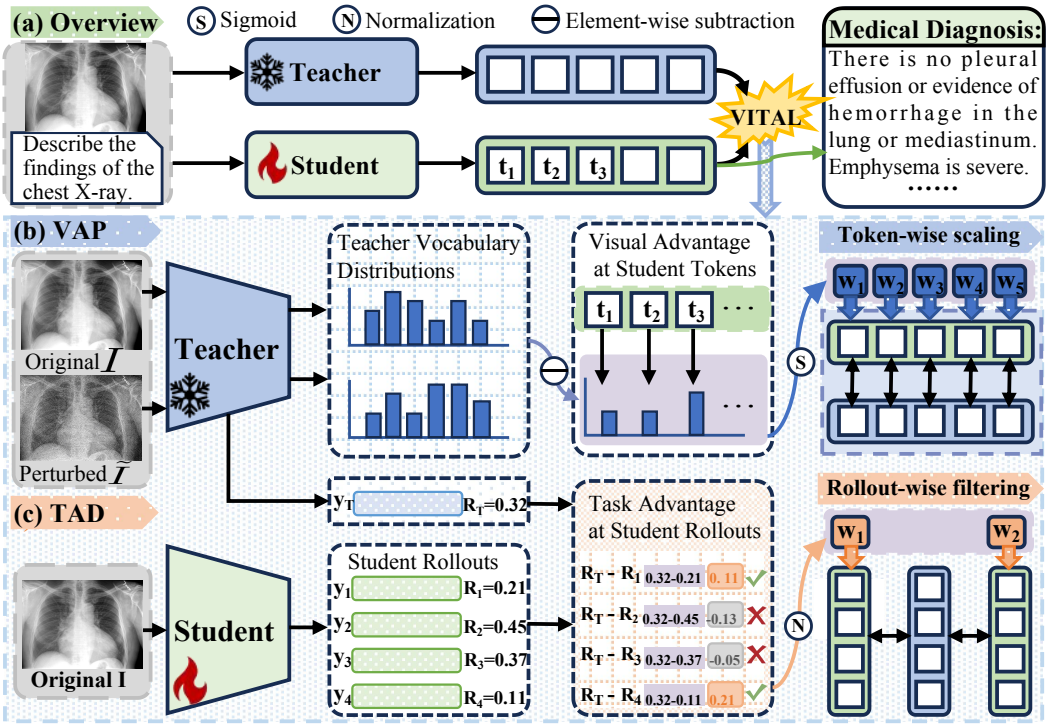


Figure 2: (a) Overview of VITAL, consisting of VAP and TAD for on-policy distillation in medical diagnosis. (b) VAP scales token-level distillation strength to align the student's visual dependence with the teacher's. (c) TAD retains supervision only for rollouts where teacher outperforms student.

4 METHOD

4.1 OVERVIEW

We study on-policy distillation for the diagnosis of medical diseases from chest X-rays. Given a chest radiograph image I and a task instruction q, we form the multimodal prompt $x = (I, q)$ and feed the same input to a frozen teacher policy π_T and a compact student policy π_S. Following on-policy distillation, we sample n responses (referred to as rollouts) from the student:

$$y_i \sim \pi_S(\cdot \mid x), \qquad i = 1, \ldots, n. \tag{3}$$

Each student rollout is evaluated under both the teacher and student policies. For each generated token $y_{i,t}$, where t indexes tokens within rollout, we compute its log-probability under each policy:

$$\log \pi_T \left(y_{i,t} \mid x, y_{i,<t}\right), \qquad \log \pi_S \left(y_{i,t} \mid x, y_{i,<t}\right). \tag{4}$$

Building on this on-policy setup, VITAL (Fig. 2(a)) reduces the teacher–student log-probability gap through two complementary mechanisms: VAP scales token-level updates by the teacher's visual sensitivity, while TAD retains rollouts with positive teacher task advantage.

4.2 VISUAL ADVANTAGE POLICY OPTIMIZATION

To address the student's misaligned, rather than simply insufficient, visual dependence, we propose Visual Advantage Policy Optimization (VAP), which uses teacher visual sensitivity to scale token-level on-policy updates. Following Med-OPD Qian et al. (2026), we estimate teacher visual sensitivity through controlled image perturbations. Given an RGB image $I \in [0,1]^{H \times W \times 3}$, obtained by dividing the original pixel intensities by 255, we add spatially independent Gaussian noise shared across the RGB channels and clip the result to $[0, 1]$:

$$\tilde{I} = \text{clip}\left(I + \epsilon,\, 0,\, 1\right), \qquad \epsilon \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2_{\text{noise}}), \tag{5}$$

where σ_{noise} is the noise standard deviation before clipping.

Let $x = (I, q)$ and $\tilde{x} = (\tilde{I}, q)$ denote the original and perturbed multimodal prompts, respectively. For each student-generated rollout $y_i = (y_{i,1}, \ldots, y_{i,T_i})$, we evaluate the frozen teacher's log-probability of each student token $y_{i,t}$ under both prompts, while keeping the student-generated prefix $y_{i,<t}$ fixed to isolate the visual effect. The resulting log-probability change measures the teacher's visual dependence at each student token, with larger changes indicating stronger visual sensitivity. We then transform these token-level visual-dependence scores into continuous weights $w_{\mathrm{vis},i,t}$ through rollout-wise z-score normalization, sigmoid gating, and unit-mean normalization:

$$v^{\mathrm{T}}_{i,t} = \left| \log \pi_{\mathrm{T}}(y_{i,t} \mid x, y_{i,<t}) - \log \pi_{\mathrm{T}}(y_{i,t} \mid \tilde{x}, y_{i,<t}) \right|, \; w_{\mathrm{vis},i,t} = \mathrm{MeanNorm}\left[\sigma\big(z(v^{\mathrm{T}}_{i,t})/\tau\big)\right], \quad (6)$$

where z and $\mathrm{MeanNorm}$ apply rollout-wise z-score and unit-mean normalization respectively, σ is the sigmoid function, and $\tau > 0$ is the temperature parameter.

To incorporate this visual dependency into OPD, we use the teacher–student log-probability ratio as a token-level advantage signal and scale it by $w_{\mathrm{vis},i,t}$. The resulting advantage $A^{\mathrm{vis}}i, t$ guides the student's policy updates: the log-probability ratio determines the update direction, while the positive visual weight controls its strength. We stop gradients through the advantage during optimization. To account for policy lag in asynchronous rollout generation, we also define the importance ratio $ri, t(\theta)$ relative to the behavior policy π_{old} that generated the rollout:

$$A^{\mathrm{vis}}_{i,t} = \mathrm{sg}\Big(w_{\mathrm{vis},i,t} \log \tfrac{\pi_{\mathrm{T}}(y_{i,t}|x,y_{i,<t})}{\pi_{\mathrm{S},\theta}(y_{i,t}|x,y_{i,<t})}\Big), \; r_{i,t}(\theta) = \tfrac{\pi_{\mathrm{S},\theta}(y_{i,t}|x,y_{i,<t})}{\pi_{\mathrm{old}}(y_{i,t}|x,y_{i,<t})}. \quad (7)$$

Higher teacher sensitivity therefore strengthens the signal to encourage or suppress sampled token, depending on discrepancy's sign. We optimize student using a clipped policy-gradient objective:

$$\mathcal{L}_{\mathrm{VAP}} = -\tfrac{1}{n}\textstyle\sum_{i=1}^{n} \tfrac{1}{T_i} \sum_{t=1}^{T_i} \min\Big(r_{i,t}(\theta)A^{\mathrm{vis}}_{i,t}, \mathrm{clip}\left(r_{i,t}(\theta), 1-\epsilon_{\mathrm{clip}}, 1+\epsilon_{\mathrm{clip}}\right)A^{\mathrm{vis}}_{i,t}\Big), \quad (8)$$

where ϵ_{clip} controls the clipping range.

4.3 Task Advantage Distillation

VAP corrects the student's misaligned visual dependence, but does not assess whether teacher supervision benefits each student response. Since the student can match or surpass the teacher on individual rollouts, indiscriminate imitation may suppress effective student behaviors rather than correct errors. This motivates a selective distillation mechanism based on the teacher's relative task advantage. We therefore introduce TAD, which retains teacher guidance only for student responses with lower task scores than the teacher's, avoiding unnecessary imitation on competitive rollouts. Since diagnostic responses take the form of descriptive paragraphs, we use ROUGE-L Lin (2004) as a proxy for task performance, which shows a strong positive correlation with clinical F1 on NeoCXR (Pearson's $r = 0.751$). For each prompt, we compute the teacher score R_{T} and student scores $R_{\mathrm{S},i}$ by comparing their responses y_{T} and y_i against the same reference y^*. The score difference $R_{\mathrm{T}} - R_{\mathrm{S},i}$ serves as an estimate of the teacher's relative task advantage. Based on this advantage, we apply rejection-based filtering to retain student rollouts with positive teacher task advantage and reject those with zero or negative advantage, focusing distillation on responses where the teacher scores higher. The resulting accepted rollout set for each input prompt is denoted by $\mathcal{I}_+$:

$$R_{\mathrm{T}} = \text{ROUGE-L}(y_{\mathrm{T}}, y^*), \qquad R_{\mathrm{S},i} = \text{ROUGE-L}(y_i, y^*), \quad (9)$$

$$A_i = \max\left\{R_{\mathrm{T}} - R_{\mathrm{S},i},\, 0\right\}, \qquad \mathcal{I}_+ = \left\{i \in \{1,\ldots,n\} \mid A_i > 0\right\}. \quad (10)$$

We weight accepted rollouts by their task advantages and normalize these weights to unit mean:

$$\hat{w}_{\mathrm{task},i} = \frac{A_i}{\frac{1}{|\mathcal{I}_+|}\sum_{j\in\mathcal{I}_+} A_j} \quad \text{if } i \in \mathcal{I}_+, \text{ and 0 otherwise.} \quad (11)$$

We then use $\hat{w}_{\mathrm{task},i}$ to scale the distillation of each rollout based on task advantage. Specifically, we compute the token-level log-probability gap $\Delta_{i,t}$ between the teacher and the student, and aggregate the resulting divergence over the sequence and across all n rollouts:

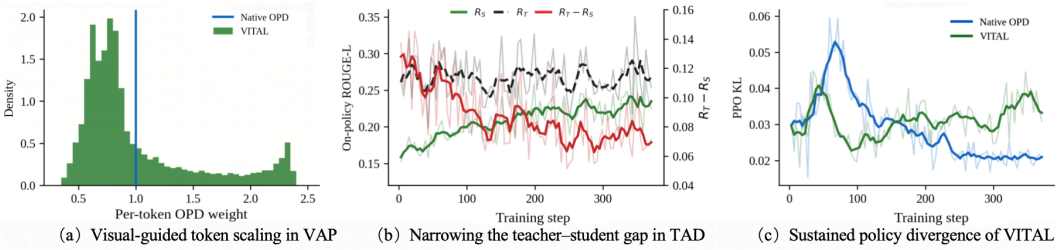
$$\Delta_{i,t} = \log \pi_{\mathrm{T}}\left(y_{i,t} \mid x, y_{i,<t}\right) - \log \pi_{\mathrm{S},\theta}\left(y_{i,t} \mid x, y_{i,<t}\right), \quad (12)$$

$$\mathcal{L}_{\mathrm{TAD}} = \frac{1}{n}\sum_{i=1}^{n} \hat{w}_{\mathrm{task},i}\left[\frac{1}{T_i}\sum_{t=1}^{T_i}\left(\exp(\Delta_{i,t}) - \Delta_{i,t} - 1\right)\right], \quad (13)$$

where the surrogate $\exp(\Delta_{i,t}) - \Delta_{i,t} - 1$ is pointwise nonnegative for all $\Delta_{i,t} \in \mathbb{R}$, so positive and negative teacher–student gaps cannot cancel during token aggregation.

Table 1: Main results on adult disease diagnosis on MIMIC-CXR (in-domain) and CheXpert Plus (out-of-domain) datasets. **Bold** denotes the best results among 2B student models.

Dataset	Model	R-L (↑)	R-1 (↑)	MET (↑)	RaTE (↑)	P (↑)	R (↑)	F1 (↑)	Avg. (↑)
MIMIC-CXR *In-Domain*	Qwen3-VL-8B Teacher	30.17	31.98	22.71	50.37	39.87	55.48	46.40	39.57
	Qwen3-VL-2B Student	17.73	19.25	20.17	49.08	35.27	24.87	29.17	27.93
	Native OPD	27.93	29.31	**21.09**	50.34	**42.79**	39.81	41.25	36.07
	MedOPD	24.39	25.67	18.34	45.91	37.16	31.39	34.03	30.98
	SKD	27.01	28.24	20.71	49.62	42.18	37.96	39.96	35.10
	VITAL (Ours)	**27.94**	**29.33**	20.54	**51.14**	42.68	**40.14**	**41.37**	**36.16**
CheXpert Plus *Out-of-Domain*	Qwen3-VL-8B Teacher	25.30	26.17	16.32	44.46	30.37	52.61	38.51	33.39
	Qwen3-VL-2B Student	18.92	19.98	17.28	42.76	36.45	38.71	37.55	30.24
	Native OPD	24.78	25.59	**17.46**	41.95	34.40	37.47	35.87	31.07
	MedOPD	21.73	22.52	13.91	38.43	33.89	34.99	34.43	28.56
	SKD	20.23	20.94	16.62	**43.78**	34.95	35.12	35.03	29.53
	VITAL (Ours)	**25.03**	**25.88**	17.02	43.26	**38.27**	**42.93**	**40.47**	**33.27**



(a) Visual-guided token scaling in VAP (b) Narrowing the teacher–student gap in TAD (c) Sustained policy divergence of VITAL

Figure 3: Mechanism analysis. (a) VAP adjusts token weights using visual sensitivity. (b) TAD narrows the teacher–student reward gap. (c) VITAL maintains higher policy KL in later training.

4.4 Training Objective

By integrating VAP and TAD, VITAL tackles visual and task deficits through a unified objective:

$$\mathcal{L}_{\text{VITAL}} = \mathcal{L}_{\text{VAP}} + \alpha \mathcal{L}_{\text{TAD}}, \tag{14}$$

where α serves as a hyperparameter to trade off gradient contributions between VAP and TAD.

4.5 Experimental Setup

Training Setup. We use Qwen3-VL-8B-Instruct as the teacher and Qwen3-VL-2B-Instruct as the student. Both models first undergo cold-start SFT for 1 epoch, after which the teacher is further post-trained with GSPO to serve as a robust reasoning oracle. For fair comparison, students across all distillation methods are initialized from the exact same cold-start checkpoint. During distillation, both models adopt full-parameter tuning with frozen vision encoders, while the teacher's weights remain strictly fixed. The cold-start and distillation datasets maintain a size ratio of 1:3.

During on-policy distillation, the student samples $n = 4$ report rollouts per prompt from its current policy. For the specific mechanisms in VITAL, we empirically set the Gaussian noise standard deviation $\sigma_{\text{noise}} = 0.1$, the visual gating temperature $\tau = 0.5$, and the TAD loss weight $\alpha = 2$. Comprehensive data construction and training hyperparameters (e.g., learning rates, batch configurations) are detailed in Appendix A and B.

Datasets and Evaluation Protocol. We evaluate VITAL on four chest X-ray disease diagnosis datasets spanning adult and neonatal settings, with one in-domain and one out-of-domain dataset per setting. For adult disease diagnosis, we use MIMIC-CXR Johnson et al. (2019) for training and its official test split for in-domain evaluation. For OOD evaluation, we use the official validation split of CheXpert Plus Chambon et al. (2024). For Neonatal disease diagnosis, we train on NeoCXR Liu et al. (2026b) and evaluate in-domain performance on its test split. For OOD evaluation, we use NeoCXR-EV Liu et al. (2026b) collected from another clinical source.

Table 2: Main results on neonatal disease diagnosis on NeoCXR and NeoCXR-EV datasets. VITAL demonstrates strong OOD generalization while maintaining competitive in-domain performance.

Dataset	Model	R-L (↑)	R-1 (↑)	MET (↑)	RaTE (↑)	P (↑)	R (↑)	F1 (↑)	Avg. (↑)
NeoCXR *In-Domain*	Qwen3-VL-8B Teacher	61.40	62.10	60.23	65.18	57.90	78.51	66.65	64.57
	Qwen3-VL-2B Student	28.45	28.60	18.25	26.96	21.18	3.27	5.67	18.91
	Native OPD	44.85	45.34	44.14	60.40	**55.94**	54.17	55.04	51.41
	MedOPD	46.13	46.68	45.27	59.51	55.83	54.83	**55.33**	51.94
	SKD	45.98	46.47	45.43	**60.88**	54.51	**55.23**	54.87	51.91
	VITAL (Ours)	**46.32**	**46.80**	**45.76**	59.91	55.19	55.19	55.19	**52.05**
NeoCXR-EV *Out-of-Domain*	Qwen3-VL-8B Teacher	40.68	41.91	40.25	51.55	52.66	57.11	54.79	48.42
	Qwen3-VL-2B Student	24.36	24.67	16.34	20.31	11.93	1.27	2.30	14.45
	Native OPD	31.63	32.52	31.18	49.75	50.23	43.44	46.59	40.76
	MedOPD	29.68	30.62	28.66	48.43	50.59	37.28	42.93	38.31
	SKD	32.58	33.57	32.31	**50.37**	48.65	**45.79**	47.18	41.49
	VITAL (Ours)	**32.91**	**33.84**	**32.43**	50.33	**55.14**	45.68	**49.97**	**42.90**

Table 3: Ablation study on CheXpert Plus and NeoCXR-EV.

Dataset	VAP	TAD	R-L (↑)	R-1 (↑)	MET (↑)	RaTE (↑)	P (↑)	R (↑)	F1 (↑)	Avg. (↑)
CheXpert Plus			18.92	19.98	17.28	42.76	36.45	38.71	37.55	30.24
		✓	24.75	25.74	16.99	41.59	34.07	36.42	35.21	30.68
	✓		24.23	25.17	16.76	43.12	34.51	37.98	36.16	31.13
	✓	✓	**25.03**	**25.88**	17.02	**43.26**	**38.27**	**42.93**	**40.47**	**33.27**
NeoCXR-EV			24.36	24.67	16.34	20.31	11.93	1.27	2.30	14.45
		✓	30.46	31.39	30.29	50.12	46.94	42.46	44.59	39.46
	✓		28.81	29.73	28.01	48.67	46.38	34.16	39.34	36.44
	✓	✓	**32.91**	**33.84**	**32.43**	**50.33**	**55.14**	**45.68**	**49.97**	**42.90**

We evaluate diagnosis quality using ROUGE-L Lin (2004), ROUGE-1 Lin (2004), METEOR Banerjee & Lavie (2005), and assess clinical factuality with RaTE Zhao et al. (2024). Following prior work Müller et al. (2024), we measure Clinical Efficacy using micro-averaged precision, recall, and F1.

4.6 Comparison with State-of-the-Art Methods

We compare VITAL against representative distillation baselines, including Native OPD Agarwal et al. (2024), SKD Xu et al. (2025b) for LLM-based OPD, and MedOPD Qian et al. (2026) for multimodal OPD. As shown in Tables 1 and 2, VITAL consistently improves the 2B student across all four datasets, increasing the average score from 27.93% to 36.16% on MIMIC-CXR, from 30.24% to 33.27% on CheXpert Plus, from 18.91% to 52.05% on NeoCXR, and from 14.45% to 42.90% on NeoCXR-EV. Compared with existing distillation methods, VITAL achieves the best overall in-domain average performance on both MIMIC-CXR and NeoCXR, while obtaining the highest F1 and average scores on both OOD datasets. On CheXpert Plus, VITAL achieves an F1 of 40.47%, outperforming the strongest competing baseline by 4.60 percentage points. On NeoCXR-EV, it reaches an F1 of 49.97%, surpassing SKD by 2.79 percentage points. VITAL also improves the average score over the strongest competing baselines by 2.20 and 1.41 percentage points on CheXpert Plus and NeoCXR-EV, respectively. These consistent gains demonstrate the effectiveness of VITAL under both in-domain and distribution-shifted settings. Additional qualitative comparisons and detailed analyses are provided in Appendix C.

4.7 Ablation Studies

Ablation of VAP and TAD. We evaluate the individual and joint contributions of VAP and TAD on CheXpert Plus and NeoCXR-EV, as shown in Table 3. We use the student model as baseline to isolate the contribution of each component. On CheXpert Plus, the TAD-only and VAP-only variants improve the average score from 30.24% to 30.68% and 31.13%, respectively, while combining them further raises it to 33.27%. Similar trends are observed on NeoCXR-EV, where TAD-only and VAP-only achieve average scores of 39.46% and 36.44%, substantially outperforming the baseline score

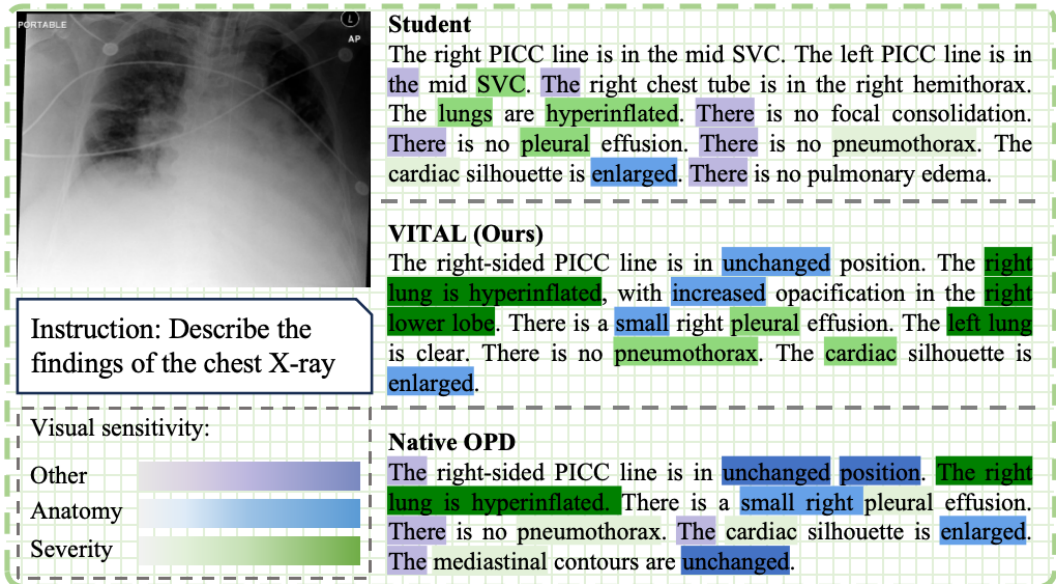


Figure 4: Effectiveness of VAP on the MIMIC-CXR dataset. VITAL concentrates visual sensitivity on clinical tokens (e.g., anatomy and severity) rather than non-clinical fillers like "the".

Table 4: Performance under different noise standard deviations σ_{noise} on MIMIC-CXR dataset.

σ_{noise}	**R-L** (↑)	**R-1** (↑)	**MET** (↑)	**RaTE** (↑)	**P** (↑)	**R** (↑)	**F1** (↑)	**Avg.** (↑)
0.05	26.86	28.08	20.32	49.51	43.31	38.41	40.71	35.32
0.10	**27.94**	**29.33**	**20.54**	**51.14**	**42.68**	**40.14**	**41.37**	**36.16**
0.15	26.04	27.21	19.62	47.73	41.74	36.97	39.21	34.07
0.20	25.36	26.50	19.17	47.54	41.12	34.96	37.79	33.21

of 14.45%. Their combination yields the best average score of 42.90%, exceeding the stronger individual variant by 3.44 percentage points. These results demonstrate that VAP and TAD provide complementary benefits when jointly optimized within VITAL.

Mechanism Analysis of VAP and TAD. Fig. 3 dissects the scaling behavior of VAP alongside the training dynamics of TAD and the overall VITAL framework. As shown in Fig. 3(a), unlike the rigid uniform token weighting (weight = 1.0) employed in Native OPD, VAP adaptively redistributes token weights based on visual sensitivity. It actively downweights tokens lacking visual grounding while assigning a pronounced long-tail of higher weights (> 1.0) to tokens with strong visual dependence. This dynamic scaling encourages the student to ground its reasoning in genuine visual evidence rather than linguistic priors. Furthermore, Fig. 3(b) illustrates the training progression under TAD: the student's on-policy ROUGE-L score (R_S) steadily climbs, driving a continuous narrowing of the teacher–student reward gap ($R_T - R_S$). This convergence aligns with TAD's objective of learning from the teacher's relative advantages. Finally, Fig. 3(c) reveals a key property of VITAL's policy evolution. The rapid decrease in policy KL during early training suggests that TAD accelerates convergence by focusing teacher supervision on underperforming student rollouts. In later stages, VITAL maintains a higher KL than Native OPD, suggesting that VITAL maintains greater divergence from the teacher rather than converging toward uniform imitation.

Empirical Validation of VAP and TAD. Fig. 4 illustrates the impact of the VAP module on token-level visual dependency. Unlike the student and Native OPD model, which waste visual dependence on generic, non-clinical tokens (e.g., "The"), VITAL successfully recalibrates the model's attention. It highly concentrates sensitivity on clinically relevant expressions—specifically anatomical regions (green) and disease severity (blue). This contrast confirms that VAP guides the student to rely on genuine visual evidence for medical reasoning, rather than hallucinating from linguistic priors (detailed definitions and extended comparisons are in Appendix C). Fig. 5 dissects TAD's effectiveness by decoupling samples based on relative teacher-student expertise. On teacher-better samples

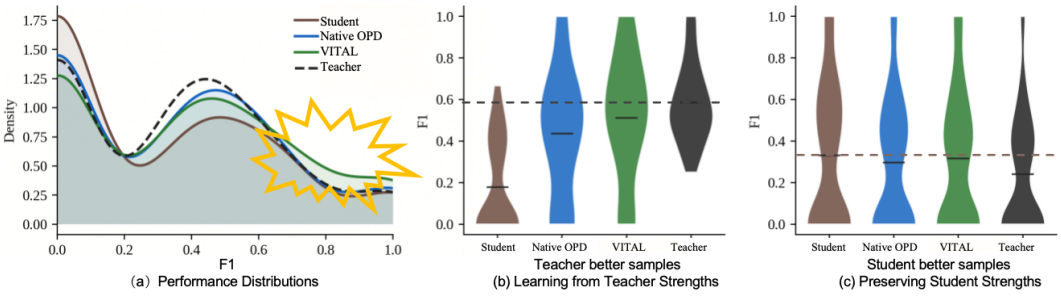


Figure 5: Effectiveness of TAD on CheXpert Plus measured by F1. VITAL surpasses native OPD on student-better samples, supporting TAD's preservation of beneficial student behaviors.

Table 5: Sensitivity analysis of the loss weight α on NeoCXR and NeoCXR-EV datasets.

Dataset	Loss weight α	R-L (↑)	R-1 (↑)	MET (↑)	RaTE (↑)	P (↑)	R (↑)	F1 (↑)	Avg. (↑)
NeoCXR	0.5	44.81	45.27	44.17	**60.14**	55.38	55.43	55.40	51.51
	1.0	44.87	45.38	44.28	59.41	56.14	**56.04**	**56.09**	51.74
	2.0	**46.32**	**46.80**	**45.76**	59.91	55.19	55.19	55.19	**52.05**
	4.0	44.35	44.82	43.65	59.24	**56.33**	53.77	55.02	51.03
NeoCXR-EV	0.5	32.36	33.30	32.18	50.33	45.32	43.79	44.54	40.26
	1.0	30.73	31.60	30.20	50.03	48.38	42.05	44.99	39.71
	2.0	**32.91**	**33.84**	**32.43**	50.33	**55.14**	**45.68**	**49.97**	**42.90**
	4.0	30.67	31.54	30.59	**50.54**	47.12	43.17	45.06	39.81

(Fig. 5b), VITAL exhibits a notably higher clinical F1 median and a denser high-score distribution than Native OPD, demonstrating superior capability in acquiring knowledge from the teacher's strengths. Conversely, on student-better samples (Fig. 5c), Native OPD suffers severe degradation, dragged down by inferior teacher signals. In stark contrast, VITAL successfully maintains a higher center and preserves the student's beneficial distribution. These results confirm that TAD dynamically acts as a selective filter: effectively assimilating teacher advantages while defending against flawed guidance to preserve inherent student strengths.

Hyperparameter Analysis. We examine VITAL's sensitivity to two key hyperparameters through controlled parameter sweeps: the Gaussian noise standard deviation σ_{noise} and the TAD coefficient α. We first vary σ_{noise} from 0.05 to 0.20. As shown in Table 4, VITAL achieves the best overall performance at $\sigma_{\text{noise}} = 0.10$. Furthermore, Table 5 reveals that the model is notably more sensitive to α on NeoCXR-EV (ranging from 39.71% to 42.90%) than on NeoCXR (51.03% to 52.05%). This distinct behavior suggests that TAD has a greater impact on OOD performance, further highlighting its role in improving model generalization under substantial distribution shifts. Setting $\alpha = 2$ yields the optimal average score on both datasets, fully justifying our empirical choice.

5 Conclusion and Limitations

In this paper, we identify two critical bottlenecks when distilling large multimodal models for medical diagnosis: the visual deficit, where students hallucinate from linguistic priors rather than relying on visual evidence, and the task deficit, where forcing students to mimic flawed teacher behaviors degrades performance. To address these, we propose VITAL, a novel on-policy distillation framework. Specifically, VAP mitigates the visual deficit by adaptively scaling token-level supervision based on visual reliance, while TAD overcomes the task deficit by selectively supervising student rollouts only when the teacher exhibits a positive task advantage. Together, they tailor the distillation process to the student's actual needs instead of indiscriminately imitating the teacher. Extensive experiments across four adult and neonatal chest X-ray datasets demonstrate that VITAL achieves competitive overall in-domain performance and strong out-of-domain generalization.

Limitations. Despite encouraging results, VITAL has primarily been evaluated on chest X-ray diagnosis, and its applicability to other medical modalities and tasks warrants further exploration across broader clinical settings. Moreover, VAP characterizes visual dependence mainly through textual signals. Extending it to region- or pixel-level pathological grounding is a promising direction that requires addressing challenges in fine-grained visual understanding and annotation costs.

AI USE STATEMENT

Generative AI was used solely for language refinement, including improving grammar, clarity, and readability. The authors reviewed all edits and take full responsibility for the final content.

REPRODUCIBILITY STATEMENT

Detailed experimental setups and dataset configurations are provided in Section 4.5 and Appendices A and B. The source code is anonymously available at `https://anonymous.4open.science/r/VITAL-for-Medical-Diagnosis-CDD4`.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Walid Bousselham, Hilde Kuehne, and Cordelia Schmid. Vold: Reasoning transfer from llms to vision-language models via on-policy distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26209–26218, 2026.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P Langlotz. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *arXiv preprint arXiv:2405.19538*, 2024.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, et al. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 7346–7370, 2024.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Xuechen Guo, Wenhao Chai, Shi-Yan Li, and Gaoang Wang. Llava-ultra: Large chinese language and vision assistant for ultrasound. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 8845–8854, 2024.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. *arXiv preprint arXiv:2402.03898*, 2024.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. Distillm-2: A contrastive approach boosts the distillation of llms. *arXiv preprint arXiv:2503.07067*, 2025.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in neural information processing systems*, 36:28541–28564, 2023.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Ruiqi Liu, Xiaolei Lv, Gengsheng Li, Ximo Zhu, Zhiheng Wang, Zhengbo Zhang, Junkai Chen, Zhiheng Li, Bo Li, Jun Gao, et al. Visual-advantage on-policy distillation for vision-language models. *arXiv preprint arXiv:2605.21924*, 2026a.

Yinan Liu, Hongtai Xia, Haoran Xu, Jiankang Hong, Jingkuan Song, and Ye Luo. Neored: A knowledge-logic-alignment multimodal large language model for neonatal respiratory disease diagnosis. *arXiv preprint arXiv:2609.03527*, 2026b.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine learning for health (ML4H)*, pp. 353–367. PMLR, 2023.

Philip Müller, Georgios Kaissis, and Daniel Rueckert. Chex: Interactive localization and region description in chest x-rays. In *European Conference on Computer Vision*, pp. 92–111. Springer, 2024.

Yunhang Qian, Jiaquan Yu, Jiawei Liu, Meng Wang, Hongwei Bran Li, and Xiaobin Hu. Medopd: Improving medical vision-language models via evidence-aware on-policy distillation. *arXiv preprint arXiv:2607.16303*, 2026.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024.

Shravan Venkatraman, Omkar Thawakar, Ritesh Thawkar, Abdelrahman Shaker, and Rao Muhammad Anwer. Perception before supervision: Self-contained visual distillation from counterfactual blind spots. *arXiv preprint arXiv:2608.09931*, 2026.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866, 2025.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025a.

Wenda Xu, Rujun Han, Zifeng Wang, Long Le, Dhruv Madeka, Lei Li, William Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. In *International Conference on Learning Representations*, volume 2025, pp. 64616–64646, 2025b.

Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, et al. Advancing multimodal medical capabilities of gemini. *arXiv preprint arXiv:2405.03162*, 2024.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, et al. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16(1):3108, 2025.

Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nature medicine*, 30(11):3129–3141, 2024.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Ratescore: A metric for radiology report generation. *arXiv preprint arXiv:2406.16845*, 2024.

Table 6: Dataset sizes and usage across training and evaluation stages. Counts refer to entries in the corresponding split lists. GSPO and distillation reuse the same training examples.

Dataset	Cold-start SFT	GSPO	Distillation	Evaluation	Evaluation setting
MIMIC-CXR	1,000	3,000	3,000	3,398	Adult ID
CheXpert Plus[†]	—	—	—	202	Adult OOD
NeoCXR	1,054	3,165	3,165	1,434	Neonatal ID
NeoCXR-EV	—	—	—	1,065	Neonatal OOD

[†] Despite its relatively small size (202 instances), CheXpert Plus evaluation set is widely adopted as benchmark by authoritative medical models (e.g., Lingshu Xu et al. (2025a), HuluMedJiang et al. (2025)).

A Additional Data Details

Table 6 summarizes dataset sizes and usage across training and evaluation stages. For MIMIC-CXR, we sample 4,000 examples with random seed 42 and partition them into disjoint subsets: 1,000 for cold-start SFT and 3,000 shared by GSPO and distillation. We use 4,908 development examples for validation and the official frontal test split of 3,398 examples for evaluation. For NeoCXR, the SFT and GSPO splits contain 1,054 and 3,165 examples, respectively, with the latter also used for distillation. For OOD evaluation, we use the official CheXpert Plus validation set of 202 examples and 1,065 external evaluation examples from NeoCXR-EV. Neither OOD dataset is used for training.

B Additional Training Details

Teacher training. We train the Qwen3-VL-8B-Instruct teacher in two stages. First, supervised fine-tuning minimizes the negative log-likelihood of reference diagnosis:

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(x,y^*)\sim\mathcal{D}_{\mathrm{SFT}}}\left[\frac{1}{T^*}\sum_{t=1}^{T^*}\log \pi_\theta(y_t^* \mid x, y_{<t}^*)\right], \tag{15}$$

where y^* is the reference diagnosis and T^* is its length.

We then apply Group Sequence Policy Optimization (GSPO). For each prompt, we sample G responses from the behavior policy $\pi_{\theta_{\mathrm{old}}}$ and compute group-normalized advantages:

$$\hat{A}_i = \frac{r(x,y_i) - \bar{r}}{\mathrm{Std}_{j=1}^{G}[r(x,y_j)] + \varepsilon}, \qquad \bar{r} = \frac{1}{G}\sum_{j=1}^{G} r(x,y_j), \tag{16}$$

where the task reward is $r(x,y_i) = 0.5\,\mathrm{ROUGE\text{-}L}(y_i,y^*) + 0.5\,\mathrm{F1}(y_i,y^*)$, combining textual overlap and clinical F1 against the reference diagnosis y^*, and $\varepsilon > 0$ ensures numerical stability. GSPO uses a length-normalized sequence-level importance ratio:

$$\rho_i(\theta) = \exp\left[\frac{1}{T_i}\sum_{t=1}^{T_i}\log\frac{\pi_\theta(y_{i,t} \mid x, y_{i,<t})}{\pi_{\theta_{\mathrm{old}}}(y_{i,t} \mid x, y_{i,<t})}\right]. \tag{17}$$

The clipped policy objective is:

$$\mathcal{J}_{\mathrm{GSPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\min\left(\rho_i(\theta)\hat{A}_i,\ \mathrm{clip}\left(\rho_i(\theta), 1-\epsilon_{\mathrm{low}}, 1+\epsilon_{\mathrm{high}}\right)\hat{A}_i\right)\right]. \tag{18}$$

Teacher training hyperparameters are summarized in Table 7.

Student training. We initialize Qwen3-VL-2B-Instruct through cold-start SFT using the objective in Eq. 15. All distillation methods start from the same SFT student checkpoint. The vision encoder remains frozen during both SFT and distillation, while the remaining model parameters are updated. Student training hyperparameters are summarized in Table 8.

We also provide the complete distillation procedure of VITAL in Algorithm 1. At each step, the student generates on-policy rollouts, which are scored by the frozen teacher under original and perturbed images. VAP uses teacher visual sensitivity to weight token-level updates, while TAD applies additional distillation to rollouts with positive teacher task advantage. The student is updated using the combined objective $\mathcal{L}_{\mathrm{VAP}} + \alpha\mathcal{L}_{\mathrm{TAD}}$.

Table 7: Training hyperparameters for the 8B teacher (cold-start SFT then GSPO).

Hyperparameter	**Teacher SFT**	**Teacher GSPO**
Backbone	Qwen3-VL-8B-Instruct	Qwen3-VL-8B-Instruct
Initialization	Pretrained checkpoint	Cold-start checkpoint
Training strategy	Full-param. (freeze ViT)	Full-param. (freeze ViT)
Training duration	1 epoch	600 steps
Learning rate	5×10^{-6}	2×10^{-7}
Batch configuration	Global 16, micro 1	Train 2, rollout group 8
Mini-batch size	—	4
Optimizer	AdamW ($\beta = [0.9, 0.95]$)	AdamW ($\beta = [0.9, 0.95]$)
Weight decay	0.1	0.1
Learning-rate schedule	10% warm-up + cosine	Constant
Gradient clip norm	1.0	1.0
PPO clip ratio	—	3×10^{-4} / 4×10^{-4}
Max. prompt length	8192	8192
Max. response length	—	4096
Entropy coefficient	—	0.001
GPU	4×A800	4×A800

Table 8: Training hyperparameters for the 2B student (cold-start SFT then OPD).

Hyperparameter	**Student SFT (Cold Start)**	**OPD**
Backbone	Qwen3-VL-2B-Instruct	Qwen3-VL-2B-Instruct
Initialization	Pretrained checkpoint	Cold-start checkpoint
Training strategy	Full-param. (freeze ViT)	Full-param. (freeze ViT)
Training duration	1 epoch	1500 rollout steps
Learning rate	5×10^{-6}	5×10^{-7}
Batch configuration	Global 4, micro 1	Gen 1, PPO mini 4, n=4
Mini-batch size	—	4
Optimizer	AdamW ($\beta = [0.9, 0.95]$)	AdamW ($\beta = [0.9, 0.95]$)
Weight decay	0.1	0.1
Learning-rate schedule	10% warm-up + cosine	Constant
Gradient clip norm	1.0	—
PPO clip ratio	—	0.2
Max. prompt length	8192	8192
Max. response length	—	4096
Entropy coefficient	—	0
GPU	2×A800	4×A800

C Additional Analysis of VITAL

Additional Analysis of TAD. We complement the F1-based analysis in the main text with a ROUGE-L-based evaluation on CheXpert Plus (Figure 6). Compared with Native OPD, VITAL effectively learns from the teacher on teacher-better samples and achieves a higher median score to better preserve student's performance on student-better samples. These findings consistently support TAD's motivation: maximizing knowledge acquisition from teacher strengths while robustly preserving effective student behaviors.

Semantic Token Categorization. We align diagnosis generated from original and perturbed images using token-level Levenshtein distance to extract replacement, deletion, and insertion spans. Each span is assigned to the first matching category according to predefined lexicons, case-insensitive regular expressions, and a fixed priority order. Table 9 summarizes the rules for the three displayed categories. We further extend the visual-dependence analysis of NativeOPD in the main text to SKD (Figure 7) and MedOPD (Figure 8). Compared with both methods, VITAL shows greater visual dependence on clinically relevant anatomy and severity tokens and less dependence on other tokens.

Additional Qualitative Results. Figures 9–12 highlight VITAL's practical superiority over SKD and MedOPD. Unlike baselines, VITAL exhibits heightened sensitivity in adult cases, accurately capturing subtle abnormalities (e.g., mild edema). Furthermore, it effectively prevents severe misdiagnoses in complex neonatal scenarios by reliably distinguishing overlapping conditions like NRDS and TTN, demonstrating robust clinical accuracy.

Algorithm 1 VITAL: Visual and Task Advantage Learning for On-Policy Distillation

Require: Dataset $\mathcal{D}$; frozen teacher π_{T}; cold-start student $\pi_{\mathrm{S},\theta}$
Require: Steps K; rollouts n; learning rate η; noise scale σ_{noise}; gate parameters (β, τ); TAD weight α; clipping parameter ϵ_{clip}
Ensure: Trained student $\pi_{\mathrm{S},\theta}$

for $k = 1, \ldots, K$ **do**
  Sample a mini-batch $\mathcal{B} \subset \mathcal{D}$
  for all $(I, q, y^*) \in \mathcal{B}$ **do**
    Set $x \leftarrow (I, q)$; obtain rollouts $\{y_i\}_{i=1}^{n}$ from their associated behavior policy π_{old} and generate y_{T} from $\pi_{\mathrm{T}}(\cdot \mid x)$
    Store $\ell_{i,t}^{\mathrm{old}} \leftarrow \mathrm{sg}[\log \pi_{\mathrm{old}}(y_{i,t} \mid x, y_{i,<t})]$
    Construct $\tilde{x} \leftarrow (\mathrm{clip}(I + \epsilon, 0, 1), q)$, where $\epsilon_{h,w} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{\mathrm{noise}}^2)$ is shared across RGB channels
    Evaluate, for all valid (i, t):

$$\ell_{i,t}^{\mathrm{T}} = \log \pi_{\mathrm{T}}(y_{i,t} \mid x, y_{i,<t}), \quad \tilde{\ell}_{i,t}^{\mathrm{T}} = \log \pi_{\mathrm{T}}(y_{i,t} \mid \tilde{x}, y_{i,<t}), \quad \ell_{i,t}^{\mathrm{S}} = \log \pi_{\mathrm{S},\theta}(y_{i,t} \mid x, y_{i,<t})$$

    $\Delta_{i,t} \leftarrow \ell_{i,t}^{\mathrm{T}} - \ell_{i,t}^{\mathrm{S}}$

  VAP: Visual Advantage Policy Optimization

    $v_{i,t}^{\mathrm{T}} \leftarrow |\ell_{i,t}^{\mathrm{T}} - \tilde{\ell}_{i,t}^{\mathrm{T}}|$; $\boldsymbol{w}_{\mathrm{vis},i} \leftarrow \mathrm{MeanNorm}\left[\mathrm{sigmoid}\left(\frac{z(\boldsymbol{s}_i^{\mathrm{T}}) - \beta}{\tau}\right)\right]$
    $A_{i,t}^{\mathrm{vis}} \leftarrow \mathrm{sg}(w_{\mathrm{vis},i,t}\Delta_{i,t})$; $r_{i,t} \leftarrow \exp(\ell_{i,t}^{\mathrm{S}} - \ell_{i,t}^{\mathrm{old}})$
    $\mathcal{L}_{\mathrm{VAP}}(x) \leftarrow -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{T_i}\sum_{t=1}^{T_i}\min\big(r_{i,t}A_{i,t}^{\mathrm{vis}}, \mathrm{clip}(r_{i,t}, 1-\epsilon_{\mathrm{clip}}, 1+\epsilon_{\mathrm{clip}})\, A_{i,t}^{\mathrm{vis}}\big)$

  TAD: Task-Advantage Distillation

    $R_{\mathrm{T}} \leftarrow \text{ROUGE-L}(y_{\mathrm{T}}, y^*)$; $R_{\mathrm{S},i} \leftarrow \text{ROUGE-L}(y_i, y^*)$
    $A_i \leftarrow \max(R_{\mathrm{T}} - R_{\mathrm{S},i}, 0)$; $\mathcal{I}_+ \leftarrow \{i : A_i > 0\}$
    Initialize $\hat{w}_{\mathrm{task},i} \leftarrow 0$ for all i
    if $\mathcal{I}_+ \neq \varnothing$ **then**
      $\hat{w}_{\mathrm{task},i} \leftarrow \frac{|\mathcal{I}_+| A_i}{\sum_{j \in \mathcal{I}_+} A_j}, \quad i \in \mathcal{I}_+$
    end if
    $\mathcal{L}_{\mathrm{TAD}}(x) \leftarrow \frac{1}{n}\sum_{i=1}^{n}\hat{w}_{\mathrm{task},i}\left[\frac{1}{T_i}\sum_{t=1}^{T_i}\left(e^{\Delta_{i,t}} - \Delta_{i,t} - 1\right)\right]$
  end for
  $\mathcal{L}_{\mathrm{VITAL}} \leftarrow \frac{1}{|\mathcal{B}|}\sum_{(I,q,y^*)\in\mathcal{B}}[\mathcal{L}_{\mathrm{VAP}}(x) + \alpha\mathcal{L}_{\mathrm{TAD}}(x)]$
  $\theta \leftarrow \theta - \eta\nabla_\theta\mathcal{L}_{\mathrm{VITAL}}$
end for
return $\pi_{\mathrm{S},\theta}$

Table 9: Rule-based definitions of the three semantic categories shown in visual-sensitivity analysis.

Category	Matching rule
Anatomy	Matches anatomical structures or locations, including *lung, pleura, heart, mediastinum, diaphragm, rib, apex*, and *lobe*, together with word forms covered by the regular expressions.
Severity	Matches expressions of degree or change, including *mild, moderate, severe, small, increased, decreased, improved, stable*, and *unchanged*. This category covers both severity and changes in status.
Other	Includes function words, templated expressions, and content not covered by the predefined semantic lexicons.

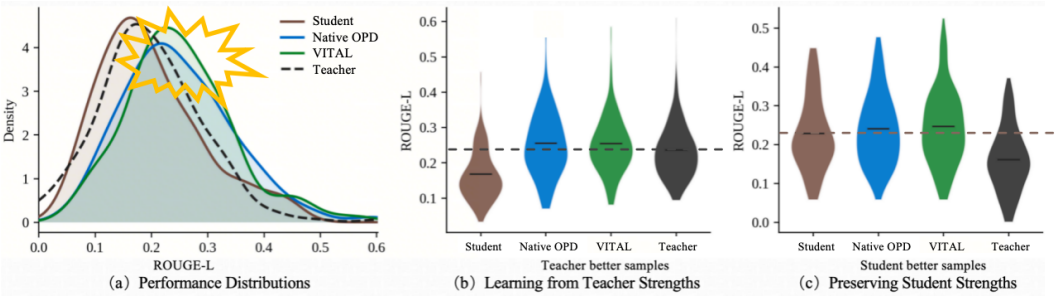


Figure 6: Additional ROUGE-L based analysis of TAD on CheXpert Plus. Compared with Native OPD, VITAL exhibits a consistent distribution shift toward higher scores, successfully acquiring teacher strengths while robustly preserving student performance.

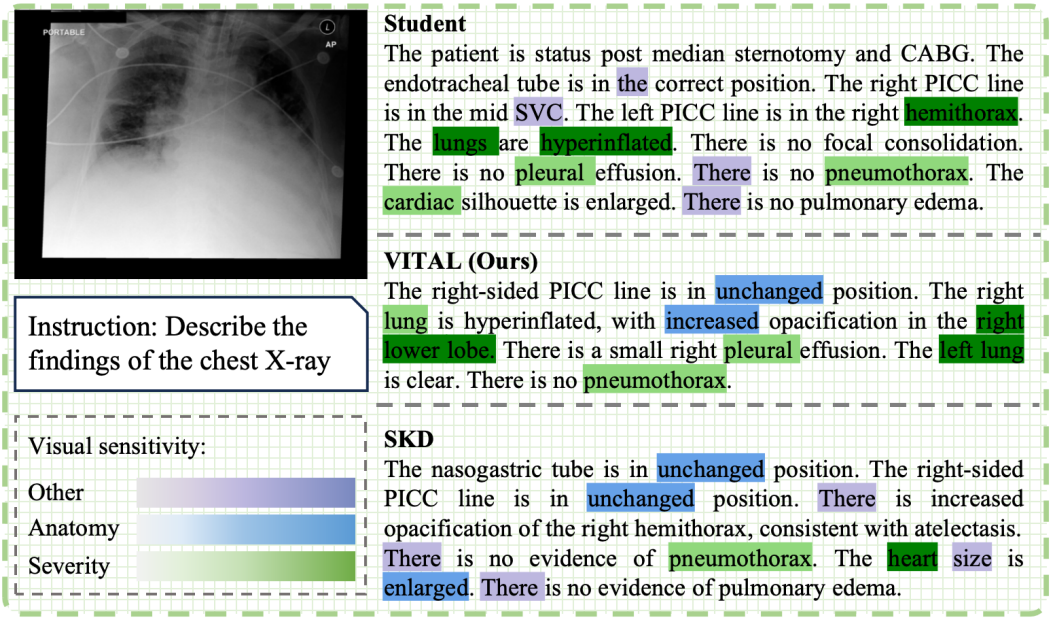


Figure 7: Visual dependency comparison of the student, VITAL, and SKD on MIMIC-CXR.

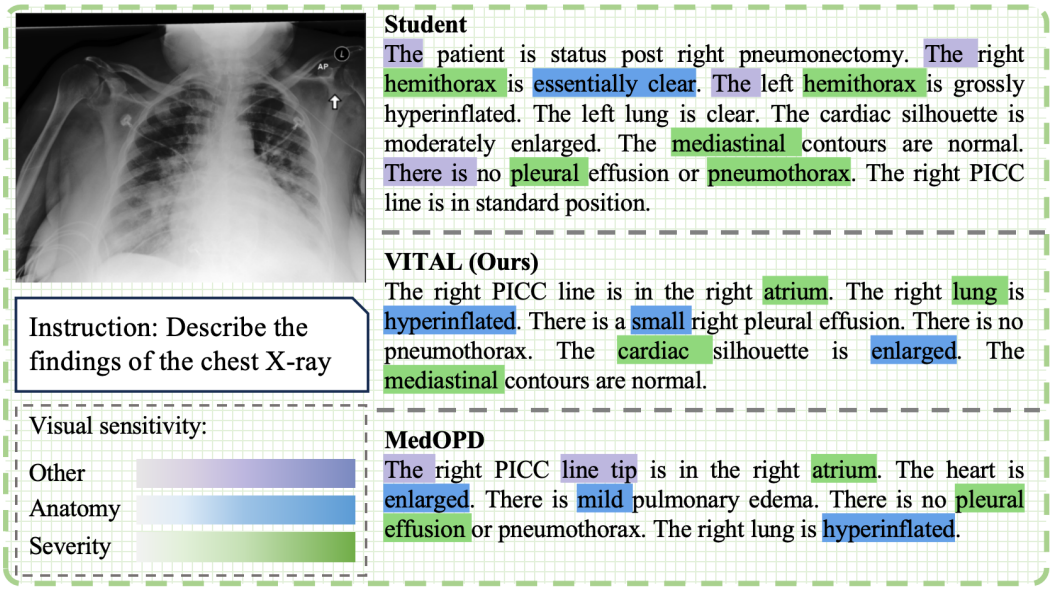


Figure 8: Visual dependency comparison of the student, VITAL, and MedOPD on MIMIC-CXR.

Prompt:
You are a helpful assistant. Please generate a report for the given chest X-ray image. Describe the findings of the chest X-ray.



Findings: The heart remains mildly enlarged. There is mild interstitial pulmonary edema similar to prior. No large effusion. Eventration of the right hemidiaphragm. Mediastinal contour stable. No focal consolidation. Bony structures intact. **GT**

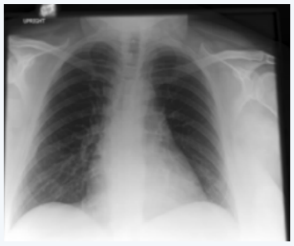
Findings: PA and lateral views of the chest were obtained. The heart size is enlarged. The mediastinal and hilar contours are normal. There is no pleural effusion or pneumothorax. The lungs are clear. **SKD**

Findings: PA and lateral views of the chest were obtained. The heart size is enlarged. The mediastinal and hilar contours are unremarkable. There is no pleural effusion or pneumothorax. The lungs are clear. **MedOPD**

Findings: PA and lateral views of the chest were obtained. The heart size is enlarged. There is mild pulmonary edema. There is no pleural effusion or pneumothorax. The mediastinal contours are normal. **VITAL**

Figure 9: Qualitative comparison on MIMIC-CXR. VITAL identifies cardiomegaly and mild pulmonary edema, whereas SKD and MedOPD miss the edema.

Prompt:
You are a helpful assistant. Please generate a report for the given chest X-ray image. Describe the findings of the chest X-ray.



Impression: Cardiac silhouette remains mildly enlarged. Lungs are clear. Healed right-sided rib fracture is stable. **GT**

Findings: The lungs are clear. There is no pleural effusion or pneumothorax. The cardiac silhouette is normal. The mediastinal contours are normal. Impression: No acute pulmonary or cardiac abnormality. **SKD**

Findings: The lungs are clear. The heart size is normal. There is no pleural effusion or pneumothorax. The mediastinal contours are normal.
Impression: No acute pulmonary or cardiac abnormality. **MedOPD**

Findings: The lungs are clear. There is no pleural effusion or pneumothorax. The cardiac silhouette is enlarged. The mediastinal contours are normal.
Impression: No pleural effusion or pneumothorax. Mild cardiomegaly. **VITAL**

Figure 10: Qualitative comparison on CheXpert-Plus. VITAL correctly identifies mild cardiomegaly, whereas SKD and MedOPD incorrectly report a normal cardiac silhouette.

Prompt:
You are a senior neonatologist. Please generate a report based on the Chest X-ray and the Clinical context, including both Imaging conclusion and Disease diagnosis.



Clinical context:
<dev> : Gestational age (27 weeks), birth weight (1070g).
<peri>: cesarean delivery, premature rupture of membranes.
<phys> : Apgar 1min (4), Apgar 5min (7), body temperature (36.0° C), pulse (146 bpm), respiration rate (62 bpm).

Imaging conclusion: Both lungs were hypoplastic with honeycomb changes and exudative shadows. No obvious abnormalities were observed in the abdomen.
Disease diagnosis: Neonatal Respiratory Distress Syndrome (NRDS), Neonatal Pneumonia — GT

Imaging conclusion: Increased texture on both lungs. There were no obvious abnormalities in the abdomen.
Disease diagnosis: Neonatal Pneumonia — SKD

Imaging conclusion: Increased texture on both lungs. There were no obvious abnormalities in the abdomen.
Disease diagnosis: Neonatal Pneumonia — MedOPD

Imaging conclusion: Increased texture on both lungs. There were no obvious abnormalities in the abdomen.
Disease diagnosis: Neonatal Respiratory Distress Syndrome (NRDS), Neonatal Pneumonia — VITAL

Figure 11: Qualitative comparison on NeoCXR. VITAL correctly diagnoses both NRDS and neonatal pneumonia, while SKD and MedOPD identify only neonatal pneumonia.

Prompt:
You are a senior neonatologist. Please generate a report based on the Chest X-ray and the Clinical context, including both Imaging conclusion and Disease diagnosis.



Clinical context:
<dev>: Gestational age (35 weeks), birth weight (2330g), body length (48.0cm), head circumference (33.0cm).
<peri> : vaginal delivery, premature rupture of membranes, gestational diabetes.
<phys> : Apgar 1min (9), Apgar 5min (9), Apgar 10min (9), body temperature (35.2° C), pulse (140 bpm), respiration rate (65 bpm), blood sugar (5.15 mmol/L).

Imaging conclusion: The texture in both lungs is increased, and intestinal dynamics are changed. Please follow up.
Disease diagnosis: Neonatal Transient Tachypnea (TTN) — GT

Imaging conclusion: Increased texture on both lungs. There were no obvious abnormalities in the abdomen.
Disease diagnosis: Neonatal Respiratory Distress Syndrome (NRDS), Neonatal Pneumonia — SKD

Imaging conclusion: Increased texture on both lungs. There were no obvious abnormalities in the abdomen.
Disease diagnosis: Neonatal Pneumonia — MedOPD

Imaging conclusion: Increased texture on both lungs. There were no obvious abnormalities in the abdomen.
Disease diagnosis: Neonatal Transient Tachypnea (TTN) — VITAL

Figure 12: Qualitative comparison on NeoCXR-EV. VITAL correctly diagnoses TTN, while SKD and MedOPD misdiagnose NRDS or neonatal pneumonia.