Supplementary Information

Performance evaluation on the In-Distribution (ID) dataset

For in-distribution (ID) evaluation, we conducted experiments on the CE-Bench-Test dataset to compare our proposed CE-R1 with current state-of-the-art VLMs, including Gemma-3-12B [36], LLaVA-NeXT-Video [37], QVQ-72B-Preview [38], Qwen-2.5-VL-7B [31], Qwen-2.5-VL-32B [31], Llama-3.2-11B-Vision [39], and MedGemma-4B [40]. As shown in Table S1, the CE-Bench-Test dataset encompasses five main clinical tasks: anatomy identification (including organ and landmark identification), endoscopic findings (covering both abnormal and normal findings), disease diagnosis, report generation, and treatment planning. Table S1 presents a comprehensive breakdown of model performance across all 44 specific clinical sub-tasks, revealing CE-R1's consistent superiority across the diverse clinical workflow with an overall accuracy of 68.15%±10.99%, substantially outperforming the best baseline model, Qwen-2.5-VL-32B (19.62%±8.50%), by a margin of 48.53 percentage points. In anatomy identification, CE-R1 demonstrates exceptional precision in distinguishing specific anatomical structures, achieving near-perfect performance in challenging landmark recognition tasks including the ileocecal valve (99.5%±2.4%), pylorus (97.2%±5.5%), and large intestine (96.3%±6.3%), while baseline models largely fail at these fine-grained anatomical distinctions with most achieving accuracies below 15%. For endoscopic findings, CE-R1 maintains superior detection capabilities across the full spectrum of pathological conditions, excelling in identifying critical findings such as active bleeding (91.0%±9.6%), ulcers (95.8%±6.7%), and foreign bodies (91.5%±9.3%), while also demonstrating robust performance in detecting subtle abnormalities including lymphangiectasia (73.9%±14.6%), angiectasia (84.6%±12.0%), and elevated lesions (68.7%±15.5%), as well as achieving 97.4%±5.3% accuracy in identifying normal clean mucosa. In disease diagnosis, CE-R1 consistently outperforms baseline models across diverse gastrointestinal conditions with particularly strong performance in duodenal bleeding (70.0%±3.3%), chronic gastritis (68.5%±8.8%), and duodenal ulcer (56.4%±6.3%), and for complex inflammatory conditions such as Crohn's disease and ulcerative colitis, CE-R1 achieves accuracies of 45.2%±2.5% and 45.7%±7.1%, respectively, substantially exceeding baseline models which rarely surpass 40%. For the higher-level clinical tasks of treatment planning and report generation, CE-R1 demonstrates significant advantages with accuracies of 76.7%±5.9% and 81.4%±6.4%, respectively, substantially outperforming the best baseline model Qwen-2.5-VL-32B (35.1%±4.5% for treatment planning and 38.2%±5.8% for report generation). This comprehensive analysis underscores that CE-R1's advantages extend beyond aggregate performance metrics to encompass reliable competency across the nuanced spectrum of clinical scenarios encountered in capsule endoscopy practice, establishing its potential for comprehensive clinical deployment.

Supplementary Table S1: Performance comparison among different methods on CE-Bench-Test dataset. S.I. is the abbreviation for small intestine.

Main Category	Sub-Task	Gemma -3-12B	LlaMA-3.2 -11B-Vision	LLaVA-NeXT -Video	MedGemma -4B	QVQ-72B -Preview	Qwen2.5-VL -32B	Qwen2.5-VL -7B	**CE-R1**
Anatomy Identification	Stomach	9.97±9.99	27.93±14.96	9.86±9.94	20.25±13.40	8.02±9.05	2.56±5.26	13.68±11.46	**95.77±6.71**
	Esophagus	26.98±14.80	36.08±16.01	12.27±10.93	39.11±16.27	4.04±6.56	4.76±7.10	11.40±10.59	**50.51±16.67**
	Small intestine	12.47±11.01	27.63±14.91	14.42±11.71	13.57±11.41	43.93±16.54	89.76±10.10	20.22±13.39	**94.31±7.72**
	Large intestine	15.28±11.99	1.81±4.44	12.98±11.20	18.27±12.88	2.31±5.01	0.17±1.38	21.45±13.68	**96.29±6.30**
	Duodenal bulb	2.34±5.04	9.78±9.90	0.53±2.42	0.32±1.88	9.14±9.61	11.90±10.79	3.51±6.13	**80.00±13.33**
	Duodenal papilla	0.00±0.00	12.50±11.02	12.50±11.02	0.00±0.00	12.50±11.02	0.00±0.00	0.00±0.00	**37.50±16.14**
	Ileocecal valve	9.65±9.84	3.25±5.91	0.63±2.64	0.16±1.32	4.61±6.99	6.24±8.06	3.25±5.91	**99.48±2.40**
	Antrum	2.39±5.09	2.26±4.95	0.39±2.07	0.32±1.89	0.77±2.92	0.13±1.20	2.84±5.54	**93.96±7.94**
	Pylorus	5.59±7.65	13.30±11.32	2.39±5.10	0.27±1.72	1.06±3.42	6.65±8.30	6.65±8.30	**97.24±5.46**
Endoscopic Finding	Active bleeding	0.27±1.74	0.68±2.75	0.27±1.74	1.50±4.06	1.78±4.41	8.34±9.22	0.14±1.23	**90.97±9.55**
	Fresh blood stains	0.00±0.00	0.86±3.08	0.00±0.00	3.45±6.08	0.86±3.08	15.52±12.07	0.00±0.00	**75.00±14.43**
	Old blood stains	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	**66.67±15.71**
	Erosion	2.40±5.10	0.82±3.00	0.31±1.84	0.31±1.84	1.79±4.41	9.85±9.93	2.60±5.31	**69.08±15.41**
	Ulcer	4.83±7.15	6.90±8.45	0.73±2.85	2.17±4.86	10.95±10.41	20.31±13.41	4.73±7.08	**95.79±6.70**
	Congestion	0.00±0.00	0.13±1.19	0.13±1.19	0.38±2.06	0.76±2.90	0.51±2.37	0.76±2.90	**71.21±15.09**
	Erythema	2.33±5.02	0.00±0.00	0.00±0.00	9.30±9.68	0.00±0.00	2.33±5.02	2.33±5.02	**44.19±16.55**
	Inflammation lesion	2.17±4.86	10.87±10.38	4.35±6.80	19.57±13.22	15.22±11.97	34.78±15.88	15.22±11.97	**34.78±15.88**
	Foreign body	0.00±0.00	10.11±10.05	1.06±3.42	14.36±11.69	2.13±4.81	3.19±5.86	1.06±3.42	**91.49±9.30**
	Parasitosis	0.00±0.00	6.74±8.36	0.00±0.00	0.00±0.00	3.37±6.02	4.49±6.91	1.12±3.51	**82.02±12.80**
	Lymph Follicle Hyperplasia	0.27±1.74	0.00±0.00	0.00±0.00	0.00±0.00	0.27±1.74	0.00±0.00	0.00±0.00	**79.35±13.49**
	Lymphangiectasia	0.58±2.53	0.29±1.79	0.00±0.00	0.00±0.00	0.29±1.79	0.00±0.00	0.00±0.00	**73.91±14.64**
	Elevated lesion	2.04±4.71	4.76±7.10	1.36±3.86	7.48±8.77	10.20±10.09	16.33±12.32	6.12±7.99	**68.71±15.46**
	Polypoid lesion	15.82±12.16	15.82±12.16	2.55±5.26	1.53±4.09	12.76±11.12	18.37±12.91	11.73±10.73	**64.29±15.97**
	Diverticula	0.00±0.00	0.00±0.00	0.00±0.00	12.50±11.02	0.00±0.00	0.00±0.00	0.00±0.00	**56.25±16.54**
	Angiectasia	0.43±2.17	1.28±3.75	0.43±2.17	0.00±0.00	5.56±7.64	10.68±10.30	0.00±0.00	**84.62±12.03**
	Vascular abnormalities	3.51±6.13	1.66±4.26	0.18±1.43	0.37±2.02	9.04±9.56	11.99±10.83	0.00±0.00	**77.31±13.96**
	Normal clean mucosa	0.02±0.47	19.65±13.25	9.38±9.72	2.61±5.32	12.43±11.00	24.33±14.30	13.05±11.23	**97.43±5.27**
Disease Diagnosis	Colonic bleeding	6.25±3.61	18.75±3.61	0.00±0.00	18.75±3.61	18.75±6.91	18.75±6.91	6.25±3.61	**50.00±5.89**
	Colonic ulcer	40.74±8.83	26.85±7.02	8.33±5.56	33.33±7.52	37.04±9.81	42.59±9.57	22.22±7.75	**45.37±8.70**
	Ulcerative colitis	40.00±10.89	18.33±8.68	0.00±0.00	44.00±9.49	33.00±7.20	23.33±6.67	35.33±6.80	**45.67±7.09**
	Duodenal bleeding	10.00±3.33	0.00±0.00	0.00±0.00	10.00±3.33	10.00±3.33	40.00±0.00	20.00±6.67	**70.00±3.33**
	Duodenal inflammation	17.98±7.71	14.11±5.76	4.17±3.44	22.14±7.96	20.60±6.63	25.83±7.37	18.57±7.51	**46.25±7.13**
	Duodenal polyp	15.00±5.27	25.00±7.07	17.50±7.18	7.50±3.57	14.17±5.13	35.83±10.02	31.67±10.96	**40.00±11.10**
	Duodenal ulcer	11.67±4.24	5.56±4.14	0.00±0.00	3.33±2.48	2.78±2.07	16.39±5.65	16.39±5.65	**56.39±6.26**
	Chronic gastritis	23.13±7.46	27.51±8.59	5.05±4.61	21.62±7.47	28.21±8.45	40.53±8.45	29.22±9.54	**68.54±8.75**
	Hemorrhagic gastritis	0.00±0.00	6.25±3.61	0.00±0.00	20.83±4.17	5.00±2.89	16.25±5.70	12.50±7.22	**56.67±2.36**
	Crohn's disease	38.10±5.72	23.81±7.19	0.00±0.00	33.33±7.27	26.19±7.78	33.33±7.27	19.05±7.53	**45.24±2.51**
	S.I. bleeding	3.33±2.48	13.33±6.57	0.00±0.00	0.00±0.00	13.33±6.57	33.33±7.86	0.00±0.00	**50.00±0.00**
	S.I. enteritis	35.29±7.53	27.33±8.28	1.86±2.03	37.81±7.51	31.00±7.59	35.76±8.41	25.43±8.48	**52.57±8.78**
	S.I. mass	21.67±6.48	26.67±5.39	10.00±4.08	26.67±5.39	26.67±9.20	26.67±9.20	15.00±10.00	**46.67±5.72**
	S.I. ulcer	33.40±7.83	18.78±7.92	6.22±4.50	26.86±8.31	27.56±8.77	31.41±8.78	18.46±8.01	**48.97±7.44**
Treatment Planning		41.32±5.14	10.05±3.12	5.15±1.79	32.11±4.70	28.20±4.80	35.13±4.54	19.16±3.65	**76.68±5.92**
Report Generation		19.52±4.73	25.98±6.26	8.99±5.24	20.72±5.68	23.21±6.25	38.16±5.83	36.64±6.28	**81.43±6.37**
Overall Accuracy		10.83±6.34	11.44±7.60	3.50±4.89	11.97±6.87	11.81±7.27	19.62±8.50	10.63±7.12	**68.15±10.99**

Performance evaluation on the Out-Of-Distribution (OOD) datasets

To assess generalization capacity across diverse clinical settings, we evaluated VLM performance on four external datasets from independent hospitals (YPH, TSH, RJH, and PWH). Tables S2 and S3 present comparative results across primary clinical tasks.

External validation on TSH and RJH datasets. Table S2 presents results across three clinical tasks of anatomy identification, endoscopic finding detection, and report generation. CE-R1 achieved superior performance on both datasets, with accuracy of 87.98±7.48% and 65.96±15.88% for anatomy identification on TSH and RJH respectively, and 87.74±7.61% and 84.42±9.30% for finding detection. The model's low variance indicates robust prediction stability across heterogeneous clinical scenarios and varying acquisition protocols. For report generation, performance gaps narrowed considerably. CE-R1 led both datasets with accuracies of 69.98±1.22% on TSH and 60.15±0.75% on RJH, followed closely by QVQ-72B-Preview (63.34±2.29%) and Qwen2.5-VL-32B-Instruct (64.01±1.48%) on TSH. This convergence suggests that large-scale language pre-training partially compensates for limited domain exposure in high-level reasoning tasks. Conversely, general-purpose models collapsed on fine-grained visual tasks. LLaVA-NeXT-Video achieved only 10.58±6.69% for anatomy identification on TSH, representing an 87.9% performance deficit compared to CE-R1. Most striking were the catastrophic failures of ostensibly medical-oriented models. MedGemma-4B and Gemma-3-12B registered near-zero accuracies for finding detection on TSH (0.60±0.42% and 0.28±0.19%) and RJH (2.16±1.50% and 2.32±1.60%), demonstrating that superficial medical knowledge integration proves insufficient for specialized imaging modalities requiring deep visual-clinical reasoning alignment.

External validation on YPH dataset. YPH evaluation encompassed five tasks of anatomy identification, finding detection, disease diagnosis, report generation, and treatment planning (Table S3). CE-R1 achieved peak performance across all tasks with accuracies of 79.07±11.70%, 87.20±7.89%, 50.71±6.11%, 70.25±1.91%, and 60.24±4.08% respectively, suggesting successful internalization of hierarchical medical knowledge rather than task-specific pattern matching. General-purpose VLMs exhibited task-dependent competency. QVQ-72B-Preview and Qwen2.5-VL-32B-Instruct performed credibly on abstract reasoning tasks, achieving 68.34±2.03% and 64.81±2.11% for report generation and 44.21±7.11% and 38.84±5.64% for diagnosis. However, these models experienced precipitous decline in perceptually demanding tasks. Performance gaps in specialized visual recognition were particularly pronounced. CE-R1 outperformed the best general-purpose model, LlaMa-3.2-11B-Vision, by 237% in anatomy identification (79.07±11.70% versus 23.47±12.70%) and by 257% in finding detection (87.20±7.89% versus 24.40±13.05% achieved by Qwen2.5-VL-32B-Instruct). These disparities represent the boundary between clinically actionable systems and those inadequate for deployment, establishing domain-specialized training as fundamentally necessary rather than merely beneficial.

External validation on PWH dataset. PWH validation confirmed CE-R1's cross-institutional robustness (Table S3). The model dominated visual discrimination tasks with anatomy identification achieving 72.17±14.20% and finding detection achieving 88.04±7.44%, exceeding nearest competitors by margins of 58.55 and 76.08 percentage points respectively. Disease diagnosis showcased CE-R1's most impressive performance at 70.00±0.12%, nearly doubling the accuracy of Qwen2.5-VL-32B-Instruct (37.26±5.87%). The remarkably low standard deviation of 0.12% indicates exceptional prediction consistency, which is critical for clinical decision support where erratic behavior compromises patient safety. Competing models exhibited substantially higher vari-

Supplementary Table S2: Performance Results for RJH and TSH Datasets

Methods	Anatomy Identification	Endoscopic Finding	Report Generation	Overall Accuracy
TSH Dataset				
Gemma-3-12B	10.04±6.39	0.28±0.19	62.61±1.86	24.31±2.81
LlaMa-3.2-11B-Vision	23.83±12.84	17.64±10.27	51.17±2.12	30.88±8.41
LLaVA-NeXT-Video	10.58±6.69	6.61±4.37	23.81±3.29	13.67±4.78
MedGemma-4B	18.16±10.51	0.60±0.42	59.10±1.84	25.95±4.26
QVQ-72B-Preview	10.48±6.64	13.64±8.33	63.34±2.29	29.15±5.75
Qwen2.5-VL-32B	9.82±6.26	18.88±10.83	64.01±1.48	30.90±6.19
Qwen2.5-VL-7B	14.43±8.73	13.87±8.45	61.55±1.81	29.95±6.33
CE-R1	**87.98±7.48**	**87.74±7.61**	**69.98±1.22**	**81.90±5.44**
RJH Dataset				
Gemma-3-12B	12.59±7.78	2.32±1.60	49.57±2.44	21.49±3.94
LlaMa-3.2-11B-Vision	15.13±9.08	16.62±9.80	40.30±2.53	24.02±7.14
LLaVA-NeXT-Video	15.25±9.14	10.91±6.87	19.20±2.38	15.12±6.13
MedGemma-4B	30.44±14.97	2.16±1.50	47.26±2.21	26.62±6.23
QVQ-72B-Preview	12.96±7.98	24.52±13.09	48.43±1.93	28.64±7.67
Qwen2.5-VL-32B	12.72±7.85	26.94±13.92	49.33±2.27	29.66±8.01
Qwen2.5-VL-7B	12.85±7.92	25.15±13.31	44.20±2.03	27.40±7.75
CE-R1	**65.96±15.88**	**84.42±9.30**	**60.15±0.75**	**70.18±8.64**

ance ranging from 3.38% to 6.46%, suggesting unstable decision boundaries. Report generation revealed an intriguing deviation where Qwen2.5-VL-32B-Instruct achieved comparable performance to CE-R1 (42.39±4.13% versus 43.09±1.72%). This likely reflects the task's greater reliance on language modeling capabilities where general-purpose pre-training provides compensatory advantages. However, CE-R1 reclaimed decisive leadership in treatment planning with 51.99±2.53% compared to 27.37±1.44% for Qwen2.5-VL-32B-Instruct, achieving a 90% relative improvement. This demonstrates that while general-purpose models generate linguistically coherent narratives, they struggle to synthesize multimodal evidence into actionable therapeutic recommendations.

Cross-institutional validation establishes CE-R1's robust generalization across geographically diverse datasets with varying imaging protocols and disease distributions. Performance hierarchies exhibit clear task dependency where specialized visual discrimination creates insurmountable challenges for general-purpose models, while higher-level reasoning tasks show narrower gaps. CE-R1's combination of high accuracy with low variance across diverse scenarios positions it as the only system approaching clinical deployment viability, with particular strength in fine-grained anatomical and pathological characterization that is most critical for diagnostic accuracy.

Supplementary Table S3: Performance Results for PWH and YPH Datasets.

Methods	Anatomy Identification	Endoscopic Finding	Disease Diagnosis	Report Generation	Treatment Planning	Overall Accuracy
YPH Dataset						
Gemma-3-12B	9.65±6.17	0.30±0.21	37.15±5.35	61.53±1.59	35.37±1.92	28.80±3.05
LlaMa-3.2-11B-Vision	23.47±12.70	16.37±9.68	29.24±5.40	51.40±2.61	4.26±0.22	24.95±6.12
LLaVA-NeXT-Video	13.03±8.01	2.98±2.04	7.91±2.22	27.13±3.30	7.81±0.26	11.77±3.17
MedGemma-4B	20.50±11.52	0.00±0.00	30.08±4.21	55.75±2.06	26.42±1.25	26.55±3.81
QVQ-72B-Preview	9.52±6.09	16.37±9.68	44.21±7.11	68.34±2.03	26.51±1.47	32.99±5.28
Qwen2.5-VL-32B	9.12±5.86	24.40±13.05	38.84±5.64	64.81±2.11	33.26±1.50	34.09±5.63
Qwen2.5-VL-7B	16.47±9.73	14.88±8.96	39.27±5.78	59.00±2.05	17.61±0.33	29.45±5.37
CE-R1	**79.07±11.70**	**87.20±7.89**	**50.71±6.11**	**70.25±1.91**	**60.24±4.08**	**69.49±6.34**
PWH Dataset						
Gemma-3-12B	12.17±7.56	4.35±2.94	27.99±4.43	25.77±2.83	21.92±1.43	18.44±3.84
LlaMa-3.2-11B-Vision	10.14±6.45	6.52±4.31	17.46±3.38	18.76±1.88	5.94±1.54	11.76±3.51
LLaVA-NeXT-Video	4.93±3.31	3.26±2.23	10.76±4.90	11.32±1.25	5.26±0.50	7.11±2.44
MedGemma-4B	7.25±4.75	7.61±4.97	21.15±5.34	16.72±2.20	19.27±0.94	14.40±3.64
QVQ-72B-Preview	11.88±7.40	14.13±8.58	26.34±5.67	30.40±3.81	16.86±1.14	19.92±5.32
Qwen2.5-VL-32B	13.62±8.32	11.96±7.44	37.26±5.87	42.39±4.13	27.37±1.44	26.52±5.44
Qwen2.5-VL-7B	5.51±3.68	9.78±6.24	28.70±6.46	29.89±3.45	12.60±0.68	17.30±4.10
CE-R1	**72.17±14.20**	**88.04±7.44**	**70.00±0.12**	**43.09±1.72**	**51.99±2.53**	**65.06±5.20**

Effectiveness of dynamic router

We investigated whether dynamic routing can enhance clinical reasoning by comparing three architectural variants: CE-R1 (adaptive routing), CE-R1-Lite (shallow reasoning only), and CE-R1-Deep (deep reasoning only). The adaptive CE-R1 model employs a difficulty-aware router that channels straightforward queries to CE-R1-Lite while directing challenging cases to CE-R1-Deep. We assessed these variants on five clinical tasks from CE-Bench-Test, as shown in Table S4. Our analysis reveals that reasoning depth requirements vary significantly across task types. Simple visual recognition tasks—anatomy identification and endoscopic finding—do not benefit from deep reasoning. In fact, CE-R1-Deep's performance drops markedly on endoscopic finding tasks (83.5±9.7%), underperforming both CE-R1 (94.0±4.0%) and CE-R1-Lite (93.5±4.3%) by approximately 10 percentage points. This decline suggests that over-complicated reasoning pathways can introduce unnecessary noise in tasks where direct pattern matching is optimal. For anatomy identification, all variants achieve strong performance above 91%, with CE-R1 reaching 95.0±3.3%. Conversely, cognitively demanding tasks demonstrate the critical value of adaptive depth selection. Disease diagnosis exemplifies this pattern most clearly: CE-R1 achieves 58.2±5.2% accuracy, surpassing CE-R1-Lite by 12.7 percentage points and CE-R1-Deep by 3.9 percentage points. This

Supplementary Table S4: Effectiveness of dynamic router.

Methods	Anatomy Identification	Endoscopic Finding	Disease Diagnosis	Report Generation	Treatment Planning	Overall Accuracy
CE-R1-Lite	93.76±4.14	93.46±4.32	45.49±4.54	78.56±3.70	76.04±2.44	84.34±4.13
CE-R1-Deep	91.96±5.23	83.54±9.72	54.27±5.36	45.80±2.93	65.88±3.61	75.00±6.60
CE-R1	**95.03±3.34**	**94.04±3.96**	**58.16±5.19**	**81.43±2.58**	**76.68±2.23**	**86.72±3.68**

superiority extends to report generation, where CE-R1 (81.4±2.6%) dramatically outperforms the deep-only variant (45.8±2.9%) by 35.6 percentage points. Interestingly, CE-R1-Lite performs reasonably well on report generation (78.6±3.7%), suggesting the router intelligently classifies many reporting tasks as relatively straightforward. Treatment planning shows a similar trend, with both CE-R1 (76.7±2.2%) and CE-R1-Lite (76.0±2.4%) exceeding CE-R1-Deep (65.9±3.6%) by over 10 percentage points. The aggregate performance metrics confirm the router's efficacy: CE-R1 achieves 86.72±3.68% overall accuracy, outperforming CE-R1-Lite by 2.38 points and CE-R1-Deep by 11.72 points. These findings demonstrate that adaptive complexity matching—rather than uniformly shallow or deep reasoning—optimizes both performance and computational resource allocation across diverse clinical reasoning scenarios.

Performance comparison on simple and difficult questions

Clinical examination tasks demonstrate substantial variation in complexity, necessitating flexible reasoning approaches tailored to question difficulty. We employed a systematic categorization method for the CE-Bench-Test dataset: CE-R1-Lite generated predictions across multiple temperature configurations (0.6, 0.7, 0.8, and 0.95), with questions achieving mean accuracy below 75% designated as difficult cases. Table S5 reveals distinct performance patterns across difficulty levels and task types. For simple questions, CE-R1-Lite dominates in four out of five tasks, achieving exceptionally high accuracy in Anatomy Identification (98.91%±1.07), Endoscopic Finding (98.16%±1.80), Report Generation (93.48%±1.60), and Treatment Planning (82.12%±1.15). Notably, CE-R1-Deep shows substantial performance degradation on these straightforward cases, with Endoscopic Finding accuracy dropping to 86.98%±11.33 and Report Generation plummeting to 48.14%±4.24—a 45.34 percentage point decline. This suggests that excessive reasoning depth may introduce unnecessary complexity and potential error propagation in simple scenarios. The performance landscape shifts dramatically for difficult questions. CE-R1-Deep demonstrates marked improvements, particularly in Anatomy Identification (37.27%±23.38 vs. 17.07%±14.15) and Endoscopic Finding (31.21%±21.47 vs. 21.70%±16.99), representing relative gains of 118% and 55% respectively. Disease Diagnosis emerges as a consistently challenging task where deeper reasoning proves advantageous across both difficulty levels—CE-R1-Deep achieves 54.19%±7.38 on simple questions and maintains 54.47%±8.06 on difficult ones, outperforming CE-R1-Lite by approximately 9 percentage points in both cases. Interestingly, Report Generation exhibits an in-

Supplementary Table S5: Performance Comparison on the Simple and Difficult Questions.

Methods	Difficulty	Anatomy Identification	Endoscopic Finding	Disease Diagnosis	Report Generation	Treatment Planning
CE-R1-Lite	Simple	**98.91±1.07**	**98.16±1.80**	45.67±6.51	**93.48±1.60**	**82.12±1.15**
CE-R1-Deep		95.63±4.18	86.98±11.33	**54.19±7.38**	48.14±4.24	68.23±4.74
CE-R1-Lite	Difficult	17.07±14.15	21.70±16.99	45.04±6.18	**62.04±4.05**	55.28±5.96
CE-R1-Deep		**37.27±23.38**	**31.21±21.47**	**54.47±8.06**	43.19±3.91	**57.69±5.47**

verse pattern: CE-R1-Lite maintains superior performance (62.04%±4.05 vs. 43.19%±3.91) even on difficult questions, suggesting this task may benefit more from concise, direct reasoning than elaborate analytical processes. The variance in scores provides additional insights. CE-R1-Deep exhibits substantially higher standard deviations on difficult questions, particularly in Anatomy Identification (±23.38%) and Endoscopic Finding (±21.47%), indicating less stable performance when applying deep reasoning to challenging cases. This variability underscores the importance of adaptive routing mechanisms. These contrasting patterns validate our dynamic routing approach. By intelligently selecting between Lite and Deep reasoning pathways based on question difficulty, CE-R1 with routing achieves superior performance over both standalone variants, effectively capturing the benefits of lightweight reasoning for straightforward tasks while leveraging deep analytical capabilities for complex clinical scenarios.