PancCADx: A Multimodal Framework for Pancreatic Cancer Diagnosis

Shan Hu[1⋆], Changhong Xiao[1⋆], Xianzheng Qin[2], Bin Mei[3], Bin Cheng[2], and Zhongyuan Wang[1⋆⋆]

[1] School of Computer Science, Wuhan University, Wuhan, China
hushanhn@gmail.com, xch67690002@gmail.com, wzy_hope@163.com
[2] Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China
277099992@qq.com, b.cheng@tjh.tjmu.edu.cn
[3] Zhongnan Hospital of Wuhan University, Wuhan, China
neuromei20@whu.edu.cn

Abstract. Endoscopic ultrasonography (EUS) interpretation for pancreatic cancer is highly operator-dependent, leading to significant diagnostic variability and limited reasoning transparency. We present PancCADx, an interpretable multimodal framework leveraging Qwen3-VL-8B-Thinking with a Chain-of-Thought (CoT) mechanism to align model reasoning with expert clinical logic. To ensure clinical alignment and minimize hallucinations, we utilized a two-stage alignment strategy: Supervised Fine-Tuning (SFT) followed by Direct Preference Optimization (DPO). Developed on an internal cohort ($n = 438$) and validated on multi-center external cohorts ($n = 191$), PancCADx achieved 95.74% external sensitivity, significantly outperforming state-of-the-art benchmarks (92.55%) with superior cross-institutional generalizability. In a prospective sequential validation involving 12 clinicians, AI assistance significantly enhanced diagnostic accuracy (+7.7%, $p < 0.001$) and decision confidence ($p < 0.001$). Notably, PancCADx effectively bridged the novice-expert gap and improved specificity for senior clinicians (up to 76%). By harmonizing high diagnostic performance with transparent, evidence-based reasoning, PancCADx provides a robust, clinically-aligned solution for objective and trustworthy pancreatic cancer diagnosis.

Keywords: Pancreatic Cancer · Endoscopic Ultrasonography · Multimodal Large Language Model · Chain of Thought · Clinical Alignment.

1 Introduction

Pancreatic cancer remains a global health crisis with a 5-year survival rate of merely 8.5%, demanding effective early detection tools [2,18]. While endoscopic ultrasonography (EUS) serves as a high-resolution modality for assessing pancreatobiliary diseases, its diagnostic efficacy is limited by high operator dependence

⋆ These authors contributed equally to this work.
⋆⋆ Corresponding author: wzy_hope@163.com

and the difficulty in distinguishing subtle early-stage pathological changes, leading to substantial inter-observer variability and potential misdiagnosis [3,19,10].

Deep learning has achieved remarkable success in medical image analysis, with Convolutional Neural Networks (CNNs) establishing themselves as the standard for expert-level lesion detection [21,12]. However, traditional CNN architectures face limitations hindering broad clinical adoption. First, their "black-box" nature poses a significant trust barrier, providing prediction probabilities without the semantic reasoning required for high-stakes clinical decision-making [17]. Second, as unimodal systems isolated from rich clinical context, they fail to replicate the holistic diagnostic process of human experts [15].

The rapid evolution of Multimodal Large Language Models (MLLMs) has catalyzed domain-specific adaptations, such as LLaVA-Med and Med-PaLM, which align visual encoders with biomedical knowledge [11,20]. However, deploying these models into specialized workflows like pancreatic EUS reveals critical bottlenecks. First, data scarcity hinders sub-specialty mastery: high-quality, expert-annotated datasets for granular domains are rare. Consequently, most models rely on massive, scraped biomedical corpora suffering from heterogeneous quality and noise, limiting the precision required for fine-grained diagnostics [5,14]. Second, a methodological disconnect exists between training paradigms and clinical reality. Existing adaptations predominantly rely on Supervised Fine-Tuning and are evaluated on broad-spectrum benchmarks [8,24,6,9]. This "SFT-only" approach, lacking post-training alignment via Reinforcement Learning (RL), yields models that may achieve high generalist scores but fail to adhere to complex clinical guidelines or minimize hallucinations in practice. Finally, a cognitive gap persists: models primarily rely on surface-level pattern matching rather than deductive reasoning, failing to generate the step-by-step rationale necessary for effective clinical decision support [23].

To bridge this gap, we introduce PancCADx, a novel reasoning-based diagnostic framework leveraging the Qwen3-VL-8B-Thinking architecture [1]. Unlike traditional "black-box" classifiers or unaligned MLLMs, PancCADx enhances clinical trust by integrating a Chain-of-Thought (CoT) mechanism. To rigorously align this reasoning with expert logic, we employ a two-stage alignment strategy—Supervised Fine-Tuning (SFT) followed by Direct Preference Optimization (DPO) [16]—efficiently implemented via Low-Rank Adaptation (LoRA) [7] using single-center data.

Our contributions are threefold. First, we propose a specialized multimodal reasoning framework that harmonizes visual perception with clinical logic through a novel SFT-DPO pipeline, transforming the model from a passive classifier into an interpretable agent capable of minimizing hallucinations. Second, we demonstrate the system's superior diagnostic accuracy and cross-center robustness through rigorous validation on diverse external cohorts from three independent institutions, where it achieves higher sensitivity than state-of-the-art benchmarks while providing interpretable reasoning and greater data efficiency. Finally, transcending in-silico metrics, we provide empirical evidence of clinical utility via a prospective self-controlled sequential study, quantifying PancCADx's ability to

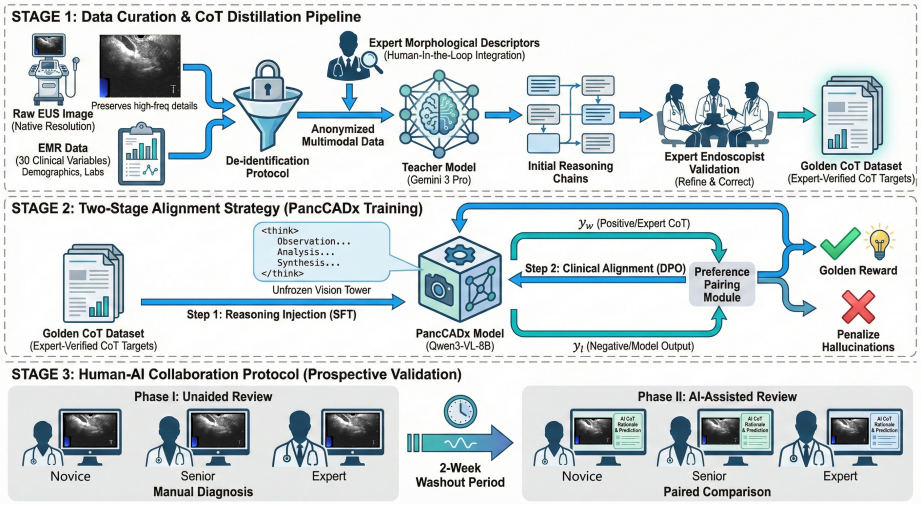


Fig. 1: PancCADx Framework Overview. Built on the Qwen3-VL-8B-Thinking backbone, the model synthesizes EUS images with clinical data. Training employs a two-stage strategy: (1) Reasoning Injection (SFT) via expert CoT paths; and (2) Clinical Alignment (DPO), contrasting expert-verified vs. flawed outputs to ensure rigorous logic and minimize hallucinations.

bolster diagnostic confidence and function as an interpretable second reader in realistic workflows.

2 Methods

2.1 The PancCADx Framework

The proposed framework (Fig. 1) leverages the Qwen3-VL-8B-Thinking architecture, selected for its Naive Dynamic Resolution mechanism. Unlike traditional MLLMs enforcing fixed-resolution inputs via down-sampling, this architecture processes images of arbitrary aspect ratios by mapping them into variable visual token sequences. For EUS imaging, preserving high-frequency spatial details—such as subtle parenchymal heterogeneity or jagged carcinoma margins—is critical. The 8-billion parameter scale offers an optimal balance between deep diagnostic reasoning and computational efficiency.

Multimodal Serialization and Privacy Protocol. We developed a streamlined serialization pipeline aligning raw visual data with clinical context. The input comprises two parallel streams: native-resolution EUS images and structured prompts containing 30 key clinical variables from electronic medical records (EMR). Prior to serialization, a rigorous De-identification Protocol was implemented. All Protected Health Information (PHI) was redacted or tokenized, while

EUS images were stripped of metadata and assigned randomized hash codes. Anonymized clinical features were transformed into natural language prompts, encoding missing values as "null" tokens to maintain structural integrity.

2.2 Chain-of-Thought Distillation Pipeline

To endow PancCADx with interpretable reasoning, we constructed a specialized CoT dataset via a "Distill-and-Refine" pipeline. In the distillation phase, Gemini 3 Pro generated reasoning chains, bridging visual perception and linguistic logic by integrating Expert-Annotated Morphological Descriptors (e.g., echogenicity) with raw EUS images and EMR data.

To enforce precision, we adopted a "Conclusion-First" protocol. Unlike traditional deduction, our CoT positions the diagnosis first, followed by retrospective morphological analysis. This strategy anchors reasoning to the outcome, mitigating logic drift [13]. Furthermore, by mimicking expert recognition-primed decision-making—aligning intuition with System 2 verification [22]—this structure enhances diagnostic efficiency. These chains were refined by three experts ($>$ 15 years experience) to standardize terminology and eliminate hallucinations.

For DPO, we targeted clinically rigorous, hallucination-free reasoning. We constructed a preference dataset $D = \{(x, y_w, y_l)\}$ by capturing discrepancies between SFT model outputs and expert consensus. Generated CoTs diverging from the expert standard served as negative samples (y_l), paired with verified paths as positive samples (y_w). This strategy yielded 386 high-quality preference pairs targeting residual cognitive errors.

2.3 Two-Stage Alignment Strategy

We implemented a two-stage alignment via Low-Rank Adaptation (LoRA). First, Reasoning Injection (SFT) fine-tuned the model on the CoT dataset to maximize the likelihood of the expert reasoning path y given input x:

$$\mathcal{L}_{SFT} = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{CoT}} \left[\log P_\theta(y|x)\right] \tag{1}$$

This transforms the model from a passive classifier into an active reasoner. Subsequently, Clinical Alignment (DPO) was applied to enforce adherence to clinical logic by optimizing the policy to prefer expert-curated reasoning y_w over flawed outputs y_l:

$$\mathcal{L}_{DPO} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right] \tag{2}$$

where π_{ref} represents the SFT model, and β controls the deviation penalty. This process effectively penalizes hallucinations and aligns the model with expert diagnostic logic.

2.4 Human-AI Collaboration Protocol

To evaluate clinical utility, we designed a prospective two-phase sequential validation study with 130 patients. We recruited twelve endoscopists stratified into three tiers: experts (annually performed 300 EUS procedures with > 10 years of experience), seniors (annually performed 150 EUS procedures with > 5 years of experience), and novices (with > 1 year of experience in EUS).

The trial adopted a self-controlled design to minimize inter-observer variability. In Phase I (Unaided), all endoscopists reviewed the cohort using conventional methods. Following a strict 2-week Washout Period, the same cohort was reviewed in Phase II (AI-Assisted), with participants provided PancCADx predictions and CoT reasoning. This paired design allows for direct quantification of the AI intervention's impact on diagnostic performance and confidence.

3 Experiments

3.1 Datasets and Experimental Setup

Datasets and Cohorts. Strictly adhering to STARD guidelines, we curated a multicenter cohort of pancreatobiliary EUS examinations from four tertiary referral centers (Centers A–D). Retrospective data were collected between January 2014 and December 2022, with a prospective cohort enrolled between January and June 2023. This study was approved by the Institutional Review Board (IRB) of the participating centers (No. TJ-IRB202409042). Given the use of fully anonymized data, the requirement for informed consent was waived. Eligibility was limited to patients with solid pancreatic lesions confirmed by definitive histopathology or rigorous clinical-radiological consensus ($\geq$6 months follow-up). The high-volume Center A ($n = 438$) served as the development cohort (stratified 90:10 for training/validation), while data from three independent institutions (Centers B, C, and D) formed the external testing cohort ($n = 191$). Comprehensive characteristics are detailed in Table 1. To validate clinical utility, we further curated a dedicated Challenge Cohort of 130 prospectively enrolled cases (including challenging presentations such as lesions < 2 cm and chronic pancreatitis background) for a two-phase sequential validation study engaging twelve stratified endoscopists under blind review protocols.

Implementation and Evaluation. The framework was implemented in PyTorch on an NVIDIA A100 (40GB) GPU. The training pipeline utilized the AdamW optimizer with a two-stage strategy: SFT for 20 epochs ($lr = 1e^{-4}$, batch=8, cosine scheduler) followed by DPO for 10 epochs ($lr = 1e^{-7}$, $\beta = 0.3$). LoRA was applied to linear layers ($r = 128, \alpha = 256$). Performance was assessed using composite endpoints for diagnostic accuracy (sensitivity, accuracy) and interpretability. Additionally, a Subjective Assessment was conducted using a 5-point Likert questionnaire across five dimensions: Descriptive Fidelity, Diagnostic Confidence, Reasoning Acceptance, CoT Utility, and Willingness to Adopt.

Table 1: Demographic and Clinical Characteristics of the Retrospective Cohorts. Non-Malignant includes AIP, CP, NET, SPT, etc.Each patient corresponds to a single EUS image. AIP = Autoimmune Pancreatitis; CP = Chronic Pancreatitis; NET = Neuroendocrine Tumor; SPT = Solid Pseudopapillary Tumor.

Cohort	Total *N*	Malignant *N* (%)	Non-Malignant *N* (%)	Age (Mean ± SD)	Gender (M / F)
Development Cohort					
Center A	438	270 (61.6)	168 (38.4)	56.7 ± 12.5	283 / 155
- Training	395	243 (61.5)	152 (38.5)	56.7 ± 12.9	258 / 137
- Internal Val	43	27 (62.8)	16 (37.2)	56.2 ± 8.3	25 / 18
External Validation Cohorts					
Center B	23	15 (65.2)	8 (34.8)	61.1 ± 9.8	13 / 10
Center C	65	27 (41.5)	38 (58.5)	53.5 ± 15.7	36 / 29
Center D	103	52 (50.5)	51 (49.5)	58.2 ± 13.6	69 / 34
Total	629	364 (57.9)	265 (42.1)	56.7 ± 13.1	401 / 228

Note that threshold-independent metrics (e.g., AUC) are not applicable here, as PancCADx produces categorical diagnostic conclusions through reasoning chains rather than continuous probability scores.

3.2 Comparison with State-of-the-Art

Baselines and Fairness of Comparison. We benchmarked PancCADx against the SOTA Joint-AI model [4]. To ensure a rigorous upper-bound comparison, we utilized the official Joint-AI model trained on a cohort superset (multi-frame video from the same population). In contrast, PancCADx was restricted to a single-frame constraint. This comparison highlights our architecture's data efficiency, aiming to match a baseline that benefits from significantly richer temporal and spatial redundancy.

Diagnostic Performance. Diagnostic efficacy was evaluated on internal and external cohorts (Table 2). On the internal Center A hold-out ($n = 43$), PancCADx demonstrated moderate stability (accuracy 83.72%, 95% CI [70.0, 91.9]; F1 87.72%). Given the small holdout size, these point estimates carry inherent statistical uncertainty. Notably, benchmarking reveals severe overfitting in Joint-AI: despite near-perfect internal accuracy (98.00%), its performance suffered a precipitous drop externally (88.48%), highlighting a substantial generalization gap (ΔAcc $\approx -10\%$). Conversely, PancCADx demonstrated superior external robustness ($n = 191$), effectively bridging domain shifts to achieve a pooled sensitivity of 95.74% (95% CI [89.6, 98.3]), exceeding Joint-AI's sensitivity (92.55%),

Table 2: Benchmarking Diagnostic Performance Against SOTA Methods. PancCADx (Ours) vs. SOTA Joint-AI [4]. †Joint-AI 'Total Ext.' metrics are pooled recalculations based on the absolute sample sizes of the perfectly matched external cohorts to ensure a rigorous comparison.

Dataset	Model	Performance Metrics (%)					
		Acc	Sens	Spec	PPV	NPV	F1
Internal Validation (Center A)							
Center A	Joint-AI	98.00	99.00	98.00	98.00	98.00	98.50
	PancCADx	83.72	92.59	68.75	83.33	84.62	87.72
External Validation (Multicenter)							
Center B	Joint-AI	91.00	93.00	88.00	93.00	88.00	93.00
	PancCADx	91.30	100.00	75.00	88.24	100.00	93.75
Center C	Joint-AI	84.00	92.00	78.00	76.00	94.00	83.24
	PancCADx	81.54	96.30	71.05	70.27	96.43	81.25
Center D	Joint-AI	90.00	92.00	88.00	89.00	92.00	90.48
	PancCADx	91.26	94.23	88.24	89.09	93.75	91.59
Total Ext.	Joint-AI†	88.48	92.55	84.54	85.29	92.13	88.78
	PancCADx	87.96	95.74	80.41	82.57	95.12	88.67

a clinically critical advantage given that missed pancreatic malignancies carry far greater risk than false positives in a disease with merely 8.5% five-year survival [2]. Pooled specificity was 80.41% (95% CI [71.4, 87.1]) and accuracy 87.96% (95% CI [82.6, 91.8]). This robustness was confirmed via subgroup analysis: in the largest cohort (Center D, $n = 103$), PancCADx outperformed the benchmark in accuracy (91.26% vs. 90.00%) and sensitivity (94.23% vs. 92.00%), achieving 100.00% sensitivity in Center B ($n = 23$). Although Joint-AI showed marginally higher specificity (84.54% vs. 80.41%), PancCADx maintained a superior Negative Predictive Value (NPV) of 95.12%, highlighting its utility as a reliable rule-out tool for screening.

3.3 Clinical Utility and Workflow Impact

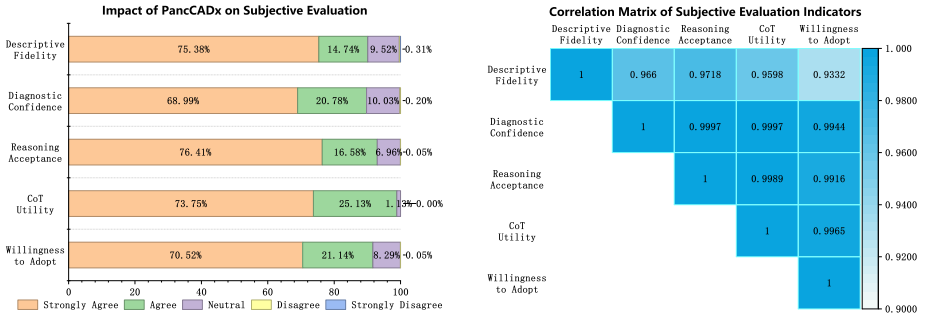
To validate real-world applicability, we conducted a prospective self-controlled sequential study (Table 3). For novice endoscopists ($n = 6$), AI assistance acted as a critical equalizer, boosting sensitivity from 77% to 88% ($p < 0.001$) and accuracy from 66% to 84% ($p < 0.001$), effectively aligning their performance with unassisted experts. Senior endoscopists ($n = 4$) experienced a significant rise in specificity from 40% to 76% ($p = 0.003$), suggesting that the model's Chain-of-Thought (CoT) reasoning mitigated false-positive biases. Expert endoscopists ($n = 2$) maintained high baseline performance with no statistically

Table 3: Comparison of Endoscopists' Performance With and Without AI Assistance: Joint-AI vs. PancCADx. Values represent mean performance metrics. 95% Confidence Intervals are omitted for brevity. P-values are calculated based on paired tests comparing results with and without AI assistance for each group. Joint-AI results are referenced from [4].

Group	Condition	Sensitivity		Specificity		Accuracy	
		Joint-AI	PancCADx	Joint-AI	PancCADx	Joint-AI	PancCADx
Expert ($n = 2$)	Without AI	0.88	0.86	0.88	0.82	0.88	0.84
	With AI	0.80	0.95	0.85	0.80	0.82	0.90
	p-value	$> .99$	.215	$> .99$	$> .99$	$> .99$	.154
Senior ($n = 4$)	Without AI	0.72	0.94	0.84	0.40	0.77	0.75
	With AI	0.78	0.96	0.84	0.76	0.80	0.89
	p-value	.39	.418	$> .99$	**.003**	.23	$<$ **.001**
Novice ($n = 6$)	Without AI	0.61	0.77	0.81	0.46	0.69	0.66
	With AI	0.91	0.88	0.88	0.77	0.90	0.84
	p-value	$<$ **.001**	$<$ **.001**	.34	$<$ **.001**	$<$ **.001**	$<$ **.001**

significant changes (all $p > 0.15$), suggesting that AI primarily reinforced diagnostic confidence rather than altering clinical decisions at this experience level. Notably, the limited subgroup size ($n = 2$) restricts statistical power for this stratum. Crucially, unlike Joint-AI which failed to yield statistically significant improvements for experienced readers ($p > 0.05$), the prevalence of extremely low p-values with PancCADx confirms its superior clinical applicability and robustness in augmenting decision-making across the expertise spectrum.

We further analyzed subjective clinician feedback (Fig. 2). The stacked bar chart (Fig. 2a) displays rating distributions across five key dimensions, indicating high user satisfaction with the vast majority of responses falling into "Strongly Agree" or "Agree." Notably, "Diagnostic Confidence" and "Reasoning Acceptance" achieved the highest positive ratings, confirming that the AI's transparent logic effectively aligns with clinical expectations. To understand the interplay between metrics, we computed the Pearson correlation matrix (Fig. 2b). The heatmap reveals strong positive correlations, particularly between "Descriptive Fidelity" and "Diagnostic Confidence," suggesting that accurate morphological description is the foundational driver of clinician trust. Furthermore, the strong correlation between "CoT Utility" and "Willingness to Adopt" confirms interpretability is a decisive factor for integration into routine workflows.



(a) (b)

Fig. 2: Subjective evaluation results. (a) Average ratings stratified by endoscopist expertise. (b) Correlation matrix among the five evaluation metrics.

4 Conclusion

We introduce PancCADx, a reasoning-driven framework bridging the gap between black-box AI and clinical interpretability. Leveraging a novel SFT-DPO alignment, the model achieves 95.74% sensitivity across multicenter cohorts, outperforming multi-frame SOTA baselines using only single-frame inputs, which underscores its superior data efficiency. Our sequential validation study confirms its utility in bridging the novice-expert accuracy gap and correcting specificity bias in senior endoscopists. Crucially, the interpretable Chain-of-Thought reasoning provides dual benefits: reinforcing expert confidence while functioning as an educational scaffold for novices. Thus, PancCADx establishes a new paradigm for transparent, semantically aligned clinical decision support.

Limitations. Several limitations warrant discussion. First, the expert subgroup ($n = 2$) restricts statistical power for evaluating AI benefit at the highest experience level; future studies should recruit larger expert cohorts. Second, the Conclusion-First CoT protocol, while grounded in recognition-primed decision-making theory, carries an inherent risk of confirmation bias—the model may selectively attend to evidence supporting its initial impression. Although DPO partially mitigates this by penalizing incomplete reasoning chains, dedicated prospective studies are needed to quantify residual bias. Third, the internal validation set ($n = 43$) yields statistically fragile point estimates (95% CI $\approx$ [72%, 92%] for accuracy); larger holdout sets would strengthen internal performance characterization. Fourth, the current framework is validated exclusively on pancreatic EUS; generalizability to other endoscopic modalities or organ sites remains undemonstrated. Finally, CoT quality was assessed via subjective Likert scales rather than independent factual accuracy audits of reasoning chains, which represents a direction for future work.

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

Acknowledgements. This work was supported by the Transformation Fund Project of Scientific and Technological Achievements of Zhongnan Hospital of Wuhan University (2023CGZH-ZD006).

Data Availability. The source code and model weights are publicly available at https://github.com/hill-hu/PancCADx. Patient-level clinical data cannot be shared publicly due to institutional ethics requirements but are available from the corresponding authors upon reasonable request with appropriate ethical approval.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R.L., Soerjomataram, I., Jemal, A.: Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **74**(3), 229–263 (2024)
3. Ciaravino, V., D'Onofrio, M.: Pancreatic ultrasound: state of the art. Journal of Ultrasound in Medicine **38**(5), 1125–1137 (2019)
4. Cui, H., Zhao, Y., Xiong, S., Feng, Y., Li, P., Lv, Y., Chen, Q., Wang, R., Xie, P., Luo, Z., et al.: Diagnosing solid lesions in the pancreas with multimodal artificial intelligence: a randomized crossover trial. JAMA Network Open **7**(7), e2422454–e2422454 (2024)
5. He, S., Nie, Y., Chen, Z., Cai, Z., Wang, H., Yang, S., Chen, H.: Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. CoRR (2024)
6. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300 (2020)
7. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
8. Jin, D., Pan, E., Oufattole, N., Weng, W.H., Fang, H., Szolovits, P.: What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Applied Sciences **11**(14), 6421 (2021)
9. Kung, T.H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., et al.: Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models. PLoS digital health **2**(2), e0000198 (2023)
10. Lee, D., Jesry, F., Maliekkal, J.J., Goulder, L., Huntly, B., Smith, A.M., Khaled, Y.S.: Application of artificial intelligence in pancreatic cyst management: A systematic review. Cancers **17**(15), 2558 (2025)
11. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)

12. Li, J., Jing, X., Zhang, Q., Wang, X., Wang, L., Shan, J., Zhou, Z., Jiang, D., Yan, Y., Liu, L., et al.: Interpretable deep learning model diagnoses gastrointestinal stromal tumors and lesion characteristics with microprobe endoscopic ultrasonography. Scientific Reports **15**(1), 34366 (2025)
13. Lin, T., Zhao, X., Zhang, X., Long, R., Xu, Y., Jiang, Z., Su, W., Zheng, B.: Ravr: Reference-answer-guided variational reasoning for large language models. arXiv preprint arXiv:2510.25206 (2025)
14. Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., Xie, W.: Pmc-clip: Contrastive language-image pre-training using biomedical documents. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 525–536. Springer (2023)
15. Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P.: Foundation models for generalist medical artificial intelligence. Nature **616**(7956), 259–265 (2023)
16. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. Advances in neural information processing systems **36**, 53728–53741 (2023)
17. Rajpurkar, P., Chen, E., Banerjee, O., Topol, E.J.: Ai in health and medicine. Nature medicine **28**(1), 31–38 (2022)
18. Siegel, R.L., Giaquinto, A.N., Jemal, A.: Cancer statistics, 2024. CA: a cancer journal for clinicians **74**(1) (2024)
19. Tacelli, M., Lauri, G., Tabacelia, D., Tieranu, C.G., Arcidiacono, P.G., Săftoiu, A.: Integrating artificial intelligence with endoscopic ultrasound in the early detection of bilio-pancreatic lesions: Current advances and future prospects. Best Practice & Research Clinical Gastroenterology **74**, 101975 (2025)
20. Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.C., Carroll, A., Lau, C., Tanno, R., Ktena, I., et al.: Towards generalist biomedical ai. Nejm Ai **1**(3), AIoa2300138 (2024)
21. Wang, Y.Y., Liu, B., Wang, J.H.: Application of deep learning-based convolutional neural networks in gastrointestinal disease endoscopic examination. World Journal of Gastroenterology **31**(36), 111137 (2025)
22. Weston, J., Sukhbaatar, S.: System 2 attention (is something you might need too). arXiv preprint arXiv:2311.11829 (2023)
23. Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.C., Liu, Z., Wang, L.: The dawn of lmms: Preliminary explorations with gpt-4v (ision). arXiv preprint arXiv:2309.17421 (2023)
24. Zuo, Y., Qu, S., Li, Y., Chen, Z., Zhu, X., Hua, E., Zhang, K., Ding, N., Zhou, B.: Medxpertqa: Benchmarking expert-level medical reasoning and understanding. arXiv preprint arXiv:2501.18362 (2025)