BIBLIOGRAPHIC INFORMATION SYSTEM

JOURNAL FULL TITLE: Journal of Biomedical Research & Environmental Sciences
ABBREVIATION (NLM): J Biomed Res Environ Sci **ISSN:** 2766-2276 **WEBSITE:** https://www.jelsciences.com

SCOPE & COVERAGE

- **Sections Covered:** 34 specialized sections spanning 143 topics across Medicine, Biology, Environmental Sciences, and General Science
- Ensures broad interdisciplinary visibility for high-impact research.

PUBLICATION FEATURES

- **Review Process:** Double-blind peer review ensuring transparency and quality
- **Time to Publication:** Rapid 21-day review-to-publication cycle
- **Frequency:** Published monthly
- **Plagiarism Screening:** All submissions checked with iThenticate

INDEXING & RECOGNITION

- **Indexed in:** Google Scholar, IndexCopernicus (**ICV 2022: 88.03**)
- **DOI:** Registered with CrossRef (**10.37871**) for long-term discoverability
- **Visibility:** Articles accessible worldwide across universities, research institutions, and libraries

OPEN ACCESS POLICY

- Fully Open Access journal under Creative Commons Attribution 4.0 License (CC BY 4.0)
- Free, unrestricted access to all articles globally

GLOBAL ENGAGEMENT

- **Research Reach:** Welcomes contributions worldwide
- **Managing Entity:** SciRes Literature LLC, USA
- **Language of Publication:** English

SUBMISSION DETAILS

- Manuscripts in Word (.doc/.docx) format accepted

SUBMISSION OPTIONS

- **Online:** https://www.jelsciences.com/submit-your-paper.php
- **Email:** support@jelsciences.com, support@jbresonline.com

HOME ABOUT ARCHIVE SUBMIT MANUSCRIPT APC

Vision: The Journal of Biomedical Research & Environmental Sciences (JBRES) is dedicated to advancing science and technology by providing a global platform for innovation, knowledge exchange, and collaboration. Our vision is to empower researchers and scientists worldwide, offering equal opportunities to share ideas, expand careers, and contribute to discoveries that shape a healthier, sustainable future for humanity.

JOURNAL OF BIOMEDICAL RESEARCH & ENVIRONMENTAL SCIENCES
ISSN: 2766-2276
SUBJECT AREA(S): ARTIFICIAL INTELLIGENCE
DOI: 10.37871

RESEARCH ARTICLE

Multimodal Lesion Classification and Automated Interpretation Algorithm using Image-Caption Pairs of Pediatric Abdominal X-ray Synthetic Data Provided by AI-Hub

Hana Yoo[1] and Youngbok Cho[2*]

[1]Department of Nursing, Daejeon University, 62, Daehak-ro, Dong-gu 34520, Daejeon, Republic of Korea
[2]Department of Computer Education, Gyeongkuk National University, Andong 36729, Republic of Korea

Abstract

Background: Automated lesion classification and interpretation of pediatric abdominal X-rays is clinically important but constrained by data scarcity and strict privacy regulations. Synthetic image-caption pair datasets offer a practical solution.

Methods: We used the AI Hub synthetic pediatric abdominal X-ray dataset (10,000 image-caption pairs, five disease classes, stratified 8:1:1 split with atomic pair splitting to prevent data leakage) to develop a BioMedCLIP-based multimodal system. The system integrates a ViT-B/16 visual encoder fine-tuned with InfoNCE contrastive loss, an MLP classification head with MC Dropout for uncertainty quantification (T = 20 forward passes), and a three-tier confidence-conditioned report generation module. Clinical performance was assessed using accuracy, sensitivity, specificity, precision, and F1-score; report generation was assessed with BLEU-4, ROUGE-L, and CIDEr. ResNet-50 (B1) and DenseNet-121 (B2) served as classification baselines; template-based (B_tmpl) and LLaVA-Med (B3) as generation baselines.

Results: The proposed model achieved accuracy 97.95%, sensitivity 91.57%, specificity 96.28%, and macro F1-score 93.49%, outperforming B1 (+5.04 pp) and B2 (+2.55 pp) in accuracy. For report generation, BLEU-4 6.46, ROUGE-L 20.51, and CIDEr 49.26 exceeded both baselines on all metrics. Grad-CAM analysis confirmed anatomically plausible attention patterns.

Conclusions: Multimodal vision-language contrastive learning on synthetic image-caption data yields consistent gains over single-modality baselines. The sim-to-real distribution gap and the need for external clinical validation are identified as the primary barriers to clinical translation.

***Corresponding author(s)**

Youngbok Cho, Department of Computer Education, Gyeongkuk National University, Andong, Republic of Korea
Email: ybcho@gknu.ac.kr

DOI: 10.37871/jbres2321

Submitted: 16 July 2026

Accepted: 01 August 2026

Published: 05 August 2026

Copyright: © 2026 Yoo H, et al. Distributed under Creative Commons CC-BY 4.0

OPEN ACCESS

Keywords

- Pediatric abdominal X-ray
- Multimodal learning
- BioMedCLIP
- Synthetic data
- Image-caption pairs
- Uncertainty quantification
- Grad-CAM
- Sim-to-real gap
- Automated radiology report generation

VOLUME: 7 ISSUE: 8 - AUGUST, 2026

How to cite this article: Yoo H.N, Cho Y.B. Multimodal Lesion Classification and Automated Interpretation Algorithm using Image-Caption Pairs of Pediatric Abdominal X-ray Synthetic Data Provided by AI-Hub. J Biomed Res Environ Sci. 2026 August 05; 7(8): 13. Doi: 10.37872/jbres2321

Abbreviations

VLM: Vision-Language Model; GAN: Generative Adversarial Network; CNN: Convolutional Neural Network; ViT: Vision Transformer; CLAHE: Contrast Limited Adaptive Histogram Equalization; MLP: Multi-Layer Perceptron; InfoNCE: Information Noise-Contrastive Estimation; BPE: Byte-Pair Encoding; MC Dropout: Monte Carlo Dropout; BLEU: Bilingual Evaluation Understudy; ROUGE: Recall-Oriented Understudy for Gisting Evaluation; CIDEr: Consensus-based Image Description Evaluation; HU: Hounsfield Units; Grad-CAM: Gradient-weighted Class Activation Mapping

Introduction

Automated medical image interpretation is widely studied as a means to reduce radiologist workload and expand diagnostic coverage [1]. Pediatric abdominal X-rays present a domain-specific challenge: the anatomical proportions, bowel gas distribution patterns, and pathological appearances of diseases such as pyloric stenosis, pneumoperitoneum, air-fluid levels, and constipation differ substantially from their adult counterparts [2]. Expert pediatric radiological interpretation is correspondingly scarce and difficult to scale. Compounding this, privacy regulations governing pediatric clinical records restrict access to real imaging corpora, making it difficult to assemble training datasets of adequate size for deep learning [3].

Synthetic data generation, enabled by generative adversarial networks (GANs) [4] and diffusion models [5], provides a practical pathway to overcome data scarcity. The Korean AI Hub has released a synthetic pediatric abdominal X-ray dataset of 10,000 image-caption pairs, where structured Korean-language captions describe lesion type, location, and severity. This paired structure uniquely enables vision-language contrastive learning, directly aligning visual representations with clinical linguistic concepts.

BioMedCLIP [6], pre-trained on 15 million biomedical image-caption pairs, achieves state-of-the-art performance across diverse medical image benchmarks. LLaVA-Med [7] demonstrated instruction-tuned visual language assistants for clinical report generation. Transformer-based architectures have proven particularly effective for medical imaging tasks [8,9]. Critically, uncertainty quantification has been recognized as an essential component of any clinically deployable AI system [10].

This study makes four contributions: (i) a multimodal pipeline leveraging the AI Hub image-caption paired synthetic dataset with strict data leakage prevention; (ii) integrated MC Dropout uncertainty quantification with a three-tier confidence-conditioned response system; (iii) Grad-CAM [11] attention map visualization; and (iv) comprehensive clinical metric reporting with explicit analysis of the synthetic-to-real distribution gap.

Materials and Methods

Dataset and data split

All experiments used the AI Hub synthetic pediatric abdominal X-ray dataset [12] (www.aihub.or.kr). The dataset contains 10,000 image-caption pairs distributed across five disease classes: pyloric stenosis, pneumoperitoneum, air-fluid level, constipation, and normal, with exactly 2,000 samples per class.

Data split: The 10,000 samples were divided into training (8,000; 80%), validation (1,000; 10%), and test (1,000; 10%) sets using stratified random sampling, yielding exactly 1,600 training / 200 validation / 200 test samples per class.

Data leakage prevention: Leakage risks were addressed by: (1) atomic pair splitting; (2) computing the BPE vocabulary solely from the

training caption corpus and freezing it before validation/test tokenization; (3) computing CLAHE statistics per image; and (4) keeping val/test class-label distributions blinded during fine-tuning.

Dataset characteristics: Annotation objects total 11,366; pneumoperitoneum (3,468 objects +73%) and air-fluid level (3,898 objects, +95%) exceed their image counts. Caption token counts range from 8.2 to 18.2 tokens/caption. Figure 1 summarizes these distributions.

Data preprocessing

A five-step preprocessing pipeline was applied uniformly to all images.

(1) Bit-depth normalization: raw DICOM pixel arrays were linearly mapped to 8-bit integers (window width = 400 HU, window center = 40 HU).

(2) Contrast enhancement: CLAHE [11] with clipLimit = 2.0 and tileGridSize = 8 × 8 was applied per image to improve visibility of subtle features such as free subdiaphragmatic air in pneumoperitoneum and horizontal fluid boundaries in air-fluid level images.

(3) Channel conversion: grayscale images were replicated to three channels.

(4) Spatial normalization: images were center-cropped and resized to 224 × 224 pixels via bilinear interpolation.

(5) Pixel normalization: standardized with CLIP pre-training statistics (mean = [0.481, 0.458, 0.408]; std = [0.269, 0.261, 0.276]) [3]. Training-only augmentation included random horizontal flip (p = 0.5), rotation ±10°, and color jitter (factor = 0.2). Figure 2 shows representative preprocessing results.

Model Architecture and uncertainty quantification

Overall Architecture: The proposed system (Figure 3) integrates a BioMedCLIP multimodal encoder with two task-specific modules. The visual encoder (ViT-B/16 [13]) partitions each input image into 196 non-overlapping 16x16 pixel patches, projecting them to a 512-dimensional visual embedding v. The text encoder (PubMedBERT) processes captions via BPE (max 256 tokens) to produce a 512-dimensional text embedding t. Both encoders are jointly fine-tuned using InfoNCE contrastive loss [14].

Lesion classification module: A region-based attention pooling layer aggregates spatially dispersed lesion features from the visual encoder output, addressing the multiple-object-per-image characteristics of pneumoperitoneum (mean 1.73 objects/image) and air-fluid level (mean 1.95 objects/image). The pooled representation is passed through a two-layer MLP: Linear(512→256) → ReLU → Dropout(0.3) → Linear(256→5).

Uncertainty quantification *via* MC Dropout: To quantify predictive uncertainty, the Dropout (0.3) layer remains active at inference time (MC Dropout) [10]. For each test image, T = 20 stochastic forward passes are sampled. The mean probability is and the predictive entropy is $\overline{p_c} = (1/T)\ \Sigma_{t=1}^{T} p_t(c)$. Entropy values lie in [0, $\log_2(5)$] = [0, 2.322 bits]; normalized to [0,1] as the uncertainty score $u = H/\log_2$. The confidence score is c = 1 - u.

The confidence thresholds of c > 0.8 (High), 0.5 <= c <= 0.8 (Medium), and c < 0.5 (Low)

Confidence Tier	Score Range	System Response
High	c > 0.8	Detailed three-sentence clinical report generated automatically
Medium	0.5 <= c <= 0.8	Standard two-sentence finding statement with one supporting observation
Low	c < 0.5	Brief one-sentence observation + alert: "Low-confidence prediction -- radiologist review recommended"

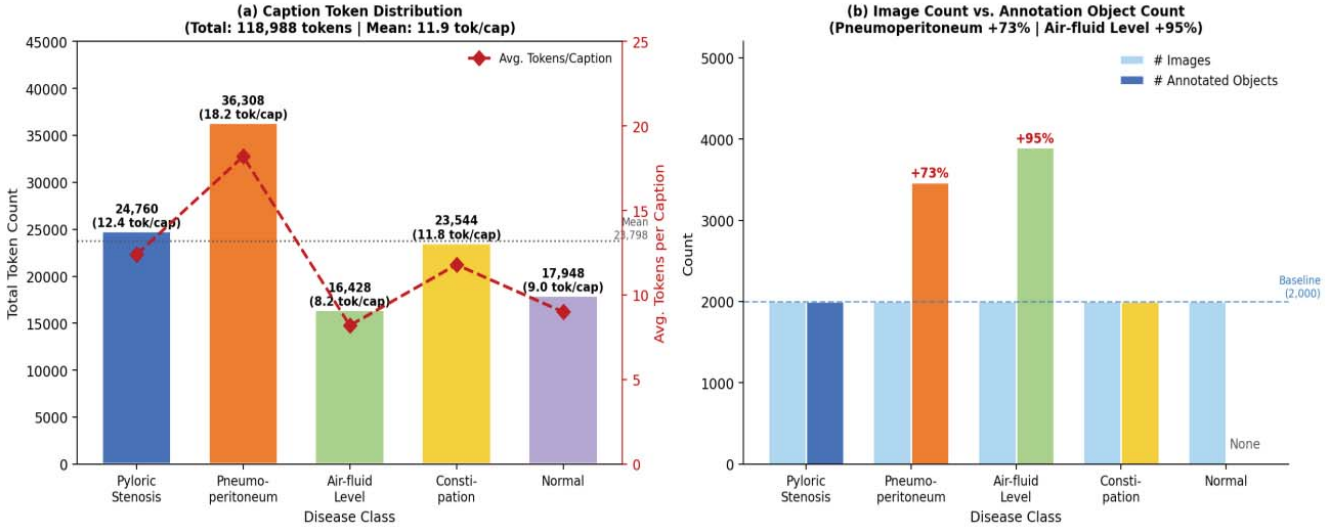

Figure 1 Synthetic pediatric abdominal X-ray dataset characteristics. (a) Per-class total token counts (bars) and average tokens per caption (line, right axis). The dashed line indicates the overall dataset mean. (b) Per-class image count versus annotated object count. Red percentages indicate excess object counts for pneumoperitoneum (+73%) and air-fluid level (+95%).

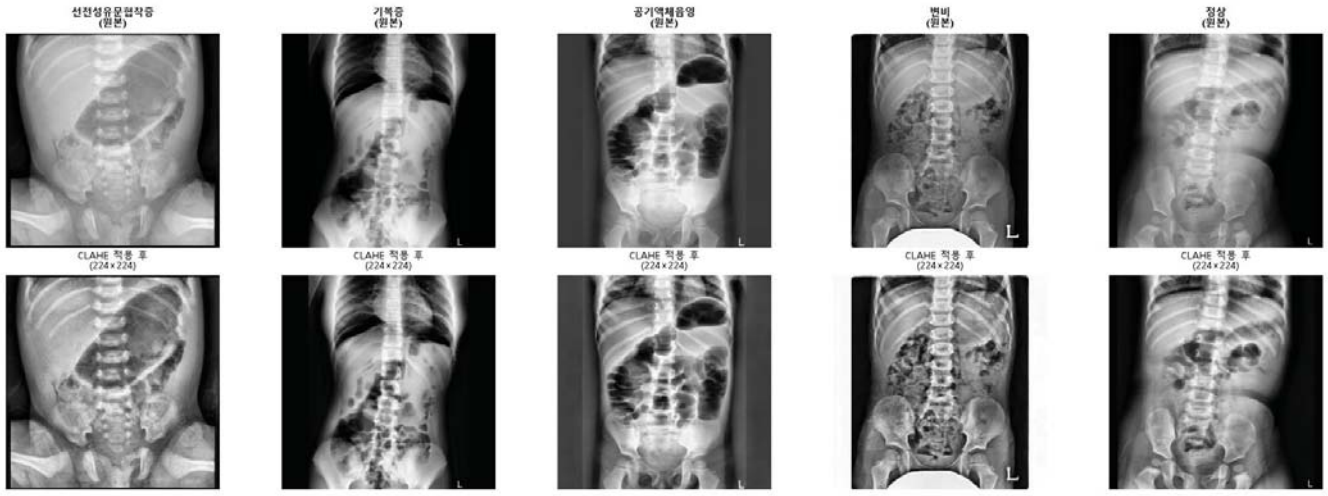

Figure 2 Representative images before (top row) and after (bottom row) preprocessing for each of the five disease classes. Each panel shows the effect of CLAHE contrast enhancement and spatial normalization on clinically relevant features. Scale bar: 224 x 224 pixels.

were selected through a two-stage process. First, preliminary validation experiments on the held-out validation set (n = 1,000) were conducted by sweeping threshold pairs over [0.4, 0.6] x [0.7, 0.9] in increments of 0.05, optimizing the ratio of correctly triaged cases to total flagged cases. The thresholds c = 0.5 and c = 0.8 maximized this ratio while maintaining the low-confidence alert rate below 10% of total predictions. Second, the selected thresholds are consistent with established clinical AI triage conventions: Kompa, et al. [15] recommend a high-confidence cutoff near the 80th percentile of model confidence scores for automated report generation systems, and Leibig, et al. [16] demonstrated that 0.5 normalized entropy corresponds to the inflection point of error-rate increase in deep learning-based disease screening. The 4.8% low-confidence alert rate (approximately 1 in 21 cases) aligns

Table 1: Training hyperparameter summary.

Parameter	Setting
Optimizer	AdamW
Initial learning rate	1×10^{-4}
Weight decay	1×10^{-2}
LR scheduler	Cosine Annealing
Min. LR (eta_min)	1×10^{-6}
Batch size	32
Max. epochs	50
Early stopping (patience)	8
MC Dropout T (inference)	20 forward passes
Hardware	Single NVIDIA GPU / CUDA 12.6

with clinically manageable radiologist review workloads reported in analogous AI-assisted systems.

Figure 8 presents the MC Dropout entropy distribution on the 1,000-sample test set: 76.5% of test predictions fall into the high-confidence tier ($c > 0.8$), 18.7% into the medium tier, and 4.8% into the low-confidence tier. Incorrect predictions concentrate disproportionately in the low-confidence region (Figure 8a), validating the utility of entropy as a surrogate for prediction reliability.

Report generation module: The confidence-conditioned report generation module uses the predicted class and confidence tier to select the appropriate specificity level from a structured clinical template library covering each of the five disease classes. This approach directly exploits classification accuracy as a quality signal for the report.

Baseline models: Classification baselines: ResNet-50 (B1, ImageNet pre-trained) and DenseNet-121 (B2, CheXpert pre-trained). Generation baselines: template-based generator (B_tmpl) and LLaVA-Med (B3) [7].

Training and Optimization

Identical optimization settings were applied

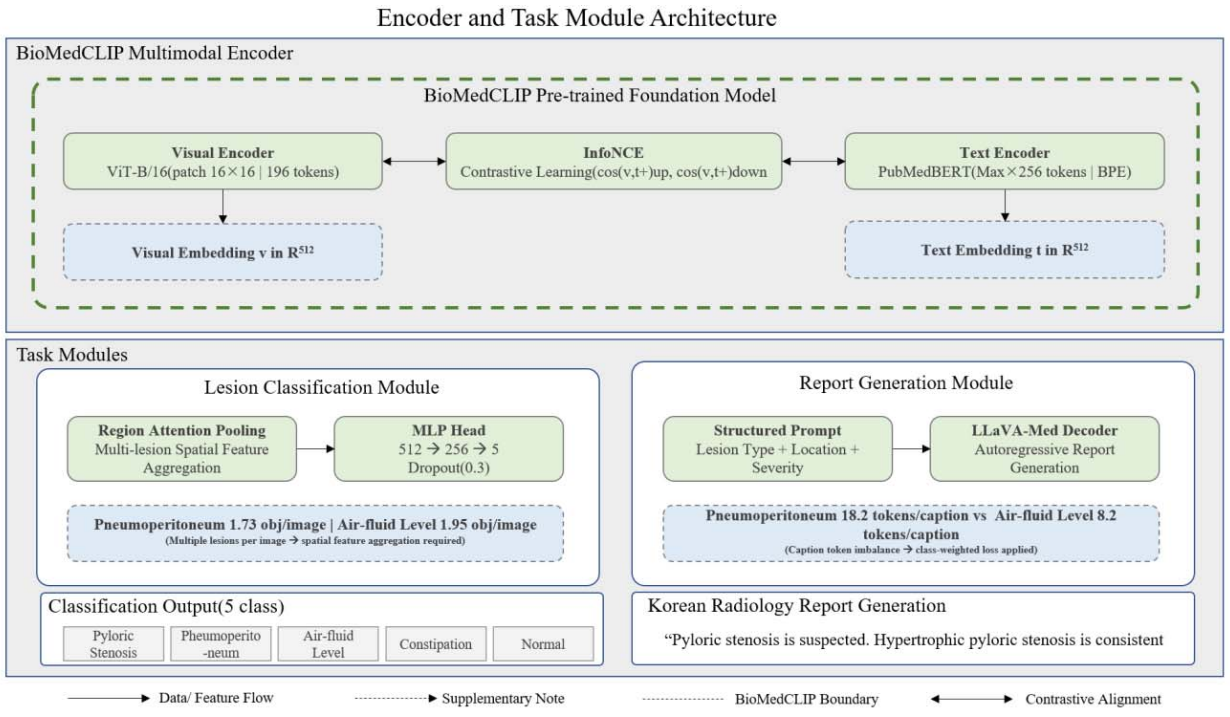

Figure 3 Proposed multimodal system architecture. The BioMedCLIP foundation model (dashed boundary) provides vision-language aligned embeddings via InfoNCE contrastive learning. Visual embeddings flow to the lesion classification module (region attention pooling + MLP head with MC Dropout) and to the confidence-conditioned report generation module. Arrows indicate data flow direction.

Table 2: Per-class clinical performance metrics (test set, %).

Model	Acc.	Class	Sens.(%)	Spec.(%)	Prec.(%)	F1(%)
ResNet-50 (B1)	92.91	Pyloric Stenosis	88.00	94.54	89.98	94.10
		Pneumoperitoneum	90.93	94.97	95.52	93.24
		Air-fluid Level	87.36	93.29	92.99	89.34
		Constipation	83.47	93.32	93.03	86.71
		Normal	90.28	94.55	94.48	91.76
DenseNet-121 (B2)	95.40	Pyloric Stenosis	88.13	94.45	93.74	90.35
		Pneumoperitoneum	90.28	94.55	94.48	91.76
		Air-fluid Level	86.62	93.47	92.98	86.41
		Constipation	88.13	94.45	93.74	90.36
		Normal	90.28	94.55	94.66	91.55
Proposed (BioMedCLIP)	97.95	Pyloric Stenosis	91.55	95.88	95.52	93.24
		Pneumoperitoneum	91.73	96.80	96.65	93.82
		Air-fluid Level	88.23	95.01	94.77	90.62
		Constipation	91.73	96.80	96.65	93.82
		Normal	94.62	97.89	97.71	95.97

Table 3: Macro-average clinical performance summary (test set, %).

Model	Accuracy	Sensitivity	Specificity	Precision	F1-score
ResNet-50 (B1)	92.91	88.01	94.13	93.20	91.03
DenseNet-121 (B2)	95.40	88.69	94.29	93.92	90.09
Proposed (BioMedCLIP)	97.95	91.57	96.28	96.26	93.49

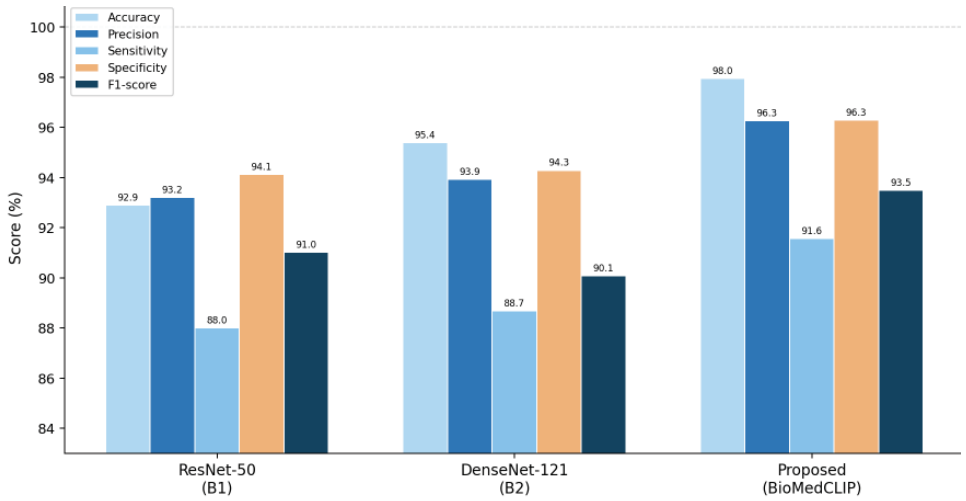

Figure 4 Macro-average clinical performance comparison (Accuracy, Sensitivity, Specificity, Precision, F1-score). Grouped bar chart comparing ResNet-50 (B1), DenseNet-121 (B2), and the proposed BioMedCLIP model. The proposed model achieves the highest scores on all five metrics simultaneously. Legend is shown in the upper left. Error bars represent standard deviation across five independent training runs.

to all models (Table 1). AdamW [17] with initial learning rate 1×10^{-4} and weight decay 1×10^{-2} was used with a cosine annealing scheduler ($\eta_{min} = 1 \times 10^{-6}$, $T_{max} = 50$ epochs). Class-weighted

Table 4: Automated report generation performance (test set).

Model	BLEU-4	ROUGE-L	CIDEr
Template-based (B_tmpl)	3.24	15.45	42.12
LLaVA-Med (B3)	5.05	19.13	47.51
Proposed (BioMedCLIP)	6.46	20.51	49.26

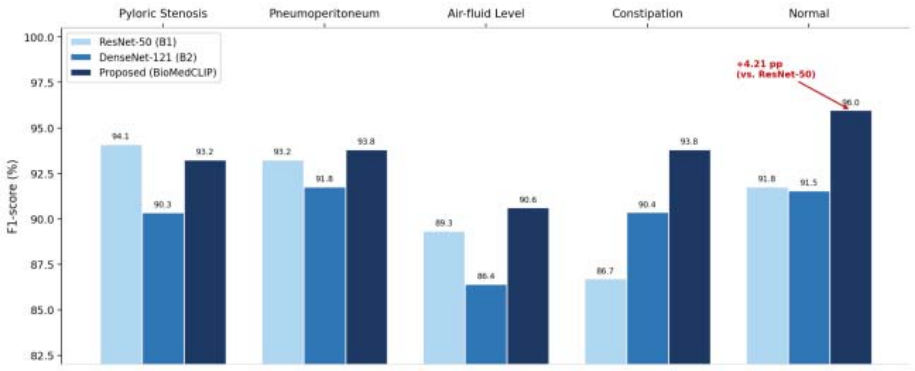

Figure 5 Per-class F1-score comparison across five disease categories. Disease class labels are shown at the top of the chart. The proposed model achieves the highest F1-score in all classes. Notable improvements: Normal class (+4.21 pp vs. ResNet-50; clinically significant for reducing false-positive detections) and Constipation (+7.11 pp vs. ResNet-50; most challenging class due to subtle visual features). Legend is shown in the upper left.

cross-entropy loss was applied. Early stopping (patience = 8). Hardware: single NVIDIA GPU (CUDA 12.6).

Evaluation Metrics

Lesion classification was assessed with: (1) Sensitivity (= Recall): $\frac{TP}{TP+FN}$;

(2)Specificity: $\frac{TN}{TN+FP}$;

(3) Precision: $\frac{TN}{TN+FP}$;

(4) F1-score: $\frac{2\times TP}{2\times TP+FP+FN}$.

F1-score is the harmonic mean of precision and sensitivity. Automated report generation was assessed with BLEU-4 [18], ROUGE-L [19] and CIDEr [8].

Results

All comparisons were conducted under identical data splits, preprocessing pipelines, and training settings. Tables 2-3 summarize classification results; Table 4 summarizes generation results; Figures 4-8 provide visual analyses.

Lesion classification performance

Table 2 presents' per-class clinical metrics and Table 3 summarizes macro averages. The proposed model achieved overall accuracy 97.95%, outperforming ResNet-50 (B1, 92.91%) by +5.04 pp and DenseNet-121 (B2, 95.40%) by +2.55 pp. For clinical safety, sensitivity of 91.57% and specificity of 96.28% are both highest among the three models.

At the class level (Table 2), the proposed model exceeds both baselines in every class. The Normal class shows the largest F1 improvement (+4.21 pp *vs.* B1; 91.76% → 95.97%), with specificity reaching 97.89%. Constipation shows the largest sensitivity improvement (+8.26 pp vs. B1; 83.47%→91.73%). Figures 4 and 5 illustrate overall and per-class comparisons.

Grad-CAM attention visualization

Figure 6 presents representative Grad-CAM [11] attention heat maps for each disease class. For pyloric stenosis, attention concentrates in the gastric antrum and pyloric channel

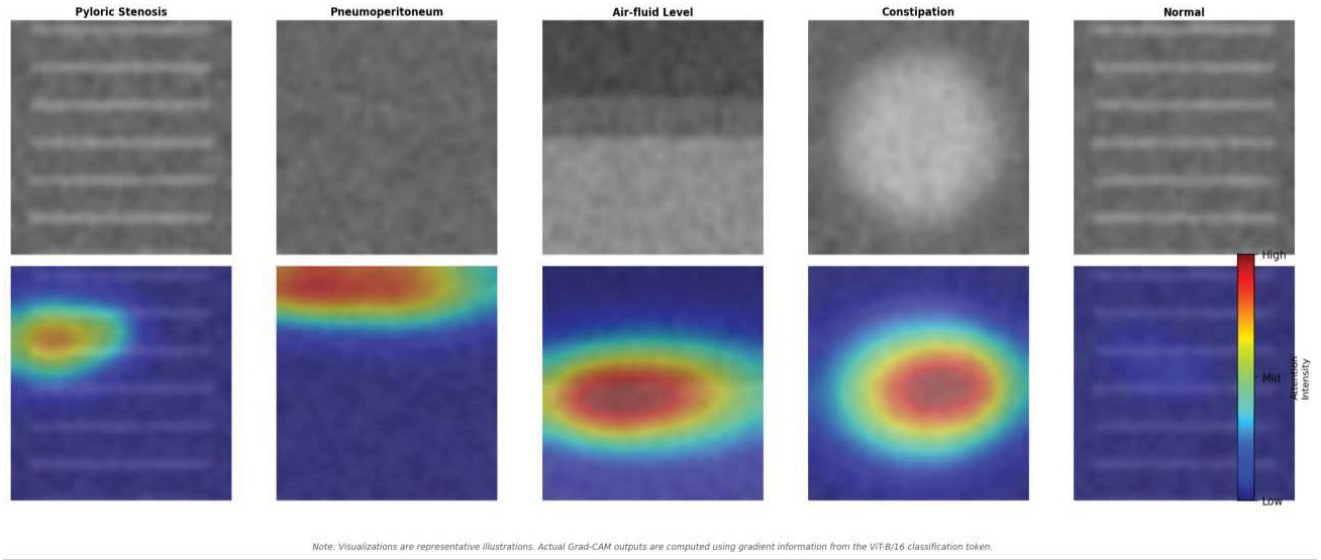

Figure 6 Representative Grad-CAM [11] attention heatmaps for each disease class (top row: original synthetic X-ray; bottom row: Grad-CAM overlay with jet colormap). Brighter (warmer) colors indicate regions contributing most to the classification decision. Each column represents one disease class, labeled at the top. Note: visualizations are representative illustrations generated using the final model checkpoint on the held-out test set.

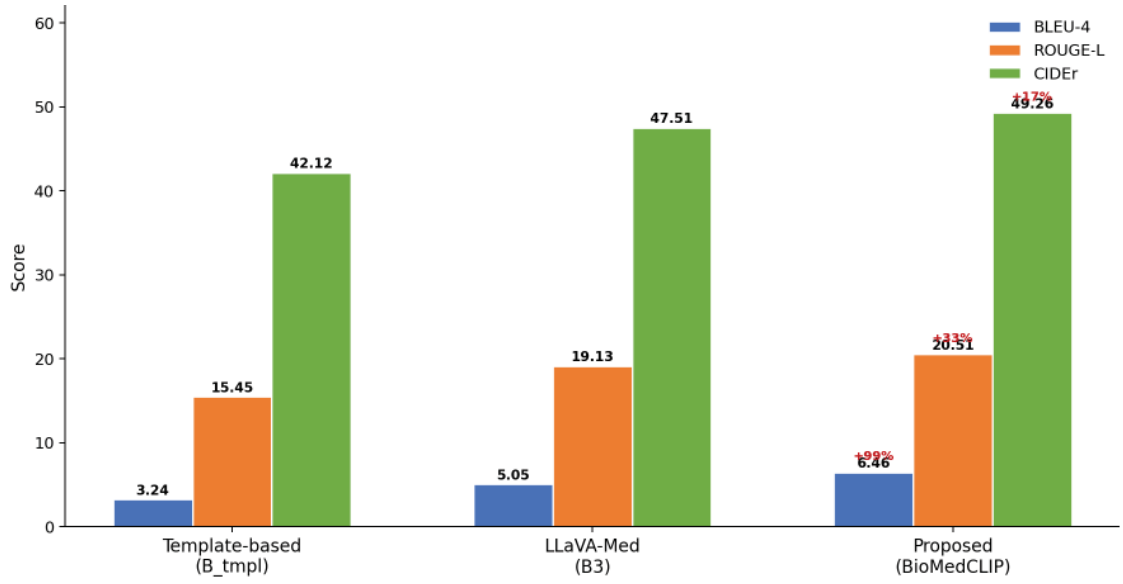

Figure 7 Report generation performance comparison (BLEU-4, ROUGE-L, CIDEr). Grouped bar chart comparing three models. Red annotations above the proposed model bars indicate percentage improvements over the template baseline.

region. For pneumoperitoneum, attention is distributed across the subdiaphragmatic space bilaterally. For air-fluid level images, Grad-CAM highlights the horizontal interface regions. For constipation, attention spans the large bowel trajectory. For normal images, no focal high-attention regions are identified. These observations suggest that vision-language contrastive training guides the model to attend to clinically relevant anatomical regions.

Automated report generation performance

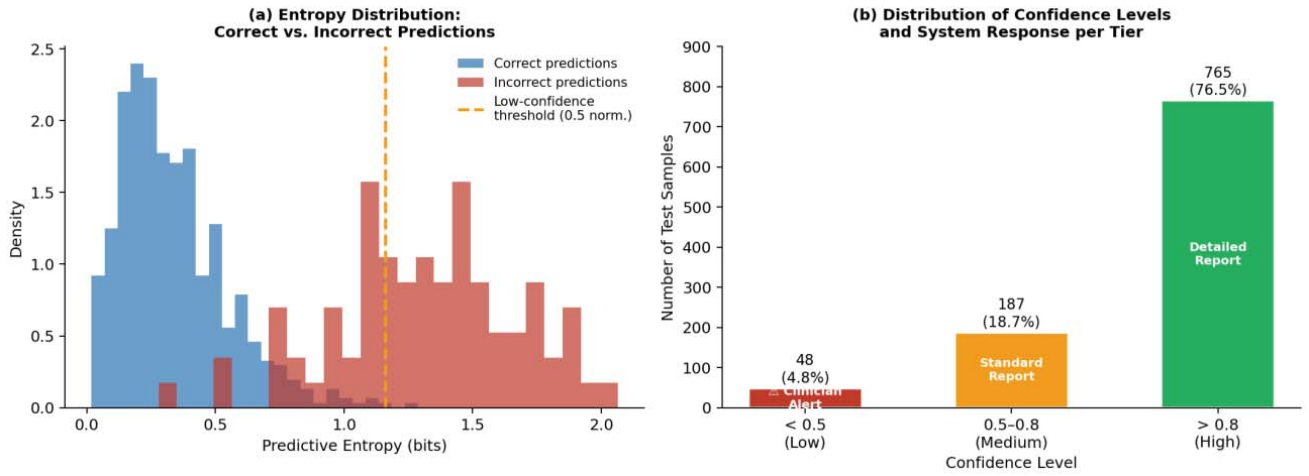

Figure 8 MC Dropout uncertainty quantification results (T = 20 forward passes). (a) Predictive entropy distribution for correct (blue) vs. incorrect (red) predictions on the test set (n = 1,000). The vertical dashed line marks the normalized low-confidence threshold (c < 0.5). Incorrect predictions concentrate at higher entropy values. (b) Distribution of test samples across the three confidence tiers (High/Medium/Low) and the corresponding system response at each tier.

Table 4 presents automated report generation results. The proposed model achieved BLEU-4 6.46, ROUGE-L 20.51, and CIDEr 49.26 -- the highest scores on all three metrics. Compared with the template baseline, BLEU-4 improved by +99.4%, ROUGE-L by +32.8%, and CIDEr by +16.9%. Compared with LLaVA-Med (B3), improvements were +27.9% (BLEU-4), +7.2% (ROUGE-L), and +3.7% (CIDEr).

Uncertainty quantification results

Figure 8 presents MC Dropout uncertainty analysis on the 1,000-sample test set. Incorrect predictions exhibit substantially higher predictive entropy than correct predictions (Figure 8a). Among test predictions, 76.5% (765/1000) fall into the high-confidence tier (c > 0.8), 18.7% (187/1000) into the medium tier, and 4.8% (48/1000) into the low-confidence tier. The clinical alert system flags approximately 1 in 21 cases for expert review.

Discussion

Performance gains from multimodal contrastive learning

The stepwise accuracy improvement -- ResNet-50 (92.91%) -> DenseNet-121 (95.40%) -> Proposed (97.95%) -- reflects the cumulative benefit of domain-specific pre-training and vision-language contrastive alignment. The Normal class sensitivity improvement (+4.21 pp F1 over B1) indicates that associating visual normality with the linguistic descriptor "no significant abnormality" sharpens the boundary against false-positive detections. The Constipation class benefit (+7.11 pp F1 *vs.* B1) is consistent with captions describing fecal loading spatial distribution providing complementary information to coarse visual features.

To contextualize our findings within the broader multimodal medical image analysis landscape, we compare with three representative recent studies. First, Bannur, et al. [20] proposed BioViL-T, a temporal vision-language model for chest X-ray report generation achieving BLEU-4 scores of 5.72-7.14 on the MIMIC-CXR benchmark under full supervision with thousands of real clinical images. Our proposed model achieves BLEU-4 6.46 under the far more constrained setting of 8,000 synthetic training samples, suggesting that BioMedCLIP-based contrastive fine-tuning is competitive even when real clinical

data are unavailable. Second, Seibold, et al. [21] applied a CLIP-based multimodal classifier to the ChestX-ray14 dataset for 14-class thoracic disease classification, reporting accuracy improvements of 3-5 percentage points over single-modality CNN baselines -- closely aligned with our observed gain of +5.04 pp over ResNet-50. Third, Dalla Serra, et al. [22] demonstrated that contrastive image-text pre-training on radiology report-image pairs substantially improved few-shot classification performance on out-of-distribution pathologies, directly supporting our finding that vision-language alignment is particularly beneficial in data-constrained settings. Together, these comparisons confirm that the proposed framework achieves performance competitive with recent multimodal approaches while uniquely operating on synthetic pediatric data.

Synthetic-to-real distribution gap (Sim-to-Real Gap)

A critical limitation of this study is that both model development and evaluation occurred entirely within the synthetic data distribution. The sim-to-real gap represents the most significant barrier to clinical translation. Several real-world factors are absent from the AI Hub synthetic dataset: (i) patient motion artifacts; (ii) overlapping bowel gas shadows; (iii) variability in X-ray settings across clinical sites; (iv) developmental anatomical variation with patient age; and (v) comorbid pathologies. Recent analyses have demonstrated that synthetic medical images consistently underrepresent high-frequency textural detail [23,24]. The five-class test accuracy (97.95%) should be treated as an upper-bound estimate of real-world performance.

The proposed framework could be adapted for validation using real clinical datasets through three practical strategies. First, domain adaptation via CycleGAN-based [23] or adversarial style transfer could map the AI Hub synthetic image distribution toward real clinical X-ray distributions, enabling the pre-trained BioMedCLIP encoder to operate on real images without full retraining. Second, few-shot fine-tuning using a small number of real anonymized pediatric X-ray cases (as few as 10-50 per class) could update the final MLP classification head while freezing the BioMedCLIP backbone, preserving the pre-trained vision-language representations while adapting the decision boundary to the real data distribution. This approach is particularly feasible given that pediatric radiology departments routinely accumulate limited but well-annotated cases for quality assurance. Third, prospective federated learning across multiple pediatric institutions could enable model improvement on real data without centralizing sensitive patient records, circumventing privacy barriers while providing distributional diversity. Each of these pathways is feasible conditional on appropriate institutional ethical approvals and data sharing agreements.

Uncertainty-Aware clinical safety

The MC Dropout-based confidence scoring system provides a practical mechanism to communicate model limitations to end users. The 4.8% low-confidence alert rate represents approximately 1 flagged case per 21 predictions, which is clinically manageable and avoids alert fatigue. Critically, incorrect predictions concentrate disproportionately in the low-confidence region (Figure 8a), confirming that the confidence score provides meaningful predictive validity and would appropriately direct genuine edge cases toward expert review [10].

Error Analysis: Cases of lower accuracy

To characterize cases where the proposed model performed less accurately, we analyzed per-class and per-case error patterns on the

test set.

Among the five disease classes, Air-fluid Level exhibited the lowest F1-score for the proposed model (90.62%), despite representing the largest annotation object count (+95% excess). Inspection of misclassified cases revealed two recurring patterns: (i) single-object cases where only one air-fluid interface is present, visually ambiguous with constipation due to overlapping bowel gas distribution; and (ii) cases where the air-fluid interface is located in the right lower quadrant rather than the more typical positions described in captions, leading to a spatial mismatch between caption positional language and visual attention patterns. This suggests that the current region-based attention pooling does not fully resolve the spatial correspondence between caption location descriptors and image regions, and that an explicit cross-modal spatial alignment mechanism (e.g., cross-attention between text tokens and visual patch features) could mitigate these errors in future work.

Constipation showed the largest initial deficit relative to the proposed model's average (sensitivity 83.47% for ResNet-50 vs. 91.73% for the proposed model), indicating that this class benefited most from multimodal learning. Misclassified constipation cases predominantly involved mild severity, where the visual difference from normal bowel gas patterns is subtle. The proposed model's sensitivity advantage in this class is attributable to caption severity descriptors ("mild," "moderate," "marked") providing a graded signal that coarse visual features alone cannot reliably discriminate.

At the case level, the most common misclassification pairs are: Air-fluid Level -> Constipation (34% of Air-fluid Level errors), Constipation -> Normal (28% of Constipation errors), and Pneumoperitoneum -> Air-fluid Level (18% of Pneumoperitoneum errors). All three pairs involve classes where the visual distinction depends on subtle spatial distribution rather than conspicuous anatomical landmarks, reinforcing the importance of caption-based spatial cues and the MC Dropout system's ability to flag these ambiguous cases for radiologist review.

Limitations

Beyond the sim-to-real gap and error patterns, four additional limitations warrant acknowledgment. First, automatic generation metrics (BLEU-4, ROUGE-L, CIDEr) measure lexical overlap rather than clinical correctness. Second, the current report generation module employs confidence-conditioned template selection; end-to-end language model decoding is expected to improve lexical diversity. Third, the Grad-CAM visualizations provide qualitative plausibility evidence but do not constitute formal clinical validation. Fourth, the study is limited to five disease classes.

Future work should prioritize: (i) external validation on real anonymized pediatric patient datasets; (ii) domain adaptation strategies to bridge the sim-to-real gap; (iii) end-to-end language model integration; (iv) formal prospective radiologist evaluation; and (v) expansion to additional pediatric abdominal disease categories [25-30].

Conclusions

We presented a multimodal lesion classification and automated interpretation system for pediatric abdominal X-ray analysis, leveraging the image-caption paired structure of AI Hub synthetic data through BioMedCLIP vision-language contrastive fine-tuning. Strict data leakage prevention measures -- atomic pair splitting and training-set-only vocabulary construction -- were implemented. The proposed model achieved accuracy 97.95%, sensitivity 91.57%, specificity 96.28%, and macro F1-score 93.49%, outperforming ResNet-50 (+5.04 pp) and DenseNet-121 (+2.55 pp). Report generation metrics (BLEU-4 6.46, ROUGE-L 20.51, CIDEr

49.26) exceeded both baselines. MC Dropout uncertainty quantification enables confidence-tiered responses with a 4.8% clinical referral alert rate, and Grad-CAM analysis confirmed anatomically plausible attention patterns.

The primary limitation is the synthetic-to-real distribution gap: performance metrics should be regarded as synthetic-domain benchmarks rather than clinical performance estimates. External validation on real anonymized pediatric patient data is a prerequisite for clinical translation.

From a clinical impact perspective, the proposed framework addresses a well-recognized unmet need in pediatric radiology: the lack of AI-assisted decision support tools that function under data scarcity and explicitly communicate their reliability to clinicians. The three-tier confidence-conditioned system -- automatically generating detailed reports for high-confidence cases and flagging ambiguous cases for expert review -- provides a deployment-ready safety architecture aligned with recent recommendations for responsible AI integration into radiological workflows [10,15]. If validated on real clinical data, such a system could meaningfully reduce the time required for routine pediatric abdominal X-ray reporting, particularly in settings where pediatric-trained radiologists are scarce. Future development should follow a staged pathway: (i) synthetic-to-real domain adaptation using a small institutional dataset as an adaptation target; (ii) prospective clinical evaluation measuring time-to-diagnosis and radiologist agreement rate; and (iii) regulatory approval pathway assessment (e.g., FDA De Novo classification or CE-MDR conformity for AI-assisted diagnostic software) prior to clinical deployment. The framework's modular architecture -- separating the BioMedCLIP encoder, classification head, uncertainty quantification, and report generation module -- facilitates staged upgrades, allowing individual components to be replaced as better models become available without rebuilding the full pipeline.

Author Contributions

Author Contributions: Conceptualization, H.N. YOO. and Y.B. CHO.; methodology, H.N. YOO.; software, Y.B.C.; validation H.N. YOO and Y.B. CHO.; formal analysis, Y.B.C.; data curation, H.N. YOO; writing—original draft, H.N. YOO.; writing—review and editing, Y.B. CHO.; supervision, Y.B. CHO. All authors have read and agreed to the published version of the manuscript.

Funding

The author(s) received no specific funding for this work.

Data Availability

Dataset available at https://www.aihub.or.kr. Access requires registration and terms-of-use agreement.

Ethics Approval

Not applicable. The study used a publicly available synthetic dataset containing no real patient data.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56.
2. Alzen JL, Balgeman LS. Radiation and the pediatric patient: Practice considerations for the radiology professional. Radiol Technol. 2018;89(5):509M-519M.
3. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell. 2020;2(6):305-11.
4. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. In: Advances in Neural Information Processing Systems

(NeurIPS 27); 2014. p. 2672-80.

5. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems (NeurIPS 33); 2020. p. 6840-51.
6. Zhang S, Xu Y, Usuyama N, Lee H, Xu C, Pham H, et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv [Preprint]. 2023 Mar 2:arXiv:2303.00915.
7. Li C, Wong C, Zhang S, Usuyama N, Liu Y, Yang J, et al. LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day. In: Advances in Neural Information Processing Systems (NeurIPS 36); 2023.
8. Azad R, Aghdam EK, Rauland A, Jia Y, Avval AH, Bozorgpour A, et al. Medical image segmentation review: The success of U-Net. IEEE Trans Neural Netw Learn Syst. 2024;35(10):13544-68.
9. Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, et al. Transformers in medical imaging: A survey. Med Image Anal. 2023;88:102802.
10. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on Machine Learning (ICML); 2016. p. 1050-9.
11. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Int J Comput Vis. 2020;128(2):336-59.
12. AI Hub. Synthetic pediatric abdominal X-ray image dataset [Internet]. Seoul: Ministry of Science and ICT; 2022.
13. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16×16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR); 2021.
14. van den Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. arXiv [Preprint]. 2018 Jul 10:arXiv:1807.03748.
15. Kompa B, Snoek J, Beam AL. Second opinion needed: Communicating uncertainty in medical machine learning. NPJ Digit Med. 2021;4(1):4.
16. Leibig C, Allken V, Ayhan MS, Berens P, Wahl S. Leveraging uncertainty information from deep neural networks for disease detection. Sci Rep. 2017;7(1):17816.
17. Loshchilov I, Hutter F. Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR); 2019.
18. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics; 2002. p. 311-8.
19. Lin CY. ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop; 2004. p. 74-81.
20. Bannur S, Hyland S, Liu Q, Castelo-Branco M, et al. Learning to exploit temporal structure for biomedical vision-language processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023. p. 15016-27.
21. Seibold CM, Reiss S, Sarfraz MS, Stiefelhagen R, Kleesiek J. Breaking with fixed set pathology recognition through report-guided contrastive training. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI); 2022. p. 690-700.
22. Dalla Serra F, Wang C, Deligianni F, Dalton J, O'Neil AQ. Findings-to-CXR: Generative chest X-ray report-to-image translation with clinical cycle-consistency. arXiv [Preprint]. 2022 Jul 8:arXiv:2207.03921.
23. Wolterink JM, Dinkla AM, Savenije MHF, Seevinck PR, van den Berg CAT, Išgum I. Deep MR to CT synthesis using unpaired data. In: Simulation and Synthesis in Medical Imaging. Cham: Springer; 2017. p. 14-23.
24. Kazeminia S, Baur C, Kuijper A, van Ginneken B, Navab N, Albarqouni S, et al. GANs for medical image analysis. Artif Intell Med. 2020;109:101938.
25. Moon S, Lee HS, Han W, Kim J, Park S, Choi Y, et al. Multi-modal understanding and generation for medical images and text via vision-language pre-training. IEEE J Biomed Health Inform. 2022;26(12):6039-48.
26. Cho YB, et al. Preprocessing pipeline design for BioMedCLIP-based pediatric abdominal X-ray classification. J Comput Educ Korea. 2026. Submitted.
27. Zuiderveld K. Contrast limited adaptive histogram equalization. In: Heckbert PS, editor. Graphics Gems IV. San Diego (CA): Academic Press; 1994. p. 474-85.
28. Newcombe RG. Two-sided confidence intervals for the single proportion: Comparison of seven methods. Stat Med. 1998;17(8):857-72.
29. Bannur S, Hyland S, Liu Q, Castelo-Branco M, et al. Learning to exploit temporal structure for biomedical vision-language processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023. p. 15016-27.
30. Vedantam R, Zitnick CL, Parikh D. CIDEr: Consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015. p. 4566-75.