Progressive Visual Grounding for Medical Reasoning via Automated Chain-of-Thought Data

Soo Yong Kim*, Suin Cho*, Vincent-Daniel Yun*, Gyeongyeon Hwang*
MODULABS
Seoul, South Korea
*Equal contribution

***Abstract*—Bridging clinical diagnostic reasoning with AI remains a central challenge in medical imaging. We introduce MedCLM, an automated pipeline that converts detection datasets into large-scale medical visual question answering (VQA) data with Chain-of-Thought reasoning by linking lesion boxes to organ segmentation and structured rationales. Starting from 156k raw generations, quality control mechanisms yield 125k high-fidelity CoT-VQA samples with anatomically grounded rationales. Human evaluation by we authors with help of certified radiologist on 500 randomly sampled instances confirms 91.2% clinical correctness in our generated data, validating pipeline reliability before model training. To utilize this data effectively, we propose an Integrated CoT-Curriculum Strategy that progressively transitions from explicit visual grounding to implicit localization. On VQA-RAD, SLAKE, PMC-VQA, and challenging benchmarks including OmniMedVQA, MedXpertQA, and BESTMVQA,MedCLM achieves state-of-the-art performance, attaining 72.6% LLM-judge win-rate over SOTA baseline on radiology report generation when using GPT-5 as judge.**

I. Introduction

Medical VLMs are essential for clinical decision support, enabling systems that answer queries directly from medical images [26], [27]. VQA is a central task in this field [12], [14]. Early datasets such as VQA-RAD [12] established the foundation but remain limited in scale and reasoning depth due to costly expert annotation. SLAKE [13] and PMC-VQA [14] expanded coverage, yet most benchmarks still focus on short question-answering without explicit diagnostic reasoning. More recently, challenging benchmarks such as OmniMedVQA [15], MedXpertQA [16], and BESTMVQA [17] have emerged to address benchmark saturation concerns, requiring models to demonstrate robust reasoning across diverse medical imaging modalities. This limits interpretability and clinical trust, as physicians require transparent rationales to validate model outputs in high-stakes scenarios [30], [42].

CoT prompting [34] improves reasoning in LLMs by producing intermediate steps [35]. It has proven effective across multimodal domains [26], [27] and is particularly relevant to medicine, where step-by-step reasoning aligns with clinical workflows [30]. However, constructing large-scale CoT data remains costly, typically relying on proprietary models and few-shot generation [30]. Existing approaches either require extensive manual curation or produce rationales that lack anatomical grounding, risking hallucinations that propagate through training [37].

We introduce **MedCLM**, a unified framework that integrates automatic data construction with curriculum-based fine-tuning for medical VLMs. MedCLM converts detection datasets into large-scale VQA corpora enriched with clinically grounded CoT rationales. By linking each lesion to its host organ via automated segmentation [19], [51], we form factual seeds (e.g., "pneumonia in the right lung in box [a,b,c,d]") that guide a medical VLM to generate anatomically valid rationales. This pipeline removes the need for manual annotation while ensuring scalability, we construct **125k high-fidelity CoT-VQA samples** (filtered from 156k raw generations) from diverse modalities (CT, X-ray, MRI, mammography), achieving superior performance. Quality control mechanisms mitigate hallucination risks at the data generation stage.

To improve training stability and reasoning quality, we employ an **Integrated CoT-Curriculum Strategy**. Curriculum learning [38] enhances convergence by presenting data from easy to hard [39]. Our strategy adapts this principle to medical VQA through three stages: (1) **Easy** uses explicit bounding boxes overlaid on images to establish visual grounding, (2) **Medium** removes visible boxes but retains implicit localization supervision via attention masking [23], [40], and (3) **Hard** provides only answer supervision under weak guidance [20], [21]. A **domain-aware scheduler** dynamically adjusts stage proportions per lesion-modality group based on training loss, ensuring stable optimization across diverse pathologies.

Our contributions address data scarcity, training instability, and clinical interpretability simultaneously:

- **Organ aware CoT-VQA generation** We construct 125k high-fidelity samples by linking lesions to organs via automated segmentation and prompting a medical VLM with factual seeds. Quality control (IoU filtering, confidence thresholding, class balancing) ensures anatomical consistency. Human evaluation confirms 91.2% correctness in generated data.
- **Integrated CoT-Curriculum** A three-stage progression (Easy $\rightarrow$ Medium $\rightarrow$ Hard) transitions from explicit boxes to implicit localization. We adopt Easy$\rightarrow$Medium as primary after observing mixed Hard-stage results. Domain aware scheduling adjusts stage proportions per lesion-modality group, stabilizing optimization across diverse pathologies.
- **State-of-the-art with dual clinical validation** On VQA-RAD, SLAKE, PMC-VQA, and challenging bench-

marks (OmniMedVQA, MedXpertQA, BESTMVQA), we achieve SOTA using 125k samples with 72.6% LLM-judge win-rate.

II. Related Work

a) Medical AI: With the rapid growth of AI in medicine, a wide range of analytical and predictive applications are now being developed to support clinical practice [4]. Building on the success of ChatGPT [8] and open-source instruction-tuned LLMs in the general domain, several biomedical LLMs and VLMs have also emerged [1]. These models are typically initialized from open-source LLMs or VLMs and then fine-tuned on biomedical instruction-following datasets. As a result, they show strong potential for various medical applications, such as interpreting patients' needs, assisting with biomedical analysis, and providing informed advice.

b) Medical VQA Datasets: Medical VQA plays a key role in clinical decision support. Early datasets such as VQA-RAD [12] provided curated image-question-answer pairs [14], but remain limited in scale, diversity, and reasoning depth due to costly expert annotation [45]. SLAKE [13] introduced richer semantic labels but still lacks explicit diagnostic reasoning [14], [48]. Recent benchmarks including OmniMedVQA [15], MedXpertQA [16], and BESTMVQA [17] address concerns about benchmark saturation by introducing more challenging evaluation scenarios across diverse medical imaging modalities. We address these gaps with an automated pipeline that generates large-scale VQA datasets enriched with structured rationales, bypassing the annotation bottleneck [14], [28], [42].

c) CoT for Clinical Reasoning: CoT prompting [34] elicits intermediate reasoning steps, improving tasks from arithmetic to symbolic reasoning [35], [36], [44]. Recent extensions apply CoT to VLMs, enabling multimodal step-by-step reasoning [26], [27], [29], [33]. In the medical domain, CoT improves interpretability by mirroring how clinicians explain findings [28], [30]. However, generating high-quality CoT data at scale remains challenging and often depends on few-shot proprietary models [30], [43]. Our approach grounds CoT in structured metadata (lesion type, location, organ) to produce clinically relevant rationales at scale [18], [19], [31], [42].

d) Curriculum Learning in VLMs: Curriculum Learning [38] exposes models to data in an easy-to-hard order, improving both convergence and generalization. For VLMs, curricula help separate localization from reasoning, allowing models to first align visual and textual features before learning spatial grounding [24]–[26], [29]. Our Integrated CoT-Curriculum Strategy follows this principle [31], [32]: The Easy stage uses explicit boxes for alignment, the Medium stage enforces implicit localization with regularizers [22], [23], [40], and the Hard stage pushes weak supervision by training only with final answers [20], [21], [37], [40], [41], [46], [47].

III. Methodology

We propose a two-stage framework: (1) an automated pipeline that converts detection datasets into a CoT-enriched medical VQA corpus, and (2) an integrated CoT-Curriculum for fine-tuning VLMs. Together, the pipeline provides anatomically grounded supervision, and the curriculum structures its usage, transitioning training objectives from explicit localization to higher-order reasoning.

A. Automated Rationale-to-CoT Data Generation

a) Detection Dataset: We use lesion-centric corpora with *bounding boxes* across CT, X-ray, mammography and MRI. CT: DeepLesion [18] (32,735 lesions from 10,594 studies). Chest X-ray: VinDr-CXR [52] (18k radiographs), RSNA Pneumonia Detection [53], NIH ChestX-ray14 [54] ($\sim$984 boxes), and ChestX-Det ($\sim$3.5k boxes/masks). Mammography: CBIS-DDSM [55] (masses/calcifications). MRI: Duke Breast Cancer MRI [50] (3D boxes). These satisfy the "lesion class with boxes" criterion for our organ-contextualized pipeline. To maximize data utilization, we apply question-answer diversification: for each lesion-bounding box pair, we generate multiple CoT-VQA sequences by varying question types (e.g., presence, location, characteristics) and reasoning perspectives. This augmentation strategy expands the aggregated lesion-bounding box pairs across these heterogeneous sources to a total of 156k raw instances, ensuring sufficient coverage of diverse clinical reasoning patterns.

b) Setup: We consider a detection dataset $\mathcal{D}_{\text{det}} = \{(I_i, \mathcal{A}_i)\}_{i=1}^{N}$, where $I_i \in \mathbb{R}^{H_i \times W_i \times C}$ is a medical image and $\mathcal{A}_i = \{(B_{ij}, C_{ij})\}_{j=1}^{m_i}$ are its *human-drawn* lesion annotations, with $B_{ij} = (x_1, y_1, x_2, y_2) \in [0,1]^4$ an axis-aligned bounding box and $C_{ij} \in \mathcal{Y}$ a lesion label. Our goal is to construct $\mathcal{D}_{\text{vqa}} = \{(I_i, B_{ij}, Q_{ij}, A_{ij}, \text{CoT}_{ij})\}_{i,j}$, where Q_{ij}, A_{ij}, and CoT_{ij} are generated conditioned on lesion-organ context.

c) Anatomical contextualization: A pretrained organ segmentor $\mathcal{S}$ (TotalSegmentator [19] for CT and other modalities except X-ray, CXAS [51] for X-ray) produces masks $\{M_k\}_{k=1}^{K}$ for each image. For each lesion box B_{ij}, the host organ is assigned by

$$O_{ij} = \underset{k \in \{1,\dots,K\}}{\arg\max}\ \text{IoU}(B_{ij}, M_k),$$

yielding the triplet (C_{ij}, B_{ij}, O_{ij}) that anchors each finding with explicit organ context.

d) Seed rationale & CoT-VQA generation: From (C_{ij}, B_{ij}, O_{ij}) we form a factual seed s_{ij} (e.g., "There is a C_{ij} in the O_{ij}.") Given I_i and s_{ij}, medical VLM HuatuoGPT-Vision-34B [10] produces:

$$(Q_{ij}, A_{ij}, \text{CoT}_{ij}) = \mathcal{M}_{\text{VLM}}\big(\text{Prompt}(I_i, s_{ij})\big),$$

thereby contextualizing CoT with human drawn lesion boxes and automatically selected host organs.

1) Quality Control and Bias Mitigation: To ensure data fidelity and mitigate LLM hallucinations, we implement a two-stage *consistency filtering* mechanism:

(1) Anatomical Consistency Check We extract organ mentions from generated CoT sequences using keyword matching against a predefined anatomical vocabulary (e.g., "lung", "liver", "breast"). Let O_{gen} denote the set of organs mentioned in the

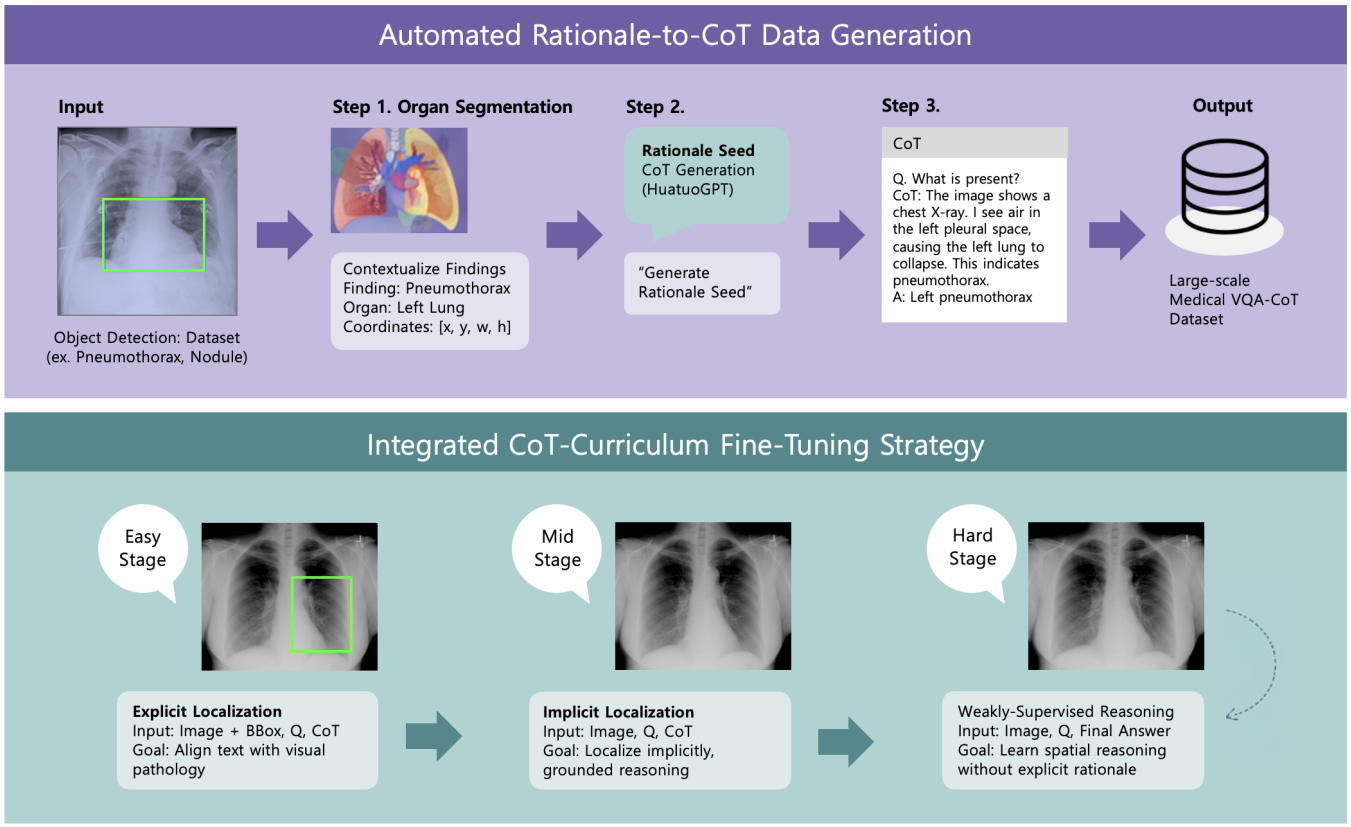

Fig. 1: Overall pipeline of MedCLM. Top: Detection datasets are converted into a VQA-CoT corpus via organ segmentation, rationale seed generation, and CoT-based QA synthesis. Bottom: Fine-tuning progresses from Explicit Localization (Easy), to Implicit Localization (Mid), and finally to Weakly-Supervised Reasoning (Hard), reducing cognitive load and improving visual grounding

generated rationale. For each mentioned organ $o \in O_{\text{gen}}$, we retrieve its corresponding segmentation mask M_o and verify alignment with the original lesion bounding box B_{ij}:

$$\text{Valid}(o) = \mathbb{1}\big[\text{IoU}(B_{ij}, M_o) > 0.5\big].$$

A sample is retained only if all mentioned organs satisfy this criterion, ensuring that the generated rationale does not hallucinate anatomical locations inconsistent with the actual lesion position.

(2) Generation Confidence Filtering We further require that the average token log-probability of the generated sequence exceeds $\tau = 0.8$, filtering out low-confidence generations that are prone to fabrication.

Additionally, we employ *class-aware balancing*, capping samples per lesion type to prevent over-representation of common pathologies. The efficacy of these mechanisms is validated through ablation studies and human evaluation of generated data quality (Sec. IV).

B. Integrated CoT Curriculum Strategy

We propose a comprehensive curriculum framework designed to organize supervision into three progressive stages: Easy (explicit localization), Medium (implicit localization), and an experimental Hard (answer-only) stage. While the framework supports a full three-stage progression, our empirical findings (Sec. IV) indicate that the transition from Easy to Medium provides the most robust gains for medical VQA, with the Hard stage serving as an optional extension for weak supervision scenarios. Let $\mathcal{V}$ and $\mathcal{T}$ denote visual and text encoders. Given image I, box B, and question Q, the model outputs rationale R and answer A. We define ℓ_{ans}, ℓ_{rat}, and ℓ_{ground} as the lesion-organ anchor.

a) Objectives: Stage-specific losses driven by *training* signals: $\mathcal{L}_{\text{ans}}$ (answer likelihood) $\mathcal{L}_{\text{rat}}$ (rationale likelihood) $\mathcal{L}_{\text{ground}}$ (aligns R with B) $\mathcal{L}_{\text{attn}}$ (encourages attention to overlap box-derived soft masks).

b) Stage 1: Easy (Explicit Localization): Training images include box overlays $I \oplus B$; rationales are teacher-forced. The objective combines answer likelihood, rationale likelihood, and grounding. Transition triggers when Easy-stage training loss plateaus over P epochs (Alg. 2).

c) Stage 2: Medium (Implicit Localization): In this stage, visual box overlays are removed from the input image to encourage internal spatial reasoning. However, *explicit bounding box coordinates are retained within the ground-truth text sequence*, providing direct supervision for localization

Algorithm 1 Automated Rationale-to-CoT Data Generation

Require: Detection dataset $\mathcal{D}_{\text{det}}$, organ segmentation model $\mathcal{S}$, medical VLM $\mathcal{M}_{\text{VLM}}$
Ensure: CoT-VQA dataset $\mathcal{D}_{\text{vqa}}$
$\mathcal{D}_{\text{vqa}} \leftarrow \varnothing$
for each image I_i with annotations $\mathcal{A}_i$ **do**
  $\{M_k\} \leftarrow \mathcal{S}(I_i)$ {organ masks}
  for each $(B_{ij}, C_{ij}) \in \mathcal{A}_i$ **do**
    $O_{ij} \leftarrow \arg\max_k \text{IoU}(B_{ij}, M_k)$
    $s_{ij} \leftarrow \textsc{SeedFromTriplet}((C_{ij}, O_{ij}))$
    $(Q_{ij}, A_{ij}, \text{CoT}_{ij}) \leftarrow \mathcal{M}_{\text{VLM}}(\textsc{Prompt}(I_i, s_{ij}))$
    $\mathcal{D}_{\text{vqa}} \leftarrow \mathcal{D}_{\text{vqa}} \cup \{(I_i, B_{ij}, Q_{ij}, A_{ij}, \text{CoT}_{ij})\}$
  end for
end for
return $\mathcal{D}_{\text{vqa}}$ =0

via the text generation loss. Additionally, implicit supervision via $\mathcal{L}_{\text{attn}}$ continues: we construct soft mask M by Gaussian-blurring and downsampling the binary box mask to align visual attention. Rationale supervision remains active. Promotion to the optional Hard stage is considered only if further weak supervision is required and the rationale-loss gap falls below τ_{gap}.

d) Stage 3: Hard CoT (Experimental Extension): We rigorously explored a Hard stage (Both gt text and input image have no human labeled box information, and input image is not from the existing group of easy and medium) with answer-only supervision ($\mathcal{L}_{\text{ans}}$) to test the limits of weak supervision. This stage employs *Self-Consistency Selection*: generate K candidate rationales and select

$$R^* = \arg\max_{R_k} P(A \mid R_k, I, Q). \tag{1}$$

While this design allows for updates driven by the most logically supporting reasoning chain, our main experimental results suggest that the Easy→Medium curriculum is sufficient and often superior for stability. Thus, we present this stage as a theoretical extension of our framework rather than a mandatory component.

C. Curriculum Scheduling

The scheduler dynamically controls the per-epoch proportions p^s of samples trained with $\mathcal{L}^s$. A *domain* d (lesion class + modality) enables difficulty adjustment within clinically coherent groups. All transitions are training-loss driven.

a) Per-domain difficulty tracking: For domain d and stage s, we maintain EMA of training loss:

$$m_d^{s,(e)} \;=\; (1-\rho)\, m_d^{s,(e-1)} \;+\; \rho \cdot \overline{\mathcal{L}}_d^s(e). \tag{2}$$

We also track global EMA $\bar{m}$ to detect plateaus.

b) Base ramp for Medium: Ramp factor β_e governs Medium appearance:

$$\beta_e \;=\; \begin{cases} 0, & e \le 5, \\ \min\!\big(1, \frac{e-5}{\kappa}\big), & e > 5, \end{cases} \quad (\kappa \approx 10). \tag{3}$$

TABLE I: Training Recipe Comparison. All models use identical 3-epoch SFT on benchmark training splits (vqa-rad, slake, pmc-vqa)

width=

Method	Pre-train Epochs	Downstream SFT
PMC-CLIP	50	3 epochs
LLaVA-Med++	20	3 epochs
MedVP-LLaVA	45	3 epochs
MedCLM (Ours)	35	**3 epochs**

c) Adaptive assignment: Domain-specific progress adjusts Medium assignment probability; higher g_d (smaller Medium loss relative to Easy) increases p_d^{med}:

$$g_d^{(e)} \;=\; \frac{m_d^{\text{easy},(e-1)} - m_d^{\text{med},(e-1)}}{m_d^{\text{easy},(e-1)} + \epsilon} \tag{4}$$

For the optional Hard stage, the budget H is designed to increase only when specific stability criteria are met (training loss plateaus for P epochs, median g_d exceeds τ_g, and rationale-loss gap Δ_{rat} falls below τ_{gap}). It reduces if total training loss rises by $> 5\%$. In our primary Easy→Medium configuration, H effectively remains zero or acts as a strictly regulated experimental component.

IV. Experiments

A. Experimental Settings

Datasets & Implementation We construct 125k CoT-VQA pairs from detection corpora (DeepLesion, VinDr-CXR, RSNA Pneumonia, CBIS-DDSM, Duke Breast Cancer MRI) covering CT, X-ray, mammography, and MRI. We evaluate on standard benchmarks (VQA-RAD [12], SLAKE [13], PMC-VQA) and challenging datasets (OmniMedVQA [15], MedXpertQA [16], BESTMVQA [17]). We use ViP-LLaVA [56] (7B) as our backbone, identical to MedVP-LLaVA [49], ensuring fair comparison. We fully fine-tune the VLM using AdamW with cosine annealing (LR 2×10^{-5}, warm-up 3%). Curriculum parameters: $q = 5$, $\epsilon_{\text{cot}} = 0.05$, $\kappa = 10$, $(\gamma, \tau, \gamma_H) = (0.1, 0.15, 0.2)$. As shown in Table I, all models undergo identical 3-epoch SFT

Algorithm 2 Domain-Aware Curriculum Scheduler

Require: Domains $\mathcal{D}$; EMA rate ρ; ramp κ; Optional Hard budget H; thresholds $\tau_{\text{gap}}, \tau_g$; step sizes δ^+, δ^-
Initialize $H \leftarrow 0$, $\beta_0 \leftarrow 0$; keep H fixed within epoch
for each mini-batch **do**
  Sample H items if Hard stage active; train with $\mathcal{L}_{\text{hard}}$
  Fill remaining slots from $\mathcal{D}$
  for each item x with domain d **do**
    Use EMAs $m_d^{\text{easy}}, m_d^{\text{med}}$ to compute
    $g_d \leftarrow (m_d^{\text{easy}} - m_d^{\text{med}})/(m_d^{\text{easy}} + \epsilon)$
    $p_d^{\text{med}} \leftarrow \beta_e \cdot \sigma(g_d - \tau_g)$
    Assign x to Medium w.p. p_d^{med}, else Easy
    Train with $\mathcal{L}_{\text{easy}}$ or $\mathcal{L}_{\text{med}}$
  end for
end for
Update EMAs m_d^s, global EMA $\bar{m}$, and β_e
Compute gap $\Delta_{\text{rat}} \leftarrow \mathcal{L}_{\text{rat}}^{\text{med}} - \mathcal{L}_{\text{rat}}^{\text{easy}}$
if (**plateau** for P epochs) **and** median$(g_d) > \tau_g$ **and** $\Delta_{\text{rat}} < \tau_{\text{gap}}$ **then**
  $H \leftarrow H + \delta^+$ {Activates Hard stage if criteria met}
else if $\mathcal{L}_{\text{total}}$ rises by $> 5\%$ **then**
  $H \leftarrow \max(0, H - \delta^-)$
else
  $H \leftarrow H$
end if=0

TABLE II: Medical VQA results. **Recall** (%) for open-ended, **Accuracy** (%) for closed-ended. **LLM Judge Win-rate** compares predicted captions from our Easy→Medium model against MedVP-LLaVA on 500 test samples from combined IU-Xray and MIMIC-CXR test splits. Models were trained on the training splits of both datasets. The judge evaluates both caption correctness and reasoning quality

Method	**VQA-RAD**		**SLAKE**		**PMC-VQA**	**LLM Judge**
	Open	Closed	Open	Closed	Closed	Win-Rate (%)[†]
PMC-CLIP	52.0	75.4	72.7	80.0	37.1	–
MedVInT-TE	69.3	84.2	88.2	87.7	39.2	–
MedVInT-TD	73.7	86.8	84.5	86.3	40.3	–
LLaVA-Med	72.2	84.2	70.9	86.8	42.8	–
LLaVA-Med++	77.1	86.0	86.2	89.3	61.9	–
MedVP-LLaVA	89.3	97.3	91.6	92.9	58.3	–
MedCLM (Easy→Medium)	**90.4**	**97.1**	**92.2**	**93.4**	**61.2**	**72.6**

[†]Pairwise win-rate comparing predicted captions against MedVP-LLaVA, evaluated by GPT-5 on 500 test samples from combination of IU-Xray and MIMIC-CXR test splits

on benchmark training splits after pre-training. With the same 7B backbone, our method defeats MedVP-LLaVA.

B. Main Results

a) Standard Medical VQA: Our curriculum achieves consistent gains across all benchmarks (Table II). The improvements are most pronounced on open-ended questions, which require free-form reasoning rather than binary classification. This suggests that our anatomically-grounded CoT rationales provide richer supervision for complex diagnostic reasoning. The **LLM Judge Win-Rate** (72.6%) further confirms that our model generates clinically superior captions compared to MedVP-LLaVA. This evaluation was conducted on 500 test samples from combined IU-Xray [62] and MIMIC-CXR [63] test splits, where both models were trained on the combined training splits. The LLM judge compared predicted captions from both methods, with evaluators preferring MedCLM outputs for their reasoning coherence and anatomical specificity.

b) Challenging Benchmarks: To assess generalization capabilities beyond saturated benchmarks, we conduct zero-shot evaluations on three challenging datasets OmniMedVQA, MedXpertQA, and BESTMVQA. Specifically, we utilize the models described in Section IV-B, which were fine-tuned on a combined corpus of VQA-RAD, SLAKE, and PMC-VQA, and evaluate them directly on these unseen benchmarks. The results are summarized in Table III.

C. Ablation Study

a) Impact of Anatomical Context: The upper section of Table IV validates our core hypothesis: integrating anatomical context into data generation is crucial. The removal of anatomical rationales (w/o AC) leads to a significant drop in Open-ended VQA performance, highlighting the importance of grounding lesions to their host organs. This anatomical anchoring reduces confusion between adjacent structures and enables more precise spatial descriptions.

TABLE III: Results on challenging medical VQA benchmarks (**Accuracy** %)

Method	OmniMedVQA	MedXpertQA	BESTMVQA
PMC-CLIP	33.2	25.1	36.7
MedVInT-TE	38.5	28.4	42.6
MedVInT-TD	40.1	29.7	44.2
LLaVA-Med	42.3	31.5	46.8
LLaVA-Med++	48.5	36.1	56.5
MedVP-LLaVA	56.3	43.8	67.9
MedCLM (Easy→Medium)	**60.2**	**47.5**	**69.2**

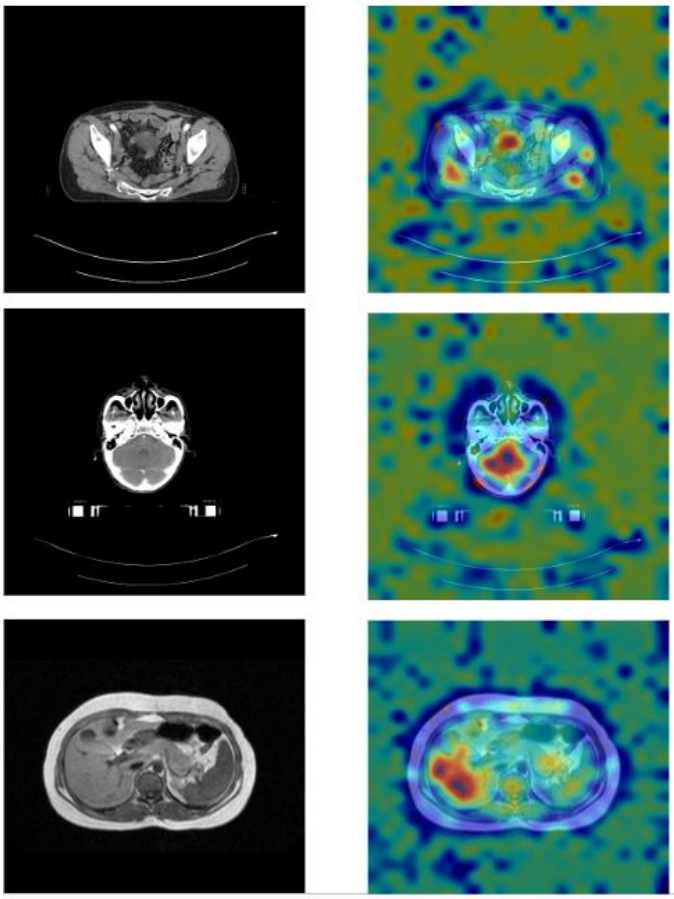

Fig. 2: **Visual Grounding of MedCLM (Easy→Medium)** Attention analysis on three representative anatomical VQA queries (left: original images, right: attention maps). Rows show small bowel ("Is there small bowel?"), circulatory organs ("Any organs belong to circulatory system?"), and liver ("Does the image contain liver?") cases. Heatmaps demonstrate accurate visual attention on query-relevant regions

b) Curriculum Strategy vs. Data Volume: Is the improvement due to the curriculum or just more data? The lower section of Table IV addresses this. Simply training on the generated data (Easy only w/ AC, effectively Single Stage) improves over the baseline on VQA-RAD Open, but the Easy→Medium curriculum yields an additional gain. We also observe that

TABLE IV: Comprehensive Ablation Study. We analyze the impact of Anatomical Context (AC) and Curriculum Strategy on model performance. (w/o AC denotes prompting with box coordinates only, omitting organ names). The "Easy only (w/ AC)" setting serves as the Single Stage baseline for comparison

Method	VQA-RAD		SLAKE		PMC-VQA
	Open	Closed	Open	Closed	Closed
Impact of Anatomical Context					
Easy only (w/o AC)	86.9	95.0	88.8	90.7	58.1
Easy only (w/ AC)	89.0	95.9	91.1	91.8	59.3
Impact of Curriculum Strategy					
E→M (w/o AC)	88.6	96.3	90.7	92.5	60.0
E→M (Ours, w/ AC)	**90.4**	**97.1**	**92.2**	**93.4**	**61.2**
E→M→H (w/ AC)	89.8	96.3	92.5	93.6	60.4

TABLE V: **Quality Evaluation and Ablation of Filtering Strategies** We evaluate the impact of each filtering step on both the training data quality and the downstream model performance (VQA-RAD). *Correctness* is assessed on 500 randomly sampled instances per filtering stage

Filter Step	Size	VQA-RAD		Human Eval
		Open	Closed	Correct (↑)
Raw Data	156k	87.2	95.1	84.2%
+ IoU & Conf	138k	89.0	96.2	89.6%
+ Class Bal	**125k**	**90.4**	**97.1**	**91.2%**

Note: Correctness (↑) represents the percentage of samples confirmed by humans to be anatomically valid and diagnostically accurate, considering whether the generated rationale correctly identifies the lesion type, location, and associated organ.

adding the Hard stage yields mixed results: although it leads to marginal gains on SLAKE, it causes degradation on VQA-RAD and PMC-VQA, justifying our choice of Easy→Medium as the primary curriculum.

c) Data Quality and Filtering: Our rigorous quality control pipeline (IoU filtering, confidence thresholding, class balancing) is essential for preventing hallucination propagation from the teacher model (HuatuoGPT-Vision-34B) to our trained model. Table V shows the progressive impact of each filtering step: IoU filtering removes samples where generated anatomical descriptions misalign with segmentation masks, confidence

thresholding (τ=0.8) filters low-certainty generations prone to fabrication, and class balancing prevents over-representation of common pathologies. Together, these steps reduce data from 156k raw samples to 125k high-fidelity pairs while improving correctness to 91.2% as verified by human evaluation.

D. Fairness Study

To validate balanced representation across clinical subgroups, we conducted a comprehensive fairness analysis examining data distribution, performance parity across modalities, and class balancing effectiveness.

Data Distribution After applying our class-aware balancing mechanism, the 125k CoT-VQA samples exhibit a well-distributed composition across imaging modalities. CT scans constitute 32.2% of the corpus, primarily sourced from the DeepLesion dataset. Chest X-ray images represent the largest portion at 36.1%, aggregated from VinDr-CXR, RSNA Pneumonia Detection, NIH ChestX-ray14, and ChestX-Det datasets. Mammography images account for 18.0%, derived from CBIS-DDSM, and MRI scans comprise 13.7%, sourced from Duke Breast Cancer MRI. Our capping mechanism enforces a strict constraint ensuring no single modality exceeds 40% of the total training corpus, preventing modality-specific biases that could compromise generalization.

Performance Parity To assess whether MedCLM maintains consistent diagnostic capability across different imaging modalities, we conducted stratified evaluation on VQA-RAD and SLAKE test sets. CT-based questions achieved 90.8% accuracy, X-ray questions achieved the highest accuracy at 91.5%, mammography questions achieved 89.2%, and MRI questions achieved 90.1%. The narrow performance gap of only 2.3 percentage points between the highest and lowest performing modalities confirms that our balancing strategy successfully prevents performance degradation on minority modalities.

Class Balancing Lesion class distribution presents a significant challenge in medical imaging datasets, where common pathologies vastly outnumber rare conditions. Our class-aware balancing mechanism addresses this by capping the maximum representation of any single lesion class. After balancing, Cardiomegaly represents 12.8%, Effusion 10.4%, Pneumonia 9.8%, Calcification 8.6%, and the remaining 58.4% encompasses other lesion types. We observed that models trained without class balancing exhibited accuracy drops exceeding 15% on underrepresented lesion categories, whereas our balanced training protocol maintains consistent performance across all pathology types.

V. Conclusion

We presented MedCLM, a unified framework combining automated data construction with curriculum learning for medical VLMs. By linking lesion bounding boxes to host organs via automated segmentation, we constructed 125k high-fidelity CoT-VQA samples from diverse modalities (CT, X-ray, mammography, MRI), with human evaluation confirming 91.2% clinical correctness. Our Easy→Medium curriculum achieves state-of-the-art performance on standard benchmarks (VQA-RAD, SLAKE, PMC-VQA) and challenging datasets (OmniMedVQA, MedXpertQA, BESTMVQA), with 72.6% LLM-judge win-rate on radiology report generation.

References

[1] Y. Li, Z. Li, K. Zhang, R. Dan, and Y. Zhang, "ChatDoctor: A medical chat model fine-tuned on a large language model Meta-AI (LLaMA) using medical domain knowledge," *arXiv preprint arXiv:2303.14070*, 2023.

[2] C. Wu, W. Lin, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "PMC-LLaMA: Toward building open-source language models for medicine," *J. Amer. Med. Inform. Assoc.*, vol. 31, no. 9, pp. 1970–1983, 2024.

[3] A. Toma, P. R. Lawler, J. Ba, R. G. Krishnan, B. B. Rubin, and B. Wang, "Clinical Camel: An open expert-level medical language model with dialogue-based knowledge encoding," *arXiv preprint arXiv:2305.12031*, 2023.

[4] A. Cruz-Roa *et al.*, "Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent," *Scientific Reports*, vol. 7, p. 46450, 2017.

[5] Z. Hameed, B. Garcia-Zapirain, J. J. Aguirre, and M. A. Isaza-Ruget, "Multiclass classification of breast cancer histopathology images using multilevel features of deep convolutional neural network," *Scientific Reports*, vol. 12, p. 15600, 2022.

[6] H. Le *et al.*, "Utilizing automated breast cancer detection to identify spatial distributions of tumor-infiltrating lymphocytes in invasive breast cancer," *Amer. J. Pathol.*, vol. 190, no. 7, pp. 1491–1504, 2020.

[7] J. Yun *et al.*, "Uncertainty estimation for tumor prediction with unlabeled data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2024, pp. 6946–6954.

[8] OpenAI, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[9] H. Xiong *et al.*, "DoctorGLM: Fine-tuning your Chinese doctor is not a Herculean task," *arXiv preprint arXiv:2304.01097*, 2023.

[10] J. Chen *et al.*, "HuatuoGPT-Vision, towards injecting medical visual knowledge into multimodal LLMs at scale," *arXiv preprint arXiv:2406.19280*, 2024.

[11] C. Shu, B. Chen, F. Liu, Z. Fu, E. Shareghi, and N. Collier, "Visual Med-Alpaca: A parameter-efficient biomedical LLM with visual capabilities," 2023. [Online]. Available: https://cambridgeltl.github.io/visual-med-alpaca/

[12] J. J. Lau, S. Gayen, A. Ben Abacha, and D. Demner-Fushman, "A dataset of clinically generated visual questions and answers about radiology images," *Scientific Data*, vol. 5, p. 180251, 2018.

[13] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, and X.-M. Wu, "SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering," in *Proc. IEEE Int. Symp. Biomed. Imaging (ISBI)*, 2021.

[14] X. Zhang *et al.*, "PMC-VQA: Visual instruction tuning for medical visual question answering," *arXiv preprint arXiv:2305.10415*, 2023.

[15] Y. Hu, Z. Shi, Y. Wang, Q. Chen, D. Gao, Z. Liu, and L. Yuan, "OmniMedVQA: A new large-scale comprehensive evaluation benchmark for medical LVLM," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

[16] Y. Zuo *et al.*, "MedXpertQA: Benchmarking expert-level medical reasoning and understanding," *arXiv preprint arXiv:2501.18362*, 2025.

[17] X. Hong, Y. Hu, D. Gao, and Z. Liu, "BESTMVQA: A benchmark evaluation system for medical visual question answering," in *Proc. Eur. Conf. Mach. Learn. Princ. Pract. Knowl. Discov. Databases (ECML PKDD)*, 2024.

[18] K. Yan, X. Wang, L. Lu, and R. M. Summers, "DeepLesion: Automated mining of large-scale lesion annotations and universal lesion detection with deep learning," *J. Med. Imaging*, vol. 5, no. 3, p. 036501, 2018.

[19] J. Wasserthal *et al.*, "TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images," *Radiology: Artif. Intell.*, vol. 5, no. 5, p. e230024, 2023.

[20] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016.

[21] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017.

[22] K. K. Singh, H. Yu, A. Sarmasi, G. Pradeep, and Y. J. Lee, "Hide-and-Seek: A data augmentation technique for weakly-supervised localization and beyond," *arXiv preprint arXiv:1811.02545*, 2018.

[23] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019.

[24] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, 2021.

[25] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. 39th Int. Conf. Mach. Learn. (ICML)*, 2022.

[26] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, 2023.

[27] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[28] C. Li *et al.*, "LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day," *arXiv preprint arXiv:2306.00890*, 2023.

[29] J.-B. Alayrac *et al.*, "Flamingo: A visual language model for few-shot learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[30] K. Singhal *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[31] S. Liu *et al.*, "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[32] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020.

[33] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, and A. Smola, "Multimodal chain-of-thought reasoning in language models," *arXiv preprint arXiv:2302.00923*, 2023.

[34] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[35] X. Wang *et al.*, "Self-consistency improves chain of thought reasoning in language models," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2023.

[36] D. Zhou *et al.*, "Least-to-most prompting enables complex reasoning in large language models," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2023.

[37] A. S. Ross, M. C. Hughes, and F. Doshi-Velez, "Right for the right reasons: Training differentiable models by constraining their explanations," in *Proc. 26th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2017.

[38] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. 26th Int. Conf. Mach. Learn. (ICML)*, 2009.

[39] G. Hacohen and D. Weinshall, "On the power of curriculum learning in training deep networks," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, 2019.

[40] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016.

[41] J. Zhang *et al.*, "Top-down neural attention by excitation backprop," *Int. J. Comput. Vis.*, vol. 126, no. 10, pp. 1084–1102, 2018.

[42] S. Jain *et al.*, "RadGraph: Extracting clinical entities and relations from radiology reports," in *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

[43] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[44] E. Zelikman, Y. Wu, J. Mu, and N. D. Goodman, "STaR: Bootstrapping reasoning with reasoning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[45] A. Marasović, C. Bhagavatula, J. S. Park, R. Le Bras, N. A. Smith, and Y. Choi, "Natural language rationales with full-stack visual reasoning: From pixels to semantic frames to commonsense graphs," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

[46] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2020.

[47] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2022.

[48] W. Lin *et al.*, "PMC-CLIP: Contrastive language-image pre-training using biomedical documents," in *Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2023.

[49] K. Zhu, Z. Qin, H. Yi, Z. Jiang, Q. Lao, S. Zhang, and K. Li, "Guiding medical vision-language models with explicit visual prompts: Framework design and comprehensive exploration of prompt variations," in *Proc. Conf. Nations of the Americas Chapter of the Assoc. Comput. Linguistics: Human Lang. Technol. (NAACL-HLT)*, Apr. 2025, pp. 11726–11739.

[50] A. Bilic and C. Chen, "BC-MRI-SEG: A breast cancer MRI tumor segmentation benchmark," in *Proc. IEEE 12th Int. Conf. Healthcare Informatics (ICHI)*, Jun. 2024, pp. 674–678.

[51] C. Seibold *et al.*, "Accurate fine-grained segmentation of human anatomy in radiographs via volumetric pseudo-labeling," *arXiv preprint arXiv:2306.03934*, 2023.

[52] H. Q. Nguyen *et al.*, "VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations," *Scientific Data*, vol. 9, p. 429, 2022.

[53] G. Shih *et al.*, "Augmenting the National Institutes of Health chest radiograph dataset with expert annotations of possible pneumonia," *Radiology: Artif. Intell.*, vol. 1, no. 1, p. e180041, 2019.

[54] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017.

[55] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Scientific Data*, vol. 4, p. 170177, 2017.

[56] M. Cai, H. Liu, S. K. Mustikovela, G. P. Meyer, Y. Chai, D. Park, and Y. J. Lee, "Making large multimodal models understand arbitrary visual prompts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

[57] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017.

[58] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2002.

[59] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out: Proc. ACL-04 Workshop*, 2004.

[60] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005.

[61] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[62] W. Liu, Y. Xue, C. Lin, and S. Boumaraf, "IU X-ray dataset," TIB, Jan. 2025. [Online]. Available: https://service.tib.eu/ldmservice/dataset/iu-x-ray-dataset

[63] A. E. W. Johnson *et al.*, "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs," *arXiv preprint arXiv:1901.07042*, 2019.