MedCLM: Learning to Localize and Reason via a CoT-Curriculum in Medical Vision-Language Models

Anonymous ACL submission

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 (CoT) reasoning by linking lesion boxes to organ segmentation and structured rationales. These contextual signals enable medical vision-language models to generate question-answer pairs with step-by-step reasoning. To utilize this data effectively, we propose an Integrated CoT-Curriculum Strategy composed of an Easy stage with explicit lesion boxes for visual grounding, a Medium stage that encourages implicit localization, and a Hard stage for weakly supervised reasoning. Experimental results demonstrate that MedCLM attains state-of-the-art performance on several medical VQA benchmarks, providing a scalable framework for developing clinically aligned medical vision-language models. The GitHub repository will be released upon paper acceptance at: https://github.com/anonymous/medclm

1 Introduction

Medical Vision Language Models (VLMs) are essential for clinical decision support. They enable systems that answer queries directly from medical images. Medical Visual Question Answering (VQA) is a central task in this field (Lau et al., 2018; He et al., 2020; Zhang et al., 2023b). Early datasets such as VQA RAD (Lau et al., 2018) and PathVQA (He et al., 2020) established the foundation but remain limited in scale and reasoning depth due to costly expert annotation. SLAKE (Liu et al., 2021) and PMC VQA (Zhang et al., 2023b) expanded coverage yet most benchmarks still focus on short question answering without explicit diagnostic reasoning. This limits interpretability and clinical trust.

Chain of Thought (CoT) prompting (Wei et al., 2022) improves reasoning in Large Language Models (LLMs) by producing intermediate steps (Wang et al., 2023). It has been effective across multimodal domains (Liu et al., 2023a; Li et al., 2023b) and is particularly relevant to medicine where reasoning aligns with clinical workflows (Singhal et al., 2023). However constructing large scale CoT data remains costly due to dependence on proprietary models and few shot generation.

We introduce **MedCLM**, a unified framework that integrates automatic data construction and curriculum based fine tuning for medical VLMs. MedCLM converts detection datasets into large scale VQA corpora enriched with clinically grounded CoT rationales. Structured metadata such as lesion type, location and organ provides factual seeds that guide VLMs to generate valid rationales (Yan et al., 2018; Jain et al., 2021). This removes the need for manual annotation and ensures scalability.

To improve stability during training we employ an Integrated CoT Curriculum Strategy. Curriculum learning (CL) (Bengio et al., 2009) enhances convergence by presenting data from easy to hard. Our strategy follows this principle. The Easy stage uses explicit boxes for grounding. The Medium stage applies implicit localization with regularizers (Bilen and Vedaldi, 2016; Yun et al., 2019). The Hard stage trains only on final answers under weak supervision (Zhou et al., 2016; Selvaraju et al., 2017). This gradual supervision reduces cognitive load and promotes spatial reasoning without direct annotation.

Contributions We summarize our work in three main components—data construction, training strategy, and empirical validation—and instantiate them as follows:

- **Organ-aware VQA-CoT generation.** From detection datasets, we build a large VQA-CoT

corpus by linking each lesion to its host organ, forming factual seeds, and prompting a medical VLM—no manual annotation.

- **Integrated CoT-Curriculum with scheduling.** A three-stage recipe (Easy → Medium → Hard) separates grounding from reasoning; a domain-aware scheduler and implicit-localization regularizers stabilize training under weak supervision.

- **Effectiveness & interpretability.** The approach improves standard medical VQA benchmarks and radiology report generation, while producing concise, anatomically grounded rationales without extra labels.

2 Related Work

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 (Cruz-Roa et al., 2017; Le et al., 2020; Hameed et al., 2022; Yun et al., 2024). Building on the success of ChatGPT (OpenAI, 2023) and open-source instruction-tuned LLMs in the general domain, several biomedical LLMs and VLMs have also emerged (Li et al., 2023c; Shu et al., 2023; Wu et al., 2023; Toma et al., 2023; Xiong et al., 2023; Chen et al., 2024; Li et al., 2023a; Zhu et al., 2025). 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.

Medical VQA Datasets. Medical VQA plays a key role in clinical decision support. Early datasets such as VQA-RAD (Lau et al., 2018) and PathVQA (He et al., 2020) provided curated image-question-answer pairs, (Zhang et al., 2023b,a) but remain limited in scale, diversity, and reasoning depth due to costly expert annotation (Marasović et al., 2020). SLAKE (Liu et al., 2021) introduced richer semantic labels but still lacks explicit diagnostic reasoning (Zhang et al., 2023b; Lin et al., 2023). We address these gaps with an automated pipeline that generates large-scale VQA datasets enriched with structured rationales, bypassing the annotation bottleneck.(Jain et al., 2021; Zhang et al., 2023a; Li et al., 2023a)

CoT for Clinical Reasoning. CoT prompting (Wei et al., 2022) elicits intermediate reasoning steps, improving tasks from arithmetic to symbolic reasoning (Wang et al., 2023; Zhou et al., 2023; Zelikman et al., 2023). Recent extensions apply CoT to VLMs, enabling multimodal step-by-step reasoning (Zhang et al., 2024; Liu et al., 2023a; Li et al., 2023b; Alayrac et al., 2022). In the medical domain, CoT improves interpretability by mirroring how clinicians explain findings (Singhal et al., 2023; Li et al., 2023a). However, generating high-quality CoT data at scale remains challenging and often depends on few-shot proprietary models (Singhal et al., 2023; Ouyang et al., 2022). Our approach grounds CoT in structured metadata (lesion type, location, organ) to produce clinically relevant rationales at scale (Yan et al., 2018; Wasserthal et al., 2023; Jain et al., 2021; Liu et al., 2023b).

Curriculum Learning in VLMs. Curriculum Learning (CL) (Bengio et al., 2009) exposes models to data in an easy-to-hard order, improving both convergence and generalization (Hacohen and Weinshall, 2019). For VLMs, curricula help separate localization from reasoning, allowing models to first align visual and textual features before learning spatial grounding (Radford et al., 2021; Li et al., 2022, 2023b; Alayrac et al., 2022). Our Integrated CoT-Curriculum Strategy follows this principle (Liu et al., 2023b; Carion et al., 2020): the Easy stage uses explicit boxes for alignment, the Medium stage enforces implicit localization with regularizers (Bilen and Vedaldi, 2016; Singh and Lee, 2018; Yun et al., 2019), and the Hard stage pushes weak supervision by training only with final answers (Bilen and Vedaldi, 2016; Zhou et al., 2016; Selvaraju et al., 2017; Gu et al., 2022; Abnar and Zuidema, 2020; Zhang et al., 2018; Ross et al., 2017).

3 Methodology: MedCLM

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 the training objectives from explicit localization to higher-order reasoning.

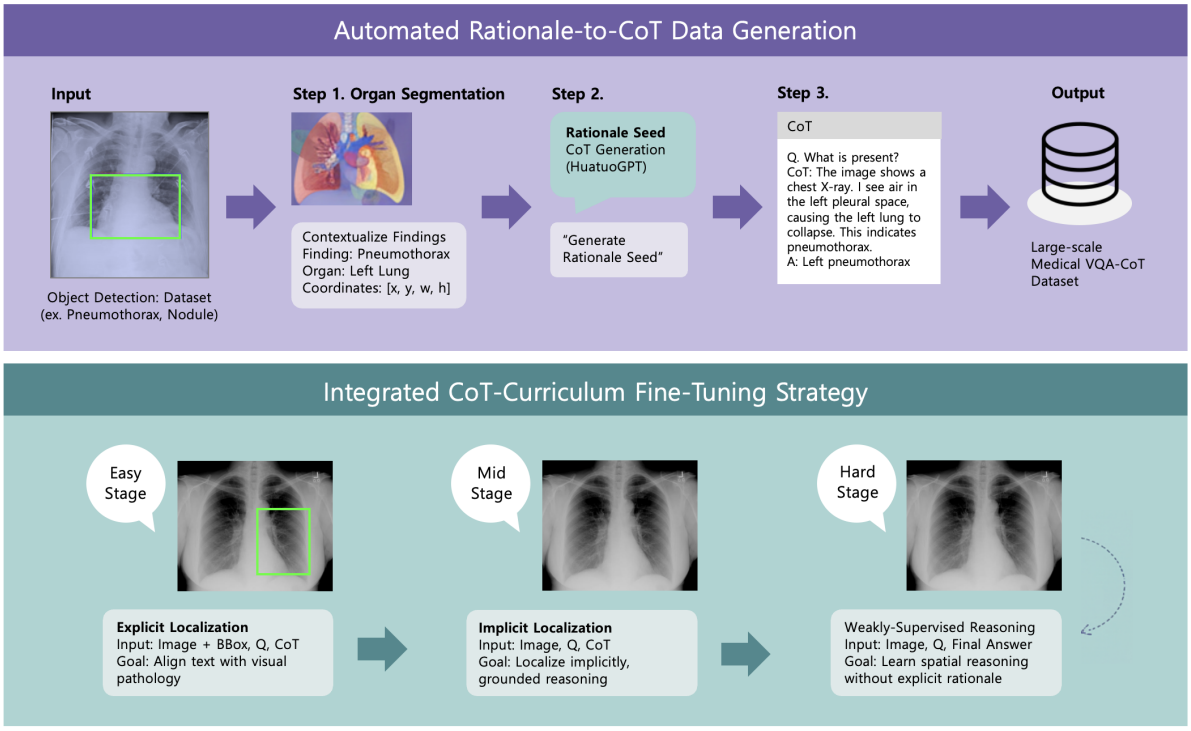


Figure 1: Overall pipeline of MedCLM. Top: Detection datasets are converted into a VQA-CoT corpus via organ segmentation, rationale seed generation, and CoT-based QA synthesis. Bottom: Fine-tuning progresses from Explicit Localization (Easy), to Implicit Localization (Mid), and finally to Weakly-Supervised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

3.1 Automated Rationale-to-CoT Data Generation

Detection Dataset. We use lesion-centric corpora with *bounding boxes* across CT, X-ray, mammography and MRI. CT: DeepLesion (Yan et al., 2018) (2D boxes; 32,735 lesions from 10,594 studies; +21k later annotations). Chest X-ray: VinDr-CXR (Nguyen et al., 2022) (18k radiographs with radiologist local labels), RSNA Pneumonia Detection (Shih et al., 2019) (pneumonia-region boxes), NIH ChestX-ray14 (Wang et al., 2017) (official bbox subset ∼984), and community ChestX-Det (∼3.5k instance-level boxes/masks). Mammography: CBIS-DDSM (Lee et al., 2017) (updated ROIs and *bounding boxes* for masses/calcifications). MRI: Duke Breast Cancer MRI (Saha et al., 2021) (radiologist-drawn *3D bounding boxes*). These sources satisfy the "lesion class with boxes" criterion and plug directly into our organ-aware seeding and CoT-generation pipeline.

We aggregate lesion-centric detection corpora with radiologist-drawn bounding boxes across major modalities—CT, chest X-ray, mammography, and breast MRI. Representative sources include DeepLesion for CT (Yan et al., 2018); VinDr-CXR, RSNA Pneumonia Detection, and the bbox subset of NIH ChestX-ray14 for chest radiography (Nguyen et al., 2022; Shih et al., 2019; Wang et al., 2017); CBIS-DDSM for mammography (Lee et al., 2017); and Duke Breast MRI for 3D annotations (Saha et al., 2021). All datasets satisfy the "lesion class with boxes" criterion and integrate directly with our organ-aware seeding and CoT generation pipeline, where boxes serve only to anchor organ assignment and rationale prompts.

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-annotated* (radiologist-drawn) lesion annotations, with $B_{ij} = (x_1, y_1, x_2, y_2) \in [0,1]^4$ an axis-aligned bounding box (normalized by image size) and $C_{ij} \in \mathcal{Y}$ a lesion label. Our goal is to construct a VQA-CoT corpus $\mathcal{D}_{\text{vqa}} = \{(I_i, B_{ij}, Q_{ij}, A_{ij}, \text{CoT}_{ij})\}_{i,j}$, where Q_{ij}, A_{ij}, and CoT_{ij} are generated *conditioned on* the lesion-organ context derived below.

Anatomical contextualization. A pretrained organ/structure segmentor $\mathcal{S}$ (we use TotalSegmentator (Wasserthal et al., 2023), CXAS (Seibold et al.,

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 $\mathcal{D}_{\text{vqa}}$

```
D_vqa ← ∅
for each image I_i with annotations A_i do
    {M_k} ← S(I_i)                                          ▷ organ masks
    for each (B_ij, C_ij) ∈ A_i do
        O_ij ← arg max_k IoU(B_ij, M_k)
        s_ij ← SeedFromTriplet((C_ij, O_ij))
        (Q_ij, A_ij, CoT_ij) ← M_VLM(Prompt(I_i, s_ij))
        D_vqa ← D_vqa ∪ {(I_i, B_ij, Q_ij, A_ij, CoT_ij)}
    end for
end for
return D_vqa
```

2023)) produces masks $\{M_k\}_{k=1}^{K}$ for each image I_i. For each *human* lesion box B_{ij}, the host organ is assigned by

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

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

Seed rationale & CoT-VQA generation. From (C_{ij}, B_{ij}, O_{ij}) we form a factual seed sentence s_{ij} (e.g., "There is a C_{ij} in the O_{ij}."). Given I_i and s_{ij}, a medical VLM, HuatuoGPT-Vision-34B (Chen et al., 2024) produces a localized question, a consistent answer, and a brief rationale:

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

thereby grounding CoT in the *human* lesion box and the *automatically selected* host organ.

3.2 Integrated CoT–Curriculum Strategy

The curriculum organizes supervision in three stages—explicit localization, implicit localization, and answer-only. Let g and h denote the visual and text encoders. Given image I, box B, and question Q, the model outputs a rationale CoT and answer A. We define $I' = \mathrm{draw_box}(I, B)$, $r_B = \mathrm{ROIAlign}(g(I), B)$, and $t_{\ell,o} = h($"[lesion=ℓ] in [organ=o]"$)$ as the lesion-organ anchor.

Objectives. We use stage-specific losses driven by *training* signals. Here, $\mathcal{L}_{\text{ans}}$ is answer likelihood; $\mathcal{L}_{\text{cot}}$ is rationale likelihood (teacher-forced when provided); $\mathcal{L}_{\text{ground}}$ aligns r_B with $t_{\ell,o}$; and $\mathcal{L}_{\text{attn-mask}}$ encourages model attention to overlap soft masks derived from B.

Easy (explicit localization). Training images include overlays I', and rationales are teacher-forced. The objective combines (1) answer likelihood, (2) rationale likelihood, and (3) grounding of r_B to $t_{\ell,o}$. Transition away from Easy is triggered when an EMA of the *Easy-stage training loss* plateaus over q consecutive epochs (see Alg. 2).

Medium (implicit localization). Boxes are *not* rendered to the model (no overlays visualized on image), but their masks remain in the supervision signal via $\mathcal{L}_{\text{attn-mask}}$. Concretely, we construct a soft mask m_B from the box B by Gaussian-blurring the binary box mask and downsampling it to the attention-map resolution, and add an alignment term $\mathcal{L}_{\text{attn-mask}} = \mathrm{KL}(\text{attn}\|m_B)$. Rationale supervision continues. Promotion toward Hard is considered once the *Medium-stage training loss* stabilizes and a *training-time rationale-loss gap* between Easy and Medium falls below a preset margin ϵ_{cot}.

Hard CoT (answer-only reasoning). In the hard stage, we train on the answer only ($\mathcal{L}\text{hard} = \mathcal{L}\text{ans}$). During training, we may sample multiple rationale candidates and condition on the one that maximizes $p(A \mid I, Q, \mathrm{CoT})$; the rationale itself is not supervised.

3.3 Curriculum Scheduling

The scheduler controls the per-epoch proportions $(\lambda_E^{(e)}, \lambda_M^{(e)}, \lambda_H^{(e)})$ of samples trained with $\mathcal{L}_{\text{easy}}, \mathcal{L}_{\text{medium}}, \mathcal{L}_{\text{hard}}$ in Sec. 3.2. A *domain* d is defined by lesion class and modality so that difficulty is adjusted within clinically coherent groups rather than globally. All transitions are *training-loss–driven*.

Algorithm 2 Domain-Aware Curriculum Scheduler (per epoch e)

Require: Domains $\{d\}$ (lesion class, modality); EMA rate ρ; ramp β_e; Hard budget $\lambda_H^{(e)}$; thresholds $(\gamma, \tau, \gamma_H, \epsilon_{\text{plat}}, q, \epsilon_{\text{cot}}, \delta_{\text{rise}})$; step sizes $(\eta_\uparrow, \eta_\downarrow)$; losses $\mathcal{L}_{\text{easy}}, \mathcal{L}_{\text{medium}}, \mathcal{L}_{\text{hard}}$

Initialize realized proportions $\lambda_E^{(e)} \leftarrow 0$, $\lambda_M^{(e)} \leftarrow 0$; keep $\lambda_H^{(e)}$ fixed within epoch
for each mini-batch **do**
  Sample $\lfloor \lambda_H^{(e)} B \rfloor$ Hard items from $\mathcal{D}_{\text{hard}}$; train with $\mathcal{L}_{\text{hard}}$
  Fill remaining slots from $\mathcal{D}_{\text{vqa}}$
  for each item x with domain d **do**
    Use EMAs $m_d^{\text{easy},(e-1)}, m_d^{\text{med},(e-1)}$ of *training* losses to compute
    $g_d^{(e)} \leftarrow \frac{m_d^{\text{easy},(e-1)} - m_d^{\text{med},(e-1)}}{m_d^{\text{easy},(e-1)} + \epsilon}$
    $P_{\text{med}} \leftarrow \beta_e \cdot \sigma\big((g_d^{(e)} - \gamma)/\tau\big)$
    Assign x to Medium w.p. P_{med}, else to Easy
    Train with $\mathcal{L}_{\text{medium}}$ or $\mathcal{L}_{\text{easy}}$ accordingly
  end for
end for
Update per-domain EMAs $m_d^{s,(e)}$ from epoch-mean *training* losses $\overline{\mathcal{L}}_d^s(e)$; update global EMA $\bar{m}^{(e)}$ and $\Delta\bar{m}^{(e)} \leftarrow \bar{m}^{(e)} - \bar{m}^{(e-1)}$
Compute training-time rationale gap $\text{gap}_{\text{cot}}^{(e)} \leftarrow \overline{\mathcal{L}}_{\text{cot}}^{\text{med}}(e) - \overline{\mathcal{L}}_{\text{cot}}^{\text{easy}}(e)$
if (**plateau**: $|\Delta\bar{m}^{(e')}| \le \epsilon_{\text{plat}}$ for last q epochs) **and** $\text{median}_d\, g_d^{(e)} \ge \gamma_H$ **and** $\text{gap}_{\text{cot}}^{(e)} \le \epsilon_{\text{cot}}$ **then**
  $\lambda_H^{(e+1)} \leftarrow \min\big(\lambda_H^{(e)} + \eta_\uparrow,\ \lambda_{H,\max}\big)$ ▷ increase Hard only from training-loss signals
else if $\Delta\bar{m}^{(e)} \ge \delta_{\text{rise}}$ **then**
  $\lambda_H^{(e+1)} \leftarrow (1 - \eta_\downarrow)\, \lambda_H^{(e)}$ ▷ reduce Hard if total training loss rises
else
  $\lambda_H^{(e+1)} \leftarrow \lambda_H^{(e)}$
end if

Per-domain difficulty tracking. For domain d and stage $s \in \{\text{easy}, \text{med}\}$, we maintain an EMA of the *training* loss:

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

where $\overline{\mathcal{L}}_d^s(e)$ is the epoch-mean of the stage-s objective used for items from domain d. We also track a global EMA $\bar{m}^{(e)}$ of total training loss to detect plateaus and regressions.

Base ramp for Medium. A ramp factor β_e governs when Medium samples appear:

$$\beta_e \;=\; \begin{cases} 0, & e \le 5, \\ \min\big(1, \frac{e-5}{\kappa}\big), & e > 5, \end{cases} \quad (\kappa \approx 10). \quad (2)$$

Adaptive assignment. Domain-specific progress adjusts the probability of assigning a sample to Medium; higher $g_d^{(e)}$ (i.e., smaller Medium loss relative to Easy) increases P_{med}, shifting mass toward implicit localization.

$$g_d^{(e)} \;=\; \frac{m_d^{\text{easy},(e-1)} - m_d^{\text{med},(e-1)}}{m_d^{\text{easy},(e-1)} + \epsilon} \quad (3)$$

The Hard budget $\lambda_H^{(e)}$ is increased only when the *training* loss plateaus for q epochs, the median $g_d^{(e)}$ across domains exceeds γ_H, and the training-time rationale-loss gap $\text{gap}_{\text{cot}}^{(e)}$ falls below ϵ_{cot}; it is reduced if the total training loss rises by at least δ_{rise}.

4 Experiments

4.1 Experimental Settings

Datasets. We construct the CoT-VQA dataset using diverse detection datasets, leveraging its diverse lesion annotations across CT, MRI, X-Ray images. For anatomical contextualization, we employ organ segmentation models (Wasserthal et al., 2023; Seibold et al., 2023). VQA performance is evaluated on three standard benchmarks: VQA-RAD (Lau et al., 2018), PMC-VQA, and SLAKE (Liu et al., 2021), covering different modalities (radiology, pathology) and both open- and closed-ended questions. We also evaluate the report generation performance on IU-Xray (Chen et al., 2020) and MIMIC-CXR (Johnson et al., 2019) to assess report quality-factual consistency and clinical completeness. We include both VQA and report generation datasets as they assess complementary aspects of clinical image understanding: VQA benchmarks provide short-form supervision with

Table 1: Main results on standard medical VQA benchmarks. We report **Recall** (%) for open-ended and **Accuracy** (%) for closed-ended questions. Our curriculum-based method achieves state-of-the-art performance across all datasets.

Method	VQA-RAD		SLAKE		PMC-VQA
	Open	Closed	Open	Closed	Closed
PMC-CLIP (Lin et al., 2023)	52.0	75.4	72.7	80.0	37.1
MedVInT-TE (Zhang et al., 2023a)	69.3	84.2	88.2	87.7	39.2
MedVInT-TD (Zhang et al., 2023a)	73.7	86.8	84.5	86.3	40.3
LLaVA-Med (Li et al., 2023a)	72.2	84.2	70.9	86.8	42.8
LLaVA-Med++ (Li et al., 2023a)	77.1	86.0	86.2	89.3	**61.9**
MedVP-LLaVA (Zhu et al., 2025)	89.3	**97.3**	91.6	92.9	58.3
MedCLM (Easy stage only) (Ours)	89.0	95.9	91.1	91.8	59.3
MedCLM (Easy → Medium) (Ours)	**90.4**	97.1	**92.2**	**93.4**	61.2

Table 2: Main results on radiology report generation. Comparisons across two widely-used benchmark datasets, *IU-Xray* and *MIMIC-CXR*, using standard evaluation metrics (BLEU, ROUGE, and METEOR).

Method	IU-Xray			MIMIC-CXR		
	BLEU	ROUGE	METEOR	BLEU	ROUGE	METEOR
PMC-CLIP (Lin et al., 2023)	8.57	10.90	7.30	10.76	11.60	9.92
MedVInT-TE (Zhang et al., 2023a)	9.96	12.66	8.48	12.51	13.48	11.53
MedVInT-TD (Zhang et al., 2023a)	10.04	12.76	8.55	12.61	13.59	11.62
LLaVA-Med (Li et al., 2023a)	9.64	12.26	8.21	12.11	13.05	11.16
LLaVA-Med++ (Li et al., 2023a)	10.82	13.76	9.21	13.59	14.64	12.52
MedVP-LLaVA (Zhu et al., 2025)	11.60	14.75	9.88	14.57	15.70	13.43
MedCLM (Easy stage only) (Ours)	11.54	14.67	9.82	14.49	15.62	13.36
MedCLM (Easy → Medium) (Ours)	**11.73**	**14.92**	**9.99**	**14.74**	**15.88**	**13.58**

explicit correctness criteria; report-generation corpora provide document-style supervision that emphasizes discourse coherence. Using both yields a balanced evaluation across structured QA and narrative reporting settings.

Implementation We build on **VIP-LLaVA**(Cai et al., 2024) which is 7B parameters and train with AdamW under a cosine-annealing schedule with linear warm-up (initial LR 2×10^{-5}, $\eta_{\min} = 10^{-6}$, warm-up ratio 3%), weight decay 0.05, and $(\beta_1, \beta_2) = (0.9, 0.98)$. We apply gradient clipping at 1.0 and mixed precision (bfloat16 when supported, otherwise fp16); the batch size is 1 per GPU. For our curriculum scheduler, we set the plateau patience to $q = 5$ epochs and the rationale-loss gap margin to $\epsilon_{\text{cot}} = 0.05$. Training begins with an Easy-only warm-up for ~ 5 epochs, after which harder samples are gradually introduced.

4.2 Main results

Medical VQA. Our Integrated CoT-Curriculum achieves strong and consistent gains across VQA-RAD, SLAKE, and PMC-VQA (Table 1(Lau et al., 2018; Liu et al., 2021; Zhang et al., 2023b)). The largest improvements appear on open-ended questions, where our method sets new state-of-the-art scores on VQA-RAD (Open) and SLAKE (Open/Closed), while remaining near-state-of-the-art on VQA-RAD (Closed) and competitive on PMC-VQA (Closed). We attribute this to the staged design(Bengio et al., 2009; Hacohen and Weinshall, 2019): the Easy stage secures robust visual grounding, and the Medium stage enforces reasoning without explicit location cues, mitigating vague or unsupported responses while preserving accuracy on closed-ended formats.

Report generation. As shown in Table 2, the Easy→Medium curriculum improves report quality on IU-Xray and MIMIC-CXR(Chen et al., 2020; Johnson et al., 2019) over strong baselines, with consistent gains in BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005). Although the gains are numerically modest, they are robust across datasets and metrics, indicating that the curriculum strategy improves the model's ability to generate factually grounded and coherent text. Qualitative analysis further

shows that the model trained with our method more reliably identifies and describes lesion locations and their likely causes, moving beyond the generation capability from the generic templates toward clinically meaningful, organ aware narratives(Wasserthal et al., 2023; Seibold et al., 2023).

4.3 Ablation study

Anatomical Rationale in Data. Integrating anatomical context into the data generation pipeline—by explicitly linking each lesion to its host organ—proved to be a crucial factor in improving model performance(Wasserthal et al., 2023; Jain et al., 2021). This contextual enrichment delivered uniform benefits across all datasets and training stages. As shown in Table 3, the most significant gains were observed in open-ended question-answering on VQA-RAD and SLAKE, where the added anatomical grounding helps the model formulate more precise and relevant responses (Lau et al., 2018; Liu et al., 2021). We observed that this approach effectively reduces errors arising from anatomical confusion, such as misattributing a finding in the lungs to the mediastinum (Jain et al., 2021; Seibold et al., 2023; Wasserthal et al., 2023).

Table 3: Ablation study on the effect of incorporating Anatomical Rationales. Performance comparisons on three benchmark datasets (*VQA-RAD*, *SLAKE*, and *PMC-VQA*), reporting results on both open and closed-ended questions where applicable. AC denotes *Anatomical Context* as defined in prior work (Lau et al., 2018; Wasserthal et al., 2023; Seibold et al., 2023).

Method	VQA-RAD		SLAKE		PMC-VQA
	Open	Closed	Open	Closed	Closed
Easy stage only (w/o AC)	86.9	95.0	88.8	90.7	58.1
Easy stage only (w/ AC)	89.0	95.9	91.1	91.8	59.3
Easy → Medium (w/o AC)	88.6	96.3	90.7	92.5	60.0
Easy → Medium (w/ AC)	**90.4**	**97.1**	**92.2**	**93.4**	**61.2**

By providing organ-aware seeds, we successfully constrain the model's explanation-generation process, steering it toward clinically plausible rationales without overfitting to the specific geometry of segmentation masks (Ross et al., 2017).

Effect of Hard COT. The introduction of the weakly supervised Hard CoT stage, which relies solely on final answer supervision, yielded mixed results (Wei et al., 2022; Wang et al., 2023; Zhang et al., 2024). On the SLAKE dataset, this stage acted as an effective regularizer, leading to minor improvements in performance by encouraging more concise and focused rationales (Liu et al., 2021) as shown in Table 4. However, on the VQA-RAD and PMC-VQA benchmarks, we observed a slight decline in accuracy (Lau et al., 2018; Zhang et al., 2023b). This suggests that while the Hard stage can refine reasoning when visual grounding is already robust, it may compromise answer calibration in scenarios with larger domain shifts or stronger textual priors (Lin et al., 2023). Given these findings, we adopted the more stable and consistently high-performing Easy-to-Medium curriculum for our main results, demonstrating its reliability across diverse medical VQA challenges (Bengio et al., 2009; Hacohen and Weinshall, 2019).

Table 4: Ablation study on the effect of introducing the Hard CoT stage. Model performances with and without Hard CoT supervision across three standard benchmarks (*VQA-RAD*, *SLAKE*, and *PMC-VQA*)

Method	VQA-RAD		SLAKE		PMC-VQA
	Open	Closed	Open	Closed	Closed
w/o Hard COT	**90.4**	**97.1**	92.2	93.4	**61.2**
w/ Hard COT	89.8	96.3	**92.5**	**93.6**	60.4

4.4 Qualitative results

Our Integrated CoT-Curriculum yields concise, anatomically consistent narratives by fostering internal spatial reasoning without overlays through staged Easy→Medium→Hard supervision and brief CoT steps (Bengio et al., 2009; Hacohen and Weinshall, 2019; Wei et al., 2022; Wang et al., 2023).

In binary QA (Fig. 2) across *VQA-RAD*, *SLAKE*, and *PMC-VQA*, the model correctly localizes pathology whether or not the question references a box, while baselines (LLaVA-Med++ and MedVP-LLaVA) fail in at least one case despite explicit visual prompts (Lau et al., 2018; Liu et al., 2021; Zhang et al., 2023b; Li et al., 2023a; Zhu et al., 2025).

For free-form description, our outputs align with key report findings (e.g., heart size at the upper limit of normal; stable mild pulmonary oedema; right-predominant bibasilar atelectasis with minimal left improvement; right IJ catheter at the cavo-atrial junction; no pneumothorax/effusion), avoiding over-calls and omissions observed in the baselines and marking progress toward clinically useful medical VLMs (Li et al., 2023a; Zhu et al., 2025; Singhal et al., 2023).

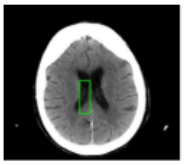
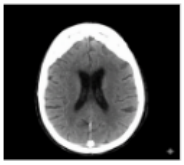
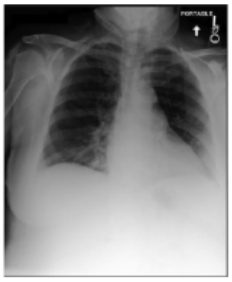
	Question	According to the box, is potential pathology indicated?
	LLaVA-Med ++	No
	MedVP-LLaVA	Yes
	Ours	Yes
	Answer	Yes
	Question	Is potential pathology indicated in image?
	LLaVA-Med ++	No
	MedVP-LLaVA	No
	Ours	Yes
	Answer	Yes
	Question	Describe this image
	LLaVA-Med ++	This is a frontal chest X-ray taken in an emergency department. The heart appears enlarged and there is fluid in both lungs, suggesting pulmonary oedema or pleural effusion. Atelectasis, which is the collapse or closure of lung tissue resulting in reduced gas exchange, may be present. Atelectasis can be caused by blockage of the airways or external pressure on the lung. Additionally, several tubes and lines are visible. The lungs otherwise appear clear without signs of pneumonia or pneumothorax.
	MedVP-LLaVA	The frontal chest radiograph shows a heart size at the upper limit of normal. Mild pulmonary oedema is present and there is bibasilar atelectasis, more pronounced on the right but slightly improved on the left. A right internal jugular venous catheter projects to the cavo-atrial junction. There is no pneumothorax or pleural effusion.
	Ours	Inspecting the cardiac silhouette, the heart appears within the upper limits of normal. The pulmonary vasculature and lower lung zones show mild oedema that is unchanged from prior studies. Bibasilar atelectasis is present, larger on the right and minimally improved on the left. A right internal jugular catheter is present with its tip at the cavo-atrial junction. There is no pneumothorax or pleural effusion.
	Answer	Cardiac size is top normal. Mild pulmonary edema is grossly unchanged. Bibasilar atelectasis larger on the right have minimally improved on the left. Right IJ catheter tip is in the cavoatrial junction. There is no pneumothorax or pleural effusion.

Figure 2: Qualitative comparison of model outputs on binary and descriptive medical VQA tasks. The first two rows show binary QA cases with and without explicit box references, where our method correctly identifies pathology while baselines fail in at least one instance. The third row shows a free-form description task on a chest X-ray: our model produces a clinically faithful report aligned with the reference, whereas LLaVA-Med++ introduces extraneous findings and MedVP-LLaVA omits key stability details.

5 Conclusion

We presented an automated framework that transforms detection datasets into medical VQA samples with clinically grounded Chain-of-Thought (CoT) reasoning and a structured curriculum that progresses from explicit grounding to implicit localization. This unified design encourages models to learn spatial reasoning gradually while maintaining alignment between visual evidence and textual interpretation. The framework achieves strong performance on medical VQA benchmarks, especially in open ended settings, and also improves radiology report generation by producing concise and anatomically consistent descriptions. Using a 7B backbone (ViP-LLaVA 7B), our method surpasses comparable 7B models such as MedVP-LLaVA 7B and remains competitive with larger 13B variants including LLaVA-Med++. These results demonstrate that the improvements stem from the structure of the curriculum and anatomy based CoT reasoning rather than the scale of parameters.

6 Limitations

Our approach depends on lesion-box supervision and organ segmentation quality; errors or gaps in these inputs can propagate to CoT generation and training signals. While the Hard-stage CoT can act as a weak regularizer, its benefits are dataset-sensitive, and the most reliable default remains the Easy→Medium schedule. Finally, we did not exhaustively benchmark parity-sized 13B variants or clinically validate in prospective workflows, leaving systematic size-controlled comparisons and real-world evaluation to future work.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, and 1 others. 2022. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*.

Hakan Bilen and Andrea Vedaldi. 2016. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. 2024. Making large multimodal models understand arbitrary visual prompts. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. 2024. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Angel Cruz-Roa, Hannah Gilmore, Ajay Basavanhally, Michael Feldman, Shridar Ganesan, Natalie N.C. Shih, John Tomaszewski, Anant Madabhushi, and Fabio González. 2017. Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent. *Scientific Reports*, 7:46450.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2022. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*.

Guy Hacohen and Daphna Weinshall. 2019. On the power of curriculum learning in training deep networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*.

Zobia Hameed, Begonya Garcia-Zapirain, J. J. Aguirre, and 1 others. 2022. Multiclass classification of breast cancer histopathology images using multilevel features of deep convolutional neural network. *Sci Rep*, 12:15600.

Yixuan He, Xin Yang, Yiqing Shi, and 1 others. 2020. Pathvqa: 30000+ questions for medical visual question answering. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

Satya Jain, Pranav Upadhyaya, Michael Chen, and 1 others. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *Preprint*, arXiv:1901.07042.

Joyce Lau, Tirthankar Gayen, Asma Ben-Abacha, and Dina Demner-Fushman. 2018. A dataset for medical visual question answering (vqa-rad). In *IEEE BigComp Workshops*.

Han Le, Rajarsi Gupta, Le Hou, Shahira Abousamra, Danielle Fassler, Luke Torre-Healy, Richard A. Moffitt, Tahsin Kurc, Dimitris Samaras, Rebecca Batiste, Tianhao Zhao, Arvind Rao, Alison L. Van Dyke, Ashish Sharma, Erich Bremer, Jonas S. Almeida, and Joel Saltz. 2020. Utilizing automated breast cancer detection to identify spatial distributions of tumor-infiltrating lymphocytes in invasive breast cancer. *The American Journal of Pathology*, 190(7):1491–1504.

Rebecca S. Lee, Francisco Gimenez, Assaf Hoogi, and Daniel L. Rubin. 2017. A curated breast imaging subset of DDSM. *Scientific Data*.

Dongxu Li, Junnan Li, Kai Zhang, and 1 others. 2023a. LLaVA-Med: Training a large language-and-vision assistant for biomedicine. *arXiv preprint arXiv:2308.02463*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, and 1 others. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, and You Zhang. 2023c. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL-04 Workshop on Text Summarization Branches Out*.

Wenhui Lin, Ming Yang, Zhi Huang, and 1 others. 2023. PMC-CLIP: Contrastive language-image pre-training using biomedical documents. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Also available as arXiv preprint.

Haotian Liu, Chunyuan Li, Qifeng Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Pengfei Liu, Kun Yuan, Xian Fu, and 1 others. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical vqa. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

Shilong Liu, Zhaoyang Zeng, Tianhe Zhang, and 1 others. 2023b. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Ana Marasović, Trenton Jiang, and Noah A. Smith. 2020. Natural language rationales with full supervision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ha Q. Nguyen, Hieu H. Pham, Minh C. Nguyen, and 1 others. 2022. VinDr-CXR: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint*.

Long Ouyang, Jeff Wu, Xu Jiang, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

Andrew Sklar Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*.

A. Saha, M. R. Harowicz, L. J. Grimm, J. Weng, E. H. Cain, C. E. Kim, S. V. Ghate, R. Walsh, and M. A. Mazurowski. 2021. Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations. [Data set].

Constantin Seibold, Alexander Jaus, Matthias A. Fink, Moon Kim, Simon Reiß, Ken Herrmann, Jens Kleesiek, and Rainer Stiefelhagen. 2023. Accurate fine-grained segmentation of human anatomy in radiographs via volumetric pseudo-labeling.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, and 1 others. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

George Shih, Carol C. Wu, Safwan S. Halabi, and 1 others. 2019. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. In *Radiological Society of North America (RSNA)*.

Chang Shu, Baian Chen, Fangyu Liu, Zihao Fu, Ehsan Shareghi, and Nigel Collier. 2023. Visual med-alpaca: A parameter-efficient biomedical llm with visual capabilities.

Krishna Kumar Singh and Yong Jae Lee. 2018. Hide-and-seek: A data augmentation technique for weakly-supervised localization and beyond. *arXiv preprint arXiv:1811.02545*.

Karan Singhal, Shekoofeh Azizi, Tu Tu, and 1 others. 2023. Large language models encode clinical knowledge (Med-PaLM). *Nature Medicine*.

Augustin Toma, Patrick R. Lawler, Jimmy Ba, Rahul G. Krishnan, Barry B. Rubin, and Bo Wang. 2023. Clinical camel: An open expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint*.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *arXiv preprint arXiv:1705.02315*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, and 1 others. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*.

Thomas Wasserthal, Henning Breit, Felix Meyer, and 1 others. 2023. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct. *Scientific Reports*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yikuan Wu, Hongyi Ye, Yutao Liu, and 1 others. 2023. PMC-LLaMA: Further finetuning LLaMA on medical papers. *arXiv preprint arXiv:2304.14454*.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint*.

Ke Yan, Xiaosong Wang, Le Lu, and Ronald M. Summers. 2018. Deeplesion: Automated mining of large-scale lesion annotations and universal lesion detection in ct. *Journal of Medical Imaging*.

Juyoung Yun, Shahira Abousamra, Chen Li, Rajarsi Gupta, Tahsin Kurc, Dimitris Samaras, Alison Van Dyke, Joel Saltz, and Chao Chen. 2024. Uncertainty estimation for tumor prediction with unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 6946–6954.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Eric Zelikman, Yuhuai Wu, Niklas Muennighoff, and 1 others. 2023. Star: Bootstrapping reasoning with reasoning. *arXiv preprint arXiv:2302.06161*.

S. Zhang, Alexander A. S. A. Kuo, Z. Lin, and Heung-Yeung Shum. 2018. Top-down neural attention by excitation backprop. *International Journal of Computer Vision (IJCV)*.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023a. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

Yusheng Zhang, Zhen Lu, Yifan Liu, and 1 others. 2023b. Pmc-vqa: Visual question answering on biomedical literature. In *NeurIPS Datasets and Benchmarks Track*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024. Multimodal chain-of-thought reasoning in language models.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yuhuai Zhou, Quoc V. Le, and Graham Neubig. 2023. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations (ICLR)*.

Kangyu Zhu, Ziyuan Qin, Huahui Yi, Zekun Jiang, Qicheng Lao, Shaoting Zhang, and Kang Li. 2025. Guiding medical vision-language models with diverse visual prompts: Framework design and comprehensive exploration of prompt variations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11726–11739.

A More qualitative results

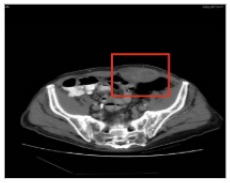
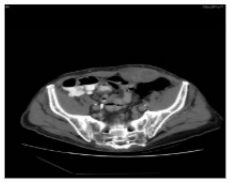
	Question	According to the box, is there any mass?
	MedVInT-TD	No
	MedVP-LLaVA	Yes
	Ours	Yes
	Answer	Yes
	Question	Is there any mass?
	MedVInT-TD	No
	MedVP-LLaVA	No
	Ours	Yes
	Answer	Yes
	Question	Describe this image
	MedVInT-TD	Cardiac silhouette moderately enlarged. Bilateral pleural effusions with basilar atelectasis. Perihilar interstitial/alveolar edema pattern; no visible pneumothorax.
	MedVP-LLaVA	Moderate bilateral pleural effusions with dependent bibasilar atelectasis. Perihilar hazy opacities and vascular indistinctness compatible with moderate pulmonary edema. Cardiomediastinal silhouette appears enlarged. No pneumothorax.
	Ours	Cardiac size is difficult to assess given moderate bilateral pleural effusions but appears at least moderately enlarged. Mediastinal contours are grossly unremarkable. Perihilar haziness with vascular indistinctness and diffuse alveolar opacities are compatible with moderate pulmonary edema. Bibasilar compressive atelectasis is present. No pneumothorax. Moderate multilevel degenerative changes of the thoracic spine.
	Answer	Heart size is difficult to assess given the presence of moderate to large bilateral pleural effusions, but appears at least moderately enlarged. The mediastinal contours are grossly unremarkable. Perihilar haziness with vascular indistinctness and diffuse alveolar opacities are compatible with moderate pulmonary edema. Bibasilar compressive atelectasis is demonstrated. No pneumothorax is seen. Moderate multilevel degenerative changes are noted in the thoracic spine.

Figure 3: **Robustness to prompt form and supervision** MEDCLM remains consistent across question types—answering mass-presence queries and producing comprehensive radiographic summaries that match the reference impression. In contrast, baselines fluctuate between false negatives/positives and terse or incomplete descriptions. The qualitative gains mirror our training design that decouples explicit grounding from answer-only reasoning while preserving clinical fidelity

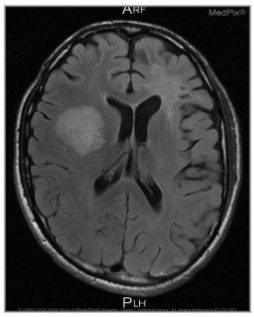
	Question	Check if there is any abnormality, if so what is the location?
	MedVInT-TE	The scan appears to be within normal limits
	MedVInT-TD	Yes, a lesion is observed. It is located in the left parietal lobe
	LLaVA-Med	Abnormal finding is present. The location is the left temporal lobe
	LLaVA-Med++	An abnormal lesion is identified in the right frontal lobe
	MedVP-LLaVA	There is a suspected lesion within the frontal region of the right cerebral hemisphere
	Ours	Conspicuous lesion in the right frontal lobe
	Answer	Yes, right frontal lobe
	Question	Describe this image
	MedVInT-TE	Subacute infarction involving the left PCA territory centered in the occipital cortex.
	MedVInT-TD	Acute–subacute infarct along the left PCA distribution, subtle posterior corpus callosum involvement suspected.
	LLaVA-Med	Acute ischemia in the right posterior cerebral artery territory.
	LLaVA-Med++	Left PCA-territory infarct with possible extension toward the temporal lobe; mild mass effect cannot be excluded.
	MedVP-LLaVA	Left occipital infarct consistent with PCA territory, no hemorrhage identified
	Ours	Infarct in the left PCA distribution with an additional small focus in the left splenium of the corpus callosum
	Answer	The image shows acute or subacute left PCA distribution infarct. Additional smaller infarct within the left splenium of the corpus callosum is also seen.

Figure 4: MEDCLM correctly localizes the lesion (right frontal lobe) and produces territory-aware descriptions consistent with the reference. Competing models disagree on lobe/hemisphere or vascular territory, reflecting unstable localization. These results illustrate our curriculum's benefit for implicit spatial reasoning and clinically faithful phrasing

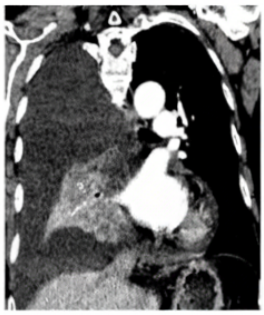
	Question	Is cardiomegaly shown?
	HealthGPT	Yes
	HuatuoGPT-Vision	Yes
	Ours	Yes
	Answer	Yes
	Question	Check if there is any abnormality, if so what is the location?
	HealthGPT	Abnormality exists in the right lung
	HuatuoGPT-Vision	No abnormality present
	Ours	An abnormality is identified in the right lung
	Answer	There is abnormality in the right lung
	Question	Describe this image
	HealthGPT	Large right pleural effusion with compressive atelectasis of the right middle and lower lobes. A bulging, hypo-enhancing nodular lesion in the right lower lobe is highly suggestive of primary bronchogenic malignancy
	HuatuoGPT-Vision	Right pleural effusion causing collapse of the right middle and lower lobes. A mass-like, low-enhancing nodule in the right lower lobe raises concern for bronchogenic carcinoma
	Ours	Large right pleural effusion with volume loss of the right middle and lower lobes within the right lower lobe, a bulging hypo-enhancing nodule consistent with a primary bronchogenic tumor
	Answer	Computed tomography (CT) of the chest confirms a large right pleural effusion with collapse of the middle and lower lobe.The bulging nodular hypoenhancing mass (*) in the right lower lobe is suspicious for primary bronchogenic malignancy.

Figure 5: Although comparing to recent sota models like Huatuo and HealthGPT, MedCLM still shows remarkable performance. For yes/no and localization queries, MEDCLM aligns with the reference answers across cases. For descriptive prompts, it generates concise, clinically grounded reports without unsupported claims. (HuatuoGPT-Vision is 8B , HealthGPT is 14B parameters here)