This CVPR Findings paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

Step-CoT: Stepwise Visual Chain-of-Thought for Medical Visual Question Answering

Lin Fan[1], Yafei Ou[2,3] *, Zhipeng Deng[4], Pengyu Dai[2,5], Chongxian Hou[6], Jiale Yan[7], Yaqian Li[7], Kaiwen Long[7], Xun Gong[1], Masayuki Ikebe[3], Yefeng Zheng[4]
[1] Southwest Jiaotong University [2] RIKEN [3] Hokkaido University [4] Westlake University
[5] The University of Tokyo [6] Shenzhen People's Hospital [7] Li Auto Inc.

Abstract

Chain-of-thought (CoT) reasoning has advanced medical visual question answering (VQA), yet most existing CoT rationales are free-form and fail to capture the structured reasoning process clinicians actually follow. This work asks: Can traceable, multi-step reasoning supervision improve reasoning accuracy and the interpretability of Medical VQA? To this end, we introduce Step-CoT, a large-scale medical reasoning dataset with expert-curated, structured multi-step CoT aligned to clinical diagnostic workflows, implicitly grounding the model's reasoning in radiographic evidence. Step-CoT comprises more than 10K real clinical cases and 70K VQA pairs organized around diagnostic workflows, providing supervised intermediate steps that guide models to follow valid reasoning trajectories. To effectively learn from Step-CoT, we further introduce a teacher-student framework with a dynamic graph-structured focusing mechanism that prioritizes diagnostically informative steps while filtering out less relevant contexts. Our experiments show that using Step-CoT can improve reasoning accuracy and interpretability.

Benchmark: *github.com/hahaha111111/Step-CoT*

Dataset Card: *huggingface.co/datasets/fl-15o/Step-CoT*

1. Introduction

Medical Visual Question Answering (Med-VQA) has emerged as a critical topic in healthcare AI, leveraging multi-modal deep learning to answer clinical questions related to images posed in natural language [6, 7, 12–14, 25]. Leveraging the ability to generate coherent responses while incorporating extensive medical domain knowledge, Med-VQA has demonstrated practical utility across diverse tasks such as computer-aided diagnosis and fine-grained tumor attribute recognition [16, 25, 45]. Recent advances in Med-VQA models have evolved from focusing on performance enhancement through scaling model architectures and expanding pre-training datasets [44], such as ExGra-Med [29], LLaVA-Med [20], MedGemma [34], and LLaVA-Tri [42], to improving interpretability by decomposing complex diagnostic tasks into step-by-step reasoning processes through CoT mechanisms [43, 48], including ReasonMed [3], HVCR [11] and MedCoT [26]. By explicitly revealing their intermediate reasoning steps, CoT methods enhance both predictive accuracy and interpretability, rendering them particularly valuable for high-stakes domains such as healthcare.

Recent studies have explored automatic generation of CoT reasoning data for Large vision language models (LVLMs) to improve reasoning accuracy and enhance interpretability. Approaches such as MedCoT [26], MedThink [15], ReasonMed [36], and the HVCR [11] provide textual reasoning traces that enable models to produce rationales alongside predictions. To further connect reasoning with visual evidence, datasets like V2T-CoT [41], Med-GRIT-270k [19], and MedTrinity-25M [42] pair CoT statements with image annotations (as shown in Table 1). Although these approaches improve data availability, their effectiveness in modeling clinical reasoning remains limited by two major issues: (i) These datasets lack a structured, stepwise diagnostic protocol. They either provide free-form rationales or automatically generated reasoning chains that fail to align with real clinical workflows and omit intermediate diagnostic states reflecting radiologists' sequential decision-making. (ii) Most CoT datasets rely heavily on GPT-4.1-based synthetic rationales derived from existing image-text pairs, raising serious concerns about factual inconsistency.

Beyond dataset construction, training paradigms for CoT in LVLMs further enlarge this gap. Most CoT training paradigms for LVLMs, which rely on Supervised Fine-Tuning (SFT) or Reinforcement Learning (RL), are inherently non-interactive and perceptually static [5, 50], primarily using a static image plus question input and reporting final-answer outputs without explicit action-level reason-

*Corresponding author: Yafei Ou (yafei.ou@riken.jp)

Table 1. Comparison of recent medical reasoning datasets with Step-CoT.

Dataset	Instances	CoT Construction	Expert Involvement	Clinical Workflow
PMC-VQA [47]	227K	×No CoT	✓Yes	×No
Med-GRIT-270k [19]	270k QA pairs	✓GPT-generated	×No experts	×No
MedCoT [26]	50k Extends Med-VQA	✓Expert validation	✓Yes	✓Partial
ReasonMed [3]	370k reasoning samples	✓Agent-generated	✓Multi-agent validation	×No
HVCR [11]	31k QA pairs	Expert validation	✓Yes	×No
V2T-CoT [41]	39k examples	✓GPT-generated CoTs	×No experts	×No
MedThink [15]	50k Extends Med-VQA	✓Decision-making rationales	✓Semi-auto + human validation	×No
Step-CoT (Ours)	70K QA pairs with reasoning	✓Structured multi-step	✓Semi-auto + human validation	✓Yes

ing chains [20], limiting the model to the initial input image and preventing it from actively gathering new information or refining its perception. For example, BLIP-2 [23] has achieved strong open-domain VQA performance but lacks mechanisms for interactive perceptual refinement or multi-step visual reasoning, while emerging specialized medical LVLMs such as LLaVA-Med [20] demonstrate domain adaptation but still without action-level reasoning chains [20]. Similarly, the recent MedVLM-R1 [30] framework uses RL to incentivize reasoning paths in medical image analysis. Yet, the perceptual input remains static, and the model cannot execute intermediate actions that change what it perceives.

These limitations underscore the need for a structured reasoning annotation framework that not only integrates visual reasoning and aligns multi-step inference with multimodal clinical evidence, but also enables traceable, stepwise reasoning interactions with visual data, allowing each reasoning step to dynamically update and refine diagnostic understanding in a clinically consistent manner. This leads to the central question of this work:
Can traceable, multi-step reasoning supervision improve reasoning accuracy and the interpretability of Medical VQA?

To solve this problem, we introduce Step-CoT, a novel medical dataset for multi-step vision-language reasoning. Step-CoT organizes clinical diagnostic workflows with clinician-curated intermediate reasoning steps, providing the necessary supervision for models to learn dynamic, multi-step reasoning through actionable sequences that can alter perceptual input. By structuring reasoning as a series of clinical actions, our dataset facilitates a transition from static reasoning to a multi-step problem-solving paradigm. Along with the dataset, we release benchmark evaluations, pretrained baselines, and a teacher-student framework to support supervised intermediate reasoning and context-conditioned inference. To summarize, this paper makes the following contributions:

- We present Step-CoT, a dedicated medical visual CoT dataset comprising more than 10K real clinical cases and 70K question-answer pairs. Each instance contains clinically grounded reasoning chains. Our experiments show that Step-CoT progressively aligns multi-step reasoning with visual evidence, guides models to follow valid and traceable diagnostic trajectories, enhancing the accuracy, interpretability, and clinical relevance of Med-VQA.
- To effectively learn from Step-CoT, we propose an innovative teacher-student CoT reasoning framework that distills complex clinical reasoning knowledge from the dataset into a lightweight student model, thereby improving generalization and adaptability to diverse diagnostic tasks.
- We introduce a visual CoT benchmark for Med-VQA that evaluates performance in scenarios requiring clinically faithful reasoning chains and evidence-based justifications for supporting diagnosis.

2. Step-CoT Dataset

We constructed the medical visual CoT dataset Step-CoT (see Fig. 1). Step-CoT formalizes the clinical diagnostic trajectory as a seven-step, sequential reasoning process and applies full supervision across the entire diagnostic pipeline, including ground-truth answers and intermediate reasoning annotations for each step. Each sample comprises one medical image, seven clinically relevant reasoning questions, and their corresponding answers with intermediate supervisory signals. Every answer is accompanied by an explanatory reasoning chain that articulates the thought process behind it.

2.1. Data Collection

This work comprises original Chest X-ray (CXR) images and diagnostic text drawn from three public sources (totaling 10,068 CXR studies): (i) IU X-Ray [9], we use the subset of 3749 CXR studies. (ii) PadChest-GR [4], we use the subset of 3230 CXR studies. (iii) Med-Image-Reports [1], we use the subset of 3089 CXR studies. Appendix Sec. B provides more detailed information regarding the dataset, including the sample selection criteria and dataset structure, among other details.

[1] https://huggingface.co/datasets/zirui3/med-image-reports

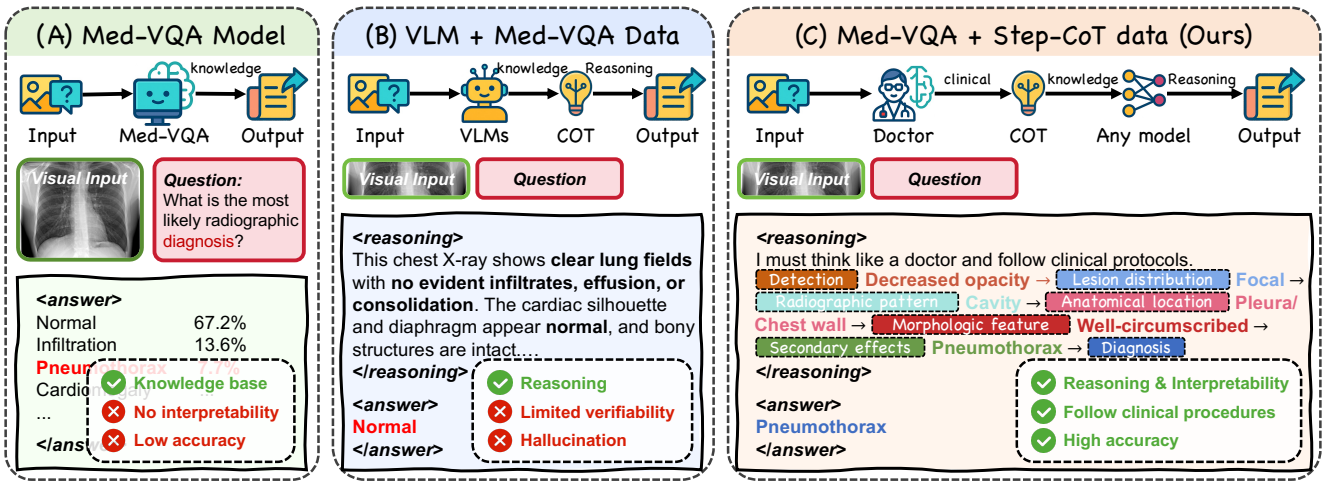


Figure 1. Overview of the Step-CoT dataset. (A) Conventional Med-VQA approaches, where models take an image and a question as input, perform multi-modal feature fusion and output a diagnostic answer. Although leveraging multi-modal knowledge, this paradigm lacks interpretability and often yields limited diagnostic accuracy. (B) Enhances interpretability by integrating large language models with CoT reasoning to generate intermediate explanations; however, such reasoning is often unreliable. (C) Our proposed Step-CoT dataset and training framework, which introduces explicit intermediate supervision. By guiding the model to learn structured clinical reasoning steps, Step-CoT not only improves interpretability through trustworthy intermediate reasoning but also enhances diagnostic accuracy.

2.2. Data Annotation Process

This study employed a structured analytical framework to evaluate chest X-ray radiology reports using DeepSeek-R1, an LLM that has demonstrated robust performance in clinical reasoning tasks [32, 39]. For annotation, we first collected the original clinician-authored, free-text radiology reports provided with the datasets. These unstructured reports were submitted to DeepSeek-R1 (prompts are detailed in Appendix Sec. A) to extract structured image findings and attributes. The model outputs were then mapped onto the predefined multi-step reasoning schema, producing per-step question-answer entries and explicit reasoning chains. The dataset consisted of systematically processed radiology reports paired with chest X-ray images, ensuring comprehensive case representation. Each case included three structured components: a sequential set of clinical questions, corresponding answer logic, and explicit reasoning chains. This design provided the methodological foundation for evaluating both diagnostic accuracy and interpretability in medical visual question answering. The structured components of the analysis framework are summarized in the Appendix Sec. B.

The case analysis protocol was constructed to emulate established radiological diagnostic workflows [2, 8, 33]. For each patient case, clinical questions progressed in a stepwise manner, beginning with abnormality detection, moving through pattern characterization and spatial assessment, and concluding with diagnostic synthesis. This seven-step cascade ensured that each analytical stage logically built upon prior conclusions, thereby maintaining clinical coherence and mirroring the reasoning structure employed by expert radiologists. The prompt engineering implemented a seven-step analytical cascade that maintains contextual continuity across diagnostic stages:

- **Abnormal Radiodensity Detection (Detection step)** Determine the presence or absence of abnormal radiodensity in the lungs or surrounding thoracic structures (e.g., increased, decreased, or mixed opacity).
- **Appearance Survey (Lesion distribution step and Radiographic pattern step)** Characterizes any detected abnormality by assessing its spatial distribution (e.g., focal, diffuse) and its predominant basic radiographic pattern (e.g., reticular, consolidation).
- **Feature Analysis (Anatomical location step, Morphologic feature step, and Secondary effects / associated signs step)** Refines the description by specifying the precise anatomical location of the lesion (e.g., right upper lobe), its margins and internal morphology (e.g., well-circumscribed), and any secondary effects on surrounding structures or lung volumes (e.g., mediastinal shift).
- **Diagnostic Synthesis (Diagnosis step)** Integrates all previous findings to formulate a comprehensive radiographic diagnosis or impression (e.g., atelectasis).

This cascading reasoning structure ensures that each analytical step logically builds upon previous conclusions, creating a coherent diagnostic narrative that reflects expert clinical thought processes. The prompt was specifically instructed to reference prior step conclusions in subsequent reasoning, thereby maintaining diagnostic continuity and reducing contextual fragmentation. All diagnostic proce-

dures and data labels were reviewed and verified by a board-certified physician, who obtained his Medical Practitioner Certificate in 2018.

2.3. Statistics of Step-CoT

We performed a statistical analysis of the Step-CoT dataset, examining data sources, disease prevalence, answer distributions, reasoning lengths, and other key attributes. Detailed statistics are summarized in Fig. 2. From the diagnostic-label perspective (Fig. 2 (A)), IU-CXR is enriched with normal, effusion, and cardiomegaly cases; BIMCV is dominated by Normal and Nodule cases; and Med-Image-Reports exhibits a comparatively broader spread, including less frequent conditions such as mass and pneumothorax. These inter-dataset differences, together with step-level applicability signals, provide a valuable substrate for assessing model robustness and generalizability under varying label priors and reporting styles. Data drawn from the three sources show a consistent pattern: instances labeled Normal constitute more than 50% of each cohort, while other disease categories are relatively balanced. This class imbalance has been widely documented in clinical cohorts and aligns with the clinical findings reported by Alshanketi [1]. Preserving this natural imbalance, rather than performing artificial class balancing, can help avoid a common pitfall in AI studies and ensure that training and evaluation results retain clinical relevance.

In terms of textual complexity (Fig. 2 (B)), the mean word count per reasoning step varies across datasets (IU-CXR: 15-19 words; Hugging: 15-25 words; BIMCV: 9-11 words), indicating heterogeneous linguistic complexity and reasoning depth. The modest increase in sentence length toward later steps suggests that more elaborate, conclusive statements tend to be produced after successive reasoning stages.

Analysis of the multi-step answer distribution across the dataset (Fig. 2 (C)) indicates that many samples follow relatively consistent reasoning trajectories, in which a limited number of principal diagnostic flows account for a substantial portion of reasoning transitions. This pattern suggests that the seven-step question-answer sequences tend to form stable, clinically coherent pathways that broadly align with structured diagnostic reasoning processes observed in clinical practice [10]. Notably, the dominance of reasoning flows implies that the final diagnostic conclusions are strongly conditioned by earlier reasoning steps, reflecting the inherent interdependence between preliminary observations and ultimate diagnostic judgments. The observed regularity implies that the dataset can decompose complex diagnostic reasoning into sequential and interpretable subtasks, where each step contributes complementary clinical information that collectively supports the final diagnostic interpretation. The concentration of reasoning flow within a few central nodes further supports the procedural consistency and clinical plausibility of the Step-CoT design, suggesting that it provides a structured approximation of step-by-step medical reasoning rather than a collection of isolated decision points.

3. Enhancing Med-VQA Method with Step-CoT Dataset

We introduce a visual CoT framework for the Step-CoT dataset, which provides stepwise supervision corresponding to clinically meaningful sub-tasks. Steps in Step-CoT have variable lengths and non-linear dependencies across diseases [33]. To capture this, each step is modeled as a node and clinical dependencies as edges, forming a graph. A global memory node dynamically aggregates information, preserving contextual coherence and enabling interpretable multi-step reasoning. This framework adopts a collaborative teacher-student paradigm. Further implementation details are provided in Appendix Sec. D.

Teacher Model Overview. The teacher processes S sequential clinical questions (steps) and an explicit *memory node* that aggregates cross-step information. For each step $s \in \{1, \ldots, S\}$ the model: (1) encodes the step prompt with a shared text encoder, (2) forms a node set $\{\mathbf{t}_1, \ldots, \mathbf{t}_S, \mathbf{m}\}$ where $\mathbf{t}_s$ is the CLS embedding of step s and $\mathbf{m}$ is the learnable memory, (3) updates node states with a multi-head Graph Attention Network (GAT), (4) composes a step context by fusing the step node and the memory node, and (5) predicts the step label using a step-specific classifier. After prediction, the teacher writes a compact prediction embedding back to the memory via a gated GRU update, enabling information flow to later steps.

GAT Memory. The GAT implements multi-head attention between nodes. For a single head, after a linear map W, we compute attention scores as Eq. 1, where $\mathbf{h}_i$ and $\mathbf{h}_j$ denote the input features of node i and node j, respectively, W is a learnable linear projection matrix, and $\mathbf{a}_{\text{src}}$, $\mathbf{a}_{\text{dst}}$ are learnable attention vectors for source and destination nodes.

$$e_{ij} = \text{LeakyReLU}\big(\mathbf{a}_{\text{src}}^{\top}(W\mathbf{h}_i) + \mathbf{a}_{\text{dst}}^{\top}(W\mathbf{h}_j)\big), \quad (1)$$

The normalized attention coefficient is defined as Eq. 2, where α_{ij} represents the attention weight from node i to node j. The memory node corresponds to the last node in the graph and thus receives aggregated information from all step nodes.

$$\alpha_{ij} = \text{softmax}_j(e_{ij}), \quad (2)$$

Student Model and Distillation. While the teacher graph is explicitly designed to capture rich, inter-step reasoning

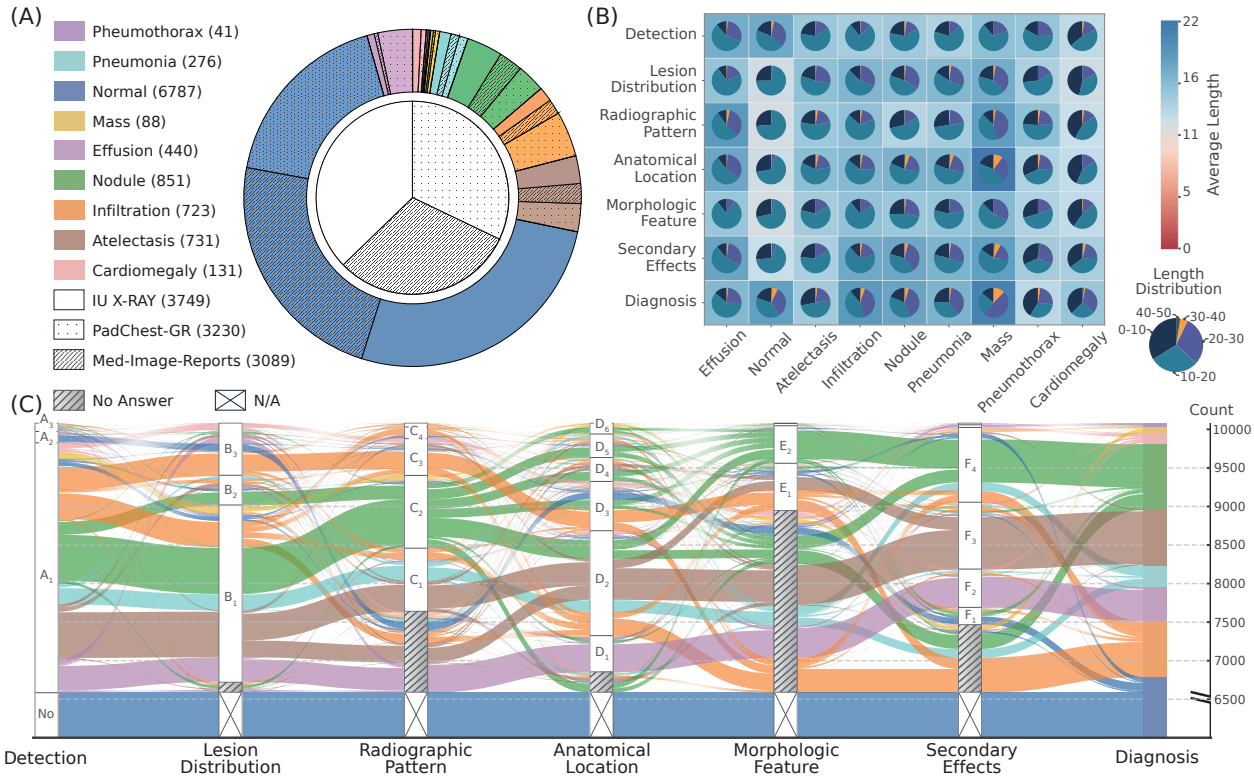


Figure 2. Distribution and statistics for the data sources, disease prevalence, answer distributions, and reasoning lengths in the Step-CoT dataset. (A) The inner ring illustrates the proportional distribution across different datasets, while the outer ring represents the distribution of various disease categories within the datasets. (B) This confusion matrix, organized by disease categories and reasoning steps, visualizes the average reasoning chain length. Each cell contains a pie chart representing the statistical distribution of samples across different chain lengths, while the marginal histograms on the axes display the sample count distributions by chain length for individual steps (x-axis) and disease categories (y-axis). (C) This diagram presents the outcome transition statistics between consecutive reasoning steps, mapping the flow of diagnostic conclusions throughout the clinical reasoning pathway. The name of each annotation (e.g., A1) can be referred to in the dataset description section of Appendix Sec. B.

dependencies on our Step-CoT dataset, these learned relations are often highly structured and dataset-specific. To enable practical deployment and cross-dataset transfer, we therefore train a lightweight student model via knowledge distillation. The student aims not only to preserve the teacher's reasoning behavior but also to learn a compressed, more generalizable representation of those complex relations. This reduces inference cost, simplifies deployment, and mitigates adaptation problems that arise when the teacher's dataset-specific reasoning patterns are applied to datasets with different reasoning complexity or distributional characteristics.

The student model is a compact chain model that only uses image features and a sequence of light-weight heads. We distill from the teacher to the student using three complementary losses per step:

Hard Supervision: cross-entropy loss on labeled examples is defined as Eq. 3, where N is the number of labeled examples and $p(y_i)$ is the predicted probability for the true label y_i.

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\log p(y_i), \tag{3}$$

Soft KD: Kullback-Leibler divergence between softened teacher and student logits is defined as Eq.4, where ℓ_t and ℓ_s denote the teacher and student logits respectively, $\sigma(\cdot)$ denotes the softmax function, and T is the temperature controlling softening of logits.

$$\mathcal{L}_{\mathrm{KD}} = T^2 \cdot \mathrm{KL}\big(\sigma(\ell_s/T)\,\|\,\sigma(\ell_t/T)\big), \tag{4}$$

Channel/Relation Alignment (CH): an HSIC-inspired [27] inter-example similarity alignment loss is defined as Eq. 5, where K_U and K_V are the softmax-normalized projected feature matrices of the teacher and student, respectively, and w_{fw} is a similarity weighting factor computed from the alignment of their centered Gram matrices, reflecting how well teacher-student feature

Table 2. Test results of diagnosis step on different models using Step-CoT(%). Entries are reported as $a(+b)$, where a is the performance without Step-CoT and b is the improvement with it. The best results in each column are highlighted in **bold**, and the second-best values are underlined.

Model	Accuracy	mAUC	Sensitivity	Specificity
LVLMs				
LLaVA-Med [20]	42.7	58.3	42.7	79.4
Med-Flamingo [28]	30.1	61.2	28.4	89.8
Visual foundation models				
VisualBERT [24]	56.2(+9.3)	48.8(+14.3)	8.5(+1.6)	89.1(+2.7)
CLIP [31]	64.7(+4.5)	48.8(+3.8)	10.1(+1.9)	87.5(+2.1)
ALBEF [21]	68.1(+3.9)	53.9(+21.2)	16.3(+2.2)	91.5(+1.8)
BLIP [22]	66.4(+4.6)	53.2(+21.7)	15.5(+1.7)	90.8(+2.1)
FLAVA [35]	62.5(+4.6)	50.2(+14.0)	9.3(+1.6)	90.7(+1.6)
biomedclip [46]	69.3(+3.8)	55.6(+20.4)	19.4(+2.3)	91.8(+1.7)
Ours (Teacher)	**78.3**	89.5	**46.0**	**96.6**
Ours (Student)	77.5	**90.0**	41.8	96.0

relations align.

$$\mathcal{L}_{\mathrm{CH}} = w_{\mathrm{fw}} \cdot \mathrm{KL}\big(\log K_V \,\|\, K_U\big), \tag{5}$$

Training Recipe. We train with separate optimizers for teacher and student. The teacher may be optionally pretrained for several epochs with supervised loss only, then both teacher and student are trained: the teacher receives supervised CE updates, and the student is trained to minimize as Eq. 6, where α_{KD} and α_{CH} are weighting coefficients that balance the contributions of soft KD and CH alignment losses.

$$\mathcal{L}^{(s)}_{\mathrm{student}} = \mathcal{L}^{(s)}_{\mathrm{CE}} + \alpha_{\mathrm{KD}}\mathcal{L}^{(s)}_{\mathrm{KD}} + \alpha_{\mathrm{CH}}\mathcal{L}^{(s)}_{\mathrm{CH}}, \tag{6}$$

4. Experiments

4.1. Benchmark Establishment

We establish performance baselines using LVLMs and visual foundation models. Table 2 reveals the results. Multimodal models achieve good results but still exhibit sensitivity–specificity imbalance. Among them, BiomedCLIP [46] attains the best overall performance, suggesting that domain adaptation enhances domain alignment. In contrast, LVLMs demonstrate limited transfer accuracy on this benchmark (30–40%), reflecting the gap between generic multi-modal pretraining and domain-specific reasoning requirements. Despite their impressive open-ended generation ability, these LVLMs often rely on surface-level correlations and struggle to maintain factual precision or structured reasoning consistency across clinical steps. Regarding the Step-CoT setup, this setting specifies whether Step-CoT is enabled, allowing the model to leverage structured stepwise reasoning, or disabled, in which case the model is trained without any intermediate reasoning guidance. The results show that incorporating CoT leads to improvements

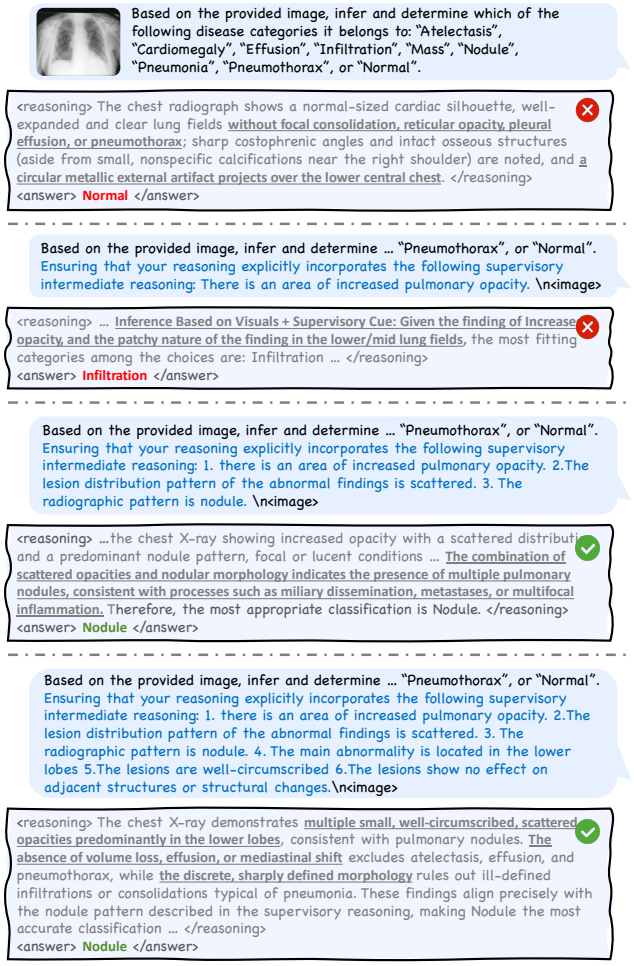


Figure 3. Fine-tuning an LVLM with Step-CoT-based intermediate constraints under verifiable instructions. The testing of different steps is conducted in independent dialogue sessions. The model progressively adjusts its stepwise reasoning, produces coherent intermediate steps, and converges to the correct final diagnosis; this demonstrates that Step-CoT's structured intermediate constraints strengthen model reasoning and reliably guide it to accurate conclusions.

across visual foundation models. The proposed Step-CoT framework introduces stepwise supervision over intermediate reasoning processes, enabling the model to accumulate and verify evidence across diagnostic stages. This design not only improves factual precision but also leads to significant gains in interpretability.

4.2. Stepwise Effectiveness Study of the Step-CoT Dataset

This experiment was designed to verify the effectiveness of the stepwise reasoning framework introduced in the Step-CoT dataset. To this end, we employed the verifiable instruction framework proposed in Instruction-Following

Eval (IFEval) [49] as an objective tool to assess whether introducing explicit step reasoning enables the model to generate more accurate final answers. Specifically, we first tasked Gemini [38] to directly output the diagnosis and corresponding reasoning process. Subsequently, we progressively introduced intermediate reasoning constraints based on Step-CoT, requiring Gemini to base its reasoning on these provided structured constraints. Each reasoning step was governed by corresponding verifiable instructions, thereby enabling an evaluation of whether Step-CoT's intermediate reasoning constraints can enhance the model's reasoning capability and guide it toward correct conclusions.

As shown in Fig. 3, the baseline model often produced ill-defined or incorrect predictions, whereas the model guided by Step-CoT reasoning exhibited coherent intermediate reasoning and correctly localized and classified the lesions. Quantitative evaluation under the same verification framework further demonstrated that the step-supervised model achieved higher accuracy across multiple reasoning categories. These findings confirm that the stepwise reasoning mechanism embedded in Step-CoT effectively enhances model reasoning reliability and diagnostic accuracy, providing empirical evidence for the value of structured, verifiable reasoning in medical visual question answering.

4.3. Cross-Dataset Generalization Evaluation

To evaluate the generalization capability and robustness of models trained on our Step-CoT dataset, we conduct a cross-dataset evaluation. Following the training on Step-CoT, models are directly evaluated on the ChestX-ray8 dataset [40] without any further fine-tuning. This rigorous assessment aims to verify whether the multi-step reasoning skills acquired from Step-CoT can transfer effectively to a different clinical benchmark. The ChestX-ray8 [40], released by the U.S. National Institutes of Health (NIH), is a large-scale chest X-ray image database containing over 100k images from tens of thousands of patients. It is annotated with common thoracic diseases such as atelectasis, cardiomegaly, and effusion. In this study, we utilized the single-label classification subsets defined in the original dataset as our research subjects. It contains 985 single-labeled samples belonging to the following categories: atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, and pneumothorax.

The results in Table 3 demonstrate that networks trained with Step-CoT supervision consistently outperform their non-step counterparts when transferred to the ChestX-ray8 benchmark. The results indicate that the Step-CoT training regime improves both discrimination and clinical relevance after cross-dataset transfer. The fact that both the distilled Student and the Teacher, each trained under the Step-CoT paradigm, perform competitively on ChestX-ray8 suggests that the stepwise knowledge is transferable through distillation and that the learned stepwise reasoning is robust to dataset shift. These results support that Step-CoT enables the model to learn structured, stepwise diagnostic logic rather than relying on end-to-end correlations.

Table 3. Transfer performance comparison with and without using Step-CoT on a different test dataset (%). Entries are reported as $a(+b)$, where a is the performance without Step-CoT and b is the improvement with it. The best value in each column is **bold** and the second best is underlined.

Model	Accuracy	mAUC	Sensitivity	Specificity
VisualBERT [24]	25.7(+6.3)	52.1(+4.3)	30.8(+4.7)	73.9(+3.9)
CLIP [31]	26.8(+7.3)	53.3(+4.6)	18.4(+3.3)	79.7(+3.7)
ALBEF [21]	30.4(+5.2)	58.3(+3.1)	38.4(+3.5)	81.5(+2.9)
BLIP [22]	27.5(+6.1)	52.7(+3.7)	17.9(+2.9)	78.2(+3.5)
FLAVA [35]	28.9(+5.5)	56.8(+2.9)	35.5(+4.1)	77.3(+3.0)
biomedclip [46]	33.6(+5.8)	**61.9(+3.6)**	**42.7(+3.9)**	**85.6(+2.6)**
Ours (Teacher)	**40.3**	61.3	20.2	86.1
Ours (Student)	39.1	57.4	23.5	80.7

Table 4. Comparison of memory ablation results across four of seven reasoning steps (%). The best results in each column are highlighted in **bold**, and the second-best values are underlined.

	Detection	Distribution	Location	Diagnosis
w/o Memory	73.7	69.6	63.2	65.5
w/o Text	81.5	76.1	69.3	72.1
Ours (Teacher)	**91.8**	**84.6**	**77.1**	**78.3**
Ours (Student)	**91.8**	83.4	76.9	77.5

4.4. Ablation Experiments

We evaluate the proposed GAT-based teacher model with knowledge distillation against established baselines.

Comparative Performance Analysis. To evaluate the proposed GAT-Memory framework, we compare it with two ablated variants. The *w/o Memory* variant removes the memory module and GRU update, disabling cross-step information accumulation and thus breaking the continuity of multi-step reasoning. The *w/o Text* variant omits textual prompts, relying only on visual features. As reported in Table 4, both ablations cause consistent performance drops across steps: removing the memory produces the largest accuracy decline (65.45%), highlighting the necessity of temporal state propagation for synthesizing intermediate evidence, while excluding textual prompts also reduces performance (72.10%), confirming the role of linguistic guidance in grounding visual interpretation. The full Teacher model attains the highest accuracy (78.26%), and the distilled Student closely matches this performance (77.53%) with reduced complexity. Taken together, these findings highlight key insights: (i) the memory module is indispensable for maintaining contextual reasoning across diagnostic steps, enabling the model to "think with images" in a clinically coherent manner; (ii) the student model effectively inherits the teacher's structured diagnostic logic while achieving

Table 5. Effect of step ordering (Sequential vs. Randomized) (%). The best results in each column are highlighted in **bold**, and the second-best values are underlined.

Step setting	Model	Accuracy	mAUC	Sensitivity	Specificity
Random (train & test)	Teacher	61.2	72.1	29.4	82.7
	Student	59.0	70.3	27.1	80.9
Random (only test)	Teacher	69.1	82.4	33.8	90.4
	Student	72.8	84.9	39.2	93.1
Step-CoT	Teacher	**78.3**	89.5	**46.0**	**96.6**
	Student	77.5	**90.0**	41.8	96.0

Table 6. Step-CoT vs. GPT free-form CoT (%). The best results in each column are highlighted in **bold**, and the second-best values are underlined.

Model	Accuracy	Sensitivity	Specificity	F1-Score
Teacher (free)	62.1	18.7	78.9	22.5
Student (free)	59.8	20.4	75.6	23.0
Teacher (Step-CoT)	**79.8**	**41.6**	**95.1**	**41.3**
Student (Step-CoT)	68.5	35.9	93.9	34.6

lightweight, transferable inference.

Validation of Structured Supervision and Sequential Dependencies in Step-CoT. We establish a visual CoT framework to probe the source of Step-CoT's performance gains from both knowledge transfer and structural necessity perspectives. Specifically, the teacher abstracts step-level dependencies, while the student evaluates their generalizability over dataset patterns. Results in Table 5 and Table 6 show that such stepwise supervision consistently improves performance, suggesting that Step-CoT effectively transfers structured reasoning knowledge. To move beyond correlation and examine causality, we design controlled experiments that explicitly manipulate the sequential structure. By disrupting the step order (Table 6), we isolate sequential logic as the independent variable. The resulting performance degradation demonstrates that Step-CoT's ordered structure is functionally necessary, rather than a superficial prompting artifact. Finally, to rule out the possibility that improvements stem merely from formatting or verbosity, we compare Step-CoT with GPT-generated free-form CoT under the same conditions. Step-CoT consistently outperforms free-form CoT for both teacher and student models. Together with the randomized step-order ablation, these results confirm that the gains originate from its structured, clinically consistent design, rather than incidental characteristics of the prompt.

4.5. Visualization of the Reasoning Steps

Fig. 4 presents the seven stepwise attention maps, which visualize how the model's focus evolves during reasoning. Across the sequence, the attention progressively concentrates from broad, image-level saliency to fine-grained, lesion-specific regions: early steps highlight global abnormality and distribution patterns, middle steps emphasize modality-specific imaging cues and precise lesion localization, and the final steps concentrate on diagnostic features.

This ordered sharpening of attention provides direct, interpretable evidence that the model engages in image-guided, multi-step reasoning rather than a single opaque decision. The maps serve two roles: (i) they demonstrate the model's emerging multi-step capability by showing how intermediate visual evidence is assembled into later diagnostic judgments; (ii) they make the diagnostic chain traceable, enabling qualitative inspection of which image regions and which reasoning stages contribute to the final prediction.

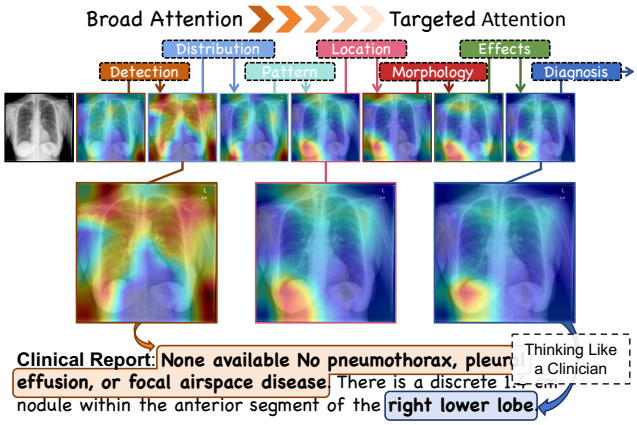


Figure 4. The feature attention visualization across multi-step reasoning demonstrates an evolution from broad attention in the initial query steps to highly targeted attention in the final diagnostic step, reflecting the multi-step capability of Step-CoT and visually verifying the effectiveness of the reasoning chain.

5. Conclusion

In this work, we presented Step-CoT, a large-scale, clinically grounded medical reasoning dataset designed to bring multi-step reasoning into Med-VQA. By explicitly supervising intermediate reasoning with expert-curated diagnostic steps, Step-CoT enables models to think with multi-step, progressively aligning visual attention and linguistic inference to follow clinically valid diagnostic pathways, enhancing interpretability while maintaining diagnostic precision. Building upon this foundation, our teacher-student CoT framework effectively learn from Step-CoT, enhancing the efficiency of multi-step reasoning. Extensive experiments confirm that Step-CoT establishes a structured and credible paradigm for medical reasoning, bridging the gap between human clinical cognition and AI-based decision-making. We expect Step-CoT to serve as a cornerstone resource for developing next-generation, trustworthy medical VQA systems.

References

[1] Faisal Alshanketi, Abdulrahman Alharbi, Mathew Kuruvilla, Vahid Mahzoon, Shams Tabrez Siddiqui, Nadim Rana, and Ali Tahir. Pneumonia detection from chest x-ray images

using deep learning and transfer learning for imbalanced datasets. *Journal of imaging informatics in medicine*, 38(4): 2021–2040, 2025. 4

[2] Taranpal Bansal and Richard Beese. Interpreting a chest x-ray. *British Journal of Hospital Medicine*, 80(5):C75–C79, 2019. 3

[3] Wenrui Cai, Chengyu Wang, Junbing Yan, Jun Huang, and Xiangzhong Fang. Reasoning with omnithought: A large cot dataset with verbosity and cognitive difficulty annotations. *arXiv preprint arXiv:2505.10937*, 2025. 1, 2

[4] Daniel C Castro, Aurelia Bustos, Shruthi Bannur, Stephanie L Hyland, Kenza Bouzid, Maria Teodora Wetscherek, Maria Dolores Sánchez-Valverde, Lara Jaques-Pérez, Lourdes Pérez-Rodríguez, Kenji Takeda, et al. Padchest-gr: A bilingual chest x-ray dataset for grounded radiology report generation. *arXiv preprint arXiv:2411.05085*, 2024. 2, 3

[5] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025. 1

[6] Pengyu Dai, Yafei Ou, Yuqiao Yang, Dichao Liu, Masahiro Hashimoto, Masahiro Jinzaki, Mototaka Miyake, and Kenji Suzuki. Sasamim: synthetic anatomical semantics-aware masked image modeling for colon tumor segmentation in non-contrast abdominal computed tomography. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 567–578. Springer, 2024. 1

[7] Pengyu Dai, Yafei Ou, Yuqiao Yang, Ze Jin, and Kenji Suzuki. Goca: Trustworthy multi-modal rag with explicit thinking distillation for reliable decision-making in med-lvlms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 251–261. Springer, 2025. 1

[8] Gerald De Lacey, Simon Morley, and Laurence Berman. *The chest X-ray: a survival guide*. Elsevier Health Sciences, 2012. 3

[9] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016. 2, 3

[10] Frank C Detterbeck, Peter J Mazzone, David P Naidich, and Peter B Bach. Screening for lung cancer: diagnosis and management of lung cancer: American college of chest physicians evidence-based clinical practice guidelines. *Chest*, 143 (5):e78S–e92S, 2013. 4

[11] Chao Ding, Mouxiao Bian, Pengcheng Chen, Hongliang Zhang, Tianbin Li, Lihao Liu, Jiayuan Chen, Zhuoran Li, Yabei Zhong, Yongqi Liu, et al. Building a human-verified clinical reasoning dataset via a human llm hybrid pipeline for trustworthy medical ai. *arXiv preprint arXiv:2505.06912*, 2025. 1, 2

[12] Lin Fan, Xun Gong, Cenyang Zheng, and Yafei Ou. Tri-vqa: triangular reasoning medical visual question answering for multi-attribute analysis. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1485–1488. IEEE, 2024. 1

[13] Lin Fan, Xun Gong, Cenyang Zheng, Xuli Tan, Jiao Li, and Yafei Ou. Cycle-vqa: A cycle-consistent framework for robust medical visual question answering. *Pattern Recognition*, 165:111609, 2025.

[14] Lin Fan, Pengyu Dai, Zhipeng Deng, Haolin Wang, Xun Gong, Yefeng Zheng, and Yafei Ou. Evolving medical imaging agents via experience-driven self-skill discovery. *arXiv preprint arXiv:2603.05860*, 2026. 1

[15] Xiaotang Gai, Chenyi Zhou, Jiaxiang Liu, Yang Feng, Jian Wu, and Zuozhu Liu. Medthink: A rationale-guided framework for explaining medical visual question answering. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7438–7450, 2025. 1, 2

[16] Haifan Gong, Guanqi Chen, Mingzhi Mao, Zhen Li, and Guanbin Li. Vqamix: Conditional triplet mixup for medical visual question answering. *IEEE Transactions on Medical Imaging*, 41(11):3332–3343, 2022. 1

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 12

[18] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 12

[19] Xiaoshuang Huang, Haifeng Huang, Lingdong Shen, Yehui Yang, Fangxin Shang, Junwei Liu, and Jia Liu. A refer-and-ground multimodal large language model for biomedicine. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 399–409. Springer, 2024. 1, 2

[20] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023. 1, 2, 6

[21] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 6, 7, 8

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 6, 7, 8, 12

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2

[24] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and perfor-

mant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 6, 7, 8, 12

[25] Bo Liu, Ke Zou, Li-Ming Zhan, Zexin Lu, Xiaoyu Dong, Yidi Chen, Chengqiang Xie, Jiannong Cao, Xiao-Ming Wu, and Huazhu Fu. Gemex: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21310–21320, 2025. 1

[26] Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Tianyi Zhou, and Zuozhu Liu. Medcot: Medical chain of thought via hierarchical expert. *arXiv preprint arXiv:2412.13736*, 2024. 1, 2

[27] Wan-Duo Kurt Ma, JP Lewis, and W Bastiaan Kleijn. The hsic bottleneck: Deep learning without back-propagation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 5085–5092, 2020. 5

[28] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023. 6

[29] Duy MH Nguyen, Nghiem T Diep, Trung Q Nguyen, Hoang-Bao Le, Tai Nguyen, Tien Nguyen, TrungTin Nguyen, Nhat Ho, Pengtao Xie, Roger Wattenhofer, et al. Enriched instruction-following graph alignment for efficient medical vision-language models. 1

[30] Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 337–347. Springer, 2025. 2

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 6, 7, 8, 12

[32] Sarah Sandmann, Stefan Hegselmann, Michael Fujarski, Lucas Bickmann, Benjamin Wild, Roland Eils, and Julian Varghese. Benchmark evaluation of deepseek large language models in clinical decision-making. *Nature Medicine*, pages 1–1, 2025. 3, 8

[33] James A Scott. The lung scan and the abnormal chest x-ray: difficult diagnoses. *Nuclear medicine communications*, 25 (11):1137–1141, 2004. 3, 4

[34] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025. 1

[35] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15638–15650, 2022. 6, 7, 8, 12

[36] Yu Sun, Xingyu Qian, Weiwen Xu, Hao Zhang, Chenghao Xiao, Long Li, Yu Rong, Wenbing Huang, Qifeng Bai, and Tingyang Xu. Reasonmed: A 370k multi-agent generated dataset for advancing medical reasoning. *arXiv preprint arXiv:2506.09513*, 2025. 1

[37] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 12

[38] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 7

[39] Mickael Tordjman, Zelong Liu, Murat Yuce, Valentin Fauveau, Yunhao Mei, Jerome Hadjadj, Ian Bolger, Haidara Almansour, Carolyn Horst, Ashwin Singh Parihar, et al. Comparative benchmarking of the deepseek large language model on medical tasks and clinical reasoning. *Nature medicine*, pages 1–1, 2025. 3, 8

[40] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 7

[41] Yuan Wang, Jiaxiang Liu, Shujian Gao, Bin Feng, Zhihang Tang, Xiaotang Gai, Jian Wu, and Zuozhu Liu. V2t-cot: From vision to text chain-of-thought for medical reasoning and diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 658–668. Springer, 2025. 1, 2

[42] Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, et al. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. *arXiv preprint arXiv:2408.02900*, 2024. 1

[43] Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2087–2098, 2025. 1

[44] Songxiao Yang, Haolin Wang, Yao Fu, Ye Tian, Tamotsu Kamishima, Masayuki Ikebe, Yafei Ou, and Masatoshi Okutomi. Ram-w600: A multi-task wrist dataset and benchmark for rheumatoid arthritis. *Advances in neural information processing systems*, 2025. 1

[45] Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2345–2354, 2020. 1

[46] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023. 6, 7, 8

[47] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023. 2

[48] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1702–1713, 2025. 1

[49] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023. 7

[50] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019. 1