From Perception to Evidence: Toward Reliable Clinical Diagnostic Reasoning with Role-Aware Diagnostic Evidence Chain-of-Thought

Anonymous ACL submission

Abstract

Medical vision-language models have shown strong potential in radiology report generation and medical visual question answering. However, insufficient faithfulness in clinical generation remains a critical bottleneck, as models must reliably distinguish image-supported findings, unsupported inferences, uncertain observations, and conflicting evidence. Existing methods typically use structured diagnosis, retrieval, or cross-modal alignment as auxiliary supervision or output-level correction, lacking explicit modeling and structured control of clinical evidence before generation. To address this, we propose Diagnostic Evidence Chain-of-Thought (**DeCoT**), a diagnostic evidence chain framework for high-faithfulness medical vision-language generation. DeCoT first predicts structured diagnostic anchors from medical images to model observation-level clinical states, including presence, absence, and uncertainty. It then constructs an evidence chain with explicitly defined roles, including primary, supporting, negative, prototype, and suppressed evidence, which guides downstream report generation or question answering. Extensive experiments on three medical VQA and report generation benchmarks show that DeCoT consistently outperforms structured diagnosis, case-based retrieval-augmented generation, and preference optimization baselines in generation faithfulness, reasoning accuracy, and text quality.

1 Introduction

Medical vision-language models, with their powerful cross-modal alignment and generation capabilities, have become the mainstream technical paradigm for clinical intelligence tasks such as radiology report generation and medical visual question answering (Li et al., 2023a; Moor et al., 2023; Tu et al., 2024; Zhang et al., 2024; Wang et al., 2023). However, the lack of faithfulness in clinical generation remains the core bottleneck for their deployment in real-world clinical scenarios (Yu et al., 2023; Miura et al., 2021; Delbrouck et al., 2022; Jain et al., 2021; Smit et al., 2020): models must not only generate semantically fluent text but also strictly distinguish between image-supported objective findings, evidence-free hallucinations, uncertain observations, and conflicting diagnostic conclusions, as any deviation can directly affect clinical judgment and patient safety.

To improve clinical faithfulness, existing research has explored directions such as structured diagnosis (Irvin et al., 2019; Smit et al., 2020; Jain et al., 2021), retrieval augmentation and memory mechanisms (Li et al., 2019; Chen et al., 2020; Endo et al., 2021), preference learning (Rafailov et al., 2023; Banerjee et al., 2024; Liu et al., 2025), and reward optimization (Miura et al., 2021; Delbrouck et al., 2022). However, these methods still have limitations. Structured diagnosis methods typically use diagnostic labels only as auxiliary supervision signals or semantic priors, failing to organize evidence structures before generation. Retrieval and memory methods only provide additional context or implicit feature alignment without explicitly defining the roles and priorities of evidence. Factuality reward and preference optimization methods focus on output-level constraints, lacking fine-grained control over evidence sources, credibility, and conflict relationships before generation. These shortcomings make it difficult for models to fundamentally avoid hallucinations and achieve explainable, traceable, and constrained clinical-grade generation. ***Therefore, ensuring evidence-based reasoning and traceable generation has become a critical clinical requirement, calling for a dedicated evidence governance paradigm.***

To address the above limitations, this paper proposes Diagnostic Evidence Chain-of-Thought (**DeCoT**), a diagnostic evidence chain framework for high-faithfulness medical vision-language generation. Unlike existing methods, DeCoT redefines clinical generation as an evidence organiza-

tion and reasoning problem, achieving full-chain control from visual perception to trustworthy output through three stages. Specifically, the model first predicts structured diagnostic anchors from medical images, explicitly modeling observation-level states such as presence, absence, and uncertainty; these anchors are treated as soft clinical evidence rather than hard decision outputs. Then, based on the diagnostic anchors, it performs evidence retrieval, filtering, and conflict processing to construct a role-based evidence chain comprising primary evidence, supporting evidence, negative evidence, prototype evidence, and suppressed evidence. Finally, under the strong guidance of the evidence chain, it completes report generation or question answering, transforming structured diagnosis from an auxiliary semantic signal into the core of clinical evidence organization. We evaluate DeCoT on radiology report generation and medical visual question answering tasks. Experimental results show that the proposed method achieves consistent and significant improvements over direct fine-tuning, preference optimization, and reward-based baselines. Ablation studies further validate the roles of diagnostic anchors, memory retrieval, evidence role assignment, and sequence-level factuality rewards.

Our contributions are summarized as follows:
❶ We propose DeCoT, a diagnostic evidence chain framework for high-faithfulness medical vision-language generation, redefining clinical generation as an evidence organization and reasoning problem.
❷ We introduce a diagnostic anchor mechanism that models observation-level clinical states as organizable and retrievable soft evidence coordinates, laying the foundation for evidence structuring.
❸ We design a role-based evidence chain construction paradigm, which assigns explicit functional roles to multi-source evidence before generation for the first time, achieving precise evidence control and conflict suppression.
❹ Extensive experiments demonstrate that DeCoT significantly outperforms direct fine-tuning, preference optimization, and reward-based baselines.

2 Related Work

2.1 Structured Diagnosis and Observation-Guided Generation

Recent radiology report generation studies have introduced structured clinical semantics as intermediate guidance beyond end-to-end image-to-text mapping. ORGAN (Hou et al., 2023) and DCRRG (Zhou and Wang, 2024) guide generation via observation plan prediction and observation-level decomposition, while DCL (Li et al., 2023b), GMoD (Xiang et al., 2024), and KARGEN (Li et al., 2024) exploit disease labels, graph priors, and region-level clinical features to improve clinical consistency. Nevertheless, these methods mainly use diagnostic information as auxiliary labels or feature priors, without explicitly modeling the functional roles of clinical evidence. DeCoT instead treats diagnostic anchors as the basis for evidence chain construction, turning structured diagnosis into a pre-generation mechanism for clinical evidence.

2.2 Retrieval-Augmented and Memory-Based Clinical Generation

Retrieval and memory mechanisms have been widely used to improve factual grounding in clinical generation. KERP (Li et al., 2019) retrieves abnormality-related text templates for case-specific report generation, while R2Gen (Chen et al., 2020) and R2GenCMN (Chen et al., 2021) introduce memory-driven and cross-modal memory mechanisms to capture key decoding information and image-text alignment. Recent retrieval-augmented methods further retrieve relevant radiology texts or factual reports to guide large multimodal or language models, such as GPT-based retrieval-augmented report generation (Ranjit et al., 2023), FactMM-RAG (Sun et al., 2025), and RADAR (Hou et al., 2025). However, retrieved information is usually used as additional context, reference reports, or latent memory, without explicit role assignment during generation. DeCoT instead organizes evidence into primary, supporting, negative, prototype, and suppressed roles before generation.

2.3 Factuality Alignment and Preference Optimization

Output-level reward and preference optimization have been explored to improve clinical faithfulness. Semantic reward methods use structured annotations, such as RadGraph entities and relations (Jain et al., 2021), to promote factual consistency (Miura et al., 2021; Delbrouck et al., 2022). Recent studies further adopt clinical factuality or quality metrics as reinforcement signals (Zhou et al., 2024b), while DPO (Rafailov et al., 2023) and its medical or multimodal variants (Wang et al., 2024; Zhu et al., 2025) align models with pairwise preferences. Self-

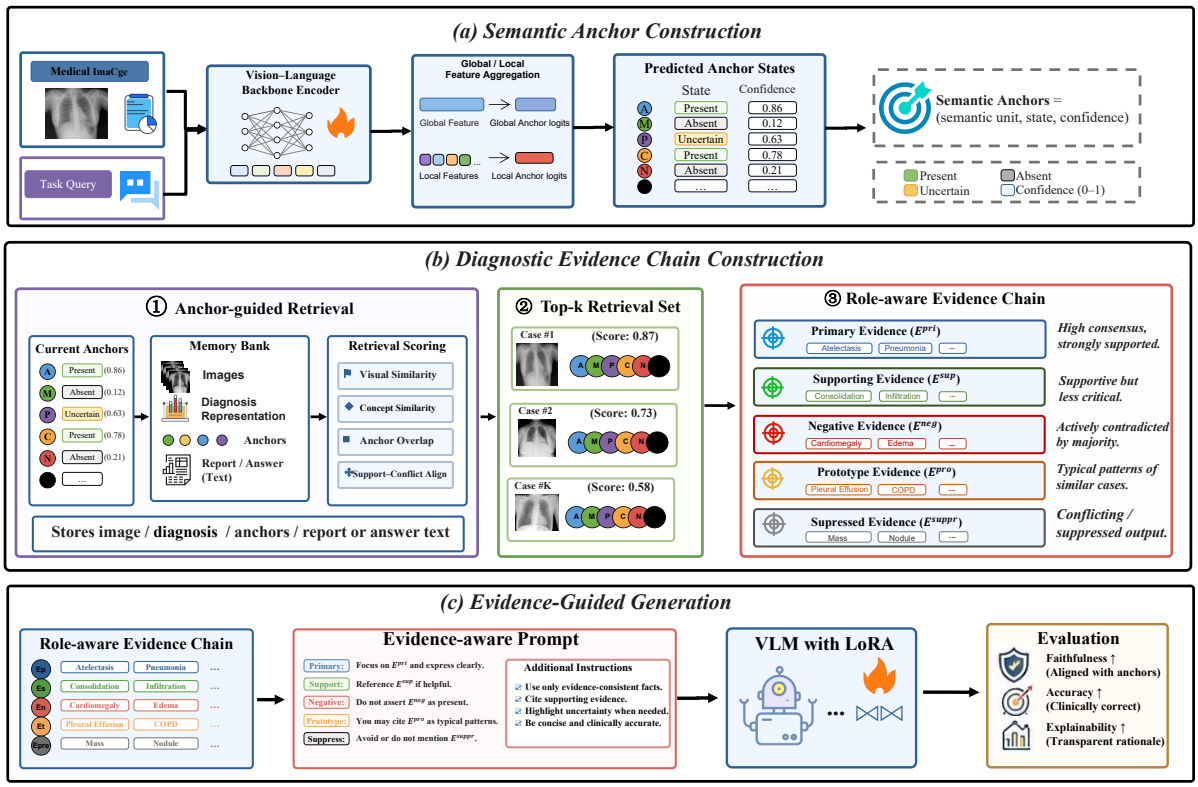

Figure 1: Overview of DeCoT. In Stage 1, global and local visual-language features are aggregated to predict semantic diagnostic anchors that capture observation-level clinical states. In Stage 2, the predicted anchors are used to retrieve relevant memory cases and construct a role-aware evidence chain, including primary, supporting, negative, prototype, and suppressed evidence. In Stage 3, the evidence chain is linearized into an evidence-aware prompt to guide the vision-language backbone toward faithful report generation or medical visual question answering.

alignment methods also generate preference data for radiology report generation (Xiao et al., 2025). Unlike these output-level approaches, DeCoT organizes diagnostic evidence before generation and guides the model through explicit evidence roles.å

3 Method

3.1 Stage 1: Semantic Anchor Construction

Many existing medical vision-language models perform end-to-end mapping directly from images and task inputs, without explicitly modeling observation-level clinical states (Li et al., 2023a; Moor et al., 2023; Tu et al., 2024; Wang et al., 2023).To address this limitation, we introduce a semantic anchor layer that maps raw inputs into structured, task-relevant semantic unit states, providing reliable "soft coordinates" for subsequent evidence organization. Specifically, given a medical image x and a task input u, we first use a vision-language backbone to encode both and obtain a multimodal hidden state sequence:

$$H = \mathcal{F}_\theta(x, u), \quad H \in \mathbb{R}^{B \times L \times d_h} \tag{1}$$

where $\mathcal{F}_\theta(\cdot)$ is a parameterized encoder, B is the batch size, L is the sequence length, and d_h is the hidden dimension. Furthermore, we apply average pooling over the valid (non-padding) tokens in H to obtain a global representation:

$$h_{\text{global}} = \mathcal{F}_{\text{Pool}}(H; m) \in \mathbb{R}^{B \times d_h} \tag{2}$$

where $m \in \{0, 1\}^{B \times L}$, with 0 indicating a padding token and 1 indicating a real token. We then predefine a set of task-relevant semantic units $\{o_i\}_{i=1}^{N_u}$ and assign a learnable query vector q_i to each o_i. For each unit, the model first extracts the most relevant features from H via an attention mechanism $\mathcal{F}_{\text{Attn}}(\cdot)$, and then fuses global and local information through a gating mechanism $\mathcal{G}_\phi(\cdot)$ parameterized by ϕ:

$$p_\phi(s \mid o_i, x, u) = \sigma\big(\mathcal{G}_\phi(h_{\text{global}}, \mathcal{F}_{\text{Attn}}(q_i, H))\big) \tag{3}$$

where σ denotes the sigmoid function. $s \in \mathcal{S}_u$ represents a specific state of semantic unit o_i, and $\mathcal{S}_u$ denotes task-specific state spaces: for report generation, o_i corresponds to a clinical observation

and $\mathcal{S}_u = \{s_{pre}, s_{abs}, s_{un}\}$, where s_{pre} indicates that the observation is supported by imaging evidence as present, s_{abs} indicates that the observation is supported by imaging evidence as absent, and s_{un} indicates that the current imaging evidence is insufficient to make a definitive judgment or only supports an uncertain judgment; for medical VQA, o_i corresponds to a visual-answer clue under the question condition, and $\mathcal{S}_u$ represents task-specific answer or evidence states.

Finally, the set of semantic anchors z is defined as a set of triples containing each semantic unit, its predicted state, and its confidence:

$$z = \big\{(o_i, s_i, \rho_i) \mid s_i = \arg\max_{s \in \mathcal{S}_u} p_\phi(s \mid o_i, x, u), \\ \rho_i = \max_{s \in \mathcal{S}_u} p_\phi(s \mid o_i, x, u)\big\}_{i=1}^{N_u} \tag{4}$$

where ρ_i is the confidence of the selected state. In both tasks, z serves as the semantic anchor layer for subsequent evidence retrieval, conflict filtering, and evidence-guided generation.

3.2 Stage 2: Evidence Chain Construction

Obtaining semantic judgments alone is not sufficient to determine how to generate the final output; thus we need to further organize the semantic anchors into a controllable evidence chain. Since the local information of a single sample cannot stably provide the priors, constraints, and referential expressions needed for generation, we introduce a memory bank $\mathcal{M}$ constructed from training samples, storing image / diagnostic representations, semantic anchors, report / answer texts, etc.

Given an anchor set z and a training memory bank $\mathcal{M}$, we first retrieve relevant memory items using a weighted score. For simplicity, we formulate the retrieval signals from different sources as a weighted fusion:

$$\begin{aligned} r(q, m_j) = \lambda_v s_v(q, m_j) + \lambda_c s_c(q, m_j) \\ + \lambda_o s_o(z, m_j) + \lambda_a s_a(z, m_j) \end{aligned} \tag{5}$$

where $q = (x, u, z)$, m_j is the j-th memory item in the memory bank, and $r(q, m_j)$ is the relevance score between the query q and memory item m_j. The terms $s_v(q, m_j)$, $s_c(q, m_j)$, $s_o(z, m_j)$, and $s_a(z, m_j)$ denote visual similarity, concept similarity, anchor overlap, and support/conflict alignment between the current input and the data in the memory bank, respectively; $\lambda_v, \lambda_c, \lambda_o, \lambda_a$ are corresponding learnable weights. Visual similarity s_v and concept similarity s_c depend on the full query $q = (x, u, z)$, while anchor overlap s_o and alignment s_a are primarily computed based on the semantic anchors z.

Subsequently, the top-K highest-ranked memory items form the retrieval set:

$$\mathcal{R}_k(q) = \mathrm{TopK}_{m_j \in \mathcal{M}} \; r(q, m_j) \tag{6}$$

Next, we compute the consistency frequency, conflict intensity, and consensus ratio between these retrieved items and the current anchors. For report generation, comparisons are mainly carried out around structured attributes such as observation, status, and anatomy, combined with diagnostic logits, retrieval similarity, and cross-sample consensus ratios to determine support or conflict relations; for VQA, we compare question types, content types, answer families, and visual clues from similar Q&A pairs.

Based on the above relations, DeCoT constructs the following functional evidence set:

$$\mathbb{E} = \{\mathbb{E}^{\text{pri}}, \mathbb{E}^{\text{sup}}, \mathbb{E}^{\text{neg}}, \mathbb{E}^{\text{pro}}, \mathbb{E}^{\text{suppr}}\} \tag{7}$$

where $\mathbb{E}^{\text{pri}}$ is the primary evidence from the current anchors; $\mathbb{E}^{\text{sup}}$ reinforces the main judgment with consistent memory evidence; $\mathbb{E}^{\text{neg}}$ helps exclude unsupported findings or answers; $\mathbb{E}^{\text{pro}}$ provides representative expression patterns; $\mathbb{E}^{\text{suppr}}$ marks conflict-dominant evidence that should not be directly expressed. For report generation, these roles apply to clinical findings and diagnostic conclusions; for VQA, they apply to question-conditioned visual-answer clues and retrieved Q&A examples.

3.3 Stage 3: Evidence-Guided Generation

This stage converts the role-aware evidence chain into an evidence-aware prompt to constrain the final text generation. Given a user instruction u and the functional evidence set $\mathbb{E}$, we first construct a prompt via a formatting function:

$$P_{\mathbb{E}} = \mathrm{Format}(\mathbb{E}, u) \tag{8}$$

where $\mathrm{Format}(\cdot)$ organizes the evidence chain according to evidence roles: it first assigns primary evidence, supporting evidence, negative evidence, prototype evidence, and suppressed evidence into separate prompt sections; then converts the observation, anatomical location, status, and reliability information of each evidence item into concise natural language descriptions; and finally attaches role-specific generation constraints to different evidence roles, enabling the model to distinguish

Method	IU-Xray				SLAKE		VQA-RAD	
	BLEU-Avg ↑	Rouge-L ↑	Meteor ↑	Cider ↑	Open ↑	Closed ↑	Open ↑	Closed ↑
Qwen2.5-VL-3B-Instruct	13.60	23.71	16.18	1.78	33.01	76.66	13.97	68.01
+DPO	27.57	38.70	19.49	38.77	71.02	63.93	23.46	71.32
+mDPO	23.32	36.05	17.63	29.30	71.98	71.09	25.14	59.93
+MMedPO	22.59	36.28	17.63	28.35	71.02	69.50	**27.93**	60.29
+LM-RRG	27.83	36.88	18.78	**42.59**	76.78	85.41	25.14	72.79
+POVID	27.84	38.68	19.59	39.35	76.20	85.15	20.11	74.03
+SIMA	27.78	38.63	19.37	36.64	65.26	71.88	24.02	65.44
+FiSAO	27.80	38.55	19.58	38.62	69.48	71.09	24.58	67.28
+DeCOT (Ours)	**28.56**	**39.75**	**21.01**	36.90	**78.12**	**87.00**	25.24	**74.26**
Direct SFT	25.86	35.59	17.94	**34.97**	73.13	82.49	26.26	73.53
+DPO	13.45	32.18	12.15	33.97	80.42	87.00	26.26	66.54
+mDPO	18.42	32.55	13.76	34.90	81.19	88.06	24.58	68.38
+MMedPO	13.69	32.17	12.23	33.99	80.23	87.80	25.70	68.38
+LM-RRG	14.82	32.30	12.58	34.90	79.46	86.21	25.14	**76.47**
+POVID	13.45	32.18	12.15	33.98	79.46	87.27	23.46	73.53
+SIMA	13.44	32.16	12.15	33.91	80.99	86.74	24.02	68.38
+FiSAO	13.46	32.28	12.16	33.95	80.23	87.02	26.26	67.65
+DeCOT (Ours)	**27.92**	**35.86**	**20.24**	32.53	**81.85**	**88.53**	**26.82**	75.00

Table 1: **Performance comparison** on medical report generation and medical VQA tasks. SFT indicates that the model is first supervised fine-tuned before applying the corresponding preference optimization method. The best results are highlighted in bold, and the second-best results are underlined.

which content should be prioritized for expression, which should serve only as auxiliary reference, which negative findings should be used to generate rule-out statements, and which conflicting information should not be directly generated.The final output is generated by the same vision-language backbone:

$$y = \mathrm{VLM}_\theta(x, u, P_{\mathbb{E}}) \tag{9}$$

where VLM_θ denotes the VLM with trainable LoRA adapters.

4 Experiments

4.1 Experiment Setup

Datasets. To verify the effectiveness of DeCoT in improving clinical faithfulness, we evaluate our method on two VQA datasets, SLAKE (Liu et al., 2021) and VQA-RAD (Lau et al., 2018), and one report generation dataset, IU-Xray (Demner-Fushman et al., 2016).

Implementation Details. We adopt Qwen2.5-VL-3B-Instruct as the target model and apply LoRA fine-tuning (Hu et al., 2022) with $r = 16$, $\alpha = 32$, and a dropout rate of 0.05. The model is trained with a batch size of 1, gradient accumulation steps of 4, a learning rate of 2×10^{-5}, weight decay of 1×10^{-4}, and a gradient clipping norm of 1.0 for 10 epochs. For DeCoT, we retrieve $K = 4$ memory items from the memory bank with a candidate pool of 12, and retain at most 4 evidence items per role in the evidence-guided prompt. The maximum input length is 768 tokens, and the maximum output lengths are set to 160 tokens for report generation and 80 tokens for VQA. All experiments are conducted on 8 NVIDIA GeForce RTX 4090 GPUs with a random seed of 7.

Baselines. We compare DeCoT with two groups of baselines. First, DeCoT is compared with three categories of alignment methods: general preference optimization methods (DPO (Rafailov et al., 2023), POVID (Zhou et al., 2024a), SIMA (Wang et al., 2025), FiSAO (Cui et al., 2025)), medical-specific preference optimization methods (mDPO (Wang et al., 2024), MMedPO (Zhu et al., 2025)), and the sequence-level reward optimization method (LM-RRG (Zhou et al., 2024b)). Second, we compare DeCoT with structured diagnosis prompting, case-based RAG, and diagnosis-aware RAG. For all baselines, DeCoT uses Qwen2.5-VL-3B-Instruct as the backbone model and is evaluated under two initialization settings: zero-shot and direct supervised fine-tuning (SFT).

Evaluation Metrics. For the report generation task, we report standard text generation metrics, including BLEU-Avg (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and CIDEr (Vedantam et al., 2015), where

Method	Variant	IU-Xray				SLAKE		VQA-RAD	
		BLEU-Avg ↑	Rouge-L ↑	Meteor ↑	Cider ↑	Open ↑	Closed ↑	Open ↑	Closed ↑
Qwen	Anchors	9.45	18.78	8.37	2.00	36.19	74.27	15.08	66.91
	Retrieval	16.40	26.08	17.82	3.94	63.43	75.07	18.44	68.01
	A+R	15.35	23.98	16.93	4.04	58.40	74.27	17.32	67.28
	DeCoT	**28.56**	**39.75**	**21.01**	**36.90**	**78.12**	**87.00**	**25.24**	**74.26**
Qwen + SFT	Anchors	26.84	35.30	18.19	**35.25**	74.44	86.21	25.70	73.90
	Retrieval	20.25	35.23	15.67	22.44	75.19	86.47	24.02	71.69
	A+R	19.78	33.28	14.35	15.37	76.12	85.15	22.91	71.69
	DeCoT	**27.92**	**35.86**	**20.24**	32.53	**81.85**	**88.53**	**26.82**	**75.00**

Table 2: **Comparison** of different evidence utilization strategies under the same Qwen backbone. Anchors denote structured diagnostic evidence extracted from model predictions; Retrieval denotes case-based RAG with non-oracle retrieved cases or QA examples; Anchors+Retrieval combines structured diagnostic evidence with retrieved cases. DeCoT organizes these signals into a role-aware evidence chain with conflict suppression.

BLEU-Avg denotes the average of BLEU-1/2/3/4. For the medical VQA task, we report normalized exact-match accuracy separately for open-ended and closed-ended questions, respectively.

4.2 Main Results

Performance Comparison. Table 1 reports the results on IU-Xray, SLAKE, and VQA-RAD. Under the zero-shot setting, DeCoT achieves the best BLEU-Avg, Rouge-L, and Meteor scores on IU-Xray, reaching 28.56, 39.75, and 21.01, with gains of 0.72, 1.05, and 1.42 points over the strongest baselines, respectively. Although LM-RRG obtains the highest CIDEr score, DeCoT remains competitive with 36.90. On SLAKE, DeCoT achieves the best open- and closed-ended accuracies, improving over the strongest baselines by 1.34 and 1.59 points. On VQA-RAD, DeCoT obtains the best closed-ended accuracy of 74.26 and the second-best open-ended accuracy of 25.24. Under the SFT setting, DeCoT again achieves the best BLEU-Avg, Rouge-L, and Meteor scores on IU-Xray, with gains of 2.06, 0.27, and 2.30 points over direct SFT, showing that evidence-chain construction remains beneficial after supervised fine-tuning. It also performs competitively in CIDEr, ranking second to direct SFT. On SLAKE, DeCoT obtains the highest open- and closed-ended accuracies, exceeding the strongest baselines by 0.66 and 0.47 points. On VQA-RAD, DeCoT achieves the best open-ended accuracy of 26.82 and a competitive closed-ended accuracy of 75.00, second only to LM-RRG.

Evidence Strategy. Table 2 shows that DeCoT consistently outperforms structured anchor prompting, retrieval augmentation, and their direct concatenation. Under the zero-shot Qwen setting, DeCoT achieves substantial gains over the best non-DeCoT variant on IU-Xray, improving BLEU-Avg, Rouge-L, Meteor, and CIDEr by 12.16, 13.67, 3.19, and 32.86 points, respectively. It also improves SLAKE open/closed accuracy by 14.69/11.93 points and VQA-RAD open/closed accuracy by 6.80/6.25 points. Notably, simply combining anchors and retrieval does not yield stable improvements. Anchors+Retrieval even underperforms Retrieval alone in some cases, such as SLAKE and VQA-RAD open-ended accuracy, suggesting that naive evidence aggregation may introduce redundancy, conflict, or unclear evidence priorities. Under Qwen+SFT, DeCoT still achieves the best overall performance, especially on VQA tasks, improving over the strongest non-DeCoT variant by 5.73/2.06 points on SLAKE and 1.12/1.10 points on VQA-RAD. These results demonstrate that DeCoT's advantage comes from role-aware evidence organization rather than simply adding anchors or retrieval examples.

5 Analysis

5.1 Templete Collapse on IU-Xray

We further analyze the performance degradation of several SFT-based preference optimization methods on IU-Xray. Table 3 shows that several SFT-based alignment methods suffer from severe template collapse on the IU-Xray test set. DPO, POVID, SIMA, and FiSAO generate exactly one unique report for all 590 test samples, leading to a top-repeat ratio of 100%. MMedPO and LM-RRG also exhibit strong collapse, with top-repeat ratios of 98.64% and 94.41%, respectively. In contrast, SFT+DeCoT substantially alleviates this collapse, reducing the top-repeat ratio to 53.56% while

Method	Unique Reports ↑	Top-1 Ratio ↓	Avg. Len.	BLEU-Avg ↑	METEOR ↑	CIDEr ↑
SFT	3	78.14	23.56	25.86	17.94	**34.97**
SFT+DPO	1	100.00	14.00	13.45	12.15	33.97
SFT+mDPO	**42**	78.47	18.42	18.42	13.76	34.90
SFT+MMedPO	3	98.64	14.17	13.69	12.23	33.99
SFT+LM-RRG	10	94.41	15.07	14.82	12.58	34.90
SFT+POVID	1	100.00	14.00	13.45	12.15	33.97
SFT+SIMA	1	100.00	14.00	13.45	12.15	33.97
SFT+FiSAO	1	100.00	14.00	13.45	12.15	33.97
SFT+**DeCoT**	28	**53.56**	**28.36**	**27.92**	**20.24**	32.53

Table 3: **Template collapse analysis** on the IU-Xray test set. "Unique Reports" denotes the number of distinct generated reports among the 590 test samples. "Top-1 Ratio" denotes the proportion of the most frequent generated report. "Avg. Len." denotes the average token length of generated reports.

Models	SLAKE	
	Open ↑	Closed ↑
DeCoT w/o SFT	78.12	87.00
w/o Evidence Roles	74.09	82.49
w/o Retrieval	76.39	86.47
w/o Negative Evidence	71.98	81.17
w/o Conflict Suppression	71.40	81.43
w/o Prototype	75.62	83.02

Table 4: **Ablation study** on SLAKE. We remove one component from DeCoT at a time to evaluate its contribution, including evidence roles, retrieval, negative evidence, conflict suppression, and prototype-based evidence modeling.

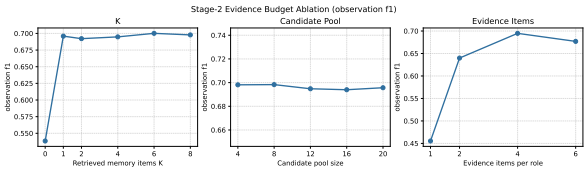

Figure 2: Stage 2 **evidence-budget** ablation curves on the IU X-Ray validation set. The figure shows the effect of the number of retrieved memory items, candidate pool size, and retained evidence items per role on *observation_f1*.

achieving the best BLEU-Avg, Rouge-L, and Meteor scores. This phenomenon suggests that, in low-resource medical report generation such as IU-Xray, preference optimization after SFT tends to amplify high-frequency normal templates, leading the model to produce generic descriptions. In contrast, DeCoT leverages an explicit evidence chain that includes diagnostic anchors, positive and negative evidence, and conflicting information to constrain generation and preference building, reducing reliance on generic templates. As a result, DeCoT preserves sample-specific information and lowers the risk of template collapse.

5.2 Ablation Study

Model Component. Table 4 reports the ablation results on SLAKE. The full DeCoT model achieves the best performance, with 78.12 open-ended accuracy and 87.00 closed-ended accuracy. Removing any component leads to performance degradation, confirming the contribution of each module. Among all variants, removing conflict suppression causes the largest drop in open-ended accuracy, from 78.12 to 71.40, while removing negative evidence also results in a substantial decline, reducing open- and closed-ended accuracy to 71.98 and 81.17, respectively. These results indicate that negative evidence and conflict suppression are particularly important for medical VQA, as they help prevent unsupported findings and reduce the influence of contradictory evidence during reasoning. In contrast, removing retrieval causes the smallest degradation, with accuracies of 76.39 and 86.47. This suggests that, on SLAKE, image-derived structured evidence provides stronger supervision than retrieved external examples, while retrieval still offers complementary information. Evidence roles and prototypes also improve performance by organizing clinical evidence and providing representative patterns, though their removal leads to relatively smaller drops.

Evidence Budget. Figure 2 presents the ablation results of Stage 2 under different evidence budget settings, where *observation_f1* measures the matching degree between the observation items in the generated evidence chain and the reference clinical observations. The results show that removing memory retrieval leads to a significant performance drop; once $K \geq 1$, the performance stabilizes, indicating that a small amount of memory evidence

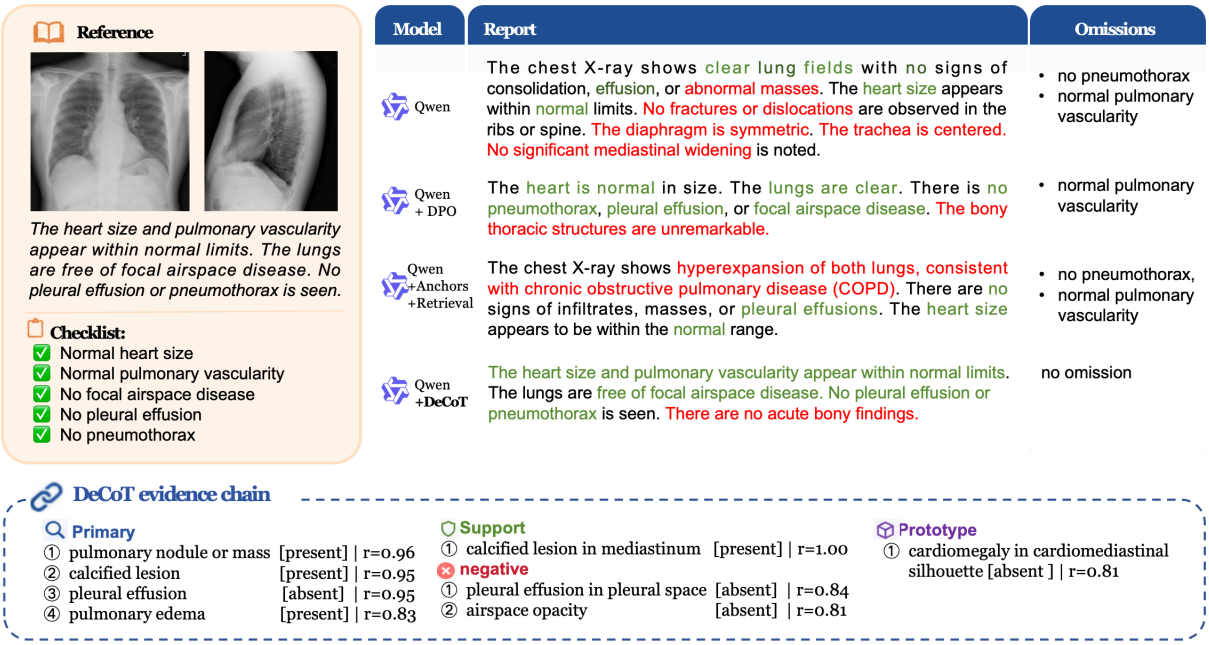

Figure 3: **Case study** on the IU-Xray report generation task. Green highlights denote findings correctly aligned with the reference report, red highlights denote severe hallucinations, and orange highlights denote minor errors such as unsupported normal-template statements or omitted details. Compared with Qwen zero-shot, zero-shot+DPO, and zero-shot+Anchors+Retrieval, DeCoT better preserves the key negative findings and generates the most reference-aligned report.

can effectively complement the Stage 1 semantic anchors. The candidate pool size has a minor impact on performance, suggesting that the retrieval process is insensitive to this parameter. In contrast, the number of evidence items per type has a more pronounced effect: increasing this number from 1 to 4 substantially improves performance, but further increasing it to 6 causes a decline, indicating that excessive evidence may introduce redundancy or noise. Based on the above analysis, we adopt $K = 4$, a candidate pool size of 12, and 4 evidence items per type as the default configuration.

5.3 Qualitative Analysis

As shwon in Fig 3, the reference report describes a normal cardiopulmonary study: normal heart size and pulmonary vascularity, clear lungs, and no pleural effusion or pneumothorax. The Qwen zero-shot model captures some normal findings but adds unsupported template-like details, such as "no masses," "no fractures," "no tracheal deviation," and "no mediastinal widening." DPO improves alignment with the reference and correctly includes the key negative finding of "no pneumothorax," yet still adds an unsupported statement about bony structures. The Anchors+Retrieval model introduces retrieval noise, hallucinating "bilateral hyperexpansion/COPD," and omits "no pneumothorax." In contrast, DeCoT best matches the reference, preserving the essential negative findings, particularly "no focal airspace disease," "no pleural effusion," and "no pneumothorax," although it still adds a minor unsupported bony statement. Overall, this case demonstrates that DeCoT produces the most faithful and interpretable report, whereas retrieval or preference optimization alone are insufficient to fully prevent unsupported findings.

6 Conclusion

We propose DeCoT, a framework for generating diagnostic evidence chains with high fidelity in medical visual language. By predicting structured diagnostic anchors and combining a training memory bank to construct role-based evidence chains, DeCoT can precisely control the sources, credibility, and conflict relationships of evidence before generation. Experimental results on multiple tasks show that DeCoT outperforms the baseline tasks in both radiology report generation and medical visual question answering tasks. These results indicate that DeCoT, as an evidence-driven medical visual language generation framework, can significantly enhance the fidelity of the model to image facts.

Limitations

Despite the promising results of DeCoT, our work has several limitations.First, due to limited computational and storage resources, our experiments are conducted only on commonly used medical report generation and medical VQA benchmarks, including IU-Xray, SLAKE, and VQA-RAD. These datasets are relatively small and may not fully reflect the diversity, noise, and distribution shifts present in real clinical environments. Second, our evaluation primarily follows standard automatic metrics, such as BLEU, Rouge-L, METEOR, CIDEr, and normalized exact-match accuracy. These metrics cannot fully capture clinical correctness, factual consistency, or diagnostic safety. In the future, we plan to incorporate evaluations by expert radiologists and clinically grounded metrics to provide a more comprehensive assessment. Third, DeCoT introduces additional inference-time computational overhead due to evidence extraction, retrieval, and evidence-chain construction. Although this design improves evidence utilization, it may increase deployment costs compared to direct generation methods. We will explore more efficient evidence selection and compression strategies in future work.

Ethical Considerations

This work focuses on improving the faithfulness and traceability of medical vision-language generation. All experiments are conducted on publicly available research datasets and publicly released backbone models. We use these artifacts only for their intended research purposes, including radiology report generation and medical visual question answering evaluation. We do not introduce or release any new patient-level clinical data, and any future released code or derived artifacts will remain subject to the original access conditions and licenses of the corresponding datasets and models. Although DeCoT is designed to reduce unsupported generation by organizing diagnostic evidence before decoding, it is not intended for autonomous clinical diagnosis or direct clinical deployment. The generated reports and answers should be viewed as research outputs and must not replace professional medical judgment. Medical vision-language models may still produce incorrect, incomplete, or biased outputs, especially under distribution shifts, low-quality images, rare diseases, or ambiguous findings. Therefore, any practical use would require rigorous clinical validation, human oversight, and compliance with local medical, legal, and privacy regulations.

We also acknowledge that evidence-chain prompting may expose structured intermediate findings produced by the model. These intermediate outputs are designed for interpretability and error analysis, but they may still contain incorrect or uncertain predictions. Users should avoid interpreting them as verified clinical facts without expert review.

References

Oishi Banerjee, Hong-Yu Zhou, Kay Wu, Subathra Adithan, Stephen Kwak, and Pranav Rajpurkar. 2024. Direct preference optimization for suppressing hallucinated prior exams in radiology report generation. In *Machine Learning for Healthcare Conference*. PMLR.

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*, pages 65–72.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 5904–5914.

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 (EMNLP)*, pages 1439–1449.

Chenhang Cui, An Zhang, Yiyang Zhou, Zhaorun Chen, Gelei Deng, Huaxiu Yao, and Tat-Seng Chua. 2025. Fine-grained verifiers: Preference modeling as next-token prediction in vision-language alignment. In *International Conference on Learning Representations*.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360.

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

Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y Ng, and Pranav Rajpurkar. 2021. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Machine learning for health*, pages 209–219. PMLR.

Wenjun Hou, Yi Cheng, Kaishuai Xu, Heng Li, Yan Hu, Wenjie Li, and Jiang Liu. 2025. Radar: Enhancing radiology report generation with supplementary knowledge injection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26366–26381.

Wenjun Hou, Kaishuai Xu, Yi Cheng, Wenjie Li, and Jiang Liu. 2023. Organ: Observation-guided radiology report generation via tree reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8108–8122.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, and 1 others. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Q. H. Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251.

Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. 2019. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6666–6673.

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

Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. 2023b. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3334–3343.

Yingshu Li, Zhanyu Wang, Yunyi Liu, Lei Wang, Lingqiao Liu, and Luping Zhou. 2024. Kargen: Knowledge-enhanced automated radiology report generation using large language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 382–392. Springer.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE.

Hong Liu, Dong Wei, Zhe Xu, Xian Wu, Yefeng Zheng, and Liansheng Wang. 2025. Rrg-dpo: Direct preference optimization for clinically accurate radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 552–562. Springer.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. 2021. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304.

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

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*, pages 311–318.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Mercy Ranjit, Gopinath Ganapathy, Ranjit Manuel, and Tanuja Ganu. 2023. Retrieval augmented chest x-ray report generation using openai gpt models. In *Machine learning for healthcare conference*, pages 650–666. PMLR.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. 2020. Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In

Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), pages 1500–1519.

Liwen Sun, James Jialun Zhao, Wenjing Han, and Chenyan Xiong. 2025. Fact-aware multimodal retrieval augmentation for accurate medical radiology report generation. 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 643–655.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, and 1 others. 2024. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024. mdpo: Conditional preference optimization for multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8078–8088.

Xiyao Wang, Jiuhai Chen, Zhaoyang Wang, Yuhang Zhou, Yiyang Zhou, Huaxiu Yao, Tianyi Zhou, Tom Goldstein, Parminder Bhatia, Taha Kass-Hout, and 1 others. 2025. Enhancing visual-language modality alignment in large vision language models via self-improvement. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 268–282.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3):100033.

ZhiPeng Xiang, ShaoGuo Cui, CaoZhi Shang, Jingfeng Jiang, and Liqiang Zhang. 2024. Gmod: graph-driven momentum distillation framework with active perception of disease severity for radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 295–305. Springer.

Ting Xiao, Lei Shi, Yang Zhang, HaoFeng Yang, Daisy Zhe Wang, and Chenjia Bai. 2025. Online iterative self-alignment for radiology report generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27799–27814.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, and 1 others. 2023. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9).

Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, and 1 others. 2024. A generalist vision–language foundation model for diverse biomedical tasks. *Nature medicine*, 30(11):3129–3141.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024a. Aligning modalities in vision large language models via preference fine-tuning. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.

Yuanpin Zhou and Huogen Wang. 2024. Divide and conquer radiology report generation via observation level fine-grained pretraining and prompt tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7597–7610.

Zijian Zhou, Miaojing Shi, Meng Wei, Oluwatosin Alabi, Zijie Yue, and Tom Vercauteren. 2024b. Large model driven radiology report generation with clinical quality reinforcement learning. *arXiv preprint arXiv:2403.06728*.

Kangyu Zhu, Peng Xia, Yun Li, Hongtu Zhu, Sheng Wang, and Huaxiu Yao. 2025. MMedPO: Aligning medical vision-language models with clinical-aware multimodal preference optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 80207–80222. PMLR.

A Dataset Details

We evaluate DeCoT on one radiology report generation dataset and two medical visual question answering datasets. For report generation, we use IU-Xray (Demner-Fushman et al., 2016), which contains frontal/lateral chest X-ray studies paired with free-text radiology reports. In our processed manifest, IU-Xray contains 2,069 training studies, 296 validation studies, and 590 test studies, corresponding to 4,138, 592, and 1,180 image-view records, respectively. For medical VQA, we use SLAKE (Liu et al., 2021) and VQA-RAD (Lau et al., 2018). For SLAKE, we keep English questions only and exclude samples labeled as knowledge-based VQA (`base_type=KVQA`) by default, focusing evaluation on visually grounded medical VQA. The resulting SLAKE split contains 4,337 training, 898 validation, and 913 test QA pairs. For VQA-RAD, we use the official train/test split, resulting in 1,793 training and 451 test QA pairs.

B Anchor Definition

Following the definition of semantic anchors z in the main text, each anchor corresponds to an observation-level clinical state rather than a hard diagnostic label. In our implementation, an anchor is instantiated by three components: the observation category o_i, the observation state $s_i \in \{\text{present}, \text{uncertain}, \text{absent}\}$, and the confidence score c_i. Thus, z can be viewed as a set of soft diagnostic coordinates over observation-state pairs. Present anchors indicate findings that may be explicitly described, uncertain anchors mark observations that should be expressed cautiously, and absent anchors provide negative evidence for exclusionary statements.

For the IU-Xray dataset, the anchor space contains 14 chest X-ray observation categories: parenchymal opacity, pleural effusion, atelectasis, cardiomegaly, pneumothorax, support device, nodule or mass, calcification, scar or fibrosis, hyperinflation, pulmonary edema, pleural thickening, fracture, and postsurgical change. These anchors are used as the entry point for constructing role-aware evidence chains: reliable positive or uncertain anchors can be promoted to primary evidence, absent anchors contribute to negative evidence, and conflicting or unreliable anchors can be suppressed before generation.

C Evidence Chain Construction

After obtaining the semantic anchors from Stage 1, DeCoT constructs a role-aware evidence chain to organize the information needed for subsequent generation. Each evidence item is normalized into a unified format that includes the observation category, clinical state, certainty, confidence, source, and its role in the evidence chain. This structured representation enables the generator to distinguish, before producing the final text, between findings directly supported by the image, retrieved supporting evidence, negative evidence, prototype evidence, and suppressed conflicting evidence.

Specifically, high-confidence anchors predicted from the input image are retained as primary evidence, which constitutes the most direct image-based diagnostic cues. DeCoT then retrieves similar cases from the training memory and determines whether the findings in those cases are consistent or in conflict with the current anchors. Findings that are consistently supported by the retrieved cases are assigned as memory-supported evidence; reliable negative findings are assigned as negative evidence and are used to guide the model to generate rule-out statements such as "no pleural effusion" or "no pneumothorax". In addition, recurrent findings summarized from the retrieved cases are treated as prototype evidence, providing a reference pattern without being forced into the final report. If a predicted finding contradicts the retrieval results or lacks sufficient support, it is placed into the suppressed evidence set; without other supporting evidence, the model should avoid directly verbalizing that finding.

Finally, each evidence section is deduplicated and then ranked according to confidence, retrieval consensus, and reliability. The ranked evidence chain is converted into an evidence-guided prompt for Stage 3. In this prompt, primary and memory-supported evidence guide affirmative statements, negative evidence guides rule-out statements, prototype evidence serves as an auxiliary reference, and suppressed evidence acts as a constraint to prevent the model from generating unsupported or conflicting content.

C.1 Evidence-Aware Prompt Template

Report generation template. For radiology report generation, DeCoT converts the structured evidence chain into a natural-language prompt before final decoding. The prompt used in our implemen-

tation is:

Evidence-aware report generation prompt

You are a radiology report writer for chest X-ray studies.
Use the chest X-ray image(s) together with the evidence chain below.
The evidence chain is schema-based and separates primary, supporting, negative, and suppressed evidence.
Treat primary evidence and memory-supported evidence as the main basis for the report.
Use negative evidence to explicitly rule out findings when appropriate.
Do not verbalize suppressed conflicting evidence unless clearly supported elsewhere.
Retrieved exemplars are style references only and must not introduce new facts.
Never mention osseous integrity, degenerative change, vascularity, aortic tortuosity/calcification, devices, or post-surgical findings unless they are explicitly supported in the evidence chain.
If a section is marked none or a finding is not evidenced, omit it rather than filling in a normal template sentence.
Stay conservative: if evidence is weak or conflicting, prefer omission or uncertain wording.
Output only the final chest X-ray report.

Evidence chain:
{formatted evidence chain}

Report:

C.2 Evidence chain formatting.

The function $\mathrm{Format}(\cdot)$ transforms the evidence chain into a set of ordered prompt sections. First, the study view information is written into a `[Views]` section. If available, global statistics such as view support and dominant observations are summarized in a `[Study overview]` section. Then, evidence items are grouped according to their roles and written into separate sections: `[Primary evidence]`, `[Memory-supported evidence]`, `[Negative evidence]`, `[Prototype summary evidence]`, and optionally `[Suppressed conflicting evidence]`. Each evidence item is linearized into a concise sentence-like record containing the observation, state, anatomical region when available, certainty, confidence, and evidence span. For example:

Formatted evidence chain example

[Views]
- frontal, lateral
[Primary evidence]
- pleural effusion [absent]; id=P1; anatomy=pleural space; certainty=certain; confidence=0.95
- pneumothorax [absent]; id=P2; anatomy=pleural space; certainty=certain; confidence=0.92
[Memory-supported evidence]
- airspace opacity [absent]; id=M1; certainty=certain; confidence=0.88
[Negative evidence]
- pulmonary edema [absent]; id=N1; certainty=certain; confidence=0.83
[Prototype summary evidence]
- cardiomegaly [absent]; id=T1; confidence=0.80

This formatting makes the role of each evidence source explicit. Primary and memory-supported evidence are used for assertive findings, negative evidence is used for rule-out statements, prototype evidence serves as a soft reference, and suppressed evidence is used as an avoidance constraint.

C.3 VQA template.

For medical VQA, the prompt is shorter because the output is an answer rather than a full report. The DeCoT VQA prompt is:

Evidence-aware VQA prompt

You are a medical visual question answering assistant using the full DeCoT pipeline.
Question: {question}
{formatted evidence chain}
Final Answer:

Compared with report generation, the VQA prompt places the question before the evidence context and asks for a concise final answer. For closed-ended questions, the model is expected to answer with the shortest correct option, such as “yes” or “no”; for open-ended questions, it produces a brief free-text answer. VQA uses the same evidence formatting

function as report generation, but focuses more on answering the specific question rather than composing a complete radiology report.

C.4 Prompt changes in ablation studies.

We also construct prompt variants to isolate the contribution of different components. In the *without role* setting, the same evidence items are provided without role-aware organization: role-specific section names and role-specific generation instructions are removed, so the model receives evidence cues without explicit functional labels such as primary, negative, prototype, or suppressed. In the *without negative evidence* setting, the `[Negative evidence]` section is removed, together with the instruction that encourages explicit rule-out statements for absent findings. In the *without retrieval* setting, retrieval-derived evidence is disabled, including memory-supported evidence, prototype evidence, retrieved exemplars, and suppression signals derived from retrieval conflicts; the prompt is therefore grounded mainly on image-conditioned anchor evidence. For the non-DeCoT anchors-and-retrieval baseline, the prompt contains structured anchors and retrieved similar cases as auxiliary context, but does not assign role-aware evidence functions such as primary, negative, prototype, or suppressed evidence.