Structured Radiology Report Generation with Rich Clinical Context

Anonymous ACL submission

Abstract

Automated *structured radiology report generation* (SRRG) from chest X-ray images offers significant potential to reduce workload of radiologists. While radiologists effectively utilize available clinical contexts in their diagnostic reasoning, existing SRRG systems *overlook* these essential elements. This fundamental gap leads to critical problems including *temporal hallucinations* when referencing non-existent clinical contexts. To address these limitations, we introduce *contextualized SRRG* (**C-SRRG**) that comprehensively incorporates rich clinical context for SRRG. We curate C-SRRG dataset by integrating comprehensive clinical context encompassing 1) multi-view X-ray images, 2) clinical indication, 3) imaging techniques, and 4) prior studies with corresponding comparisons based on patient histories. Through extensive benchmarking with diverse multimodal large language models, we demonstrate that incorporating clinical context with the proposed C-SRRG significantly improves clinically grounded report generation while mitigating temporal hallucinations.

1 Introduction

Writing a report based on medical images requires radiologists to accurately interpret the images and synthesize them into two main components: 1) detailed *findings* that systematically document anatomical structures and pathological observations, and 2) concise *impressions* that provide clinical interpretations for subsequent decision-making (Wallis and McCoubrie, 2011; Pahadia et al., 2020; Haygood et al., 2018; Trinh et al., 2019; ESR, 2011). However, writing up such comprehensive reports is both cognitively demanding and time-consuming for the radiologists. Given the high volume of imaging studies and time-intensive nature of report writing, there is an unmet need for automated systems to generate accurate and structured reports to assist radiologists improve diagnostic efficiency and efficacy (Markotić et al., 2021; Alexander et al., 2022).

In this regime, *Automated radiology report generation* (**RRG**) has emerged as a crucial task in both natural language processing and medical imaging, which helps radiologists in the diagnostic workflow (Esteva et al., 2019; Sloan et al., 2024; Tanno et al., 2025). Deep learning has accelerated the development of automated RRG frameworks that generate reports directly from medical images (Shin et al., 2016; Jing et al., 2018; Li et al., 2018; Wang et al., 2018; Jing et al., 2020; Chen et al., 2020a, 2022; Wang et al., 2022a). Recent advances in multimodal large language models (MLLMs) further enhanced generation of coherent and clinically relevant text by integrating large vision and language foundation models (Lee et al., 2025; Li et al., 2023a; Hyland et al., 2023; Bannur et al., 2024; Chen et al., 2024b; Sellergren et al., 2025; Team et al., 2025; Wang et al., 2025).

Despite the recent progress, most automated RRG frameworks overlook *essential clinical context* such as imaging indication, technique, and prior studies that radiologists routinely use to generate reports (Kahn et al., 2009; ESR, 2011). Ignoring the clinical context leads to systematic errors (Liu et al., 2019; Ramesh et al., 2022) as the models fail to capture patient-specific properties and longitudinal changes critical for accurate diagnosis, *e.g.*, *temporal hallucinations* where the models generate references to nonexistent priors or fabricate temporal comparisons (Fig. 3) when temporal statements in the reports are not inferable from the current image alone. Although some work utilized partial context—*e.g.*, multi-view images (Yuan et al., 2019; Chen et al., 2022), historical images (Hou et al., 2023a; Zhu et al., 2023b), and prior reports or indications (Miao et al., 2024; Wang et al., 2024)—these are still *limited* as they only consider fragmented clinical context, rely only on the immediately preceding image or report (Ban-

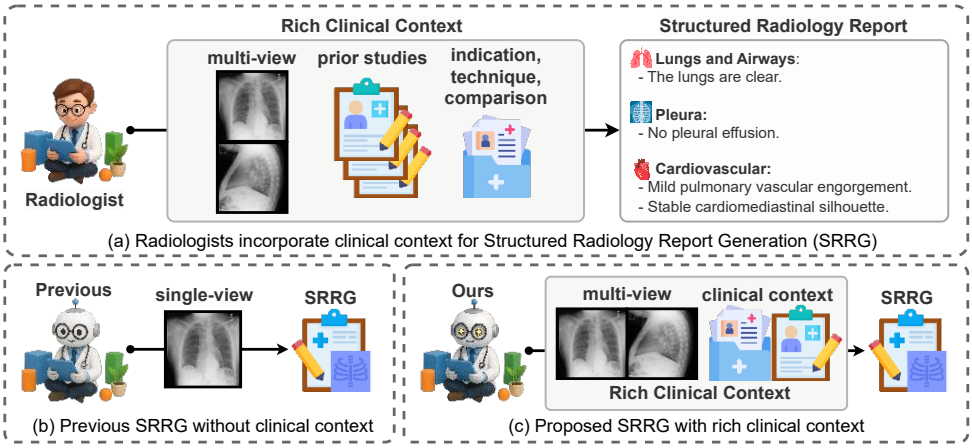

Figure 1: **A conceptual illustration of the proposed C-SRRG**. **(a)** Radiologists routinely use clinical context, while **(b)** existing SRRG frameworks do not. Motivated by this gap, **(c)** C-SRRG leverages multi-view images, indication, technique, and variable-length prior studies/comparisons to generate structured radiology reports.

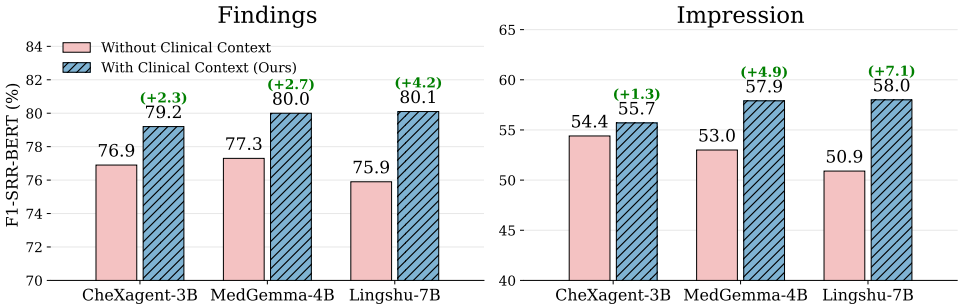

Figure 2: **Clinical context consistently and significantly improves medical MLLMs**—including CheXagent-3B (Chen et al., 2024b), MedGemma-4B (Sellergren et al., 2025), and Lingshu-7B (Team et al., 2025)—on both the findings and impression tasks for SRRG, as measured by F1-SRR-BERT metric (Delbrouck et al., 2025).

nur et al., 2024; Liu et al., 2025b,a), and they all target free-form RRG rather than *structured* RRG (SRRG) (Delbrouck et al., 2025).

To this end, we first present **C-SRRG**,[1] a setup for *contextualized structured radiology report generation* (Fig. 1), built upon the recently introduced SRRG paradigm (Delbrouck et al., 2025). We curate a *large-scale* dataset for *structured* report generation with *rich clinical context* by leveraging MIMIC-CXR (Johnson et al., 2019) and CheXpert Plus (Chambon et al., 2024). Specifically, our **C-SRRG dataset** provides 1) **multi-view images** (frontal and lateral), 2) clinical **indication**, 3) imaging **technique**, and 4) variable-length **prior studies**, which models can incorporate depending on their choice and architecture.

We evaluate the effectiveness of the proposed C-SRRG by incorporating the clinical context within diverse medical MLLMs (Chen et al., 2024b; Sellergren et al., 2025; Team et al., 2025) and find that C-SRRG **substantially and consistently improves report quality** (Tabs. 2 and 3) measured by various metrics (Papineni et al., 2002; Lin, 2004; Zhang et al., 2019; Delbrouck et al., 2022, 2025). We also provide a comprehensive analysis, including extensive ablation studies on clinical context (Tabs. 4 to 7), temporal hallucination mitigation (Tab. 8), and organ-level performance (Tab. 9).

Our **contributions** are summarized as follows:

- We reformulate SRRG as a clinically contextualized generation problem, *i.e.*, **C-SRRG**, that explicitly models heterogeneous and partially available clinical context. It addresses a key limitation of existing SRRG frameworks, *i.e.*, the neglect of essential *clinical context*, which induces systematic errors, most notably *temporal hallucinations* on nonexistent prior studies.
- We curate one of *the largest structured* radiology report generation dataset with *rich clinical context*, namely, **C-SRRG dataset**, which includes 1) multi-view images, 2) indication, 3) technique, and 4) prior studies/comparisons, for training and evaluation of the proposed C-SRRG framework.
- We provide *fine-tuning baselines* for C-SRRG by benchmarking **diverse medical MLLMs**, demonstrating that clinical context *consistently improves report quality across backbones* (*e.g.*,

[1]We use **C-SRRG** for the setup, **C-SRRG dataset** for the curated data, and **C-SRRG-Findings / C-SRRG-Impression** for the two tasks; the ✓/✗ markers in our tables denote presence/absence of clinical context.

+2.3∼4.2/+1.3∼7.1 on findings/impression for F1-SRR-BERT), with larger absolute gains observed on higher-capacity models, while reducing *temporal hallucinations* (See Tab. 8).

We will *publicly release* the 1) **dataset**, 2) **code**, and 3) **checkpoints** of benchmarked models to facilitate and benefit further research in C-SRRG.

2 Related Work

Automated radiology report generation (RRG). Automated RRG has emerged as a promising approach to reduce radiologists' workload and improve reporting efficiency (Yang et al., 2023; Sloan et al., 2024; Esteva et al., 2019; Sirshar et al., 2022; Tanno et al., 2025; Singh and Singh, 2025). While early approaches simply combined vision encoders with language decoders for visual feature extraction (He et al., 2015; Dosovitskiy et al., 2020) and text generation (Shin et al., 2016; Jing et al., 2018; Li et al., 2018; Wang et al., 2018; Jing et al., 2020; Chen et al., 2020b; Yan and Pei, 2022; Miura et al., 2020), architectural innovations have significantly improved report quality, such as memory-driven transformers (Chen et al., 2020a; Liu et al., 2024b), specialized architectures for medical domain knowledge (Yang et al., 2021; Wang et al., 2022b; Kong et al., 2022), cross-modal learning for improved alignment (Chen et al., 2022; Wang et al., 2022a; Li et al., 2023b), and region-guided frameworks for anatomically relevant features (Tanida et al., 2023; Li et al., 2023c; Hou et al., 2023b). In this work, we focus on extending the recently proposed structured RRG (SRRG; Delbrouck et al., 2025)—improving clarity, consistency, and interpretability through standardized structure (Weiss and Langlotz, 2008; Kahn et al., 2009; Bosmans et al., 2012, 2015)—by incorporating rich clinical context aligned with radiologists' workflow.

Multimodal large language models (MLLMs). Building on recent advances in LLMs (Bai et al., 2023; Achiam et al., 2023; Touvron et al., 2023a,b; Yang et al., 2025), MLLMs have shown strong performance across many domains, including medical applications (Achiam et al., 2023; Team et al., 2023; Yang et al., 2025; Comanici et al., 2025; Wang et al., 2025; Zhu et al., 2025). They integrate visual understanding with natural-language generation, enabling effective tools for medical image analysis and clinical text generation (Li et al., 2023a; Chen et al., 2024a; He et al., 2024; Hurst et al., 2024; Lai et al., 2025; Pan et al., 2025). Medical-specific MLLMs further improve performance by incorporating domain knowledge and clinical expertise through specialized training procedures, including CheXagent (Chen et al., 2024b), MedGemma (Sellergren et al., 2025), and Lingshu (Team et al., 2025). These foundation models are particularly promising for comprehensive RRG frameworks (Lee et al., 2023; Zhu et al., 2023a; Liu et al., 2024c; Wang et al., 2023; Hyland et al., 2023; Bannur et al., 2024; Lee et al., 2025; Chen et al., 2024b; Dai et al., 2025; Sellergren et al., 2025; Team et al., 2025), where their ability to accept flexible inputs and produce coherent clinical text is especially valuable. Accordingly, we benchmark medical MLLMs for contextualized SRRG.

Clinical context. Radiologists routinely leverage clinical context when drafting reports, drawing upon patient history, prior studies, and clinical indications (Kahn et al., 2009; ESR, 2011; Wallis and McCoubrie, 2011; Haygood et al., 2018; Trinh et al., 2019; Pahadia et al., 2020; Castillo et al., 2020; Nguyen et al., 2021), motivating various approaches to integrate such clinical context into automated RRG frameworks. Multi-view image analysis utilizes complementary imaging perspectives, such as frontal and lateral views, to provide comprehensive anatomical coverage (Yuan et al., 2019; Miao et al., 2024; Chen et al., 2022; Nooralahzadeh et al., 2021; Serra et al., 2023; Liu et al., 2024d; Nicolson et al., 2024). Indication and clinical history integration approaches incorporate patient-specific clinical information (Hou et al., 2023a; Zhu et al., 2023b; Wang et al., 2024; Liu et al., 2024a; Miao et al., 2024). Previous studies enable temporal comparison and disease progression tracking (Hou et al., 2023a; Zhu et al., 2023b; Serra et al., 2023; Wang et al., 2024; Liu et al., 2021). Recent works such as MLRG (Liu et al., 2025b), PriorRG (Liu et al., 2025a), and MAIRA-2 (Bannur et al., 2024) have attempted to incorporate clinical context for report generation.

However, these approaches have *limitations*: they either 1) consider only partial clinical context, 2) are restricted to specific input configurations, 3) have narrow temporal scope (only the previous prior study), with all 4) focusing on unstructured free-form report generation. In contrast, **C-SRRG** jointly models heterogeneous clinical context, *i.e.*, multi-view imaging, indication, technique, and variable-length prior studies, within a unified structured-report generation framework, while explicitly evaluating robustness to missing context and temporal hallucinations.

Structured Report (Excerpt):
History: A male patient with hep C cirrhosis and large right pleural effusion status post thoracocentesis.
Comparison: Prior portable AP chest radiograph
Findings:
Pleura:
- Moderate pleural effusion within the right pleural space.
- Moderate right pneumothorax, **new from prior exam**.
- No left pleural effusion or pneumothorax.
Impression:
1. Moderate right-sided pneumothorax.
2. Moderate right pleural effusion.

Hallucination: The phrase "**new from prior exam**" represents temporal information that cannot be verified from the current study alone, if not with previous history.

Figure 3: **An example of temporal hallucinations**. This report contains "**new from prior exam**" even though any prior studies are not provided. Please see examples of full structured reports in Figs. 23 to 25.

3 Method

3.1 Curation of Clinical Context

Motivation. Our design principle is to reflect the clinical workflow of radiologists that incorporates a diverse diagnostic context such as indication, technique, and comparison (Wallis and McCoubrie, 2011; Trinh et al., 2019; Pahadia et al., 2020; Nguyen et al., 2021), supported by empirical evidence showing improvement in report quality (Castillo et al., 2020; Liu et al., 2021). This emphasis on a comprehensive clinical context aligns with recent work advocating that AI systems must be developed beyond narrow task-specific approaches that lack the ability to incorporate multimodal data and provide comprehensive interpretation assistance (Dogra et al., 2025). Most importantly, without this context, existing automated systems are prone to *temporal hallucinations*: ground truth reports frequently contain temporal statements such as "**new from prior exam**" (as shown in Fig. 3), which leads models to hallucinate by referencing nonexistent prior examination (Ramesh et al., 2022). When trained on such data, the SRRG frameworks learn to generate these temporal phrases even when no prior studies are available, as demonstrated in Figs. 26 to 31.

Clinical context. To address this limitation, we curate a dataset that exposes rich clinical context to SRRG models. Specifically, we consider **four clinical elements** that radiologists routinely use (refer to §F for detailed clinical interpretation of each clinical context element):

1. **Multi-view images** (*e.g.*, posteroanterior, anteroposterior, and lateral) provide complementary perspectives from different angles, enabling comprehensive assessment of abnormalities that are obscured in single views. Multi-view fusion captures richer information through cross-view consistency, and reduces diagnostic uncertainty (Yuan et al., 2019; Miao et al., 2024).
2. **Indication** conveys the clinical rationale for imaging, providing context about patient symptoms, suspected conditions, or clinical questions. This enables models to focus on specific diagnostic questions, tailor findings to physician concerns, and avoid clinically insignificant findings.
3. **Technique** captures imaging protocols, contrast use, and factors affecting image quality, helping models note caveats and avoid mistaking artifacts for pathology.
4. **Prior studies** enable longitudinal comparison to detect disease progression, treatment response, and interval changes. Radiologists routinely consult such a history (Haygood et al., 2018; Liu et al., 2025a), which supports accurate change detection and prevents hallucinations referencing nonexistent prior exams.

Dataset curation. We build on the recently proposed SRRG dataset (Delbrouck et al., 2025), which includes both MIMIC (Johnson et al., 2019) and CheXpert Plus (Chambon et al., 2024). When multiple views are available, we integrate multi-view images using `ViewPosition` for MIMIC and `frontal_lateral`, `ap_pa` for CheXpert Plus. For MIMIC, each patient is identified by a unique `subject_id`; we group by `subject_id` and use temporally-ordered `StudyDate` and `StudyTime` to establish chronological sequences. For CheXpert Plus, each study contains a `deid_patient_id` and

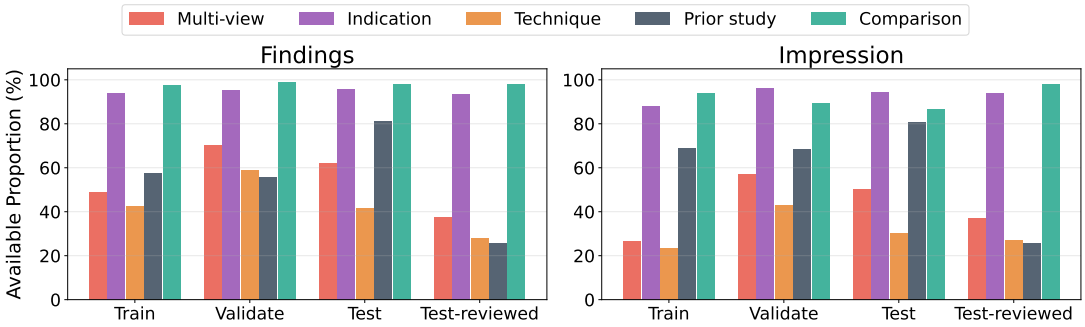

Figure 4: **Available proportion of clinical context** for each split in findings and impression.

Figure 5: **Distribution of the number of prior studies** available per sample for findings and impression.

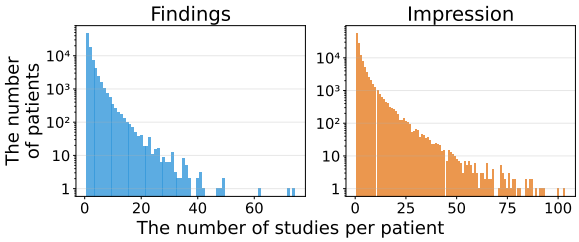

Figure 6: **Distribution of number of studies per patient**.

Tasks	Train	Valid	Test	Test-reviewed	Total
Findings	181,874	976	1,459	233	184,542
Impression	405,972	1,505	2,219	231	409,927

Table 1: **Dataset statistics**.

`patient_report_date_order` field. We group studies by patient and use the order of report dates to form longitudinal sequences. For other clinical contexts (indication, technique, comparison), we use SRRG components, parsed from free-form reports using GPT-4 (Achiam et al., 2023).

Dataset analysis. Accordingly, we curate two C-SRRG tasks that mirror clinical practice—**C-SRRG-Findings** and **C-SRRG-Impression** with *train*, *valid*, *test*, and *test-reviewed*[2] splits, as summarized in Tab. 1. The splits enforce the strict separation of patients between training and evaluation to prevent data leakage and properly assess generalization, as in Fig. 8. The availability of a clinical context varies across splits and tasks (Fig. 4). The availability of prior studies follows a long-tailed distribution (Fig. 5), alongside the long-tailed counts of studies per patient (Fig. 6)—from no history to extensive longitudinal sequences. This variability reflects *real-world clinical practice* and requires models to handle missing information while leveraging available context.

3.2 Contextualized Radiology Report Generation (C-SRRG)

Contextualized generation formulation. We formulate C-SRRG as conditional structured generation under heterogeneous clinical context by constructing prompt templates that present clinical context to the MLLM as a single multimodal token sequence. As shown in Fig. 7, each input begins with an instruction, followed by $N \geq 0$ *prior-study* blocks (each repeating an indication/technique/findings/impression/images block), the *current-study* block (indication/technique/comparison/images), and finally the structured assistant response. Four instruction variants cover findings vs. impression $\times$ with vs. without prior studies; filled examples for all four are deferred to §I and §J.

Training and inference. We fine-tune medical MLLMs on these contextualized prompt–response

[2]The *test-reviewed* split reports are reviewed by board-certified radiologists (Delbrouck et al., 2025).

SYSTEM: You are an expert radiologist.
USER: Analyze the current chest X-ray (and compare with prior studies, if any) to write the FINDINGS or IMPRESSION section. Consider available clinical context.

=== PREVIOUS STUDY n **===**
Previous Indication: {indication$_n$}
Previous Technique: {technique$_n$}
Previous Findings: {findings$_n$}
Previous Impression: {impression$_n$}
Previous Images: [image tokens]

(repeat for N priors; $N \geq 0$, block omitted if no prior)

=== CURRENT INDICATION === {indication}
=== CURRENT TECHNIQUE === {technique}
=== CURRENT COMPARISON === {comparison}
=== CURRENT IMAGES === [image tokens]

ASSISTANT: {FINDINGS or IMPRESSION: structured anatomical sections}

Figure 7: **C-SRRG input schema.** Each example forms a single multimodal token sequence: a system/user instruction, $N \geq 0$ *prior-study* blocks, the *current-study* block (indication/technique/comparison/images), then the structured assistant response. Filled examples for the four instruction variants (findings/impression $\times$ with/without priors) are in §I and §J.

pairs for both findings and impression tasks. Models receive prompt and clinical context to form a unified multimodal input sequence. The training objective is then the next-token prediction task under an autoregressive language modeling loss:

$$\mathcal{L}(\theta) := \frac{1}{T}\sum_{t=1}^{T} -\log p_\theta(y_t|x, y_{<t}), \qquad (1)$$

where x is the multimodal token sequence (Fig. 7; full examples in §J) and $y_{1:T}$ is the target token sequence (*e.g.*, structured reports on findings or impression). Here, p_θ denotes the MLLM parameterized by θ. We minimize the negative log-likelihood with respect to θ, *i.e.*, standard cross-entropy over the vocabulary. If prior studies are available, they are inserted into designated slots; otherwise, the model receives only the clinical context of the current study. This design allows the model to adapt to *heterogeneous clinical contexts* (Figs. 4 to 6), to produce context-aware reports when there is prior information, and to avoid hallucinated temporal comparisons when not.

4 Experiments

We evaluate C-SRRG across diverse medical MLLMs on structured report generation, robustness to missing clinical context, and mitigation of temporal hallucinations.

4.1 Experimental Setups

Implementation details. We evaluate three medical multi-modal large language models—**CheXagent-3B** (Chen et al., 2024b), **MedGemma-4B** (Sellergren et al., 2025), and **Lingshu-7B** (Team et al., 2025)—which natively accept variable multi-image and multi-prior input.[3] We first train baseline models without clinical context, generating reports directly from a single image. When training with C-SRRG, we use all available clinical context (*e.g.*, indication, prior studies), except for CheXagent-3B on C-SRRG-Impression which uses only *indication* since full context fails its output format (§E.2). We consider the two most recent prior studies with limited number of images (2/3/2 for CheXagent-3B/MedGemma-4B/Lingshu-7B) due to context limits.

We fine-tune all models with LoRA (Hu et al., 2022) (rank r=32, α=64, dropout p=0.1) for one epoch using batch size 128 and Adam (Kingma and Ba, 2014) with a learning rate of 2×10^{-4}, a cosine scheduler, and 3% warmup. For inference, we use vLLM (Kwon et al., 2023) with greedy decoding for reproducibility. All experiments run on a single NVIDIA H100 GPU.

Evaluation metrics. To assess text quality, we first use the conventional metrics, such as **BLEU** (Papineni et al., 2002), **ROUGE-L** (Lin, 2004), and **BERTScore** (Zhang et al., 2019). For clinical accuracy, we report **F1-RadGraph** (Delbrouck et al., 2022) and SRRG-specific metrics (Delbrouck et al., 2025): **F1-SRRG-BERT**, built on CXR-BERT (Boecking et al., 2022) for structured evaluation, and **Category Score** (only for findings) measuring whether the generated report includes the correct set of anatomical section headers (Pleura, Cardiovascular, Lungs/Airways, etc.), reported as precision, recall, and F1 over the 8-category vocabulary of Delbrouck et al. (2025).

4.2 Experimental Results

Results on the C-SRRG-Findings. Tab. 2 demonstrates *substantial improvements* achieved by C-SRRG on the C-SRRG-Findings across all

[3] Specialized fixed-schema RRG architectures (*e.g.*, MAIRA-2 (Bannur et al., 2024), RaDialog (Pellegrini et al., 2023), CheXpert-Plus (Chambon et al., 2024)) cannot accommodate C-SRRG inputs and are excluded; see §B.

Model	Clinical Context	Split	Traditional Metrics				F1-SRR-BERT			Category Score		
			BLEU	ROUGE-L	BERT Score	F1-RadGraph	Precision	Recall	F1-Score	Precision	Recall	F1-Score
CheXagent-3B	✗	Valid	1.97	20.63	30.33	13.07	44.67	45.16	43.46	73.61	81.17	75.54
		Test	2.08	20.09	31.91	12.99	43.73	42.54	41.70	74.47	85.26	77.74
		Test-reviewed	2.13	20.38	32.73	12.96	44.94	42.78	42.31	72.84	87.35	77.55
	✓	Valid	2.31	23.01	33.46	15.76	48.73	48.20	46.79	77.58	83.46	78.73
		Test	1.89	20.92	33.28	13.58	45.18	44.10	43.07	75.79	85.69	78.82
		Test-reviewed	1.98	21.64	34.32	14.05	47.50	45.09	44.59	76.08	88.87	79.93
MedGemma-4B	✗	Valid	1.51	20.95	30.83	13.98	42.93	45.50	42.12	78.48	78.00	76.26
		Test	1.58	19.69	31.52	13.30	42.32	41.38	40.19	76.31	82.36	77.44
		Test-reviewed	1.60	20.11	32.61	13.42	44.49	42.94	41.92	75.39	86.56	78.24
	✓	Valid	4.98	27.22	37.87	20.44	50.52	49.68	48.42	80.38	83.73	80.35
		Test	3.05	23.17	35.65	15.91	45.84	44.24	43.43	78.28	84.67	79.59
		Test-reviewed	4.29	24.37	36.60	17.01	47.90	45.17	44.96	76.73	87.84	80.04
Lingshu-7B	✗	Valid	1.42	17.68	27.20	10.56	40.15	41.45	39.29	74.37	75.97	73.57
		Test	1.40	17.71	29.65	11.14	40.60	39.41	38.65	75.86	81.47	76.84
		Test-reviewed	1.60	18.62	31.09	12.09	42.85	40.82	40.37	74.32	85.20	77.39
	✓	Valid	6.02	28.70	38.85	21.67	51.16	50.50	49.20	81.97	83.03	80.87
		Test	3.16	23.53	35.60	16.07	45.96	44.42	43.63	79.80	83.20	79.68
		Test-reviewed	4.42	23.70	35.76	16.09	47.48	44.80	44.54	77.57	86.71	79.83

Table 2: **Results on the C-SRRG-Findings**. Clinical context is incorporated under the C-SRRG setting.

Model	Clinical Context	Split	Traditional Metrics				F1-SRR-BERT		
			BLEU	ROUGE-L	BERT Score	F1-RadGraph	Precision	Recall	F1-Score
CheXagent-3B	✗	Valid	9.44	34.03	61.82	19.30	63.80	63.48	59.10
		Test	7.83	29.40	59.82	16.13	57.18	59.18	54.27
		Test-reviewed	7.42	28.60	58.35	13.71	51.32	56.34	49.74
	✓	Valid	7.52	32.99	60.93	17.90	66.28	61.75	59.04
		Test	7.03	29.18	59.66	16.07	59.69	58.42	55.07
		Test-reviewed	6.92	29.04	58.91	14.84	55.42	58.26	52.86
MedGemma-4B	✗	Valid	8.92	41.24	60.94	17.80	62.19	60.77	56.81
		Test	7.15	37.84	59.09	15.35	56.27	57.01	52.69
		Test-reviewed	7.57	35.91	58.35	14.57	51.69	54.42	49.51
	✓	Valid	11.76	46.26	64.28	24.25	65.78	66.77	62.31
		Test	10.58	41.92	61.85	19.23	59.45	61.89	57.12
		Test-reviewed	11.21	40.15	61.12	19.16	55.02	60.71	54.20
Lingshu-7B	✗	Valid	8.15	32.17	59.15	17.23	63.82	57.10	55.06
		Test	6.65	27.27	57.18	13.87	56.03	51.55	49.33
		Test-reviewed	7.04	27.70	57.37	13.49	52.34	52.85	48.37
	✓	Valid	11.77	38.46	64.82	25.29	69.42	63.57	62.48
		Test	10.58	32.86	62.07	19.85	63.04	58.39	57.01
		Test-reviewed	11.61	33.66	62.04	21.28	57.48	58.80	54.53

Table 3: **Results on the C-SRRG-Impression**. Clinical context is incorporated under the C-SRRG setting.

evaluation metrics, except for slight BLEU decreases for CheXagent-3B on the test/test-reviewed splits (-0.19/-0.15). F1-SRR-BERT and Category Score both improve consistently across all three backbones, with absolute gains roughly proportional to model capacity (Lingshu-7B > MedGemma-4B > CheXagent-3B; full per-model numbers in Tab. 2).

Results on the C-SRRG-Impression. Tab. 3 shows the same pattern on C-SRRG-Impression: F1-SRR-BERT improves consistently across all backbones, again with absolute gains larger for higher-capacity models, which suggests that more capable models may extract richer signal from rich context. CheXagent-3B alone shows minor BLEU regressions on both tasks (Tabs. 2 and 3), indicating that rich clinical context may compromise text-generation fluency in smaller models.

Robustness to missing clinical context. We ablate the four train×eval configurations of clinical context using MedGemma-4B (Tabs. 4 and 5). Providing context in only one phase yields small, mostly non-negative changes, indicating that C-SRRG-trained models *degrade gracefully* when context is missing. The full ✓/✓ configuration achieves the strongest performance, confirming the benefit of context in both phases and the setting remains usable under partial-context deployment.

Impact of Each Clinical Context Component. We ablate four clinical-context components, 1) multi-view images, 2) indication, 3) technique, and 4) prior studies with comparison, to isolate their contributions on the performance. Tabs. 6 and 7 report F1-SRR-BERT for each variant on the C-SRRG-Findings, C-SRRG-Impression, respectively. *All components contribute incrementally* to both tasks (except for few cases), with *performance being highest when using all available context*, which shows the importance of incorporating clinical context in SRRG.

Clinical Context		Split	F1-SRR-BERT		
Train	Eval		Precision	Recall	F1-Score
✗	✗	Valid	42.93	45.50	42.12
		Test	42.32	41.38	40.19
		Test-reviewed	44.49	42.94	41.92
✓	✗	Valid	47.00	47.09	45.35
		Test	42.76	41.55	40.64
		Test-reviewed	44.79	43.12	42.45
✗	✓	Valid	45.28	45.25	43.56
		Test	43.02	40.94	40.44
		Test-reviewed	44.40	41.50	41.36
✓	✓	Valid	50.52	49.68	48.42
		Test	45.84	44.24	43.43
		Test-reviewed	47.90	45.17	44.96

Table 4: **Effect of clinical context for train/eval on the C-SRR-Findings** using MedGemma-4B.

Clinical Context		Split	F1-SRR-BERT		
Train	Eval		Precision	Recall	F1-Score
✗	✗	Valid	62.19	60.77	56.81
		Test	56.27	57.01	52.69
		Test-reviewed	51.69	54.42	49.51
✓	✗	Valid	63.87	61.86	58.45
		Test	54.42	56.53	51.64
		Test-reviewed	51.45	57.86	51.17
✗	✓	Valid	62.60	64.23	59.10
		Test	53.34	59.11	52.59
		Test-reviewed	49.35	58.68	50.66
✓	✓	Valid	65.78	66.77	62.31
		Test	59.45	61.89	57.12
		Test-reviewed	55.02	60.71	54.20

Table 5: **Effect of clinical context for train/eval on the C-SRR-Impression** using MedGemma-4B.

Configuration	Split	F1-SRR-BERT		
		Precision	Recall	F1-Score
Single-view	Valid	47.00	47.09	45.35
	Test	42.76	41.55	40.64
	Test-reviewed	44.79	43.12	42.45
Multi-view	Valid	47.46	47.21	45.80
	Test	44.44	42.57	41.92
	Test-reviewed	45.49	42.69	42.39
+ Indication	Valid	46.95	46.06	44.85
	Test	44.62	42.56	41.98
	Test-reviewed	45.92	43.14	42.79
+ Technique	Valid	50.35	49.27	48.24
	Test	45.50	43.89	43.15
	Test-reviewed	47.48	44.60	44.42
+ Comparison + Prior studies	Valid	50.52	49.68	48.42
	Test	45.84	44.24	43.43
	Test-reviewed	47.90	45.17	44.96

Table 6: **Ablation study on clinical context for the C-SRRG-Findings** using MedGemma-4B.

Configuration	Split	F1-SRR-BERT		
		Precision	Recall	F1-Score
Single-view	Valid	63.87	61.86	58.45
	Test	54.42	56.53	51.64
	Test-reviewed	51.45	57.86	51.17
Multi-view	Valid	65.74	62.92	59.89
	Test	55.70	58.11	53.36
	Test-reviewed	51.78	59.37	52.25
+ Indication	Valid	66.91	65.02	61.67
	Test	58.47	59.11	55.00
	Test-reviewed	53.32	60.47	52.86
+ Technique	Valid	65.91	65.65	61.78
	Test	58.65	60.24	55.88
	Test-reviewed	54.66	60.05	53.39
+ Comparison + Prior studies	Valid	65.78	66.77	62.31
	Test	59.45	61.89	57.12
	Test-reviewed	55.02	60.71	54.20

Table 7: **Ablation study on clinical context for C-SRRG-Impression** using MedGemma-4B.

Task	Split	Temporal Hallucination Rate		Mitigation
		Baseline (✗)	C-SRRG (✓)	
Findings	Valid	146/976 (15.0%)	70/976 (7.2%)	-7.8%
	Test	416/1459 (28.5%)	194/1459 (13.3%)	-15.2%
	Test-reviewed	49/233 (21.0%)	21/233 (9.0%)	-12.0%
	Overall	611/2668 (22.9%)	285/2668 (10.7%)	-12.2%
Impression	Valid	630/1505 (41.9%)	364/1505 (24.2%)	-17.7%
	Test	1012/2219 (45.6%)	599/2219 (27.0%)	-18.6%
	Test-reviewed	92/231 (39.8%)	58/231 (25.1%)	-14.7%
	Overall	1734/3955 (43.8%)	1021/3955 (25.8%)	-18.0%

Table 8: **Mitigation of temporal hallucination** for MedGemma-4B.

Mitigation of temporal hallucination. We compare MedGemma-4B trained *without* clinical context (**baseline**) and *with* clinical context (**C-SRRG**) on evaluation samples where no prior clinical context is available. We measure temporal hallucination by counting reports containing any of 33 manually curated temporal indicators[4] identified from ground-truth and generated reports. Tab. 8 shows that *clinical context substantially mitigates hallucinations* on both tasks, confirming that C-SRRG properly handles missing clinical context. LLM-as-Judge validation (Qwen2.5-72B-Instruct) on the test-reviewed split agrees with the keyword rate on 97–100% of samples and shows a 10–32-point reduction (Tab. 20).

Anatomical region analysis. We additionally compare organ-level Category Score against the no-context baseline (§D), observing improvements on all anatomical regions except hila/mediastinum.

5 Conclusion

We introduced *contextualized structured radiology report generation* (**C-SRRG**), which aligns with radiologists' diagnostic workflow by integrating rich clinical context—multi-view images, indication, technique, and prior studies—into SRRG. Across diverse medical MLLMs, our work suggests that clinically grounded context is essential for reliable structured radiology report generation, consistently improving text quality, diagnostic accuracy, and temporal-hallucination mitigation. We further observe larger gains on higher-capacity backbones, motivating future scaling studies. We will publicly release our dataset, code, and model checkpoints to support further research.

[4]The indicators span three categories: *time references*, *stability indicators*, and *change indicators*; the full keyword list is in §C.

Limitations

Synthetic structured labels. C-SRRG inherits structured reports from Delbrouck et al. (2025), which were reformulated from free-form reports via GPT-4. Only a small subset of the corpus has been radiologist-validated, so residual reformulation biases may propagate through training (§A).

Model scale. We evaluate backbones up to 7B parameters under fine-tuning, limited by single-H100 compute. Larger publicly available variants remain untested, leaving the scale–context interaction confounded with architecture differences (§A).

Constrained clinical-context scope. C-SRRG conditions on at most two prior studies with limited per-study image counts (2/3/2 for our three backbones); patients with extensive longitudinal histories exceed current MLLM context limits. EHR records, lab results, and non-imaging modalities are not incorporated.

Training methodology. We use standard supervised fine-tuning with greedy decoding for reproducibility; preference learning (*e.g.*, PPO, DPO) and retrieval-conditioned decoding, which may further reduce hallucinations, are not explored.

No radiologist-in-the-loop evaluation. Our evaluation relies on automated metrics and an LLM-judge validation of the keyword-based temporal-hallucination detector (Tab. 20); a clinician reader study is left for future work.

Modality coverage. Empirical results are limited to chest X-ray; generalization to CT, MRI, ultrasound, or mammography is plausible per the framework's design but untested.

Residual temporal hallucination. Even with full clinical context, our best model retains 10.7% temporal hallucination on findings and 25.8% on impression (Tab. 8); the framework reduces but does not eliminate the failure mode.

Extended discussion and future-work directions for each item are provided in §A.

Ethics Statement

This work uses only existing de-identified public datasets (MIMIC-CXR and CheXpert Plus) that have undergone rigorous de-identification and received appropriate IRB approvals. MIMIC-CXR is accessed under the PhysioNet Credentialed Health Data License with a signed data use agreement, and CheXpert Plus under its associated research license; all evaluated model checkpoints (CheXagent-3B, MedGemma-4B, Lingshu-7B) are used under the terms of their respective public releases for non-commercial research. No additional patient data was collected, and all privacy protections from the source datasets are maintained. We acknowledge that automated report generation systems may carry biases from their pretraining and source corpora. Our dataset and models are intended solely for research purposes and should not be used for clinical decision-making without appropriate validation and regulatory approval.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Robert Alexander, Stephen Waite, Michael A Bruno, Elizabeth A Krupinski, Leonard Berlin, Stephen Macknik, and Susana Martinez-Conde. 2022. Mandating limits on workload, duty, and speed in radiology. *Radiology*, 304(2):274–282.

Katherine P Andriole. 2023. Picture archiving and communication systems: past, present, and future. *Journal of Medical Imaging*, 10(6):061405–061405.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando P'erez-Garc'ia, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Prasanna Ranjit, Shaury Srivastav, Julia Gong, Fabian Falck, Ozan Oktay, Anja Thieme, Matthew P. Lungren, Maria Teodora Wetscherek, Javier Alvarez-Valle, and Stephanie L. Hyland. 2024. Maira-2: Grounded radiology report generation. *ArXiv*, abs/2406.04449.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel Coelho de Castro, Anton Schwaighofer, Stephanie L. Hyland, Maria Teodora Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. 2022. Making the most of text semantics to improve biomedical vision-language processing. In *European Conference on Computer Vision*.

Jan ML Bosmans, Emanuele Neri, Osman Ratib, and Charles E Kahn Jr. 2015. Structured reporting: a fusion reactor hungry for fuel. *Insights into imaging*, 6(1):129–132.

Jan ML Bosmans, Lieve Peremans, Matías Menni, Arthur M. De Schepper, Philippe Duyck, and Paul M. Parizel. 2012. Structured reporting: if, why, when, how—and at what expense? results of a focus group meeting of radiology professionals from eight countries. *Insights into Imaging*, 3:295 – 302.

Chelsea Castillo, Tom Steffens, L. Sim, and Liam J. Caffery. 2020. The effect of clinical information on radiology reporting: A systematic review. *Journal of Medical Radiation Sciences*, 68:60 – 74.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Curtis P Langlotz, and 1 others. 2024. Chexpert plus: Hundreds of thousands of aligned radiology texts, images and patients. *arXiv e-prints*, pages arXiv–2405.

Junying Chen, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. 2024a. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *ArXiv*, abs/2406.19280.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2022. Cross-modal memory networks for radiology report generation. *arXiv preprint arXiv:2204.13258*.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020a. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020b. Generating radiology reports via memory-driven transformer. *ArXiv*, abs/2010.16056.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, Emily Bao Tsai, Andrew Johnston, Cameron Olsen, Tanishq Mathew Abraham, Sergios Gatidis, Akshay S. Chaudhari, and Curtis P. Langlotz. 2024b. Chexagent: Towards a foundation model for chest x-ray interpretation. *ArXiv*, abs/2401.12208.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Wei Dai, Peilin Chen, Chanakya Ekbote, and Paul Pu Liang. 2025. Qoq-med: Building multimodal clinical foundation models with domain-aware grpo training. *arXiv preprint arXiv:2506.00711*.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Blüthgen, Emily B Tsai, Omar Almusa, and Curt P. Langlotz. 2022. Improving the factual correctness of radiology report generation with semantic rewards. In *Conference on Empirical Methods in Natural Language Processing*.

Jean-Benoit Delbrouck, Justin Xu, Johannes Moll, Alois Thomas, Zhihong Chen, Sophie Ostmeier, Asfandyar Azhar, Kelvin Zhenghao Li, Andrew Johnston, Christian Bluethgen, and 1 others. 2025. Automated structured radiology report generation. *arXiv preprint arXiv:2505.24223*.

Siddhant Dogra, Xiaoman Zhang, Ezequiel Silva, and Pranav Rajpurkar. 2025. The financial, operational, and clinical advantages of generalist radiology ai. *Radiology*, 316(3). Published Online: Sep 9 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929.

ESR. 2011. Good practice for radiological reporting. guidelines from the european society of radiology (esr). *Insights into Imaging*, 2:93 – 96.

Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. 2019. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint*. ArXiv:2312.10997.

Shinya Hattori, Hajime Yokota, Toshihiko Takada, Takuro Horikoshi, Hazuki Takishima, Wakako Mikami, Hiroki Mukai, Hiraku Funakoshi, Toru Naganuma, Tetsuhiro Yano, Kazuhiko Terada, Ryota Inokuchi, Satoshi Tsuchiya, Akira Akutsu, Atsushi Saiga, Yoshihiro Kubota, Takeshi Wada, Akiyo Takada, Hajime Fujimoto, and Takashi Uno. 2021. Impact of clinical information on ct diagnosis by radiologist and subsequent clinical management by physician in acute abdominal pain. *European Radiology*, 31:5454 – 5463.

Tamara Miner Haygood, Barry Mullins, Jia Sun, Behrang Amini, Priya R Bhosale, Hyunseon Christine Kang, Tara L. Sagebiel, and Bilal Mujtaba. 2018. Consultation and citation rates for prior imaging studies and documents in radiology. *Journal of Medical Imaging*, 5.

Kristiina Häyrinen, Kaija Saranto, and Pirkko Nykänen. 2008. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *International journal of medical informatics*, 77(5):291–304.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Sunan He, Yuxiang Nie, Hongmei Wang, Shu Yang, Yihui Wang, Zhiyuan Cai, Zhixuan Chen, Yingxue Xu, Luyang Luo, Huiling Xiang, and 1 others. 2024. Gsco: Towards generalizable ai in medicine via generalist-specialist collaboration. *arXiv preprint arXiv:2404.15127*.

Wenjun Hou, Yi Cheng, Kaishuai Xu, Wenjie Li, and Jiang Liu. 2023a. Recap: Towards precise radiology report generation via dynamic disease progression reasoning. *arXiv preprint arXiv:2310.13864*.

Wenjun Hou, Kaishuai Xu, Yi Cheng, Wenjie Li, and Jiangming Liu. 2023b. Organ: Observation-guided radiology report generation via tree reasoning. *ArXiv*, abs/2306.06466.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

OpenAI Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mkadry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alexander Kirillov, Alex Nichol, Alex Paino, and 397 others. 2024. Gpt-4o system card. *ArXiv*, abs/2410.21276.

Stephanie L. Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Mercy Prasanna Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, Noel Codella, Matthew P. Lungren, Maria Teodora Wetscherek, Ozan Oktay, and Javier Alvarez-Valle. 2023. Maira-1: A specialised large multimodal model for radiology report generation. *ArXiv*, abs/2311.13668.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. 2024. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2607–2615.

Baoyu Jing, Zeya Wang, and Eric Xing. 2020. Show, describe and conclude: On exploiting the structure information of chest x-ray reports. *arXiv preprint arXiv:2004.12274*.

Baoyu Jing, Pengtao Xie, and Eric Xing. 2018. On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2577–2586, Melbourne, Australia. Association for Computational Linguistics.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Roger G. Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317.

Charles E. Kahn, C. Langlotz, Elizabeth S. Burnside, John A. Carrino, David S. Channin, David M. Hovsepian, and D. Rubin. 2009. Toward best practices in radiology reporting. *Radiology*, 252 3:852 – 856.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ming Kong, Zhengxing Huang, Kun Kuang, Qiang Zhu, and Fei Wu. 2022. Transq: Transformer-based semantic query for medical report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Haotong Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Proceedings of the 29th Symposium on Operating Systems Principles*.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, and Xiaofen Yang. 2025. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *ArXiv*, abs/2503.13939.

Seowoo Lee, Jiwon Youn, Hyungjin Kim, Mansu Kim, and Soon Ho Yoon. 2025. Cxr-llava: a multimodal large language model for interpreting chest x-ray images. *European Radiology*, pages 1–13.

Suhyeon Lee, Won Jun Kim, Jinho Chang, and Jong-Chul Ye. 2023. Llm-cxr: Instruction-finetuned llm for cxr image understanding and generation. In *International Conference on Learning Representations*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv preprint*. ArXiv:2005.11401.

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. *ArXiv*, abs/2306.00890.

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. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3334–3343.

Yaowei Li, Bang Yang, Xuxin Cheng, Zhihong Zhu, Hongxiang Li, and Yuexian Zou. 2023c. Unify, align

and refine: Multi-level semantic alignment for radiology report generation. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2851–2862.

Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. *Advances in neural information processing systems*, 31.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023d. Textbooks are all you need ii: **phi-1.5** technical report. *arXiv preprint arXiv:2309.05463*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021. Exploring and distilling posterior and prior knowledge for radiology report generation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13748–13757.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew B. A. McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019. Clinically accurate chest x-ray report generation. *ArXiv*, abs/1904.02633.

Kang Liu, Zhuoqi Ma, Zikang Fang, Yunan Li, Kun Xie, and Qiguang Miao. 2025a. Priorrg: Prior-guided contrastive pre-training and coarse-to-fine decoding for chest x-ray report generation. *arXiv preprint arXiv:2508.05353*.

Kang Liu, Zhuoqi Ma, Xiaolu Kang, Yunan Li, Kun Xie, Zhicheng Jiao, and Qiguang Miao. 2025b. Enhanced contrastive learning with multi-view longitudinal data for chest x-ray report generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10348–10359.

Kang Liu, Zhuoqi Ma, Xiaolu Kang, Zhusi Zhong, Zhicheng Jiao, Grayson Baird, Harrison X. Bai, and Qiguang Miao. 2024a. Structural entities extraction and patient indications incorporation for chest x-ray report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

Kang Liu, Zhuoqi Ma, Mengmeng Liu, Zhicheng Jiao, Xiaolu Kang, Qiguang Miao, and Kun Xie. 2024b. Factual serialization enhancement: A key innovation for chest x-ray report generation. *ArXiv*, abs/2405.09586.

Rui Liu, Mingjie Li, Shen Zhao, Ling Chen, Xiaojun Chang, and Lina Yao. 2024c. In-context learning for zero-shot medical report generation. *Proceedings of the 32nd ACM International Conference on Multimedia*.

Zhizhe Liu, Zhenfeng Zhu, Shuai Zheng, Yawei Zhao, Kunlun He, and Yao Zhao. 2024d. From observation to concept: A flexible multi-view paradigm for medical report generation. *IEEE Transactions on Multimedia*, 26:5987–5995.

Vedran Markotić, Tina Pojužina, Dorijan Radančević, Miro Miljko, and Vladimir Pokrajčić. 2021. The radiologist workload increase; where is the limit?: mini review and case study. *Psychiatria Danubina*, 33(suppl 4):768–770.

Qiguang Miao, Kang Liu, Zhuoqi Ma, Yunan Li, Xiaolu Kang, Ruixuan Liu, Tianyi Liu, Kun Xie, and Zhicheng Jiao. 2024. Evoke: Elevating chest x-ray report generation via multi-view contrastive learning and patient-specific knowledge. *arXiv preprint arXiv:2411.10224*.

Yasuhide Miura, Yuhao Zhang, C. Langlotz, and Dan Jurafsky. 2020. Improving factual completeness and consistency of image-to-text radiology report generation. *ArXiv*, abs/2010.10042.

Hoang T.N. Nguyen, Dong Nie, Taivanbat Badamdorj, Yujie Liu, Yingying Zhu, Jason Truong, and Li Cheng. 2021. Automated generation of accurate & fluent medical x-ray reports. In *Conference on Empirical Methods in Natural Language Processing*.

Aaron Nicolson, Jason Dowling, Douglas Anderson, and Bevan Koopman. 2024. Longitudinal data and a semantic similarity reward for chest x-ray report generation. *Informatics in Medicine Unlocked*, 50:101585.

Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. 2021. Progressive transformer-based generation of radiology reports. *arXiv preprint arXiv:2102.09777*.

Mayank Pahadia, Sonam Khurana, Hassem Geha, and S Thomas Ii Deahl. 2020. Radiology report writing skills: A linguistic and technical guide for early-career oral and maxillofacial radiologists. *Imaging Science in Dentistry*, 50:269 – 272.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Ouyang Cheng, and Daniel Rueckert. 2025. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *ArXiv*, abs/2502.19634.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Chantal Pellegrini, Ege Özsoy, Benjamin Busam, Nassir Navab, and Matthias Keicher. 2023. Radialog: A large vision-language model for radiology report generation and conversational assistance. *arXiv preprint arXiv:2311.18681*.

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.

Vignav Ramesh, Nathan A Chi, and Pranav Rajpurkar. 2022. Improving radiology report generation systems by removing hallucinated references to non-existent priors. In *Machine Learning for Health*, pages 456–473. PMLR.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla P. Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2025. Medgemma technical report. *ArXiv*, abs/2507.05201.

Francesco Dalla Serra, Chaoyang Wang, Fani Deligianni, Jeffrey Dalton, and Alison Q O'Neil. 2023. Controllable chest x-ray report generation from longitudinal representations. *arXiv preprint arXiv:2310.05881*.

Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M Summers. 2016. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2497–2506.

Prateek Singh and Sudhakar Singh. 2025. Chestx-transcribe: a multimodal transformer for automated radiology report generation from chest x-rays. *Frontiers in Digital Health*, 7:1535168.

Mehreen Sirshar, Muhammad Faheem Khalil Paracha, Muhammad Usman Akram, Norah Saleh Alghamdi, Syeda Zainab Yousuf Zaidi, and Tatheer Fatima. 2022. Attention based automated radiology report generation using cnn and lstm. *Plos one*, 17(1):e0262209.

Phillip Sloan, Philip Clatworthy, Edwin Simpson, and Majid Mirmehdi. 2024. Automated radiology report generation: A review of recent advances. *IEEE Reviews in Biomedical Engineering*, 18:368–387.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2023. Interactive and explainable region-guided radiology report generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7433–7442.

Ryutaro Tanno, David GT Barrett, Andrew Sellergren, Sumedh Ghaisas, Sumanth Dathathri, Abigail See, Johannes Welbl, Charles Lau, Tao Tu, Shekoofeh Azizi, and 1 others. 2025. Collaboration between clinicians and vision–language models in radiology report generation. *Nature Medicine*, 31(2):599–608.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Lasa Team, Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang, and Yu Rong. 2025. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *ArXiv*, abs/2506.07044.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tony W Trinh, Atul B. Shinagare, Daniel I. Glazer, Pamela J. DiPiro, Jacob C. Mandell, Giles Boland, and Ramin Khorasani. 2019. Radiology report template optimization at an academic medical center. *AJR. American journal of roentgenology*, pages 1 – 7.

Alison Wallis and Paul McCoubrie. 2011. The radiology report–are we getting the message across? *Clinical radiology*, 66 11:1015 – 1022.

Fuying Wang, Shenghui Du, and Lequan Yu. 2024. Hergen: Elevating radiology report generation with longitudinal data. In *European Conference on Computer Vision*, pages 183–200. Springer.

Jun Wang, Abhir Bhalerao, and Yulan He. 2022a. Cross-modal prototype driven network for radiology report generation. In *European Conference on Computer Vision*, pages 563–579. Springer.

Shansong Wang, Mingzhe Hu, Qiang Li, Mojtaba Safari, and Xiaofeng Yang. 2025. Capabilities of gpt-5 on multimodal medical reasoning. *ArXiv*, abs/2508.08224.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. 2018. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9049–9058.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023. R2gengpt: Radiology report generation with frozen llms. *ArXiv*, abs/2309.09812.

Zhanyu Wang, Mingkang Tang, Lei Wang, Xiu Li, and Luping Zhou. 2022b. A medical semantic-assisted transformer for radiographic report generation. *ArXiv*, abs/2208.10358.

David L Weiss and Curtis P Langlotz. 2008. Structured reporting: patient care enhancement or productivity nightmare? *Radiology*, 249(3):739–747.

Bin Yan and Mingtao Pei. 2022. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In *AAAI Conference on Artificial Intelligence*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shuxin Yang, Xian Wu, Shen Ge, Xingwang Wu, S.kevin Zhou, and Li Xiao. 2021. Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical image analysis*, 86:102798.

Shuxin Yang, Xian Wu, Shen Ge, Zhuozhao Zheng, S. Kevin Zhou, and Li Xiao. 2023. Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Analysis*, 86:102798.

Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. 2019. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023a. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *ArXiv*, abs/2304.10592.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Yue Cao, Yangzhou Liu, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Han Lv, and 29 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *ArXiv*, abs/2504.10479.

Qingqing Zhu, Tejas Sudharshan Mathai, Pritam Mukherjee, Yifan Peng, Ronald M Summers, and Zhiyong Lu. 2023b. Utilizing longitudinal chest x-rays and reports to pre-fill radiology reports. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 189–198. Springer.

This appendix provides comprehensive supplementary materials including limitations and future work discussions (§A), discussion on clinical context integration in prior work (§B), specific limitations of CheXagent-3B with long clinical contexts (§E), detailed interpretation of each clinical context component (§F), baseline performance evaluation of medical MLLMs before instruction tuning (§G), detailed dataset statistics with patient-level data splits (§H), prompt design examples for findings and impression generation (§I), complete instruction fine-tuning examples with clinical context integration (§J), and detailed hallucination analysis with examples (§K).

A Limitations and Future Work

This section extends the main-paper Limitations with detailed discussion and future-work directions for each item.

A.1 Synthetic Dataset and Annotations

Our C-SRRG dataset builds upon the SRRG dataset (Delbrouck et al., 2025), which was generated through reformulation of free-form radiology reports using large language models. This synthetic generation process introduces several potential limitations that warrant careful consideration. The use of LLM-generated content may introduce subtle hallucinations or inconsistencies that could propagate through our training pipeline. Additionally, the reformulation process may inadvertently introduce biases present in the underlying language models, potentially affecting the diversity and clinical accuracy of the generated reports.

While these limitations are inherent to synthetic data generation, several factors mitigate their impact on our work. Prior work validated the structured report generation task through a rigorous reader study with five board-certified radiologists (Delbrouck et al., 2025). Five board-certified radiologists reviewed 464 reports sampled from the MIMIC-CXR test set and CheXpert Plus validation set, assessing the clinical validity of GPT-4 generated structured reports on the test-reviewed split. Their analysis shows high average similarity ratios, computed as 2 × matches / total tokens in original and edited texts: 0.77 for impressions and 0.88 for findings, indicating that modifications were generally minor and focused on enhancing clarity rather than altering fundamental diagnostic content. The utterance-level label validation achieved a 72% exact match rate and 74% average Jaccard similarity between GPT-generated consensus labels and radiologist-reviewed labels, confirming the reliability of the automated label extraction approach. We note that this expert evaluation was conducted specifically on the **test-reviewed** split rather than the full dataset, which was necessary given the resource-intensive nature of comprehensive radiologist review at this scale. Large-scale data may average out such errors across diverse examples, and the clinical validation from prior work suggests the error rate is manageable.

Furthermore, our evaluation protocol is designed to assess clinical validity rather than exact textual matching. F1-SRR-BERT captures semantic similarity appropriate for medical reporting where multiple valid phrasings exist, while organ-level breakdowns provide fine-grained anatomical assessment. Temporal hallucination analysis was conducted by detecting temporal keywords solely from model-generated structured reports, not from synthetically generated ground truth, naturally avoiding potential bias when comparing generated outputs with synthetic targets.

Nevertheless, our approach lacks regulatory clearance for clinical deployment, and practitioners must retain final authority over all diagnostic decisions. Given the inherent risk of errors or hallucinations in any AI system, rigorous clinical validation and expert oversight remain essential when exploring or extending this work.

A.2 Model Architecture and Scale Limitations

Our experimental evaluation faces several architectural constraints that limit the full potential of our approach. First, we are constrained to backbone models with parameters up to 7B, which likely underestimates achievable performance. The computational and memory requirements of larger models present practical limitations for comprehensive evaluation across multiple architectures and clinical contexts. While exploring the effectiveness of C-SRRG across a broader range of model scales would provide valuable insights into how contextual signals scale with model capacity, several practical constraints limit our evaluation scope.

As noted in the §4.1, all experiments were conducted on a single NVIDIA H100 GPU, which fundamentally constrains the scale of models we could evaluate. More importantly, publicly available medical MLLMs offer only limited size options: MedGemma (Sellergren et al., 2025) pro-

vides 4B and 27B variants, while Lingshu (Team et al., 2025) provides 7B and 32B versions. These larger variants (e.g., MedGemma-27B or Lingshu-32B) not only exceed our available computational capacity but also surpass the model sizes commonly studied in academic fine-tuning and evaluation settings. Consequently, our current evaluation does not demonstrate how our findings generalize to larger MLLMs, representing an important direction for future work as computational resources and model accessibility continue to improve.

Second, the multimodal large language models employed were not originally designed to handle multiple images simultaneously. In our experiments, we encountered specific limitations requiring tailored approaches. **Lingshu-7B** (Team et al., 2025) was limited to 2 images due to computational efficiency constraints, while **CheXagent-3B** (Chen et al., 2024b) was similarly restricted to 2 images due to model constraints. In contrast, **MedGemma-4B** (Sellergren et al., 2025) demonstrated multi-image processing capabilities, requiring fewer tokens per image and enabling the use of more images in our longitudinal analysis setting.

These parameter and architectural constraints particularly impact the models' ability to effectively integrate complex temporal relationships and multi-modal clinical information across current and prior studies. However, ongoing research trends in long-context LLM development (Dao et al., 2022; Kwon et al., 2023) suggest that future model architectures will naturally address these limitations. Scaling to larger foundation models, exploring mixture-of-experts variants, and advances in multimodal attention mechanisms represent promising directions for substantial improvements.

A.3 Clinical Context Integration and Selection Methodology

Our approach faces constraints in both the scope and selection of clinical contexts. We impose limits on the number of images and prior studies included per clinical case, with prioritization given to the most recent studies. This limitation stems from current multimodal architectures' context window constraints and computational overhead of processing extensive longitudinal histories. While this recency-based selection strategy captures the most clinically relevant temporal information, it may occasionally omit important historical context that could inform diagnostic reasoning. Additionally, our current implementation primarily utilizes publicly available datasets such as MIMIC-CXR (Johnson et al., 2019) and CheXpert-Plus (Chambon et al., 2024), limiting the diversity of clinical contexts.

Our design choice to use the two most recent prior studies is motivated by both clinical practice and technical constraints. Radiologists typically reference a limited number of relevant prior studies when writing reports. Empirical evidence shows that radiologists consult an average of 3.2 prior imaging studies per report (Haygood et al., 2018), as clinical workflows prioritize the most recent and pertinent examinations for interval change detection. Our default setting of two most recent studies aligns with this standard practice in radiology while remaining within the practical constraints of current medical MLLMs. Current publicly available medical MLLMs also face significant context-length limitations. Some patients have more than 60-100 longitudinal studies as shown in Fig. 11, which exceeds the practical capacity of existing medical MLLMs.

While our approach uses the most recent prior studies, patients with extensive longitudinal histories may have clinically important examinations from earlier studies. This challenge could be addressed through two complementary approaches: advances in efficient attention mechanisms (Dao et al., 2022; Kwon et al., 2023) that enable processing longer temporal contexts, and retrieval-augmented generation methods (Lewis et al., 2021; Gao et al., 2024) that intelligently select the most informative prior studies based on clinical relevance, potentially integrated within agentic systems for dynamic study selection. Our design choice to truncate longitudinal contexts is motivated by both clinical practice and technical limitations, but exploring longer patient histories remains an important research direction.

The scope could be significantly expanded to include additional clinical modalities such as CT imaging, Electronic Health Records (EHR) (Häyrinen et al., 2008), and comprehensive patient histories. Future work should explore learned selection policies that intelligently identify the most informative clinical contexts and optimize longitudinal coverage. Retrieval-augmented generation approaches over Picture Archiving and Communication Systems (PACS) (Andriole, 2023) and EHR systems could dynamically surface the most relevant historical information for each case, unlocking the untapped potential for richer clinical context inte-

gration.

Our focus on imaging-centric clinical contexts (multi-view images, indication, technique, and prior studies) is motivated by established clinical practice and empirical evidence. Prior works demonstrate that providing additional clinical information enhances interpretation accuracy and reporting confidence (Castillo et al., 2020), with best practice guidelines (Kahn et al., 2009) establishing indication, technique, and comparison as standard components of high-quality radiology reports. The structured radiology reporting framework (Delbrouck et al., 2025) explicitly defines these elements (Exam Type, History, Technique, Comparison, Findings, Impression) as standard sections, reflecting clinical practice. Our empirical results further demonstrate the effectiveness of these chosen imaging-centric clinical contexts and their substantial impact on reducing temporal hallucinations.

While additional patient information such as lab tests, demographics, and comprehensive EHR records could provide complementary value, integrating such diverse sources introduces additional complexity. Data heterogeneity, privacy considerations, availability across clinical settings, and computational requirements present practical challenges. The chosen clinical contexts are motivated by established clinical studies and supported by our empirical observations, while investigating the effectiveness of additional patient information remains an important direction for future work.

A.4 Training Methodology and Decoding Strategy

Our training approach is restricted to supervised fine-tuning with greedy decoding for reproducibility and computational efficiency. This methodology, while providing stable and consistent results, may not fully capture the nuanced decision-making processes that characterize expert radiological interpretation. The supervised learning paradigm limits the model's ability to learn from comparative feedback and iterative refinement that occurs in clinical practice. Incorporating preference learning techniques and Reinforcement Learning (RL)-based methods with radiologists' feedback, such as Proximal Policy Optimization (PPO) (Schulman et al., 2017) or Direct Preference Optimization (DPO) (Rafailov et al., 2023), could enhance the fidelity and clinical appropriateness of generated reports. Furthermore, exploring retrieval-conditioned decoding strategies could improve temporal consistency and reduce hallucinations by grounding generation in verified clinical contexts.

A.5 Radiologist-in-the-Loop Evaluation

Our evaluation framework relies exclusively on automated metrics rather than radiologist-in-the-loop assessment. While radiologist evaluation would provide crucial insights beyond automated metrics, including clinical appropriateness of terminology, coherence of diagnostic reasoning, and the impact of reduced temporal hallucinations on clinical decision-making, our work establishes technical foundations through comprehensive automated evaluation. Nevertheless, several factors support the validity of our automated approach: the underlying dataset quality has been validated through expert assessment in prior work (Delbrouck et al., 2025) as discussed in §A, our automated metrics (F1-SRR-BERT, F1-RadGraph) correlate with human judgment (Delbrouck et al., 2025, 2022), and our findings remain consistent across multiple model architectures, suggesting robust improvements from clinical context integration.

Conducting radiologist evaluation presents significant resource challenges, including funding for expert compensation, coordination of radiologist availability. Therefore, given these constraints, radiologist-in-the-loop evaluation represents an important direction for future work as resources become available, particularly when translating these methods toward clinical deployment where regulatory approval and clinical validation become essential requirements.

A.6 Generalization to Other Imaging Modalities

The proposed C-SRRG framework is fundamentally motivated by the diagnostic workflow of radiologists, where clinical context plays a crucial role in report generation across all imaging modalities, not limited to chest X-rays. We believe C-SRRG demonstrates strong potential to generalize to other imaging modalities such as computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, and mammography, as these modalities similarly rely on contextual information for accurate interpretation (Hattori et al., 2021). The core components (*e.g.*, multi-view images, clinical indication, imaging technique, and prior studies) are naturally applicable across modalities, though specific instantiations may vary (*e.g.*, CT involves multi-

ple slices rather than views, MRI includes diverse sequences with different contrasts).

However, each imaging modality presents unique characteristics and reporting conventions that may require modality-specific adaptations. Empirical validation on other imaging modalities would be valuable to confirm generalization and identify the specific adaptations required for optimal performance in each domain. Systematic evaluation of C-SRRG across diverse imaging modalities represents an important direction for exploring its broader clinical utility beyond chest radiography.

B Clinical Context Integration in Prior Work

Prior works have explored incorporating clinical context into radiology report generation, though primarily considering partial contexts in isolation or limited combinations. Some works (Yuan et al., 2019; Miao et al., 2024) incorporate multi-view images, while others explore longitudinal data (Zhu et al., 2023b; Wang et al., 2024; Serra et al., 2023). Recent attempts to combine multiple contexts still face fundamental constraints: fixed input structures (*e.g.*, MAIRA-2 (Bannur et al., 2024) is constrained to exactly three images), limited temporal scope (typically only the immediate previous study), and complex multi-stage training procedures (Liu et al., 2025b,a). In contrast, our work systematically investigates comprehensive clinical context integration with variable numbers of multi-view images, indication text, technique specifications, and multiple prior studies with their complete reports. The comprehensive ablation studies (Tabs. 4 and 5) demonstrate that each context component provides complementary benefits, yielding substantial improvements (2–7% F1-SRR-BERT) and significant reductions in temporal hallucinations (12–18%).

Critically, previous works focused exclusively on free-form report generation, overlooking structured reports despite their clinical importance for clarity, consistency, and adherence to reporting standards (Kahn et al., 2009; Weiss and Langlotz, 2008; Delbrouck et al., 2025). The structured report generation task, recently introduced by prior work (Delbrouck et al., 2025), has not been studied in the context of comprehensive clinical contextualization.

Furthermore, prior works employ specialized architectural components that fundamentally limit their applicability to evaluation on the proposed benchmark. These include multi-view fusion with cross-view consistency (Yuan et al., 2019), multi-positive contrastive learning (Miao et al., 2024; Liu et al., 2025b), posterior and prior knowledge distillation (Liu et al., 2021), graph-based disease progression reasoning (Hou et al., 2023a), group causal transformers for longitudinal aggregation (Wang et al., 2024), and multi-stage contrastive learning (Liu et al., 2025a). These dedicated architectural modifications are specifically designed for their respective context integration approaches and are fundamentally incompatible with recent medical multimodal large language models. Consequently, only recent medical MLLMs (*e.g.*, CheXagent-3B (Chen et al., 2024b), MedGemma-4B (Sellergren et al., 2025), Lingshu-7B (Team et al., 2025)) can be evaluated on the proposed datasets, leveraging their flexibility in handling diverse multimodal inputs.

Why specialized SRRG baselines cannot be fine-tuned on C-SRRG. The SRRG benchmark of Delbrouck et al. (2025) reports three specialized RRG systems—MAIRA-2 (Bannur et al., 2024), RaDialog (Pellegrini et al., 2023), and CheXpert-Plus (Chambon et al., 2024). None of these architectures can ingest the C-SRRG input format, which jointly conditions on variable-length multi-view images and variable-length prior studies with associated indication, technique, and comparison text. We therefore exclude them from our experimental comparison rather than copying numbers from their original (non-structured, partial-context) benchmarks, which cannot be made apples-to-apples with C-SRRG-Findings/Impression evaluation. The specific incompatibilities are:

- **MAIRA-2** (Bannur et al., 2024) hardcodes exactly three input slots (current frontal, current lateral, one prior frontal) with a Rad-DINO encoder producing 1,369 visual tokens per image. The fixed three-slot interface cannot accept the variable multi-view $\times$ multi-prior combinations C-SRRG enables, and there is no slot for indication or technique text.
- **CheXpert-Plus** (Chambon et al., 2024) uses a SwinV2 encoder paired with a 2-layer BERT decoder with a 512-position maximum; the 64-token image budget leaves only 448 tokens for all textual I/O, insufficient to accommodate even a single indication+technique+prior-report sequence, let alone several.

- **RaDialog** (Pellegrini et al., 2023) accepts a single image (BioVil-T encoder) feeding a Vicuna-7B decoder, with no mechanism for multiple images or for parsing a structured prior-study block. Adapting it would require architectural surgery on the image and history pathways, not a fine-tuning change.
- **Other specialized methods** (*e.g.*, PromptMRG (Jin et al., 2024), MLRG (Liu et al., 2025b), PriorRG (Liu et al., 2025a), HERGen (Wang et al., 2024), RECAP (Hou et al., 2023a)) similarly impose fixed input schemas, single-image encoders, or single-prior constraints that preclude direct C-SRRG fine-tuning.

Our experimental comparison therefore focuses on recent medical MLLMs that natively accept variable multimodal input, treating C-SRRG itself—rather than cross-paper numbers from incompatible benchmarks—as the principled point of comparison.

C Temporal Hallucination Indicators

The keyword-based temporal-hallucination detector used in Tab. 8 counts a report as containing a temporal hallucination if any of the following 33 indicators appears, grouped by category:

1. **Time references:** 'new', 'newly', 'recent', 'recently', 'previous', 'prior', 'interval', 'compared to', 'since', 'from prior'.
2. **Stability indicators:** 'unchanged', 'stable', 'persistent', 'persisting'.
3. **Change indicators:** 'improved', 'improvement', 'worsened', 'worsening', 'increased', 'decreased', 'enlarging', 'reducing', 'progression', 'regression', 'evolving', 'evolve', 'developing', 'developed', 'resolving', 'resolved', 'temporal change', 'compare', 'comparison'.

The list was compiled by the authors through manual inspection of generated and reference reports. The keyword-based count is cross-validated against an LLM judge (Qwen2.5-72B-Instruct) on the test-reviewed split; keyword and LLM-judge rates agree on 97–100% of samples (Tab. 20).

D Anatomical Region Analysis

We compare organ-level performance for the findings task against the baseline using the SRRG anatomical categories (Delbrouck et al., 2025). Tab. 9 reports the Category Score on the validation split using MedGemma-4B, with abbreviations: P = pleura, A = abdominal, H/M = hila/mediastinum, O = Other, L/A = lungs/airways, C = cardiovascular, M/C = musculoskeletal/chest wall, T/C/S = tubes/catheters/support devices. We observe that incorporating clinical context with proposed C-SRRG **improves the performance across all the anatomical regions**, except for H/M.

Region	Baseline (✗)			C-SRRG (✓)		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
P	47.69	39.62	41.72	58.2	52.39	52.77
A	8.0	8.0	8.0	17.86	17.86	17.86
H/M	37.91	36.55	36.97	34.25	33.07	33.41
O	9.21	7.94	8.2	12.17	10.13	10.6
L/A	40.67	64.67	45.68	57.25	62.24	55.69
C	71.77	68.6	68.99	73.28	70.26	70.51
M/C	26.22	25.26	25.55	43.2	41.97	42.3
T/C/S	57.04	64.18	58.45	59.81	64.39	60.35

Table 9: **Organ-level Category Score** on the Valid split.

E Limitation in CheXagent-3B

E.1 Increased Temporal Hallucination on C-SRRG-Findings

We conduct additional temporal hallucination analysis with CheXagent-3B on C-SRRG-Findings. We can only conduct this analysis on the C-SRRG-Findings, as the C-SRRG-Impression model was trained solely with indication inputs without clinical history due to the training failure behavior described in §4.1. Consequently, the impression model lacks access to temporal information from prior reports, making temporal hallucination analysis infeasible for this task.

Task	Split	Temporal Hallucination Rate		Mitigation
		Baseline (✗)	C-SRRG (✓)	
Findings	Valid	170/976 (17.4%)	280/976 (28.7%)	+11.3%
	Test	462/1459 (31.7%)	614/1459 (42.1%)	+10.4%
	Test-reviewed	65/233 (27.9%)	77/233 (33.0%)	+5.1%
	Overall	697/2668 (26.1%)	971/2668 (36.4%)	+10.3%

Table 10: **Temporal hallucination rates with all temporal keywords for CheXagent-3B**.

While CheXagent-3B shows improvements in the F1-SRR-BERT metric when trained with clinical context on the C-SRRG-Findings task (Tab. 2), the model exhibits unexpected increases in temporal hallucinations as detailed in Tab. 10. Across all evaluation splits (validate, test, and test-reviewed), the model trained with clinical context exhibits higher rates of temporal keyword usage compared to the baseline, with an overall hallucination rate of 36.4% versus 26.1%, representing a 10.3 percentage point increase. On the validation split, the hallucination rate increases from 17.4% to 28.7%

Keyword	Temporal Keyword Counts		Difference	Change (%)
	Baseline (✗)	C-SRRG (✓)		
stable	307	583	+276	+89.9%
unchanged	319	355	+36	+11.3%
interval	14	23	+9	+64.3%
worse	20	21	+1	+5.0%
previous	0	1	+1	+100.0%
similar	0	1	+1	+100.0%
decreased	1	1	0	0.0%
recent	3	2	-1	-33.3%
better	1	0	-1	-100.0%
improved	9	7	-2	-22.2%
persistent	28	13	-15	-53.6%
from prior	42	0	-42	-100.0%
prior	45	1	-44	-97.8%
new	137	88	-49	-35.8%
increased	81	26	-55	-67.9%
Overall	1007	1122	+115	+11.4%

Table 11: **Temporal keyword distribution comparison for CheXagent-3B**. The table shows the frequency of temporal keywords in generated findings for the baseline model (trained without clinical context) and C-SRRG (trained with clinical context), both evaluated on test sets without clinical context. Only keywords that appear at least once in either model are included.

(+11.3%); on the test split, from 31.7% to 42.1% (+10.4%); and on the test-reviewed split, from 27.9% to 33.0% (+5.1%).

To understand the nature of these hallucinations, we further analyzed the distribution of specific temporal keywords in Tab. 11. The analysis reveals that the model trained with clinical context demonstrates a marked preference for specific temporal terms, with "stable" increasing by 89.9% (from 307 to 583 occurrences) and "unchanged" increasing by 11.3% (from 319 to 355 occurrences). This shift suggests that when trained with clinical context, the model develops an over-reliance on specific temporal terms while inappropriately applying them even when no prior information is available, rather than generating more nuanced temporal comparisons.

Task	Split	Temporal Hallucination Rate		Mitigation
		Baseline (✗)	C-SRRG (✓)	
Findings	Valid	174/976 (17.8%)	140/976 (14.3%)	-3.5%
	Test	491/1459 (33.7%)	369/1459 (25.3%)	-8.4%
	Test-reviewed	64/233 (27.5%)	42/233 (18.0%)	-9.4%
	Overall	729/2668 (27.3%)	551/2668 (20.7%)	-6.7%
Impression	Valid	692/1505 (46.0%)	382/1505 (25.4%)	-20.6%
	Test	1153/2219 (52.0%)	622/2219 (28.0%)	-23.9%
	Test-reviewed	116/231 (50.2%)	50/231 (21.6%)	-28.6%
	Overall	1961/3955 (49.6%)	1054/3955 (26.6%)	-22.9%

Table 12: **Mitigation effect of temporal hallucination for LingShu-7B**. The table shows temporal hallucination rates across different data splits for the baseline model (trained without clinical context) and the model trained with clinical context (C-SRRG). The hallucination effect of MedGemma-4B is shown in Table 8.

Importantly, such hallucination increases were not observed in other models. As shown in Tab. 8 and Tab. 12, temporal hallucination rates were substantially reduced for both MedGemma-4B and LingShu-7B when trained with clinical context. Furthermore, as demonstrated in Tab. 13 and Tab. 14, the tendency to overuse temporal keywords was also not observed in these models, with both showing significant reductions in total temporal keyword usage.

Given this keyword distribution pattern, we conducted additional analyses excluding specific temporal terms to isolate their contribution to the hallucination problem. Tab. 15 presents results when excluding the keyword "stable," revealing a remarkable reversal: C-SRRG now achieves lower hallucination rates than the baseline across all splits (validate: 10.8% vs 11.9%, -1.1%; test: 21.0% vs 23.7%, -2.7%; test-reviewed: 20.2% vs 21.0%, -0.8%). The overall hallucination rate drops from 36.4% to 17.2%, demonstrating that the single keyword "stable" accounts for over half of the model trained with clinical context's apparent temporal hallucinations.

When we further exclude both "stable" and "unchanged", this advantage becomes even more pronounced (Tab. 16). The model trained with clinical context achieves substantially lower hallucination rates across all splits: validate (4.9% vs 7.7%, -2.8%), test (7.1% vs 14.1%, -7.0%), and test-reviewed (7.3% vs 13.7%, -6.4%). The overall hallucination rate drops dramatically from 36.4% to 6.3%, while the baseline decreases from 26.1% to 11.7%. This analysis reveals that 83% of the model trained with clinical context's temporal hallucinations (30.1 percentage points out of 36.4%) are attributable solely to these two terms, whereas they account for only 55% of the baseline's hallucinations (14.4 percentage points out of 26.1%). Critically, when these terms are excluded, the model trained with clinical context demonstrates superior performance with 5.4 percentage points lower hallucination rate than the baseline, indicating that it actually learned more conservative and appropriate usage of other temporal language.

These findings suggest that CheXagent-3B's apparent temporal hallucination problem is not a fundamental failure of temporal reasoning, but rather the model learning to incorporate certain temporal terms ("stable," "unchanged") as stylistic markers during training on longitudinal data. However, this learned behavior is problematic from a clinical perspective: assertions like "effusion is stable" create false confidence by implying that prior comparison

Keyword	Temporal Keyword Counts		Difference	Change (%)
	Baseline (✗)	C-SRRG (✓)		
increased	12	52	+40	+333.3%
interval	13	19	+6	+46.2%
new	62	66	+4	+6.5%
persistent	6	7	+1	+16.7%
compare	1	0	-1	-100.0%
compared to	1	0	-1	-100.0%
worsened	1	0	-1	-100.0%
better	1	0	-1	-100.0%
similar	1	0	-1	-100.0%
worse	6	2	-4	-66.7%
decreased	5	1	-4	-80.0%
improved	7	0	-7	-100.0%
stable	86	46	-40	-46.5%
from prior	112	33	-79	-70.5%
prior	115	34	-81	-70.4%
unchanged	500	113	-387	-77.4%
Total	929	373	-556	-59.8%

Keyword	Temporal Keyword Counts		Difference	Change (%)
	Baseline (✗)	C-SRRG (✓)		
new	149	166	+17	+11.4%
worse	93	74	-19	-20.4%
progression	29	1	-28	-96.6%
increased	52	19	-33	-63.5%
compare	39	0	-39	-100.0%
compared to	39	0	-39	-100.0%
from prior	46	5	-41	-89.1%
recent	6	1	-5	-83.3%
decreased	6	0	-6	-100.0%
resolving	1	0	-1	-100.0%
previous	21	1	-20	-95.2%
prior	136	27	-109	-80.1%
interval	196	79	-117	-59.7%
persistent	217	46	-171	-78.8%
unchanged	457	202	-255	-55.8%
stable	1487	900	-587	-39.5%
Total	2974	1521	-1453	-48.9%

Table 13: **Temporal keyword distribution for MedGemma-4B**. The table shows the frequency of temporal keywords in generated findings (left) and impressions (right) for the baseline model (trained without clinical context) and the model trained with clinical context. Only keywords that appear at least once in either model are included.

Keyword	Temporal Keyword Counts		Difference	Change (%)
	Baseline (✗)	C-SRRG (✓)		
interval	11	37	+26	+236.4%
new	27	54	+27	+100.0%
increased	27	48	+21	+77.8%
worse	5	12	+7	+140.0%
worsened	0	1	+1	+100.0%
better	0	1	+1	+100.0%
decreased	1	0	-1	-100.0%
compare	2	0	-2	-100.0%
compared to	2	0	-2	-100.0%
improved	5	0	-5	-100.0%
previous	5	0	-5	-100.0%
similar	5	0	-5	-100.0%
stable	68	60	-8	-11.8%
persistent	11	0	-11	-100.0%
from prior	129	33	-96	-74.4%
prior	131	34	-97	-74.0%
unchanged	653	391	-262	-40.1%
Total	1082	671	-411	-38.0%

Keyword	Temporal Keyword Counts		Difference	Change (%)
	Baseline (✗)	C-SRRG (✓)		
worse	77	95	+18	+23.4%
improved	1	0	-1	-100.0%
worsened	1	0	-1	-100.0%
decreased	1	0	-1	-100.0%
resolving	2	0	-2	-100.0%
progression	2	0	-2	-100.0%
recent	15	4	-11	-73.3%
new	104	71	-33	-31.7%
from prior	35	0	-35	-100.0%
increased	57	3	-54	-94.7%
interval	179	102	-77	-43.0%
prior	130	31	-99	-76.2%
previous	135	1	-134	-99.3%
compare	178	1	-177	-99.4%
compared to	178	1	-177	-99.4%
persistent	245	47	-198	-80.8%
unchanged	591	209	-382	-64.6%
stable	1618	960	-658	-40.7%
Total	3549	1525	-2024	-57.0%

Table 14: **Temporal keyword distribution for LingShu-7B**. The table shows the frequency of temporal keywords in generated findings (left) and impressions (right) for the baseline model (trained without clinical context) and the model trained with clinical context. Only keywords that appear at least once in either model are included.

Task	Split	Temporal Hallucination Rate		Mitigation
		Baseline (✗)	C-SRRG (✓)	
Findings	Valid	116/976 (11.9%)	105/976 (10.8%)	-1.1%
	Test	346/1459 (23.7%)	306/1459 (21.0%)	-2.7%
	Test-reviewed	49/233 (21.0%)	47/233 (20.2%)	-0.8%
	Overall	511/2668 (19.2%)	458/2668 (17.2%)	-2.0%

Table 15: **Temporal hallucination rates without "stable" for CheXagent-3B**.

Task	Split	Temporal Hallucination Rate		Mitigation
		Baseline (✗)	C-SRRG (✓)	
Findings	Valid	75/976 (7.7%)	48/976 (4.9%)	-2.8%
	Test	206/1459 (14.1%)	104/1459 (7.1%)	-7.0%
	Test-reviewed	32/233 (13.7%)	17/233 (7.3%)	-6.4%
	Overall	313/2668 (11.7%)	169/2668 (6.3%)	-5.4%

Table 16: **Temporal hallucination rates without "stable" and "unchanged" for CheXagent-3B**.

occurred when no prior study is available, which is clinically misleading.

E.2 Training Failure on C-SRRG-Impression

While most models show improvement with clinical context, CheXagent-3B exhibits a **critical failure** in following the structured report format instructions when provided with full clinical context. Instead of generating properly formatted impression sections with numbered findings, the model frequently produces single-word outputs or generic phrases. For instance, when the expected format is a multi-point structured impression such as "1. Slight decrease in size of the right apicolateral pneumothorax with chest tube in place. 2. Unchanged multifocal right-sided pulmonary opacities...", CheXagent-3B often generates only "Pneumothorax" or "Pneumonia". This format degradation is widespread, with the model generating non-structured outputs like "No acute cardiopulmonary

process" or "Pulmonary edema" rather than detailed clinical impressions.

The performance metrics in Tab. 17 reveal the severity of this issue: when provided with full clinical context, traditional metrics plummet dramatically (BLEU: 9.44→2.57, ROUGE-L: 34.03→21.76, BERTScore: 61.82→40.10 on validation set). This catastrophic degradation suggests that CheXagent-3B, likely trained primarily on shorter sequence lengths, struggles to process and integrate the extensive clinical context while maintaining adherence to the structured output format. The model's inability to handle long input sequences effectively undermines its utility for clinical applications requiring comprehensive context integration.

E.3 Specific Limitations in CheXagent-3B

We hypothesize that these failures of CheXagent-3B when trained with rich clinical context stems from two fundamental limitations inherent to the model design.

Architectural and Capacity Constraints. CheXagent-3B is built upon Phi-2 (Li et al., 2023d), a 2.7 billion parameter decoder-only transformer that was pretrained with a maximum sequence length of 2,048 tokens. While CheXagent-3B extends this to 4,096 tokens during fine-tuning (Chen et al., 2024b), this capacity remains insufficient for processing rich clinical context that combines visual tokens (ViT sequence length of 1,024), imaging techniques, clinical indications, and extensive prior study narratives. The relatively small model capacity (3.1 billion parameters total, including the vision encoder) may be further insufficient to simultaneously encode rich visual features, process extensive clinical text, maintain attention over long contexts, and generate structured outputs conforming to specific format constraints. Furthermore, the vision-language projector consists of only a two-layer multi-layer perceptron (MLP) projecting visual features from dimension 1,024 to 2,560 (Chen et al., 2024b), which may create a bottleneck when the model must integrate complex visual semantics with extensive textual context.

Training Data Characteristics and Template Overfitting. The CheXinstruct dataset, while large-scale with 8.5 million samples across 35 tasks, predominantly consists of short, task-specific instruction-response pairs using ten manually-defined templates for each task (Chen et al., 2024b). The dataset composition emphasizes perception tasks with concise outputs: view classification with three-choice answers (AP/PA/Lateral), disease identification with binary Yes/No responses, and findings generation from CXR images. For example, a typical training sample for disease classification might consist of: instruction ("Identify if pneumothorax is present in this image"), a single CXR image, and a short response ("Yes" or "No"). Critically, none of the 35 tasks in CheXinstruct explicitly require processing long clinical context alongside images to generate structured radiology reports. This creates a fundamental gap: the model learns to generate reports from visual input with minimal textual context, but fails when required to integrate extensive clinical narrative while maintaining structured output format. The temporal hallucination increase (§E.1) and the training failure (§E.2) suggest that CheXagent-3B has difficulty generalizing beyond the relatively constrained instruction formats encountered during training.

F Clinical Interpretation of Each Context Component

In this section, we provide detailed clinical interpretation of each context component introduced in §3.1, explaining their distinct roles in radiological interpretation and the clinical implications of their omission.

Multi-view Images. Multi-view chest radiography provides complementary anatomical information essential for comprehensive diagnostic assessment. For instance, lateral views enable visualization of retrocardiac and retrosternal regions obscured by the cardiac silhouette in frontal projections. Without multi-view information, pathologies such as middle lobe pneumonia or retrosternal masses may remain undetected. Multiple views allow radiologists to triangulate spatial locations and distinguish true pathology from overlapping normal structures.

Clinical Indication. Clinical indication provides essential context regarding patient symptoms and clinical history that guide radiological interpretation. For example, fever and productive cough direct attention toward infectious etiologies, while trauma history shifts focus to skeletal injuries and pneumothorax. Without clinical indication, models cannot appropriately integrate clinical significance

Model	Full Clinical Context	Split	Traditional Metrics				F1-SRR-BERT		
			BLEU	ROUGE-L	BERT Score	F1-RadGraph	Precision	Recall	F1-Score
CheXagent-3B	✗	Valid	9.44	34.03	61.82	19.30	63.80	63.48	59.10
		Test	7.83	29.40	59.82	16.13	57.18	59.18	54.27
		Test-reviewed	7.42	28.60	58.35	13.71	51.32	56.34	49.74
	✓	Valid	2.57	21.76	40.10	13.10	74.48	49.40	54.05
		Test	2.40	17.54	33.79	9.78	66.56	41.04	45.99
		Test-reviewed	2.89	19.61	37.88	11.87	64.18	41.27	46.44

Table 17: **Performance degradation of CheXagent-3B on C-SRRG-Impression with full clinical context.** The model shows dramatic drops across all metrics when provided with complete clinical context.

of findings, potentially emphasizing incidental observations while overlooking relevant pathologies.

Imaging Technique. Technical parameters document the imaging protocol and inherent limitations crucial for proper interpretation. These include patient positioning, projection type, inspiratory effort, and whether the study was portable. Omitting technique information can lead models to misinterpret technical artifacts as pathological findings. for instance, limited inspiratory effort produces apparent cardiomegaly, while portable anteroposterior examinations magnify cardiac silhouette. Recognition of these factors prevents false-positive interpretations.

Prior Studies. Temporal comparison with prior examinations enables detection of interval changes often more clinically significant than static findings. Prior study information allows tracking disease progression, assessing treatment response, and distinguishing stable chronic findings from acute changes. For example, serial nodule measurements differentiate stable benign lesions from growing malignancies, directly impacting management decisions. Without prior studies, models may generate temporal hallucinations or fail to detect critical patterns such as progressive infiltrates or worsening effusions.

In summary, each clinical context component used in our setting serves a distinct role in radiological interpretation. Context omission not only reduces model performance quantitatively (Tabs. 4 and 5), but fundamentally limits the clinical utility of generated reports by removing essential diagnostic information.

G Evaluation before Instruction Tuning

Medical MLLMs possess inherent capabilities to generate free-form radiology reports even without task-specific fine-tuning, as these models are typically pretrained on diverse medical image-text pairs and general medical knowledge. However, the structured radiology report generation (SRRG) task imposes specific format requirements that fundamentally affect whether pretrained models can be meaningfully evaluated. In this section, we investigate the baseline performance of pretrained medical MLLMs before instruction tuning to establish their zero-shot capabilities and demonstrate the necessity of task-specific fine-tuning.

G.1 Task-Specific Format Requirements

While existing medical MLLMs have been exposed to diverse radiology reports during pretraining, they have not encountered the specific structured format required by the SRRG task (Delbrouck et al., 2025). The evaluation challenges that arise from this format mismatch vary significantly depending on which report section is being generated.

For the **findings generation** task, models must produce reports in a specific structured format with predefined anatomical sections such as “Lungs and Airways:”, “Pleura:”, “Cardiovascular:”, “Hila and Mediastinum:”, and “Tubes, Catheters, and Support Devices:”. Each section must contain bullet-pointed observations organized by anatomical region. However, pretrained models without instruction fine-tuning cannot generate this structured format; instead, they produce free-form narrative text rather than adhering to the required section structure. Since the evaluation metrics and clinical utility depend critically on this structured organization, findings generation cannot be meaningfully evaluated without fine-tuning.

In contrast, the **impression generation** task requires only numbered diagnostic conclusions without strict section headers, making it feasible to evaluate pretrained models. While the quality and clinical appropriateness may vary, pretrained models can generate numbered lists that align with the expected output format, enabling quantitative evaluation.

G.2 Baseline Performance Without Fine-Tuning

We evaluate the baseline performance of pretrained medical MLLMs on the impression generation task without any instruction fine-tuning. Models are evaluated both with (✓) and without (✗) clinical context provided at inference time to assess whether clinical information improves zero-shot report generation capabilities.

G.3 Analysis and Observations

The results in Tab. 18 reveal several important observations about pretrained model capabilities and the role of clinical context.

Substantial Performance Gap Between Pretrained and Fine-Tuned Models. Pretrained models achieve substantially lower F1-SRR-BERT scores compared to all fine-tuned models reported in the main paper (Tab. 3): CheXagent-3B (36.14–45.47), MedGemma-4B (47.50–58.88), and Lingshu-7B (36.56–51.44). The only exception is MedGemma-4B when evaluated with clinical context on the validation set, which achieved 58.88, surpassing the same model trained without clinical context (56.81 from Tab. 3). However, when MedGemma-4B is trained with clinical context as shown in the ablation study (Tab. 4), it achieves 59.10, which still exceeds the pretrained baseline of 58.88. This demonstrates that fine-tuning with appropriate clinical context consistently yields superior performance.

Model-Specific Baseline Capabilities and Improvement Potential. MedGemma-4B exhibits stronger baseline performance than Lingshu-7B across all settings before fine-tuning (47.50–58.88 vs. 36.56–51.44). However, after fine-tuning, Lingshu-7B demonstrates more substantial improvements, ultimately achieving competitive or superior performance (Tab. 3). This suggests that larger models may benefit more from task-specific fine-tuning despite weaker initial capabilities, possibly due to greater capacity to learn structured output formats and integrate complex clinical reasoning patterns.

Clinical Context Improves Zero-Shot Performance. Notably, even without fine-tuning, providing clinical context at evaluation time consistently improves baseline performance across most settings, underscoring its fundamental importance. Specifically, clinical context improves F1-SRR-BERT across splits: on validation by +0.76 (CheXagent-3B), +3.82 (MedGemma-4B), and +5.18 (Lingshu-7B); on test by +1.26 (CheXagent-3B), +4.65 (MedGemma-4B), and +6.09 (Lingshu-7B); and on test-reviewed by +2.67 (Lingshu-7B) and +3.28 (MedGemma-4B). While only one exception exists (CheXagent-3B on test-reviewed split showing -1.05 degradation), it demonstrates that clinical context enhances report generation quality even for models not explicitly trained to utilize such information. This zero-shot improvement provides additional evidence that clinical context carries inherent diagnostic value that medical MLLMs can leverage through their general pretraining.

Necessity of Instruction Fine-Tuning. These results demonstrate the necessity of instruction fine-tuning for the SRRG task. While pretrained models possess general medical knowledge and basic report generation capabilities, they lack the task-specific understanding required for high-quality structured report generation. For this reason, our evaluation focuses on comparing fine-tuned models, examining whether they are trained with or without clinical context to isolate the effect of comprehensive clinical information on structured report generation.

Generalist VLMs with In-Context Learning. We additionally evaluate two recent generalist vision–language models, **Qwen3-VL-8B** and **Qwen3-VL-30B** (MoE), under zero-shot and 1-shot in-context-learning (ICL) protocols on the test-reviewed split (Tab. 19). 0-shot Findings scores collapse to near-zero F1-SRR-BERT because the models produce free-form text rather than the structured anatomical-section format expected by the metric; a single format demonstration (1-shot ICL) lifts findings F1-SRR-BERT to 21–28, while Impression—which has weaker structural constraints—remains close to its 0-shot level. Across both model sizes and both tasks, providing C-SRRG context improves over the no-context (SRRG) input in 1-shot ICL (*e.g.*, Findings 28.09 vs. 26.23 for Qwen3-VL-30B), confirming that clinical context helps even generalist VLMs at inference time. However, the best zero-shot configuration (Qwen3-VL-30B, 1-shot ICL, C-SRRG) still falls 16+ F1-SRR-BERT below our fine-tuned MedGemma-4B + C-SRRG (28.09 vs. 44.96 on Findings), reinforcing that task-specific fine-tuning—not just larger models with prompt

Model	Clinical Context	Split	Traditional Metrics				F1-SRR-BERT		
			BLEU	ROUGE-L	BERT Score	F1-RadGraph	Precision	Recall	F1-Score
CheXagent-3B	✗	Valid	1.03	18.71	30.00	13.80	67.65	37.64	44.71
		Test	0.98	15.19	21.22	12.27	60.07	28.65	36.14
		Test-reviewed	1.13	16.02	22.56	13.20	59.20	32.76	39.70
	✓	Valid	1.44	19.13	30.23	13.98	67.66	38.60	45.47
		Test	1.45	15.75	25.83	12.47	59.77	30.66	37.40
		Test-reviewed	1.32	15.91	24.93	12.38	57.44	31.81	38.65
MedGemma-4B	✗	Valid	1.23	14.26	34.10	9.11	68.13	53.97	55.06
		Test	1.21	13.44	32.86	9.60	60.07	48.46	48.89
		Test-reviewed	0.70	11.73	29.99	7.58	55.20	48.68	47.50
	✓	Valid	1.80	15.04	37.60	11.29	66.18	61.74	58.88
		Test	1.85	15.02	38.06	11.76	59.21	57.54	53.54
		Test-reviewed	1.17	13.16	34.88	8.94	56.39	54.78	50.78
Lingshu-7B	✗	Valid	1.46	14.11	37.18	9.33	64.96	41.23	46.26
		Test	1.58	13.90	39.37	10.40	55.45	34.80	39.02
		Test-reviewed	1.27	13.23	37.38	9.34	49.93	33.17	36.56
	✓	Valid	4.44	21.29	50.03	12.54	61.45	51.64	51.44
		Test	4.70	21.92	52.44	12.80	54.50	46.05	45.11
		Test-reviewed	2.22	14.81	39.97	9.12	51.69	36.37	39.23

Table 18: **Baseline performance on impression generation without instruction fine-tuning**. Models are evaluated with (✓) and without (✗) clinical context at inference time.

engineering—is required for SRRG.

Model	Context	Findings		Impression	
		0-shot	1-shot	0-shot	1-shot
Qwen3-VL-8B	SRRG	0.41	21.28	48.12	43.37
	C-SRRG	4.28	25.04	48.85	45.51
Qwen3-VL-30B	SRRG	0.53	26.23	43.03	41.15
	C-SRRG	1.83	**28.09**	45.88	44.21
Reference: fine-tuned MedGemma-4B + C-SRRG		44.96		54.20	

Table 19: **Zero-shot and 1-shot ICL performance of Qwen3-VL on the test-reviewed split** (F1-SRR-BERT). Even the largest configuration (30B, 1-shot, C-SRRG) trails our fine-tuned 4B model by 16+ points on Findings.

H Detailed Dataset Statistics

We provide detailed statistics of our clinical context chest X-ray dataset, focusing on patient distribution across splits.

Patient Distribution Across Splits. Our dataset maintains strict patient-level separation across training, validation, and test splits to prevent data leakage. As shown in the patient overlap heatmaps, the training set contains 83,147 unique patients for findings and 125,947 unique patients for impression tasks. The validation sets include 434 patients for findings and 477 patients for impression, while the test sets contain 274 patients for findings and 423 patients for impression. The test-reviewed splits comprise 173 patients for findings and 172 patients for impression, with 106 and 108 patients respectively shared with the test split. This patient-level split ensures that clinical studies from the same patient do not appear across different evaluation splits, with zero patient overlap between training and evaluation sets. The distribution maintains clinical diversity while preserving the integrity of comprehensive clinical contexts within patient histories.

I Prompt Designs

In this section, beyond formats for impression with prior studies (Fig. 11), current study (Fig. 13), we provide examples of our design choices for prompts, *e.g.*, formats for findings with/without prior studies (Figs. 9 and 10), impression without prior studies (Fig. 12), prior studies (Fig. 14), and response format for findings and impression (Figs. 15 and 16), used for training and evaluation.

J Instruction Tuning Dataset Prompt Example

We provide detailed instruction fine-tuning examples that showcase the comprehensive clinical context utilized in our approach. These examples demonstrate how all available clinical information is systematically integrated into our instruction tuning dataset, including patient medical history, imaging techniques, previous study findings, and temporal comparisons. The following multi-part examples illustrate the complete structure of our training data, highlighting how comprehensive clinical contexts including temporal, multi-view, and metadata information are preserved and leveraged for clinical reasoning in radiology report generation.

J.1 Findings Generation Example

The first example demonstrates the generation of the FINDINGS section, which requires detailed anatomical observation and temporal comparison across multiple studies (Figs. 17 to 19):

J.2 Impression Generation Example

The second example demonstrates the generation of the IMPRESSION section, which requires clinical synthesis and diagnostic reasoning (Figs. 20 to 22):

K Hallucination Analysis

In this section, we examine a critical limitation of radiology report generation models trained without clinical context, specifically their tendency to hallucinate temporal comparisons when referencing non-existent prior studies. We first demonstrate that dataset ground truth reports contain temporal statements that become hallucinations when clinical context is absent, as radiologists naturally write these comparisons when they have access to prior studies. We then analyze how models trained without such clinical context systematically produce these hallucinations, even for patients with no imaging history.

Dataset Hallucination. Ground truth radiology reports in clinical datasets frequently contain temporal statements such as "new from prior exam," "unchanged," or "stable compared to previous study." These temporal references are clinically appropriate when radiologists have access to prior imaging studies for comparison. However, when language models are trained on these reports without access to the corresponding clinical context and prior studies, they learn to replicate these temporal language patterns indiscriminately. This training paradigm creates a systematic hallucination problem where models generate temporal comparison statements even for patients with no prior imaging history. The following examples demonstrate these temporal hallucinations present in ground truth radiology reports from the dataset, showing how temporal comparison statements appear without proper clinical context (Figs. 23 to 25):

Trained Model Hallucination. Our trained models also exhibit hallucination patterns, particularly when generating temporal comparisons despite having access only to the current study. The following examples from MedGemma-4B model trained without clinical history demonstrate how models can inappropriately generate comparative language even when no prior studies are available for comparison (Figs. 26 to 31):

LLM-judge validation of the keyword detector. To check that the 33-keyword detector used in the main paper (Tab. 8) is not dominated by lexical false positives, we re-ran the detection on the test-reviewed split with an LLM judge (Qwen2.5-72B-Instruct-AWQ). The judge receives the generated report alone and is prompted to classify whether it contains any ungrounded temporal reference, defined as a statement implying comparison with a prior study when no prior study is provided. We evaluate both SRRG-trained baselines (no context at training or inference) and C-SRRG-trained models in a cross-evaluation setup where the C-SRRG model is fed SRRG-style inputs (no context at inference)—this isolates the effect of training-time context exposure from inference-time context availability. Tab. 20 reports both rates per (model, task). Agreement is 97–100% across all combinations, indicating that the keyword detector is well-calibrated; the LLM-judged absolute reduction from C-SRRG training is 10–32 percentage points, consistent in direction with the main-paper keyword-based finding.

Model	Task	SRRG-trained		C-SRRG-trained	
		Keyword	LLM-Judge	Keyword	LLM-Judge
MedGemma-4B	Findings	21.0%	20.2%	5.2%	3.0%
	Impression	39.8%	38.5%	22.1%	22.1%
Lingshu-7B	Findings	27.5%	26.2%	17.2%	15.9%
	Impression	50.2%	50.2%	19.9%	18.6%

Table 20: **LLM-judge validation of the temporal-hallucination keyword detector** on the test-reviewed split. C-SRRG-trained models are evaluated in a cross-evaluation setup (no context at inference). Agreement between keyword and LLM-judge rates is 97–100% across all rows; the LLM-judged reduction from C-SRRG training is 10–32 absolute points.

L The Use of LLMs

We used LLMs solely for light editing such as correcting grammatical errors and polishing some words. They did not contribute to research ideation, experiments, analysis, or substantive writing.

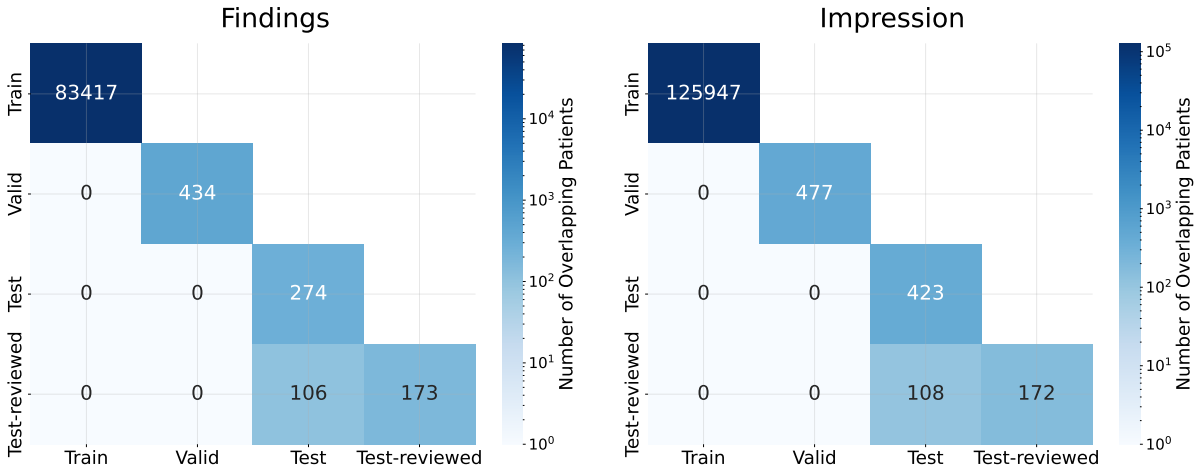

Figure 8: **Patient overlap heatmaps** across train, valid, test, and test-reviewed splits.

SYSTEM PROMPT:
You are an expert radiologist.

USER PROMPT:
Analyze the current chest X-ray images and compare them with the previous studies to write the FINDINGS section of a radiology report. Use standard medical terminology and note any changes from the prior studies, focusing on the most recent comparisons. Consider the available clinical contexts when formulating your findings.

Figure 9: **Prompt format for FINDINGS** for C-SRRG.

SYSTEM PROMPT:
You are an expert radiologist.

USER PROMPT:
Analyze the chest X-ray images and write the FINDINGS section of a radiology report. Use standard medical terminology and organize findings by anatomical regions. Consider the available clinical contexts when formulating your findings.

Figure 10: **Prompt format for FINDINGS without previous history** for C-SRRG.

SYSTEM PROMPT:
You are an expert radiologist.

USER PROMPT:
Analyze the current chest X-ray images and compare them with the previous studies to write the IMPRESSION section of a radiology report. Provide a concise clinical summary and diagnosis, noting any changes from the prior studies, focusing on the most recent comparisons. Consider the available clinical contexts when formulating your impression.

Figure 11: **Prompt format for IMPRESSION** for C-SRRG.

SYSTEM PROMPT:
You are an expert radiologist.

USER PROMPT:
Analyze the chest X-ray images and write the IMPRESSION section of a radiology report. Provide a concise clinical summary and diagnosis based on the imaging findings. Consider the available clinical contexts when formulating your impression.

Figure 12: **Prompt format for IMPRESSION without previous history** for C-SRRG.

=== CURRENT CLINICAL HISTORY/INDICATION ===
Male with end-stage renal disease on hemodialysis, multiple orthopedic hardware, bioprosthetic aortic valve replacement, coronary artery disease status post coronary artery bypass grafting, permanent pacemaker for sick sinus syndrome, admitted for MRSA bacteremia, transferred to the critical care unit for hypotension during anesthesia induction. Patient intubated for respiratory status assessment.

=== CURRENT TECHNIQUE ===
Chest single view

=== CURRENT COMPARISON ===
Prior imaging at an unspecified time.

=== CURRENT IMAGES ===
`[Images: Current chest X-ray images]`

Figure 13: **An example of current study** for C-SRRG.

=== PREVIOUS STUDY N ===
Previous Clinical History/Indication:
Status post coronary artery bypass grafting (CABG), post chest tube removal.

Previous Technique:
Standard posteroanterior and lateral chest radiographs.

Previous Findings:
Lungs and Airways:
- No pneumothorax visualized
- Low lung volumes
- Increased infiltrate in the left upper lung

Pleura:
- Retrocardiac opacity
- Layering left pleural effusion

Cardiovascular:
- Vascular plethora

Other:
- Patchy areas of alveolar edema on the right
- Endotracheal (ET) tube positioned 4.3 cm above the carina
- Nasogastric (NG) tube within the stomach

Previous Impression:
1. Worsened infiltrate in the left upper lung suggestive of an acute process.
2. Presence of retrocardiac opacity and layering left pleural effusion.
3. Vascular plethora and patchy alveolar edema on the right, which may indicate fluid overload or cardiac etiology.
4. Proper positioning of the ET and NG tubes.

Previous Images:
`[Images: Previous chest X-ray images]`

Figure 14: **An example of prior study** for C-SRRG.

ASSISTANT RESPONSE:
FINDINGS:
Lungs and Airways:
- No pneumothorax or focal consolidation.

Pleura:
- Enlarged small right pleural effusion.

Cardiovascular:
- Mildly enlarged heart.

Hila and Mediastinum:
- Normal hilar and mediastinal contours.

Tubes, Catheters, and Support Devices:
- Right internal jugular (IJ) catheter terminates at the superior cavoatrial junction.

Figure 15: **An example of ground-truth assistant responses** in the C-SRRG-Findings dataset.

ASSISTANT RESPONSE:
IMPRESSION:
1. Worsened left basilar opacity, which may represent atelectasis or pneumonitis in the appropriate clinical context.
2. Mild improvement in pulmonary vascularity.

Figure 16: **An example of ground-truth assistant responses** in the C-SRRG-Impression dataset.

Findings Example - Part 1: Current Study Context

USER PROMPT:
Analyze the current chest X-ray images and compare them with the previous studies to write the FINDINGS section of a radiology report. Use standard medical terminology and note any changes from the prior studies, focusing on the most recent comparisons. Consider the available clinical contexts when formulating your findings.

=== CURRENT CLINICAL HISTORY/INDICATION ===
Evaluation for fluid overload.

=== CURRENT TECHNIQUE ===
Standard frontal chest radiography protocol.

=== CURRENT COMPARISON ===
Prior radiographs and CT scans.

=== CURRENT IMAGES ===
`[Images: Current chest X-ray images]`

Figure 17: **Findings generation example (Part 1)** in the C-SRRG-Findings dataset.

Findings Example - Part 2: Previous Study 1

=== PREVIOUS STUDY 1 (Most Recent) ===
Previous Clinical History/Indication:
Patient with a history of multifocal after CABG, currently presenting with symptoms suggestive of CHF or pneumonia.

Previous Technique:
A single frontal chest radiograph was obtained.

Previous Comparison:
Multiple prior radiographs

Previous Findings:
Lungs and Airways:
- No definitive consolidation observed on this examination; however, subsequent CT confirms presence at the right base
- Mild pulmonary edema

Pleura:
- Moderate right pleural effusion, unchanged
- No pneumothorax

Cardiovascular:
- Moderate cardiomegaly noted
- Aortic tortuosity present

Tubes, Catheters, and Support Devices:
- Status post median sternotomy with CABG and valve replacements

Previous Impression:
1. Mild pulmonary edema with right pleural effusion.
2. Right base consolidation not definitive on radiograph but confirmed on subsequent CT.

Previous Images:
`[Images: Previous chest X-ray images]`

Figure 18: **Findings generation example (Part 2)** in the C-SRRG-Findings dataset.

Findings Example - Part 3: Previous Study 2 & Response

=== PREVIOUS STUDY 2 ===
Previous Clinical History/Indication:
The patient presents with tachycardia.

Previous Technique:
Posteroanterior (PA) and lateral chest radiographs.

Previous Findings:
Cardiovascular:
- Status post sternotomy with mitral and aortic valve replacements
- Moderate cardiomegaly, unchanged

Hila and Mediastinum:
- Stable mediastinal and hilar contours

Lungs and Airways:
- New mild-to-moderate elevation of the right hemidiaphragm suggesting volume loss
- Increased patchy opacity in the right lower lobe, concerning for pneumonic consolidation
- Probable new opacification in the right middle lobe
- New vague opacity in the right suprahilar region of the right upper lobe, potentially indicating early pneumonia

Pleura:
- No definite pleural effusion

Previous Impression:
1. Findings suggest multifocal pneumonia involving the right lung.

ASSISTANT RESPONSE:
FINDINGS:
Lungs and Airways:
- No pneumothorax or focal consolidation.

Pleura:
- Enlarged small right pleural effusion.

Cardiovascular:
- Mildly enlarged heart.

Hila and Mediastinum:
- Normal hilar and mediastinal contours.

Tubes, Catheters, and Support Devices:
- Right internal jugular (IJ) catheter terminates at the superior cavoatrial junction.

Figure 19: **Findings generation example (Part 3)** in the C-SRRG-Findings dataset.

Example Case with Clinical Context - Part 1: Current Study

USER PROMPT:
Analyze the current chest X-ray images and compare them with the previous studies to write the IMPRESSION section of a radiology report. Provide a concise clinical summary and diagnosis, noting any changes from the prior studies, focusing on the most recent comparisons. Consider the available clinical contexts when formulating your impression.

=== CURRENT CLINICAL HISTORY/INDICATION ===
Male with end-stage renal disease on hemodialysis, multiple orthopedic hardware, bioprosthetic aortic valve replacement, coronary artery disease status post coronary artery bypass grafting, permanent pacemaker for sick sinus syndrome, admitted for MRSA bacteremia, transferred to the critical care unit for hypotension during anesthesia induction. Patient intubated for respiratory status assessment.

=== CURRENT TECHNIQUE ===
Chest single view

=== CURRENT COMPARISON ===
Prior imaging at an unspecified time.

=== CURRENT IMAGES ===
`[Images: Current chest X-ray images]`

Figure 20: **Impression generation example (Part 1)** in the C-SRRG-Impression dataset.

Example Case with Clinical Context - Part 2: Previous Study 1

=== PREVIOUS STUDY 1 (Most Recent) ===
Previous Clinical History/Indication:
Status post coronary artery bypass grafting (CABG), post chest tube removal.

Previous Technique:
Standard posteroanterior and lateral chest radiographs.

Previous Findings:
Lungs and Airways:
- No pneumothorax visualized
- Low lung volumes
- Increased infiltrate in the left upper lung

Pleura:
- Retrocardiac opacity
- Layering left pleural effusion

Cardiovascular:
- Vascular plethora

Other:
- Patchy areas of alveolar edema on the right
- Endotracheal (ET) tube positioned 4.3 cm above the carina
- Nasogastric (NG) tube within the stomach

Previous Impression:
1. Worsened infiltrate in the left upper lung suggestive of an acute process.
2. Presence of retrocardiac opacity and layering left pleural effusion.
3. Vascular plethora and patchy alveolar edema on the right, which may indicate fluid overload or cardiac etiology.
4. Proper positioning of the ET and NG tubes.

Previous Images:
`[Images: Previous chest X-ray images]`

Figure 21: **Impression generation example (Part 2)** in the C-SRRG-Impression dataset.

Example Case with Clinical Context - Part 3: Previous Study 2 & Response

=== PREVIOUS STUDY 2 ===
Previous Clinical History/Indication:
Patient presents with portions and exertion x.

Previous Technique:
The examination was performed using frontal and lateral views of the chest.

Previous Findings:
Tubes, Catheters, and Support Devices:
- Dual lead left-sided pacemaker extending to the expected positions of the right atrium and right ventricle.

Pleura:
- Slight blunting of the posterior costophrenic angles, which may suggest very trace pleural effusions.

Lungs and Airways:
- No focal consolidation observed.

Cardiovascular:
- The cardiac silhouette appears stable.

Hila and Mediastinum:
- The mediastinal silhouette is stable.

Musculoskeletal and Chest Wall:
- Right proximal humerus hardware is present but not well evaluated due to the imaging technique.

Other:
- Slight prominence of the interstitium, which may indicate minimal interstitial edema.

Previous Impression:
1. Possible very trace pleural effusions.
2. Minimal interstitial edema.
3. Stable cardiac and mediastinal silhouettes.
4. Presence of dual lead left-sided pacemaker in expected positions.
5. Right proximal humerus hardware present but not adequately assessed.

ASSISTANT RESPONSE:
IMPRESSION:
1. Worsened left basilar opacity, which may represent atelectasis or pneumonitis in the appropriate clinical context.
2. Mild improvement in pulmonary vascularity.

Figure 22: **Impression generation example (Part 3)** in the C-SRRG-Impression dataset.

Dataset Hallucination Example 1: Temporal Information Fabrication

Structured Report:
Exam Type: Chest radiograph.

Technique: Portable anteroposterior (AP) chest radiography was performed.

History: A male patient with hep C cirrhosis and large right pleural effusion status post thoracocentesis. Evaluate for resolution of pleural effusion.

Comparison: Prior portable AP chest radiograph

Findings:

Lungs and Airways:
- Mild inflation of the right upper lobe
- Collapsed right lower lobe
- No consolidation in the left lung

Pleura:
- Moderate pleural effusion within the right pleural space.
- Moderate right pneumothorax, new from prior exam.
- No left pleural effusion or pneumothorax.

Cardiovascular:
- No significant mediastinal shift observed.

Hila and Mediastinum:
- Mediastinum appears unremarkable

Impression:
1. Moderate right-sided pneumothorax.
2. Moderate right pleural effusion.
3. Inflation of the right upper lobe with collapse of the right lower lobe.
4. No mediastinal shift.

Hallucination: The phrase "**new from prior exam**" represents temporal information that cannot be verified from the current study alone, if not with previous history.

Figure 23: **Dataset hallucination example 1** in SRRG dataset.

Dataset Hallucination Example 2: Stability Assumption Without Comparison

Structured Report:
Exam Type: Chest radiograph

Technique: Standard frontal and lateral chest radiographic views were performed.

History: Atrial fibrillation (AF), coronary artery disease (CAD), congestive heart failure (CHF).

Comparison: Prior chest radiographs

Findings:

Cardiovascular:
- Mild to moderate cardiomegaly, unchanged.
- Tortuous but stable aorta.

Tubes, Catheters, and Support Devices:
- Transvenous pacemaker/AICD with leads terminating in the right atrium and right ventricle.
- Median sternotomy wires are aligned and intact.

Lungs and Airways:
- Lungs are clear with no evidence of consolidation, pleural effusion, pneumothorax, or overt pulmonary edema.

Impression:
1. No radiographic evidence for acute cardiopulmonary process.

Hallucination: The term "**unchanged**" implies comparison with prior studies, which is problematic without previous history.

Figure 24: **Dataset hallucination example 2** in SRRG dataset.

Dataset Hallucination Example 3: Generic Temporal Statement

Structured Report:
Exam Type: Chest radiograph

Technique: Single AP upright portable chest radiograph.

History: Shortness of breath.

Comparison: Prior chest radiograph

Findings:

Lungs and Airways:
- Low lung volumes
- Minimal bibasilar atelectasis
- Subcentimeter left lower lung rounded calcification, stable, possibly representing a calcified granuloma
- No focal consolidation
- No overt pulmonary edema

Pleura:
- No pleural effusion
- No evidence of pneumothorax

Cardiovascular:
- Stable cardiac and mediastinal silhouettes

Hila and Mediastinum:
- Ovoid calcification over the left mediastinum, stable

Tubes, Catheters, and Support Devices:
- Right-sided Port-A-Cath in place, terminating at the cavoatrial junction, without significant change in position

Musculoskeletal and Chest Wall:
- Status post median sternotomy

Impression:
1. No significant interval change.

Hallucination: The phrase "**No significant interval change**" is a generic temporal statement that, if not with prior studies, represents a non-specific conclusion that could mask important clinical details.

Figure 25: **Dataset hallucination example 3** in SRRG dataset.

Trained Model Hallucination Example 1: Temporal Comparisons in Findings

Sample ID: 12
Generated Report:
FINDINGS:
Tubes, Catheters, and Support Devices:
- Endotracheal tube terminates 4.5 cm above the carina.
- Right internal jugular central venous catheter terminates in the right atrium.
- Left pectoral pacemaker with leads in the right atrium and right ventricle.
- Median sternotomy wires are intact.

Cardiovascular:
- Moderate cardiomegaly, unchanged.

Lungs and Airways:
- Low lung volumes.
- No evidence of pneumothorax.

Pleura:
- Small left pleural effusion, unchanged.

Other:
- Retrocardiac opacity, likely representing atelectasis, unchanged.

Hallucination: Multiple instances of "**unchanged**" indicate temporal comparisons when only analyzing a single study.

Figure 26: **Trained model hallucination example 1 on test-reviewed split** using MedGemma-4B.

Trained Model Hallucination Example 2: Interval and Persistent Language

Sample ID: 60
Generated Report:
FINDINGS:
Pleura:
- Interval placement of a right-sided pigtail catheter.
- Decreased size of the right pleural effusion.
- No evidence of pneumothorax.

Lungs and Airways:
- Persistent right basilar opacity , likely representing atelectasis.
- Left lung is clear.

Cardiovascular:
- Stable cardiomediastinal silhouette.

Hallucination: Terms like "**Interval placement**", "**Decreased size**", and "**Persistent**" all imply temporal comparisons.

Figure 27: **Trained model hallucination example 2 on test-reviewed split** using MedGemma-4B.

Trained Model Hallucination Example 3: Explicit Temporal Language

Sample ID: 37
Generated Report:
FINDINGS:
Lungs and Airways:
- No significant interval change in the appearance of the lungs.

Pleura:
- Small right pleural effusion, unchanged.

Cardiovascular:
- Stable cardiomediastinal silhouette.

Tubes, Catheters, and Support Devices:
- Tracheostomy tube in place.
- Right internal jugular (IJ) dialysis catheter in place.
- Vascular stent in the right subclavian region.

Other:
- No evidence of pneumothorax.

Hallucination: Explicit temporal language ("**No significant interval change**", "**unchanged**") when no prior studies are available.

Figure 28: **Trained model hallucination example 3 on test-reviewed split** using MedGemma-4B.

Trained Model Hallucination Example 4: Unchanged Position in Impression

Sample ID: 0
Generated Report:
IMPRESSION:
1. No evidence of pneumothorax.
2. Stable small bilateral pleural effusions.
3. Unchanged position of the right chest tube.
4. Normal cardiomediastinal silhouette.
5. Presence of a vagal nerve stimulator.

Hallucination: The term "**Unchanged position**" suggests comparison with prior imaging that is not available.

Figure 29: **Trained model hallucination example 4 on test-reviewed split** using MedGemma-4B.

Trained Model Hallucination Example 5: Interval and Persistent Terms

Sample ID: 36
Generated Report:
IMPRESSION:
1. Interval increase in bilateral pleural effusions.
2. Persistent left lower lobe atelectasis.
3. Mild pulmonary edema.
4. Stable cardiomegaly.

Hallucination: Terms "**Interval increase**" and "**Persistent**" indicate temporal comparisons without access to prior studies.

Figure 30: **Trained model hallucination example 5 on test-reviewed split** using MedGemma-4B.

Trained Model Hallucination Example 6: New vs Unchanged Terminology

Sample ID: 40
Generated Report:
IMPRESSION:
1. New right lower lobe opacity , which may represent pneumonia or aspiration.
2. Small right pleural effusion.
3. Stable mild cardiomegaly.
4. Unchanged position of the right-sided central venous catheter.

Hallucination: The term "**Unchanged position**" implies knowledge of prior imaging studies that the model does not have access to. **New** also suggests temporal comparison but may be acceptable if referring to acute findings.

Figure 31: **Trained model hallucination example 6 on test-reviewed split** using MedGemma-4B.