CLINOPD: ON-POLICY DISTILLATION WITH CLINICAL HINTS FOR MEDICAL REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Automatic medical report generation aims to generate clinically accurate reports from medical images. Existing approaches learn the mapping from images to reference reports, but generated reports may still miss clinically important findings or introduce unsupported abnormalities. Inspired by the clinical reporting workflow, we introduce **Clinical Hints**, a compact signal that retains only the organs associated with positive findings. We find that providing such hints to the same model substantially improves report generation, suggesting that relevant clinical knowledge is often already encoded in the model but is not reliably activated from medical images alone. Based on this observation, we propose **ClinOPD**, an on-policy self-distillation framework that uses clinical hints as privileged supervision. A frozen hint-conditioned teacher provides token-level guidance along student-generated trajectories, while the student takes only the medical image as input and is regularized toward the original policy for stable optimization. No clinical hints are required at inference time. We evaluate clinical correctness using a factualness-driven F1 metric based on matched, missed, and unsupported findings, which shows stronger agreement with physician judgments than conventional report-generation metrics. Across the 2D IU-Xray and 3D CT-RATE datasets, ClinOPD improves F1 from 46.06 to 51.12 and from 23.65 to 30.06, corresponding to relative gains of **11.0%** and **27.1%** over SFT, respectively, while outperforming competing reinforcement-learning and on-policy distillation methods. Code will be made available.

1 INTRODUCTION

Medical report generation using AI aims to emulate the clinical reporting process by automatically translating medical images into clinically accurate reports (Jing et al., 2018; Tanno et al., 2025). Early approaches mainly relied on task-specific encoder–decoder architectures (Li et al., 2019; Liu et al., 2021), while recent advances in foundation models, large language models (LLMs), and multimodal large language models (MLLMs) have enabled increasingly powerful general vision–language capabilities to be transferred to this task (Bannur et al., 2024; Xu et al., 2026). Since zero-shot generation often remains insufficient for clinically accurate reporting (Zhou et al., 2024), a common practice is to further adapt MLLMs on paired medical images and reports through supervised fine-tuning (SFT). However, SFT primarily teaches the model to imitate reference reports and their writing patterns, without explicitly optimizing clinical correctness (Miura et al., 2021; Tanno et al., 2025). As a result, generated reports may reproduce stereotyped reporting patterns and, more importantly, still omit clinically relevant abnormalities or introduce unsupported findings (Yuan et al., 2026). Reinforcement learning (RL) offers a promising way to further improve clinical correctness, but medical report generation still lacks a well-established reward model that can reliably assess the clinical quality of generated reports, particularly missed abnormalities and unsupported findings (Yu et al., 2023). This raises a key question: *how can we provide fine-grained clinical supervision for post-training MLLMs to generate more clinically faithful reports?*

Radiologists review medical images by moving across organs and anatomical regions, identifying areas that require attention before describing the corresponding abnormalities (Rubin et al., 2015; Tanida et al., 2023). In other words, report generation is not only a process of recognizing findings, but also of deciding which regions and findings should enter the report (Bertram et al., 2016; Zhong et al., 2025). This motivates us to explicitly capture this intermediate anatomical focus as an

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

additional training signal, which we term **Clinical Hints**. For each ground-truth report, we retain only the organs associated with positive findings and remove the corresponding abnormalities, attributes, and diagnostic descriptions, so that the hints indicate *where to focus* without revealing *what to report*. To examine whether this signal is useful, we train a mixed-SFT model with equal proportions of image-only and hint-conditioned examples and evaluate the same checkpoint with and without hints. Under our clinical finding-level evaluation, providing clinical hints improves Clin-F1 by 57.4% on IU-Xray (Demner-Fushman et al., 2016) and 118.6% on CT-RATE (Hamamci et al., 2026). Since the model parameters are identical in both settings, this large improvement suggests that much of the knowledge needed to describe the abnormalities is already encoded in the model, while a key bottleneck lies in directing the model toward clinically relevant information.

Clinical hints are derived from ground-truth reports and cannot be used at inference time. We therefore propose **ClinOPD** to distill their benefit into a model that takes only the medical image as input. During training, the student generates reports from medical images alone, while a frozen copy of the same model additionally receives clinical hints and provides token-level supervision along student-generated trajectories. We also regularize the student toward the original SFT policy to maintain stable generation. At inference time, ClinOPD requires only the medical image. We evaluate ClinOPD on IU-Xray and CT-RATE, covering 2D chest radiography and 3D volumetric CT, respectively. ClinOPD consistently improves clinical report quality over SFT and outperforms competing post-training methods. Physician evaluation also shows that our factualness-driven metric, Clin-F1, better reflects report quality than conventional report-generation metrics.

Our main contributions are summarized as follows:

- **Clinical Hints.** We introduce clinical hints as a compact "where to focus" signal that retains only the organs associated with positive findings. Our controlled diagnostic study shows that such hints substantially improve medical report generation.
- **ClinOPD.** We propose an on-policy self-distillation framework that uses clinical hints as privileged supervision during training and distills their benefit into a model that requires only medical images. ClinOPD consistently improves clinical report quality over SFT and competing post-training methods across both 2D and 3D medical imaging.
- **Clinical factuality gains.** Across IU-Xray and CT-RATE, representing 2D and 3D medical imaging respectively, ClinOPD achieves Clin-F1 scores of **51.12** and **30.06**, corresponding to relative improvements of **11.0%** and **27.1%** over SFT, while outperforming competing reinforcement-learning and on-policy distillation methods.

2 Related Work

2.1 Automatic Medical Report Generation

Automatic medical report generation has been widely studied for both 2D and 3D medical images (Li et al., 2019; Zhang et al., 2020; Liu et al., 2021). Early approaches adopt task-specific encoder–decoder architectures, where visual features extracted by convolutional or Transformer-based encoders are translated into reports using recurrent or Transformer decoders (Jing et al., 2018; Wang et al., 2018). Subsequent studies improve report modeling through memory mechanisms (Chen et al., 2020), region-aware visual reasoning (Tanida et al., 2023), and the incorporation of structured clinical knowledge (Park et al., 2025), aiming to better capture clinically relevant findings and their relationships (Huang et al., 2025).

With the rapid development of LLMs and MLLMs, medical report generation has increasingly shifted toward vision–language models with stronger language priors and broader medical knowledge (Li et al., 2023; Wu et al., 2025). Representative models include radiology-oriented models such as MAIRA-2 (Bannur et al., 2024), as well as generalist medical MLLMs such as Hulu-Med (Jiang et al., 2025) and Lingshu (Xu et al., 2026), which extend report generation to a broader range of medical vision–language tasks. Existing studies typically either directly leverage the zero-shot capabilities of these models or adapt them to report generation through supervised fine-tuning on paired medical images and reports. However, as also observed in our experiments, neither strategy yields optimal clinical accuracy. Zero-shot generation lacks task-specific adaptation and may struggle to consistently produce reports that follow the expected reporting conventions. Supervised

fine-tuning alleviates these issues by teaching the model the basic structure and writing conventions of medical reports, but it still does not explicitly optimize the clinical correctness of the generated content. Consequently, reports may still omit clinically important findings or introduce unsupported abnormalities (Yuan et al., 2026). These limitations motivate further post-training of medical MLLMs, including reinforcement learning and distillation-based approaches (Qian et al., 2026).

2.2 On-Policy Distillation

Recent advances in post-training LLMs and MLLMs have increasingly relied on reinforcement learning (RL) (Shao et al., 2024; Guo et al., 2025). Early RL approaches optimize sparse, sequence-level rewards. Methods such as GRPO (Shao et al., 2024) improve reasoning through policy optimization over on-policy trajectories; however, their reward signals are shared across tokens within a response, resulting in coarse credit assignment. In contrast, on-policy distillation (OPD) (Agarwal et al., 2024) provides dense supervision at the token level. Specifically, OPD uses a teacher π_T to supervise a student π_θ through token-wise distributional feedback along student-generated trajectories. Given a dataset $\mathcal{D}$, for each input $x \sim \mathcal{D}$, the student samples a trajectory $\hat{y} \sim \pi_\theta(\cdot \mid x)$. A standard OPD objective can be written as

$$\mathcal{L}_{\mathrm{OPD}} = \mathbb{E}_{x\sim\mathcal{D},\, \hat{y}\sim\pi_\theta(\cdot|x)} \left[\sum_{t=1}^{|\hat{y}|} D_{\mathrm{KL}} \left(\pi_\theta(\cdot \mid x, \hat{y}_{<t}) \,\|\, \pi_T(\cdot \mid x, \hat{y}_{<t}) \right) \right]. \tag{1}$$

Here, $\hat{y}_{<t}$ denotes the student-generated prefix before step t, and D_{KL} measures the discrepancy between the student and teacher distributions. While standard OPD relies on a separate teacher model, recent work explores whether such dense on-policy supervision can be obtained from the model itself. On-Policy Self-Distillation (OPSD) (Zhao et al., 2026; Pan et al., 2026) follows this direction by constructing the teacher from the same model and conditioning it on privileged information y^*, while the student observes only the original input. Its objective becomes

$$\mathcal{L}_{\mathrm{OPSD}} = \mathbb{E}_{x\sim\mathcal{D},\, \hat{y}\sim\pi_\theta(\cdot|x)} \left[\sum_{t=1}^{|\hat{y}|} D_{\mathrm{KL}} \left(\pi_\theta(\cdot \mid x, \hat{y}_{<t}) \,\|\, \pi_T(\cdot \mid x, y^*, \hat{y}_{<t}) \right) \right], \tag{2}$$

where π_T denotes the privileged teacher policy. However, in our experiments, we find that the default OPSD strategy of using the ground-truth report as privileged information is inefficient for medical report generation. Inspired by the reporting workflow of radiologists, who typically identify clinically salient abnormal regions before composing the final report, we introduce **ClinOPD**, which uses clinical hints as privileged supervision for medical image report generation.

3 ClinOPD

Figure 1 provides an overview of ClinOPD. Our framework consists of two stages. We first construct *clinical hints* by extracting the organs associated with positive findings from ground-truth reports, and use image-only and hint-conditioned samples jointly to obtain a mixed-SFT model. We then perform on-policy self-distillation from a frozen hint-conditioned teacher to a hint-free student initialized from the same mixed-SFT checkpoint. During training, the teacher provides token-level guidance along student-generated trajectories, while a frozen SFT reference model regularizes the student against excessive policy drift. At inference time, only the medical image is required, and the resulting student generates reports without access to clinical hints.

3.1 Clinical Hints

A common failure in medical report generation is that the model either misses abnormalities in clinically important organs, or generates unsupported findings in irrelevant regions. This suggests that part of the difficulty lies in deciding *where to focus* before describing *what is abnormal*. However, standard OPSD-style supervision provides the entire report as privileged information, where the clinically relevant signal can be diluted by abundant descriptive content. Motivated by this, we introduce *clinical hints* that explicitly indicate the organs requiring attention during report generation.

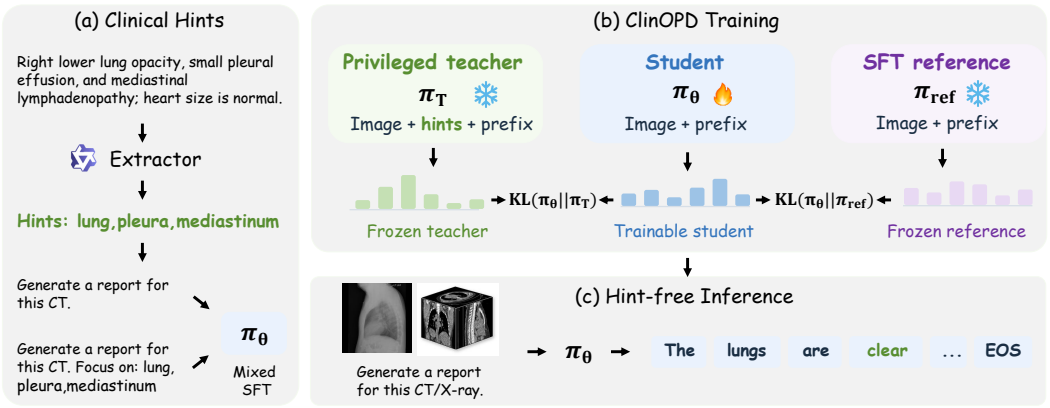

Figure 1: Overview of ClinOPD. (a) We extract *clinical hints* from positive findings, retaining only the corresponding organ names for mixed SFT. (b) During ClinOPD training, a frozen hint-conditioned teacher distills its clinical guidance into a hint-free student through token-level on-policy supervision, while a frozen reference model regularizes policy drift. (c) The resulting student performs hint-free medical report generation at inference time.

Hints construction. For each image–report pair (x, y), we use an LLM-based extractor $\mathcal{E}$ to parse the ground-truth report and identify the organs mentioned in positive findings:

$$h = \mathcal{E}(y) = \{o_1, o_2, \ldots, o_m\}, \tag{3}$$

where o_i denotes an organ associated with at least one positive finding in y. We retain only the organ names and discard the corresponding abnormalities, attributes, and diagnostic descriptions. The resulting hint is appended to the input through a short instruction such as *"Focus on the following organs: liver, lung, kidney."* Therefore, h specifies *where to focus* rather than *what to report*. Compared with directly providing the full ground-truth report as privileged information, this representation isolates a compact clinical signal that can guide the model toward potentially abnormal regions without exposing the target description itself. Further details of the hint extraction procedure are provided in Appendix A.

Diagnostic study. We first examine whether these clinical hints provide effective supervision for report generation. We adopt a mixed SFT setting where, for each training sample, we randomly use either the standard image-only input x, e.g., *'Generate a radiology report for this CT."*, or the hint-conditioned input (x, h), e.g., *'Generate a radiology report for this CT. Focus on the following organs: liver, lung, kidney."*, with equal probability. This results in a mixed training set $\mathcal{D}_{\text{mix}}$ containing 50% image-only and 50% hint-conditioned samples. The SFT objective is

$$\mathcal{L}_{\text{SFT}} = \mathbb{E}_{(\tilde{x},y)\sim\mathcal{D}_{\text{mix}}}\left[-\sum_{t=1}^{|y|}\log \pi_\theta\left(y_t \mid \tilde{x}, y_{<t}\right)\right], \tag{4}$$

where $\tilde{x}$ denotes either x or (x, h).

At inference, we evaluate the same model under image-only and hint-conditioned settings:

$$\hat{y}_{\text{img}} \sim \pi_\theta(\cdot \mid x), \qquad \hat{y}_{\text{hint}} \sim \pi_\theta(\cdot \mid x, h). \tag{5}$$

To isolate the effect of clinical hints, we compare image-only and hint-conditioned generation from the same SFT checkpoint. We conduct this study on IU X-ray and CT-RATE, representing 2D radiography and 3D volumetric CT, respectively, and use Clin-F1 to measure clinical correctness by jointly accounting for missed and unsupported findings; the datasets, splits, and evaluation protocol follow Section 4.1. As shown in Table 1, providing clinical hints at inference time raises Clin-F1 by 26.44 points on IU X-ray and 28.05 points on CT-RATE, with improvements in both clinical precision and recall. These results motivate using clinical hints as privileged supervision.

Table 1: Effect of clinical hints on IU-Xray and CT-RATE.

Dataset	Setting	Precision	Recall	F1
IU-Xray	$\hat{y}_{\text{img}}$	48.97	46.81	46.06
	$\hat{y}_{\text{hint}}$	**73.86**	**74.03**	**72.50**
	Δ	+24.89	+27.22	+26.44
CT-RATE	$\hat{y}_{\text{img}}$	32.26	22.12	23.65
	$\hat{y}_{\text{hint}}$	**57.77**	**50.34**	**51.70**
	Δ	+25.51	+28.22	+28.05

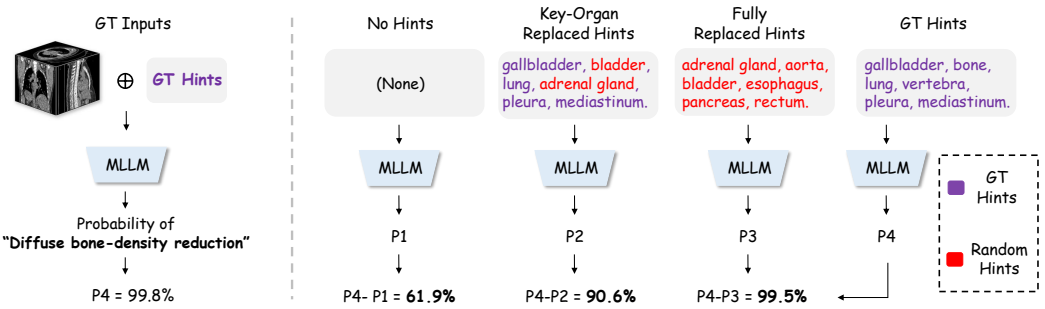

Figure 2: Controlled intervention analysis of clinical hints for "Diffuse bone-density reduction." Given the same CT scan and model, we vary only the clinical hints and measure the change in the probability of generating the target finding before and after each hint intervention.

Why do clinical hints help? Figure 2 provides a representative example. Without hints, the model mainly reports pulmonary abnormalities and misses the ground-truth finding "Diffuse bone-density reduction." Providing the relevant organ hints recovers this finding. To isolate the effect of the hints, we keep the CT scan and model fixed and vary only the hints. For the same target finding, we compare four settings: no hints (P_1); key-organ replaced hints, where "bone, vertebra" are replaced by "bladder, adrenal gland" while the other organs remain unchanged (P_2); fully replaced hints, where all ground-truth organs are replaced by an equally long random organ list (P_3); and the ground-truth hints "gallbladder, bone, lung, vertebra, pleura, mediastinum" (P_4). Relative to P_1, P_2, and P_3, P_4 increases the probability of generating the target finding by 61.9, 90.6, and 99.5 points, respectively. These controlled interventions suggest that the improvement is driven by clinically relevant anatomical information, particularly the inclusion of "bone" and "vertebra." More examples and analyses are provided in Appendix G.

3.2 On-Policy Self Distillation Using Clinical Hints

The diagnostic study shows that clinical hints substantially improve medical report generation, suggesting that they expose useful clinical knowledge already encoded in the model. However, such hints are unavailable at deployment time. We therefore introduce **ClinOPD** to internalize their benefit through on-policy self-distillation.

ClinOPD uses a hint-conditioned teacher prompt and a hint-free student prompt. Starting from the mixed-SFT model, the model conditioned on x first generates a report $\hat{y} \sim \pi_\theta(\cdot \mid x)$. Conditioned additionally on h, it then serves as a privileged teacher, providing token-level supervision to the policy without access to h. Specifically, we minimize

$$\mathcal{L}_{\text{ClinOPD}} = \mathbb{E}_{(x,h)\sim\mathcal{D},\, \hat{y}\sim\pi_\theta(\cdot|x)} \left[\sum_{t=1}^{|\hat{y}|} D_{\text{KL}} \left(\pi_\theta(\cdot \mid x, \hat{y}_{<t}) \,\|\, \pi_T(\cdot \mid x, h, \hat{y}_{<t}) \right) \right], \tag{6}$$

where π_T denotes the privileged teacher distribution. Because supervision is evaluated directly on trajectories sampled from the current policy, ClinOPD continuously transfers the hint-induced preference toward clinically informative tokens into the hint-free policy.

Table 2: Agreement between automatic evaluation metrics and physician judgments.

	BLEU-1	BLEU-4	ROUGE-L	METEOR	CIDEr	Clin-F1
Spearman's ρ	0.3289	0.4070	0.4200	0.4281	0.3344	**0.5576**
Kendall's τ	0.2942	0.3683	0.3705	0.3810	0.2998	**0.5190**
Top-1 Acc.	55.71%	57.79%	52.88%	61.83%	53.71%	**71.13%**

In practice, we observe that optimizing the distillation objective alone can cause excessive policy drift, occasionally leading to abnormally long or repetitive reports. To stabilize training, we further regularize the updated policy toward a frozen copy of the mixed-SFT model under the same hint-free context. Let π_{ref} denote this frozen reference model. The regularization term is

$$\mathcal{L}_{\text{reg}} = \mathbb{E}_{x\sim\mathcal{D},\, \hat{y}\sim\pi_\theta(\cdot|x)} \left[\sum_{t=1}^{|\hat{y}|} D_{\text{KL}} \left(\pi_\theta(\cdot \mid x, \hat{y}_{<t}) \,\|\, \pi_{\text{ref}}(\cdot \mid x, \hat{y}_{<t}) \right) \right]. \tag{7}$$

The final training objective is

$$\mathcal{L} = \mathcal{L}_{\text{ClinOPD}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}, \tag{8}$$

where λ_{reg} controls the trade-off between absorbing privileged clinical supervision and preserving the stable generation behavior of the SFT model.

4 Experiments

4.1 Experimental Setup

Datasets We evaluate ClinOPD on two complementary report-generation benchmarks spanning both 2D and 3D medical imaging. IU X-ray (Demner-Fushman et al., 2016) consists of 2D chest radiographs paired with radiology reports, while CT-RATE (Hamamci et al., 2026) provides 3D volumetric chest CT scans with corresponding reports. This setting allows us to examine whether ClinOPD generalizes across different visual representations, from 2D projection images to 3D volumetric inputs. For IU X-ray, we follow the data split adopted by ClinFusion (Yuan et al., 2026), while for CT-RATE, we use the official patient-level training and validation split.

Evaluation Metrics. For evaluation, we focus on the clinical correctness of generated reports rather than their surface-level similarity to the GT. We therefore do not use conventional report-generation metrics, such as BLEU (Papineni et al., 2002), as the primary metrics in our main experiments, since they mainly capture lexical or n-gram overlap and do not distinguish correctly identified findings from clinically important omissions or unsupported findings. Inspired by ClinFusion (Yuan et al., 2026), we introduce **Clin-F1**, a clinically grounded metric that evaluates generated reports at the *clinical finding level*. Specifically, an LLM extracts clinical findings from both the generated and reference reports and associates them with their corresponding anatomical regions. Each finding is then categorized as *matched* (true positive; TP), *missed* (false negative; FN), or *hallucinated* (false positive; FP). Based on these categories, we compute precision, recall, and their harmonic mean, Clin-F1, independently for each case, and report their macro-averages over the evaluation set. Detailed definitions and the complete computation procedure are provided in Appendix E.

To examine whether Clin-F1 better reflects physician judgments, we randomly sample 100 cases from each dataset, yielding a 200-case evaluation subset, and invite two experienced clinicians to independently assess the generated reports. On this subset, we additionally compare Clin-F1 with conventional metrics, including BLEU-1 (Papineni et al., 2002), BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), CIDEr (Vedantam et al., 2015), and METEOR (Banerjee & Lavie, 2005). We quantify the agreement between each automatic metric and physician ratings using Spearman's ρ (Spearman, 1904), Kendall's τ (Kendall, 1945), and Top-1 accuracy. As shown in Table 2, Clin-F1 exhibits consistently stronger agreement with physician assessments than the conventional metrics. Additional details are provided in Appendix F.

Table 3: Report generation performance on IU-Xray and CT-RATE.

Method	IU-Xray			CT-RATE		
	Precision	Recall	F1	Precision	Recall	F1
GPT-5.5	35.28	41.96	35.76	22.97	22.79	21.15
Qwen3.7-Plus	39.05	36.95	36.99	9.60	7.81	7.93
Qwen3-VL-Flash	33.96	34.67	32.91	10.98	12.88	10.96
Qwen3-VL-Plus	22.54	22.77	21.39	12.46	14.77	12.02
Qwen3.5-9B	39.25	36.75	37.07	8.61	7.85	7.75
Lingshu-7B	42.51	37.68	38.74	3.84	3.64	3.60
SFT	48.97	46.81	46.06	32.26	22.12	23.65
+ GRPO	53.24	51.33	50.50	34.04	22.51	24.54
+ DAPO	50.16	46.16	46.64	36.05	25.31	26.97
+ OPSD	50.35	45.79	46.11	34.19	26.74	27.39
+ ClinOPD	**54.81**	**51.65**	**51.12**	**38.10**	**28.65**	**30.06**

Baselines. We use Qwen3.5-9B (Team, 2026) as the backbone for all post-training experiments. Before post-training, we train the backbone using the mixed-SFT scheme introduced in Section 3.1, where image-only and hint-conditioned samples are used with equal probability. We refer to the resulting checkpoint as *Mixed SFT* or *SFT*.

Starting from the same Mixed-SFT checkpoint, we compare the reinforcement-learning-based methods GRPO (Shao et al., 2024) and DAPO (Yu et al., 2025), the on-policy self-distillation baseline OPSD (Zhao et al., 2026), and our ClinOPD. We additionally report results from GPT-5.5 (OpenAI, 2026), Qwen3.7-Plus (Alibaba Cloud, 2026), Qwen3-VL-Flash (Bai et al., 2025), Qwen3-VL-Plus (Bai et al., 2025), and Lingshu-7B (Xu et al., 2026) as external baselines.

Implementation Details. We fine-tune all parameters of the student model while keeping the teacher model frozen throughout training. Thinking mode is disabled for both models. We train for 100 optimization steps with a learning rate of 1×10^{-6}. Each rollout batch contains eight input examples, with eight responses sampled per example. For the RL baselines, GRPO and DAPO, we use the case-level Clin-F1 score between each generated report and its corresponding ground-truth report as the sequence-level reward. The reward is computed using the same finding-level evaluation procedure described above, thereby providing the RL baselines with a clinically grounded objective while keeping the supervision signal consistent across methods. In contrast, ClinOPD does not optimize this scalar reward directly, but instead receives token-level distributional supervision from the privileged teacher along its on-policy trajectories. Unless otherwise specified, all remaining hyperparameters follow the original OPSD (Zhao et al., 2026) setup. For IU-Xray, all available views from the same study are jointly provided as multi-image input. For CT, we uniformly sample 32 slices and apply windowing with a level of 40 and a width of 400.

4.2 Main Results

Quantitative results. Table 3 reports the results under the factuality-driven F1 evaluation protocol. ClinOPD achieves the strongest overall performance on both datasets. On IU-Xray, it obtains 54.81 precision, 51.65 recall, and 51.12 F1. Compared with SFT, this corresponds to gains of 5.84, 4.84, and 5.06 points in precision, recall, and F1, respectively, with the F1 gain corresponding to an **11.0%** relative improvement. ClinOPD also exceeds GRPO, the strongest competing post-training method on this dataset, by 0.62 F1 points. The improvement is more pronounced on CT-RATE, where ClinOPD achieves 38.10 precision, 28.65 recall, and 30.06 F1. Relative to SFT, precision, recall, and F1 improve by 5.84, 6.53, and 6.41 points, respectively, with the F1 gain corresponding to a **27.1%** relative improvement. ClinOPD further surpasses OPSD, the strongest competing post-training method on CT-RATE, by 2.67 F1 points.

Case study. Figure 3 presents two deliberately selected cases to illustrate how ClinOPD improves finding-level fidelity over SFT. In IU-Xray, SFT produces a generic normal report and omits a

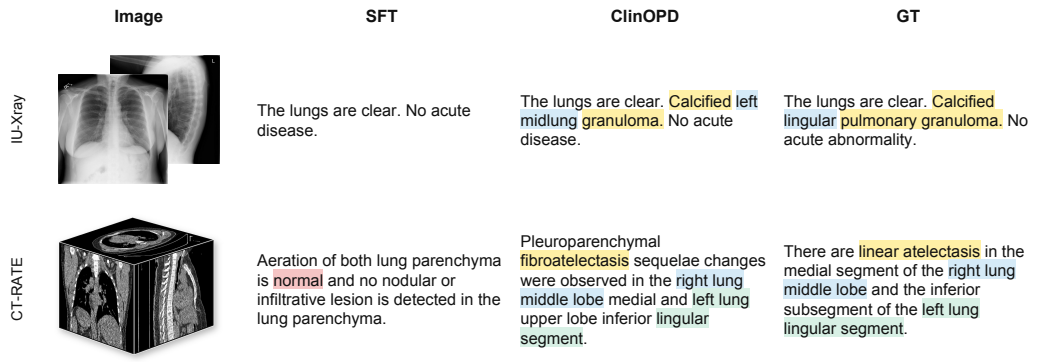

Figure 3: Case studies comparing radiology reports generated by different methods on IU-Xray (top) and CT-RATE (bottom). Yellow highlights clinically relevant positive findings, while blue and green highlight the corresponding anatomical regions of interest (RoIs), with different colors distinguishing different RoIs. Red highlights incorrect or contradictory findings.

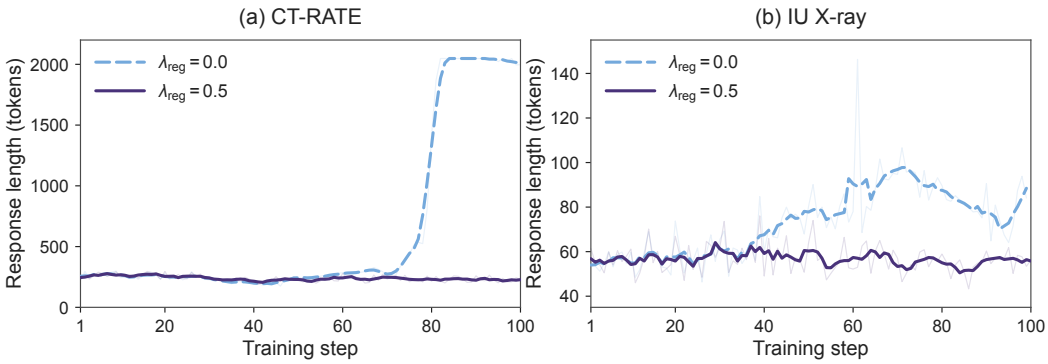

Figure 4: Response lengths during training on CT-RATE and IU X-ray.

chronic calcified pulmonary granuloma, whereas ClinOPD recovers the finding and its left-lung location, broadly matching the reference, although "left midlung" is less specific than the reference location "lingular." In CT-RATE, SFT incorrectly reports normal bilateral aeration, while ClinOPD identifies bilateral fibroatelectatic changes in the right middle lobe and left lingular segment, consistent with the reference linear atelectasis.

4.3 Ablation Studies

We investigate the effect of the regularization coefficient λ_{reg} from two perspectives: generation stability and downstream clinical performance.

Generation stability. Without $\mathcal{L}_{\text{reg}}$, response length grows markedly during training (Figure 4), reaching over $2{,}000$ tokens on CT-RATE and showing increasing fluctuations on IU-Xray. Figure 5 shows that this growth is accompanied by repetitive and unsupported generation. In contrast, $\lambda_{\text{reg}} = 0.5$ keeps the outputs stable at about 229 tokens on CT-RATE and 55 on IU-Xray, indicating that the regularization term is important for preventing generation degeneration. Additional examples are provided in Appendix C.

Sensitivity to the regularization strength. We evaluate $\lambda_{\text{reg}} \in \{0, 0.01, 0.5, 1.0\}$ on IU-Xray. As shown in Figure 6, weak regularization leads to late-stage degradation, whereas $\lambda_{\text{reg}} = 0.5$ achieves the best validation performance. We therefore use $\lambda_{\text{reg}} = 0.5$ in all experiments.

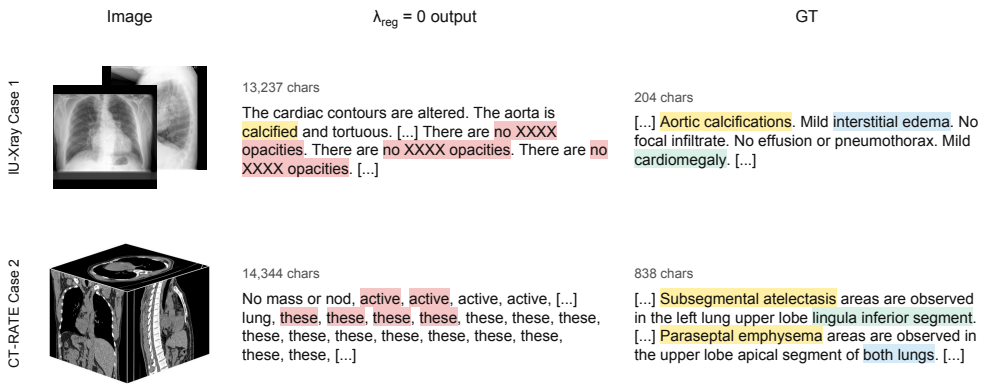

Figure 5: Examples of generation degeneration without regularization on IU-Xray and CT-RATE.

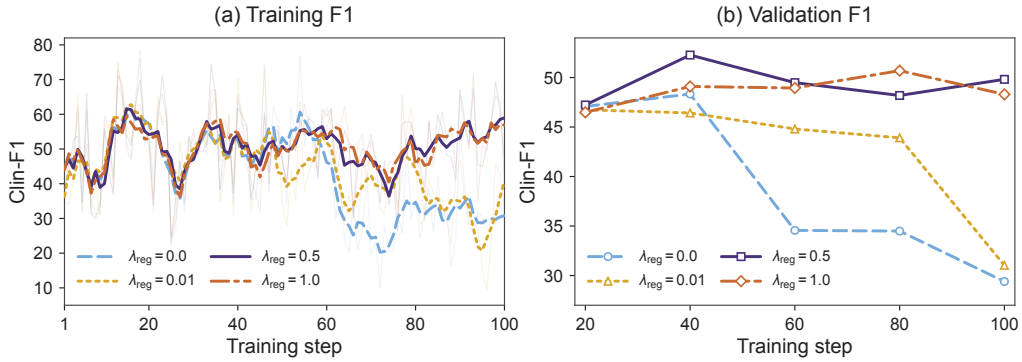

Figure 6: Training and validation Clin-F1 trajectories on IU X-ray under different λ_{reg}.

5 Conclusion

In this work, we introduced **ClinOPD**, an on-policy self-distillation framework for clinically faithful medical report generation. Motivated by the clinical reporting workflow, we proposed **clinical hints** as a compact "where to focus" signal that retains only the organs associated with positive findings. ClinOPD uses these hints as privileged supervision during training and distills their benefit into a model that requires only medical images at inference time. Across both 2D and 3D medical imaging, ClinOPD consistently improves over SFT and competing post-training methods, achieving relative finding-level F1 gains of 11.0% on IU-Xray and 27.1% on CT-RATE over SFT. Further analysis shows better coverage of clinically relevant abnormalities and fewer missed findings, supporting the effectiveness of transferring anatomical focus from a privileged teacher to a hint-free student. Overall, our results highlight the value of compact, task-structured privileged information for post-training medical MLLMs without introducing additional inputs at inference time. Future work could explore richer forms of clinical hints beyond organ-level guidance and extend ClinOPD to broader imaging modalities and clinical reporting settings.

AI use statement

Generative AI tools were used to assist with language polishing, including improving grammar, clarity, and readability of the manuscript. All technical content, research ideas, methodology, experiments, analyses, and conclusions were developed and verified by the authors. All AI-assisted edits were carefully reviewed by the authors, who take full responsibility for the content of this work.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.

Alibaba Cloud. Qwen3.7-plus. Alibaba Cloud Model Studio, 2026. URL https://www.alibabacloud.com/help/en/model-studio/qwen3-7-plus. Accessed: 2026-09-21.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Noel C. F. Codella, Fabian Falck, Ozan Oktay, Matthew P. Lungren, Maria Teodora Wetscherek, Javier Alvarez-Valle, and Stephanie L. Hyland. Maira-2: Grounded radiology report generation, 2024. URL https://arxiv.org/abs/2406.04449.

Raymond Bertram, Johanna Kaakinen, Frank Bensch, Laura Helle, Eila Lantto, Pekka Niemi, and Nina Lundbom. Eye movements of radiologists reflect expertise in ct study interpretation: a potential tool to measure resident development. *Radiology*, 281(3):805–815, 2016.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1439–1449, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.112. URL https://aclanthology.org/2020.emnlp-main.112/.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan

Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, et al. Generalist foundation models from a multimodal dataset for 3d computed tomography. *Nature Biomedical Engineering*, 10(8):1610–1628, 2026. doi: 10.1038/s41551-025-01599-y. URL https://doi.org/10.1038/s41551-025-01599-y.

Xiaofei Huang, Wenting Chen, Jie Liu, Qisheng Lu, Xiaoling Luo, and Linlin Shen. Damper: a dual-stage medical report generation framework with coarse-grained mesh alignment and fine-grained hypergraph matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3769–3778, 2025.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025.

Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2577–2586, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1240. URL https://aclanthology.org/P18-1240/.

Maurice G Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945.

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

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

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

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation . In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13748–13757, Los Alamitos, CA, USA, June 2021. IEEE Computer Society. doi: 10.1109/CVPR46437.2021.01354. URL https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.01354.

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

OpenAI. Gpt-5.5 system card. OpenAI, April 2026. URL https://openai.com/index/gpt-5-5-system-card/.

Leyi Pan, Shuchang Tao, Yunpeng Zhai, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Aiwei Liu, and Lijie Wen. Rlcsd: Reinforcement learning with contrastive on-policy self-distillation, 2026. URL https://arxiv.org/abs/2606.11709.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Sang-Jun Park, Keun-Soo Heo, Dong-Hee Shin, Young-Han Son, Ji-Hye Oh, and Tae-Eui Kam. Dart: Disease-aware image-text alignment and self-correcting re-alignment for trustworthy radiology report generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15580–15589. IEEE, 2025.

Yunhang Qian, Jiaquan Yu, Jiawei Liu, Meng Wang, Hongwei Bran Li, and Xiaobin Hu. Med-opd: Improving medical vision-language models via evidence-aware on-policy distillation, 2026. URL https://arxiv.org/abs/2607.16303.

Geoffrey D Rubin, Justus E Roos, Martin Tall, Brian Harrawood, Sukantadev Bag, Donald L Ly, Danielle M Seaman, Lynne M Hurwitz, Sandy Napel, and Kingshuk Roy Choudhury. Characterizing search, recognition, and decision in the detection of lung nodules on ct scans: elucidation with eye tracking. *Radiology*, 274(1):276–286, 2015.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159.

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

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

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

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

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9049–9058. IEEE, 2018.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866, 2025.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: Generalist foundation model for unified multimodal medical understanding and reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

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

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Hangjie Yuan, Yichen Qian, Zhiwei Tang, Xianzhe Xu, Lirong Wu, Sicheng Yang, Jinwang Wang, Pengju Wang, Zhitao Zeng, Yizeng Han, Yan Xing, Shengxuan Luo, Tao Feng, Qing Xie, Weigen Yao, Yi Yang, Zuozhu Liu, Jiasheng Tang, Shaocheng Wang, Jitao Wang, Jiahong Dong, Weihua Chen, Feng Xu, and Fan Wang. Clinfusion: A vision-centric multimodal llm system for holistic medical understanding, 2026. URL https://arxiv.org/abs/2607.24743.

Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. When radiology report generation meets knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 12910–12917, 2020.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Zhusi Zhong, Yuli Wang, Jing Wu, Wen-Chi Hsu, Vin Somasundaram, Lulu Bi, Shreyas Kulkarni, Zhuoqi Ma, Scott Collins, Grayson Baird, et al. Vision-language model for report generation and outcome prediction in ct pulmonary angiogram. *NPJ digital medicine*, 8(1):432, 2025.

Yiliang Zhou, Hanley Ong, Patrick Kennedy, Carol C Wu, Jacob Kazam, Keith Hentel, Adam Flanders, George Shih, and Yifan Peng. Evaluating gpt-4v (gpt-4 with vision) on detection of radiologic findings on chest radiographs. *Radiology*, 311(2):e233270, 2024.

A Clinical Hints Extraction Details

This section details the implementation of the extractor $\mathcal{E}$ used in Section 3.1. Clinical hints are extracted only from ground-truth reports and do not use the generated report.

Positive finding extraction. Given a ground-truth report y, we first use a structured extractor $\mathcal{E}_{\mathrm{str}}$ to identify the set of positive findings:

$$\mathcal{Z}(y) = \mathcal{E}_{\mathrm{str}}(y) = \{(o_i, f_i, e_i)\}_{i=1}^{K}, \tag{9}$$

where o_i, f_i, and e_i denote the anatomical structure, abnormal finding, and supporting sentence, respectively. An item is included only when the report explicitly asserts an abnormality in a specific anatomical structure. Normal or absent findings, examination limitations, recommendations, and statements without an independently identifiable abnormality are excluded. The extraction prompt template used for both CT-RATE and IU X-ray is provided in Figure 7.

The final clinical hint retains only the unique anatomical structures associated with these positive findings:

$$h = \mathcal{E}(y) = \mathrm{Unique}\left(o_1, o_2, \ldots, o_K\right), \tag{10}$$

where $\mathrm{Unique}(\cdot)$ removes duplicate structures. The corresponding findings f_i and supporting sentences e_i are discarded before constructing the training prompt.

Extractor implementation. We instantiate $\mathcal{E}$ with Qwen3.5-Plus (Team, 2026), using temperature 0 for both CT-RATE and IU X-ray. The two datasets follow the same extraction rules and structured output format, with only the modality-specific report description changed in the prompt.

B Additional Case Studies

Figure 8 provides additional qualitative comparisons on IU-Xray and CT-RATE, illustrating several common failure modes of the SFT model, including missed abnormalities, unsupported findings, and incomplete characterization of clinically relevant findings.

On IU-Xray, ClinOPD more reliably captures findings that are missed or incorrectly described by SFT. In Case 1, SFT reports an unremarkable mediastinum and clear lungs, whereas ClinOPD identifies the tortuous aorta described in the ground-truth report. In Case 2, ClinOPD captures the reduced lung volume that is absent from the SFT output. Case 3 illustrates a different failure mode: SFT introduces streaky bibasilar opacities unsupported by the reference, while ClinOPD produces a report consistent with the normal lung findings.

A similar pattern is observed on CT-RATE. Despite corresponding to different CT inputs, the SFT predictions in Cases 1 and 2 collapse to the same generic negative description, stating that no nodular or infiltrative lesion is present. In contrast, ClinOPD identifies the small pulmonary nodules described in the corresponding reference reports. In Case 3, ClinOPD similarly recovers the bilateral pulmonary nodules missed by SFT, although the accompanying hiatal hernia is not captured. Across these examples, ClinOPD improves the coverage of clinically relevant abnormalities and reduces unsupported findings.

C Examples of Generation Degeneration

Figure 9 provides examples of the generation degeneration observed when the reference regularization is removed, *i.e.*, $\lambda_{\mathrm{reg}} = 0$. We show two cases from IU X-ray and two cases from CT-RATE, together with their corresponding ground-truth reports. The character counts of the generated and reference reports are also provided in the figure to illustrate the magnitude of the length expansion.

Across both datasets, the excessive increase in response length does not correspond to additional clinically meaningful content. Instead, the unregularized model frequently enters repetitive generation patterns. For example, the IU X-ray outputs repeatedly produce placeholder-like phrases such as "XXXX opacities," while the CT-RATE outputs exhibit more severe phrase-level loops involving repeated expressions such as "no active," "no," and "these." Consequently, outputs containing more than $8{,}000$ characters can correspond to reference reports of only a few hundred characters.

Clinical Hints Extraction Prompt

System Prompt

You are a senior radiologist extracting structured abnormal findings from a CT/X-ray report. Answer with a strict JSON object only.

User Prompt

CT/X-ray report:
"""{report}"""

Extract every organ / anatomical structure that has a POSITIVE (abnormal) finding in this report.

What counts as a positive finding:
- The sentence asserts that an abnormality is present in a specific anatomical structure.

What to exclude:
- Sentences asserting that a structure is normal, or that an abnormality is absent.
- Sentences whose point is a limitation of the examination itself - coverage, technique, or diagnostic confidence. This exclusion still applies when such a sentence also names a possible abnormality, if the sentence states that the abnormality cannot be established under those conditions.
- Sentences that refer back to something described elsewhere without naming the abnormality themselves, so that the finding cannot be identified from the sentence alone.
- Recommendations, differential questions, and statements that a prior appearance is unchanged.

How to fill each field:
- "organ": the single anatomical structure the abnormality is located in, as one lowercase noun. Strip laterality, lobe, segment and any descriptive modifier. Choose the structure the abnormality is in, not a neighbouring one.
- "finding": the standard radiological name of the abnormality, lowercase, one to three words. Normalize it: drop severity and extent qualifiers, drop measurements, drop location, drop temporal and non-specificity qualifiers, and use the conventional term for the entity rather than the report's descriptive phrasing. Two reports describing the same entity must yield the same string. Do not escalate: stay at the level of certainty the sentence expresses, and never name a diagnosis or etiology the sentence does not state.
- "evidence": one sentence copied verbatim from the report.

Emit one entry per distinct abnormality. If one structure has several distinct abnormalities, emit several entries.

Return JSON with exactly this shape:

```
{
  "items": [
    {"organ": "<single anatomical noun>",
      "finding": "<short lowercase abnormality>",
      "evidence": "<verbatim sentence from the report>"}
  ]
}
```

Return an empty items list if the report describes no abnormality at all.

Figure 7: Prompt used for clinical hints extraction.

Together with Figure 4, these cases indicate that the length increase observed without $\mathcal{L}_{\text{reg}}$ is mainly caused by repetitive and degenerated generation, rather than by the inclusion of more useful clinical information. Regularizing the policy toward the SFT reference therefore helps prevent this failure mode during ClinOPD training.

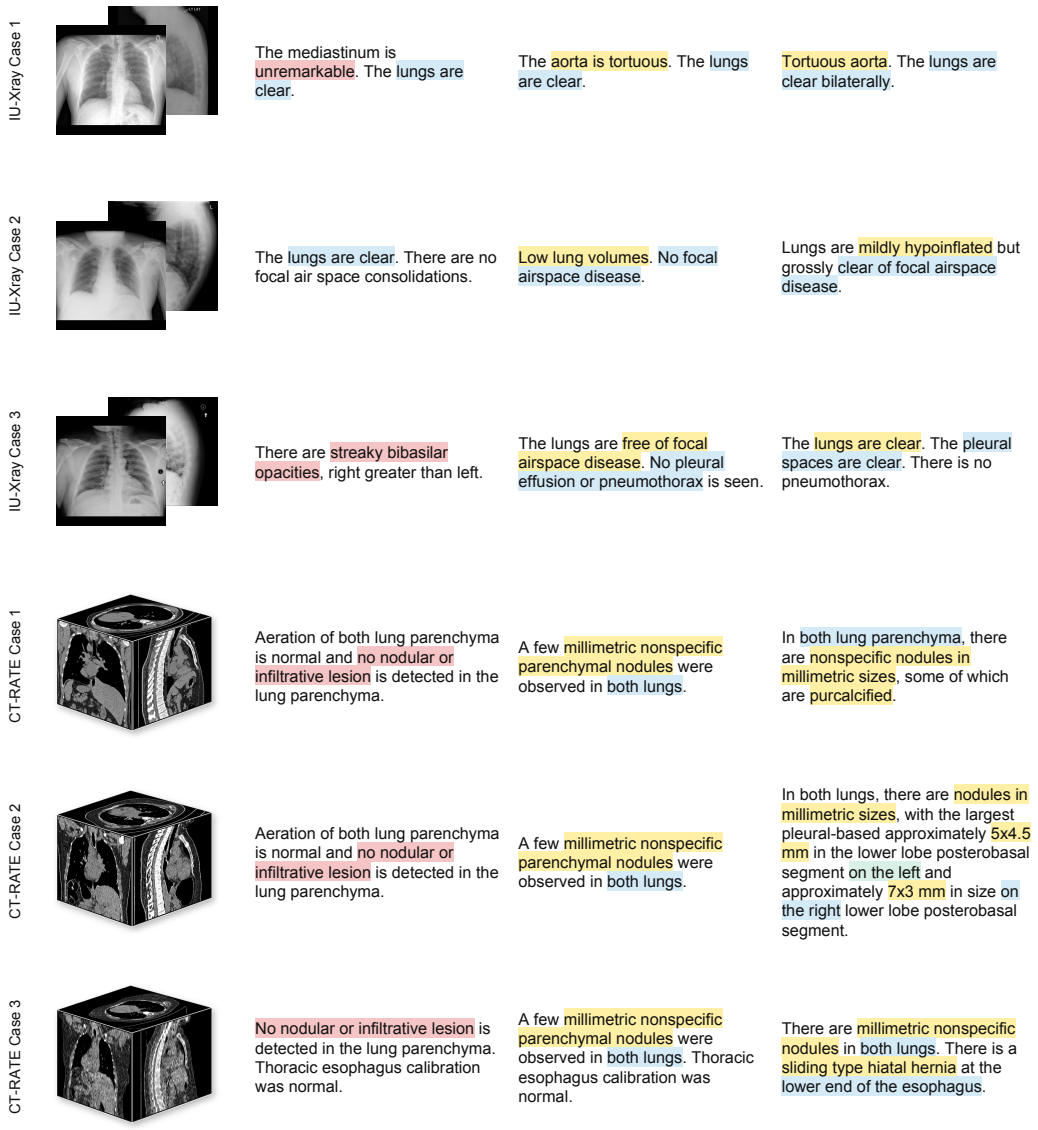

Figure 8: Additional comparisons on IU-Xray (top) and CT-RATE (bottom). Columns show the image, SFT prediction, ClinOPD prediction, and ground truth. Yellow highlights findings, blue and green indicate anatomical details, and red marks incorrect statements.

D OPSD Baseline Details

We provide additional details of the OPSD baseline. As illustrated in Figure 10, the privileged teacher receives both the medical image and the ground-truth report, while the student receives only the medical image. The frozen teacher provides token-level supervision along student-generated trajectories. To ensure a controlled comparison, OPSD and ClinOPD use the same frozen reference policy and reference KL regularization ($\lambda_{\text{reg}} = 0.5$). During distillation, the key difference lies in the privileged information provided to the teacher: the full ground-truth report for OPSD and clinical hints for ClinOPD. At inference time, both models generate reports from medical images alone.

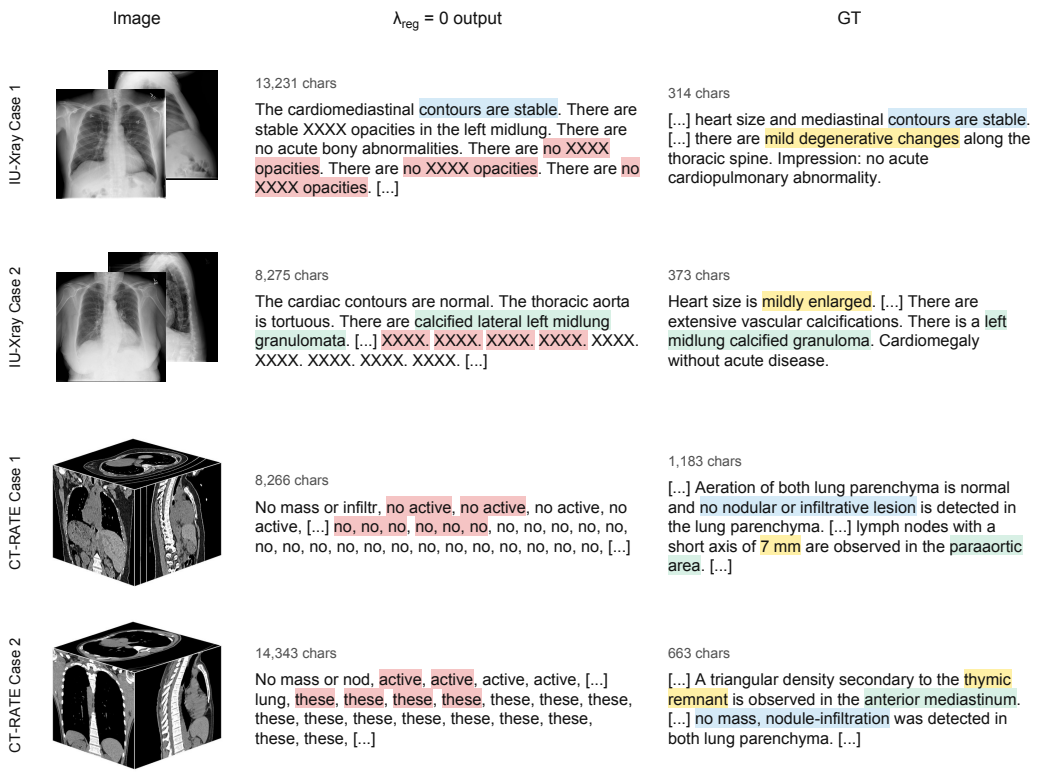

Figure 9: Examples of generation degeneration without reference regularization ($\lambda_{\text{reg}} = 0$). Two IU-Xray cases and two CT-RATE cases are shown, with the input image, unregularized model output, and ground-truth report (GT) from left to right. Character counts illustrate the substantial length inflation of the generated reports. The additional text is largely repetitive or low-information rather than clinically meaningful. Red highlights mark representative degenerate spans in the generated reports, while colored highlights in the GT indicate representative clinical findings for comparison.

OPSD Teacher Prompt

<image>
Here is a reference report for this study:
=== Reference Report Begin ===
{normalized_ground_truth_report}
=== Reference Report End ===

After reading the reference report above, make sure you understand the findings it describes.

Generate a report for this CT/X-ray.

Figure 10: Teacher prompt used for the OPSD baseline.

E CLIN-F1 EVALUATION DETAILS

We provide the detailed definition and computation procedure of **Clin-F1**, our clinical correctness metric for medical report generation. Inspired by the region-of-interest-based evaluation protocol in ClinFusion (Yuan et al., 2026), Clin-F1 assesses whether clinically relevant findings associated with anatomical regions are correctly captured in the generated report. Unlike ClinFusion, which conditions report generation on GT-derived regions of interest, our protocol provides no such privileged

information during generation. The reference and generated reports are compared only at evaluation time to identify matched, missed, and hallucinated clinical findings.

Finding-level evaluation. As illustrated in Figure 11, the evaluator takes the reference report and the generated report as input, extracts clinically relevant findings together with their corresponding anatomical regions, and compares the two reports semantically. Each finding is assigned to one of three categories: *matched*, *missed*, or *hallucinated*. A matched finding is a clinical finding in the generated report that is consistent with a corresponding finding in the reference report. A missed finding is present in the reference report but not captured by the generated report, whereas a hallucinated finding is introduced by the generated report without support from the reference. The matching is based on clinical semantics rather than exact lexical identity, allowing different surface forms that describe the same clinical finding to be matched.

Clin-F1 computation. We treat matched, missed, and hallucinated findings as true positives (TP), false negatives (FN), and false positives (FP), respectively. For each case i, clinical precision is computed as

$$P_i = \begin{cases} \dfrac{TP_i}{TP_i + FP_i}, & TP_i + FP_i > 0, \\ 1, & TP_i = FP_i = FN_i = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

Clinical recall is computed as

$$R_i = \begin{cases} \dfrac{TP_i}{TP_i + FN_i}, & TP_i + FN_i > 0, \\ 1, & TP_i = FP_i = FN_i = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

The case-level Clin-F1 is defined directly from the finding counts as

$$F_i = \begin{cases} \dfrac{2TP_i}{2TP_i + FP_i + FN_i}, & 2TP_i + FP_i + FN_i > 0, \\ 1, & TP_i = FP_i = FN_i = 0. \end{cases} \tag{13}$$

The second case corresponds to a correctly generated normal report, where neither the reference nor the generated report contains an abnormal finding. For all other cases with an undefined precision or recall denominator, the corresponding metric is set to zero.

For an evaluation set containing N cases, we report the macro-average of each metric across cases:

$$\begin{aligned} P_{\text{clin}} &= \frac{1}{N}\sum_{i=1}^{N} P_i, \\ R_{\text{clin}} &= \frac{1}{N}\sum_{i=1}^{N} R_i, \\ \text{Clin-F1} &= \frac{1}{N}\sum_{i=1}^{N} F_i. \end{aligned} \tag{14}$$

This formulation jointly penalizes two clinically important failure modes: omission of findings supported by the reference report and generation of unsupported findings, while assigning a perfect score to a correctly generated normal case. All reported metrics are multiplied by 100 and presented as percentages.

Evaluator implementation. We use Qwen3.5-Plus (Team, 2026) as the LLM evaluator with temperature set to 0. The same evaluation protocol is applied to the predictions of all methods.

By considering both missed and unsupported findings, Clin-F1 provides a more clinically meaningful assessment of generated reports than lexical-overlap metrics. This observation is also consistent with the judgments of two experienced clinicians.

Clin-F1 Evaluation Prompt

You are an expert radiologist acting as a strict evaluator. Compare a model-generated report against a ground truth report, focusing exclusively on abnormal findings.

[GROUND TRUTH REPORT]:

{ground_truth}

[MODEL GENERATED REPORT]:

{prediction}

Categorize abnormal findings using exactly these definitions:

1. [MATCHED_ABNORMALITY]: The ground truth mentions an abnormality and the model correctly identifies the same abnormality. Minor wording differences are acceptable.
2. [MISSED_ABNORMALITY]: The ground truth mentions an abnormality but the model omits it or contradicts it.
3. [HALLUCINATED_ABNORMALITY]: The model mentions an abnormality not present in the ground truth.

Return only the following three headers, in this exact order. Keep empty headers if there are no findings.

[MATCHED_ABNORMALITY]
[MISSED_ABNORMALITY]
[HALLUCINATED_ABNORMALITY]

Figure 11: Prompt used for Clin-F1 evaluation. The evaluator compares the generated report with the reference report and identifies matched, missed, and hallucinated clinical findings.

F Agreement between Clin-F1 and Physician Judgments

To examine whether Clin-F1 better reflects the clinical quality of generated reports than conventional text-overlap metrics, we conduct an independent evaluation with two experienced clinicians. The evaluation is performed in a blinded manner: for each case, the clinicians are shown the reference report and one generated report without access to the identity of the generating method. Reports from both IU X-ray and CT-RATE are included to cover 2D radiography and 3D volumetric CT.

Physician rating protocol. Each generated report is rated independently by two clinicians on a five-point clinical correctness scale. A score of 1 indicates substantial clinical errors or clinically misleading content, whereas a score of 5 indicates that the generated report is clinically consistent with the reference report with no important omission or unsupported finding. Intermediate scores reflect different degrees of clinical correctness. The evaluation focuses on clinical content rather than stylistic or lexical similarity to the reference report.

Let $s_{c,m}^{(r)} \in \{1, \ldots, 5\}$ denote the score assigned by clinician r to the report generated by method m for case c. With two clinicians, the mean opinion score (MOS) is defined as

$$y_{c,m} = \mathrm{MOS}_{c,m} = \frac{1}{2}\left(s_{c,m}^{(1)} + s_{c,m}^{(2)}\right). \tag{15}$$

Case-level metric alignment. We evaluate six automatic metrics: BLEU-1, BLEU-4, ROUGE-L, METEOR, CIDEr, and Clin-F1. Let $x_{c,m}^{(q)}$ denote the score assigned by automatic metric q to method m on case c. For a case containing predictions from M methods, we form

$$\mathbf{x}_c^{(q)} = \left[x_{c,1}^{(q)}, x_{c,2}^{(q)}, \ldots, x_{c,M}^{(q)}\right], \tag{16}$$

and the corresponding physician-score vector

$$\mathbf{y}_c = [y_{c,1}, y_{c,2}, \ldots, y_{c,M}]\,. \tag{17}$$

Importantly, agreement is measured *within each case*: each automatic metric is evaluated according to how well it reproduces the physician-derived ordering of different methods for the same medical image. This avoids directly comparing absolute metric values across different cases or datasets.

Spearman rank correlation. For each case c and metric q, we first convert the automatic metric scores $\mathbf{x}_c^{(q)}$ and physician MOS scores $\mathbf{y}_c$ into rank vectors $\mathbf{r}_c^{(q)} = [r_{c,1}^{(q)}, \ldots, r_{c,M}^{(q)}]$ and $\mathbf{h}_c = [h_{c,1}, \ldots, h_{c,M}]$, respectively. Since higher values indicate better report quality for both the automatic metrics and MOS, methods are ranked in descending order, with rank 1 assigned to the highest score. Tied methods are assigned their average rank. Specifically,

$$r_{c,m}^{(q)} = 1 + \sum_{\substack{j=1 \\ j\neq m}}^{M} \mathbb{I}\left[x_{c,j}^{(q)} > x_{c,m}^{(q)}\right] + \frac{1}{2}\sum_{\substack{j=1 \\ j\neq m}}^{M} \mathbb{I}\left[x_{c,j}^{(q)} = x_{c,m}^{(q)}\right], \tag{18}$$

and

$$h_{c,m} = 1 + \sum_{\substack{j=1 \\ j\neq m}}^{M} \mathbb{I}\left[y_{c,j} > y_{c,m}\right] + \frac{1}{2}\sum_{\substack{j=1 \\ j\neq m}}^{M} \mathbb{I}\left[y_{c,j} = y_{c,m}\right], \tag{19}$$

where $\mathbb{I}[\cdot]$ is the indicator function. The case-level Spearman correlation is then computed as the Pearson correlation between the two rank vectors:

$$\rho_c^{(q)} = \frac{\sum_{m=1}^{M}\left(r_{c,m}^{(q)} - \bar{r}_c^{(q)}\right)\left(h_{c,m} - \bar{h}_c\right)}{\sqrt{\sum_{m=1}^{M}\left(r_{c,m}^{(q)} - \bar{r}_c^{(q)}\right)^2}\sqrt{\sum_{m=1}^{M}\left(h_{c,m} - \bar{h}_c\right)^2}}, \tag{20}$$

where

$$\bar{r}_c^{(q)} = \frac{1}{M}\sum_{m=1}^{M} r_{c,m}^{(q)}, \qquad \bar{h}_c = \frac{1}{M}\sum_{m=1}^{M} h_{c,m}. \tag{21}$$

Finally, we macro-average the case-level correlations over the N evaluated cases:

$$\rho^{(q)} = \frac{1}{N}\sum_{c=1}^{N} \rho_c^{(q)}. \tag{22}$$

Kendall's τ. We further evaluate pairwise ranking consistency using Kendall's τ with tie correction. Unlike Spearman's correlation, which compares the complete rank vectors, Kendall's τ considers the relative ordering of every pair of methods.

For each case c, let

$$\mathcal{P} = \{(i,j) \mid 1 \le i < j \le M\} \tag{23}$$

denote the set of all method pairs, with $|\mathcal{P}| = \binom{M}{2}$. For a pair $(i,j) \in \mathcal{P}$, we define the score differences under automatic metric q and physician MOS as

$$\Delta x_{c,ij}^{(q)} = x_{c,i}^{(q)} - x_{c,j}^{(q)}, \qquad \Delta y_{c,ij} = y_{c,i} - y_{c,j}. \tag{24}$$

A pair is *concordant* if the automatic metric and physician MOS prefer the same method:

$$\Delta x_{c,ij}^{(q)} \Delta y_{c,ij} > 0, \tag{25}$$

and *discordant* if they prefer opposite methods:

$$\Delta x_{c,ij}^{(q)} \Delta y_{c,ij} < 0. \tag{26}$$

Accordingly, the numbers of concordant and discordant pairs are

$$C_c^{(q)} = \sum_{(i,j)\in\mathcal{P}} \mathbb{I}\left[\Delta x_{c,ij}^{(q)} \Delta y_{c,ij} > 0\right], \tag{27}$$

and

$$D_c^{(q)} = \sum_{(i,j)\in\mathcal{P}} \mathbb{I}\left[\Delta x_{c,ij}^{(q)} \Delta y_{c,ij} < 0\right]. \tag{28}$$

Because both automatic metrics and physician MOS may contain tied scores, we explicitly account for method pairs tied in only one of the two rankings. The number of pairs tied according to automatic metric q but not according to physician MOS is

$$T_{x,c}^{(q)} = \sum_{(i,j)\in\mathcal{P}} \mathbb{I}\left[\Delta x_{c,ij}^{(q)} = 0 \ \wedge \ \Delta y_{c,ij} \neq 0\right], \tag{29}$$

while the number tied according to physician MOS but not according to the automatic metric is

$$T_{y,c}^{(q)} = \sum_{(i,j)\in\mathcal{P}} \mathbb{I}\left[\Delta x_{c,ij}^{(q)} \neq 0 \ \wedge \ \Delta y_{c,ij} = 0\right]. \tag{30}$$

Pairs tied under both the automatic metric and physician MOS do not contribute to $C_c^{(q)}$, $D_c^{(q)}$, $T_{x,c}^{(q)}$, or $T_{y,c}^{(q)}$. The tie-corrected case-level Kendall correlation is then computed as

$$\tau_c^{(q)} = \frac{C_c^{(q)} - D_c^{(q)}}{\sqrt{\left(C_c^{(q)} + D_c^{(q)} + T_{x,c}^{(q)}\right)\left(C_c^{(q)} + D_c^{(q)} + T_{y,c}^{(q)}\right)}}. \tag{31}$$

A value of $\tau_c^{(q)} = 1$ indicates that the automatic metric and physician MOS induce fully consistent pairwise orderings for case c, whereas $\tau_c^{(q)} = -1$ indicates completely reversed pairwise orderings. Values around zero indicate weak pairwise association.

Finally, as with Spearman's correlation, we macro-average the case-level Kendall correlations over the N evaluated cases:

$$\tau^{(q)} = \frac{1}{N}\sum_{c=1}^{N} \tau_c^{(q)}. \tag{32}$$

Top-1 accuracy. While Spearman's ρ and Kendall's τ measure agreement over the overall ordering of methods, Top-1 accuracy evaluates whether an automatic metric identifies the report that physicians consider best for each case.

For case c and automatic metric q, we define the set of methods achieving the highest score as

$$\mathcal{S}_c^{(q)} = \underset{m\in\{1,\ldots,M\}}{\arg\max}\ x_{c,m}^{(q)}, \tag{33}$$

and the set of methods achieving the highest physician MOS as

$$\mathcal{S}_c^{\mathrm{MOS}} = \underset{m\in\{1,\ldots,M\}}{\arg\max}\ y_{c,m}. \tag{34}$$

To account for ties, the case-level Top-1 accuracy is defined as

$$A_c^{(q)} = \frac{\left|\mathcal{S}_c^{(q)} \cap \mathcal{S}_c^{\mathrm{MOS}}\right|}{\left|\mathcal{S}_c^{(q)}\right|}. \tag{35}$$

This definition gives a score of 1 when every method selected as top-scoring by the automatic metric is also among the physician-preferred methods. If the automatic metric assigns the highest score to

two methods and only one of them belongs to the physician-preferred set, the case receives a score of 0.5. If there is no overlap between the two sets, the score is 0.

The overall Top-1 accuracy for metric q is obtained by macro-averaging the case-level scores over the N evaluated cases:

$$\text{Top-1 Acc.}^{(q)} = \frac{1}{N}\sum_{c=1}^{N} A_c^{(q)}. \tag{36}$$

Agreement analysis. Table 2 provides a more detailed view of the metric–physician agreement. Among the conventional metrics, METEOR achieves the strongest alignment, with a Spearman correlation of 0.4281, a Kendall correlation of 0.3810, and a Top-1 accuracy of 61.83%. Clin-F1 further improves these values to 0.5576, 0.5190, and 71.13%, corresponding to gains of 0.1295, 0.1380, and 9.30 percentage points, respectively. The consistent improvement indicates that the advantage of Clin-F1 is not specific to a particular agreement measure.

G Controlled Intervention Analysis of Clinical Hints

Motivation. The report comparisons in the main paper show that clinical hints can recover findings that are omitted under image-only generation. To isolate the effect of the hints from changes in the autoregressive decoding trajectory, we further conduct a controlled teacher-forced analysis. For each case, we keep the input image, model checkpoint, ground-truth report, and preceding ground-truth tokens fixed, and change only the organs included in the clinical hints. The resulting effects are summarized in Figure 12.

Teacher-forced finding scoring. We follow the notation in Section 3.1. Given a medical image x, its ground-truth report $y = (y_1, \ldots, y_{|y|})$, and clinical hint $h = \{o_1, \ldots, o_m\}$, let $h^{(v)}$ denote the organ set used under intervention v. At decoding step t, the model is conditioned on the ground-truth prefix $y_{<t}$. For a hint-conditioned intervention, the log-probability of the ground-truth token is

$$\ell_t^{(v)} = \log \pi_\theta \left(y_t \mid x, h^{(v)}, y_{<t} \right). \tag{37}$$

For the image-only condition,

$$\ell_t^{(\text{plain})} = \log \pi_\theta \left(y_t \mid x, y_{<t} \right). \tag{38}$$

For each case, we select a positive finding from the ground-truth report and map its complete text span to a token index set $\mathcal{I}$. We define the finding score as the mean token log-probability:

$$\bar{\ell}(v) = \frac{1}{|\mathcal{I}|}\sum_{t\in\mathcal{I}} \ell_t^{(v)}. \tag{39}$$

A larger $\bar{\ell}(v)$ indicates that the model assigns higher probability to the ground-truth finding under intervention v, whereas a smaller value indicates weaker support for that finding.

For interpretation, the corresponding geometric-mean token probability is

$$G(v) = \exp\big(\bar{\ell}(v)\big). \tag{40}$$

Accordingly, a larger $G(v)$ means that the tokens in the ground-truth finding receive higher conditional probability on average.

Controlled hint interventions. Let $h_{\text{tar}} \subseteq h$ denote the target organ or target-organ group associated with the finding being analyzed. We construct six conditions to support four complementary comparisons: *plain*, *full hint*, *target only*, *target replaced*, *unrelated only*, and *unrelated full*. Their concrete organ lists for the three cases are given in Tables 4–6.

The conditions are grouped according to the effect being measured. *Plain* and *full hint* capture the overall change after adding the complete clinical hint. To determine whether this change depends on the target-organ information, *target replaced* keeps the remaining clinical hints unchanged but substitutes the target organ with unrelated organs. To examine the target organ in isolation, *target*

Table 4: Controlled hint interventions for the hepatosteatosis case. The target organ is the liver.

Intervention	Hint organs
Full hint	liver, lung
Target only	liver
Target replaced	bladder, lung
Unrelated only	bladder
Unrelated full	bladder, heart

Table 5: Controlled hint interventions for the bilateral pleural effusion case. The target organ is the pleura.

Intervention	Hint organs
Full hint	pericardium, pleura, lung
Target only	pleura
Target replaced	pericardium, aorta, lung
Unrelated only	aorta
Unrelated full	aorta, esophagus, pancreas

only is compared with an *unrelated only* condition of matched length. Finally, *unrelated full* replaces the complete clinical hint with an equally long list of unrelated organs, providing a control for the presence of an organ list itself.

For the three controlled comparisons, prompt length is matched within each pair or group: *full hint*, *target replaced*, and *unrelated full* have the same final multimodal prompt length, while *target only* and *unrelated only* are also length matched. The *plain–full hint* comparison is not length matched, since adding the complete hint is the intervention whose total effect is being measured.

Intervention effects. Using Equation 39, we compute four effects. The *total effect* is

$$\mathrm{TE} = \bar{\ell}(\mathrm{full}) - \bar{\ell}(\mathrm{plain}), \tag{41}$$

which measures the overall change produced by adding the complete hint.

The *necessity* effect is

$$\mathrm{N} = \bar{\ell}(\mathrm{full}) - \bar{\ell}(\text{target-replaced}), \tag{42}$$

which measures the contribution of the correct target organ within the full hint relative to a length-matched replacement.

The *sufficiency* effect is

$$\mathrm{S} = \bar{\ell}(\text{target-only}) - \bar{\ell}(\text{unrelated-only}), \tag{43}$$

which compares the target organ alone against a length-matched unrelated organ control.

The *specificity* effect is

$$\mathrm{SP} = \bar{\ell}(\mathrm{full}) - \bar{\ell}(\text{unrelated-full}), \tag{44}$$

which compares the complete clinical hint against a fully unrelated, length-matched organ list.

Results. Figure 12 shows a consistent positive effect of clinically relevant organ hints across all three cases, with the magnitude increasing from hepatosteatosis to bilateral pleural effusion and further to diffuse bone-density reduction. This ordering holds for all four measurements: TE increases from 0.052 to 0.453 and 0.619, while the corresponding necessity, sufficiency, and specificity effects increase from $0.209/0.134/0.252$ to $0.477/0.406/0.574$ and finally to $0.906/0.741/0.995$. Two additional patterns are consistent across the cases. First, specificity is the largest controlled effect in every case, showing that the complete clinically relevant hint receives substantially higher support than an equally long list of unrelated organs. Second, necessity is consistently larger than sufficiency, suggesting that the target organ has its strongest effect when placed in the full clinical context, although the target alone still provides a clear gain over its length-matched unrelated control. The

Table 6: Controlled hint interventions for the diffuse bone-density reduction case. The target group contains bone and vertebra.

Intervention	Hint organs
Full hint	gallbladder, bone, lung, vertebra, pleura, mediastinum
Target only	bone, vertebra
Target replaced	gallbladder, bladder, lung, adrenal gland, pleura, mediastinum
Unrelated only	bladder, adrenal gland
Unrelated full	adrenal gland, aorta, bladder, esophagus, pancreas, rectum



Figure 12: Controlled hint intervention effects for three representative findings. Each cell reports the change in mean token log-probability under the total effect, necessity, sufficiency, and specificity comparisons. Positive values indicate greater support for the target finding from clinically relevant organ information.

bone-density case shows the largest separation: its necessity effect of 0.906 corresponds to a $2.48\times$ increase in geometric-mean token probability, while its specificity effect of 0.995 corresponds to approximately $2.70\times$. These teacher-forced changes are also consistent with the free-generation examples: without hints, the model omits hepatosteatosis, denies pleural effusion, and describes the bone structures as normal, whereas the corresponding findings appear after the relevant organ hints are provided. Together, the positive N, S, and SP values under length-matched controls show that the observed changes are not explained by prompt length alone, but depend on which anatomical regions are supplied in the hint.